Speech Adaptation Modeling for Statistical Machine Translation

By

NICHOLAS W. RUIZ

International Doctoral School of
Information and Communication Technologies
UNIVERSITY OF TRENTO

A dissertation submitted to the University of Trento in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in the Faculty of Information and Communication Technologies.

APRIL 2017

Abstract

Spoken language translation (SLT) exists within one of the most challenging intersections of speech and natural language processing. While machine translation (MT) has demonstrated its effectiveness on the translation of textual data, the translation of spoken language remains a challenge, largely due to the mismatch between the training conditions of MT and the noisy signal that is output by an automatic speech recognition (ASR) system. In the interchange between ASR and MT, errors propagated from noisy speech recognition outputs may become compounded, rendering the speech translation to be unintelligible. Additionally, aspects such as stylistic differences between written and spoken registers can lead to the generation of inadequate translations. This scenario is predominantly caused by a mismatch between the training conditions of ASR and MT. Due to the lack of training data that couples speech audio with translated transcripts, MT systems in the SLT pipeline must rely predominantly on textual data that does not represent well the characteristics of spoken language. Likewise, independence assumptions between each sentence results in ASR and MT systems that do not yield consistent outputs.

In this thesis develop techniques to overcome the mismatch between speech and textual data by improving the robustness of the MT system. Our work can be divided into three parts. First we analyze the effects the difference between spoken and written registers has on SLT quality. We additionally introduce a data analysis methodology to measure the impact of ASR errors on translation quality. Secondly, we propose several approaches to improve the MT component's tolerance of noisy ASR outputs: by adapting its models based on the bilingual statistics of each sentence's neighboring context, and through the introduction of a process by which textual resources can be transformed into synthetic ASR data to use when training a speech-centric MT system. In particular, we focus on the translation from spoken English to French and German – the two parent languages of English – and demonstrate that information about the types and frequency of ASR errors can improve the robustness of machine translation for SLT. Finally, we introduce and motivate several challenges in spoken language translation with neural machine translation models that are specific to their modeling architecture.

Keywords: natural language processing, spoken language translation, statistical machine translation, automatic speech recognition, error analysis

DEDICATION AND ACKNOWLEDGMENTS

I am glad to have had the opportunity to work at the Human Language Technologies group at the Fondazione Bruno Kessler. I would first like to express my gratitude to my advisor, Marcello Federico, for his support, encouragement, and advice during the completion of the thesis. Marcello's expertise and excitement for statistical machine translation helped me find my research interests. I greatly appreciate the opportunity to work alongside him and other colleagues at FBK-irst and look forward to future research collaboration.

I would also like to express my gratitude for the following people at FBK:

- Nicola Bertoldi for constantly having an open office door and for his willingness not only to answer an immediate question, but also to immerse himself in the research problems of others to provide insightful suggestions;
- Matteo Negri and Marco Turchi for their endless excitement for new research and useful feedback;
- Mauro Cettolo, Roldano Cattoni, Luisa Bentivogli, Daniele Falavigna, and Roberto Gretter for providing datasets to support my research;
- Arianna Bisazza for being a local standard of excellence as a PhD researcher;
- Mattia Antonino di Gangi for collaboration with Neural Machine Translation experiments;
- Prashant Mathur and M. Amin Farajian for sharing an office with me and for being the first to hear my next crazy idea;
- My other student colleagues: José Guilherme Camargo de Souza, Shahab Jalalvand, Gözde Özbal, Serra Sinem Tekiroglu, Ngoc Phuoc An Vo, Simone Magnolini, Anna Feltracco, Rajen Chatterjee, and Mohammed Qwaider.

Thanks to all of you for making Trento feel more like home.

I would also like to thank Microsoft Research for an unforgettable internship experience. In particular, I would like to thank Qin Gao, William Lewis, Anthony Aue, Hany Hassan, Xiaodong He, Jonathan Clark, and Arul Menezes for inviting me into an exciting industrial research environment where the barriers are constantly being pushed in spoken language translation.

I would like to thank my dissertation committee: Loïc Barrault, Stefan Riezler, and François Yvon, for their insightful discussion and useful feedback.

I would like to thank the Chiesa Evangelica di Trento for being a welcoming place for foreigners of all languages and nationalities and for being intentional about interacting with foreigners like me, even when language barriers made it awkward.

I would like to thank my family for their constant support, from what was supposed to be a two European Masters degree, but grew into a six year period in Europe. Thanks for not forgetting about us while we were so far away.

Thanks to Jennifer, my pillar of support and my anchor; my home whenever I felt homeless; my reason for joy and contentment in the midst of challenges. Thank you for your love and sacrifice and for starting a family and home for us right in Trento. I am overwhelmed by God's blessings every time I look at you and our children.

Finally, I thank God for where He has taken me. I had never imagined a season of research in Europe, nor how the people I would meet along the way would change the way I look at the world. I am thankful that although these steps in isolation might seem unexpected and absurd, I can see a clear path and a story unfolding. Thanks for leading me to Trento, and thanks for the places You are leading me now.

TABLE OF CONTENTS

List of Tables

List of Figures

FORWARD

I was sitting in a chair during a church service one day, thinking about how difficult cross-lingual communication can be. I recall the pastor was sharing an interesting message, but I can't remember its contents because I was too busy thinking about how I could make it possible for other non-native speakers to understand what was being said without our interpreter being present. About 10 foreign students and refugees were crowding around an older woman who volunteered to interpret during a 90 minute service, despite lacking formal training from a translation school. After every service she would be exhausted from the cognitive efforts of simultaneously listening and translating to a message without much context known in advance.

I had thought about giving interpretation a try, but I, too, was a non-native speaker and would not have been able to convey all of the message in an adequate message. After all, I was a PhD student living in Trento, Italy, who was also working hard to understand and participate in speeches and discussions in Italian.

These translation scenarios and the future possibility of applying my research in spoken language translation to bridge linguistic divides was the primary motivation for spending five years in Italy, working at the University of Trento and the Fondazione Bruno Kessler laboratory. I didn't just want to address the research problems from within a lab – I wanted to experience the frustrations of not being able to communicate my needs, wants, and desires with the same finesse and fluency that I was able to do in the United States as a native English speaker. I wanted to enter communication scenarios where I would be forced to humble myself and to spend more time listening than speaking. I wanted to meet other people who shared in those struggles, whether it was a voluntary or forced decision and to hear their stories about how hard it was to even do the simple day-to-day societal activities.

Spoken language translation is not only a research problem to be solved, but also a societal problem that continues to grow with each year. The five years devoted to this research consisted in understanding the societal needs for spoken language translation technology and analyzing the drawbacks that are preventing current approaches from

being widely adopted. While the problem of spoken language translation has a long way to go before being fully solved for even the simplest language pairs, the completion of the research in this thesis motivates me even more to work toward the goal of delivering translation technologies that enable the linguistically disadvantaged to thrive in a multilingual world.

CHAPTER 1

INTRODUCTION

Multilingualism is entering the Western world at an accelerating pace. The United States Center for Immigration Studies reported that, as of 2015, 21.5 percent of United States residences speak a foreign language at home (Camarota and Zeigler, 2016). The increase of foreign language speakers settling in the West and overall globalization has resulted in an increase in demand for multilingual technologies to bridge the linguistic divide. While linguistic barriers often coincided with political and geographic barriers in the past, global migration has shifted this problem to local communities.

As a result, there is an ever-growing need for translation to ensure that all members of a society can consume and distribute information. The need for widespread translation is seen from the perspective of individuals, to government entities, such as the European Commission Directorate-General for Translation, which reported translating nearly two million pages of content across the 24 official European Union languages in 2015 (EC DGT, 2016). In particular, there is a growing need for translation of the following mediums of communication:

- Text (e.g. literature, legal and bureaucratic documents, online content);
- Media (e.g. news broadcasts, movie subtitling, closed captioning);
- Live speeches (e.g. political speeches, lecture translation);
- Human to human communication (e.g. Skype™, face-to-face conversations)

While professional translators and interpreters work to alleviate some of the need for translation, translation projects are costly and time-consuming. In addition, there are various modalities of human communication that are disrupted by involving a human mediator, such as Skype™ calls and face-to-face conversations.

Two promising technologies that have seen increasing adoption are statistical machine translation (SMT) and spoken language translation (SLT). Statistical methods for translation have benefited society by augmenting human translation and interpretation with a scalable, versatile, and cost-effective alternative. As opposed to human translation, which requires translators to not only acquire fluency in each natural language, but also to specialize in the generation of language that meets top literacy standards, the core approaches of SMT vary little with respect to each language pair. This allows a SMT architecture to generalize across combinations of language pairs spanning multiple language families, with the only requirement being the availability of a sufficient amount of language bitexts to use for training. Although machine translation has not reached parity with human translation, it is helping us bridge linguistic divides by making information available beyond the language barrier.

While statistical machine translation has demonstrated its effectiveness on textual data such as newswire, government proceedings, patents, and Internet texts, the translation of spoken language has been challenging. SLT minimally consists of two main components: an automatic speech recognition (ASR) system, which decodes a speaker's speech in a source language into a sequence of words, and a machine translation (MT) system, which translates the transcribed words into a target language. Each of these components has its own unique challenges that the underlying models need to overcome.

Some of the problems faced in SLT are shared with human interpreters and professional translators. Given the task to communicate the content from a foreign language into a target language, the interpreter must process the sequence of spoken words she hears and plan a sequence of utterances to communicate the information received to the recipient according to the time restrictions and quality expectations required for the task. In some language pairs, this process can be localized, allowing an interpreter to translate based on a localized context of words. Likewise, machine translation systems must process an input source, whether it be text or speech, and generate an adequate response in a target language. However, ASR introduces more challenging problems over the auditory process of word recognition. Even in the best of audio recording scenarios, ASR must deal with signal noise, speaker dialect and pronunciation variations, disflu-

encies in the form of filler words (“um”, “uh”), repeated syllables, and under-articulation. The best performing ASR systems still produce errors and ungrammatical outputs, so how can the machine translation component identify these errors and work around them in order to provide a faithful translation of speech?

A further exacerbation of the problem is due to the fact that most SLT systems are a pipeline in which ASR and MT systems are largely trained independently from one another – mostly due to the lack of data that covers the recording of human speech and its professional translation. The majority of the available training data stops at the production of source language transcripts, or it begins with the written word and provides its professional translation, but lacks an audio recording for speech processing. An ideal SLT training scenario consist of a substantially large corpus of data consisting of audio, source transcript, and target translation tuples in order to use for training both the ASR and MT systems to ensure their underlying models are optimized for the target SLT task. However, this data is not readily available due to the time and costs involved in collecting the data. As a result, SLT systems are impeded by the problem domain mismatch; while the ASR system is trained on data largely representing conversational registers, the MT system is trained on a majority of written register data with clean, syntactically-correct transcripts. Statistical differences between the speech and text training corpora create a scenario where MT may not be properly trained to anticipate ASR errors, as well as differences in registers. As a result, MT will compound the errors already made by an ASR system, rendering many speech translation outputs to be unintelligible. This result is unacceptable in communities that have increasing reliance on translation technologies to ensure the spread of information across linguistic barriers.

In this thesis, we aim to address the problem of domain mismatch and lack of error tolerance in machine translation. In particular, we focus on the problem of context adaptation and ASR error modeling in spoken language translation. By context adaptation, we mean two things: adapting MT systems to anticipate the difference between spoken and written registers to tolerate the kinds of input coming from speech recognition hypotheses, and adapting MT models to remember previous inputs and decisions made by the MT engine. By error modeling, we intend to model channel noise as a process that estimates confusions made by the ASR system as a source of data to enhance the MT system’s tolerance of ASR errors.

SRC_1	You know, cadaver dissection **is** the traditional way of **learning** human **anatomy**.
ASR_1	**Seeing a**, cadaver dissection **and ease** the traditional way of **loaning** human **and that to me**.
REF_1	Vous savez, la dissection cadavérique est la manière traditionnelle d'apprentissage de l'anatomie humaine.
MT_1	Vous savez, la dissection de la cadaver est la façon traditionnelle d'apprendre l'anatomie humaine.
SLT_1	En voyant une dissection, la dissection et la facilité la façon traditionnelle d'être humain et de moi.
SRC_2	For students, it's quite an experience, but for **a** school, it **could** be very difficult or expensive to maintain.
ASR_2	For students, it's quite an experience, but for school, it **can** be very difficult or expensive to maintain.
REF_2	Pour les étudiants, c'est une véritable expérience, mais pour une école, ça pourrait être très difficile ou coûteux à entretenir.
MT_2	Pour les étudiants, c'est plutôt une expérience, mais pour une école, ça pourrait être très difficile ou cher à entretenir.
SLT_2	Pour les étudiants, c'est plutôt une expérience, mais pour l'école, ça peut être très difficile ou cher à entretenir.
SRC_3	So we learned the majority of anatomic classes taught, they do not have a cadaver dissection lab.
ASR_3	So **real and promote unity Obama panic class thought**, they do not have a **kind of a** dissection lab.
REF_3	Donc nous avons appris que la majorité des classes anatomiques n'a pas de laboratoire de dissection cadavérique.
MT_3	Nous avons donc appris la majorité des cours anatomiques qu'ils ont appris, ils n'ont pas un laboratoire de dissection de cadaver.
SLT_3	Si vrai et promouvoir l'unité de la classe de panique d'Obama, ils n'ont pas une sorte de laboratoire de dissection.

Figure 1.1: An example of spoken language translation errors. SLT errors caused by ASR errors are highlighted in red. MT-related errors are highlighted in blue.

1.1 Motivating Example

When evaluating the performance of a SLT system, we can isolate the sources of errors as being predominantly caused by either the ASR or the MT system. As we analyze an except from a TED talk in Fig. 1.1, we observe several interesting SLT errors that motivate our work. The examples come from the 2012 International Workshop on Spoken Language Translation (IWSLT) evaluation campaign, which focused on the translation of lectures.

The ASR output of the first sentence (ASR_1) of Figure 1.1 is unintelligible. Recognition errors that transform content words from SRC_1, such as "learning"⇒"loaning" and "anatomy"⇒ "and that to me" corrupt the meaning of the sentence, making an accurate translation impossible. Many speech recognition errors result in sequences of

words that sound similar to the expected ASR reference but change the meaning of the utterance. While the MT system (MT_1) attempts to correct some of the errors in the beginning of the sentence, it duplicates "dissection" in the process and drops "cadaver". MT_1 attempts to translate the remainder of the ASR errors literally, but makes additional translation errors in the process, such as in "to me"⇒"de moi" (myself).

The second sentence has an ASR hypothesis (ASR_2) containing minimal function word errors that do not have much impact on translation quality. MT_2 makes a minor error caused by a literal translation for "quite an experience", but the MT and SLT translations are still understandable. The third sentence is corrupted again by ASR, where the context of the previous sentences could have helped the ASR decoder to recognize the words "anatomic" and "cadaver". Although "anatomy" was already misrecognized in ASR_1, the acoustic information around "cadaver" was discarded in the current recognition context and a phonetically similar sequence of words was hypothesized instead ("kind of a"). As a result, ASR_3 is unintelligible and the SLT result bears no resemblance to the translation MT_3 in a perfect recognition scenario.

While adaptation to ASR models is outside the scope of this thesis, it is important to highlight scenarios where the ASR system should have been able to find the correct recognition result. It is in these cases where we look for opportunities for the MT system to compensate. A machine translation system might have been able to compensate for errors such as "cadaver"⇒"kind of a", if it possessed a mechanism to *model the previous decisions* of the ASR and MT decoders, and to *model noisy ASR outputs during system training*. In order to model the errors made by an ASR system to make a strong SLT system, we must *analyze the impact of ASR errors on translation quality*.

1.2 Contributions

The main research contributions of this thesis are as follows:

- we compare and contrast the difficulties of machine translation for textual domains versus speech domains;
- we provide a statistical data analysis framework to analyze the impact of speech recognition errors on machine translation quality;
- we introduce a novel technique to inexpensively adapt machine translation models based on small discourse context windows;

- we introduce an *ASR damaging channel* that provides synthetic noisy ASR training data for machine translation;

- we evaluate the effects of speech recognition errors on the recent state-of-the-art neural machine translation systems against conventional log-linear phrase-based machine translation approaches and suggest further areas of research for neural spoken language translation.

1.3 Translation versus Interpretation

Whereas translation involves the transfer of meaning from a source language to a target language with the use of materials such as phrase books and dictionaries, interpretation involves an interaction between monolingual speakers and a bilingual mediator who simultaneously processes speech and contextualizes it to make it understandable in the recipient's target language. Speech interpreters do not have the time to consult phrase books or dictionaries to aid in translation; instead they are constrained by limited time and memory to convey the most important information in speech – as such, adequacy is more important than fluency or literal accuracy. Automatic spoken language translation does not have the same limitations as an interpreter. A SLT system learns representations and transfers them from one language to another and can quickly look up the necessary information to generate translations.

While language translation in interpretation scenarios is an interesting research problem, the scope of this thesis is to explore the problem of spoken language *translation*.

1.4 Structure of this Thesis

The remainder of this thesis is structured as follows. Chapter 2 introduces the fundamental concepts of ASR and MT and illustrates their composition to form spoken language translation. We outline the de-facto evaluation measures for each component and outline promising directions of research in the field of SLT.

Chapter 3 analyzes the differences between spoken and written registers and outlines how these differences provide unique challenges in machine translation research. While focusing on the difficulties of translating spoken language content, we reserve the discussion of the impact of ASR errors on the problem of spoken language translation for

Chapter 4. In Chapter 4, we use *linear mixed-effects models* to measure how the ASR error types in popular evaluation metrics exacerbate the English-French translation problem and contribute to the degradation of translation quality.

Chapter 5 introduces a novel technique to adjust the translation and language model scores in a statistical machine translation system based on topic adaptation using bilingual latent semantic models. As a log-linear approximation to conventional adaptation techniques, it can efficiently be used in the SLT scenario. We demonstrate its ability to improve translation results using a monolingual or bilingual context window of only a few sentences in machine translation.

Chapter 6 introduces a noisy-channel model designed to transform textual data into synthetic ASR outputs in order to compensate for the lack of translated ASR data. Evaluated on a number of English-X language pairs (e.g. French, Spanish, Italian, Mandarin), the method is shown to improve translation quality without introducing additional training corpora to the model, and without affecting the latency of decoding.

Chapter 7 introduces and motivates interesting problems one faces when considering the translation of ASR outputs on neural machine translation (NMT) systems for English-French. We compare the robustness of NMT's encoder-decoder modeling against a state-of-the-art PBMT system when translating noisy speech input. Chapter 8 concludes this thesis with a summary of the major findings of the thesis and suggests future research directions.

1.5 Evaluation Data

The majority of experiments in this thesis involve the translation of TED Talks:[1] a popular SLT data source that has been used for over 6 years in the International Workshop of Spoken Language Translation. TED talks are a collection of short speeches covering a variety of topics. A large crowd-sourcing community of annotators actively captions and translates TED talks, thus far providing a rich corpus covering over 80 languages (Cettolo et al., 2012). In this thesis, we focus primarily on the translation of TED talks from English to French or English to German.

1.6 Relevant Publications

Parts of Chapter 3 were published in:

[1] http://www.ted.com/talks

- Nicholas Ruiz and Marcello Federico. “Complexity of Spoken Versus Written Language for Machine Translation” appeared in the proceedings of the 17th Conference of the European Association for Machine Translation (Ruiz and Federico, 2014a).

Chapter 4 combines the interpretation of ASR errors on SLT quality from the following papers:

- Nicholas Ruiz and Marcello Federico. “Assessing the Impact of Speech Recognition Errors on Machine Translation Quality” appeared in the proceedings of the International Workshop on Spoken Language Translation (Ruiz and Federico, 2014b).
- Nicholas Ruiz and Marcello Federico. “Phonetically-Oriented Word Error Alignment for Speech Recognition Error Analysis in Speech Translation” appeared in the proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (Ruiz and Federico, 2015).

Parts of Chapter 5 were published in the following papers:

- Nicholas Ruiz and Marcello Federico. “Topic Adaptation for Lecture Translation through Bilingual Latent Semantic Models” appeared in the proceedings of the Sixth Workshop on Statistical Machine Translation (Ruiz and Federico, 2011).
- Nicholas Ruiz and Marcello Federico. “MDI Adaptation for the Lazy: Avoiding Normalization in LM Adaptation for Lecture Translation” appeared in the proceedings of the International Workshop on Spoken Language Translation (Ruiz and Federico, 2012).

Parts of Chapter 6 were published in:

- Nicholas Ruiz, Qin Gao, William Lewis, and Marcello Federico. “Adapting Machine Translation Models toward Misrecognized Speech with Text-to-Speech Pronunciation Rules and Acoustic Confusability” appeared in the proceedings of Interspeech (best student paper award) (Ruiz et al., 2015).

The machine translation systems described in the various chapters were submitted in various IWSLT evaluation campaigns, with the description papers listed below:

- N. Ruiz, A. Bisazza, F. Brugnara, D. Falavigna, D. Giuliani, Diego, S.Jaber, Suhel, R. Gretter, and M. Federico. “FBK @ IWSLT 2011” appeared in the proceedings of the International Workshop on Spoken Language Translation (Ruiz et al., 2011).

- N. Ruiz, A. Bisazza, R. Cattoni, and M. Federico. “FBK’s Machine Translation Systems for IWSLT 2012’s TED Lectures” appeared in the proceedings of the International Workshop on Spoken Language Translation (Ruiz et al., 2012).
- N. Bertoldi, A. Farajian, P. Mathur, N. Ruiz, and M. Federico. “FBK’s Machine Translation Systems for the IWSLT 2013 Evaluation Campaign” appeared in the proceedings of the International Workshop on Spoken Language Translation (Bertoldi et al., 2013b).
- A. Aue and Q. Gao and H. Hassan and X. He and G. Li and N. Ruiz and F. Seide. “MSR-FBK IWSLT 2013 SLT System Description” appeared in the proceedings of the International Workshop on Spoken Language Translation (Aue et al., 2013).
- N. Bertoldi, P. Mathur, N. Ruiz, and M. Federico. “FBKs Machine Translation and Speech Translation Systems for the IWSLT 2014 Evaluation Campaign” appeared in the proceedings of the International Workshop on Spoken Language Translation (Bertoldi et al., 2014).

CHAPTER 2

SPOKEN LANGUAGE TRANSLATION MODELING

Spoken language translation (SLT) consists of multiple components: an Automatic Speech Recognition (ASR) system, which processes an audio signal and transcribes it into a sequence of corresponding words, and a Machine Translation (MT) system, which takes the words of the speaker and translates them into a sequence of words in a target language. In this chapter, we introduce and summarize each of the components of a baseline SLT system and describe the main evaluation metrics for each component. Then, we compare two general approaches to combining ASR and MT: either as a sequential pipeline, or as a tightly coupled SLT system.

Both ASR and Statistical Machine Translation (SMT) are based on Shannon (1948)'s *noisy-channel model*, which assumes that the output received by a human is a corruption of a source input that was passed through a noisy channel. In ASR this formulation is easily understood by observing that during the language production process, a speaker's planned utterance may be distorted by factors including external noise and variations in the speaker's articulation. In SMT, the model is applied under Warren Weaver's famous assumption that translation is a decoding process (Weaver, 1949/1955). The noisy-channel descriptions in this chapter refer to the translation of spoken words from a foreign language into English. Note that the same construction applies without loss of generality for SLT tasks from any source language to any target language.

This chapter is organized as follows: in Section 2.1 we introduce the formulation for automatic speech recognition and outline its evaluation metrics. In Section 2.2 we outline the log-linear approach to statistical machine translation and briefly introduce is-

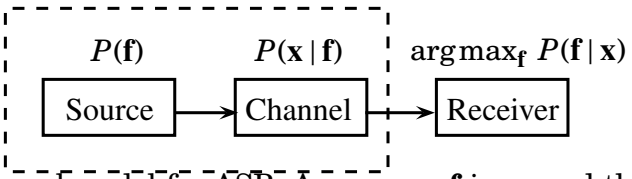


Figure 2.1: Noisy-channel model for ASR. A message $\mathbf{f}$ is passed through a noisy channel, which causes a "corruption" of the message to an audio signal $\mathbf{x}$. The message is reconstructed via a source model $P(\mathbf{f})$ and a channel model $P(\mathbf{x} \mid \mathbf{f})$.

sues in translation modeling and decoding, as well as introduce the log-linear modeling paradigm. In Section 2.3 we introduce the task of language modeling, whose principles are shared in common with ASR and MT modeling. In Section 2.4 we introduce Neural Machine Translation, an alternative to the log-linear SMT approach described in Section 2.2. In Section 2.5 we discuss the evaluation metrics for MT. Finally, in Section 2.6 we discuss the approaches to combine ASR and MT to form SLT and provide a summary of the chapter discussion in Section 2.8.

2.1 Automatic Speech Recognition

In the noisy channel formulation, ASR decoding is a generative process in which words are recognized from an audio signal. Figure 2.1 outlines the process, which assumes that a sequence of words $\mathbf{f}$ were received by the hearer as a sequence of observations $\mathbf{x}$ from an audio signal.

The statistical approach to speech recognition is based on the decision rule:

$$\mathbf{f}^* = \arg\max_{\mathbf{f}} P(\mathbf{x} \mid \mathbf{f}) \cdot P(\mathbf{f}). \tag{2.1}$$

The decoding is performed over all possible source word hypotheses for a given spoken utterance $\mathbf{f}$. $P(\mathbf{x} \mid \mathbf{f})$ corresponds to an intersection between an *acoustic model* (AM) that estimates the conditional probability for a (context-dependent) sequence of phonemes given acoustic observations $\mathbf{x}$ and a *pronunciation dictionary* (PD) that defines the words that can be legally formed from the sequence of phonemes. $P(\mathbf{f})$ is a *language model* (LM) that computes the prior probability over the sequence of words in each recognition hypothesis.

As a brief overview, automatic speech recognition occurs as follows. A *signal processing* component converts an input audio signal into a sequence of salient feature vectors $\mathbf{x}$ that can be processed by the acoustic model. In order to extract accurate features, the component first converts an input audio signal from the time domain to the

frequency-domain. The input signal is binned into overlapping frames, or Hamming windows, of about 25 milliseconds and a window shift of about 10 ms. Feature coefficients are extracted from within each window and also by measuring the change in coefficients (delta coefficients) across adjacent Hamming windows. Two discriminative feature sets with higher accuracy are Mel Frequency Cepstral Coefficients (Davis and Mermelstein, 1980) and Perceptual Linear Prediction components (Hermansky, 1990). The feature vectors are transformed to account for signal noise and other channel distortions through techniques such as Linear Discriminant Analysis (Haeb-Umbach and Ney, 1992), vocal tract length normalization (Acero, 1991; Wooters and Stolcke, 1994; Lee and Rose, 1996; Giuliani et al., 2006), and neural network-based speaker mapping (Huang, 1992); as well as speaker adaptation techniques, such as maximum likelihood linear regression (Leggetter and Woodland, 1995). In order to account for the temporal variability of speech, most acoustic models utilize hidden Markov models (HMMs) (Rabiner and Juang, 1986; Huang et al., 1990). While in the past Gaussian mixture models (GMMs) were used to determine how well each HMM state fits a window of speech frames, discriminative hierarchical models such as deep neural networks (DNNs) (Dahl et al., 2012; Hinton et al., 2012) have demonstrated a marked increase in acoustic modeling accuracy (Seide et al., 2011).

2.1.1 Evaluation

Word error rate (WER) is one of the most important automatic evaluation measures for ASR accuracy. The WER for an ASR hypothesis, compared against a human reference transcript is computed as:

$$WER = \frac{S + D + I}{L}, \tag{2.2}$$

where S, D, and I are the number of word substitutions, deletions, and insertions in the Levenshtein alignment (Levenshtein, 1966) between the hypothesis and its reference transcript, and L is the ASR reference length (in words). A *substitution* occurs when a word is incorrectly substituted for the correct word; a *deletion* is the omission of a correct word in the hypothesis; and an *insertion* occurs when a word is incorrectly introduced in the hypothesis.

2.2 Statistical Machine Translation

The objective in statistical machine translation (SMT) is to find the most probable target sentence that translates a given source sentence. SMT is most often formulated as a generative model that maximizes the posterior probability $P(\mathbf{e} \mid \mathbf{f})$ of $\mathbf{e} = e_1, e_2, ..., e_{l_\mathbf{e}}$ given the observed foreign message $\mathbf{f} = f_1, f_2, ..., f_{l_\mathbf{f}}$. Using Bayes' decision rule, this is expressed as:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} P(\mathbf{e} \mid \mathbf{f}) = \arg\max_{\mathbf{e}} P(\mathbf{f} \mid \mathbf{e}) P(\mathbf{e}). \tag{2.3}$$

$P(\mathbf{f} \mid \mathbf{e})$ is referred to as the *translation model* (TM), which models the conditional probability of English words (or phrases) given foreign words. The translation model is constructed using *bitexts* which are aligned at the sentence or clause level. Similar to the ASR decision rule, $P(\mathbf{e})$ is represented as a *language model*, which measures the fluency of a translated text by computing the probability of a sequence of target language words, independent of the input sentence $\mathbf{f}$.

The decomposition into exactly two components in (2.3) was directly used in early approaches to SMT as word-based MT (Brown et al., 1990, 1993), in which $P(\mathbf{f} \mid \mathbf{e})$ is cast as a *lexical translation model* that translates words in isolation. More recently, SMT models were enriched with contextual information: Phrase-Based Statistical Machine Translation (PBMT) (Zens et al., 2002; Och, 2002; Koehn et al., 2003) models the translation process in terms of phrases, i.e. sequences of contiguous words; Hierarchical Machine Translation (Yamada and Knight, 2001; Chiang, 2005) and Syntax-based Machine Translation, based on formulations such as synchronous context-free grammars (Chiang, 2005; Wu, 1997), allow discontiguous translation decisions.

In the last few years, the Deep Neural Network paradigm was applied with large success to SMT, creating a branch of SMT research known as Neural Machine Translation (NMT) (Sutskever et al., 2014; Cho et al., 2014b; Bahdanau et al., 2015). Section 2.4 summarizes its main concepts.

In Sections 2.2.1 and 2.2.2 we briefly introduce the main concepts of the lexical and phrase-based translation models, respectively. Section 2.2.3 illustrates the combination of translation model, language model, and reordering model as weighted features in a generalized log-linear model that allows additional features to be incorporated. Section 2.2.4 summarizes the decoding process that searches for the best translation of a foreign sentence. An overview of Hierarchical and Syntax-based machine translation are not provided, as they are not used in this thesis.

2.2.1 Lexical Translation Models

Originating from Brown et al. (1993), a bilingual dictionary (or translation table) is constructed by deriving a probability distribution over alignments between words in the source and target languages. If our bitexts contained word alignments, we could simply use *maximum likelihood estimation* to calculate the probability distribution of the data; however, in normal translation scenarios, we do not have this information. Thus, to build the lexical translation model, we learn an *alignment model* $P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$, where the marginalized variable $\mathbf{a}$ is obtained from an alignment function ($a : j \rightarrow i$) that links a source input word at position j to a target output word at position i.

The alignment function allows one-to-many alignments from source to target. Although unknown at the time of translation, the alignment of words is an important clue in defining the best translation. The lexical translation model is expressed as:

$$P(\mathbf{f} \mid \mathbf{e}) = \sum_{a} P(\mathbf{f}, a \mid \mathbf{e}), \tag{2.4}$$

which marginalizes over all possible word alignments between source and target sentences. Such a model favors translations that better follow the alignment rules of the translation pair.

The model can be further expressed as:

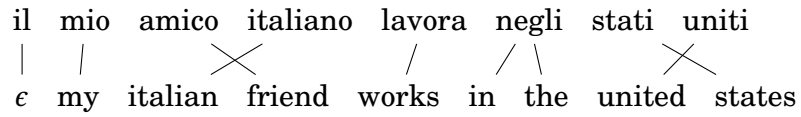
$$P(\mathbf{f}, a \mid \mathbf{e}) = Z \prod_{j=1}^{l_f} q(a(j) \mid j, \mathbf{f}, \mathbf{e})\ t(f_j \mid e_{a(j)}) \tag{2.5}$$

i.e. in a composition of the actual lexical translation probabilities $t(f \mid e)$, and of the alignment probabilities $q(a \mid j, \mathbf{f}, \mathbf{e})$ of aligning the source word at position j with the target word at position a. Z is a normalization factor.

In the IBM Candide project, Brown et al. (1990, 1993) derive several alignment models with increasing dependency assumptions that are used in sequence to learn lexical translation and alignment probabilities. The models are trained using *Expectation Maximization* (Dempster et al., 1977).

IBM Model 1 assumes that each lexical translation decision is independent from one another, thus the alignment distribution is uniform and depends only on the source and target length. As such, the translation probability for a foreign sentence $\mathbf{f}$ of length l_f to an English sentence $\mathbf{e}$ of length l_e is defined as:

$$P(\mathbf{f}, a \mid \mathbf{e}) = Z\ \frac{1}{(l_e + 1)^{l_f}} \prod_{j=1}^{l_f} t(f_j \mid e_{a(j)}), \tag{2.6}$$



$$a : \{1 \rightarrow 0, 2 \rightarrow 1, 3 \rightarrow 3, 4 \rightarrow 2, 5 \rightarrow 4, 6 \rightarrow 5, 6 \rightarrow 6, 7 \rightarrow 7, 8 \rightarrow 8\}$$

Figure 2.2: Example Italian to English word alignment in IBM Model 1. A null token ϵ is introduced to capture source words that are dropped during lexical translation.

where Z is a normalization factor.

IBM Model 2 extends Model 1 by incorporating local alignment through an enhanced alignment probability distribution: $q(a(j) \mid j, \mathbf{f}, \mathbf{e}) = q(i \mid j, l_e, l_f)$, which models the likelihood that an arbitrary foreign sentence of length l_f aligns position j with position i in any English translation of length l_e, without considering the actual words. Model 2 is defined as:

$$P(\mathbf{f}, a \mid \mathbf{e}) = Z \prod_{j=1}^{l_f} t(f_j \mid e_{a(j)}) q(a(j) \mid j, l_e, l_f). \tag{2.7}$$

During training, each subsequent IBM model (IBM Models 3-5) initializes the lexical translation and alignment probabilities $t(f_j \mid f_{e(j)})$ and $q(a(j) \mid j, \mathbf{f}, \mathbf{e})$ to the results of the previous model after convergence has been reached.

Figure 2.2 provides an example illustrating several alignment types from source to target for IBM Model 1 – in this case, for Italian to English. During alignment a null token ϵ is introduced into the target sentence at at position 0 to cover dropped source words (e.g. "il" $\Rightarrow \epsilon$). Source words can have multiple corresponding target words ("negli" $\Rightarrow$ "in the") and alignments can be discontiguous ("mio amico" $\Rightarrow$ "my ... friend").

IBM Models 3-5 reduce the deficiencies of earlier models by introducing *fertility* (word duplication to enforce one-to-one alignments), null (ϵ) word insertion, and word deletions, as well as increasing context-dependency and constraining against overlapping alignment positions. Alternative alignment models include an HMM-based word alignment that models relative alignment positions (Vogel et al., 1996) and a "fast aligner" that re-parameterizes IBM Model 2 in a manner that prefers monotonic alignments, particularly working well for languages with the same word order (Dyer et al., 2013).

2.2.2 Phrase-Based Models

A major disadvantage of word-based approaches to machine translation is that the context of each word is not taken into account. PBMT (Zens et al., 2002; Och, 2002; Koehn et al., 2003) incorporates additional context by enabling chunks of words to be translated at a time. These chunks do not have to correspond to phrasal constituents. These chunks are translated independently from one another and are scored by the language model.

PBMT assumes that foreign and English sentences are decomposed into exactly K phrases governed by the alignment variable b_1^K, defined as:

$$b_1^K = ((J_1, I_1), (J_2, I_2), ..., (J_K, I_K)), \tag{2.8}$$

where $I_1, ..., I_K$ are the contiguous intervals partitioning $\mathbf{e}$, and $J_1, ..., J_K$ are the corresponding alignment positions for the target word positions in each chunk. Note that the J intervals cover a contiguous span of source words, but their positions may be permuted due to alignments.

To a certain extent the phrase alignment b_1^K replaces the word alignment $a_1^{l_f}$. Thus, similarly to the lexical translation model in (2.5), the phrase translation model is decomposed into a *phrase translation model* ϕ and a phrase *distortion* (or *reordering*) *model* d as:

$$P(\mathbf{f}, b_1^K \mid \mathbf{e}) = d(b_1^K)\ \phi(\mathbf{f} \mid \mathbf{e}, b_1^K) = \prod_{k=1}^{K} d(J_{k-1}, J_k)\ \phi(\tilde{f}_k \mid \tilde{e}_k), \tag{2.9}$$

where $\tilde{f}_k$ are the source words covered by alignment partition J_k and $\tilde{e}_k$ are the target words covered by I_k. The standard distortion model $d(\cdot)$ in (2.9) assigns an exponentially decaying cost function for the number of words skipped in either direction from the position of the previous phrase (Och and Ney, 2004). Alternate and more complex reordering models include lexicalized (or hierarchical) phrase orientation (Tillmann, 2004; Koehn and Monz, 2005; Galley and Manning, 2008), pairwise lexicalized distortion (Al-Onaizan and Papineni, 2006), and reordered source n-grams (Feng et al., 2010).

The bidirectional Viterbi alignments (source-to-target and target-to-source) estimated by the lexical translation model are first symmetrized using heuristics (Och et al., 1999; Koehn et al., 2003), and then used to extract phrase translation candidates that allow one-to-one, one-to-many, and many-to-many alignments. From the symmetrized alignment a phrase pair $(\tilde{f}, \tilde{e})$ is valid and hence it is extracted, if and only if all words of $\tilde{f}$ are aligned with any word of $\tilde{e}$ or to the null token ϵ, and vice-versa.

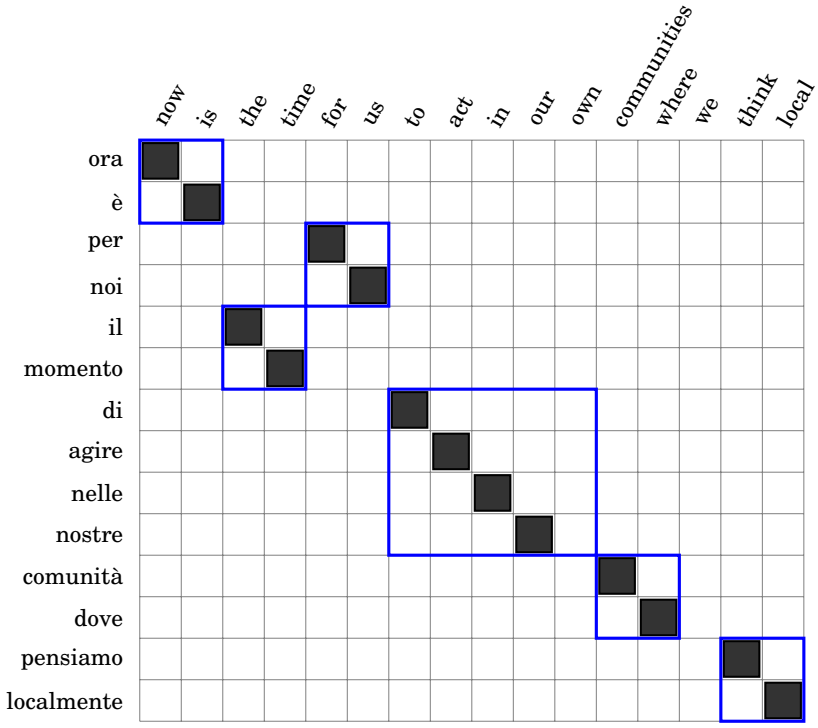


Figure 2.3: An example of a word alignment and of phrase pairs extracted from a training sentence pair. Blue borders highlight some of the phrase pairs that may be extracted.

An example of an alignment between two sentences and a subset of possible phrase pairs is shown in Figure 2.3. Note that all valid phase pairs are extracted even if they overlap with other previously extracted pairs; for instance, “per noi il momento”$\Longleftrightarrow$“the time for us” is also a valid phrase pair, although it overlaps with “il momento”$\Longleftrightarrow$"the time”.

Translation table probabilities are estimated via maximum likelihood estimation, given the counts $n(\cdot)$ of the extracted phrase pairs $(\tilde{f},\tilde{e})$ within the training corpora:

$$\phi(\tilde{f} \mid \tilde{e}) = \frac{n(\tilde{f},\tilde{e})}{\sum_{\tilde{f}_i} n(\tilde{f}_i,\tilde{e})}. \tag{2.10}$$

2.2.3 Log-Linear Translation Model

Given a translation model $P(\mathbf{f} \mid \mathbf{e})$ consisting of a phrase translation table $\phi(\bar{f} \mid \bar{e})$ and a reordering model $d(\cdot)$, the phrase-based generative model is factorized as:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \prod_{k=1}^{K} \phi(\tilde{f}_k \mid \tilde{e}_k) d(b_k | b_1 \ldots b_{k-1}) \prod_{i=1}^{|\mathbf{e}|} p_{\mathrm{LM}}(e_i \mid e_1 \ldots e_{i-1}) \tag{2.11}$$

when combining the translation model with the language model. Although (2.11) assumes that each component has equal weight, Och (2002) introduces a log-linear modeling framework for SMT, which allows for the integration of additional components as weighted feature functions. The subsequent decision rule becomes:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e},\mathbf{f}), \tag{2.12}$$

where $h_m(\mathbf{e},\mathbf{f})$ are the M feature functions of the model and λ_m are their associated weights. The feature weights may be optimized (or tuned) using discriminative training approaches such as *minimum error rate training* (MERT) for dense feature functions (Och, 2003) or the Margin Infused Relaxed Algorithm (MIRA) for sparse feature functions (Crammer et al., 2006; Watanabe et al., 2007; Chiang et al., 2008; Cherry and Foster, 2012) against a global evaluation criterion.

State-of-the-art PBMT systems such as Moses (Koehn et al., 2007) and cdec (Dyer et al., 2010) contain the following features:

- A statistical language model $p_{LM}(\mathbf{e})$,
- Direct and inverse phrase translation models $\phi(\mathbf{f} \mid \mathbf{e},\mathbf{b})$ and $\phi(\mathbf{e} \mid \mathbf{f},\mathbf{b})$,
- Direct and inverse lexical translation models $lex(\mathbf{f} \mid \mathbf{e},\mathbf{b})$ and $lex(\mathbf{e} \mid \mathbf{f},\mathbf{b})$,
- A reordering model $d(\mathbf{b})$,
- A phrase insertion penalty I, penalizing translations with little context,
- A word insertion penalty L, penalizing translations with many words.

In addition to these features, a number of additional features have been empirically shown to increase machine translation performance on a number of translation tasks, including:

- Operation sequence modeling (Durrani et al., 2011);
- Discriminative Word Lexicon (Niehues and Waibel, 2013);
- Bilingual word embeddings (Zou et al., 2013).

