Paraphrasing and Translation

Chris Callison-Burch

Doctor of Philosophy

Institute for Communicating and Collaborative Systems

School of Informatics

University of Edinburgh

2007

Abstract

Paraphrasing and translation have previously been treated as unconnected natural language processing tasks. Whereas translation represents the preservation of meaning when an idea is rendered in the words in a different language, paraphrasing represents the preservation of meaning when an idea is expressed using different words in the same language. We show that the two are intimately related. The major contributions of this thesis are as follows:

- We define a novel technique for automatically generating paraphrases using bilingual parallel corpora, which are more commonly used as training data for statistical models of translation.

- We show that paraphrases can be used to improve the quality of statistical machine translation by addressing the problem of coverage and introducing a degree of generalization into the models.

- We explore the topic of automatic evaluation of translation quality, and show that the current standard evaluation methodology cannot be guaranteed to correlate with human judgments of translation quality.

Whereas previous data-driven approaches to paraphrasing were dependent upon either data sources which were uncommon such as multiple translation of the same source text, or language specific resources such as parsers, our approach is able to harness more widely parallel corpora and can be applied to any language which has a parallel corpus. The technique was evaluated by replacing phrases with their paraphrases, and asking judges whether the meaning of the original phrase was retained and whether the resulting sentence remained grammatical. Paraphrases extracted from a parallel corpus with manual alignments are judged to be accurate (both meaningful and grammatical) 75% of the time, retaining the meaning of the original phrase 85% of the time. Using automatic alignments, meaning can be retained at a rate of 70%.

Being a language independent and probabilistic approach allows our method to be easily integrated into statistical machine translation. A paraphrase model derived from parallel corpora other than the one used to train the translation model can be used to increase the coverage of statistical machine translation by adding translations of previously unseen words and phrases. If the translation of a word was not learned, but a translation of a synonymous word has been learned, then the word is paraphrased

and its paraphrase is translated. Phrases can be treated similarly. Results show that augmenting a state-of-the-art SMT system with paraphrases in this way leads to significantly improved coverage and translation quality. For a training corpus with 10,000 sentence pairs, we increase the coverage of unique test set unigrams from 48% to 90%, with more than half of the newly covered items accurately translated, as opposed to none in current approaches.

Acknowledgements

I had the great fortune to be doing research in machine translation at a time when the subject was just beginning to flourish at Edinburgh. When I began my graduate work, I was the only person working on the topic at the university. As I leave, there are five other PhD students, three full-time researchers, and two faculty members all striving towards the same goal. The School of Informatics is undoubtedly the best place in the world to be studying computational linguistics, and the intellectual community here is simply amazing. I am grateful to every member of that community but would like to single out the following people to whom I am especially indebted:

- My PhD supervisor, Miles Osborne, whose data-intensive linguistics class opened my eyes to statistical NLP and played a crucial role in my deciding to stay at Edinburgh for the PhD. His endlessly creative ideas and boundless enthusiasm made our weekly meetings in his office (and at the pub) a true joy. As much as it is due to any one person, my success at Edinburgh is due to Miles.

- My best friend and business partner, Colin Bannard, without whom I would not have founded Linear B. One of my fondest memories of Edinburgh is sitting in our living room trying to name the company. *Linear B* was perfect since it allowed us to convey to investors that we use clever methods to decipher foreign languages, while at the same time tacitly acknowledging that it might take us decades to do so.

- Josh Schroeder, who is the primary reason that it did not take decades to achieve all that we did at Linear B. Josh lived in the boxroom in my flat for a year, intrepidly writing code so elegant and easy to maintain that I still use it to this day. Linear B put me in the enviable position of having two full-time programmers working for me during my PhD. The quality and amount of research that I was able to produce as a result far outstripped what I would have been able do alone.

- Philipp Koehn joined the faculty at Edinburgh after I hounded him to apply and then lobbied the head of the school to allow student input into the hiring decision (a diplomatic means of me getting my way). When Philipp arrived at the university he became the center of gravity for the machine translation group and allowed us to form a coherent whole. He has been a wonderful collaborator and I value the time that I had to work with him.

- I owe much to the other outstanding members of the machine translation group: Abhi Arun, Amittai Axelrod, Lexi Birch, Phil Blunsom, Trevor Cohn, Loïc Dugast, Hieu Hoang, Josh Schroeder, and David Talbot, along with many visitors and master's students. I must also thank my academic brothers Markus Becker and Andrew Smith, who were always willing to form an impromptu support group over coffee on the odd occasion that we needed to complain about our supervisor.
- Thank you to Mark Steedman for providing so much sage advice during my PhD. Thank you to Aravind Joshi, Mitch Marcus, and Fernando Pereira for lending me an office at Penn to write up my thesis when I needed to escape Edinburgh's distractions (although Philadelphia provided wonderful things to replace them). Thank you to Bonnie Webber and Kevin Knight for being such an exceptional thesis committee. Somehow my thesis defense was an enjoyable experience – it felt like an engaging conversation rather than an ordeal.

Outside of Edinburgh, I had the opportunity to collaborate with a number of superb researchers in the EuroMatrix project and at a summer workshop at Johns Hopkins. It was a wonderful learning experience writing the EuroMatrix proposal with Andreas Eisele, Philipp Koehn and Hans Uszkoreit, and a pleasure working with Cameron Shaw Fordyce. I'd like to take this opportunity thank the CLSP workshop participants Nicola Bertoldi, Ondrej Bojar, Alexandra Constantin, Brooke Cowan, Chris Dyer, Marcello Federico, Evan Herbst, Hieu Hoang, Christine Moran, Wade Shen, and Richard Zens, and to apologize to them for suggesting *Moses* as the name for our open source software, which was meant to lead people away from the *Pharaoh* decoder. I thought it was clever at the time.

I am exceptionally grateful (and still amazed) that at the end of the summer workshop David Yarowksy invited me to apply for a faculty position at Johns Hopkins. In no small part due to David's championing my application, I am now an assistant research professor at JHU! I will work my damnedest to live up to his high expectations.

Not least, thank you to all my friends who made the past six years in Edinburgh so wonderful: Abhi, Akira, Alexander, Amittai, Amy, Andrew, Anna, Annabel, Bea, Beata, Ben, Brent, Casey, Colin, Daniel, Danielle, Dave, Eilidh, Hanna, Hieu, Jackie, Josh, Jochen, John, Jon, Kate, Mark, Matt, Markus, Marco, Natasha, Nikki, Pascal, Pedro, Rojas, Sam, Sebastian, Soyeon, Steph, Tom, Trevor, Ulrike, Viktor, Vera, Zoe, and many, many others.

Finally, thank you to my family. I am who I am because of you.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Chris Callison-Burch*)

I dedicate this work to my grandparents for showing me the world, and for making so many things possible that would not have been possible otherwise.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Paraphrasing and translation have previously been treated as unconnected natural language processing tasks. Whereas translation represents the preservation of meaning when an idea is rendered in the words of a different language, paraphrasing represents the preservation of meaning when an idea is expressed using different words in the same language. We show that the two are intimately related. We intertwine paraphrasing and translation in the following ways:

- We show that paraphrases can be generated using data that is more commonly used to train statistical models of translation.
- We show that statistical machine translation can be significantly improved by integrating paraphrases to alleviate sparse data problems.
- We show that paraphrases are crucial to evaluating translation quality, and that current automatic evaluation metrics are insufficient because they fail to account for this.

In this thesis we define a novel mechanism for generating paraphrases that exploits *bilingual parallel corpora*, which have not hitherto been used for paraphrasing. This is the first time that this type of data has been used for the task of paraphrasing. Previous data-driven approaches to paraphrasing have used *multiple translations*, *comparable corpora*, or parsed *monolingual corpora* as their source of data. Examples of corpora containing multiple translations are collections of classic French novels translated into English by several different translators, and multiple reference translations prepared for evaluating machine translation. Comparable corpora can consist of newspaper articles published about the same event written by different papers, for instance, or of

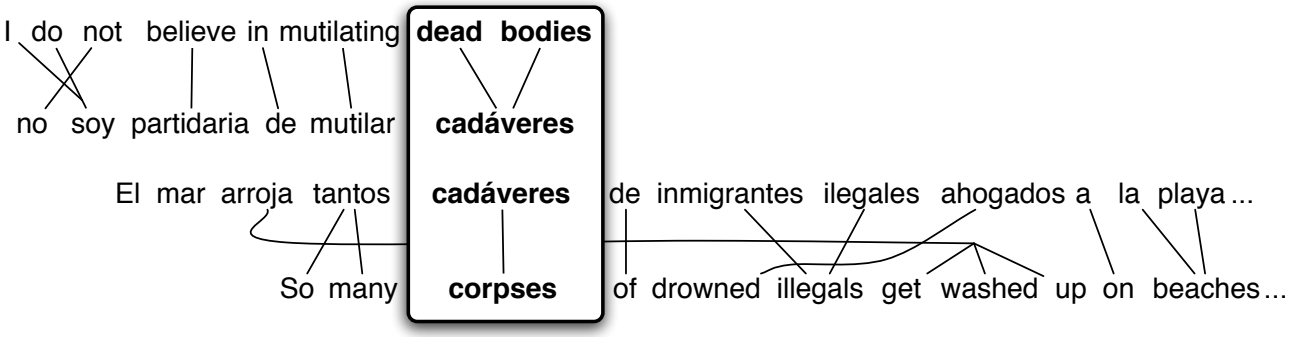


Figure 1.1: The Spanish word *cadáveres* can be used to discover that the English phrase *dead bodies* can be paraphrased as *corpses*.

different encyclopedias' articles about the same topic. Since they are written by different authors items in these corpora represent a natural source for paraphrases – they express the same ideas but are written using different words. Plain monolingual corpora are not a ready source of paraphrases in the same way that multiple translations and comparable corpora are. Instead, they serve to show the distributional similarity of words. One approach for extracting paraphrases from monolingual corpora involves parsing the corpus, and drawing relationships between words which share the same syntactic contexts (for instance, words which can be modified by the same adjectives, and which appear as the objects of the same verbs).

We argue that previous paraphrasing techniques are limited since their training data are either relatively rare, or must have linguistic markup that requires language-specific tools, such as syntactic parsers. Since parallel corpora are comparatively common, we can generate a large number of paraphrases for a wider variety of phrases than past methods. Moreover, our paraphrasing technique can be applied to more languages since it does not require language-specific tools, because it uses language-independent techniques from statistical machine translation.

Word and phrase alignment techniques from statistical machine translation serve as the basis of our data-driven paraphrasing technique. Figure 1.1 illustrates how they are used to extract an English paraphrase from a bilingual parallel corpus by pivoting through foreign language phrases. An English phrase that we want to paraphrase, such as *dead bodies*, is automatically aligned with its Spanish counterpart *cadáveres*. Our technique then searches for occurrences of *cadáveres* in other sentence pairs in the parallel corpus, and looks at what English phrases they are aligned to, such as *corpses*. The other English phrases that are aligned to the foreign phrase are deemed to be paraphrases of the original English phrase. A parallel corpus can be a rich source

of paraphrases. When a parallel corpus is large there are frequently multiple occurrences of the original phrase and of its foreign counterparts. In these circumstances our paraphrasing technique often extracts multiple paraphrases for a single phrase. Other paraphrases for *dead bodies* that were generated by our paraphrasing technique include: *bodies*, *bodies of those killed*, *carcasses*, *the dead*, *deaths*, *lifeless bodies*, and *remains*.

Because there can be multiple paraphrases of a phrase, we define a probabilistic formulation of paraphrasing. Assigning a paraphrase probability $p(e_2|e_1)$ to each extracted paraphrase e_2 allows us to rank the candidates, and choose the best paraphrase for a given phrase e_1. Our probabilistic formulation naturally falls out from the fact that we are using parallel corpora and statistical machine translation techniques. We initially define the paraphrase probability in terms of phrase translation probabilities, which are used by phrase-based statistical translation systems. We calculate the paraphrase probability, $p(corpses|dead\ bodies)$, in terms of the probability of the foreign phrase given the original phrase, $p(cadáveres|dead\ bodies)$, and the probability of the paraphrase given the foreign phrase, $p(corpses|cadáveres)$. We discuss how various factors which can affect translation quality –such as the size of the parallel corpus, and systematic errors in alignment– can also affect paraphrase quality. We address these by refining our paraphrase definition to include multiple parallel corpora (with different foreign languages), and show experimentally that the addition of these corpora markedly improve paraphrase quality.

Using a rigorous evaluation methodology we empirically show that several refinements to our baseline definition of the paraphrase probability lead to improved paraphrase quality. Quality is evaluated by substituting phrases with their paraphrases and judging whether the resulting sentence preserves the meaning of the original sentence, and whether it remains grammatical. We go beyond previous research by substituting our paraphrases into many different sentences, rather than just a single context. Several refinements improve our paraphrasing method. The most successful are: reducing the effect of systematic misalignments in one language by using parallel corpora over multiple languages, performing word sense disambiguation on the original phrase and only using instances of the same sense to generate paraphrases, and improving the fluency of paraphrases by using the surrounding words to calculate a language model probability. We further show that if we remove the dependency on automatic alignment methods that our paraphrasing method can achieve very high accuracy. In ideal circumstances our technique produces paraphrases that are both grammatical and have the correct

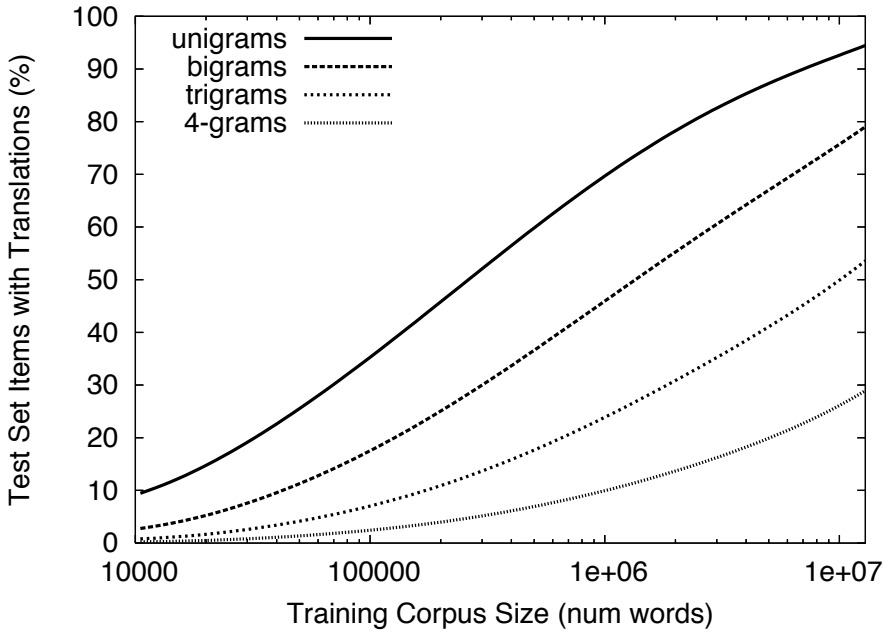


Figure 1.2: Translation coverage of unique phrases from a test set

meaning 75% of the time. When meaning is the sole criterion, the paraphrases reach 85% accuracy.

In addition to evaluating the quality of paraphrases in and of themselves, we also show their usefulness when applied to a task. We show that paraphrases can be used to improve the quality of statistical machine translation. We focus on a particular problem with current statistical translation systems: that of coverage. Because the translations of words and phrases are learned from corpora, statistical machine translation is prone to suffer from problems associated with sparse data. Most current statistical machine translation systems are unable to translate source words when they are not observed in the training corpus. Usually their behavior is either to drop the word entirely, or to leave it untranslated in the output text. For example, when a Spanish-English system is trained on 10,000 sentence pairs (roughly 200,000 words) is used to translate the sentence:

Votaré en favor de la aprobación del proyecto de reglamento.

It produces output which is partially untranslated, because the system's default behaior is to push through unknown words like *votaré*:

Votaré in favor of the approval of the draft legislation.

The system's behavior is slightly different for an unseen phrase, since each word in it might have been observed in the training data. However, a system is much less likely

votaré	I will be voting
voy a votar	I will vote / I am going to vote
voto	I am voting / he voted
votar	to vote

mejores prácticas	best practices
buenas prácticas	best practices / good practices
mejores procedimientos	better procedures
procedimientos idóneos	suitable procedures

Table 1.1: Examples of automatically generated paraphrases of the Spanish word *votaré* and the Spanish phrase *mejores prácticas* along with their English translations

to translate a phrase correctly if it is unseen. For example, for the phrase *mejores prácticas* in the sentence:

Pide que se establezcan las mejores prácticas en toda la UE.

Might be translated as:

It calls for establishing **practices in the best** throughout the EU.

Although there are no words left untranslated, the phrase itself is translated incorrectly. The inability of current systems to translate unseen words, and their tendency to fail to correctly translate unseen phrases is especially worrisome in light of Figure 1.2. It shows the percent of unique words and phrases from a 2,000 sentence test set that the statistical translation system has learned translations of for variously sized training corpora. Even with training corpora containing 1,000,000 words a system will have learned translation for only 75% of the unique unigrams, fewer than 50% of the unique bigrams, less than 25% of unique trigrams and less than 10% of the unique 4-grams.

We address the problem of unknown words and phrases by generating paraphrases for unseen items, and then translating the paraphrases. Figure 1.1 shows the paraphrases that our method generates for *votaré* and *mejores prácticas*, which were unseen in the 10,000 sentence Spanish-English parallel corpus. By substituting in paraphrases which have known translations, the system produces improved translations:

I will vote in favor of the approval of the draft legislation.
It calls for establishing **best practices** throughout the EU.

While it initially seems like a contradiction that our paraphrasing method –which itself relies upon parallel corpora– could be used to improve coverage of statistical machine translation, it is not. The Spanish paraphrases could be generated using a corpus other than the Spanish-English corpus used to train the translation model. For instance the Spanish paraphrases could be drawn from a Spanish-French or a Spanish-German corpus.

While any paraphrasing method could potentially be used to address the problem of coverage, our method has a number of features which makes it ideally suited to statistical machine translation:

- It is language-independent, and can be used to generate paraphrases for any language which has a parallel corpus. This is important because we are interested in applying machine translation to a wide variety of languages.
- It has a probabilistic formulation which can be straightforwardly integrated into statistical models of translation. Since our paraphrases can vary in quality it is natural to employ the search mechanisms present in statistical translation systems.
- It can generate paraphrases for multi-word phrases in addition to single words, which some paraphrasing approaches are biased towards. This makes it good fit for current phrase-based approaches to translation.

We design a set of experiments that demonstrate the importance of each of these features.

Before presenting our experimental results, we first examine the problem of evaluating translation quality. We discuss the failings of the dominant methodology of using the Bleu metric for automatically evaluating translation quality. We examine the importance of *allowable variation in translation* for the automatic evaluation of translation quality. We discuss how Bleu's overly permissive model of variant phrase order, and its overly restrictive model of alternative wordings mean that it can assign identical scores to translations which human judges would easily be able to distinguish. We highlight the importance of correctly rewarding valid alternative wordings when applying paraphrasing to translation – since paraphrases are by definition alternative wordings. Our results show that despite measurable improvements in Bleu score that the metric significantly underestimates our improvements to translation quality. We conduct a targeted manual evaluation in order to better observe the actual improvements to translation quality in each of our experiments. Bleu's failure to correspond to

human judgments have wide-ranging implications for the field that extend far beyond the research presented in this thesis.

Our experiments examine translation from Spanish to English, and from French to English – thus necessitating the ability to generate paraphrases in multiple languages. Paraphrases are used to increase coverage by adding translations of previously unseen source words and phrases. Our experiments show the importance of integrating a paraphrase probability into the statistical model, and of being able to generate paraphrases for multi-word units in addition to individual words. Results show that augmenting a state-of-the-art phrase-based translation system with paraphrases leads to significantly improved coverage and translation quality. For a training corpus with 10,000 sentence pairs we increase the coverage of unique test set unigrams from 48% to 90%, with more than half of the newly covered items accurately translated, as opposed to none in current approaches. Furthermore the coverage of unique bigrams jumps from 25% to 67%, and the coverage of unique trigrams jumps from 10% to nearly 40%. The coverage of unique 4-grams jumps from 3% to 16%, which is not achieved in the baseline system until 16 times as much training data has been used.

1.1 Contributions of this thesis

The major contributions of this thesis are as follows:

- We present a novel technique for automatically generating paraphrases using bilingual parallel corpora and give a probabilistic definition for paraphrasing.
- We show that paraphrases can be used to improve the quality of statistical machine translation by addressing the problem of coverage and introducing a degree of generalization into the models.
- We explore the topic of automatic evaluation of translation quality, and show that the current standard evaluation methodology cannot be guaranteed to correlate with human judgments of translation quality.

1.2 Structure of this document

The remainder of this document is structured as follows:

- Chapter 2 surveys other data-driven approaches to paraphrases, and reviews the aspects of statistical machine translation which are relevant to our paraphrasing technique and to our experimental design for improved translation using paraphrases.

- Chapter 3 details our paraphrasing technique, illustrating how parallel corpora can be used to extract paraphrases, and giving our probabilistic formulation of paraphrases. The chapter examines a number of factors which affect paraphrase quality including alignment quality, training corpus size, word sense ambiguities, and the context of sentences which paraphrases are substituted into. Several refinements to the paraphrase probability are proposed to address these issues.

- Chapter 4 describes our experimental design for evaluating paraphrase quality. The chapter also reports the baseline accuracy of our paraphrasing technique and the improvements due to each of the refinements to the paraphrase probability. It additionally includes an estimate of what paraphrase quality would be achievable if the word alignments used to extract paraphrases were perfect, instead of inaccurate automatic alignments.

- Chapter 5 discusses one way that paraphrases can be applied to machine translation. It discusses the problem of coverage in statistical machine translation, detailing the extent of the problem and the behavior of current systems. The chapter discusses how paraphrases can be used to expand the translation options available to a translation model and how the paraphrase probability can be integrated into decoding.

- Chapter 6 discusses the dominant evaluation methodology for machine translation research, which is to use the Bleu automatic evaluation metric. We show that Bleu cannot be guaranteed to correlate with human judgments of translation quality because of its weak model of allowable variation in translation. We discuss why this is especially pertinent when evaluating our application of paraphrases to statistical machine translation, and detail an alternative manual evaluation methodology.

- Chapter 7 lays out our experimental setup for evaluating statistical translation when paraphrases are included. It decribes the data used to train the paraphrase and translation models, the baseline translation system, the feature functions used in the baseline and paraphrase systems, and the software used to set their

parameters. It reports results in terms of improved Bleu score, increased coverage, and the accuracy of translation as determined by human evaluation.

- Chapter 8 concludes the thesis by highlighting the major findings, and suggesting future research directions.

1.3 Related publications

This thesis is based on three publications:

- Chapters 3 and 4 expand "Paraphrasing with Bilingual Parallel Corpora." which was published in 2005. The paper appeared the proceedings of the 43rd annual meeting of the Association for Computational Linguistics and was joint work with Colin Bannard.

- Chapters 5 and 7 elaborate on "Improved Statistical Machine Translation Using Paraphrases" which was published in 2006 in the proceedings the North American chapter of the Association for Computational Linguistics.

- Chapter 6 extends "Re-evaluating the Role of Bleu in Machine Translation Research" which was published in 2006 in the proceedings of the European chapter of the Association for Computational Linguistics.

Chapter 2

Literature Review

This chapter reviews previous paraphrasing techniques, and introduces concepts from statistical machine translation which are relevant to our paraphrasing method. Section 2.1 gives a representative (but by no means exhaustive) survey of other data-driven paraphrasing techniques, including methods which use training data in the form of multiple translations, comparable corpora, and parsed monolingual texts. Section 2.2 reviews the concepts from the statistical machine translation literature which form the basis of our paraphrasing technique. These include word alignment, phrase extraction and translation model probabilities. This section also serves as background material to Chapters 5–7 which describe how SMT can be improved with paraphrases.

2.1 Previous paraphrasing techniques

Paraphrases are alternative ways of expressing the same content. Paraphrasing can occur at different levels of granularity. *Sentential* or *clausal* paraphrases rephrase entire sentences, whereas *lexical* or *phrasal* paraphrases reword shorter items. Paraphrases have application to a wide range of natural language processing tasks, including question answering, summarization and generation. Over the past thirty years there have been many different approaches to automatically generating paraphrases. McKeown (1979) developed a paraphrasing module for a natural language interface to a database. Her module parsed questions, and asked users to select among automatically rephrased questions when their questions contained ambiguities that would result in different database queries. Later research examined the use of formal semantic representation and intentional logic to represent paraphrases (Meteer and Shaked, 1988; Iordanskaja et al., 1991). Still others focused on the use of grammar formalisms such as syn-

chronous tree adjoining grammars to produce paraphrase transformations (Dras, 1997, 1999a,b). In recent years there has been a trend towards applying statistical methods to the problems of paraphrasing (a trend which has been embraced broadly in the field of computational linguistics as a whole). As such, most current research is data-driven and does not use a formal definition of paraphrases. By and large most current data-driven research has focused on the extraction of lexical or phrasal paraphrases, although a number of efforts have examined sentential paraphrases or large paraphrasing templates (Ravichandran and Hovy, 2002; Barzilay and Lee, 2003; Pang et al., 2003; Dolan and Brockett, 2005). This thesis proposes a method for extracting lexical and phrasal paraphrases from bilingual parallel corpora. As such we review other data-driven approaches which target a similar level of granularity – we neglect sentential paraphrasing and methods which are not data-driven.

2.1.1 Data-driven paraphrasing techniques

One way of distinguishing between different data-driven approaches to paraphrasing is based on the kind of data that they use. Hitherto three types of data have been used for paraphrasing: *multiple translations*, *comparable corpora*, and *monolingual corpora*. Sources for multiple translations include different translations of classic French novels into English, and test sets which have been created for the Bleu machine translation evaluation metric (Papineni et al., 2002), which requires multiple translations. Comparable corpora are comprised of documents which describe the same basic set of facts, such as newspaper articles about the same day's events but written by different authors, or encyclopedia articles on the same topic taken from different encyclopedias. Standard monolingual corpora have also been applied to the task of paraphrasing. In order to be used for the task this type of data generally has to be marked up with some additional information such as dependency parses.

Each of these three types of data has advantages and disadvantages when used as a source of data for paraphrasing. The pros and cons of data-driven paraphrasing techniques based on multiple translations, comparable corpora, and monolingual corpora are discussed in Sections 2.1.2, 2.1.3, and 2.1.4, respectively.

2.1.2 Paraphrasing with multiple translations

Barzilay (2003) suggested that multiple translations of the same foreign source text were a source of "naturally occurring paraphrases" because they are samples of text

Emma burst into tears **and he tried to** comfort **her,** saying things to make her smile **.**
Emma cried, **and he tried to** console **her,** adorning his words with puns **.**

Figure 2.1: Barzilay and McKeown (2001) extracted paraphrases from multiple translations using identical surrounding substrings

which convey the same meaning but are produced by different writers. Indeed multiple translations do seem to be a natural source for paraphrases. Since different translators have different ways of expressing the ideas in a source text, the result is the essence of a paraphrase: different ways of wording the same information.

Multiple translations were first used for the generation of paraphrases by Barzilay and McKeown (2001), who assembled a corpus containing two to three English translations each of five classic novels including *Madame Bovary* and *20,000 Leagues Under the Sea*. They began by aligning the sentences across the multiple translations by applying sentence alignment techniques (Gale and Church, 1993). These were tailored to use token identities within the English sentences as additional guidance. Figure 2.1 shows a sentence pair created from different translations of *Madame Bovary*. Barzilay and McKeown extracted paraphrases from these aligned sentences by equating phrases which are surrounded by identical words. For example, *burst into tears* can be paraphrased as *cried*, *comfort* can be paraphrased as *console*, and *saying things to make her smile* can be paraphrased as *adorning his words with puns* because they appear in identical contexts. Barzilay and McKeown's technique is a straightforward method for extracting paraphrases from multiple translations.

Pang et al. (2003) also used multiple translations to generate paraphrases. Rather than equating paraphrases in paired sentences by looking for identical surrounding contexts, Pang et al. used a syntax-based alignment algorithm. Figure 2.2 illustrates this algorithm. Parse trees were merged by grouping constituents of the same type (for example the two noun phrases and two verb phrases in the figure). The merged parse trees were mapped onto word lattices, by creating alternative paths for every group of merged nodes. Different paths within the word lattices were treated as paraphrases of each other. For example, in the word lattice in Figure 2.2 *people were killed*, *persons died*, *persons were killed*, and *people died* are all possible paraphrases of each other.

