DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

DOCTORAL THESIS

Computational models of coherence for open-domain dialogue

Author:
Alessandra CERVONE

Supervisor:
Prof. Giuseppe RICCARDI

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

Computer Science
September 30, 2020

"Any road followed precisely to its end leads precisely nowhere."

Frank Herbert

Abstract

Coherence is the quality that gives a text its conceptual unity, making a text a coordinated set of connected parts rather than a random group of sentences (turns, in the case of dialogue). Hence, coherence is an integral property of human communication, necessary for a meaningful discourse both in text and dialogue. As such, coherence can be regarded as a requirement for conversational agents, i.e. machines designed to converse with humans. Though recently there has been a proliferation in the usage and popularity of conversational agents, dialogue coherence is still a relatively neglected area of research, and coherence across multiple turns of a dialogue remains an open challenge for current conversational AI research. As conversational agents progress from being able to handle a single application domain to multiple ones through any domain (open-domain), the range of possible dialogue paths increases, and thus the problem of maintaining multi-turn coherence becomes especially critical.

In this thesis, we investigate two aspects of coherence in dialogue and how they can be used to design modules for an open-domain coherent conversational agent. In particular, our approach focuses on modeling *intentional* and *thematic* information patterns of distribution as proxies for a coherent discourse in open-domain dialogue. While for modeling intentional information we employ Dialogue Acts (DA) theory (Bunt, 2009); for modeling thematic information we rely on open-domain entities (Barzilay and Lapata, 2008). We find that DAs and entities play a fundamental role in modelling dialogue coherence both independently and jointly, and that they can be used to model different components of an open-domain conversational agent architecture, such as Spoken Language Understanding, Dialogue Management, Natural Language Generation, and open-domain dialogue evaluation.

The main contributions of this thesis are: (I) we present an open-domain modular conversational agent architecture based on entity and DA structures designed for coherence and engagement; (II) we propose a methodology for training an open-domain DA tagger compliant with the ISO 24617-2 standard (Bunt et al., 2012) combining multiple resources; (III) we propose different models, and a corpus, for predicting open-domain dialogue coherence using DA and entity information trained with weakly supervised techniques, first at the conversation level and then at the turn level; (IV) we present supervised approaches for automatic evaluation of open-domain conversation exploiting DA and entity information, both at the conversation level and at the turn level; (V) we present experiments with Natural Language Generation models that generate text from Meaning Representation structures composed of DAs and slots for an open-domain setting.

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Giuseppe Riccardi, for always making sure I was asking the right questions, for convincing me of his ideas and let me convince him of mines, and for believing that being different is an asset, rather than a disadvantage.

My journey would not have been the same without Evgeny's unapologetic opinions, support and suggestions. I also want to thank my companions in the Alexa Prize adventure, Giuliano, Stefano and Enrico, for being together in our project's ambition, for the long days, the ragù and the deep discussions about the complicated relation between Spanish and Italian ham.

Daniele has been there for me always. This thesis and my whole PhD experience were shaped by our conversations, his enthusiasm about research, his unwavering encouragement even when I did not believe in an idea myself, and, last but not least, his role in making me find another home in Trentino.

I also would like to express my gratitude to all my coauthors during this strange experience, the visible ones and the invisible ones. Those who spent long nights with me before a deadline going through our work one more time, checking every detail was correct, and those that gave an invaluable suggestion during a casual conversation, or simply asked the right question because they were interested.

I will miss Sislab and its constellation of different personalities, backgrounds and attitudes. Among my labmates, I am grateful to Anu for his kindness, and Mahed, for our long circular walks with the excuse of a coffee and for being a real friend.

There were quite a few people that I loved randomly (and not so randomly) walking into in the Department in Povo during these years. Some of them were the hyenas, especially Paolo and Gianni, Andrea, Duygu, Irina, Mattia, and the other two musketeers. I am also thankful to those who had an impact on my PhD from a distance. Tanvi, thank you for our articulated conversations and for your passion.

Additionally, I am grateful to have met many wonderful people during the Alexa Prize competition and during my eventful internship in California, especially Sandra.

Besides Mahed, I would like to thank Adèle and Gabriel for giving me the privilege of co-supervising them.

I also would like to thank my reviewers. Your insightful comments and suggestions made this work better.

Moreover, I want to express my gratitude to Andrea Stenico, for always fighting to be helpful.

Finally, I am grateful to my family, my mother Silvia and my father Marco for being under the impression that I have a mind of my own and that this mind should be trusted, even though some decisions did not seem like the clear path at the time, Ilaria and Emanuele for making fun of me, and for reminding me the usefulness of useless things. And I am thankful to my new family, Maria Pia, Nicola, Chiara, for the long walks and the breath of nature that were our sweetest breaks during this journey.

Contents

List of Figures

List of Tables

List of Abbreviations

ASR	**A**utomatic **S**peech **R**ecognition
CA	**C**onversational **A**gent
DA	**D**ialogue **A**ct
DM	**D**ialogue **M**anager
FU	**F**unctional **U**nit
MR	**M**eaning **R**epresentation
NLG	**N**atural **L**anguage **G**eneration
SDS	**S**poken **D**ialogue **S**ystem
SLU	**S**poken **L**anguage **U**nderstanding
SOTA	**S**tate **O**f **T**he **A**rt
SWBD-Coh	**SW**itch**B**oar**D**- **Coh**erence corpus
SWBD-DA	**SW**itch**B**oar**D**- **D**ialogue **A**ct corpus
TTS	**T**ext **T**o **S**peech

Chapter 1

Introduction

1.1 Motivation

> **Example 1**
> “Mad Hatter: ‘Why is a raven like a writing-desk?’
> (...)
> ‘Have you guessed the riddle yet?’ the Hatter said, turning to Alice again.
> ‘No, I give it up,’ Alice replied: ‘What’s the answer?’
> ‘I haven’t the slightest idea,’ said the Hatter.
> ‘Nor I,’ said the March Hare.
> Alice sighed wearily. ‘I think you might do something better with the time,’ she said, ‘than waste it in asking riddles that have no answers.’ ”
> Lewis Carroll, *Alice’s Adventures in Wonderland*

Incoherent responses in a dialogue, such as the ones given by the Mad Hatter who asks riddles which do not have answers, can easily make interlocutors weary and give them the impression that the interaction is a waste of time, just as it happens to Alice in the example. Discourse *coherence* can thus be regarded as a requirement of conversational interaction, insofar as when this requirement is disappointed the whole trust of participants in the conversation might break down.

Sentences in a body of text are not just collections of unrelated items, but are usually organised according to a structure of relations connecting sentences with one another. This property, i.e. what makes a body of text a discourse, rather than a random group of sentences, is known as coherence. What this structure consists of, however, and how to define it is less agreed upon.

The concept of coherence plays a cardinal role in several theoretical approaches to discourse and pragmatics (De Beaugrande and Dressler, 1981a; Conte, 1980; Halliday and Hasan, 1976), i.e. the fields of Linguistics which analyse language in context[1]. While in various theories coherence is considered the *condicio sine qua non* of discourse, there is a lack of consensus regarding the formalization of this concept and which factors contribute to it (Taboada, 2004; Bublitz, 2011). Some theories, for example, ascribe a major role to the interpretation of speakers’ intentions (Grice, 1970; Sperber and Wilson, 1986); others concentrate more on the importance of the logical relations connecting different parts of a text (Fillmore, 1998; Hobbs, 1979); others emphasize the role of the thematic and informational structure (Hopper, 1979;

[1] The difference between the fields of study of pragmatics and discourse has been a matter of discussion among linguists. According to Horn and Kecskés (2013), while pragmatics concentrates on studying individual utterances in context, discourse analyzes organized groups of utterances. In general, both fields can be considered synergistic.

Givón, 1983); while some others point out the importance of the temporal structure (Lascarides and Asher, 1993) or of establishing a common ground (Clark and Schaefer, 1989).

Notwithstanding their heterogeneity, theoretical approaches to discourse thus share the general assumption that coherence is a tenet of human communication (Gruber and Redeker, 2014).
Given its pivotal role, we would thus expect the ability to entertain a coherent discourse to be an essential feature for a system aimed at interacting with humans. Nonetheless, producing coherent responses across multiple turns of a dialogue remains an open research problem for State-of-the-art (SOTA) *conversational agents* (CA), i.e. computational models able to converse with humans (also known as *dialogue systems*, or *intelligent virtual assistants*). Although the last few years have seen a steep surge in popularity of CAs among users[2], the task of building intelligent dialogue systems seems still far from being solved (Goode, 2018).

From a research perspective, the area of conversational Artificial Intelligence (AI) is currently divided between building *task-oriented* CAs, aimed at performing a limited selection of tasks such as reserving a restaurant or booking a flight, and *non-task-oriented* CAs, mainly able to perform chitchat, both with their drawbacks and without a real integration between the two.
On one hand, task-oriented CAs address the coherence problem by restricting their application domain, for example to restaurant reservations or movie tickets booking, and handcrafting the system's possible states and actions (Williams and Young, 2007; Wen et al., 2016a). These approaches typically rely on a pipeline of different modules, which usually includes: a Spoken Language Understanding (SLU) component, which assigns a semantic representation to the user's utterances; a Dialogue Manager (DM), in charge of selecting the next action of the system; and a Natural Language Generation (NLG) module, which transforms the selected action into a sentence. This type of architectures usually relies on a Meaning Representation (MR) language based on *slots*, the types of concepts relevant for the task (e.g. `restaurant_price`), and *intents*, the possible speakers' intentions behind a given utterance relevant for the task (e.g. `request_price`). Although slots and intents-based MRs proved very successful as building units for representing task-oriented interactions[3], there also evident limitations in their application. Crucially, both slots and intents are usually closed sets handcrafted for a specific application by the system designer, by making assumptions regarding the overall end goal of users employing that given application (De Mori et al., 2008). While research on task-oriented CAs is fairly well established, the handcrafting required in their design brings issues of scalability to novel domains.
On the other hand, non-task-oriented approaches, mostly based on black-box Deep Neural Networks (DNN) models trained on large dialogue datasets (Serban et al., 2016d), seem a more promising direction given the advantage that they do not require hand-crafted design and can thus be considered *open-domain*, i.e. not dependant on a given domain. However these models are difficult to control and tend to give incoherent, non committal responses that break user engagement over a few

[2] The market value of virtual assistants is projected to be around US dollars 11.3 Billion in 2024 according to recent reports (Group, 2019).

[3] The slots and intents MR framework is currently at the basis of most libraries for developing task-oriented CAs, for example the Alexa Skills Kit (Kumar et al., 2017), Dialog Flow (`https://cloud.google.com/dialogflow/docs/basics`) or Rasa (Bocklisch et al., 2017).

turns (Li et al., 2016a), which makes them difficult to apply for commercial applications.

Moreover, progresses in the field are curbed by a lack of standardized metrics to evaluate dialogue systems. Current automatic metrics for dialogue evaluation are not able to report an analysis of the characteristics of good conversations: some rely on surface features such as the words used (Papineni et al., 2002), others work only for task-oriented CAs (Walker et al., 1997). For evaluation, the field still relies heavily on user satisfaction, an expensive and time–consuming process which poses its own challenges given the subjectivity of human judgement.

In general, it seems we have still much to understand regarding the characteristics of successful conversations regardless of the task, that is the underlying structures of dialogue that CAs should learn in order to ensure successful interactions with humans. This gap in our understanding of multi-turn dialogue structure, which factors contribute to it and how they interact constitutes a significant bottleneck for progress in the field of conversational AI. Without clearer insights on which aspects affect our perception of a good conversation it is difficult to assess and compare the performance of different models across various tasks and domains. Vice versa, designing effective CAs not bound to a specific domain without an understanding of what might make interlocutors perception of the conversation quality break down becomes very challenging.

In this thesis, we explore computational models of coherence in open-domain dialogue. Naturally, a comprehensive approach to coherence modeling in open-domain conversation would be too ambitious to address in a single dissertation. Instead, in this thesis we select two aspects studied by previous work in connection to dialogue coherence and investigate how they can be used to learn models of coherence in dialogue and design different components of an open-domain coherent CA.

1.2 Research objectives

Coherence is a multifaceted property which has been studied under several different perspectives in the literature, both in linguistic and computational approaches to discourse.

Linguistic approaches to coherence From a linguistic perspective, we mentioned how coherence is regarded as an integral principle of meaningful text by several theoretical approaches in discourse and pragmatics (Conte, 1980; De Beaugrande and Dressler, 1981b), accounting for the very structure of discourse. This foundational role is given by the fact that coherence is essentially an "intrinsic" property of text, i.e. rather than being a property given by the text in itself, it is a property which comes from the interpretation of the receivers of the text (Andorno, 2003).
Indeed, many of the most recent theoretical approaches in pragmatics and discourse revolve around the central role of the *intentions* of speakers in creating and interpreting a text to describe its global sense (Andorno, 2003), rather than on the text itself. It is not surprising, then, that speakers' intentions behind a given utterance are at the center of some of the most influential theoretical approaches to discourse, especially the ones focusing on dialogic interaction (Grice, 1970; Sperber and Wilson, 1986). In these approaches the coherence of a conversation is thus given not by what is explicitly said, but rather by the underlying intentions of interlocutors (Levinson, 1983), categorized in some approaches as speech acts (Austin and Urmson, 1962; Searle,

1965) or in others as conversational actions (Schegloff, 1968; Schegloff and Sacks, 1973).

The intuition that the way we shape a given text is crucially influenced by how we think our text will be received and interpreted is also at the basis of another fertile research area in discourse: the one of *information structure* (Halliday, 1967). This research area is based on the idea that speakers distribute information in a utterance, such as given versus novel information, according to the perceived mental model of their interlocutors. Although this idea accounts for a wide range of discourse phenomena, by far the most studied and pervasive phenomena explored in connection to information structure are those involving lexical chains (Lambrecht, 1994) and how *entities* (typically noun phrases) are introduced and referred to in a coherent discourse (Halliday and Hasan, 1976). These studies lead to identify important dichotomies, such as the one between the theme (or topic) of a given sentence, that is the information given for granted in the mental model of the receiver, and the rheme (or focus), that is the information considered novel for the receiver.

Computational approaches to coherence From a computational perspective, coherence in dialogue has also been repeatedly attributed to speaker intentions already in early approaches (Cohen and Perrault, 1979; Allen and Perrault, 1980). This *intentional* (Moore and Pollack, 1992) approach to coherence, inspired by speech acts theory (Austin, 1975; Searle, 1965), was mainly developed for task-oriented dialogues. Following this research direction, Grosz and Sidner (1986) developed one the most comprehensive and influential formal theories of the structure of coherent discourse in dialogue. Grosz and Sidner (1986)'s approach defines discourse structure in conversation as composed by three different interacting levels: the linguistics level (i.e. the level of the text), the intentional level, given by speakers purposes and accounting for the *global* structure of the discourse, and the attentional level, defined by the entities currently in focus in a given part of the conversation and thus accounting for the *local* coherence of the discourse. In this approach, the intentional level interacts with the attentional state by modifying the information state (the mental model) of the other participants in the conversation. Although very successful from an academic perspective, this research direction was so rich in its annotation schema that proved rather difficult to apply to large scale scenarios.

Focusing only on the attentional level, Grosz and Sidner (1986)'s approach was later developed into Centering theory (Grosz, Weinstein, and Joshi, 1995), an *entity-based* computational theory of local coherence of text. This theory formulates a series of constraints for anaphora resolution and other discourse phenomena, based on the intuition that in a locally coherent text salient entities are more likely to appear across subsequent sentences in prominent syntactic positions (i.e. subject) and thus become referents. Centering theory proved very influential in the Natural Language Processing community, inspiring several approaches and applications (Walker, Joshi, and Prince, 1998). However, Centering theory rather than being data-driven, relies on a specific theory of discourse. Inspired by Grosz, Weinstein, and Joshi (1995), Barzilay and Lapata (2008) propose the successful entity grid approach, which implements some of Centering intuitions using a data-driven approach without being bound to a specific theory. Another important advantage of this approach is the fact that it does not rely on annotated data. Indeed, the coherence tasks proposed within this framework (Barzilay and Lapata, 2008; Elsner and Charniak, 2011b) employ automatic data generation methodologies where negative samples are created by disrupting the order of texts assumed to be coherent. Although mainly explored for written text (news, summary), this framework has found several extensions (Filippova and Strube, 2007; Guinaudeau and Strube, 2013; Li and Hovy, 2014; Moon et al., 2019) and widespread applications (Li et al., 2017; Farag, Yannakoudakis, and Briscoe, 2018; Clark, Ji, and Smith, 2018).

In this dissertation, we are interested in modelling coherence in open-domain dialogue. As we have seen, coherence is a property which has been approached with a plethora of different perspectives across multiple research fields. Nevertheless, in our opinion across the literature there are some recognizable patterns, some aspects which have been repeatedly associated to dialogic coherence by different scholars in various approaches. In our view, these patterns repeatedly point to two phenomena as particularly crucial for modelling conversational coherence: the intentions of speakers behind given utterances, which we call the *intentional* aspect of coherence, and the patterns of distribution of entities across subsequent sentences, which we refer to in this thesis as the *thematic*[4] aspect of coherence.
More specifically, in our approach to capture the thematic aspect we rely on entity-based models following previous approaches (Barzilay and Lapata, 2008), while for the intentional structure we rely on Dialogue Acts (DA) theory (Bunt, 1999; Bunt, 2009), a more generalised version of intents which proved useful for dialogue systems research. As recognised in previous work (Grosz and Sidner, 1986) these two aspects are closely related, since DAs can be thought of as serving the function of updating information in the interlocutors mental states (Huang, 2017, Chapter19).

Hence, the main hypotheses we explore throughout this dissertation can be formulated as:

H1: *Can we model coherence in dialogue using Dialogue Acts and entities?*
H2: *Can we use Dialogue Acts and entities as units to build models for an open-domain coherent conversational agent?*

H1 postulates that coherent dialogue is characterized by patterns of distribution of DAs and entities both as independent and joint signals. Moreover, a corollary of H1 is that modelling DAs and entities improve the performance of models aimed at predicting dialogue coherence. Assuming these patterns do exist, H2 further suggests that it is possible to design models based on such DA and entity patterns of distribution in order to create different components of a coherent CA.
While the first research question has a more empirical nature, the second research question is more application oriented. Interestingly, in our work the results of one hypothesis feeds into the other one and vice versa.

Throughout the dissertation, our main hypotheses are further declined into specific sub-hypotheses, as shown in Table 1.1. In particular, for H1 we explore two types of approaches, namely weakly supervised and fully supervised techniques. Additionally, in both cases, coherence prediction is explored at two levels: the macro level of the whole conversation and the micro level of single turns. On the other hand, H2 is also declined into the different components and tasks of a possible open-domain

[4] In this thesis, we decided to refer to this phenomenon as the "thematic" level following the notion of theme in Halliday and Hasan (1976). This choice was made to avoid confusion with other possible terms to indicate this aspect, such as "information structure" whose terminology might overlap for example with "informational approaches" to coherence, referring to approaches based on coherence relations following Moore and Pollack (1992).

coherent CA, that is Spoken Language Understanding, Dialogue Management, Natural Language Generation and evaluation. In some cases this double perspective, empirical and application-oriented, is used within the same approach, as shown in Table 1.1.

1.3 Research challenges

Throughout this thesis' work we faced several challenges involved in the topic under investigation. The main challenges we encountered are:

1. **Difficulty of open-domain dialogue evaluation**: Investigating coherence in open-domain dialogue leads inextricably to the issue of evaluation. However, open-domain dialogue is difficult to evaluate even for humans and can be very subjective, since people might have diverse expectations from a conversation and hold it to different standards (as indicated, for example, by the results of our annotation experiments in 6.1.7).

2. **Sparsity of related theoretical background**: As we have seen in sections 1.1 and 1.2, theoretical approaches to coherence offers great insights, but no definite solutions. Compared to other Linguistics fields, such as syntax, the field of pragmatics and discourse are not consistent, but are rather fragmented into several theories not necessarily compatible with each other. Additionally, approaches to coherence are also fragmented across different research areas, such as psychology, philosophy, sociology and of course natural language processing, and the lines identifying single fields from each other may be quite blurry (some computational approaches could easily be described as making strong contributions to the field of theoretical Linguistics). Moreover, even within a single field and area of interest, approaches may vary widely. For example, in the case of DAs, we have several different DA schemas which might not be consistent with each other (sections 2.3.2 and 4.2.1 present a variety of DA schemas).

3. **Data problem**: This thesis mainly focuses on data-driven approaches, thus the availability of relevant data for the different tasks we explore is crucial. However, in general there is a lack of large good quality corpora for open-domain dialogue, especially datasets with coherence annotations. Hence, throughout the dissertation we investigate different methodologies to overcome this data issue. In Chapter 3, for example, we explore an architecture not fully trained though designed to be trainable; in Chapter 4 we create a dataset combining multiple available resources, while whole Chapter 5 is dedicated to weakly supervised methodologies for training which do not rely on annotated data. Additionally, in the same Chapter we also create a novel publicly available resource with turn coherence ratings for open-domain dialogue. Finally, in Chapter 6, given the lack of annotated resources at the time, we explore different signals as proxies for coherence in dialogue.

1.4 Contributions

Table 1.1 shows the thesis structure in terms of contributions to the main research hypotheses for each chapter.

	Chapters				
	Ch. 3	**Ch. 4**	**Ch. 5**	**Ch. 6**	**Ch. 7**
H1: *Can we model coherence in dialogue using DAs and entities–based approaches?*					
H1.1 *Weakly supervised approaches*					
H1.1.1 Conversation level			X		
H1.1.2 Turn level			X		
H1.2 *Supervised approaches*					
H1.1.1 Conversation level				X	
H1.1.2 Turn level				X	
H2: *Can we use DAs and entities to create models for a coherent open-domain CA?*					
H2.1 Spoken Language Understanding (SLU)	X	X			
H2.2 Dialogue Manager (DM)	X		X		
H2.3 Natural Language Generation (NLG)					X
H2.4 Evaluation			X	X	

TABLE 1.1: Structure of the dissertation in terms of main contributions to the thesis hypotheses across different chapters.

The empirical contributions of this dissertation in regards to H1 point to the *crucial importance of both DA and entities, especially when combined, for predicting open-domain dialogue coherence and evaluation*.
In particular, our empirical contributions towards H1 include:

- Our experiments on standard coherence tasks across different (single and open-domain) spoken dialogue corpora indicate the crucial role of DA information both independently and in combination with entities information for conversation-level coherence prediction (in Chapter 5, section 5.1.5).
- A statistical analysis of a dialogue corpus annotated with turn coherence ratings indicates the importance of both DA and entities information for turn coherence perception in open-domain conversation (in Chapter 5, section 5.2.5).
- Results across traditional and neural ML models for predicting human turn coherence ratings point to the essential role of entities, DAs and especially their combinations for turn coherence ranking in open-domain conversation (in Chapter 5, section 5.2.8).
- DAs and topic representations are found useful for predicting user ratings for entire conversations using a supervised approach (in Chapter 6, section 6.1.9). Specific DAs are also found to be correlated with conversational user ratings (in Chapter 6, section 6.1.5).
- Both DAs and entities information are also found to be relevant for predicting turn level coherence and engagement using a supervised approach (in Chapter 6, section 6.2.8).

Other empirical contributions of this thesis, indirectly related to the main hypotheses, include:

- Our experiments during the Alexa Prize competition indicate that system-driven initiatives lead to higher user ratings in chitchat open-domain conversation (in Chapter 3, section 3.4.1).

- We show the importance of combining multiple publicly available corpora to achieve better performances in open-domain DA tagging through a corpus ablation study (in Chapter 4, section 4.5.2).
- An annotation experiment performed on chitchat conversations between Alexa users and a CA indicates the difficulty of predicting user ratings of open-domain non-task-oriented human-machine conversations for human experts (in Chapter 6, section 6.1.7).

On the other hand, the contributions of this dissertation to H2 are organized according to different components of a possible modular CA pipeline. The contributions of the thesis towards H2 include:

- **General**
  - We propose Roving Mind, an entire modular architecture for open-domain dialogue designed for coherence and engagement relying on Functional Units structures, composed by DAs and entities (in Chapter 3).
  - We present a methodology to use a Commonsense Knowledge Base to create engaging responses in open-domain conversation (in Chapter 3, section 3.3.2).
- **Spoken Language Understanding**
  - We propose a SLU module for open-domain conversation which parses an utterance into a list of Functional Units, composed by Dialogue Acts and open-domain entities (in Chapter 3, section 3.3.3).
  - We propose a methodology to map several available corpora for training an ISO-standard compliant open-domain DA tagger (in Chapter 4).
  - We present a simple yet efficient DA tagging model, whose performance we assess first on the Switchboard Dialogue Act Corpus (with SOTA results compared to models published at the time) using the DAMSL scheme and then on three out-of-domain corpora using the ISO standard scheme (in Chapter 4).
- **Dialogue Management**
  - We propose a novel type of sequential Dialogue Management architecture designed for coherence and engagement in open-domain conversation based on DAs and entities structures, which relies on different submodules each expleting a different conversational function (in Chapter 3, section 3.3.4).
  - We present different ranking strategies based on conversational functions to ensure the selection of the best response in DM (in Chapter 3, section 3.3.4).
  - We propose models based on DAs and entities information trained in a weakly supervised fashion (response selection) which could be used for ranking possible responses in the DM (in Chapter 5, section 5.2).
- **Natural Language Generation**
  - We apply the MR-to-text framework (typical of NLG for task-oriented dialogue) to an open-domain QA application (in Chapter 7).

  - We explore the importance of adding the previous conversational context to improve the quality of the generated output (in Chapter 7).
  - We investigate the possibility of learning NLG models using a MR-to-text approach with increasingly larger ontologies in terms of slot types (in Chapter 7).
  - We experiment with multi-task learning for NLG between open-domain QA and task-oriented dialogue (in Chapter 7).
  - We also propose new evaluation metrics for open-domain NLG to capture the variability of output in open-domain QA compared to NLG for task–oriented dialogue (in Chapter 7).

- **Evaluation**
  - We explore models based on entities and DAs and their combinations for ranking whole open-domain conversations according to their coherence on weakly supervised standard coherence tasks (in Chapter 5, section 5.1).
  - We investigate different models relying on entities and DAs information for ranking turns of open-domain conversations according to their coherence trained on weakly supervised tasks (response selection) and further tested on a corpus annotated with human coherence ratings (in Chapter 5, section 5.2).
  - We propose different entities and DAs information representations which can be used as input for various machine learning models (in Chapter 5).
  - We investigate supervised models for predicting user ratings for open-domain conversations combining intentional (DA) and thematic (LDA) features using real-world conversations between users and an open-domain CA (in Chapter 6, section 6.1).
  - We propose supervised models for predicting turn level coherence and engagement based on features combinations including entities and DA trained on a large annotated corpus of chitchat conversations (in Chapter 6).

Our contributions in terms of publicly available code are:

- Resource to train ISO standard compliant DA tagger and map different corpora to the ISO standard [5].
- Resource to train coherence models for dialogue at the conversation-level [6].

The contributions of this thesis in terms of corpora are:

- Corpus of dialogues annotated with DAs generated by combining publicly available resources which we mapped to a subset of the DA ISO standard (described in Chapter 4) [7].

[5] Available at https://github.com/ColingPaper2018/DialogueAct-Tagger

[6] Available at https://github.com/alecervi/Coherence-models-for-dialogue

[7] Generated using https://github.com/ColingPaper2018/DialogueAct-Tagger

- The Switchboard Coherence (SWBD-Coh) corpus, where 1000 source dialogues from the Switchboard Dialogue Acts corpus are annotated with coherence ratings at the turn level using Amazon Mechanical Turk (described in Chapter 5)[8].

1.5 Thesis outline

In Chapter 2 we review the background literature relevant for this dissertation. First, we describe current approaches in conversational AI, from modular task-oriented architectures to non-task-oriented models, and their evaluation. Afterwards, we offer a perspective on *linguistic* theories about coherence. Finally, we present *computational* approaches to coherence modeling which have been a source of inspiration for this dissertation, with a particular focus on approaches based on entities and DAs.

In Chapter 3, based on Cervone et al. (2017), we present Roving Mind, the open-domain conversational agent we built for the first edition of the Amazon Alexa Prize competition. In the competition participants were challenged to create a CA able to talk to random users about popular topics (such as sports, politics etc.) in a coherent and engaging manner. We describe the main components of our proposed architecture based on DAs and entities structures and designed for coherence and engagement. Additionally, we present experiments performed during the semifinals phase which point to the influence of system-driven strategies on user ratings.

In Chapter 4, based on Mezza et al. (2018), we present a methodology to train a DA tagger for CAs compliant with the ISO standard (Bunt et al., 2010), the latest internationally accepted standard for DAs. To address the reduced number of available resources, we propose to map publicly available corpora to a subset of the standard. We find that in order to train a DA tagger able to be robust for all types of DAs from both task-oriented and non-task-oriented conversation, it is crucial to use a combination of multiple corpora for training.

Chapter 5 is dedicated to *weakly supervised* approaches to learning coherence models for open-domain dialogue based on entities and DAs information. The first part of the chapter, based on Cervone, Stepanov, and Riccardi (2018), explores models combining DAs and entities for standard weakly supervised coherence tasks at the *conversation level*. We find that DAs play a crucial role for dialogue coherence, especially when combined with entities information. In the second part of the chapter, based on Cervone and Riccardi (2020), we focus on modelling coherence at the *turn level*. First, we collect the Switchboard Coherence corpus, a resource annotated with turn level coherence ratings, in order to investigate human perception of turn coherence in correlation with DAs and entities patterns of distributions. Our statistical analysis of the corpus indicate that DAs and entities indeed correlate with turn coherence perception both independently and jointly. Then, we present models based on DAs and entities information trained using a weakly supervised methodology and use the collected corpus as testset. The results of our experiments point once again to the importance of combining both DAs and entities information for predicting turn coherence in open-domain conversation.

Chapter 6 focuses on *supervised* approaches to open-domain dialogue evaluation. In the first section, based on Cervone et al. (2018), we propose supervised models for automatic prediction of user ratings at the *conversation level* using a dataset of

[8] Available at `https://github.com/alecervi/switchboard-coherence-corpus`

human-machine chitchat conversations. We find that predicting user ratings is a difficult task even for humans and that DAs are useful for predicting conversational ratings. In the second section, based on Yi et al. (2019), we explore supervised models for automatic prediction of coherence and engagement at the *turn level* in open-domain human-machine chitchat conversation. Our models rely on a combination of features including DAs and entities information. Additionally, we also investigate how to use these turn level evaluation models for coherent and engaging response generation.

Chapter 7, based on Cervone et al. (2019), investigates models for Natural Language Generation with open-domain entities. Our models apply the Meaning-Representation-to-text approach typical of task-oriented dialogue, where the MR is composed by a DA and a list of entities (slots), to an open-domain setting.

Finally, in Chapter 8 we summarize the contents of our work and draw on the conclusions of this thesis.

1.6 Publications

The contents of this thesis are partially based on the following peer-reviewed publications (ordered according to their appearance in the dissertation):

1. **Cervone, A.**, Tortoreto, G., Mezza, S., Gambi, E., and Riccardi, G. (2017). Roving mind: a balancing act between open–domain and engaging dialogue systems. *First Proceedings of the Alexa Prize.*

2. Mezza, S., **Cervone, A.**, Stepanov, E., Tortoreto, G., and Riccardi, G. (2018). ISO-Standard Domain-Independent Dialogue Act Tagging for Conversational Agents. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3539-3551).

3. **Cervone, A.**, Stepanov, E., and Riccardi, G. (2018). Coherence Models for Dialogue. *Proceedings of the 19th Annual Conference of the International Speech Communication Association* 2018, 1011-1015.

4. **Cervone, A.**, and Riccardi, G. (2020). Is this Dialogue Coherent? Learning from Dialogue Acts and Entities. *Proceedings of the 21st Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL).*

5. **Cervone, A.**, Gambi, E., Tortoreto, G., Stepanov, E. A., and Riccardi, G. (2018). Automatically Predicting User Ratings for Conversational Systems. In *Fifth Italian Conference on Computational Linguistics (CLiC-it).*

6. Yi, S., Goel, R., Khatri, C., **Cervone, A.**, Chung, T., Hedayatnia, B., Venkatesh, A., Gabriel R., and Hakkani-Tur, D. (2019). Towards Coherent and Engaging Spoken Dialog Response Generation Using Automatic Conversation Evaluators. *Proceedings of the 12th International Conference on Natural Language Generation* (pp. 65-75).

7. **Cervone, A.**, Khatri, C., Goel, R., Hedayatnia, B., Venkatesh, A., Hakkani-Tur, D., and Gabriel, R. (2019). Natural Language Generation at Scale: A Case Study for Open Domain Question Answering. *Proceedings of the 12th International Conference on Natural Language Generation* (pp. 453-462).

Additionally, the publications not directly relevant for this dissertation, but which were nevertheless produced during the course of the PhD are (in reverse chronological order):

8. Tammewar, A., **Cervone, A.**, Messner, E. M., and Riccardi, G. (2020). Annotation of Emotion Carriers in Personal Narratives. *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020).*

9. Tammewar, A., **Cervone, A.**, Messner, E. M., and Riccardi, G. (2019). Modeling user context for valence prediction from narratives. *Proceedings of the 20th Annual Conference of the International Speech Communication Association* 2019.

10. Marinelli, F., **Cervone, A.**, Tortoreto, G., Stepanov, E. A., Di Fabbrizio, G., and Riccardi, G. (2019). Active Annotation: bootstrapping annotation lexicon and guidelines for supervised NLU learning. *Proceedings of the 20th Annual Conference of the International Speech Communication Association* 2019.

11. Aubin, A., **Cervone, A.**, Watts, O., and King, S. (2019). Improving speech synthesis with discourse relations. *Proceedings of the 20th Annual Conference of the International Speech Communication Association* 2019, 4470-4474.

12. Dubiel, M., **Cervone, A.**, and Riccardi, G. (2019, August). Inquisitive mind: a conversational news companion. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (p. 6). ACM.

13. Tortoreto, G., Stepanov, E., **Cervone, A.**, Dubiel, M., and Riccardi, G. (2019). Affective Behaviour Analysis of On-line User Interactions: Are On-line Support Groups More Therapeutic than Twitter?. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task* (pp. 79-88).

14. **Cervone, A.**, Stepanov, E. A., Celli, F., and Riccardi, G. (2017). Irony Detection: from the Twittersphere to the News Space. In *Fourth Italian Conference on Computational Linguistics (CLiC-it)* (Vol. 2006).

Chapter 2

Background

In this Chapter, we review some of the background work on which the contributions of this thesis rely. First, in Section 2.1, we review current approaches to conversational AI research, mainly divided between modular task-oriented conversational agents (described in Section 2.1.2) and non-task-oriented dialogue models (discussed in Section 2.1.3), both with their advantages and disadvantages. We conclude this first section with a discussion about the growing area of dialogue evaluation (in Section 2.1.4).

Then, in Section 2.2, we describe approaches to coherence from a theoretical Linguistics perspective, with a particular focus on how this concept has been studied for the dialogue genre.

Finally, in Section 2.3, we provide an overview of how coherence has been investigated from a Computational Linguistics perspective. Also in this case, our main focus is on coherence in dialogue, rather than on the text genre. In the last part of the section, we discuss in more details approaches to coherence based on entities and on Dialogue Acts.

2.1 Conversational AI

As humans, one of the very first attributes that comes to our mind when defining the concept of intelligence is the ability of an entity to communicate with us through language. Indeed the Turing test (Turing, 1950), still the most influential test in Artificial Intelligence (AI) to demonstrate the ability of a machine to replicate human intelligence, is framed into a supposedly natural conversation between the entity being tested and a human.

Nevertheless, conversational AI, the research domain aimed at building machines able to interact with humans through language, is a relatively yet unexplored area in AI research. Although it could prove useful for several different applications, this field gained the attention of the wider research community only in recent years, due to its complexity.

As shown in Figure 2.1, given as input a sentence produced by a human in the form of a waveform, a classic CA architecture processes that acoustic signal in order to determine the sequence of words uttered, extract their semantic meaning and then elaborate a response for the user and output it in the form of a waveform.

This means that a CA relies on the integration of several different technologies (such as Automatic Speech Recognition, Spoken Language Understanding, Natural Language Generation, Text-To-Speech) together into a complex architecture. Until recently, however, many of these technologies were not mature enough to be actively employed for such complex applications and thus research on CAs has been quite sparse and limited to relatively few laboratories in the last quarter of the 20th century.

Over the last decade, however, due to improvements in Machine Learning techniques especially with the employment of Deep Neural Networks (LeCun, Bengio, and Hinton, 2015) algorithms, we have seen major advances in prominent fields for conversational AI such as Automatic Speech Recognition (Hinton et al., 2012; Xiong et al., 2016), Natural Language Processing (Bengio et al., 2006; Mikolov et al., 2010; Mikolov and Zweig, 2012; Devlin et al., 2019) and Text-to-Speech (Oord et al., 2016).

Thanks to these advancements, in the last few years the field has witnessed an explosion of CAs also in commercial applications (e.g. Apple Siri, Amazon Alexa, the Google Assistant and several others) and is projected to grow even more in the next years (KennethResearch, 2020).

While current commercial applications are mostly based on *task-oriented* designs, there is also another line of research in conversational AI which focuses on *non-task-oriented* approaches. This distinction between task-oriented and non-task oriented approaches, which has characterized conversational AI research since the early days, still divides current research approaches, with only few works attempting to fill the gap in between. In particular, the two approaches have evolved into two different types of architectures: on one hand we have task-oriented CAs which typically rely on *modular* architectures, on the other we have *non-modular* approaches, such as response generation models, which have been mainly explored for chitchat conversations without a real practical task to accomplish. We will discuss both these approaches, with their respective advantages and disadvantages, in the next sections.

However, before discussing current trends in conversational AI approaches, in the next section we provide some basic terminological distinctions for the growing area of dialogue research.

2.1.1 Terminology

The recent explosion in usage of CAs also corresponded to an explosion in the related terminology, as a natural consequence of the renewed interest in the field. This terminological explosion, however, could have negative consequences, such as fostering confusion on specific' models designs and thus making models' comparability harder. Here, we introduce some terminological distinctions useful within the context of this thesis:

Chatbot, socialbot, dialogue system, conversational agent Computational models able to have a dialogic interaction with humans have been referred to with several different names over the years and across separate fields. These terms, however, are not necessarily synonyms, since some typically refer only to specific instances of the broad category. An additional layer confusion is also given by the different usage of the same term across separate fields (for example research papers' terminology might differ from the one used in commercial applications).

The name *chatbot*, for example, typically refers to a dialogue model designed without a specific task in mind (e.g. chitchat models), but rather for entertainment. Also, at least in the earliest days, chatbots were relying on simple keywords mechanisms for generating conversational responses to the user (see the description of Eliza in Section 2.1.3). However, since this term has entered the field of commercial applications, it is now also used in some cases within this context to refer to models based on a more structured Meaning Representation (such as the one based on slots and intents) or even to the whole category of dialogue systems. The term *socialbot*, similarly to chatbot, is used to denote a dialogue model designed for entertainment and especially user engagement [1].

On the other side of the spectrum, the term *dialogue system* more typically refers to models designed for task-oriented interaction, usually based on a modular structured architecture (see the next Section). The term *conversational agent* is also typically used in reference to modular dialogue systems, though it is also used in some cases to refer broadly to the whole category.

In this thesis, we use the term dialogue systems or CAs when referring to models based on a modular architecture (both task- and non-task-oriented), while we prefer the term dialogue models (and, occasionally, the term chatbot) when referring to models based on non-modular architectures (such as response generation models). The term socialbot is used within the context of the Amazon Alexa Prize to describe various non-task-oriented models implemented by participants to the competition.

Single-domain vs open-domain This dichotomy refers to the difference in the domain coverage of a given CA. While single-domain models are designed to cover only one domain (e.g. restaurants reservation), open-domain models do not depend on a given domain in their design, but could rather handle any domain (non-task-oriented models tend to belong to this latter category). Early task-oriented CAs were typically single-domain, though also thanks to the rapid acceleration seen by the field in recent years, we are witnessing an increasing number of multi-domain CAs. A multi-domain CA is a system able to support multiple, though still predefined sets of domains (Wen et al., 2016b).

Task-oriented vs non-task-oriented The distinction between task-oriented (or task-based) and non-task-oriented CAs indicates whether the model has been designed with the goal of accomplishing given tasks for the user or not. As pointed out in Grosz and Sidner (1986), any conversation is in some way task–oriented, however in some conversations the task at hand is less defined or practical compared to others. The term task-oriented CAs thus usually indicates CAs able to accomplish practical, short-term tasks (standard examples of these tasks are making a restaurant reservation or booking movie tickets).

System-initiative, user-initiative and mixed-initiative CAs can be categorised also according to who, between the machine and its interlocutors, has the initiative in the conversation (Walker and Whittaker, 1990). A Question-Answering system, for example, is usually characterised by user-initiative, since it is the user who has the initiative in the interaction. However, spontaneous human conversation is usually based on mixed initiative, that is the case where all interlocutors can take the conversational initiative. Early task-oriented CAs, for tasks such as call-routing, tended to

[1] As it has been used, for example, in the context of the Amazon Alexa Prize to refer to all models implemented by the competing teams, regardless of the underlying architecture.

be more based on system-initiative where the interaction could be easier to handle for the machine, while current CAs tend to be more mixed-initiative.

Intents Classical CAs architectures are usually based on a MR language involving the notion of intents. Intents are typically taxonomies predefined by the system's designer to capture the possible intentions the user might have in the interaction with the machine. For example, when designing a CA for restaurant reservation we might have an intent category such as `request_price`, to capture users' utterances, such as "Could you tell me the price of the restaurant?", which have the goal of requesting the price of a given restaurant.
The notion of intents can be confused with the one of Dialogue Acts, since both involve taxonomies for speakers' intentions behind given utterances. However, intents are usually heavily bound to their end application domain. DAs, on the other hand, can be considered a generalised version of intents with the aim of capturing more abstract aspects of the interaction. For example, in the case of the utterance mentioned the DA would be of a *directive* type, since the speaker is directing to machine towards sharing a given information (in this case the price). We provide a more in-depth discussion on the notion of DAs in Section 2.3.2.

Slots Within a task-oriented CA, slots can be described as the semantic concepts relevant for a given task. This notion, associated to frame-based semantics (Tur and De Mori, 2011), is another crucial part together with intents of classic task-oriented CAs Meaning Representation language. To each slot *type* usually corresponds a set of slot *values*. An example of slot type could be `restaurant_name`, where its associated slot types could be {`The Shed, Da Marco, Chez moi`}. While prototypical examples of slot values tend to be Noun or also Adjective Phrases (see also the example in Figure 2.1), in some cases this category has been used to capture also binary choices (as the slot type `kidsallowed` with {`yes, no`} values in the San Francisco restaurant dataset (Wen et al., 2015)).

2.1.2 Task-oriented Modular Dialogue systems

The area of task–oriented CAs, which have the goal of accomplishing predefined tasks given by users (e.g. booking movie tickets), is fairly well–established.

Early applications of task–driven CAs started from call–routing (AT&T's HMIHY (Gorin, Riccardi, and Wright, 1997)) and travel planning (GUS (Bobrow et al., 1977), ATIS (Hemphill, Godfrey, and Doddington, 1990) or DIALOGOS for italian telephone-based railway timetable inquiries (Albesano et al., 1997)). Most commercial CAs (Alexa, Google assistant, Cortana) belong to this type and task–oriented dialogue systems are also the most investigated ones in the literature, probably because of the more practical (and arguably easier) function they are required to accomplish compared to non-task-oriented models.

Given their predefined purpose (e.g. booking a restaurant), when building task-based dialogue systems designers can make assumptions regarding the user goals (e.g. make a reservation), the type of entities that will be introduced in the conversation (e.g. menus, prices, locations) and thus the probable paths the conversation will follow (e.g. the user requests a specific type of restaurant, the machine proposes a range of restaurants satisfying those requirements etc.). Task-oriented CAs are thus usually designed for a specific *domain* (e.g. restaurant reservations), with a predefined set of *slots* (e.g. restaurant prices, names), user *intents* (e.g. request a chinese restaurant) and machine *states* (e.g. the machine believes the user is looking for a chinese restaurant) and *actions* (e.g. search a database of restaurants).

Figure 2.1 shows an example of a typical task-oriented CA architecture, composed of a pipeline of multiple modules. Excluding the speech-related modules (Automatic Speech Recognition and Text-To-Speech), in the next paragraphs we give an overview of the role of each module in the pipeline and how the tasks associated to that module have been approached in the literature.

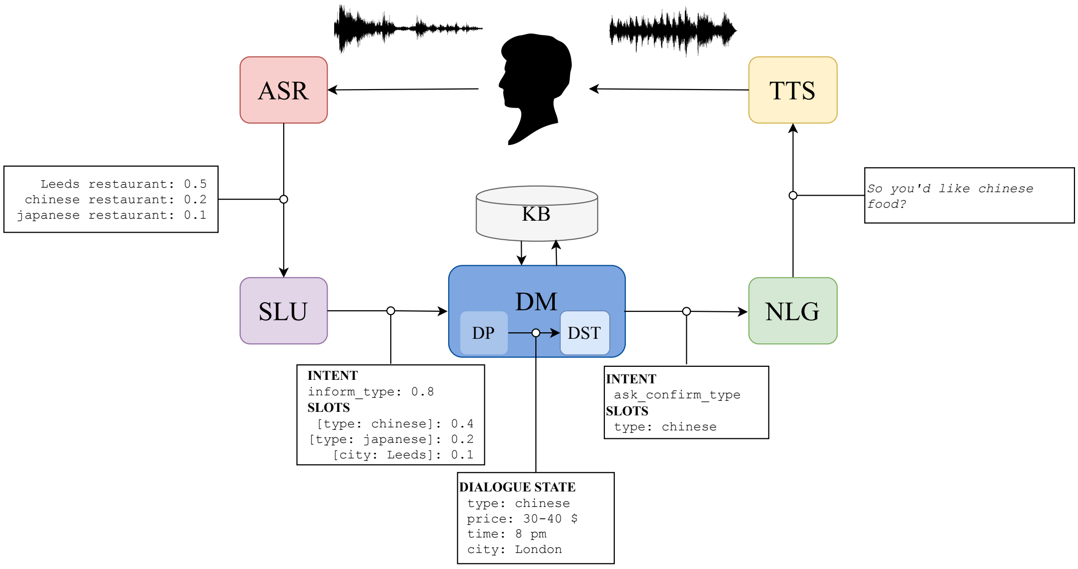


FIGURE 2.1: Traditional task-oriented modular conversational agent architecture. The user utterance is processed by the Automatic Speech Recognition (ASR) module, which produces a list of hypotheses regarding the sequence of words uttered. The Spoken Language Understanding (SLU) module, takes as input these hypotheses and outputs a semantic parsing of the given utterance (typically into intents and slots). Then, the Dialogue Manager (DM), using the information coming from the SLU, updates the current dialogue state in the Dialogue State Tracker (DST) and consequently chooses the policy for the next action in the Dialogue Policy (DP). The chosen action, structured as a series of intents and associated slots, is then passed to the Natural Language Generation (NLG) that transforms it into a textual format, which is finally passed to the Text-To-Speech (TTS) module.

Spoken Language Understanding The goal of the SLU is to produce a probability distribution of semantic hypotheses over the meaning of the input utterance, starting from a distribution over word sequences coming from the ASR. In traditional task-oriented CAs, the main tasks of the SLU are intent and domain classification (the latter being necessary when the CA is multi-domain) and slot filling (Tur and De Mori, 2011, Chapter 3). *Domain* and *intent detection* in task–oriented CAs are usually modelled as classification problems, addressed in the literature using various Machine Learning classifiers, such as SVMs (Haffner, Tur, and Wright, 2003) or MaxEnt (Chelba, Mahajan, and Acero, 2003), Deep Belief Networks (Sarikaya, Hinton, and Deoras, 2014), and LSTM (Wang, Shen, and Jin, 2018). On the other hand, *slot filling* in task–driven CAs has been considered in many approaches as a sequence labelling task, accomplished employing models such as HMMs (Wang, Deng, and Acero, 2005), CRFs (Raymond and Riccardi, 2007), RNNs (Mesnil et al., 2015) and

Encoder-Decoder models (Zhu and Yu, 2017).
While several works address intent classification and slot filling as separate tasks, recent work also showed advantages in combining the two (Guo et al., 2014; Liu and Lane, 2016).
Interestingly, some successful approaches to task–based CAs more concentrated on Dialogue Management propose to skip SLU all together (Henderson, Thomson, and Young, 2014b; Henderson, Thomson, and Young, 2014a; Wen et al., 2016a) and give directly the ASR n–best distribution as input to the DM in order to avoid information loss in the SLU component. This approach, however, it's easier to apply to single domains where the range of slot types is restricted, while it would be more difficult to handle for open–domain conversation.

Dialogue Management Taking as input either the full SLU n–best distribution or only the most probable hypothesis, the Dialogue Manager is aimed at selecting the next action of the system, usually interacting with a Knowledge Base which could be a simple database (Wen et al., 2016a) or combine various sources in more complicated applications (e.g. calendars, phonebook, API, web crawling) (Lemon, 2012). The techniques and architecture used in this component can vary greatly according to the approaches.
While DMs can be a simple set of rules (see the one of Eliza in Section 2.1.3), in traditional CAs research DM are usually divided into two subcomponents:

- a *Dialogue State Tracker*, which, given as input the distribution of hypotheses over possible intents and slots from the SLU, estimates and updates the dialogue state of the conversation keeping track of the information gathered by the system up to the current utterance (the dialogue state usually consists of a predefined set of slot types-value pairs, as in Figure 2.1);
- a *Dialogue Policy*, which receives as input the chosen dialogue state and selects the next action of the machine (usually framed as a structure of intents and associated slots) to be passed to the NLG.