2.2.4 Decoding

The search for the best translation, called *decoding*, consists of finding the optimal sequence of words $\mathbf{e}^*$ that translates a foreign sentence $\mathbf{f}$. In the case of PBMT, the decoding process comprises three general operations: (i) partitioning the input into phrases $J_1,...,J_K$, (ii) deciding the permutation of J_1^K, and (iii) deciding the translation $\tilde{e}_k$ for source phrase $\tilde{f}_k$. The process of decoding in SMT is defined as:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} P(\mathbf{e}) \sum_{a} P(\mathbf{f}, a \mid \mathbf{e}). \tag{2.13}$$

The *Viterbi approximation* is generally used, which approximates (2.13) as:

$$\mathbf{e}^* \approx \arg\max_{\mathbf{e}} P(\mathbf{e}) \max_{a} P(\mathbf{f}, a \mid \mathbf{e}), \tag{2.14}$$

which allows the use of dynamic programming algorithms, such as DP beam-search (Tillmann and Ney, 2003) and A* search (Och et al., 2001) for PBMT and chart parsing (Zollmann and Venugopal, 2006) for hierarchical MT.

The DP beam-search algorithm incrementally constructs hypotheses that consist of partial translations of the input sentence. Beginning with an empty hypothesis, the hypothesis is expanded by selecting each translation option that generates the initial phrase in the English sentence. Expanded hypotheses are placed in a stack that corresponds to the number of English words covered by the hypothesis (i.e. if hypothesis h contains i translated words, it is placed in the ith stack). Each hypothesis has an associated cost, determined by its current cost and its future cost. The current cost for a partial hypothesis is determined from the probability of the phrases already in the hypothesis, which, in the case of a log-linear model, follows (2.12). A low probability corresponds to a high cost. The future cost is the expected minimum cost of translating the rest of the sentence.

Pruning techniques are used to limit the number of hypotheses per stack. In histogram pruning, a maximum number of n hypotheses with the lowest cost are preserved in each stack. In threshold pruning, hypotheses with scores that are worse than the best hypothesis in its corresponding stack by a specific threshold α are pruned.

2.3 Language Modeling

Language modeling is an important component for scoring the fluency of a system's decoded output. In the case of ASR, it is composed with the ASR pronunciation dictionary

and acoustic model to rank likely recognition hypotheses. It is also a component in the log-linear model for SMT approaches as described in Section 2.2.3, which guides the decoding process by suggesting word ordering and lexical choices in translation. Given a sequence of l words $\mathbf{w} = w_1, w_2, ..., w_l$, the language model computes the joint probability of every word w_i in the sequence. The chain rule factorizes the probability of the sequence as the conditional probability of each word, given a history of preceding words in the sequence:

$$P(\mathbf{w}) = P(w_1) \cdot \prod_{i=2}^{l} P(w_i \mid w_1^{i-1}) \approx \prod_{i=1}^{l} P(w_i \mid w_{i-n+1}^{i-1}). \tag{2.15}$$

A *Markov assumption* is made in (2.15) for n-gram language modeling which assumes that the history is limited to a window of $n-1$ preceding words. In their simplest form, n-gram language models can be computed according to maximum likelihood estimation.

While ASR operates under the *closed vocabulary* assumption, where words that do not exist in the pronunciation dictionary cannot be predicted, MT must be able to assign translation scores for out-of-vocabulary (OOV) words. In order to account for OOVs, *discounting* techniques have been proposed to assign some of the probability mass of more frequently observed words to out-of-vocabulary words. These techniques include Good-Turing (Good, 1953), Witten-Bell smoothing (Witten and Bell, 1991), and Kneser-Ney smoothing (Kneser and Ney, 1991, 1993, 1995). In addition, the statistics for higher-order n-gram models are sparse, thus *interpolation* and *back-off* techniques are used in conjunction with discounting to smooth higher-order n-gram counts by lower-order observations.

An interpolated language model (Jelinek and Mercer, 1980) is a linear combination of n-gram language models of varying size, recursively defined as:

$$P_n^I(w_i \mid h_{i,n}) = \lambda_{h_{i,n}} P_n^I(w_i \mid h_{i,n}) + (1 - \lambda_{h_{i,n}}) P_{n-1}^I(w_i \mid h_{i,n-1}), \tag{2.16}$$

where $h_{i,n} = w_{i-n+1}, ..., w_{i-1}$ is the n-gram history of word w_i, with associated interpolation weights $\lambda_{h_{i,n}}$ that are optimized with Expectation Maximization.

Back-off modeling (Katz, 1987) alternatively relies on lower-order counts only if the particular n-gram's history is not observed in training. Otherwise, a discounted n-gram probability $P^*(w_i \mid h_{i,n})$ is used. The back-off model is defined as a system of equations:

$$P_n^{\mathrm{BO}}(w_i \mid h_{i,n}) = \begin{cases} P^*(w_i \mid h_{i,n}) & \text{if } \mathrm{count}_n(h_{i,n}) > 0, \\ z(h_{i,n})^{-1} P_{n-1}^{\mathrm{BO}}(w_i \mid h_{i,n-1}) & \text{otherwise} \end{cases} \tag{2.17}$$

where $z(h_{i,n})$ normalizes the back-off probability.

In addition to n-gram language modeling, neural network approaches to language modeling, such as recurrent neural network models (Mikolov et al., 2010), have shown to be promising. We do not describe them in detail because they are not used by the MT systems in our experiments.

2.3.1 Evaluation

The most common metric for evaluating the quality of language models is *perplexity*, which measures how well a language model predicts a sequence of words in a test set. Perplexity refers to the average number of equally probable words the language model must choose from when predicting the next word in a sequence. Thus, a lower perplexity implies that the language model assigns higher probabilities to the test set. The perplexity (PP) measure is based on the principle of cross-entropy, which is defined as:

$$H(P_{\text{LM}}) = -\frac{1}{l}\sum_{i=1}^{l}\log P_{\text{LM}}(w_i \mid w_{i-n+1}^{i-1}) \tag{2.18}$$

for a sequence of length l, and a n-gram LM. The perplexity is simply the exponential of the cross-entropy:

$$PP = 2^{H(P_{\text{LM}})}. \tag{2.19}$$

2.4 Neural Machine Translation

As an alternate representation of SMT, the objective of Neural Machine Translation (NMT) is to find a target sentence $\mathbf{e}$ that maximizes the conditional probability of $\mathbf{e}$ given a source sentence $\mathbf{f}$. While traditional statistical MT models require a collection of complex features that are combined together in a log-linear framework, NMT fits a parameterized model to maximize the conditional probability of a collection of bitexts. NMT simplifies the modeling paradigm of machine translation by casting the entire problem into a *sequence-to-sequence* model that does not require any feature engineering. Let $\mathbf{f}$ and $\mathbf{e}$ be the source and target sentences. NMT directly addresses the conditional probability defined as:

$$P(\mathbf{e} \mid \mathbf{f}) = \prod_{i=1}^{l_e} P(e_i \mid \mathbf{e}_{<i}, \mathbf{f}), \tag{2.20}$$

In particular, NMT acts like a language model that can be used to incrementally predict each word of the translation $\mathbf{e}$. This predictive model is actually implemented through

the combination of two recurrent neural networks (RNNs), called encoder and decoder models, and by a feed-forward neural network called an *attention model*.

2.4.1 Encoder, Decoder and Attention Models

The state of the art models in NMT (Sutskever et al., 2014; Cho et al., 2014b; Bahdanau et al., 2015) consists of (i) an *encoder* RNN that reads and encodes the source sentence **f** word by word into a sequence of hidden states **h**; (ii) a *decoder* RNN that generates word by word the target sentence **e**; and (iii) an *attention model* that at each step provides the decoder with a context vector computed over the encoder's hidden state sequence **h**. Source and target words are assumed to be represented with one-hot vectors, i.e. unit vectors with one single component, corresponding to the represented word, set to one and all the others set to zero.

More formally, the decoder and attention models can be described with the following equations:

$$P(e_i \mid \mathbf{e}_{<i}, \mathbf{f}) = g(e_i, e_{i-1}, s_i, c_i), \tag{2.21}$$

where the output distribution g is computed on the decoder hidden state s_i

$$s_i = f(s_{i-1}, e_{i-1}, c_i), \tag{2.22}$$

and the context vector c_i is a convex combination over the encoder's hidden states:

$$c_i = \sum_{j=1}^{l_f} \alpha_{i,j}\, h_j. \tag{2.23}$$

The weights to compute the context vector are provided by the attention model:

$$\alpha_{i,j} \propto t(s_i, h_j) \tag{2.24}$$

which is implemented with a simple feed-forward network coupled with a soft-max layer. Regarding the encoder, the sequence of hidden state vectors h_j is a combination of bidirectional hidden state sequences, i.e.:

$$h_j = \left[\overrightarrow{h}_j, \overleftarrow{h}_j\right] \quad j = 1, \ldots, l_f \tag{2.25}$$

where each directional sequence is generated by two distinct RNNs:

$$\begin{aligned} \overrightarrow{h}_j &= \overrightarrow{g}(f_j, \overrightarrow{h}_{j-1}) \\ \overleftarrow{h}_j &= \overleftarrow{g}(f_j, \overleftarrow{h}_{j+1}) \end{aligned} \tag{2.26}$$

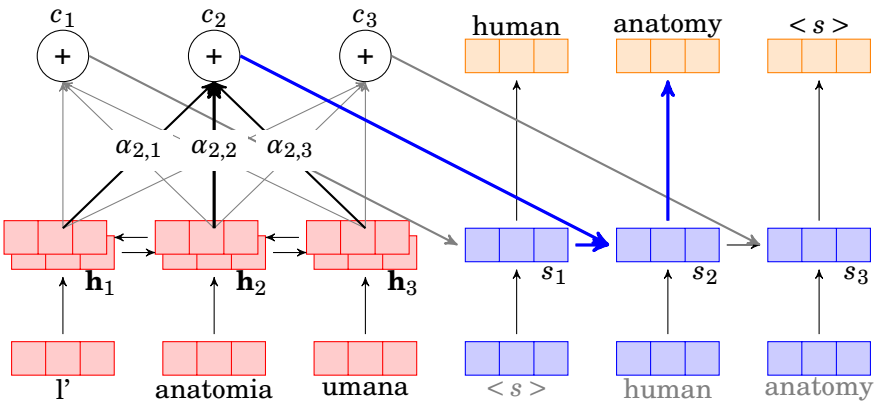


Figure 2.4: Translation of "l'anatomia umana" from Italian-English using Bahdanau et al. (2015)'s Encoder-Decoder RNN with an attention mechanism. The output of y_2 depends on the previous RNN hidden state s_1 and context vector c_2 that is computed from the attention weights $\alpha_{2,j}$ for each input word position j.

It is worth mentioning that the recurrent networks of both the encoder and decoder start with a word embedding layer, that maps the one-hot vectors into smaller and dense vectors, which are passed on to one recurrent layer implemented with gated recurrent units (GRUs) (Cho et al., 2014b) that generate the internal state vectors. The decoder is enriched with an output layer and a softmax layer that computes a distribution over the target vocabulary.

Figure 2.4 provides a graphical example with the translation of "l'anatomia umana" in Italian to "human anatomy" in English using Bahdanau et al. (2015)'s RNN encoder-decoder-attention model. The first word is generated by initializing the output sequence with the start symbol <s>, which is mapped to an embedding and further encoded into a recursive hidden state s_1 that depends on c_1, a linear combination of the bidirectional hidden states of the encoder. The English translation's second word, "anatomy" is generated from an RNN decoder that depends on the context vector c_2 and the previous decoding hidden state s_1. This process is carried out until the end of sentence symbol <s> is emitted.

2.4.2 Beam Search

The presented NMT architecture can generate a target sentence by simply sampling the most probable word from the output distribution of the decoder until the delimiter string is reached. Better translation quality can be reached by replacing this greedy local strategy with a beam-search method. At each step i, the most promising k transla-

tions of the previous step $i-1$ are used as input alternatives. Then, the corresponding output distributions are computed and the overall top k translations are to be used during the next step $i+1$. This simple beam-search procedure, which introduces some extra book-keeping to the search procedure, has been shown to significantly reduce search errors even with small values of k, between 10 and 20.

2.4.3 Training

Source and target vocabularies of about 50K words defines a NMT architecture with hundreds of thousands of parameters. These include weights and biases of all the layers of its three component networks, the encoder, the decoder and the attention models. All parameters can be jointly trained on a parallel corpus with maximum likelihood estimation or other criteria. In particular, stochastic gradient descent (SGD) (Goodfellow et al., 2016) can be applied on random mini-batches of the training data. This from one side gives faster convergence and, from the other side, permits to perform the needed calculations with high-parallelism on a graphical processing unit (GPU). SGD for NMT is implemented with the *back-propagation through time* algorithm (Goodfellow et al., 2016) which usually requires several training epochs, i.e. iterations over the entire training set through the mini-batches. Training is stopped when optimal performance is reached on a cross-validation set.

2.5 Machine Translation Evaluation

Several automatic evaluation metrics exist to assess the quality of machine translation output, and generally fall into categories such as precision-based string matching metrics (e.g. *BLEU* (Papineni et al., 2001), *NIST* (Doddington, 2002), *TER* (Snover et al., 2006)) and semantic matching metrics (e.g. *METEOR* (Banerjee and Lavie, 2005), MEANT (Lo and Wu, 2011)). We describe BLEU and TER, the metrics used in this thesis, below.

2.5.1 BLEU

The most widely used translation metric, BLEU (BiLingual Evaluation Understudy) (Doddington, 2002) is a precision-based metric that uses n-gram based matching to measure the similarity between MT outputs and one or more reference translations. The geometric average of modified n-gram precisions p_n are computed, using n-grams up to

length N (typically 4) and positive weights w_n attributed to each n-gram level, which sum to one. For typical evaluations, uniform weights are assumed. A *brevity penalty* is introduced to ensure that exceedingly short translations are not favored over longer translations which is defined as:

$$\text{BP} = \begin{cases} 1 & \text{if } L_{\text{sys}} > L_{\text{ref}} \\ e^{(1-L_{\text{ref}})/L_{\text{sys}}} & \text{if } L_{\text{sys}} \leq L_{\text{ref}}, \end{cases} \tag{2.27}$$

where L_{sys} is the candidate translation length and L_{ref} is the reference translation length. Thus,

$$\text{BLEU}_N = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right). \tag{2.28}$$

BLEU scores range between the interval [0,1], based on the n-gram similarity between the candidate and reference. It should be noted that BLEU scores are relative to the translation task and thus cannot be compared universally. An enhanced version of BLEU can benefit from multiple references.

2.5.2 TER

Similar to WER in ASR evaluation, translation edit rate (TER) (Snover et al., 2006) computes the minimum number of string edit operations required to transform a MT hypothesis into its reference translation. In addition to the insertion, deletion, and substitution Levenshtein error types described in Sec. 2.1.1, TER includes a *shift* operation that moves a contiguous sequence of words within the MT hypothesis to another position that aligns better with the reference. The shift operation is used to prevent the over-penalization of reordering errors. Because edit distance alignments with shift operations are NP-complete (Shapira and Storer, 2002), a greedy search is required to select a minimal alignment.

TER is additionally a suitable metric for computer assisted translation evaluation, as it correlates well with the amount of effort required to post-edit a machine translation into a human-acceptable output (Federico et al., 2014).

Several variants of TER exist, including *Human-targeted TER* (HTER), which computes TER between a machine translation output and its human post-edited version, and *Multi-reference TER* (mTER), which computes the TER score against the closest translation among a collection of multiple translation references.

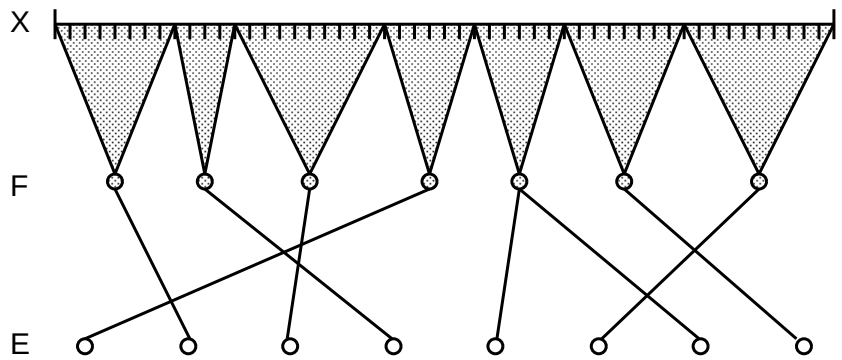


Figure 2.5: An illustration of the search problem in SLT.

2.6 Incorporating ASR in Spoken Language Translation

Spoken language translation (SLT) utilizes automatic speech recognition to enable machine translation systems to translate speech. SLT consists of two layers, as depicted in Figure 2.5.[1] An input signal $\mathbf{x}$ in the form of acoustic vectors is received. The ASR layer combines the features extracted from waveforms in $\mathbf{x}$ into words that compose a source-language utterance $\mathbf{f}$. The words in $\mathbf{f}$ are translated and reordered in the MT layer into a sequence of translated words $\mathbf{e}$.

This process is broken down as a pipeline in Figure 2.6,[2] which begins with the recognition of source words in an acoustic input signal and ends with translation. The recognized words are segmented and processed to include missing punctuation prior to decoding in the MT layer. Ideally, if an oracle ASR system was available, the speech recognition process would always produce a perfect transcription of an acoustic signal. Unfortunately, ASR is riddled with ambiguities. As opposed to translating the single most likely transcription, providing MT with multiple speech recognition hypotheses is may help overcome noise in the ASR output.

Formally, we define the SLT process that combines ASR and MT with the Bayes'

[1]Figure 2.5 is adapted from Casacuberta et al. (2008).
[2]Figure 2.6 is adapted from Stuker et al. (2012).

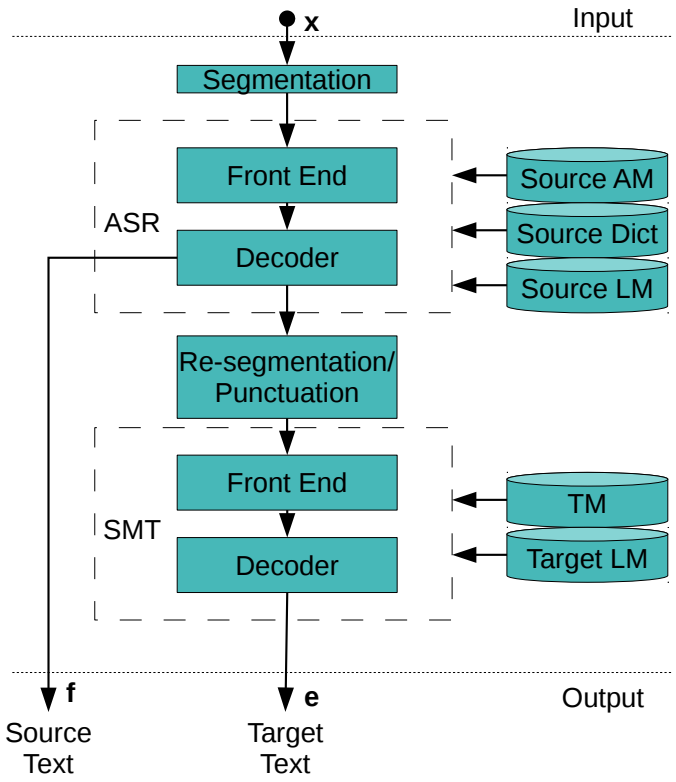


Figure 2.6: The basic speech translation pipeline, combining ASR with SMT.

decision rule as proposed by Ney (1999):

$$
\begin{aligned}
\mathbf{e}^* &= \arg\max_{\mathbf{e}} p(\mathbf{e} \mid \mathbf{x}) \\
&= \arg\max_{\mathbf{e}} \sum_{\mathbf{f} \in \mathscr{F}(\mathbf{x})} p(\mathbf{e}, \mathbf{f} \mid \mathbf{x}) \\
&\approx \arg\max_{\mathbf{e}} \sum_{\mathbf{f} \in \mathscr{F}(\mathbf{x})} p(\mathbf{e} \mid \mathbf{f}) p(\mathbf{x} \mid \mathbf{f}) p(\mathbf{f}).
\end{aligned} \tag{2.29}
$$

Since ASR transcripts are notoriously ambiguous and are a by-product of the SLT task, the Bayes' decision rule reduces the importance of a single speech hypothesis by marginalizing over all source transcriptions $\mathbf{f}$. In practice, it is intractable to explore all transcription hypotheses $\mathscr{F}(x)$ because the number of hypotheses are too large and a word graph is not easy to manage during search.

There are several considerations that must be made when combining ASR and downstream MT to form end-to-end spoken language translation. As ASR hypotheses are produced, the outputs must be preprocessed to fit the orthographic form expected by the MT system. As parts of the preprocessing step:

- *Sentence segmentation* converts ASR utterances into sentence-like units that bet-

ter fit the syntactic structure modeled during MT training (Stolcke and Shriberg, 1996; Lavie et al., 1997; Shriberg et al., 2000; Matusov et al., 2006b; Rao et al., 2007).

- *Streaming segmentation* is an alternative to sentence segmentation that attempts to segment utterances into minimal units that may be translated adequately while simultaneously minimizing the latency of the translation system (Fügen et al., 2007; Sridhar et al., 2013; Oda et al., 2014).

- *Disfluency removal* deletes filler words, speech noise, and normalizes restarts. It is often performed alongside sentence segmentation (Stolcke et al., 1998; Heeman and Allen, 1999; Liu et al., 2006).

- *Punctuation insertion* is used to ensure the input string matches the training conditions of the MT system (Beeferman et al., 1998; Huang and Zweig, 2002). PBMT translation tables contain statistics that score the translations of punctuated source segments into target segments; the reordering model implicitly uses punctuation to help decide how phrases are organized in the translation. NMT attention mechanisms use source token slots containing punctuation to guide fluent output generation.

- *Word recasing* (Kim and Woodland, 2004; Chelba and Acero, 2006). (e.g. "*i*"⇒"*I*") improves the discriminative power of the MT system; for example, it allows the MT system to implicitly adjust the translation of named entities from their common noun counterparts (e.g. *Apple* (corporation) versus *apple* (fruit)).

- *Orthographic normalization* ensures that other lexical forms such as numbers are represented as in-vocabulary items. This includes cases where ASR lexicon entries do not match the MT vocabulary (Adda et al., 1997; Sproat et al., 1999, 2001).

- *Word tokenization* converts single words like *don't* and *it's* into multiple components, such as "*do n't*" or "*don 't*" and "*it 's*" or "*it is*" to separate the syntactic/semantic information and improve word alignment quality. While MT uses word tokenization to improve modeling accuracy, the tokenized forms are not typically used in ASR because they are hard to detect as isolated acoustic events.

In addition to text processing, composition techniques of ASR and MT fall into two research approaches. The first approach addresses the SLT problem as a pipeline which concatenates an ASR and a SMT system as a sequence of dependent processes. The

other approach considers SLT as a joint problem of simultaneously recognizing and translating utterances. We discuss them in greater detail below.

2.6.1 SLT as a Sequential Pipeline

Treating SLT as a sequential pipeline of processes involves simplifying the marginalization of all ASR transcripts in the decision rule defined in (2.29) by considering a list of the top N speech recognition hypotheses from the ASR decoder. This top N can passed to the machine translation decoder as a list of utterances, a *word lattice* (Saleem et al., 2004; Zhang et al., 2005; Dyer et al., 2008; Schroeder et al., 2009), or as a *confusion network* (Mangu et al., 2000; Bertoldi and Federico, 2005). While a word lattice represents a pruned decoding graph from the ASR decoder weighted by its acoustic and language model scores, a confusion network is a compact lattice where each arc represents a word label and a posterior probability. A path in the CN must pass through all nodes in the graph. Both graphs provide information about the ASR search space that can be utilized during MT decoding. Fig. 2.7 provides simple examples of a word lattice and a confusion network. In the alignment process, the CN introduces null arcs that represent gaps. While the CN can be read as input more efficiently than a word lattice, it may permit paths that generate word sequences that are not originally present in the ASR lattice (e.g. "*they offer for us [-] weed ratifying fantasies*") and may stretch the input by inserting an increasing number of epsilons as the word lattice becomes more complex.

Jiang et al. (2011); Ahmed et al. (2012) propose a closer integration by representing the MT system as a phoneme to word translation system. The phoneme to word models are constructed by first modeling word to word translation and reordering models and subsequently applying a grapheme to phoneme transducer to represent the source words in translation table as phonemes. ASR phoneme confusion networks are passed to the MT system for decoding.

In each of these approaches the MT decoder processes an input lattice or CN by constructing a translation lattice which pre-fetches all translation options from the translation model whose constituents appear in any of the paths of the lattice. This idea is rooted in the concept of multi-source translation (Och and Ney, 2001). Beam search is used to prune unlikely paths through the input graph.

An advantage of this decoupled approach is that the ASR system can be trained on monolingual data and enjoys standard acoustic- and translation model adaptation techniques. Additionally, decoupling substitutes the statistical dependence between the

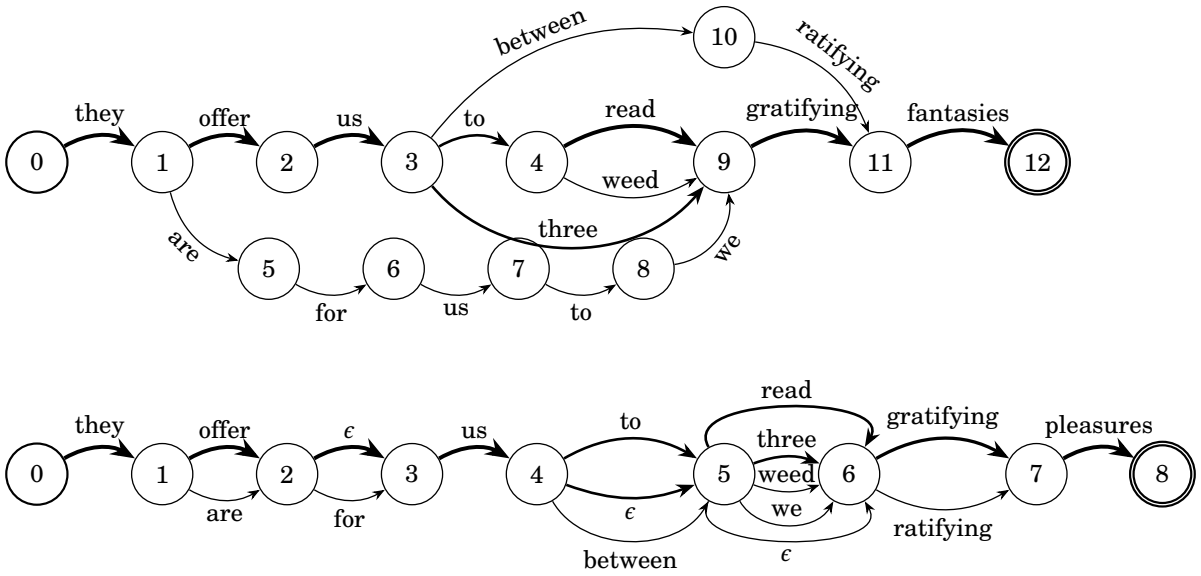


Figure 2.7: Examples of a word lattice (top) and a confusion network (bottom). Edges are weighted by ASR model scores in the word lattice, and posterior scores in the confusion network.

speech signal and SMT with word lattices or confusion networks, reducing the search space for each decoder. On the other hand, He et al. (2011a) show that automatic speech recognition evaluation metrics such as WER are not a good metric for ASR in the SLT scenario through experiments that highlight the mismatch between WER and BLEU. While a WER score of zero implies a perfectly recognized ASR hypothesis, a nonzero score does not provide any indication how the misrecognized words affect the context of the entire utterance. For example, the deletion or substitution of a content word is scored by WER the same as that of a function word; however, as shown in the examples from Figure 1.1, isolated function word substitutions by phonetically similar alternatives may not affect the adequacy of the translation (e.g. "*could*"⇒"*can*"), in the same way that the substitution of a content word changes the meaning of the utterance (e.g. "learning"⇒"loaning").

2.6.2 Unified Spoken Language Translation

Unified SLT approaches represent the search space over all ASR hypotheses as an intermediate step, enabling the marginalization over all possible source text hypotheses $\mathbf{f}$. For simplicity, the marginalization in (2.29) is typically simplified by replacing this summation with a $\arg\max$ operator to find the most likely ASR hypothesis. By representing

the translation and lexical reordering models as weighted finite state transducers (WFSTs), Bangalore and Riccardi (2001, 2002) compose the MT system directly with the WFST represented by ASR, which has the additional benefit of efficient decoding. Zhou et al. (2007) integrates the works of Bangalore and Riccardi (2002) and Jiang et al. (2011) by integrating a fully finite-state phrase-based SMT framework that constructs source word sequence, source segmentation, phrase translation, phrase-to-word transduction, and the target language model machines. SLT decoding is treated as a best path search through the WFST after composing the WFST network with the ASR system. While the WFST approach allows for the efficient coupling of ASR and MT, the complexity of the subcomponents of the PBMT system requires strong constraints on the reordering distance allowed during translation, as well as a limitation on the length of phrases in the phrase table. Thus, a fast-performing decoder will have to sacrifice accuracy for speed.

He et al. (2011a) test a "truncated log-linear model" which adds the ASR components as features into a hierarchical phrase-based English-Chinese SMT system. By only optimizing the ASR feature weights in the log-linear model, they demonstrate the efficacy of unified SLT with BLEU optimization over sequential pipelining with WER evaluation for ASR.

Zhang et al. (2011) use decision-feedback learning to optimize ASR and SMT models for SLT. Using the Bayes' decision rule in (2.29), a discriminant function $D(\cdot)$ is defined,

$$D(\mathbf{e},\mathbf{f};\mathbf{x}) = \log(p(\mathbf{e} \mid \mathbf{f})p(\mathbf{x} \mid \mathbf{f})p(\mathbf{f})), \tag{2.30}$$

which scores each source-language speech input $\mathbf{x}$ to its recognition hypothesis $\mathbf{f}$ and translation hypothesis $\mathbf{e}$. The optimal translation hypothesis $\mathbf{e}^0$ is selected by finding the hypothesis with a maximal BLEU score against its reference. The optimal speech recognition hypothesis is selected by

$$\mathbf{f}^0 = \arg\max_{\mathbf{f}^i} D(\mathbf{e}^0, \mathbf{f}^i; \mathbf{x}). \tag{2.31}$$

The optimal hypotheses are used to compute a loss function that defines a parameter updating scheme for source- and target LMs, as well as the translation model.

2.7 Machine translation error modeling approaches

Pérez et al. (2012) compare the performance of tightly coupled and decoupled approaches to SLT and discover that while integrating the models keep good quality translation hy-

potheses in the decoding process, re-scoring models have not been able to exploit the them to improve translation quality over decoupled approaches.

Ananthakrishnan et al. (2013) modify a phrase-based SMT decoder to include penalties for bilingual phrase pairs spanning erroneous and error-free regions of input, and target language model (LM) likelihoods in the vicinity of source errors. While observing significant improvements, their approach assumes the presence of a high-performing ASR error detection process, which is a non-trivial task.

Tsvetkov et al. (2014) model the presence of ASR errors in MT training by replicating translation model entries and damaging the words with acoustically confusable words through a sequence of WFST operations and report significant improvements over baseline TED talk translation tasks.

2.8 Chapter Summary

In this chapter we provided an overview of automatic speech recognition and statistical machine translation for spoken language translation.

We first outlined the process of ASR, from the identification of speech units to the recognition of words and described its primary components, the acoustic model, the pronunciation dictionary, and the language model. We introduced Word Error Rate, the de-facto automatic evaluation metric for ASR.

We subsequently introduced statistical machine translation and defined a generative model for SMT based on the noisy-channel model. We outlined the original IBM models for word alignment, which assume that words are individually translated in a sentence. We then extended the discussion to phrase-based models, which leverage word alignments from the IBM models to construct a phrase translation table and a richer reordering model. We then factorized the phrase-based translation model into a log-linear model composed of a phrase translation table feature, a reordering feature, and a language modeling feature that can be assigned different weights. We additionally summarized the decoding process in which an input sentence is translated into an output sentence and outlined techniques to reduce the search space into a tractable problem via a beam search.

We introduced language modeling as a component of both ASR and MT: during ASR decoding language models score the fluency of a recognized hypothesis, while during MT decoding it is used to score the target language hypotheses as they are being generated. We introduced n-gram language modeling as a technique used in this thesis for SLT

language modeling and described smoothing techniques to improve its robustness for modeling low frequency n-grams. Next, we introduced neural machine translation NMT as an alternative representation of SMT, which composes the translation problem as a unified encoder-decoder task that begins by encoding a sequence of source words into a sequence of hidden states, a decoder that generates target words, and an attention model that uses the context of hidden states representing the input words to guide the decoding decisions. Finally, we outlined MT evaluation metrics, which are shared by SLT and outlined techniques to combine ASR and MT to form an SLT system.

In this thesis, we focus primarily on the single-best outputs from the automatic speech recognizer, primarily since our SLT experiments rely on the ASR outputs of systems spanning multiple research laboratories.

CHAPTER 3

LANGUAGE COMPLEXITY OF TEXT VERSUS SPEECH

As we consider spoken language translation, we must consider attributes which distinguish spoken language from text and how these attributes impact translation quality. While speech recognition errors are prime contributors of errors in machine translation, there are other properties of spoken language that make a SMT system trained on only text insufficient for speech translation.

Specia et al. (2011) outline three categories of discriminative features that are relevant for models trained to perform MT quality estimation: *confidence* indicators derived from MT models, *complexity* indicators that measure the difficulty of translating the source text, and *fluency* indicators that measure the grammaticality of a translation. Likewise, the difficulty of a translation task can be estimated by analyzing source complexity and target language features that indicate the capacity of a statistical system to generate fluent translations.

In this chapter, we focus on *language complexity* and how the transition from a written to a spoken register poses challenges for machine translation. The majority of psycholinguistics research on language complexity focuses on language acquisition and generation by native speakers or second language learners, with a primary focus on a single language. But what makes a linguistic interaction understandable by humans? Audience members rely most on extralinguistic information, which includes prior world knowledge and their familiarity with the topics mentioned within a discourse to interpret its meaning. The conveyor of information often uses a variety of linguistic devices, such as anaphoric mentions and grounding to prime a recipient's extralinguistic knowl-

edge. Additionally, the audience must be able to organize the information received from the discourse into coherent blocks.

Graesser et al. (1994) claim that the audience routinely attempts to construct coherent meanings and connections among constituents in a discourse unless the quality of the discourse is too poor. This *coherence assumption* forms one the core hypotheses in the constructivist theory of discourse comprehension. As a result, many complexity analysis tools attempt to detect coherence and cohesion through syntax, semantic, and discourse connectives (Graesser et al., 2004; Mitchell et al., 2010; Newbold and Gillam, 2010).

3.1 Spoken versus Written Registers

It is often assumed that people write in the same manner as they speak. However, the manner in which an interlocutor conveys her message depends on both the medium and the context. Finegan (2014) highlights four general distinctions observable in most spoken and written registers.[1]

1. **Oral communication can exploit intonation and voice pitch to convey information.** Written text must rely on word choice, syntax, rhetorical structure, and punctuation to provide cues to the reader, while oral communication uses multiple channels simultaneously to convey information. For example, Finegan highlights prosodic and tonal features in speech that convey irony and sarcasm.

2. **Speakers and addressees often have visual contact with one another.** In addition to speech, speakers and addresses may use multimodal features, including body language, to convey information. A speaker may gesture to refer to entities located in vicinity of the conversation or use referents such as “this” and “that”, without introducing the entities in spoken dialogue.

3. **Speech and writing differ in the amount of planning that is possible.** Most written registers provide sufficient time for the author to plan and modify her text.

[1]While these tenants are true in the majority of written and spoken contexts, there are some registers that overlap. For example, chat scenarios are dialogue-based, where communicators interact more closely to real-time and thus do not plan their utterances in the same way as other written registers. Prepared speeches allow the opportunity to revise and redact the form and content of the speech in a similar process as written articles; although the speaker may deviate slightly during execution as he receives feedback from his audience.

Although the same can be said about prepared speeches, the majority of spoken interactions occur spontaneously or with minimal planning. Finegan remarks that as a result of planning, written registers typically show a richer and more varied vocabulary than spoken registers. In spontaneous speech, the use of richer vocabulary may include more pause time as the speaker selects the vocabulary in real-time.

4. **Written registers tend to rely less on the context of interaction than spoken registers do.** While a speaker can gauge the addressee's comprehension or acceptance of her utterance and can use speech acts to achieve a particular goal, a writer must construct her discourse based on prior assumptions about the knowledge of the reader that cannot be amended through real-time feedback.

Biber (1988) and follow-up work by researchers investigated the variation in cohesion across text and speech corpora. In particular, Biber (1988) used factor analysis to divide 23 written and spoken registers into several categories, based on their linguistic features. Louwerse et al. (2004) extended the factor analysis approach of Biber (1988) by performing a multi-dimensional analysis to identify particular linguistic features that divide written and spoken genres across several registers. Their results show variance between speech and writing corpora on a variety of factors, including type frequency, polysemy, pronoun density, abstract noun usage, type-token ratios for nouns, and stem overlap. These features divide the written and spoken genres into subdomains posing unique challenges in comprehension (e.g. prepared speeches versus conversational speech; news broadcasts versus legal documents). In addition to distinguishing between speech versus writing, the results of Biber (1988) and Louwerse et al. (2004) demonstrate that spoken and written sub-genres share a number of linguistic properties in common, based on the following dimensions.

- **Informational versus declarative registers.** Informational registers (e.g. spontaneous speeches, broadcast news) are characterized by a higher occurrence of temporal cohesion, imageability, and concreteness, but a low occurrence of causality, as opposed to declarative registers (e.g. planned speeches, academic writing).

- **Factual versus situational registers.** Situational registers (e.g. telephone conversations, editorials) contain a higher frequency of imageability and a lower frequency of clarification and causal connectives than factual registers (e.g. spontaneous speech, academic writing).

- **Topic consistency versus topic variation.** Louwerse et al. remark that written letters and spontaneous speech, for example, often have a similar set of topics that are used, while face-to-face conversations, public debates, and editorials have greater topic variation, marked by lower cohesion.

- **Elaborative versus constrained registers.** Elaborative registers tend to be more opinion-based (e.g. personal letters, press reviews), while constrained registers are more concise and factual (e.g. professional letters, press reviews).

- **Narrative versus non-narrative registers.** Narration of events is more prominent in fiction writing and biographies, as well as face-to-face conversations, as opposed to press reviews and professional letters, for example.

3.2 Language Complexity in Machine Translation

Given the differences between linguistic features used in various spoken and written registers, what aspects of language pose difficulties for natural language processing tasks, such as machine translation? And how would a mismatch in training and testing conditions affect a machine translation system's ability to translate?

Given the variance within spoken and written registers, we attempt to focus on complexity issues that are irrespective of a particular text, speaker, or language pair and focus on issues that are relevant to the machine translation task. We can categorize these issues into three general areas: the lexicon, syntax, and semantics. When considering the lexicon, we can observe effects of vocabulary size, morphological variations, and both lexical and translation ambiguity as key impacts affecting the ability of the statistical models to cover the words in the language (Carpuat and Wu, 2007). On the syntax level, sentence length, structure complexity, and structural dependencies affect the decoding search space. On the semantic level, phenomena such as idiomatic expressions, figures of speech, anaphora, and elliptical expressions define intrinsic limitations of syntactic models. While we can observe nearly all of these language features on the monolingual level, many of these issues have a greater impact when transferring linguistic information in the process of translation. Between distant language pairs, the effects of these linguistic features cause a cumulative increase in the difficulty of MT.

Although discourse-based machine translation takes into account intersentential factors affecting translation quality (Carpuat, 2009; Foster et al., 2010; Xiao et al., 2011),

the majority of MT systems treat each sentence independently, ruling out additional context.

3.3 Corpus Analysis of Language Complexity

Following our exposition in Ruiz and Federico (2014b), we compare two sources of spoken and written language, derived from the de-facto evaluation campaigns in speech and text translation. From the International Workshop on Spoken Language Translation (IWSLT), we utilize as spoken language corpora the English TED talk transcripts[2] from the IWSLT 2013 evaluation campaign (Cettolo et al., 2013). TED talks fall under the prepared speech register, thus minimizing the effects of spontaneous speech and dialogue on a spoken register in our analysis. As a written register, we select the News Commentary texts from the translation evaluation campaign in the 2009 Workshop on Machine Translation (Callison-Burch et al., 2009).

Both types of texts cover a variety of topics whose content is produced by several authors and represent corpora from the de-facto evaluation tasks for text and speech translation. Although these types of texts correspond to different genres, they are popular representatives of spoken and written language investigated in MT, while belonging to similar domains. Both genres consist of speakers or authors with similar communication goals: namely, the mass distribution of information and ideas delivered by subject matter experts. At the same time, TED speakers have the additional objective of selling ideas through persuasive speeches. We focus on the English-German language pair, which belong to the same language family, but have marked differences in levels of inflection, morphological variation, verb ordering, and pronoun cases. In practice, the top performing MT systems use many of the same training and decoding approaches in these evaluations. But are the WMT and IWSLT translation tasks just different flavors of the same translation problem? Are the strategies used to translate written language directly applicable to the genre of spoken language – in particular, prepared speeches? Our goal is to investigate the qualitative and quantitative differences between the genres of news texts and prepared speech that explain differences in MT system performance across translation tasks, as well as the types of errors occurring often in MT systems trained on text and speech corpora.

In our experiments, we sample approximately two million words from both the English TED and WMT News Commentary corpora, as well as the German translations

[2] http://www.ted.com/talks

Measure	TED-EN	WMT-EN	TED-DE	WMT-DE
Word Count	2,000,018	2,000,016	1,890,106	2,046,071
Line Count	103,588	82,256	103,588	82,256
Surface forms	46,001	50,129	86,787	95,922
Stems	34,417	36,904	62,929	66,735
Words/Line	19.31	24.31	18.25	24.87
Stem/Surface	0.748	0.736	0.725	0.696

Table 3.1: Statistics for two million word TED and WMT News Commentary corpora samples.

of their sentences. In our continued discussion, we refer to WMT News Commentary as simply "WMT". Rather than randomly sampling sentences from the corpora, we sequentially read the sample to allow us to preserve the underlying discourse. Sentences containing more than 80 words are excluded. We additionally subdivide the sampled corpora into blocks of 100,000 words to measure statistics on vocabulary growth rate.

We use TreeTagger (Schmid, 1994) to lemmatize and assign part-of-speech tags using the Penn Treebank (Marcus et al., 1993) and STTS (Schiller et al., 1995) tagsets for English and German, respectively. Some simple corpora statistics are provided in Table 3.1.

3.4 Word statistics

3.4.1 Sentence length

Since the unconstrained search space in MT is exponential with respect to the length of the source sentence, we examine the distribution of sentence lengths between the TED and WMT News Commentary corpora, as shown in Figure 3.1. On average, TED consists of lines containing around 19 words, while News Commentary averages five more words per line. Forty percent of the sentences in TED have between six and 15 words, while the majority of the sentences in News Commentary contain over 20 words. This suggests that TED is less susceptible to length-dependent decoding issues such as long distance reordering.