While multiple translations contain paraphrases by their nature, there is an inherent disadvantage to any paraphrasing technique which relies upon them as a source of data:

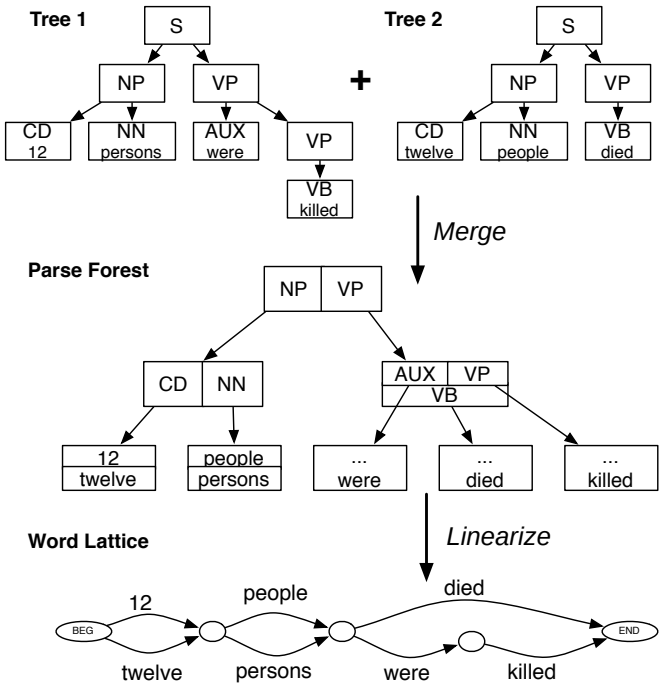


Figure 2.2: Pang et al. (2003) extracted paraphrases from multiple translations using a syntax-based alignment algorithm

multiple translations are a rare resource. The corpus that Barzilay and McKeown assembled from multiple translations of novels contained 26,201 aligned sentence pairs with 535,268 words on one side and 463,959 on the other. Furthermore, since the corpus was constructed from literary works, the type of language usage which Barzilay and McKeown paraphrased might not be useful for applications which require more formal language, such as information retrieval, question answering, etc. The corpus used by Pang et al. was similarly small. They used a corpus containing eleven English translations of Chinese newswire documents, which were commissioned from different translation agencies by the Linguistics Data Consortium, for use with the Bleu machine translation evaluation metric (Papineni et al., 2002). A total of 109,230 English-English sentence pairs can be created created from all pairwise combinations of the 11 translations of the 993 Chinese sentences in the data set. There are total of 3,266,769 words on either side of these sentence pairs, which initially seems large. However, it is still very small when compared to the amount of data available in bilingual parallel corpora.

Let us put into perspective how much more training data is available for paraphrasing techniques that draw paraphrases from bilingual parallel corpora rather than from

multiple translations. The Europarl bilingual parallel corpora (Koehn, 2005) used in our paraphrasing experiments has a total of 6,902,255 sentence pairs between English and other languages, with a total of 145,688,773 English words. This is 34 times more than the combined totals of the corpora used by Barzilay and McKeown and Pang et al. Moreover, the LDC provides corpora for Arabic-English and Chinese-English machine translation. This provides a further 8,389,295 sentence pairs, with 220,365,680 English words. This increases the relative amount of readily available bilingual data by 86 times the amount of multiple translation data that was used in previous research. The implications of this discrepancy are than even if multiple translations are a natural source of paraphrases, techniques which use it as a data source will be able to generate only a small number of paraphrases for a restricted set of language usage and genres. Since many natural language processing applications require broad coverage, multiple translations are an ineffective source of data for "real-world" applications. The availability of large amounts of parallel corpora also means that the models may be better trained, since other statistical natural language processing tasks demonstrate that more data leads to better parameter estimates.

2.1.3 Paraphrasing with comparable corpora

Whereas multiple translation are extremely rare, *comparable corpora* are much more common by comparison. Comparable corpora consist of texts about the same topic. An example of something that might be included in a comparable corpus is encyclopedia articles on the same subject but published in different encyclopedias. The most common source for comparable corpora are news articles published by different newspapers. These are generally grouped into clusters which associate articles that are about the same topic and were published on the same date. The reason that comparable corpora may be a rich source of paraphrases is the fact that they describe the same set of basic facts (for instance that *a tsunami caused some number of deaths* and that *relief efforts are undertaken by various countries*), but different writers will express these facts differently.

Comparable corpora are like multiple translations in that both types of data contain different writers' descriptions of the same information. However, in multiple translations generally all of the same information is included, and pairings of sentences is relatively straightforward. With comparable corpora things are more complicated. Newspaper articles about the same topic will not necessarily include the same informa-

tion. They may focus on different aspects of the same events, or may editorialize about them in different ways. Furthermore, the organization of articles will be different. In multiple translations there is generally an assumption of linearity, but in comparable corpora finding equivalent sentences across news articles in a cluster is a difficult task.

A primary focus of research into using comparable corpora for paraphrasing has been how to discover pairs of sentences within a corpus that are valid paraphrases of each other. Dolan et al. (2004) defined two techniques to align sentences within clusters that are potential paraphrases of each other. Specifically, they find such sentences using: (1) a simple string edit distance filter, and (2) a heuristic that assumes initial sentences summarize stories. The first technique employs string edit distance to find sentences which have similar wording. The second technique uses a heuristic that pairs the first two sentences from news articles in the same clusters.

Here are two examples of sentences that are paired by Dolan et al.'s heuristics. Using string edit distance the sentence:

> Dzeirkhanov said 36 people were injured and that four people, including a child, had been hospitalized.

is paired with:

> Of the 36 wounded, four people including one child, were hospitalized, Dzheirkhanov said.

Using the heuristic which pairs the first two sentences across news stories in the same cluster, Dolan et al. matched:

> Two men who robbed a jeweler's shop to raise funds for the Bali bombings were each jailed for 15 years by Indonesian courts today.

with

> An Indonesian court today sentenced two men to 15 years in prison for helping finance last year's terrorist bombings in Bali by robbing a jewelry store.

Dolan et al. used the two heuristics to assemble two corpora containing sentences pairs such as these. It is only after distilling sentences pairs from a comparable corpus that it can be used for paraphrase extraction. Before applying the heuristics there is no way of knowing which portions of the corpus describe the same information.

Quirk et al. (2004) used the sentences which were paired by the string edit distance method as a source of data for their automatic paraphrasing technique. Quirk et al. treated these pairs of sentences as a 'parallel corpus' and viewed paraphrasing as

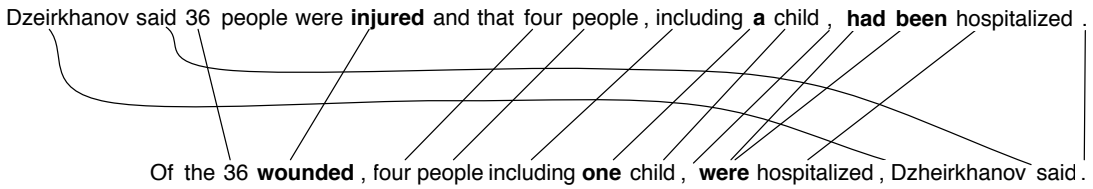


Figure 2.3: Quirk et al. (2004) extracted paraphrases from word alignments created from a 'parallel corpus' consisting of pairs of similar sentences from a comparable corpus

'monolingual machine translation.' They applied techniques from SMT (which are described in more detail in Section 2.2) to English sentences aligned with other English sentences, rather than applying these techniques to the bilingual parallel corpora that they are normally applied to. Rather than discovering the correspondences between English words and their foreign counterparts, Quirk et al. used statistical translation to discover correspondences between different English words. Figure 2.3 shows an automatic word alignment for one of the sentence pairs in the corpus, where each line denotes a correspondence between words in the two sentences. These correspondences include not only identical words, but also pairs non-identical words such as *wounded* with *injured*, and *one* with *a*. Non-identical words and phrases that were connected via word alignments were treated as paraphrases.

While comparable corpora are a more abundant source of data than multiple translations, and while they initially seem like a ready source of paraphrases since they contain different authors' descriptions of the same facts, they are limited in two significant ways. Firstly, there are difficulties associated with drawing pairs of sentences with equivalent meaning from comparable corpora that were not present in multiple translation corpora. Dolan et al. (2004) proposed two heuristics for pairing equivalent sentences, but the "first two sentences" heuristic was not usable in the paraphrasing technique of Quirk et al. (2004) because the sentences were not sufficiently close. Secondly, the heuristics for pairing equivalent sentences have the effect of greatly reducing the size of the comparable corpus, thus minimizing its primary advantage. Dolan et al.'s comparable corpus contained 177,095 news articles containing a total of 2,742,823 sentences and 59,642,341 words before applying their heuristics. When they apply the string edit distance heuristic they winnow the corpus down to 135,403 sentence pairs containing a total of 2,900,260 words. The "first two sentences" heuristic yields 213,784 sentence pairs with a total of 4,981,073 words. These numbers pale

in comparison to the amount of bilingual parallel corpora. Even when they are combined the size of the two corpora still barely tops the size of the multiple translation corpora used in previous research.

2.1.4 Paraphrasing with monolingual corpora

Another data source that has been used for paraphrasing is plain monolingual corpora. Monolingual data is more common than any other type of data used for paraphrasing. It is clearly more abundant than multiple translations, than comparable corpora, and than the English portion of bilingual parallel corpora, because all of those types of data constitute subsets of plain monolingual data. Because of its abundance, plain monolingual data should not be affected by the problems of availability that are associated with multiple translations or filtered comparable corpora. However, plain monolingual data is not a "natural" source of paraphrases in the way that the other two types of data are. It does not contain large numbers of sentences which describe the same information but are worded differently. Therefore the process of extracting paraphrases from monolingual corpora is more complicated.

Data-driven paraphrasing techniques which use monolingual corpora are based on a principle known as the Distributional Hypothesis (Harris, 1954). Harris argues that synonymy can determined by measuring the distributional similarity of words. Harris (1954) gives the following example:

> If we consider *oculist* and *eye-doctor* we find that, as our corpus of utterances grows, these two occur in almost the same environments. If we ask informants for any words that may occupy the same place as *oculist* in almost any sentence we would obtain *eye-doctor*. In contrast, there are many sentence environments in which *oculist* occurs but *lawyer* does not. ... It is a question of whether the relative frequency of such environments with *oculist* and with *lawyer*, or of whether we will obtain *lawyer* here if we ask an informant to substitute any word he wishes for *oculist* (not asking what words have the same meaning). These and similar tests all measure the probability of particular environments occurring with particular elements ... If A and B have almost identical environments we say that they are synonyms, as is the case with *oculist* and *eye-doctor*.

Lin and Pantel (2001) extracted paraphrases from a monolingual corpus based on Harris's Distributional Hypothesis using the distributional similarities of dependency relationships. They give the example of the words *duty* and *responsibility*, which share similar syntactic contexts. For example, both *duty* and *responsibility* can be modified by adjectives such as *additional, administrative, assumed, collective, congressional,*

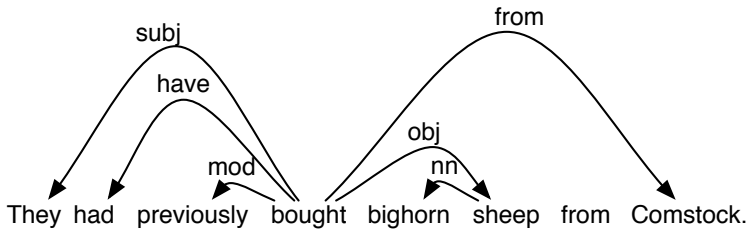


Figure 2.4: Lin and Pantel (2001) extracted paraphrases which had similar syntactic contexts using dependancy parses like this one

constitutional, and so on. Moreover they both can be the object of verbs such as *accept, assert, assign, assume, attend to, avoid, breach,* and so forth. The similarity of *duty* and *responsibility* is determined by analyzing their common contexts in a parsed monolingual corpus. Lin and Pantel used Minipar (Lin, 1993) to assign dependency parses like the one shown in Figure 2.4 to all sentences in a large monolingual corpus. They measured the similarity between *paths* in the dependency parses using *mutual information*. Paths with high mutual information, such as *X finds solution to Y* $\approx$ *X solves Y*, were defined as paraphrases.

The primary advantage of using plain monolingual corpora as a source of data for paraphrasing is that they are the most common kind of text. However, monolingual corpora don't have paired sentences as with the previous two types of texts. Therefore paraphrasing techniques which use plain monolingual corpora make the assumption that similar things appear in similar contexts. Techniques such as Lin and Pantel's method defines "similar contexts" through the use of dependency parses. In order to apply this technique to a monolingual corpus in a particular language, there must first be a parser for that language. Since there are many languages that do not yet have parsers, Lin and Pantel's paraphrasing technique can only be applied to a few languages.

Whereas Lin and Pantel's paraphrasing technique is limited to a small number of languages because it requires language-specific parsers, our paraphrasing technique has no such constraints and is therefore is applicable to a much wider range of languages. Our paraphrasing technique uses *bilingual parallel corpora*, a source of data which has hitherto not been used for paraphrasing, and is based on techniques drawn from statistical machine translation. Because statistical machine translation is formulated in a language-independent way, our paraphrasing technique can be applied to any language which has a bilingual parallel corpus. The number of languages which have

English	French
Spain declined to confirm that Spain declined to aid Morocco.	L' Espagne a refusé de confirmer que l' Espagne avait refusé d' aider le Maroc.
We note that the situation is changing every day.	Force est de constater que la situation évolue chaque jour .
We see that the French government has sent a mediator.	Nous voyons que le gouvernement français a envoyé un médiateur .
Mr. President, I would like to ask a question.	Monsieur le président, je voudrais poser une question.
Can we ask the bureau to look into this fact?	Nous voudrions demander au bureau d ' examiner cette affaire?
. . .	. . .

Figure 2.5: Parallel corpora are made up of translations aligned at the sentence level

such a resource is certainly far greater than the number of languages that have dependency parsers, and thus our paraphrasing technique can be applied to a much larger number of languages. This is useful when paraphrasing is integrated into other natural language processing tasks such machine translation (as detailed in Chapter 5).

The nature of bilingual parallel corpora and they way that they are used for statistical machine translation is explained in the next section. Chapter 3 then details how bilingual parallel corpora can be used for paraphrasing.

2.2 The use of parallel corpora for statistical machine translation

Parallel corpora consist of sentences in one language paired with their translations into another language, as in Figure 2.5. Parallel corpora form basis for data-driven approaches to machine translation such as example-based machine translation (Nagao, 1981), and statistical machine translation (Brown et al., 1988). Both approaches learn sub-sentential units of translation from the sentence pairs in a parallel corpus and reuse these fragments in subsequent translations. For instance, Sato and Nagao (1990) showed how an example-based machine translation (EBMT) system can use phrases in a Japanese-English parallel corpus to translate a novel input sentence like *He buys a book on international politics.* If the parallel corpus includes a sentence pair that

contains the translation of the phrase *he buys*, such as:

He buys a notebook.
Kare ha nouto *wo kau*.

And another which contains the translation of *a book on international politics*, such as:

I read *a book on international politics*.
Watashi ha *kokusaiseiji nitsuite kakareta hon* wo yomu

The EBMT system can use these two sentence pairs to produce the Japanese translation *(Kare ha) (kokusaiseiji nitsuite kakareta hon) (wo kau).* One of the primary tasks for both EBMT and SMT is to identify the correspondence between sub-sentential units in their parallel corpora, such as *a notebook* → *nouto*.

In Sections 2.2.1 and 2.2.2 we examine the mechanisms employed by SMT to align words and phrases within parallel corpora. We focus on the techniques from statistical machine translation because they form the basis of our paraphrasing method, because SMT has become the dominant paradigm in machine translation in recent years and repeatedly has been shown to achieve state-of-the-art performance. For an overview of EBMT and an examination of current research trends in that area, we point the interested reader to Somers (1999) and Carl and Way (2003), respectively.

2.2.1 Word-based models of statistical machine translation

Brown et al. (1990) proposed that translation could be treated as a probabilistic process in which every sentence in one language is viewed as a potential translation of a sentence in the other language. To rank potential translations, every pair of sentences $(\mathbf{f}, \mathbf{e})$ is assigned a probability $p(\mathbf{e}|\mathbf{f})$. The best translation $\hat{\mathbf{e}}$ is the sentence that maximizes this probability. Using Bayes' theorem Brown et al. decomposed the probability into two components:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) \tag{2.1}$$

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} p(\mathbf{e})p(\mathbf{f}|\mathbf{e}) \tag{2.2}$$

The two components are $p(\mathbf{e})$ which is a *language model* probability, and $p(\mathbf{f}|\mathbf{e})$ which is a *translation model* probability. The language model probability does not depend on the foreign language sentence $\mathbf{f}$. It represents the probability that the $\mathbf{e}$ is a valid sentence in English. Rather than trying to model valid English sentences in terms

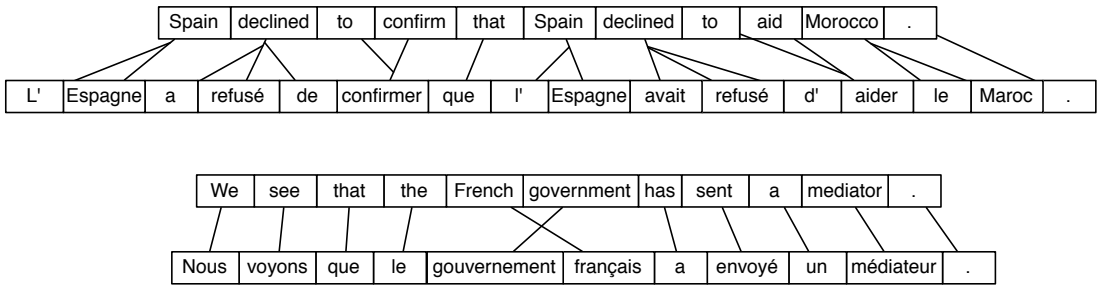


Figure 2.6: Word alignments between two sentence pairs in a French-English parallel corpus

of grammaticality, Brown et al. borrow n-gram language modeling techniques from speech recognition. These language models assign a probability to an English sentence by examining the sequence of words that comprise it. For $\mathbf{e} = e_1\ e_2\ e_3 ...\ e_n$, the language model probability $p(\mathbf{e})$ can be calculated as:

$$p(e_1\ e_2\ e_3 ...\ e_n) = p(e_1)p(e_2|e_1)p(e_3|e_1\ e_2)...p(e_n|e_1\ e_2\ e_3 ...\ e_{n-1}) \tag{2.3}$$

This formulation disregards syntactic structure, and instead recasts the language modeling problem as one of computing the probability of a single word given all of the words that precede it in a sentence. At any point in the sentence we must be able to determine the probability of a word, e_j, given a history, $e_1\ e_2\ ...\ e_{j-1}$. In order to simplify the task of parameter estimation for n-gram models, we reduce the length of the histories to be the preceding $n-1$ words. Thus in an trigram model we would only need to be able to determine the probability of a word, e_j, given a shorter history, $e_{j-2}\ e_{j-1}$. Although n-gram models are linguistically simpleminded they have the redeeming feature that it is possible to estimate their parameters from plain monolingual data.

The design of a translation model has similar trade-offs to the design of a language model. In order to create a translation model whose parameters can be estimated from data (which in this case is a parallel corpus) Brown et al. eschew linguistic sophistication in favor of a simpler model. They ignore syntax and semantics and instead treat translation as a word-level operation. They define the translation model probability $p(\mathbf{f}|\mathbf{e})$ in terms of possible word-level alignments, $\mathbf{a}$, between the sentences:

$$p(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} p(\mathbf{f}, \mathbf{a}|\mathbf{e}) \tag{2.4}$$

Just as n-gram language models can be defined in such a way that their parameters can be estimated from data, so can $p(\mathbf{f}, \mathbf{a}|\mathbf{e})$. Introducing word alignments simplifies the

translation probabilities	$t(f_j\|e_i)$	The probability that a foreign word f_j is the translation of an English word e_i.
fertility probabilities	$n(\phi_i\|e_i)$	The probability that a word e_i will expand into ϕ_i words in the foreign language.
spurious word probability	p	The probability that a spurious word will be inserted at any point in a sentence.
distortion probabilities	$d(p_i\|i,l,m)$	The probability that a target position p_i will be chosen for a word given the index of the English word that this was translated from i, and the lengths l and m of the English and foreign sentences.

Table 2.1: The IBM Models define translation model probabilities in terms of a number of parameters, including translation, fertility, distortion, and spurious word probabilities.

problem of determining whether a sentence is a good translation of another into the problem of determining whether there is a sensible mapping between the words in the sentences, like in the alignments in Figure 2.6.

Brown et al. defined a series of increasingly complex translation models, referred to as the IBM Models, which define $p(\mathbf{f},\mathbf{a}|\mathbf{e})$. IBM Model 3 defines word-level alignments in terms of four parameters. These parameters include a word-for-word translation probability, and three less intuitive probabilities (fertility, spurious word, and distortion) which account for English words that are aligned to multiple foreign words, words with no counterparts in the foreign language, and word re-ordering across languages. These parameters are explained in Table 2.1. The probability of an alignment $p(\mathbf{f},\mathbf{a}|\mathbf{e})$ is calculated under IBM Model 3 as:[1]

$$p(\mathbf{f},\mathbf{a}|\mathbf{e}) = \prod_{i=1}^{l} n(\phi_i|e_i) * \prod_{j=1}^{m} t(f_j|e_i) * \prod_{j=1}^{m} d(j|a_j,l,m) \tag{2.5}$$

If a bilingual parallel corpus contained explicit word-level alignments between its sentence pairs, like in Figure 2.6, then it would be possible to directly estimate the parameters of the IBM Models using maximum likelihood estimation. However, since word-aligned parallel corpora do not generally exist, the parameters of the IBM Models must be estimated without explicit alignment information. Consequently, alignments

[1]The true equation also includes the probabilities of spurious words arising from the "NULL" word at position zero of the English source string, but it is simplified here for clarity.

are treated as hidden variables. The expectation maximization (EM) framework for maximum likelihood estimation from incomplete data (Dempster et al., 1977) is used to estimate the values of these hidden variables. EM consists of two steps that are iteratively applied:

- The E-step calculates the posterior probability under the current model of every possible alignment for each sentence pair in the sentence-aligned training corpus;
- The M-step maximizes the expected likelihood under the posterior distribution, $p(\mathbf{f}, \mathbf{a}|\mathbf{e})$, with respect to the model's parameters.

While EM is guaranteed to improve a model on each iteration, the algorithm is not guaranteed to find a globally optimal solution. Because of this the solution that EM converges on is greatly affected by initial starting parameters. To address this problem Brown et al. first train a simpler model to find sensible estimates for the t table, and then use those values to prime the parameters for incrementally more complex models which estimate the d and n parameters described in Table 2.1. IBM Model 1 is defined only in terms of word-for-word translation probabilities between foreign words f_j and the English words e_{a_j} which they are aligned to:

$$p(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.6}$$

IBM Model 1 produces estimates for the the t probabilities, which are used at the start EM for the later models.

Beyond the problems associated with EM and local optima, the IBM Models face additional problems. While Equation 2.4 and the E-step call for summing over all possible alignments, this is intractable because the number of possible alignments increases exponentially with the lengths of the sentences. To address this problem Brown et al. did two things:

- They performed approximate EM wherein they sum over only a small number of the most probable alignments instead of summing over all possible alignments.
- They limited the space of permissible alignments by ignoring many-to-many alignments and permitting one-to-many alignments only in one direction.

Och and Ney (2003) undertook systematic study of the IBM Models. They trained the IBM Models on various sized German-English and French-English parallel corpora

and compare the most probable alignments generated by the models against reference word alignments that were manually created. They found that increasing the amount of data improved the quality of the automatically generated alignments, and that the more complex of the IBM Models performed better than the simpler ones.

Improving alignment quality is one way of improving translation models. Thus word alignment remains an active topic of research. Some work focuses on improving on the training procedures used by the IBM Models. Vogel et al. (1996) used Hidden Markov Models. Callison-Burch et al. (2004) re-cast the training procedure as a partially supervised learning problem by incorporating explicitly word-aligned data alongside the standard sentence-aligned training data. Fraser and Marcu (2006) did similarly. Moore (2005); Taskar et al. (2005); Ittycheriah and Roukos (2005); Blunsom and Cohn (2006) treated the problem as a fully supervised learning problem and apply discriminative training. Still others have focused on improving alignment quality by integrating linguistically motivated constraints (Cherry and Lin, 2003).

The most promising direction in improving translation models has been to move beyond word-level alignments to phrase-based models. These are described in the next section.

2.2.2 From word- to phrase-based models

Whereas the original formulation of statistical machine translation was word-based, contemporary approaches have expanded to phrases. Phrase-based statistical machine translation (Och and Ney, 2002; Koehn et al., 2003) uses larger segments of human translated text. By increasing the size of the basic unit of translation, phrase-based SMT does away with many of the problems associated with the original word-based formulation. In particular, Brown et al. (1993) did not have a direct way of translating phrases; instead they specified the *fertility* parameter which is used to replicate words and translate them individually. Furthermore, because words were their basic unit of translation, their models required a lot of reordering between languages with different word orders, but the *distortion* parameter was a poor explanation of word order. Phrase-based SMT eliminated the fertility parameter and directly handled word-to-phrase and phrase-to-phrase mappings. Phrase-based SMT's use of multi-word units also reduced the dependency on the distortion parameter. In phrase-based models less word re-ordering needs to occur since local dependencies are frequently captured. For example, common adjective-noun alternations are memorized, along with other fre-

quently occurring sequences of words. Note that the 'phrases' in phrase-based translation are not congruous with the traditional notion of syntactic constituents; they might be more aptly described as 'substrings' or 'blocks' since they just denote arbitrary sequences of contiguous words. Koehn et al. (2003) showed that using these larger chunks of human translated text resulted in high quality translations, despite the fact that these sequences are not syntactic constituents.

Phrase-based SMT calculates a phrase translation probability $p(\bar{f}|\bar{e})$ between an English phrase $\bar{e}$ and a foreign phrase $\bar{f}$. In general the phrase translation probability is calculated using maximum likelihood estimation by counting the number of times that the English phrase was aligned with the French phrase in the training corpus, and dividing by the total number of times that the English phrase occurred:

$$p(\bar{f}|\bar{e}) = \frac{count(\bar{f},\bar{e})}{count(\bar{e})} \tag{2.7}$$

In order to use this maximum likelihood estimator it is crucial to identify phrase-level alignments between phrases that occur in sentence pairs in a parallel corpus.

Many methods for identifying phrase-level alignments use word-level alignments as a starting point. Och and Ney (2003) defined one such method. Their method first creates a word-level alignment for each sentence pair in the parallel corpus by outputting the alignment that is assigned the highest probability by the IBM Models. Because the IBM Models only allow one-to-many alignments in one language direction they have an inherent asymmetry. In order to overcome this, Och and Ney train models in both the E→F and F→E directions, and symmetrize the word alignments by taking the union of the two alignments. This is illustrated in Figure 2.7. This creates a single word-level alignment for each sentence pair, which can contain one-to-many alignments in both directions. However, these symmetrized alignments do not have many-to-many correspondences which are necessary for phrase-to-phrase alignments.

