Multifunctional ISO standard Dialogue Act tagging in Italian

Gabriel Roccabruna[1], **Alessandra Cervone**[2]*, **Giuseppe Riccardi**[1]
[1]Signals and Interactive Systems Lab, University of Trento, Italy, [2]Amazon Alexa AI
gabriel.roccabruna@studenti.unitn.it, giuseppe.riccardi@unitn.it

Abstract

English. The task of Dialogue Act (DA) tagging, a crucial component in many conversational agents, is often addressed assuming a single DA per speaker turn in the conversation. However, speakers' turns are often multifunctional, that is they can contain more than one DA (i.e. "I'm Alex. Have we met before?" contains a 'statement', followed by a 'question'). This work focuses on multifunctional DA tagging in Italian. First, we present iLISTEN2ISO, a novel resource with multifunctional DA annotation in Italian, created by annotating the iLISTEN corpus with the ISO standard. We provide an analysis of the corpus showing the importance of multifunctionality for DA tagging. Additionally, we train DA taggers for Italian on iLISTEN (achieving State of the Art results) and iLISTEN2ISO. Our findings indicate the importance of using a multifunctional approach for DA tagging.

1 Introduction

Dialogue Acts (DAs), a linguistically motivated model of speakers' intentions in a conversation, play a crucial role for several conversational AI tasks. DAs have been successfully used as part of conversational agents components, for example for Spoken Language Understanding (Zhao and Feng, 2018) or Natural Language Generation, and for response generation (Hedayatnia et al., 2020). Moreover, DAs have been shown to be important features to learn the intentional structure of conversations (Allen and Perrault, 1980; Cervone and Riccardi, 2020; Cervone et al., 2018).

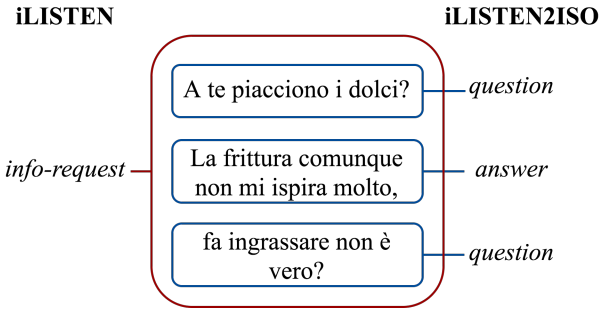


Figure 1: Example of the same turn with iLISTEN annotation versus our iLISTEN2ISO *multifunctional* annotation following the ISO standard. In this example, without the multifunctional approach a Conversational Agent would not understand that two different questions are asked.

One of the bottlenecks for current research on DAs is the lack of publicly available resources with DA annotation. While this is true also for English, it is even more important for languages with fewer resources, such as Italian. For Italian, the only publicly available resource with DA annotation is currently the iLISTEN corpus (Basile and Novielli, 2018), released for EVALITA in 2018.

While useful, this resource relies on an annotation scheme which assumes only one single DA per conversational turn (see Figure 1). However, ISO 24617-2 (Bunt et al., 2010; Bunt et al., 2020), the latest accepted standard for DA annotation, posits that conversational turns can be multifunctional in a sequential way, i.e. speakers' turns can be composed of multiple DAs in sequence (Huang, 2017).

In this work, we investigate the task of multifunctional DA tagging in Italian. The contributions of this paper are: (1) we create iLISTEN2ISO, to the best of our knowledge the first

* Work done while at University of Trento, prior to joining Amazon.

publicly available resource with DA annotation in Italian which uses a *multifunctional* approach and is *ISO-standard compliant*; (2) we present an analysis of iLISTEN2ISO showing the importance of multifunctional DA annotation; (3) we propose baseline DA tagging models for Italian trained on iLISTEN (achieving, to the best of our knowledge, SOTA results) and iLISTEN2ISO.[1]