3.4.2 Predictability: Perplexity and new words

Perplexity measures the similarity of n-gram distributions between a training set and a test set. Source and target language n-gram distributions govern a Phrase-based SMT

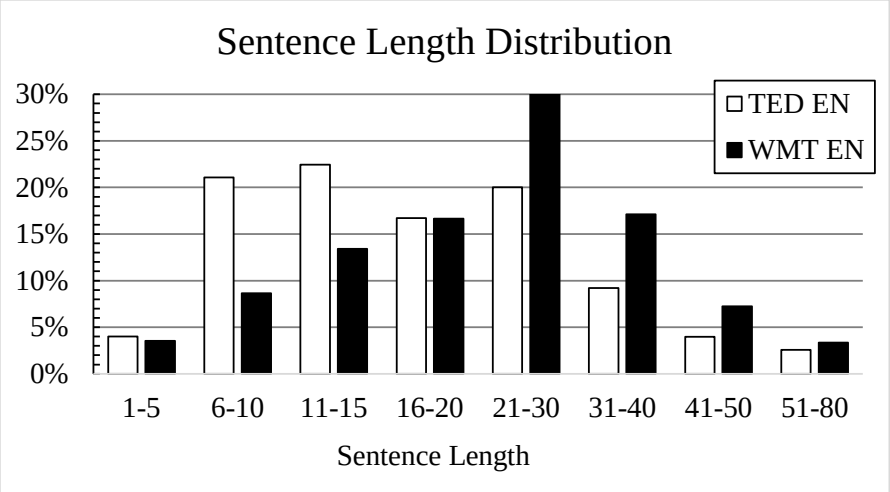


Figure 3.1: Sentence length statistics for English. TED talk sentences are shorter – typically between 6 and 30 words, while WMT News Commentary sentences are usually 11 to 40 words long.

(PBMT) system's capacity to adequately translate a sequence of words with its phrase table and language model (LM). Likewise, the out-of-vocabulary (OOV) rate estimates the amount of source words that are impossible to translate with the given training data. We measure these notions of complexity by constructing English and German language models and evaluating their predictive power against in-domain data. Using our 2 million word corpora samples, we incrementally add 100,000 words to each corpus and evaluate its perplexity and OOV rate against a held-out 100,000 word sample from

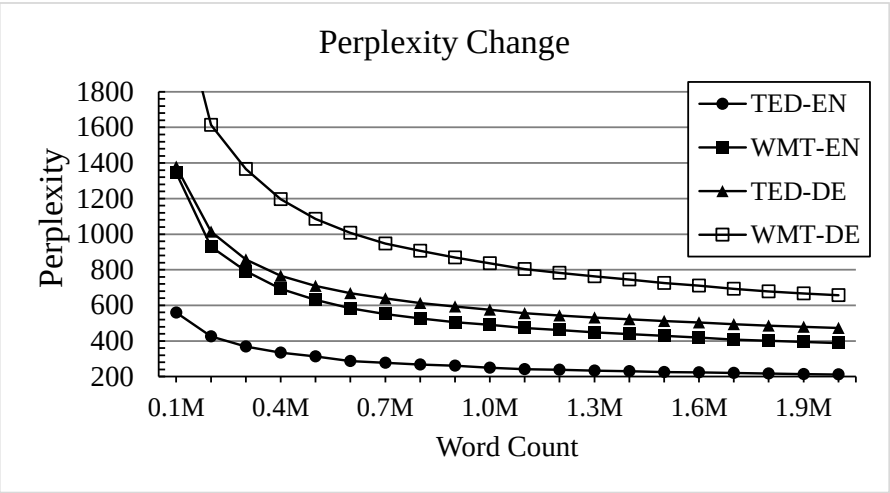


Figure 3.2: Perplexity change as corpus size increases for English and German.

each training corpus. Using IRSTLM (Federico et al., 2008), we construct trigram LMs, using improved Kneser-Ney smoothing, no pruning, and a fixed vocabulary size of 10 million words.

According to the results shown in Figure 3.2, TED consistently has lower trigram perplexity rates (-46% with the full data for English, -28% for German). We observe no significant differences in OOV between TED and News Commentary. The results suggest TED is more capable of being modeled than News Commentary with the same amount of training data and the translation of TED is more regular than the translation of News Commentary.

3.5 Lexical ambiguity

Two measurements of lexical ambiguity are word polysemy and translation entropy. We analyze the ambiguity of noun and verb lemmas, which as content words carry the most important information needed to understand a sentence. We only consider the types that contain sense information in WordNet (Fellbaum, 1998). We take the top 100 lists of verbs and nouns from each corpus and measure their ambiguity, as described in the sections below. We compare the results against measurements on the full set of nouns or verbs and additionally measure the overlapping lemmas in the corpora.

3.5.1 Polysemy

As an upper-bound measure of word ambiguity, we measure the number of senses each English word in the corpus can express, as reported by WordNet. While not every sense may be observed in our corpora, this measure estimates how ambiguous a corpus is for a statistical system that considers each sense to be equally likely for a given word. Figure 3.3 provides a comparison between the top 100 verb and noun lemmas in the two corpora. On a global scale, we do not observe significant differences in the number

Lemma	# Senses	TED	WMT
tell	8	2159	362
learn	6	1102	336
hear	5	875	187
read	11	529	110

Table 3.2: Common polysemic verbs and their occurrence frequencies in TED and WMT News Commentary.

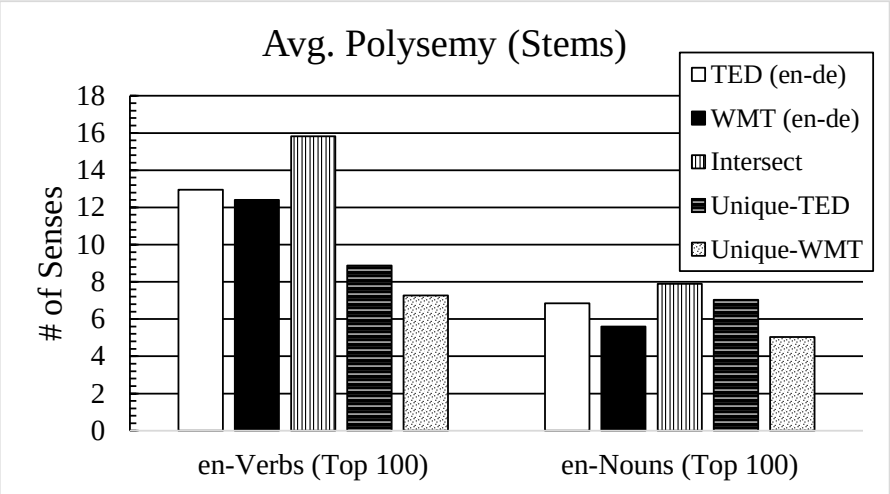


Figure 3.3: Average number of senses per verb/noun for the 100 most English frequent words in each corpus, as well as the types shared in common (Intersect), and those unique to the respective corpus (Unique-TED and Unique-WMT).

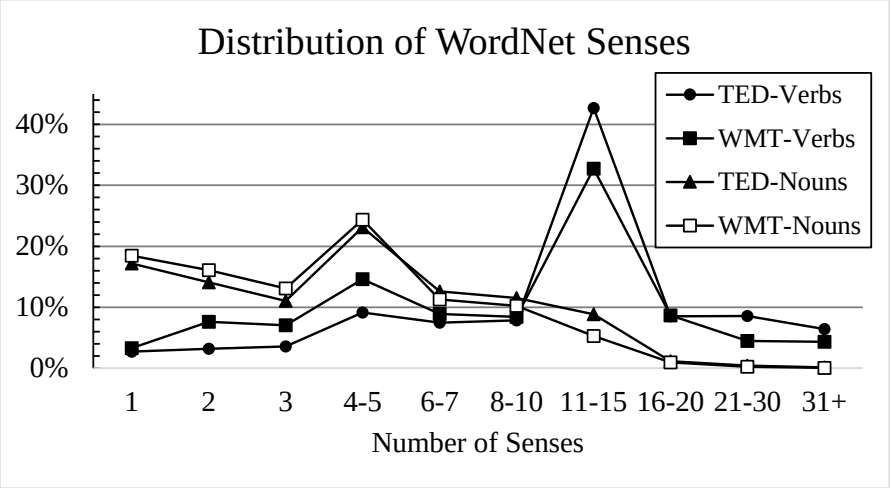


Figure 3.4: Distribution of WordNet senses for all English nouns and verbs in TED and WMT News Commentary, weighted by observation frequency. Frequencies are bucketed to highlight differences between the corpora.

of senses over the entire set of verbs and nouns in the corpora. By focusing on the top 100 lists, we observe that while the nouns and verbs shared in common between TED and WMT explain the majority of the ambiguity with respect to polysemy, the non-overlapping lemmas demonstrate TED's higher ambiguity through the use of common verbs and nouns. By isolating the lemmas that are unique to each corpus' top 100 list, we see that TED's verbs and nouns exhibit 1.5 and 2 more senses respectively than

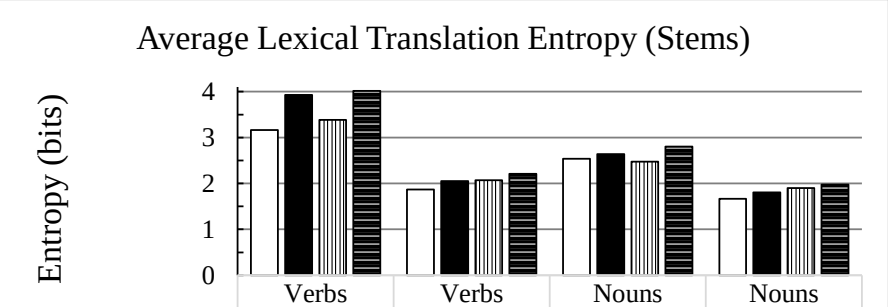
those of WMT.

In order to measure the overall effects of polysemy on the corpora, we weight the noun and verb senses by their corpora frequencies. Figure 3.4 shows how the distributional frequency of noun and verb senses varies over TED and WMT vary. For verbs, we observe that TED exhibits fewer tokens with low ambiguity and a significant increase in tokens with over 11 word senses. The noun senses behave in a similar manner, though the differences are not as pronounced.

These results demonstrate that TED favors the use of common, expressive verbs. Examples are shown in Table 3.2. We find that this is the case when combining these observations with the perplexity measures in Section 3.4.2.

3.5.2 Lexical translation entropy

If the results in Section 3.5.1 suggest that TED talks are more ambiguous through the use of common verbs and nouns, does this transfer to the problem of MT? To address this question, we analyze the lexical translation table provided by Moses and MGIZA through the word alignment process. We again compare TED and WMT both on the top 100 lists and the full sets of noun and verb lemmas. We train a word alignment model using MGIZA on the lemmatized corpora to build an English-German lexical translation table. In order to control the effects of alignment noise, we consider the German lexical



	Verbs (Top 100)	Verbs (All)	Nouns (Top 100)	Nouns (All)
TED	3.166	1.870	2.536	1.667
WMT	3.925	2.051	2.639	1.804
TED-intersect	3.385	2.071	2.471	1.902
WMT-intersect	4.013	2.210	2.804	1.974

Figure 3.5: Average lexical translation entropy (bits) on English noun and verb stems, computed from the top 95% threshold in the lexical translation table generated by MGIZA.

translations of each English lemma that cover the top 95% of the probability mass. Figure 3.5 compares TED and WMT in terms of lexical entropy.

Translating the top 100 verbs is much less ambiguous in the TED talk translation task (3.2 bits versus 3.9 bits). Most of the entropy is explained by the set of verbs TED and WMT share in common. WMT suffers from underspecification of these primarily common verbs. For example, the verb "bring", which occurs over 800 times in both corpora, exhibits an entropy of 4.04 bits and 170 translation options in TED, as opposed to 4.39 bits and 210 translation options for WMT. In terms of translation perplexity, the translation difficulty is as hard as deciding between 16 equally likely translations in TED, versus 21 in WMT. As a word with 11 senses in WordNet, this implies that fewer senses are actually being considered during translation in TED. A similar behavior can be observed for the common nouns. These results indicate that while TED has potentially higher English noun and verb polysemy, the common nouns and verbs are used more regularly than in WMT.

3.5.3 Pronominal anaphora

Hardmeier and Federico (2010) demonstrate that differences in the pronominal systems of a source and target language often results in the mistranslation of pronouns. For example, German has four personal pronoun cases, while English only has two. In cases of high ambiguity, it is up to models that depend on local context, such as n-gram LMs to determine the correct pronoun to use in the translation. If the local features of the sentence cannot resolve the ambiguity, the output pronoun is up to chance. We highlight two additional problems outlined by Hardmeier (2012): the difficulty for anaphora resolution systems to resolve pronouns (e.g. expletive pronouns), and translation divergences, such as when a pronoun is replaced with its referent in the translation.

Using the POS tags assigned by TreeTagger, we identify the English and German pronouns for TED and WMT and report statistics in Table 3.3. TED contains three

Person	Pronouns	TED	WMT	Diff	Rel Diff
1st	10	3.85%	0.48%	3.37%	699.2%
2nd	4	1.68%	0.06%	1.63%	2776.5%
3rd	24	4.06%	2.56%	1.50%	58.6%
Total	38	9.59%	3.10%	6.49%	209.5%

Table 3.3: Percent of English pronoun tokens in the 2 million word TED and WMT samples. Pronouns are grouped by grammatical person.

Field	TED	WMT	Difference	Relative Difference
Idioms/1K	1.541	2.122	-0.581	-27%
Avg. Length	2.896	2.695	0.201	7.46%
Types	494	556	-62	-11%
Singletons	289	271	18	7%

Table 3.4: The average rate of idioms per 1,000 words, idiom length, and the number of idiom and singleton types in each corpus sample.

times as many pronouns than WMT. While WMT contains few first and second person anaphoric mentions, TED consists of talks in which the speaker often refers to himself and to the audience. In particular, TED and WMT share seven pronominal translations for the English pronoun "you", based on the context of the sentence. At times, "you" may be translated as an indefinite pronoun ("man", "jemand", "eine"), or can be replaced with a different grammatical person ("wir", "sie"). TED contains additional ambiguity which may be attributed to word alignment errors, resulting in high translation entropy (1.53 bits). Likewise the indefinite and ambiguous pronoun "it" occurs twice as often in TED.

3.5.4 Idiomatic expressions

Low frequency idiomatic expressions pose challenges for MT systems. We crawled a list of English idioms generated by an online user community[3]. We manually scanned and pruned a handful of submitted entries that were likely to suggest more false positives than actual idiomatic expressions. In total, we collected 3,720 distinct idiomatic expressions. We perform a greedy idiom search on the surface representation of each corpus, favoring long idioms and ensuring that idioms do not overlap one another. Some statistics are reported in Table 3.4.

TED and WMT share 237 idioms in common, such as "at the end of the day", "in the face of", and "on the table". These signify expressions that cross genres and are likely to be easily represented with statistical models. Some TED-specific expressions include "beeline for", "bells and whistles", "up the wall", and "warm and fuzzy" – expressions that may be difficult to translate in MT systems trained on news genres. While TED uses fewer idioms overall, nearly 60% of the idiom types appear only once, compared to nearly 50% in WMT.

[3]http://www.usingenglish.com/reference/idioms/

3.6 Word reordering

One of the most notorious problems in phrase-based statistical machine translation is word reordering (Birch et al., 2009). Expressing the reordering problem as a task of searching through a set of word permutations for a given source sentence **f**, we arrange each source word f_i according to the mean of the target positions $\bar{a}_i$ aligned to it, as suggested by Bisazza and Federico (2013). Unaligned words are assigned the mean of their neighboring words' alignment positions. We then compute a word-after-word distortion length histogram between adjacent source words in their projection to the target language (Brown et al., 1990). To eliminate the effects of sentence length, we randomly sample 100 sentences with replacement for each observable sentence length in each corpus. A histogram is computed for each sentence length, whose results are averaged together.

Figure 3.6 compares the reordering behaviors of TED and WMT after stratified random sampling. Word permutations are computed from the symmetrized word alignments on English and German stems, using the *grow-diag-final-and* heuristic in Moses. To visualize the results better, we consider the absolute value of the relative distortion positions. In the figure, Bucket #1 corresponds to discontiguous reordering jumps one position forward (i.e. e_i __ e_{i+1}) or backward (i.e. e_{i+1} e_i), and so on. For example, "*we could communicate*" is translated once as "*wir kommunizieren können*" and yields re-

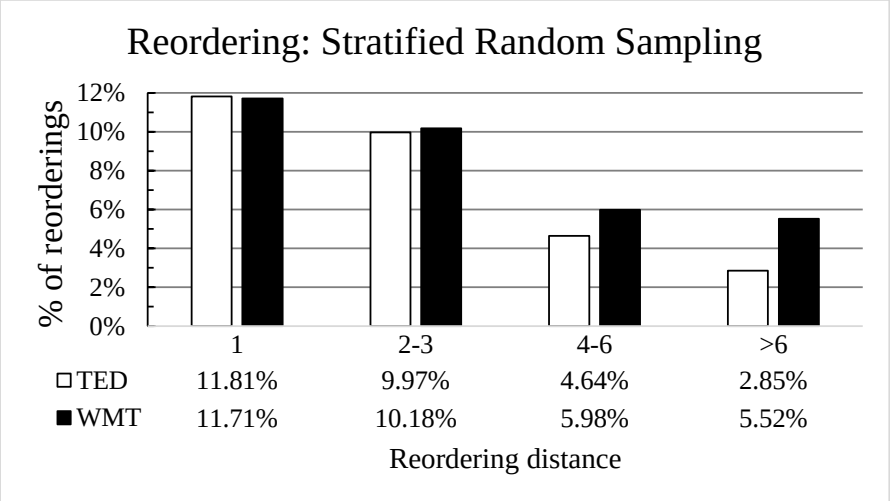


Figure 3.6: Discontiguous word reordering percentage by reordering distance for English-German. Statistics are computed on reordering buckets of ±1, ±[2,3], ±[4,6], and ±[7,∞).

ordering jumps of (+1,-1), which are both binned into Bucket #1. For English-German, monotonic reorderings account for 70.73% and 66.63% for TED and WMT, respectively. This 4% absolute increase in monotonic reorderings for TED is accounted for by the reduction in long distance reorderings of four positions or more.

3.7 Machine Translation performance

Thus far, we have identified several linguistic factors that distinguish the TED translation task from that of WMT News Commentaries. We continue our analysis with a head-to-head comparison of MT performance. Since we cannot directly compare BLEU scores from the two official evaluation tasks, we create a small scale baseline evaluation that fixes the corpora sizes. Using the same two million word samples, we train separate PBMT systems on TED and WMT, and tune two held-out samples of 100,000 words. We average the results of three MERT runs to reduce random effects. Each MT system is trained with the default training parameters of Moses (Koehn et al., 2007). We construct separate 4-gram LMs on the German side of the training data with IRSTLM, using a similar configuration as in Section 3.4.2. To evaluate, we control the effects of sentence length by focusing on sentences containing between 10 and 20 words (after tokenization). For each unique sentence length, we sample 200 sentences with replacement from 300,000 word segments of the TED and WMT corpora. We evaluate using the Translation Edit Rate (TER) metric (Snover et al., 2006). Results are reported in Figure 3.7 for PBMT systems trained with 500K, 1M, and 2M words.

Due to the limited amount of TED data, we cannot measure the effects of additional training data on translation quality, but we attempt to extrapolate the learning curve by looking at smaller training sets. While we cannot explicitly say that TED translation yields higher translation quality than that of WMT, we do observe a growth in the absolute TER difference from 6.4% to 6.8% with 500K words and 2M words, respectively. Likewise, TED has fewer phrase table entries (3.5M vs. 3.7M) and LM entries (1.68M vs. 1.91M 4-grams) than WMT. These results suggest that the characteristics of TED allow better modeling of the translation task with less training data.

3.8 Summary

We have studied several phenomena that indicate differences between speech and text that affect machine translation. Both TED and WMT News Commentary are good sand-

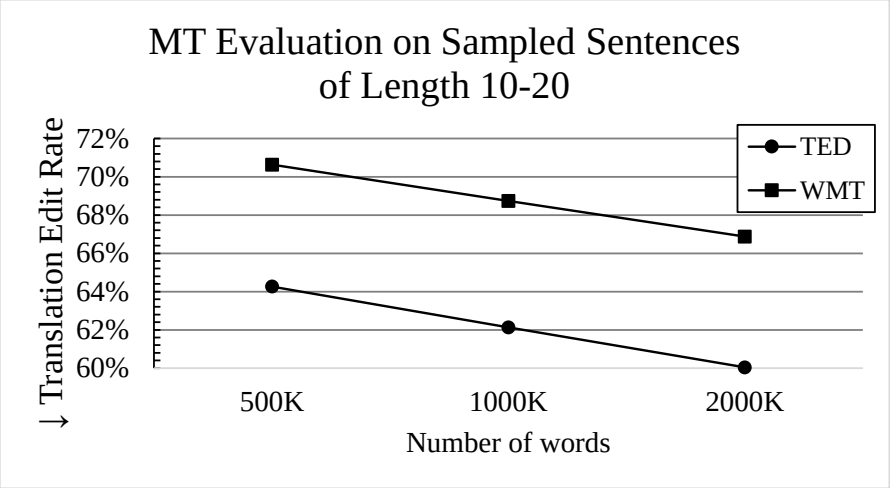


Figure 3.7: Phrase-based MT results for sampled sentences of length 10-20 in TED and WMT. PBMT systems are trained with 500K, 1M, and 2M words.

boxes for evaluating specific aspects of MT. Our experimental results identify several distinct linguistic phenomena that distinguish each genre's usefulness on specific areas of MT research.

TED talks enjoy performance advantages due to a MT system's ability to translate their content reasonably well with a surprisingly small amount of training data. While TED has lower lexical ambiguity than WMT in terms of translation entropy, it uses significantly more common and thus more ambiguous expressions. Because of this, it is a good candidate for evaluating semantically-informed translation models. The key issue for TED talks is the problem of pronominal anaphora. With over three times as many pronouns than WMT and twice as many third person mentions, the ability for MT systems to handle context is crucial. This makes it is an excellent task for investigating the translation of anaphoric expressions through discourse-aware MT, while at the same time managing the complexity of the system.[4]

As WMT consists of longer sentences with more frequent cases of long distance reordering, it is a better task for measuring differences between hierarchical and linear phrase-based SMT. Additionally, with a lower German-English sentence length ratio, noun and verb compound detection may be a larger issue in WMT. WMT also suffers from higher perplexity scores than TED, suggesting that it can be a good benchmark for evaluating language modeling strategies with large amounts of readily-available

[4]The translation of pronouns is an active area of research for discourse-based machine translation. Due to its high ratio of pronouns to tokens, TED talks have consistently been used in the DiscoMT shared task on pronoun translation (Hardmeier et al., 2015). `https://www.idiap.ch/workshop/DiscoMT`

in-domain data. Both TED and WMT are good candidates for research on handling idiomatic expressions during translation.

Some linguistic features do not correspond well with the problem of translation difficulty. As shown with our comparison of WordNet polysemy and lexical translation entropy, the challenge of disambiguating between a high number of noun and verb senses lessens during the word alignment process. This could be one of the reasons why previous work on word sense disambiguation in MT has yet to achieve significant improvements in automatic evaluations (Carpuat and Wu, 2007).

It should also be mentioned that while TED appears to be a simpler MT task overall, we have not addressed the larger problem of TED talk translation: the integration with automatic speech recognition. The linguistic features of TED make it a perfect candidate for speech translation, allowing researchers to focus on problems of translating content that may have been corrupted by speech recognition errors.

3.9 Chapter Summary

We have shown that the TED spoken language corpus and WMT News Commentary machine translation corpora exhibit differences in several linguistic features that each warrant dedicated research in machine translation. By sampling two million words from TED and WMT, we compared the two corpora on a number of linguistic aspects, including word statistics, such as sentence length and language model perplexity, lexical ambiguity, pronominal anaphora, idiomatic expressions, and word reordering. We observe that while TED consists of shorter sentences with less reordering behavior and stronger predictability through language model perplexity and lexical translation entropy, it has increased occurrences of pronouns that may refer to antecedents in the transcript and a high amount of polysemy through common verbs and nouns. In a small MT experiment, we evaluated a subset of sentence lengths in TED and WMT with MT systems trained on a comparable amount of data and show that TED can be modeled more compactly and accurately.

Finally, we have outlined linguistic features that distinguish the two corpora and propose suggestions to the MT community to focus their attention on TED or WMT, depending on their research goals. While both tasks are interesting for MT research, characteristics of spoken versus written texts provide different challenges to overcome. In the subsequent chapters, we turn our attention to the translation of TED talks in the presence of speech recognition errors.

CHAPTER 4

Speech Recognition Errors and Spoken Language Translation Quality

In the previous chapter, we studied linguistic phenomena that demonstrate differences between speech and text. These differences were shown to affect the way machine translation handles spoken language versus written language registers. While the differences outlined in the previous chapter are interesting, they leave out a crucial piece to spoken language translation: automatic speech recognition (ASR).

In spoken language translation, it is crucial that an ASR system produces outputs that can be adequately translated by a machine translation system. The introduction of ASR errors is the single greatest point of failure in machine translation, not only by affecting the translation of individual words, but also the entire context surrounding each misrecognized word. A phrase-based MT (PBMT) system's translation model relies on statistics governing the translation of contiguous sequences of words from one language to another. If a single word is misrecognized, not only does the individual word get translated wrong, but it is also no longer possible to select long phrases surrounding that word in the translation model during decoding. Additionally, the reordering behavior of the entire sentence can be affected, as the decisions of each reordering move in the reordering model is dependent on the previous reordering decisions.

In this chapter, we outline a statistical framework for analyzing the impact of specific ASR error types on translation quality in a speech translation pipeline, using a representative sample of ASR systems trained for speech recognition on lectures. Our

approach is based on *linear mixed-effects regression models* (Searle, 1973), which we use to take into account the variability of ASR systems and the difficulty of each speech utterance being translated in a specific experimental setting, while holding the particular SMT system fixed. We additionally take a second look at the Word Error Rate (WER) metric and its subsequent Levenshtein alignment of the words belonging to the reference and hypothesis transcripts and demonstrate that the alignment heuristics of the conventional alignment algorithms in WER can introduce variance that can skew the results of our analysis. Instead, we propose a modification to the alignment algorithm that leverages sub-word information to improve the alignment accuracy which enables greater insight into the error types that impact downstream natural language processing tasks such as machine translation.

We focus again on the translation of TED talks in order to track the impact of speech recognition errors on spoken language translation quality. After we introduce our statistical analysis framework, which is based on our reports in Ruiz and Federico (2014a, 2015), we provide a survey of previous findings in error analysis for machine translation and summarize the research performed in a related field: quality estimation for ASR and MT.

4.1 Experimental setup

As mentioned earlier, the individual components of a SLT system are trained and evaluated independently against local optimization metrics that fit each statistical model to its local task, but they do not generalize to overall SLT quality. Our goal is to analyze the impact of ASR errors on machine translation quality.

Word Error Rate (WER) (2.2) is the de-facto evaluation metric for ASR, which categorizes ASR errors as insertions, substitutions, or deletions corresponding to the Levenshtein distance alignments between a hypothesis and its reference. Using WER as a metric for ASR quality, how do errors in recognizing speech utterances affect the accuracy of a machine translation system that assumes that each source sentence is clean and well-formed?

We perform our experiments on an intersection of the ASR and MT results of the IWSLT 2013 evaluation campaign (Cettolo et al., 2013), which focused on the translation of TED talks. We collect each submitter's English ASR hypotheses on the tst_{2012} test set and take the subset of the ASR hypotheses that correspond to the reference set for the English-French MT track. A subset of the MT outputs of each system in the

MT track was manually post-edited by professionals. These post-editions served as multiple human references for automatic evaluation. Using these post-edited translations, we construct 3-way data consisting of eight English ASR hypotheses for 580 utterances, a single unpunctuated reference transcript from the ASR track, and the human post-edited translations from the English-French MT track.

We will use Translation Edit Rate (TER) (Snover et al., 2006) as a sentence-level MT quality metric, as it models the original post-editing scenario of the evaluation campaign by estimating the amount of effort required to correct machine translation output. In order to analyze the impact of ASR errors on MT quality, we construct experiments to address the following questions:

- Does ASR's WER correlate with SMT's automatic quality metrics (e.g. TER)?
- Do higher WER scores cause a degradation in MT quality with respect to translations on perfectly recognized utterances (ΔTER)?
- Which types of ASR errors have the strongest impact on translation quality?

In Section 4.1.1, we discuss the preprocessing steps for each ASR hypothesis and in Section 4.1.2, we discuss how machine translation outputs are generated for each ASR hypothesis.

4.1.1 ASR data processing

IWSLT's ASR submissions are in lowercase, lack punctuation, and do not have embedded segmentation. We use the segmentation file provided in the SLT track to induce segmentation. After segmentation, we use the documentation provided in the IWSLT evaluation campaign to find and match each source transcript and ASR hypothesis with the tst_{2012} set from the MT track.

In the past, ASR evaluations such as DARPA Hub-4 (Pallett et al., 1998) and subsequent ASR evaluations such as NIST's Rich Transcription tasks (Garofolo et al., 2002) used an evolving normalization script to prevent penalization for minor orthographic variations such as multiple spellings (e.g. British vs. American English), compound words (e.g. "storyline" vs. "story line"), and contractions (e.g. "it's" vs. "it is"). Assuming that a phrase-based SMT system in the SLT pipeline is trained on ASR reference transcripts, orthographic variances in ASR outputs can result in out-of-vocabulary words or under-represented source language n-grams in the translation model, further degrading

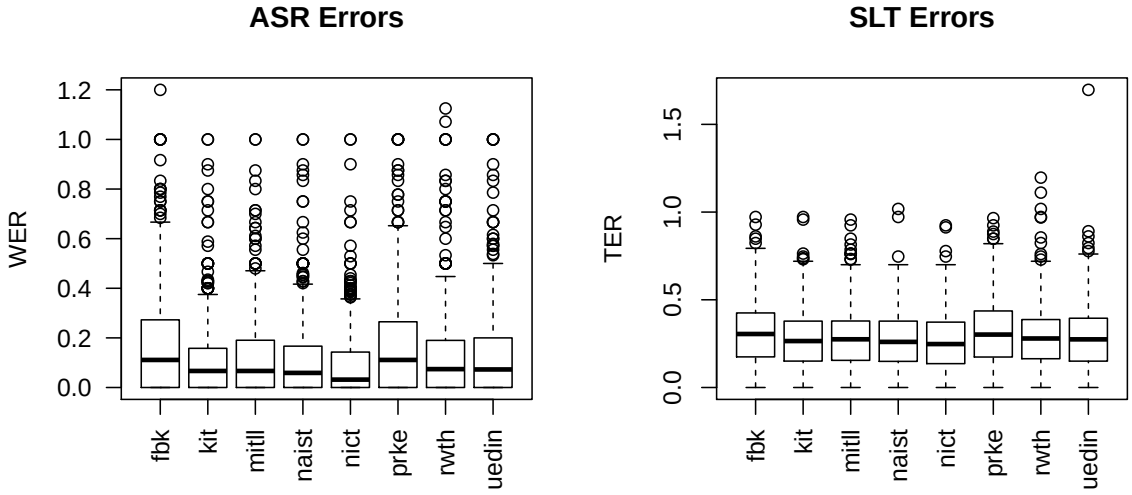


Figure 4.1: Boxplots describing the distribution of ASR errors (WER) and their impact on translation errors (TER) by ASR system and utterance. An extended analysis is provided in Appendix A.

machine translation quality. Although both the ASR hypotheses and the reference transcripts were normalized in prior evaluations, our experiments require the ASR reference to remain unmodified in order to properly evaluate the translation of ASR outputs against the translation of the original TED transcripts. Instead, we wrote a supervised word compounding script that splits or compounds words, depending on the word form in the reference transcript. Afterward we applied a bare-bones version of the normalization file provided by IWSLT which only maps British English to American English, since we observed anomalies including inconsistent mappings in the filters used for previous evaluations and as a second pass used a rule-based text normalization tool to correct other British English words. For extended details on the text normalization approach, see Appendix A.2.

We observed a $\pm 0.3\%$ absolute difference between our WER measurements after normalization and the scores reported in the official IWSLT evaluation task (Cettolo et al., 2013). The rankings of each system remained consistent. In Table 4.1 we report the performance of each ASR system, before and after orthographic normalization. Note that 5% of the errors for each system are attributed to normalization issues of compounding or word form (e.g. British English instead of American English). The majority of the errors are related to word compounding. The left-hand side of Fig. 4.1 shows a system-by-system comparison of ASR error distributions. Only a couple of ASR systems have significantly different error distributions from one another.

		ASR WER % ↓				MT TER % ↓	
ASR System	Norm	All	S	D	I	Post-edit	REF
fbk	none	21.4	13.3	2.9	5.2	33.70	54.68
	+COMP	16.8	10.8	3.0	3.0	32.84	54.10
	+NORM	16.5	10.5	3.1	2.9	32.71	54.09
kit	none	15.3	9.2	1.6	4.5	29.86	52.07
	COMP	10.4	6.6	1.7	2.1	28.83	51.40
	+NORM	10.1	6.3	1.7	2.1	28.73	51.40
mitll	none	16.4	9.6	2.0	4.8	30.13	52.17
	COMP	11.6	7.0	2.1	2.5	29.36	51.53
	+NORM	11.4	6.8	2.2	2.4	29.32	51.58
naist	none	15.7	9.1	2.2	4.4	29.86	51.88
	COMP	10.9	6.5	2.3	2.0	28.94	51.31
	+NORM	10.6	6.3	2.3	2.0	28.82	51.28
nict	none	14.5	8.7	1.4	4.4	28.92	51.43
	COMP	9.5	6.0	1.5	2.0	27.94	50.75
	+NORM	9.2	5.8	1.5	1.9	27.84	50.77
prke	none	21.3	13.2	2.8	5.3	33.79	54.83
	COMP	16.9	10.8	2.9	3.1	33.09	54.42
	+NORM	16.6	10.6	2.9	3.1	33.01	54.42
rwth	none	16.5	10.1	1.7	4.7	30.93	52.66
	COMP	11.9	7.7	1.8	2.4	29.93	52.08
	+NORM	11.7	7.5	1.8	2.4	29.84	52.06
uedin	none	17.2	10.2	2.1	4.9	30.84	52.66
	COMP	12.6	7.7	2.3	2.7	29.99	52.04
	+NORM	12.3	7.4	2.2	2.6	29.94	52.05
gold	none	0.0	0.0	0.0	0.0	21.27	46.46

Table 4.1: ASR outputs used as English-French MT evaluation input data on the human evaluation task of IWSLT 2013. ASR outputs are evaluated with no additional normalization, oracle word compounding (COMP), or compounding with word form normalization (NORM). Translated ASR outputs are tokenized and evaluated against the reference translation (Auto) and a combination of the human post-edited sentences from the MT task (Post-edit).

4.1.2 MT data processing

Since we are evaluating the impact of ASR errors on translation quality, we use a fixed SMT system trained on TED talk transcripts from the ASR track. We use FBK's primary phrase-based SMT system used in the English-French MT track (Bertoldi et al., 2013b). The normalized ASR hypotheses are post-processed by inserting punctuation and applying recasing. We insert the punctuation as closely as possible to the position

dictated in the reference in order to control the impact of punctuation on translation output. This is done by computing the Levenshtein alignments between the unpunctuated TED transcripts and each ASR hypothesis, using SCLITE[1]. We train and apply a recaser model using the standard Moses tools (Koehn et al., 2007) with IWSLT 2013's TED training data to all of the newly-punctuated ASR outputs.

After introducing punctuation and recasing the ASR output, we translate each ASR output and evaluate the results using TER over the seven human post-edited translations. Translation results are contrasted with FBK's primary MT submission on the bottom (gold) of the right-hand side of Table 4.1. We observe over a 6% absolute increase in TER for each of the translations of ASR hypotheses against the post-edited translation references. Most of the ASR transcripts' translations yield a TER score around 30% against the post-edited references. Turchi et al. (2013) empirically determine that translations with a TER score above 40% against a set of post-edited references require the translator to re-translate the source sentence from scratch, while lower scores imply that it is productive for the translator to post-edit the MT output. Likewise, our reported TER scores suggest that the translations of the ASR hypotheses are of good enough quality to be used in a post-editing scenario. The right-hand side of Fig. 4.1 shows a system-by-system comparison of SLT error distributions, measured in TER. In particular, we observe less variance among ASR systems as their hypotheses are translated by the SMT system. Futher details on data preparation and outlier removal, as well as measurements using BLEU and METEOR, are found in Appendices A.2 and A.3.

4.2 Phonetically-Oriented Word Alignment

Before we begin our study of the impact of ASR errors on SLT quality, we perform an analysis on the ASR errors themselves. Let $\mathbf{r} = r_1, r_2, ..., r_m$ and $\mathbf{h} = h_1, h_2, ..., h_n$ be the reference and hypothesis strings for an ASR example. The Levenshtein alignment $\mathbf{a}$ between $\mathbf{r}$ and $\mathbf{h}$ is $\mathbf{a} = a_1, a_2, ..., a_l$, where $\max(n,m) \leq l \leq n+m$. We define a *substitution error span*, SS, as a maximal contiguous sequence of two or more alignment errors

[1] http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

WER	POWER
i set up my studio in the red-light district and i set up a studio in the red light district and S I S	*i set up my studio in the red-light district and* i set up a studio in the "red light" district and S SS (1:2)
obsessively wrapped myself obsessive the rat myself I S S	*obsessively wrapped myself* "obsessive the" rat myself SS (1:2) S
traditional way of learning human anatomy traditional way of loaning human and that to me S I I I S	*traditional way of learning human anatomy* traditional way of loaning human "and that to me" S SS (1:4)
we developed with a Dr. Brown in Stanford we developed with doctor brahmin stamp or S S S S D	*we developed with a Dr. "Brown in" Stanford* we developed with doctor brahmin "stamp or" D S SS (2:1) SS (1:2)
the majority of anatomic classes taught promote unity obama panic class thought S S S S S S	*"the majority" "of anatomic" classes taught* "promote unity" "obama panic" class thought SS (2:2) SS (2:2) S S

Figure 4.2: Error alignment differences between the reference (top) and hypothesis (bottom) for WER and POWER. *S*ubstitution, *D*eletion, and *I*nsertion errors are annotated, as well as Substitution Spans (*SS*) with the ratio of hypothesis to reference words for POWER. POWER aligns homophonic errors such as *anatomy → and that to me*, while WER rate only aligns single words (e.g. *anatomy→me*).

containing at least one substitution error and any other error type. In other words,

$$
\begin{aligned}
(4.1) \qquad \text{SS} \quad := \quad & a_{i:j} \mid 1 \le i < j \le a_l \\
& \text{and} \quad \exists x \in (a_i, a_j) \mid x = \text{“Substitution”} \\
& \text{and} \quad \exists y \in (a_i, a_j) \mid y \in \{\text{“Insertion”}, \text{“Deletion”}\} \\
& \text{and} \quad \forall z \in (a_i, a_j), z \neq \text{“Correct”} \\
& \text{and} \quad a_{i-1} = \text{“Correct” if } i > 1 \\
& \text{and} \quad a_{j+1} = \text{“Correct” if } j < l.
\end{aligned}
$$

If we analyze the SCLITE implementation of Levenshtein alignments, we observe that the quality of the alignments is highly dependent on the alignment heuristic used for substitution error spans.

As shown by several alignment examples on the left-hand side of Fig. 4.2, suboptimal word alignments may be produced in adjacent error spans that consist of one or more substitution errors. Although there may be multiple alignment paths that yield the same edit distance score, the SCLITE implementation of the Levenshtein alignment algorithm aligns a substitution error span containing k substitution errors by assigning the k right-most words the substitution error labels. The remaining alignment positions on the left-hand side are annotated as deletions or insertions.[2] For example, in the first utterance in Fig. 4.2, the hypothesis word "obsessive" within alignment span $a_{12:14}$ could

[2]While we highlight the alignment heuristics for error spans in SCLITE, many Levenshtein alignment

have been aligned to the adverb "obsessively" as a substitution error given that it is the root form of the reference word. Instead, the alignment heuristic aligned a determiner to an adverb, instead of the correct noun ⇒ adverb alignment. For the purposes of computing an error rate score, this has no bearing; however, the misalignments themselves lead one to believe that a noun was improperly inserted during ASR decoding.

We also observe ambiguous alignments such as in the second example of Fig. 4.2, where "me" aligns with "anatomy". It is not clear how to best align "anatomy" to a single word in the error span. However, if we read out the utterance aloud, we observe that "anatomy" approximately covers the entire phonetic sequence of "and that to me". Fig. 4.3 computationally demonstrates this behavior via a Levenshtein alignment on the phrases' phonemes. The frequent observation of this behavior leads us to consider a third alignment error type that is distinct from the error types used in WER: the *phonetic substitution span*. Phonetic substitution spans are a special case of the alignment spans defined in (4.2) which consist of one-to-many or many-to-many error alignments that are best described as phonetic misrecognitions of the audio. They yield three additional substitution error cases, with examples from the right-hand side of Fig. 4.2:

- Hypothesis span ⇒ Reference word (e.g. "*and that to me*" ⇒ *anatomy*)
- Hypothesis word ⇒ Reference span (e.g. *brahmin* ⇒ "*Brown in*")
- Hypothesis span ⇒ Reference span (e.g. "*Obama panic*" ⇒ "*of anatomic*")

In the sections below we outline our Phonetically-Oriented Word Error Rate (POWER) alignment algorithm that adapts WER to allow for phonetic substitution error spans. We not only use POWER to align hypothesis words with respect to the reference, but we can also use it to adjudicate mismatches between the ASR hypothesis and reference. POWER uses phonetic transcriptions generated by the Festival TTS system trained with the CMU English pronunciation dictionary (Black and Taylor, 1997) to convert words into phonemes.

4.2.1 Alignment algorithm

Our phonetically-oriented word alignment algorithm is divided into two stages. First, we capture error spans whose error labels are likely to be ambiguous. The reference

implementations suffer similar deficiencies. Most authors do not treat this problem because they are only concerned about the quantity of errors and the aggregation of their score as opposed to the exact positions of the errors.

‖	#	*ax*	*n*			#		*ae*	*t*		#		*ax*	*m*		#		*iy*	‖
‖	#	ae	n	d	‖	#	dh	ae	t	‖	#	t	ax		‖	#	m	iy	‖
		S		I	I		I			I		I		D	I		I		

Figure 4.3: Phonetically-oriented alignment of *anatomy* to *and that to me*, with word (‖) and syllable (#) boundaries.

and hypothesis words in each span are transcribed into phonemes by a text-to-speech (TTS) analyzer. Each phoneme is treated as an independent token and word and syllable boundary tokens are introduced. The reference and hypothesis tokens are aligned using a variant of the Levenshtein alignment algorithm that introduces the following constraints:

1. Boundary tokens may not be substituted.

2. Vowel phonemes can only be aligned to other vowels (including r-colored vowels, but not semivowels).

3. Consonant phonemes can only be aligned to other consonants (including semivowels).

The boundary tokens provide an implicit distance constraint, penalizing adjacent phonemes within the same syllable when they are aligned far from one another.