Och and Ney (2004) defined a method for extracting incrementally longer phrase-to-phrase correspondences from a word alignment, such that the phrase pairs are *consistent* with the word alignment. Consistent phrase pairs are those in which all words within the source language phrase are aligned only with the words of the target language phrase and the words of the target language phrase are aligned only with the words of the source language phrase. Och and Ney's phrase extraction technique is illustrated in Figure 2.8. In the first iteration, bilingual phrase pairs are extracted directly from the word alignment. This allows single words to translate as phrases, as with *grandi* → *grown up*. Larger phrase pairs are then created by incorporating ad-

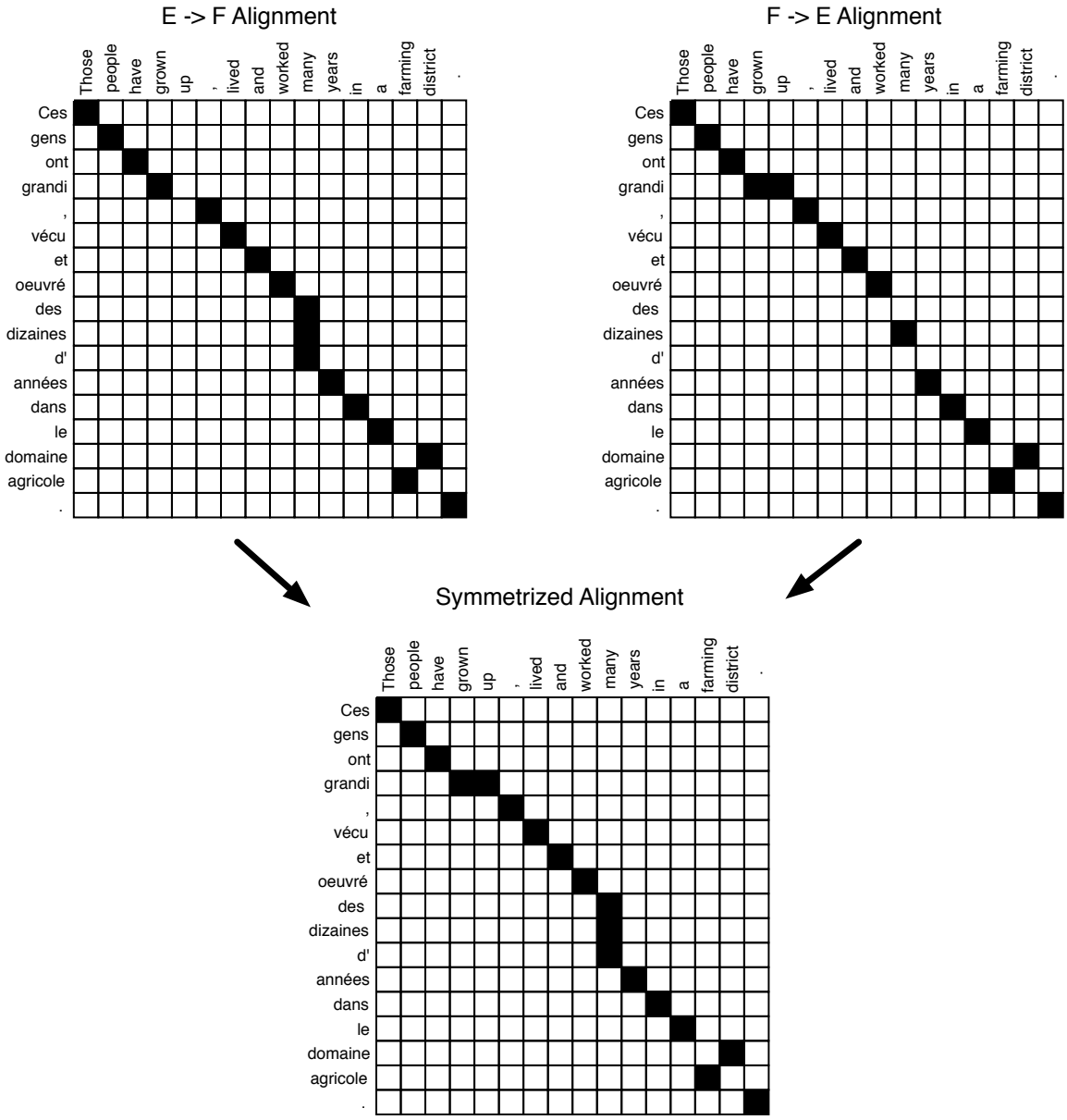


Figure 2.7: Och and Ney (2003) created ‘symmetrized’ word alignments by merging the output of the IBM Models trained in both language directions

jacent words and phrases. In the second iteration the phrase *a farming* does not have a translation since there is not a phrase on the foreign side which is consistent with it. It cannot align with *le domaine* or *le domaine agricole* since they have a point that fall outside the phrase alignment (*domaine, district*). On the third iteration *a farming district* now has a translation since the French phrase *le domaine agricole* is consistent with it.

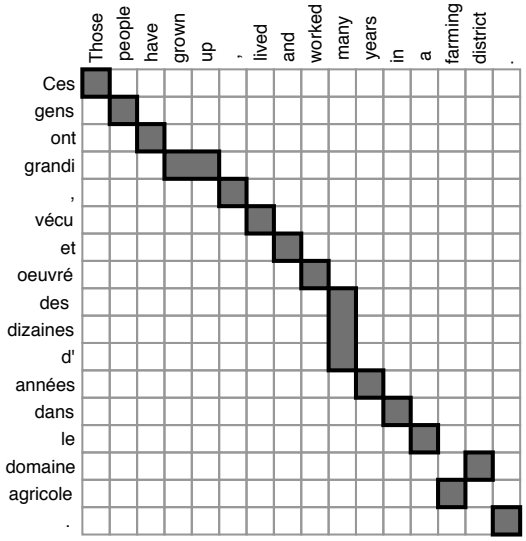
To calculate the maximum likelihood estimate for phrase translation probabilities the phrase extraction technique is used to enumerate all phrase pairs up to a certain length for all sentence pairs in the training corpus. The number of occurrences of each of these phrases are counted, as are the total number of times that pairs co-occur. These are then used to calculate phrasal translation probabilities, using Equation 2.7. This process can be done with Och and Ney's phrase extraction technique, or a number of variant heuristics. Other heuristics for extracting phrase alignments from word alignments were described by Vogel et al. (2003), Tillmann (2003), and Koehn (2004).

As an alternative to extracting phrase-level alignments from word-level alignments, Marcu and Wong (2002) estimated them directly. They use EM to estimate phrase-to-phrase translation probabilities with a model defined similarly to IBM Model 1, but which does not constrain alignments to be one-to-one in the way that IBM Model 1 does. Because alignments are not restricted in Marcu and Wong's model, the huge number of possible alignments makes computation intractable, and thus makes it impossible to apply to large parallel corpora. Recently, Birch et al. (2006) made strides towards scaling Marcu and Wong's model to larger data sets by putting constraints on what alignments are considered during EM, which shows that calculating phrase translation probabilities directly in a theoretically motivated may be more promising than Och and Ney's heuristic phrase extraction method.

The phrase extraction techniques developed in SMT play a crucial role in our data-driven paraphrasing technique which is described in Chapter 3.

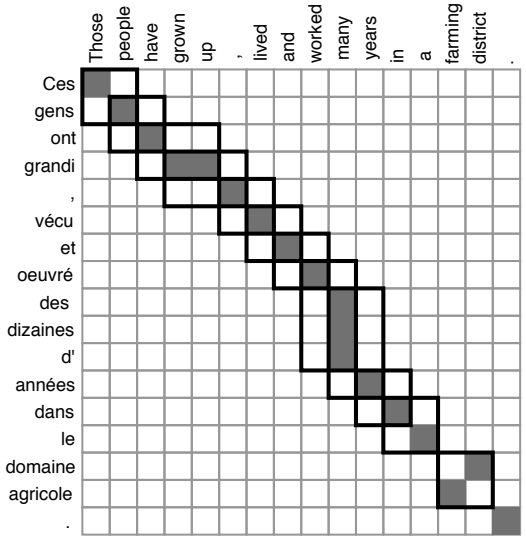
2.2.3 The decoder for phrase-based models

The decoder is the software which uses the statistical translation model to produce translations of novel input sentences. For a given input sentence the decoder first breaks it into subphrases and enumerates all alternative translations that the model has learned for each subphrase. This is illustrated in Figure 2.9. The decoder then chooses among these phrasal translations to create a translation of the whole sentence. Since



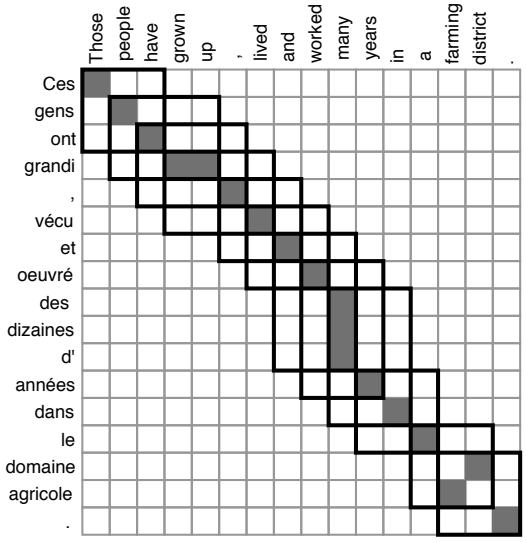
Phrase pairs extracted on iteration 1:

Ces	Those
gens	people
ont	have
grandi	grown up
,	,
vécu	lived
et	and
oeuvré	worked
des dizaines d'	many
années	years
dans	in
le	a
domaine	district
agricole	farming
.	.



Iteration 2:

Ces gens	Those people
gens ont	people have
ont grandi	have grown up
grandi ,	grown up ,
, vécu	, lived
vécu et	lived and
et oeuvré	and worked
oeuvré des dizaines d'	worked many
des dizaines d' années	many years
années dans	year in
dans le	in a
domaine agricole	farming district



Iteration 3:

Ces gens ont	Those people have
gens ont grandi	people have grown up
ont grandi ,	have grown up ,
grandi , vécu	grown up , lived
, vécu et	, lived and
vécu et oeuvré	lived and worked
et oeuvré des dizaines d'	and worked many
oeuvré des dizaines d' années	worked many years
des dizaines d' années dans	many years in
années dans le	years in a
le domaine agricole	a farming districtle
domaine agricole .	farming district .

Figure 2.8: Och and Ney (2004) extracted incrementally larger phrase-to-phrase correspondences from word-level alignments

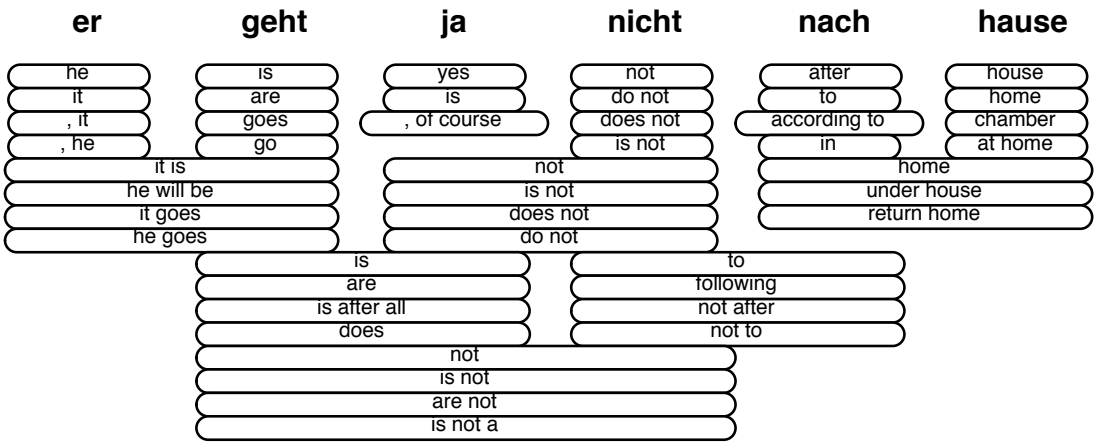


Figure 2.9: The decoder enumerates all translations that have been learned for the subphrases in an input sentence

there are many possible ways of combining phrasal translations the decoder considers a large number of partial translations simultaneously. This creates a *search space* of hypotheses, as shown in Figure 2.10. These hypotheses are *ranked* by assigning a cost or a probability to each one. The probability is assigned by the statistical translation model.

Whereas the original formulation of statistical machine translation (Brown et al., 1990) used a translation model that contained two separate probabilities:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) \tag{2.8}$$

$$= \arg\max_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e}) \tag{2.9}$$

contemporary approaches to SMT instead employ a log linear formulation (Och and Ney, 2002), which breaks the probability down into an arbitrary number of weighted feature functions:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) \tag{2.10}$$

$$= \arg\max_{\mathbf{e}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e},\mathbf{f}) \tag{2.11}$$

The advantage of the log linear formulation is that rather than just having a translation model probability and a language model probability assign costs to translation, we can now have an arbitrary number of feature functions, $h(\mathbf{e},\mathbf{f})$ which assign a cost to a translation. In practical terms this gives us a mechanism to break down the assignation of cost in a modular fashion based on different aspects of translation. In current

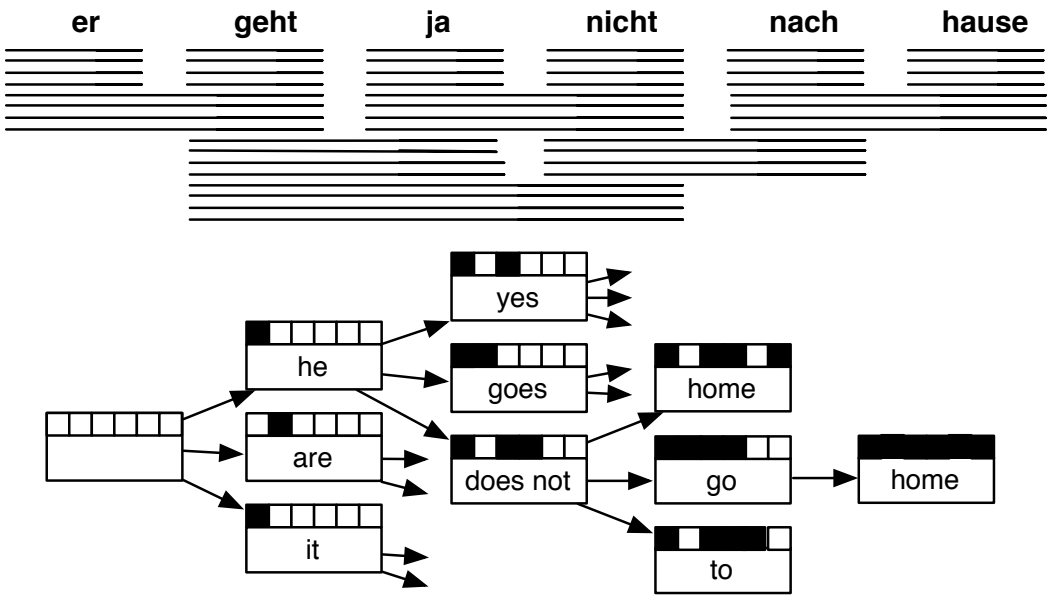


Figure 2.10: The decoder assembles translation alternatives, creating a search space over possible translations of the input sentence. In this figure the boxes represents a *coverage vector* that shows which source words have been translated. The best translation is the hypothesis with the highest probability when all source words have been covered.

systems the feature functions that are most commonly used include a language model probability, a phrase translation probability, a reverse phrase translation probability, lexical translation probability, a reverse lexical translation probability, a word penalty, a phrase penalty, and a distortion cost.

The weights, λ, in the log linear formulation act to set the relative contribution of each of the feature functions in determining the best translation. The Bayes' rule formulation (Equation 2.9) assigns equal weights to the language model and the translation model probabilities. In the log linear formulation these may play a greater or lesser role depending on their weights. The weights can be set in an empirical fashion in order to maximize the quality of the MT system's output for some development set (where human translations are given). This is done through a process known as minimum error rate training (Och, 2003), which uses an objective function to compare the MT output against the reference human translations and minimizes their differences. Modulo the potential of over-fitting the development set, the incorporation of additional feature functions should not have a detrimental effect on the translation quality

because of the way that the weights are set.

2.2.4 The phrase table

The decoder uses a data structure called a *phrase table* to store the source phrases paired with their translations into the target language, along with the value of feature functions that relate to translation probabilities.[2] The phrase table contains an exhaustive list of all translations which have been extracted from the parallel training corpus. The source phrase is used as a key that is used to look up the translation options, as in Figure 2.9, which shows the translation options that the decoder has for subphrases in the input German sentence. These translation options are learned from the training data and stored in the phrase table. If a source phrase does not appear in the phrase table, then the decoder has no translation options for it.

Because the entries in the phrase table act as basis for the behavior of the decoder – both in terms of the translation options available to it, and in terms of the probabilities associated with each entry – it is a common point of modification in SMT research. Often people will augment the phrase table with additional entries that were not learned from the training data directly, and show improvements without modifying the decoder itself. We do similarly in our experiments, which are explained in Chapter 7.

2.3 A problem with current SMT systems

One of the major problems with SMT is that it is slavishly tied to the particular words and phrases that occur in the training data. Current models behave very poorly on unseen words and phrases. When a word is not observed in the training data most current statistical machine translation systems are simply unable to translate it. The problems associated with translating unseen words and phrases are exacerbated when only small amounts of training data are available, and when translating with morphologically rich languages, because fewer of the word forms will be observed. This problem can be characterized as a *lack of generalization* in statistical models of translation or as one of *data sparsity*.

[2] Alternative representations to the phrase table have been proposed. For instance, Callison-Burch et al. (2005) described a suffix array-based data structure, which contains an indexed representation of the complete parallel corpus. It looks up phrase translation options and their probabilities on-the-fly during decoding, which is computationally more expensive than a table lookup, but which allows SMT to be scaled to arbitrarily long phrases and much larger corpora than are currently used.

A number of research efforts have tried to address the problem of unseen words by integrating language-specific morphological information, allowing the SMT system to learn translations of base word forms. For example, Koehn and Knight (2003) showed how monolingual texts and parallel corpora could be used to figure out appropriate places to split German compound words so that the elements can be translated separately. Niessen and Ney (2004) applied morphological analyzers to English and German and were able to reduce the amount of training data needed to reach a certain level of translation quality. Goldwater and McClosky (2005) found that stemming Czech and using lemmas improved the word-to-word correspondences when training Czech-English alignment models. de Gispert et al. (2005) substituted lemmas for fully-inflected verb forms to partially reduce the data sparseness problem associated with the many possible verb forms in Spanish. Kirchhoff et al. (2006) applied morpho-syntatic knowledge to re-score Spanish-English translations. Yang and Kirchhoff (2006) introduced a back-off model that allowed them to translate unseen German words through a procedure of compound splitting and stemming. Talbot and Osborne (2006) introduced a language-independent method for minimizing what they call "lexical redundancy" by eliminating certain inflections used in one language which are not relevant when translating into another language. Talbot and Osborne showed improvements when their method is applied to Czech-English and Welsh-English translation.

Other approaches have focused on ways of acquiring data in order to overcome problems with data sparsity. Resnik and Smith (2003) developed a method for gathering parallel corpora from the web. Oard et al. (2003) described various methods for quickly gathering resources to create a machine translation system for a language with no initial resources.

In this thesis we take a different approach to address problems that arise when a particular word or phrase does not occur in the training data. Rather than trying to introduce language-specific morphological information as a preprocessing step or trying to gather more training data, we instead try to introduce some amount of *generalization* into the process through the use of *paraphrases*. Rather than being limited to translating only those words and phrases that occurred in the training data, external knowledge of paraphrases is used to produce new translations. Thus if the translation of a word has not been learned, but a translation of its synonym has been learned, then we will be able to translate it. Similarly, if we haven't learned the translation of a phrase, but have learned the translation of a paraphrase of it, then we are able to translate it accurately.

Chapter 3

Paraphrasing with Parallel Corpora

Paraphrases are useful in a wide variety of natural language processing tasks. In natural language generation the production of paraphrases allows for the creation of more varied and fluent text (Iordanskaja et al., 1991). In multidocument summarization the identification of paraphrases allows information repeated across documents to be recognized and for redundancies to be eliminated (McKeown et al., 2002). In the automatic evaluation of machine translation, paraphrases may help to alleviate problems presented by the fact that there are often alternative and equally valid ways of translating a text (Zhou et al., 2006). In question answering, paraphrased answers may provide additional evidence that an answer is correct (Ibrahim et al., 2003; Dalmas, 2007). Because of this wide range of potential applications, a considerable amount of recent research has focused on automatically learning paraphrase relationships (see Section 2.1 for a review of recent paraphrasing research). All data-driven paraphrasing techniques share the need for large amounts of data in the form of pairs or sets of sentences that are likely to exhibit paraphrase alternations. Sources of data for previous paraphrasing techniques include multiple translations, comparable corpora, and parsed monolingual texts.

In this chapter[1] we define a novel paraphrasing technique which utilizes *parallel corpora*, a type of data which is more commonly used as training data for statistical machine translation, and which has not previously been used for paraphrasing. In Section 3.1 we detail the challenges of using this resource which were not present with previous resources, and describe how we extract paraphrases using techniques from phrase-based statistical machine translation. In Section 3.2 we lay out a probabilistic

[1]Chapters 3 and 4 extend the exposition and analysis presented in Bannard and Callison-Burch (2005) which was joint work with Colin Bannard. The experimental results are the same as in the previously published work.

treatment of paraphrasing, which allows alternative paraphrases to be ranked by their likelihood. Having a mechanism for ranking paraphrases is important because our technique extracts multiple paraphrases for each phrase, and because the quality and accuracy of paraphrases can vary depending on the contexts that they are substituted into. In Section 3.3 we discuss a number of factors which influence paraphrase quality within our setup. In Section 3.4 we describe how we can take these factors into account by refining the paraphrase probability. Chapter 4 delineates the experiments that we conducted to investigate the quality of the paraphrases generated by our technique.

3.1 The use of parallel corpora for paraphrasing

Parallel corpora are very different from the types of data that have been used in other paraphrasing efforts. Parallel corpora consist of sentences in one language paired with their translations into another language (as illustrated in Figure 2.5). Multiple translation corpora and filtered comparable corpora also consist of pairs of sentences that are equivalent in meaning. However, their sentences are in a single language, making them a natural source for paraphrases. Simple heuristics can be used to extract paraphrases from such data, like Barzilay and McKeown's rule of thumb that phrases which are surrounded by identical words in their paired sentences are good paraphrases (illustrated in Figure 2.1). The process of extracting paraphrases from parallel corpora is less obvious, since their sentence pairs are in different languages and since they do not contain identical surrounding contexts.

Instead of extracting paraphrases directly from a single pair of sentences, our paraphrasing technique uses many sentence pairs. We use phrases in the other language as pivots. To extract English paraphraseswe look at what foreign language phrases the English translates to, find all occurrences of those foreign phrases, and then look at what other English phrases they originated from. We treat the other English phrases as potential paraphrases. Figure 3.2 illustrates how a German phrase can be used to discover that *in check* is a paraphrase of *under control*. To align English phrases with their German counterparts we use techniques from phrase-based statistical machine translation, which are detailed in Section 2.2.2.[2]

[2]The phrase extraction techniques that we adopt in this work operate on contiguous sequences of words. Recent work has extended statistical machine translation to operate on hierarchical phrases which allow embedded or discontinuous elements (Chiang, 2007). We could extend our method to hierarchical phrases, which would allow us to extract paraphrases with variables like *rub X on Y* $\Leftrightarrow$ *apply X to Y*, which are not currently handled by our framework.

Note that while the examples in this chapter illustrate how parallel corpora can be used to generate English paraphrases there is nothing that limits us to English. Chapters 5 and 7 give example Spanish and French paraphrases. All methods presented here can be applied to any other languages which have parallel corpora, and will work to the same extent that the language-independent mechanisms of statistical machine translation do.

Rather than extracting paraphrases directly from a single pair of English sentences with equivalent meaning (as in previous paraphrasing techniques), we use foreign language phrases as pivots and search across the entire corpus. As a result, our method frequently extracts more than one possible paraphrase for each phrase, because each instance of the English phrase can be aligned to a different foreign phrase, and each foreign phrase can be aligned to different English phrases. Figure 3.1 illustrates this. The English phrase *military force* is aligned with the German phrases *truppe*, *streikräfte*, *streikräften*, and *friedenstruppe* in different instances. At other points in the corpus these German phrases are aligned to other English phrases including *force*, *armed forces*, *forces*, *defense* and *peace-keeping personnel*. We treat all of these as potential paraphrases of the phrase *military force*. Moreover each German phrase can align to multiple English phrases, as with *streikräfte*, which connects with *armed forces* and *defense*.

Given that we frequently have multiple possible paraphrases, and given that the paraphrases are not always as good as those for *military force*, it is important to have a mechanism for ranking candidate paraphrases. To do this we define a paraphrase probability, which can be used to rank possible paraphrases and select the best one.

3.2 Ranking alternatives with a paraphrase probability

We define a paraphrase probability, $p(e_2|e_1)$, in a way that fits naturally with the fact that we use parallel corpora to extract paraphrases. Just as we are able to use alignment techniques from phrase-based statistical machine translation, we can take advantage of its translation model probabilities. We can define $p(e_2|e_1)$ in terms of the translation model probabilities $p(f|e_1)$, that the original English phrase e_1 translates as a particular phrase f in the other language, and $p(e_2|f)$, that the candidate paraphrase e_2 translates as that foreign language phrase. Since e_1 can translate as multiple foreign language

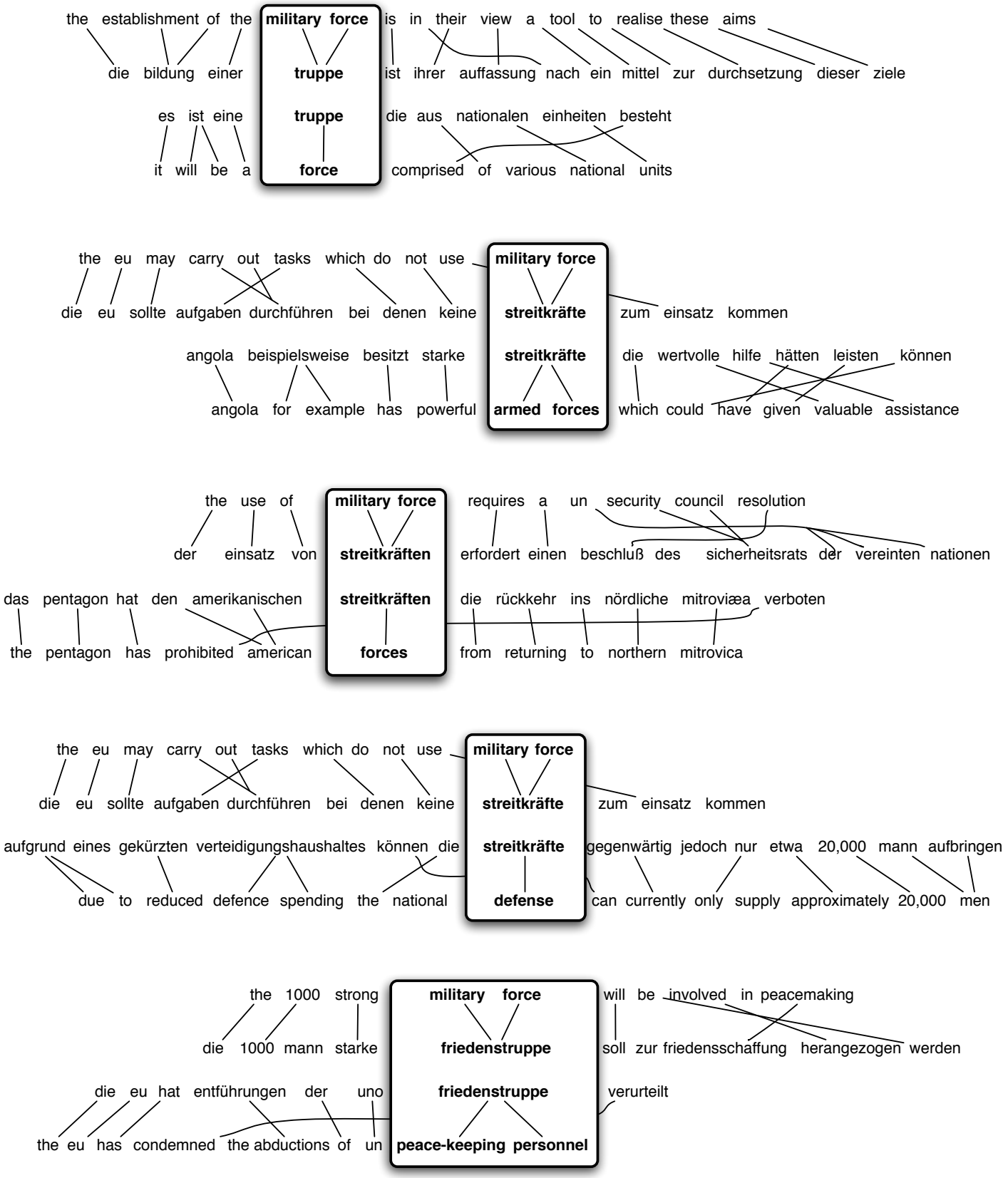


Figure 3.1: A phrase can be aligned to many foreign phrases, which in turn can be aligned to multiple possible paraphrases

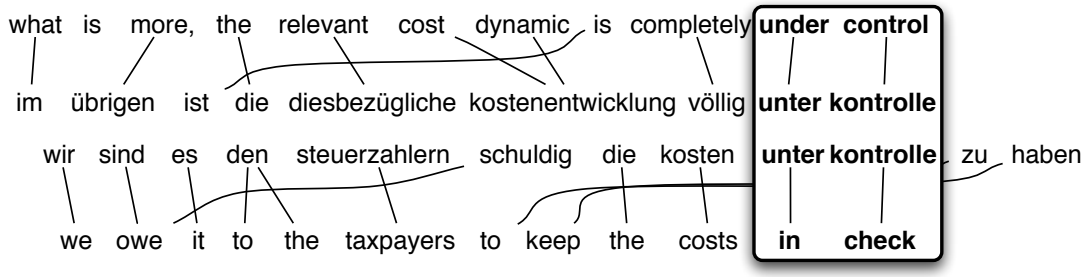


Figure 3.2: Using a bilingual parallel corpus to extract paraphrases

phrases, we sum over f:

$$\hat{e_2} = \arg\max_{e_2 \neq e_1} p(e_2|e_1) \quad (3.1)$$

$$= \arg\max_{e_2 \neq e_1} \sum_f p(f|e_1)p(e_2|f) \quad (3.2)$$

The translation model probabilities can be computed using any formulation from phrase-based machine translation including maximum likelihood estimation (as in Equation 2.7). Thus $p(f|e_1)$ and $p(e_2|f)$ can be calculated as:

$$p(f|e_1) = \frac{count(f,e_1)}{count(e_1)} \quad (3.3)$$

$$p(e_2|f) = \frac{count(e_2,f)}{count(f)} \quad (3.4)$$

Figure 3.3 gives counts for how often the phrase *military force* aligns with its German counterparts, and for how often those German phrases align with various English phrases in a German-English corpus. Based on these counts we can get the following values for $p(f|e_1)$:

$$\begin{aligned}
p(\textit{militärische gewalt} \mid \textit{military force}) &= 0.222 \\
p(\textit{truppe} \mid \textit{military force}) &= 0.222 \\
p(\textit{streitkräften} \mid \textit{military force}) &= 0.111 \\
p(\textit{streitkräfte} \mid \textit{military force}) &= 0.111 \\
p(\textit{militärischer gewalt} \mid \textit{military force}) &= 0.111 \\
p(\textit{friedenstruppe} \mid \textit{military force}) &= 0.111 \\
p(\textit{militärische eingreiftruppe} \mid \textit{military force}) &= 0.111
\end{aligned}$$