2 Related work

Dialogue act corpora Most publicly available resources with DA annotation are hardly compatible, given that each resource is typically tagged with its own different scheme tailored for a given domain (Carletta et al., 1997). This prevents both meaningful comparisons among different resources, and the possibility of experimenting with cross-corpora training of DA taggers. ISO 24617 (Bunt et al., 2010), the latest universally accepted standard for DA annotation, represents an attempt to overcome this fragmentation by providing a domain- and task-independent taxonomy, useful for both task- and non-task-oriented dialogue. Compared to previous schemes, the ISO standard is *multifunctional*, both from a sequential perspective (the same turn can contain multiple DAs in sequence) and from a simultaneous perspective (a text span can have multiple DA tags at once). Moreover, the ISO standard is a hierarchical taxonomy, rather than a flat one, which enables it to capture similarities among different tags. Sequential multifunctionality is also present in the DAMSL schema (Core and Allen, 1997), although this definition is not commonly applied to corpora that adopted DAMSL (Chowdhury et al., 2016), with the consequent possibility of introducing ambiguities and a lack of precision in understanding the communicative functions of text spans.
While for English there have recently been successful attempts to create publicly available resources mapped to ISO 24617-2 (Mezza et al., 2018); datasets mapped to ISO are scarcely available for other languages, see for example (Ngo et al., 2018) for Vietnamese and (Yoshino et al., 2018) for Japanese. For the Italian language, the only corpus with a subset of dialogues tagged with ISO in a multifunctional way is LUNA (Chowdhury et al., 2016), which is currently not publicly available.

Dialogue Act tagging DA tagging is the task of assigning a DA tag to a given utterance in a dialogue. The definition of utterance depends on the schema used: in some schemes (Dinarelli et al., 2009), the utterance corresponds to a turn, while in others (Jurafsky, 1997) to segments of a turn. DA tagging is usually framed as text classification (Lee and Dernoncourt, 2016; Mezza et al., 2018) or as a sequence tagging problem (Quarteroni et al., 2011; Chen et al., 2018; Colombo et al., 2020).

3 iLISTEN2ISO: Mapping iLISTEN to ISO standard

The iLISTEN corpus (Basile and Novielli, 2018) is a dataset of dyadic dialogues about food and dietary issues in Italian annotated with DAs, used during the 2018 EVALITA competition for a DA classification task. The corpus consists of 60 dialogues, with 1576 user turns and 1611 system turns. Dialogues were collected with a Wizard of Oz procedure using either written (30) or spoken (30) interactions. The system side mimics a diet therapist, asking questions about users diets or answering to users' questions. The DA schema adopted is a refined version of DAMSL (Core and Allen, 1997). As reported in Table 1, the number of DAs in the schema is 15, where 7 are reserved only to users, 6 only to the system and the remaining 2 are in common.

In iLISTEN the turn DA annotation is not multifunctional, i.e. each turn is assigned one single DA. However, not tackling the turn DAs with a multifunctional approach could result in loss of information, with the DA tag capturing only the most dominant function of a turn. In Figure 1, for example, tagging the entire turn with one DA would prevent the system from understanding that two different questions are asked.

In order to create the iLISTEN2ISO annotation, each turn from iLISTEN was annotated with a *multifunctional* approach following the ISO standard. This process involved first segmenting turns into functional units (FUs), defined as minimal stretches of communicative behaviour that have a communicative function (Bunt et al., 2010); and then annotating each FU with a DA tag. The subset of ISO schema used for mapping iLISTEN to ISO was build incrementally, since an a-priori definition was impossible due to the fact that many communicative functions were hidden by the lack

[1] iLISTEN2ISO and the code of our experiments are available at: `https://github.com/BrownFortress/Multifunctional-Dialogue-Act-tagging-in-Italian`.

of segmentation. This annotation process involved user and system turns, since system turns are used as context in the prediction phase. The annotation of system turns was done only on unique turns, given the repetitiveness of system turns (only 430 of 1611 are unique).

Because of the lack of resources, the segmentation and mapping process of iLISTEN was conducted by one single annotator, under the supervision of a second annotator with previous training in ISO standard annotation. In order to ensure a reliable annotation process, after the creation of the guidelines, the second annotator repeatedly assessed a sample (100 utterances) of the annotated data. This sample was built through a stratified random sample, where for each DA tag, 20% of examples of that class was randomly sampled. This evaluation and reassessment was performed twice. In the first round, performed after the first annotation of the data, some issues regarding the usage of some DAs arose and were discussed; in the second examination, performed after the second phase of annotation of the data, no problem was found.