In the second stage, we recombine the phonetic alignments into word alignments by performing a left-to-right scan of the alignment sequence. Substitution alignments are identified by considering the words covered by the aligned phonemes contained between two "correct"-aligned word boundary markers in the reference and hypothesis. Single word substitutions (S) are distinguished from phonetic substitution spans (SS) containing multiple words in the reference or the hypothesis. If a sequence of reference phonemes are terminated with a word boundary, but no hypothesis words have been scanned, the reference word is marked as a deletion (D). Likewise, a hypothesis word with no aligned reference word is marked as an insertion (I).

Returning to Fig. 4.2, the Levenshtein aligner used in WER could have alternatively aligned the reference word *anatomy* to any one of the hypothesis words currently marked as insertion errors. However, *anatomy* is pronounced similarly to the entire sequence of the four hypothesis words in the error span. The phonetically-oriented alignment in Fig. 4.3 captures this phenomenon by aligning the smallest word boundary closure across the entire span of reference and hypothesis words, thereby identifying

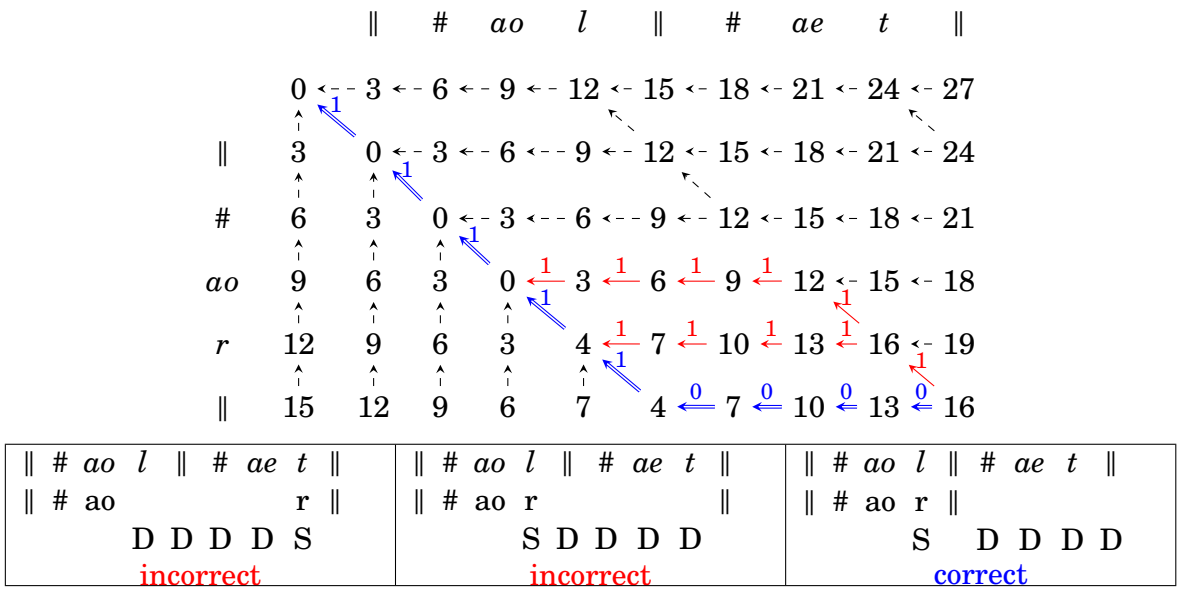


Figure 4.4: POWER alignments for *all at→or*. The Levenshtein backtrack matrix shows three alignments with the same optimal edit distance scores (16). The third and correct alignment (highlighted in the backtrack matrix) compactly aligns *all→or* with a span size of 5 symbols, while the others greedily align *or* to multiple reference words, with span sizes of 9 symbols.

anatomy⇒and that to me as a phonetic substitution span and yielding the alignment on the right-hand side of Fig. 4.2. Likewise, while WER may have considered slightly better word alignments like *Brown⇒brahmin* and *Stanford⇒or*, it is incapable of capturing relationships such as *Stanford⇒stamp or*.

4.2.2 Word alignment heuristics

While the phonetically-oriented alignments provide better phonetically-grounded alignments, its underlying Levenshtein alignment algorithm must also decide between multiple equally-weighted best paths.

In particular, for alignments with large differences in the number of reference and hypothesis syllables, our implementation tends to align the first and last word boundaries close to the beginning and end of the alignment sequence. For example, Fig. 4.4 shows three candidate alignments for the error span *all at⇒or* that minimize the edit distance. Two out of three alignments attempt to align *or* to the entire two-syllable reference. However, only *all* should align to *or* as a substitution, and *at* should be considered a deletion error. We resolve ambiguities like these by finding the alignment that

minimizes the number of alignment gaps between the first and last word boundaries in both the reference and hypothesis. In practice, we do this by encoding the best paths in the Levenshtein backtrack matrix into an edge-weighted graph and use Dijkstra's algorithm to find the best path.

Since there still remains some noise in the phonetic alignments, we introduce a couple of heuristics to prevent the aligner from overzealously marking single-syllable words as members of a phonetic substitution span, when in reality they do not have a phonetic correspondence on the other side. When annotating a phonetic substitution span, we keep a record of the number of reference and hypothesis syllables. If there is an extra syllable in the reference or hypothesis, we check if it is the first syllable of a new word. If so, we mark this word as a deletion or insertion error, respectively.

4.2.3 Scoring

Our Phonetically-Oriented Word Error Rate (POWER) score is defined nearly identically to WER as:

$$\text{POWER} = \frac{S + D + I + SS}{L}, \tag{4.2}$$

$$SS = \sum_{span} \max(|span_{ref}|, |span_{hyp}|),$$

where L is the length of the reference and S, D, and I are the number of word-level substitution, deletion, and insertion labels, respectively. SS is the count of phonetic substitution spans, weighted by the maximum number of words in each span. These one-to-many or many-to-many word alignments indicate phonetic confusability as the cause of the error. In principle, WER and POWER provide scores that are virtually the same as one another. Although the phonetically-oriented alignment process may occasionally introduce insertion or deletion errors based on violations in the phonetic alignment rules described above, the resulting score is statistically similar to those of WER.

4.2.4 Error Analysis Comparison

Using the experimental setup from Section 4.1, we compare POWER to WER. by looking at the basic ASR error types (S, D, I, and SS), which implicitly contain no linguistic information. Fig. 4.5 shows the contribution of the basic Levenshtein error types toward the error rate score for each ASR system. According to WER, substitutions intuitively

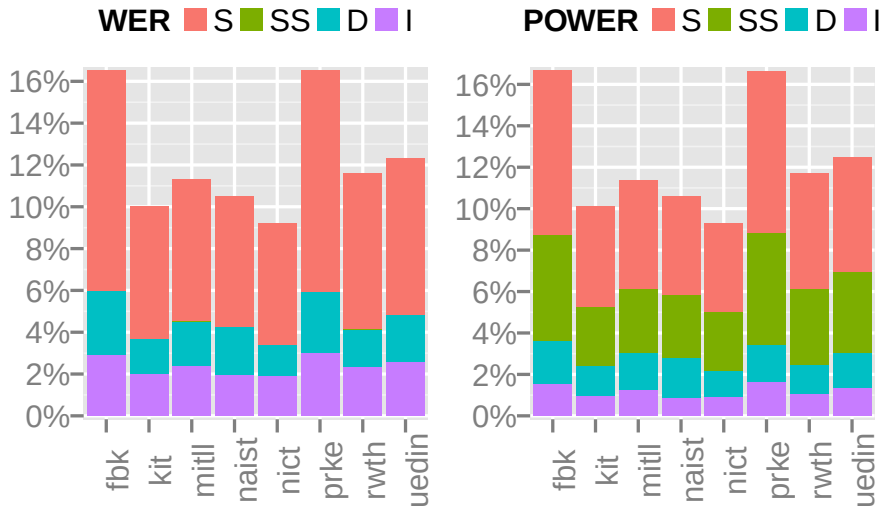


Figure 4.5: Distribution of error types for WER (left) and POWER (right) for each IWSLT 2013 ASR evaluation participant.

make up the majority of error types (62.3% ± 0.7%). Across all ASR systems, WER suggests that the number of deletions are slightly lower than the number of insertion errors (17.9% ± 0.7% deletions and 19.8% ± 0.5% insertions).

However, POWER suggests that roughly half of these alleged insertion errors (10.0%± 0.3%) are instances where a larger reference word is being hypothesized as a homophonic sequence of shorter words. Likewise, a portion of "deletion" errors are instances where multiple reference words were hypothesized as a longer homophonic word (4.1%± 0.5%). Since these phonetic substitution span errors are typically cases of one-to-many alignments, the number of reported word-level substitution errors are reduced. As such, POWER claims that 30.0%(±0.7%) of the errors are substitution spans involving homophony, leaving 13.8%(±0.8%) of the remaining errors as deletions and only 9.8%(±0.3%) as insertions whose pronunciations do not align to any words – both measures are substantially lower than those reported by WER. The remaining 46.4%(±0.5%) are word-level substitutions.

We can corroborate this by observing in Table 4.2 that, across all ASR systems, 70.4%(±2.5%) of the phonetic substitution spans involve multiple hypothesis words, while only 34.8%(±2.8%) contain multiple reference words. The first figure may be explained by the presence of out-of-vocabulary words in the ASR reference, as well as the effects of domain variation on the evaluation data. The alignment of multiple reference words to a single hypothesis word may be indicative of mispronunciations and/or

underarticulation by the speaker.

Table 4.3 provides confusion pair examples from FBK's ASR system output that demonstrate the utility of POWER. Word confusion pairs such as *a*⇒*today* are likely errors induced by an ASR language model that biases the acoustic model to artificially recognize non-existant phonemes. Likewise, POWER is able to provide insight that *crude*⇒*crudely* is not a morphological error, but rather another language model-induced bias that considers *leaf* an unlikely successor to *crude*. Other confusion pairs include word normalizations, affix errors, and phonetic confusions.

4.3 Do ASR errors correlate with SMT errors?

Having addressed the word error alignment issue in Section 4.2, we continue our assessment of ASR errors and their impact on translation quality. Using the WER and POWER metrics, how do ASR errors correlate with SMT errors? We split this question into two related inquiries. First, is there any relation between an ASR system's difficulty to recognize a speech utterance and the difficulty of translating the utterance, assuming it was recognized perfectly? The answer may seem obvious, since an ASR model could be trained poorly and generate hypotheses that have no bearing with their references. However, as described earlier, each of the ASR systems used in the IWSLT evaluation are capable of producing translations that can be efficiently post-edited by a professional translator. Second, do ASR errors correspond directly with translation quality? In other words, does the increase or decrease in WER correlate with the number of translation errors in the speech translation pipeline? We address these questions by

SysID	SS.ref	SS.hyp	SS: ref>1	SS: hyp>1	SS: ref>1 & hyp>1
fbk	0.036	0.040	0.450	0.615	0.065
kit	0.017	0.025	0.254	0.800	0.054
mitll	0.019	0.026	0.321	0.714	0.036
naist	0.019	0.026	0.336	0.715	0.051
nict	0.018	0.025	0.305	0.763	0.069
prke	0.039	0.042	0.488	0.585	0.073
rwth	0.023	0.032	0.310	0.737	0.047
uedin	0.025	0.032	0.317	0.700	0.017

Table 4.2: Left: Percentage of reference/hypothesis words appearing in a phonetic substitution span. Right: Percentage of phonetic substitution spans containing multiple reference words, multiple hypothesis words, or both.

WER		POWER	
Reference	Hypothesis	Reference	Hypothesis
a	today	a day	today
ascending	and	ascending	and sending
anesthetize	and	anesthetize	and decent size
butchering	the	butchering	maturing
centigrade	cents	centigrade	cents a great
crude	crudely	crude leaf	crudely
cyclones	soy	cyclones	soy clones
face-to-face	face	face-to-face	face to face
of	obama	of anatomic	obama panic

Table 4.3: Confusion pair examples using WER and POWER.

analyzing the correlation between the independent variable (WER) and the dependent variables (TER on translations of ASR references and ASR hypotheses, respectively) in Section 4.3.1, followed by constructing linear regression models to test for statistical significance in Section 4.3.2.

4.3.1 Correlation

We first measure the correlation between the WER scores of each ASR system and the TER acquired by translating each corresponding ASR reference. The Pearson correlation coefficient, r, measures the linear dependence between two variables. For our experiments, we control the effects of sentence length by binning the ASR hypotheses from each system into buckets corresponding to the quartiles of the reference length. Since much of the skewness of ASR errors shown in Fig. 4.1 is related to ASR reference length, we take correlation measurements on the 2nd and 3rd length quartiles, corresponding to reference lengths of 9-15 and 16-22. Using all ASR systems, we observe r values of 0.039 and 0.091, on the respective reference lengths, implying no correlation. Using only the observations of NICT's primary system (which had the lowest WER in the ASR evaluation track), we observe r values of -0.031 and 0.049, respectively.

We repeat the experiment, this time comparing ASR errors to their corresponding translation errors. Using all ASR systems, we observe r values of 0.672 and 0.632, respectively, implying strong correlation. We observe a similar trend when considering NICT's system alone. Again, these results are not surprising, since a machine translation system depends on the speech recognition output in order to generate a translation. It is important to note that while there is naturally a strong correlation between ASR

outputs and the quality of their translations, translation quality is not solely dependent on ASR quality. The missing 30% includes phenomena related to the problem of transferring content from the source language (English) to the target language (French), which take into consideration the lexical, syntactic, and semantic properties of each language (Vilar et al., 2006; He et al., 2011b; Ruiz and Federico, 2014b).

4.3.2 Linear Regression

To verify whether the correlation results in the previous section imply dependence, we fit univariate linear regression models using a single ASR system to evaluate the contribution of WER to the corresponding translation's TER score. We focus our attention on the observations of NICT's primary system. The response variable is the TER score computed against seven post-edited translation references. TER is computed either on the ASR references or on the translations of NICT's ASR hypotheses. Again, WER is computed on the uncased, unpunctuated output of the ASR system. Translations are performed using FBK's primary MT submission.

Our model treats the TER of the translated ASR hypotheses as the response variable. WER significantly predicts TER scores, $\beta = 0.696, t(578) = 18.42, p < 10^{-4}$ and explains a significant proportion of variance in TER scores ($r^2 = 0.369, F(1,225) = 339.4, p < 10^{-4}$). However, much of the variance remains unexplained by the model. Without accounting for the reference transcript's utterance length, WER cannot intrinsically anticipate the difficulty of translating the utterance, since the length of the input affects the search space and the hypothesis pruning decisions made by the decoder. As evidence, we sample two utterances recognized by NICT's ASR system, both with WER scores of 20% but having a different number of words in the reference (5 and 25, respectively). The TER scores of their translations are 46.7% and 28.4%, respectively. WER also assumes that each error contributes independently towards the error metric and thus does not measure interactions between multiple errors in an utterance. In phrase-based SMT, the position and density of ASR errors can hinder the translation model's ability to select proper target phrases, as well as affect the reordering model's ability to properly arrange the phrases in the target language.

4.4 WER scores and translation quality

Our previous experiments in Section 4.3.1 measured the relationship between WER of ASR hypotheses and TER. While WER is a significant predictor of TER in our simple regression model, it fails to capture the variance in TER associated with the innate difficulty of translating the utterance. WER cannot make predictions about translation difficulty for perfectly recognized utterances; instead it can be used to estimate the relative increase of SLT errors caused by ASR errors. Thus, we minimize the bias of the intrinsic translation difficulty by measuring the difference between the TER associated with translating the perfect ASR reference and the TER associated with translating the ASR hypothesis, labeled as ΔTER:

$$\Delta\text{TER} = \text{TER}_{\text{ASR}} - \text{TER}_{\text{gold}}, \tag{4.3}$$

where TER_{gold} is the TER score for a perfectly recognized utterance, and TER_{ASR} is the TER score on the translation of the ASR hypothesis. By using ΔTER, we assume that TER_{gold} is the upper-bound on translation quality with the given SMT system. In other words, we assume that a SMT system cannot translate transcripts containing errors better than clean transcripts. We check this assumption in our observation data and note 64 violations out of a total of 4,640 observations covering the outputs of the eight ASR systems (1.4% of the time). As a sanity check, we had two native French speakers evaluate the translation quality of several scenarios where $\Delta\text{TER} < -0.1$. In all cases, the native speakers preferred the MT outputs of translated ASR references over the translations of ASR hypotheses. These violations are likely due to the greedy alignment heuristics used by the TER algorithm to accommodate reordering shifts in the Levenshtein alignment (Snover et al., 2006). Extended details of our outlier removal process are outlined in Appendix A.3.

We first measure the correlation between WER and ΔTER using Pearson's r. Following the same approach as Section 4.3.1, we observe strong correlations on the observations with reference lengths in the middle 50% length quartiles: 0.780 and 0.756 using all ASR systems for utterance lengths of 9-15 and 16-22, respectively, and scores of 0.786 and 0.778 using only NICT's ASR system.

We next verify ΔTER's dependence on WER using *linear mixed-effects models*, which have been effectively used on linguistic data (Baayen et al., 2008). Mixed-effects models allow us to take into consideration random effects caused by an ASR system and the particular features of each ASR utterance. For a sample of n observations with p fixed

	All ASR		NICT+FBK	
Fixed effects	β	Std. Error	β	Std. Error
(Intercept)	8.72e-03	3.14e-03 ∘	1.01e-02	3.93e-03 ∘
WER	6.30e-01	8.55e-03 •	6.16e-01	1.46e-02 •
Random effects	Variance	Std. Dev.	Variance	Std. Dev.
UttID (Inter)	4.50e-03	0.067	3.89e-03	0.062
SysID (Inter)	0.000	0.000	2.57e-06	0.002
Residual	3.74e-03	0.061	3.99e-03	0.063

Table 4.4: Fixed and random effects for the null model (WER-only), which measures the effect of WER on ΔTER for English-French SLT. The model is constructed with observations from all ASR systems in IWSLT 2013's ASR Track on the left-hand side and only NICT and FBK's ASR systems on the right. Fixed effects coefficients (β) and standard errors are reported. Random intercepts account for variances by utterance (*UttID*) and ASR system (*SysID*). Statistical significance at $p < 10^{-4}$ is marked with • and $p < 10^{-2}$ is marked with ∘.

effects and q grouping variables, the regression equation is of the following form:

$$y = \mathbf{X}\beta + \mathbf{Z}b + \epsilon, \tag{4.4}$$

where $\mathbf{X}$ is a $n \times p$ fixed effects design matrix, $\mathbf{Z}$ is a $n \times q$ random effects design matrix, β and b are the p fixed and q random effects coefficient vectors, respectively, and ϵ is a $n \times 1$ observation error vector. We analyze the relationship between WER as an independent variable and ΔTER as the response variable. We provide random intercepts for the utterance (labeled as *UttID*) and ASR system (labeled as *SysID*), reflecting the 580 speech utterances transcribed by eight ASR systems. An illustration is provided in Appendix A.1. In total, the model is trained on 4,640 observations. The models are fit by maximum likelihood, using the R (R Core Team, 2013) implementation of linear mixed-effects models in the *lme4* library (Bates et al., 2014). Fixed effect coefficients and random effects variance for the WER-only model are reported in Table 4.4.[3]

Both WER and the intercept are observed as statistically significant. The coefficients suggest that if there are no ASR errors, TER will increase by 0.87%. However, for each percentage point of WER, the TER will further increase by roughly $0.63 \times 0.01 = 0.0063$ (0.63%). We observe a r^2 value of 0.840 for the model, 0.154 of which is attributed to the fixed effects.

As a random effect, *SysID* was not significant, as it has a standard deviation near zero. This behavior is also evident in the boxplots of Fig. 4.1, implying that the differ-

[3]Note that the WER and TER values in Table 4.1 are listed as percentages, while our regression models express the values between 0 and 1.

ences between the emitted WER scores and translation TER scores for each ASR system are not significantly different from one another. In order to verify that the random intercept associated with the ASR system is indeed insignificant, we repeat the mixed-effects analysis, using two systems with significantly different WER scores; namely NICT and FBK. Statistics on the fixed and random effects are also listed in Table 4.4. We again observe near-zero variance for *SysID* and do not observe significant differences in the fixed effects coefficients, implying that the *SysID* random effect has no impact on the model.

4.5 ASR Levenshtein error types and translation quality

Now that we have verified that an increase in WER significantly increases TER, are there significant differences between the effects of individual ASR error types on translation quality? We hypothesize that not all ASR errors are treated equally when ASR hypotheses are used in the speech translation pipeline. To demonstrate this, we construct new mixed-effects models which factorize the WER and POWER metrics into the components used to compute its score. WER is factorized into three independent variables, corresponding to the number of occurrences of each error type, normalized by the reference length, according to (2.2), while POWER contains the extra phonetic substitution span component. We continue to use the utterance ID and the ASR system ID as random effects. Our null hypothesis states that all length-normalized ASR error types (S,D,I, as well as SS for POWER) contribute equally to ΔTER, which is the same as our WER-only model specification in Section 4.4.

4.5.1 Basic error types

We begin by looking at the basic ASR error types (S, D, I, and SS), which implicitly contain no linguistic information. We use the counts for S, D, I, and SS errors, normalized by the reference length L as fixed effects and maintain the same random effects as the WER-only model. To simplify the notation in our models, labeled WER_{basic} and $POWER_{\text{basic}}$, we refer to the length-normalized error types in shorthand form (e.g. *WER.S*, *WER.SS*).

The coefficients of the fixed effects of the fitted models (2) and (3) are shown in Table 4.5. We observe a significant difference between WER_{basic} and the baseline, rejecting

	WER	WER_{basic}	$POWER_{\text{basic}}$
	(1)	(2)	(3)
WER	0.630***		
	(0.586,0.674)		
WER.D		0.564***	0.615***
		(0.506,0.622)	(0.556,0.674)
WER.I		0.707***	0.829***
		(0.642,0.772)	(0.753,0.906)
WER.S		0.624***	0.649***
		(0.578,0.671)	(0.601,0.696)
WER.SS			0.535***
			(0.487,0.584)
Constant	0.001	0.001	−0.0001
	(-0.003,0.004)	(-0.002,0.004)	(-0.003,0.003)
Observations	4,640	4,640	4,640
Log Likelihood	6,172.170	6,180.631	6,194.288
Akaike Inf. Crit.	-12,330.340	-12,343.260	-12,368.580
Bayesian Inf. Crit.	-12,285.240	-12,285.280	-12,304.150
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4.5: Fixed effects coefficients and 95% confidence intervals for the first three mixed-effects models, which measure the effect of ASR error types on ΔTER for English-French SLT. The baseline encapsulates all error types in a single WER measure, while the subsequent models use WER and POWER-aligned error types.

the null hypothesis that each basic ASR error type contributes equally to translation quality, in terms of ΔTER ($\chi^2(2) = 16.922, p < 2.12 \times 10^{-4}$). We additionally observe a significant difference between the standard WER-aligned error types (WER_{basic}) and the POWER-aligned error types ($POWER_{\text{basic}}$) that include substitution spans ($\chi^2(1) = 27.314, p = 1.73 \times 10^{-7}$), indicating that substitution spans are a significant predictor of translation quality. As shown in Table 4.5, while the impact of substitution errors remains in principle the same, the impact of insertions increase sharply, both due to the higher quality of the error labels and their lower frequency. $POWER_{\text{basic}}$ indicates that an utterance with a WER (or equivalently, POWER) score of 10% as insertion errors would expect an increase in TER by $0.1 \times 0.829 - 0.0001 = 8.3\%$, while 10% in substitution errors would correspond to a TER increase of $0.1 \times 0.649 - 0.0001 = 6.49\%$.

4.5.2 Word classes and morphology

In Section 4.5.1, we verified that individual Levenshtein error types have different effects on translation quality, suggesting that the breakdown of WER into length- normalized Levenshtein alignment types better models the relationship between ASR errors and translation quality (in ΔTER). Yet, research literature has shown that particular linguistic classes of words are problematic, either for ASR or MT. In ASR, researchers such as **?** have identified function words (also known as closed class words) as problematic for speech recognition. Oftentimes a speaker may alter the pronunciation of high frequency function words, such as prepositions and articles, by underarticulating or dropping phonemes. While a human can predict these words with high accuracy, an ASR system relies on phoneme or triphone recognition as an intermediate step toward recognizing words. Content words (also known as open class words) are generally simpler to recognize, as they often contain more syllables and cover a larger amount of speaking time within an utterance. On the other hand, open class words might not be represented in a speech lexicon, rendering them impossible to be generated by an ASR system. Aside from the issue of out-of-vocabulary words, SMT systems have the opposite problem. Researchers such as Vilar et al. (2006) demonstrate that missing content words contribute more toward translation errors than missing function words.

Taking this into account, since we have already observed differences between Levenshtein error types, we now look at differences between how misrecognitions of open and closed class words affect translation outputs. We use TreeTagger (Schmid, 1994) to assign part-of-speech (POS) tags on the ASR references using the Penn Treebank (Marcus et al., 1993). Using the Levenshtein alignments between each ASR hypothesis and its reference, we annotate deletion and substitution errors with their POS tags. We do not annotate insertion errors, as an insertion error indicates that no reference word is available to tag. We manually map each POS tag associated with a substitution and deletion error to its class (open or closed), using the mapping rules outlined in Table A.3 in the Appendix.

We annotate each Levenshtein alignment error type by the word classes of the reference and hypothesis words, respectively, as follows:

$$S \Rightarrow S \circ \{\text{closed}, \text{open}\} \times \{\text{closed}, \text{open}\} \tag{4.5}$$

$$D \Rightarrow D \circ \{\text{closed}, \text{open}\} \times \{\varnothing\} \tag{4.6}$$

$$I \Rightarrow I \circ \{\varnothing\} \times \{\text{closed}, \text{open}\} \tag{4.7}$$

$$SS \Rightarrow SS \circ \{\text{closed}, \text{open}, \text{span}\} \times \{\text{closed}, \text{open}, \text{span}\}, \tag{4.8}$$

where "span" is short-hand for a phonetic substitution span. For example, the WER.S error type is factorized into four error types as a cross-product of {closed, open} × {closed, open}. The new mixed-effects models are labeled in Table A.6 as WER_{wc} and $POWER_{\text{wc}}$. Thus, the WER reference-hypothesis alignment of *anatomy→the* corresponds to a *S.open_closed* error type under (4.5), and the POWER alignment of *anatomy→"and that to me"* corresponds to a *SS.open_span* error under (4.8) in the second example of Table 4.2.

We now extend the mixed-effects models from Section 4.5.1 with fixed effects corresponding to our new annotations and use the same random effects. Statistics on the fixed effects as well as the 95% confidence intervals of their coefficients are shown on the right-hand side of Table A.6. Likelihood ratio tests between $POWER_{\text{wc}}$ and $POWER_{\text{basic}}$ indicate that the Levenshtein error types grouped by word class better measure the impact of ASR errors on translation quality ($\chi^2(10) = 120.58, p = 2.20 \times 10^{-16}$).[4] Our results confirm that all word class-specific ASR error types are significant at the $p < 10^{-4}$ level, with insertions of open class words having the highest detrimental impact on translation scores (95% confidence interval of [0.92,1.15]).

4.6 Discussion

The results of our analysis demonstrate that particular ASR error types more problematic in spoken language translation than others. For example, Section 4.5.2 demonstrates that, all other factors held constant, a standard phrase-based machine translation system is apparently more tolerant of ASR deletion errors on function words than towards substitution errors on function words. This is most commonly due to cases where a function word is recognized as another function word from a different lexical category: for example, a preposition misrecognized as a determiner.

Although the fixed effects coefficients for our mixed-effects models reported in Table A.6 show the expected increase in TER for each percentage of WER associated with a particular error type, an error type with a high coefficient but a low frequency may not be important from an error correction standpoint. Ideally, we wish to measure which ASR errors are particularly problematic for a given SLT task – in this case, TED talks in the IWSLT evaluation. Considering the fixed and random effect scores on each utterance, we measure the average weighted contribution of each ASR error type toward the

[4] Likelihood ratio tests between WER_{wc} and WER_{basic} also indicate that word class-annotated error types better describe SLT quality ($\chi^2(5) = 75.033, p = 9.16 \times 10^{-15}$).

All utterances				Error present in utterance			
ErrorType	coef	$\mu_{\Delta\text{TER}}$	Rank	ErrorType	coef	$\mu_{\Delta\text{TER}}$	Rank
WER.S.open_open	0.590	0.0175	1	WER.SS.closed_span	0.713	0.102	1
WER.S.closed_closed	0.757	0.0132	2	WER.SS.span_span	0.687	0.100	2
WER.SS.open_span	0.546	0.0123	3	WER.SS.open_span	0.546	0.085	3
WER.D.closed	0.548	0.0069	4	WER.SS.span_closed	0.553	0.084	4
WER.I.closed	0.723	0.0059	6	WER.I.open	1.036	0.077	5
WER.S.open_closed	0.585	0.0057	6	WER.D.open	0.663	0.071	6
WER.D.open	0.663	0.0048	7	WER.SS.span_open	0.451	0.064	7
WER.I.open	1.036	0.0044	9	WER.I.closed	0.723	0.058	8
WER.S.closed_open	0.802	0.0036	9	WER.S.closed_closed	0.757	0.058	8
WER.SS.span_open	0.451	0.0038	9	WER.S.open_open	0.590	0.056	10
WER.SS.span_closed	0.553	0.0016	12	WER.S.closed_open	0.802	0.051	11
WER.SS.span_span	0.687	0.0015	12	WER.S.open_closed	0.585	0.047	12
WER.SS.closed_span	0.713	0.0011	13	WER.D.closed	0.548	0.044	13

Table 4.6: Ranking of word class-annotated POWER ASR error types by their mean frequency-weighted contributions toward translation ΔTER ($\mu_{\Delta\text{TER}}$) in the tst_{2012} evaluation sample. Error types are ranked according to their contribution to machine translation errors in the TED talk translation task. Left: Frequency-weighted scores across all utterances. Right: Frequency-weighted scores only for utterances containing the error type.

ΔTER measure. In other words, if we observe one ASR error of a particular type, how much is it expected to degrade the translation quality?

Table 4.6 reports the mean increase in TER for each weighted error type, using the word class-annotated error types provided by POWER. The left-hand side computes the scores and rank across the entire tst_{2012} evaluation sample, while the right-hand side focuses only on the utterances which contain a given error type. Naturally, the ASR error types with the highest frequencies have the greatest contribution to overall translation quality. As shown in the ASR error frequency rankings in Table A.4 (see the Appendix), within-class substitution errors have the highest frequency-weighted contribution toward ΔTER.

Our analysis ranks phonetic substitution span (SS) errors as the leading contributor to high TER scores. Although only SS errors on open class reference words occur frequently, the presence of any SS error type has a serious effect on translation quality. The translation of Utterance #1 in Fig. 4.6 demonstrates the effects of misrecognizing *anatomy* as *"and that to me"*, which renders the translation unintelligible.

In a similar vein, Utterances #57 and #71 demonstrate the impact of open class word-level substitution errors. #57 misrecognizes a common noun as a proper noun; although

1	ASR_{Ref}	*cadaver dissection is the traditional way of learning human*
	ASR_{Hyp}	cadaver dissection and ease the traditional way of loaning human
		I:CC S:VBZ→VV S:VVG→VVG
		anatomy
		"and that to me"
		SS:NN→span
	MT_{Ref}	cadaver dissection est la manière traditionnelle de l' apprentissage humain de l' anatomie
	MT_{Hyp}	cadaver dissection et facilité la manière traditionnelle de **prêt** humaine **et qui pour moi**
	$Trans_{Ref}$	la dissection de cadavres est la manière traditionnelle d' apprendre l' anatomie humaine
16	ASR_{Ref}	*Maybe I can cut there see the brain and I can change my cut*
	ASR_{Hyp}	maybe i can cut theirs is the brain can change my cut
		S:RB→PP S:VB→VBZ D:CC D:PP
	MT_{Ref}	Peut-être que je peux coupé là , vous voyez le cerveau , et je peux changer ma couper .
	MT_{Hyp}	Peut-être que je peux **couper leurs** , **c' est** le cerveau , **peuvent** changer ma couper .
	$Trans_{Ref}$	Peut-être que je peux couper là , voir le cerveau , et je peux changer ma coupe .
57	ASR_{Ref}	*Your mic wasn't off during sound check*
	ASR_{Hyp}	your mike was a knock during sound checks
		S:NN→NP S:VBD→VBD I:DT S:RB→NN S:JJ→NNS
	MT_{Ref}	" Votre micro n' était pas de pendant son vérifier .
	MT_{Hyp}	" votre **Mike était frappe** pendant son **équilibre** .
	$Trans_{Ref}$	" Votre micro n' était pas éteint lors du contrôle de son .
71	ASR_{Ref}	*I call myself a body architect*
	ASR_{Hyp}	i call myself a bloody architects
		S:NN→JJ S:NN→NNS
	MT_{Ref}	J' appelle moi-même un corps architecte .
	MT_{Hyp}	J' appelle **un foutu architectes** .
	$Trans_{Ref}$	Je me considère comme un architecte du corps .
84	ASR_{Ref}	*A maybe that could take the form of a gas or a liquid*
	ASR_{Hyp}	maybe that could take the form of like a gas or liquid
		D:DT I:IN D:DT
	MT_{Ref}	Un peut-être qui pourrait prendre la forme d' un liquide ou un gaz .
	MT_{Hyp}	Peut-être que pourrait prendre la forme **de** liquide ou un gaz .
	$Trans_{Ref}$	Un peut-être qui pourrait prendre la forme d' un gaz ou d' un liquide .

Figure 4.6: Effects of FBK's ASR errors, automatically annotated with POS tags, on machine translation output.

in this case, a text-to-speech synthesis system may pronounce it correctly to a listener, the translation is further corrupted from the loss of the negation in *wasn't→was*. #71 misrecognizes a common noun as an adverb.

The deletion of closed class words happens often in our ASR samples, but their individual impacts are usually small, with the exception of deleting pronouns. In Utterance #16 from Fig. 4.6, closed-class deletion of the personal pronoun *I* causes the auxillary verb *can* to incorrectly attach *brain* as its subject before translating the utterance, while the multiple deletions of the determiner *a* in Utterance #84 have negligible quality degradation. Covering 16.2%(±0.6%) of all the ASR errors in our experiment, we consider insertion and deletion errors on closed class words to be low-hanging fruit for

correction. Although closed class words can be under-articulated in ASR, the number of alternative words are small and an ASR system's language model would more likely have sufficient statistics to recover them in alternative ASR hypotheses from the lattice.

4.7 Related Work

Goldwater et al. (2010) used mixed-effects regression models to analyze the effects of each word's lexical and prosodic features on each individual word's contribution to WER (IWER). Evaluating with two English ASR systems for conversational telephone speech they reported that acoustically similar words with similar language model probabilities have strong effect on the contribution of individual words on WER. This is similar to the analysis outcome we discovered in our analysis through the use of phonetic substitution error spans within our POWER metric. Additionally, while measuring the lexical and prosodic features in a similar fashion as Goldwater et al. would be useful, it would require highly accurate alignments between the words in an ASR hypothesis and the translated outputs. Due to the interdependency between words during translation, it is not possible to decompose the TER score on an individual word level to measure the effects of many of the lexical and prosodic features in Goldwater et al. (2010), which is our primary motivation for analyzing the sentence-level effects of ASR errors on translation quality.

Quality estimation is an active area of research that is related to our analysis, which has the primary goal of predicting the response variable used to evaluate the quality of ASR or MT. We briefly introduce each in the sections below.

4.7.1 Quality Estimation for Machine Translation

The goal of quality estimation in machine translation is to assign a quality score to machine translation outputs in the absence of a reference translation. This score could be a confidence score (Ueffing et al., 2003; Specia et al., 2009; Bach et al., 2011), a prediction of an automatic measure such as BLEU or TER, or it can predict a human-interpretable score, such as the reviewer's quality score (Soricut and Echihabi, 2010) or the time taken to post-edit a machine translation output (Specia, 2011). Quality estimation systems use machine learning techniques such as regression models (Albrecht and Hwa, 2007), neural networks (Buck, 2012), online and multitask learning (de Souza et al., 2014) to predict a quality score, given a combination of fluency and adequacy fea-

tures derived from the linguistic properties (Felice and Specia, 2012; Buck, 2012) of the source sentence and the MT system's translation hypotheses, and optionally features derived from the translation system's internal models. Our linear mixed-effects model approach, on the other hand, is used to measure the effects of ASR error types in an evaluation scenario where the quality of the translation is already known.

4.7.2 Quality Estimation for Automatic Speech Recognition

The idea of estimating ASR quality began with confidence estimation, which uses ASR-internal features, such as acoustic stability, hypothesis density, and likelihood and posterior scores to score each predicted word (Jiang, 2005). As an alternative, *quality estimation* (Negri et al., 2014) focuses on the prediction of ASR quality scores in the absence of reference transcripts. and internal ASR model features typically used for confidence estimation. Negri et al. (2014); Jalalvand et al. (2016) identify 72 features, comprising signal features, lexical features, language model features, and part-of-speech features that are relevant to the task and provide an evaluation tool. Jalalvand et al. (2015) additionally use sentence-level quality estimation to rank an n-best list of ASR hypotheses prior to combining them with system combination techniques such as ROVER.

4.8 Chapter Summary

In this chapter, we presented a statistical data analysis framework based on linear mixed-effects regression models to measure the impact of ASR errors on spoken translation language quality. We introduced a variant of WER, dubbed *POWER*, that captures and scores phonetic substitution error spans (i.e. errors involving multiple words that sound phonetically similar to its aligned counterpart) in addition to the standard substitution, deletion, and insertion errors from WER. By annotating ASR errors with POWER, we observed that the WER metric's error annotations lead one to believe that an ASR system generates more recognition errors on content word types than is actually the case. For example, substitution errors between content words and function words does not frequently happen. Instead these misalignments are often cases where a long reference word was misrecognized as a sequence of shorter hypothesis words that sound similar to the reference word.

As we applied the ASR errors annotated by POWER to our mixed-effects models, we confirmed that substitution words on similar word classes has the greatest impact

on SLT quality, followed by our newly annotated phonetic error spans, which capture error regions that were likely caused by conflicts between the acoustic model and the language model during ASR decoding. The frequency and severity of these error types inspires us to research error modeling strategies that attempt to model ASR errors caused by acoustic confusions directly in the machine translation system to increase error tolerance in SLT, which we will address in Chapter 6.

CHAPTER 5

CONTEXT ADAPTATION USING BILINGUAL LATENT SEMANTIC MODELS

Traditional statistical machine translation (SMT) systems operate on the assumption that each sentence to be translated is independent from one another. In the case of discourse, whether it is speech or the written word, this assumption seldom holds true. As a result, traditional machine translation systems rely only on local observations to translate a sentence and often provide suboptimal translations. Were the MT system able to use statistics derived from the context of neighboring sentences, the system could use semantically related words that better match the topical structure of the discourse, improving accuracy and cohesion. Likewise, in the spoken language translation scenario, context can aid in both the automatic speech recognition (ASR) and the MT components. For ASR the decoder may take into account the decisions made in previous utterances within a dialogue or discourse. MT in the spoken language translation pipeline can benefit both from the decisions made by the ASR decoder, as well as the previous translation decisions made by the decoder, potentially allowing the MT system to be more forgiving of ASR errors.

In this chapter, we discuss a simple, yet novel approach to incorporating context during machine translation decoding using an technique called Minimum Discrimination Information (MDI) adaptation, where we compute word counts estimated with bilingual topic modeling approaches. Our variant, dubbed "Lazy MDI", approximates MDI adaptation in a manner that is suitable for real-time translation scenarios by avoiding the

computation of the normalization term in conventional language model adaptation approaches that requires all n-grams to be re-estimated. This context modeling approach is motivated by language model adaptation techniques, but it expands into a general adaptation technique that fits well in the log-linear framework of MT architectures such as phrase-based machine translation. We observe that Lazy MDI performs comparably to classic MDI in topic adaptation for SMT, but possesses the desired scalability features for real-time adaptation of large-order n-gram LMs. We demonstrate its effectiveness on the translation of TED talks.

We begin the chapter by providing an overview on topic modeling in Section 5.1 as well as introducing techniques for language model adaptation in Section 5.2. We additionally review our work in Ruiz and Federico (2011) on language model adaptation via bilingual topic modeling and follow up with a demonstration on how to modify the adaptation framework as a log-linear feature, which we dub "Lazy MDI": an efficient approximation of MDI adaptation (Ruiz and Federico, 2012) (Section 5.3). We then provide experimental results on its use for model adaptation given windows of context in two paradigms: one where we take source language information from a sliding window, and the other where we use bilingual information by using look-ahead and look-behind context (Section 5.4), followed by a brief survey of relevant previous work in Section 5.5. In Section 5.6 we summarize our findings.

5.1 Topic adaptation

Topic adaptation is used as a technique to adapt language models based on small contexts of information that concentrate on the temporal focus of a discourse, rather than reflecting an entire domain or genre. In scenarios such as lecture translation, it is advantageous to perform language model adaptation on the fly to reflect topical changes in a discourse. In these scenarios, general purpose domain adaptation techniques fail to capture the local behavior of a discourse; while domain adaptation works well in modeling newspapers and government texts which contain a limited number of subtopics, the genres of lectures and speech may cover a virtually unbounded number of topics that change over time. Instead of general purpose adaptation, adaptation should be performed on smaller windows of context.

Most domain adaptation techniques require the re-estimation of an entire language model to leverage the use of out-of-domain corpora in the construction of robust models. While efficient algorithms exist for domain adaptation, they are in practice intended

to adapt language models globally over a new translation task. Topic adaptation, on the other hand, intends to adapt language models as relevant contextual information becomes available. For speech, the relevant contextual information may come in sub-minute intervals. Well-established and efficient techniques such as Minimum Discrimination Information adaptation (Della Pietra et al., 1992; Federico, 1999) are unable to perform topic adaptation in real-time scenarios for large order n-gram language models. In practice, new contextual information is likely to be available before techniques such as MDI have finished LM adaptation from earlier contexts. Thus spoken language translation systems are typically unable to use such as technique for the purpose of topic adaptation.

5.1.1 Topic modeling

Many topic adaptation approaches leverage *topic modeling* to estimate the topic distribution of documents. This is done with a generative process that assumes that each word w in a document d is generated by considering the distribution of *topics* z that are expressed in the document. Using Bayes' formula, this process is defined as:

$$P(w \mid d) = \sum_{z \in Z} P(w \mid z) P(z \mid d). \tag{5.1}$$

In the context of topic modeling, the latent topics $z \in Z = \{z_1, ..., z_k\}$ are class variables used to derive probabilistic distributions of words $w \in W = \{w_1, ..., w_m\}$ in a document $d \in D = \{d_1, ..., d_n\}$ with $k << n$.