We get the following values $p(e_2|f)$:

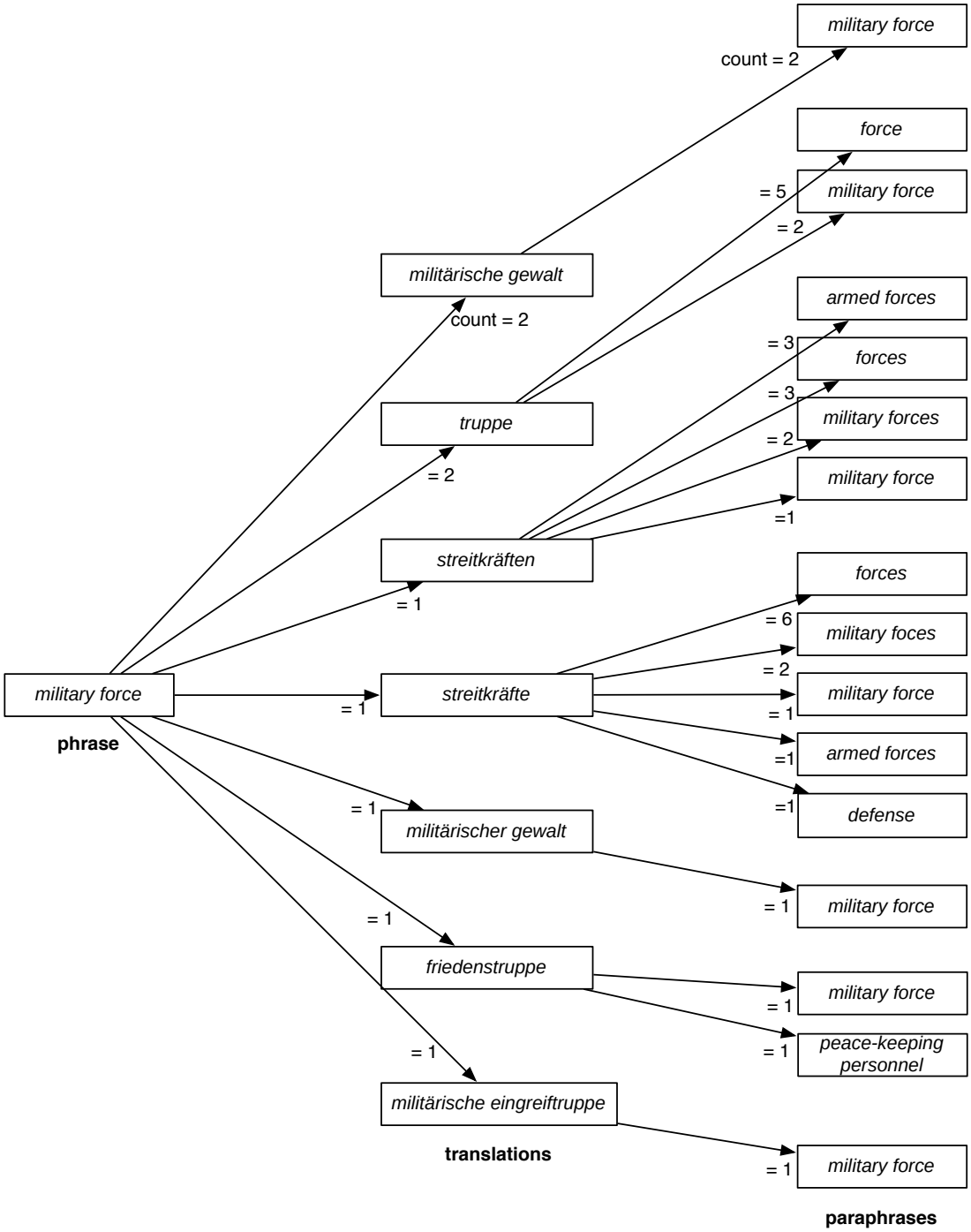


Figure 3.3: The counts of how often the German and English phrases are aligned in a parallel corpus with 30,000 sentence pairs. The arrows indicate which phrases are aligned and are labeled with their counts.

$$
\begin{aligned}
p(\textit{military force} \mid \textit{militärische gewalt}) &= 1.0 \\
p(\textit{force} \mid \textit{truppe}) &= 0.714 \\
p(\textit{military force} \mid \textit{truppe}) &= 0.286 \\
p(\textit{armed forces} \mid \textit{streitkräften}) &= 0.333 \\
p(\textit{forces} \mid \textit{streitkräften}) &= 0.333 \\
p(\textit{military forces} \mid \textit{streitkräften}) &= 0.222 \\
p(\textit{military force} \mid \textit{streitkräften}) &= 0.111 \\
p(\textit{forces} \mid \textit{streitkräfte}) &= 0.545 \\
p(\textit{military forces} \mid \textit{streitkräfte}) &= 0.181 \\
p(\textit{military force} \mid \textit{streitkräfte}) &= 0.09 \\
p(\textit{armed forces} \mid \textit{streitkräfte}) &= 0.09 \\
p(\textit{defense} \mid \textit{streitkräfte}) &= 0.09 \\
p(\textit{military force} \mid \textit{militärischer gewalt}) &= 1.0 \\
p(\textit{military force} \mid \textit{friedenstruppe}) &= 0.5 \\
p(\textit{peace-keeping personnel} \mid \textit{friedenstruppe}) &= 0.5 \\
p(\textit{military force} \mid \textit{militärische eingreiftruppe}) &= 1.0
\end{aligned}
$$

The values for the two translation model probabilities allow us to calculate the paraphrase probability $p(e_2|e_1)$ using Equation 3.1:

$$
\begin{aligned}
p(\textit{military force} \mid \textit{military force}) &= 0.588 \\
p(\textit{force} \mid \textit{military force}) &= 0.158 \\
p(\textit{forces} \mid \textit{military force}) &= 0.096 \\
p(\textit{peace-keeping personnel} \mid \textit{military force}) &= 0.055 \\
p(\textit{armed forces} \mid \textit{military force}) &= 0.047 \\
p(\textit{military forces} \mid \textit{military force}) &= 0.046 \\
p(\textit{defense} \mid \textit{military force}) &= 0.01
\end{aligned}
$$

Thus for the initial definition of the paraphrase probability given in Equation 3.2, the e_2 which maximizes $p(e_2|e_1)$ such that $e_2 \neq e_1$ would be the phrase *force*. We specify that $e_2 \neq e_1$ to ensure that the paraphrase is different from the original phrase. Notice that the sum of all the paraphrase probabilities is one. This is necessary in order for the paraphrase probability to be a proper probability distribution. This property is guaranteed based on the formulations of the translation model probabilities. Given the formulation in Equation 3.1 the values for $p(e_2|e_1)$ will always sum to one for any phrase e_1 when we use a single parallel corpus to estimate the parameters of the probability function.

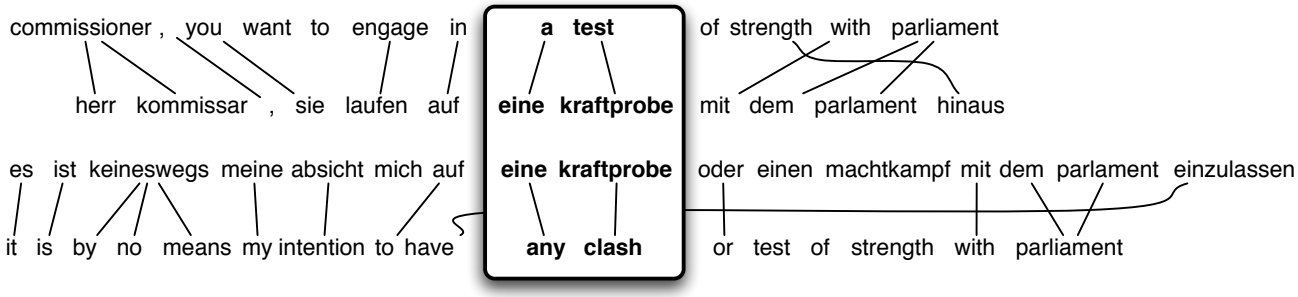


Figure 3.4: Incorrect paraphrases can occasionally be extracted due to misalignments, such as here, where *kraftprobe* should be aligned with *test of strength*

In the next section we examine some of the factors that affect the quality of the paraphrases that we extract from parallel corpora. In Section 3.4 we use these insights to refine the paraphrase probability in order to pick out better paraphrases.

3.3 Factors affecting paraphrase quality

There are a number of factors which can affect the quality of paraphrases extracted from parallel corpora. There are factors attributable to the fact that we are borrowing methods from SMT, and others which are associated with the assumptions we make when using parallel corpora. There are still more factors that are not specifically associated with our paraphrasing technique alone, but which apply more generally to all paraphrasing methods.

3.3.1 Alignment quality and training corpus size

Since we rely on statistical machine translation to align phrases across languages, we are dependent upon its *alignment quality*. Just as high quality alignments are required in order to produce good translations (Callison-Burch et al., 2004), they are also required to produce good paraphrases. If a phrase is misaligned in the parallel corpus then we may produce spurious paraphrases. For example Figure 3.4 shows how incorrect word alignments can lead to incorrect paraphrases. We extract *any clash* as a paraphrase of *a test* because the German phrase *kraftprobe* is misaligned (it should be aligned to *test of strength* in both instances). Since we are able to rank paraphrases based on their probabilities, occasional misalignments should not affect the best paraphrase. However, misalignments that are systematic may result in poor estimates of the two translation probabilities in Equations 3.3 and 3.4 and thus result in a different

$\hat{e_2}$ maximizing the paraphrase probability.

One way to improve the quality of the paraphrases that our technique extracts is to improve alignment quality. A significant amount of statistical machine translation research has focused on improving alignment quality by designing more sophisticated alignment models and improving estimation techniques (Vogel et al., 1996; Melamed, 1998; Och and Ney, 2003; Cherry and Lin, 2003; Moore, 2004; Callison-Burch et al., 2004; Ittycheriah and Roukos, 2005; Taskar et al., 2005; Moore et al., 2006; Blunsom and Cohn, 2006; Fraser and Marcu, 2006). Other research has also examined various ways of improving alignment quality through the automatic acquisition of large volumes of parallel corpora from the web (Resnik and Smith, 2003; Wu and Fung, 2005; Munteanu and Marcu, 2005, 2006). Small training corpora may also affect paraphrase quality in a manner unrelated to alignment quality, since they are plagued by sparsity. Many words and phrases will not be contained in the parallel corpus, and thus we will be unable to generate paraphrases for them.

In Section 3.4.1 we describe a method that helps to alleviate the problems associated with both misalignments and small parallel corpora. We show that paraphrases can be extracted from parallel corpora in multiple languages. Using a parallel corpus to learn a translation model necessitates a single language pair (English-German, for example). For paraphrasing we can use multiple parallel corpora. For instance, if we were creating English paraphrases we could use not only the English-German parallel corpus, but also parallel corpora between English and other languages, such as Arabic, Chinese, or Spanish. Using multiple languages minimizes the effect of systematic misalignments in one language. It also increases the number of words and phrases that we observe during training, thus effectively reducing sparsity.

3.3.2 Word sense

One fundamental assumption that we make when we extract paraphrases from parallel corpora is that phrases are *synonymous* when they are aligned to the same foreign language phrase. This is the converse of the assumption made in some word sense disambiguation literature which posits that a word is *polysemous* when it is aligned to different words in another language (Brown et al., 1991; Dagan and Itai, 1994; Dyvik, 1998; Resnik and Yarowksy, 1999; Ide, 2000; Diab, 2000; Diab and Resnik, 2002). Diab illustrates this assumption using the classic word sense example of *bank*, which can be translated into French either with the word *banque* (which corresponds

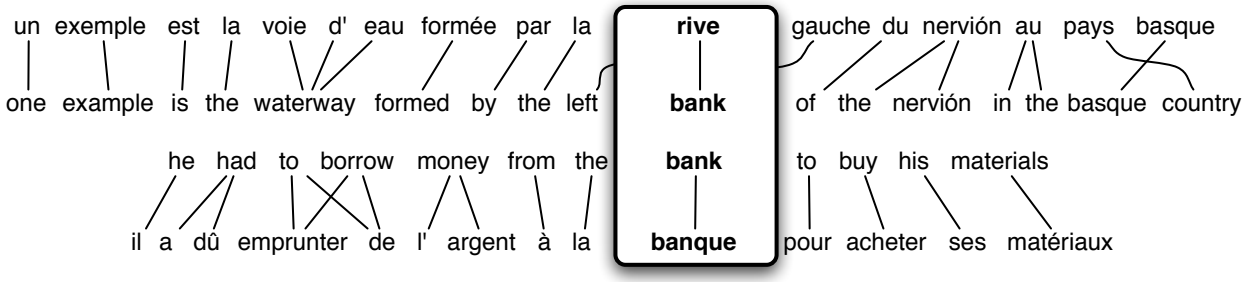


Figure 3.5: A polysemous word such as *bank* in English could cause our paraphrasing technique to extract incorrect paraphrases, such as equating *rive* with *banque* in French

to the *financial institution* sense of *bank*), or the word *rive* (which corresponds to the *riverbank* sense of *bank*). This example is used to motivate using word-aligned parallel corpora as source of training data for word sense disambiguation algorithms, rather than relying on data that has been manually annotated with WordNet senses (Miller, 1990). While constructing training data automatically is obviously less expensive, it is unclear to what extent multiple foreign words actually pick out distinct senses.

The assumption that a word which aligns with multiple foreign words has different senses is certainly not true in all cases. It would mean that *military force* should have many distinct senses, because it is aligned with many different German words in Figures 3.1. However there is only one sense given for *military force* in WordNet: *a unit that is part of some military service*. Therefore, a phrase in one language that is linked to multiple phrases in another language can sometimes denote synonymy (as with *military force*) and other times can be indicative of polysemy (as with *bank*). If we did not take multiple word senses into account then we would end up with situations like the one illustrated in Figure 3.5, where our paraphrasing method would conflate *banque* with *rive* as French paraphrses. This would be as nonsensical as saying that *financial institution* is a paraphrase of *riverbank* in English, which is obviously incorrect.

Since neither the assumption underlying our paraphrasing work, nor the assumption underlying the word sense disambiguation literature holds uniformly, it would be interesting to carry out a large scale study to determine which assumption holds more often. However, we considered such a study to be outside the scope of this thesis. Instead we adopted the pragmatic view that both phenomena occur in parallel corpora, and we adapted our paraphrasing method to take different word senses into account. We attempted to avoid constructing paraphrases when a word has multiple senses by modifying our paraphrase probability. This is described in Section 3.4.2.

3.3.3 Context

One factor that determines whether a particular paraphrase is good or not is the *context* that it is substituted into. For our purposes context means the sentence that a paraphrase is used in. In Section 3.2 we calculate the paraphrase probability without respect to the context that paraphrases will appear in. When we start to use the paraphrases that we have generated, context becomes very important. Frequently we will be substituting a paraphrase in for the original phrase – for example, when paraphrases are used in natural language generation, or in machine translation evaluation. In these cases the sentence that the original phrase occurs in will play a large role in determining whether the substitution is valid. If we ignore the context of the sentence, the resulting substitution might be ungrammatical, and might fail to preserve the meaning of the original phrase.

For example, while *forces* seems to be a valid paraphrase of *military force* out of context, if we were substitute the former for the later in a sentence, the resulting sentence would be ungrammatical because of agreement errors:[3]

> The invading **military force** is attacking civilians as well as soldiers.
> ∗The invading **forces** is attacking civilians as well as soldiers.

Because the paraphrase probability that we define in Equation 3.2 does not take the surrounding words into account it is unable to distinguish that a singular noun would be better in this context.

A related problem arises when generating paraphrases for languages which have grammatical gender. We frequently extract morphological variations as potential paraphrases. For instance, the Spanish adjective *directa* is paraphrased as *directamente*, *directo*, *directos*, and *directas*. None of these morphological variants could be substituted in place of the singular feminine adjective *directa*, since they are an adverb, a singular masculine adjective, a plural masculine adjective, and a plural feminine noun, respectively. The difference in their agreement would result in an ungrammatical Spanish sentence:

> Creo que una acción **directa** es la mejor vacuna contra futuras dictaduras.
> ∗Creo que una acción **directo** es la mejor vacuna contra futuras dictaduras.

It would be better instead to choose a paraphrase, such as *inmediata*, which would agree with the surrounding words.

[3] In these examples we denote grammatically ill-formed sentences with a star, and disfluent or semantically implausible sentences with a question mark. This practice is widely used in linguistics literature.

The difficulty introduced by substituting a paraphrase into a new context is by no means limited to our paraphrasing technique. In order to be complete any paraphrasing technique would need to account for what contexts its paraphrases can be substituted into. However, this issue has been largely neglected. For instance, while Barzilay and McKeown's example paraphrases given in Figure 2.1 are perfectly valid in the context of the pair of sentences that they extract the paraphrases from, they are invalid in many other contexts. While *console* can be valid substitution for *comfort* when it is a verb, it is an inappropriate substitution when *comfort* is used as a noun:

George Bush said Democrats provide **comfort** to our enemies.
∗George Bush said Democrats provide **console** to our enemies.

Some factors which determine whether a particular substitution is valid are subtler than part of speech or agreement. For instance, while *burst into tears* would seem like a valid replacement for *cried* in any context, it is not. When *cried* participates in a verb-particle construction with *out* suddenly *burst into tears* sounds very disfluent:

She **cried** out in pain.
∗She **burst into tears** out in pain.

Because *cried out* is a phrasal verb it is impossible to replace only part of it, since the meaning of *cried* is distinct from *cried out*.

The problem of multiple word senses also comes into play when determining whether a substitution is valid. For instance, if we have learned that *shores* is a paraphrase of *bank*, it is critical to recognize when it may be substituted in for *bank*. It is fine in:

Early civilization flourished on the **bank** of the Indus river.
Early civilization flourished on the **shores** of the Indus river.

But it would be inappropriate in:

The only source of income for the **bank** is interest on its own capital.
∗The only source of income for the **shores** is interest on its own capital.

Thus the meaning of a word as it appears in a particular context also determines whether a particular paraphrase substitution is valid. This can be further illustrated by showing how the words *idea* and *thought* are perfectly interchangeable in one sentence:

She always had a brilliant **idea** at the last minute.
She always had a brilliant **thought** at the last minute.

But when we change that sentence by a single word, the substitution seems marked:

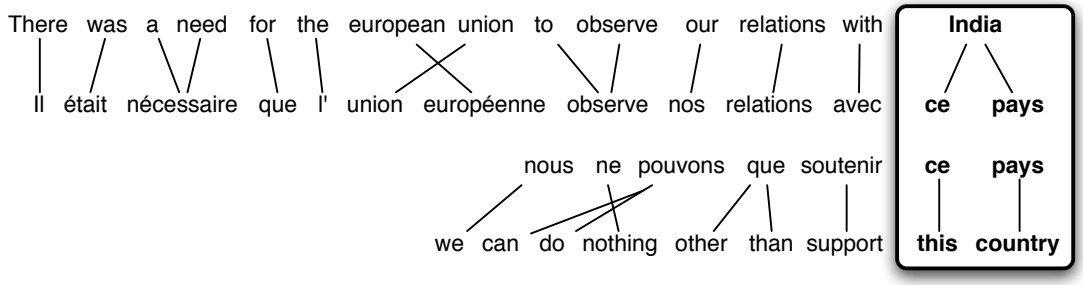


Figure 3.6: Hypernyms can be identified as paraphrases due to differences in how entities are referred to in the discourse.

She always got a brilliant **idea** at the last minute.
?She always got a brilliant **thought** at the last minute.

The substitution is strange in the slightly altered sentence due to the fact that *get an idea* is sounds fine, whereas *get a thought* sounds strange. The lexical selection of *get* doesn't hold for *have*.

Section 3.4.3 discusses how a language model might be used in addition to the paraphrase probability to try to overcome some of the lexical selection and agreement errors that arise when substituting a paraphrase into a new context. It further describes how we could constrain paraphrases based on the grammatical category of the original phrase.

3.3.4 Discourse

In addition to local context, sometimes more global context can also affect paraphrase quality. Discourse context can play a role both in terms of what paraphrases get extracted from the training data, and in terms of their validity when they are being used. Figure 3.6 illustrates how the hypernym *this country* can be extracted as a paraphrase for *India* since the French sentence makes references to the entity in different ways than the English.[4] Using a hypernym might be a valid way of paraphrasing its hyponym in some situations, but larger discourse constraints come into play. For instance, *India* should not be replaced with *this country* if it were the first or only instance of *India*.

In addition hyponym / hypernym paraphrases, differences in how entities are referred across two languages can lead to other sorts of paraphrases. For instance, dis-

[4]While the French phrase *ce pays* aligns with hypernyms of *India* such as *this country*, *that country*, and *the country*, it also aligns with other country names. In our corpus it aligned once each with *Afghanistan*, *Azerbijan*, *Barbados*, *Belarus*, *Burma*, *Moldova*, *Russia*, and *Turkey*. These would therefore be treated as potential paraphrases of *India* under our framework, albeit with very low probability.

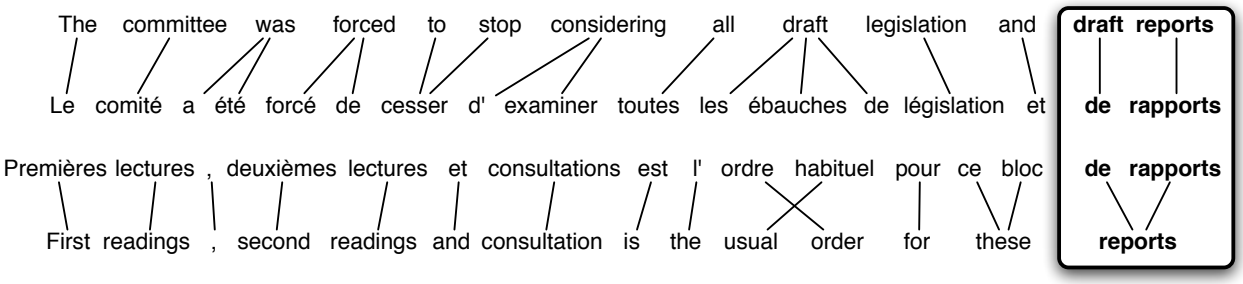


Figure 3.7: Syntactic factors such as conjunction reduction can lead to shortened paraphrases.

course factors such as reduced reference can lead to shortened paraphrases. This can lead us to result in paraphrases groups such as *U.S. President Bill Clinton*, *the U.S. president*, *President Clinton*, and *Clinton*. Variation in paraphrase length can also arise from syntactic factors such as conjunction reduction. Figure 3.7 illustrates how adjective modification can differ between two languages. In the illustration the adjective *draft* is repeated for the coordinated nouns in English, but the corresponding French *ébauches* is not repeated. This difference leads to *reports* being extracted as a potential paraphrase of *draft reports*.

Paraphrasing discourse connectives also presents potential problems. Many connectives, such as *because*, are sometimes explicit and sometimes implicit. Our technique extracts *because otherwise* as a potential paraphrase of *otherwise*, but has no mechanism for determining when the connective should be used (when it occurs as a clause-initial adverbial). The problem of when such connectives should be realized also holds for the intensifiers *actually* and *in fact* (which are extracted as paraphrases of each other, and of *because*). These can sometimes be implicit, or explicit, or doubly realized (*because in fact*). We acknowledge the difficulty in paraphrasing such items, but leave it as an avenue for future research.

While it would be possible to refine our paraphrase probability to utilize discourse constraints, this is not something that we undertook. Very few of the paraphrases exhibited these problems in our experiments (which are presented in the next chapter). Paraphrases such as hyponyms generally had a low probability (due to the fact that they occurred less frequently), and thus were generally not selected as the best paraphrase, and therefore were not used. We therefore focused instead on refining our model to address more common problems.

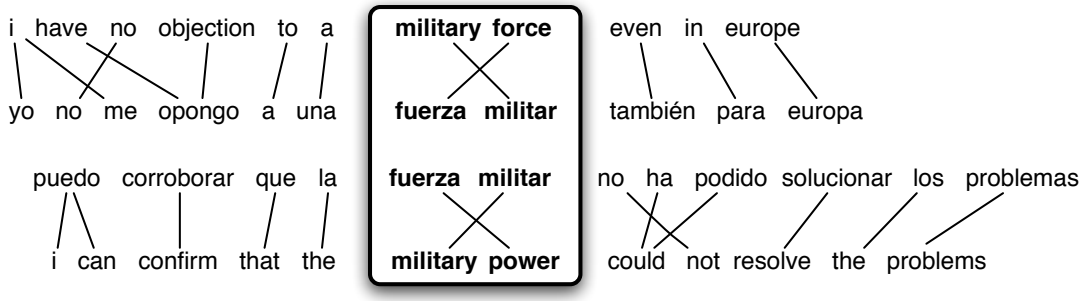


Figure 3.8: Other languages can also be used to extract paraphrases

3.4 Refined paraphrase probability calculation

In this section we introduce refinements to the paraphrase probability in light of the various factors that can affect paraphrase quality. Specifically, we look at different ways of modifying the calculation of the paraphrase probability in order to:

- Incorporate multiple parallel corpora to reduce problems associated with systematic misalignments and sparse counts
- Constrain word sense in an effort to account for the fact that sometimes alignments are indicative of polysemy rather than synonymy
- Add constraints to what constitutes a valid paraphrase in terms of syntactic category, agreement, etc.
- Rank potential paraphrases using a language model probability which is sensitive to the surrounding words

Each of these refinements changes the way that paraphrases are ranked in the hope that they will allow us to better select paraphrases from among the many candidates which are extracted from parallel corpora.

3.4.1 Multiple parallel corpora

As discussed in Section 3.3.1, systematic misalignments in a parallel corpus may cause problems for paraphrasing. However, there is nothing that limits us to using a single parallel corpus for the task. For example, in addition to using a German-English parallel corpus we might use a Spanish-English corpus to discover additional paraphrases of *military force*, as illustrated in Figure 3.8. If we redefine the paraphrase probability

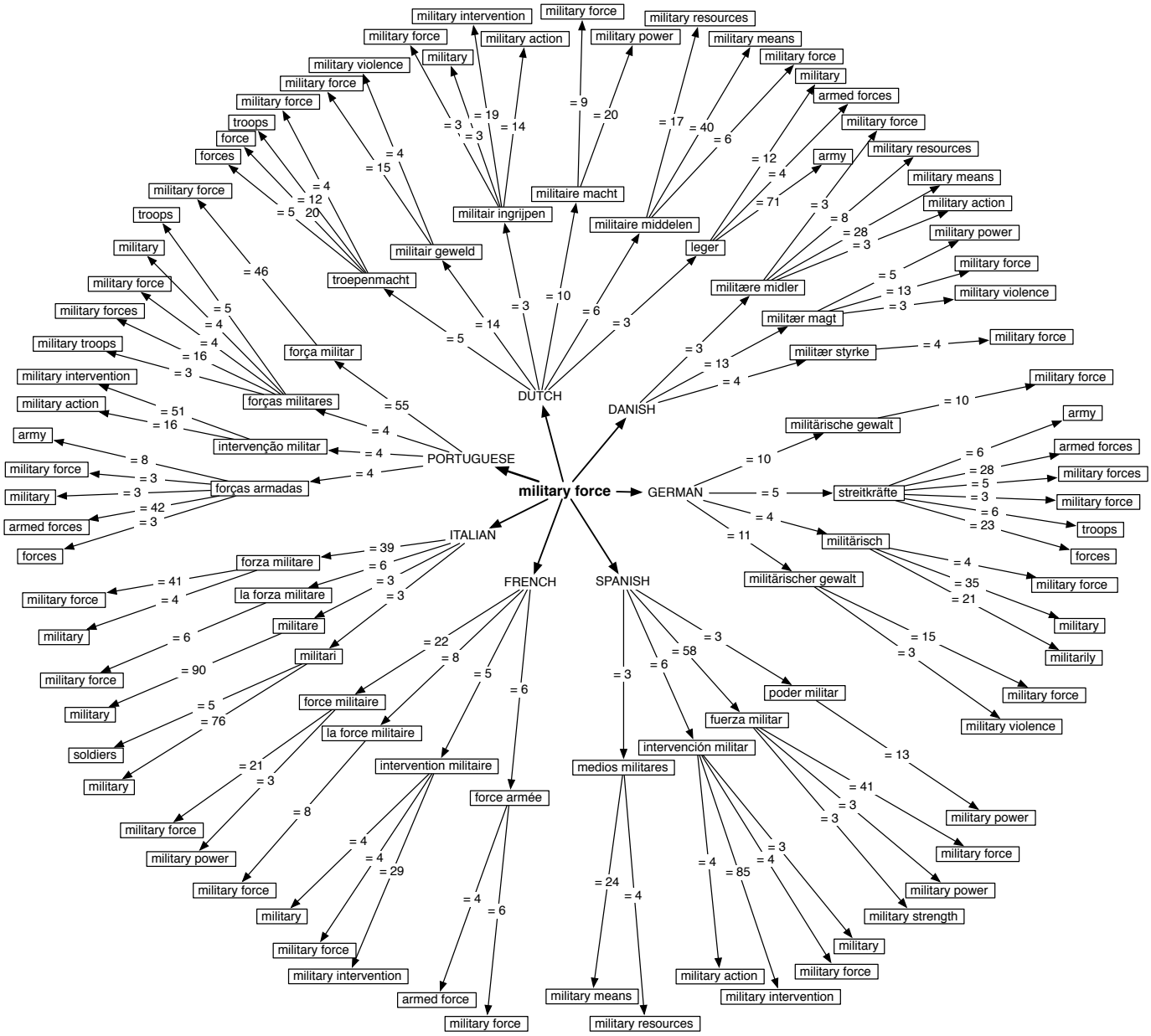


Figure 3.9: Parallel corpora for multiple languages can be used to generate paraphrases. Here counts are collected from Danish-English, Dutch-English, French-English, German-English, Portuguese English and Spanish-English parallel corpora.

so that it collected counts over a set of parallel corpora, C, then we need to normalize in order to have a proper probability distribution for the paraphrase probability. The most straightforward way of normalizing is to divide by the number of parallel corpora that we are using:

$$p(e_2|e_1) = \frac{\sum_{c \in C} \sum_{f \text{ in } c} p(f|e_1) p(e_2|f)}{|C|} \tag{3.5}$$

where $|C|$ is the cardinality of C. This normalization could be altered to include variable weights λ_c for each of the corpora:

$$p(e_2|e_1) = \frac{\sum_{c \in C} \lambda_c \sum_{f \text{ in } c} p(f|e_1) p(e_2|f)}{\sum_{c \in C} \lambda_c} \tag{3.6}$$

Weighting the contribution of each of the parallel corpora would allow us to place more emphasis on larger parallel corpora, or on parallel corpora which are in-domain or are known to have good word alignments.