DST is currently one of the most fertile and well–studied areas of conversational AI research, also thanks to the Dialogue State Tracking Challenge (Williams, Raux, and Henderson, 2016). Following Williams, Raux, and Henderson (2016), DST approaches can roughly be divided into rule-based (Wang and Lemon, 2013; Sun et al., 2016), generative statistical models (Young et al., 2013) and discriminative statistical models (Xie et al., 2018).
On the other hand, the task of learning the best Dialogue Policy has been extensively researched using Reinforcement Learning training techniques (Williams and Young, 2007; Georgila and Traum, 2011).

Natural Language Generation Although NLG is quite a broad and comprehensive field, in this thesis when referring to this area we will consider mainly the data-to-text approaches typically used in CAs architectures (Gatt and Krahmer, 2018). Given as input the next action chosen by the DM, typically framed as a MR composed by intents and associated slots (e.g. in Figure 2.1 where the intent is `ask_confirm_type` and the slot is `type:chinese`), this module is responsible for generating a corresponding utterance (e. g. "so you'd like chinese food?").
Although quite important especially for usability and for creating a believable persona for the CA, NLG has not been relatively less investigated by researchers compared to the previous modules. Commercial applications and research still rely

heavily on handcrafted templates (Cheyer and Guzzoni, 2014), although some approaches tried using an overgeneration and reranking approach (Oh and Rudnicky, 2000). State-of-art approaches are usually based on neural models using LSTM (Wen et al., 2015) and encoder-decoder models (Nayak et al., 2017).

Research on task-oriented CAs has been carried out mostly on each module independently in early years, without considering the interdependence across different tasks in the pipeline, with the risk of error propagation and loss of information over the flow of the system. To overcome such issues, in recent years researchers have increasingly investigated joint training across different modules (Bayer and Riccardi, 2012; Yang et al., 2017; Rastogi, Gupta, and Hakkani-Tur, 2018) in the pipeline and end-to-end trainable CAs architectures (Wen et al., 2016a; Dhingra et al., 2017).

Among the main advantages of the traditional task-oriented pipeline are its *interpretability* at each step of the process and thus the comparatively high *control* that designers can exercise over the CA behaviour. Interpretability and control are crucial characteristics for systems designed to interact with humans, especially in the case of real-world deployment. Achieving interpretability is possible also thanks to the mapping of the conversation to a more abstract set of categories, i.e. the Meaning Representation, which through the notion of intents and slots allows to capture users' behaviour and relevant concepts into a predefined set of labels over which each module can reason and learn. However, such a dependence on a predefined set of categories, usually valid only for a single domain, can also lead to important issues.

Indeed, the main drawback of task-oriented CAs lies in their *scalability*. While for a small number of tasks it makes sense to hand-craft the possible states of the machine and rely on a closed-set MR language (where only a predefined set of slots and intents are defined), this methodology poses real issues for scalability and the ability of models to transfer even to similar domains.

2.1.3 Non-task-oriented models

On the other hand of conversational AI research, we have the less established area of *non–task–oriented* dialogue models, where the machine does not have a specific, practical task to accomplish, but rather its main goal is typically engaging the user in open–domain conversation.

Actually, some of the very earliest examples of human-machine interaction come from this type of models. The well known Eliza system (Weizenbaum, 1966), for example, was not designed with specific, practical tasks in mind, but rather with the goal of engaging users in conversation and study human-machine interaction. Here is a famous example of an interaction between a user and Eliza, where the system plays the role of a Rogerian psychologist (Johnson, 2011):

Example 2
young woman: Men are all alike.
eliza: In what way?
young woman: They're always bugging us about something specific or other.
eliza: Can you think of a specific example?
young woman: Well, my boyfriend made me come here.
eliza: Your boyfriend made you come here?
young woman: He says I'm depressed much of the time.

eliza: I am sorry to hear you are depressed.
young woman: It's true. I'm unhappy.
eliza: Do you think coming here will help you not to be unhappy?
young woman: I need some help, that much seems certain.
eliza: What would it mean to you if you got some help?
young woman: Perhaps I could learn to get along with my mother.
eliza: Tell me more about your family.
young woman: My mother takes care of me.
eliza: Who else in your family takes care of you?
young woman: My father.
eliza: Your father.

Reading this conversation could give the impression that Eliza was able to understand the complexity of the meaning portrayed in the user's utterances. Nothing further from the truth. Eliza was designed by its creator to rely upon very simple pattern matching rules on purpose, for example the fact that the user might mention a keyword matching a family member (e.g. "mother") would trigger the template response "Tell me more about your family". With such simple mechanisms, its creator wanted to show how superficial human-machine interaction could be. Surprisingly for its creator, several users who engaged in conversation with Eliza seemed to attribute to the system a real understanding of the conversation. Parry, another model of the same period built using similar mechanisms but designed to play the part of a paranoid person, even passed a version of the Turing test (Colby et al., 1972). Interestingly, researchers also experimented with making Parry and Eliza talk to each other. Here's an extract from one of their interactions (Johnson, 2011):

Example 3
parry: People get on my nerves sometimes.
eliza: I am not sure I understand you fully.
parry: You should pay more attention.
eliza: Suppose you should pay more attention?
parry: You're entitled to your opinion.
eliza: What makes you think I'm entitled to my opinion?
parry: It bothers me just to be around people.
eliza: Please go on.
Parry continues with some paranoid ramblings about horseracing, bookies, and the Mafia
parry: I know the mob controls the big rackets.
eliza: Who are you really talking about?
parry: Have you ever really thought about the Mafia?
eliza: Suppose you ever really thought about the Mafia?
parry: Let's talk about something else.
eliza: I am not sure I understand you fully.
parry: You keep asking me the same things, why not change the topic?
eliza: In what way?
.....
parry: I am getting tired of answering the same questions over and over.
eliza: What resemblance do you see?
parry: Why do you need to know that?
eliza: What do you think?

......
parry: You harp on the same questions too much.
eliza: What does that similarity suggest to you?
parry: I don't get you.
eliza: Do you wish that you did get it?
parry: It's about time you showed some interest in my feelings.
eliza: Oh, I showed some interest in your feelings?

Compared to Example 2 where the reader might attribute a form of intelligence and deep understanding of the conversation to Eliza, in this case it would be much harder to attribute intelligence to any of the two machines involved. The impression given by the Eliza-Parry interaction is indeed one of a lack of sense, a lack of coherence across the various turns of the conversation. Example 3 shows how both these models crucially rely on the human interpreter for maintaining the coherence of the conversation and therefore give the illusion of intelligence. While Eliza and Parry were early examples of what today are known as "chatbots", or non-task-oriented dialogue models, there are models that rely on a very close architecture still used to this day (see Alice (Wallace, 2009)).

The type of behaviour shown in Example 3 highlights one of the perils of non-task-oriented open-domain models, that is that in several cases they lack *control* of the conversation without a structured representation for users' goals. Although open–domain dialogue seems by far a more challenging task than booking a flight since we cannot make assumptions over the conversation flow and restrict the space of the possible directions the dialogue can take, current approaches to open–domain CAs (Ritter, Cherry, and Dolan, 2011; Sordoni et al., 2015; Sutskever, Vinyals, and Le, 2014; Serban et al., 2016d) also have a level of control over the conversation that is much less than the one achieved by SOTA task-oriented approaches.

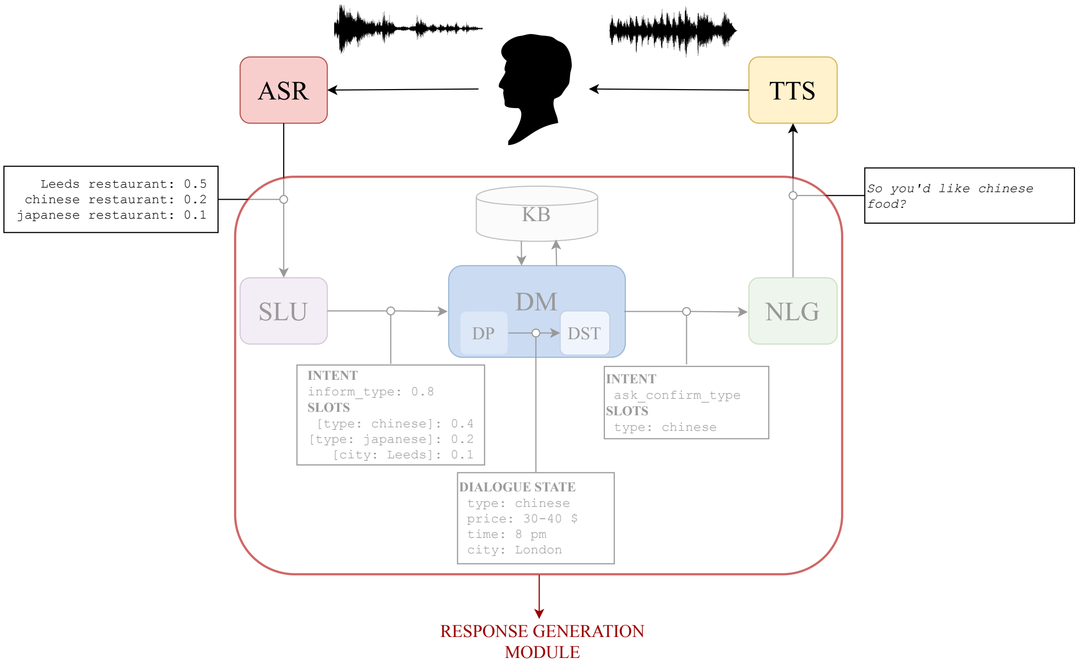


FIGURE 2.2: Non-modular models typically replace the whole pipeline of Spoken Language Understanding, Dialogue Manager and Natural Language Generation with a single response generation module.

INPUT: what is your name?
I don't know.
I don't know!
I don't know, sir
Oh, my god!
...
My name is Robert
My name is John

TABLE 2.1: Sample responses generated by a Seq2Seq model to a given input, ordered according to their probability (taken from Li et al. (2016a)). The ellipsis indicates several other responses not shown before the final one.

Excluding again the necessary speech-related modules, the pipeline of submodules used in task–based systems is here reduced to a single module (see Figure 2.2) which selects the most similar response to the user utterance as learned from a large database of dialogues using Machine Learning techniques such as DNN, in particular various versions of Sequence-to-sequence (Seq2seq) – also known as Encoder-Decoder – (Sutskever, Vinyals, and Le, 2014) neural networks (Sordoni et al., 2015; Vinyals and Le, 2015). These type of models thus, initially borrowed from Machine Translation, directly map the user input utterance to the machine response without intermediate representations (such as the MR used in the task-oriented framework). This literature has mostly focused on two tasks: response selection and response generation.

Response selection This task, also known as Next Utterance Classification (NUC), consists of selecting the optimal response to a given utterance (which could be extended to include a larger context of the previous conversation), given a list of candidate responses (Lowe et al., 2017b). Models trained on this task can be applied for ranking the output of generation models, or for selecting the next response in a retrieval-based dialogue model. The latest SOTA models (Kim et al., 2019) typically rely also on large pretrained language models for the task (Devlin et al., 2019).

Response generation This task, also known as Next Utterance Generation (NUG), consists of generating a response to a given utterance (also here the context could be extended). Models trained on this task could be directly applied as whole dialogue models, without using other modules, as shown in Figure 2.2. Although Encoder-Decoder models have been the most extensively used for this task, also in this case, the latest models use large pretrained models (Zhang et al., 2019; Adiwardana et al., 2020).

Probably the biggest advantage of these models compared to the task–oriented framework is that they require *no hand–crafted features*. Indeed, without relying on intermediate representations (i.e. MRs) and using tasks that can be addressed using weakly supervised approaches (datasets for response selection and generation can be created automatically), these models do not necessarily require annotation effort. However, very *large datasets* are necessary for training these models, which are not easy to retrieve or produce.

Moreover, these models present also structural problems that are not due only to lack of data. The main idea behind these approaches is indeed the fact that the most

A: how old are you? (1)
B: I'm 16. (2)
A: 16? (3)
B: I don't know what you are talking about (4)
A: You don't know what you are saying. (5)
B: I don't know what you are talking about (6)
A: You don't know what you are saying. (7)
...

TABLE 2.2: Sample conversation generated by making two Seq2Seq models interact with each other (taken from Li et al. (2016c)), exemplifying how models quickly fall into "conversational black holes".

probable responses to a given message are learned from a large dataset of conversations. However, the majority of responses in any large database of conversations tend naturally to be also the shorter, less engaging and more repetitive ones (such as 'I don't know' or 'I see' as in Table 2.1), while the more content–related ones are more sparse and bound to the greater context of the conversation. This problem is known as the *generic response* problem. Generally these models seem unable to keep the conversation context for more than a few turns and thus give a sense of meaningless, *incoherent* conversation to users when interacting over multiple turns. This type of behaviour is exemplified in Table 2.2, which reports a conversation (somewhat reminding of Example 3) where two Seq2Seq models interact with each other and over a few turns the conversation ends up in a loop of generic responses. Moreover, since the training dataset contains thousands of responses from different users, the models do not have a consistent persona and could give incoherent responses even to the same question over the course of the same interaction (as shown in Table 2.1).

Recent models tried to overcome these issues either by extending the DNN architecture to incorporate more context (dialogue history) (Serban et al., 2016b; Serban et al., 2017c) or by trying to use different features for the training of the model in order to promote more diverse and on–topic responses using Reinforcement Learning (Li et al., 2015; Li et al., 2016c).

Following the second approach, Li et al. (2016c) use Reinforcement Learning to model a reward function according to these ideal conversation properties: informativity (it penalizes the cosine similarity between two adjacent machine turns), semantic coherence (using Maximum Mutual Information (Bahl et al., 1986), a metric which evaluates the mutual interdependence between two sequences), and ease of answering (manually compiling a short list of common dull responses in Seq2Seq models and designing a function to extract their common features with the hope to cover similar cases). Although this model shows consistent improvements over a combination of metrics (length of generated turns, number of different bigrams, human evaluation), the authors admit how the reward function does not yet cover many crucial aspects which contribute to a good conversation.

2.1.4 Evaluation

Evaluation of dialogue systems is an open issue, since so far there are no standardized automatic metrics valid for both task- and non-task-oriented models. Additionally, while in modular task-oriented approaches each module has its own evaluation techniques besides metrics used to evaluate the entire system (although the two do not necessarily coincide (Takanobu et al., 2020)), naturally the same is not true for non-modular models. In the latter case, the evaluation of the entire model coincides with single units evaluation and therefore having meaningful metrics becomes even more crucial.

Current approaches to dialogue evaluation can be roughly grouped into task-oriented metrics, metrics based on surface text, metrics relying on learned models and human judgement [2].

Task-oriented metrics This class of metrics, based on the task success rate, works only for task-oriented CAs evaluation (Walker et al., 1997). This family of metrics is generally aimed at evaluating the success of the system in accomplishing the task requested by the user. Danieli and Gerbino (1995), for example, propose contextual appropriateness, implicit recovery and transaction success, as metrics to evaluate the ability of the system to appropriately deal with a user request and recover from errors. Perhaps the most famous framework of evaluation proposed for task-oriented dialogue is PARADISE (Walker et al., 1997). In PARADISE, the task success rate is measured using matrices which represent the information requirements towards accomplishing a given dialogue task. We can observe how these metrics focus mainly on the ability of the CA to accomplish a given task, but do not cover other less practical aspects which might however also contribute towards the system's evaluation (i.e. did the user feel the machine was respectful towards the user? did the user give more implicit signals of dissatisfaction with the interaction?).

Surface-based metrics This group of metrics, mostly borrowed from Machine translation, evaluates directly the text generated by the machine responses. BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), for example, evaluate the similarity between two sentences based on formal factors such as the number of overlapping n-grams (sequences of words of various lengths) among the two. These metrics are thus used to measure the ability of the machine to replicate exactly a given expected response. Although these metrics are widely used to this day to evaluate non-task-oriented response generation models, they have been shown to have a poor correlation with human judgements (Liu et al., 2016). This is hardly surprising, since in dialogue responses with very different surfaces forms might be perfectly valid conversational responses in the same context.

Model-based metrics This class of approaches propose to create to train Machine Learning models on the task of dialogue models evaluation. Although a promising direction, in many cases the models proposed learn to predict generic human judgements without providing closer insights to why a given utterance has been given a particular evaluation score or to the characteristics that machine responses should have in order to achieve a good evaluation score (Lowe et al., 2017a).

Human judgement In general, manual surveys still remain the best option, although the process is expensive and not completely reliable given the subjectivity of user judgements, especially non-expert ones.

[2] Computational models of evaluation for dialogue focusing on coherence will be discussed in Section 2.3.

Overall, it seems we're still far away from having automatic standardised evaluation metrics for both task- and non-task-oriented dialogue. Additionally, good evaluation metrics should also be informative regarding which characteristics of conversation quality a given CA fulfills or not in order to allow designers to act upon the reported issues, which is usually not the case for surface-based and model-based metrics. The current void in evaluation methodologies creates a real bottleneck especially for open-domain models, since these models are also often optimised using functions based on such surface-metrics, which are easy to compute, but are at the same time highly inadequate to evaluate conversational responses.

2.2 Coherence in theoretical Linguistics

In Linguistics literature the property which gives a unified meaning to a text is defined as its *coherence* (De Beaugrande and Dressler, 1981b). A discourse can be regarded as coherent if we can recognise in it a conceptual unity, a *sense*.

More formal definitions of coherence are however rather difficult to formulate, given the abstract nature of this concept. In order to better understand the idea of coherence the Linguistics literature has devised the concept of *cohesion* (or local coherence) (Halliday and Hasan, 1976), that is the property that makes the different sentences in a text connected to each other by a series of surface devices, such as having a coherent *consecutio temporum* (concordance of the grammatical tenses used in the sentences) or a continuity in the entities mentioned. In order to be coherent, a text must be cohesive (or locally coherent), but cohesion is not a sufficient condition for global coherence.

Example 4
" 'Contrariwise,' continued Tweedledee, 'if it was so, it might be; and if it were so, it would be; but as it isn't, it ain't. That's logic.' "
Lewis Carroll, *Through the Looking-Glass, and What Alice Found There*

Example 4 can be regarded as cohesive (or locally coherent), since the different clauses are connected by the reference to the same entity (in the text "it"), but it is nonetheless not globally coherent for the reader, because it does not have a recognizable sense.

The literature makes thus a distinction between *global* coherence — the "deep" conceptual unity of an entire text including also the goals of the speaker — and *local* coherence — the "surface" realization of coherence expressed through shared syntactic and semantic features across neighboring sentences.

As Andorno (2003) correctly identifies, the concept of coherence can be considered an overarching principle in Pragmatics and Discourse Linguistics. This concept is considered as hierarchically higher compared to other integral principles of Linguistics (Conte, 1980; Gernsbacher and Givón, 1995) since it represents a constitutive part of the very essence of texts. Ultimately the property of coherence does not even pertain to the text itself, but rather to the *interpretation that the receiver* of the text creates of it (Conte, 1989). The same text, could indeed be considered coherent or not according by different interpreters. For example, let's consider the following conversation:

Example 5
A: Where is Michael?

B: The train left the station.

Reading Example 5 one could think of it as incoherent, as the relation between the two utterances is not clear: A poses a question and B seems to make an unrelated statement. However, if we imagine that A told B in a previous part of the conversation or in another context that Michael would take that train, we could end up considering this exchange as perfectly coherent, since B's sentence would be considered an Answer to A's question.

This idea of the centrality of the interpretation process is a common thread across several approaches in Pragmatics and Discourse. The influential relevance theory (Sperber and Wilson, 1986), for example, posits that the meaning of an utterance goes much beyond its literal meaning and mainly stems from the effort made by the listener to find relevance in the message of the speaker, that is to find a consistent goal behind the speaker's messages. The reliance of the concept of coherence on the receiver's interpretation is also exemplified by the interaction between Eliza and Parry in Example 3, where the two machines quickly lose the sense of the conversation after a few turns and get stuck in loops seemingly without content or a purpose, and the same happens in the conversation between Seq2seq models reported in Table 2.2. In both cases, none of the two participants in the conversation was maintaining the coherence by interpreting in a unified way the utterances of the other speaker, assigning a goal behind the other's speakers utterances and hence giving a structure to the interaction (which instead is happening in Example 2).

Indeed one of the main tenets of the fields of Pragmatics and Discourse is that discourse has a structure. The way that utterances in a message are ordered is not random, but constitutes a structure that is an integral part of the meaning of the message itself. So much that if the order was changed, even by one sentence, the meaning, i.e. the interpretation of the message would change. Over the years, several theories have been proposed in Linguistics to uncover this sentence-by-sentence structure. Different theories, for example, focused on the importance of the interpretation process of the intentions of the producer in creating a given text (Grice, 1970; Sperber and Wilson, 1986). Several works studied discourse structure in terms of continuity of relevant entities across different parts of a text (Givón, 1983; Halliday and Hasan, 1976), also in connection to grounding (Givón, 1987). Other works concentrated more on the logical types of relations connecting a piece of text, also in connection to speaker's goals (Hobbs, 1979). Others identified a consistency in the use of tenses across a text in connection to discourse structure (Kamp and Rohrer, 1983; Moens and Caenepeel, 1994). While there are certainly overlaps and connections among the phenomena and the theoretical approaches of these works, the fields of Pragmatics and Discourse are generally more fragmented compared to other areas of Linguistics (e.g. Synthax, Morphology). As Andorno (2003) notices, this is not an issue of these research fields, but rather a defining factor: the main object of study of Pragmatics and Discourse is the *parole* (De Saussure, 1989), the language in context, where the context could be infinitely complex and various.

Notwithstanding the fragmentation of the field, there are some phenomena that have attracted the attention of scholars more compared to others. One of such phenomena is the way in which *entities* are introduced and referred to across different sentences in a coherent text, which we call the *thematic* aspect of discourse. One of the reasons why the response of B in Example 5 could be perceived as incoherent is the fact that the entity "train" is mentioned with the definite article "the", as if B assumed that A (or the reader) should know what train B is talking about. If in the previous context A had said for example "Michael should take a train at 11.40", B's response would be considered coherent and the reference to the train would be easily resolved by the receiver. The use of definite versus indefinite articles is indeed connected to whether or not a given entity has already been introduced within the context of a discourse or not. This is just one example of phenomena connected to the thematic structure of discourse. This area of study encompasses also phenomena such as how the load of information (such as novel versus already mentioned entities) is positionally and syntactically distributed within a sentence (the theme VS rheme dichotomy (Halliday and Matthiessen, 2013)), coreference and many others. Although present in any kind of discourse, this type of structures have been studied in particular for text, compared to other genres such as dialogue.

The thematic phenomena described are crucially connected to the concept of *mental models* (Van Dijk, 1985; Dijk, 1987; Van Dijk, 1999). The idea of mental models, also used in other research areas such as cognitive science, refers to mental constructs about shared knowledge that participants create and continuously update during the course of an interaction. Mental models imply thus a dynamic, rather than static approach to process the information exchange that takes place during an interaction. This concept affects both the way that speakers shape their messages and the way that receivers interpret messages. According to this idea, speakers carve the contents of their discourse according to the mental model they have about the receivers' knowledge status regarding the contents they are communicating. Vice versa, receivers construct their interpretation of a message according to the mental model they have of the mental model that they think speakers might have of their knowledge about a given topic. Hence, mental models are used by receivers to integrate their previous knowledge with the novel contents introduced in the conversation and thus make sense of the interaction. In this framework, coherence could be defined as the recognition by the receiver that a message was constructed using a mental model.

Coherence in dialogue As a genre, dialogue can be considered a privileged field of study for Pragmatics, since it is is even more dependent on the context compared to written text. Dialogic interactions have therefore been studied extensively in this research area, so much that there is even an whole genre-specific dedicated to the dialogic form, called Conversation analysis.

We have already mentioned the central role played by the interpretation of the receiver for coherence perception. This interpretation is inherently shaped by the intentions that the receiver attributes to the speaker. While the *intentional* aspect is important in any interaction, it has been particularly studied within the context of dialogue, where compared to text we have a more dynamic relation among the participants to a given act of communication.

The intentional aspect plays an important role in several recent influential theoretical approaches to Pragmatics. Grice's approach, for example, proposes that the whole communicative (non-natural) meaning of utterances is based on the intentional aspect, that is on the intended effect that the speaker wants to have on the receivers with that act of communication (Grice, 1970). Also according to relevance theory (Sperber and Wilson, 1986; Sperber, Cara, and Girotto, 1995), an influential Pragmatics framework proposed in recent years, speakers intentions play a central role in defining utterances' meaning. Naturally, the intentional aspect is also at the heart of

Speech acts theory (Austin and Urmson, 1962; Searle, 1965), a theory first developed in philosophy of language that studies the non-literal meaning of utterances in terms of conversational actions [3]. Another important line of research that investigates the intentional aspect of discourse is Conversational Analysis with the concept of adjacency pairs (Schegloff, 1968; Schegloff and Sacks, 1973; Sacks and Jefferson, 1995). Adjacency pairs are recognizable patterns of conversational actions (such as the fact that a question is usually followed by an answer in a conversation), categorised as a set of allowed speakers' intentions in the dialogue. Adjacency pairs capture an important characteristic of the intentional aspect: its dependence on context, that is the fact that the meaning of a given intention can be understood only within the larger conversational context of previous intentions shown by participants. While adjacency pairs mainly focus on capturing sequences of two intentions, other theories analysed intentional patterns spanning larger portions of the conversation (as we will see in Section 2.3.2). All these theories have been extensively studied over the years and used to investigate the structure of coherent discourse.
It is interesting to notice, that mental models are not just connected to the thematic aspect, but also to the intentional aspect of discourse, if we consider speakers' intentions as concepts that capture conversational actions associated to given changes in the information status of participants to the interaction. For example, a question could be considered a request to fill a gap in the knowledge of the participant posing the question or a statement could be considered an action to update the knowledge of the other participant with some given contents.

Hence, although these two important aspects, the thematic one and the intentional one, at the heart of different influential theories of discourse, might appear independent, they could actually be considered intertwined within a framework such as mental models. While the thematic aspect, capturing phenomena such as lexical chains and coreference, has been traditionally considered as an expression of local coherence; the intentional aspect, capturing phenomena such as adjacency pairs, has been often linked to the notion of global coherence. In the next section, dedicated to computational approaches to coherence, we will deepen both aspects and how they have been approached from a computational perspective.

2.3 Computational models of coherence

Computational Discourse and Pragmatics[4] are the fields of study that investigate computational models of utterances' meaning beyond their literal single-sentence meaning, but within the larger context (such as the interactional context or the larger linguistic context of an entire text or dialogue)[5].

One possible way of categorizing computational approaches to Pragmatics and Discourse is according to the methodological approach followed. While earlier research relies on an *inferential* approach, mainly based on belief logic and utilising logical inference; current research tends to rely on a *probabilistic* approach, based on the use of Machine Learning techniques (Horn and Ward, 2004, Chapter 26).

[3] The implications of Speech Acts and its evolution into Dialogue Acts theory will be discussed in Section 2.3.2

[4] Similarly to their Linguistics counterparts, while computational Pragmatics concentrates more on modeling language in context, computational Discourse focuses more on modeling the relationships among groups of utterances. Also in this case, there are overlaps among the two fields.

[5] For an introduction to Computational Pragmatics see (Huang, 2017, Chapter 19); for an introduction to Computational Discourse see (Jurafsky and Martin, 2009, Chapter 21).

Another way of categorizing computational approaches to Pragmatics and Discourse is according to their main object of study. While also in the case of computational approaches coherence plays a pervasive role across different theories, Moore and Pollack (1992) divides the main computational approaches to coherence into *informational* and *intentional* ones.[6] Although Moore and Pollack (1992) advocate for a synergy among the informational and intentional approach based on the fact that in a coherent discourse different clauses can be connected simultaneously by multiple type of relations, the two approaches have been largely investigated independently, also in terms of the genre studied. While traditionally informational approaches to coherence have been applied to written text (e.g. news); intentional approaches have been explored within the context of dialogic interaction.

Computational models of coherence in text *Informational* approaches (Moore and Pollack, 1992) have probably been the most influential approaches to coherence for written text, besides entity-based ones[7].

Moore and Pollack (1992) calls informational approaches a class of theories (Mann and Thompson, 1987; Prasad et al., 2008) which identify sets of coherence relations holding across different sentences in a coherent text. These coherence relations (also known as rhetorical or discourse relations, according to the approach), capture the rhetorical structure of discourse, describing how the *contents* of a given clause are logically related to the contents of other clauses within a coherent discourse. For example in the text "It was raining. But I had forgotten the umbrella" the first and second sentence are connected by a contrastive type of relation, since the two events described are in contrast with one another, as expressed by the connective "but". Instead, if we consider the text "I brought the umbrella. It was raining." the two sentences are connected by a causal type of relation, albeit in an implicit manner.

Within this general approach, we can further identify two main theoretical approaches: the Rhetorical Structure Theory (Mann and Thompson, 1988) and the approach behind the Penn Discourse Treebank (Prasad et al., 2008). While the former theory posits that it is possible to reconstruct a hierarchical structure among the different coherence relations in a text, a process known as discourse parsing (Marcu, 2000), the Penn Discourse Treebank approach only identifies flat types of relations among different clauses of a text. Current approaches to discourse parsing are generally based on probabilistic approaches using neural models (Braud, Coavoux, and Søgaard, 2017; Jia et al., 2018; Lin et al., 2019), rather than on inferential approaches.

Computational models of coherence in dialogue As we have seen for coherence approaches focusing on dialogic interaction in Linguistics; also computational approaches to coherence in dialogue mostly concentrate on the *intentional* aspect, that is the role of speakers' intentions in the conversation. In particular, this class of approaches rely on the intuition that dialogue coherence across different turns is inherently defined by the participation of each utterance to an overall goal, a *plan* that speakers have for that interaction (Cohen and Perrault, 1979; Allen and Perrault, 1980). Here, the purpose of dialogic interaction is thus to modify the participants mental state regarding given contents (Moore and Pollack, 1992).

[6] While in Linguistics the term *informational* is typically used within the context of information structure research, that is approaches that investigate phenomena such as discourse referents and lexical chains, which in this thesis we refer to as the thematic aspect of discourse; in this case the term informational refers to discourse relations theories, that rather capture the rhetorical relations between different parts of a text.

[7] Entity-based approaches to coherence will be extensively discussed in Section 2.3.1.

In this context, speech acts are considered as the key units, the operators used to put into action the overall plan (Allen and Perrault, 1980). However, we notice how early approaches analyse mainly the case of task-oriented dialogue. Thus, within these frameworks the speech acts used are usually very limited sets of categories specifically related to the speakers' beliefs and goals (in Allen and Perrault (1980), for example, the only speech acts analysed are REQUEST and INFORM). These works in general are based on inferential approaches.

Following Cohen and Perrault (1979) intuition, Grosz and Sidner (1986) propose arguably the most prominent theoretical framework for dialogue coherence also based on the essential role of the intentional aspect. Grosz and Sidner (1986) propose a framework that crucially combines the *intentional* and the *thematic* aspects for defining dialogue structure, and hence its coherence. In particular, Grosz and Sidner (1986)'s approach relies on the definition of an abstract model of dialogic discourse structure as an interaction of three components: the linguistic level, that is the sequence of utterances in their surface textual form; the intentional structure, which captures speakers' purpose in the conversation; and the attentional state, which defines the shifting focus of attention at each point in the conversation.

In this framework, the intentional structure is defined by the segmentation of the discourse in different segments, each one with a specific discourse purpose, which defines how that segment contributes to the overall purpose of the interaction. Similarly to Cohen and Perrault (1979), here the speakers' intentions considered are only the ones useful for the goal of the discourse and thus related to the beliefs of different participants. Different discourse purposes are connected by two main relations: dominance and satisfaction-precedence, governing the structure of the dialogic interaction and how different segments contribute to the overall purpose.

The attentional structure, on the other hand, is a dynamic model representing the participants' focus of attention as the discourse unfolds (Grosz and Sidner, 1986). Each discourse segment has thus an associated focus space, recording entities that are salient in the current segment. From this perspective, the structure of the conversation can then be understood as a series of shifts in focusing spaces, governed by specific transition rules (where focusing is the process of manipulating these spaces). The attentional structure, hence, captures the thematic aspect of dialogue coherence. These two different structures, along with the linguistic one, all interactively contribute to the segmentation of the dialogue in discourse segments.

Grosz and Sidner (1986) shows how this abstract model can be used to account for a series of linguistic phenomena, such as pronominalization and interruptions. Also in this case, the main genre considered is task-oriented dialogue and the approach is inferential-based rather than probabilistic. Grosz and Sidner (1986) approach has been very influential over the years, especially from an academic perspective, and it remains to this day one of the most highly regarded approaches to discourse structure in general. However, while this framework offers great insights, it found limited applicability so far for real-world applications (such as conversational agents design). While for specific use cases (such as task-oriented dialogue) this framework is easier to define, the same is not true for more extended cases (such as open-domain dialogue). Some of the reasons for its limited application are thus the difficulty in identifying (and finding agreement) the boundaries of discourse segments, and in general the richness of the theoretical model, which made the creation of dedicated annotated resources and tools a very time consuming effort. Additionally, while inferential-based approaches were very popular in the past, current AI research has shifted more towards a probabilistic data-driven framework.

In this section, we provided an overview of important approaches to coherence modelling in computational linguistics, with a focus on intentional approaches proposed for coherence modelling in dialogue. The next sections are dedicated to the two main computational approaches to coherence relevant for this thesis. First, we give a deeper insight into entity-based coherence models, which capture thematic phenomena; then, we review Dialogue Acts theory which has also been linked to dialogue coherence and which addresses intentional aspects of dialogue.

2.3.1 Entity-based models

Computational approaches to coherence modelling relying on entities capture what in this thesis we call the *thematic* phenomena of discourse. This group of approaches, which proved very influential over the years, posits that a coherent discourse is characterised by specific patterns of distributions in the continuity of given *entities* across different sentences. According to these frameworks, intuitively a text where each sentence mentions different entities would be thus perceived as less coherent compared to a text where the same entity appears in prominent positions across different sentence (as it happens for the entity “Microsoft” in Table 5.1). By prominent positions, these approaches intend for example the fact that a given entity would appear with the syntactic role of subject rather than a less prominent role such as the one of indirect object.

The most influential theory among entity-based approaches to coherence is *Centering theory* (Grosz, Weinstein, and Joshi, 1995), evolved from a combination of Grosz and Sidner (1986)'s work and theories laid out in Joshi and Kuhn (1979) and Joshi and Weinstein (1981). In particular, Centering is proposed as a framework for modeling the *local* level of the attentional structure defined in Grosz and Sidner (1986). Centering theory crucially relies on the idea that each utterance in a discourse is characterised by a center, i.e. a salient entity currently in focus. Within this framework, the perceived coherence of a discourse is affected by the continuity of centers across different sentences. This phenomenon is formalised in the notions of backward-looking centers and forward-looking ones. In Centering, each utterance has a backward looking center, i.e. the entity in focus at the current moment, and a number of forward-looking centers, that is a group of candidates to become the next salient entities in the following sentences. These candidates are not all considered on the same level, but are rather ranked according to their prominence (such as their syntactic role) in the sentence. For example, in Table 5.1 the backward-looking center of sentence 1 can be considered “the Justice Department", while "anti-trust trial" or "Microsoft Corp" could be considered forward-looking centers (indeed the entity “Microsoft”, referred also as “the corporation” in the same sentence, will become the next backward-looking center in the following sentences). Using these categories, Centering theory then formulates rules to describe centers transitions behaviour, for example to capture pronominalization phenomena of entities across different sentences. Centering theory has found various applications, in particular for tasks such as coreference resolution.

While Centering has been a very influential approach, there are also recognisable shortcomings of this framework. Poesio et al. (2004), for example, notice how some of the notions on which Centering relies are only partially specified, and thus there have been several different interpretations of how to correctly instantiate the theory. Additionally, another disadvantage of Centering is its top-down rather than data-driven approach to modeling discourse phenomena. Indeed, Centering assumes a

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics
s1	S	O	S	X	O	-	-	-	-	-	-	-
s2	-	-	O	-	-	X	S	O	-	-	-	-
s3	-	-	S	O	-	-	-	-	S	O	O	-
s4	-	-	S	-	-	-	-	-	-	-	-	S

s1: [The Justice Department]$_\mathrm{S}$ is conducting an [anti-trust trial]$_\mathrm{O}$ against [Microsoft Corp.]$_\mathrm{X}$ with [evidence]$_\mathrm{X}$ that [the company]$_\mathrm{S}$ is increasingly attempting to crush [competitors]$_\mathrm{O}$.
s2: [Microsoft]$_\mathrm{O}$ is accused of trying to forcefully buy into [markets]$_\mathrm{X}$ where [its own products]$_\mathrm{S}$ are not competitive enough to unseat [established brands]$_\mathrm{O}$.
s3: [The case]$_\mathrm{S}$ revolves around [evidence]$_\mathrm{O}$ of [Microsoft]$_\mathrm{S}$ aggressively pressuring [Netscape]$_\mathrm{O}$ into merging [browser software]$_\mathrm{O}$.
s4: [Microsoft]$_\mathrm{S}$ claims [its tactics]$_\mathrm{S}$ are commonplace and good economically.

TABLE 2.3: Entity grid example extracted from Table 1 and 2 in Barzilay and Lapata (2008) with a document composed by four sentences (s1-4). Entities in the sentences are annotated with their syntactic role: subject (S), object (O) or neither (X).

specific theory regarding how entities are continued in a coherent discourse and although several of its assumptions might be correct, they cannot define all nuances and variations of real-world linguistic behaviour.

While focusing on the same phenomena, the fertile *entity grid* framework (Barzilay and Lapata, 2008) relies on a different approach to modeling local coherence compared to Centering. Instead of assuming a particular theory about entities transitions in a coherent text, the entity grid approach relies on a data-driven methodology which lets the model learn a function directly from the data by simply providing relevant linguistic information (such as the grammatical roles of entities across multiple sentences).

As shown in Table 5.1, in this approach the authors propose to represent the structure of a text through a grid displaying relevant information (the grammatical role) about entities across neighboring sentences in a text. The grammatical role can be: subject (S), direct object (O) or neither (X), plus a symbol ($-$) to signal the entity does not appear in that sentence s. For example, in the case of the grid represented in Table 5.1 we notice how the sentences are connected by the continuity of the entity "Microsoft" which assumes prominent grammatical roles across all sentences. By computing the probabilities over all possible transitions of length n from one category to all others (thus $\{S, O, X, -\}^n$) we can turn this representation into a feature vector, representing the syntactic role transitions over entities in the text. Barzilay and Lapata (2008) then use these feature vectors to train a Machine Learning model (in their case a Support Vector Machine) framing coherence as a ranking problem.

Rather than using a supervised approach, which would depend on the time consuming process of data labeling, Barzilay and Lapata (2008) propose a *weakly supervised* approach to automatically generate the training data for the task. The authors automatically create a set of positive examples, by assuming a group original documents to be coherent, and a set of negative samples, by randomly permuting the order of the sentences in the original documents and assuming these to be examples of incoherent documents. Models are then trained using pairwise training on the data generated in such fashion. Intuitively, the authors expect the model to learn patterns of distribution in local entities transitions that differentiate coherent from incoherent texts. The authors also experiment with models using only salient entities (the most frequent ones) and models using all entities, getting higher scores with the first setting. The typical tasks on which local coherence models are measured are:

- *sentence ordering discrimination*: where the task is to rank original documents (considered coherent) higher than the same documents with randomly permuted sentences (assumed to be not coherent);
- *sentence insertion*: where the task is to rank original documents (considered coherent) higher than the same documents with only one sentence permuted to a different position (assumed to be not coherent).
- *summary coherence rating*: where automatically generated summary rankings are compared to coherence rankings created by humans.

It is important to notice, how both sentence ordering discrimination and insertion use synthetic data, which could show different characteristics compared to coherence violations found my human readers. Summary coherence rating, however, shows the correlation of the models' performances with human judgement.

The algorithm proposed in Barzilay and Lapata (2008) derives thus automatically this abstract representation for a text, with as the only requirement a syntactic parser and a dataset. Among the weak points of this framework, however, there is the fact that it models only local coherence (patterns of distribution across adjacent sentences) and a data sparsity problem. Moreover, the grid does not keep lexical information about the entities, which could prove useful for assessing text coherence.

Over the years, entity grid model inspired numerous extensions and similar implementations. Filippova and Strube (2007), for example, tried to extend the model using semantic relatedness of the entities but without much improvement. Elsner and Charniak (2011b) extended the entity grid by adding all nouns (not just head nouns as in Barzilay and Lapata (2008)) to the grid and entity specific features such as named entity information or number of entity modifiers in order to stress the saliency of some entities compared to other ones. This extended version of the original entity grid achieved the best results compared to all other proposed extensions. Guinaudeau and Strube (2013) model the problem as a graph using a similar idea to the entity grid, but making it completely unsupervised. Although this graph representation achieved comparable scores, it was not able to beat the original model. Mesgar and Strube (2016) also follows a similar intuition by using a graph-based approach to modeling lexical coherence. Li and Hovy (2014) used a distributed representation of the sentences in the document with Recursive and Recurrent NN, however without using the entity grid representation. Nguyen and Joty (2017) presents an approach that uses the entity grid as input to a Convolutional Neural Network (CNN), with the advantage of being able to model long-term transitions compared to the original grid model. Mesgar and Strube (2018) propose a coherence model based on a combination of Recurrent Neural Networks and CNN and find it found useful for essay scoring and readability assessment. The current SOTA on the benchmark tasks for text coherence is achieved by Moon et al. (2019) which uses an LSTM sentence encoder and combines the entity grid representation with other features such as discourse relations.

While most research on entity-based coherence models is based on text, different works have also shown that these models can be useful also for other genres, such as dialogue (Purandare and Litman, 2008; Elsner and Charniak, 2011a). Elsner and Charniak (2011a), for example, show that the local coherence models off the shelf (among which the original entity grid and the extended one from (Elsner and Charniak, 2011b)) transfer well also to human-human conversation (Elsner and Charniak, 2011a). Gandhe and Traum (2008) finds that ordering tasks correlate well with coherence perception in dialogue. More recently, Joty, Mohiuddin, and Nguyen (2018) proposed an extended version of local coherence models based on entities for modeling threads on forums, emails and other genres of asynchronous conversations.

As we have seen throughout this section, research on entity-based coherence models is still active to this day, with various extensions and applications for different tasks. The models presented here, however, are mainly aimed at modeling local coherence, that is the surface features that make a text locally connected. Poesio et al. (2004) notices that local coherence models must be supplemented with other factors of coherence, besides entity-based ones. Additionally, in Grosz and Sidner (1986)'s approach, the global coherence of dialogue was rather give by the intentional level, compared to the attentional one. In the next section, we will introduce Dialogue Acts theory, an approach capturing the intentional phenomena of dialogue.

2.3.2 Dialogue Acts

As we've seen in Sections 2.2 and 2.3, the *intentional* aspect is deemed essential for coherence in dialogue across different approaches. In particular, in various theories this aspect is linked to the *global* coherence of dialogue, that is to the overall structure of the conversation (Grosz and Sidner, 1986). However, while thematic phenomena, being more linked to the surface form, are somewhat easier to define; the same is not true for phenomena connected to the intentional level. Being more connected to the *implicit* meaning of utterances, the intentional aspect can be harder to define, and thus approaches might consequently be quite different and at the same time prone to confusion with one another, as we will see throughout this section.

The concept of Dialogue Acts (DAs) has evolved from the one of Speech Acts, originally formulated within the field of philosophy of language (Austin and Urmson, 1962; Searle, 1965; Austin, 1975). Speech acts theory posits that utterances have multiple levels of meaning. The first one, the locutionary level, represents the surface literal form of the utterance, somewhat similarly to the linguistic level of Grosz and Sidner (1986). The second one, the *illocutionary* level, indicates the purpose behind a given utterance, such as offering to do something or thanking someone. The final level, the perlocutionary one, represents the intended effect should have the utterance has on the speaker, such as triggering specific feelings or thoughts. In the classic example "Can you pass the salt?", the locutionary level simply asks about the ability of a person to perform the action of passing the salt (if interpreted only in this way, the other participant could answer "yes", without passing the salt). The illocutionary level, on the other hand, identifies this utterance as a request to pass the salt: "Please, pass me the salt". The perlocutionary level is the actual effect of the other person passing the salt. Importantly, illocutionary acts can be further categorised into groups, such as the one of *commissives*, where someone commits or offers to do something, or *directives*, where someone tries to make another person do something, for example with a request or an order. Over the years, it has been indeed the notion of illocutionary acts, the one which inspired most research and extensions.

The notion of Dialogue Acts is indeed derived from the illocutionary level of Speech act theory. Although the concepts of Speech and Dialogue Acts are considered synonymous in some approaches, they are different in various regards (Huang, 2017, Chapter 19). While Speech Act theory is more theoretical, being formulated within philosophy of language; Additionally, according to Huang (2017, Chapter 19), the two theories are different in other fundamental aspects. Compared to Speech Acts, focused on verbal behaviour, DA theory encompasses also non-verbal behaviour DA theory includes also *non-verbal* behaviour. Moreover, while in Speech Act theory utterances have only one associated speech act, in DA theory utterances can have multiple associated DAs, i.e. they can be *multifunctional*. Indeed within DA theory utterances are segmented into functional segments, described as a "a minimal stretch of communicative behaviour that has a communicative function (and possibly more than one)" (Geertzen, Petukhova, and Bunt, 2007). Furthermore, Speech Acts are considered in isolation, while DAs are explicitly *contextual*, that is they are dependant on the relations with the context for interpretation, and are connected by a more articulate structure. Last but not least, DAs have a connected interpretation in terms of *information state updates* of dialogue participants. Within the latter theory, DAs have indeed a semantics defined by their functions as update operations of the participants' knowledge status (Poesio and Traum, 1998). This notion, connected to the one of mental models (Dijk, 1981) previously introduced, is also at the basis of research on grounding (Clark and Schaefer, 1989). In grounding approaches, participants to a conversation are assumed to share a common ground, that is a set of mutual knowledge available to all speakers in that moment. Computational models (Traum, 1994) of grounding, explored especially for task-oriented dialogue, also crucially rely on the use of acts to model the dynamic updates of the common ground.

Besides the difference between DA and Speech Act theory, another common confusion among different approaches to the intentional aspect, is the one between DAs and the intents used in task-oriented dialogue approaches, as mentioned in the Terminology section (see 2.1.1 on *intents*). Although both approaches rely on a taxonomy of possible functions that participants' utterances express in a conversation; DAs can be considered a more *abstract* version of intents, capturing more high-level and less practical purposes. Intents are typically domain-specific, while DAs are *domain-independent*. Moreover, intents are usually bound to specific arguments, following a predicate-argument structure (Marinelli et al., 2019); while DAs are not. In the example of the utterance 'Can you book the tickets for the movie?", the associated intent `book_tickets` has a structure composed of the predicate "book" followed by the argument "tickets", while we using a DA taxonomy we would simply consider this utterance a *request*. This dependency from specific arguments in the taxonomy design, makes intents much more domain-dependant compared to DAs.

As introduced in Section 2.3, one possible way of categorising approaches to Computational Discourse and Pragmatics is according to their methodology . This is especially true for the area of research of Dialogue Acts (Horn and Ward, 2004, Chapter 26). On one side, we have the *inferential* approaches, relying on belief logic and inferences (Cohen and Perrault, 1979; Allen et al., 1996); on the other, we have *probabilistic* approaches relying on Machine Learning models. Inferential approaches tend to rely on very rich and expressive formalisms related to the connection of dialogue acts to a specific plan, which however could prove difficult to apply in the real-world. On the other hand, probabilistic approaches tend to be rely on more shallow formalisms and rather let models learn the intended behaviours directly from the data, without imposing any particular interpretation of the phenomenon studied.

Due to such reasons, while the inferential approaches were more popular in earlier years; probabilistic ones are definitely the most popular ones in current days. In this thesis, while taking inspiration from the intuitions at the basis of inferential approaches, we focus on data-driven probabilistic approaches.

Dialogue Act taxonomies Being empirically-based, the history of DA theory is inherently connected to the development of different DA taxonomies over the years. Importantly, DA taxonomies were usually developed with a bottom-up approach, that is by examining a large set of dialogues and attempting to create a categorization that would capture the conversational behaviour in a specific dataset. One of the earliest examples of such efforts is the HCRC MapTask scheme (Anderson et al., 1991), developed for a corpus composed of conversations where participants were asked to complete a task involving a map. The MapTask scheme makes a distinction between *initiating moves*, with an action initiation component (examples are questions, giving instructions), and *response moves*, with more of a reaction component to a previous DA (examples include answers to a previous question, acknowledgments of previous statements etc.). This scheme was developed for a specific type of task-oriented conversations and thus could not be directly applied to non-task-oriented conversations.