One topic modeling approach is Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999), which is a statistical model built exactly on (5.1). Thus, the objective of PLSA is to learn $P(z \mid d)$ and $P(w|z)$ by maximizing the log-likelihood function:

$$L(W, D) = \sum_{d \in D} \sum_{w \in W} n(w, d) \log P(w \mid d), \tag{5.2}$$

where $n(w, d)$ is the term frequency of w in d. Using the Expectation Maximization (EM) algorithm (Dempster et al., 1977), the parameters $P(z|d)$ and $P(w|z)$ are estimated via an iterative process. A document-topic distribution $\hat{\theta}$ can be inferred on a new document d' via the decision rule:

$$\hat{\theta} = \arg\max_{\theta} \sum_{w} n(w, d') \log \sum_{z} P(w \mid z) \theta_{z,d'}, \tag{5.3}$$

$$\theta_{z,d'} = P(z \mid d').$$

Topic models can be actually applied to infer a bag-of-word distribution for the new document d'. In fact, the inferred $P(z \mid d')$ topic distribution can be used in conjunction with the $P(w \mid z)$ distribution learned during training to estimate a word unigram distribution. It is worth noticing that this method permits to infer full unigram distributions even from very small documents (Federico, 2002). Other topic modeling approaches are based on *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003), which adds a Dirichlet prior to the formulation in (5.1) that helps the model generalize to unseen documents.

5.1.2 Extending to bilingual contexts

In order to infer an adaptation bag-of-word model in the target language from a source text, several works have leveraged bilingual topic modeling approaches, including Tam et al. (2007); Zhao and Xing (2008); Ruiz and Federico (2011). While most bilingual topic modeling approaches tend to construct separate topic models for the source and target language and induce a mapping (Tam et al., 2007), Ruiz and Federico (2011) introduce a simplified form of bilingual topic modeling based on PLSA, which treats source and target bitext sentences $(\mathbf{f}, \mathbf{e})$ as "monolingual" documents with vocabulary $V_{FE} = V_F \cup V_E$. The underlying assumption is that the topics in a parallel text share the same semantic meanings across languages; thus, their vocabularies can be merged into a "super language". In order to ensure the uniqueness between word tokens between languages, we mark each token with the language it comes from. We perform PLSA training to estimate word-topic distributions $P(w|z), w \in V_F \cup V_E$. In normal inference scenarios, we would again use a bilingual sentence tuple to infer the distribution of words in the source and target languages. However, the target language is unobservable during MT decoding. Instead, bilingual PLSA is used to perform inference only from the tokens in the source language. Nevertheless, the model is capable of inferring a full unigram distribution for the vocabulary V_{FE}. In the context of language model adaptation for MT, the source words in V_F are pruned, and the remaining statistics on V_E are re-normalized.

5.2 MDI Adaptation

MDI adaptation was original designed by Della Pietra et al. (1992) as a means for domain adaptation of a generic background n-gram language models from a domain specific 1-gram model. In practice, MDI adaptation scales the probabilities of a background

language model, $P_B(h,w)$, by a factor determined by a ratio between the unigram statistics observed in an adaptation text A versus the same statistics observed in the background corpus B:

$$\alpha(w) = \left(\frac{\hat{P}_A(w)}{P_B(w)}\right)^{\gamma}, \quad 0 < \gamma \leq 1 \tag{5.4}$$

where γ is suitable smoothing factor. The adapted language model $P_A(h,w)$ is computed as:

$$P_A(h,w) = P_B(h,w)\alpha(w), \tag{5.5}$$

where h is the n-gram history of word w. As outlined in Federico (2002), the adapted language model can also be written recursively in an interpolated conditional form with discounted frequencies $f^*(w|h)$ and reserved probabilities for out-of-vocabulary words $\lambda(h)$:

$$P_A(w|h) = f_A^*(w|h) + \lambda_A(h)P_A(w|h'), \tag{5.6}$$

$$f_A^*(w|h) = \frac{f_B^*(w|h)\alpha(w)}{z(h)}, \tag{5.7}$$

$$\lambda_A(h) = \frac{\lambda_B(h)z(h')}{z(h)}, \tag{5.8}$$

with a normalization term

$$z(h) = \left(\sum_{w:N_B(h,w)>0} f_B^*(w|h)\alpha(w)\right) + \lambda_B(h)z(h'), \tag{5.9}$$

that efficiently computes the normalization term for high order n-grams recursively simply by summing over the observed n-grams in A. The recursion ends with the following initial values for the empty history ϵ:

$$z(\epsilon) = \textstyle\sum_w P_B(w)\alpha(w), \tag{5.10}$$

$$P_A(w|\epsilon) = P_B(w)\alpha(w)z(\epsilon)^{-1}. \tag{5.11}$$

While MDI has been applied for domain adaptation on both language models (Federico, 1999) and translation models (Tam et al., 2007), its re-estimation requires the computation of the normalization term outlined in (5.9) that makes it unsuitable for real-time ASR or spoken language translation. In topic adaptation scenarios, it is desirable to rapidly adapt a background language model using small adaptation contexts

consisting of few sentences. One method of inferring unigram statistics for MDI adaptation given sparse data is to perform bilingual topic modeling (Tam et al., 2007; Mimno et al., 2009; Ruiz and Federico, 2011). While it has been shown that the combination of topic modeling and MDI adaptation yield a significant improvement in translation adequacy, the approach of adapting non-overlapping contexts of size C requires M/C full LM re-estimations on a translation task with M sentences, with each re-estimation requiring the expensive computation of the normalization term.

5.3 Lazy MDI Alternative for SMT

The goal of MDI adaptation is to construct an adapted language model that minimizes its Kullback-Leibler divergence from the background LM, which is effectively performed via the unigram ratio scaling method described in (5.4) and (5.5). We seek to loosely approximate this KL divergence in statistical machine translation by adapting only n-grams that appear as translation options for a given sentence. As such, we seek to avoid computing the normalization term in (5.9) that requires observing the probabilities of all high- and lower-order n-grams in the LM. Since the ratio of unigram probabilities is defined across the range $[0,+\infty]$, we explore smoothing functions that bind the ratio to a finite range.

5.3.1 Smoothing unigram ratios

In machine learning, sigmoid activation functions are typically used to constrain functions in the range of $[0,a]$ or $[-a,a]$ to reduce the bias of a few data points within a training set. Likewise we explore the use of sigmoid functions to reward n-gram probabilities across the range of $[0,a]$. However, since we are scaling ratios in general, we desire the following properties of our smoothing function f:

$$\begin{aligned} f(0) &= 0; \quad \lim_{x\to+\infty} f(x) = a \\ f(1) &= 1; \quad \lim_{x\to-\infty} f(x) = -a \end{aligned}$$

In particular the $f(1) = 1$ constraint ensures that background LM probabilities remain fixed when the ratio is balanced. A *fast sigmoid* approximation (Georgiou, 1992) of the form:

$$f(x,a) = \frac{ax}{a+|x|-1}, \quad a > 1 \tag{5.12}$$

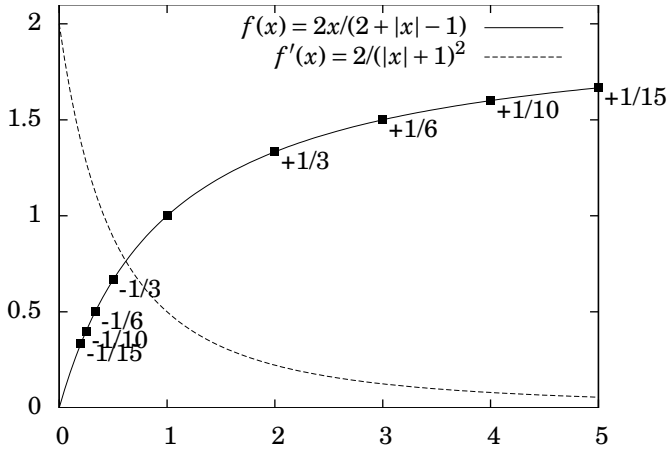


Figure 5.1: A plot of the transformed fast sigmoid function for positive ratios in (5.12) and its first derivative, evaluated at $a = 2$. The relative changes in $f(x)$ are labeled, centered at $f(1)$. The changes in $f(x)$ are symmetric with respect to each ratio and inverse ratio.

satisfies these properties. Additionally, its first-order derivative is symmetric with respect to inverted ratios, relative to the center at $x = 1$. Fig. 5.1 demonstrates this for $a = 2$, where a ratio of 2:1 (i.e. twice as many observations of a word x in an adaptation context window over its background statistics) yields a scale of $1 + \frac{1}{3}$, while a ratio of 1:2 yields a scale of $1 - \frac{1}{3}$.

5.3.2 Log-linear feature

Since we are no longer normalizing n-gram probabilities, we can consider the smoothed unigram probabilities as a function that rewards or penalizes translation options based on the likelihood that the words composing the target phrase should appear in the translation. We treat the smoothed unigram probabilities as a new feature in the discriminative log-linear model of the decoder. While our new feature is independent from any language model features, we can logically consider the adaptation of a background language model as a log-linear combination of the LM feature and the Lazy MDI feature as:

$$\hat{P}_{LM}(\mathbf{e}) = P_{LM}(\mathbf{e})^{\gamma_1} \cdot \prod_{i=1}^{l_e} \hat{\alpha}(e_i)^{\gamma_2}, \tag{5.13}$$

where $P_{LM}(\mathbf{e})$ computes the language model probabilities of target sentence $\mathbf{e}$; $\hat{\alpha}(e_i)$ is the Lazy MDI adaptation feature on the ith target word in $\mathbf{e}$, defined as:

$$\hat{\alpha}(e) = f\left(\frac{P_A(e)}{P_B(e)}\right). \tag{5.14}$$

By rearranging terms, we arrive at our unnormalized log-linear approximation of (5.5):

$$\hat{P}_{LM}(\mathbf{e}) = \prod_{i=1}^{l_e} P_{LM}(e_i \mid e_{i-2}, e_{i-1})^{\gamma_1} \cdot \hat{\alpha}(e_i)^{\gamma_2}. \tag{5.15}$$

In practice, only translation hypotheses suggested by the translation model are scored by the language model, thus limiting the number of unigram ratios to consider. Additionally, for computational efficiency, calculations are performed in log space. For $a = 2$, our fast sigmoid function can be rewritten as:

$$f(x,2) = 2 \cdot \left(1 + e^{-\ln(x)}\right)^{-1}, \quad x > 0, \tag{5.16}$$

which allows us to compute log probability ratios as $\ln P_A(w) - \ln P_B(w)$.

5.3.3 Sparsity considerations

If we treat the background and adaptation unigram statistics as unigram language models, we can use smoothing to reserve probability for out-of-vocabulary words. However, due to the sparsity of unigram features in adaptation texts, it is possible that the adapted unigram statistics are missing words that appear in the background LM. Assuming that there are insufficient adaptation statistics to reliably scale the probabilities of n-grams containing these words, we instead leave the background probabilities intact by fixing the unigram probability ratio to 1.

A similar problem can arise in the scenario that the adaptation text contains unigrams that are not observed in the background LM. One possible solution is to limit the vocabulary of the adaptation statistics to the same as that of the background.

5.3.4 Inferring unigrams via bilingual topic modeling

Since an adaptation text is in practice too small to directly compute reliable unigram statistics, we resort to the topic modeling approach described in Section 5.1.2 to infer full unigram probabilities. We then convert the word-document probabilities into pseudo-counts via a scaling function:

$$n(w \mid d) = \frac{P(w \mid d)}{\max_{w'} P(w' \mid d)} \cdot \Delta, \tag{5.17}$$

where Δ is a scaling factor to raise the probability ratios above 1. Since our goal is to generate a unigram language model on the target language for adaptation, we remove the source words generated in (5.17) prior to building the language model.

5.4 Experiments

We evaluate the utility of Lazy MDI in two scenarios. We first compare in Section 5.4.1 the performance of Lazy MDI against the original MDI system of Ruiz and Federico (2011) on the adaptation of a lowercased MT system with case-insensitive unigram statistics from both the adaptation text and the background text. Secondly, in Section 5.4.2, we evaluate the effects of context window size and the presence of previously translated segments on Lazy MDI adaptation.

Both experiments are conducted on the IWSLT 2012 TED English-French MT shared task (Federico et al., 2012). All machine translation systems are phrase-based and built upon the Moses open-source SMT toolkit (Koehn et al., 2007). In the classic MDI adaptation experiments the background LM is replaced by its MDI-adapted counterpart offline for SMT decoding, and the original LM log-linear feature weight is preserved. For Lazy MDI, adaptation is integrated into the Moses decoder as a feature function, using the same context unigrams as in MDI to compute on-the-fly adaptation ratios. MERT is used to tune each MT system on the dev_{2010} development set using context-adapted models. All models are evaluated using Multeval 0.3 (Clark et al., 2011).

5.4.1 Lazy MDI versus MDI adaptation

We first evaluate the performance of our fast sigmoid-smoothed Lazy MDI adaptation against both an unadapted baseline and the bilingual MDI adaptation approach described in Ruiz and Federico (2011). Our first baseline (FBK-TED) is a simple lowercased phrase-based SMT system based upon the Moses open-source SMT toolkit (Koehn et al., 2007)[1] and trained only on TED data.

MDI adaptation is applied using the unigram word ratios computed within context windows of length 5.[2] We experiment with adaptation via unigram ratios computed before and after smoothing with our transformed fast sigmoid function. Words not in the adaptation unigram LM are fixed with a 1:1 ratio to prevent their effect on the global translation hypothesis score. The evaluation results of each system are averaged over their three MERT optimizations.

In Table 5.1, we observe nearly identical MDI and smoothed Lazy MDI scores across each of the three MERT runs. The systems respectively yield statistically significant

[1] http://www.statmt.org/moses/

[2] The five-line contexts were determined by the segmentation provided by the IWSLT evaluators. In a traditional SLT scenario, we would need to determine the size of the context window.

System	Metric	Opt 1	Opt 2	Opt 3	Avg
FBK-TED	BLEU	27.64	28.20	28.20	28.0
MDI	BLEU	28.49	28.07	28.16	**28.2**
Lazy MDI (unsmoothed)	BLEU	27.14	*17.80*	28.40	24.4
	weight	0.1537	*0.4096*	0.0445	*0.3361*
Lazy MDI (smoothed)	BLEU	28.27	28.39	28.17	**28.3**
	weight	0.0132	0.0177	0.0138	0.0149

Table 5.1: Lowercased evaluation runs for the TED baseline and Lazy MDI adaptations for the IWSLT 2010 test set across three tuning instances. Unsmoothed Lazy MDI yields unstable adaptation feature weights across each run. "Opt 2" overpowers the log-linear model, yielding a drop in over 10 BLEU. "Opt 3" provides the best generalization to the test set by reducing the effects of the adaptation. For fast sigmoid-smoothed Lazy MDI, the adaptation weights remain consistent across all runs.

average improvements of 0.2 and 0.3 BLEU over the FBK-TED baseline. As predicted, unsmoothed Lazy MDI adaptation performs poorly as the unigram ratios between the background and context LMs often diverge greatly. If we observe the feature weights for the unsmoothed Lazy MDI model in Table 5.1, we see divergent values across each MERT instance, suggesting that the unbounded adaptation ratios yield unpredictable results during decoding. The smoothed Lazy MDI model, on the other hand, has relatively stable weights.

5.4.2 Context window size and bilingual context

Given the positive results in the experiment above, we evaluate the effects of the context-window size and the use of previously translated sentences on adaptation quality. Assuming that translation is occurring near real-time, the SMT system can consider sliding context windows that cover up to nine transcript lines before the current line, as well as optional future context that simulates latency in machine translation.

Our second baseline, (FBK-FULL) is FBK's primary phrase-based SMT system from IWSLT 2012, described as follows. A single phrase and reordering table were constructed using the fill-up technique (Bisazza et al., 2011) in a cascaded fashion in the order of TED, Giga French-English, and Europarl. The system consists of translation and reordering models trained from the in-domain TED[3] corpus, as well as out-of-domain Giga French-English[4] and Europarl v7 (Koehn, 2002) corpora. All out-of-domain par-

[3] https://wit3.fbk.eu/mt.php?release=2012-03-test

[4] 10^9 French-English data set provided by the WMT 2012 translation task (Callison-Burch et al., 2012).

Window		Bilingual		Monolingual	
Prev	Next	BLEU	NIST	BLEU	NIST
2	0	**33.27**	**7.562**	32.74	7.487
2	2	**33.18**	**7.550**	32.69	7.488
4	0	**33.19**	**7.559**	32.83	7.503
4	2	32.77	7.496	32.76	7.501
9	0	32.73	7.486	32.73	7.488
9	2	32.74	7.490	32.76	7.496
Unadapted		-	-	32.42	7.443

Table 5.2: Lazy MDI adaptation results on the IWSLT tst_{2010} English-French test set.

allel and monolingual corpora were domain-adapted by aggressive filtering using the cross-entropy difference scoring techniques described by Moore and Lewis (2010); Axelrod et al. (2011) on the French side and optimizing the perplexity against the (French) TED training data by incrementally adding sentences.

A domain-adapted 5-gram mixture language model was constructed with IRSTLM from the TED, Giga French-English, Gigaword French v2 AFP[5], and WMT News Commentary v7 corpora.

Table 5.2 provides evaluation results on tst_{2010} using adaptation contexts containing the current transcript line as well as a "look-behind" of 2, 4, or 9 previously translated transcript lines and optionally a "look-ahead" of two lines. In the bilingual modality, translated French hypotheses are provided only for the previous transcript lines. The monolingual contexts are treated as the baseline in the experiment.

Monolingual contexts Similar to the results reported by Ruiz and Federico (2012), we observe improvements of roughly +0.3 BLEU over the unadapted baseline for monolingual contexts varying from length 3 to 12. This improvement is over a baseline that was already domain-adapted toward the TED translation task. Each adapted system improves the translation scores by a similar amount, regardless of the orientation of the context window (look-behind only, versus look-behind with look-ahead).

Bilingual contexts As shown in Table 5.2, we observe significant improvements of around +0.4 BLEU over our monolingual adaptation experiments with small adaptation context windows of 3-5 sentences when including the translation hypotheses generated by the previous decoding decisions within the look-behind context window.

[5]http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2009T28

System	Text	TER
SRC	Gross has several companies, including one called eSolar that has some great solar thermal technologies.	*#771*
REF	Gross possède plusieurs sociétés, ... qui maîtrise quelques importantes **technologies thermiques solaires**.	
FBK-FULL	Bill Gross a plusieurs entreprises, dont un appelé eSolar qui a une grande **solaire thermique**.	0.500
MDI-M	Bill Gross a plusieurs entreprises, dont un appelé eSolar qui a un grand **technologies solaire thermique**.	(-0.06)
MDI-B	Bill Gross a plusieurs entreprises, dont un appelé eSolar qui a un grand **technologies thermiques solaire**.	(-0.11)
SRC	Now we can do it in a more precise way.	*#1497*
REF	Maintenant, **on peut** être plus précis .	
FBK-FULL	Maintenant, **nous pouvons** le faire d'une manière plus précise.	1.000
MDI-M	Maintenant, **nous pouvons** le faire d'une manière plus précise.	(-0.00)
MDI-B	Maintenant, **on peut** le faire d'une manière plus précise.	(-0.25)
SRC	... thousands of pink flamingos, a literal pink carpet for as far as you could see.	*#1018*
REF	... des milliers de flamants roses, **un véritable tapis rose**, s'étalant aussi loin que porte la vue.	
FBK-FULL	... des milliers de flamants roses, **une moquette rose** littéral pour autant que vous pouvez voir.	0.641
MDI-M	... des milliers de flamants roses, **une moquette rose** littéral pour autant que vous pouvez voir.	(-0.05)
MDI-B	... des milliers de flamants roses, **un tapis roses** littéral pour autant que l'on pouvait voir.	(-0.08)

Figure 5.2: Effects of bilingual Lazy MDI adaptation using the previous four sentences as context on the IWSLT 2010 English-French TED talk translation test set. REF refers to the reference translation, FBK-FULL refers to an unadapted baseline, MDI-M refers to a monolingual adaptation without translation hypothesis context, MDI-B performs adaptation with translation hypothesis contexts. The sentence-level TER scores are listed by each hypothesis and the difference is listed in parentheses by the reference.

Figure 5.2 shows examples of adaptation improvements using bilingual contexts with a look-back of 4, and reports sentence-level TER evaluation scores. The adaptation in sentence #771 focuses on the proper translation of "*solar thermal technologies*". Our baseline system misses the word "*technologies*" altogether, while our MDI-adapted systems recover it. The monolingual MDI-M recovers the missing word, but the adjectives do not agree in number with the noun. MDI-B manages to correct "*thermiques*" and correctly reorders the entire phrase as "*technologies thermiques solaire*", but still deficiently models the morphology.

Line #1497 provides an example of the bilingual model's improvement of fluency for colloquial speech. Our baseline and MDI-M systems provide a direct translation of the source sentence; while it is adequate for written translation, it is too formal for the style of the speaker. In addition, the sense of the translation is "we are able to do it more precisely," which draws attention toward the actor rather than the theme. The reference, on

the other hand, changes the articulation using "*on peut être*," which alters the meaning of the sentence to "it may be more precise." MDI-B replaces "*nous pouvons*" to "*on peut*" in a way that emphasizes the focus of the sentence toward the object to be changed. Each translation is grammatically correct, but for a TED talk, MDI-B's translation is preferred. MDI-B was able to make this lexical decision by drawing on the context of the previous sentence: "We got rid of the stuff that didn't [work]," in which the reference translation uses the same trick of substituting the first person plural "nous" with the third person singular form, "on". The translation hypothesis context counteracts the penalized unadapted unigram ratio assigned by MDI-M (0.798) by increasing it to nearly a 1:1 ratio (0.959).

Sentence #1018 provides an example of improved lexical choice for terminologies. Drawing on the context of the previous sentence: "At that moment, it was as if a film director called for a set change," the speaker uses a metaphor of a "pink carpet" to describe the flamingos in a farm. The baseline and MDI-M systems use "une moquette", which is a type of carpet that is permanently fixed in a location, while MDI-B uses "un tapis", which best captures a sense similar to the expression "rolling out the red carpet".

5.5 Related Work

This work is based on Ruiz and Federico (2011), who combine MDI adaptation with bilingual topic modeling on small adaptation contexts for lecture translation. Adaptation texts are drawn from source language inputs and are leveraged for language model adaptation. A bilingual Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) model is constructed by combining parallel training texts, allowing for inference on monolingual source texts for MDI adaptation by removing source language unigram statistics.

A similar approach is considered by Tam et al. (2007) in domain adaptation by constructing two hierarchical LDA models from parallel document corpora and enforcing a one-to-one correspondence between the models by learning the hyperparameters of the variational Dirichlet posteriors in one LDA model and bootstrapping the second model by fixing the hyperparameters. The bilingual LSA framework is also applied to adapt translation models. Other bilingual topic modeling approaches include Hidden Markov Bilingual Topic AdMixtures (Zhao and Xing, 2008) and Polylingual Topic Models (Mimno et al., 2009).

The literature focuses primarily on domain adaptation, using techniques such as

information retrieval to select similar sentences in training corpora for adaptation, either through interpolation (Zhao et al., 2004) or corpora filtering (Sethy et al., 2006), or mixture model adaptation approaches (Foster and Kuhn, 2007; Koehn and Schroeder, 2007).

An alternative to MDI adaptation is proposed by Chen et al. (1998), which uses a log-linear combination of binary features $f_i(h,w)$ to scale LM probabilities $P(w \mid h)$:

$$\hat{P}(w \mid h) = \exp\left(\sum_i f_i(h,w)\lambda_i\right) P(w \mid h).$$

Normalization is avoided by simply dividing $\hat{P}(w \mid h)$ by $\hat{P}(w \mid h) + 1$.

Bertoldi et al. (2013a) introduce a cache-based language model for adaptation that is proven useful for professional post-editing of machine translation outputs in a computer assisted translation scenario. The model acts as a local model that rewards the n-grams found in translation post-edits. The model incorporates freshness by applying an exponential decaying function to penalize and eventually drop older n-grams. This work is likely complimentary to ours as the cache-based model could directly use the unigram ratios generated by Lazy MDI.

5.6 Chapter Summary

We have presented a simplified framework for approximating MDI adaptation in an online manner for lecture translation as a viable strategy for incorporating context during machine translation and spoken language translation. We avoid normalization computations that prevent classic MDI from being used in speech translation scenarios. Lazy MDI adaptation acts as a separate log-linear feature that doesn't directly adapt LM probabilities – instead, it rewards or penalizes the scores of each translation hypothesis by observing the unigram probabilities inferred an adaptation context and compares it to the background in a smoothed ratio. The smoothing is performed by a conservative fast sigmoid function that favors 1:1 ratios and prevents ratios from growing above a magnitude a. The smoothing function is required to ensure the stability of the unigram adaptation feature in the PBMT log-linear model.

We conducted adaptation experiments on TED talk data from IWSLT 2012 and demonstrate a significant improvement in terms of BLEU, NIST, and TER over two baselines: a lowercased TED-only system, and a state-of-the-art cased system that combines in-domain and out-of-domain data. We demonstrate that Lazy MDI adaptation

has cumulative adaptation effects on already-adapted language models. We additionally analyzed the effect of incorporating translation hypotheses in the context window and observed significant improvements for small context windows on the TED talk translation task.

A potential weakness in our approach is that it uses topic models that do not filter stop-words and it performs unigram adaptation directly on the surface words. For morphologically-rich languages, such as German or Arabic, the vocabulary sizes can increase greatly due to word splitting. For languages such as Arabic, it may be beneficial to apply a morphological segmenter prior to constructing the bilingual topic model.

CHAPTER 6

AUTOMATIC SPEECH RECOGNITION DAMAGING CHANNEL

As discussed in the introduction and in Chapter 4, machine translation (MT) systems that are trained solely on written bitexts do not adapt well to spoken language translation (SLT) scenarios. This is the greatest drawback of the conventional approach of treating SLT as a cascade of automatic speech recognition (ASR) and MT systems, whereby speech recognition is performed and the results are subsequently translated by a MT system (Matusov et al., 2006a; Bertoldi et al., 2007; Casacuberta et al., 2008). The loose coupling between ASR and MT training data creates a scenario where the MT system may not be able to anticipate structural differences between ASR outputs and the input data observed during training time, as well as the presence of speech recognition errors that make the source content untrustworthy. A statistical machine translation (SMT) system trained on written bitexts has statistics covering artifacts that differ from those of ASR outputs. Thus, the ASR output that is passed into a SMT system that lacks sufficient statistical information to model spoken registers may lack fluency. Additionally, the ASR system may generate errors due to (a) signal noise, or (b) modeling deficiencies that prevent the transcription of words not present in the ASR system's pronunciation dictionary. Without observing training data that contains these types of errors, a conventional SMT system has no recourse to recover from them at decoding time. These issues result in the large inventory of translation examples learned in the translation models of well-performing SMT systems being underutilized

in SLT, as many of the paths in the MT search space are inaccessible due to the mismatch between noisy ASR outputs and MT inputs.

Ideally, the structural mismatch between speech and text could be overcome by training the SMT system on speech corpora that have been both transcribed and translated, using a combination of human transcriptions to accurately model the speech register. However, few corpora exist with a sufficient amount of human-transcribed audio with corresponding target language translations – and this type of three-way data is expensive to construct. As a result, SMT systems are limited to being trained with a combination of large amounts of written bitexts and a small amount of translated speech transcripts as adaptation data.

To overcome the dearth of bilingual speech training data, an alternative solution would be to convert the source side of a bitext to ASR-like outputs. Considering the ASR system as a noisy channel that converts the actual transcripts of the speech input to error-prone outputs, we can model the production of ASR errors and apply it on a large amount of bitexts to introduce possible ASR errors, either as training data for SMT system training (Ruiz et al., 2015) or directly injecting examples into the translation model (Tsvetkov et al., 2014). A straightforward method is to actually pronounce every source language sentence in the corpus into a microphone and pass the audio signal through the actual ASR system that will be used in the pipeline. However, this method is also costly and time-consuming. Instead of mapping the text to a signal representation and back to text, we can leverage our knowledge of the noisy channel model used by conventional ASR systems to generate an intermediate representation. Recall in Section 2.1 that an ASR system is minimally composed of an acoustic model, which transforms audio into phoneme sequences, a pronunciation model, which converts phoneme sequences into words, and a language model, which scores sequences of words based on their fluency. Likewise, our model can stop at the phoneme level and generate (1) phonetic confusion between phonemes; and (2) ambiguity within phoneme sequences, using actual ASR outputs recognized from a small amount of transcribed speech audio.

In this chapter, we propose a novel technique to simulate the errors generated by an ASR system, using the ASR system's pronunciation dictionary and language model. Treating the generation of ASR errors as a translation problem, we construct a phoneme-to-word translation model by converting lexical entries in the pronunciation dictionary into phoneme sequences using a text-to-speech (TTS) analyzer. The translation model and ASR language model are combined into a phoneme-to-word MT system that "damages" clean texts to look like ASR outputs based on acoustic confusions. Training texts

are TTS-converted and damaged into synthetic ASR data for use as adaptation for training a spoken language translation system. Our proposed technique yields consistent improvements in translation quality on a number of English-X language pairs, both for lecture and conversational data. Our description expands on the findings in Ruiz et al. (2015).

6.1 Damaging Channel

Our SLT system is a standard cascading ASR-MT pipeline, where the MT system accepts as input a single-best hypothesis from an ASR system. The MT input is recased, punctuated, and tokenized prior to translation.

Our goal is to build a pipeline that converts written text into ASR-like utterances. This pipeline requires (1) a reliable conversion from written text to phonemes, and (2) a modeling technology that can optimize towards a small development set of actual ASR output. For (1), we can either use solely the word-to-pronunciation rules listed in the ASR pronunciation dictionary (PD), or employ the text analysis component of a TTS engine, which dictates written text based on a combination of PD rules, letter-to-sound (LTS) rules, and context-dependent pronunciation rules for numbers, ordinals and acronyms. A discussion of the merits of using TTS is given in Section 6.1.1. For (2), we use phrase-based SMT (Koehn et al., 2003) and tune the system on a small sample of actual ASR outputs. The resulting system is used to generate large volumes of SLT training data to improve the SMT system's robustness toward speech recognition phenomena by "damaging" the source-side of written bitexts to look like ASR-transcribed utterances.

We can conceptually divide the ASR damaging channel pipeline into two stages, as shown in Fig. 6.1. In the first stage, the damaging channel learns how to transform clean source language texts into outputs that contain synthetic ASR errors. Each word in an ASR system's PD is converted into a sequence of phonemes using the LTS rules provided by a TTS analyzer. The mapping between phoneme sequences and their corresponding lexical forms are entered into a phoneme-to-word phrase table with uniform forward and backward probabilities. This phrase table is combined with the language model used by the original ASR system. Since the SMT system which combines phonemes into words is monotonic, no reordering table is required. The phoneme-to-word SMT system is tuned using Minimum Error Rate Training (MERT) (Och, 2003), using a small supervised set of source language speech transcripts and the corresponding single-best

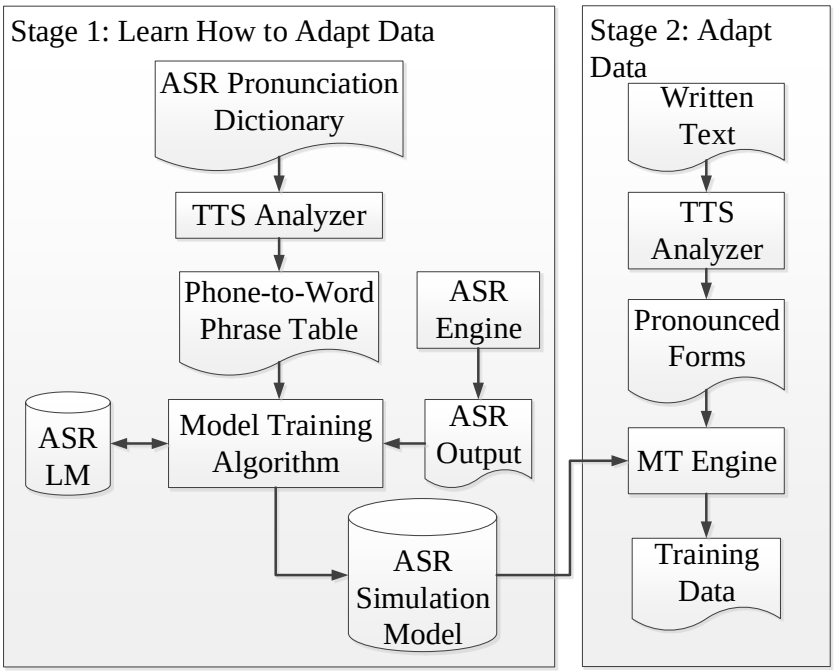


Figure 6.1: ASR damaging channel pipeline. Source language texts are transformed into phoneme sequences and translated back into words, corresponding to a phoneme-to-word SMT system that models errors performed during ASR decoding.

hypotheses from the ASR system. Due to the existence of homophones and other pronunciation anomalies, such a table may have multiple entries for a single phoneme sequence. For example, the phoneme sequence /T UW/ may be mapped to *two*, *to* and *too*.

In the second stage, the source side of the training bitexts are again transformed into phoneme sequences by the TTS analyzer, which are subsequently translated by the phoneme-to-word SMT system to generate synthetic ASR outputs for training the MT component of the SLT pipeline. In practice, all training bitexts are duplicated prior to "damaging"; in this manner the MT component can be trained simultaneously on bitexts with clean source language texts and synthetic ASR output.

6.1.1 TTS-based pronunciation generation

While it is possible to directly use the phoneme sequences present in an ASR pronunciation dictionary to construct a phoneme-to-word phrase table, there are some severe drawbacks:

1. **No coverage for out-of-vocabulary (OOV) words.** The large volume of written

bitexts used to train conventional SMT systems contain many words that are not represented in an ASR system's pronunciation dictionary. In normal speech recognition scenarios, these words are transcribed as a sequence of phonetically similar words or phrases that cover the phones used by the speaker to utter them. In order to generate pronunciations for OOV words and capture their misrecognitions, we must rely on LTS rules.

2. **No pronunciation rules for some acronyms** (e.g. *ADHD*, *MTV*) **and numeric sequences** (e.g. *1998* or *$275,000*). In the currency example, "$" is uttered as "dollars" after the numeric sequence, and the commas are ignored. The remaining numeral is uttered as a sequence of words. We need to apply external rules to correctly "pronounce" these tokens.

3. **Context dependency.** Words may contain different pronunciations given their context (i.e. *record* in *to record music* vs. *a music record*).

Instead of reinventing the wheel, we borrow the text analysis module from a TTS system. While the text analysis module can provide a pronunciation hypothesis for any word, the output of our damaging channel should be ASR-like, meaning it should not contain any OOV words with respect to the ASR PD. Thus, the phoneme-to-word phrase table is restricted to contain only entries in ASR PD.

Another issue is that TTS analyzers may use different phoneme sets from the ASR PD, or they may have been trained on different dialects. For this reason we also pass the lexical entries in the ASR PD through the TTS analyzer. To account for multiple pronunciations of words inside the ASR PD, we may also collect alternative pronunciations of words from the written text and augment the phoneme-to-word phrase table.

6.1.2 Phoneme-level confusion

Our damaging channel pipeline described in Section 6.1 models acoustic confusability in a decoding process similar to the model described in Fig. 6.1. Thus far, we have assumed that the PD contains only valid transcriptions. As such, the decoding process undergone by the phoneme-to-word SMT system defines segmentation boundaries on a sequence of phonemes to reconstruct words. However, this pipeline does not account for phonetic confusability. In the actual ASR scenario, a sequence of phonemes is generated based on the acoustic properties of an input signal and the most likely sequence is generated through a Viterbi search though a series of HMMs. During ASR decoding,

phonemes may be missing or distorted in the input signal, rendering the decoder likely to misrecognize parts of the actual utterance. In response, we introduce an additional step in the damaging channel pipeline which introduces distortions into a sequence of phonemes, based on the observed decoding behavior of an ASR system. Fig. 6.2 outlines this process as a phoneme-to-phoneme SMT pipeline, similar to that of Tan et al. (2010).

A phoneme-to-phoneme phrase table is estimated on a set of phoneme-transcribed source language transcripts and their single-best ASR hypotheses. Optionally, a small lexicalized reordering model may be estimated to allow the swapping of adjacent phonemes.[1] A phoneme language model is estimated on the phoneme sequences of the ASR hypotheses. The weights of the models are optimized using MERT on a held-out development set. The trained phoneme-to-phoneme SMT system can perform the following operations: (1) delete one or more potentially silent or unrecognizable phonemes[2] (2) insert one or more adjacent phonemes[3]; and (3) exchange phonemes that have similar context. The resulting system is applied to each lexical entry in the ASR PD to generate n distorted pronunciation alternatives which are used to expand the dictionary.

6.2 Experiments

We perform experiments on the English-French SLT task from the IWSLT 2014 evaluation campaign (Cettolo et al., 2014), which involves the translation of TED talks in a lecture scenario.

Our baseline SLT system is a cascaded ASR-MT pipeline. The ASR system is described in BabaAli et al. (2014). As a brief summary, the acoustic model is trained on TED talk videos released before December 31, 2010, corresponding to 820 talks and about 144 hours of speech after filtering. It uses a deep neural network (DNN) that is trained using the Karel setup of the open-source Kaldi ASR toolkit (Povey et al., 2011). It is trained over acoustic features generated in the second pass after having applied LDA-MLLT-fMLLR transformations with SAT HMMs. An eleven-frame context window of LDA-MLLT-fMLLR features (5 frames at each side) is used as input to form a 440-dimensional feature vector. The DNN has 6 hidden layers each with 2048 neurons and is pre-trained with Restricted Boltzmann Machines (RBM), followed by mini-batch Stochastic Gradient Descent training, and sequence-discriminative training such

[1]This handles the uncommon scenario where a speaker mispronounces a word (i.e. *nuclear* /N UW K L IY ER/ → /N UW K UW L ER/).

[2]Deletion examples: *arctic* /AA R K T IH K/ → /AA R T IH K/; *figure* /F IH G Y ER → /F IH G ER/

[3]Insertion examples: *realtor* /R IY AH L T ER/ → /R IY AH L AH T ER/; *taut* /T AO T/ → /T AO N T/

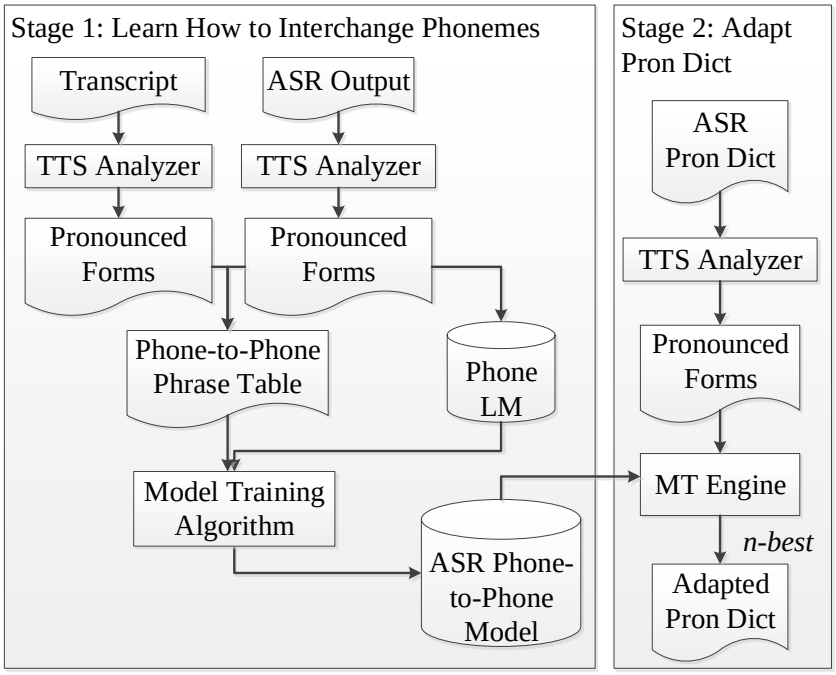


Figure 6.2: Phoneme damaging channel pipeline. Phoneme-to-phoneme SMT maps phoneme sequences from reference transcripts to ASR outputs. Acoustic confusability is modeled directly in the ASR PD.

as Minimum Phone Error (MPE) and state-level Minimum Bayes Risk (sMBR). The single-best ASR hypotheses are punctuated, recased, and tokenized prior to being translated by the MT system. Our ASR system yields a word error rate (WER) of 11.7% on tst_{2012}.

The baseline MT component of our SLT system is a phrase-based Moses system (Koehn et al., 2003, 2007), trained on the TED talk training set permitted in the IWSLT 2014 evaluation. Our baseline system features a statistical log-linear model including a phrase-based translation model (TM) and a lexicalized phrase-based reordering model (RM), both trained on TED data, a 5-gram language model (LM) trained with IRSTLM (Federico et al., 2008) on the French side of the TED training corpus, and distortion, word, and phrase penalties.

6.2.1 Damaging channel

The monotonic phoneme-to-word SMT system is trained on one of three PD configurations: (1) the ASR pronunciations (*lex*); (2) a TTS-generated set of pronunciations for each word (*tts*); or (3) a union of the two (*lex+tts*). In *tts* configurations, each word in

Pron		Phoneme Confusion n-best		
Dict	Types	0	5	10
lex	0.31 M	0.33 M	1.45 M	3.30 M
tts	0.31 M	0.31 M	1.36 M	3.15 M
lex+tts	0.31 M	0.48 M	2.08 M	4.83 M

Table 6.1: Number of word pronunciations modeled in each damaging channel configuration. Phoneme confusability is introduced by translating the PD with a phoneme-to-phoneme SMT system and appending the n-best pronunciations to the dictionary.

the original PD is converted into phonemes using the Festival TTS system (Black and Taylor, 1997) with the English CMU pronouncing dictionary.[4] Statistics on the number of pronunciations modeled by each dictionary type are listed in Table 6.1.

The ASR system's language model is included in the phoneme-to-word SMT system and all model weights are tuned via MERT and evaluated on bitexts that map clean source-language transcripts to our ASR system's single-best hypotheses (tst_{2010} and tst_{2012}, respectively). The clean transcripts are transcribed into phonemes, either using the original PD or by running Festival's TTS analysis component. The ASR hypotheses maintain their lexical form.

We additionally augment the pronunciation dictionaries described above with phoneme confusions using the approach described in Section 6.1.2. The phoneme-to-phoneme SMT system is trained on English bitexts from tst_{2010}. In this case, both the clean transcripts and their ASR hypotheses are converted into phoneme sequences using Festival. A 4-gram language model is estimated on the ASR phoneme sequences using IRSTLM. The model weights are tuned on dev_{2010}. Either a 5- or 10-best list of phoneme sequences is generated for each word in the PD by translating each TTS-generated phoneme sequence into damaged phonemes.

The resulting damaging channel configurations are used to mirror and generate SMT adaptation data from the TED training bitexts, where the source-side transcripts are processed through the damaging channel to generate synthetic ASR output. The synthetic outputs are tokenized, recased, and punctuated prior to being included as training data.

[4] In practice, we could also have run the TTS analyzer on an entire corpus to extract additional word pronunciations, but there were practical issues with implementing this with Festival that are beyond the scope of this experiment.