The use of multiple parallel corpora lets us lessen the risk of retrieving bad paraphrases because of systematic misalignments, and also allows us access to a larger amount of training data. We can use as many parallel corpora as we have available for the language of interest. In some cases this can mean a significant increase in training data. Figure 3.9 shows how we can collect counts for English paraphrases using a number of other European languages.

3.4.2 Constraints on word sense

There are two places where word senses can interfere with the correct extraction of paraphrases: when the phrase to be paraphrased is polysemous, and when one or more of the foreign phrases that it aligns to is polysemous. In order to deal with these potential problems we can treat each word sense as a distinct item. So rather than collecting counts over all instances of a polysemous word such as *bank*, we only collect counts for those instances which have the same sense as the instance of the phrase that we are paraphrasing. This has the effect of partitioning the space of alignments, as illustrated in Figure 3.11. If we want to paraphrase an instance of *bank* which corresponds to the *riverbank* sense (labeled $bank_2$), then we can collect counts over our parallel corpus for instances of $bank_2$. None of those instances would be aligned to the French word *banque*, and so we would never get *banking* as a potential paraphrase for $bank_2$. Similarly, if we treat the different word senses of the foreign words as distinct items we can further narrow the range of potential paraphrases. In Figure 3.11

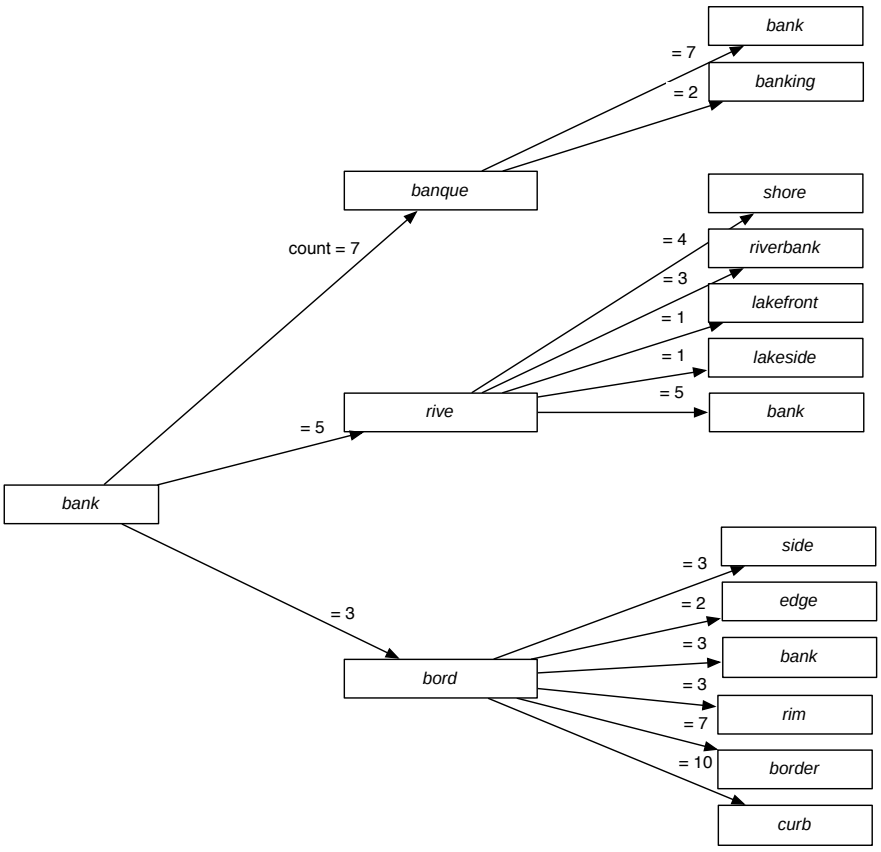


Figure 3.10: Counts for the alignments for the word *bank* if we do not partition the space by sense

note that $bank_2$ is only ever aligned to $bord_1$, which corresponds to the *water's edge* sense, and never to $bord_2$, which corresponds to a more general sense of *delineation*.

We can calculate the paraphrase probabilities for the word *bank* if we did not treat each of its word senses as a distinct element using the counts given in Figure 3.10. Based on these counts we get the following values for $p(f|e_1)$:

$$\begin{aligned} p(banque \mid bank) &= 0.466 \\ p(rive \mid bank) &= 0.333 \\ p(bord \mid bank) &= 0.2 \end{aligned}$$

And the following values for $p(e_2|f)$:

$$
\begin{aligned}
p(bank \mid banque) &= 0.777 \\
p(banking \mid banque) &= 0.222 \\
p(shore \mid rive) &= 0.286 \\
p(riverbank \mid rive) &= 0.214 \\
p(lakefront \mid rive) &= 0.071 \\
p(lakeside \mid rive) &= 0.071 \\
p(bank \mid rive) &= 0.357 \\
p(side \mid bord) &= 0.107 \\
p(edge \mid bord) &= 0.071 \\
p(bank \mid bord) &= 0.107 \\
p(rim \mid bord) &= 0.107 \\
p(border \mid bord) &= 0.25 \\
p(curb \mid bord) &= 0.357
\end{aligned}
$$

These allow us to calculate the paraphrase probabilities for *bank* as follows:

$$
\begin{aligned}
p(bank \mid bank) &= 0.503 \\
p(banking \mid bank) &= 0.104 \\
p(shore \mid bank) &= 0.093 \\
p(riverbank \mid bank) &= 0.071 \\
p(lakefront \mid bank) &= 0.024
\end{aligned}
$$

$$
\begin{aligned}
p(lakeside \mid bank) &= 0.024 \\
p(side \mid bank) &= 0.021 \\
p(edge \mid bank) &= 0.014 \\
p(rim \mid bank) &= 0.021 \\
p(border \mid bank) &= 0.05 \\
p(curb \mid bank) &= 0.071
\end{aligned}
$$

The phrase e_2 which maximizes the probability and is not equal to e_1 is *banking*. When we ignore word sense we can make contextual mistakes in paraphrasing by generating *banking* as a paraphrase of *bank* when it has a different sense. Notice that in this case the word *curb* is an equally likely paraphrase of bank as *riverbank*.

If we treat each word sense as a distinct item then we can calculate the following probabilities for the second sense of *bank*. The $p(f|e_1)$ values work out as:

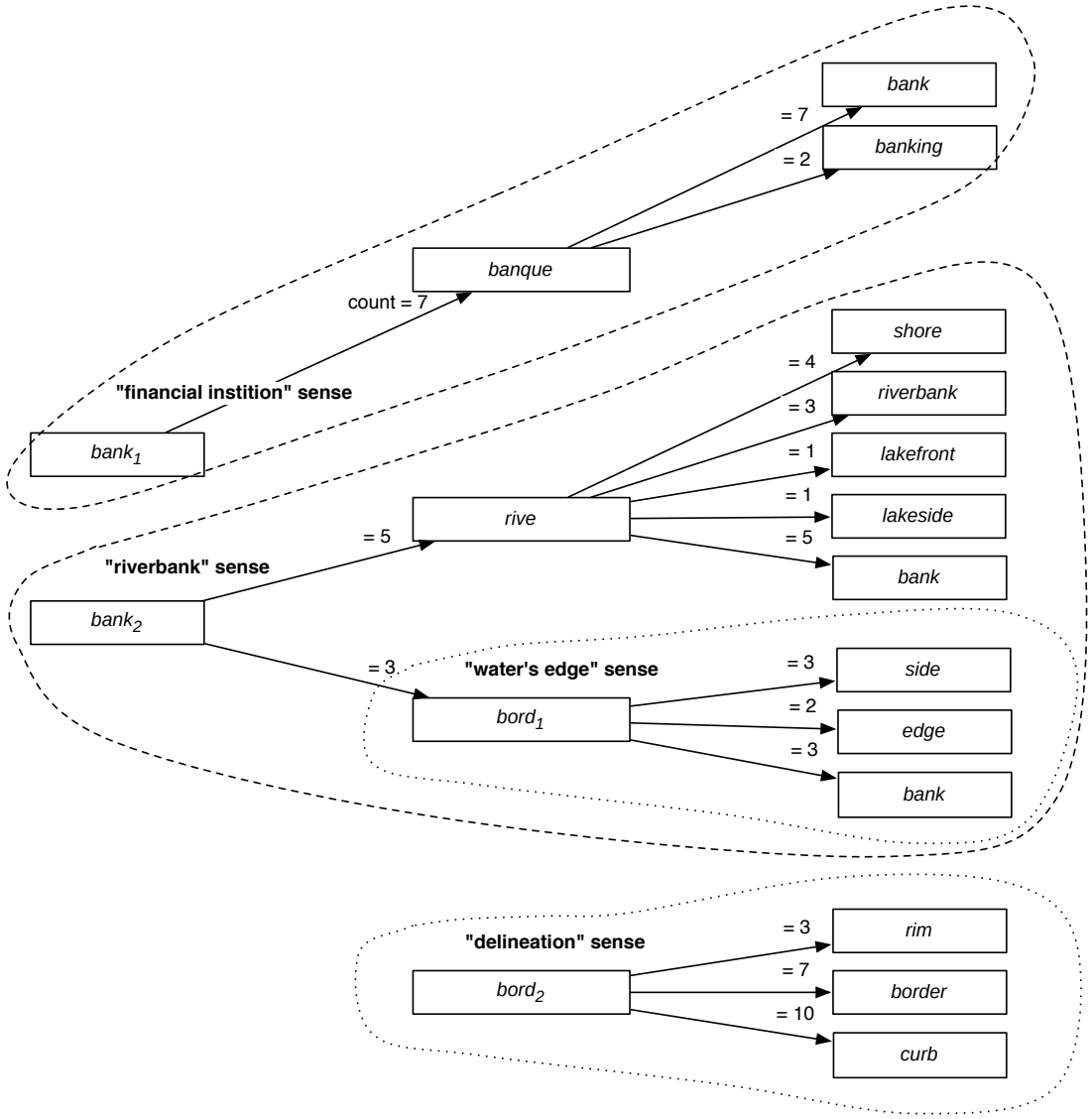


Figure 3.11: If we treat words with different senses as different items then their alignments are partitioned. This allows us to more draw more appropriate paraphrases, if we are given the word sense of the original phrase.

$$\begin{aligned} p(banque \mid bank_2) &= 0 \\ p(rive \mid bank_2) &= 0.625 \\ p(bord_1 \mid bank_2) &= 0.375 \\ p(bord_2 \mid bank_2) &= 0 \end{aligned}$$

The $p(e_2|f)$ that change are:

$$\begin{aligned} p(side \mid bord_1) &= 0.375 \\ p(edge \mid bord_1) &= 0.25 \\ p(bank \mid bord_1) &= 0.375 \end{aligned}$$

The revised paraphrase probabilities when word sense is taken into account are:

$$\begin{aligned} p(bank \mid bank_2) &= 0.364 \\ p(banking \mid bank_2) &= 0 \\ p(shore \mid bank_2) &= 0.179 \\ p(riverbank \mid bank_2) &= 0.134 \\ p(lakefront \mid bank_2) &= 0.045 \\ p(lakeside \mid bank_2) &= 0.045 \\ p(side \mid bank_2) &= 0.1406 \\ p(edge \mid bank_2) &= 0.094 \\ p(rim \mid bank_2) &= 0 \\ p(border \mid bank_2) &= 0 \\ p(curb \mid bank_2) &= 0 \end{aligned}$$

When we account for word sense we get *shore* rather than *banking* as the most likely paraphrase for the *river* sense of *bank*. The treatment of foreign word senses for *bord* also eliminates the spurious paraphrases *rim*, *border* and *curb* from consideration and thus more accurately distributes the probability mass.

In the experiments presented in Section 4.3.4, we extend these "word sense" controls to phrases. We show that this helps us select among the paraphrases for polysemous phrases like *at work*, which can mean either *at the workplace* or *functioning* depending on the context.

3.4.3 Taking context into account

Note that the paraphrase probability defined in Equation 3.1 returns the single best paraphrase, $\hat{e_2}$, irrespective of the context in which e_1 appears. Since the best paraphrase may vary depending on information about the sentence that e_1 appears in, we

can extend the paraphrase probability to include that sentence. In the experiments described in Chapter 4 we explore one way of using the contextual information provided by the sentence: we use a simple language model probability, which additionally ranks e_2 based on the probability of the sentence formed by substituting e_2 for e_1 in the sentence.

Ranking candidate paraphrases with a language model probability in addition to our paraphrase probability allows us to distinguish between things that are strongly lexicalized. For instance, if we were deciding between using *strong* or *powerful* the context could dictate which is better. In one context *powerful* might be preferable to *strong*:

> ? He decided that a **strong** computer is what he needed.
> He decided that a **powerful** computer is what he needed.

And in another *strong* might be preferable to *powerful*:

> He decided that a **strong** drug is what he needed.
> ? He decided that a **powerful** drug is what he needed.

A simple trigram language model is sufficient to tell us that *a strong computer* is a less probable phrase in English than *a powerful computer* is, and that *a strong drug* is a more probable phrase than *a powerful drug*. A trigram language model might also facilitate local agreement problems, such as the ungrammatical phrase *the forces is* discussed in Section 3.3.3.

Having contextual information available also lets us take other factors into account like the syntactic type of the original phrase. We may wish to permit only paraphrases that are the same syntactic type as the original phrase, which we could do by extending the translation model probabilities to count only phrase occurrences of that type.

$$p(e_2|e_1,type(e_1)) = \sum_f p(f|e_1,type(e_1))p(e_2|f,type(e_1)) \tag{3.7}$$

We can use this type information to refine the the calculation of the translation model probability given in Equation 3.3. For example, when $type(e_1)$ = NP, we could calculate it as:

$$p(f|e_1,type = \text{NP}) = \frac{count_{e_1=\text{NP}}(f,e_1)}{count_{e_1=\text{NP}}(e_1)} \tag{3.8}$$

and

$$p(e_2|f,type = \text{NP}) = \frac{count_{e_2=\text{NP}}(e_2,f)}{count(f)} \tag{3.9}$$

Now we collect counts over a smaller set of events: instead of gathering counts of all instances of e_1 we now only count those instances which have the specified syntactic type, and further only gather counts when e_2 is of the same syntactic type.

3.5 Discussion

In this chapter we developed a novel paraphrasing technique that uses parallel corpora, a data source that has not hitherto been used for paraphrasing. By drawing on techniques from phrase-based statistical machine translation, we are able to align phrases with their paraphrases by pivoting through foreign language phrases. This frees us from the need for pairs of equivalent sentences (which were required by previous data-driven paraphrasing techniques), and allows us to extract a range of possible paraphrases. Because we frequently extract many possible paraphrases of a single phrase we would like a mechanism to rank them. We show how paraphrasing can be treated as a probabilistic mechanism, and define a paraphrase probability which naturally arises naturally from the fact that we are using parallel corpora and alignment techniques from statistical machine translation. We discuss a wide range of factors which can potentially affect the quality of our paraphrases – including alignment quality, word sense and context – and show how the paraphrase probability can be refined to account for each of these.

In the next chapter we delve into the topic of evaluating the quality of our paraphrases. We design a number of experiments which allow us to empirically determine the accuracy of our paraphrases. We examine each of the refinements that we made to the paraphrase probability, and demonstrate their effectiveness in choosing the best paraphrase. These experiments focus on the quality of paraphrases in and of themselves. In Chapter 5 we investigate the usefulness of our paraphrases when they are applied to a particular task. The task that we choose is improving machine translation. This task allows us to showcase the fact that our paraphrasing technique is language-independent in that it can easily be applied to any language for which we have a parallel corpus. Rather than generating English paraphrases, as we have shown in this chapter, we use our technique to generate French and Spanish paraphrases. While the main focus of this thesis is on the generation of lexical and phrasal paraphrases, we address the issue of how parallel corpora may be used to generate more sophisticated structural paraphrases in Chapter 8.

Chapter 4

Paraphrasing Experiments

In this chapter we investigate how well our proposed paraphrasing technique can do, with particular focus on each of the factors which can potentially affect paraphrase quality. Prior to presenting our experiments we first delve into the issue of how to properly evaluate paraphrase quality. Section 4.1 presents our evaluation criteria and methodology. Section 4.2 presents our experimental design and data. Section 4.3 presents our results. Section 4.4 puts these into the context of previous data-driven approaches to paraphrasing.

4.1 Evaluating paraphrase quality

There is no standard methodology for evaluating paraphrase quality directly. As such task-based evaluation is frequently employed, wherein paraphrases are applied to another task which has a more concrete evaluation methodology. The usefulness of paraphrases is demonstrated by showing that they can measurably improve performance on the other task. Duboue and Chu-Carroll (2006) demonstrated the usefulness of their paraphrases by showing that they could potentially improve question answering systems. In Chapters 7.1 and 7.2 we show that our paraphrases improve machine translation quality. In this chapter we examine the quality of the paraphrases themselves, rather than inferring their usefulness indirectly by way of an external task. In order to evaluate the quality of paraphrases directly, we needed to develop a set of criteria to judge whether a paraphrase is correct or not. Though this would seem to be relatively simple, there is no consensus even about how this ought to be done. Barzilay and McKeown (2001) asked judges whether paraphrases had "approximate conceptual equivalence" when they were shown independent of context and when shown substi-

Adequacy

How much of the meaning expressed in the reference translation is also expressed in the hypothesis translation?

5 = All

4 = Most

3 = Much

2 = Little

1 = None

Fluency

How do you judge the fluency of this translation?

5 = Flawless English

4 = Good English

3 = Non-native English

2 = Disfluent English

1 = Incomprehensible

Figure 4.1: In machine translation evaluation the following scales are used by judges to assign adequacy and fluency scores to each translation

tuted into the original context that they were extracted from. Pang et al. (2003) asked judges to make a distinction as to whether a paraphrase is correct, partially correct, or incorrect in the context of the sentence group that it was generated from. Ibrahim et al. (2003) evaluated their paraphrase system by asking judges whether the paraphrases were "roughly interchangeable given the genre."

4.1.1 Meaning and grammaticality

Because we generate phrasal paraphrases we believe that the most natural way of assessing their correctness is through substitution, wherein we replace an occurrence of the original phrase with the paraphrase. In our evaluation we asked judges whether the paraphrase retains the same meaning as the phrase it replaced, and whether the resulting sentence remains grammatical. The reason that we ask about both meaning and grammaticality is the fact that what constitutes a "good" paraphrase is largely dictated by the intended application. For applications like information retrieval it might not matter if some paraphrases are syntactically incorrect, so long as most of them are

semantically correct. Other applications, like natural language generation, might require that the paraphrases be both syntactically and semantically correct. We evaluated both dimensions and reported scores for each so that our results would be as widely applicable as possible.

Rather than write our own instructions for how to manually evaluate the meaning and grammaticality, we used existing guidelines for evaluating *adequacy* and *fluency*. The Linguistic Data Consortium developed two five point scales for evaluating machine translation quality (LDC, 2005). These well-established guidelines have been used in the annual machine translation evaluation workshop which is run by the National Institute of Standards in Technology in the United States (Przybocki, 2004; Lee and Przybocki, 2005). Figure 4.1 gives the five point scales and the questions that are presented to judges when they evaluate translation quality. We adapted these questions for paraphrase evaluation:

- How much of the meaning of the original phrase is expressed in the paraphrase?
- How do you judge the fluency of the sentence?

Paraphrases were considered it to be 'correct' when they were rated at a 3 or higher on each of the scales. Therefore, a paraphrase was accurate if it contained all, most, or much of the meaning of the original phrase and if the sentence was judged to be flawless English, good English or non-native English. A paraphrase was inaccurate if it contained little or none of the meaning of the original phrase, or if the sentence that it was in was judged to be disfluent or incomprehensible. In Section 4.3 we report the 'accuracy' of our paraphrases under a number of different conditions. We define 'accuracy' to be the average number of paraphrases that were judged to be 'correct'. We also report the average number of times that our paraphrases were judged to have the correct meaning under each scenario. Correct meaning is defined as being rated 3 or higher on the *adequacy* scale, and it ignores *fluency*.

4.1.2 The importance of multiple contexts

One further refinement that we made in our evaluation methodology was to judge paraphrases when they were substituted into multiple different contexts. As discussed in Section 3.3.3, context can play a major role in determining whether a particular paraphrase is valid. This is something that has been largely ignored by past research. For

You should investigate whether criminal activity is **at work** here, and whether it is linked to trafficking in forced prostitution.
The most important issue is developing mature interpersonal relationships in the family, **at work**, and in society.
The European Union was traumatised by its powerlessness in the face of the violent disintegration **at work** in the Balkans.
Smart cards could be the best way to regulate the hours during which truck drivers are on the road and **at work**.
That means that we need to pursue with vigour the general framework on information and consultation **at work**.
Despite considerable progress for women, there are still considerable differences, especially discrimination **at work** and different wages for the same job.
A second directive on discrimination **at work** is to be examined shortly.

Table 4.1: To address the fact that a paraphrase's quality depends on the context that it is used, we compiled several instances of each phrase that we paraphrase. Here are the seven instances of the phrase *at work* which we paraphrased and then evaluated.

instance, Barzilay and McKeown solicited judgments about their paraphrases by substituting them into a single context. Worse yet, that context was the original sentence that they were extracted from. For example, Figure 2.1 shows how their system learned that *comfort* is a paraphrase of *console*. When evaluating the paraphrase they showed it substituted into same sentence:

> Emma cried and he tried to **console** her, adorning his words with puns.
> Emma cried and he tried to **comfort** her, adorning his words with puns.

Because of the way that Barzilay and McKeown's extraction algorithm works, substituting paraphrases into the original context is likely to result in a falsely high performance estimate. It would be more accurate to choose multiple instances of the original phrase randomly and substitute paraphrases in for those occurrences.

In order to be more rigorous in our evaluation methodology we substituted our paraphrases into multiple sentences. Table 4.1 shows seven sentences containing the phrase *at work*, which we paraphrased and replaced with our paraphrases. Notice that by sampling a number of sentences we manage to extract different senses of the phrase – some of the sentences represent the *in the workplace* sense, and some represent the sense of *something taking place*. Because of this different paraphrases will be valid in

Original sentence: You should investigate whether criminal activity is **at work** here, and whether it is linked to trafficking in forced prostitution.		
Adequacy	**Fluency**	**Paraphrased sentence**
2	5	You should investigate whether criminal activity is **at stake** here, and whether it is linked to trafficking in forced prostitution.
5	**4**	You should investigate whether criminal activity is **working** here, and whether it is linked to trafficking in forced prostitution.
1	2	You should investigate whether criminal activity is **workplace** here, and whether it is linked to trafficking in forced prostitution.
2	5	You should investigate whether criminal activity is **to work** here, and whether it is linked to trafficking in forced prostitution.

Original sentence: The most important issue is developing mature interpersonal relationships in the family, **at work**, and in society.		
Adequacy	**Fluency**	**Paraphrased sentence**
5	**3**	The most important issue is developing mature interpersonal relationships in the family, **the work**, and in society.
1	1	The most important issue is developing mature interpersonal relationships in the family, **at**, and in society.
5	**4**	The most important issue is developing mature interpersonal relationships in the family, **employment**, and in society.
5	**3**	The most important issue is developing mature interpersonal relationships in the family, **work**, and in society.
3	2	The most important issue is developing mature interpersonal relationships in the family, **working**, and in society.
5	**5**	The most important issue is developing mature interpersonal relationships in the family, **at the workplace**, and in society.
5	**3**	The most important issue is developing mature interpersonal relationships in the family, **workplace**, and in society.

Table 4.2: The scores assigned to various paraphrases of the phrase **a**t work when they are substituted into two different contexts. Bold scores indicate items that were judged to be ‘correct’.

Original sentence: The European Union was traumatised by its powerlessness in the face of the violent disintegration **at work** in the Balkans.		
Adequacy	**Fluency**	**Paraphrased sentence**
2	2	The European Union was traumatised by its powerlessness in the face of the violent disintegration **the work** in the Balkans.
2	1	The European Union was traumatised by its powerlessness in the face of the violent disintegration **at** in the Balkans.
1	5	The European Union was traumatised by its powerlessness in the face of the violent disintegration **at stake** in the Balkans.
5	**5**	The European Union was traumatised by its powerlessness in the face of the violent disintegration **working** in the Balkans.
1	1	The European Union was traumatised by its powerlessness in the face of the violent disintegration **workplace** in the Balkans.
3	**5**	The European Union was traumatised by its powerlessness in the face of the violent disintegration **held** in the Balkans.
5	**3**	The European Union was traumatised by its powerlessness in the face of the violent disintegration **took place** in the Balkans.

Original sentence: Smart cards could be the best way to regulate the hours during which truck drivers are on the road and **at work**.		
Adequacy	**Fluency**	**Paraphrased sentence**
3	2	Smart cards could be the best way to regulate the hours during which truck drivers are on the road and **the work**.
2	2	Smart cards could be the best way to regulate the hours during which truck drivers are on the road and **employment**.
3	2	Smart cards could be the best way to regulate the hours during which truck drivers are on the road and **work**.
5	**5**	Smart cards could be the best way to regulate the hours during which truck drivers are on the road and **working**.
3	**3**	Smart cards could be the best way to regulate the hours during which truck drivers are on the road and **workplace**.

Table 4.3: The scores assigned to various paraphrases of the phrase **a**t work when they are substituted into two more contexts. Bold scores indicate items that were judged to be ‘correct’.

the different contexts. Tables 4.2 and 4.3 show what adequacy and fluency scores were assigned by one of our judges for paraphrases of *at work*. The paraphrases given in the tables were generated for our different experimental conditions (which are explained in Section 4.2).

4.1.3 Summary and limitations

Our evaluation methodology can be summarized by the following key points:

- We evaluated paraphrase quality by replacing phrases with their paraphrases, soliciting judgments about the resulting sentences.
- We evaluated both meaning and grammaticality so that our results would be as generally applicable as possible. We used established guidelines for evaluating adequacy and fluency, rather than inventing ad hoc guidelines ourselves.
- We choose multiple occurrences of the original phrase and substituted each paraphrase into more than one sentences. We choose 2–10 sentences that the original phrase occurred, with an average of 6.3 sentences per phrase.
- We had two native English speakers produce judgments of each paraphrase, and measured their agreement on the task using the Kappa statistic. The inter-annotator agreement for these judgements was $\kappa = 0.605$, which is conventionally interpreted as "good" agreement.

We acknowledge that our evaluation methodology is limited in two ways: Firstly, the adequacy scale might be slightly inappropriate for judging the meaning of our paraphrases. The adequacy scale only allows for the possibility that a paraphrased sentence contains less information than in the original sentence, but in some circumstances paraphrases may add *more information* (for instance, if *force* were paraphrased as *military force*). It would be worthwhile to have a category that reflected whether information was added, and possibly a separate judgment about whether it was acceptable given the context.