4 Analysis of iLISTEN2ISO

The annotation layer of iLISTEN2ISO does not change only the legacy iLISTEN schema, but also the internal structure of turns due to the segmentation process. On average in iLISTEN2ISO we have 1.61 FUs per turn, which become 1.81 on system side and 1.5 on user side if we consider them separately (this difference is justified by the fact that the system turns are on average longer than user turns). In Figure 2, we report for each legacy DA (user side), the number of segments per turn on average. Furthermore, inside the bars we list the 3 most common sequences of ISO DAs to which each legacy DA is mapped. In Table 1, we compare the number of DA tags between iLISTEN and iLISTEN2ISO. We notice that iLISTEN2ISO has a larger number of DAs in total, compared to iLISTEN. Additionally, while in iLISTEN the number of DAs in common (2) is much lower compared to either user or system DAs, in iLISTEN2ISO the common DAs between system and user are larger than the independent ones, with the advantage of potential better generalization across the two. Looking at the distribution of the ISO DAs regarding user turns, it can be noticed that the four most common DAs are: *inform* 24.5%, *question* 21.3%, *answer* 15.3% and *auto-positive* 7%. Moreover, the DA distribution has a tail composed of 19 DAs with a frequency below the 5%. However, this is not a drawback of the scheme since it gives us a fine-grained representation of the actions performed by the user. Additionally, iLISTEN2ISO can be used in conjunction with any other corpus annotated with ISO standard thus, giving the possibility of augmenting the samples for a specific low-represented class.

	User	**System**	**Common**	**Total**
iLISTEN	7	6	2	15
iLISTEN2ISO	10	2	15	27

Table 1: Number of Dialogue Acts (DAs) used by the system and the user in iLISTEN and iLISTEN2ISO (multifunctional). "Common" reports the number of DAs used by both system and user.

5 Models

In this section, we describe the two baseline models for Dialogue Act (DA) classification used in our experiments. The first model is a Support Vector Machine (SVM) (Vapnik, 1995) with linear kernel, with One versus One strategy. The features used are: FastText word embeddings, Part-Of-Speech (POS) and dependency parsing tags (DEP) (retrieved using Spacy), and the previous DA tag. For word embeddings, the utterance representation is computed using the average of the relative word embeddings. The model was implemented using scikit-learn (Pedregosa et al., 2011). Hyperparameters and features selection was performed using 3 folds cross-validation. The feature vector that gave the best results for iLISTEN is the concatenation of word embeddings, POS tags, DEP tags and the previous DA. For iLISTEN2ISO, the feature vector that gave the best performances is the concatenation of word embeddings and the previous DA.

Our second model is a Convolutional Neural Network (CNN), following (Lee and Dernoncourt, 2016). The utterance representation is computed using a CNN taking as input FastText word embeddings. This representation is then concatenated with the previous DA and passed through a linear and a softmax output layer. We use cross entropy loss optimized with Adam and early stopping according to best Macro F1 on a randomly generated development set (6 dialogues), chosen for the

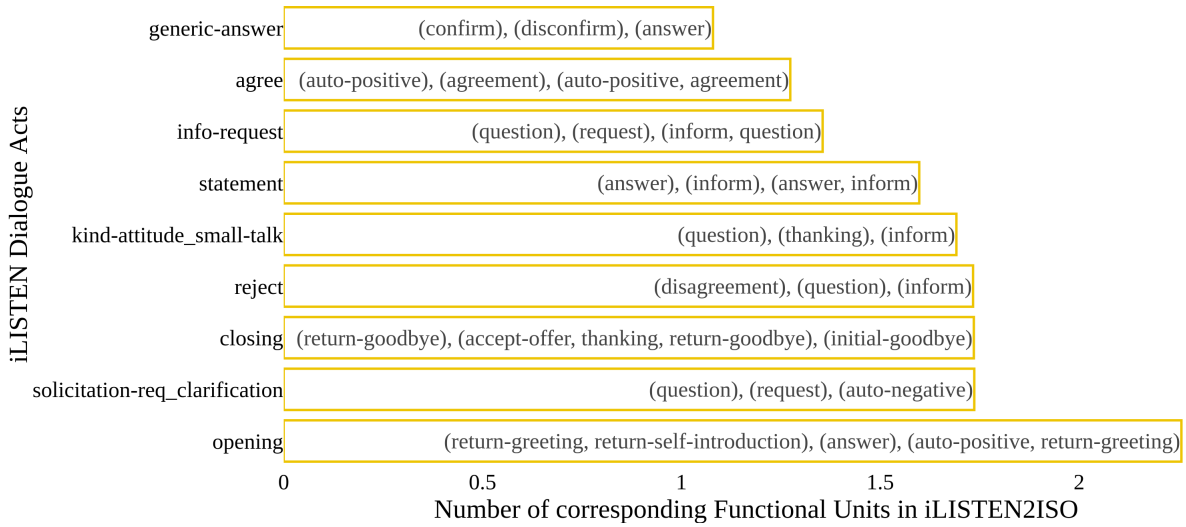


Figure 2: For each iLISTEN user DA, we report the corresponding average number of corresponding Functional Units and (inside the bars) the three most common DA sequences in iLISTEN2ISO.

lowest tags distribution difference compared to the full dataset. The learning rate is set to 10^{-3} and the batch size to 128. The number of filters is 200 and the filters sizes are 1,2,3 and 4 times the word embedding dimension (300).