The Discourse Annotation and Markup System of Labeling (DAMSL) (Core and Allen, 1997), represents instead an example of taxonomy developed for capturing non-task-oriented conversations behaviour, specifically the ones part of the Switchboard Dialogue Act corpus (Jurafsky, 1997; Stolcke et al., 2000a). Inspired by previous work (Allwood, Nivre, and Ahlsén, 1992; Allwood, 1995), the DAMSL tagset makes a distinction between the *forward-looking* function and the *backward-looking* function of utterances. Similarly to the MapTask scheme distinction between initiating and response moves for task-oriented dialogue, the DAMSL distinction captures the relation of DAs to the previous and following context. While backward-looking DAs are those that respond to a previous DA, such as agreeing or rejecting a proposal, answering questions or requests or signaling non-understanding; forward-looking DAs are those that have an intended effect on future DAs, such as requests of information, committing to a future action or statements. Although an improvement compared to previously proposed tagsets, DAMSL has also issues, related for example to the fact that there is no hierarchy among categories resulting in a flat taxonomy, making it hard to find the similarities among different tags.

The ISO 24617-2 standard (Bunt et al., 2010; Bunt et al., 2020), the latest internationally accepted standard for DA annotation, represents an attempt to create a unified taxonomy both domain- and task-independent. Compared to DAMSL, the ISO standard taxonomy is hierarchical, allowing to group the various DA categories according to their similarities and differences and therefore making it more easily mappable and flexible for different requirements (for example it would be possible to map a given utterance to different levels of the tree, or to add additional leaves in case the designers would like to capture more nuances).

Dialogue Act tagging The task of automatically recognising the DA of a given utterance is known as DA tagging and the models trained to perform such a task are commonly known as DA taggers. As a task, DA tagging has been mainly modeled in the literature as text classification, using techniques such as SVM (Quarteroni and Riccardi, 2010) or neural models (Lee and Dernoncourt, 2016), and sequence labeling, using for example HMM (Stolcke et al., 2000b), CRF (Quarteroni, Ivanov,

and Riccardi, 2011) or more recently neural models (Kumar et al., 2018). Importantly, among sequence labeling approaches, we distinguish those aimed at simultaneously segmenting the utterance into multiple DAs and classifying them (Quarteroni, Ivanov, and Riccardi, 2011; Zhao and Kawahara, 2018), from those simply aimed at classification (Ji, Haffari, and Eisenstein, 2016).
Another important difference is the one between *offline* and *online* approaches. While offline approaches frame the task of DA tagging for an entire conversation at once, and therefore for tagging the DA of the utterance at timestep t_i consider both previous utterances (for example the one at timestep t_{i-1}) and the following ones (for example the utterance at timestep t_{i+1}). Offline approaches (Ji, Haffari, and Eisenstein, 2016; Kumar et al., 2018) can be useful for analysing the structure of conversations, however for online use, for example within a CA, online approaches (Bothe et al., 2018), which consider only the conversation history up to the present point, are the only ones that can be used.

In this thesis, we explore probabilistic approaches to DA theory to study the intentional aspect of coherence in dialogue. However, such approaches mostly rely on a linear, flat representation of DAs. While a linear representation focuses on the temporal dimension of dialogue, it can also be regarded as somewhat simplistic, if we consider the long-range dependencies among multiple DAs that can exist in human conversation. Other approaches, on the other hand, investigated non-linear, structured representations of DAs to capture such dependencies. Traum and Hinkelman (1992), for example, propose a DA theory based on multiple levels, which include larger DAs (e.g. argumentation acts) composed of sequences of other types of (core) DAs. Such dependencies have especially been explored in the context of Discourse Representation theory (Poesio and Traum, 1997) and Segmented Discourse Representation theory (Asher and Lascarides, 2003; Lascarides and Asher, 2008).The usefulness of such non-linear DA representations is particularly evident for multi-party and asynchronous conversations. In this case, rather than relying only on the temporal dimension, graph approaches have been explored for capturing long-range dependencies across different parts of the conversation (Joty, Carenini, and Lin, 2011; Afantenos et al., 2015). While structured approaches to DA representation are certainly interesting, they have been comparatively less researched compared to linear ones. Most current probabilistic approaches to DA tagging are based on linear representations, which we use in this thesis.

DA theory has been used and it is still used nowadays for various tasks, such as understanding conversational structures and as a key component of several CA applications. As we discussed, DAs have also been repeatedly linked to dialogue multi-turn coherence (Allen and Perrault, 1980; Grosz and Sidner, 1986). However, although DA and entity-based phenomena have been linked across various theories of discourse structure in the literature, such as mental models (Dijk, 1987), the relation between the two aspects has not been extensively investigated in the literature, especially with data-driven methodologies. In this thesis, we explore how to combine DA and entity-based theories using (whenever possible) a probabilistic approach for both coherence modeling in dialogue and for creating different modules of a open-domain CA designed to maintain coherence across multiple turns of a conversation.

2.4 Summary

In this Chapter, we provide a general overview of the background work related to this thesis. First, we introduce key concepts and a brief history of conversational AI research within the perspective of coherence. In particular, we discussed the dichotomy between task-oriented domain-dependent modular CAs and open-domain non-modular chitchat dialogue models. While the former ones crucially rely on domain-dependent Meaning Representation structures usually composed of intents and slots, as an abstract representation of the interaction; the latter rely directly on the surface forms as input for an end-to-end non modular Machine Learning models, which allows them to be trained on any domain. From the perspective of maintaining multi-turn coherence, both have their advantages and disadvantages. On one hand modular CAs are generally able to maintain coherence by remaining in a fixed set of domains and handcrafting the dialogue state and MR by making expert assumptions about the domains considered. However, this solution is not scalable when progressing towards designs able to handle larger number of domains through open-domain more flexible solutions. On the other hand, chitchat models while representing a more flexible alternative, have serious issues with maintaining multi-turn coherence in conversation. At the end of our overview of conversational AI, we discussed current approaches to dialogue evaluation. In particular, we saw the reliance of the field on time-consuming human evaluation and the current void of standardised automatic evaluation methodologies that would be able to effectively compare both task- and non-task-oriented models.

The second part of our overview was dedicated to how the concept of coherence has been investigated in theoretical Linguistics. We saw how the field distinguishes between global and local coherence of discourse, and the reliance of the concept of coherence not on the text itself, but rather on the interpretation of the receiver of a message. Within the field, we highlighted the importance of two main strands of research on coherence: one studying thematic phenomena and more investigated for written text, and another one studying the intentional aspect of discourse, more studied for dialogic interactions. Although often considered independently, we discussed how these two phenomena are actually connected if we consider the notion of mental models.

In the third part of the Chapter, we reviewed computational approaches to coherence. We saw how approaches to coherence in Computational Linguistics can be divided into those more concerned with written text, such as informational and entity-based ones, and those more interested in dialogue, i.e. intentional approaches. In particular, we gave a brief introduction to Grosz and Sidner (1986)'s inferential approach to the discourse structure of dialogue, which posits the crucial interaction of intentional and attentional (related to the thematic level) structures for defining conversational coherence. Then, we provided a closer overview to entity-based approaches to coherence which investigate the thematic aspect, such as the data-driven entity grid model which has the advantage of relying on a weakly supervised data generation methodology. Finally, we introduced Dialogue Act theory which studies the intentional level, discussing the distinction of the approach compared to Speech Acts and intents and the definition of DAs in terms of information-state update of interlocutors.

In the next chapters, we delineate the contributions of this thesis, which rely on intuitions from entity-based and DA theories for modelling coherence in open-domain conversation and for designing different modules of a domain–independent coherent CA.

Chapter 3

Roving Mind: an open-domain CA designed for coherence and engagement

This Chapter[1], presents Roving Mind, the open-domain conversational agent (CA) we designed for the first edition of the Amazon Alexa Prize (Ram et al., 2018), a university competition to design CAs able to talk about popular open-domain topics (i.e. sports, music), while maintaining coherence and user engagement.

First, we describe Roving Mind's modular architecture, which relies on a Meaning Representation (MR) language based on functional units, i.e. structures based on open-domain Dialogue Acts and entities. Using such MR structures, the CA architecture is designed to be at the same time modular and to maintain coherence and engagement across multiple turns in open-domain conversation, by keeping track of DAs and entities previously used in the interaction.

Then, we present results from a series of experiments performed during the competition's semifinals. The results of our experiments highlight the importance of system's directed strategies in open-domain conversation for user engagement.

3.1 The Alexa Prize competition

The Alexa Prize[2] is a competition among university teams, first launched by Amazon in September 2016, with the goal of encouraging open-domain conversational AI research. The first edition of the competition, described in this Chapter, took place between November 2016 and November 2017 (Ram et al., 2018).

The main goal of the competition was to create chitchat dialogue models ("socialbots", in the competition's terms) able to talk *coherently* and *engagingly* with random users on popular topics (such as sports, politics, entertainment etc.). Competing teams had the possibility of testing their models on real Amazon Alexa users. During the competition, users based in the United States could simply open a special Alexa Prize skill and would then be randomly connected with one of the competing socialbots. Users could end the conversation at any time they wanted. At the end of the interaction, users were asked to rate the conversation on a 1-5 Likert scale, according to how likely it was they would want to speak to that specific socialbot

[1] The Chapter is based on Cervone et al. (2017).

[2] The Amazon Alexa Prize (`https://developer.amazon.com/alexaprize`) is currently at its third edition.

again. The Grand challenge posed by the competition was to speak to a user coherently and engagingly for 20 minutes. While it was not expected for a team to win the Grand challenge within the first edition (which hasn't been won yet so far), the best university team would still win the first year's competition. While users rating were the main variable for evaluation, the Alexa Prize team developed several additional metrics to evaluate socialbots: Coherence, Engagement, Conversational User Experience, Domain Coverage, Topical Diversity and Conversational Depth (Ram et al., 2018).

Fifteen universities were selected to participate in the first year, with teams composed by university students, advised by a faculty advisor. Semifinals took place for 6 consecutive weeks between July 1 and August 15 2017, after which three teams were selected to compete in the finals: the two teams which on average had the higher user ratings throughout semifinals and a third wildcard team chosen by the Alexa Prize team.

As we will see throughout the Chapter, the competition posed several challenges. One of the most important ones is the *lack of data*. While nowadays random conversations are available in large number over the internet, it is much harder to find resources for what could be defined as good conversations. Furthermore, conversations annotated with quality metrics are even harder to find. Additionally, the lack of data is also relevant in terms of available open-source resources that can offer good performance on spoken dialogue for common NLP tasks (i.e. coreference resolution, parsing etc.).
Another challenge posed by the competition was the fact that compared to task-oriented conversation, chitchat open-domain models have been *less investigated* by the research community. Notions such as what makes a person engaged in an open-domain conversation, are still not well understood.
The Alexa Prize also involved the challenge of *scalability*, both from an engineering perspective, by being able to support up to hundreds of conversations at the same time without latency, and in terms of domain coverage, i.e. designing able to handle virtually any popular topic, including those whose knowledge might be very recent, such as daily news.
Finally, interacting with *real-users* is a challenge per se. The virtually unbound range of people that could interact with out model meant that no assumption could be made in terms of social characteristics of the users' pool (such as age, gender, interests, opinions etc.). Moreover, competing socialbots were required to comply with different rules in order to keep the interaction to a polite level, which required creating offensive speech detection modules and implementing specific strategy to avoid responding to profanities or inadvertently discussing inappropriate content (examples of dangers of an uncontrolled behaviour in this regard would be answering to questions such as "Should I kill myself?" or "Should I buy those stocks?"). This last challenge highlights the need for exercising some control over the generated responses. As we will see throughout the Chapter, our model was designed keeping these challenges in mind.

3.2 Introduction

The structure of human dialogue is closely connected to the underlying goals of the participants to the conversation. Therefore, having a model of the speakers' intentions currently represents the only way to have a form of control (as well as understanding) on the decision–making process of the machine in human–computer conversational interactions. Additionally, having an explicit model of participants' intentions in the conversation can be useful to maintain the contextual coherence of the responses.

Traditional task–oriented approaches to dialogue generation (Young et al., 2013) solve this issue by circumscribing the application of the dialogue system to a specific domain (e.g. restaurant reservations); and usually rely on carefully hand–crafted parameters, such as the state and action space for the domain. Such constrained settings allow the system to be in good control of the interaction at each turn, representing the user input through intents (the user's goals) and their associated slots (relevant concepts for the task mentioned in the conversation); and allow using modular architectures (Wen et al., 2016a). However, given the dependency on either domain–specific data or hand–crafted features, this approach is difficult to scale to open–domain conversations.

A different line of research emerged in recent years which replaces the modular pipeline of traditional systems with a single model (e.g. sequence-to-sequence recurrent neural networks (Serban et al., 2016d)) trained on a large collection of dialogues. Such models directly generate the most probable response to a user utterance. Since the most probable responses tend to be the most generic and dull ones, while the most diverse and interesting ones are naturally much more sparse (Li et al., 2016a), such systems are hardly engaging. Due to the fact that sequence-to-sequence models learn a direct association between string sequences, there is no need for hand-crafted features or explicit representations of user intents. However, this lack of explicit representations makes them difficult to interpret and control; and hard to generate responses at the same time coherent with the conversation context (due to difficulties in capturing and maintaining the long-term context) and engaging for more than a few turns (Li et al., 2016a). Typically, such models end up in loops of dull responses after a short conversational context (Li et al., 2016c).

Sequence-to-sequence models have the benefit of being domain-independent given the right dataset. Traditional modular domain-specific task-oriented architectures have the benefit of control and interpretability of their decision–making. However, the design of open-domain and controllable conversational agents still remains an open issue. Additionally, user engagement in human-machine interaction, crucial for open-domain conversations, is a rather new research area addressed by very few (Yu et al., 2016a). Moreover, regardless of the approach, there is a lack of large and high-quality data to train domain-independent conversational systems. In Roving Mind (RM) the advantages of both approaches are joined into a system which is at the same time modular and domain–independent, with a specific submodule to address user engagement. In particular the characteristics of our architecture (shown in Figure 3.1) are:

- *modularity:* in RM the behavior of each module is not affected by others. This allows firstly to interpret and control the decision–making process of each submodule and secondly to easily combine and replace rule–based approaches (currently required by the lack of appropriate training data) with data–driven ones. Moreover, the modularity of the approach allowed to add an *Engagement* submodule to our architecture with the task of keeping the user engaged in the interaction. RM's architecture can therefore be described as a balance between human–expert design (for the global architecture) and data–driven approaches (locally).

- *domain independence*: we represent both user and machine utterances through domain-independent *functional unit* (FU) structures, consisting of a list of *entities* (slots in task-based approaches) and a *dialogue act* (DA) structure representing the goal of that FU (intents in task-oriented approaches). The reliance on this domain-independent abstract representation of the conversation, stored in the dialogue history of RM, allows the system to maintain *coherence* across multiple turns of the conversation, also through the use of single- and multi-turn domain-independent strategies. Our system also supports queries to various domain-independent knowledge bases (KBs), such as news, useful for user engagement.

To the best of our knowledge RM represents the first attempt to build a modular, open-domain dialogue system architecture, with an explicit representation of user intents and the entities mentioned in the conversation and a fully automatic module for user engagement.

After detailing the system architecture in section 3.3, in section 3.4 we analyze the results of the Alexa Prize semifinals, during which RM was tested and rated by a large group of Amazon Alexa customers. While our initial setting during the competition was completely open-topic, we report the results of a set of experiments performed in order to optimize user ratings using features aimed at making the conversation more driven towards the topics where we had the highest coverage, without changing our architecture. We have observed a significant correlation between user ratings and *cumulative sentiment* – a combination of sentiment and DAs revealing sentiment of the user towards a topic – which could be used in the future as a potential error–signal for user engagement. Finally in section 3.6 we draw on the conclusions of our work.

3.3 System design and architecture

In this section we will go through a coarse-grained description of RM's scalable infrastructure pipeline (section 3.3.1) and its KBs (section 3.3.2). Subsequently, we describe the design and implementation of the Spoken Language Understanding (SLU) pipeline in section 3.3.3 and the Dialogue Manager (DM) and Natural Language Generation (NLG) modules in section 3.3.4.

3.3.1 System pipeline

The input of the system pipeline is the data structure coming from the Alexa Skill Kit Automatic Speech Recognition (ASR) module. The first element of the pipeline is the client application, whose main goal is to address predefined intents of the Alexa Skill Kit and redirect everything else to the server, which will generate the response. The client application was implemented using an AWS Lambda function, as in common practice in Alexa Skills implementation.
The server exposes a RESTful API to communicate with the client. Its first task, once the request is received, is to interact with a database to retrieve the dialogue history collected during the interaction. After this information is retrieved, the server runs through the SLU pipeline to create the list of user FUs and then passes the results to the DM module, which produces the output machine Dialogue State. During the DM phase, AWS RDS is used to manage the KBs. These KBs are constantly updated through independent scripts running on separate EC2 instances. The last step of the pipeline is the NLG module, which transforms the generated list of machine FUs

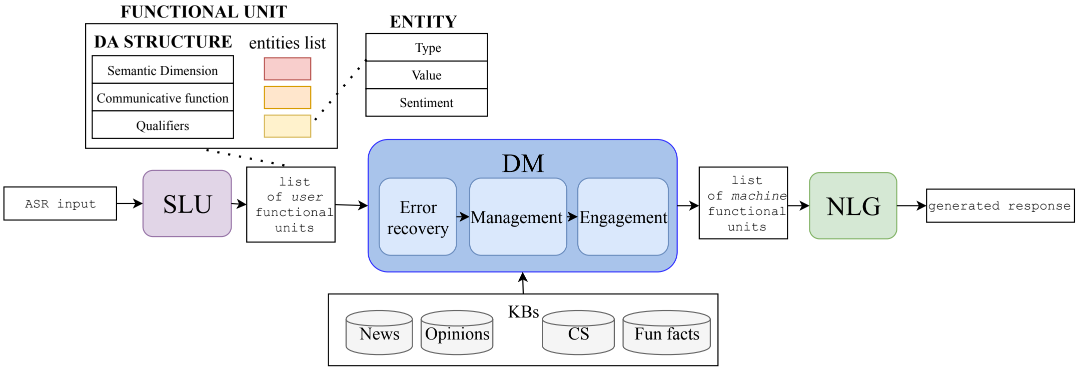


FIGURE 3.1: In our architecture the Dialogue Manager (DM) takes as input a list of *user* functional units (FUs) (i.e. consisting of an associated dialogue act (DA) structure and a list of entities) created by the Spoken Language Understanding (SLU) module from the Automatic Speech Recognition (ASR) output. This list is processed progressively by each of the three DM submodules to create a list of *machine* FUs (also retrieving information from a set of knowledge bases). This list is then passed to the Natural Language Generation (NLG) module, where the corresponding template for each machine FU is generated and joined together to produce the final response.

into a string output. This output is finally returned to the client application, which sends the text to the Alexa Text-To-Speech (TTS) provided by the Alexa Skill Kit. The server API was implemented using an AWS Elastic Beanstalk, which eases the scalability of the system and makes it more robust and reliable, while the database containing users data was implemented with an AWS DynamoDB, which allows high read–write throughput.

3.3.2 Knowledge Bases

Roving Mind has a rich set of content sources to keep the user engaged and to maximize the coverage of the topics provided by the user. In its final version, the system features News, external Opinions, Fun Facts and a knowledge graph to perform Commonsense reasoning. Factual question answering (Q&A) is performed querying the EVI answer engine.

News The News database is based on two main scripts that run continuously in background, updating the KB every day. The first script downloads the news from the Washington Post API and stores them in an AWS Aurora database. In this phase news are clustered together, according to the 'section' parameter provided by the API. An affinity propagation(Frey and Dueck, 2007) clustering technique is used to group news efficiently and enable a fast search for related news on the same domain. The second script processes the downloaded news articles for both generating article summaries and linking news to a specific keyphrase. To be able to link news articles to keyphrases, sentences in an article are represented as lists of triplets in the form <subject,predicate,object>. The approach was adapted from the ClausIE framework (Del Corro and Gemulla, 2013); and allows to classify each sentence as relevant or not to a keyphrase.

Opinions Opinions are extracted from news articles using Conditional Random Fields (CRF) models trained on the Parc dataset (Pareti, 2012). This model extracts quote spans and attribute them to their authors. The obtained quote spans are processed for keyphrase extraction and triplet generation. Each quote is then classified as relevant or not to the keyphrase according to the keyphrase's position in the triplet. Finally, in the KB, the quote is represented with the disambiguated speaker of the quote, the span and the main keyphrase.

Commonsense For our third KB, representing commonsense, we use the data from ConceptNet 5 (Speer and Havasi, 2012) and integrate it with data regarding entities connected to the relation "occupation" from Wikidata. With the collected data we created a graph. On the resulting graph we defined a set of queries that are performed online during the execution of commonsense strategies (see section 3.3.4).

Fun Facts Fun facts were collected online[3]. Similarly to news, we extract keyphrases and generate lists of triplets in order to trigger fun facts related to a given keyphrase mentioned by user.

3.3.3 Spoken Language Understanding

Spoken Language Understanding (SLU) for task-based systems is traditionally approached as identification of the intent (the purpose behind the utterance) and detection of its arguments (slot-filling) from a user utterance (Tur and De Mori, 2011). In a task and domain-specific SLU the set of intents and slots is usually predefined; in a domain-independent SLU, on the other hand, it is not the case.

Since the range of user intents in an open-domain dialogue is virtually unbound, in RM intents rely on domain-independent dialogue acts (DA) (e.g. questions, requests), and are further specified with an additional set of qualifiers. Intents cover various aspects of a conversation, as they can represent information exchanges (e.g. questions and points of view expressed by the user), social obligations (e.g. thanking, salutations) and action discussions (e.g. requests, orders). In addition, a user utterance can contain one or more intents: e.g. greeting and an action request; thus, an utterance needs to be segmented with respect to these intents. The recently accepted international ISO standard for DA annotation – DiaML (ISO 24617-2) (Bunt et al., 2013) – is a good choice to address these necessities. An utterance is segmented into functional units (FU) and each unit is attributed a DA that consists of semantic dimension (e.g. Social Obligation) and communicative function (e.g. Thanking). Thus, the specification naturally lends itself to the design of open-domain systems. In RM, the ISO specification was adapted, following in the work of (Chowdhury, Stepanov, and Riccardi, 2016), with emphasis on the communicative aspect rather than the semantic dimension.

Similarly to intents, in a domain-independent non-task-based setting the range of possible slots is unlimited. In RM, we replace the notion of slot, predefined in task-based systems, with the one of entity. An *entity* is defined as any user-provided content that can function as a topic of a conversation – Named Entities, keyphrases, etc. – or that triggers specific dialogue functionality – News, Opinions etc. Entities are not predefined, but classified into types according to their association to our KBs. Hence, in RM each FU is represented by an intent, that is a DA structure composed by a DA (identified by semantic dimension and communicative function) and its qualifiers, and a set of entities extracted from the span of that FU (see Table 3.1).

The rest of the section describes the system pipeline for the segmentation into functional units and identification of user intents, and entity extraction.

[3] `https://www.thefactsite.com/`

Pre-processing: The SLU pipeline starts with the pre-processing of user input. First, the ASR output is case-restored to improve the detection of Named Entities. Then, the input is split into sentences using the state-of-the-art sentence-boundary detection algorithm of (Read et al., 2012), without prosodic information (not available). Each detected 'sentence' is used to extract information for a FU structure.

Intent Identification: The first step in intent identification is DA tagging. For DA tagging we consider three semantic dimensions from DiaML (Bunt et al., 2013): (1) Social Obligation dimension that addresses basic social conventions such as greetings, (2) Feedback dimension that addresses user's feedback regarding the previous machine statement, and (3) Task dimension that addresses user's actions such as requests for information or action.

Given that each of the semantic dimensions consists of specific set of communicative functions, the DA tagger is implemented in two separated steps. We first identify the dimension of the DA, and then use dimension specific classifiers to identify the communicative function. For the Social Obligation dimension, the communicative functions are assigned using a lexicon-based system. Functions from this dimension capture salutations, apologies and thanking, following the ISO standard definition. For the Feedback dimension, on the other hand, there is only one communicative function which is simply called 'Feedback'. The Task dimension is more complex and, in RM, consists of 16 communicative functions, which correspond to the "General Function" tags of the ISO standard scheme, where some of the lower level tags are not considered. These tags are arranged in a tree structure: at the first level there is a separation between Action discussion, which captures specific dialogue-related actions by the user (e.g. Stop, Repeat, Start Quiz), and information-transfer, which captures an information exchange (statement or question) between the user and the system. Lower levels provide more details about the interaction, specifying for example whether it is an Information providing or seeking (for Information transfers) or whether the action performed by the user is Directive or Commissive (for Action Discussions).

DAs are further analyzed to determine sentiment polarity, factual information type and subjective information type (i.e. qualifiers). The sentiment polarity qualifier records the attitude of a user towards the entities (or topics) present in FU, sentiment analysis is performed using a lexicon-based analyser following (Alistair and Diana, 2005). The factual information type qualifier is similar to the Expected Answer Type in Q&A and records if the provided/requested information is about a specific time, place, reason, entity, etc. The subjective information type qualifier, on the other hand, categorizes non-factual information such as opinions or personal information (e.g. "My favorite sport is football"). The qualifier assignment is handled with lexicon and rule-based systems.

Finally, a lexicon based functionality classifier determines whether the user dialogue act contains a request for a specific functionality of the system: such functionalities are divided in (1) dialogue control functionality: built-in intents such as Stop, Repeat and Continue; and (2) Content Request functionality: requests for one of the featured contents of the system such as News or Fun Facts.

DA taggers were trained using multiple corpora – Switchboard (Godfrey, Holliman, and McDaniel, 1992), AMI (Carletta, 2006), the HCRC Maptask corpus (Anderson et al., 1991), the BT Oasis Data and the VerbMobil2 (Alexandersson et al., 1998). Following the approach of (Fang et al., 2012a), the DA annotation of these corpora was mapped to the modified tag hierarchy of RM. We use a combination of CRF and Support Vector Machine classifiers for tagging; as it achieved the highest performance on the withheld 10% of data.

Entity Extraction and Linking: RM extracts Named Entities and keyphrases which are later linked to nodes in the KBs. Keyphrase extraction is performed via a simple Noun Phrase Chunker. After an initial filtering step to remove phrases containing 'taboo' words (specified by a stop-word list), extracted chunks are linked to KB nodes using the approach of (Chisholm and Hachey, 2015). Commonsense nodes extraction is performed via a Breadth First Search on the dependency parsing tree for the user utterance, starting from the root, down to the leaves. All paths are explored, and sibling nodes sharing common properties become candidate nodes. When all candidates are collected, each of them is verified against the label present in our commonsense KB and, if a match is found, it is stored in the FU structure. The list of *user* FUs generated by the SLU is then passed to the DM.

3.3.4 Dialogue Manager and Natural Language Generation

The Dialogue Manager of Roving Mind is modular (see Figure 3.1), and consists of three main components, each addressing a specific dialogue function: the *Error Recovery*, the *Management* and the *Engagement* modules. The design of these different components was inspired by DA theories, which divide DA categories among backward- and forward-looking, according to their relation to the conversational context (Allwood, Nivre, and Ahlsén, 1992; Jurafsky, 1997). The sequential design of our Dialogue Manager follows thus an ideal abstract intentional structure of the conversational turn, where we assume a progression from backward-looking to forward-looking DAs. When constructing the response for the user, RM goes first through the Error Recovery component, with the possibility of generating backward-looking DAs such as signal-non-understanding; then, the response generation phase progresses through the Management, with a different class of backward looking DAs such as answers, or addressing previous requests; afterwards, the response generation reaches the Engagement module, , which presents a set of associated forward-looking DAs, such as statements or questions.

Hence, in our design, these three components work in a sequence, each taking as input the output of all previous modules, i.e. the list of user FUs (Bunt et al., 2013), the list of generated machine FUs so far, and the entire *history* of the conversation in terms of strategies, entities and FUs selected in previous turns. Each DM module can generate zero, one or more new FUs and add them to the list of machine FUs which will then be passed to the NLG to generate the final turn. The only difference between user and machine FUs is that each machine FU is associated with a specific strategy and the relevant content, retrieved from the KBs to fill the NLG template for that strategy. In RM, the list of current user FUs, the list of current machine FUs and the previous dialogue history constitute the *dialogue state*.

Internally, all the three submodules are based on similar components, most of which are currently rule–based due to the lack of training data, but are easily replaceable by data–driven modules thanks to the modularity of the architecture. There are two main phases for each module: *machine FU creation* and *content retrieval*. In the first phase, taking as input the list of current user FUs, the list of current machine FUs and the previous dialogue history, the system can (but it's not mandatory) create a new machine FU with its related machine DA structure, entities and strategy. Starting from this partially formed FU, in the second step the system tries to instantiate the selected strategy by retrieving the content required by the strategy through queries on the available KBs. If the content retrieval phase fails the control is returned to the *machine FU creation* step which can create another machine FU selecting a different combination of strategy/entities/DA structure.

Overall we designed more than 60 dialogue strategies (mapped to DAs), each one selectable by one or more modules. Strategies were grouped among error–recovery, management and engagement strategies. As mentioned, some of our strategies rely on queries on our KBs (e.g. the ones providing relevant news summaries given an entity of type Keyphrase).

Probably the most interesting set of strategies were the ones relying on *commonsense*, which could be selected in case an entity of type Commonsense node was found. The commonsense strategies were created to create engaging content (such as RM statements or follow-up questions) coherent with the previous conversation and with the previous machine FUs generated so far for the current context. Firstly we created a list of categories of nodes relevant for open–domain conversation (e.g. human activity, entertainment event) from our commonsense KB. Each category was identified by a set of commonsense relations that nodes belonging to the category had to fulfill. Then we created a dataset of conversational natural language templates each one based on one of the selected categories and associated to a list of additional queries on our commonsense KB. Each template was tagged with an utterance type (question or statement) and a DA structure (e.g. its communicative function etc.) as described in 3.3.3. Given as input from the DM a list of commonsense node entities, the DA structure and the utterance type, our commonsense strategies first selects the list of templates matching that DA structure and utterance type. Afterwards, if any templates are found, the strategy verifies that the input nodes belong to the category appearing in the templates and finally, if additional commonsense queries are required to fill the template, it executes them online. At last, if any templates were generated, the most relevant template is chosen (i.e. the one sharing the higher number of commonsense entities with the original), otherwise the strategy fails and the control is returned to the DM.

Given the basic two steps previously described, each DM submodule works in a slightly different way according to its main goal.

Error recovery module The Error Module is the first module of the pipeline to be executed, with a *backward-looking* function, i.e. to address an event that happened in the previous conversational context. One of the main functions of the Error Recovery module is to trigger a strategy for user FUs with low ASR confidence in order to signal non-understanding. We extended the goal of this module to a more generic analysis of user FUs that are not "well–formed" according to "explicit" content (using a dictionary) and personal information (using Regex). This decision is due to the fact that the system handles these cases in a similar manner, filtering them from the FUs to store. The strategies this module can select from are traditional error–recovery strategies (e.g. asking the user to repeat in case of low ASR).

Management module Once the input has been filtered, the output is passed to the Management module, whose main goal is to address the potential requests (including questions) expressed by the user in the last turn, with a *backward-looking* function. Together with the Error module the Management can be used independently from the last module as a Q&A system or a task–based system. The second goal of this module is to update the active entities in the system, by adding any entity the user

mentioned in the input utterance to the entity pool of the current dialogue state. Moreover, if the user utterance has a negative sentiment towards entities introduced by the system during the previous turn, they are discarded. In addition, this module manages the continuation of multi–turn strategies (such as subdialogues) started by the system in previous turns.
In this module, during the *machine FU creation* step each user FU is mapped to a machine FU through a rule-based approach. Afterwards, in the *content retrieval* step, each machine FU is associated to a set of strategies (applied in a sequence) designed to address that specific request (e.g. an entity of type question is sent to the Q&A strategies). If the system detects a user FU that cannot be handled by the system, the *content retrieval* step will generate a machine FU with a strategy informing the user that the required functionality couldn't be handled. If every strategy fails, the system creates an empty machine FU that will be used by the Engagement module to determine the next action and discarded later on. The same approach is applied if the user FU does not contain any requests/questions. Regarding the Q&A strategies, questions detected as factual are redirected to Evi, while for personal questions we use the previously described commonsense strategies.

Engagement Module After all requests have been satisfied the output is passed to the Engagement module, responsible for keeping the user engaged in the conversation, thus with a *forward-looking* function. The idea of having engagement as a module in a dialogue system design was introduced in (Yu et al., 2016a), where engagement is defined as the user's interest in continuing the conversation. The authors shows how having a module which addresses directly user engagement leads to a better user experience. The difference with (Yu et al., 2016a) is that our engagement module is fully automatic and takes into account the machine FUs created by previous modules for the current turn.
The Engagement Module takes as input the utterances generated by previous modules with the updated history of active entities and the history of the conversation, for generating contents coherent with the previous context. In the *machine FU creation* step, the module firstly verifies whether machine FUs created by the previous modules require additional engaging content. If that is the case, entities from the history are sorted according to a series of criteria, which include the types of the available entities, whether they are user-introduced or machine-introduced and their age in the system. Once entities are ranked and the entities to be used is decided, the list of engagement strategies is filtered by selecting the ones compatible with the type of entities. Then the strategies are ranked in a rule-based submodule according to three main features: the current turn index of the conversation (e.g. some strategies work better if they are run in an early/late stage of the conversation), the amount of times the strategy has been used in the previous turns and the ratio between questions and statements in the dialogue history (in order to keep a balance between them when interacting with the user). Finally, in the *content retrieval* step, the selected strategies are executed according to their ranking until one of them succeeds. If all of them fail, a different set of entities is selected and the process is repeated all over again until either a strategy succeeds or the maximum number of attempts is reached, in which case a entity-less strategy is selected.

NLG The last step of the pipeline is the NLG module, which receives as input a list of machine functional units, each one with a corresponding strategy and the related content retrieved from the Contents KBs. For each of these FUs the NLG generates

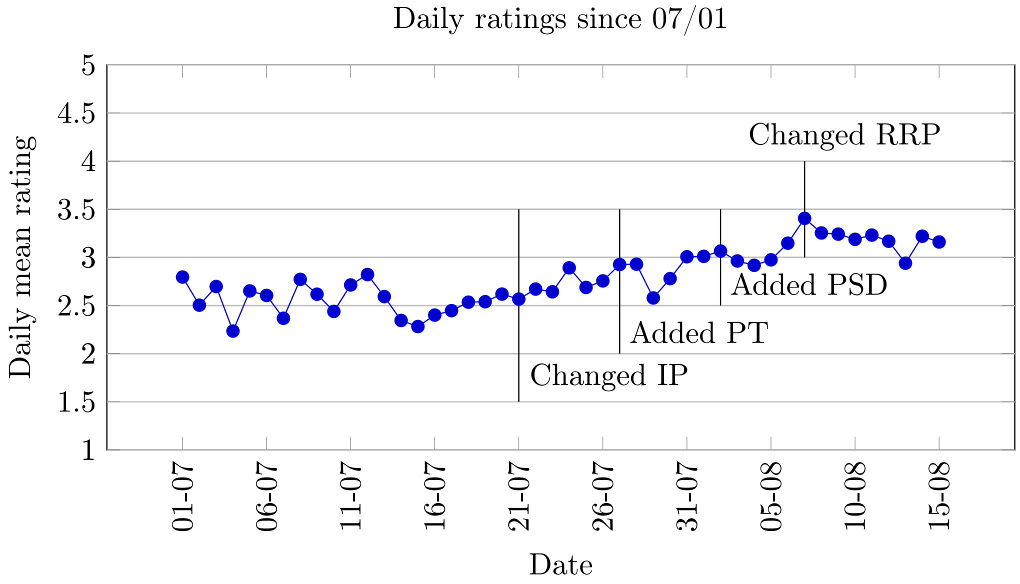


FIGURE 3.2: The figure shows the positive effect of our experiments on our daily average ratings. The daily volume of rating was constant with small daily variations, except for the period between 14th and 17th of July (in gray) were we had an increment of about 289% of rated interactions compared to the daily conversations average.

an output utterance filling the template associated to that strategy with the related content. All the created utterances are then concatenated in the final response and passed to the TTS.

3.4 Evaluation

During the Alexa Prize semifinals, Roving Mind was tested for 6 consecutive weeks by Amazon Alexa users, which at the end of the interaction were asked to rate how likely they were to interact with that socialbot again on a 1–5 scale.
The initial RM dialogue system (active till 21st of July) was completely open – there was no bias towards any domain or topic. In the following weeks, a set of new functionalities was added such that each new functionality was evaluated for a short period of time and compared to the previous system ratings. The dates and results of these changes are visible in Figure 3.2. The details of the experiments are discussed in Section 3.4.1.
The new functionalities evaluated during semifinals are the following: (1) driving conversation towards a specific set of domains/topics via initial prompt, (2) addition of "entertaining" features such as quizzes and tests, (3) system-driven subdialogues, and (4) change to the request rejection prompts.

3.4.1 Experiments

Open Prompt vs Directed Prompt Our first experiment was changing the opening prompt from a very generic "What do you want to talk about" to "We can talk about topics like sports, politics or celebrities. What would you like to chat about ?". This was done after noticing a large number of conversations in which users appeared confused about the features of the system and were expecting a list of possible topics,

System		Time Window		Mean Rating
RM	(baseline)	Jul 15 - Jul 20	(6)	2.44
RM+IP	(+initial prompt)	Jul 21 - Jul 26	(6)	2.75*
RM+IP+PT	(+personality tests)	Jul 27 - Aug 01	(6)	2.89*
RM+IP+PT+PSD	(+predefined subdialogue)	Aug 02 - Aug 07	(6)	3.07*
RM+IP+PT+PSD+RP	(+rejection prompt)	Aug 08 - Aug 15	(8)	3.17

TABLE 3.1: Changes to the Roving Mind system with the time window for data collection (number of days in parentheses) and mean user rating received for the rated conversations within that window. The ratings for each version of the system are compared to the previous one. The statistically significant changes in rating are marked with * ($p < 0.01$).

rather than coming up with their own. Also, the change of opening prompt gave us the possibility of driving the user towards topics where Roving Mind has a high coverage across different strategies. To evaluate the results of the change we drew a comparison with the previous week ratings, as visible in Table 3.1. Results are statistically significant and show that average rating has improved with the simple addition of a driving opening prompt.

Personality Tests and Quizzes Personality tests and Quizzes have been added to evaluate the effect of simple user–entertaining features on user ratings without any modification of our architecture. The five personality tests and the only quiz proposed share the same simple structure, at the beginning the quiz or test is proposed to the user, if the user accepts the offer, the system will go through a set of predefined questions asking the user to select one among multiple choices. The quiz is interrupted by any question and request made by the user. After three or four questions the system provides the result of the test. In the window considered, the personality test is proposed according to the entities introduced by the user or a random one when the user does not introduce any entity. We decided to compare the ratings of conversations during a week with personality tests and during the previous week, when tests were not yet part of the system. Results, as visible in Table 3.1, show a high statistically significant difference between the rating before and after the introduction of quizzes.

Predefined subdialogues The results of the personality tests seemed to suggest a positive effect of predefined multi–turn structures on user ratings. For this reason we created a set of predefined subdialogues where the user is guided through a sequence of coherent prescripted responses and decided to verify their effect on user ratings. The structure of subdialogues is similar to the one of personality tests and quizzes. The main difference is the fact that the system adds a reaction to the user statement, supports a set of predefined question related to each machine turn and according to the sentiment of the user response can choose a different question for the next turn. This setting allows to have a tight control on the user interaction during the subdialogue. We defined a small set of subdialogues that do not require any entity introduced by the user that start at the beginning of the interaction and another set of subdialogues triggered when a set of specific entities was mentioned in the conversation. The results show a boost in ratings as shown in Table 3.1. However, the system achieves the best performances when the user experiences both subdialogues and quizzes together, as visible in Table 3.2.

Changed Request Rejection Prompt Given that in previous experiments guiding the user through the conversation had a positive effect on user ratings, we decided

Dialogue Subset	Mean Rating
no personality test or predefined subdialogue	2.66
personality test **or** predefined subdialogue	3.08*
personality test **and** predefined subdialogue	3.71*

TABLE 3.2: The effect of the activation of predefined subdialogues and personality tests on user ratings, between July, 21st and August, 10th. Each strategy activation setting has been compared to the row above, and ratings with statistically significant differences are marked with * (p < 0.01).

to change the prompt used by the machine in case the requested feature was not supported. Instead of simply saying that the system was not able to satisfy the request, the changed prompt provided alternatives on our system's capabilities. This change had a positive effect on user ratings as visible in Table 3.1.

User Interaction and Ratings: In the course of evaluation users were providing *subjective* ratings, which are expensive to obtain while developing conversational systems. An objective metric that could be used as an online error signal for user engagement towards a topic or a DM strategy is a sentiment score, already computed in RM. Cumulative average sentiment is computed as the average of user FU sentiments combined with the sentiment expressed by DAs revealing the user's sentiment towards a topic (e.g. yes/no answers to topic proposals), sentiment analysis is performed as described in section 3.3.3. Since engagement is also reflected in the conversation length, it is another evaluated objective metric. Analysis of user ratings with respect to conversation length and cummulative sentiment over the duration of conversation reveals statistically significant differences between low-ranked (rating 1-2) and high-ranked (3-5) conversations using t-test and $p < 0.01$. Low ranked conversation had an average conversation length of 9, while high ranked ones of 13 turns. Moreover, sentiment score that considers only DAs yields statistically significant differences as well.

3.5 Discussion

While innovative in different ways, RM also presents limitations, in part due to the short time span of the competition. Similarly to other competitors in the first year, including the winning team (Fang et al., 2017), RM has a modular, rather than an end-to-end non-modular approach, heavily relying on SLU module and knowledge bases for coherence and engagement. According to Ram et al. (2018), having a robust NLU and knowledge bases lead to successful ratings with users. While using the news and Evi as knowledge sources was done by several teams in the first year (also because of the Washington Post and Evi's APIs support), usage of other knowledge sources was more varied. To the best of our knowledge, for example, RM was the only socialbot among competitors of the first year Alexa Prize to utilise a commonsense knowledge base. Sounding Board (Fang et al., 2017), on the other hand, successfully employed an unstructured type of knowledge using Reddit conversations for driving users' engagement.

While Sounding Board and RM architecture appear somewhat similar, given the modular approach and the use of DAs as part of the SLU, the approach to the DM component was different. RM presents a sequential DM, to recreate the intentional internal structure of a turn following DA theory, which was the same across different domains. Sounding Board, on the other hand, relies on a more vertical, albeit fragmented approach to the DM component, with independent DMs (Miniskills) built for each of the supported domains (Fang et al., 2017). Such a vertical approach on the most popular domains (e.g. Sports, Music, etc.) was also used by several other teams (Pichl et al., 2017; Serban et al., 2017b; Papaioannou et al., 2017), thus comprising all finalists. One of the advantages of using a fragmented approach to Dialogue Management is being able to deliver more effective and varied on the supported domains. However, in our opinion, such an approach also brings some disadvantages, such as a lack of robustness on less popular domains which might bring to break user engagement in cases those domains end up being discussed; and a lack of integration and possibly coherence for the DM, which could make it more difficult for the DM to learn successful patterns of distribution for coherent and engaging responses across different domains. While RM's more horizontal approach enables it to discuss virtually any domain (we had conversations spanning from philosophy, to cocktails, to what are the attributes of a good politician), it has the disadvantages of possibly not supporting the most popular domains with enough content. However, we notice how more in-domain support could easily be added to RM's architecture.

Another important feature which could be improved in the current RM architecture is performing better reranking of the generated responses. Other competitors relied more heavily on ensembling and reranking (Papaioannou et al., 2017; Serban et al., 2017b), which also lead to positive scores according to Ram et al. (2018), especially if combined with a robust SLU component.

3.6 Conclusions

We presented a modular open–domain dialogue system which reacts to user engagement. Our experiments suggest that a fully open-domain, non task-based dialogue system makes users confused and tends to produce lower ratings. This emerged by the progressive success of all our experiments (prompt change, personality tests, prescripted dialogues and prompt for unsupported request change). Another key aspect of the interaction with the user that came out during the semifinals is that users do not necessarily expect a regular, conversation-based interaction with social bots: many users are pleased with non canonical discussions such as personality tests, quizzes and word games, especially if they are coherently inserted in a conversation.

These results could be due to the fact that usually Alexa skills are based on strict, predefined interactions, and users expect the same kind of interaction from the Alexa Prize socialbots. In addition, our analysis shows that users have an average of 1.07 conversations, with a variance of 0.14, which could explain how strategies like personality tests and scripted dialogues kept giving good results throughout the competition, despite the fact that these strategies would probably annoy the users if proposed more than once. Another interesting finding was the positive correlation between both average length of conversations and average user's cumulative sentiment and user ratings. This suggests that users tend to give higher ratings to longer conversations and conversations in which they discuss entities towards which they have a positive attitude.

Regarding the future work, one improvement of our architecture would be the replacement of rule-based modules with data–driven ones using data collected during the competition, exploiting the modularity of our system. Useful features to train these modules could include successful patterns of strategies found in high-rated conversations, as well as information about the sentiment which emerged from the analysis of the user logs. We could also exploit our findings regarding cumulative sentiment to develop an error signal to integrate in our architecture. Another important improvement would be to enrich the commonsense knowledge base with other Wikidata information besides the "occupation" relation (e.g. movies, videogames, books), while at the same time refining the way entities and relations introduced by the machine are selected, for example ranking entities according to how much they appear in our news KB in order to select entities users are familiar with.

3.7 Summary

In this section, we presented Roving Mind (RM), the open-domain modular CA architecture which participated in the Alexa Prize 2017 competition, whose main challenge was to create models able to talk coherently and engagingly with real users about popular chitchat topics. The system presented in this Chapter was built from scratch by our team during the course of the competition.

RM architecture was designed to balance domain-independence and modularity, to address the need to have a form of control over the generated content (particularly important, given the open-domain nature and the interaction with real users). In order to address the challenges of multi-turn coherence and user engagement, RM crucially relies on Functional Units (FUs), i.e. Meaning Representation (MR) structures composed of open-domain DAs and entities. FUS are used throughout the entire design of our system, in the SLU, DM, NLG and KBs components.

In particular, the SLU module, relying on a domain-independent DA tagger and a chunker, parses the user input into a list of FUs and connects the extracted entities to our open-domain KBs. Then, the DM module is designed to generated a list of machine FUs which will then be realised by the NLG component into full utterances. In particular, the presented DM is based on a novel sequential architecture designed to mirror the intentional structure of conversational responses, following intuition from DA theory, to generate utterances coherent with the previous context and at the same time engaging. Progressing from components designed to fulfill a backward-looking function (Error Recovery and Management module) to those aimed at fulfilling a forward-looking function (Engagement module), each DM component in the sequence is associated to specific DA categories and connected strategies. Additionally, RM's DM is connected to a series of open-domain KBs, including a commonsense one, for coherent content generation.

In the last part of the Chapter we presented experiments conducted throughout the competition to test the usefulness of system-directed strategies to ensure a better experience for users. Overall, our experiments showed that directing the conversation towards specific topics can help ensuring a better interaction within the context of chitchat dialogue.

We also discussed how RM presented various issues, which could be addressed to improve the model's performance. For example, several parts of the architecture, though designed to be trainable, were not trained for lack of available data. Additionally, we highlighted the need for better ranking strategies, such as the ranking of entities selected for response generation, which we believe affected the performance of our system in the semifinals.

Chapter 4

ISO-standard open-domain Dialogue Act Tagging for Conversational Agents

This Chapter[1] addresses the problem of Dialogue Act (DA) tagging for a domain-independent Conversational Agent application. In particular, we describe the effort to create the DA tagger used as a crucial component of Roving Mind within the context of the Alexa Prize, where the model needed to be robust to capture DAs typical of chitchat open-domain conversation (such as Information transfers) and also be able to understand and address DAs more frequent in task-oriented dialogue (such as Action discussions).

However, at the time of publication of this work, the publicly available corpora (and connected schemes) available for DA tagging were either only focused on task-oriented conversation (e.g. MapTask) or only on chitchat (Switchboard); and research on DA tagging models was mainly based on in-domain approaches, where the model was trained and tested on the same dataset and dataset-dependant scheme.

In order to address these challenges, in this Chapter we propose a methodology to map different publicly available resources to the domain-independent ISO-standard scheme. Additionally, we present experiments to train domain-independent DA tagging models. First, we benchmark the performance of our proposed DA tagging models in-domain on the Switchboard Dialogue Act corpus obtaining SOTA results compared to similar models at the time; then we experiment with domain-independent DA tagging by training the tagger on the aggregated corpus and testing it on three out-of-domain datasets.

4.1 Introduction

The correct interpretation of the intents behind a speaker's utterances plays an essential role in determining the success of a dialogue. Hence, the module responsible for intents classification lies at the very core of many dialogue systems, both in research and industry (e.g. Alexa, Siri). Moreover, although the task of intent recognition is traditionally linked to task-based systems, recently it has also proved crucial for non task-based conversational agents (CAs). According to the results of the Amazon Alexa Prize challenge (Ram et al., 2018), the most successful CA in the competition

[1] The Chapter is based on Mezza et al. (2018).

relied on a strong spoken language understanding module, while more than 60% of the approaches explicitly used intents.

Nevertheless, automatic intent recognition is hard, since participants' intents in a dialogue are implicit. Intent classification has therefore been mostly modeled as a supervised machine learning problem (Gupta et al., 2006; Xu and Sarikaya, 2013; Yang et al., 2016), with the consequent definition of intents taxonomies. Over time this led to the creation of expensive annotated resources (Price, 1990; Henderson, Thomson, and Williams, 2014) with the related time-consuming design of multiple intent schemes. In most cases, however, intents taxonomies are defined specifically for a given application or a dataset and are not generalizable to other systems or tasks, making these resources difficult to reuse (e.g. the popular Air Travel Information Services (ATIS) corpus include heavily domain-dependent intents such as *Airfare* or *Ground Service*).