Secondly, testing paraphrases through substitution might be limiting, because a change in one part of the sentence may require a change in another part of the sentence in order to be correct. While our method does not make such transformations, it has bearing on techniques which produce sentential paraphrases. Judging sentential paraphrases rather than lexical and phrasal paraphrases is more complicated since they

potentially change different parts and differing amounts of a sentence. This would add another dimension to the evaluation process when comparing different two sentential paraphrases. For the purpose of evaluating paraphrases of the level of granularity that our technique produces, the substitution test is sufficient.

4.2 Experimental design

We designed a set of experiments to test our paraphrasing method. We examined our technique's performance in relationship to the various factors discussed in Section 3.3. Specifically, we investigated the effect of word alignment quality on paraphrase quality, the usefulness of extracting paraphrases from multiple parallel corpora, the extent to which controlling word sense can improve quality, and whether language models can be used to select fluent paraphrases. Section 4.2.1 details our experimental conditions. Section 4.2.2 describes the data sets that we used to train our paraphrase models, and how we prepared the training data. Section 4.2.3 lists the phrases that we paraphrased, and describes the sentences that we substituted our paraphrases into when evaluating them. The results of our experiments are presented in Section 4.3.

4.2.1 Experimental conditions

We had a total of eight experimental conditions. Each used a different mechanism to select the best paraphrase from the candidate paraphrases extracted from a parallel corpus. The conditions were:

1. **The simple paraphrase probability, as given in Equation 3.1.** In this case we choose the paraphrase $\hat{e_2}$ such that

$$\hat{e_2} = \arg\max_{e_2 \neq e_1} \sum_f p(f|e_1)p(e_2|f) \tag{4.1}$$

   For this condition we calculated the translation model probabilities $p(f|e_1)$ and $p(e_2|f)$ using a German-English parallel corpus, with the word alignments calculated automatically using standard techniques from statistical machine translation.

2. **The simple paraphrase probability when calculated with manual word alignments.** We repeated the first condition but with an idealized set of word alignments. For a 50,000 sentence portion of the German-English parallel corpus

we manually aligned each English phrase e_1 with its German counterpart f, and each occurrence of f with its corresponding e_2. Our data preparation is described in the next section. By calculating the paraphrase probability with manual word alignments we were able to assess the extent to which word alignment quality affects paraphrase quality, and we were able to determine how well our method could work *in principle* if we were not limited by the errors in automatic alignment techniques.

3. **The paraphrase probability calculated over multiple parallel corpora, as given in Equation 3.5.** In this case we choose the paraphrase $\hat{e_2}$ such that

$$\hat{e_2} = \arg\max_{e_2 \neq e_1} \sum_{c \in C} \sum_{f \text{ in } c} p(f|e_1)p(e_2|f) \tag{4.2}$$

Where C contained four parallel corpora: the German-English corpus used in the first experimental condition plus a French-English corpus, an Italian-English corpus and a Spanish-English corpus. These are described in Section 4.2.2. Under this experimental condition we again used automatic word alignments, since we did not have the resources to manually align four parallel corpora.

4. **The paraphrase probability when controlled for word sense.** As discussed in Sections 3.3.2 and 3.4.2 we sometimes extract false paraphrases when the original phrase e_1 or the foreign phrase f is polysemous. Under this experimental condition we controlled for the word sense of e_1 by specifying which sense it took in each evaluation sentence.[1] Rather than performing real word sense disambiguation, we instead used Diab and Resnik (2002)'s assumption that an aligned foreign language phrase can be indicative of the word sense of an English phrase. Since our test sentence are drawn from a parallel corpus (as described in Section 4.2.3), we know which foreign phrase f is aligned with each instance of the phrase e_1 that we evaluated. We use the foreign phrase as an indicator of the word sense. Rather than summing our f like we do in Equation 4.1, we use the single foreign language phrase.

$$\hat{e_2} = \arg\max_{e_2 \neq e_1} p(f|e_1)p(e_2|f) \tag{4.3}$$

By limiting ourselves to paraphrases which arise through the particular f, we control for phrases which have that sense. This is equivalent to knowing that

[1] Note that we treat *phrases* as potentially having multiple senses, and treat the problem of disambiguating them in the same way that *word* sense is treated.

a particular instance of the word *bank* which we were evaluating is aligned to *rive*. Thus, we would calculate the probability of $p(e_2|bank)$ for only those paraphrases e_2 which were aligned to *rive*. Using the counts from Figure 3.10 the $\hat{e_2}$ would be *shore* rather than *banking*, which would is the best paraphrase of *bank* in the first condition.

This is not a perfect mechanism for testing word sense, since it ignores the possibility of polysemous foreign phrases f and since real word sense disambiguation systems might make different predictions about what the word senses of our phrases e_1 are. That being said, it is sufficient to give us an idea of the role of word sense in paraphrase quality. In the word sense condition we used automatic word alignments and the single German-English parallel corpus.

5–8. **We repeated each of the four above cases using a combination of the paraphrase probability and a language model probability**, rather than the paraphrase probability alone. In conditions 1–3 above the paraphrase probability ignores context and always selects the same paraphrase $\hat{e_2}$ regardless of what sentence the phrase e_1 occurs in. In condition 4 the context of the sentence plays a role in determining what the word sense of e_1 is. In conditions 5–8 we use the words surrounding e_1 to help determine how good each e_2 is when substituted into the test sentence. We use a trigram language model and thus only cared about the two words preceding e_1, which we denote w_{-2} and w_{-1}, and the two words following e_1, which we denote w_{+1} and w_{+2}. We then choose the best paraphrase as follows:

$$\hat{e_2} = \arg\max_{e_2 \neq e_1} p(e_2|e_1)p(w_{-2}\ w_{-1}\ e_2\ w_{+1}\ w_{+2}) \tag{4.4}$$

Where $p(w_{-2}\ w_{-1}\ e_2\ w_{+1}\ w_{+2})$ is calculated using a trigram language model. Note that since e_2 is itself a phrase it can represent multiple words, and therefore there are three or more trigrams. We combine their probabilities by taking their product.

As an example of how this language model is used in this way, consider the paraphrases of *at work* when they were substituted into the test sentence:

> You should investigate whether criminal activity is **at work** here, and whether it is linked to trafficking in forced prostitution.

We would calculate $p($*activity is at stake here ,*$)$, $p($*activity is working here ,*$)$, $p($*activity is workplace here ,*$)$, and so on for each of the potential paraphrases

e_2. Each of these would be calculated using a trigram language model, as

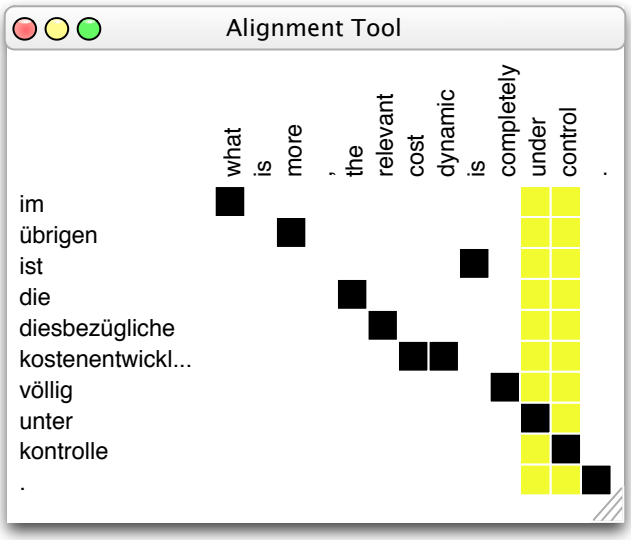
$$\begin{aligned}
p(\textit{activity is at stake here , }) &= p(\textit{at}|\textit{activity is}) * \\
&\quad p(\textit{stake}|\textit{is at}) * \\
&\quad p(\textit{here}|\textit{at stake}) * \\
&\quad p(,|\textit{stake here}) \\
p(\textit{activity is working here , }) &= p(\textit{working}|\textit{activity is}) * \\
&\quad p(\textit{here}|\textit{is working}) * \\
&\quad p(,|\textit{working here}) \\
p(\textit{activity is workplace here , }) &= p(\textit{workplace}|\textit{activity is}) * \\
&\quad p(\textit{here}|\textit{is workplace}) * \\
&\quad p(,|\textit{workplace here})
\end{aligned}$$

These language model probabilities are combined with the paraphrase probability $p(e_2|e_1)$ to rank the candidate paraphrases. In our experiments the language model and paraphrase probabilities were equally weighted. It would also be possible to set different weights for the two, for instance, using a log linear formulation.

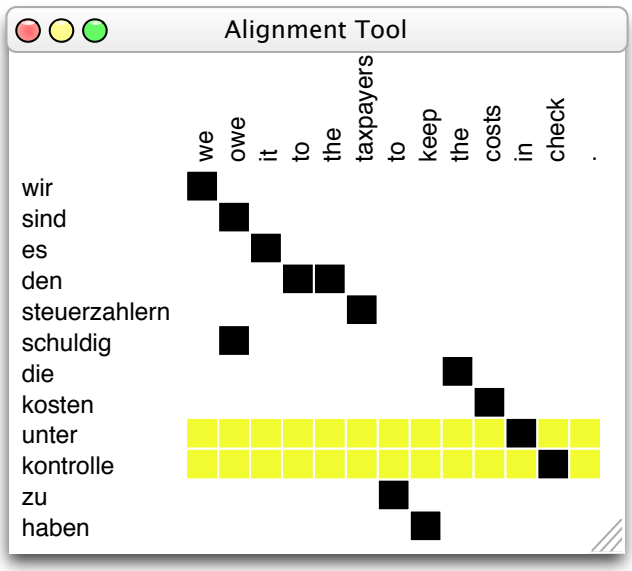
4.2.2 Training data and its preparation

Parallel corpora serve as the training data for our models of paraphrasing. In our experiments we drew our corpora from the Europarl corpus, version 2 (Koehn, 2005). The Europarl corpus consists of parallel texts between eleven different European languages. We used a subset of these in our experiments. We used the German-English parallel corpus to train the paraphrase models which used only a single parallel corpus. For the conditions where we extracted paraphrases from multiple parallel corpora we use three additional corpora from the Europarl set: the French-English corpus, the Italian-English corpus, and the Spanish-English corpus. Table 4.4 gives statistics about the size of each of these parallel corpora. When we combine them all in conditions 3 and 7, we are able to draw paraphrases from nearly 60 million words worth of English text. This is considerably larger than the 16 million words contained in German-English corpus alone, which are used in conditions 1, 4, 5 and 8.

We created automatic word-alignments for each of the parallel corpora using Giza++ (Och and Ney, 2003), which implements the IBM word alignment models (Brown



(a) First, each instance of the English phrase to be paraphrased is aligned to its German counterparts



(b) Next, each occurrence of its German translations is aligned back to other English phrases

Figure 4.2: To test our paraphrasing method under ideal conditions we created a set of manually aligned phrases. This was done by having a bilingual speaker align each instance of an English phrase with its German counterparts, and then align each of the German phrases with other English phrases.

Corpus	Sentence Pairs	English Words	Foreign Words
French-English	688,032	13,808,507	15,599,186
German-English	751,089	16,052,704	15,257,873
Italian-English	682,734	14,784,374	14,900,783
Spanish-English	730,741	15,222,507	15,725,138
Totals:	2,852,596	59,868,092	61,482,980

Table 4.4: The parallel corpora that were used to generate English paraphrases under the multiple parallel corpora experimental condition

et al., 1993). These served as the basis for the phrase extraction heuristics that we use to align an English phrase with its foreign counterparts, and the foreign phrases with the candidate English paraphrases. The phrase extraction techniques are described in Section 2.2.2. Because we wanted to test our method independently of the quality of word alignment, we also developed gold standard word alignments for the set of phrases that we paraphrased. The gold standard word alignments were created manually for a sample of 50,000 sentence pairs. For every instance of our test phrases we had a bilingual individual annotate the corresponding German phrase. This was done by highlighting the original English phrases and having the annotator modify an automatic alignment so that it was correct, as shown in Figure 4.2(a). After all instances of the English phrase had been correctly aligned with their German counterparts, we repated the process aligning every instance of the German phrases with other English phrases, which themselves represented potential paraphrases. The alignment of the German phrases with English paraphrases is shown in Figure 4.2(b). In the 50,000 sentences, each of the 46 original English phrases (described in the next section) could be aligned to between 1–11 German phrases, with the English phrases aligning to an average of 3.9 German phrases. There were a total of 637 instances of the original English phrases, and 3,759 instances of their German counterparts.[2] The annotators changed a total of 4,384 alignment points from the automatic alignments.

The language model that was used in experimental conditions 5–8 was trained on the English portion of the Europarl corpus using the CMU-Cambridge language modeling toolkit (Clarkson and Rosenfeld, 1997).

[2]The annotators skipped alignments for 8 generic German words (*in, zu, nicht, auf, als, an zur,* and *nur*, which were aligned with the original phrases *concentrate on, turn to,* and *other than* in some loose translations). Including instances of these common German phrases would have added an additional 54,000 instances to hand align.

a million, as far as possible, at work, big business, carbon dioxide, central america, close to, concentrate on, crystal clear, do justice to, driving force, first half, for the first time, global warming, great care, green light, hard core, horn of africa, last resort, long ago, long run, military action, military force, moment of truth, new world, noise pollution, not to mention, nuclear power, on average, only too, other than, pick up, president clinton, public transport, quest for, red cross, red tape, socialist party, sooner or later, step up, task force, turn to, under control, vocational training, western sahara, world bank

Table 4.5: The phrases that were selected to paraphrase

4.2.3 Test phrases and sentences

We extracted 46 English phrases to paraphrase (shown in Table 4.5), randomly selected from multiword phrases in WordNet which also occured multiple times in the first 50,000 sentences of our bilingual corpus. We selected phrases from WordNet because we initially intended to use the synonyms that it listed as one measure of paraphrase quality. However, it subsequently became clear that the WordNet synonyms were incomplete, and furthermore, were not necessarily appropriate to our data sets. We therefore did not conduct a comparison to WordNet.

For each of the 46 English phrases we extracted test sentences from the English side of the small German-English parallel corpus. Extracting test sentences from a parallel corpus allowed us to perform word sense experiments using foreign phrases as proxies for different senses. Because the acccuracy of paraphrases can vary depending on context, we substituted each set of candidate paraphrases into 2–10 sentences which contained the original phrase. We selected an average of 6.3 sentences per phrase, for a total of 289 sentences. We created sentences to be evaluated by substituting the paraphrases that were generated by each of the experimental conditions for the original phrase (as illustrated in Tables 4.2 and 4.3). We avoided duplicating evaluation sentences when different experimental conditions selected the same paraphrase. All told we created a total of 1,366 unique sentences through substitution. Each of these was evaluated for its fluency and adequacy by two native speakers of English, as described in Section 4.1.

4.3 Results

We begin by presenting the results of our paraphrasing under ideal conditions. Section 4.3.1 examines the paraphrases that were extracted from a manually word-aligned parallel corpus. The results show that in principle our technique can extract very high quality paraphrases. Because these results employ idealized alignments they may be thought of as an upper bound on the potential performance of our technique (or at least an upper bound when context is ignored). The remaining sections examine more realistic scenarios involving automatic word alignments. Section 4.3.2 contrasts the quality of paraphrases extracted using 'gold standard' alignments with paraphrases extracted from a single automatically aligned parallel corpus. This represents the baseline performance of our method. Sections 4.3.3, 4.3.4, and 4.3.5 attempt to improve upon these results by using multiple parallel corpora, controlling for word sense, and integrating a language model. Summary results are given in Tables 4.7 and 4.8.

4.3.1 Manual alignments

Table 4.6 gives a set of example paraphrases extracted from the gold standard alignments. Even without rigorously evaluating these paraphrases in context it is clear that the method is able to extract high quality paraphrases. All of the extracted items are closely related to phrases that they paraphrase – ranging from items that are generally interchangeable like *nuclear power* with *atomic energy*[3] or the abbreviation of *carbon dioxide* to *CO2*, to items that have more abstract relationships like *green light* and *signal*. In some cases we extract multiple paraphrases which are morphological variants of each other, as with the paraphrases of *step up*: *increase / increased / increasing* and *strengthen / strengthening*. The choice of which of these variants to use depends upon the context in which it is used (as discussed in Section 3.3.3).

We applied the evaluation methodology discussed in Section 4.1 to these paraphrases. For this experimental condition, we substituted the italicized paraphrases in Table 4.6 into a total of 289 different sentences and judged their adequacy and fluency. The italicized paraphrases were assigned the highest probability by Equation 3.2, which chooses a single best paraphrase without regard for context. The paraphrases were judged to be accurate (to have the correct meaning and to remain grammatical) an

[3]Note that even for these seemly perfectly interchangeable items, there are some contexts in which they are not transposed. For instance *Pakistan has become a nuclear power* cannot be changed to *Pakistan has become an atomic energy*.

a million	*one million*
at work	at the workplace, employment, held, operate, organised, taken place, took place, *working*
carbon dioxide	*CO2*
close to	a stone's throw away, almost, around, densely, *close*, in the vicinity, near, next to, virtually
crystal clear	all clarity, clear, clearly, no uncertain, *quite clear*, quite clearly, very clear, very clear and comprehensive, very clearly
driving force	capacity, driver, *engine*, force, locomotive force, motor, potential, power, strength
first half	*first six months*
great care	*a careful approach*, attention, greater emphasis, particular attention, special attention, specific attention, very careful
green light	approval, call, *go-ahead*, indication, message, sign, signal, signals, formal go-ahead
long ago	a little time ago, a long time, *a long time ago*, a while ago, a while back, for a long time, long, long time, long while
long run	duration, lasting, long lived, *long term*, longer term, permanent fixture, permanent one, term
military action	military activity, military activities, *military operation*
military force	armed forces, defence, *force*, forces, military forces, peacekeeping personnel
nuclear power	atomic energy, *nuclear*
pick up	*add*, highlight, point out, say, single out, start, take, take over the baton, take up
public transport	field of transport, *transport*, transport systems
quest for	ambition to, benefit, concern, efforts to, endeavor to, favor, strive for, rational of, *view to*
sooner or later	*at some point*, eventually
step up	enhanced, increase, increased, increasing, more, strengthen, *strengthening*, reinforce, reinforcement
under control	checked, curbed, *in check*, limit, slow down

Table 4.6: Paraphrases extracted from a manually word-aligned parallel corpus. The italicized paraphrases have the highest probability according to Equation 3.2.

	Correct Meaning & Grammatical	Correct Meaning
Manual Alignments	75.0%	84.7%
Automatic Alignments	48.9%	64.5%
Using Multiple Corpora	54.9%	65.4%
Word Sense Controlled	57.0%	69.7%

Table 4.7: Paraphrase accuracy and correct meaning for the four primary data conditions

average of 75% of the time. They were judged to have the correct meaning 84.7% of the time. The difference between the two numbers shows that sometimes a paraphrase substitution can have the correct meaning but not be grammatically correct. Sometimes a substitution holds up to both criteria. For instance:

> I personally thought this problem was resolved **long ago**.
> I personally thought this problem was resolved **a long time ago**.

In other contexts that same substitution might have the correct meaning but be disfluent. For example:

> French mayors used bulldozers against immigrants not so **long ago**.
> ∗French mayors used bulldozers against immigrants not so **a long time ago**.

In this case the expression *not so long ago* is not something that can be internally modified.[4] There are cases where the reverse holds true; where a paraphrase substitution is grammatical but has the wrong meaning. Consider the example of *first half* and *first six months*. In many cases it is a perfectly valid substitution:

> The youth council will hold national meetings in the **first half** of 2007.
> The youth council will hold national meetings in the **first six months** of 2007.

But in other cases the substitution is fluent, but wrong:

> Armies clashed throughout the **first half** of the century.
> Armies clashed throughout the **first six months** of the century.

In some cases there is the syntactic role of the paraphrases vary from the original phrase For example, the noun *reinforcement* is posited as a potential paraphrase of the verb *step up*, but would not be an allowable substitute, although *reinforce* would be:

[4] Although the whole multiword expression might be paraphrased as *not such a long time ago*.

We must begin to **step up** the security at our unprotected ports.
∗We must begin to **reinforcement** the security at our unprotected ports.
We must begin to **reinforce** the security at our unprotected ports.

In other cases the paraphrases themselves have same syntactic role as the original phrase, but differ in the kinds of arguments that they take. For instance *quest for* and *endeavor to* take different types of compliments, making the substitution of one for the other impossible without transforming subsequent words in the sentence:

The **quest for** readability is never ending.
∗The **endeavor to** readability is never ending.
The **endeavor to make this readable** is never ending.

The language model probability analyzed in Section 4.3.5 may filter out some of examples with the wrong syntactic type (since, the trigram *to reinforcement the* would have a much lower probability than *to reinforce the*). However, problem might be better addressed directly by accounting for the syntactic types of phrases and their arguments, as proposed in Section 3.4.3.

By and large our paraphrases have very good quality. On average 85% have correct meaning. However, we must keep in mind that this is in an idealized setting. In the next section we examine quality when we use automatic word alignments which are error prone, and therefore may introduce errors into the paraphrases.

4.3.2 Automatic alignments (baseline system)

In this experimental condition paraphrases were extracted from a set of automatic alignments produced by running Giza++ over a set of 751,000 German-English sentence pairs (roughly 16,000,000 words in each language). When the single best paraphrase (irrespective of context) was used in place of the original phrase in the evaluation sentence the accuracy reached 48.9% which is quite low compared to the 75% of the manually aligned set. Many of these errors are due to misalignments where the paraphrases are only off by one word. For example, for paraphrases of *green light* the best paraphrase extracted from the manually aligned corpus is *go ahead*, but for the automatic alignments it is missing the word *go*, which renders it incorrect:

This report would give the **green light** to result-oriented spending.
This report would give the **go-ahead** to result-oriented spending.
∗This report would give the **ahead** to result-oriented spending.

A similar thing happens for paraphrases of the phrase *military action*:

I won't make value judgments about a specific NATO **military action**.
I won't make value judgments about a specific NATO **military operation**.
∗I won't make value judgments about a specific NATO **military**.

In this data condition it seems that we are selecting phrases which frequently have the correct meaning (64.5%) but are not grammatical – partially due to the misalignments. These results suggest two things: that improving the quality of automatic alignments would lead to more accurate paraphrases, and that there is room for improvement in limiting the paraphrases by their context. We address these points below.

4.3.3 Using multiple corpora

Work in statistical machine translation suggests that, like many other machine learning problems, performance increases as the amount of training data increases. Och and Ney (2003) show that the accuracy of alignments produced by Giza++ improve as the size of the training corpus increases. Since we used the whole of the German-English section of the Europarl corpus, we were prevented from trying to improve the alignments by simply adding more German-English training data. However, another way of effectively increasing the amount of training data used for paraphrasing is to extract paraphrases from multiple parallel corpora. For this condition we used Giza++ to align the French-English, Spanish-English, and Italian-English portions of the Europarl corpus in addition to the German-English portion, for a total of nearly 3,000,000 sentence pairs in the training data. This also has the advantage of potentially diminishing problems associated with systematic misalignments in one language pair. The extent to which this holds is variable. For example, for the *green light* example above the multiple parallel corpora do not contain the *ahead / go-ahead* misalignment but instead have a different misalignment which introduces *green* as a paraphrase:

∗This report would give the **ahead** to result-oriented spending.
? This report would give the **green** to result-oriented spending.

In other cases the multiple corpora manage to overcome the problem of misalignments in a single language pair:

∗I won't make value judgments about a specific NATO **military**.
I won't make value judgments about a specific NATO **military intervention**.

Overall the accuracy of paraphrases extracted over multiple corpora increased from 49% to 55%. These could be further improved by including other English parallel

corpora, such as the remainder of the Europarl set, the GALE Chinese-English and Arabic-English corpora, or the Canadian Hansards. The improvements for meaning alone were less dramatic, increasing by only 1%. In the next section we shall see that word sense disambiguation has the potential to improve both meaning and accuracy more effectively.

4.3.4 Controlling for word sense

As discussed in Section 3.3.2, the way that we extract paraphrases is the converse of the methodology employed in word sense disambiguation work that uses parallel corpora (Diab and Resnik, 2002). The assumption made in the word sense disambiguation work is that if a source language word aligns with different target language words then those words may represent different word senses. This can be observed in the paraphrases for *at work* in Table 4.6. The paraphrases *at the workplace, employment,* and *in the work sphere* are a different sense of the phrase than *operate, held,* and *holding*, and they are aligned with different German phrases.

When we calculate the paraphrase probability we sum over different target language phrases. Therefore the English phrases that are aligned with the different German phrases (which themselves may be indicative of different word senses) are mingled. Performance may be degraded since paraphrases that reflect different senses of the original phrase, and which therefore have a different meaning, are included in the same candidate set. We performed an experiment to see whether improvement could be achieved by limiting the candidate paraphrases to the same sense as the original phrase in each test sentence. To do this, we used the fact that our test sentences were drawn from a parallel corpus. We limited phrases to the same word sense by constraining the candidate paraphrases to those that aligned with the same target language phrase. The paraphrase probability for this condition was calculated using Equation 4.3. Using the foreign language phrase to identify the word sense is obviously not applicable in monolingual settings, but acts as a convenient stand-in for a proper word sense disambiguation algorithm here.

When word sense is controlled in this way, the accuracy of the paraphrases extracted from the automatic alignments rises dramatically from 48.9% to 57%. The percent of items with correct meaning also jumps significantly from 64.5% to 69.7%, a much more dramatic increase than when integrating multiple parallel corpora. Moreover, these methods could potentially be combined for further improvements.

4.3.5 Including a language model probability

In order to allow the surrounding words in the sentence to have an influence on which paraphrase was selected, we re-ranked the paraphrase probabilities based on a trigram language model trained on the entire English portion of the Europarl corpus. Table 4.8 presents the results for each of the conditions when the language model probability is combined with the paraphrase probability. By comparing the numbers in Table 4.8 to those in Table 4.7 we can see how effective the language model is at making the output sentences more fluent. In most cases it improves fluency, as reflected in an increase in the percent of time the annotators judged the paraphrases to both have the correct meaning and be grammatical. For the automatic alignment condition accuracy jumps by 6.4%, when using multiple parallel corpora it increases by 2.4%, and when controlling for word sense it increases by 4.9%. In the case of the manual alignments accuracy dips from 75% to 71.8%.

In most cases the language model also seems to lead to decreased performance when meaning is the sole criterion, dropping by 3.7% for manual and automatic alignments, by 2.1% for multiple parallel corpora, and essentially remaining unchanged for the word sense condition.

Some of the errors in meaning are introduced when the language model probability is high for an inaccurate paraphrase created through misalignment. For instance, *on* is extracted as a potential paraphrase of *on average* due to errors in the automatic alignments. Substituting *on* for *on average* in some situations still results in a grammatical sentence, but it does not reflect the meaning of the original phrase:

> This leads **on average** to higher returns.
> This leads **on** to higher returns.

A similar situation arises when inaccurate alignments allow *red cross* to be paraphrased as *cross*:

> The symbol of the **red cross** brings hope to battlefields worldwide.
> The symbol of the **cross** brings hope to battlefields worldwide.

These examples suggests that the language model does quite a good job at selecting well-formed sentences, but that random, inaccurate paraphrases give it too much latitude for constructing such sentences. This problem might be ameliorated in a number of ways: the possible set of paraphrases could be filtered to try to eliminate inaccurate paraphrase (such as the substrings shown above), or the language model could be weighted differently.

	Correct Meaning & Grammatical	Correct Meaning
Manual Alignments	71.8%	81.0%
Automatic Alignments	55.3%	60.8%
Using Multiple Corpora	57.3%	63.5%
Word Sense Controlled	61.9%	70.5%

Table 4.8: Percent of time that paraphrases were judged to be correct when a language model probability was included alongside the paraphrase probability

4.4 Discussion

In this chapter we presented experiments which evaluated the quality of paraphrases that were extracted by our paraphrasing technique. We showed that in principle our method can achieve very high quality paraphrases with 85% having the correct meaning and 75% also being grammatical in context. In more realistic scenarios we are able to achieve paraphrases that retain correct meaning more than 70% of the time and are grammatical nearly two thirds of the time. Barzilay and McKeown (2001) reported an average precision of 86% at identifying paraphrases out of context, and of 91% when the paraphrases are substituted into the original context of the aligned sentence, based on "approximate conceptual equivalence". Ibrahim et al. (2003) produced paraphrases which were "roughly interchangeable given the genre" an average of 41% of the time on a set of 130 paraphrases. Our evaluation criteria were stricter and our methodology was more rigorous so our numbers compare quite favorably.

In the next chapter we explore an application of paraphrases which takes advantages of some of the additional features of our technique which were not explored in this chapter. We show that paraphrases can be used to improve the quality of statistical machine translation by reducing problems associated with coverage. The application of our paraphrasing technique is greatly facilitated by the fact that it can be easily applied to any language, can extract paraphrases for a wide range of phrases, and has a probabilistic formulation.