6 Experiments

In this section, we report the results of Dialogue Act (DA) tagging experiments, using our proposed baselines on both legacy (using iLISTEN) and *multifunctional* ISO standard DA schemes (using iLISTEN2ISO).

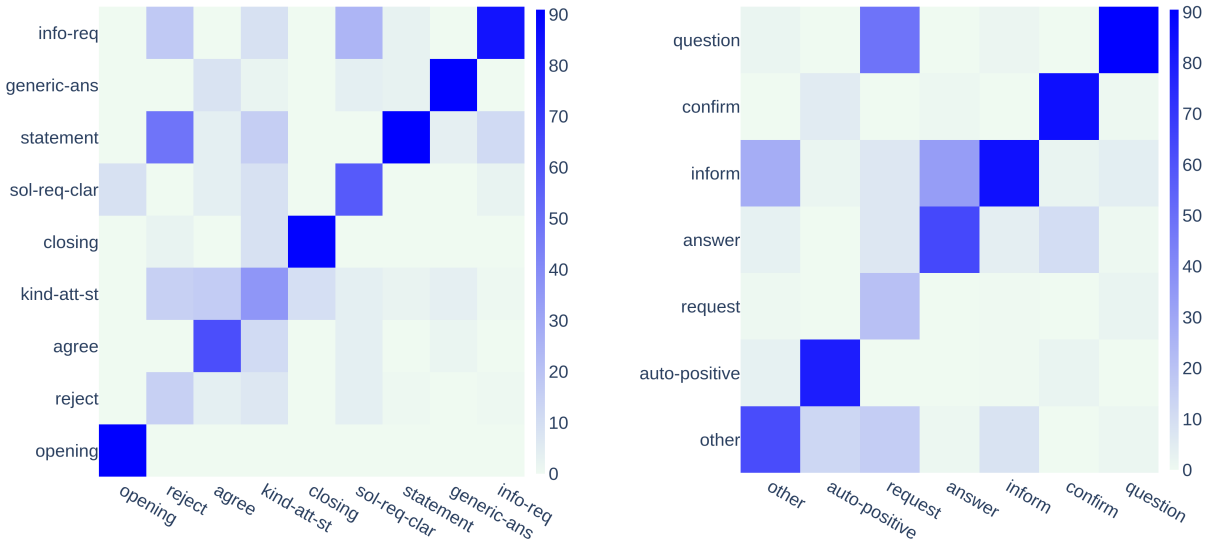
Experimental setup For comparison with previous work, we follow the competition rules and report results considering only user DAs, using official splits. Additionally, we do not assume gold DAs for the context for testing (which might not be available at inference time), rather we use predicted ones. In order to do this we train a separate model for tagging system DAs used only during inference. The performances of system models are: Micro F1 96.1% and Macro F1 96.6% on iLISTEN; Micro F1 97.5% and Macro F1 96.3% on iLISTEN2ISO. For iLISTEN, the obtained classification results are compared with Unitor, the winner of the EVALITA competition (Basile and Novielli, 2018) and to the best of our knowledge the SOTA on iLISTEN (we could not perform comparisons for iLISTEN2ISO, as the code is not publicly available). Given the larger number of DA tags with few examples in iLISTEN2ISO, for comparison with the legacy scheme we group the least frequent DA tags to the label "Other". The final DA scheme for iLISTEN2ISO consists of 7 DAs. In iLISTEN the number of examples in training and testing is 1097 and 479 respectively; in iLISTEN2ISO we have 1609 and 777 respectively.

Dataset	Model	Macro F1	Micro F1
iLISTEN	Unitor	63.7	73.2
	SVM	67.3	**75.1**
	CNN	68.0	75.0
iLISTEN2ISO	SVM	69.3	74.8
	CNN	**71.4**	74.9

Table 2: Results of Dialogue Act (DA) tagging using iLISTEN legacy annotation and iLISTEN2ISO multifunctional annotation.