Transcript	**We're** about to go from six and a half to **9** billion people over the next **40** years
LEX PHONEMES	**w axr** ax b aw t t ax g ow **f ah m** s ih k s **ae n ah** hh ae f t ax **9** b ih l iy ax n p iy **p el** ...
LEX-DAMAGE	**we're** about to go from six and a half to **9** billion people over the next **40** years
TTS PHONEMES	**w er** ax b aw t t ax g ow **f r ah m** s ih k s **ae n d ax** hh ae f t ax **n ay n** b ih l y ax n p iy **p ax l** ...
TTS-DAMAGE	**were** about to go from six and a half to **nine** billion people over the next **forty** years
Transcript	**Their** hunters could smell animal urine at **40 paces and** tell you what **left it behind**
LEX PHONEMES	**dh axr** hh ah n **t axr z** k uh d s m eh l ae n ax **m el y uh r ih n** ae t **40** p ey s ax z ...
LEX-DAMAGE	**they're** hunters could smell animal urine at **40 paces and** tell you what species **left it behind**
TTS PHONEMES	**dh eh r** hh ah n **t er z** k uh d s m eh l ae n ax **m ax l y er ax n** ae t **f ao r t iy** p ey s ax z ...
TTS-DAMAGE	**their** hunters could smell animal urine at **forty paisa Zand** tell you what species left **Iturbe a hind**
TTS-DAMAGE-P2P	**their** hunters could smell animal urine at **forty paces as and** tell you what species **left it behind**

Table 6.2: Example damaging channel output on dev_{2010}, using the original ASR pronunciation dictionary and TTS.

Phone Trans	Pron Dict	ASR BLEU	ASR TER	Transcript BLEU	Transcript TER
lex	lex	74.68	16.20	98.37	0.81
	tts*	20.39	79.53	25.79	72.05
	lex+tts	74.67	16.22	98.27	0.87
tts	lex*	27.97	58.13	33.57	52.00
	tts	47.84	40.37	57.26	32.23
	lex+tts	51.73	35.82	61.94	27.15

*Mismatch between pronunciations.

Table 6.3: Damaging channel models, converting English transcripts into ASR-like outputs, evaluated on dev_{2010}. Phoneme conversion uses either the ASR PD (*lex*) or TTS (*tts*). Evaluated on the target ASR texts and the original transcripts. Mismatches between conversion types are also evaluated.

6.2.2 Synthetic ASR outputs

No phoneme confusions. We first measure how well the damaging channel converts reference transcripts into ASR hypotheses, compared to how much it diverges from itself. A large divergence from the ASR output indicates that the damaging channel is not modeling ASR errors well, while a small divergence from the original transcript indicates that the damaging channel is just reconstructing the original input. Table 6.3 evaluates the effects of phoneme-to-word translation, without factoring in phonetic confusability, both on the ASR hypotheses and the original, unpunctuated transcripts.[5]

[5]While we report both BLEU and TER scores, the TER metric better measures this divergence and it is closely correlated with conventional WER metrics in ASR evaluation.

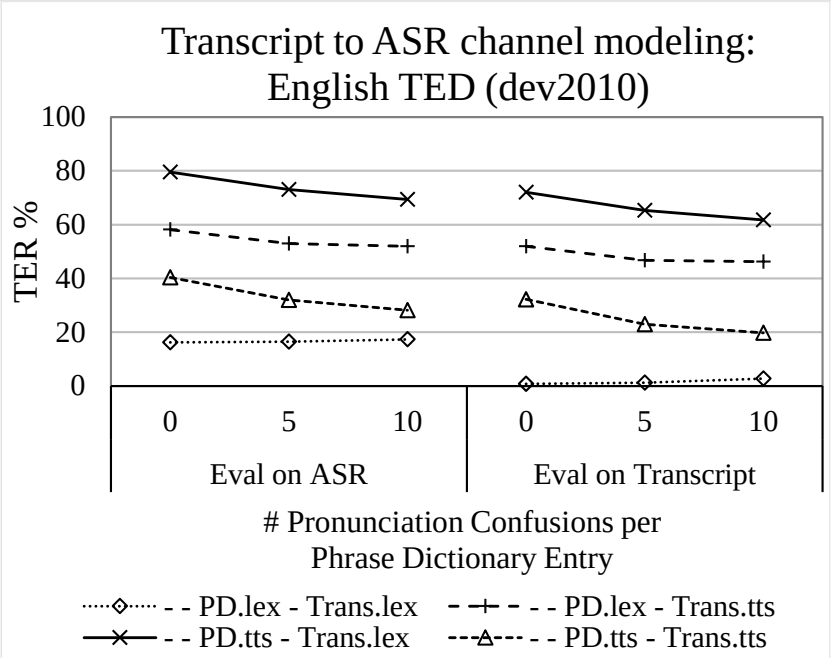


Figure 6.3: Effects of augmenting the PD with phoneme confusions.

While damaging channel models trained on the original ASR PD (*lex*) yield TER scores around 16% against the ASR hypotheses, the damaged texts are virtually the same as the originals; thus, the damaging channel does not model acoustic confusability well enough to transform them. On the other hand, TTS-generated pronunciations yield TER scores around 40% on ASR hypotheses and a similar amount on the original transcripts. We find similar results when combining the *tts* and *lex* pronunciations with a 5% absolute TER improvement. Mismatches between phoneme converters (i.e. transcribing transcripts with *lex* and damaging with a *tts*-trained damaging system and vice-versa) yield abysmal results.

Phoneme confusions. Fig. 6.3 shows the effects of phoneme transduction on the damaging channel. This is done for each of the damaging channel configurations. In nearly every case, adding up to 10 distorted phoneme sequences to each PD before training the damaging channel yields nearly a 10% absolute improvement in TER, both against the ASR outputs and the original transcripts. We also observe that the effects of merging *tts* and *lex* dictionaries becomes insignificant when phoneme confusions are introduced, since the phoneme-to-phoneme SMT system generates the valid pronunciation variants as well.

Table 6.2 provides some examples of synthetic ASR outputs on dev_{2010}. In the first ex-

ample, the numbers 9 and 40 are not in the original ASR PD. The PD-driven damaging channel (LEX-DAMAGE) treats these as OOV words and dumps the numbers back in the damaged outputs. During SMT model training, phrases containing these numbers will never be used in the SLT pipeline. The TTS-driven damaging channel (TTS-DAMAGE) successfully converts them to phoneme sequences and reconstructs their lexical form. Additionally, *we're* is pronounced with a northwestern accent (/W ER/) by the TTS analyzer, resulting in the word being reconstructed as *were*.

The second example demonstrates cases where the TTS-driven damaging channel's TM may give higher scores to low frequency words than common words. *Paces and* is converted to *paisa Zand*. This is because the TM assigns uniform probabilities to phoneme-to-word and word-to-phoneme entries. Since the PD was generated in a data-driven fashion, junk entries appear that usually are never encountered during ASR decoding. However, by introducing phoneme confusions through the phoneme-to-phoneme SMT system (TTS-DAMAGE-P2P), the TM scores are smoothed with the addition of 5 pronunciations per lexical entry. TTS-DAMAGE-P2P assigns *paces* a pronunciation with a dropped "s" (/P EY S AX/) and duplicates /AX/, rendering the damaged output as *paces as and* (/P EY S AX AX Z AE N D/). We discuss the TM issue in more detail in Section 6.2.4.

6.2.3 SLT evaluation

We conduct two sets of TED-only experiments to simulate two domain adaptation scenarios. In the first set of experiments, the damaged TED transcripts and their translations are concatenated with the clean TED training data to estimate the translation model and reordering model (CONCAT). In the second set of experiments, a separate phrase table is estimated on the damaged bitexts. The source phrases in the damaged phrase-table that are not present in the baseline TED phrase-table are appended using the fill-up technique (Nakov, 2008; Bisazza et al., 2011) with a provenance feature that marks the phrase as synthetic (FILL-UP). The overlapping synthetic phrase pairs are discarded. To control for optimizer instability (Clark et al., 2011), we run MERT three times on each experiment and evaluate the performance of each system using the MultEval toolkit[6]. Table 6.4 evaluates the adaptation techniques reports on the tst_{2012} data set.

We observe statistically significant improvements in BLEU, ranging from 0.6-0.8

[6]https://github.com/jhclark/multeval

	Phoneme Confusion n-best					
System	0	5	10	0	5	10
Baseline	28.44	–	–	28.44	–	–
lex	29.19	29.04	28.92	29.06	29.02	28.83
tts	29.08	29.24	29.06	28.90	28.54	28.94
lex+tts	28.91	29.13	29.20	28.90	28.84	28.77
	CONCAT			FILL-UP		

Table 6.4: Evaluation results on tst_{2012} (in BLEU). Damaged TED transcripts are either CONCATenated with clean transcripts or used to generate new FILL-UP phrase table entries on the baseline TED phrase-table.

English ref	Since **it's** digital, we can do reverse dissection.
ASR output	Since **its** digital we can do reverse dissection .
Baseline MT	**Depuis que** nous pouvons faire **son numérique** inverser sentinelles .
LEX-DAMAGE	**Puisque c' est que** nous pouvons faire renverser dissection **du numérique** .
TTS-DAMAGE	**Depuis ses digital** , nous pouvons faire régresser axillaire .
TTS-DAMAGE-P2P	**Depuis ses numérique** , nous pouvons faire renverser axillaire .
French Ref	**Puisque c' est numérique** , nous pouvons faire une dissection à l' envers .
English Ref	... I've studied **technologies of mobile communication** ...
ASR output	... I 've studied **technologies , of mobile , communication** ...
Baseline MT	... j' ai étudié **technologies , de téléphones , la communication** ...
LEX-DAMAGE	... j' ai étudié **les technologies de communication , de technologie mobile** ...
TTS-DAMAGE	... j' ai étudié **les technologies , de technologie mobile , la communication** ...
TTS-DAMAGE-P2P	... j' ai étudié **les technologies de communication , de portable** ...
French Ref	... j' ai étudié **les technologies de communication mobile** ...

Table 6.5: Example SLT outputs from tst2012, using damaging channel output as concatentated training data.

for our CONCAT and 0.4-0.6 for FILL-UP ($p < 0.01$), with the exception of the TTS-trained damaging channel. The fill-up results are weaker due to the lack of training data to estimate each phrase table, yielding less reliable count statistics and subsequent phrase probabilities. With larger amounts of training data, concatenating corpora generally causes the larger pool of out-of-domain corpora to dominate the TM (Koehn and Schroeder, 2007).

Table 6.5 provides examples of end-to-end SLT English-French translations on tst_{2012}, generated by the baseline SMT system and SMT systems trained with LEX-DAMAGE, TTS-DAMAGE, and TTS-DAMAGE-P2P. In the first example, the contraction *it's* is misrecognized as the possessive pronoun *its*. While all damaging channel systems permit the error-tolerant mapping of *its* to *c'est*, only LEX-DAMAGE applies it successfully. However, it comes at the cost of splitting the source phrase *it's digital* into two separate phrases and *digital* is reordered incorrectly to the end of the sentence.

The second example demonstrates punctuation errors that change a segment's meaning. *Technologies of mobile communication* becomes a list of three items. The baseline and TTS-DAMAGE-P2P systems translate *mobile* either as a physical telephone device or a portable object. LEX-DAMAGE and TTS-DAMAGE-P2P generate translations related to *communication technologies*, which captures part of the original meaning. TTS-DAMAGE, on the other hand, generates a translation for *mobile technology*. While imperfect, each damaging channel-trained system manages to reorder phrase pairs in order to cross the erroneous punctuation boundaries, thereby improving the translation quality.

6.2.4 Analysis

Our damaging channel's phoneme-to-word TM suffers from forward probability dilution when multiple pronunciations for a word exist. For instance, LEX-DAMAGE has 12 pronunciations for *intercontinental*, each with a forward score of 0.077. The problem is exacerbated when introducing phoneme confusions. The 12 original pronunciations inflate to 34 and 69 when adding the 5- and 10-best phoneme confusions, respectively, while a word with a single pronunciation gains a quantity proportional to n. This behavior may result in junk word sequences like *in ter continent tall* to be favored, in spite of the word penalty feature and the low LM probabilities. The impact of this issue may be reduced by weighting the probability distribution by corpus frequencies, or pruning infrequent junk words.

Additionally, using a single TTS pronunciation for each word proves to be detrimental to the damaging channel. Gerund words such as *doing* and *creating* in the PD are transcribed with a /IH NG/ suffix in isolation by Festival, but in context they are commonly transcribed as /AX NG/ in context.[7] No valid pronunciations exist in the phrase table, causing the damager to back off to nonsense constructions with junk entries like *due a ng* and *create ng*. Phoneme-to-phoneme pronunciation expansions minimize this effect, at the cost of diluting phrase table scores. Instead, the TTS analyzer should generate additional pronunciations by processing sentences from a corpus and segmenting the resulting phoneme sequences.

[7]This issue is not restricted to words belonging to a particular lexical class. It may occur anytime there is a mismatch between TTS and the entries in the PD.

Lang	Corpus	Segments	Types	Tokens	WER
English	APP	1111	2772	14186	16.79
English	SKP	290	578	2824	17.36

Table 6.6: Statistics on internal conversational test sets.

6.2.5 Experiments on conversational data

We conducted additional experiments to analyze the utility of our damaging channel on conversational data from English to several target languages. The ASR system is described in Aue et al. (2013). The MT system is a typical phrase-based SMT system, similar to that of Koehn et al. (2007), trained on a large collection of in-house speech and text data. LM rescoring is applied to lattices of ASR hypotheses, and single-best ASR hypotheses are preprocessed following the approach of Hassan et al. (2014), prior to translation.

Our damaging channel is a phoneme-to-word MT system as described in Section 6.1. The phoneme-to-word phrase table is populated by pronouncing each word in the ASR PD with an in-house TTS recognition engine. We augment the pronunciations by also running TTS on a large set of text corpora and mapping pronunciations to the words in the PD. OOV word pronunciations are discarded. All available training corpora are duplicated, damaged, and trained as separate phrase tables to be used together with phrase tables trained with clean data.

We evaluate on two sets of in-house conversational data. Statistics on the English transcripts from the corpora and WER scores are provided in Table 6.6. Our ASR system yields WER scores around 17% on the internal corpora. We compare the performance of our SMT system which includes damaged data (DAMAGE) against a baseline (BASE) that only uses clean data. Results are shown in Table 6.7, both on the translation of ASR hypotheses and reference transcripts. We have two references for English-Mandarin and one reference for other test sets. All SLT experiments that utilize damaged training data perform statistically as well or better than the baseline. We observe significant improvements on the SKP dataset for English-Italian (+0.3 BLEU), and English-Mandarin (+1.1 BLEU) and improvements for English-Spanish (+1.3 BLEU) on the APP dataset.

In general, we also observe that utilizing the damaging channel not only improves the translation quality of ASR outputs, but in many cases also increases the phrase table coverage on reference transcripts, as well. The increase in coverage yields similar benefits to augmenting a phrase table with paraphrases (Callison-Burch et al., 2006).

		ASR		Transcript	
Lang	Corpus	Base	Damage	Base	Damage
English-Spanish	APP	31.54	**32.89**	40.87	**43.89**
English-Spanish	SKP	24.94	24.93	39.26	**40.72**
English-German	SKP	9.96	10.16	**17.72**	17.17
Engish-French	SKP	22.86	22.71	**36.46**	35.86
English-Italian	SKP	14.35	**14.67**	22.33	**22.71**
English-Chinese (Mandarin)	SKP	29.56	**30.73**	34.08	**35.40**

Table 6.7: Evaluation results on internal test sets (in BLEU) for multiple language pairs.

6.3 Related work

Techniques to generate synthetic ASR errors have been used for discriminative language modeling (Kurata et al., 2009; Jyothi and Fosler-Lussier, 2010; Sagae et al., 2012), ASR error prediction (Jyothi and Fosler-Lussier, 2009), and speech translation (Aue et al., 2013; Tsvetkov et al., 2014).

Kurata et al. (2009, 2011) use a weighted finite state transducer (WFST) compiled from an ASR PD to convert phoneme sequences back into words. The ASR system's acoustic model is used to measure confusability between phonemes. Sagae et al. (2012) propose a variant to phoneme transduction by estimating phoneme substitution probabilities using maximum likelihood estimates on Levenshtein alignments between the reference transcript and a n-best list of ASR hypotheses. In both methods n-best outputs were generated and utilized in discriminative LM training.

Tan et al. (2010) implement a similar phoneme-to-phoneme transducer, modeled as a SMT system and propose its use in conjunction with a FST-based phoneme-to-word transducer to damage texts. However, they assume that no OOVs are present in the texts to damage and they did not apply their work on actual MT training data. Our method uses a TTS analyzer to bridge the crucial gap between the ASR PD and the MT data. Tsvetkov et al. (2014) extend the method by using a phone confusion transducer. The transducer allows substitutions based on phone clusters, consonant deletions, vowel duplications, and suffix insertions. Like Tan et al. (2010), they compose the transducer with the ASR PD and LM. They apply the transducer on each entry in the SMT phrase table, generating alternative source phrases.

Our approach is an extension and deeper analysis of the text normalization approach of Aue et al. (2013), which uses a text-to-speech engine to introduce phonetic confusability by generating alternative pronunciations for existing words in an ASR lexicon and

using phoneme-to-word SMT to reconstruct word sequences constrained in the lexicon.

6.4 Chapter Summary

We have constructed several variants of a damaging channel that utilizes principles of acoustic and phonetic confusability to model the conversion of sequences of phonemes to synthetic ASR outputs containing potential errors. In particular, clean texts are converted to phoneme sequences by a TTS analyzer and are subsequently "translated" back into words based on the observed behavior of an ASR system. Our TTS-driven approach successfully converts OOV words, acronyms, and numeric sequences into words belonging to a ASR PD and can be used to generate synthetic speech data to adapt a MT system to the SLT task. Our experiments show that MT systems adapted with damaged texts are better suited to receive ASR outputs as input than systems trained only on bitexts.

In its current state, the TTS-driven damaging channel performs similarly to configurations which directly use the ASR PD. However, the TTS-driven approach is capable of generating synthetic texts that diverge further from the original transcripts in such a way that utilizing multiple damaged hypotheses could improve error coverage during MT training.

One of the problems of the phoneme damaging channel is that the introduction of acoustic confusions is applied on a discrete set of symbols. GMM/DNN approaches to acoustic modeling implicitly allow tolerance in recognizing phoneme sequences due to the fact that it must recognize discrete phonemes from acoustic events in the form of continuous features. While ASR systems use allophone-based acoustic modeling and speaker adaptation to to model the variability of the acoustic realizations of a phone (Yi and Fung, 2003), our model maps discrete symbols to phonemes as a grapheme to phoneme transduction problem. In the process we lose information about the acoustic distance between phonemes that acoustic models learn. One could try to map the grapheme to phoneme transduction problem used by our approach to convert the problem into a continuous space through the use of word embeddings and sequence-to-sequence modeling. However, the process would require more data than that available in our experiments.

CHAPTER 7

NEURAL SPOKEN LANGUAGE TRANSLATION EVALUATION

A substantial amount of progress has been made in Neural Machine Translation (NMT) for text documents. Research has shown that the encoder-decoder model with an attention mechanism generates high quality translations that exploit long range dependencies in an input sentence (Bahdanau et al., 2015). While NMT has proven to yield significant improvements for text translation over log-linear approaches to MT such as phrase-based machine translation (PBMT), it has yet to be shown the extent to which gains purported in the literature generalize to the scenario of spoken language translation (SLT), where the input sequence may be corrupted by noise in the audio signal and uncertainties during automatic speech recognition (ASR) decoding. Are NMT models implicitly better at modeling and mitigating ASR errors than the former state-of-the-art approaches to machine translation? As the final chapter in this thesis, we analyze the impact of ASR errors on neural machine translation quality by studying the properties of the translations provided by an encoder-decoder NMT system with an attention mechanism, against a strong baseline PBMT system that rivals the translation quality of Google Translate™ on TED talks.

In this chapter, we address the following questions regarding NMT:

1. How do NMT systems react when ASR transcripts are provided as input?
2. Do the ASR error types discussed in this thesis impact SLT quality the same for NMT as PBMT? Or is NMT implicitly more tolerant against ASR errors?

3. Which types of sentences does NMT handle better than PBMT, and vice-versa?

In order to address these questions, we explore the impact of feeding ASR hypotheses, which may contain noise, disfluencies, and different representations on the surface text, to a NMT system that has been trained on TED talk transcripts that do not reflect the noisy conditions of ASR. Our experimental framework is similar to that of Chapter 4, with the addition of a ranking experiment to evaluate the quality of NMT against our PBMT baseline. These experiments are intended as an initial analysis with the purpose to suggesting directions to focus on in the future.

7.1 Neural versus Statistical MT

Before beginning our analysis, we summarize some of the biggest differences between NMT and other forms of statistical machine translation, such as PBMT, and highlight some works in the literature that provide some preliminary analyses.

Bentivogli et al. (2016) compare neural machine translation against three top- performing statistical machine translation systems in the TED talk machine translation track from IWSLT 2015.[1] The evaluation set consists of 600 sentences and 10,000 words, which were post-edited by five professional translators by applying the minimal edits required to transform each sentence into a fluent output with the same meaning as the source sentence. In addition to reporting a 26% relative improvement in multi-reference TER (mTER), Luong and Manning (2015)'s encoder-decoder attention-based NMT system trained on full words outperformed well-established state of the art SMT systems on English-German, a language pair known to have issues with morphology and whose syntax differs significantly from English in subordinate clauses. Bentivogli et al.'s analysis yields the following observations:

- *Precision versus Sentence length*: The NMT system outperformed every comparable log-linear MT system, regardless of the sentence length; however, Bentivogli et al. confirmed Cho et al. (2014a)'s findings that translation quality deteriorates rapidly as the sentence length approaches 35 words.
- *Morphology*: NMT generates translations that have better case, gender, and number agreement than PBMT systems.
- *Lexical choice*: NMT made 17% fewer lexical errors than any PBMT system.

[1]The International Workshop of Spoken Language Translation.

- *Word order*: NMT yielded fewer *shift* errors in the TER alignment between its outputs and its post-edited reference than any SMT system. NMT yielded significantly higher Kendall Reordering Score (KRS) values than any PBMT system. In particular, NMT generated 70% fewer verb order errors than the second-best performing hybrid phrase and syntax-based (PBSY) system .

While NMT performs better than the log-linear SMT approaches in the areas listed above, there are several modeling challenges that are exacerbated in NMT. The first significant difference is that log-linear SMT translation models can handle word vocabularies that are orders of magnitude larger than those of NMT systems. Since each in-vocabulary token increases the size of the network, including its hidden layers, careful modeling is required in NMT to represent a sufficiently large set of tokens that adequately covers the source and language vocabularies of the translation task, while simultaneously maintaining the overall network size to ensure that the model remains small enough to train on existing architectures. Due to the vocabulary size limitations imposed by NMT, Hirschmann et al. (2016) observe that only 69% of German nouns are covered when encoding the English-German WMT 2014 training data into a fixed-size vocabulary of 30,000 words.[2] Although noun compound splitting works well in the German→English direction, English→German model performance not improve significantly. In particular named entities (e.g. persons, organizations, and locations) are underrepresented under the reduced vocabulary size restrictions of modern NMT models.

On the other hand, one major advantage of NMT is the ability to model subword units such as characters (Chung et al., 2016) or coarser grained segmentations on low frequency words (Sennrich et al., 2016) without substantial changes to the system architecture. The main downside of fine-grained segmentation such in as character-based models is that the length of the input (or output) string is affected. While both the encoder and decoder hidden states are modeled by recurrent neural networks architectures that can model long distance dependencies over an entire sequence of observations, sub-word models increase the time and resources required for training and decoding. A PBMT system such as that described by Koehn and Schroeder (2007) requires strong constraints on the distance of reordering operations, as well as the number of tokens permissible in the source or target language side of the translation model. Likewise, differences in the orthographic representation of the source and target strings can be problematic, for example, in Japanese-English or Mandarin-English phrase-based MT

[2]WMT 2014 training data consists primarily of news texts, European parliament proceedings, and web crawled data. `http://www.statmt.org/wmt14/translation-task.html`

(Chang et al., 2008; Wang et al., 2007). On the other hand, works such as Nakov and Tiedemann (2012); Neubig et al. (2013) attempt to build translation models that combine full words and substring units. The downside of fine-grained segmentations such as in character-based models for NMT is that the length of the input affects training and decoding time, as both the attention and hidden state models are modeled by recurrent neural network architectures that model long distance dependencies over the entire sequence of observations in an input string. Nevertheless, the representation of subword units allow the attention model to decide the length and grain of the input sequence that is useful for decoding each target position.

Firat et al. (2016) have additionally demonstrated NMT's ability to translate between multiple language pairs with a neural translation model trained with a single attention mechanism. While there have been attempts to model multilingual translation in traditional forms of statistical machine using multi-source machine translation (Och and Ney, 2001; Schroeder et al., 2009) or pivoting (Cohn and Lapata, 2007; Bertoldi et al., 2008), most phrase-based and hierarchical MT systems require training on individual language pairs.

Although NMT models translate with higher precision, it comes at a large cost: models are slow to train even with the most powerful graphical processing units – often taking weeks for the strongest systems to complete training. On the other hand, large order PBMT systems trained in the latest ModernMT framework[3] may be trained within a few hours and can be adapted in near real-time with translation memories containing translation post-editions by professional translators. We discuss the translation system and summarize the ModernMT project in Section 7.2.2.

7.2 Research Methodology

Similar to our experimental framework in Section 4.1, we collect the English ASR hypotheses from eight research laboratories, which correspond to their English ASR submissions on the tst_{2012} test set. These ASR hypotheses are segmented into sentences that match the reference set of the English-French MT track of the evaluation campaign. Coupled with the reference translations from the MT track, we construct our spoken language translation dataset, consisting of the eight English ASR hypotheses for 1,124 utterances (8,992 in total), a single unpunctuated reference transcript from the ASR track, and the reference translations from the English-French MT track. The

[3] `http://www.modernmt.eu`

English ASR hypotheses and reference transcript are normalized and punctuated according to the same approach as we described in Section 4.1. We use BLEU (Papineni et al., 2002) and Translation Edit Rate (TER) (Snover et al., 2006) both as global evaluation metrics. For sentence-level MT quality assessments in our subsequent experiments, we measure the increase in sentence-level TER, ΔTER as a result of ASR errors in the source language input. We compute automatic translation scores, sentence-level system ranking, and take a closer look at the types of errors observed in the data.

In the sections below, we briefly describe the MT systems used in this experiment.

7.2.1 Neural MT system

Our NMT system is based on FBK's primary MT submission to the IWSLT 2016 evaluation for English-French TED talk translation (Farajian et al., 2016). The system is based on the sequence-to-sequence encoder-decoder architecture proposed in Bahdanau et al. (2015) and further developed by Luong and Manning (2015); Sennrich et al. (2016). The network architecture closely follows the system described in our exposition on NMT in Section 2.4 and is trained on full-word text units to allow a direct comparison with our PBMT counterpart. We refer to this system as NMT-FULL for the remainder of our experiments.

7.2.2 Phrase-based MT system

Our phrase-based MT system is built upon the ModernMT framework, designed under the European Union's Horizon 2020 research and innovation programme.[4] ModernMT is an extension of the phrase-based MT framework introduced in Koehn and Schroeder (2007) that enables context-aware translation in a framework that allows for fast and incremental training. Context-aware translation is achieved through the partitioning of the training data in homogeneous domains by a *context analyzer*, which permits the rapid construction of domain-specific translation and language models. The context analyzer also permits a rapid interpolation of the translation, reordering, and language sub-models based on the context received during a decoding run and the underlying models may be modified online. Given an input sentence and its context,[5] the context analyzer returns a list of semantically similar example instances which are used to customize domain-adapted models on-the-fly that represent the domain of the transla-

[4] `https://ec.europa.eu/programmes/horizon2020/`

[5] The context usually consists of the full document to which the input sentence belongs.

tion task, The adaptive translation and reordering models compute scores exploiting suffix-array and ranked sampling driven by the context analyzer output. The adaptive language model computes its score as a weighted linear combination of domain-specific LMs, where weights are again provided by the context analyzer. The decoder also exploits other standard features (phrase, word, distortion, and unknown word penalties) and performs cube-pruning search. A detailed description of the ModernMT project can be found in Bertoldi et al. (2016). We refer to this system as MMT for the remainder of our experiments.

7.3 SLT Evaluation

We first report the translation results on the evaluation task in Table 7.1. The results compare NMT-FULL's translation performance against the ModernMT version of PBMT (MMT). NMT outperforms our best PBMT system by 4.5 BLEU in the absence of ASR errors (*gold*). In the presence of ASR errors, we note that while the NMT-FULL results outperform those of MMT, the lead is reduced to approximately 3 BLEU across all ASR hypothesis inputs. Given the high starting point of NMT-FULL on the gold standard, we expected it to outperform MMT. Overall, the introduction of ASR errors results in decreases in BLEU by 5.5(±0.8) and 5.4(±0.8) and TER increases of 6.0(±0.9) and 6.2(±0.9) for MMT and NMT-FULL, respectively.

Table 7.2 provides a sentence-level view of the evaluation results by providing the average sentence level TER scores, as well as average sentence-level ΔTER scores, which

		MMT		NMT-FULL	
ASR system	WER	BLEU	TER	BLEU	TER
gold	0.0	43.40	39.50	47.90	35.40
fbk	16.5	35.60	48.40	38.50	45.60
kit	10.1	38.10	45.00	41.80	41.70
mitll	11.4	37.70	45.80	41.40	42.40
naist	10.6	38.10	45.00	41.80	41.50
nict	9.2	38.70	44.70	42.50	41.10
prke	16.6	34.90	48.70	38.10	45.80
rwth	11.7	37.20	46.10	41.40	42.30
uedin	12.3	37.30	46.10	40.80	42.90

Table 7.1: A comparison of Neural MT versus Phrase-based MT on the SLT evaluation of TED talks (tst_{2012}) from the IWSLT 2012 evaluation campaign. Evaluation results are compared to a *gold* standard that assumes that no ASR errors occur.

	MMT		NMT-FULL		DIFFERENCE	
SysID	TER	ΔTER	TER	ΔTER	TER	ΔTER
gold	39.6	0.0	35.6	0.0	-4.0	0.0
fbk	49.3	9.7	46.6	11.0	-2.7	1.3
kit	45.9	6.3	42.7	7.1	-3.2	0.8
mitll	46.8	7.2	43.7	8.1	-3.1	0.9
naist	45.6	6.1	42.1	6.5	-3.5	0.5
nict	45.1	5.5	41.9	6.3	-3.1	0.9
prke	49.4	9.8	46.5	10.9	-2.9	1.1
rwth	47.0	7.4	43.2	7.6	-3.8	0.2
uedin	46.7	7.1	43.8	8.2	-2.9	1.1

Table 7.2: Average utterance-level translation TER and ΔTER scores for the MMT and Neural MT systems. The average Neural MT TER scores are an absolute 3% better than the PBMT counterpart.

report the degradation of SLT quality by the presence of ASR errors. We observe that although the average TER scores from the MMT outputs are higher, the ΔTER scores are lower than their NMT-FULL counterparts, implying that the MMT SLT outputs are closer to their gold standard MT outputs. This may imply that PBMT may be less sensitive to local changes to an input caused by minor ASR errors.

7.3.1 MT system ranking

These results above lead us to raise the question: "Are there ASR error conditions in which PBMT remains a better solution than NMT, and if so, what are the properties of these utterances that makes them difficult for NMT?" To address this question, we take a closer look at the sentence-level translation scores by ranking the performance of each MT system on the utterances where ASR errors exist, in order to understand how each MT system handles noisy input. For each utterance, we rank the systems based on their the sentence-level TER scores computed on their translation outputs over each ASR hypothesis. We also mark ties, in which both systems yield the same TER score. Results are provided in Table 7.3.

NMT-FULL consistently produces better scoring translations in terms of TER over 47% of the utterances better than MMT. The NMT-FULL and MMT scores are tied on over 20% of the utterances. The middle of Table 7.3 contains counts and percentage of the wins by each system. For the better performing ASR systems (e.g. NICT, KIT), we observe a slightly higher proportion of utterances with better NMT translations and

				TER (avg)	
Lab	Winner	Count	Percentage	MMT	NMT-FULL
fbk	MMT	257	32.4	51.1	64.0
	NMT-FULL	373	47.1	58.9	44.5
	Tie	162	20.5	54.1	54.1
kit	MMT	213	30.6	46.3	59.1
	NMT-FULL	347	49.9	55.0	41.1
	Tie	135	19.4	52.9	52.9
mitll	MMT	194	27.6	48.4	61.4
	NMT-FULL	351	49.9	55.5	41.5
	Tie	159	22.6	52.2	52.2
naist	MMT	189	28.3	43.9	56.5
	NMT-FULL	342	51.2	54.8	41.2
	Tie	137	20.5	52.5	52.5
nict	MMT	184	31.8	46.3	58.4
	NMT-FULL	286	49.4	54.0	40.8
	Tie	109	18.8	56.0	56.0
prke	MMT	256	31.6	48.0	60.3
	NMT-FULL	378	46.7	57.7	44.1
	Tie	175	21.6	55.8	55.8
rwth	MMT	221	29.9	47.0	59.2
	NMT-FULL	383	51.8	55.5	41.3
	Tie	135	18.3	55.0	55.0
uedin	MMT	219	30.6	47.9	59.2
	NMT-FULL	348	48.7	56.5	42.8
	Tie	148	20.7	52.2	52.2

Table 7.3: Ranked evaluation of the SLT utterances containing ASR errors in tst_{2012}. (Left) Counts of the winner decisions and the percentage of all of the decisions that were influenced by ASR errors. (Right) Mean TER scores across each sentence in the ranked set. The remainder of winner decisions are made on error-free ASR transcripts.

a reduced number of ties. On the right-hand side of the table we report the average TER scores within each ranking partition of the data. For example, for the utterances that are translated better by MMT, we observe that the average TER scores for NMT-FULL have an absolute average improvement of 10% in TER over MMT. The converse is also true, suggesting that there is a subset of utterances that MMT translates better than NMT-FULL. We now focus on these utterances and determine whether the ranking decisions made under noisy scenarios matches the scenario where the utterance was recognized perfectly.

The distribution of wins reported in Table 7.3 suggests that systems with more ASR errors seem to have a greater percentage of sentences that are either scored higher by MMT or are ties. We look into this more closely by plotting the changes in MT system ranking as we shift from a gold standard ASR to the actual ASR results from each system in the evaluation. Figure 7.1 shows the distribution of ranking decision pairs

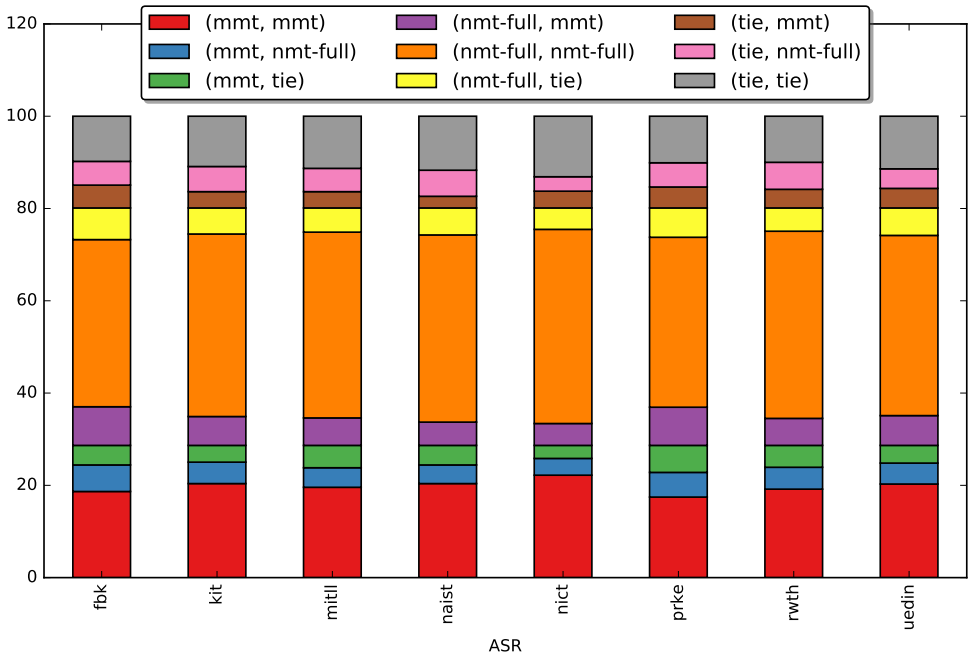


Figure 7.1: Changes in MT system rankings as ASR errors are introduced. Tuples are labeled by (MT rank, SLT rank).

by ASR system. Across all ASR outputs, 70.2% of the MT evaluation ranking decisions remain the same when ASR errors create noisy input. Of the 29.8% of changes, an average of 40.8% favor NMT-FULL and 29.8% favor MMT across all systems. There are fewer cases of mismatches between the gold ranking decisions as the WER of the ASR system decreases.

7.3.2 Translation examples

We provide three examples of key differences between in how NMT-FULL and MMT mitigate FBK's ASR errors Figure 7.3. In utterance U4, NMT-FULL is missing the translation of two content words from its vocabulary. In the absence of errors NMT-FULL$_{gold}$ passes the source word "embody" through to its output without translating it. During ASR, "embody" is misrecognized as "body", which could either be a source word passed through to the translation, or potentially a French piece of lingerie. We find it strange that "body" was not translated as "corps", given that other utterances containing "body" receive that translation. After investigating further, we came across other cases of gold transcripts where "body" was not translated at all. Utterances U212, U214, and U242

U212	I call myself a body architect.	je m' appelle un corps architecte .
U214	As a body architect, I fascinate with the human body	en tant qu' architecte , je me suis retrouvé avec le corps humain
U242	As a body architect, I've created	en tant qu' architecte , j' ai créé

Figure 7.2: Examples where NMT translates "body architect" differently, based on its context. U214 and U242 drop the word "body" altogether.

have the phrase "body architect", but only U212 has a translation for the word "body" (shown in Figure 7.2). It is likely that *NMT may not be able to fully translate contextual patterns it hasn't observed before.* MMT on the other hand provides valid translation for both words; although the meaning of the sentence is lost due to the translation of ASR errors. As a PBMT system, it will translate phrases consistently, as long as there is not another overlapping phrase pair in the translation model that leads to a path in the search graph with a higher translation score.

Utterance U85 shows longer range effects of ASR errors on translation in NMT. In the translation of ASR_{gold}, both MT systems translate the expression "stepped back" in the sense of "returned". MMT_{gold} reorders "centre" incorrectly. ASR_{hyp} has a single error where the past tense suffix "-ed" on "step" was lost. NMT-FULL_{ASR} provides an adequate translation as "je recule", but in the process, the attention mechanism seems to have taken the incorrect source word and translation as context that corrupts the remainder of the translation. While MMT_{ASR} makes a translation error at the beginning of the sense, the remainder of the translated sentence remains the same as its gold translation. This suggests that *ASR errors may have longer range effects on NMT systems* in languages that are even observable in sentences that lack long distance dependencies.

Utterance U296 demonstrates an example where *misrecognitions of short function words can cause the duplication of content words in NMT*. While MMT handles the misrecognition "and"⇒"an" by backing off by translating it independently from other phrases in the sentence, NMT-FULL, attaches "photo" both to the article "an" and additionally outputs "photo" at its usual position. As an innocuous closed-class word error that could happen often in ASR outputs, this could be a potentially significant problem in NMT.

		TER	ΔTER
ASR_{gold}	I embody the central paradox.		(U4)
ASR_{hyp}	I **body** the central paradox.		
Trans	j' incarne le paradoxe central .		
MMT_{ASR}	je **corps au** paradoxe central .	50.0	50.0
$\text{NMT-FULL}_{\text{ASR}}$	je **body** le paradoxe central .	33.33	16.66
MMT_{gold}	j' incarne le paradoxe central .	0.0	
NMT-FULL_{gold}	j' **embody** le paradoxe central .	16.67	
ASR_{gold}	But when I stepped back, I felt myself at the cold, hard center of a perfect storm.		(U85)
ASR_{hyp}	But when I **step** back, I felt myself at the cold, hard center of a perfect storm.		
Trans	mais quand j' ai pris du recul , je me suis sentie au centre froid , et dur d' une tempête parfaite .		
MMT_{ASR}	mais quand j' ai *[]* du recul , je me sentais au froid , dur centre d' une tempête parfaite .	21.74	-17.39
$\text{NMT-FULL}_{\text{ASR}}$	mais quand **je recule** , je me sentais **dans le froid et le centre** d' une tempête parfaite .	47.83	13.05
MMT_{gold}	mais quand **je suis revenu** , je me sentais **au froid** , dur centre d' une tempête parfaite .	39.13	
NMT-FULL_{gold}	mais quand **je suis revenu** , je me sentais au centre froid et dur d' une tempête parfaite .	34.78	
ASR_{gold}	And he emailed me this picture.		(U296)
ASR_{hyp}	An emailed me this picture.		
Trans	il m' a envoyé cette photo .		
MMT_{ASR}	**un** m' a envoyé cette photo .	14.29	0.0
$\text{NMT-FULL}_{\text{ASR}}$	**une photo** m' a envoyé cette photo .	28.57	14.28
MMT_{gold}	et il m' a envoyé cette photo .	14.29	
NMT-FULL_{gold}	et il m' a envoyé cette photo .	14.29	

Figure 7.3: Three examples of changes in NMT errors (NMT-FULL) caused by ASR errors: (1) the effects of unobserved context; (2) long distance effects of local ASR errors; (3) duplication of content words caused by substitution errors on short function words. Alternative translations are provided by MMT. TER and ΔTER scores are reported for each sentence translated by NMT-FULL and MMT.