Chapter 5

Improving Statistical Machine Translation with Paraphrases

In this chapter[1] we describe one way in which statistical machine translation can be improved using paraphrases. Specifically, we focus on the problem of coverage. To increase coverage we apply paraphrases to source language phrases that are unseen in the training data (as described below). However, this is by no means the only way of improving translation using paraphrases. We could also apply paraphrasing when the target is unseen, or when the source or target is seen. Using paraphrases in each of these possible cases could potentially improve a different aspect of statistical machine translation:

- Paraphrasing unseen target phrases could come into play when there is no way for a system to produce a reference translation given its training data. Paraphrasing the reference sentence could allow the system to better match it, which might be beneficial during minimum error rate training or when automatically evaluating system output.

- Paraphrasing seen source and/or target phrases potentially help with alignment. Paraphrasing could be used to group words and phrases in the training set which have similar meaning. These equivalence classes might allow an alignment algorithm to converge on better alignments than when the relationship between words is unspecified.

- Paraphrasing seen source phrases might allow us to transform an input sentence

[1] Chapters 5 and 7 extend Callison-Burch et al. (2006a). Chapter 5 adds additional exposition about how we extend SMT with paraphrases, and Chapter 7 does additional analysis of experimental results.

onto something that is easier to translate. In this chapter we propose paraphrasing and then translating unseen source phrases. Doing the same with phrases which occurred in the training data below some threshold might have a similar benefit, since phrases which occurred infrequently are less likely to translate correctly.

Any of the above scenarios could be a potential application of paraphrases to machine translation. A number of these scenarios have been explored by other researchers using our paraphrasing method: Owczarzak et al. (2006) and Zhou et al. (2006) use it to extend machine translation evaluation metrics, and Madnani et al. (2007) use it to augment minimum error rate training. Other researchers applied different paraphrasing techniques to problems in machine translation. Kanayama (2003) uses manually crafted paraphrasing rules to create a canonical representation for evaluation data, and Kauchak and Barzilay (2006) use WordNet paraphrases to facilitate automatic evaluation.

In this chapter we apply paraphrases to a different aspect of machine translation, applying them to statistical machine translation to address the problem of coverage. Coverage is a significant problem because SMT learns translations from data which is often limited in size. Therefore many source words and phrases that occur in test data may not occur in the training data. Current systems handle this situation poorly.

5.1 The problem of coverage in SMT

Statistical machine translation made considerable advances in translation quality with the introduction of phrase-based translation. By increasing the size of the basic unit of translation, phrase-based machine translation does away with many of the problems associated with the original word-based formulation of statistical machine translation (Brown et al., 1993). For instance, some words which are ambiguous in translation are less so when adjacent words are considered. Furthermore, with multi-word units less re-ordering needs to occur since local dependencies are frequently captured. For example, common adjective-noun alternations are memorized. However, since this linguistic information is not explicitly and generatively encoded in the model, unseen adjective noun pairs may still be handled incorrectly.

Thus, having observed phrases in the past dramatically increases the chances that they will be translated correctly in the future. However, for any given test set, a huge

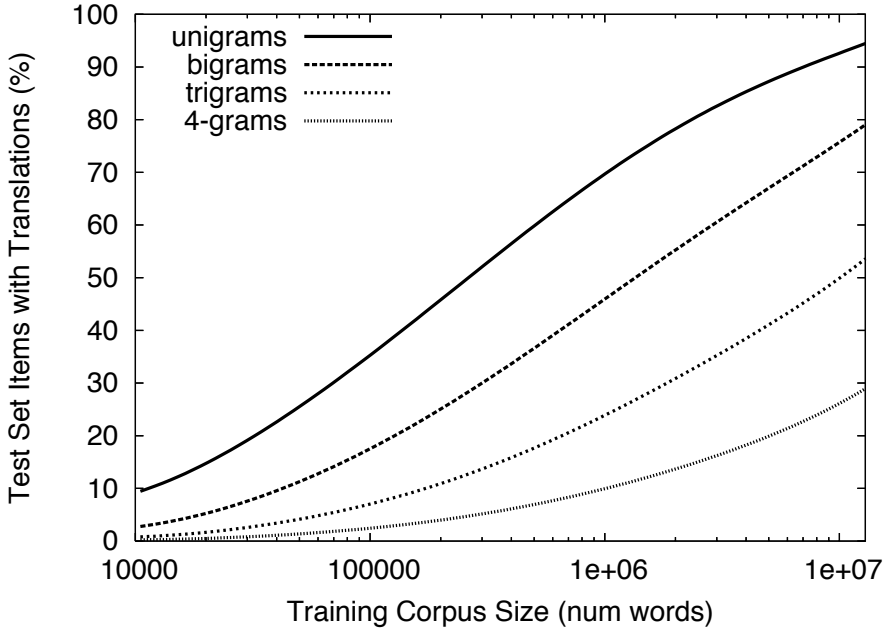


Figure 5.1: Percent of unique unigrams, bigrams, trigrams, and 4-grams from the Europarl Spanish test sentences for which translations were learned in increasingly large training corpora

amount of training data has to be observed before translations are learned for a reasonable percentage of the test phrases. Figure 5.1 shows the extent of this problem. For a training corpus containing 10,000 words translations will have been learned for only 10% of the unigrams (*types*, not tokens). For a training corpus containing 100,000 words this increases to 30%. It is not until nearly 10,000,000 words worth of training data have been analyzed that translation for more than 90% of the vocabulary items have been learned. This problem is obviously compounded for higher-order n-grams (longer phrases).

The problem of coverage is also exacerbated in a number of other situations. It is especially problematic when we are dealing with so-called *low density* languages which do not have very large parallel corpora. Coverage is also related to the morphological complexity of a language, since morphologically rich languages have a greater number of word forms and therefore a larger amount of data is required to observe them all. Coverage also makes it difficult to translate texts that are outside the domain of the training data, since specialized terminology will not be covered.

encargarnos	to ensure, take care, ensure that
garantizar	guarantee, ensure, to ensure, ensuring, guaranteeing
velar	ensure, make sure, safeguard, protect, ensuring
procurar	ensure that, try to, ensure, endeavour to
asegurarnos	ensure, secure, make certain
usado	used
utilizado	used, use, spent, utilized
empleado	used, spent, employee
uso	use, used, usage
utiliza	used, uses, used, being used
utilizar	to use, use, used

Table 5.1: Example of automatically generated paraphrases for the Spanish words *encargarnos* and *usado* along with their English translations which were automatically learned from the Europarl corpus

5.2 Handling unknown words and phrases

Currently many statistical machine translation systems are simply unable to handle unknown words. There are two strategies that are commonly employed when an unknown source word is encountered. Either the source word is simply omitted when producing the translation, or alternatively it is passed through untranslated, which is a reasonable strategy if the unknown word happens to be a name (assuming that no transliteration need be done). Neither of these strategies is satisfying, because information is lost when words are deleted, and words passed through untranslated are unhelpful since users of MT systems generally do not have competency in the source language.

When a system is trained using 10,000 sentence pairs (roughly 200,000 words) there will be a number of words and phrases in a test sentence which it has not learned the translation of. For example, the Spanish sentence:

> *Es positivo llegar a un acuerdo sobre los procedimientos, pero debemos encargarnos de que este sistema no sea susceptible de ser usado como arma política.*

may translate as:

> It is good reach an agreement on procedures, but we must *encargarnos* that this system is not susceptible to be *usado* as arms policy.

Table 5.1 gives example paraphrases of the unknown source words along with their translations. If we had learned a translation of *garantizar* we could translate it instead of *encargarnos*, and similarly we could translate *utilizado* instead of *usado*. This would allow us to produce an improved translation such as:

> It is good reach an agreement on procedures, but we must **guarantee** that this system is not susceptible to be **used** as arms policy.

Thus the previously untranslated source words can be translated appropriately.

We extend this strategy so that in addition to substituting paraphrases in for unknown words we do the same for *unknown phrases* as well. This allows us to take advantage of the fact that using longer phrases generally results in higher quality translations since they have additional context. For example, while the translation model might contain translations for the Spanish words *arma* and *política* individually, it might not contain a translation for the two word phrase *arma política*. While *arma* can be correctly translated as *arms* in some contexts and while it is acceptable to render *política* as *policy* in most contexts, when they occur together as a phrase they should be translated as *political weapon* instead of *arms policy*. We can attempt to improve the translation by paraphrasing the phrase *arma política*. Just as we use parallel corpora to generate paraphrases for single words, we can also use them to generate paraphrases for multiword phrases. Table 5.1 gives example paraphrases for *arma política* along with their translations. If we had learned a translation of *recurso político* we could translate it instead of *arma política*, and the resulting translation would be better:

> It is good reach an agreement on procedures, but we must guarantee that this system is not susceptible to be used as **political weapon**.

Thus substituting paraphrases for unknown phrases may lead to improved translation quality within phrase-based SMT.

While any paraphrasing method could potentially be used to increase the coverage of statistical machine translation, the method that we defined in Chapter 3 has several features that make it an ideal candidate for incorporation into statistical machine translation system. It is language independent, in that it can easily be applied to any language for which we have one or more parallel corpora, making it an appropriate paraphrasing technique for the task of machine translation. It has high recall, in that it is able to generate paraphrases for many phrases, making it appropriate for the problem of coverage. It defines a mechanism for assigning probabilities to paraphrases, allowing it to be incorporated into the probabilistic framework of SMT.

arma política	political weapon, political tool
recurso político	political weapon, political asset
instrumento político	political instrument, instrument of policy, policy instrument, policy tool, political implement, political tool
arma	weapon, arm, arms
palanca política	political lever
herramienta política	political tool, political instrument

Table 5.2: Example of paraphrases for the Spanish phrase *arma política* and their English translations

5.3 Increasing coverage of parallel corpora with parallel corpora?

Our technique extracts paraphrases from parallel corpora. While it may seem circular to try to alleviate the problems associated with small parallel corpora using paraphrases generated from parallel corpora, it is not. The reason that it is not is the fact that paraphrases can be generated from parallel corpora between the source language and languages *other than* the target language. For example, when translating from English into a minority language like Maltese we will have only a very limited English-Maltese parallel corpus to train our translation model from, and will therefore have only a relatively small set of English phrases for which we have learned translations. However, we can use many other parallel corpora to train our paraphrasing model. We can generate English paraphrases using the English-Danish, English-Dutch, English-Finnish, English-French, English-German, English-Italian, English-Portuguese, English-Spanish, and English-Swedish from the Europarl corpus. The English side of the parallel corpora does not have to be identical, so we could also use the English-Arabic and English-Chinese parallel corpora from the DARPA GALE program. Thus translation from English to Maltese can potentially be improved using parallel corpora between English and any other language.

Note that there is an imbalance since translation is only improved when translating from the resource rich language into the resource poor one. Therefore additional English corpora are not helpful when translating from Maltese into English. In the scenario when we are interested in translating from Maltese into English, we would need some other mechanism for generating paraphrases. Since Maltese is resource poor,

the paraphrasing techniques which utilize monolingual data (described in Section 2.1) may also be impossible to apply. There are no parsers for Maltese, ruling out Lin and Pantel's method. There are not ready sources of multiple translations into Maltese, ruling out Barzilay and McKeown's and Pang et al.'s techniques. It is unlikely there are enough newswire agencies servicing Malta to construct the comparable corpus that would be necessary for Quirk et al.'s method.

5.4 Integrating paraphrases into SMT

The crux of our strategy for improving translation quality is this: replace unknown source words and phrases with paraphrases for which translations are known. There are a number of possible places that this substitution could take place in an SMT system. For instance the substitution could take place in:

- A preprocessing step whereby we replace each unknown word and phrase in a source sentence with their paraphrases. This would result in a set of many paraphrased source sentences. Each of these sentences could be translated individually.
- A post-processing step where any source language words that were left untranslated were paraphrased and translated subsequent to the translation of the sentence as a whole.

Neither of these is optimal. The first would potentially generate too many sentences to translate because of the number of possible permutations of paraphrases. The secould would give no way of recognizing unknown phrases. Neither would give a way of choosing between multiple outcomes. Instead we have an elegant solution for performing the substitution which integrates the different possible paraphrases into decoding that takes place when producing a translation, and which takes advantage of the probabilistic formulation of SMT. We perform the substitution by expanding the *phrase table* used by the decoder, as described in the next section.

5.4.1 Expanding the phrase table with paraphrases

The decoder starts by matching all source phrases in an input sentence against its phrase table, which contains some subset of the source language phrases, along with their translations into the target language and their associated probabilities. Figure 5.2

garantizar

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
guarantee	0.38	0.32	0.37	0.22	2.718
ensure	0.21	0.39	0.20	0.37	2.718
to ensure	0.05	0.07	0.37	0.22	2.718
ensuring	0.05	0.29	0.06	0.20	2.718
guaranteeing	0.03	0.45	0.04	0.44	2.718

velar

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
ensure	0.19	0.01	0.37	0.05	2.718
make sure	0.10	0.04	0.01	0.01	2.718
safeguard	0.08	0.01	0.05	0.03	2.718
protect	0.03	0.03	0.01	0.01	2.718
ensuring	0.03	0.01	0.05	0.04	2.718

recurso político

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
political weapon	0.01	0.33	0.01	0.50	2.718
political asset	0.01	0.88	0.01	0.50	2.718

arma

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
weapon	0.65	0.64	0.70	0.56	2.718
arms	0.02	0.02	0.01	0.02	2.718
arm	0.01	0.06	0.01	0.02	2.718

Figure 5.2: Phrase table entries contain a source language phrase, its translations into the target language, and feature function values for each phrase pair

gives example phrase table entries for the Spanish phrases *garantizar*, *velar*, *recurso político*, and *arma*. In addition to their translations into English the phrase table entries store five feature function values for each translation:

- $p(\bar{e}|\bar{f})$ is the phrase translation probability for an English phrase $\bar{e}$ given the Spanish phrase $\bar{f}$. This can be calculated with maximum likelihood estimation as described in Equation 2.7, Section 2.2.2.
- $p(\bar{f}|\bar{e})$ is the reverse phrase translation probability. It is the phrase translation probability for a Spanish phrase $\bar{f}$ given an English phrase $\bar{e}$.
- $lex(\bar{e}|\bar{f})$ is a lexical weighting for the phrase translation probably. It calculates the probability of translation of each individual word in the English phrase given the Spanish phrase.
- $lex(\bar{f}|\bar{e})$ is the lexical weighting applied in the reverse direction.
- the phrase penalty is a constant value ($exp(1) = 2.718$) which helps the decoder regulate the number of phrases that are used during decoding.

The values are used by the decoder to guide the search for the best translation, as described in Section 2.2.3. The role that they play is further described in Section 7.1.2.

The phrase table contains *the complete set of translations* that the system has learned. Therefore, if there is a source word or phrase in the test set which does not

have an entry in the phrase table then the system will be unable to translate it. Thus a natural way to introduce translations of unknown words and phrases is to expand the phrase table. After adding the translations for words and phrases they may be used by the decoder when it searches for the best translation of the sentence. When we expand the phrase table we need two pieces of information for each source word or phrase: its translations into the target language, and the values for the feature functions, such as the five given in Figure 5.2.

Figure 5.3 demonstrates the process of expanding the phrase table to include entries for the Spanish word *encargarnos* and the Spanish phrase *arma política* which the system previously had no English translation for. The expansion takes place as follows:

- Each unknown Spanish item is paraphrased using parallel corpora other than the Spanish-English parallel corpus, creating a list of potential paraphrases along with their paraphrase probabilities, $p(\bar{f}_2|\bar{f}_1)$.
- Each of the potential paraphrases is looked up in the original phrase table. If any entry is found for one or more of them then an entry can be added for the unknown Spanish item.
- An entry for the previously unknown Spanish item is created, giving it the translations of each of the paraphrases that existed in the original phrase table, with appropriate feature function values.

For the Spanish word *encargarnos* our paraphrasing method generates four paraphrases. They are *garantizar*, *velar*, *procurar*, and *asegurarnos*. The existing phrase table contains translations for two of those paraphrases. The entries for *garantizar* and *velar* are given in Figure 5.2. We expand the phrase table by adding a new entry for the previously untranslatable word *encargarnos*, using the translations from *garantizar* and *velar*. The new entry has ten possible English translations. Five are taken from the phrase table entry for *garantizar*, and five from *velar*. Note that some of the translations are repeated because they come from different paraphrases.

Figure 5.3 also shows how the same procedure can be used to create an entry for the previously unknown phrase *arma política*.

5.4.2 Feature functions for new phrase table entries

To be used by the decoder each new phrase table entry must have a set of specified probabilities alongside its translation. However, it is not entirely clear what the val-

paraphrases *existing phrase table entries* *new phrase table entry*

encargarnos

paraphrases	p(f2lf1)
garantizar	0.07
velar	0.06
procurar	0.04
asegurarnos	0.01

+ garantizar

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty	p(f2lf1)
guarantee	0.38	0.32	0.37	0.22	2.718	1.0
ensure	0.21	0.39	0.20	0.37	2.718	1.0
to ensure	0.05	0.07	0.37	0.22	2.718	1.0
ensuring	0.05	0.29	0.06	0.20	2.718	1.0
guaranteeing	0.03	0.45	0.04	0.44	2.718	1.0

+ velar

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty	p(f2lf1)
ensure	0.19	0.01	0.37	0.05	2.718	1.0
make sure	0.10	0.04	0.01	0.01	2.718	1.0
safeguard	0.08	0.01	0.05	0.03	2.718	1.0
protect	0.03	0.03	0.01	0.01	2.718	1.0
ensuring	0.03	0.01	0.05	0.04	2.718	1.0

= encargarnos

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty	p(f2lf1)
guarantee	0.38	0.32	0.37	0.22	2.718	0.07
ensure	0.21	0.39	0.20	0.37	2.718	0.07
to ensure	0.05	0.07	0.37	0.22	2.718	0.07
ensuring	0.05	0.29	0.06	0.20	2.718	0.07
guaranteeing	0.03	0.45	0.04	0.44	2.718	0.07
ensure	0.19	0.01	0.37	0.05	2.718	0.06
make sure	0.10	0.04	0.01	0.01	2.718	0.06
safeguard	0.08	0.01	0.05	0.03	2.718	0.06
protect	0.03	0.03	0.01	0.01	2.718	0.06
ensuring	0.03	0.01	0.05	0.04	2.718	0.06

paraphrases *existing phrase table entries* *new phrase table entry*

arma política

paraphrases	p(f2lf1)
recurso político	0.08
instrumento político	0.06
arma	0.04
palanca política	0.04
herramienta política	0.02

+ arma

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty	p(f2lf1)
weapon	0.65	0.64	0.70	0.56	2.718	1.0
arms	0.02	0.02	0.01	0.02	2.718	1.0
arm	0.01	0.06	0.01	0.02	2.718	1.0

+ recurso político

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty	p(f2lf1)
political weapon	0.01	0.33	0.01	0.50	2.718	1.0
political asset	0.01	0.88	0.01	0.50	2.718	1.0

= arma política

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty	p(f2lf1)
political weapon	0.01	0.33	0.01	0.50	2.718	0.08
political asset	0.01	0.88	0.01	0.50	2.718	0.08
weapon	0.65	0.64	0.70	0.56	2.718	0.04
arms	0.02	0.02	0.01	0.02	2.718	0.04
arm	0.01	0.06	0.01	0.02	2.718	0.04

Figure 5.3: A phrase table entry is generated for a phrase which does not initially have translations by first paraphrasing the phrase and then adding the translations of its paraphrases.

ues of feature functions like the phrase translation probability $p(\bar{e}|\bar{f})$ should be for entries created through paraphrasing. What value should be assign to the probability $p(guarantee \mid encargarnos)$, given that the pair of words were never observed in our training data? We can no longer rely upon maximum likelihood estimation as we do for observed phrase pairs.

Yang and Kirchhoff (2006) encounter a similar situation when they add phrase table entries for German phrases that were unobserved in their training data. Their strategy was to implement a back off model. Generally speaking, backoff models are used when moving from more specific probability distributions to more general ones. Backoff models specify under which conditions the more specific model is used and when the model "backs off" to the more general distribution. When a particular German phrase was unobserved, Yang and Kirchhoff's backoff model moves from values for a more specific phrase (the fully inflected, compounded German phrases) to the more general phrases (the decompounded, uninflected versions). They assign their backoff probability for

$$p_{BO}(\bar{e}|\bar{f}) = \begin{cases} d_{\bar{e},\bar{f}} p_{orig}(\bar{e}|\bar{f}) & \text{If } count(\bar{e},\bar{f}) > 0 \\ p(\bar{e}|stem(\bar{f})) & \text{Otherwise} \end{cases}$$

where $d_{\bar{e},\bar{f}}$ is a discounting factor. The discounting factor allows them to borrow probability mass from the items that were observed in the training data and divide it among the phrase table entries that they add for unobserved items. Therefore the values of translation probabilities like $p(\bar{e}|\bar{f})$ for observed items will be slightly less than their maximum likelihood estimates, and the $p(\bar{e}|\bar{f})$ values for the unobserved items will some fractional value of the difference.

We could do the same with entries created via paraphrasing. We could create a backoff scheme such that if a specific source word or phrase is not found then we back off to a set of paraphrases for that item. It would require reducing the probabilities for each of the observed word and phrases items and spreading their mass among the paraphrases. Instead of doing that, we take the probabilities directly from the observed words and assign them to each of their paraphrases. We do not decrease probability mass from the unparaphrased entry feature functions, $p(\bar{e}|\bar{f})$, $p(\bar{f}|\bar{e})$ etc., and so the total probability mass of these feature functions will be greater than one. In order to compensate for this we introduce a new feature function to act as a *scaling factor* that down-weights the paraphrased entries.

The new feature function incorporates the paraphrase probability. We designed the paraphrase probability feature function (denoted by h) to assign the following values

to entries in the phrase table:

$$h(\mathbf{e},\mathbf{f_1}) = \begin{cases} p(\mathbf{f_2}|\mathbf{f_1}) & \text{If phrase table entry } (\mathbf{e},\mathbf{f_1}) \\ & \text{is generated from } (\mathbf{e},\mathbf{f_2}) \\ 1 & \text{Otherwise} \end{cases}$$

This means that if an entry existed prior to expanding the phrase table via paraphrasing, it would be assigned the value 1. If the entry was created using the translations of a paraphrase then it is given the value of the paraphrase probability. Since the translations for a previously untranslatable entry can be drawn from more than one paraphrase the value of $p(\mathbf{f_2}|\mathbf{f_1})$ can be different for different translations. For instance, in Figure 5.3 for the newly created entry for *encargarnos*, the translation *guarantee* is taken from the paraphrase *garantizar* and is therefore given the value of its paraphrase probability which is 0.07. The translation *safeguard* is taken from the paraphrase *velar* and is given its paraphrase probability which is 0.06.

The paraphrase probability feature function has the advantage of distinguishing between entries that were created by way of paraphrases which are very similar to the unknown source phase, and those which might be less similar. The paraphrase probability should be high for paraphrases which are good, and low for paraphrases which are less so. Without incorporating the paraphrase probability, translations which are borrowed from bad paraphrases would have equal status to translations which are taken from good paraphrases.

5.5 Summary

This chapter gave an overview of how paraphrases can be used to alleviate the problem of coverage in SMT. We increase the coverage of SMT systems by locating previously unknown source words and phrases and substituting them with paraphrases for which the system has learned a translation. In Section 5.2 we motivated this by showing how substituting paraphrases in before translation could improve the resulting translations for both words and phrases. In Section 5.4 we described how paraphrases could be integrated into a SMT system, by performing the substitution in the phrase table. In order to test the effectiveness of the proposal that we outlined in this chapter, we need an experimental setup. Since our changes effect only the phrase table, we require no modifications to the inner workings of the decoder. Thus our method for improving the coverage of SMT with paraphrases can be straightforwardly tested by using an existing decoder implementation such as Pharaoh (Koehn, 2004) or Moses (Koehn et al., 2006).

The Chapter 7.1 gives detailed information about our experimental design, what data we used to train our paraphrasing technique and our translation models, and what experiments we performed to determine whether the paraphrase probability plays a role in improving quality. Chapter 7.2 presents our results that show the extent to which we are able to improve statistical machine translation using paraphrases. Before we present our experiments, we first delve into the topic of how to go about evaluating translation quality. Chapter 6 describes the methodology that is commonly used to evaluation translation quality in machine translation research. In that chapter we argue that the standard evaluation methodology is potentially insensitive to the types of translation improvements that we make, and present an alternative methodology which is sensitive to such changes.

Chapter 6

Evaluating Translation Quality

In order to determine whether a proposed change to a machine translation system is worthwhile some sort of evaluation criterion must be adopted. While evaluation criteria can measure aspects of system performance (such as the computational complexity of algorithms, average runtime speeds, or memory requirements), they are more commonly concerned with the *quality of translation*. The dominant evaluation methodology over the past five years has been to use an automatic evaluation metric called Bleu (Papineni et al., 2002). Bleu has largely supplanted human evaluation because automatic evaluation is faster and cheaper to perform. The use of Bleu is widespread. Conference papers routinely claim improvements in translation quality by reporting improved Bleu scores, while neglecting to show any actual example translations. Workshops commonly compare systems using Bleu scores, often without confirming these rankings through manual evaluation. Research which has not show improvements in Bleu scores is sometimes dismissed without acknowledging that the evaluation metric itself might be insensitive to the types of improvements being made.

In this chapter[1] we argue that Bleu is not as strong a predictor of translation quality as currently believed and that consequently the field should re-examine the extent to which it relies upon the metric. In Section 6.1 we examine Bleu's deficiencies, showing that its model of allowable variation in translation is too crude. As a result, Bleu can fail to distinguish between translations of significantly different quality. In Section 6.2 we discuss the implications for evaluating whether paraphrases can be used to improve translation quality as proposed in the previous chapter. In Section 6.3 we present an alternative evaluation methodology in the form of a focused manual evaluation which

[1]This chapter elaborates upon Callison-Burch et al. (2006b) with additional discussion of allowable variation in translation, and by presenting a method for targeted manual evaluation.

targets specific aspects of translation, such as improved coverage.

6.1 Re-evaluating the role of BLEU in machine translation research

The use of Bleu as a surrogate for human evaluation is predicated on the assumption that it correlates with human judgments of translation quality, which has been shown to hold in many cases (Doddington, 2002; Coughlin, 2003). However, there are questions as to whether improving Bleu score always guarantees genuine translation improvements, and whether Bleu is suitable for measuring all types of translation improvements. In this section we show that under some circumstances an improvement in Bleu is *not sufficient* to reflect a genuine improvement in translation quality, and in other circumstances that it is *not necessary* to improve Bleu in order to achieve a noticeable (subjective) improvement in translation quality. We argue that these problems arise because Bleu's model of *allowable variation in translation* is inadequate. In particular, we show that Bleu has a weak model of variation in phrase order and alternative wordings. Because of these weaknesses, Bleu admits a huge amount of variation for identically scored hypotheses. Typically there are millions of variations on a hypothesis translation that receive the same Bleu score. Because not all these variations are equally grammatically or semantically plausible, there are translations which have the *same* Bleu score but would be judged *worse* in a human evaluation. Similarly, some types of changes are indistinguishable to Bleu, but do in fact represent genuine improvements to translation quality.

6.1.1 Allowable variation in translation

The rationale behind the development of automatic evaluation metrics is that human evaluation can be time consuming and expensive. Automatic evaluation metrics, on the other hand, can be used for frequent tasks like monitoring incremental system changes during development, which are seemingly infeasible in a manual evaluation setting. The way that Bleu and other automatic evaluation metrics work is to compare the output of a machine translation system against reference human translations. After a reference has been produced then it can be reused for arbitrarily many subsequent evaluations. The use of references in the automatic evaluation of machine translation is complicated by the fact that there is a degree of *allowable variation* in translation.

Machine translation evaluation metrics differ from metrics used in other tasks, such as automatic speech recognition, which use a reference. The difference arises because there are many equally valid translations for any given sentence. The word error rate (WER) metric that is used in speech recognition can be defined in a certain way because there is much less variation in its references. In speech recognition, each utterance has only a single valid reference transcription. Because each reference transcription is fixed, the WER metric can compare the output of a speech recognizer against the reference using string edit distance which assumes that the transcribed words are unambiguous and occur in the fixed order (Levenshtein, 1966). In translation, on the other hand, there are different ways of wording a translation, and some phrases can occur in different positions in the sentence without affecting its meaning or its grammaticality. Evaluation metrics for translation need some way to correctly reward translations that deviate from a reference translation in acceptable ways, and penalize variations which are unacceptable.

Here we examine the consequences for an evaluation metric when it poorly models allowable variation in translation. We focus on two types of variation that are most prominent in translation:

- Variation in the wording of a translation – a translation can be phrased differently without affecting its translation quality.
- Variation in phrase order – some phrases such as adjuncts can occur in a number of possible positions in a sentence.

Section 6.1.2 gives the details of how Bleu scores translations by matching them against *multiple* reference translations, and how it attempts to model variation in word choice and phrase order. Section 6.1.3 discusses why its model is poor and what consequences this has for the reliability of Bleu's predictions about translation quality. Section 6.2 discusses the implications for evaluating the type of improvements that we make when introducing paraphrases into translation.