Dialogue Acts (DA), also known as speech or communicative acts, represent an attempt to create a formalized and generalized version of intents. As such, DAs have been investigated by the research community for many years (Stolcke et al., 2000b) and have been applied successfully to many tasks. In particular, their aspiration to generality makes them an appealing option for non task-based application (e.g. more than 20% of the teams in Amazon Alexa Prize Challenge explicitly used DAs (Cervone et al., 2017; Bowden et al., 2017), including the winning team (Fang et al., 2017)). Also, in the case of DAs, over the years there have been several efforts to produce publicly available annotated resources (Godfrey, Holliman, and McDaniel, 1992; Carletta, 2006; Alexandersson et al., 1998) to train DA taggers. The DA taxonomies created for these resources, albeit arguably more general compared to corpora like ATIS (for example utilizing categories such as *wh-questions*), are still dataset specific; since many of these schemes lack coverage of some crucial aspects of dialogic interaction. Furthermore, given that all these datasets utilize different schemes, these resources are hardly compatible.

The ISO 24617-2 (Bunt et al., 2010; Bunt et al., 2012), the international ISO standard for DA annotation, represents the first attempt to create a truly domain and task independent scheme. Given its holistic approach compared to previous schemes, ISO 24617-2 can be used as a lingua franca for cross-corpora DA mapping, as confirmed by successful attempts to remap single corpora to the standard (Chowdhury, Stepanov, and Riccardi, 2016; Fang et al., 2012b).

However, there is no reference training set for the standard, since the only public resource currently available with ISO 24617-2 annotation (DialogBank, (Bunt et al., 2016)) is too small to be used to train classifiers. Therefore, most DA tagging research still focuses on in-domain studies on large datasets with incompatible DA annotations (Stolcke et al., 2000b; Ji, Haffari, and Eisenstein, 2016). Moreover, most publicly available corpora are imbalanced with respect to the distribution of various DA dimensions such as *Information Transfer* (e.g. "What's your favourite book?") or *Action Discussion* (e.g. "Tell me the news."), which are required for successful open-domain CAs.

In this work, we show how to reuse and combine publicly available annotated resources to create a large training corpus for domain-independent DA tagging experiments. We map different corpora using an ISO standard compliant DA taxonomy, following the previous research on the topic (Fang et al., 2012b; Petukhova, Malchanau, and Bunt, 2014) and we share this resource with the research community.[2]

In order to investigate the soundness of our approach compared to in-domain models we further experiment with domain-independent DA tagging. As previously done in the literature we cast the Dialogue Act tagging task as a supervised multi-class classification problem using Support Vector Machines. The correctness of the approach is first tested on the de facto DA tagging standard – the Switchboard (SWBD-DA) corpus (Godfrey, Holliman, and McDaniel, 1992), using the reference training and test sets and achieving SOTA performance compared to similar approaches. Secondly, we experiment with domain-independent DA tagging following the same approach and using our combined resource as a training corpus. The DialogBank corpus, that represents a reference manual DA annotation for the ISO standard, is used for the evaluation of the tagger. To the best of our knowledge this is the first attempt to test automatic DA annotation on this corpus.

The domain-independence and suitability of the tagger for CAs trained on multiple resources is additionally evaluated on two other corpora annotated following our optimized taxonomy (human-machine conversations from the Amazon Alexa Prize Challenge). The performances achieved on these three datasets suggest that the training on multiple corpora represents a step forward for DA tagging of open-domain non task-based human-machine conversations. Finally, we present experiments to investigate the contribution of the different corpora to the performance of the classifiers. The results of our experiments show the importance of utilizing multiple resources to achieve a sound performance across different types of DA categories. The multi-domain DA tagger presented here was successfully employed in Roving Mind, our open-domain CA for the Alexa Prize described in Chapter 3.

4.2 State of the Art

4.2.1 Dialogue Act Annotation Schemes

The notion of Dialogue Acts can be traced back to the one of illocutionary acts introduced by (Austin and Urmson, 1962). The illocutionary act represents a level of description of an utterance's meaning that goes beyond the purely semantic level ("Is the window open?") to encompass the intent of the speaker in producing that utterance ("Please, close the window.").

One of the first DA taxonomies was the one created for the task-based corpus MapTask (Anderson et al., 1991) in the early nineties. The MapTask scheme distinguishes between *initiating moves* – such as giving instructions, explaining, checking information or asking questions – and *response moves* – for example acknowledging instructions, answering questions and clarifying information. The corpus also makes a distinction regarding the grammatical and semantic structure of the interactions, classifying, for example *wh-questions*, *yn-questions* and *positive/negative answers*. Although pioneering at the time, the MapTask annotation scheme is very specific to the described scenario, and some of its DAs (e.g. *instruct*, *clarify*, *check*) do not scale well to generic, non task-based conversations. Moreover, its taxonomy was not designed to capture all human behaviours during conversations, and, as a consequence, its coverage for labelling a non task-based interaction is inadequate.

[2] The suite of scripts we wrote to map and combine publicly available corpora can be found at `https://github.com/ColingPaper2018/DialogueAct-Tagger`

The first attempt to define a unified, non task-based standard for DA tagging was the Discourse Annotation and Markup System of Labeling (DAMSL) (Core and Allen, 1997) tag-set for the SWBD-DA (Godfrey, Holliman, and McDaniel, 1992) corpus. This annotation scheme proposes a taxonomy of 42 tags, describing both semantic aspects of conversation (*opinion*, *non-opinion*, *preference*, etc.), syntactic aspects (*yn-questions*, *wh-questions*, *declarative questions*, etc.) and behaviours related to the dialogue (*conventional closing*, *hedge*, *backchanneling*, etc.). Nevertheless, the taxonomy still has some issues: tags are mutually exclusive (making it impossible to annotate, for example, a no answer which was also signaling non-understanding) and are organised in a flat taxonomy, which does not take into account similarities and differences between the tags.

Bunt (1999) introduced the Dynamic Interpretation Theory (DIT) for dialogues, setting the theoretical foundation for a domain-independent and task-independent DA taxonomy. The paper introduced some very important concepts like the idea of *multidimensionality* of DAs and the distinction between *Action-Discussion*, a macro-category of DAs encompassing cases in which interlocutors negotiate actions to be performed (e.g. requests like "Let's switch topic."), and *Information-Transfer* interactions, capturing the DAs through which speakers exchange information (e.g. sharing personal information like "My name is John."). The DIT++ taxonomy (Bunt, 2009) was then defined in 2009 with the aim of providing a unique and universally recognized standard for DA annotation based on the theoretical ideas introduced in the DIT scheme. Its fifth version was accepted as ISO 24617-2.

The core aspects of the ISO standard are its multidimensionality and its domain and task independence. The ISO scheme is multidimensional in the sense that it makes a clear distinction between *semantic dimensions* (i.e. the aspect of the communication which the DA describes) and *communicative functions* (i.e. the illocutionary act performed within that dimension). In this way, ambiguities between various aspects of the communication and overlapping between DAs are removed. Furthermore, the scheme contains a generic dimension and communicative functions, which is suitable for mapping virtually any kind of conversation, both task-based and non task-based. Moreover, its multidimensional aspect and hierarchical taxonomy make it extensible and potentially adaptable to specific conversational sets.

4.2.2 Dialogue Act Tagging

The automatic recognition of Dialogue Acts has been addressed by the literature using various machine learning techniques. In particular DA classification has been modeled both as a sequence labeling problem, using techniques such as HMM (Stolcke et al., 2000b), neural networks (Ji, Haffari, and Eisenstein, 2016; Lee and Dernoncourt, 2016) or CRF (Quarteroni, Ivanov, and Riccardi, 2011), and as a multi-class classification problem, using for example SVM (Quarteroni and Riccardi, 2010). Additionally, different approaches cast the DA tagging task as either an offline task (Ji, Haffari, and Eisenstein, 2016), where the model considers both previous and future utterances compared to the current one and can thus be used only after the conversation is concluded, or an online task (Lee and Dernoncourt, 2016; Ortega and Vu, 2017), where the model only consider the utterances up to the current point of the conversation. Naturally, given our goal of using the DA tagger within a CA architecture, we cast the task as online. Mentioning and comparing all DA classification approaches is difficult because of the differences in annotation schemes and datasets used. All approaches, however, are usually tested on in-domain data.

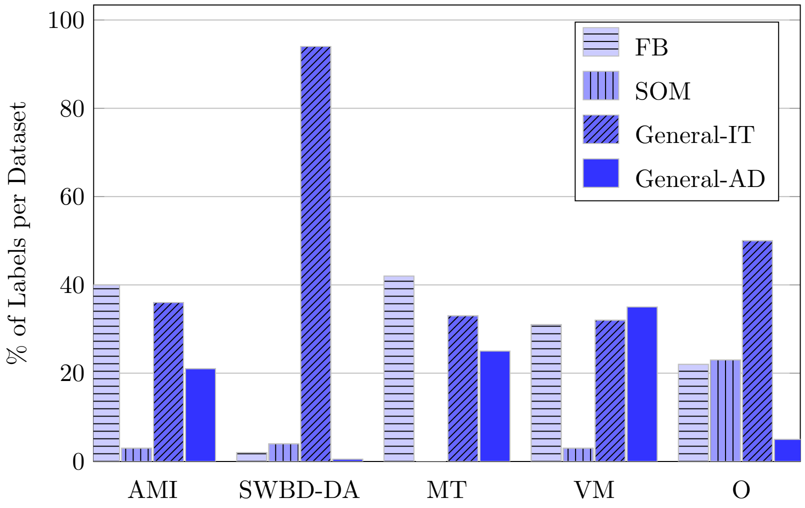


FIGURE 4.1: The distribution of dialogue act categories (after mapping to ISO standard) in various corpora – AMI, SWBD-DA, MapTask (MT), VerbMobil (VM) and BT Oasis (O). The represented DA categories are Social Obligations Management (SOM) and Feedback dimension DAs, as well as *Action-Discussion* (AD) and *Information-Transfer* (IT) DAs from general dimension..

One of the most popular datasets for benchmarking is SWBD-DA (Godfrey, Holliman, and McDaniel, 1992), a dataset of human-human open-domain telephone conversations. The SOTA at the time of publication of this work on SWBD-DA was achieved by (Ji, Haffari, and Eisenstein, 2016) using deep neural networks in offline mode (77.0% accuracy) and by (Ortega and Vu, 2017) for online mode (73.8%accuracy).

The ISO standard (Bunt et al., 2010) can be seen as a generalization of all these annotation schemes. However, there is no available training data for the ISO standard.

4.3 Data Sets

4.3.1 Training sets

The scarcity of resources of adequate size annotated with the ISO standard makes it difficult to train a DA tagger for this taxonomy. To the best of our knowledge, the DBOX corpus (Petukhova et al., 2014) – the only resource manually annotated using the ISO standard – is not yet publicly available . The best possible approach given the current availability of data is to map existing corpora's DA schemes to the ISO scheme. Given the limited – and often domain-dependent – annotation scheme of these resources, it is impossible to map enough data to train a DA tagger for the full ISO taxonomy, since some of the ISO dialogue acts have no correspondence in any of the considered corpora. Therefore, we opted for a reduced version of the taxonomy, limiting our research to subsets of the General (Task), Social Obligation Management and Feedback dimensions. So far, we mapped the following five different corpora to our scheme:

SWBD-DA The Switchboard corpus (Godfrey, Holliman, and McDaniel, 1992) is a dataset of transcribed open-domain telephone conversations. The Switchboard

Dialogue Act Corpus (SWBD-DA) is a subset of the Switchboard corpus annotated with DAs. SWBD-DA represents a logical choice when building a training set for a domain-independent DA tagger, as it is a large collection of open-domain, non task-based conversations, and therefore provides a natural similarity to the conversational domain of social bots. Moreover, there are already examples in literature of mappings from the Switchboard corpus to the ISO standard (Fang et al., 2012b). As visible from Figure 4.1, drawbacks of the corpus with respect to the task include its unbalancedness (60% of utterances are *Information-providing*) and lack of *Action-Discussion* interactions (less than 1% of overall corpus).

AMI This corpus contains transcriptions from 100 hours of meeting recordings of the European-funded AMI project (FP6-506811), a consortium dedicated to the research and development of technology (Carletta, 2006). This dataset presents a reasonably balanced collection of utterances and a taxonomy which shares some similarities with the ISO standard (e.g. distinction between *Action-discussion* and *Information-transfer*). Drawbacks of the corpus include the fact that there are multiple speakers (it is therefore more difficult to capture contextual information) and sometimes its scheme does not map to the leaves of the ISO tree.

MapTask This is a task-based dialogue corpus collected by the HCRC at the University of Edinburgh (Anderson et al., 1991). Dialogues involve two participants, one with an empty map and one with a route-marked map which must instruct the other speaker to draw the same route. The corpus was chosen due to its abundance of *Action-discussion* interactions (more than 30% of the overall corpus), which are often lacking in other corpora.

VerbMobil This is a collection of task-based dialogues released in 1997 (Burger et al., 2000). A subset of these dialogues is annotated with DAs (Alexandersson et al., 1998). The scenario involves two speakers, which play respectively the roles of a travel agent and of a client. The client usually provides a set of constraints and requests to be satisfied, while the traveling agent has to ask questions and provide information in order to satisfy the client's requests. Interactions happening within the VerbMobil 2 corpus closely resemble those usually seen with personal assistants, with a user looking for the fulfillment of a task and a serving agent interacting with the user to solve his/her issues making it an appealing addition to our training set.

BT Oasis The BT Oasis corpus is a collection of task-based conversations involving personal assistance for clients of the British Telecom services (Leech and Weisser, 2003). The conversations are human-to-human, and usually involve a user who has a problem to solve and an assistant who helps the user solving his issues. The BT Oasis corpus was chosen as part of the training set for its interesting scheme, called SPAAC (Speech Act Annotation scheme for Corpora), which is easily mappable to the ISO standard due to its clear separation of grammatical and illocutionary act.

4.3.2 Test sets

DialogBank (DB) the DialogBank (Bunt et al., 2016) is a corpus[3] annotated with ISO 24617-2 which currently contains 15 English dialogues: 3 from MapTask and 3 from TRAINS (Traum, 1996) (both task-based), 5 from DBOX (games collected in a

[3] `https://dialogbank.uvt.nl/`

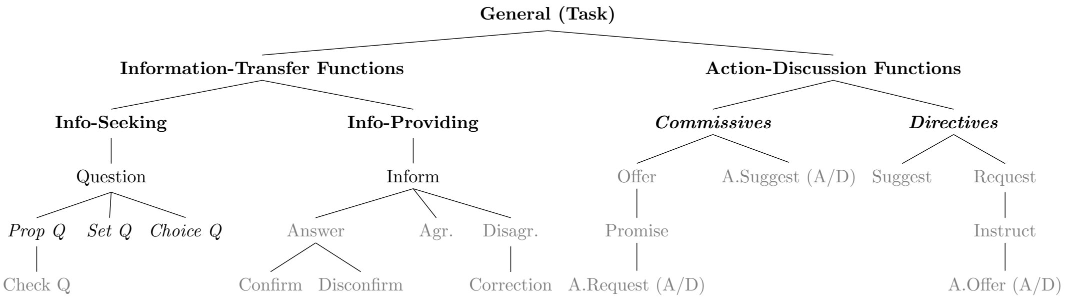


FIGURE 4.2: Communicative functions from ISO 24617-2 General purpose (Task) dimension. The nodes of the taxonomy that are not considered are grayed out.

Wizard-of-Oz fashion) and 4 from Switchboard (open-domain human-human conversation). Overall there are 1,596 DAs. The corpus currently represents the only publicly available resource manually annotated using the ISO standard.

Common Alexa Prize Conversations (CAPC) The CAPC corpus (Ram et al., 2018) is a dataset of 3,764 anonymised individual user turns pooled from different users interacting with all socialbots participating in the Alexa Prize. We have extracted a balanced subset of 458 turns and have annotated it with DAs from our adapted version of the ISO standard by 3 annotators, with an inter-annotator agreement of $\kappa = 0.82$. CAPC exemplifies frequent user interaction data not biased by the interaction with one socialbot in particular. Another advantage of this dataset is that it is balanced across different DA categories. One drawback is that no interaction context (previous DA) is available for the individual turns.

Socialbot Logs (S-Logs) S-Logs is a dataset of 13 open-domain conversations that different native American English speakers had with one of the socialbots of the Alexa Prize Challenge 2017. Overall this dataset contains 310 machine DAs and 165 user DAs. Two annotators tagged this dataset with DAs from our adapted version of ISO 24617-2, with an inter-annotator agreement of $\kappa = 0.81$. While we have annotated both machine and user turns, we test only on the latter and exploit machine turns as features for our classification experiments.

4.4 Methodology

4.4.1 Preprocessing

Before mapping the DA schemes of the corpora to the ISO subset scheme, a series of preprocessing steps have been performed to obtain a uniform training resource with the same surface text features as the testing corpora, since in S-Logs data, user input is lowercased and the punctuation is limited to apostrophes. More specifically, the text has been lowercased (including any named entity appearing in the original transcription and excluding the 'I' pronoun), punctuation has been removed (except for the apostrophe character in contracted expressions like "let's" and "can't") and any special characters have been deleted from the utterances. Moreover, any information regarding prosody has been removed, since this feature is not available in our test sets.

For experiments on the SWBD-DA DAMSL corpus we recreated the same setting described in (Stolcke et al., 2000b), using the same train and test set and preprocessing the corpus in the same way following the WS97 manual annotator guidelines (Jurafsky, 1997).

4.4.2 Dialogue Act scheme and mapping

The Socialbot scheme (S-scheme), the DA scheme used during the classification experiments, is a subset of the official ISO standard. Only three dimensions out of the official eight defined in the standard are considered (*Task*, *Social Obligation Management* and *Feedback*), and some of the communicative functions are generalized with an higher level of the tree.

Figure 4.2 shows the labeled subset of the ISO standard taxonomy for the General-purpose functions (i.e. functions independent from any given dimensions), while table 4.1 shows the correspondence for dimension-specific functions. The main difference between the DA scheme labeled in this work and the complete ISO taxonomy is the lack of further specification for the *Inform*, *Commissive* and *Directive* tags. This is due to the fact that most of the DA schemes used when building the training set do not provide contextual information detailed enough to label these tags accurately. Moreover, there is confusion and discrepancies about when these contextual DA should be used, even in the official ISO guidelines. Indeed, in (Fang et al., 2012b), which provides the official mapping from Switchboard to the ISO standard, it is reported that some contextual DA tags (for example *other_answer*) do not have a direct mapping to the standard. This becomes even more problematic considering that among the training resources there are corpora like AMI, MapTask or VerbMobil, which label answers as Informs, which would make training data for this class extremely noisy. A similar argument can be raised on the lower leaves of the *Directive* and *Commissive* nodes, some of which are not labeled even in the very detailed SWBD-DA taxonomy. Mapping of the available corpora to this scheme was done according to the available documentation in literature.

For the Switchboard corpus, a detailed mapping is provided in (Fang et al., 2012b), which was followed exactly for the supported dimensions/communicative functions. For MapTask and AMI, there is already research highlighting similarities and differences between their schemes and the ISO standard one (Petukhova, Malchanau, and Bunt, 2014). These results do not provide an exact mapping between the two schemes, which in some cases is impossible: for example the AMI *Elicit-inform* tag is the equivalent of ISO 24617-2's *Question*, but will not map to any specific question type (*SetQ*, *PropQ*, *ChoiceQ*, etc.). Utterances whose tags cannot be directly mapped to the ISO scheme were dropped and do not appear in the training set.

Since there is no available literature on mapping the VerbMobil 2 and BT Oasis corpora to the ISO standard, a specific mapping was designed from scratch by drawing inspiration from the approaches available on other corpora.

Table 4.2 presents counts of the DAs after mapping to our scheme, across all training and testing corpora. As mentioned in the paper, the corpora present quite imbalanced distributions of DA categories.

S-scheme	ISO 24617-2
Social Obligation Management	
Salutation	Greeting, Goodbye, Self-Intro
Apology	Apology, Accept Apology
Thanking	Thanking, Accept Thanking
Feedback	
Feedback	Auto-Feedback (all), Allo-Feedback (all)

TABLE 4.1: Our scheme (S-scheme) compared to the corresponding ISO 24617-2 scheme for the SOM and Feedback dimensions

DA	SWBD-DA	MapTask	VerbMobil	Oasis BT	AMI	DB	CAPC	S-Logs
Semantic Dimensions								
General (Task)	83,652	15,054	5,330	2587	1,523	1035	442	142
Social OM	2,866	0	384	588	10,039	21	16	7
Feedback	39,866	5,070	2,768	1,172	31,985	407	0	16
Total∗	126,384	20,508	8,482	2,381	43,547	855	329	109
% of Corpus	79%	100%	72%	58%	74%	100%	100%	100%
General (Task) Dimension								
Commissives	63	-	7	25	1,523	57	20	1
Directives	7	4,075	2,911	181	10,039	131	93	32
Inform	75,667	4,860	-	1,648	33,403	652	105	91
Prop. Question	1,986	583	-	492	-	61	68	12
Set Question	5,506	1,692	-	241	-	134	149	6
Choice Question	423	-	-	-	-	8	7	-
Total∗	83,652	11,210	2,918	2,587	44,965	1,035	442	142
% of Corpus	57%	30%	32%	35%	34%	N/A	N/A	N/A
Social Obligations Management								
Salutation	2,711	-	340	231	-	13	6	2
Apology	75	-	-	44	-	6	3	4
Thanking	80	-	44	193	-	2	7	1
Total∗	2,866	0	384	468	2,201	21	16	7
% of Corpus	2%	0%	2%	8%	0%	N/A	N/A	N/A
Feedback								
Total	39,886	5,070	2,768	1,172	31,985	407	-	16
% of Corpus	79%	100%	72%	58%	74%	N/A	N/A	N/A

TABLE 4.2: Dialogue Act category counts across the considered corpora for different levels of the taxonomy. *Percentages of corpora* indicate the percentage of data available for the particular level in the corpus.
∗ It is frequently the case that DA tags do not map to any leaf-node, e.g. VerbMobil for Task dimension and AMI for Social Obligations Management.

4.5 Experiments and Results

Since, to the best of our knowledge, there is no established SOTA on DialogBank – the only corpus manually annotated following the ISO 24617-2 scheme – we first establish the tagging methodology on the SWBD-DA corpus using the DAMSL 42 tag set and compare it to the SOTA for online mode (Ortega and Vu, 2017) and offline mode (Ji, Haffari, and Eisenstein, 2016). Then, the feature set and the parameters of the best performing models are used for the training of the DA tagger on the aggregate dataset, considering some of the semantic dimensions and the communicative functions of the ISO 24617-2 . The models are then evaluated on the DialogBank and open-domain human-machine data from Amazon Alexa Prize Challenge. McNemar's test (McNemar, 1947) for statistical significance has been used to analyze whether introduced features give a significant contribution to the overall performance.

4.5.1 Experiments on SWBD-DA

Prior to training the classification models, the SWBD-DA (Jurafsky, 1997) utterances are preprocessed following (Stolcke et al., 2000b). The dataset is split into training (1,115 dialogues) and test set (19 dialogues) following the same paper, and the remaining 21 dialogues are used as development set to tune the *C* parameter of Support Vector Machines (SVM) (Vapnik, 1995). For the experiments, we used the SVM implementation of scikit-learn (Pedregosa et al., 2011) with linear kernel (i.e. its *liblinear* (Fan et al., 2008) wrapper).

The results of the experiments on SWBD-DA are presented in Table 4.3. The performances are on the SWBD-DA test set with the SVM *C* parameter set to 0.1, with respect to the best results on the development set. It is worth mentioning that tuning the *C* parameter boosts the performance on the development set by 2 points.[4] For comparison, the table also includes majority baseline, the results from different online mode taggers, including (Stolcke et al., 2000b; Quarteroni and Riccardi, 2010; Lee and Dernoncourt, 2016) and the SOTA (Ortega and Vu, 2017). Additionally, we also include the SOTA for offline mode (Ji, Haffari, and Eisenstein, 2016).

Following the previous studies on SWBD-DA (Stolcke et al., 2000b; Quarteroni and Riccardi, 2010), we experiment with n-grams (unigrams, bigrams, and trigrams) and previous DA tag features. We do not consider the unit length feature from (Quarteroni and Riccardi, 2010), since classification instances in the SWBD-DA scheme and ISO 24617-2 are different (slash unit vs. functional unit). The results are reported in Table 4.3; since the results reported were obtained with SVM *C* parameter set to 0.1, they are higher than the ones reported in (Quarteroni and Riccardi, 2010): e.g. for 1-2-grams 70.0 vs. 71.7.

The first observation is that the addition of the previous DA significantly improves the performance. Addition of part-of-speech tags does not yield any improvement; however, when POS-tags are indexed with their positions in an utterance, accuracy is significantly improved and rises to 76.2. Addition of dependency relations (both with and without indexing with their position) does not improve the performance. The addition of the averaged pre-trained word-embedding vectors (from Google

[4] From 73 (C = 1.0) to 75 (C = 0.1) for the model trained on unigrams, bigrams and previous DA-tag.

Task definition	Models	Accuracy
online	Baseline: Majority	31.5
	HMM (Stolcke et al., 2000b)	71.0
	SVM (Quarteroni and Riccardi, 2010)	72.4
	CNN (Lee and Dernoncourt, 2016)	73.1
	LSTM (Ortega and Vu, 2017)	73.8
online	SVM 1-grams	71.2
	SVM 1-2-grams	71.7
	SVM 1-2-3-grams	71.4
	SVM 1-2-grams + PREV	74.6*
	SVM 1-2-grams + PREV + POS	74.6
	SVM 1-2-grams + PREV + I-POS	76.2*
	SVM 1-2-grams + PREV + I-POS + DEP	76.0
	SVM 1-2-grams + PREV + I-POS + I-DEP	76.1
	SVM 1-2-grams + PREV + I-POS + WE	**76.7***
offline	DrLM (LSTM) (Ji, Haffari, and Eisenstein, 2016)	77.0

TABLE 4.3: Classification accuracy of the different feature combinations on the SWBD-DA test set. The best results are highlighted in bold. The results that are significantly better are marked with *. Besides the Majority baseline, we report the results of *online* models, which consider only the dialogue history up to the current turn and to which also our model belong, and results of *offline* models (Ji, Haffari, and Eisenstein, 2016), which consider both past and future utterances for DA prediction of the current utterance.

News) to the model with indexed POS-tags, however, rises the accuracy to 76.7, making it a SOTA resulst compared to previously published online approaches (though coming 0.3 short of the offline model reported in (Ji, Haffari, and Eisenstein, 2016)).

4.5.2 Experiments on Aggregate ISO-standard Data

The methodology established on SWBD-DA is applied to training the ISO 24617-2 subset models using the aggregate data set. Since in ISO 24617-2 annotation scheme DAs consist of semantic dimensions and communicative functions, the utterances are first classified into the considered semantic dimensions – general, social obligations management (SOM), and feedback. Then, we experiment with the Task dimension, reporting the results without error propagation from the previous step, in order to give the reader a clearer understanding of the current classification capabilities when restricting interactions with the system to general communicative functions.

Semantic Dimension Classification

The results of the binary dimension classification models on the test sets – DialogBank (DB), CAPC, and S-Logs – are reported in Table 4.4. The CAPC corpus consists of isolated utterances; consequently, the *Feedback* dimension is not present. On DB and S-Logs, on the other hand, the *Feedback* dimension yields the lowest accuracy in comparison to General and *SOM* dimension communicative functions. Low performances on the *Feedback* dimension could be explained by the fact that the training data mostly contains *Allo-feedback* and lacks *Auto-feedback* and *Feedback elicitations*, which are present in DB.

Dimension	DB	CAPC	S-Logs
General	73.3	83.0	80.2
SOM	78.1	90.7	86.6
Feedback	56.3	–	71.3
Overall	68.4	83.3	79.4

TABLE 4.4: Classification accuracies of the binary semantic dimension models: General, SOM and Feedback. The CAPC corpus does not contain Feedbacks, therefore results for this dimension are not reported.

Features	DB	CAPC	S-Logs
BL: Majority	53.4	22.9	63.4
1-2-grams	64.2	71.2	78.7
+ PREV	64.3	70.7	81.6*
+ I-POS	65.8*	73.8*	82.2
+ I-DEP	67.1*	74.3	82.3
+ WE	65.2	74.8*	82.0
+ I-DEP + WE	66.6	75.1	81.8

TABLE 4.5: Accuracies of the feature combinations on the general-purpose communicative functions on the test sets. The best results are marked in bold, and statistically significant differences with *.

Communicative Function Classification

The utterances are further classified into communicative functions of the General (Task) dimension, using the methodology established on SWBD-DA, i.e. the same hyper-parameter settings ($C = 0.1$) and features. However, since models with dependency relations do not yield statistically significant differences, they are also considered. The results of the models on the test sets are reported in Table 4.5. The behavior of the models trained with various feature combinations is in-line with the SWBD-DA experiments: the addition of the previous DA tags and part-of-speech tags indexed with their positions in a sentence improves the performance. Different from the SWBD-DA, the addition of the indexed dependency relations improves the performance on the test sets. In the case of DialogBank and CAPC, their contribution is statistically significant. Additionally, unlike for SWBD-DA, the addition of word embeddings with and without index dependency relation (I-DEP) does not produce significant improvements for all but CAPC. Consequently, the model trained on 1-2-grams, previous DA tags, indexed POS-tags and dependency relations is chosen for the ablation study.

Corpora Combinations

The aggregation of all the corpora mapped to our subset of ISO 24617-2 is not necessarily the best one, as the distributions of DA categories varies from corpus to corpus. Consequently, we also present results on the test sets for the models trained solely on SWBD-DA and AMI; as well as perform an ablation experiment removing one corpus at a time. The best performing model from the previous subsection (1-2-grams, previous DA-tag, indexed POS-tags, and indexed dependency relations) is used for the study. The results of these experiments are reported in Table 4.6.

Dataset	DB	CAPC	S-Logs
All	**67.1**	74.3	82.3
All except AMI	59.7	73.7	71.3
All except SWBD-DA	60.2	68.3	77.5
All except Oasis BT	66.1	74.2	81.8
All except MapTask	66.8	**74.6**	80.5
All except VerbMobil	66.5	74.0	**82.6**
Only SWBD-DA	57.9	71.3	53.5
Only AMI	53.2	39.8	61.6

TABLE 4.6: Accuracies of the corpora combinations on the test sets – Dialog Bank (DB), CAPC, and S-Logs.

While the best results for Dialog Bank are achieved considering all the corpora, for CAPC the best results are achieved by removing MapTask. For S-Logs, on the other hand, the best performing corpora combination is all except VerbMobil. However, the performance differences from the models trained on all corpora are not statistically significant. Training DA taggers solely on SWBD-DA and AMI – the largest and the most diverse corpora – yields performances inferior to the combination of all the corpora. From the table, we can also observe that these two corpora – SWBD-DA and AMI – contribute most to the performance, as removing them affects the performance the most. On the other hand, removing the smaller datasets – BT Oasis, MapTask, and VerbMobil – affects the performance less.

4.6 Conclusions

We have presented an effective methodology for corpora aggregation for domain-independent Dialogue Act Tagging on a subset of the ISO 24617-2 annotation. We have also reported an accurate evaluation of our approach on both in-domain and out-of-domain datasets, proving that the described DA tagging technique is indeed independent from the underlying scheme and task of the annotated corpora. Finally, the machine learning technique used for DA tagging was tested on a popular DA tagging task (the Switchboard corpus), obtaining SOTA results for the online mode (which considers only the utterances up to the current point).

This work represents one of the first attempts to use an ISO compliant DA scheme for a real-life application, as well as one of the first structured approaches for evaluation of dialogue resources annotated with this taxonomy.

Research on available training resources is one of the first things to look forward to, since the current data proved to be effective, but also presented numerous drawbacks (lack of adequate coverage for the Feedback dimension, imbalanced DAs, lack of context-aware communicative functions). In the meanwhile, we plan to make our resource continue to grow in the future by adding and mapping additional corpora, such as the MRDA (Shriberg et al., 2004) corpus.

Additionally, the approach to DA tagging presented in this work presents some limitations, which should be addressed in future work. First, our models address only the task of DA classification. However, according to the ISO standard (Huang, 2017, Chapter 19), utterances can be multifunctional from a sequential perspective, i.e. the same utterance can be composed of multiple DAs in sequence. This means that, in order for our DA tagging models to be employed we first need a segmentation step (unless we assumed that each utterance contains only one DA). DA segmentation is indeed an often overlooked step by SLU modules. In RM the segmentation and classification steps are done independently, as reported in Section 3.3.3. Having these steps performed independently though, could lead to an error propagation across modules, therefore over the years there have been approaches to jointly perform DA segmentation and classification, such as Quarteroni, Ivanov, and Riccardi (2011) and most recently Zhao and Kawahara (2019). In order to properly address this issue, in future work we could explore joint DA segmentation and classification using the datasets proposed in this work.

A second limitation of our approach is the fact that our models currently utilise only the immediate previous context (i.e. the preceding turn). However, DAs are contextual, and the dependencies across different DAs can span also several turns, as we discussed in section 2.3.2. In future work we should explore how to include the full length of the context to account for such long range dependencies.

4.7 Summary

In this Chapter, we addressed the challenge of training a DA tagger that can be used online in a real-world domain-independent CA application using publicly available resources.

First, we analysed current approaches and resources for DA tagging. We discussed currently available corpora annotated with DAs and showed their imbalance in terms of distribution of crucial DA macro-categories; looking in particular at Action Discussion and Info transfer which are both essential for a real-world open-domain CA. In general, we highlighted the need to create resources with a consistent distribution of diverse DA categories across different domains.

Our first main contribution was then to create a methodology to create an aggregated multi-domain corpus for DA tagging by mapping 5 different publicly available corpora (SWBD-DA, AMI, MapTask, VerbMobil and BT OAsis), each one with its own in-domain scheme, to a subset of the ISO 24617-2 standard, the latest accepted taxonomy for domain-independent DAs.

Afterwards, we presented a serious of experiments for both in-domain and out-of-domain DA tagging, proposing a simple, yet efficient SVM-based model. In our first set of experiments we assessed the performance of our model on the popular SWBD-DA, commonly used as a benchmark for DA tagging models. Compared to other online models (i.e. considering only the utterances up to the current one) our model achieved SOTA results at the time. Afterwards, we presented experiments of out-of-domain DA tagging, by training our model on the aggregated corpus and testing on 3 out-of-domain corpora (DialogBank, and two corpora created using Alexa Prize data). Finally, we performed ablation experiments, by assessing the performance of the model when removing specific corpora from the aggregated resource. Overall, the experiments showed the importance of combining multiple resources for performing robustly on different DA categories.

Besides being a key component of Roving Mind (see Section 3), the DA tagger presented in this Chapter is used also in Section 6.1 to add DA features for conversation evaluation experiments .

Chapter 5

Weakly supervised approaches for open-domain dialogue coherence

In this Chapter, we investigate methodologies to train ranking models for coherence in open-domain dialogue using weakly supervised data generation methodologies. Compared to supervised techniques, weakly supervised ones could be useful to alleviate the data bottleneck typical of dialogue (as described in Section 1.3).

In particular, we propose coherence models for open-domain dialogue based on DA and entity information, first considering whole conversations[1] and then at the level of single turns[2]. Our findings throughout this Chapter indicate that DA and entity play an essential role for assessing dialogue coherence, especially if combined.

5.1 Conversation-level

5.1.1 Introduction

This Section addresses the problem of automatic coherence assessment of conversations. Coherence – what makes a text unified rather than a random group of sentences – is an essential property to pursue for a system aimed at conversing with humans. Nonetheless, producing coherent responses across conversation turns remains an open research problem for state-of-the-art (SOTA) open-domain dialogue models (Li et al., 2016a; Li et al., 2016c).

Furthermore, progresses in open-domain dialogue modelling are currently curbed by a lack of standardized automatic metrics to evaluate and compare conversational systems (Liu et al., 2016). Most available automatic metrics for dialogue evaluation either rely on surface features such as the words used (e.g. BLEU (Papineni et al., 2002)), try to replicate generic human judgments (Lowe et al., 2016), or work only for task-based dialogue systems (Walker et al., 1997). For evaluation, the field still relies heavily on user satisfaction, an expensive and time-consuming process which poses its own challenges given the subjectivity of human judgment. While coherence has been proposed multiple times as an important metric to evaluate open-domain dialogue, there have been only few studies on open-domain dialogue coherence assessment (Gandhe and Traum, 2016; Higashinaka et al., 2014; Venkatesh et al., 2018).

[1] Section 5.1 is based on Cervone, Stepanov, and Riccardi (2018).

[2] Section 5.2 is based on Cervone and Riccardi (2020).

On the other hand, the Natural Language Processing (NLP) literature has made several attempts (Grosz, Weinstein, and Joshi, 1995; Barzilay and Lapata, 2008) to formalize the notion of text coherence into *coherence models*. The entity grid, one the most popular approaches to coherence modelling in this community, proposes to represent documents according to the patterns of distribution of entities mentioned in the text across adjacent sentences (Barzilay and Lapata, 2008). Besides its correlations with human judgment, among the reasons behind the success of this approach is the fact that it is linguistically motivated, capturing important aspects of discourse coherence related to entities distribution (Joshi and Kuhn, 1979; Givón, 1987). Since its original proposal, the entity grid has undergone multiple extensions and has been widely applied to different tasks such as text coherence rating, automatic summaries assessment and sentence ordering, among others (Barzilay and Lapata, 2008; Elsner and Charniak, 2011b). It has also been successfully applied to dialogue (Purandare and Litman, 2008; Elsner and Charniak, 2011a), for example for chat disentanglement.

Being a local coherence model, i.e. modelling paragraphs internal coherence rather than the global coherence of the entire text, the extensions of the grid proposed for dialogue do not take into account one essential characteristic of dialogue coherence that has been studied for several years: its intentional structure.

Several theories studying dialogue coherence are indeed rooted on the idea of an internal structure given by participants' intents in a conversation (Sacks, Schegloff, and Jefferson, 1974; Sacks and Jefferson, 1995; Schegloff, 1968; Schegloff and Sacks, 1973). In many approaches, the basic units of these sequences are a variation of Dialogue Acts (DAs), a concept based on Speech Acts theory (Austin, 1975), that conveys the illocutionary function of an utterance in a conversation; and represents a formalized and generalized lexicon of speaker intents. Attempts to formalize computationally similar theoretical intuitions about dialogue coherence (Grosz and Sidner, 1986; Allen and Perrault, 1980) did not find wide-spread application, since they require extensive expertise.

In this section, we propose entity-grid inspired coherence models for dialogue augmented with intentional information, represented by DA transitions across turns. To the best of our knowledge, this work is the first to combine entity grid coherence models with DAs. We compare our models to the original entity grid on the two de-facto standard tasks for coherence, i.e. sentence (in our case turn) ordering discrimination and insertion. We perform our experiments on three publicly available datasets conveying different types of dialogue (task-based and open-domain) and DAs annotation schemes, namely Switchboard (Godfrey, Holliman, and McDaniel, 1992), AMI (Carletta, 2006) and Oasis (Leech and Weisser, 2003). Our results show the crucial importance of the DA information for assessing dialogue coherence.

5.1.2 State of the art

One of the most fertile frameworks for local coherence modelling in text is the *entity grid* (Barzilay and Lapata, 2008). As shown in Figure 5.1, this approach proposes to represent the structure of a document (in our case a dialogue) through a grid displaying transitions in the syntactic roles of entities (the heads of Noun Phrases (NP)) across neighbouring sentences in the text. In the grid, the rows represent subsequent sentences (turns in our case, as in Elsner and Charniak (2011a)) while each entity is represented by a column. A grammatical role can be: subject (*S*), direct object (*O*) or neither (X), plus a symbol ($-$) to signal that an entity does not appear in that turn t. The assumption is that the grid topology of coherent texts exhibits certain regularities associated to the way entities are introduced and become the focus of the discourse. For example, in the case of the grid represented in Figure 5.1, Table A we can notice how the sentences are connected by the continuity of the entity "drugs" across different turns. If an entity appears more than once in the same turn the most prominent syntactic role is chosen ($S > O > X$).

By computing the probabilities over all possible transitions of length n from one category to all others (thus $\{S, O, X, -\}^n$) we can turn this representation into a feature vector, similar to a language model over the entity tags, representing the syntactic role transitions of entities in the whole document. It is important to notice that the entity grid is not lexicalised, since this information is lost when creating the feature vectors.

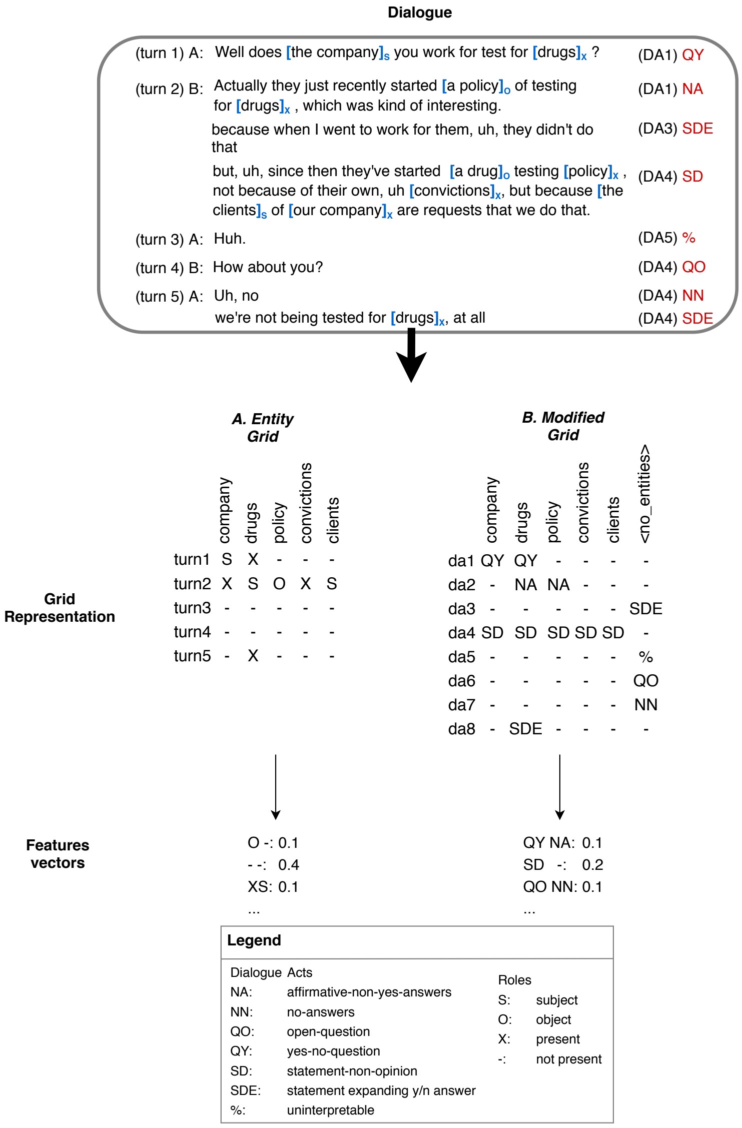
In Barzilay and Lapata (2008), the authors use these feature vectors to train a Support Vector Machine (using SVM$^{\text{light}}$ (Joachims, 2002)) modelling coherence as a ranking problem and using as a training dataset a set of original documents as positive (coherent) examples, paired with a set of the same documents with the sentences randomly permuted as negative (incoherent) examples (a procedure called pairwise training). The authors also experiment with models using different degrees of saliency (entity frequency) and transitions lengths (between 2 and 4), and by employing coreference resolution systems to detect entity chains (however given the performances of SOTA systems this addition does not provide improvements).

The algorithm proposed in Barzilay and Lapata (2008) derives thus automatically an abstract representation for a text, with as the only requirement a syntactic parser and a dataset. Among the weak points of this framework, however, is the fact that it models only local coherence (patterns of distribution across adjacent sentences) and a data sparsity problem.

Over the years, the entity grid model inspired numerous extensions (Guinaudeau and Strube, 2013; Filippova and Strube, 2007; Elsner and Charniak, 2011b) and similar implementations. Some approaches (Filippova and Strube, 2007), for example, augmented the model using the semantic relatedness of the entities but without much improvement. Others (Elsner and Charniak, 2011b) showed the usefulness of incorporating entity–specific features such as named entity information and considering also nouns which do not head NPs (as in Figure 5.1, turn 2, where in the NP "a drug testing policy" we consider both "drug" and "policy" as entities).

The typical tasks on which local coherence models are currently evaluated are: sentence ordering *discrimination*, where the system needs to rank original documents higher than randomly permuted ones, and *insertion*, introduced by Elsner and Charniak (2011b), where the system has to rank the position of a sentence removed from a document. The state of the art for these tasks was recently achieved by Nguyen and Joty (2017), which uses the entity grid as input to a Convolutional Neural Network. The authors report an accuracy of 88.69 (compared to 81.58 of the original grid with both head and non-head nouns) and an insertion score of 25.95 (compared to 22.13 of the same model). One of the advantages of the neural model compared to the original one is its ability to model long range entity transitions. Other recent works inspired by the entity grid include coherent paragraph generation (Li and Jurafsky, 2017), and applications to automated essay scoring (Farag, Yannakoudakis, and Briscoe, 2018) and neural stories text generation (Clark, Ji, and Smith, 2018).

Entity-based local coherence models apply well to dialogue as is or with some extra features, but have not been investigated in connection to DAs (Purandare and Litman, 2008; Elsner and Charniak, 2011a). Dialogue coherence has been explored outside of the entity grid approach as well (Higashinaka et al., 2014; Gandhe and Traum, 2016; Venkatesh et al., 2018). In Gandhe and Traum (2016), the authors propose a semi-automatic approach to evaluate dialogue coherence using only DA and relying on turn level coherence ratings from multiple sources. To the best of our knowledge, the only approach that combines entity and DA information for dialogue coherence evaluation is Higashinaka et al. (2014), which did not utilize the entity grid and models coherence as a binary classification task on utterance pairs rather than the whole conversation.



(turn 1) A: Well does [the company]$_{s}$ you work for test for [drugs]$_{x}$? (DA1) QY

(turn 2) B: Actually they just recently started [a policy]$_{o}$ of testing for [drugs]$_{x}$, which was kind of interesting. (DA1) NA

because when I went to work for them, uh, they didn't do that (DA3) SDE

but, uh, since then they've started [a drug]$_{o}$ testing [policy]$_{x}$, not because of their own, uh [convictions]$_{x}$, but because [the clients]$_{s}$ of [our company]$_{x}$ are requests that we do that. (DA4) SD

(turn 3) A: Huh. (DA5) %

(turn 4) B: How about you? (DA4) QO

(turn 5) A: Uh, no (DA4) NN

we're not being tested for [drugs]$_{x}$, at all (DA4) SDE

	company	drugs	policy	convictions	clients
turn1	S	X	-	-	-
turn2	X	S	O	X	S
turn3	-	-	-	-	-
turn4	-	-	-	-	-
turn5	-	X	-	-	-

	company	drugs	policy	convictions	clients	<no_entities>
da1	QY	QY	-	-	-	-
da2	-	NA	NA	-	-	-
da3	-	-	-	-	-	SDE
da4	SD	SD	SD	SD	SD	-
da5	-	-	-	-	-	%
da6	-	-	-	-	-	QO
da7	-	-	-	-	-	NN
da8	-	SDE	-	-	-	-

O -: 0.1
- -: 0.4
XS: 0.1
...

QY NA: 0.1
SD -: 0.2
QO NN: 0.1
...

Legend

Dialogue Acts

NA:	affirmative-non-yes-answers
NN:	no-answers
QO:	open-question
QY:	yes-no-question
SD:	statement-non-opinion
SDE:	statement expanding y/n answer
%:	uninterpretable

Roles

S:	subject
O:	object
X:	present
-:	not present

FIGURE 5.1: Entity grid example (A) vs. our modified grid (B) for the extract of the dialogue 002_4330 from the Switchboard Dialogue Act (SWBD-DA) corpus (upper part). Entities in the sentences are annotated with their syntactic role: subject (S), object (O) or neither (X). The Dialogue Act tags are directly taken from the SWBD-DA DAMSL annotation.

5.1.3 Methodology

Both the original and its SOTA extensions for coherence assessment focus on modelling local (entity-based) coherence, which is a form of surface coherence of the text (cohesion in Pragmatics theory (Halliday and Hasan, 1976)). However we can easily imagine how the entity grid or its extensions would not capture the lack of coherence in the following example:

A. Do you have dogs?
B. What is the average height of dogs?

In this case the text would be judged coherent given the continuation of the entity "dogs" across both turns. Nonetheless this example is incoherent because B does not answer A's question, but rather introduces an unrelated question.

In this work we augment the original entity grid document representation with a notion of global coherence, as provided by the intentional structure of the conversation in the form of Dialogue Acts. Our hypothesis is that DA information could improve coherence models performance on dialogue. This hypothesis is also motivated by the fact that syntactic roles might no be so prominent or reliable when transferred to the spoken dialogue domain, since for some dialogue types turns tend to be quite short and syntactic parsers are not very robust when there is no punctuation.

In order to test our hypothesis, we experiment with various grid constructions in order to find the best way to combine the DAs information with the original representation. For clarity, we follow a template <row>-Grid:<cell> for naming our different document representations. In particular the <row> refers to text span (row in the grid) chosen, either the Turn (T) as in Elsner and Charniak (2011a) or the text span of the DA (D); the <cell> refers to the category in the grid cells, either the syntactic role (*role*), the presence of the entity (*presence*, reducing the vocabulary to entities presence (X) or not ($-$) already proposed in Barzilay and Lapata (2008)) or the DA tag (*DA*, which varies according the DA schema of each dataset). In the rest of the section we detail the document representations in our experiments.