7.4 Mixed-effects analysis

In order to quantify the effects of ASR errors on each system, we build linear mixed-effects models (Searle, 1973) in a similar manner to our mixed-effects analysis Section 4.1 in Chapter 4. We construct two sets of mixed-effects models, using the word error rate scores of the 8 ASR hypotheses as independent variables and the resulting increase in translation errors ΔTER as the response variable. The models contain random effect intercepts that account for the variance associated with the ASR system (SysID), the intrinsic difficulty of translating a given utterance (UttID), and a random effects slope accounting for the variability of word error rate scores (WER) across systems. Instead of

	WER-only (null model)			
	NMT-FULL		MMT	
Fixed effects	β	Std. Error	β	Std. Error
(Intercept)	4.35×10^{-3}	2.68×10^{-3}	-2.08×10^{-5}	1.92×10^{-3}
WER	6.09×10^{-1}	1.98×10^{-2} •	5.58×10^{-1}	1.85×10^{-2} •
Random effects	Variance	Std. Dev.	Variance	Std. Dev.
UttID (Intercept)	5.64×10^{-3}	7.51×10^{-2}	2.56×10^{-3}	5.06×10^{-2}
WER	2.33×10^{-1}	4.82×10^{-1}	2.24×10^{-1}	4.73×10^{-1}
SysID (Intercept)	0	0	0	0
Residual	5.45×10^{-3}	7.38×10^{-2}	3.82×10^{-3}	6.18×10^{-2}

	WER_{basic} (Levenshtein alignment errors)			
	NMT-FULL		MMT	
Fixed effects	β	Std. Error	β	Std. Error
(Intercept)	4.87×10^{-3}	2.69×10^{-3}	-5.76×10^{-5}	1.93×10^{-3}
WER.S	6.80×10^{-1}	2.10×10^{-2} •	5.35×10^{-1}	1.96×10^{-2} •
WER.D	4.28×10^{-1}	2.41×10^{-2} •	5.94×10^{-1}	2.20×10^{-2} •
WER.I	5.59×10^{-1}	3.01×10^{-2} •	5.98×10^{-1}	2.68×10^{-2} •
Random effects	Variance	Std. Dev.	Variance	Std. Dev.
UttID (Intercept)	5.73×10^{-3}	7.57×10^{-2}	2.58×10^{-3}	5.08×10^{-2}
WER	2.33×10^{-1}	4.83×10^{-1}	2.26×10^{-1}	4.75×10^{-1}
SysID (Intercept)	0	0	0	0
Residual	5.29×10^{-3}	7.28×10^{-2}	3.81×10^{-3}	6.17×10^{-2}

Table 7.4: Mixed-effects summary, comparing Neural MT (NMT-FULL)'s tolerance against ASR errors as compared to Phrase-based MT (MMT). Top: Fixed and random effects for models modeling the WER score as a single predictor of translation ΔTER. Bottom: Fixed and random effect for models decomposing WER into the basic alignment error operations. Random intercepts account for variances by utterance (*UttID*) with a random slope associated with the WER score, and by ASR system (*SysID*). Statistical significance at $p < 10^{-4}$ is marked with •.

treating each different MT system as a random effect in a joint mixed-effect model, we construct a mixed-effects model for each MT system with the purpose of comparing the degree to which each ASR error type explains the increase in translation difficulty. The models are build using R Core Team (2013) and the *lme4* library (Bates et al., 2014).

Our first models, (*WER-only*), use the raw WER score as a predictor of ΔTER. Our second models break WER into its substitution (S), deletion (D), and insertion (I) error alignments. Although we propose the use of phonetic substitution spans in our analysis of Chapter 4, we leave them out here because these preliminary experiments do not account for the syntactic properties of the errors. The phonetically-oriented alignment approach of Ruiz and Federico (2015) is useful for future analyses that focus the linguistic properties of the error types, which requires more precise word-level alignments than offered by the conventional bag of alignment errors computed in conventional WER.

The fixed-effects coefficients and the variance of the random effects for each model are shown in Table 7.4.

WER score only Our first NMT-FULL and MMT mixed-effects models focus on the effects of the global WER score on translation quality (ΔTER). Our fitted models claim that each point of WER yields approximately the same change in ΔTER for NMT-FULL and MMT (roughly 0.61 ± 0.02, versus 0.056 ± 1.9 for NMT-FULL versus MMT, respectively).

ASR Levenshtein error alignments Our second model breaks up WER into a bag of substitution (S), deletion (D), and insertion (I) error alignment counts, each being normalized by the length of the reference transcript. In particular, we observe that NMT-FULL assigns higher penalties to substitution errors than MMT. This observation holds for examples as shown above, where small substitution errors can cause long distance translation effects and dropped words.

7.5 Chapter Summary

We have introduced a preliminary analysis of the impact of ASR errors on SLT for models trained by neural machine translation systems. In particular, we identify the following as areas to focus on in new research in evaluating NMT for spoken language translation scenarios: (1) contextual patterns not observed during training – SMT systems usually can back off to shorter sized entries in their translation table; NMT behavior can be erratic. (2) localized and minor ASR errors can cause long distance errors in translation. (3) NMT duplicates content words when minor ASR errors cause the modification of function words.

As a preliminary work, there are many areas to expand upon. To analyze the effects of the phenomena observed above, we recommend several experiments. We observed examples in the paragraphs above where the content word “body” was either passed through during decoding or omitted altogether. Without a deeper look into the behavior of the decoder, is not clear why the NMT decoder would not attempt to translate the misrecognized phrase “I body” from Utterance U85 in Figure 7.3 similar manner as MMT’s hypothesis “je corps”, nor is it clear why multiple contexts containing “body architect” drop the word “body” altogether in the translation.

Perhaps a more interesting problem is to understand the effects of minor substitution, deletion, and insertion errors involving short function words on translation quality. In Chapter 4, we showed that these error types have a strong impact on translation quality. We showed in Section 7.3.2 an example of a noun being overproduced due to the substitution of a conjunction with a determiner. In phrase-based machine translation systems, isolated instances of these error types affect the translations of phrases in a small window of context. However, a PBMT system uses hypothesis stacks that words already covered in each step of translation decoding. Tu et al. (2016) recently identified this problem and proposed a coverage vector to keep track of the history of previously translated words to help the attention model address untranslated source words. Appendix Table A.5 consistently lists substitution errors between determiners, prepositions, pronouns, and conjunctions among the top 10 substitution error types present in the tst_{2012} dataset. We recommend an ablation test where function words are randomly inserted and dropped from sentences to measure their effects on NMT translations. This could also be a good use for our ASR damaging channel, proposed in Chapter 6, which synthesizes ASR-like errors on text data.

Thus far, we have modeled NMT-FULL, which uses full-word representations. We intend to explore the effects of input and output representations on error tolerance. For example, does subword unit or character-based modeling allow the attention mechanism to reach over ASR errors due to phonetic confusions?

We will expand this analyses in this chapter further to include mixed-effects models that account for errors that can be tolerated with subword units.

CHAPTER 8

CONCLUSION

Spoken language translation operates in the intersection between automatic speech recognition (ASR) and statistical machine translation (SMT). Currently the most widely accepted approach of combining ASR and SMT is by treating SMT as a downstream task that receives the outputs of ASR and decodes it. This approach has the advantage that enables each component to be trained independently, given the paucity of bilingual training corpora that simultaneously contains audio recordings, recording transcripts, and translated transcripts. The overall training time for reduced as each component may relax its dependency assumptions and focus on local optimization against its own evaluation metric – e.g. word error rate for ASR; BLEU (Papineni et al., 2002) or TER Snover et al. (2006) for SMT.

However, the ease of training comes at a cost. Conventional SMT systems do not model noisy input texts as part of their training process, in part due to the paucity of ASR training data that simultaneously contains a reference translation. Instead, the SMT system expects to receive clean, well-formed source language texts that match its system training conditions. While the SMT system has been trained to generate fluent output on well-formed sentences, the search space of adequate translations is inaccessible in the presence of speech recognition errors, as the statistics representative of noisy texts are too sparse in the SMT models to generate coherent sentences. Additionally, the introduction of ASR substitution errors on content words in the source utterance can transform the meaning of the utterance, rendering it impossible for SMT to provide a correct translation, largely because under normal training conditions the SMT has no

way to model phonetic confusions of similar sounding words.

In order to understand and respond to these deficiencies in existing spoken language translation (SLT) models, we performed several analyses, comparing the difficulty of translating speech versus text, and proposed techniques to (1) incorporate shallow discourse context to adapt machine translation models toward lectures, and (2) to model the noisy channel scenario in MT training by synthesizing ASR errors. Our analyses and experiments were evaluated primarily on the translation of TED talks from English to French.

In our preliminary analyses, we evaluated the differences in the task of translating TED lectures as a representative of spoken discourse versus a comparable size of news commentaries collected for evaluation at the Workshop on Machine Translation. In the translation from English to German, we observed that while spoken language registers such as TED talks are comprised of shorter sentences with less reordering behavior and stronger predictability than written news registers, the increased occurrence of anaphora require translation systems to be mindful of antecedents within the discourse to ensure that gender, number, and case agree in its translations. Likewise, the increased usage of polysemous verbs and nouns can be a boon or a curse, depending on the language pair – for languages that use similar common words to express the same concepts, this behavior makes translation easier, while for language pairs such as English-Mandarin, this can be problematic (Palmer and Wu, 1995).

We next extended our analysis into the realm of spoken language translation by analyzing the impact of speech recognition errors on the translation quality of a machine translation that has not been adapted for the English to French spoken language translation scenario. We performed a linear mixed effects regression analysis, using the submissions of eight automatic speech recognition hypothesis sets for the IWSLT 2013 TED English ASR track and evaluate a strong SMT baseline system's degradation, measured by the increase in translation edit rate (TER) scores against the translation of clean texts. We aligned the ASR word recognition errors between the ASR hypotheses and their reference transcripts by using a custom Levenshtein alignment process that involves phonetic alignments and annotated each word with its word class information. We observed a significant effect of substitution errors involving phonetically similar spans of one or more words (e.g. *anatomy⇒and that to me*) on translation quality and used this information as inspiration for error modeling in SMT.

Next we addressed the need for spoken language translation systems to use shallow discourse context to adapt the machine translation models to improve consistency and

precision through an approximate language model adaptation approach based on bilingual latent semantic analysis that fits well in the log-linear framework for conventional statistical machine translation systems. Our approach works to adapt the MT models to account for word distribution changes based on minor shifts in topic that occur during lectures such as TED talks, based on a small sliding window of context. We demonstrate that both the monolingual input strings and the translations of previous segments can improve translation accuracy.

As an error modeling strategy for SMT, we developed an *ASR damaging channel* which uses machine translation strategies to decompose textual machine translation training data into ASR-like outputs by minimally supervising the ASR errors caused by acoustic confusions on a small development set. The damaging process first normalizes all textual data into phonemes and reconstructs words by building a translation model from an ASR system's pronunciation dictionary and language model. The damaging channel was applied to entire collections of text corpora and the resulting outputs were added to the training corpora available for MT training. Evaluation on English-French TED talk translation demonstrated that the inclusion of damaged training data improves the robustness of a MT system with respect to ASR errors.

In our outlook for the future of spoken language translation, we turned our attention to neural machine translation (NMT) and performed a preliminary assessment of NMT's inherent robustness against ASR errors, in comparison to conventional phrase-based machine translation by studying the impact of ASR errors on MT, using the same TED English-French SLT dataset as described above. Our preliminary results identified NMT's sensitivity to localized ASR errors on function words; given that NMT allows long distance word dependencies to be modeled, the lexical translation of content words can be affected by minor ASR errors on constituents belonging to a different region of the sentence. We additionally observed duplications of content word translations, caused by minor ASR substitution and insertion errors. While NMT consistently yielded higher automatic translation quality scores than our strong phrase-based machine translation baseline, the errors found in our initial experiments warrant deeper analysis to understand how even minor ASR errors can cause significant adequacy issues for NMT.

As we look forward to the future of spoken language translation in light of the emergence of deep neural networks in speech recognition and machine translation. Already, ASR systems are achieving unprecedented recording in recognition accuracy on a number of tasks, including the recent report by Xiong et al. (2016) on conversational ASR. While ASR accuracy continues to increase with the latest neural architectures, there re-

mains a number of languages and environment contexts which ASR systems continue to struggle with; thus, we are confident that error modeling approaches similar to those proposed in this thesis will continue to be necessary for increased spoken translation performance.

It is exciting to be living in a time where machine translation has become a viable solution in many interaction contexts. In addition to leading research laboratories such as Microsoft Research, Google, and Systran's flagship translation applications, the consumer market is starting to fill with products including phone translation applications, wearable translation devices, and optical character recognition and translation. We are excited to be among the community of researchers who are actively working toward breaking down communication barriers through advancements in speech and language technologies.

APPENDIX A

SPOKEN LANGUAGE TRANSLATION ERROR ANALYSIS NOTES

Included are some extra details regarding our SLT error analysis experiments on the TED data from Chapter 4.

A.1 Experiment data

Fig. A.1 provides an example of the hierarchical nature of random effects in a mixed-effects model. The fixed effects are the measurable results of a particular speech recognition output; however our mixed-effects model accounts for the variance inherit in a particular ASR system or utterance. For each ASR system, we decode a sample of 580 utterances; thus the mixed-effects model needs to account for the hierarchical nature of random effects. By controlling for random variance, the model is assumed to generalize better for any arbitrary ASR system and utterance that resembles English TED talks than a regression model trained on a single ASR system.

A.2 Data preparation

IWSLT's ASR submissions are in lowercase, lack punctuation, and do not have embedded segmentation. We use the segmentation file provided in the SLT track to induce segmentation. After segmentation, we use the documentation provided in the IWSLT

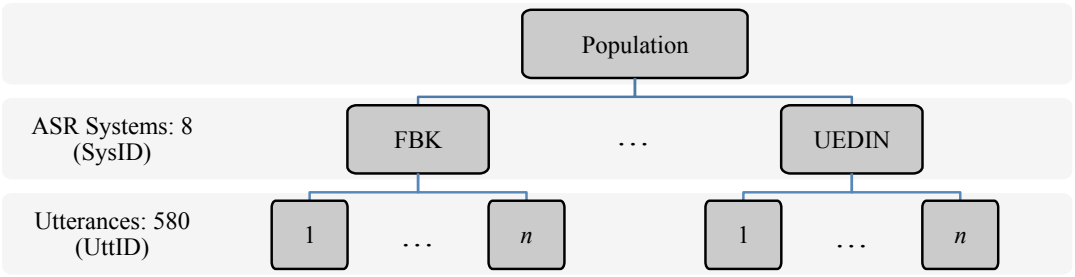


Figure A.1: Illustration of hierarchical random effects in our SLT quality experiments. The random effects represent the eight ASR systems used in the IWSLT 2014 evaluation, which are drawn from a theoretically large population of ASR systems. For each system, we transcribe 580 utterances derived from the audio samples of tst_{2012}.

evaluation campaign to find and match each source transcript and ASR hypothesis with the tst_{2012} set from the MT track. While we follow a similar approach to our previous paper, we revisit a number of preprocessing steps in order to minimize the effects of the following phonemena:

- Reference-oriented text normalization
  - British to American English lexical normalization
  - Hyphenation
  - Word compounding
  - Contractions/Expansions
  - Number and date normalization
- Word tokenization
- Punctuation
- Recasing

In the case of text normalization and punctuation, we rely on the punctuated ASR references (the original TED transcripts) to provide guidance on how to properly preprocess the ASR hypotheses. We do the data gathering process in this order:

1. Normalize British English words into American English words on the ASR hypothesis;

2. Pre-compute WER and POWER and their Levenshtein alignments, before other normalization steps;

3. Apply text normalization to the ASR hypothesis with respect to the words in the unpunctuated ASR reference, based on their Levenshtein alignment;

4. Re-compute WER and POWER and their Levenshtein alignment;

5. Perform POS tagging on the reference and hypothesis; bind the POS tags to the Levenshtein alignments;

6. Apply "oracle" punctuation insertion on the ASR hypothesis by borrowing the punctuation from the reference, based on Levenshtein alignment;

7. Tokenize the reference and hypothesis;

8. Recase the hypothesis;

9. Translate the hypothesis and reference with our baseline SMT system; compute evaluation scores (e.g. BLEU, TER, and METEOR);

10. Compute ΔBLEU, ΔTER, ΔMETEOR for each sentence;

11. Collect delta MT metrics, WER, and broken-down WER scores for WER- and POWER-based error alignments.

We explain each of the steps in further detail below.

A.2.1 Text normalization

Prior to evaluating hypotheses from ASR systems, the DARPA Hub-4 evaluation plan (Pallett et al., 1998) and subsequent ASR evaluations such as NIST's Rich Transcription tasks (Garofolo et al., 2002) used an evolving normalization script to prevent penalization for minor orthographic variations such as multiple spellings (e.g. British vs. American English), compound words (e.g. "storyline" vs. "story line"), and contractions (e.g. "it's" vs. "it is"). Assuming that a phrase-based SMT system in the SLT pipeline is trained on ASR reference transcripts, orthographic variances in ASR outputs can result in out-of-vocabulary words or under-represented source language n-grams in the translation model, further degrading machine translation quality. Although both the ASR

hypotheses and the reference transcripts were normalized in prior evaluations, our experiments require the ASR reference to remain unmodified in order to properly evaluate the translation of ASR outputs against translation of the original TED transcripts.

British to American English Although the TED talks are transcribed in American English most of the ASR submissions use British English, We use the varcon tool from SCOWL[1] to convert British English words to American English. SCOWL is a database of information on English words useful for creating word lists suitable for spell checkers of various English dialects.

Contractions, numbers, acronyms, and abbreviations In order to carry out the remainder of the text normalization, we require that the hypothesis and reference words are orthographically consistent. We use SCLITE to perform an initial alignment between the reference and hypothesis and use the alignments to find the orthographic differences in the hypothesis that need correction.

A.2.2 Re-alignment of errors

After applying the proper text normalization, we recompute WER and POWER with their associated word alignments.

A.2.3 Punctuation insertion

Again, we rely on the POWER alignment to apply punctuation on the hypothesis. We insert punctuation in two steps. First, we apply a character-based alignment between the unpunctuated and the punctuated reference to segment the punctuation, which appear as insertion errors. Then we pivot around the POWER alignment to assign the punctuation to a particular slot. If punctuation appears at the end of the reference utterance, it is forced to the end of the hypothesis utterance.

A.2.4 Recasing

We trained a recaser using the Moses tools on the IWSLT 2013 English TED training data and apply it to the normalized, punctuated ASR hypotheses.

[1] https://github.com/kevina/wordlist

A.2.5 Translation and evaluation

We use the defualt English-French TED-only system trained by the providers for the WIT^3 corpus. The system is trained on the English-French training sets for the MT track in IWSLT 2013. We translate each of the normalized, punctuated, and recased ASR hypotheses, as well as the punctuated ASR referenceerence and evaluate using MultEval v0.3 (Clark et al., 2011) to compute BLEU, TER, and METEOR, both on the sentence level and the summary level. In order to measure the effects of particular ASR errors on translation quality, we are interested in the amount each sentence-level metric changes when ASR errors are introduced.

Using TER as an example, we control for the difficulty of translating an otherwise perfect speech recognition hypothesis, we use the difference between the TER associated with translating the perfect ASR reference and the TER associated with translating the ASR hypothesis, labeled as ΔTER:

$$\Delta\text{TER} = \text{TER}_{\text{ASR}} - \text{TER}_{\text{gold}}, \tag{A.1}$$

where TER_{gold} is the TER score for a perfectly recognized utterance, and TER_{ASR} is the TER score on the translation of the ASR hypothesis. By using ΔTER, we assume that TER_{gold} is the upper-bound on translation quality with the given SMT system. The equations for ΔBLEU and ΔMETEOR are derived in the same way. Note that ΔTER is expected to be positive, while the others should be negative.

A.3 Outlier removal

Let's look at the relationship between ASR's WER and MT's ΔTER in order to better understand the data. This plot doesn't separate each ASR system (SysID) or each utterance (UttID), but it helps us to spot potential outliers. In principle, a increase in WER should yield an increase in TER. We see several outliers in the data. Some of the outliers may be due to some anomalies in the text normalization process described above, while others could be due to data that is extreme in nature. For example, a single error in a short utterance yields a high WER. However, due to deficiencies in our translation metrics, a single deletion error could actually improve the translation score, although the adequacy of the translation utterance would be poor.

Table A.1 shows some examples. The first example shows how a simple morphological error yields a better translation score; however, the main difference in the trans-

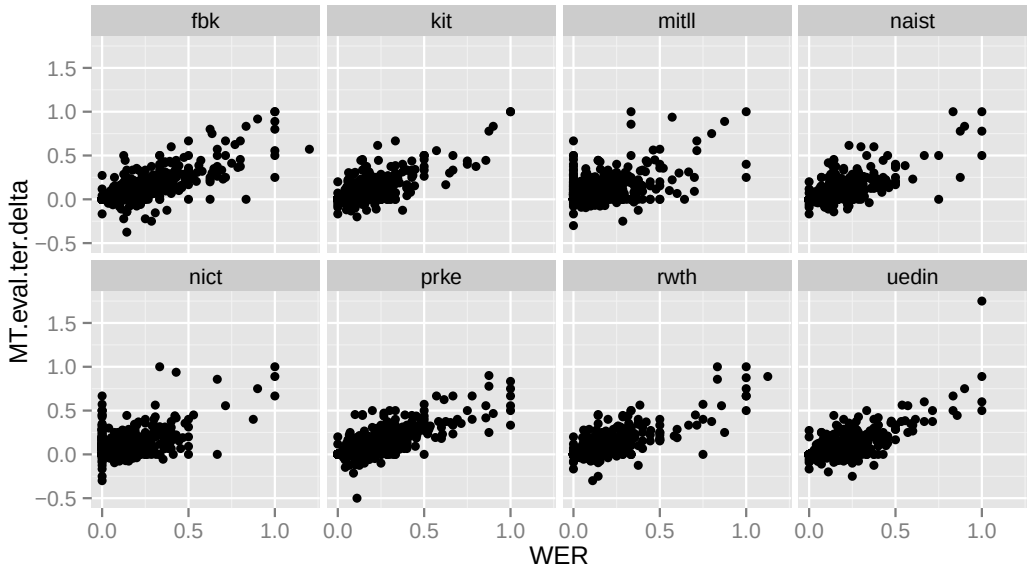


Figure A.2: ASR errors (WER) vs. change in MT errors (ΔTER) by ASR system, before outlier removal.

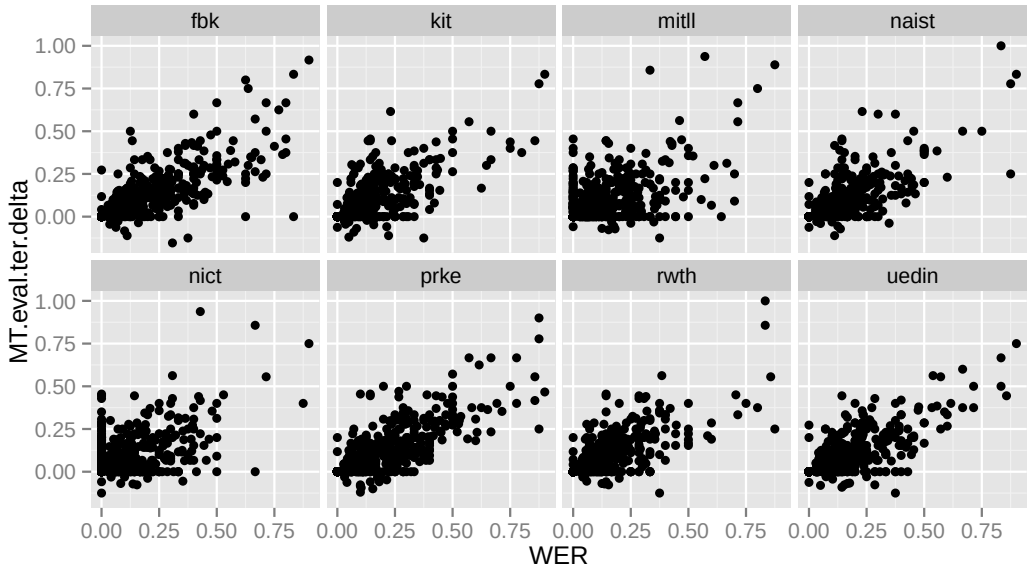


Figure A.3: ASR errors (WER) vs. change in MT errors (ΔTER) by ASR system, after outlier removal.

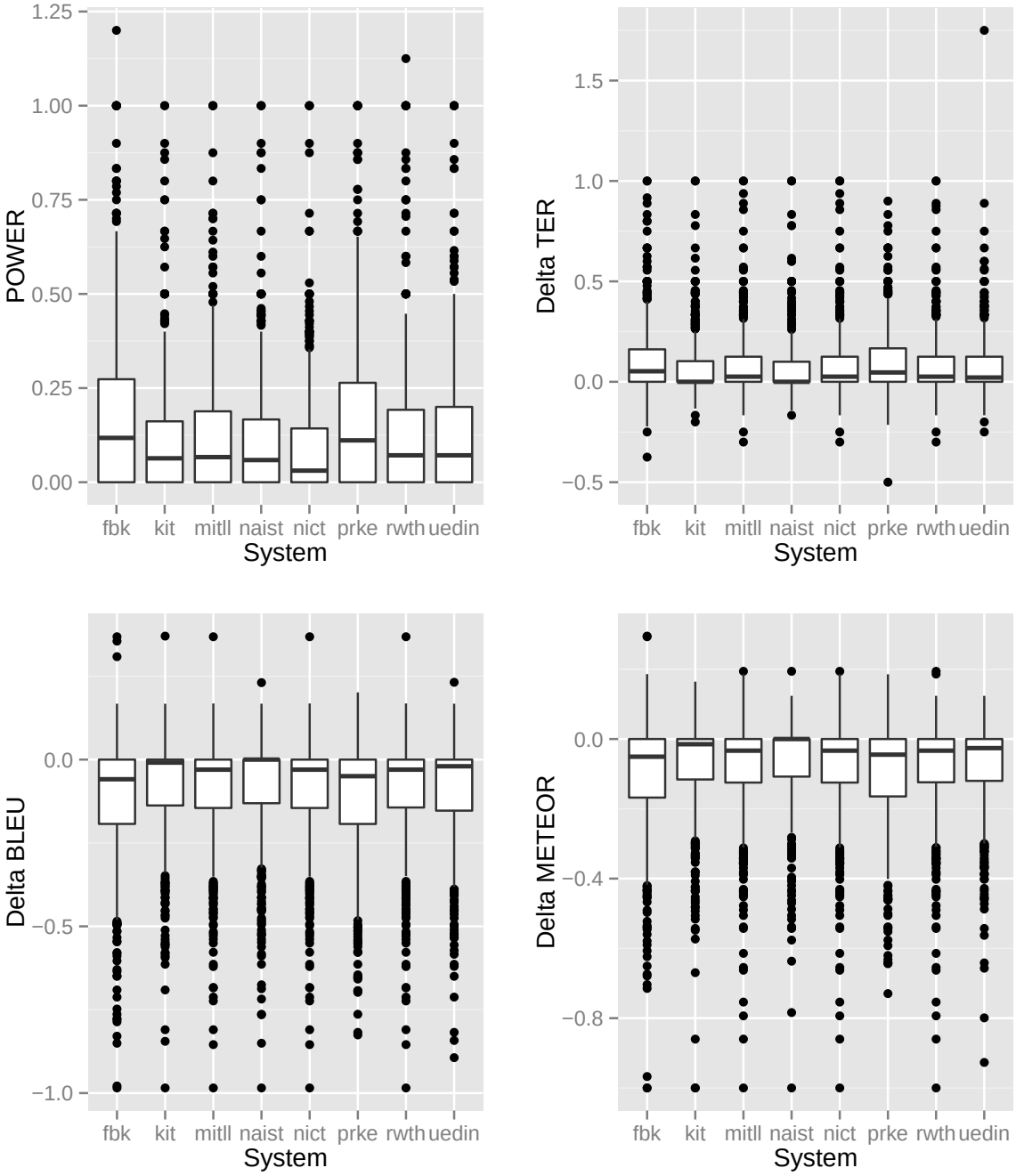


Figure A.4: (Pre-outlier removal) Boxplots describing the distribution of ASR errors (POWER) and their impact on translation errors by ASR system and utterance.

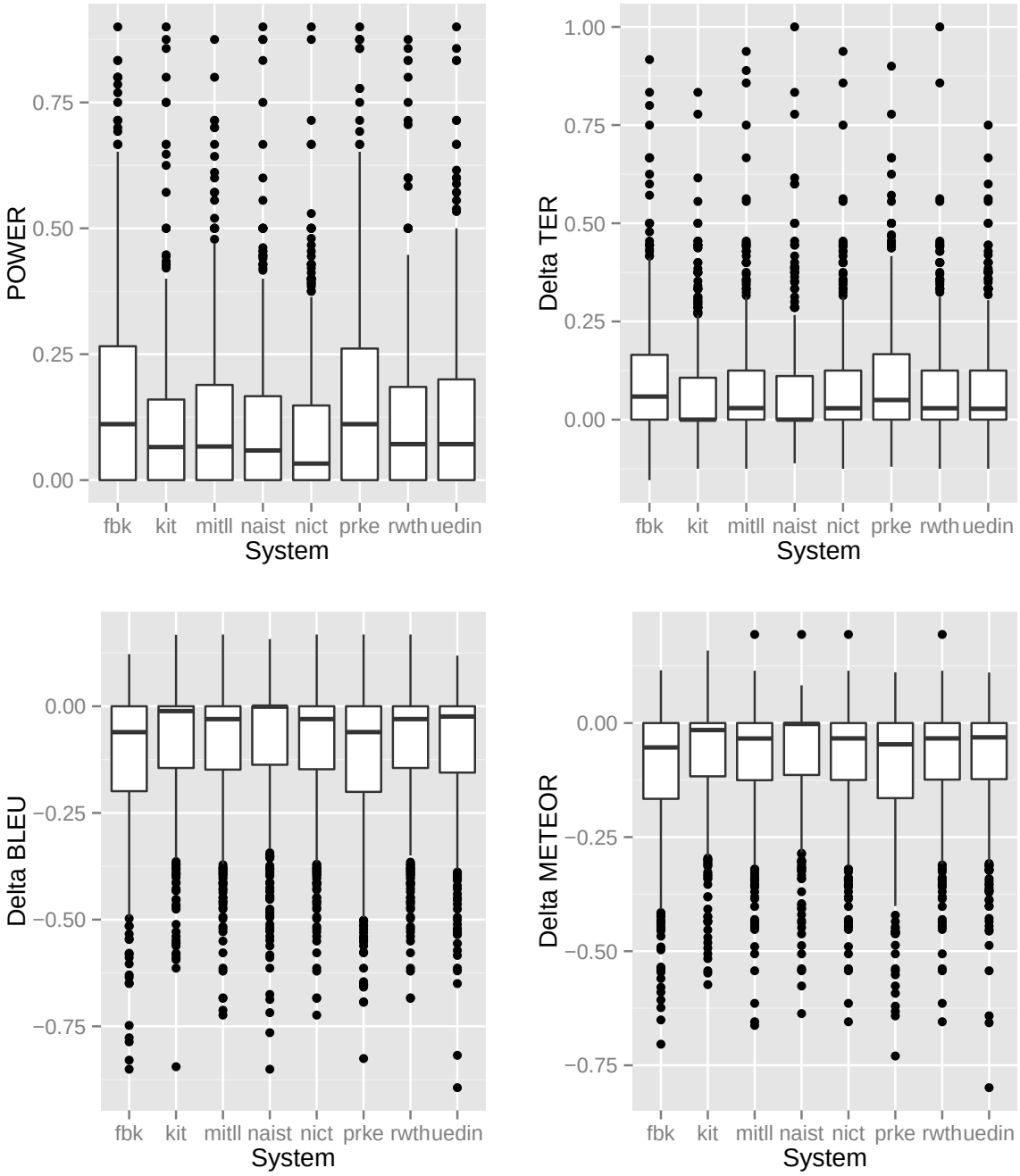


Figure A.5: (Post-outlier removal) Boxplots describing the distribution of ASR errors (POWER) and their impact on translation errors by ASR system and utterance.

	UEDIN #98 (WER: 0.25, ΔTER: -0.25)	NAIST #293 (WER: 0.14, ΔTER: -0.14)
ASR Ref	So case **closed** , right ?	**Our** disciplinary conventions were funny as well .
ASR Hyp	So case **close** , right ?	Disciplinary conventions were funny as well .
ASR Ref trans	Donc cas fermé , pas vrai ?	**Notre** disciplinary conventions étaient drôle aussi .
ASR Hyp trans	Donc cas près , non ?	Disciplinary conventions étaient drôle aussi .
French Ref	La discussion est close , non ?	**Nos** conventions disciplinaires étaient drôles aussi .

Table A.1: Outlier examples with negative Δ TER scores from tst2012.

lation is the translation of the tag question, "right?", which is likely due to a weakly-trained SMT system.

The second example demonstrates an ASR system dropping a possessive determiner. TER penalizes a mistranslation of "our" in the ASR reference ("notre"), whose object is plural: the substitution error "nos" → "notre" is penalized over a deletion error. Likewise, the ratio of matched n-grams is affected by utterance-level BLEU and METEOR.[2]

In order to correct them, we discard entries that meet one or more of the following criteria:

- $WER \geq 1$: 36 utterances
- Less than 5 words in the ASR reference: 23 utterances per lab
- Greater than 40 words in the ASR reference: 20 utterances per lab
- Delta MT metrics in the wrong direction (ΔBLEU > 0 and ΔMETEOR > 0 and ΔTER < 0): 111 utterances

In total, we remove 471 data points. Our primary summary statistics are the following are listed in Table A.2.

A.4 Word class clustering

As linguistic annotations, we consider each word's word class, as well as a generalized categorization of part-of-speech (POS) tags, which simplifies the Penn Treebank POS labels into nine classes. Table A.3 lists our mapping of Penn Treebank POS tags to open and closed classes.

For substitutions and substitution spans, linguistic annotations are reported first for the reference word(s), followed by the annotations for the hypothesis word(s). For

[2]Again, as an aside, the WIT^3 MT system is not trained with enough examples to correctly translate "disciplinary".

Statistic	Mean	St. Dev.	Min	Median	Max
ASR.hyplen	16.390	8.360	3	15	42
ASR.reflen	16.344	8.291	5	15	40
WER.basic.S	0.942	1.345	0	0	10
WER.basic.SS	0.606	1.343	0	0	13
WER.basic.D	0.254	0.672	0	0	14
WER.basic.I	0.192	0.596	0	0	7
WER	0.128	0.162	0.000	0.071	0.900
MT.eval.ter.delta	0.085	0.130	−0.154	0.030	1.000
MT.eval.bleu.delta	−0.102	0.150	−0.894	−0.033	0.168
MT.eval.meteor.delta	−0.085	0.122	−0.799	−0.033	0.194

Table A.2: Summary data of key statistics after outlier removal. Levenshtein error counts are provided instead of % contribution to POWER.

Type	Members
open	noun ∪ verb ∪ adv ∪ adj
closed	prn ∪ prep ∪ coord ∪ det ∪ aux
noun	{NN NNP NNPS NNS POS CD}
verb	{VB VBD VBG VBN VBP VBZ}
aux	{MD}
adv	{RB RBR RBS WRB}
prn	{PRP PRP$ WP WP$ EX}
coord	{CC}
prep	{IN RP TO}
det	{DT PDT WDT}
adj	{JJ JJR JJS}

Table A.3: Mapping of Penn Treebank POS tags to word classes and general POS classes.

example, the ref→hyp alignment of “know”→“a” is annotated as S.open_closed when considering word classes, and S.verb_det when considering general part-of-speech (POS) clusters. For deletions, linguistic annotations are on the deleted reference word, while for insertions, linguistic annotations are on the inserted hypothesis word.

A.5 POWER vs WER: Word class-annotated errors

Given the inconsistent error labeling in WER, which types of errors are actually being skewed by false alignments? To answer this question, we annotate the reference

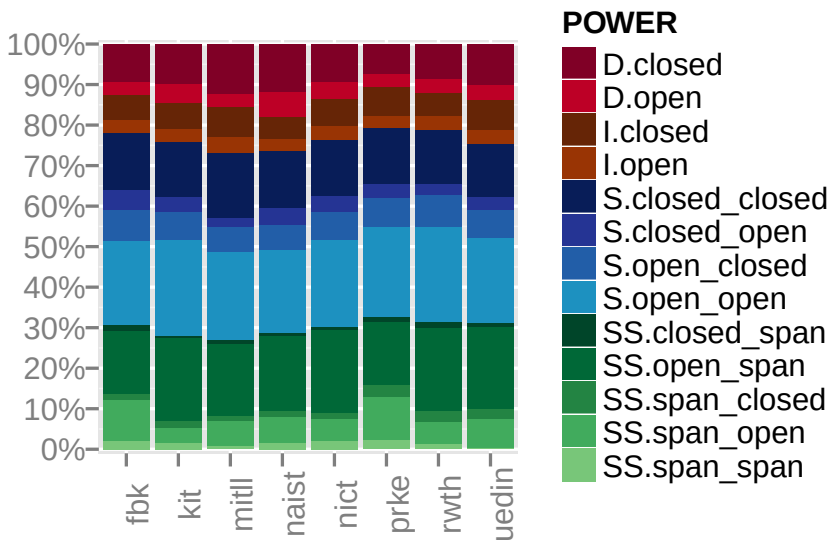


Figure A.6: Distribution of error types by word class for POWER for each IWSLT 2013 ASR evaluation participant.

and hypothesis words by their word class and observe their alignment statistics. We additionally apply lemmatization to distinguish morphological errors from other substitution types. According to the word statistics in Table **??**, the ratio of open to closed class words remains the same across each ASR hypothesis and the reference (gold).

The distribution of word class-annotated ASR errors for each IWSLT 2013 participant is shown in Fig. A.6. We observe an asymmetrical behavior by each of the ASR systems: while an ASR system will commonly misrecognize an open class word as a sequence of shorter words; the opposite rarely occurs. This is due to the fact that the misrecognized hypothesis words are usually frequently-occurring words that are well-represented by a n-gram language model, while the misrecognized reference words are often domain-specific nouns or verbs. As the ASR system more closely fits the domain of the speaker, the number of phonetic substitution error spans will decrease. We also observe that the majority of substitution errors are within the same word class.

The proportion of errors associated with each ASR error type is shown in Table A.4.

Word-level substitution errors. Both WER and POWER report that the majority of substitution errors are within the same word class. While the proportion of closed-closed class substitutions remain the same, POWER reports 8% fewer open-open class substitution errors, which are often instances of substitution error spans containing a word-level substitution error and one or more short function words (e.g. *Brown in→brahmin* from Fig. 4.2). Of the open-open class substitution errors, 5.4% are morphological errors.

ErrorType	WER	POWER	WER Rank	POWER Rank
S.open_open	0.299	0.219	1	1
SS.open_span		0.186		2
S.closed_closed	0.148	0.140	2	3
D.closed	0.112	0.097	4	4
S.open_closed	0.107	0.069	4	6
SS.span_open		0.069		6
I.closed	0.101	0.065	6	7
D.open	0.067	0.041	8	8
S.closed_open	0.069	0.036	8	8
I.open	0.096	0.033	6	10
SS.span_closed		0.019		12
SS.span_span		0.016		12
SS.closed_span		0.010		13

Table A.4: Proportion of ASR error types by word class, averaged across all ASR systems and ranked by importance. Substitution labels (S, SS) show the alignment from reference class to hypothesis class. Substitution spans (SS) contain a *span* of words aligned either to a single word or another span.

POWER likewise reports 7% fewer cross-class substitution errors, many of which are attributed to the correction of misalignments.

Table A.5 provides the 10 most common substitution errors by general POS type for each ASR system, using POWER as the evaluation metric. In particular, we observe that nouns and verbs are confused with themselves. Of the substitution errors, 17.4%(±0.7%) are instances of nouns being confused with other nouns, and 13.9%(±0.3%) are instances of verbs being confused with other verbs. We also observe that nouns and verbs are commonly confused with one another (2.8% ± 0.2% are reference verbs misrecognized as nouns and 2.8% ± 0.2% is vice-versa).

Aside from nouns and verbs, the majority of the most common substitution errors involving hard-to-recognize function words. Determiners are commonly confused with one another 4.1%(±0.2%), which is the third most common substitution error. Other confusable closed class types include prepositions to determiners (3.0% ± 0.2%) and coordinations to prepositions (3.3% ± 0.3%). Incidentally, the converse is less likely to occur. Reference determiners are less likely to be confused with prepositions (1.1%±0.2%), and reference prepositions are less likely confused with coordinations (1.8% ± 0.1%).

It is interesting to note that adjectives and nouns are commonly confused (2.1% ± 0.1% in the reference-hypothesis direction and 1.4% ± 0.2% in the opposite direction). The remaining proportion of substitution errors are tapered off by similarly common

	fbk.Err	fbk.Pct	kit.Err	kit.Pct	mitll.Err	mitll.Pct	naist.Err	naist.Pct
1	S.noun_noun	0.131	S.noun_noun	0.198	S.noun_noun	0.167	S.noun_noun	0.177
2	S.verb_verb	0.126	S.verb_verb	0.138	S.verb_verb	0.146	S.verb_verb	0.140
3	S.det_det	0.041	S.det_det	0.044	S.det_det	0.054	S.prep_prep	0.040
4	S.verb_noun	0.038	S.prn_prn	0.040	S.coord_prep	0.044	S.prn_prn	0.038
5	S.prep_det	0.033	S.coord_prep	0.037	S.prep_prep	0.042	S.det_det	0.036
6	S.verb_prep	0.032	S.verb_noun	0.029	S.prn_prn	0.035	S.coord_prep	0.034
7	S.noun_verb	0.031	S.verb_prn	0.029	S.prep_det	0.031	S.prep_det	0.034
8	S.prn_det	0.031	S.adj_noun	0.025	S.verb_prep	0.031	S.verb_prep	0.030
9	S.prep_prep	0.029	S.verb_prep	0.023	S.adj_noun	0.027	S.noun_verb	0.028
10	S.coord_prep	0.025	S.prep_prep	0.021	S.verb_noun	0.025	S.verb_prn	0.026
	nict.Err	nict.Pct	prke.Err	prke.Pct	rwth.Err	rwth.Pct	uedin.Err	uedin.Pct
1	S.noun_noun	0.184	S.noun_noun	0.168	S.noun_noun	0.183	S.noun_noun	0.183
2	S.verb_verb	0.132	S.verb_verb	0.148	S.verb_verb	0.147	S.verb_verb	0.139
3	S.coord_prep	0.040	S.prep_prep	0.047	S.det_det	0.038	S.verb_prn	0.038
4	S.det_det	0.040	S.det_det	0.039	S.prn_prn	0.034	S.verb_noun	0.037
5	S.prep_det	0.038	S.prn_prn	0.031	S.verb_prep	0.034	S.coord_prep	0.035
6	S.prep_prep	0.038	S.coord_prep	0.027	S.noun_verb	0.033	S.det_det	0.035
7	S.noun_verb	0.035	S.prep_det	0.026	S.prep_det	0.029	S.prn_prn	0.033
8	S.prn_prn	0.031	S.verb_prep	0.026	S.prep_prep	0.027	S.noun_verb	0.031
9	S.verb_noun	0.028	S.verb_noun	0.025	S.verb_prn	0.025	S.prep_det	0.031
10	S.verb_prep	0.028	S.noun_verb	0.023	S.verb_noun	0.024	S.prep_prep	0.031

Table A.5: Top 10 substitution error types for each research lab's ASR system for each system, clustered by POS tag (POWER).

errors.

Deletions and Insertions. According to WER, deletion and insertion errors account for 37.7%(±0.7%) of all errors. WER marks nearly as many open class insertions as closed class insertions, but suggests that closed class deletions are more prominent than open class ones (6.7%±0.4% open versus 11.2%±0.5% closed class deletions). However, with POWER, deletion and insertion errors only account for 23.6%(±0.8%) of all errors, with the majority of the reduction attributed to fewer open class insertion errors (3.3%±0.1%). An example of a corrected open class "deletion" is *Stanford→stamp or* from Fig. 4.4.

Substitution spans. The majority of substitution spans have a single open class reference word (18.6%±0.7%), such as *anatomy→and that to me* in Fig. 4.4; these represent the second most common POWER error type. Likewise, the presence of a substitution span in the ASR reference indicates that the hypothesis word is likely to be a content word (6.9%±0.8%). Closed class function words are unlikely to be aligned to substitution spans (2.9%±0.3%), since most have few syllables that cannot easily be mistaken for multiple words. Instead, as shown in Table A.4, closed class words are more likely to be deletion or insertion errors.