Results As shown in Table 2, our proposed models yield comparable results on both non-multifunctional (iLISTEN) and multifunctional (iLISTEN2ISO) DA tagging. On iLISTEN, our models even overcome previous SOTA performances (Unitor) on both Micro and Macro F1. We observe that while in terms of Micro F1 our models achieve very similar results on both corpora, in terms of Macro F1 they perform better on multifunctional DA tagging.

Error analysis To better understand the performance of our models on iLISTEN and iLISTEN2ISO, we look at the confusion matrices depicted in Figure 3 and Tables 3 and 4 reporting the performances computed for each DA.



(a) iLISTEN

(b) iLISTEN2ISO

Figure 3: Confusion matrices of the CNN model on iLISTEN (a) and iLISTEN2ISO (b). The presented values are in percentage. To improve the readability of Figure (a) we used some abbreviations: sol-req-clar corresponds to *solicitation-req-clarification* and kind-att-st corresponds to *kind-attitude-small-talk*.

Dialogue Acts	**F1 scores**			**Freq.**
	Unitor	**SVM**	**CNN**	
statement	83.6	83.2	83.8	34%
info-request	80.1	81.3	82.4	23.4%
generic-ans	88.8	89.5	87.0	10.9%
kind-att-st	43.8	55.8	40.5	9.2%
reject	13.0	13.0	23.0	8.1%
agree-accept	53.6	66.7	65.2	5%
sol-req-clar.	48.9	52.0	59.6	5%
opening	100.0	90.9	95.2	2.3%
closing	73.6	73.7	75.0	2.1%
Macro F1	63.7	67.3	68.0	
Micro F1	73.2	75.1	75.0	

Table 3: This table reports the F1 results for each **iLISTEN** Dialogue Act achieved by Unitor, SVM and CNN models. All the values reported are in percentage. The last column (*Freq.*) reports the frequencies of the Dialogue Acts in the test set.

Dialogue Acts	**F1 scores**		**Freq.**
	SVM	**CNN**	
inform	76.6	76.3	30.5%
other	63.4	67.4	19.8%
question	84.2	85.7	17.5%
answer	77.8	68.7	13.4%
auto-positive	80.0	83.0	7.1%
confirm	83.7	87.2	6.2%
request	22.2	31.6	5.5%
Macro F1	69.9	71.4	
Micro F1	74.8	74.9	

Table 4: This table reports the F1 scores for each **iLISTEN2ISO** Dialogue Act achieved by SVM and CNN models. All the values reported are in percentage.The last column (*Freq.*) reports the frequencies of the Dialogue Acts in the test set.

Considering the CNN performance, looking at confusion matrices in Figure 3, we notice that on iLISTEN the worst class is *reject* where 48.7% of examples are predicted as *statement*. This is probably due to the similar structure of *reject* utterances to *statement* ones, while the discriminant is the semantic content that model fails to detect. This problem can be seen also in Table 3, where the *reject* DA is predicted with the worst performances among other tags. An example of error is given by the following interaction: the system says "Mangiare ad orari fissi e' un modo per evitare di saltare i pasti e di trascurare sostanze che spesso non vengono compensate nei pasti successivi." and the user responds "purtroppo spesso il lavoro limita la possibilità di fare una dieta sana e regolare.". This user's turn is tagged with *reject* but it is predicted by the model as *statement*. As it can be seen, the structure of the user's turn is similar to a statement because the user expresses her or his opinion, in this case regarding the difficulty to follow an healthy diet.

Another interesting mismatch in iLISTEN regards *info-request*, 11.6% of which are predicted as *statement*. This is interesting because the class *info-request* is usually composed of questions, however analyzing heuristically the examples we notice that some of them contains other tags, such as answers or statements, which are hidden in the legacy annotation. In this regard, another potential source of error is the lack of punctuation as it can be seen in the utterance "è necessario fare sport per mantenersi in forma". This utterance can be interpreted as a statement, but if a question mark is added at the end of the utterance it can be interpreted as a question. This also highlights the importance of punctuation or prosodic features in order to detect the right DA.

Another problem, that can be identified looking at the iLISTEN confusion matrix in Figure 3, is that the *kind-attitude-smalltalk* DA is confused with many different others DAs. This is due to lack of segmentation since analysing the ISO DAs distribution of the turns tagged with this tag, it emerged there is not a predominant DA. In fact, the four most common ISO DAs are: *inform* 21.3%, *question* 20.9%, *thanking* 13.5% and *auto-positive* 10.8%.