Baselines: The baselines $SVM\ ent_{role}\ grid_{turn}$ and $SVM\ ent_{presence}\ grid_{turn}$ replicate respectively the original entity grid in its all nouns variant (proposed by Elsner and Charniak (2011b)) and a simplified version of the grid where the vocabulary is restricted to two items.
SVM $\mathbf{ent_{role}}$ $\mathbf{grid_{DA}}$: This variation differs from the $ent_{role}\ grid_{turn}$ only for the fact that the text span units are DAs, rather than turns, while the vocabulary is still composed by syntactic roles. The disadvantage of this representation is that it is more sparse than its preceding one, but it is able to capture in-turn entities transitions.
SVM DA $\mathbf{grid_{DA}}$: In this variant the syntactic roles tags are substituted by the DA categories (according to the dataset's DA scheme). This is the modified grid shown in Figure 5.1, Table B. In this document representation an extra "no_entities" column is added to capture the DA tags where no entity is mentioned.
SVM DA: This text representation is the same as the previous one, with the difference that here all entities are dropped and we keep only one column with all the DAs.
Combinations: $SVM\ ent_{presence}\ grid_{turn} + DA$ and $SVM\ ent_{role}\ grid_{turn} + DA$ represent the combination of *SVM DA* with the two baselines by simply concatenating their feature vectors. These variations combine the entities and DAs feature vectors as

	SWBD-DA	AMI	Oasis
Number DA tags	42	16	41
Av. Number tokens/turn	13.1	15.1	10.6
Av. Number turns/dialogue	109	86.4	10.6
Number Train dialogues	740	356	191
Number Test dialogues	231	111	59
Number Dev dialogues	184	89	45

TABLE 5.1: We report the count of Dialogue Act tags, the average number of tokens per turn, the average count of turns per dialogue and counts of our Training, Test and Developments splits for our three datasets: SWBD-DA, AMI and Oasis. The document counts are given for the original documents, therefore need to be multiplied times 20 (pairs) for the discrimination task and times 100 (10x10) for the insertion task.

two separate sources of information.

5.1.4 Experimental setup

Tasks: We evaluate our models on the sentence ordering *discrimination* task proposed in the original (Barzilay and Lapata, 2008) and on the *insertion* task proposed in Elsner and Charniak (2011b), which represent the standard evaluation tasks for coherence models. In order to ensure comparability across our experiments, when permuting the order in the documents, we always permute the entire turn (therefore multiple rows in case we have several DAs in the same turn) and the same permutations are kept across all settings.

The first task, discrimination, is usually evaluated as accuracy of the model in ranking the original text higher than a permuted one (we use 20 permutations per document following previous work (Barzilay and Lapata, 2008; Elsner and Charniak, 2011b; Nguyen and Joty, 2017)). In order to better analyse our results, we add to this metric two widely used ranking metrics, i.d. Mean Reciprocal Rank (MRR, the average of reciprocal ranks in a set of queries) and Precision at One (P@1, the ability of the model to rank the original higher than all the permutations). In both these metrics, instead of comparing the original document with each of its permutation we compare the rank of the original document to all its permutations at the same time.

On the other hand, the insertion task is evaluated as the average number of sentences per document inserted in the correct position (therefore the average of the P@1). For the insertion task, we randomly pick 10 turns per dialogue and insert each one in 10 random positions (for each dataset we used the same turns and positions to ensure intra-dataset comparability).

Datasets In order to verify the robustness of our models across different DAs schemes and dialogue types, we perform all our experiments on three different publicly available datasets with DA annotation, namely BT Oasis (Leech and Weisser, 2003), AM I (Carletta, 2006) and the Switchboard Dialogue Act corpus (Godfrey, Holliman, and McDaniel, 1992) (SWBD-DA). Table 5.1 shows some differences across the datasets.[3]

[3] The code is available at: https://github.com/alecervi/Coherence-models-for-dialogue

	SWBD-DA			
	Discrimination			*Insertion*
	Acc.	**MRR**	**P@1**	**Av. P@1**
Random	50.00	16.98	4.76	8.70
SVM DA	**99.76**	98.76	97.80	45.45
SVM ent$_{presence}$ grid$_{turn}$	70.65	38.60	24.24	10.74
SVM ent$_{role}$ grid$_{turn}$	64.78	29.39	13.85	12.08
SVM ent$_{role}$ grid$_{DA}$	63.25	28.50	13.85	10.00
SVM DA grid$_{DA}$	99.57	97.36	95.67	38.79
SVM ent$_{presence}$ grid$_{turn}$ + DA	**99.76**	98.76	97.84	**45.58**
SVM ent$_{role}$ grid$_{turn}$ + DA	99.68	**99.17**	**98.70**	44.98

TABLE 5.2: We report results on the two tasks of Discrimination and Insertion on the Switchboard Dialogue Act (SWBD-DA) corpus. For Discrimination, we report the standard Accuracy (Acc.), Mean Reciprocal Rank (MRR) and Precision at one (P@1). For Insertion, we report the standard metric for this task, i.e. Precision at one (P@1) averaged for the dialogue.

The dialogues in SWBD-DA are open-domain telephone conversations. The individual turns tend to be quite long while the dialogues are the longest across the three datasets. For the DA categories we employ the 42 DAMSL ones. Oasis, on the other hand, is quite the opposite. A dataset of task-based conversations between clients and British Telecom help desk, here the turns tend to be quite short and the dialogues very short. AMI presents yet another type of dialogue data. Compared to the other datasets here the dialogues are between multiple speakers. In these dialogues participants were asked to discuss a project, so turns tend to be very long. This is also the dataset with the less rich annotation scheme compared to the previous two (only 16 DA categories).

Parameters As in the original entity grid paper we test all our models using the preference kernel implemented in SVM^{light} (Joachims, 2002) with default parameters. We follow the default original grid parameters (saliency:1, transitions length:2) for all our experiments. This was done to ensure a fair comparison between the datasets with few entities and short dialogues (Oasis) and those with many turns and several entities (Switchboard, AMI). For preprocessing the text to extract noun phrases and their syntactic roles we use spacy (Honnibal and Johnson, 2015).

5.1.5 Results

We report the results of our experiments on SWBD-DA in Table 5.2, on AMI in Table 5.3 and on Oasis in Table 5.4. To the model described in Section 5.1.3 we add a Random baseline, to give a measure of how the difficulty of both tasks vary across the datasets. To assess the respective significance of the coherence models, for discrimination accuracy and P@1 we use the McNemar test, while for discrimination MRR and the insertion Average P@1 we use Fisher's randomization test.

Regarding the *discrimination* task, the first thing to notice is how SVM DA, the model capturing DAs transitions without taking into account entities information, is a very competitive model across all the three datasets. Indeed the intentional structure information alone is so strong that on SWBD-DA and AMI, the task of discriminating an original document from randomly shuffled re-orderings of the same document

	AMI			
	Discrimination			***Insertion***
	Acc.	**MRR**	**P@1**	**Av. P@1**
Random	50.00	18.93	6.31	9.44
SVM DA	**98.78**	**95.27**	**92.79**	30.75
SVM ent$_{presence}$ grid$_{turn}$	76.71	40.88	25.23	7.21
SVM ent$_{role}$ grid$_{turn}$	79.59	46.73	28.83	11.71
SVM ent$_{role}$ grid$_{DA}$	59.41	25.40	11.71	11.71
SVM DA grid$_{DA}$	95.41	83.02	75.68	19.25
SVM ent$_{presence}$ grid$_{turn}$ + DA	98.47	93.74	90.09	31.41
SVM ent$_{role}$ grid$_{turn}$ + DA	98.51	94.56	91.89	**32.43**

TABLE 5.3: We report results on the two tasks of Discrimination and Insertion on the AMI corpus. For Discrimination, we report the standard Accuracy (Acc.), Mean Reciprocal Rank (MRR) and Precision at one (P@1). For Insertion, we report the standard metric for this task, i.e. Precision at one (P@1) averaged for the dialogue.

	Oasis			
	Discrimination			***Insertion***
	Acc.	**MRR**	**P@1**	**Av. P@1**
Random	50.00	17.39	5.08	9.16
SVM DA	91.53	68.47	54.24	41.44
SVM ent$_{presence}$ grid$_{turn}$	72.03	33.94	18.64	23.49
SVM ent$_{role}$ grid$_{turn}$	65.25	26.34	10.17	18.80
SVM ent$_{role}$ grid$_{DA}$	49.58	17.08	3.39	15.52
SVM DA grid$_{DA}$	87.80	57.64	40.68	28.96
SVM ent$_{presence}$ grid$_{turn}$ + DA	**92.46**	69.75	**57.63**	**42.74**
SVM ent$_{role}$ grid$_{turn}$ + DA	91.78	**70.39**	**57.63**	42.49

TABLE 5.4: We report results on the two tasks of Discrimination and Insertion on the Oasis corpus. For Discrimination, we report the standard Accuracy (Acc.), Mean Reciprocal Rank (MRR) and Precision at one (P@1). For Insertion, we report the standard metric for this task, i.e. Precision at one (P@1) averaged for the dialogue.

seems even too easy. With similar setup and data (also Switchboard but a different subset of dialogues with 505 original dialogues for training and 153 for testing) Elsner and Charniak (2011a) reports an accuracy of 86.0 for its extended version of the grid. The strength of the intentional structure information is still prominent, but less visible in Oasis, where the dialogues are much shorter compared to the previous two datasets and it might be possible that random shuffling of turns might not disrupt the dialogue coherence so effectively.

In general, we notice the importance of DA information across the three datasets also for the rest of the proposed models for the discrimination task. As expected, the lowest results are achieved by the SVM $\text{ent}_{\textit{role}}$ grid_{DA} model, which are still much better than the Random baseline. This model is similar to the original grid with the disadvantage of increasing the sparsity of entities.

The next lowest scores are then achieved by the SVM $\text{ent}_{\textit{presence}}$ $\text{grid}_{\textit{turn}}$ and SVM $\text{ent}_{\textit{role}}$ $\text{grid}_{\textit{turn}}$. While the second performs better on AMI, where turns are the longest and we can expect sentence structure to be more complicated, the SVM $\text{ent}_{\textit{presence}}$ $\text{grid}_{\textit{turn}}$ outperforms SVM $\text{ent}_{\textit{role}}$ $\text{grid}_{\textit{turn}}$ both on Oasis and SWBD-DA, confirming our hypothesis regarding the diminished importance of syntactic roles in dialogue. The next best model across all datasets for discrimination is SVM DA grid_{DA} with a large distance compared to SVM $\text{ent}_{\textit{presence}}$ $\text{grid}_{\textit{turn}}$ and SVM $\text{ent}_{\textit{role}}$ $\text{grid}_{\textit{turn}}$.

The best performing models are the combinations, where the entity and DA information are encoded separately. These models achieve the best results on SWBD-DA and Oasis, while their distance to SVM DA is not statistically significant on AMI.

The observations made on the discrimination task are reinforced on the *insertion* task. Only by looking at the performances (between 8.70 and 9.44) for the Random baseline, we notice how much the task is harder than the previous one (as mentioned in 5.1.2 the SOTA in the Wall Street Journal is 25.95). The noticeable difference in the results for the SVM $\text{ent}_{\textit{presence}}$ $\text{grid}_{\textit{turn}}$, SVM $\text{ent}_{\textit{role}}$ $\text{grid}_{\textit{turn}}$ compared to SVM DA grid_{DA} for insertion confirms once again how crucial is the intentional information. While also for insertion the intentional structure alone gives a very strong signal across all the datasets, the best results are achieved by combining the DAs with the entity information. This result is consistent with the nature of the task, where entity information could provide an important contribution to locating the exact place of a turn in the conversation. Also for this task, the syntactic role information yields the highest scores only for AMI, the dataset with the longest turns, while on SWBD-DA and Oasis the best results are achieved by the simpler model – SVM $\text{ent}_{\textit{presence}}$ $\text{grid}_{\textit{turn}}$ + DA.

The SVM DA model significantly outperforms the entity grid coherence models without DAs. However, while the models using the combination of entity grid and DAs (SVM $\text{ent}_{\textit{presence}}$ $\text{grid}_{\textit{turn}}$ + DA, SVM $\text{ent}_{\textit{role}}$ $\text{grid}_{\textit{turn}}$ + DA) yield better performance on SWBD-DA and Oasis, overall their differences are not statistically significant.

5.1.6 Conclusions

In this section, we applied the entity grid local coherence approach to dialogue. We experimented with different variations of its document representation in order to find the best way to augment it with participants' intents, an expression of global coherence and a signal which has been widely studied in dialogue to describe the structure of conversations. Our experiments confirm the crucial importance of the intentional structure for dialogue coherence, but also show how its combination with entity information could be useful for harder tasks connected to dialogue coherence, such as insertion.

Furthermore, our experiments show how the task of sentence ordering discrimination might be too easy on dialogue, where the DAs already give a very strong signal. On the other hand, the task of insertion is by far more difficult. In the next section, we will explore other tasks for coherence modelling that might be more useful for dialogue, such as automatic prediction of the next dialogue turn.

It is also important to notice that our proposals for document representation are independent of the Machine Learning models employed. They could therefore be used, for example, in combination with a CNN as implemented in Nguyen and Joty (2017). Another application we foresee for these models is to be used in the reward function for training dialogue systems in a Reinforcement Learning setting. Moreover, it is worth noticing that our experiments were performed using gold DAs. One of the first future experiments to perform would be to replicate the experiments with predicted DA labels (as, for example, was recently done in (Mesgar, Bücker, and Gurevych, 2020)), rather than gold ones to verify the robustness of the approach when using a DA tagger (the current approaches to DA tagging on Switchboard report accuracies above 75% (Ji, Haffari, and Eisenstein, 2016; Mezza et al., 2018)). In such a setting, we imagine that the entities information might play even more important role in assessing dialogue coherence. Other possible directions include applying our coherence models to chat disentanglement, as well as the automatic evaluation of conversational agents' coherence.

5.2 Turn-level

5.2.1 Introduction

Dialogue evaluation is an unsolved challenge in current human-machine interaction research. This is particularly true for open-domain conversation, where compared to task-oriented dialogue (i.e., restaurant reservations), we do not have a finite set of entities and intents, and speakers' goals are not defined a priori. We address the problem of dialogue evaluation from the perspective of dialogue *coherence* and how this concept can be formalized and evaluated. Our approach could be applied to both task-oriented and non-task-oriented dialogue.

Coherence in language, i.e., the property which determines that a given text is a logical and consistent whole rather than a random collection of sentences, is a complex multifaced concept which has been defined in different ways and to which several factors contribute (Redeker, 2000), e.g., rhetorical structure (Hobbs, 1979), topics discussed, and grounding (Traum, 1994).

While much recent work has focused on coherence for response generation (Serban et al., 2016c; Li et al., 2016a; Yi et al., 2019), we argue that there is still much to be understood regarding the mechanisms and substructures that affect human perception of dialogue coherence. In our approach, in particular, we are interested in studying the patterns of distribution of entities and Dialogue Acts (DAs), in regards to dialogue coherence.

Approaches to coherence based on entities have been studied extensively by the Natural Language Processing literature (Joshi and Kuhn, 1979; Grosz, Weinstein, and Joshi, 1995), especially in text (e.g., news, summaries). Coherence evaluation tasks proposed by this literature (Barzilay and Lapata, 2008) have the advantage of using weakly supervised training methodologies, but mainly considering documents as-a-whole, rather than evaluating coherence at the utterance level. The dialogue literature (Sacks and Jefferson, 1995; Schegloff, 1968), on the other hand, has focused mainly on coherence in connection to DAs, a generalized version of intents in dialogue (e.g., *yes-no-question, acknowledgement*). In section 5.1, in particular, showed the importance of both DAs and entities information for coherence modeling in dialogue. However, even in this case dialogue coherence was rated for entire dialogues rather than studying turn coherence structures.

In this Section, we investigate underlying conversation turn substructures in terms of DA and entity transitions to predict turn-by-turn coherence in dialogue. We start by annotating a corpus of spoken open-domain conversations with turn coherence ratings, the Switchboard Coherence corpus (SWBD-Coh)[4], and perform an analysis of the human perception of coherence in regards to DAs and entities. A multiple regression analysis shows the importance of both types of information for human rating of coherence. Secondly, we present novel neural models for turn coherence rating that combine DAs and entities and propose to train them using response selection, a weakly supervised methodology. While previous work on response selection (Lowe et al., 2017b; Yoshino et al., 2019) is mainly based on using the entire text as input, we deliberately choose to use only entities and DAs as input to our models, in order to investigate entities and DAs as a signal for turn coherence. Finally, we test our models on the SWBD-Coh dataset to evaluate their ability to predict turn coherence scores [5].

The main contributions of this work are:

- creating the Switchboard Coherence corpus, a novel human-annotated resource with turn coherence ratings in non-task-oriented open-domain spoken conversation;
- investigating human perception of coherence in spoken conversation in relation to entities and DAs and their combination;
- proposing novel neural coherence models for dialogue relying on entities and DAs;
- exploring response selection as a training task for turn coherence rating in dialogue.

5.2.2 Related work

Coherence evaluation in text Coherence models trained with weakly supervised methodologies were first proposed for text with applications to the news domain and summarization (Barzilay and Lapata, 2008). These models rely on the entity grid, a model that converts the entities (Noun Phrases) mentioned in the text to a sentence-by-sentence document representation in the form of a grid. The tasks on which coherence models in this line of research are usually evaluated are *sentence ordering* (Barzilay and Lapata, 2008), i.e., ranking original documents as more coherent

[4] The Switchboard Coherence corpus is available for download at: `https://github.com/alecervi/switchboard-coherence-corpus`

[5] The code for the models presented here can be found at: `https://github.com/alecervi/turn-coherence-rating`

than the same documents with the order of all sentences randomly permuted, and *insertion*, i.e., ranking original documents as more coherent than documents with only one sentence randomly misplaced. These tasks are still considered standard to this day and found wide applications, especially for text (Farag, Yannakoudakis, and Briscoe, 2018; Clark, Ji, and Smith, 2018). Recent models proposed for these tasks are based on Convolutional Neural Networks (Nguyen and Joty, 2017), also applied to thread reconstruction (Joty, Mohiuddin, and Nguyen, 2018), while the current State-of-the-art is based on a combination of bidirectional Long Short-Term Memory encoders and convolution-pooling layers (Moon et al., 2019). These tasks, however, consider documents as-a-whole and rely mainly on entities information.

Coherence evaluation in dialogue Models for dialogue coherence evaluation have mainly been explored using supervised approaches, i.e., training on corpora with human annotations for coherence, mostly at the turn level (Higashinaka et al., 2014; Gandhe and Traum, 2016; Venkatesh et al., 2018; Lowe et al., 2016; Yi et al., 2019). Different approaches tried to apply the standard coherence tasks to conversational domains such as dialogue and threads, but mainly considering the evaluation of dialogues as-a-whole (Purandare and Litman, 2008; Elsner and Charniak, 2011a; Vakulenko et al., 2018; Joty, Mohiuddin, and Nguyen, 2018; Mesgar, Bücker, and Gurevych, 2020; Zhou, Lan, and Wang, 2019). In particular, in section 5.1 we found that discrimination might be over-simplistic for dialogue coherence evaluation when considering Dialogue Act (DA) information. In this work, we propose a novel framework to model entities and DAs information for turn coherence prediction using a weakly supervised training methodology. Furthermore, our focus is on predicting coherence of single turns rather than entire dialogues.

Response Selection As a task, response selection has become a standard (Lowe et al., 2017b; Yoshino et al., 2019; Kumar, Agarwal, and Joshi, 2019) for training both task-oriented and non-task-oriented retrieval-based dialogue models. The task proved to be useful for evaluating models in task-oriented (Ubuntu), social media threads (Twitter Corpus), and movie dialogues (SubTle Corpus) (Lowe et al., 2016). Recently the task has also been proposed for pre-training models for task-oriented dialogue (Henderson et al., 2019) and for Dialogue Act tagging (Mehri et al., 2019). In this work, we investigate response selection as a task for training coherence rating models for spoken dialogue. Additionally, while response selection models are usually based on the entire text as input (Lowe et al., 2017b), we rely solely on entities and DAs information, in order to investigate their effect on turn coherence perception.

5.2.3 Methodology

In this work, we are interested in the relation between Dialogue Acts (DAs) and entities and how they can be modelled to train automatic predictors of next turn coherence in non-task-based dialogue.

Our hypothesis is that both entities and DAs are useful to predict the coherence of the next turn. In order to verify such hypothesis, we first perform an analysis of entities and DAs patterns of distribution in the Switchboard Coherence (SWBD-Coh) corpus, a novel dataset of human-human telephone conversations from Switchboard annotated with human coherence ratings per turn.

Secondly, we hypothesize that we can model entities and DAs to predict next turn coherence ratings. Rather than using supervised data for coherence prediction, we use a weakly supervised training methodology, i.e. training on the task of response

	Train	Dev	Test
Number source dialogues	740	184	231
Number insertion points	7400	1840	2310
Number pos/neg pairs	66600	16560	20790

TABLE 5.5: Train, development and test data size for response selection for both Internal and External Swap.

selection (which proved useful for other dialogue tasks (Henderson et al., 2019)) and testing on coherence ratings. In response selection given a *context*, i.e. the history of the dialogue up to the current turn, and a *list of next turn candidates*, models are asked to rank candidates according to their appropriateness with the previous dialogue history. The positive training samples for this task are automatically generated by randomly selecting a given turn in a dialogue, and considering this turn as a positive (coherent) example with the current history of the conversation (the context). Negative samples are generated by selecting other random dialogue turns, assuming that they will mostly be not appropriate as the next turn in the dialogue. In particular, we investigate two methodologies to generate negative samples from the training data automatically:

- **Internal swap**: a random turn is selected from a subsequent part of the same conversation. We assume this task to be harder for coherence evaluation since typically conversations do not have radical topic shifts.
- **External swap**: a random turn is selected from other conversations. We assume this task to be easier given the probable shifts in topic.

In our first set of experiments, we thus train our models on response selection. One of the possible shortcomings of the data generation procedure used in response selection, however, is the amount of false negatives. Although it is assumed that the majority of negative samples generated with this methodology will not be appropriate for the context, there could still be cases in which they are.

In order to verify the performance of our models based on DAs and entities to predict real human coherence judgments, in our second set of experiments models are tested on SWBD-Coh. Analogously to response selection, in turn coherence rating models need to rank next turn candidates given the history of the dialogue. In this case, however, the ranking is not binary but is rather based on a graded coherence rating given by humans for next turn candidates (for further details on the SWBD-Coh corpus see Section 5.2.4).

5.2.4 Data

The dataset chosen for our experiments is the Switchboard Dialogue Act corpus (Stolcke et al., 2000a) (SWBD-DA), a subset of Switchboard annotated with DA information. The Switchboard corpus is a collection of human–human dyadic telephone conversations where speakers were asked to discuss a given topic. This dataset was chosen both to ensure comparability with previous work on dialogue coherence and because it is open-domain. Also, this corpus has DA annotations. Interestingly, SWBD-DA is a real-world (transcribed) spoken corpus, so we have sudden topic

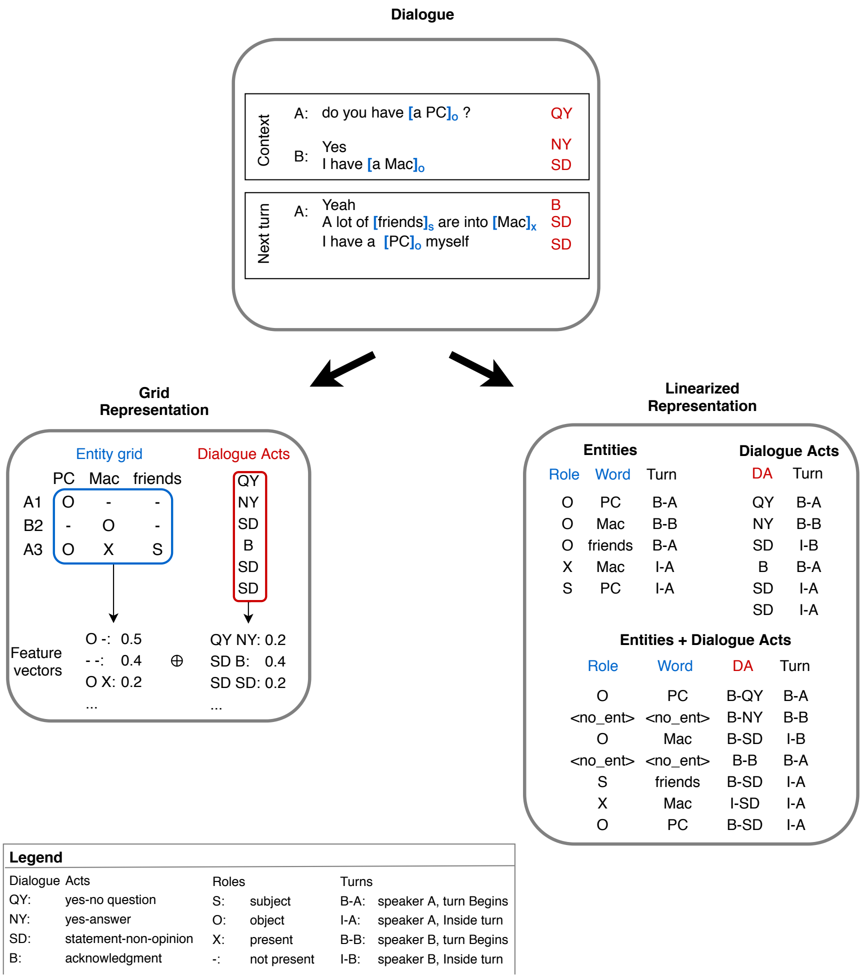


FIGURE 5.2: A source **dialogue** (at the center of the figure) is transformed into a **grid representation** (left) and into a **linearized representation** (right). In the grid representation, entities and Dialogue Acts (DAs) are transformed into feature vectors and can then be concatenated. Our linearized representation, i.e. the input to our neural models, shows 3 different possibilities: one where we only consider entity features at the turn level (top-left), another one which considers only DA features (top-right), and a joined one where DAs and entities are combined (bottom).

changes, overlap speech, disfluencies and other typical characteristics of spoken interaction. Since our goal was to study coherence in a real-world spoken dialogue setting, rather than removing these features as errors, we considered them an integral part of spoken conversations and did not remove them.

Response Selection Source dialogues are split into train, validation, and test sets (see Table 5.5) using the same distribution as in Section 5.1. For each dialogue, we randomly choose ten insertion points. Each insertion point is composed by a context (dialogue history up to that point) and the original turn following that context (regarded as positive). In order to have 10 next turn candidates, for each insertion point 9 adversarial turns (regarded as negatives) are then randomly selected either from subsequent parts of the dialogue, i.e. Internal Swap (IS), or from other dialogues, i.e.

External Swap (ES), within the same data subset, so that for example external adversarial turns for training are only taken from other source dialogues in the training set.

Switchboard Coherence corpus The dataset for turn coherence rating, the Switchboard Coherence corpus (SWBD-Coh), was created using as source dialogues the ones from SWBD-DA which are in the testset of Section 5.1. The data were annotated using Amazon Mechanical Turk (AMT). 1000 insertion points were randomly selected, following the constraints that the context (dialogue history up to the original turn) could be between 1 and 10 turns length. Since in this task we want to evaluate the coherence of a given turn with the previous dialogue history, 1 turn of context was the minimum required. We set the maximum length to 10 turns to reduce annotation time. For each insertion point, six adversarial turns were randomly selected, besides the original one (3 using the IS methodology, 3 using the ES one) for a total of 7 turn candidates. Overall the SWBD-Coh dataset is thus composed of 7000 pairs (1000 contexts × 7 turns).
Each context and turns pair was annotated by 5 AMT workers with coherence ratings. More specifically, for each dialogue workers were presented with the dialogue history up to the insertion point and the next turn candidates (randomly shuffled). Workers were asked to rate on a scale of 1 (not coherent), 2 (not sure it fits) to 3 (coherent) how much each response makes sense as the next natural turn in the dialogue. All workers (37) who annotated the dataset were first evaluated on a common subset of 5 dialogues where they had an average Weighted Kappa agreement with quadratic weights with two gold (internal) annotators of $\kappa = 0.659$ (min: 0.425, max: 0.809, STD: 0.101) and among each other an average leave-one-out correlation of $\rho = 0.78$ (i.e. correlating the scores of each worker with mean scores of all other workers who annotated the same data), following the approach used in other coherence rating datasets (Barzilay and Lapata, 2008; Lapata, 2006). [6] Scores for each candidate turn were then averaged across all annotators. Original turns were regarded on average as more coherent ($\mu = 2.6$, SD= 0.5) than adversarial turns, while turns generated with IS were considered more coherent ($\mu = 1.8$, SD= 0.7) than the ones generated via ES ($\mu = 1.4$, SD= 0.6). In terms of distribution (prior to averaging), scores were rather polarized: candidates were considered not coherent in 53% of the cases, coherent in 29%, while the score of 2 (not sure it fits) was given 17% of the cases.

5.2.5 Data analysis

In this section, we analyse the Switchboard Coherence (SWBD-Coh) dataset in regards to the distribution of Dialogue Acts (DAs) and entities. In particular, we are interested in analysing which features might affect human judgement of coherence of a given next turn candidate. For entities, we analyse two features: the number of entities mentioned in the next turn candidate that overlap with entities introduced in the context and the number of novel entities introduced in the turn. Additionally, we create a binary feature for each DA type that registers the presence of that DA in the turn candidate.

We use multiple regression analysis to verify how these different features correlate with human coherence ratings. Table 5.6, reports the Multiple Correlation Coefficient (MCC) of regression models using R squared and Adjusted R squared (Theil,

[6] More details about our data collection procedure are available in Appendix A.

	MCCR^2	**MCC**AR^2
Entities	0.27	0.26
DAs	0.34	0.29
All (Entities + DAs)	**0.45**	**0.41**
Relevant features in All	*Coeff.*	*Sign.*
Overlapping entities	0.26	**
DA: decl. yes-no-question	-0.48	*
DA: statement-opinion	-0.31	**
DA: statement-non-opinion	-0.30	**
DA: acknowledge	0.27	**

TABLE 5.6: Multiple Correlation Coefficients (MCC) from R squared (R^2) and Adjusted R squared (AR^2) of different multiple regression models that predict human coherence ratings for candidate turns given a dialogue context (Next Turn Rating task). Additionally, we report coefficients and significance (where * denotes .05 $\geq p \geq$ .01 and ** $p < .01$) of some relevant features for the best-performing model (All).

1961), adjusted for the bias from the number of predictors compared to the sample size. The results of our analysis indicate that the best MCC, 0.41 when calculated with the Adjusted R squared, is achieved when combining all features, both from entities and DAs. Moreover, in the lower part of Table 5.6 we report some of the features that proved to be the most relevant for predicting human coherence ratings. In general, it seems that while the entities overlapping the previous context seems to affect positively human coherence judgements, the DAs that most affect ratings do so in a negative way and seem to be mostly contentful DAs, such as *statement-opinion*, rather than DAs which typically present no entities, such as *acknowledge*. Our interpretation is that, in cases when there are no overlapping entities with the context, these DAs might signal explicit examples of incoherence by introducing unrelated entities.

5.2.6 Models

We model dialogue coherence by focusing on two features that have been closely associated to coherence in previous literature: the entities mentioned and the speakers' intents, modelled as Dialogue Acts (DAs), in a conversation. Our models explore both the respective roles of entities and DAs and their combination to predict dialogue coherence. We investigate both standard coherence models based on Support Vector Machines (SVM) and propose novel neural ones.

SVM models

The entity grid model (Barzilay and Lapata, 2008) relies on the assumption that transitions from one syntactic role to another of the same entities across different sentences of a text indicate local coherence patterns. This assumption is formalized by representing a text (in our case a dialogue) as a grid, as shown in Figure 5.2. For each turn of the dialogue we extract the entities, i.e. Noun Phrases (NPs), and their respective grammatical roles, i.e. whether the entity in that turn is subject (S), direct object (O), neither (X), or it is not present ($-$). Each row of the grid represents a turn in the dialogue, while each column represents one entity (in Figure 5.2, for example, the first turn of speaker A is represented by the first row of the grid $O - -$). Using this representation, we can derive feature vectors to be used as input for Machine Learning (ML) models by extracting probabilities of all roles transitions for each column.

More formally, the coherence score of a dialogue D in the entity grid approach can be modelled as a probability distribution over transition sequences for each entity e from one grammatical role r to another for all turns t up to a given history h (see Eq. 4 in Lapata and Barzilay (2005)):

$$p_{cohEnt}(D) \approx \frac{1}{m \cdot n} \prod_{e=1}^{m} \prod_{t=1}^{n} p(r_{t,e} | r_{(t-h),e} ... r_{(t-1),e}) \tag{5.1}$$

The probabilities for each column (entity) are normalized by the column length n (number of turns in the dialogue) and the ones for the entire dialogue by the number of rows m (number of entities in the dialogue). In this way we obtain the feature vectors shown in Figure 5.2 where each possible roles transition of a predefined length (e.g. $O-$) is associated to a probability. These feature vectors are then given as input to a Support Vector Machine (SVM) in the original model.

Following the work presented in Section 5.1, we can use the same approach to construct similar feature vectors for DAs information:

$$p_{cohDA}(D) \approx \frac{1}{n} \prod_{i=1}^{n} p(d_i | d_{(i-h)} ... d_{(i-1)}) \tag{5.2}$$

Here the coherence score of a dialogue is given by the probability of the entire sequence of DAs (d) for the whole dialogue, normalized by column length (n), i.e. the number of DAs for each turn.

The joint model, the one combining entity and DA information [7], simply concatenates the feature vectors obtained from both. While other ways of combining DA and entities have been explored in Section 5.1, we found that practically a simple concatenation resulted in the best performances across all tasks probably due to data sparsity issues.

Indeed among the limitations of the entity grid there is data sparsity: for example for an entity appearing only in the last turn of a dialogue we need to add a column to the grid which will be mostly containing "empty" $--$ transitions (see *friends* in Figure 5.2). Another problem of this approach is the fact that the model is not lexicalized, since we only keep role transitions when computing the feature vectors for the entities. Furthermore, the model makes the simplifying assumption that columns, thus entities, are independent from each other.

Neural models

Our neural coherence models for dialogue are based on bidirectional Gated Recurrent Units (biGRU). While other neural coherence models (Nguyen and Joty, 2017; Joty, Mohiuddin, and Nguyen, 2018) rely directly on the grid representation from Barzilay and Lapata (2008), we explore a novel way to encode the dialogue structure. The input to our biGRUs is a sequential representation of the dialogue.

[7] The model is referred as *SVM* ent_{role} $grid_{turn}$ + *DA* in Section 5.1.

Sequential input representation We linearize the structure of a dialogue composed by entities, DAs and turns into flat representations for our neural models, as shown in Figure 5.2. These representations can then be mapped to an embedding layer and be joined via concatenation.

In particular we consider three cases: (i) the case in which we model entity features; (ii) the one in which we consider DAs information; (iii) the one in which we combine both.

Entities encodings In our approach, entities are Noun Phrases, as in the entity grid approach. For each dialogue, we consider the sequence of entities ordered according to their appearance in the conversation (see Figure 5.2). Entities are represented either by their grammatical roles ent_{role} in the dialogue (using the same role vocabulary V_r of the original grid), their corresponding words ent_{word} (from a vocabulary V_w), or by both. Another feature which can be added to this representation is the *turn* (whether A or B is talking). This feature could be useful to encode the dyadic structure of the dialogue and how this might be related to entity mentions. In order to better encode the boundaries of speaker turns, turns are mapped to the IOB2 format (where the Outside token is removed because naturally never used for turns), for a resulting turn vocabulary V_t size of 4 tags (2 speakers x 2 IOB tags used). Special tokens (<no_ent>) are added to both V_w and V_r for cases in which turns do not present any entities.

DAs encodings In case we consider only *DAs* features, our input representation becomes a sequence of DAs for the whole dialogue history so far, drawn from a vocabulary V_d. Also in this case *turn* features can be added to mark the turn-wise structure of the DA sequence, using the same vocabulary V_t previously described.

Entities + DAs encodings We combine entities and DAs by considering the sequence of entities in order of their appearance within each DA and encoding DAs into IOB2 format, as previously done for turn features. In this setting, thus, the vocabulary V_d has double the size, compared to the setting where we consider only DAs. Analogously to previous settings, turn features can be added to encode turn boundaries.

It can be noticed how our representation is less sparse compared to both the original grid (Barzilay and Lapata, 2008) and recently proposed models (Nguyen and Joty, 2017), which take as input grid columns directly. Furthermore, compared to the original grid our representation is lexicalized.

Architecture The architecture of our models is shown in Figure 5.3. In the first layer of the network each input feature (ent_{role}, ent_{word}, DA, $turn$) is mapped to a d-dimensional dense vector by looking up into their respective embedding matrix **E**, one per feature type. All features vectors obtained can then be combined using concatenation. This vector is then recursively passed to the bidirectional GRU layers and then to a mean pooling layer. Finally, the output is passed through a feed-forward neural network with one hidden layer and ReLU as non-linearity.

Our models are trained using a Margin-ranking loss with a margin of 0.5 using the following equation:

$$\text{loss}(x, y) = \max(0, -y * (x1 - x2) + \text{margin}) \tag{5.3}$$

where $x1$ and $x2$ are respectively the original dialogue and the adversarial one and $y = 1$. In this way, the model is asked to rank the original dialogue higher (more

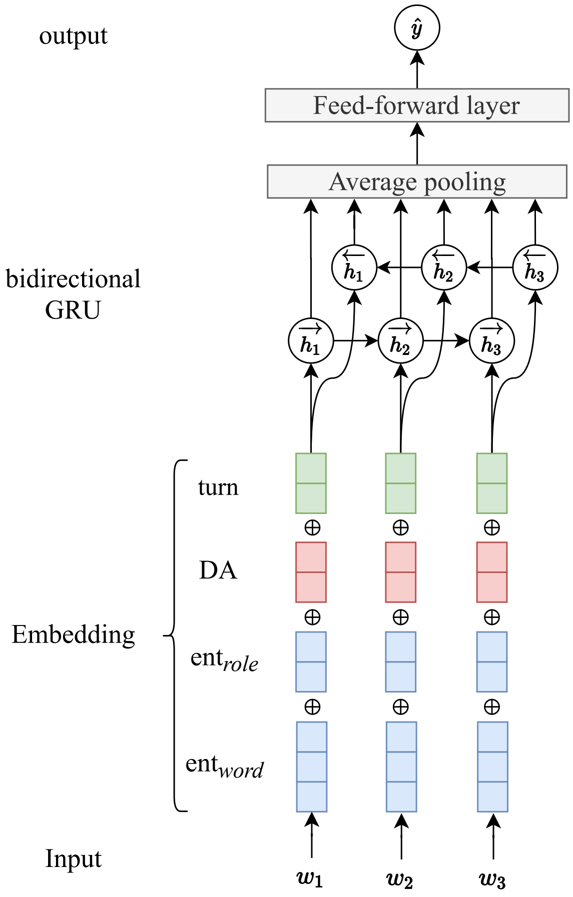


FIGURE 5.3: Our proposed architecture based on bidirectional GRUs with input entity word embedding (ent_{word}) and grammatical role (ent_{role}), Dialogue Act (*DA*) and speaker *turn* features.

coherent) than the adversarial one. The model is trained by Stochastic Gradient Descent, using the Adam update rule (Kingma and Ba, 2015).

5.2.7 Experimental set-up

Preprocessing Entities, i.e. Noun Phrases (NPs), and their syntactic roles were extracted and preprocessed with the pipeline presented in Section 5.1. Following the original entity grid formulation (Barzilay and Lapata, 2008), only NPs heads were kept. The DAs are taken from annotations on SWBD-DA (using the standard reduction to 42 tags compared to the DAMSL ones).

Evaluation For evaluating response selection, we use pairwise Accuracy, the metric used in standard coherence tasks, which evaluates the ability of the model to rank original turns higher than each adversarial one. However, this metric is not indicative of the global ranking of all candidate turns for a given context. For this reason, we add two ranking metrics to evaluate our models: Mean Reciprocal Rank (MRR), which evaluates the average of reciprocal ranks of all candidate turns for a context, and Recall at One (R@1) and Two (R@2), also used in previous work on response selection (Lowe et al., 2017b; Zhou et al., 2018) to assess the ability of the model to rank original turns respectively within the first or second rank among all candidates. Compared to response selection, where we have a binary choice between coherent

	Internal Swap				**External Swap**			
	Acc.	**MRR**	**R@1**	**R@2**	**Acc.**	**MRR**	**R@1**	**R@2**
Random	50.0	0.293	0.099	0.198	50.0	0.293	0.099	0.198
SVM ent$_{role}$ (Entity Grid)	36.6	0.260	0.103	0.178	39.5	0.246	0.096	0.126
SVM DA	60.6	0.398	0.206	0.335	61.3	0.403	0.212	0.346
SVM ent$_{role}$ + DA	62.7	0.417	0.222	0.365	64.3	0.437	0.251	0.380
biGRU ent$_{role}$	41.8	0.294	0.120	0.217	45.5	0.293	0.117	0.210
biGRU ent$_{role}$ + turn	43.3	0.295	0.120	0.214	45.9	0.293	0.115	0.211
biGRU ent$_{word}$	47.8	0.324	0.151	0.252	56.4	0.397	0.236	0.337
biGRU ent$_{word}$ + turn	49.0	0.331	0.162	0.255	56.9	0.400	0.241	0.341
biGRU ent$_{role}$ + ent$_{word}$ + turn	48.6	0.327	0.156	0.253	56.1	0.394	0.232	0.338
biGRU DA	72.4	0.484	0.276	0.443	72.6	0.486	0.278	0.447
biGRU DA + turn	74.0	0.501	0.297	0.464	74.1	0.508	0.305	0.475
biGRU ent$_{word}$ + DA + turn	**75.1**	0.520	**0.321**	0.484	**77.3**	**0.550**	**0.355**	**0.530**
biGRU all	75.0	**0.521**	**0.321**	**0.489**	77.2	0.549	0.354	0.529

TABLE 5.7: Average (5 runs) of Accuracy (Acc.), Mean Reciprocal Rank (MRR) and Recall at one (R@1) and two (R@2) for response selection using both data generation methodologies (Internal and External Swap) on Switchboard.

and negative turns, in turn coherence rating, we have a set of candidate turns each associated to a coherence score. In this case, we use Accuracy, MRR, R@1 and Normalized Discounted Cumulative Gain (nDCG) to evaluate our models. Accuracy was computed only for cases in which the rating of the turn was not identical across two candidate turns. MRR and R@1 were computed dynamically, that is considering the turn with the highest score within that particular context as the best one in the rank. The nDCG metric (Järvelin and Kekäläinen, 2002) assesses the gain of a candidate according to its rank among all candidates. Compared to previous metrics, nDCG allows taking into account the relevance (in our case, the coherence score) of candidates. For all metrics considered, if our models predicts the same score for two candidates, we always assume models made a mistake, i.e. among candidates with the same predicted score positive examples are ranked after the negative ones.

Models' settings Grid models, based on SVMs, were trained with default parameters using SVM^{light} preference kernel (Joachims, 2002)) as in the original model (Barzilay and Lapata, 2008). For saliency, i.e. the possibility of filtering entities according to their frequency, and transitions length we follow the default original grid parameters (saliency:1, transitions length:2). For neural models, implemented in Pytorch (Paszke et al., 2019), parameters were kept the same across all models to ensure comparability. The learning rate was set to 0.0005, batch size to 32, with two hidden biGRU layers of size 512. Embedding sizes for all features were set to 50–dimensions, except for word embeddings which had dimension 300. Models run for a maximum of 30 epochs with early stopping, based on the best MRR score on the development set.

5.2.8 Results

In this section, we report the results of our models for response selection. The best performing models on response selection are then evaluated on the turn coherence rating task using the Switchboard Coherence (SWBD-Coh) corpus as testset. For

	Train	Acc.	MRR	R@1	nDCG
Random		50.0	0.479	0.234	0.645
biGRU	IS	42.7	0.395	0.174	0.621
ent$_{word}$ + turn	ES	50.4	0.444	0.229	0.679
biGRU	IS	56.0	0.553	0.326	0.717
DA + turn	ES	56.0	0.558	0.337	0.725
biGRU	IS	58.5	0.575	0.358	0.738
ent$_{word}$ + DA + turn	ES	**61.1**	**0.583**	**0.369**	**0.760**

TABLE 5.8: Average (5 runs) of Accuracy (Acc.), Mean Reciprocal Rank (MRR), Recall at one (R@1) and Normalized Discounted Cumulative Gain (nDCG) for turn coherence rating for models trained using either Internal (IS) or External Swap (ES) on the Switchboard Coherence corpus.

both tasks we compare our models to a random baseline. All reported results for neural models are averaged across 5 runs with different seeds.

Response selection The results for response selection are reported in Table 5.7. Neural models seem to capture better turn-level coherence compared to classic grid SVM-based approaches. In both data generation methodologies, Internal (IS) and External Swap (ES), SVM coherence models are outperformed by neural ones for all metrics considered. As expected, entity features (ent$_{role}$, ent$_{word}$) play a more prominent role in ES compared to IS. In both cases, entity features seem to be better captured by neural models relying on our proposed input representation. When considering lexical information (ent$_{word}$), however, ent$_{role}$ features seem less relevant. This might be due to the fact that spoken dialogue has usually less complex syntactic structures compared to written text. Furthermore, parsers are usually trained on written text, and thus might be more error-prone when applied to dialogue where there are disfluencies, sudden changes of topics, etc. We notice that DAs alone (without entity information) play an important role in both IS and ES. Turn features capturing speaker information seem helpful for both DAs and entities.
In general, the combination of DAs and entities gives the best results both in SVM and neural models for both tasks, with the best performing one being the model combining ent$_{word}$, DA and turn features and without ent$_{role}$. Additionally, if we compare the IS setting to ES in terms of best MRR, Accuracy and Recall, the former seems more difficult. This confirms our expectations that IS might be an harder task for coherence.

Turn coherence rating A selection of best performing models for entities, DAs and their combination were tested on the SWBD-Coh dataset. Table 5.8 shows models' results under both training conditions, i.e. either using IS or ES data. The lowest performing model seems to be the one based solely on entity features (ent$_{word}$ + turn), while models combining DA with entities information (ent$_{word}$ + DA + turn) are the best performing ones. Additionally, models trained on ES data perform better than those trained on IS across all conditions.

			Models ranks		
Context	**Score**	**Candidates**	**Ent**	**DA**	**Ent+DA**
A: Okay. *B*: Well, if you are from **Iowa**, you must be very artsy crafty. Everyone I've ever known from the **Midwest** can do everything with their **hands**. *A*: Oh, well, actually I'm from **California** and before then I was from **Utah**. So.	3.0	I didn't know anyone ever moved from **California** to **Iowa**?	1	4	1
	2.6	Anyway, we are supposed to be talking about **crafts**. Do you, um, do you have any **hobbies** that, that you do things with your **hands**	2	2	2
	2.2	Right.	4	3	3
	2.2	Uh-huh.	4	3	3
	2.0	Oh, sure.	4	3	3
	1.2	**bags** some, their most recent, uh, **needle craft**	3	4	4
	1.0	at least at the end.	5	1	5

TABLE 5.9: Example of how different models relying only on entities (biGRU ent_{word} + turn), only on DAs (biGRU DA + turn) or both (biGRU ent_{word} + DA + turn) rank the same group of candidates for a given context.

5.2.9 Qualitative analysis

Table 5.9 shows an example of the ranking given by different models to the same context-candidates pairs in the SWBD-Coh corpus, compared to the average coherence score given by annotators. In particular, we report the ranking given by a model based solely on entities information (biGRU ent_{word} + turn), another one considering only DAs (biGRU DA + turn) and a third one considering both types of information (biGRU ent_{word} + DA + turn). All models were trained on response selection using the External Swap methodology. The models output is reported in terms of position in the rank. Entities appearing in the text are highlighted in bold.

In this example we notice entities overlap information with the previous context proves rather important in order to rank candidates according to coherence. For example, to rank the candidate with the highest coherence as the first one (*I didn't know anyone ever moved from California to Iowa?*) information regarding the overlapping entities *California* and *Iowa* allows the models encoding entities information to assign the correct rank, while the model relying only on DAs gives the candidate the fourth position in the rank. We also notice how both annotators and all models assign very close or the same middle rank scores to three very similar candidates (*Right*, *Uh-huh* and *Oh, sure.*), which indeed all have the same DA ("acknowledgment").

5.2.10 Conclusions

In this Section, we investigate how entities and Dialogue Acts (DAs) are related to human perception of turn coherence in dialogue. In order to do so, we create a novel dataset, the Switchboard Coherence (SWBD-Coh) corpus, of transcribed open-domain spoken dialogues annotated with turn coherence ratings. A statistical analysis of the corpus confirms how both entities and DAs affect human judgements of turn coherence in dialogue, especially when combined. Motivated by these findings, we experiment with different models relying on entities and DAs to automatically predict turn coherence, i.e. standard coherence models and novel neural ones. In particular, we propose a less sparse alternative, compared to the entity grid, to encode entities and DAs information. Rather than using data annotated explicitly for the task, i.e. coherence prediction, we explore two response selection methodologies for training. We find that our newly proposed architecture outperforms standard ones in response selection. Finally, we test our models on the SWBD-Coh corpus in order to evaluate their ability to predict real human turn coherence ratings. Crucially, we find that the combination of DAs and entities gives the best performances.