A.6 SLT evaluation

We now focus our attention on spoken language translation evaluation, where we seek to measure the impact of ASR errors on speech translation quality. As described in Chapter 4, we seek to measure how particular ASR error types confound a baseline MT system that has been trained on text data. To do so, we evaluate the difference between the MT quality score (in most cases, TER) computed on the punctuated and cased ASR references[3] and the ASR hypotheses of each ASR system described earlier.

We use linear mixed-effects models to measure the contribution of each ASR error type reported earlier to the change in the MT error metric. To make this more concrete, let's focus for a moment on a particular scenario, using WER as a baseline. First, we can measure the contribution of the general WER score of an ASR hypothesis to the change in the TER score for the associated translation output (ΔTER). Then, segmenting WER into its Levenshtein error types (S, D, I) and normalizing each by the ASR reference length, we have an exact construction of the features that compose the WER score. By using these in a subsequent mixed-effects model, we can test whether one type (e.g. substitution errors) have a greater impact on ΔTER than insertion errors. Taking it even further, we can deconstruct the general Levenshtein error types into error types with linguistic properties, such as word classes, similar to the process carried out in Section 4.2.4.

In the ASR-only experiments, we evaluated the frequency of error types. For our models, we will normalize the error types by their contribution to WER or POWER, in order to normalize the scale for each utterance. Table A.6 lists the fixed effects variables and their coefficients for the mixed-effects models described in Chapter 4.

[3]These correspond to the exact MT source data used in the evaluation

	WER	WER_{basic}	$POWER_{\text{basic}}$	WER_{wc}	$POWER_{\text{wc}}$
	(1)	(2)	(3)	(4)	(5)
WER	0.63***				
	(0.59,0.67)				
WER.D		0.56***	0.61***		
		(0.51,0.62)	(0.56,0.67)		
WER.D.closed				0.44***	0.55***
				(0.36,0.51)	(0.47,0.62)
WER.D.open				0.67***	0.66***
				(0.59,0.75)	(0.58,0.74)
WER.I		0.71***	0.83***		
		(0.64,0.77)	(0.75,0.91)		
WER.I.closed				0.64***	0.72***
				(0.57,0.72)	(0.63,0.81)
WER.I.open				0.81***	1.04***
				(0.72,0.90)	(0.92,1.15)
WER.S		0.62***	0.65***		
		(0.58,0.67)	(0.60,0.70)		
WER.S.closed_closed				0.68***	0.76***
				(0.61,0.76)	(0.69,0.83)
WER.S.closed_open				0.75***	0.80***
				(0.66,0.84)	(0.69,0.91)
WER.S.open_closed				0.48***	0.58***
				(0.41,0.55)	(0.51,0.66)
WER.S.open_open				0.63***	0.59***
				(0.57,0.68)	(0.53,0.65)
WER.SS			0.54***		
			(0.49,0.58)		
WER.SS.closed_span					0.71***
					(0.56,0.87)
WER.SS.open_span					0.55***
					(0.49,0.60)
WER.SS.span_closed					0.55***
					(0.47,0.64)
WER.SS.span_open					0.45***
					(0.39,0.51)
WER.SS.span_span					0.69***
					(0.59,0.78)
Constant	0.001	0.001	−0.0001	0.001	0.0002
	(-0.003,0.004)	(-0.002,0.004)	(-0.003,0.003)	(-0.002,0.004)	(-0.003,0.004)
Observations	4,640	4,640	4,640	4,640	4,640
Log Likelihood	6,172.17	6,180.63	6,194.29	6,218.15	6,237.29
Akaike Inf. Crit.	-12,330.34	-12,343.26	-12,368.58	-12,408.30	-12,436.59
Bayesian Inf. Crit.	-12,285.24	-12,285.28	-12,304.15	-12,318.10	-12,314.18

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.6: Side-by-side comparison of linear mixed-effects models. Fixed effects coefficients and 95% confidence intervals are reported for WER score, the primitive Levenshtein error types (WER_{basic}, $POWER_{\text{basic}}$), and Levenshtein errors annotated by word class (WER_{wc}, $POWER_{\text{wc}}$). Word class error types are annotated by their reference→hypothesis word alignments.

BIBLIOGRAPHY

Alejandro Acero.
Acoustical and Environmental Robustness in Automatic Speech Recognition.
PhD thesis, Pittsburgh, PA, USA, 1991.
UMI Order No. GAX91-17502.

Gilles Adda, Martine Adda-Decker, Jean-Luc Gauvain, and Lori Lamel.
Text normalization and speech recognition in French.
In *Proceedings of the European Speech Communication Association (Eurospeech)*, Rhodes, Greece, September 1997.

Zeeshan Ahmed, Jie Jiang, Julie Carson-Berndsen, Peter Cahill, and Andy Way.
Hierarchical Phrase-Based MT for Phonetic Representation-Based Speech Translation.
In *Proceedings of the tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA), San Diego, CA*, 2012.

Yaser Al-Onaizan and Kishore Papineni.
Distortion models for statistical machine translation.
In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 529–536, Sydney, Australia, July 2006. Association for Computational Linguistics.

Joshua S Albrecht and Rebecca Hwa.
Regression for Sentence-Level MT Evaluation with Pseudo References.
Association of Computational Linguistics, page 296, 2007.

Sankaranarayanan Ananthakrishnan, Wei Chen, Rohit Kumar, and Dennis Mehay.
Source-Error Aware Phrase-Based Decoding for Robust Conversational Spoken Language Translation.
In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

Anthony Aue, Qin Gao, Hany Hassan, Xiaodong He, Gang Li, Nicholas Ruiz, and Frank Seide.
MSR-FBK IWSLT 2013 SLT System Description.
In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, December 2013.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao.
Domain Adaptation via Pseudo In-Domain Data Selection.
In *Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, United Kingdom, 2011.

R. Harald Baayen, Douglas J. Davidson, and Douglas M. Bates.
Mixed-effects Modeling with Crossed Random Effects for Subjects and Items.
Journal of memory and language, 59(4):390–412, 2008.

Bagher BabaAli, Romain Serizel, Shahab Jalalvand, Daniele Falavigna, Roberto Gretter, and Diego Giuliani.
FBK @ IWSLT 2014 - ASR track.
In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, USA, December 2014.

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan.
Goodness: A Method for Measuring Machine Translation Confidence.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Techologies*, pages 211–219, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio.
Neural Machine Translation by Jointly Learning to Align and Translate.
In *5th International Conference on Learning Representations*, San Diego, USA, 2015. ICLR.

Satanjeev Banerjee and Alon Lavie.
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*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Srinivas Bangalore and Giuseppe Riccardi.
A Finite-state Approach to Machine Translation.
In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies (NAACL)*, pages 1–8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

Srinivas Bangalore and Giuseppe Riccardi.
Stochastic Finite-State Models for Spoken Language Machine Translation.
Machine Translation, 17(3):165–184, 2002.
ISSN 1573-0573.

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker.
lme4: Linear mixed-effects models using Eigen and S4, 2014.
URL `http://CRAN.R-project.org/package=lme4`.
R package version 1.1-6.

Doug Beeferman, Adam Berger, and John Lafferty.
Cyberpunc: A lightweight punctuation annotation system for speech.
In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 2, pages 689–692. IEEE, 1998.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico.
Neural versus Phrase-Based Machine Translation Quality: a Case Study.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 257–267, 2016.

Nicola Bertoldi and Marcello Federico.
A New Decoder for Spoken Language Translation based on Confusion Networks.
In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pages 86–91, San Juan, Puerto Rico, 2005.

Nicola Bertoldi, Richard Zens, and Marcello Federico.
Speech Translation by Confusion Network Decoding.
In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1297–1300, Honolulu, HA, 2007.

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni.
Phrase-based statistical machine translation with pivot languages.

In *In Proceedings of the International Workshop on Spoken Language Translation*, pages 143–149, 2008.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico.
Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation.
In *Proceedings of the MT Summit XIV*, pages 35–42, Nice, France, September 2013a.

Nicola Bertoldi, M. Amin Farajian, Prashant Mathur, Nicholas Ruiz, and Marcello Federico.
FBK's machine translation systems for the IWSLT 2013 evaluation campaign.
In *Proc. of the 10th International Workshop on Spoken Language Translation*, December 2013b.

Nicola Bertoldi, Prashant Mathur, Nicholas Ruiz, and Marcello Federico.
FBK's Machine Translation and Speech Translation Systems for the IWSLT 2014 Evaluation Campaign.
In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, USA, December 2014.

Nicola Bertoldi, Davide Caroselli, David Madl, Mauro Cettolo, and Marcello Federico.
ModernMT Second Report on Database and MT Infrastructure.
Technical Report D.32, European Union Horizon 2020 research and innovation programme, December 2016.

Douglas Biber.
Variation Across Speech and Writing.
Cambridge University Press, Cambridge, 1988.

Alexandra Birch, Phil Blunsom, and Miles Osborne.
A quantitative analysis of reordering phenomena.
In *In Proceedings of the Fourth Workshop on Statistical Machine Translation (StatMT)*, pages 197–205, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

Arianna Bisazza and Marcello Federico.
Dynamically Shaping the Reordering Search Space of Phrase-Based Statistical Machine Translation.
Transactions of the Association for Computational Linguistics, 1:327–340, 2013.

Arianna Bisazza, Nick Ruiz, and Marcello Federico.
Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation.
In *In Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 136–143, San Francisco, CA, 2011.

Alan W. Black and Paul A. Taylor.
The Festival Speech Synthesis System: System documentation.
Technical Report HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK, 1997.
Avaliable at http://www.cstr.ed.ac.uk/projects/festival.html.

David M. Blei, Andrew Ng, and Michael Jordan.
Latent Dirichlet Allocation.
Journal of Machine Learning Research, 3:993–1022, 2003.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin.
A statistical approach to machine translation.
Computational Linguistics, 16(2):79–85, 1990.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer.
The Mathematics of Statistical Machine Translation: Parameter Estimation.
Computational Linguistics, 19(2):263–312, 1993.

Christian Buck.
Black Box Features for the WMT 2012 Quality Estimation Shared Task.
In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 88–92, Montreal, Canada, June 2012. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne.
Improved Statistical Machine Translation Using Paraphrases.
In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA, June 2006. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder.
Findings of the 2009 Workshop on Statistical Machine Translation.
In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March 2009. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia.
Findings of the 2012 workshop on statistical machine translation.
In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June 2012. Association for Computational Linguistics.

Steven A. Camarota and Karen Zeigler.
Nearly 65 Million U.S. Residents Spoke a Foreign Language at Home in 2015.
Center for Immigration Studies, October 2016.
URL `http://cis.org/sites/cis.org/files/camarota-language-16.pdf`.

Marine Carpuat.
One Translation Per Discourse.
In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, pages 19–27, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
ISBN 978-1-932432-31-2.

Marine Carpuat and Dekai Wu.
Improving statistical machine translation using word sense disambiguation.
In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 61–72, 2007.

Francisco Casacuberta, Marcello Federico, Hermann Ney, and Enrique Vidal.
Recent Efforts in Spoken Language Processing.
IEEE Signal Processing Magazine, 25(3):80–88, May 2008.

Mauro Cettolo, Christian Girardi, and Marcello Federico.
WIT^3: Web Inventory of Transcribed and Translated Talks.
In *Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico.
Report on the 10th IWSLT Evaluation Campaign.
In *Proceedings of the International Workshop on Spoken Language Translation*, December 2013.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico.
Report on the 11th IWSLT Evaluation Campaign.

In *Proceedings of the International Workshop on Spoken Language Trnaslation (IWSLT)*, Lake Tahoe, USA, December 2014.

Pi-Chuan Chang, Michel Galley, and Christopher D Manning.
Optimizing Chinese word segmentation for machine translation performance.
In *Proceedings of the third Workshop on Statistical Machine Translation*, pages 224–232. Association for Computational Linguistics, 2008.

Ciprian Chelba and Alex Acero.
Adaptation of maximum entropy capitalizer: Little data can help a lot.
Computer Speech & Language, 20(4):382–399, 2006.

Stanley F. Chen, Kristie Seymore, and Ronald Rosenfeld.
Topic adaptation for language modeling using unnormalized exponential models.
In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 681–684. IEEE, 1998.

Colin Cherry and George Foster.
Batch tuning strategies for statistical machine translation.
In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 427–436, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
ISBN 978-1-937284-20-6.

David Chiang.
A Hierarchical Phrase-based Model for Statistical Machine Translation.
In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 263–270, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
doi: 10.3115/1219840.1219873.

David Chiang, Yuval Marton, and Philip Resnik.
Online Large-Margin Training of Syntactic and Structural Translation Features.
In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 224–233, Honolulu, Hawaii, 2008. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio.

On the Properties of Neural Machine Translation: Encoder-Decoder Approaches.
In *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014a.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio.
Learning phrase representations using RNN encoder-decoder for statistical machine translation.
In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014b.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio.
A Character-level Decoder without Explicit Segmentation for Neural Machine Translation.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL. Volume 1: Long Papers*, Berlin, Germany, August 2016.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith.
Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability.
In *Proceedings of the Association for Computational Lingustics*, ACL 2011, Portland, Oregon, USA, 2011. Association for Computational Linguistics.

Trevor Cohn and Mirella Lapata.
Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora.
In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer.
Online Passive-Aggressive Algorithms.
Journal of Machine Learning Research, 7:551–585, 2006.

G. E. Dahl, Dong Yu, Li Deng, and A. Acero.
Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition.
Trans. Audio, Speech and Lang. Proc., 20(1):30–42, January 2012.

ISSN 1558-7916.
doi: 10.1109/TASL.2011.2134090.

Steven B. Davis and Paul Mermelstein.
Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.
IEEE Transactions on Acoustics, Speech and Signal Processing, pages 357–366, 1980.

José G. C. de Souza, Marco Turchi, and Matteo Negri.
Towards a combination of online and multitask learning for mt quality estimation: a preliminary study.
In *Proceedings of the Workshop on interactive and adaptive machine translation*, pages 9–19, 2014.

Stephen A. Della Pietra, Vincent J. Della Pietra, Robert Mercer, and Salim Roukos.
Adaptive Language Model Estimation using Minimum Discrimination Estimation.
In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 633–636, San Francisco, CA, 1992.

A. P. Dempster, N. M. Laird, and D. B. Rubin.
Maximum-likelihood from incomplete data via the EM algorithm.
Journal of the Royal Statistical Society, B, 39:1–38, 1977.

George Doddington.
Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.
In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

Nadir Durrani, Helmut Schmid, and Alexander Fraser.
A Joint Sequence Translation Model with Integrated Reordering.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1045–1054, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
ISBN 978-1-932432-87-9.

Chris Dyer, Jonathan Weese, Hendra Setiawan, Adam Lopez, Ferhan Ture, Vladimir Eidelman, Juri Ganitkevitch, Phil Blunsom, and Philip Resnik.

Cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models.
In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 7–12, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith.
A Simple, Fast, and Effective Reparameterization of IBM Model 2.
In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648, 2013.

Christopher Dyer, Smaranda Muresan, and Philip Resnik.
Generalizing Word Lattice Translation.
In *Proceedings of ACL-08: HLT*, pages 1012–1020, Columbus, Ohio, June 2008. Association for Computational Linguistics.

M Amin Farajian, Rajen Chatterjee, Costanza Conforti, Shahab Jalalvand, Vevake Balaraman, Mattia A Di Gangi, Duygu Ataman, Marco Turchi, Matteo Negri, and Marcello Federico.
FBKs Neural Machine Translation Systems for IWSLT 2016.
In *Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, USA, December 2016.

Marcello Federico.
Efficient language model adaptation through MDI estimation.
In *Proceedings of the 6th European Conference on Speech Communication and Technology*, volume 4, pages 1583–1586, Budapest, Hungary, 1999.

Marcello Federico.
Language Model Adaptation through Topic Decomposition and MDI Estimation.
In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 703–706, Orlando, FL, 2002.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo.
IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models.
In *Proceedings of Interspeech*, pages 1618–1621, Brisbane, Australia, 2008.

Marcello Federico, Mauro Cettolo, Luisa Bentivogli, Michael Paul, and Sebastian Stüker.

Overview of the IWSLT 2012 evaluation campaign.
In *Proc. of the International Workshop on Spoken Language Translation*, December 2012.

Marcello Federico, Matteo Negri, Luisa Bentivogli, and Marco Turchi.
Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effects Models.
In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1643–1653, 2014.

Mariano Felice and Lucia Specia.
Linguistic Features for Quality Estimation.
In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 93–100, Montreal, Canada, June 2012. Association for Computational Linguistics.

Christiane Fellbaum, editor.
WordNet: an electronical Lexical Database.
MIT Press, Cambridge, MA, 1998.

Minwei Feng, Arne Mauser, and Hermann Ney.
A Source-side Decoding Sequence Model for Statistical Machine Translation.
In *Conference of the Association for Machine Translation in the Americas (AMTA)*, Denver, Colorado, USA, November 2010.

Edward Finegan.
Language: Its Structure and Use.
Cengage Learning, 2014.
ISBN 9781305162815.

Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T Yarman Vural, and Yoshua Bengio.
Multi-way, multilingual neural machine translation.
Computer Speech & Language, 2016.

George Foster and Roland Kuhn.
Mixture-Model Adaptation for SMT.
In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

George Foster, Pierre Isabelle, and Roland Kuhn.
Translating structured documents.
In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*, November 2010.

Christian Fügen, Alex Waibel, and Muntsin Kolss.
Simultaneous translation of lectures and speeches.
Machine Translation, 21(4):209–252, December 2007.

Michel Galley and Christopher D. Manning.
A simple and effective hierarchical phrase reordering model.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 848–856, Morristown, NJ, USA, October 2008. Association for Computational Linguistics.

John S. Garofolo, Jonathan G. Fiscus, Alvin F. Martin, David S. Pallett, and Mark A. Przybocki.
NIST Rich Transcription 2002 Evaluation: A Preview.
In *LREC*. European Language Resources Association, 2002.

George Michael Georgiou.
Parallel distributed processing in the complex domain.
PhD thesis, Tulane University, New Orleans, LA, USA, 1992.
UMI Order No. GAX92-29796.

Diego Giuliani, Matteo Gerosa, and Fabio Brugnara.
Improved Automatic Speech Recognition Through Speaker Normalization.
Comput. Speech Lang., 20(1):107–123, January 2006.
ISSN 0885-2308.
doi: 10.1016/j.csl.2005.05.002.

Sharon Goldwater, Daniel Jurafsky, and Christopher D. Manning.
Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates.
Speech Communication, 52(3):181–200, 2010.

I. J. Good.
The population frequencies of species and the estimation of population parameters.
Biometrika, 40:237–264, 1953.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Deep Learning.
MIT Press, 2016.

Arthur C. Graesser, Murray Singer, and Tom Trabasso.
Constructing inferences during narrative text comprehension.
Psychol Rev, 101(3):371–395, July 1994.
ISSN 0033-295X.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai.
Coh-Metrix: Analysis of text on cohesion and language.
Behavior Research Methods, Instruments, & Computers, 36(2):193–202, May 2004.

R. Haeb-Umbach and H. Ney.
Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition.
In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1 of *ICASSP'92*, pages 13–16, Washington, DC, USA, 1992. IEEE Computer Society.

Christian Hardmeier.
Discourse in Statistical Machine Translation.
Discours, (11), December 2012.

Christian Hardmeier and Marcello Federico.
Modelling Pronominal Anaphora in Statistical Machine Translation.
In *Proceedings of the Seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289, 2010.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo.
Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation.
In *Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT)*, pages 1–16, 2015.

Hany Hassan, Lee Schwartz, Dilek Hakkani-Tur, and Gokhan Tur.
Segmentation and Disfluency Removal for Conversational Speech Translation.

In *Proceedings of Interspeech*. ISCA - International Speech Communication Association, September 2014.

Xiaodong He, Li Deng, and Alex Acero.
Why word error rate is not a good metric for speech recognizer training for the speech translation task?
In *International Conference on Acoustics, Speech, and Signal Processing*, pages 5632–5635. IEEE, 2011a.
ISBN 978-1-4577-0539-7.

Xiaodong He, Li Deng, and Alex Acero.
Why Word Error Rate is not a Good Metric for Speech Recognizer Training for the Speech Translation Task.
In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. IEEE, May 2011b.

Peter A Heeman and James F Allen.
Speech repairs, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue.
Computational Linguistics, 25(4):527–571, 1999.

Hynek Hermansky.
Perceptual linear predictive (PLP) analysis of speech.
Journal of the Acoustical Society of America, 87(4):1738–1752, 1990.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury.
Deep Neural Networks for Acoustic Modeling in Speech Recognition.
IEEE Transactions on Acoustic, Speech, and Signal Processing, pages 30–42, 2012.

Fabian Hirschmann, Jinseok Nam, and Johannes Fürnkranz.
What Makes Word-level Neural Machine Translation Hard: A Case Study on English-German Translation.
In *Proceedings of the 25th International Conference on Computational Linguistics*, Osaka, Japan, December 2016.

Thomas Hofmann.
Probabilistic Latent Semantic Analysis.

In *Proceedings of the 15th Conference on Uncertainty in AI*, pages 289–296, Stockholm, Sweden, 1999.

Jing Huang and Geoffrey Zweig.
Maximum entropy model for punctuation annotation from speech.
In *INTERSPEECH*, 2002.

Xuedong Huang.
Speaker Normalization for Speech Recognition.
In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, pages 465–468, Washington, DC, USA, 1992. IEEE Computer Society.
ISBN 0-7803-0532-9.

Xuedong Huang, Yasuo Ariki, and Mervyn Jack.
Hidden Markov Models for Speech Recognition.
Columbia University Press, New York, NY, USA, 1990.
ISBN 0748601627.

Shahab Jalalvand, Matteo Negri, Daniele Falavigna, and Marco Turchi.
Driving ROVER with segment-based ASR quality estimation.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1095–1105, 2015.

Shahab Jalalvand, Matteo Negri, Marco Turchi, José GC de Souza, Daniele Falavigna, and Mohammed RH Qwaider.
TranscRater: a tool for automatic speech recognition quality estimation.
pages 43–48. Association for Computational Linguistics, Berlin, Germany, 2016.

Frederick Jelinek and Robert L. Mercer.
Interpolated estimation of Markov source parameters from sparse data.
In *Pattern Recognition in Practice*, pages 381–397, Amsterdam, Holland, 1980.

Hui Jiang.
Confidence measures for speech recognition: A survey.
Speech Communication, 45(4):455–470, 2005.

Jie Jiang, Zeeshan Ahmed, Julie Carson-Berndsen, Peter Cahill, and Andy Way.
Phonetic representation-based speech translation.
In *13th Machine Translation Summit*, Xiamen, China, 2011.

Preethi Jyothi and Eric Fosler-Lussier.
A comparison of audio-free speech recognition error prediction methods.
In *INTERSPEECH*, pages 1211–1214. ISCA, 2009.

Preethi Jyothi and Eric Fosler-Lussier.
Discriminative language modeling using simulated asr errors.
In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, editors, *INTERSPEECH*, pages 1049–1052. ISCA, 2010.

Slava M. Katz.
Estimation of probabilities from sparse data for the language model component of a speech recognizer.
IEEE Transactions on Acoustic, Speech and Signal Processing, ASSP-35(3):400–401, 1987.

Ji-Hwan Kim and Philip C Woodland.
Automatic capitalisation generation for speech input.
Computer Speech & Language, 18(1):67–90, 2004.

Reinhard Kneser and Herman Ney.
Improved clustering technique for class-based statistical language modelling.
In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, pages 973–976, Berlin, Germany, 1993.

Reinhard Kneser and Hermann Ney.
Forming word classes by statistical clustering for statistical language modelling.
In R. Köhler and B. B. Rieger, editors, *Proceedings of the First International Conference on Quantitative Linguistics*, pages 221–226, Trier, Germany, 1991. Kluwer Academic Publisher.

Reinhard Kneser and Hermann Ney.
Improved backing-off for m-gram language modeling.
In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 181–184, Detroit, MI, 1995.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst.
Moses: Open Source Toolkit for Statistical Machine Translation.
In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, 2007.

Philipp Koehn.
Europarl: A Multilingual Corpus for Evaluation of Machine Translation.
Unpublished, `http://www.isi.edu/~koehn/europarl/`, 2002.

Philipp Koehn and Christof Monz.
Shared task: Statistical machine translation between European languages.
In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 119–124, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Philipp Koehn and Josh Schroeder.
Experiments in Domain Adaptation for Statistical Machine Translation.
In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu.
Statistical phrase-based translation.
In *Proceedings of HLT-NAACL 2003*, pages 127–133, Edmonton, Canada, 2003.

G. Kurata, N. Itoh, and M. Nishimura.
Acoustically discriminative training for language models.
In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 4717–4720, April 2009.
doi: 10.1109/ICASSP.2009.4960684.

Gakuto Kurata, Nobuyasu Itoh, and Masafumi Nishimura.
Training of error-corrective model for ASR without using audio data.
In *In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 5576–5579. IEEE, 2011.
ISBN 978-1-4577-0539-7.

Alon Lavie, Donna Gates, Noah Coccaro, and Lori Levin.

Input segmentation of spontaneous speech in JANUS: A speech-to-speech translation system, pages 86–99.
Springer Berlin Heidelberg, Berlin, Heidelberg, 1997.
ISBN 978-3-540-69206-5.
doi: 10.1007/3-540-63175-5_39.

Li Lee and R. C. Rose.
Speaker Normalization Using Efficient Frequency Warping Procedures.
In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing - Volume 01*, ICASSP '96, pages 353–356, Washington, DC, USA, 1996. IEEE Computer Society.
ISBN 0-7803-3192-3.
doi: 10.1109/ICASSP.1996.541105.

C.J. Leggetter and P.C. Woodland.
Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models.
Computer Speech and Language, 9(2):171 – 185, 1995.
ISSN 0885-2308.
doi: http://dx.doi.org/10.1006/csla.1995.0010.

V. I. Levenshtein.
Binary Codes Capable of Correcting Deletions, Insertions and Reversals.
Soviet Physics Doklady, 10:707, February 1966.

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary P. Harper.
Enriching speech recognition with automatic detection of sentence boundaries and disfluencies.
IEEE Trans. Audio, Speech & Language Processing, 14(5):1526–1540, 2006.
doi: 10.1109/TASL.2006.878255.

Chi-kiu Lo and Dekai Wu.
MEANT: an inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles.
In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 220–229, 2011.

Max M. Louwerse, Phillip M. McCarthy, Daniel S. McNamara, and Arthur C. Graesser.
Variation in language and cohesion across written and spoken registers.
In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, pages 843–848, 2004.

Minh-Thang Luong and Christopher D. Manning.
Stanford neural machine translation systems for spoken language domain.
In *International Workshop on Spoken Language Translation*, Da Nang, Vietnam, 2015.

L. Mangu, E. Brill, and A. Stolcke.
Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks.
Computer, Speech and Language, 14(4):373–400, 2000.

M. Marcus, B. Santorini, and M.A. Marcinkiewicz.
Building a Large Annotated Corpus of English: the Penn Treebank.
Computational Linguistics, 19:313–330, 1993.

E. Matusov, S. Kanthak, and H. Ney.
Integrating speech recognition and machine translation: Where do we stand?
In *Proceedings of ICASSP*, pages 1217–1220, Toulouse, France, 2006a.

Evgeny Matusov, Arne Mauser, and Hermann Ney.
Automatic sentence segmentation and punctuation prediction for spoken language translation.
In *Proceedings of the International Workshop of Spoken Language Translation (IWSLT)*, pages 158–165, 2006b.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur.
Recurrent neural network based language model.
In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, pages 1045–1048, Makuhari, Chiba, Japan, September 2010.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum.
Polylingual Topic Models.
In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, August 2009.

Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller.
Syntactic and semantic factors in processing difficulty: An integrated measure.
In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 196–206, Stroudsburg, PA, USA, 2010.

Robert C. Moore and William Lewis.
Intelligent selection of language model training data.
In *ACL (Short Papers)*, pages 220–224, 2010.

Preslav Nakov.
Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Paraphrasing, Tokenization, and Recasing.
In *Workshop on Statistical Machine Translation, Association for Computational Linguistics*, 2008.

Preslav Nakov and Jörg Tiedemann.
Combining word-level and character-level models for machine translation between closely-related languages.
In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 301–305. Association for Computational Linguistics, 2012.

Matteo Negri, Marco Turchi, José G. C. de Souza, and Daniele Falavigna.
Quality estimation for automatic speech recognition.
In *COLING 2014, 25th International Conference on Computational Linguistics*, pages 1813–1823, 2014.

Graham Neubig, Taro Watanabe, Shinsuke Mori, and Tatsuya Kawahara.
Substring-based machine translation.
Machine translation, 27(2):139–166, 2013.

Neil Newbold and Lee Gillam.
The linguistics of readability: The next step for word processing.
In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, pages 65–72, Stroudsburg, PA, USA, 2010.

Hermann Ney.
Speech translation: coupling of recognition and translation.

In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, Arizona, 1999.

Jan Niehues and Alex Waibel.
An MT Error-Driven Discriminative Word Lexicon using Sentence Structure Features.
In *Proceedings of the Eighth Workshop on Statistical Machine Translation, WMT@ACL 2013, August 8-9, 2013, Sofia, Bulgaria*, pages 512–520, 2013.

Franz J. Och and Hermann Ney.
The Alignment Template Approach to Statistical Machine Translation.
Computational Linguistics, 30(4):417–450, 2004.

Franz J. Och, Christoph Tillmann, and Hermann Ney.
Improved alignment models for statistical machine translation.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June 1999.

Franz Josef Och.
Statistical machine translation: from single word models to alignment templates.
PhD thesis, RWTH Aachen University, Germany, 2002.

Franz Josef Och.
Minimum Error Rate Training in Statistical Machine Translation.
In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, 2003.

Franz Josef Och and Hermann Ney.
Statistical multi-source translation.
In *Machine Translation Summit*, pages 253–258, Santiago de Compostela, Spain, September 2001.

Franz Josef Och, Nicola Ueffing, and Hermann Ney.
An Efficient A* Search Algorithm for Statistical Machine Translation.
In *Proceedings of the Workshop on Data-driven Methods in Machine Translation - Volume 14*, DMMT '01, pages 1–8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.
doi: 10.3115/1118037.1118045.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura.
Optimizing Segmentation Strategies for Simultaneous Speech Translation.
In *Association of Computational Linguistics*, pages 551–556, 2014.

David Pallett, Jonathan G. Fiscus, John S. Garofolo, Alvin Martin, and Mark Przybocki.
The 1998 Hub-4 Evaluation Plan for Recognition of Broadcast News, in English.
`http://www.itl.nist.gov/iad/mig/tests/bnr/1998/hub4e_98_spec.html`, 1998.
Accessed: 2014-05-21.

Martha Palmer and Zhibiao Wu.
Verb semantics for english-chinese translation.
Machine translation, 10(1-2):59–92, 1995.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.
BLEU: a Method for Automatic Evaluation of Machine Translation.
Research Report RC22176, IBM Research Division, Thomas J. Watson Research Center, 2001.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.
BLEU: a method for automatic evaluation of machine translation.
In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, 2002.

Alicia Pérez, M Inés Torres, and Francisco Casacuberta.
Finite-state acoustic and translation model composition in statistical speech translation: Empirical assessment.
In *Finite-State Methods and Natural Language Processing (FSMNLP)*, pages 99–107, 2012.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely.
The kaldi speech recognition toolkit.
In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011.

R Core Team.
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria, 2013.

Lawrence R. Rabiner and Biing-Hwang Juang.
An introduction to hidden markov models.
IEEE ASSP Magazine, 1986.

Sharath Rao, Ian R. Lane, and Tanja Schultz.
Optimizing sentence segmentation for spoken language translation.
In *INTERSPEECH*, pages 2845–2848. ISCA, 2007.

Nicholas Ruiz and Marcello Federico.
Topic Adaptation for Lecture Translation through Bilingual Latent Semantic Models.
In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 294–302, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.

Nicholas Ruiz and Marcello Federico.
MDI Adaptation for the Lazy: Avoiding Normalization in LM Adaptation for Lecture Translation.
In *International Workshop on Spoken Language Translation (IWSLT)*, pages 244–251, Hong Kong, 2012.

Nicholas Ruiz and Marcello Federico.
Assessing the Impact of Speech Recognition Errors on Machine Translation Quality.
In *Association for Machine Translation in the Americas (AMTA)*, pages 261–274, Vancouver, Canada, 2014a.

Nicholas Ruiz and Marcello Federico.
Complexity of Spoken Versus Written Language for Machine Translation.
In *Proceedings of the 17th Conference of the European Association for Machine Translation (EAMT)*, pages 173–180, 2014b.

Nicholas Ruiz and Marcello Federico.
Phonetically-Oriented Word Error Alignment for Speech Recognition Error Analysis in Speech Translation.
In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, December 2015. IEEE.

Nicholas Ruiz, Arianna Bisazza, Roldano Cattoni, and Marcello Federico.
FBK's Machine Translation Systems for IWSLT 2012's TED Lectures.
In *IWSLT*, pages 61–68. Citeseer, 2012.

Nicholas Ruiz, Qin Gao, William Lewis, and Marcello Federico.
Adapting Machine Translation Models toward Misrecognized Speech with Text-to-Speech Pronunciation Rules and Acoustic Confusability.
In *Proceedings of Interspeech*, Dresden, Germany, September 2015. ISCA.

Nick Ruiz, Arianna Bisazza, Fabio Brugnara, Daniele Falavigna, Diego Giuliani, Suhel Jaber, Roberto Gretter, and Marcello Federico.
FBK @ IWSLT 2011.
In *International Workshop on Spoken Language Translation (IWSLT)*, pages 86–93, San Francisco, CA, 2011.

Kenji Sagae, Maider Lehr, Emily Tucker Prud'hommeaux, Puyang Xu, Nathan Glenn, Damianos Karakos, Sanjeev Khudanpur, Brian Roark, Murat Saraclar, Izhak Shafran, Daniel M. Bikel, Chris Callison-Burch, Yuan Cao, Keith Hall, Eva Hasler, Philipp Koehn, Adam Lopez, Matt Post, and Darcey Riley.
Hallucinated n-best lists for discriminative language modeling.
In *ICASSP*, pages 5001–5004. IEEE, 2012.
ISBN 978-1-4673-0046-9.

Shirin Saleem, Szu-Chen (Stan) Jou, Stephan Vogel, and Tanja Schultz.
Using word lattice information for a tighter coupling in speech translation systems.
In *International Conference of Spoken Language Processing*, 2004.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen.
Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS.
Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, 1995.

Helmut Schmid.
Probabilistic part-of-speech tagging using decision trees.
In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

Josh Schroeder, Trevor Cohn, and Philipp Koehn.
Word lattices for multi-source translation.
In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 719–727. Association for Computational Linguistics, 2009.

S. R. Searle.

Prediction, mixed models, and variance components.
Technical Report BU-468-M, Biometrics Unit, Cornell University, June 1973.

Frank Seide, Gang Li, and Dong Yu.
Conversational speech transcription using context-dependent deep neural networks.
In *INTERSPEECH*, pages 437–440. ISCA, August 2011.

Rico Sennrich, Barry Haddow, and Alexandra Birch.
Neural machine translation of rare words with subword units.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

Abhinav Sethy, Panayiotis Georgiou, and Shrikanth Narayanan.
Selecting relevant text subsets from web-data for building topic specific language models.
In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 145–148, New York City, USA, June 2006. Association for Computational Linguistics.

Claude E. Shannon.
A mathematical theory of communication.
Bell System Technical Journal, 27(3):379–423, 1948.

Dana Shapira and James A. Storer.
Edit distance with move operations.
In *Combinatorial Pattern Matching, 13th Annual Symposium, CPM 2002, Fukuoka, Japan, July 3-5, 2002, Proceedings*, pages 85–98, 2002.
doi: 10.1007/3-540-45452-7_9.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür.
Prosody-based automatic segmentation of speech into sentences and topics.
Speech communication, 32(1):127–154, 2000.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul.
A study of translation edit rate with targeted human annotation.
In *5th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, Massachusetts, August 2006.

Radu Soricut and Abdessamad Echihabi.

Trustrank: Inducing trust in automatic translations via ranking.
In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Lucia Specia.
Exploiting objective annotations for minimising translation post-editing effort.
In Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste, editors, *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 73–80, 2011.

Lucia Specia, Marco Turchi, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders.
Improving the confidence of machine translation quality estimates.
In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*. International Association for Machine Translation, 2009.

Lucia Specia, Najeh Hajlaoui, Catalina Hallett, and Wilker Aziz.
Predicting Machine Translation Adequacy.
In *Machine Translation Summit XIII*, pages 73–80, 2011.

Richard Sproat, Alan Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards.
Normalization of non-standard words: Ws '99 final report.
Technical report, Hopkins University, 1999.

Richard Sproat, Alan W Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards.
Normalization of non-standard words.
Computer Speech & Language, 15(3):287–333, 2001.

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan.
Segmentation Strategies for Streaming Speech Translation.
In *Proceedings of NAACL-HLT*, pages 230–238, 2013.

A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tur, and Y. Lu.
Automatic detection of sentence boundaries and disfluencies based on recognized words.

In *Proceedings of ICSLP*, volume 5, pages 2247–2250, Sydney, Australia, 1998.

Andreas Stolcke and Elizabeth Shriberg.
Automatic linguistic segmentation of conversational speech.
In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 2, pages 1005–1008. IEEE, 1996.

S. Stuker, T. Herrmann, M. Kolss, J. Niehues, and M. Wolfel.
Research Opportunities In Automatic Speech-To-Speech Translation.
Potentials, IEEE, 31(3):26–33, 2012.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le.
Sequence to sequence learning with neural networks.
In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

Yik-Cheung Tam, Ian Lane, and Tanja Schultz.
Bilingual LSA-based adaptation for statistical machine translation.
Machine Translation, 21:187–207, December 2007.
ISSN 0922-6567.
doi: 10.1007/s10590-008-9045-2.

Qun Feng Tan, Kartik Audhkhasi, Panayiotis G. Georgiou, Emil Ettelaie, and Shrikanth S. Narayanan.
Automatic speech recognition system channel modeling.
In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, editors, *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 2442–2445. ISCA, 2010.

The European Commission Directorate-General for Translation (EC DGT).
2015 Annual Activity Report.
The European Commission, April 2016.
`https://ec.europa.eu/info/publications/annual-activity-report-2015-translation_en`.

Christoph Tillmann.
A Unigram Orientation Model for Statistical Machine Translation.

In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, pages 101–104, Boston, Massachusetts, USA, 2004.

Christoph Tillmann and Hermann Ney.
Word reordering and a dynamic programming beam search algorithm for statistical machine translation.
Computational Linguistics, 29(1):97–133, 2003.

Yulia Tsvetkov, Florian Metze, and Chris Dyer.
Augmenting translation models with simulated acoustic confusions for improved spoken language translation.
In *Proceedings of the European Association of Computational Linguistics (EACL)*, pages 616–625, 2014.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li.
Modeling coverage for neural machine translation.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.

Marco Turchi, Matteo Negri, and Marcello Federico.
Coping with the subjectivity of human judgements in MT quality estimation.
In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Nicola Ueffing, K. Macherey, and Hermann Ney.
Confidence measures for statistical machine translation.
In *Proceedings of the MT Summit IX*, 2003.

David Vilar, Jia Xu, Luis Fernando dHaro, and Hermann Ney.
Error analysis of statistical machine translation output.
In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 697–702, 2006.

S. Vogel, H. Ney, and C. Tillmann.
HMM-based word alignment in statistical translation.
In *Proceedings of COLING*, pages 836–841, Copenhagen, Denmark, 1996.

Chao Wang, Michael Collins, and Philipp Koehn.

Chinese syntactic reordering for statistical machine translation.
In *EMNLP-CoNLL*, pages 737–745, 2007.

Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki.
Online Large-Margin Training for Statistical Machine Translation.
In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 764–773, 2007.

Warren Weaver.
Translation.
In William N. Locke and A. Donald Boothe, editors, *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA, 1949/1955.
Reprinted from a memorandum written by Weaver in 1949.

Ian H. Witten and Timothy C. Bell.
The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression.
IEEE Trans. Inform. Theory, IT-37(4):1085–1094, 1991.

Chuck Wooters and Andreas Stolcke.
Multiple-pronunciation lexical modeling in a speaker independent speech understanding system.
In *The 3rd International Conference on Spoken Language Processing, ICSLP 1994, Yokohama, Japan, September 18-22, 1994*, 1994.

Dekai Wu.
Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora.
Comput. Linguist., 23(3):377–403, September 1997.
ISSN 0891-2017.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang.
Document-level consistency verification in machine translation.
In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, pages 131–138. International Association for Machine Translation, 2011.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig.

Achieving human parity in conversational speech recognition.
CoRR, abs/1610.05256, 2016.

Kenji Yamada and Kevin Knight.
A Syntax-based Statistical Translation Model.
In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 523–530, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.
doi: 10.3115/1073012.1073079.

Liu Yi and Pascale Fung.
Partial change accent models for accented mandarin speech recognition.
In *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, pages 111–116, Nov 2003.
doi: 10.1109/ASRU.2003.1318413.

Richard Zens, Franz Josef Och, and Hermann Ney.
Phrase-Based Statistical Machine Translation.
In *KI 2002: Advances in Artificial Intelligence, 25th Annual German Conference on AI, KI 2002, Aachen, Germany, September 16-20, 2002, Proceedings*, pages 18–32, 2002.
doi: 10.1007/3-540-45751-8_2.

Ruiqiang Zhang, Gen-ichiro Kikui, Hirofumi Yamamoto, and Wai-Kit Lo.
A decoding algorithm for word lattice translation in speech translation.
In *International Workshop on Spoken Language Translation, IWSLT*, pages 23–29, Pittsburgh, PA, USA, October 2005.

Yaodong Zhang, Li Deng, Xiaodong He, and Alex Acero.
A novel decision function and the associated decision-feedback learning for speech translation.
In *ICASSP*, pages 5608–5611, 2011.

Bing Zhao and Eric P. Xing.
HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation.
In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1689–1696. MIT Press, Cambridge, MA, 2008.

Bing Zhao, Matthias Eck, and Stephan Vogel.

Language Model Adaptation for Statistical Machine Translation via Structured Query Models.
In *Proceedings of Coling 2004*, pages 411–417, Geneva, Switzerland, Aug 23–Aug 27 2004. COLING.

Bowen Zhou, Laurent Besacier, and Yuqing Gao.
On efficient coupling of ASR and SMT for speech translation.
In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 101–104, Honolulu, HA, 2007.

Andreas Zollmann and Ashish Venugopal.
Syntax Augmented Machine Translation via Chart Parsing.
In *Proceedings of the Workshop on Statistical Machine Translation*, StatMT '06, pages 138–141, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Will Y. Zou, Richard Socher, Daniel M. Cer, and Christopher D. Manning.
Bilingual Word Embeddings for Phrase-Based Machine Translation.
In *Empirical Methods of Natural Language Processing*, pages 1393–1398. ACL, 2013.