Regarding the iLISTEN2ISO confusion matrix, it can be seen that *request* is the most confused class. Indeed, 48.8% of examples are predicted as *question*, 16.3% as *other* and only 20.9% are predicted correctly. The reason behind this performance is that the model fails to distinguish a request from a question since both of them are in a question style.

Another frequently mispredicted DA in iLISTEN2ISO is *answer*, often confused with *inform*. This is due to the fact that the model has difficulties in representing and then distinguishing the semantic content. Moreover, as it can be noticed in Table 4 this problem is more highlighted in the CNN's rather than in SVM's performances.

Finally, comparing the iLISTEN2ISO results presented in Table 4 with iLISTEN results presented in Table 3, it can be seen that the *question* DA is better predicted than *info-request*. In this case, only 4.4% of *question* examples are confused with *inform*. The reason of this improvement is probably the segmentation process that highlighted the multifunctionality of the utterances augmenting the specificity of the classes.

Interestingly, if we compare confusion matrices for SVM (which we decided not to include in the paper for lack of space) and CNN, shown in figure 3, we notice that the most confused classes are the same for both models across both datasets.

7 Conclusions

We presented iLISTEN2ISO, a resource for Italian multifunctional DA tagging using ISO 24617-2. We argued the importance to consider turns as a composition of multiple communicative functions, in order to preserve important semantic information. Moreover, we presented different baseline DA tagging models, on both iLISTEN and iLISTEN2ISO.

We believe the presented resource could be useful to the research community for experimenting with multifunctional DA tagging in Italian, as well as cross-corpora DA tagging. As future work, we plan to explore joint DA segmentation and classification in Italian, for example taking inspiration from the work presented by Zhao and Kawahara (2019).

Acknowledgements

The research leading to these results has received funding from the European Union – H2020 Programme under grant agreement 826266: COADAPT.

References

James F Allen and C Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial intelligence*, 15(3):143–178.

Pierpaolo Basile and Nicole Novielli. 2018. Overview of the evalita 2018 italian speech act labe ling (iliste n) task. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:44.

Harry Bunt, 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. *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Harry Bunt, 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*, pages 549–558.

Jean Carletta, Stephen Isard, Gwyneth Doherty-Sneddon, Amy Isard, Jacqueline C Kowtko, and Anne H Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational linguistics*, 23(1):13–31.

Alessandra Cervone and Giuseppe Riccardi. 2020. Is this Dialogue Coherent? Learning from Dialogue Acts and Entities. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 162–174.

Alessandra Cervone, Evgeny Stepanov, and Giuseppe Riccardi. 2018. Coherence models for dialogue. In *Proc. Interspeech*, pages 1011–1015.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 225–234.

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

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *AAAI*, pages 7594–7601.

Mark G. Core 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*.

Marco Dinarelli, Silvia Quarteroni, Sara Tonelli, Alessandro Moschitti, and Giuseppe Riccardi. 2009. Annotating spoken dialogs: from speech segments to dialog acts and frame semantics. In *Proceedings of the 2nd Workshop on Semantic Representation of Spoken Language*, pages 34–41. Association for Computational Linguistics.

Behnam Hedayatnia, Seokhwan Kim, Yang Liu, Karthik Gopalakrishnan, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialogue systems. *arXiv preprint arXiv:2005.12529*.

Yan Huang. 2017. *The Oxford handbook of pragmatics*. Oxford University Press.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function. *Annotation, Technical Report, 97-02, University of Colorado, CO, USA*.

Ji Young Lee 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*, pages 515–520.

Stefano Mezza, 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*, pages 3539–3551.

Thi-Lan Ngo, Pham Khac Linh, and Hideaki Takeda. 2018. A vietnamese dialog act corpus based on iso 24617-2 standard. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

F. Pedregosa, 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. *Journal of Machine Learning Research*.

Silvia Quarteroni, 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*, pages 5596–5599. IEEE.

V.N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.

Koichiro Yoshino, Hiroki Tanaka, Kyoshiro Sugiyama, Makoto Kondo, and Satoshi Nakamura. 2018. Japanese dialogue corpus of information navigation and attentive listening annotated with extended iso-24617-2 dialogue act tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Lin Zhao and Zhe Feng. 2018. Improving slot filling in spoken language understanding with joint pointer and attention. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 426–431.

Tianyu Zhao and Tatsuya Kawahara. 2019. Joint dialog act segmentation and recognition in human conversations using attention to dialog context. *Computer Speech & Language*, 57:108–127.