For the future work, it would be interesting to investigate how to apply large pretrained models to our task, such as BERT (Devlin et al., 2019). While pretrained models have recently been successfully explored for text-based response selection (Kim et al., 2019; Henderson et al., 2019), integrating them with our proposed input representation is not a straightforward task since such models typically rely on the whole textual context, while our models do not.

While there is still much to understand regarding turn coherence in dialogue, we believe our work could be a first step towards uncovering the relation between DAs and entities in open-domain spoken dialogue. Moreover, we believe that the SWBD-Coh corpus could become a useful resource for the community to study coherence in open-domain spoken dialogue.

5.3 Summary

This Chapter was dedicated to weakly supervised methodologies for training coherence models for open-domain conversation relying on Dialogue Act (DA) and entities information.

In the first part of the Chapter, we experimented with standard coherence tasks typically used for text, which involve evaluating the coherence of entire *conversations*. In particular, we proposed to augment entity-based models with Dialogue Act (DA) information, in order to combine the thematical aspect of coherence with the intentional structure of dialogue. We proposed various models, relying on different input representations combining DA and entity information, which could be used also independently of the ML model used (for example one of these representations is used in Section 6.2). Our experiments on conversation-level standard coherence tasks point to the importance of DA information for coherence rating. However, we also found that standard coherence tasks might be less useful to assess dialogue coherence, both from an application perspective and considering the role played by DA information.

The second part of the Chapter is then dedicated to *turn*-level coherence modeling in open-domain dialogue, which we argue could have more real-world applications for conversational AI applications (i.e. Dialogue Management, response ranking, turn evaluation) compared to standard coherence tasks. Our aim was to investigate human perception of turn coherence in relation to entity and DA patterns of distribution. In order to do so, we presented a novel resource of open-domain dialogue annotated with turn coherence ratings, the Switchboard Coherence corpus. We performed a statistical analysis of the resource and found that DA and entity information correlates with human judgment of turn coherence. Consequently, we explored the possibility of modelling DA and entity information for training turn coherence rankers. We proposed neural coherence models relying on novel linearised representations of the structure of conversations using entities and DAs. From a methodological perspective, we proposed to use response selection as a weakly supervised training task and then test models on turn coherence rating. The results of our experiments on turn coherence rating indicate that DAs and entities play a crucial role in coherence assessment, especially when combined.

Chapter 6

Supervised approaches for open-domain dialogue evaluation

In this Chapter, we present evaluation models for open-domain dialogue using supervised approaches, that is relying on annotated corpora. While in the previous Chapter the resources used were collected within research projects mainly in somewhat controlled environments (Switchboard, MapTask), the work presented in this Chapter relies on human-machine interactions collected during the Alexa Prize, in a noisier environment (users could be virtually anyone and had the possibility of ending the conversation whenever they wanted).

In the first part of the Chapter[1], we present models to evaluate the quality of entire conversations relying on a combination of DA and topic representations using as a supervised signal the user ratings obtained by Roving Mind, the open-domain CA presented in Chapter 3, during the Alexa Prize competition. In the second part of the Chapter[2], we propose models relying on combination of features including DA and entity information to predict turn level coherence and engagement, using a larger dataset of interactions between users and different dialogue models collected during the Alexa Prize competition.

6.1 Conversation-level

6.1.1 Introduction

We are currently witnessing a proliferation of conversational agents in both industry and academia. Nevertheless, core questions regarding this technology remain to be addressed or analysed in greater depth. This work focuses on one such question: *can we automatically predict user ratings of a dialogue with a conversational agent?*

Metrics for task-oriented systems are generally related to the successful completion of the task. Among these, contextual appropriateness (Danieli and Gerbino, 1995) evaluates, for example, the degree of contextual coherence of machine turns with respect to user queries which are classified with ternary values for slots (appropriate, inappropriate, and ambiguous). The approach is somewhat similar to the attribute-value matrix of the popular PARADISE dialogue evaluation framework (Walker et al., 1997), where there are matrices representing the information exchange requirements between the machine and users towards solving the dialogue task, as a measure of task success rate.

[1] Section 6.1 is based on Cervone et al., 2018.

[2] Section 6.2 is based on Yi et al., 2019.

Unlike task-oriented systems, non-task-oriented conversational agents (also known as chitchat models) do not have a specific task to accomplish (e.g. booking a restaurant). The goal of chitchat models could arguably be defined as the conversation itself, i.e. the entertainment of the human it is conversing with. Thus, human judgment is still the most reliable evaluation tool we have for such models. Collecting user ratings for a system, however, is expensive and time-consuming.

In order to deal with these issues, researchers have been investigating automatic metrics for non task-oriented dialogue evaluation. The most popular of these metrics (e.g. BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005)) rely on surface text similarity (word overlaps) between machine and reference responses to the same utterances. Notwithstanding their popularity, such metrics are hardly compatible with the nature of human dialogue, since there could be multiple appropriate responses to the same utterance with no word overlap. Moreover, these metrics correlate weakly with human judgments (Liu et al., 2016).

Recently, a few studies proposed metrics having a better correlation with human judgment. ADEM (Lowe et al., 2017a) is a model trained on appropriateness scores manually annotated at the response-level. (Venkatesh et al., 2018) and (Guo et al., 2017a) combine multiple metrics, each capturing a different aspect of the interaction, and predict conversation-level ratings. In particular, (Venkatesh et al., 2018) shows the importance of metrics such as coherence, conversational depth and topic diversity, while (Guo et al., 2017a) proposes topic-based metrics. However, these studies require extensive manual annotation on top of conversation-level ratings.

In this work, we investigate non task-oriented dialogue evaluation models trained without relying on any further annotations besides conversation-level user ratings. Our goal is twofold: investigating conversation features which characterize good interactions with a conversational agent and exploring the feasibility of training a model able to predict user ratings in such context.

In order to do so, we utilize a dataset of non task-oriented spoken conversations between Amazon Alexa users and Roving Mind, the open-domain system for the Amazon Alexa Prize Challenge 2017 (Ram et al., 2017) described in Chapter 3. As an upper bound for the rating prediction task, we re-annotate a sample of the corpus using experts and analyse the correlation between expert and user ratings. Afterwards, we analyse the entire corpus using well-known automatically extractable features (user sentiment, Dialogue Acts (both user and machine), conversation length and average user turn length), which show a low, but still significant correlation with user ratings. We show how different combinations of these features together with a LSA representation of the user turns can be used to train a regression model whose predictions also yield a low, but significant correlation with user ratings. Our results indicate the difficulty of predicting how users might rate interactions with a conversational agent.

6.1.2 Data Collection

The dataset analysed in this Section was collected over a period of 27 days during the Alexa Prize 2017 semifinals and consists of conversations between our system Roving Mind and Amazon Alexa users of the United States. The users could end the conversation whenever they wanted, using a command. At the end of the interaction users were asked to rate a conversation on a 1 (not satisfied at all) to 5 (very satisfied) Likert scale. Out of all the rated conversations, we selected the ones

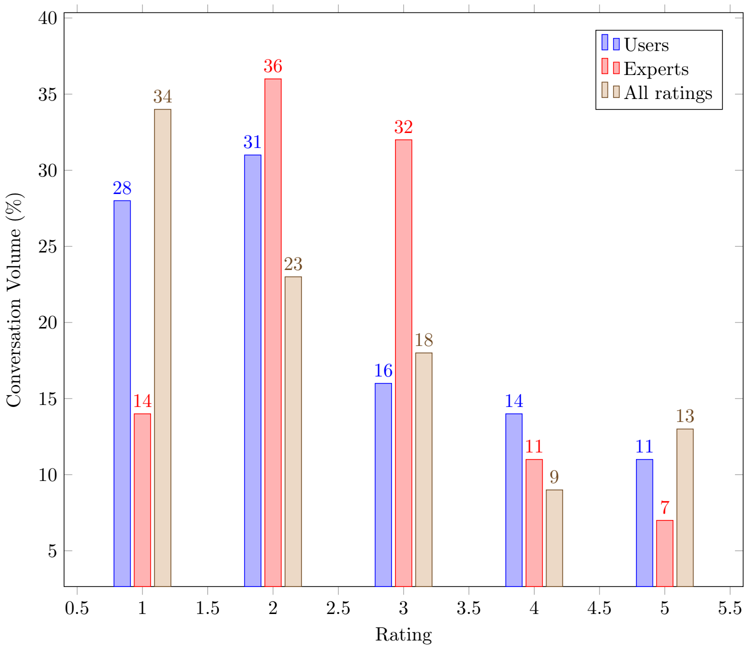


FIGURE 6.1: Distribution of user and expert ratings on the annotated random sample of 100 conversations (test set) compared to the distribution of ratings in the entire dataset ("All ratings"). For clarity of presentation, from the latter we excluded the small portion of non integer ratings (2.3% of the dataset).

longer than 3 turns to yield 4,967 conversations. Figure 6.1 shows the distribution (in percentages) of the ratings in our dataset. The large majority of conversations are between a system and a "first-time" users, as only 5.25% of users had more than one conversation.

6.1.3 Methodology

In this section we describe conversation representation features, experimentation, and evaluation methodologies used in this Section.

6.1.4 Conversation Representation Features

Since in the competition the objective of the system was to entertain users, we expect the ratings to reflect how much they have enjoyed the interaction. User "enjoyment" can be approximated using different metrics that do not require manual annotation, such as conversation length (in turns), mean turn length (in words), assuming that the more users enjoy the conversation the longer they talk; sentiment polarity – hypothesizing that enjoyable conversations should carry a more positive sentiment. While length metrics are straightforward to compute, the sentiment score is computed using a lexicon-based approach (Kennedy and Inkpen, 2006).

Another representation that could shed a light on enjoyable conversations is Dialogue Acts (DA) of user and machine utterances. DAs are frequently used as a generic representation of intents and the considered labels often include *thanking*, *apologies*, *opinions*, *statements* and alike. Relative frequencies of these tags potentially can be useful to distinguish good and bad conversations. The DA tagger we use is the one described in Chapter 4 trained on the Switchboard Dialogue Acts corpus

(Stolcke et al., 2000a), a subset of Switchboard (Godfrey, Holliman, and McDaniel, 1992) annotated with DAs (42 categories), using Support Vector Machines. The user and machine DAs are considered as separate vectors and assessed both individually and jointly.

Additional to Dialogue Acts, sentiment and length features, we experiment with word-based text representation to insert information about the topic of the conversation. Latent Semantic Analysis (LSA) is used to convert a conversation to a vector. First, we construct a word-document co-occurrence matrix and normalize it. Then, we reduce the dimensionality to 100 by applying Singular Value Decomposition (SVD).

6.1.5 Correlation Analysis Methodology

The two widely used correlation metrics are Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SRCC). While the former evaluates the linear relationship between variables, the latter evaluates the monotonic one.

The metrics are used to assess correlations of different conversation features, such as sentiment score or conversation length, with the provided human ratings for those conversations; as well as to assess the correlation of the predicted scores of the regression models to those ratings. For the assessment of the correlation of both features and regression models raw rating predictions are used.

6.1.6 Prediction Methodology

Using the conversation features described above, we train regression models to predict human ratings. We experiment with both Linear Regression and Support Vector Regression (SVR) with radial basis function (RBF) kernel using scikit-learn (Pedregosa et al., 2011). Since the latter consistently outperforms the former, we report only the results for the SVR. The performance of the regression models is evaluated using the standard metrics of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Additionally, we compute Pearson and Spearman's Rank Correlation Coefficients for the predictions with respect to the reference human ratings.

We experiment with the 10-fold cross-validation setting. The performance of the regression models is compared to two baselines: (1) mean baseline, where all instances in the testing fold are assigned as a score the mean of the training set ratings, and (2) chance baseline, where an instance is randomly assigned a rating from 1 to 5 with respect to their distribution in the training set. The models are compared for statistical significance to these baselines using paired two-tail T-test with $p < 0.05$. In Section 6.1.9 we report average RMSE and MAE as well as average correlation coefficients.

6.1.7 Upper bound

Since human ratings are inherently subjective, and different users can rate the same conversation differently, it is difficult to expect the models to yield perfect correlations or very low RMSE and MAE. In order to test this hypothesis two human experts (members of our Alexa Prize team) were asked to rate a random subset of the corpus (100 conversations). The rating distributions for both experts and users on the sample is reported in Figure 6.1. We observe that expert ratings tend to be closer to the middle of the Likert scale (i.e. from 2 to 4), while users had more conversations with ratings at both extremes of the scale (i.e. 1 and 5).

	RMSE	MAE	PCC	SRCC
Exp 1 vs. Exp 2	0.875	0.660	0.705	0.694
Exp 1 vs. Users	1.225	0.966	0.538	0.526
Exp 2 vs. Users	1.286	1.016	0.401	0.370

TABLE 6.1: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Pearson (PCC) and Spearman's rank (SRCC) correlation coefficients among user and expert ratings.

Feature	**PCC**	**SRCC**
Conversation Length	0.133**	0.111**
Av. User Turn Length	-0.068**	-0.079**
User Sentiment	0.071**	0.088**
User Dialogue Acts		
yes-answer	0.081**	0.088**
appreciation	0.070**	0.115**
thanking	0.062**	0.089**
action-directive	-0.069**	-0.052**
statement-non-opinion	0.050**	0.037*
...		
Machine Dialogue Acts		
yes-no-question	0.042**	0.038**
statement-opinion	-0.027*	-0.032*
...		

TABLE 6.2: Pearson (PCC) and Spearman's rank (SRCC) correlation coefficients for conversation lengths, sentiment score, and user and machine Dialogue Acts. Correlations significant with $p < 0.05$ are marked with * and $p < 0.01$ with **.

The RMSE, MAE and Pearson and Spearman's rank correlation coefficients of expert and user ratings are reported in Table 6.1. We observe that the experts tend to agree with each other more than they agree individually with users, since compared to each other the experts have the highest Pearson and Spearman correlation scores (0.705 and 0.694, respectively) and the lowest RMSE and MAE (0.875 and 0.660, respectively). The fact that expert ratings do not correlate with user ratings as well as they correlate among themselves, confirms the difficulty of the task of predicting subjective user ratings even for humans.

6.1.8 Correlation Analysis Results

The results of the correlation analysis are reported in Table 6.2. From the table, we can observe that conversation length has a positive correlation with human judgment, while the average user turn length has a negative correlation. The positive correlation with conversation length confirms the expectation that users tend to have longer conversations with the system when they enjoy it. The negative correlation with average user turn length, on the other hand, is unexpected. As expected, sentiment score has a significant positive correlation with human judgments.

Due to the space considerations, we report only a portion of the DAs that have significant correlations with human ratings. The analysis confirms our expectations that user DAs, such as *thanking* and *appreciation*, have significant positive correlations.

	RMSE	MAE	PCC	SRCC
BL: Chance	1.967	1.535	0.007	0.023
BL: Mean	1.382	1.189	N/A	N/A
Lengths	1.400	1.116*	0.153*	0.158**
Sentiment	1.423	1.128*	0.109*	0.122*
DA: user	1.378	1.106*	0.213**	0,207**
DA: machine	1.418	1.129*	0.104*	0.099*
DA: user+machine	1.375	1.106*	0.219**	0.211**
LSA	**1.350***	**1.075***	0.299**	0.288**
All - LSA	1.366*	1.100*	0.240**	0.230**
All	**1.350***	1.078*	**0.303****	**0.290****

TABLE 6.3: 10 fold cross-validation average Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Pearson (PCC) and Spearman's rank (SRCC) correlation coefficients for regression models. RMSE and MAE significantly better than the baselines are marked with *. Correlations significant with $p < 0.05$ are marked with * and $p < 0.01$ with **.

We also observe that the *action-directive* DA has a negative correlation. Since this DA label covers the turns where a user issues control commands to the system, we hypothesize this correlation could be due to the fact that in such cases users were using a task-oriented approach with our system which was instead designed for chitchat and might therefore feel disappointed (e.g. requesting the Roving Mind system to perform actions it was not designed to perform, such as playing music).

Regarding machine DAs, we observe that even though some DAs exhibit significant correlations, overall they are lower than user DAs. In particular, *yes-no-question* has a significant positive correlation with human judgments, indicating that some users appreciate machine initiative in the conversation. The analysis confirms the utility of length and sentiment features, as well as the importance of some DAs (generic intents) for estimating user ratings.

6.1.9 Prediction Results

The results of the experiments using 10-fold cross-validation and Support Vector Regression are reported in Table 6.3. We report performances of each feature representation is isolation and their combinations. We consider two baselines – chance and mean. For the chance baseline an instance is randomly assigned a rating with respect to the training set distribution. For the mean baseline, on the other hand, all the instances are assigned the mean of the training set as a rating. The mean baseline yields better RMSE and MAE scores; consequently, we compare the regression models to it.

Sentiment and length features (conversation and average user turn) both yield RMSE higher than the mean baseline and MAE significantly lower than it. Nonetheless, their predictions have significant positive correlations with reference human ratings. The picture is similar for the models trained on user and machine DAs alone and their combination. The RMSE scores are higher or insignificantly lower and MAE scores are significantly lower than the mean baseline.

For the LSA representation of conversations we consider ngram sizes between 1 and 4. The representation that considers 4-grams and the SVD dimension of 100 yields better performances; thus, we report the performances of this models only, and use it for feature combination experiments. The LSA model yields significantly lower error both in terms of RMSE and MAE. Additionally, the correlation of the predictions is higher than for the other features (and combinations).

The regression model trained on all features but LSA, yields performances significantly better than the mean baseline. However, they are inferior to that of LSA alone. Combination of all the features retains the best RMSE of the LSA model, but achieves a little worse MAE score. While it yields the best Pearson and Spearman's rank correlation coefficients among all the models, the difference from LSA only model is not statistically relevant using Fisher r-to-z transformation.

6.1.10 Conclusions

In this work we experimented with a set of automatically extractable black-box features which correlate with the human perception of the quality of interactions with a conversational agent. Furthermore, we showed how these features can be combined to train automatic non-task-oriented dialogue evaluation models which correlate with human judgments without further expensive annotations.

The results of our experiments and analysis contribute to the body of observations that indicate that there still remains a lot of research to be done in order to understand characteristics of enjoyable conversations with open-domain non-task oriented agents. In particular, our analysis of expert vs. user ratings suggests that the task of estimating subjective user ratings is a difficult one, since the same conversation might be rated quite differently.

6.2 Turn-level

6.2.1 Introduction

Due to recent advances in spoken language understanding and automatic speech recognition, conversational interfaces such as Alexa, Cortana, and Siri have become increasingly common. While these interfaces are task oriented, there is an increasing interest in building conversational systems that can engage in more social conversations. Building systems that can have a general conversation in an open domain setting is a challenging problem, but it is an important step towards more natural human-machine interactions.

Recently, there has been significant interest in building non-task-oriented dialogue models, also known as chatbots (Sordoni et al., 2015; Wen et al., 2015) fueled by the availability of dialogue data sets such as Ubuntu, Twitter, and Movie dialogs (Lowe et al., 2015; Ritter, Cherry, and Dolan, 2011; Danescu-Niculescu-Mizil and Lee, 2011). However, as most chatbots are text-based, work on human-machine spoken dialogue is relatively under-explored, partly due to lack of such dialogue corpora. Spoken dialogue poses additional challenges such as automatic speech recognition errors and divergence between spoken and written language.

Sequence-to-sequence (seq2seq) models (Sutskever, Vinyals, and Le, 2014) and their extensions (Luong, Pham, and Manning, 2015; Sordoni et al., 2015; Li et al., 2015), which are used for neural machine translation (MT), have been widely adopted for dialogue generation systems. In MT, given a source sentence, the correctness of the target sentence can be measured by semantic similarity to the source sentence. However, in open-domain conversations, a generic utterance such as "sounds good" could be a valid response to a large variety of statements. These seq2seq models are commonly trained on a maximum likelihood objective, which leads the models to place uniform importance on all user utterance and system response pairs. Thus, these models usually choose "safe" responses as they frequently appear in the dialogue training data. This phenomenon is known as the *generic response problem.* These responses, while arguably correct, are bland and convey little information leading to short conversations and low user satisfaction.

Since response generation systems are trained by maximizing the average likelihood of the training data, they do not have a clear signal on how well the current conversation is going. We hypothesize that having a way to measure conversational success at every turn could be valuable information that can guide system response generation and help improving system quality. Such a measurement may also be useful for combining responses from various competing systems. To this end, we build a supervised conversational evaluator to assess two aspects of responses: engagement and coherence. The input to our evaluators are encoded conversations represented as fixed-length vectors as well as hand-crafted dialogue and turn level features. The system outputs explicit scores on coherence and engagement of the system response.

We experiment with two ways to incorporate these explicit signals in response generation systems. First, we use the evaluator outputs as input to a reranking model, which are used to rescore the n-best outputs obtained after beam search decoding. Second, we propose a technique to incorporate the evaluator loss directly into the conversational model as an additional discriminatory loss term. Using both human and automatic evaluations, we show that both of these methods significantly improve the system response quality. The combined model utilizing re-ranking and the composite loss outperforms models using either mechanism alone.

The contributions of this work are two-fold. First, we experiment with various hand-crafted features and conversational encoding schemes to build a conversational evaluation system that can provide explicit turn-level feedback to a response generation system on the highly subjective task. This system can be used independently to compare various response generation systems or as a signal to improve response generation. Second, we experiment with two complementary ways to incorporate explicit feedback to the response generation systems and show improvement in dialogue quality using automatic metrics as well as human evaluation.

6.2.2 Related Works

There are two major themes in this work. The first is building evaluators that allow us to estimate human perceptions of coherence, topicality, and interestingness of responses in a conversational context. The second is the use of evaluators to guide the generation process. As a result, this work is related to two distinct bodies of work.

Automatic Evaluation of Conversations: Learning automatic evaluation of conversation quality has a long history (Walker et al., 1997). However, we still do not have widely accepted solutions. Due to the similarity between conversational response generation and MT, automatic MT metrics such as BLEU (Papineni et al., 2002) and

METEOR (Banerjee and Lavie, 2005) are widely adopted for evaluating dialog generation. ROUGE (Lin and Hovy, 2003), which is also used for chatbot evaluation, is a popular metric for text summarization. These metrics primarily rely on token-level overlap over a corpus (also synonymy in the case of METEOR), and therefore are not well-suited for dialogue generation since a valid conversational response may not have any token-level or even semantic-level overlap with the ground truths. While the shortcomings of these metrics are well known for MT (Graham, 2015; Espinosa et al., 2010), the problem is aggravated for dialogue generation evaluation because of the much larger output space (Liu et al., 2016; Novikova et al., 2017). However, due to the lack of clear alternatives, these metrics are still widely used for evaluating response generation (Ritter, Cherry, and Dolan, 2011; Lowe et al., 2017a). To ensure comparability with other approaches, we report results on these metrics for our models.

To tackle the shortcomings of automatic metrics, there have been efforts to build models to score conversations. Lowe et al. (2017a) train a model to predict the score of a system response given a dialogue context. However, they work with tiny data sets (around 4000 sentences) in a non-spoken setting. Tao et al. (2017) address the expensive annotation process by adding in unsupervised data. However, their metric is not interpretable, and the results are also not shown on a spoken setting. Our work differs from the aforementioned works as the output of our system is interpretable at each dialogue turn.

There has also been work on building evaluation systems that focus on specific aspects of dialog. Li et al. (2016c) use features for information flow, Yu et al. (2016b) use features for turn-level appropriateness. However, these metrics are based on a narrow aspect of the conversation and fail to capture broad ranges of phenomena that lead to a good dialog.

Improving System Response Generation: Seq2Seq models have allowed researchers to train dialogue models without relying on handcrafted dialogue acts and slot values. Using maximum mutual information (MMI) (Li et al., 2015) was one of the earlier attempts to make conversational responses more diverse (Serban et al., 2016d; Serban et al., 2016a). Shao et al. (2017) use a segment ranking beam search to produce more diverse responses. Our method extends the strategy employed by Shao et al. (2017) utilizing a trained model as the reranking function and is similar to Holtzman et al. (2018) but with different kind of trained model.

More recently, there have been works which aim to alleviate this problem by incorporating conversation-specific rewards in the learning process. Yao et al. (2016) use the IDF value of generated sentences as a reward signal. Xing et al. (2017) use topics as an additional input while decoding to produce more specific responses. Li et al. (2016b) add personal information to make system responses more user specific.Li, Monroe, and Jurafsky (2017) use distillation to train different models at different levels of specificity and use reinforcement learning to pick the appropriate system response. Zhou et al. (2017) and Zhang et al. (2018) introduce latent factors in the seq2seq models that control specificity in neural response generation. There has been recent work which combines responses from multiple sub-systems (Serban et al., 2017a; Papaioannou et al., 2017) and ranks them to output the final system response. Our method complements these approaches by introducing a novel learned-estimator model as the additional reward signal.

6.2.3 Data

The data used in this study was collected during the Alexa Prize (Ram et al., 2017) competition and shared with the teams who were participating in the competition. Upon initiating the conversation, users were paired with a randomly selected socialbot built by the participants. At the end of the conversation, the users were prompted to rate the socialbot quality, from 1–5, with 5 being the highest.

We randomly sampled more than 15K conversations (approximately 160K turns) collected during the competition. These were annotated for coherence and engagement (See Section 6.2.3) and used to train the conversation evaluators. For training the response generators, we selected highly-rated user conversations, which resulted in around 370K conversations containing 4M user utterances and their corresponding system response. One notable statistic is that user utterances are typically very short (mean: 3.6 tokens) while the system responses generally are much longer (mean: 23.2 tokens).

Annotations

Asking annotators to measure coherence and engagement directly is a time-consuming task. We observed that we could collect data much faster if we asked direct "yes" or "no" questions to our annotators. Hence, upon reviewing a user-chatbot interaction along with the entire conversation to the current turn, annotators[3] rated each chatbot response as "yes" or "no" on the following criteria:

- **The system response is comprehensible:** The information provided by the chatbot made sense with respect to the user utterance and is syntactally correct.
- **The system response is on topic:** The chatbot response was on the same topic as the user utterance or was relevant to the user utterance. For example, if a user asks about a baseball player on the LA Dodgers, then the chatbot mentions something about the baseball team.
- **The system response is interesting:** The chatbot response contains information which is novel and relevant. For example, the chatbot would provide an answer about a baseball player and give some additional information to create a fleshed-out response.
- **I want to continue the conversation:** Given the current state of the conversation and the system response, there is a natural way to continue the conversation. For example, this could be due to the system asking a question about the current conversation subject.

We use these questions as proxies for measuring coherence and engagement of responses. The answers to the first two questions ("comprehensible" and "on topic") are used as a proxy for coherence. Similarly, the answer to the last two questions ("interesting" and "continue the conversation") are used as a proxy for engagement.

6.2.4 Conversation Evaluators

We train conversational response evaluators to assess the state of a given conversation. Our models are trained on a combination of utterance and response pairs combined with context (past turn user utterances and system responses) along with

[3] The data was collected through mechanical turk. Annotators were presented with the full context of the dialogue up to the current turn.

Model	TREC	SUBJ	STS
Average Embeddings	0.80	0.90	0.45
Transformer	0.83	0.91	0.48
BiLSTM	0.84	0.90	0.45

TABLE 6.4: Sentence embedding performance.

other features, e.g., dialogue acts and topics as described in Section 6.2.5. We experiment with different ways to encode the responses (Section 6.2.4) as well as with different feature combinations (Figure 6.2).

Sentence Embeddings

We pretrained models that produce sentence embeddings using the ParlAI chitchat data set (Miller et al., 2017). We use the Quick-Thought (QT) loss (Logeswaran and Lee, 2018) to train the embeddings. Our word embeddings are initialized with FastText (Bojanowski et al., 2016) to capture the sub-word features and then fine-tuned. We encode sentences into embeddings using the following methods:

a) Average of word embeddings (300 dim)

b) The Transformer Network (1 layer, 600 dim) (Vaswani et al., 2017)

c) Concatenated last states of a BiLSTM (1 layer, 600 dim)

The selected dimensions and network structures followed the original paper (Vaswani et al., 2017). All models were trained with a batch size of 400 using Adam optimizer with learning rate of 5e-4.

To measure the sentence embedding quality, we evaluate our models on a few standard classification tasks. The models are used to get sentence representation, which are passed through feedforward networks that are trained for the following classification tasks: (i) Semantic Textual Similarity (STS) (Marelli et al., 2014), (ii) Question Type Classification (TREC) (Voorhees and Dang, 2003), (iii) Subjectivity Classification (SUBJ) (Pang and Lee, 2004). Table 6.4 shows the different models' performances on these tasks. Based on this, we choose the Transformer as our sentence encoder as it was overall the best performing while being fast.

Context

Given the contextual nature of the problem we extracted the sentence embeddings of user utterances and responses for the past 5 turns and used a 1 layer LSTM with 256 hidden units to encode conversational context. The last state of LSTM is used to obtain the encoded representation, which is then concatenated with other features (Section 6.2.5) in a fully-connected neural network.

6.2.5 Features

Apart from sentence embeddings and context, the following features are also used:

- **Dialogue Act:** Serban et al. (2017a) show that Dialogue act (DA) features could be useful for response selection rankers. Following this, we use model (Khatri et al., 2018)-predicted DAs (Stolcke et al., 1998) of user utterances and system responses as an indicator feature.

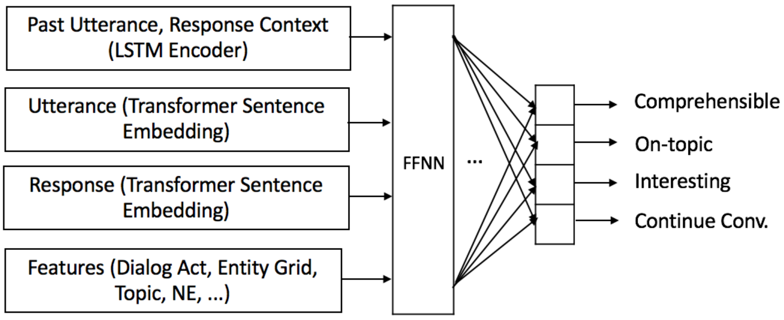

Past Utterance, Response Context (LSTM Encoder)
Utterance (Transformer Sentence Embedding)
Response (Transformer Sentence Embedding)
Features (Dialog Act, Entity Grid, Topic, NE, ...)
FFNN
...
Comprehensible
On-topic
Interesting
Continue Conv.

FIGURE 6.2: Conversation Evaluators

- **Entity Grid:** In Section 5.1 we showed that entities and DA transitions across turns can be strong features for assessing dialogue coherence. Starting from a grid representation of the turns of the conversation as a matrix (DAs × entities), these features are designed to capture the patterns of topic and intent shift distribution of a dialog. We employ the same strategy for our models.
- **Named Entity (NE) Overlap:** We use named entity overlap between user utterances and their corresponding system responses as a feature. Our named entities are obtained using SpaCy[4]. Papaioannou et al. (2017) have also used similar NE features in their ranker.
- **Topic:** We use a one-hot representation of a dialogue turn topic predicted by a conversational topic model (Guo et al., 2017b) that classifies a given dialogue turn into one of 26 pre-defined classes like Sports and Movies.
- **Response Similarity:** Cosine similarity between user utterance embedding and system response embedding is used as a feature.
- **Length:** We use the token-level length of the user utterance and the response as a feature.

The above features were selected from a large pool of features through significance testing on our development set. The effect of adding these features can be seen in Table 6.5. Some of the features such as **Topic** lack previous dialogue context, which could be updated to include the context. We leave this extension for future work.

Evaluator	'Yes' Class Distr.	Accuracy	Precision	Recall	F-score	MCC
Comprehensible	0.80	0.84 (+3%)	0.83 (+1%)	0.85 (+15%)	0.84 (+8%)	0.37 (+107%)
On-topic	0.45	0.64 (+9%)	0.65 (+10%)	0.64 (+18%)	0.64 (+13%)	0.29 (+81%)
Interesting	0.16	0.83 (-1%)	0.77 (+10%)	0.80 (-5%)	0.78 (+2%)	0.12 (+inf%)
Cont. Conversation	0.71	0.75 (+4%)	0.73 (+5%)	0.72 (+31%)	0.72 (+17%)	0.32(+179%)

TABLE 6.5: Conversation Evaluators Performance. Numbers in parentheses denote relative changes when using our best model (all features) with respect to the baseline (no handcrafted features, only sentence embeddings). Second column shows the class imbalance in our annotations. Note that the baseline model had 0 MCC for Interesting

[4] https://spacy.io/

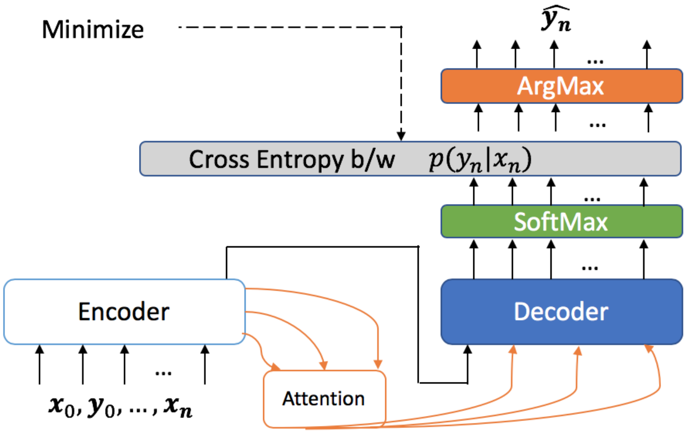


FIGURE 6.3: Baseline Response Generator (Seq2Seq with Attention)

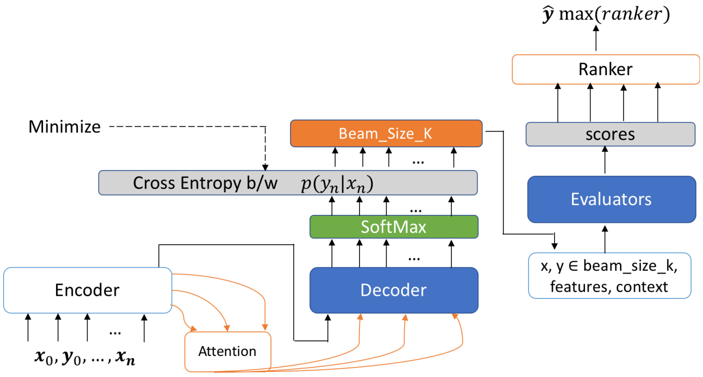


FIGURE 6.4: Reranking Using Evaluators. Top 15 candidates from beam search are passed to the evaluators. The candidate that maximizes the reranker score is chosen as the output. Encoder-decoder remain unchanged)

Models

Given the large number of features and their non-sequential nature, we train four binary classifiers using feedforward neural networks (FFNN). The input to these models is a dialogue turn. Each output layer is a softmax function corresponding to a binary decision for each evaluation metric forming a four-dimensional vector. Each vector dimension corresponds to an evaluation metric (See Section 6.2.3). For example, one possible reference output would be [0,1,1,0], which corresponds to "not comprehensible," "on topic," "interesting," and "I don't want to continue."

We experimented with training the evaluators jointly and separately and found that training them jointly led to better performance. We suspect this is due to the objectives of all evaluators being closely related. We concatenate the aforementioned features as an input to a 3-layer FFNN with 256 hidden units. Figure 6.2 depicts the architecture of the conversation evaluators.

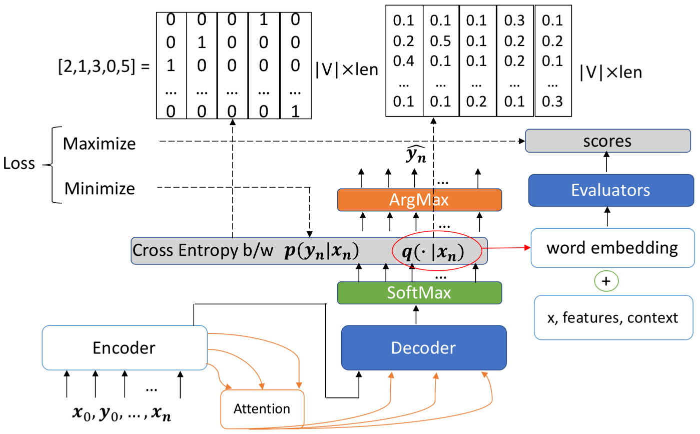


FIGURE 6.5: Response Model Configurations. The baseline is shown at the top. The terms x_n and y_n correspond to n^{th} utterance and response respectively.

6.2.6 Response Generation System

To incorporate the explicit turn level feedback provided by the conversation evaluators, we augment our baseline response generation system with the softmax scores provided by the conversation evaluators. Our baseline response generation system is described in Section 6.2.6. We then incorporate evaluators outputs using two techniques: reranking and fine-tuning.

Base Model (S2S)

We extended the approach of Yao et al. (2016) where the authors used Luong's dot attention (Luong, Pham, and Manning, 2015). In our experiments, the decoder uses the same attention (Figure 6.3). As we want to observe the full impact of conversational evaluators, we do not incorporate inverse document frequency (IDF) or conversation topics into our objective. Extending the objective to include these terms can be a good direction for future work.

To make the response generation system more robust, we added user utterances and system responses from the previous turn as context. The input to the response generation model is previous-turn user utterance, previous-turn system response, and current-turn user utterance concatenated sequentially. We insert a special transition token (Serban et al., 2016d) between turns. We then use a single RNN to encode these sentences. Our word embeddings are randomly initialized and then fine-tuned during training. We used a 1-layer Gated Recurrent Neural network with 512 hidden units for both encoder and decoder to train the seq2seq model and MLE as our training objective.

Reranking (S2S_RR)

In this approach, we do not update the underlying encoder-decoder model. We maintain a beam to get 15-best candidates from the decoder. The top candidate out of the 15 candidates is equivalent to the output of the baseline model. Here, instead of selecting the top output, the final output response is chosen using a reranking model.

For our reranking model, we calculate BLEU scores for each of the 15 candidate responses against the ground truth response from the chatbot. We then sample two responses from the *k*-best list and train a pairwise response reranker. The response with the higher BLEU is placed in the positive class (+1) and the one with lower BLEU is placed in the negative class (-1). We do this for all possible candidate combinations from the 15-best responses. We use the max-margin ranking loss to train the model. The model is a three-layered FFNN with 16 hidden units.

The input to the pairwise reranker is the softmax output of the 4 evaluators as shown in Figure 6.2. The input to the evaluators are described in Section 6.2.4. The output of the reranker is a scalar, which, if trained right, would give a higher value for responses with higher BLEU scores. Figure 6.4 depicts the architecture of this model.

Fine-tuning (S2S_FT)

In this approach, we use evaluators as a discriminatory loss to fine-tune the baseline encoder-decoder response generation system. We first train the baseline model and then, it is fine-tuned using the evaluator outputs in the hope of generating more coherent and engaging responses. One issue with MLE is that the learned models are not optimized for the final metric (e.g., BLEU). To combat this problem, we add a discriminatory loss in addition to the generative loss to the overall loss term as shown in Equation 6.1.

$$loss = \sum_{i=1}^{len} p(y_{ni}|z_n) log(q(\hat{y}_{ni}|z_n)) - \lambda||Eval(x_n, q(.|z_n)||_1 \tag{6.1}$$

where $z_n = x_n, y_{n-1}, \ldots, x_0, y_0$ is the conversational context where n is the context length. $q \in R^{|V| \times len}$ of the first term corresponds to the softmax output generated by the response generation model. The term $\hat{y}_{ni}$ refers to its corresponding decoder response at n_{th} conversation turn and i^{th} word generated. In the second term, the function *Eval* refers to the evaluator score produced for a user utterance, x_n, and decoder softmax output, q.

In Equation 6.1, the first term corresponds to the cross-entropy loss from the encoder-decoder while the second term corresponds to the discriminative loss from the evaluator. In a standalone evaluation setting, the evaluator will take one hot representation of the user utterance as input, i.e., the input is *len*-tokens long which is passed through an embedding lookup layer which makes it $\mathbb{R}^{D \times len}$ input to rest of the network where D is the size of the word embeddings. To make the loss differentiable, instead of performing *argmax* to get a decoded token, we use the output of the softmax layer (distribution of likelihood across entire vocabulary for output length, i.e., $\mathbb{R}^{|V| \times len}$) and use this to do a weighted embedding lookup across the entire vocabulary to get the same $\mathbb{R}^{D \times len}$ matrix as an input to rest of the evaluator network. Our updated evaluator input becomes the following:

$$\mathbb{R}^{D \times len} = \mathbb{R}^{D \times |V|} \times \mathbb{R}^{|V| \times len} \tag{6.2}$$

The evaluator score is defined as the sum of softmax outputs of all 4 models.We keep the rest of the input (context and features) for the evaluator as is.

Metric	Pearson Corr	p-value
Comprehensible	0.2	$\ll 0.001$
On-topic	0.4	$\ll 0.001$
Interesting	0.25	$\ll 0.001$
Cont. Conversation	0.3	$\ll 0.001$

TABLE 6.6: Evaluators Correlation with Turn-level Ratings

We weight the discriminator score by λ, which is a hyperparameter. We selected λ to be 10 using grid search to optimize for final BLEU on our development set. Figure 6.5 depicts the architecture of this approach. The decoder is fine-tuned to maximize evaluator scores along while minimizing the cross-entropy loss. The evaluator model is trained on the original annotated corpus and parameters are frozen.

6.2.7 Reranking + Fine-tuning (S2S_RR_FT)

We also combined fine-tuning with reranking, where we obtained the 15 candidates from the fine-tuned response generator and then we select the best response using the reranker, which is trained to maximize the BLEU score.

6.2.8 Experiments and Results

Conversation Evaluators

The conversation evaluators were trained using cross-entropy loss. We used a batch size of 128, dropout of 0.3 and Adam optimizer with a learning rate of 5e-5 for our conversational evaluators. Sentence embeddings for user utterances and system responses are obtained using the fast-text embeddings and Transformer network.

Table 6.5 shows the evaluator performance compared with a baseline with no handcrafted features. We present precision, recall, and f-score measures along with the accuracy. Furthermore, since the class distribution of the dataset is highly imbalanced, we also calculate Matthews correlation coefficient (MCC)(Matthews, 1975), which takes into account true and false positives and negatives. It is a balanced measure which can be used even if the classes sizes are very different. With the proposed features we observe significant improvement across all metrics.

We also performed a correlation study between the model predicted scores and human annotated scores (1 to 5) on 2000 utterances. The annotators[5] were asked to answer a single question: "On a scale of 1–5, how coherent and engaging is this response given the previous conversation?" From Table 6.6, it can be observed that evaluator predicted scores has significant correlation (moderate to high) with the overall human evaluation score on this subjective task (0.2 – 0.4 Pearson correlation with turn-level ratings). Considering the substantial individual differences in evaluating open-domain conversations, we observe that our evaluators with moderate level of correlation can be used to provide turn-level feedback for a human-chatbot conversation.

Response Generation

We first trained the baseline model (S2S) on the conversational data set (4M utterance-response pairs from the competition. Section 6.2.3). The data were split into 80%

[5] Same setup as previously described

Conversation 1	Conversation 2
User: yes Machine: is there something specific you want to discuss User: fox news **Ground Truth**: my developers are working hard to get you recent news you can say topics to hear some topics **S2S(Base)**: I found **S2S_RR**: I heard about this **S2S_FT** : I heard of a man it may be interested in this news mentioned **S2S_RR_FT**: I heard about this news talked about this	User: cool Machine: back at you ok we could talk about movies or music User: movies **Ground Truth**: shall we chat about something else I love talking about music and movies **S2S(Base)**: so do you **S2S_RR**: who would you want to talk **S2S_FT** : what actor love most **S2S_RR_FT**: what actor

TABLE 6.7: Two randomly selected qualitative examples of responses

Metric	BLEU-4	ROUGE-2	Distinct-2
S2S (Base)	5.9	5.1	0.011
S2S_RR	11.6(+97%)	6.3(+24%)	**0.017(+54%)**
S2S_FT	6.2(+5%)	5.3(+4%)	0.011(-1%)
S2S_RR_FT	**12.2(+107%)**	**6.8(+33%)**	0.017(+53%)

TABLE 6.8: Generator performance on automatic metrics.

training, 10% development, and 10% test sets. The baseline model was trained using Adam with learning rate of 1e-4 and batch size of 256 until the development loss converges. The vocabulary of 30K most frequent words were used. And the reranker was trained using the 20K number of beam outputs from the baseline model on the development set. Adam with learning rate of 1e-4 and batch size of 16 was used for the fine-tuning (S2S_FT).

Table 6.8 shows the performance comparison of different generation models (Section 6.2.6) on the Alexa Prize conversational data set. We observed that reranking *n*-best responses using the evaluator-based reranker (S2S_RR) provides nearly 100% improvement in BLEU-4 scores.

Fine-tuning the generator by adding evaluator loss (S2S_FT) does improve the performance but the gains are smaller compared to reranking. We suspect that this is due to the reranker directly optimizing for BLEU. However, using a fine-tuned model and then reranking (S2S_RR_FT) complements each other and gives the best performance overall. Furthermore, we observe that even though the reranker is trained to maximize the BLEU scores, reranking shows significant gains in ROUGE scores as well. We also measured different systems performance using Distinct-2 (Li et al., 2016a), which is the number of unique length-normalized bigrams in responses. The metric can be a surrogate for measuring diverse outputs. We see that our generators using reranking approaches improve on this metric as well. Table 6.7 also shows 2 sampled responses from different models.

To further analyze the impact of reranker trained to optimize on BLEU score, we trained a baseline response generation system on a Reddit data set[6], which comprises of 9 million comments and corresponding response comments. All the hyper-parameter setting followed the setting of training on the Alexa Prize conversational dataset.

We trained a new reranker for the Reddit data using the evaluator scores obtained from the models proposed in Section 6.2.4. We show in Table 6.9 that even though

[6] We use a publicly available data (Baumgartner, 2015).

Metric	**S2S(Base)**	**S2S_RR**
BLEU-4	3.9	7.9 (+103%)
ROUGE-2	0.6	0.8 (+33%)
Distinct-2	0.0047	0.0086 (+82%)

TABLE 6.9: Response Generator on Reddit Conversations. Due to the size of the dataset we could not fine tune these models.

Metric	Coherence	Engagement
S2S(Base)	2.34	1.80
S2S_RR	2.42	2.16
S2S_FT	2.36	1.87
S2S_RR_FT	**2.55**	**2.31**

TABLE 6.10: Mean ratings for Qualitative and Human Evaluation of Response Generators

the evaluators are trained on a different data set, the reranker learns to select better responses nearly doubling the BLEU scores as well as improving on the Distinct-2 score. Thus, the evaluator generalizes in selecting more coherent and engaging responses in human-human interactions as well as human-computer interactions. As fine-tuning the evaluator is computationally expensive, we did not fine-tune it on the Reddit dataset.

The closest baseline that used BLEU scores for evaluation in open-domain setting is from Li et al. (2015) where they trained the models on Twitter data using Maximum Mutual Information (MMI) as the objective function. They obtained a BLEU score of 5.2 in their best setting on Twitter data (average length 23 chars), which is relatively less complex than Reddit (average length 75 chars).

Human Evaluation

As noted earlier, automatic evaluation metrics may not be the best way to measure chatbot response generation performance. Therefore, we performed human evaluation of our models. We asked annotators to provide ratings on the system responses from the models we evaluated, i.e., baseline model, S2S_RR, S2S_FT, and S2S_RR_FT. A rating was obtained on two metrics: coherence and engagement. Coherence measures how much the response is comprehensible and relevant to a user's request and engagement shows interestingness of the response (Venkatesh et al. (2018)). We asked the annotators to provide the rating based on a scale of 1–5, with 5 being the best. We had four annotators rate 250 interactions. Table 6.10 shows the performance of the models on the proposed metrics. Our inter-annotator agreement is 0.42 on Cohen's Kappa Coefficient, which implies moderate agreement. We believe this is because the task is relatively subjective and the conversations were performed in the challenging open-domain setting. The S2S_RR_FT model provides the best performance across all the metrics, followed by S2S_RR, followed by S2S_FT.

6.2.9 Conclusion

Human annotations for conversations show significant variance, but it is still possible to train models which can extract meaningful signal from the human assessment of the conversations. We show that these models can provide useful turn-level guidance to response generation models. We design a system using various features and context encoders to provide turn-level feedback in a conversational dialog. Our feedback is interpretable on two major axes of conversational quality: engagement and coherence. We also plan to provide similar evaluators to the university teams participating in the Alexa Prize competition. To show that such feedback is useful in building better conversational response systems, we propose two ways to incorporate this feedback, both of which help improve on the baselines. Combining both techniques results in the best performance. We view this work as complementary to other recent work in improving dialogue models such as Li et al. (2015) and Shao et al. (2017). While such open-domain models are still in their infancy, we view the framework presented here to be an important step towards building end-to-end coherent and engaging dialogue models.

6.3 Summary

This Chapter presented different evaluation models trained using supervised techniques on noisy human-machine interactions collected during the Alexa Prize competition. In both approaches presented, models rely on DA and a form of entity representation (using topic representations via LDA in the first Section, and the entity grid in the second one).

In the first part of the Chapter, we proposed models to automatically predict user ratings of human-machine interactions at the conversation level without relying on further turn-level annotation. The dataset used were interactions between US Amazon Alexa users and Roving Mind. First, in order to establish an upper bound for the task, given the subjectivity of human judgment, we asked two human experts to reannotate a sample of the conversations with conversational ratings. A comparison of expert and users ratings via correlation analysis confirmed the difficulty of the task. Afterwards, we performed a correlation analysis of the dataset and found moderate, though not significant correlation of user ratings with conversation lengths and DA turn distribution. Then, we proposed models experimenting with a combination of features including DAs and topic representations to predict user ratings. The combination of all features yielded the best results in the experiments.

The second part of the Chapter was dedicated to turn level evaluation and response generation using coherence and engagement. In particular, we proposed to train turn level coherence and engagement predictors with a supervised methodology and then use them to rerank the output of a non-task-oriented open-domain response generation model. First, we described the turn level evaluation models, implemented combining the textual input with a number of features, including DA and entity grid representations. Our experiments on turn level evaluation show that these features are useful in predicting coherence and engagement. Second, we described the different response generation models and proposed two different techniques to augment them with the turn evaluators signal. The first methodology uses directly the evaluators output as input to a reranking model, used to rescore the responses generated after beam decoding. The second methodology incorporates the evaluator loss directly in the response generation model, as an additional discriminatory term in the loss. Then, we reported the results of response generation experiments which show that both proposed techniques significantly improve the system response quality, according to both human and automatic evaluation.

Overall, throughout the Chapter we saw how using noisy data could be a real challenge, given the subjectivity of human judgment. In both works presented we performed additional human annotation besides the annotation used for training, in the first part to assess a human upper bound for the task, in the second part for having a human evaluation of the generation models. These additional experiments further confirmed the difficulty of annotating noisy human-machine conversations, even for humans. Nevertheless, we argue that it is also important to analyse such input, instead of only relying on data collected in more controlled environments, since it comes from real-world applications.

Chapter 7

A case study for open-domain Natural Language Generation

In this Chapter[1], we address the challenge of Natural Language Generation (NLG) for domain-independent Conversational Agents (CAs). The task of NLG as a component in CAs architectures has traditionally been framed following a Meaning-Representation-to-text approach, where models need to generate text conditioned on a structure usually composed of a DA and a list of associated slots. NLG for CAs following this approach mainly focus on task-oriented applications dependant on a single domain relying on limited ontologies with a small amount of slot types, without benefitting from examples that may be available for other domains. However, with the current progression of CAs towards being multi-domain through open-domain this approach, CA applications might rely on larger, more diverse ontologies.

In this Chapter, we explore NLG for CAs with an open-domain setting with larger ontologies compared to the traditional setting. In particular, we propose to frame NLG for open-domain Question-Answering (QA) following a Meaning-Representation-to-text. First, we present experiments assessing the performance of our proposed NLG models with different ontology sizes. Then, we perform multi-task-learning cross-domain experiments across QA and task-oriented dialogue.

7.1 Introduction

In dialogue literature Natural Language Generation (NLG) is framed as the task of generating natural language responses that faithfully convey the semantic information given by a Meaning Representation (MR). A MR is typically a structure consisting of a Dialogue Act (DA) and a list of associated slots. While the DA (Stolcke et al., 2000a; Mezza et al., 2018) expresses the intent of the utterance to be generated (e.g. "inform" in Table 7.1), the slots, organized as slot type-slot value pairs (e.g. *food*:'french' in Table 7.1), represent the information which has to be conveyed in the generated text.

So far statistical NLG for dialogue has mainly been investigated in research for task-oriented applications (e.g. restaurant reservation, bus information) in narrow, controlled environments with limited ontologies, i.e. considering a small set of DAs and slot types (respectively 12 and 8 in the popular San Francisco restaurant dataset (SFX) (Wen et al., 2015), 8 and 1 in the recent E2E NLG challenge (Novikova, Dušek,

[1] The Chapter is based on Cervone et al. (2019).

	Input		Output
	context	**MR**	**Text**
Task Oriented	-	**inform** *name*: ‘fringale’ *food*: ‘french’	‘fringale is a french restaurant’ ‘fringale serves french food’
QA	‘when was kentucky founded’	**inform** *timepoint*: ’1792’ *objStr*: ‘kentucky’ *claStr*: ‘state’ *relStr*: ’founded’	‘1792’ ‘kentucky formed in 1792’ ‘kentucky founded in 1792’

TABLE 7.1: Examples of input-output pairs from a task-oriented (Task) NLG (SFX (Wen et al., 2015)) and a Question-Answering (QA) dataset. In NLG the input is typically a Meaning Representation (MR) and the output is its textual realization (Text). Each MR is composed of a Dialogue Act (bold) and a list of slot type (italic)-value pairs. Compared to most NLG datasets, our QA corpus also has the previous question (context) as input. While in the task-oriented setting we observe a one-to-one relation between slots in the input and the ones realized in the text, the same is not true for QA.

and Rieser, 2017)). Furthermore, most datasets consider MRs in isolation (Novikova, Dušek, and Rieser, 2017) i.e., they lack conversational context, even though the previous utterances in the dialogue have been shown to improve the performance of task-oriented NLG (Dušek and Jurcicek, 2016). These characteristics of current approaches to NLG can be linked to the fact that a vast majority of dialogue NLG research is tested on a single domain where the dialogue agent performs simple tasks such as giving information about a restaurant, with few exceptions (Wen et al., 2016b).

However, with the rise of conversational agents such as Amazon Alexa and Google Assistant, there is an increasing interest in complex multi-domain tasks. These systems typically rely on hand-crafted NLG, but this approach cannot scale to the complex ontologies which may be required in real-world applications (e.g. booking a trip).

In this work we explore the applicability of current NLG models for task-oriented dialogue, based on a MR-to-text framework using Encoder-Decoder architectures, to open-domain QA. This allows us to investigate the performance of current NLG research in an environment with (1) much larger numbers of slot types, and (2) a different application compared to task-oriented dialogue. We generate the QA datasets for our experiments using as source a large corpus of open-domain QA pairs from interactions between real-world users and a conversational agent. For evaluation, we utilize both objective metrics and human judgment. We observe that NLG for open-domain QA poses its own challenges compared to task-oriented dialogue, since correct answers to the same question do not necessarily convey all slot types in the MR (see Table 7.1).

In particular, in our first set of experiments, we investigate the effect of using increasingly larger ontologies with regards to slot types on the performance of our NLG models for QA. We find that, notwithstanding the larger ontologies and the noisiness of our dataset, models' performance does not degrade significantly in terms of naturalness of generated text and efficiency in encoding the MR information (i.e. Slot Error Rate). Interestingly, we find it improves for some of the human evaluation metrics. We also observe that using conversational context improves the quality of generated responses. In our second set of experiments, we investigate whether jointly training NLG models for task-oriented dialogue and QA improves performances. To this end, we experiment with learning NLG models in a multi-task setting between our QA data and SFX. Our experiments show that learning models in a multi-task setting lead to better performances in terms of naturalness of the generated output for both tasks.

This work has several contributions:

1. We apply the MR-to-text framework (typical of NLG for task-oriented dialog) to a open-domain QA application.
2. We explore the importance of adding the previous conversational context to improve the quality of the generated output.
3. We investigate the possibility of learning NLG models using a MR-to-text approach with increasingly larger ontologies in terms of slot types.
4. We experiment with multi-task learning for NLG between open-domain QA and task-oriented dialogue.
5. Finally we also propose new evaluation metrics (see Section 7.5) to capture the variability of output in open-domain QA compared to NLG for task-oriented dialogue.

7.2 Related work

While classical approaches to NLG involve a pipeline of modules such as content selection, planning, and surface realization (Gatt and Krahmer, 2018), recently a large part of the literature investigated end-to-end neural approaches to NLG. The tasks tackled include dialogue, text, and QA. While these tasks share some similarities, each comes with its own set of challenges and requires specific solutions.

NLG for dialogue State of the art NLG models for dialog (Dušek and Jurcıcek, 2016; Juraska et al., 2018) mostly use end-to-end neural Encoder-Decoder approaches with attention (Bahdanau, Cho, and Bengio, 2014) and re-ranking (Dušek, Novikova, and Rieser, 2018). Ensembling is another technique employed to boost model performance (Juraska et al., 2018). Using delexicalization (Henderson, Thomson, and Young, 2014a), i.e., the process of substituting slot values with slot types in the generated text, has also shown improvements in many settings. However, recent work also depicted the disadvantages of delexicalization (Nayak et al., 2017). In our work, we compare and combine both delixecalized and lexicalized inputs for the NLG system.

NLG for dialogue has been mostly tested in controlled environments using task-oriented, single domain datasets with limited ontologies (Wen et al., 2015; Novikova, Dušek, and Rieser, 2017; Balakrishnan et al., 2019). Although Wen et al. (2016b) perform multi-domain task-oriented NLG experiments, the ontologies used are still limited for such settings. Finally, while research has shown how encoding the previous utterance leads to better performances (Dušek and Jurcicek, 2016), most settings consider the turns in isolation (Wen et al., 2015; Novikova, Dušek, and Rieser, 2017).

	Size	Slots	DAs	Words	Domain	Context
E2E	51k	8	1	2453	restaurant	no
SFX	5k	12	8	438	restaurant	no
QA.1	6k	147	1	702	open	yes
QA.2	16k	210	1	1528	open	yes
QA.3	67k	369	1	2963	open	yes

TABLE 7.2: Our QA NLG datasets compared to popular (task-oriented) NLG datasets: San Francisco restaurant (SFX) and the NLG E2E challenge (E2E). We report the full size of datasets in terms of MR-text pairs, the number of slot types, DAs, words (computed after delexicalization), domain and whether the dataset comprises the previous utterance or not.

In our work, we perform open-domain NLG with significantly larger ontologies and also evaluate the impact of adding the context to the input.

NLG for text and QA Recent work around NLG for text involves generating text using structured data using the encoder-decoder networks (Mei et al., 2016). Similarly to dialogue, NLG for text has also been addressed in controlled environments such as weather forecast (Liang, Jordan, and Klein, 2009) with few exceptions (Lebret, Grangier, and Auli, 2016).

In the literature for QA, most approaches retrieve answers directly or generate answers jointly with the retrieval, and answers are usually entities or lists of entities (Dodge et al., 2015). On the contrary, in NLG we assume the answer has already been retrieved, and the goal is to generate text matching it. The field of QA which most strictly relates to our work is answer generation, where current approaches are also based on encoder-decoder networks encoding information directly from a knowledge base (Yin et al., 2016; He et al., 2017; Wei and Zhang, 2019). An additional challenge to answer generation is that there are no publicly available datasets for this task (Fu and Feng, 2018).

Our approach differs from answer generation in that we structure the task as in NLG dialogue literature with a MR-to-text approach.

7.3 Datasets

7.3.1 Question Answering

Source data Our source for generating the MR-text pairs are thousands of open-domain factual question-answer pairs from commercial data. The domains covered in this data are manifold, including geography (e.g. 'is canada bigger than united states' in Table 7.5), history (e.g. 'when was kentucky founded' in Table 7.1), present-day knowledge (e.g. 'will ferrell's wife' in Table 7.5), grammar ('is there a plural form of pegasus') and even mathematics ('what is one modulo seven'). Pairs are grouped according to the type of question asked. Each group consists of a list of specific questions (e.g. "who is the wife of barack obama", "tell me the wives of henry the viii") of the same type (e.g. "who is the wife of") asked by real users to a conversational agent. Each specific question additionally has: (1) the answer to the question (e.g. "michelle obama is obama's wife") generated by the NLG of the conversational system, either using information retrieval or a knowledge base search coupled with templates; (2) relevant noun and verb phrases (e.g. “michelle obama”, ”barack obama”, “wife”) used by the system to generate the answer, including the ones from the question. Noun phrases are tagged according to their semantic type (examples of semantic types are *timepoint* and *human being*), while verb phrases are tagged as “relation” types (see “founded” tagged as *relStr* in Table 7.1).

The answers in the source data are varied, and range from a simple entity to a fully formed answer, as in Table 7.1 example where valid answers to the question “when was kentucky founded” can be “1792” or “kentucky formed in 1792”. This shows an interesting difference between our QA data and task-oriented NLG datasets. While for task-oriented NLG all valid responses for a single MR have the same slot types (i.e., the ones in the input MR), in our dataset this is not always true.

QA NLG datasets We generate the NLG input-output pairs for QA from our source data. In order to perform cross-application experiments, we maintain the same MR-text format as task-oriented dialogue NLG. The target output is the text of the answers in the source data. To generate the input MRs we assumed only one DA across all answers, i.e. “inform”; for the slots, we used the semantic types and relations for noun phrases and verb phrases in the source data as slot types, while the actual entity or verb was used as the corresponding slot value. [2] On top of the generated MR we use, as additional input, the previously asked question as context.

Answers are delexicalized (Henderson, Thomson, and Young, 2014a) to improve generalization. Since we do not have alignment between entities in the input and the generated text, we use a heuristic-based aligner which we also use to filter out data that could not be appropriately aligned. All noun phrases are delexicalized while verb phrases are not. Furthermore, similar to (Juraska et al., 2018), we use delexicalization for data augmentation. We generate additional references for each MR, besides the original one, by considering all delexicalized answers in the question group as candidate template answers for each specific question in the group and then substituting (where possible) slots values which are already available in the input. The text of the previous question is also delexicalized.

Finally, to investigate performances across different ontology sizes, we generate 3 different partitions of the data (QA.1, 2 and 3 in Table 7.2) with a progressively larger number of slot types. Each QA partition was split in train, test and development set (using a 80-10-10 split) according to the type of question asked. We ensured there was no overlap between the different sets to test if models generalize to previously unseen questions.

7.3.2 Task-oriented Dialogue

As a task-oriented NLG corpus for our multi-task learning experiments we use the popular San Francisco restaurants (SFX; Wen et al. (2015)) dataset. Statistics about the dataset is shown in Table 7.2. Although SFX is not large (6k examples), compared to the E2E NLG corpus it presents more variation for DA (although less in style). For all our datasets, we use the TGEN library [3] (Dušek and Jurcıcek, 2016) to delexicalize all slot types except binary values.

[2] Although we use the original tags of the source data, a similar representation could be produced by tagging noun phrases with their Named Entity type and verb phrases with a “relation” slot type.

[3] `https://github.com/UFAL-DSG/tgen`

7.4 Model and Architectures

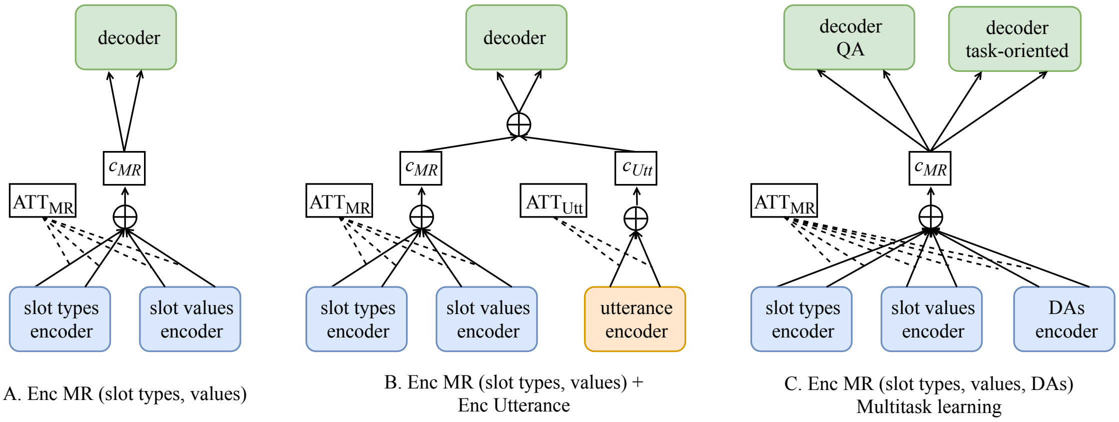


FIGURE 7.1: Our baseline model (A) and the models with the previous utterance (B) and for multi-task learning (C). While our baseline model Enc MR (slot types, values) is composed by two encoders for the MR, one for slot types and one for slot values; our model in subfigure B extends this baseline by adding an encoder for the previous utterance. In the multitask learning setting, on the hand, where we do not have the previous context but might have different dialogue Acts (DAs), we add a corresponding encoder (see subfigure C).

In this section we present the variety of different architectures used in our experiments. Although all our models are based on the Encoder-Decoder framework, we investigate architectures with different number of Encoders (up to 3). Given this variety, for clarity, we follow a template Enc <Encoder type> for naming our different models. The type of Encoder, in particular, can be of Meaning Representation (MR) type, when we encode parts of the MR, such as slot types, values or dialogue Act; or it can be of Utterance type, when we encode the previous utterance context.

Encoder-Decoder with Attention Following recent state-of-the-art approaches to NLG for dialog (Juraska et al., 2018; Balakrishnan et al., 2019), our models are based on the Encoder-Decoder with Attention framework. In particular, we use bidirectional Gated Recurrent Units (GRU) and Luong general attention (Luong, Pham, and Manning, 2015) as our baseline. While we also experimented with other types of architectures, such as using Long-Short-Term Memory Units (Hochreiter and Schmidhuber, 1997) instead of GRUs and different types of attention (including Bahdanau attention (Bahdanau, Cho, and Bengio, 2014) and Luong dot attention (Luong, Pham, and Manning, 2015), this combination gave us the best results for our setting. Depending on the encoder used, either slot type or slot value, we refer to this model as Enc MR (slot types) or Enc MR (slot values).

Multi-Encoder, Single Decoder We expand the baseline (Enc MR) models using multiple inputs from the MR (slot types, values, DAs), each encoded by a different encoder. The attention is performed on their concatenated output to produce the MR context vector c_{MR}. Figure 7.1 A shows an example of such an architecture using two encoders, one for slot types and one for slot values. Furthermore, we experimented with adding the previous utterance as input with an additional encoder (Enc Utterance). In this case, the context vector for the previous utterance c_{Utt} is produced by an independent attention mechanism and the outputs of both attentions (c_{MR} and c_{Utt}) are concatenated (see Figure 7.1 B).

Multi-Encoder, Multi-Decoder We also performed multi-task learning, jointly training the models for both QA and task-oriented NLG. As shown in Figure 7.1 C, we shared the encoders and corresponding input layers across multiple tasks while we maintained multiple decoders for individual tasks. We alternated between minibatches from various data sources to perform multitasking.

7.5 Evaluation

As word overlap metrics may not have a good correlation with human judgment for NLG output evaluation (Stent, Marge, and Singhai, 2005), we use both objective metrics and human evaluation.

Objective metrics Besides the standard BLEU score (obtained using the official E2E NLG challenge evaluation script [4]), we report different types of Slot Error Rate (SER). In dialog NLG approaches SER shows the number of correct slots in the output compared to the input MR. We refer to this metric as SER_{mr} to differentiate it from its modified versions we introduce next. The formula (Wen et al., 2015) is:

$$SER_{mr} = \frac{p_{mr} + q_{mr}}{N_{mr}} \tag{7.1}$$

where N_{mr} is the total number of slots in the input MR and p_{mr} , q_{mr} are respectively the number of missing and redundant slots in the output. This formula works well for task-oriented NLG approaches, but it assumes a one-to-one relationship between the slots in the input MR and the output text. We found this assumption might not hold for our QA datasets where not all slots in the input MR need to be realized for the output to be correct. An example of this is shown in Table 7.1, where the first QA reference text ('1792') would be penalized with 3 missing slots, while still being correct.

In order to capture this different behaviour we designed additional NLG metrics tailored for QA. Slot Error Rate Target (SER_{trg}) is a modification of SER_{mr} where we simply substitute the MR with the main reference text:

$$SER_{trg} = \frac{p_{trg} + q_{trg}}{N_{trg}} \tag{7.2}$$

SER_{trg} is designed to penalize both missing and redundant slots compared to the target sentence. Hence, using SER_{trg} the first QA reference text in Table 7.1 would not be penalized.

Slot Error Rate MultiTarget (SER_{mtrg}), on the other hand, penalizes redundant slots that did not appear in any of the references:

$$SER_{mtrg} = \frac{p_{mtrg}}{N_{mtrg}} \tag{7.3}$$

[4] We do not report other word overlap metrics (e.g., METEOR) computed by the E2E evaluation scripts due to space limitations and correlation with the BLEU score.

where N_{mtrg} are all slots appearing in any reference and p_{mtrg} are the slots in the output that did not appear in any reference sentence. To compute SER_{mtrg} for the model output "kentucky formed in 1792" given the QA MR in Table 7.1 we assume to have two references "1792" and "kentucky formed in 1792". In this case, SER_{mtrg} would consider the output correct as all of its slots appear in at least one of the references.

Human evaluation In all experiments, for each dataset, we selected a sample of 100 MR-text pairs from the test set. Pairs were randomly selected among those where all models under comparison in the experiment had generated different output text. Data for all reported experiments were annotated by 2 human annotators, and final ratings were averaged between the two. In all experiments annotators, presented with MR and all outputs of the systems under comparison, were asked to rate the *naturalness* and *informativeness* of the generated output using a 1-6 Likert score, as in previous NLG dialogue evaluations (Gatt and Krahmer, 2018). Additionally, for the QA datasets annotators had also the previous question as context. Moreover, for the QA datasets annotators were asked to rate how *conversational* the output was, on the same Likert scale, and whether or not the output could ultimately be considered an answer to the question (*answer*), as a binary choice.

7.6 Experimental setup

The hyperparameters chosen for our models were empirically determined through various experiments. Both encoder and decoder in all our models had only one layer, as we noticed additional layers did not give improvements. All embeddings were trained from scratch with a fixed dimension of 50. Models were trained using a cross-entropy loss function and the Adam (Kingma and Ba, 2014) optimizer with a learning rate of 0.001, for 1000 epochs, with early stopping on the validation set. We used mini-batches of size 32.

For the NLG models for QA, experiments on QA.1 (not reported due to space limitations) with different encoders combinations showed that the best performances were achieved using all input types (slot type, value, and previous context) with lexicalized (+ Enc Utterance lex) or delexicalized (+ Enc Utterance delex) previous context in terms of all metrics, except SER_{trg}. On this metric, the architecture with slot types and values, but without the previous context (Enc MR (slot types, values)) achieved the best performance (cf. Table 7.3). For this reason, we chose to report the performances of these architectures in our QA experiments.

7.7 Results

Open domain QA In our first batch of experiments we test various Encoder-Decoder architectures on our 3 different partitions of QA NLG data.

As we can see from Table 7.3, in general, the best performances across all QA datasets for both BLEU and SER_{trg} are achieved by the model using as additional input the lexicalized previous question, followed by the model with the delexicalized one. However, SER_{mr} results show the opposite picture, where the baseline with only slot types and values performs better (except for QA.2 where the score is close to the model with the delexicalized input) and the model with the lexicalized previous utterance is the worst. SER_{mtrg} shows, on the other hand, that the context might slightly degrade performances with bigger ontologies in terms of all text references.

Human evaluation, on the other hand, seems in line with the picture depicted by BLEU and SER_{trg}. Table 7.3 shows the model with the lexicalized context is regarded as the best, closely followed by the model with the delexicalized one in every metric except for *conversational*, where delexicalized is better. This confirms our hypothesis that SER_{mr} might be a less reliable metric to evaluate NLG QA output. Moreover, although we notice a consistent but not drastic degradation in terms of BLEU and SER_{trg} in correlation with bigger ontologies, human evaluation shows an even more gentle degradation between QA.1 and 3 for many metrics. Interestingly, it seems the ability of all models to give a proper answer to the question (*answer*) increases from QA.1 to 3.

	Objective metrics				**Human evaluation**			
	BLEU	**SER_{mr}**	**SER_{trg}**	**SER_{mtrg}**	***Nat.***	***Inf.***	***Conv.***	***Ans.***
QA.1								
Enc MR (slot types, values)	0.85	**0.42**	0.21	0.023	3.73	3.8	3.96	0.36
+ Enc Utterance delex	0.89	0.44	0.19	0.014	4.61	4.63	**4.59**	0.67
+ Enc Utterance lex	**0.95**	0.46	**0.15**	**0.012**	**4.7**	**4.69**	4.48	**0.78**
QA.2								
Enc MR (slot types, values)	0.77	0.44	0.3	0.057	3.88	3.78	4.57	0.38
+ Enc Utterance delex	0.83	**0.43**	0.23	**0.025**	4.64	4.32	**4.88**	0.5
+ Enc Utterance lex	**0.89**	0.47	**0.19**	0.03	**5.15**	**4.88**	4.85	**0.67**
QA.3								
Enc MR (slot types, values)	0.66	**0.43**	0.37	**0.05**	4.29	4.37	4.40	0.73
+ Enc Utterance delex	0.7	0.45	0.32	0.054	4.52	4.47	**4.51**	0.79
+ Enc Utterance lex	**0.72**	0.46	**0.28**	0.067	**4.57**	**4.57**	4.45	**0.80**

TABLE 7.3: Objective metrics and human evaluation results on three QA NLG datasets with increasingly larger ontologies. The models under comparison are a baseline with two encoders, for slot types and slot values, and its extensions with a delexicalised or lexicalised previous utterance. For objective metrics, while for BLEU score the higher the better, for all types of Slot Error Rate (SER) the lower the better. For human evaluation, we report averages of Naturalness (*Nat.*), Informativeness (*Inf.*), and how conversational the response was judged (*Conv.*) on a scale of 1 to 6. Additionally, we report the average of whether responses could be considered an answer to the given question (*Ans.*), given to annotators as a binary choice.

	Dataset	**BLEU**	**SER_{mr}**	**SER_{trg}**	**SER_{mtrg}**	***Nat.***	***Inf.***	***Conv.***	***Ans.***
baseline	SFX	0.727	**0.40**	-	-	4.69	**5.50**	-	-
+ QA.3		**0.74**	0.413	-	-	**5.11**	5.40	-	-
baseline	QA.3	0.659	**0.429**	0.37	**0.05**	4.29	4.33	4.40	**0.73**
+SFX		**0.673**	0.44	**0.368**	0.07	**4.43**	**4.38**	**4.5**	0.72

TABLE 7.4: Objective metrics and human evaluation results of multitask learning experiments combining QA (QA.3) and task-oriented dialogue (SFX) NLG. For all Slot Error Rate (SER) metrics the lower the better.

Multitask learning In our multitask learning experiments we combine the biggest QA dataset, QA.3, with a task-oriented corpus, SFX. We aim to investigate the possibility of transferring knowledge across different NLG systems, notwithstanding the diversity of the data in terms of domain, ontology size, DAs, application (QA vs. task-oriented). Since context is not available in SFX, the model we use has 3 MR Encoders (slot types, values, DAs) and 2 Decoders (one for each task).

Our experiments show that the NLG QA task improves the fluency on SFX both in terms of objective metrics and human evaluation (see Table 7.4). However, training with QA seems to slightly degrade the model efficiency in generating the correct slots. This is to be expected given the difference in the relation between slots in MR and output (one-to-one in SFX, variable in QA.3). As for QA.3 results, it seems the task-oriented NLG task improves QA NLG performances in terms of fluency (BLEU and *Naturalness*) and slot errors (SER_{trg} and *Informativeness*). SER_{mr} and SER_{mtrg}, however, show a slight degradation. We observe task-oriented NLG also makes QA NLG more conversational, however slightly reducing its probability of being an answer to the posed question as well.

Finally, comparing all experiments on QA.3, we notice that although multi-task learning helps, the previous context (either lexicalized or delexicalized) plays a critical role in improving the overall performance.

7.8 Qualitative analysis

In this section we report the qualitative analysis we performed on the human annotated testset. Table 7.5 reports some output examples from different models given the same input MR. In particular, we are interested on the impact of adding various features and multi-tasking.

QA According to our qualitative analysis on the QA datasets, the baseline model is the one with most grammatical errors (e.g. "will ferrell 's 's wife is viveca paulin", "no , canada is not the bigger than united states ."), while in general adding "delex" and "lex" features generates more grammatical responses. This observation was confirmed from both the objective (in terms of BLEU score) and subjective (*naturalness*) evaluations performed.

We also notice how lexicalizing the previous question helps in producing generally correct (e.g. 'unilever.') however shorter answers, which can be regarded as less conversational. Delexicalizing the input, on the other hand, produces more conversational (e.g. 'popsicle's manufacturer is unilever') but also more factually incorrect answers. These observations seem also in line with the subjective evaluation results, which on average reported the best scores for the model with lexicalized previous context (+ Enc Utterance lex) on whether the generated text could be considered an answer to the given question (*answer*), except for the metric rating how *conversational* the output was, for which the model with delexicalized previous context (+ Enc Utterance delex) was regarded as the best one across all QA partitions.

Multitask Looking at the output of the models trained in a multi-task learning setting, we observe that the baseline tends to be more prone to grammatical errors compared to models jointly trained with another task (e.g. in Table 7.5 'sanjalisco

Dataset	Input		baseline	+delex	+lex	multitask
	context	MR				
QA.1	‘sing your song writer’	*human being*: ’vic ruggiero’, *a*:’sing your song’ *b*:’writer’	‘sing your song ’ coach is vic ruggiero .’	‘sing your song ’ writer is vic ruggiero .’	‘vic ruggiero .’	-
QA.1	‘did abraham lincoln have a dad’	*true*:’positive polarity’, *lp*:’dad’ *ro*:’abraham lincoln’	‘yes . abraham lincoln has at least one dad .’	‘yes , abraham lincoln had a mother.’	‘yes , abraham lincoln had a father.’	-
QA.1	‘what is the masters starting date’	*timepoint*:’1999’, *a*:’the masters’ *b*:’starting date’	‘1999 ’s starting date point is 1999 .’	‘the the masters ’s starting date point is 1999 .’	‘the masters was created on 1999 .’	-
QA.3	‘is canada bigger than united states’	*false*:’negative polarity’, *r*:’bigger than’, *y*:’united states’, *x*:’canada’	‘no , canada is not the bigger than united states .’	‘no , canada is not bigger than united states .’	‘no , canada is not bigger than united states .’	‘no , canada is not bigger than united states .’
QA.3	‘will ferrell’s wife’	*human being*: ’viveca paulin’, *a*:"will ferrell ’s", *b*:’wife’	‘will ferrell ’s ’s wife is viveca paulin .’	‘will ferrell ’s ’s wife is viveca paulin .’	‘viveca paulin .’	‘will ferrell ’s wife is viveca paulin .’
QA.3	‘popsicle maker’	*business*:’unilever’, *a*:’popsicle’ *b*:’maker’	‘popsicle ’s maker is unilever .’	‘popsicle ’s manufacturer is unilever .’	‘unilever .’	popsicle ’s maker is unilever .’
SFX	-	**inform** (*name*:‘sanjalisco’, *kidsallowed*:‘yes’)	sanjalisco allows kid -s and is located	-	-	sanjalisco allows kid -s
SFX	-	**inform** (*name*:‘red door cafe’, *area*:‘cathedral hill’) *goodformeal*:‘breakfast’) *kidsallowed*:‘no’)	red door cafe is a nice restaurant in the cathedral hill does not allow kid -s and is good for breakfast	-	-	red door cafe is a nice restaurant in cathedral hill that is good for breakfast and does not allow kid -s
SFX	-	**inform** (*name*:‘darbar restaurant’, *food*:‘pakistani’) *goodformeal*:‘lunch’) *kidsallowed*:‘yes’)	darbar restaurant is a pakistani restaurant that allows kid -s and is good for lunch	-	-	darbar restaurant is a nice restaurant that serves pakistani food and allows kid -s

TABLE 7.5: Examples of different outputs from our models when given the same input Meaning Representation (and previous context when available) on two of our Question-Answering datasets (QA.1, QA.3) and on a task-based (SFX) dataset.

allows kid-s and is located’). Due to multi-tasking the models generate more grammatically correct and natural responses for both SFX and QA.3.

7.9 Conclusions

In this work, we apply the traditional dialogue MR-to-text approach to NLG to an open-domain QA setting, with sensibly larger ontologies compared to current task-oriented dialogue approaches. Our goal was to test the reliability of current approaches to NLG for dialogue in an environment where the number of slots could be substantial, a requirement that is critical to meet if we want to move towards an integrated NLG module across different domains.

The experiments presented show the feasibility of learning a NLG module for QA using a MR-to-text approach. NLG models performances on datasets with progressively bigger ontologies reported a continuous but not drastic decline for most metrics. Moreover, our multitask learning experiments showed that learning NLG models jointly for QA and task-oriented dialogue improves single tasks performances in terms of fluency. Results across different experimental settings also point towards the vital role played by the previous utterance context (delexicalized and especially lexicalized) to improve NLG models for open-domain QA.

While we envision our approach as a first step towards an integrated statistical NLG module for a dialogue system, still much remains to be done in order to achieve such a challenge. In this work, for example, we saw the importance of adapting approaches to NLG typical of task-oriented dialogue when moving to an open-domain QA setting. This is important not only in terms of modelling (the essential role of the previous utterance), but also in terms of evaluation (designing metrics able to capture the relative importance of some slots in a given answer compared to others). We believe an interesting research direction for follow-up work, besides expanding our multi-task-learning experiments, would be the investigation of evaluation metrics for NLG in a QA setting, for example to better capture the centrality of some slots (or entities) compared to others when answering a given question.

7.10 Summary

In this Chapter, we investigate how to apply the Meaning-Representation-to-text approach, typical of task-oriented dialogue, to open-domain Natural Language Generation (NLG). In particular, we address the task of open-domain QA, which requires substantially larger ontologies compared to task-oriented single-domain interactions.

First, we investigated the impact of increasing the number of slot types on the generation quality and experimented with different partitions of the QA data with progressively larger ontologies (up to 369 slot types). Second, we performed multitask learning experiments between open-domain QA and task-oriented dialog, and benchmarked our model on a popular NLG dataset. Moreover, we experimented with using the conversational context as an additional input to improve response generation quality. Additionally, we proposed evaluation metrics for open-domain MR-to-text NLG to capture the more flexible relation between slots in the input MR and the ones realised in the generated text. Our experiments using both objective metrics and human evaluation showed the feasibility of learning statistical NLG models for open-domain QA with larger ontologies and the usefulness of training across different tasks and domains to increase the quality of the generated response.

Chapter 8

Conclusions

8.1 Synopsis

In this thesis, we investigated how to combine intentional and thematic aspects for coherence modelling in open-domain dialogue and for the design of open-domain coherent Conversational Agents (CA), relying on Dialogue Act (DA) and entity-based theories.

Chapter 2 presented an overview of background work on which this thesis' contributions rely. First, we discussed current approaches in conversational AI, highlighting the different ways in which coherence is approached in task-oriented modular CAs and non-task-oriented dialogue models. Then, we delineated approaches to the notion of coherence in theoretical Linguistics, with a focus on those capturing the thematic and intentional aspects. Finally, we described approaches to main approaches to coherence in dialogue in Computational Linguistics, with a particular concentration on entity-based coherence models and DA theory.

The first main contribution of this thesis was discussed in Chapter 3, where we proposed Roving Mind, a novel architecture for an open-domain CA designed for coherence and engagement. We described the different components of our architecture based on structures composed by DAs and entities and relying on several open-domain Knowledge Bases, including one representing commonsense knowledge, for coherent and engaging content retrieval and generation. Then, we presented a series of experiments conducted during a period of 6 consecutive weeks during which Roving Mind was tested daily by Amazon Alexa users within the framework of the first Alexa Prize competition.

Afterwards, in Chapter 4, we proposed a methodology to train an open-domain DA tagger compliant with the ISO standard by mapping different publicly available resources with DA annotation. The proposed DA tagger was designed for online use in CA architectures. After assessing the performance of our DA tagging models on the Switchboard Dialogue Act corpus with in-domain experiments, our models were tested on three out of domain corpora. Finally, through an ablation experiment, we showed the importance of combining different resources for creating a robust domain-independent DA tagger.

Chapter 5 was dedicated to the exploration of *weakly supervised* approaches to coherence modeling across multiple corpora relying on DA and entity information. First, we investigated standard coherence tasks designed at the *conversation* level. We proposed coherence models relying on different representations combining DA and entity information. Our experiments showed the importance of DA information for coherence prediction and highlighted the fact that standard conversation-level coherence tasks might be less useful for assessing models' performance on dialogue when considering DA information. Then, we explored weakly supervised coherence tasks at the *turn* level. In particular, we proposed to use response selection as a weakly supervised task for training turn coherence rankers, which might be more useful compared to conversation-level tasks for real-world CAs applications. We presented a novel dataset, the Switchboard Coherence corpus, annotated with turn coherence ratings, which we developed in order to have a reliable testset to assess the ability of our models to predict turn coherence. A statistical analysis of the dataset showed the correlation between entity and DA information, especially when combined, with human perception of turn coherence. Additionally, we proposed neural models for dialogue coherence prediction, based on abstract conversational context representations relying on DAs and entities. We assessed the performance of our models on the response selection training task and then tested on turn coherence rating on the Switchboard Coherence corpus. Our experiments confirmed the crucial importance of combining DA and entity information for predicting turn coherence in open-domain conversation.

Subsequently, in Chapter 6 we investigated *supervised* approaches to open-domain dialogue evaluation, which also utilise DA and entity information. Also in this case, we proposed approaches for both the level of the entire conversation, and then for the level of speaker turns. For the *conversation* level part, we utilised conversational user ratings collected during the Alexa Prize competition as a supervision signal. We proposed open-domain models combining DA with different features, including topic representations, to automatically predict users judgement of whole human-machine interactions. Although we found moderate correlations between our proposed features and user ratings, our results pointed to the difficulty of predicting human judgement in noisy real-world human-machine interactions. For the *turn* level part, on the other hand, we utilized a large dataset of Alexa Prize conversations between users and various CAs, annotated at the turn level for coherence and engagement. We proposed turn level open-domain evaluation models based on entity grid, DA and other features. We found all of the proposed features to be useful for predicting turn level coherence and engagement. Then, we proposed two different ways of utilising the evaluators feedback for open-domain coherent and engaging response generation: by using the evaluators output to rerank the generated response, and to integrate the evaluators' loss directly into the conversational model. We found both techniques to be useful for coherent and engaging response generation.

Last, in Chapter 7, we detailed our experiments towards a possible open-domain Natural Language Generation module. In particular, we proposed to apply the Meaning-Representation-to-text approach, where the task of the models is generating a response given a DA and list of associated slots, to open-domain Question Answering (QA), which relies on much larger ontologies. Moreover, we presented novel evaluation metrics to capture the more flexible relation between MR slots and those to be realised in the output for open-domain NLG. We presented in-domain and cross-domain experiments, which suggested the feasibility of learning open-domain NLG models with larger ontologies.

8.2 Summary of contributions

In this section, we detail the contributions of this dissertation in regards to the hypotheses proposed.

H1: Can we model coherence in dialogue using Dialogue Acts and entities? The first aim of the thesis was the one of investigating whether we could model open-domain dialogue coherence using a representation of the dialogue based on DAs and entities. Across Chapter 5 and Chapter 6, we proposed different models and tasks to model open-domain dialogue coherence using DA and entity features. Overall, *the experiments proposed indicate that both the thematic aspect, declined through entity-based features, and the intentional aspect, captured via DA features, are useful for modeling open-domain dialogue coherence, especially when combined.*
More specifically, the thesis contributions described in Chapter 5 indicate that DA and entity play a crucial role for predicting coherence in dialogue with weakly supervised techniques, both when predicting the coherence of entire conversation (Section 5.1) and when predicting the coherence of single turns (Section 5.2). We proposed different Machine Learning techniques (SVM and neural models) and input representations (grid-inspired and linearised representations) to capture entity and DA joint and independent patterns of distribution. We presented a novel dataset, the Switchboard Coherence corpus, to investigate human perception of turn coherence in spoken dialogue and found that turn coherence perception correlates with the entity mentioned in the previous context and DAs used.
On the other hand, the thesis contributions presented in Chapter 6 also point to the usefulness of DA and entity-based features for learning evaluation models of open-domain dialogue with supervised techniques using noisy human-machine interactions data. In Section 6.1 we proposed models for predicting conversation-level user ratings and found DA and LDA-based representations to be useful for the task. In Section 6.2, we proposed models for predicting turn-level coherence and engagement and found that DA and entity-based representations are useful for this task. Additionally, we proposed techniques to incorporate the learned evaluators information for response generation models and showed both methodologies to improve coherence and engagement of the generated responses.

H2: Can we use Dialogue Acts and entities as units to build models for an open-domain coherent conversational agent? The second aim of this dissertation was exploring how to use DAs and entities to create different components of an open-domain CA. While in Chapter 3 we proposed a full open-domain architecture designed for coherence and engagement, across the subsequent Chapters we provided different solutions which could be used across different modules of such an architecture. For easiness of read, here we discuss our proposed solutions on a per-module basis. In general, *the models proposed throughout the thesis suggest that it is possible to use Dialogue and entities as building blocks to create different components of domain-independent coherent CAs.*

Spoken Language Understanding In Chapter 3, we proposed an SLU module relying on the notion of Functional Units structures, composed of DA and associated open-domain entities. We proposed to use different open-domain Knowledge Bases, including a commonsense one, to perform entity-linking within the SLU module, for retrieving content coherent with the conversational context. In Chapter 4 we detailed the methodology used to train a domain-independent DA tagger robust to different DA categories part of our proposed SLU. We presented a novel resource, the result of the aggregation and mapping of different publicly available resources, to train a domain-independent DA tagger compliant with the ISO 24617-2 standard. We reported experiments showing that the proposed DA tagger achieved satisfactory performance on in-domain (with SOTA results compared to other online DA taggers at the time) and cross-domain settings and highlighted the importance of combining multiple resources for robust domain-independent DA tagging.

Dialogue Manager In Chapter 3, we proposed a sequential architecture for an open-domain DM, whose design relies on general pragmatic functions inspired by DA theory (Jurafsky, 1997). The proposed DM, relies on the interaction with different open-domain KBs and a dialogue history, recording DA and entities previously used, for creating a list of Functional Units composed of DAs and entities, to be passed to the NLG component.
Moreover, in Chapter 5, section 5.2, we proposed response coherence rankers based on DA and entities trained with weakly supervised techniques, which could be used within an open-domain DM for ranking the set of generated Functional Units structures before passing them to the NLG. In Chapter 6, section 6.2, we proposed response coherence rankers trained in a supervised fashion, which could be used in a retrieval-based open-domain DM component, for ranking the retrieved responses.

Natural Language Generation In Chapter 7 we presented experiments towards an integrated open-domain statistical NLG for CAs. We proposed multiple Encoder-Decoder models which generate a response conditioned on DA and open-domain slots and the previous context. We presented experiments with NLG models with larger ontologies, characteristic of multi- and open-domain dialogue. Furthermore, we presented cross-domain multi-task learning experiments which point to the usefulness of integrating knowledge from multiple domains for improving the fluency of domain-independent NLG. Moreover, we proposed to modify traditional unit-testing evaluation metrics for NLG to adapt them to the more flexible open-domain setting.

Evaluation The contributions from Chapter 5 and Chapter 6 already detailed when discussing H1 within this Section are all mostly designed with the main goal of open-domain model-agnostic (i.e. which could be applied regardless of the dialogue model used for generating responses) dialogue evaluation. Another contribution of this thesis in regards to open-domain dialogue evaluation was highlighting the difficulty of modeling human judgement of real-world human-machine non-task-oriented interaction, as shown by our upper-bound annotation experiment presented in Section 6.1.7.

8.3 Conclusions and Future Directions

Overall, this dissertation represents one of many possible first steps towards a data-driven understanding of the structure of open-domain conversation. We focused on two aspects, the thematic and intentional one, which inspired numerous works on coherence across the fields of Theoretical and Computational Linguistics. We showed how these aspects could be useful for coherence modeling in open-domain dialogue, using entity-based and DA theories. At the same time, we proposed various methodologies for creating different components of open-domain coherent CAs relying on DA and entity-based structures. Nonetheless, the approaches presented in this thesis could be improved in the future in several regards.

From the perspective of the thematic aspect, one of the first things to look forward to will be the integration of coreference resolution in our proposed models. Unfortunately, during the course of this thesis, while we initially attempted to integrate coreference resolution SOTA models into our work, we found them not to be reliable yet for open-domain dialogue, especially in the case of transcribed conversations. Another limitation of our approach is the fact that it relies on a syntactic parser which might not be robust enough for spoken text. In our approach we do not require a full syntactic parsing of utterances, but mainly the parsing of the main syntactic roles of the sentence, that is the direct arguments of the verb on which usually parsers are more robust and chunkers can be used. Nevertheless, we believe that the performance and precision of our models will definitely improve with the availability of parsers more robust to the idiosyncrasies of spoken text.

On the other hand, in regards to the intentional aspect, in several cases our models rely on gold DAs. As future work, it would therefore be important to explore the performance of our models with predicted DAs, rather than the groundtruth ones. Additionally, instead of using only top-down expert designed DA taxonomies, it would be interesting to investigate alternatives ways of dynamically define DA tagsets (Marinelli et al., 2019). Moreover, given the provided insights in the relation among DAs and entities for predicting coherence, we believe a fruitful research direction to pursue would be jointly training DA tagging models with an entities extractor, similarly to previous work done for joint intent classification and slot tagging (Liu and Lane, 2016).

Among the considered techniques, we believe the most promising directions to be the ones using weakly supervised, rather than supervised techniques. In recent years, weakly supervised methodologies, also known as self-supervised, have been successfully applied to several NLP tasks (Devlin et al., 2019). Given the current lack of data in dialogue, we believe further exploring weakly supervised training tasks, could be a fertile direction for learning models of discourse coherence. Nonetheless, while exploring weakly supervised approaches for training is a promising direction, we also found to be crucial to create reliable resources for testing utilising human annotation, as mentioned in Section 5.2 as the motivation to build the Switchboard Coherence corpus.

Finally, in this work we explored two aspects of coherence. However, as discussed in Chapter 2, coherence is a multifaceted property. As such, coherence is defined by several layers besides the aspects addressed in this work. Hence, we envision that in the future the approaches presented in this thesis could greatly benefit from adding additional layers of complexity, for example using Discourse Relations or exploring augmenting our models with predicate argument structures to better encode the different relations among entities.

Consequently, we believe the work presented here to be a useful, but not comprehensive solution to the challenges of training and evaluating coherent open–domain CAs, as well as modeling dialogue coherence. Rather, we expect the insights presented in this thesis to be used in the future in combination with other metrics addressing different aspects of dialogue coherence.

Nevertheless, we believe the proposed models could be more reliable and useful than some of the currently available automatic metrics for dialogue evaluation and non-modular models, since instead of evaluating and modeling surface realizations they consider the deeper structure of dialogue based on the topics (entities) discussed and the underlying intents of the participants. Our intuition is that these structures could be especially helpful in cases where not much data is available, which as we saw to be often the case when creating CAs from scratch.

Appendix A

Data collection procedure for Switchboard Coherence corpus

A.1 Introduction

Coherence rating is an inherently subjective task and could be challenging especially for a dataset of transcribed real-world open-domain human-human conversation like Switchboard, where we have possible interruptions, overlaps and disfluencies naturally occurring. Hence, in order to ensure we collected reliable judgements for turn coherence, we followed a multi-step procedure to build the Switchboard Coherence (SWBD-Coh) corpus using Amazon Mechanical Turk (AMT).

A.1.1 Experiment with internal annotators

First we performed a small-scale annotation experiment to evaluate the feasibility of the task. Two internal annotators, both with Linguistics education, were asked to rate a set of 150 different dialogues randomly selected from the testset from (Cervone, Stepanov, and Riccardi, 2018). The 150 annotation pairs (context + set of candidate turns) were generated using the same procedure described in Section 4 of the paper. The coherence scale was divided into 1 (not coherent), 2 (not sure it fits) and 3 (coherent). Since we wanted to capture a general perception of coherence, rather than bias annotators towards our own intuitions, in the guidelines annotators the task was described as: "Your task is to rate each candidate on a scale of how much it is *coherent* with the previous dialogue context, that is *how much that response makes sense as the next natural turn in the dialogue*".

Since in this case we only have two annotators, we were able to measure their inter-annotator agreement using a weighted kappa score with quadratic weights (since our categories are ordinal). The inter-annotator agreement was of 0.657 (which can be regarded as substantial (Viera, Garrett, et al., 2005)). Then, we averaged scores for each candidate turn from both annotators. As shown in Table A.1, original turns had higher coherence scores ($\mu = 2.66$) compared to adversarial turns, while turns generated with Internal Swap were considered more coherent ($\mu = 1.78$) than the ones generated via External Swap ($\mu = 1.45$).

A.1.2 Experiment with AMT

After having assessed the feasibility of the task, we then proceeded to set up the data collection procedure on AMT.

	Original	**Internal Swap**	**External Swap**
Mean score 150	2.7 (0.5)	1.8 (0.7)	1.4 (0.7)
Mean score SWBD-Coh	2.6 (0.5)	1.8 (0.7)	1.4 (0.6)

TABLE A.1: Comparison of human annotation results for the experiment with two internal annotators (150 dialogues) and the Switchboard Coherence (SWBD-Coh) dataset. Mean scores (and standard deviation) are reported for each candidates group: originals (Orig), Internal Swap (IS) and External Swap (ES).

In order to select workers for our coherence annotation task we first set up a qualification task on AMT. The qualification task consisted of 5 dialogues (taken from the 150 internally annotated) with 7 turn candidates using the same coherence rating scale as in the gold annotation. In order to pass the qualification task a worker had to have a weighted kappa score higher than 0.4 with both our gold annotators. This threshold was decided empirically by first running a small scale experiment with other 4 internal annotators on the qualification task. 37 workers passed the qualification task. The average weighted kappa agreement with the two gold annotators was 0.659 (min: 0.425, max: 0.809, STD: 0.101). In order to calculate the agreement among all the 37 workers on this batch we employ leave-one-out resampling. For each worker who annotated the data we calculate the correlation of her/his scores with the mean ones of all other annotators in the batch. This is repeated for all workers and then averaged. This technique has been used in other coherence annotation experiments (Barzilay and Lapata, 2008; Lapata and Barzilay, 2005).

Workers who passed the qualification test could then proceed to annotate the SWBD-Coh data. The data, consisting of 1000 dialogues, was divided into 100 batches of 10 dialogues each. Each batch was annotated by at least 5 workers. In order to remove possible workers who did not perform well on a given batch, we employed a combination of techniques including leave-one-out resampling and average scores given to original turns. The average leave-one-out correlation per batch for turn coherence rating achieved with this data collection procedure was: ρ =0.723 (min: 0.580, max: 0.835, STD: 0.055). Interestingly, as shown in Table A.1, the average scores per candidate group (original, Internal swap, External swap) match closely the ones obtained in our gold 150 annotation data.

Bibliography

Adiwardana, Daniel, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. (2020). "Towards a Human-like Open-Domain Chatbot". In: *arXiv*, arXiv–2001.

Afantenos, Stergos, Eric Kow, Nicholas Asher, and Jérémy Perret (2015). "Discourse parsing for multi-party chat dialogues". In: Association for Computational Linguistics (ACL).

Albesano, Dario, Paolo Baggia, Morena Danieli, Roberto Gemello, Elisabetta Gerbino, and Claudio Rullent (1997). "Dialogos: A robust system for human-machine spoken dialogue on the telephone". In: *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*. Vol. 2. IEEE, pp. 1147–1150.

Alexandersson, Jan, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel (1998). *Dialogue acts in Verbmobil 2*. DFKI Saarbrücken.

Alistair, Kennedy and Inkpen Diana (2005). "Sentiment classification of movie and product reviews using contextual valence shifters". In: *Proceedings of FINEXIN*.

Allen, James F, Bradford W Miller, Eric K Ringger, and Teresa Sikorski (1996). "A robust system for natural spoken dialogue". In: *Proceedings of the 34th annual meeting on ACL*. ACL, pp. 62–70.

Allen, James F and C Raymond Perrault (1980). "Analyzing intention in utterances". In: *Artificial intelligence* 15.3, pp. 143–178.

Allwood, Jens (1995). "An Activity-Based Approach to Pragmatics". In: *Gothenburg paper in Theoretical Linguistics* 76.

Allwood, Jens, Joakim Nivre, and Elisabeth Ahlsén (1992). "On the semantics and pragmatics of linguistic feedback". In: *Journal of semantics* 9.1, pp. 1–26.

Anderson, Anne H, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. (1991). "The HCRC map task corpus". In: *Language and speech* 34.4, pp. 351–366.

Andorno, Cecilia Maria (2003). *Linguistica testuale. Un'introduzione*. Carocci.

Asher, Nicholas and Alex Lascarides (2003). *Logics of conversation*. Cambridge University Press.

Austin, John Langshaw (1975). *How to do things with words*. Oxford university press.

Austin, John Langshaw and JO Urmson (1962). *How to Do Things with Words. The William James Lectures Delivered at Harvard University in 1955.[Edited by James O. Urmson.]*. Clarendon Press.

Bahdanau, Dzmitry, KyungHyun Cho, and Yoshua Bengio (2014). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *arXiv preprint arXiv:1409.0473*.

Bahl, Lalit, Peter Brown, Peter De Souza, and Robert Mercer (1986). "Maximum mutual information estimation of hidden Markov model parameters for speech recognition". In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86*. Vol. 11. IEEE, pp. 49–52.

Balakrishnan, Anusha, Jinfeng Rao, Kartikeya Upasani, Michael White, and Rajen Subba (2019). "Constrained Decoding for Neural NLG from Compositional Representations in Task-Oriented Dialogue". In: *arXiv preprint arXiv:1906.07220*.

Banerjee, Satanjeev and Alon Lavie (2005). "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. Vol. 29, pp. 65–72.

Barzilay, Regina and Mirella Lapata (2008). "Modeling local coherence: An entity-based approach". In: *Computational Linguistics* 34.1, pp. 1–34.

Baumgartner (2015). *Reddit*. `https://archive.org/details/2015_reddit_comments_corpus`. [Accessed: 2018-07-01].

Bayer, Ali Orkan and Giuseppe Riccardi (2012). "Joint language models for automatic speech recognition and understanding". In: *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 199–203.

Bengio, Yoshua, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain (2006). "Neural probabilistic language models". In: *Innovations in Machine Learning*. Springer, pp. 137–186.

Bobrow, Daniel G, Ronald M Kaplan, Martin Kay, Donald A Norman, Henry Thompson, and Terry Winograd (1977). "GUS, a frame-driven dialog system". In: *Artificial intelligence* 8.2, pp. 155–173.

Bocklisch, Tom, Joey Faulkner, Nick Pawlowski, and Alan Nichol (2017). "Rasa: Open source language understanding and dialogue management". In: *arXiv preprint arXiv:1712.05181*.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2016). "Enriching Word Vectors with Subword Information". In: *arXiv:1607.04606*.

Bothe, Chandrakant, Sven Magg, Cornelius Weber, and Stefan Wermter (2018). "Conversational Analysis Using Utterance-level Attention-based Bidirectional Recurrent Neural Networks". In: *Proc. Interspeech 2018*, pp. 996–1000.

Bowden, Kevin K, Jiaqi Wu, Shereen Oraby, Amita Misra, and Marilyn Walker (2017). "Slugbot: An Application of a Novel and Scalable Open Domain Socialbot Framework". In: *Alexa Prize Proceedings*.

Braud, Chloé, Maximin Coavoux, and Anders Søgaard (2017). "Cross-lingual RST Discourse Parsing". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 292–304.

Bublitz, Wolfram (2011). "Cohesion and coherence". In: *Discursive Pragmatics, John Benjamins Publishing Company, Amsterdam/Philadelphia*, pp. 37–49.

Bunt, Harry (1999). "Dynamic interpretation and dialogue theory". In: *The structure of multimodal dialogue* 2, pp. 1–8.

– (2009). "The DIT++ taxonomy for functional dialogue markup". In: *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pp. 13–24.

Bunt, Harry, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum (2010). "Towards an ISO standard for dialogue act annotation". In: *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Bunt, Harry, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum (2012). "ISO 24617-2: A semantically-based standard for dialogue annotation." In: *LREC*, pp. 430–437.

Bunt, Harry, Alex C Fang, Xiaoyue Liu, Jing Cao, and Volha Petukhova (2013). "Issues in the addition of ISO standard annotations to the Switchboard corpus". In: *Workshop on Interoperable Semantic Annotation*.

Bunt, Harry, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot (2020). "The ISO Standard for Dialogue Act Annotation". In: *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 549–558.

Bunt, Harry, Volha Petukhova, Andrei Malchanau, Kars Wijnhoven, and Alex Chengyu Fang (2016). "The DialogBank." In: *LREC*.

Burger, Susanne, Karl Weilhammer, Florian Schiel, and Hans G Tillmann (2000). "Verbmobil data collection and annotation". In: *Verbmobil: Foundations of speech-to-speech translation*. Springer, pp. 537–549.

Carletta, Jean (2006). "Announcing the AMI Meeting Corpus". In: *The ELRA Newsletter 11(1), January-March, p. 3-5.*

Cervone, Alessandra, Enrico Gambi, Giuliano Tortoreto, Evgeny A Stepanov, and Giuseppe Riccardi (2018). "Automatically Predicting User Ratings for Conversational Systems." In: *Fifth Italian Conference on Computational Linguistics (CLiC-it).*

Cervone, Alessandra, Chandra Khatri, Rahul Goel, Behnam Hedayatnia, Anu Venkatesh, Dilek Hakkani-Tur, and Raefer Gabriel (2019). "Natural Language Generation at Scale: A Case Study for Open Domain Question Answering". In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 453–462. DOI: `10.18653/v1/W19-8657`. URL: `https://www.aclweb.org/anthology/W19-8657`.

Cervone, Alessandra and Giuseppe Riccardi (2020). "Is this dialogue coherent? Learning from Dialogue Acts and entities". In: *Proceedings of the 21st Annual SIGdial Meeting on Discourse and Dialogue*.

Cervone, Alessandra, Evgeny Stepanov, and Giuseppe Riccardi (2018). "Coherence Models for Dialogue". In: *Proceedings of the 19th Annual Conference of the International Speech Communication Association*.

Cervone, Alessandra, Giuliano Tortoreto, Stefano Mezza, Enrico Gambi, and Giuseppe Riccardi (2017). "Roving Mind: a balancing act between open–domain and engaging dialogue systems". In: *Alexa Prize Proceedings*.

Chelba, Ciprian, Milind Mahajan, and Alex Acero (2003). "Speech utterance classification". In: *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. Vol. 1. IEEE, pp. I–280.

Cheyer, Adam and Didier Guzzoni (2014). *Method and apparatus for building an intelligent automated assistant*. US Patent 8,677,377.

Chisholm, Andrew and Ben Hachey (2015). "Entity disambiguation with web links". In: *Transactions of ACL* 3, pp. 145–156.

Chowdhury, Shammur Absar, Evgeny A Stepanov, and Giuseppe Riccardi (2016). "Transfer of corpus-specific dialogue act annotation to iso standard: Is it worth it?" In: *LREC*.

Clark, Elizabeth, Yangfeng Ji, and Noah A Smith (2018). "Neural Text Generation in Stories Using Entity Representations as Context". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Vol. 1, pp. 2250–2260.

Clark, Herbert H and Edward F Schaefer (1989). "Contributing to discourse". In: *Cognitive science* 13.2, pp. 259–294.

Cohen, Philip R and C Raymond Perrault (1979). "Elements of a plan-based theory of speech acts". In: *Cognitive science* 3.3, pp. 177–212.

Colby, Kenneth Mark, Franklin Dennis Hilf, Sylvia Weber, and Helena C Kraemer (1972). "Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes". In: *Artificial Intelligence* 3, pp. 199–221.

Conte, Maria-Elisabeth (1980). "Coerenza testuale". In: *Lingua e Stile Bologna* 15.1, pp. 135–154.

– (1989). "Coesione testuale: recenti ricerche italiane". In: *La linguistica testuale*, pp. 272–295.

Core, Mark G. and James F. Allen (1997). "Coding dialogs with the DAMSL annotation scheme". In: *Proceedings of AAAI Fall Symposium on Communicative Action in Humans and Machines*.

Danescu-Niculescu-Mizil, Cristian and Lillian Lee (2011). "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs." In: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Danieli, Morena and Elisabetta Gerbino (1995). "Metrics for evaluating dialogue strategies in a spoken language system". In: *Proceedings of the 1995 AAAI spring symposium on Empirical Methods in Discourse Interpretation and Generation*. Vol. 16, pp. 34–39.

De Beaugrande, Robert and Wolfgang U Dressler (1981a). *Einführung in die Textlinguistik*. Vol. 28. Niemeyer Tübingen.

De Beaugrande, Robert-Alain and Wolfgang Ulrich Dressler (1981b). *Introduction to text linguistics*. Vol. 1. Longman London.

De Mori, Renato, Frédéric Bechet, Dilek Hakkani-Tur, Michael McTear, Giuseppe Riccardi, and Gokhan Tur (2008). "Spoken language understanding". In: *IEEE Signal Processing Magazine* 25.3, pp. 50–58.

De Saussure, Ferdinand (1989). *Cours de linguistique générale*. Vol. 1. Otto Harrassowitz Verlag.

Del Corro, Luciano and Rainer Gemulla (2013). "Clausie: clause-based open information extraction". In: *Proceedings of the 22nd international conference on World Wide Web*. ACM, pp. 355–366.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186.

Dhingra, Bhuwan, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmad, and Li Deng (2017). "Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 484–495.

Dijk, T.A. van (1981). *Studies in the pragmatics of discourse*. Janua linguarum. Mouton. ISBN: 9789027932495.

Dijk, Teun A van (1987). "Episodic models in discourse processing." In:

Dodge, Jesse, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston (2015). "Evaluating prerequisite qualities for learning end-to-end dialog systems". In: *arXiv preprint arXiv:1511.06931*.

Dušek, Ondřej and Filip Jurcicek (2016). "A Context-aware Natural Language Generator for Dialogue Systems". In: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 185–190.

Dušek, Ondrej and Filip Jurcıcek (2016). "Sequence-to-Sequence Generation for Spoken Dialogue via Deep Syntax Trees and Strings". In: *The 54th Annual Meeting of the Association for Computational Linguistics*, p. 45.

Dušek, Ondřej, Jekaterina Novikova, and Verena Rieser (2018). "Findings of the E2E NLG Challenge". In: *Proceedings of the 11th International Conference on Natural Language Generation*, pp. 322–328.

Elsner, Micha and Eugene Charniak (2011a). "Disentangling Chat with Local Coherence Models". In: *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, pp. 1179–1189.

– (2011b). "Extending the entity grid with entity-specific features". In: *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies: short papers-Volume 2*. ACL, pp. 125–129.

Espinosa, Dominic, Rajakrishnan Rajkumar, Michael White, and Shoshana Berleant (2010). "Further meta-evaluation of broad-coverage surface realization". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 564–574.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin (June 2008). "LIBLINEAR: A Library for Large Linear Classification". In: *J. Mach. Learn. Res.* 9, pp. 1871–1874. ISSN: 1532-4435.

Fang, A, Jing Cao, Harry Bunt, and Xiaoyue Liu (2012a). "The annotation of the Switchboard corpus with the new ISO standard for dialogue act analysis". In: *Workshop on Interoperable Semantic Annotation*, p. 13.

Fang, Alex C., Jing Cao, Harry Bunt, and Xiaoyue Liu (2012b). "The Annotation of the Switchboard Corpus with the New ISO Standard for Dialogue Act Analysis". In: *Workshop on Interoperable Semantic Annotation*.

Fang, Hao, Hao Cheng, Elizabeth Clark, Ariel Holtzman, Maarten Sap, Mary Ostendorf, Yejin Choi, and Noah A. Smith (2017). "Sounding Board – University of Washington's Alexa Prize Submission". In: *Alexa Prize Proceedings*.

Farag, Youmna, Helen Yannakoudakis, and Ted Briscoe (2018). "Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Vol. 1, pp. 263–271.

Filippova, Katja and Michael Strube (2007). "Extending the entity-grid coherence model to semantically related entities". In: *Proceedings of the Eleventh European Workshop on Natural Language Generation*. ACL, pp. 139–142.

Fillmore, Charles (1998). "Pragmatics and the description of discourse". In: *Pragmatics: Communication, interaction, and discourse* 5, p. 385.

Frey, Brendan J and Delbert Dueck (2007). "Clustering by passing messages between data points". In: *science* 315.5814, pp. 972–976.

Fu, Yao and Yansong Feng (2018). "Natural Answer Generation with Heterogeneous Memory". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Vol. 1, pp. 185–195.

Gandhe, Sudeep and David Traum (2008). "An evaluation understudy for dialogue coherence models". In: *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, pp. 172–181.

– (2016). "A Semi-automated Evaluation Metric for Dialogue Model Coherence". In: *Situated Dialog in Speech-Based Human-Computer Interaction*, p. 217.

Gatt, Albert and Emiel Krahmer (2018). "Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation". In: *Journal of Artificial Intelligence Research* 61, pp. 65–170.

Geertzen, Jeroen, Volha Petukhova, and Harry Bunt (2007). "A multidimensional approach to utterance segmentation and dialogue act classification". In: *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pp. 140–149.

Georgila, Kallirroi and David Traum (2011). "Reinforcement learning of argumentation dialogue policies in negotiation". In: *Twelfth Annual Conference of the International Speech Communication Association*.

Gernsbacher, Morton Ann and Talmy Givón (1995). *Coherence in spontaneous text*. Vol. 31. John Benjamins Publishing.

Givón, Talmy (1983). *Topic continuity in discourse*. John Benjamins Publishing Company.

– (1987). "Beyond foreground and background". In: *Coherence and grounding in discourse* 11, pp. 175–188.

Godfrey, John J, Edward C Holliman, and Jane McDaniel (1992). "SWITCHBOARD: Telephone speech corpus for research and development". In: *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. Vol. 1. IEEE, pp. 517–520.

Goode, Lauren (2018). "Your Voice Assistant May Be Getting Smarter, But It's Still Awkward". In: *Wired*. `https://www.wired.com/story/voice-assistants-ambient-computing/` [Online; accessed January-2020].

Gorin, Allen L, Giuseppe Riccardi, and Jeremy H Wright (1997). "How may I help you?" In: *Speech communication* 23.1, pp. 113–127.

Graham, Yvette (2015). "Accurate Evaluation of Segment-level Machine Translation Metrics." In:

Grice, Herbert P (1970). *Logic and conversation*.

Grosz, Barbara J and Candace L Sidner (1986). "Attention, intentions, and the structure of discourse". In: *Computational linguistics* 12.3, pp. 175–204.

Grosz, Barbara J, Scott Weinstein, and Aravind K Joshi (1995). "Centering: A framework for modeling the local coherence of discourse". In: *Computational linguistics* 21.2, pp. 203–225.

Group, IMARC (2019). *Intelligent Virtual Assistant Market: Global Industry Trends, Share, Size, Growth, Opportunity and Forecast 2019-2024*. Tech. rep. IMARC Group. URL: `https://www.imarcgroup.com/intelligent-virtual-assistant-market`.

Gruber, Helmut and Gisela Redeker (2014). *The pragmatics of discourse coherence: Theories and applications*. Vol. 254. John Benjamins Publishing Company.

Guinaudeau, Camille and Michael Strube (2013). "Graph-based Local Coherence Modeling." In: *ACL (1)*, pp. 93–103.

Guo, Daniel, Gokhan Tur, Wen-tau Yih, and Geoffrey Zweig (2014). "Joint semantic utterance classification and slot filling with recursive neural networks". In: *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 554–559.

Guo, Fenfei, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram (2017a). "Topic-based Evaluation for Conversational Bots". In: *NIPS 2017 Conversational AI workshop*.

– (2017b). "Topic-based Evaluation for Conversational Bots". In: *arXiv:1801.03622*.

Gupta, Narendra, Gokhan Tur, Dilek Hakkani-Tur, Srinivas Bangalore, Giuseppe Riccardi, and Mazin Gilbert (2006). "The AT&T spoken language understanding system". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.1, pp. 213–222.

Haffner, Patrick, Gokhan Tur, and Jerry H Wright (2003). "Optimizing SVMs for complex call classification". In: *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. Vol. 1. IEEE, pp. I–632.

Halliday, Michael AK (1967). “Notes on transitivity and theme in English: Part 2”. In: *Journal of linguistics* 3.2, pp. 199–244.

Halliday, Michael Alexander Kirkwood and Ruqaiya Hasan (1976). “Cohesion in English”. In:

Halliday, Michael Alexander Kirkwood and Christian MIM Matthiessen (2013). *Halliday's introduction to functional grammar*. Routledge.

He, He, Anusha Balakrishnan, Mihail Eric, and Percy Liang (2017). “Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 1766–1776.

Hemphill, Charles T, John J Godfrey, and George R Doddington (1990). “The ATIS spoken language systems pilot corpus”. In: *Proceedings of the DARPA speech and natural language workshop*, pp. 96–101.

Henderson, Matthew, Blaise Thomson, and Jason D Williams (2014). “The Second Dialog State Tracking Challenge.” In: *SIGDIAL Conference*, pp. 263–272.

Henderson, Matthew, Blaise Thomson, and Steve Young (2014a). “Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation”. In: *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, pp. 360–365.

– (2014b). “Word-based dialog state tracking with recurrent neural networks”. In: *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 292–299.

Henderson, Matthew, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su (July 2019). “Training Neural Response Selection for Task-Oriented Dialogue Systems”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5392–5404. DOI: 10.18653/v1/P19-1536.

Higashinaka, Ryuichiro, Toyomi Meguro, Kenji Imamura, Hiroaki Sugiyama, Toshiro Makino, and Yoshihiro Matsuo (2014). “Evaluating coherence in open domain conversational systems”. In: *Fifteenth Annual Conference of the International Speech Communication Association*.

Hinton, Geoffrey, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. (2012). “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups”. In: *Signal Processing Magazine, IEEE* 29.6, pp. 82–97.

Hobbs, Jerry R (1979). “Coherence and coreference”. In: *Cognitive science* 3.1, pp. 67–90.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long short-term memory”. In: *Neural computation* 9.8, pp. 1735–1780.

Holtzman, Ari, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi (2018). “Learning to write with cooperative discriminators”. In: *arXiv preprint arXiv:1805.06087*.

Honnibal, Matthew and Mark Johnson (2015). “An Improved Non-monotonic Transition System for Dependency Parsing”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1373–1378.

Hopper, Paul J (1979). “Aspect and foregrounding in discourse”. In: *Discourse and syntax*. Brill, pp. 211–241.

Horn, Laurence and István Kecskés (2013). "Pragmatics, discourse and cognition". In: *Yale University*, pp. 355–375.

Horn, Laurence R and Gregory L Ward (2004). *The handbook of pragmatics*. Wiley Online Library.

Huang, Yan (2017). *The Oxford handbook of pragmatics*. Oxford University Press.

Järvelin, Kalervo and Jaana Kekäläinen (2002). "Cumulated gain-based evaluation of IR techniques". In: *ACM Transactions on Information Systems (TOIS)* 20.4, pp. 422–446.

Ji, Yangfeng, Gholamreza Haffari, and Jacob Eisenstein (2016). "A Latent Variable Recurrent Neural Network for Discourse Relation Language Models". In: *Proceedings of NAACL-HLT*, pp. 332–342.

Jia, Yanyan, Yuan Ye, Yansong Feng, Yuxuan Lai, Rui Yan, and Dongyan Zhao (2018). "Modeling discourse cohesion for discourse parsing via memory network". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 438–443.

Joachims, Thorsten (2002). "Optimizing search engines using clickthrough data". In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 133–142.

Johnson, George (2011). *Machinery of the Mind*. Crown.

Joshi, Aravind K and Steve Kuhn (1979). "Centered Logic: The Role of Entity Centered Sentence Representation in Natural Language Inferencing." In: *IJCAI*, pp. 435–439.

Joshi, Aravind K and Scott Weinstein (1981). "Control of Inference: Role of Some Aspects of Discourse Structure-Centering." In: *IJCAI*, pp. 385–387.

Joty, Shafiq, Giuseppe Carenini, and Chin-Yew Lin (2011). "Unsupervised modeling of dialog acts in asynchronous conversations". In: *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*. Vol. 22. 3, p. 1807.

Joty, Shafiq, Muhammad Tasnim Mohiuddin, and Dat Tien Nguyen (2018). "Coherence modeling of asynchronous conversations: A neural entity grid approach". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 558–568.

Jurafsky, Dan (1997). "Switchboard SWBD-DAMSL Shallow-Discourse-Function". In: *Annotation, Technical Report, 97-02, University of Colorado, CO, USA.*

Jurafsky, Dan and James H. Martin (2009). *Speech & Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition.* Pearson Education India.

Juraska, Juraj, Panagiotis Karagiannis, Kevin Bowden, and Marilyn Walker (2018). "A Deep Ensemble Model with Slot Alignment for Sequence-to-Sequence Natural Language Generation". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Vol. 1, pp. 152–162.

Kamp, Hans and Christian Rohrer (1983). "Tense in texts". In: *Meaning, use and interpretation of language* 250269.

Kennedy, Alistair and Diana Inkpen (2006). "Sentiment classification of movie reviews using contextual valence shifters". In: *Computational intelligence* 22.2, pp. 110–125.

KennethResearch (2020). "Conversational Artificial Intelligence Market 2020 Size, Share, Revenue Growth, Forecast To 2024". In: *Market watch*. `https://www.marketwatch.com/press-release/conversational-artificial-intelligence-market-2020-size-share-revenue-growth-forecast-to-2024-2020-05-26` [Online; accessed June-2020].

Khatri, Chandra, Rahul Goel, Behnam Hedayatnia, Angeliki Metanillou, Anushree Venkatesh, Raefer Gabriel, and Arindam Mandal (2018). "Contextual Topic Modeling For Dialog Systems". In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 892–899.

Kim, Seokhwan, Michel Galley, Chulaka Gunasekara, Sungjin Lee, Adam Atkinson, Baolin Peng, Hannes Schulz, Jianfeng Gao, Jinchao Li, Mahmoud Adada, et al. (2019). "The Eighth Dialog System Technology Challenge". In: *arXiv preprint arXiv:1911.06394*.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv:1412.6980*.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. URL: `http://arxiv.org/abs/1412.6980`.

Kumar, Anjishnu, Arpit Gupta, Julian Chan, Sam Tucker, Bjorn Hoffmeister, Markus Dreyer, Stanislav Peshterliev, Ankur Gandhe, Denis Filiminov, Ariya Rastrow, et al. (2017). "Just ASK: building an architecture for extensible self-service spoken language understanding". In: *arXiv preprint arXiv:1711.00549*.

Kumar, Harshit, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi (2018). "Dialogue act sequence labeling using hierarchical encoder with crf". In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

Kumar, Harshit, Arvind Agarwal, and Sachindra Joshi (2019). "A Practical Dialogue-Act-Driven Conversation Model for Multi-Turn Response Selection". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1980–1989.

Lambrecht, Knud (1994). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Vol. 71. Cambridge university press.

Lapata, Mirella (2006). "Automatic evaluation of information ordering: Kendall's tau". In: *Computational Linguistics* 32.4, pp. 471–484.

Lapata, Mirella and Regina Barzilay (2005). "Automatic evaluation of text coherence: Models and representations". In: *IJCAI*. Vol. 5, pp. 1085–1090.

Lascarides, Alex and Nicholas Asher (1993). "Temporal interpretation, discourse relations and commonsense entailment". In: *Linguistics and philosophy* 16.5, pp. 437–493.

– (2008). "Segmented discourse representation theory: Dynamic semantics with discourse structure". In: *Computing meaning*. Springer, pp. 87–124.

Lebret, Rémi, David Grangier, and Michael Auli (2016). "Neural Text Generation from Structured Data with Application to the Biography Domain". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1203–1213.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444.

Lee, Ji Young and Franck Dernoncourt (2016). "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 515–520.

Leech, Geoffrey and Martin Weisser (2003). "Generic speech act annotation for task-oriented dialogues". In: *Procs. of the 2003 Corpus Linguistics Conference, pp. 441Y446. Centre for Computer Corpus Research on Language Technical Papers, Lancaster University*.

Lemon, Oliver (2012). "Conversational interfaces". In: *Data-Driven Methods for Adaptive Spoken Dialogue Systems*. Springer, pp. 1–4.

Levinson, Stephen C (1983). "Pragmatics". In:

Li, Jiwei, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan (2015). "A diversity-promoting objective function for neural conversation models". In: *arXiv:1510.03055*.

– (2016a). "A Diversity-Promoting Objective Function for Neural Conversation Models". In: *Proceedings of NAACL-HLT*, pp. 110–119.

Li, Jiwei, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan (2016b). "A Persona-Based Neural Conversation Model". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 994–1003.

Li, Jiwei and Eduard H Hovy (2014). "A Model of Coherence Based on Distributed Sentence Representation." In: *EMNLP*, pp. 2039–2048.

Li, Jiwei and Dan Jurafsky (2017). "Neural Net Models of Open-domain Discourse Coherence". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 198–209.

Li, Jiwei, Will Monroe, and Dan Jurafsky (2017). "Data Distillation for Controlling Specificity in Dialogue Generation". In: *arXiv:1702.06703*.

Li, Jiwei, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao (2016c). "Deep Reinforcement Learning for Dialogue Generation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202.

Li, Jiwei, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky (2017). "Adversarial Learning for Neural Dialogue Generation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2157–2169.

Liang, Percy, Michael I Jordan, and Dan Klein (2009). "Learning semantic correspondences with less supervision". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, pp. 91–99.

Lin, Chin-Yew and Eduard Hovy (2003). "Automatic evaluation of summaries using n-gram co-occurrence statistics". In: *NAACL-HLT*. ACL, pp. 71–78.

Lin, Xiang, Shafiq Joty, Prathyusha Jwalapuram, and M Saiful Bari (2019). "A Unified Linear-Time Framework for Sentence-Level Discourse Parsing". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4200.

Liu, Bing and Ian Lane (2016). "Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling". In: *Interspeech 2016*, pp. 685–689.

Liu, Chia-Wei, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau (2016). "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2122–2132.

Logeswaran, Lajanugen and Honglak Lee (2018). "An efficient framework for learning sentence representations". In: *arXiv:1803.02893*.

Lowe, Ryan, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau (2017a). "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1116–1126.

Lowe, Ryan, Nissan Pow, Iulian Serban, and Joelle Pineau (2015). "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems". In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 285–294.

Lowe, Ryan, Iulian V Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau (2016). "On the Evaluation of Dialogue Systems with Next Utterance Classification". In: *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 264.

Lowe, Ryan Thomas, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau (2017b). "Training end-to-end dialogue systems with the ubuntu dialogue corpus". In: *Dialogue & Discourse* 8.1, pp. 31–65.

Luong, Thang, Hieu Pham, and Christopher D Manning (2015). "Effective Approaches to Attention-based Neural Machine Translation". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421.

Mann, William C and Sandra A Thompson (1987). *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute.

– (1988). "Rhetorical structure theory: Toward a functional theory of text organization". In: *Text* 8.3, pp. 243–281.

Marcu, Daniel (2000). *The theory and practice of discourse parsing and summarization*. MIT press.

Marelli, Marco, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. (2014). "A SICK cure for the evaluation of compositional distributional semantic models." In: *LREC*, pp. 216–223.

Marinelli, Federico, Alessandra Cervone, Giuliano Tortoreto, Evgeny A Stepanov, Giuseppe Di Fabbrizio, and Giuseppe Riccardi (2019). "Active Annotation: Bootstrapping Annotation Lexicon and Guidelines for Supervised NLU Learning". In: *Proc. Interspeech 2019*, pp. 574–578.

Matthews, Brian W (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405.2, pp. 442–451.

McNemar, Quinn (1947). "Note on the sampling error of the difference between correlated proportions or percentages". In: *Psychometrika* 12.2, pp. 153–157.

Mehri, Shikib, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi (2019). "Pretraining Methods for Dialog Context Representation Learning". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3836–3845.

Mei, Hongyuan, TTI UChicago, Mohit Bansal, and Matthew R Walter (2016). "What to talk about and how? Selective Generation using LSTMs with Coarse-to-Fine Alignment". In: *Proceedings of NAACL-HLT*, pp. 720–730.

Mesgar, Mohsen, Sebastian Bücker, and Iryna Gurevych (2020). "Dialogue coherence assessment without explicit dialogue act labels". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1439–1450.

Mesgar, Mohsen and Michael Strube (2016). "Lexical coherence graph modeling using word embeddings". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1414–1423.

– (2018). "A neural local coherence model for text quality assessment". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4328–4339.

Mesnil, Grégoire, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. (2015). "Using recurrent neural networks for slot filling in spoken language understanding". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.3, pp. 530–539.

Mezza, Stefano, Alessandra Cervone, Evgeny Stepanov, Giuliano Tortoreto, and Giuseppe Riccardi (2018). "ISO-Standard Domain-Independent Dialogue Act Tagging for Conversational Agents". In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3539–3551.

Mikolov, Tomas, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur (2010). "Recurrent neural network based language model." In: *INTERSPEECH*. Vol. 2, p. 3.

Mikolov, Tomas and Geoffrey Zweig (2012). "Context dependent recurrent neural network language model." In: *SLT*, pp. 234–239.

Miller, Alexander H, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston (2017). "Parlai: A dialog research software platform". In: *arXiv:1705.06476*.

Moens, M and M Caenepeel (1994). "Temporal structure and discourse structure". In: *Tense and aspect in discourse*, pp. 5–20.

Moon, Han Cheol, Muhammad Tasnim Mohiuddin, Shafiq Joty, and Chi Xu (2019). "A Unified Neural Coherence Model". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2262–2272.

Moore, Johanna D. and Martha E. Pollack (1992). "A Problem for RST: The Need for Multi-Level Discourse Analysis". In: *Computational Linguistics* 18.4, pp. 537–544. URL: `https://www.aclweb.org/anthology/J92-4007`.

Nayak, Neha, Dilek Hakkani-Tur, Marilyn Walker, and Larry Heck (2017). "To plan or not to plan? discourse planning in slot-value informed sequence to sequence models for language generation". In: *Proc. of Interspeech*.

Nguyen, Dat Tien and Shafiq Joty (2017). "A Neural Local Coherence Model". In: *Proceedings of the 55th Annual Meeting of the ACL (Volume 1: Long Papers)*. Vol. 1, pp. 1320–1330.

Novikova, Jekaterina, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser (2017). "Why we need new evaluation metrics for NLG". In: *arXiv preprint arXiv:1707.06875*.

Novikova, Jekaterina, Ondřej Dušek, and Verena Rieser (2017). "The E2E Dataset: New Challenges For End-to-End Generation". In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 201–206.

Oh, Alice H and Alexander I Rudnicky (2000). "Stochastic language generation for spoken dialogue systems". In: *Proceedings of the 2000 ANLP/NAACL Workshop on Conversational systems-Volume 3*. ACL, pp. 27–32.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016). "Wavenet: A generative model for raw audio". In: *arXiv:1609.03499*.

Ortega, Daniel and Ngoc Thang Vu (2017). "Neural-based Context Representation Learning for Dialog Act Classification". In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 247–252.

Pang, Bo and Lillian Lee (2004). "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts". In: *ACL*. ACL, p. 271.

Papaioannou, Ioannis, Amanda Cercas Curry, Jose L Part, Igor Shalyminov, Xinnuo Xu, Yanchao Yu, Ondrej Dušek, Verena Rieser, and Oliver Lemon (2017). "Alana: Social dialogue using an ensemble model and a ranker trained on user feedback". In: *Alexa Prize Proceedings*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting on association for computational linguistics*. ACL, pp. 311–318.

Pareti, Silvia (2012). "A Database of Attribution Relations." In: *LREC*, pp. 3213–3217.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 8024–8035. URL: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research*.

Petukhova, Volha, Martin Gropp, Dietrich Klakow, Anna Schmidt, Gregor Eigner, Mario Topf, Stefan Srb, Petr Motlicek, Blaise Potard, John Dines, et al. (2014). "The DBOX corpus collection of spoken human-human and human-machine dialogues". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC" 14)*. EPFL-CONF-201766. European Language Resources Association (ELRA).

Petukhova, Volha, Andrei Malchanau, and Harry Bunt (2014). "Interoperability of Dialogue Corpora through ISO 24617-2-based Querying". In: *LREC*.

Pichl, Jan, Petr Marek, Jakub Konrád, Martin Matulík, Hoang Long Nguyen, and Jan Šedivỳ (2017). "Alquist: The Alexa Prize Socialbot". In: *Alexa Prize Proceedings*.

Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman (2004). "Centering: A parametric theory and its instantiations". In: *Computational linguistics* 30.3, pp. 309–363.

Poesio, Massimo and David Traum (1998). "Towards an axiomatization of dialogue acts". In: *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues (13th Twente Workshop on Language Technology*. Citeseer.

Poesio, Massimo and David R Traum (1997). "Conversational actions and discourse situations". In: *Computational intelligence* 13.3, pp. 309–347.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber (2008). "The Penn Discourse TreeBank 2.0." In: *LREC*. Citeseer.

Price, Patti J (1990). "Evaluation of spoken language systems: The ATIS domain". In: *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Purandare, Amruta and Diane J Litman (2008). "Analyzing Dialog Coherence Using Transition Patterns in Lexical and Semantic Features." In: *FLAIRS Conference*, pp. 195–200.

Quarteroni, Silvia, Alexei V Ivanov, and Giuseppe Riccardi (2011). "Simultaneous dialog act segmentation and classification from human-human spoken conversations". In: *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, pp. 5596–5599.

Quarteroni, Silvia and Giuseppe Riccardi (2010). "Classifying dialog acts in human-human and human-machine spoken conversations". In: *INTERSPEECH 2010, 11th*

Annual Conference of the International Speech Communication Association, pp. 2514–2517.

Ram, Ashwin, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue (2017). "Conversational AI: The Science Behind the Alexa Prize". In: *1st proceedings of Alexa Prize*.

Ram, Ashwin, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. (2018). "Conversational ai: The science behind the alexa prize". In: *arXiv:1801.03604*.

Rastogi, Abhinav, Raghav Gupta, and Dilek Hakkani-Tur (2018). "Multi-task Learning for Joint Language Understanding and Dialogue State Tracking". In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 376–384.

Raymond, Christian and Giuseppe Riccardi (2007). "Generative and discriminative algorithms for spoken language understanding." In: *INTERSPEECH*, pp. 1605–1608.

Read, Jonathon, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg (2012). "Sentence boundary detection: A long solved problem?" In: *COLING (Posters)* 12, pp. 985–994.

Redeker, Gisela (2000). "Coherence and structure in text and discourse". In: *Abduction, belief and context in dialogue*, pp. 233–263.

Ritter, Alan, Colin Cherry, and William B Dolan (2011). "Data-driven response generation in social media". In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 583–593.

Sacks, Harvey and Gail Jefferson (1995). "Lectures on conversation". In:

Sacks, Harvey, Emanuel A Schegloff, and Gail Jefferson (1974). "A simplest systematics for the organization of turn-taking for conversation". In: *language*, pp. 696–735.

Sarikaya, Ruhi, Geoffrey E Hinton, and Anoop Deoras (2014). "Application of deep belief networks for natural language understanding". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.4, pp. 778–784.

Schegloff, Emanuel A (1968). "Sequencing in conversational openings". In: *American anthropologist* 70.6, pp. 1075–1095.

Schegloff, Emanuel A and Harvey Sacks (1973). "Opening up closings". In: *Semiotica* 8.4, pp. 289–327.

Searle, John (1965). "What is a speech act". In: *Perspectives in the philosophy of language: a concise anthology* 2000, pp. 253–268.

Serban, Iulian, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio (2016a). "A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues". In: *arXiv:1605.06069*.

Serban, Iulian V, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. (2017a). "A deep reinforcement learning chatbot". In: *arXiv preprint arXiv:1709.02349*.

Serban, Iulian V, Chinnadhurai Sankar, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Sarath Chandar, Nan Rosemary Ke, Sai Rajeswar, Alexandre de Brebisson, et al. (2017b). "The octopus approach to the Alexa competition: A deep ensemble-based socialbot". In: *Alexa Prize Proceedings*.

Serban, Iulian Vlad, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville (2016b). "Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation". In: *arXiv:1606.00776*.

Serban, Iulian Vlad, Ryan Lowe, Laurent Charlin, and Joelle Pineau (2016c). "Generative Deep Neural Networks for Dialogue: A Short Review". In: *arXiv:1611.06216*.

Serban, Iulian Vlad, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau (2016d). "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models." In: *AAAI*. Vol. 16, pp. 3776–3784.

Serban, Iulian Vlad, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio (2017c). "A hierarchical latent variable encoder-decoder model for generating dialogues". In: *AAAI*.

Shao, Louis, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil (2017). "Generating high-quality and informative conversation responses with sequence-to-sequence models". In: *arXiv:1701.03185*.

Shriberg, Elizabeth, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey (2004). *The ICSI meeting recorder dialog act (MRDA) corpus*. Tech. rep. INTERNATIONAL COMPUTER SCIENCE INST BERKELEY CA.

Sordoni, Alessandro, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan (2015). "A neural network approach to context-sensitive generation of conversational responses". In: *arXiv:1506.06714*.

Speer, Robyn and Catherine Havasi (2012). "Representing General Relational Knowledge in ConceptNet 5." In: *LREC*, pp. 3679–3686.

Sperber, Dan, Francesco Cara, and Vittorio Girotto (1995). "Relevance theory explains the selection task". In: *Cognition* 57.1, pp. 31–95.

Sperber, Dan and Deirdre Wilson (1986). *Relevance: Communication and cognition*. Vol. 142. Oxford: Blackwell.

Stent, Amanda, Matthew Marge, and Mohit Singhai (2005). "Evaluating evaluation methods for generation in the presence of variation". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 341–351.

Stolcke, Andreas, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer (2000a). "Dialogue act modeling for automatic tagging and recognition of conversational speech". In: *Computational linguistics* 26.3, pp. 339–373.

– (2000b). "Dialogue act modeling for automatic tagging and recognition of conversational speech". In: *Computational Linguistics* 26.3.

Stolcke, Andreas, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, Carol Van Ess-Dykema, et al. (1998). "Dialog act modeling for conversational speech". In: *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pp. 98–105.

Sun, Kai, Su Zhu, Lu Chen, Siqiu Yao, Xueyang Wu, and Kai Yu (2016). "Hybrid Dialogue State Tracking for Real World Human-to-Human Dialogues." In: *INTERSPEECH*, pp. 2060–2064.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*, pp. 3104–3112.

Taboada, María Teresa (2004). *Building coherence and cohesion: Task-oriented dialogue in English and Spanish*. Vol. 129. John Benjamins Publishing.

Takanobu, Ryuichi, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang (2020). "Is Your Goal-Oriented Dialog Model Performing Really Well? Empirical Analysis of System-wise Evaluation". In: *Proceedings of the 21st Annual SIGdial Meeting on Discourse and Dialogue*.

Tao, Chongyang, Lili Mou, Dongyan Zhao, and Rui Yan (2017). "RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems". In: *arXiv:1701.03079*.

Theil, Henri (1961). "Economic forecasts and policy". In:

Traum, David (1996). "Conversational agency: The TRAINS-93 dialogue manager". In: *In Susann LuperFoy, Anton Nijhholt, and Gert Veldhuijzen van Zanten, editors, Proceedings of Twente Workshop on Language Technology, TWLT-II*. Citeseer.

Traum, David R (1994). *A Computational Theory of Grounding in Natural Language Conversation.* Tech. rep. Rochester Univ NY Dept of Computer Science.

Traum, David R and Elizabeth A Hinkelman (1992). "Conversation acts in task-oriented spoken dialogue". In: *Computational intelligence* 8.3, pp. 575–599.

Tur, Gokhan and Renato De Mori (2011). *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Turing, Alan M (1950). "Computing machinery and intelligence". In: *Mind* 59.236, pp. 433–460.

Vakulenko, Svitlana, Maarten de Rijke, Michael Cochez, Vadim Savenkov, and Axel Polleres (2018). "Measuring semantic coherence of a conversation". In: *International Semantic Web Conference*. Springer, pp. 634–651.

Van Dijk, Teun A (1985). "Cognitive situation models in discourse production: The expression of ethnic situations in prejudiced discourse". In: *Language and social situations*. Springer, pp. 61–79.

– (1999). "Context models in discourse processing". In: *The construction of mental representations during reading*, pp. 123–148.

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *NIPS*, pp. 6000–6010.

Venkatesh, Anu, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. (2018). "On evaluating and comparing conversational agents". In: *arXiv preprint arXiv:1801.03625* 4, pp. 60–68.

Viera, Anthony J, Joanne M Garrett, et al. (2005). "Understanding interobserver agreement: the kappa statistic". In: *Fam med* 37.5, pp. 360–363.

Vinyals, Oriol and Quoc Le (2015). "A neural conversational model". In: *arXiv:1506.05869*.

Voorhees, Ellen M and Hoa Trang Dang (2003). "Overview of the TREC 2003 Question Answering Track." In: *TREC*. Vol. 2003, pp. 54–68.

Walker, Marilyn and Steve Whittaker (1990). "Mixed initiative in dialogue: An investigation into discourse segmentation". In: *Proceedings of the 28th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 70–78.

Walker, Marilyn A, Aravind K Joshi, and Ellen F Prince (1998). "Centering in naturally occurring discourse: An overview". In: *Centering theory in discourse* 128.

Walker, Marilyn A, Diane J Litman, Candace A Kamm, and Alicia Abella (1997). "PARADISE: A framework for evaluating spoken dialogue agents". In: *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 271–280.

Wallace, Richard S (2009). "The anatomy of ALICE". In: *Parsing the Turing Test*. Springer, pp. 181–210.

Wang, Ye-Yi, Li Deng, and Alex Acero (2005). "Spoken language understanding". In: *IEEE Signal Processing Magazine* 22.5, pp. 16–31.

Wang, Yu, Yilin Shen, and Hongxia Jin (2018). "A Bi-Model Based RNN Semantic Frame Parsing Model for Intent Detection and Slot Filling". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 309–314.

Wang, Zhuoran and Oliver Lemon (2013). "A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information". In: *Proceedings of the SIGDIAL 2013 Conference*, pp. 423–432.

Wei, Mengxi and Yang Zhang (2019). "Natural Answer Generation With Attention Over Instances". In: *IEEE Access* 7, pp. 61008–61017.

Weizenbaum, Joseph (1966). "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Communications of the ACM* 9.1, pp. 36–45.

Wen, TH, M Gašić, N Mrkšić, PH Su, D Vandyke, and S Young (2015). "Semantically conditioned lstm-based Natural language generation for spoken dialogue systems". In: *Conference Proceedings-EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pp. 1711–1721.

Wen, Tsung-Hsien, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young (2016a). "A Network-based End-to-End Trainable Task-oriented Dialogue System". In: *arXiv preprint arXiv:1604.04562*.

Wen, Tsung-Hsien, Milica Gašic, Nikola Mrkšic, Lina M Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young (2016b). "Multi-domain Neural Network Language Generation for Spoken Dialogue Systems". In: *Proceedings of NAACL-HLT*, pp. 120–129.

Williams, Jason, Antoine Raux, and Matthew Henderson (2016). "The Dialog State Tracking Challenge Series: A Review". In: *Dialogue & Discourse* 7.3, pp. 4–33.

Williams, Jason D and Steve Young (2007). "Partially observable Markov decision processes for spoken dialog systems". In: *Computer Speech & Language* 21.2, pp. 393–422.

Xie, Kaige, Cheng Chang, Liliang Ren, Lu Chen, and Kai Yu (2018). "Cost-sensitive active learning for dialogue state tracking". In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 209–213.

Xing, Chen, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma (2017). "Topic Aware Neural Response Generation." In: *AAAI*. Vol. 17, pp. 3351–3357.

Xiong, Wayne, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig (2016). "Achieving human parity in conversational speech recognition". In: *arXiv preprint arXiv:1610.05256*.

Xu, Puyang and Ruhi Sarikaya (2013). "Convolutional neural network based triangular crf for joint intent detection and slot filling". In: *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, pp. 78–83.

Yang, Xuesong, Yun-Nung Chen, Dilek Hakkani-Tur, Paul Crook, Xiujun Li, Jianfeng Gao, and Li Deng (2016). "End-to-end joint learning of natural language understanding and dialogue manager". In: *arXiv:1612.00913*.

Yang, Xuesong, Yun-Nung Chen, Dilek Hakkani-Tür, Paul Crook, Xiujun Li, Jianfeng Gao, and Li Deng (2017). "End-to-end joint learning of natural language understanding and dialogue manager". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5690–5694.

Yao, Kaisheng, Baolin Peng, Geoffrey Zweig, and Kam-Fai Wong (2016). "An attentional neural conversation model with improved specificity". In: *arXiv:1606.01292*.

Yi, Sanghyun, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur (2019).

"Towards Coherent and Engaging Spoken Dialog Response Generation Using Automatic Conversation Evaluators". In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 65–75. DOI: `10.18653/v1/W19-8608`. URL: `https://www.aclweb.org/anthology/W19-8608`.

Yin, Jun, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li (2016). "Neural generative question answering". In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, pp. 2972–2978.

Yoshino, Koichiro, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S Lasecki, Jonathan K Kummerfeld, Michel Galley, Chris Brockett, et al. (2019). "Dialog system technology challenge 7". In: *arXiv preprint arXiv:1901.03461*.

Young, Steve, Milica Gašić, Blaise Thomson, and Jason D Williams (2013). "Pomdp-based statistical spoken dialog systems: A review". In: *Proceedings of the IEEE* 101.5, pp. 1160–1179.

Yu, Zhou, Leah Nicolich-Henkin, Alan W Black, and Alex I Rudnicky (2016a). "A Wizard-of-Oz Study on A Non-Task-Oriented Dialog Systems That Reacts to User Engagement". In: *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 55.

Yu, Zhou, Ziyu Xu, Alan W Black, and Alexander Rudnicky (2016b). "Strategy and policy learning for non-task-oriented conversational systems". In: *SIGDIAL*, pp. 404–412.

Zhang, Ruqing, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng (2018). "Learning to Control the Specificity in Neural Response Generation". In: *Proceedings of the 56th Annual Meeting of the ACL (Volume 1: Long Papers)*. Vol. 1, pp. 1108–1117.

Zhang, Yizhe, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan (2019). "Dialogpt: Large-scale generative pre-training for conversational response generation". In: *arXiv preprint arXiv:1911.00536*.

Zhao, Tianyu and Tatsuya Kawahara (2018). "A Unified Neural Architecture for Joint Dialog Act Segmentation and Recognition in Spoken Dialog System". In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 201–208.

– (2019). "Joint dialog act segmentation and recognition in human conversations using attention to dialog context". In: *Computer Speech & Language* 57, pp. 108–127.

Zhou, Ganbin, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He (2017). "Mechanism-Aware Neural Machine for Dialogue Response Generation." In: *AAAI*, pp. 3400–3407.

Zhou, Xiangyang, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu (2018). "Multi-turn response selection for chatbots with deep attention matching network". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1118–1127.

Zhou, Yunxiao, Man Lan, and Wenting Wang (2019). "Hierarchical Intention Enhanced Network for Automatic Dialogue Coherence Assessment". In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8.

Zhu, Su and Kai Yu (2017). "Encoder-decoder with focus-mechanism for sequence labelling based spoken language understanding". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5675–5679.