6.1.2 BLEU detailed

Like other automatic evaluation metrics of translation quality, Bleu compares the output of a MT system against reference translations. Alternative wordings present challenges when trying to match words in a reference translation. The fact that some words and phrases may occur in different positions further complicates the choice of

what similarity function to use. To overcome these problems, Bleu attempts to model allowable variation in two ways:

- **Multiple reference translations** – Instead of comparing the output of a MT system against a single reference translation, Bleu can compare against a set of reference translations (as proposed by Thompson (1991)). Hiring different professional translators to create multiple reference translations for a test corpus has the effect of introducing some of the allowable variation in translation described above. In particular, different translations are often worded differently. The rate of matches of words in MT output increases when alternatively worded references are included in the comparison, thus overcoming some of the problems that arise when matching against a single reference translation.

- **Position-independent n-gram matching** – Bleu avoids the strict ordering assumptions of WER's string edit distance in order to overcome the problem of variation in phrase order. Previous work had introduced a position-independent WER metric (Niessen et al., 2000) which allowed matching words to be drawn from any position in the sentence. The Bleu metric refines this idea by counting the number of *n-gram matches*, allowing them to be drawn from any position in the reference translations. The extension from position-independent WER to position-independent n-gram matching places some constraints on word order since the words in the MT output must appear in similar order as the references in order to match higher order n-grams.

Papineni et al. (2002) define Bleu in terms of *n-gram precision*. They calculate an n-gram precision score, p_n, for each n-gram length by summing over the matches for every hypothesis sentence S in the complete corpus C as:

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}$$

Bleu's n-gram precision is modified slightly to eliminate repetitions that occur across sentences. For example, even though the bigram "to Miami" is repeated across all four reference translations in Table 6.1, it is counted only once in a hypothesis translation. These is referred to as *clipped* n-gram precision.

Bleu's calculates *precision* for each length of n-gram up to a certain maximum length. Precision is the proportion of the matched n-grams out of the total number of n-grams in the hypothesis translations produced by the MT system. When evaluating natural language processing applications it is normal to calculate *recall* in addition

Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Appeared calm when he was taken to the American plane, which will to Miami, Florida.

Table 6.1: A set of four reference translations, and a hypothesis translation from the 2005 NIST MT Evaluation

to precision. If Bleu used a single reference translation, then recall would represent the proportion of matched n-grams out of the total number of n-grams in the reference translation. However, recall is difficult to define when using multiple reference translation, because it is unclear what should comprise the counts in the denominator. It is not as simple as summing the total number of clipped n-grams across all of the reference translations, since there will be non-identical n-grams which overlap in meaning which a hypothesis translation will and should only match one instance. Without grouping these corresponding reference n-grams and defining a more sophisticated matching scheme, recall would be underestimated for each hypothesis translation.

Rather than defining n-gram recall Bleu instead introduces a *brevity penalty* to compensate for the possibility of proposing high-precision hypothesis translations which are too short. The brevity penalty is calculated as:

$$\text{BP} = \begin{cases} 1 & \text{if} \quad c > r \\ e^{1-r/c} & \text{if} \quad c \leq r \end{cases}$$

where c is the length of the corpus of hypothesis translations, and r is the effective reference corpus length. The effective reference corpus length is calculated as the sum of the single reference translation from each set which is closest to the hypothesis translation.

The brevity penalty is combined with the weighted sum of n-gram precision scores to give Bleu score. Bleu is thus calculated as

$$\text{Bleu} = \text{BP} * \exp(\sum_{n=1}^{N} w_n \log p_n)$$

A Bleu score can range from 0 to 1, where higher scores indicate closer matches to the reference translations, and where a score of 1 is assigned to a hypothesis translation which exactly matches one of the reference translations. A score of 1 is also assigned to a hypothesis translation which has matches for all its n-grams (up to the maximum n measured by Bleu) in the clipped reference n-grams, and which has no brevity penalty.

To give an idea of how Bleu is calculated we will walk through what the Bleu score would be for the hypothesis translation given in Table 6.1. Counting punctuation marks as separate tokens, the hypothesis translation has 15 unigram matches, 10 bigram matches, 5 trigram matches, and three 4-gram matches (these are shown in bold in Table 6.2). The hypothesis translation contains a total of 18 unigrams, 17 bigrams, 16 trigrams, and 15 4-grams. If the complete corpus consisted of this single sentence then the modified precisions would be $p_1 = .83$, $p_2 = .59$, $p_3 = .31$, and $p_4 = .2$. Each p_n is combined and can be weighted by specifying a weight w_n. In practice each p_n is generally assigned an equal weight. The the length of the hypothesis translation is 16 words. The closest reference translation has 18 words. The brevity penalty would be calculated as $e^{1-(18/16)} = .8825$. Thus the overall Bleu score would be

$$e^{1-(18/16)} * \exp(\log .83 + \log .59 + \log .31 + \log .2) = 0.193$$

Note that this calculation is on a single sentence, and Bleu is normally calculated over a corpus of sentences. Bleu does not correlate with human judgments on a per sentence basis, and anecdotally it is reported to be unreliable unless it is applied to a test set containing one hundred sentences or more.

6.1.3 Variations Allowed By BLEU

Given that all automatic evaluation techniques for MT need to model allowable variation in translation we should ask the following questions regarding how well Bleu models it: Is Bleu's use of multiple reference translations and n-gram-based matching sufficient to capture all allowable variation? Does it permit variations which are not valid? Given the shortcomings of its model, when should Bleu be applied? Can it be guaranteed to correlate with human judgments of translation quality?

We argue that Bleu's model of variation is weak, and that as a result it is unable to distinguish between translations of significantly different quality. In particular, Bleu

1-grams: American, **Florida**, **Miami**, Orejuela, **appeared**, as, being, **calm**, carry, escorted, **he**, him, in, led, **plane**, quite, seemed, take, that, **the**, **to**, **to**, to, **was** , was, **which**, while, **will**, would, **,, .**
2-grams: American plane, **Florida .**, **Miami ,**, Miami in, Orejuela appeared, Orejuela seemed, **appeared calm**, as he, being escorted, being led, calm as, calm while, carry him, escorted to, **he was**, him to, in Florida, led to, plane that, plane which, quite calm, seemed quite, take him, that was, that would, **the American**, the plane, **to Miami**, to carry, **to the**, was being, was led, was to, **which will**, while being, will take, would take, **, Florida**
3-grams: American plane that, American plane which, **Miami , Florida**, Miami in Florida, Orejuela appeared calm, Orejuela seemed quite, appeared calm as, appeared calm while, as he was, being escorted to, being led to, calm as he, calm while being, carry him to, escorted to the, he was being, he was led, him to Miami, in Florida ., led to the, plane that was, plane that would, plane which will, quite calm as, seemed quite calm, take him to, that was to, that would take, **the American plane**, the plane that, **to Miami ,**, to Miami in, to carry him, **to the American**, to the plane, was being led, was led to, was to carry, which will take, while being escorted, will take him, would take him, **, Florida .**
4-grams: American plane that was, American plane that would, American plane which will, **Miami , Florida .**, Miami in Florida ., Orejuela appeared calm as, Orejuela appeared calm while, Orejuela seemed quite calm, appeared calm as he, appeared calm while being, as he was being, as he was led, being escorted to the, being led to the, calm as he was, calm while being escorted, carry him to Miami, escorted to the plane, he was being led, he was led to, him to Miami ,, him to Miami in, led to the American, plane that was to, plane that would take, plane which will take, quite calm as he, seemed quite calm as, take him to Miami, that was to carry, that would take him, the American plane that, the American plane which, the plane that would, **to Miami , Florida**, to Miami in Florida, to carry him to, **to the American plane**, to the plane that, was being led to, was led to the, was to carry him, which will take him, while being escorted to, will take him to, would take him to

Table 6.2: The n-grams extracted from the reference translations, with matches from the hypothesis translation in bold

places no explicit constraints on the order in which matching n-grams occur, and it depends on having many reference translations to adequately capture variation in word choice. Because of these weakness in its model, a huge number of variant translations are assigned the same score. We show that for an average hypothesis translation there are millions of possible variants that would each receive a similar Bleu score. We argue that because the number of translations that score the same is so large, it is unlikely that all of them will be judged to be identical in quality by human annotators. This means that it is possible to have items which receive identical Bleu scores but are judged by humans to be worse. It is also therefore possible to have a *higher* Bleu score *without* any genuine improvement in translation quality. This undermines Bleu's use as stand-in for manual evaluation, since it cannot be guaranteed to correlate with human judgments of translation quality.

6.1.3.1 A weak model of phrase order

Bleu's model of allowable variation in phrase order is designed in such a way that it is less restrictive than WER, which assumes that one ordering is authoritative. Instead of matching words in a linear fashion, Bleu allows n-grams from the machine translated output to be matched against n-grams from any position in the reference translations. Bleu places no explicit restrictions on word order, and instead relies on the implicit restriction that a machine translated sentence must be worded similarly to one of the references in order to match longer sequences. This allows some phrases to occur in different positions without undue penalty. However, since Bleu lacks any explicit constraints on phrase order, it allows a tremendous amount of variations on a hypothesis translation while scoring them all equally.[2] The sheer number of possible permutations of a hypothesis show that Bleu admits far more orderings than what could reasonably be considered acceptable variation.

To get a sense of just how many possible translations would be scored identically under Bleu's model of phrase order, here we estimate a lower bound on the number of permutations of a hypothesis translation that will receive the same Bleu score. Bleu's only constraint on phrase order is implicit: the word order of a hypothesis translation much be similar to a reference translation in order for it to match higher order n-grams,

[2]Hovy and Ravichandran (2003) suggested strengthening Bleu's model of phrase movement by matching part-of-speech (POS) tag sequences against reference translations in addition to Bleu's n-gram matches. While this might reduce the amount of indistinguishable variation, it is infeasible since most MT systems do not produce POS tags as part of their output, and it is unclear whether POS taggers could accurately tag often disfluent MT output.

and receive a higher Bleu score. This constraint breaks down at points in a hypothesis translation which failed to match any higher order n-grams. Any two word sequence in a hypothesis that failed to match a bigram sequence from the reference translation will also fail to match a trigram sequence if extended by one word, and so on for all higher order n-grams. We define the point in between two words which failed to match a reference bigram as a *bigram mismatch site*. We can create variations in a hypothesis translation that will be equally scored by permuting phrases around these points.

Phrases that are bracketed by bigram mismatch sites can be freely permuted because reordering a hypothesis translation at these points *will not reduce the number of matching n-grams* and thus will not reduce the overall Bleu score. Here we denote bigram mismatches for the hypothesis translation given in Table 6.1 with vertical bars:

> Appeared calm | when | he was | taken | to the American plane | , | which will | to Miami , Florida .

We can randomly produce other hypothesis translations that have the same Bleu score but have a radically different word order. Because Bleu only takes order into account through rewarding matches of higher order n-grams, a hypothesis sentence may be freely permuted around these bigram mismatch sites and without reducing the Bleu score. Thus:

> which will | he was | , | when | taken | Appeared calm | to the American plane | to Miami , Florida .

receives an identical score to the hypothesis translation in Table 6.1.

We can use the number of bigram mismatch sites to estimate a lower bound on the number of similarly scored hypotheses in Bleu. If b is the number of bigram matches in a hypothesis translation, and k is its length, then there are

$$(k-b)! \tag{6.1}$$

possible ways to generate similarly scored items using only the words in the hypothesis translation.[3] Thus for the example hypothesis translation there are at least **40,320** different ways of permuting the sentence and receiving a similar Bleu score. The number of permutations varies with respect to sentence length and number of bigram mismatches. Therefore as a hypothesis translation approaches being an identical match to one of the reference translations, the amount of variance decreases significantly. So,

[3]Note that in some cases randomly permuting the sentence in this way may actually result in a greater number of n-gram matches; however, one would not expect random permutation to increase the human evaluation.

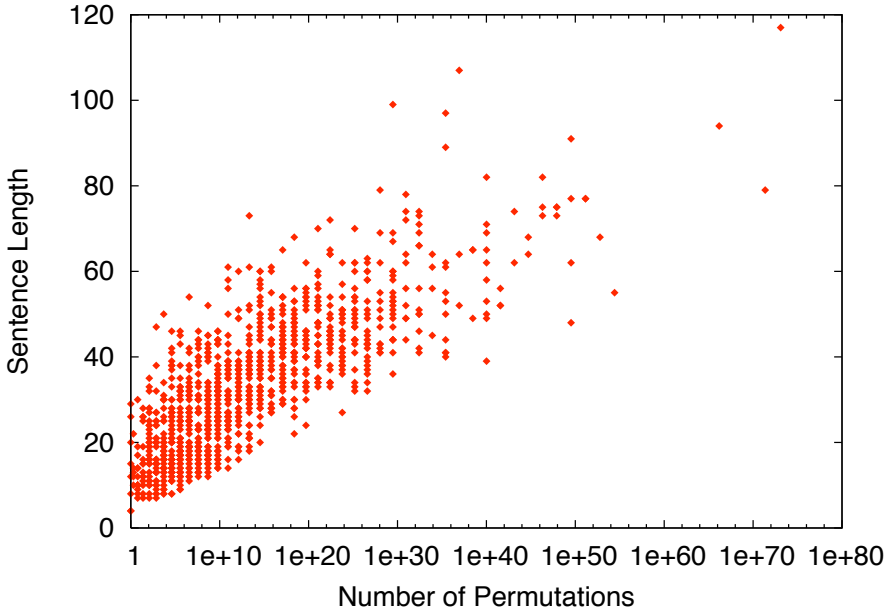


Figure 6.1: Scatterplot of the length of each translation against its number of possible permutations due to bigram mismatches for an entry in the 2005 NIST MT Eval

as translations improve, spurious variation goes down. However, at today's levels, the amount of variation that Bleu admits is unacceptably high. Figure 6.1 gives a scatterplot of each of the hypothesis translations produced by the second best Bleu system from the 2005 NIST MT Evaluation. The number of possible permutations for some translations is greater than $\mathbf{10^{73}}$.

Bleu's inability to distinguish between randomly generated variations in translation implies that it may not correlate with human judgments of translation quality in some cases. As the number of identically scored variants goes up, the likelihood that they would all be judged equally plausible goes down. This highlights the fact that Bleu is quite a crude measurement of translation quality.

6.1.3.2 A weak model of word choice

Another prominent factor which contributes to Bleu's crudeness is its model of allowable variation in word choice. Bleu is only able to handle synonyms and paraphrases if they are contained in the set of multiple reference translations. It does not have a specific mechanism for handling variations in word choice. Because it relies on the existence of multiple translation to capture such variation, the extent to which Bleu correctly recognizes hypothesis translations which are phrased differently depends on two things: the number of reference translations that are created, and the extent to

Source: *El artículo combate la discriminación y el trato desigual de los ciudadanos por las causas enumeradas en el mismo.*
Reference 1: The article combats discrimination and inequality in the treatment of citizens for the reasons listed therein.
Reference 2: The article aims to prevent discrimination against and unequal treatment of citizens on the grounds listed therein.
Reference 3: The reasons why the article fights against discrimination and the unequal treatment of citizens are listed in it.

Table 6.3: Bleu uses multiple reference translations in an attempt to capture allowable variation in translation.

which the reference translations differ from each other.

Table 6.3 illustrates how translations may be worded differently when different people produce translations for the same source text. For instance, *combate* was translated as *combats*, *flights against*, and *aims to prevent*, and *causas* was translated as *reasons* and *grounds*. These different reference translations capture some variation in word choice. While using multiple reference translations does make some headway towards allowing alternative word choice, it does not directly deal with variation in word choice. Because it is an indirect mechanism it will often fail to capture the full range of possibilities within a sentence. For instance, the multiple reference translations in Table 6.3 provide *listed* as the only translation of *enumeradas* when it could be equally validly translated as *enumerated*. The problem is made worse when reference translations are quite similar, as in Table 6.1. Because the references are so similar they miss out on some of the variation in word choice; they allow either *appeared* or *seemed* but exclude *looked* as a possibility.

Bleu's handling of alternative wordings is impaired not only if reference translations are overly similar to each other, but also if very few references are available. This is especially problematic because Bleu is most commonly used with only one reference translation. Zhang and Vogel (2004) showed that a test corpus for MT usually needs to have hundreds of sentences in order to have sufficient coverage in the source language. In rare cases, it is possible to create test suites containing 1,000 sentences of source language text and four or more human translations. However, such test sets are limited to well funded exercises like the NIST MT Evaluation Workshops (Lee and Przybocki, 2005). In most cases the cost of hiring a number of professional transla-

tors to translate hundreds of sentences to create a multi-reference test suite for Bleu is prohibitively high. The cost and labor involved undermines the primary advantage of adopting automatic evaluation metrics over performing manual evaluation. Therefore the MT community has access to very few test suites with multiple human references and those are limited to a small number of languages (Zhang et al., 2004). In order to test other languages most statistical machine translation research simply reserves a portion of the parallel corpus for use as a test set, and uses a single reference translation for each source sentence (Koehn and Monz, 2005, 2006; Callison-Burch et al., 2007).

Because it uses token identity to match words, Bleu does not allow *any* variation in word choice when it is used in conjunction with a single reference translation – not even simple morphological variations. Bleu is unable to distinguish between a hypothesis which leaves a source word untranslated, and a hypothesis which translates the source word using a synonym or paraphrase of the words in the reference. Bleu's weak model of acceptable variation in word choice therefore means that it can fail to distinguish between translations of obviously different quality, and therefore cannot be guaranteed to correspond to human judgments.

A number of researchers have proposed better models of variant word choice. Banerjee and Lavie (2005) provided a mechanism to match words in the machine translation which are synonyms of words in the reference in their Meteor metric. Meteor uses synonyms extracted from WordNet synsets (Miller, 1990). Owczarzak et al. (2006) and Zhou et al. (2006) tried to introduce more flexible matches into Bleu when using a single reference translation. They allowed machine translations to match paraphrases of the reference translations, and derived their paraphrases using our paraphrasing technique. Despite these advances, neither Meteor nor the enhancements to Bleu have been widely accepted. Papineni et al.'s definition of Bleu is therefore still the de facto standard for automatic evaluation in machine translation research.

The DARPA GALE program has recently moved away from using automatic evaluation metrics. The official evaluation methodology is a manual process wherein a human editor modifies a system's output until it is sufficiently close to a reference translation (NIST and LDC, 2007). The output is changed using the fewest number of edits, but still results in understandable English that contains all of the information that is in the reference translation. Since this is not an automatic metric, it does not have to model allowable variation in translation like Bleu does. People are able to judge what variations are allowable, and thus manual evaluation metrics are not subject to the criticism presented in this chapter.

6.1.4 Appropriate uses for BLEU

Bleu's model of allowable variation in translation is coarse, and in many cases it is unable to distinguish between translations of obvious different quality. Since Bleu assigns similar scores to translations of different quality, it is logical that a higher Bleu score may not necessarily be indicative of a genuine improvement in translation quality. Changes which fail to improve Bleu may be due to the fact that it is insensitive to such improvements. These comments do not apply solely to Bleu. Translation Error Rate (Snover et al., 2006), Meteor (Banerjee and Lavie, 2005), Precision and Recall (Melamed et al., 2003), and other such automatic metrics may also be affected to a greater or lesser degree because they are all quite rough measures of translation similarity, and have inexact models of allowable variation in translation.

What conclusions can we draw from this? Should we give up on using Bleu entirely? We think that the advantages of Bleu are still very strong; automatic evaluation metrics *are* inexpensive, and *do* allow many tasks to be performed that would otherwise be impossible. The important thing therefore is to recognize which uses of Bleu are appropriate and which uses are not. Appropriate uses for Bleu include tracking broad, incremental changes to a single system, comparing systems which employ similar translation strategies, and using Bleu as an objective function to optimize the values of parameters such as feature weights in log linear translation models, until a better metric has been proposed. Inappropriate uses for Bleu include comparing systems which employ radically different strategies, trying to detect improvements for aspects of translation that are not modeled well by Bleu, and monitoring improvements that occur infrequently within a test corpus.

6.2 Implications for evaluating translation quality improvements due to paraphrasing

Bleu's weakness are especially pertinent when we integrate paraphrases into the process of translation (as described in Chapter 5). In particular it is vital that allowable variation in word choice is correctly recognized when evaluating our approach. Because we paraphrase the source before translating it, there is a reasonable chance that the output of the machine translation system will be a paraphrase and will not be an exact match of the reference translation. This is illustrated in Figure 6.2, where the machine translation uses the phrase *ecological* rather than *environmentally-friendly*.

Source: *Estos autobuses son más* **respetuosos con el medio ambiente** *porque utilizan menos combustible por pasajero.*
Reference translation: These buses are more **environmentally-friendly** because they use less fuel per passenger.
Machine translation: These buses are more **ecological** because used less fuel per passenger.

Figure 6.2: Allowable variation in word choice poses a challenge for automatic evaluation metrics which compare machine translated sentences against reference human translations

While this alternative wording is perfectly valid, if an automatic evaluation metric does not have an adequate model of word choice then it will fail to recognize that *ecological* and *environmentally-friendly* are acceptable alternatives for each other. Because many of these instances arise in our translations, if we use an automatic metric to evaluate translation quality, it is critically important that it be able to recognize valid alternative wordings, and not strictly rely on the words in the reference translation. A problem arises when attempting to use Bleu to evaluate our translation improvements because the test sets that were available for our experiments (described in Section 7.1.1) did not have multiple translations, which rendered Bleu's already weak model of word choice totally ineffectual. Therefore we needed to take action to ensure that our evaluation was sensitive to the types of improvements that we were making. There are a number of options in this regard. We could:

- Create multiple reference translations for Bleu. This option was made difficult by a number of factors. Firstly, it is unclear how many reference translations would be required to capture the full range of possibilities (or indeed whether it is even possible to do so by increasing the number of reference translations). Secondly, because of this uncertainty the cost of hiring translators to create additional references for the test set was viewed as prohibitive.

- Use another evaluation metric such as Meteor. Despite having a better model of alternative word choice than Bleu, the fact that it uses WordNet for this model diminishes its usefulness. Since it is manually created, WordNet's range of synonyms is limited. Moreover, it contains relatively few paraphrases for multi-word expressions. Finally, WordNet provides no mechanism for determining in

which contexts its synonyms are valid substitutions.

- Conduct a manual evaluation. The problems associated with automatic metrics failing to recognize words and phrases that did not occur in reference translations can be sidestepped with human intervention. People can easily determine whether a particular phrase in the hypothesis translation is equivalent to a reference translation. Unlike WordNet they can take context into account.

Ultimately we opted to perform a manual evaluation of translation quality, which we tailored to target the particular phrases that we were interested in. Our methodology is described in the next section. The methodology in the next section is by no means the the only way to perform a manual evaluation of translation quality, and we make no claims that it is the best way. It is simply one way in which people can judge mismatches with the reference translations.

6.3 An alternative evaluation methodology

Because Bleu is potentially insensitive to the type of changes that we were making to the translations, we additionally gauged whether translation quality had improved by performing a manual evaluation. Manual evaluations usually assign values to each machine translated sentence along a scale (as given in Figure 4.1 on page 60). Instead of performing this sort of manual evaluation, we developed a *targeted* manual evaluation which allowed us to focus on a particular aspect of translation. Because we address a specific problem (coverage), we can focus on the relevant parts of each source sentence (words and phrases which were previously untranslatable), and solicit judgments about whether those parts were correctly translated after our change.

Our goal was to develop a methodology which allowed us to highlight translations of specific portions of the source sentence, and solicit judgments about whether those parts were translated accurately. Figure 6.3 shows a screenshot of the software that we used to conduct the targeted manual evaluation. In the example given in the figure, we were soliciting judgments about the translation of the Spanish word *enumeradas*, which is a word that was untranslatable prior to paraphrasing. We asked the annotator to indicate whether the phrase was correctly translated in the machine translated output. In different conditions, the phrase was translated as either *enumerated*, *as set out*, *which are listed*, or *that*. In two other conditions it was left untranslated. Rather than have the judge assign a subjective score to each sentence, we instead asked the judge

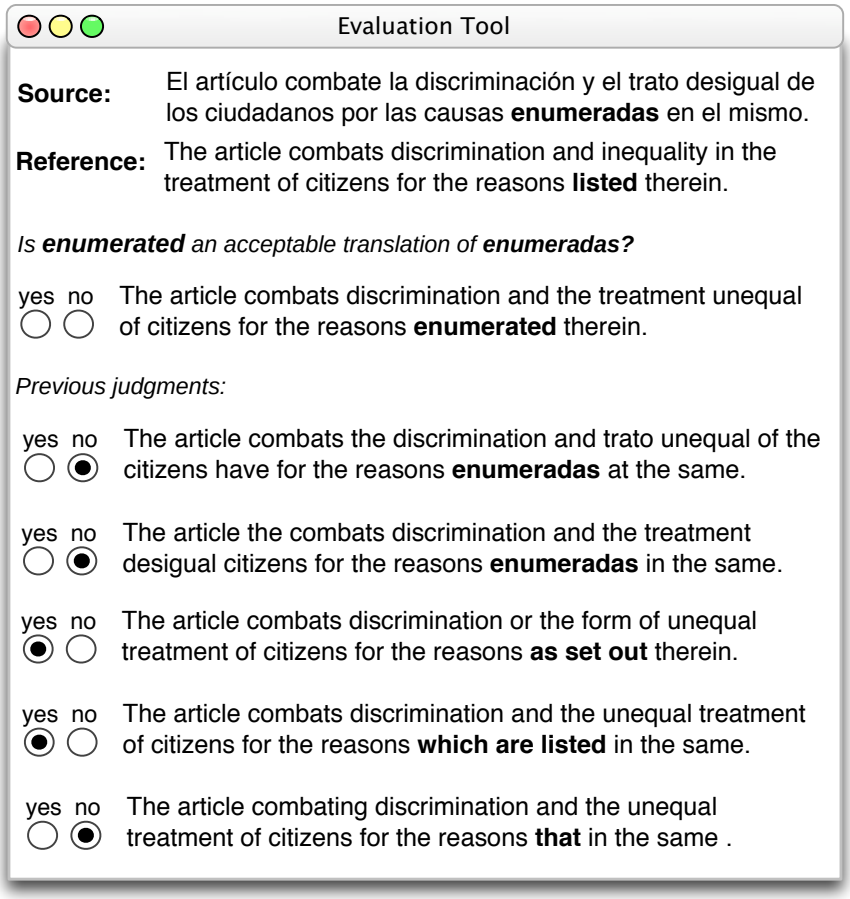


Figure 6.3: In the targeted manual evaluation judges were asked whether the translations of source phrases were accurate, highlighting the source phrase and the corresponding phrase in the reference and in the MT output.

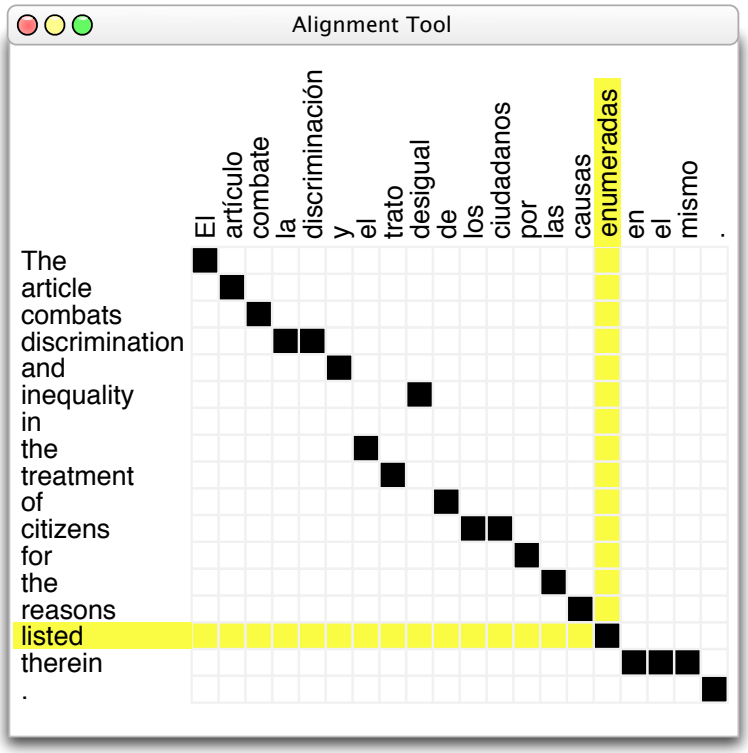


Figure 6.4: Bilingual individuals manually created word-level alignments between a number of sentence pairs in the test corpus, as a preprocessing step to our targeted manual evaluation.

to indicate whether each of the translations is acceptable, with a simple binary judgement. In addition to highlighting the source phrase and its corresponding translations in the machine translated output, we also highlighted the corresponding phrase in the reference translation to allow people who do not have a strong command of the source language to participate in the evaluation.

6.3.1 Correspondences between source and translations

In order to highlight the translations of the source phrase in the MT output and the reference translation, we need to know the correspondence between parts of the source sentence and its translations. Knowing this correspondence allows us to select a particular part of the source sentence and highlight the corresponding part of the machine translated output, thus focusing the judge's attention on the relevant part of the translation that we were interested in. We required correspondences to be specified for the MT output and the reference translations.

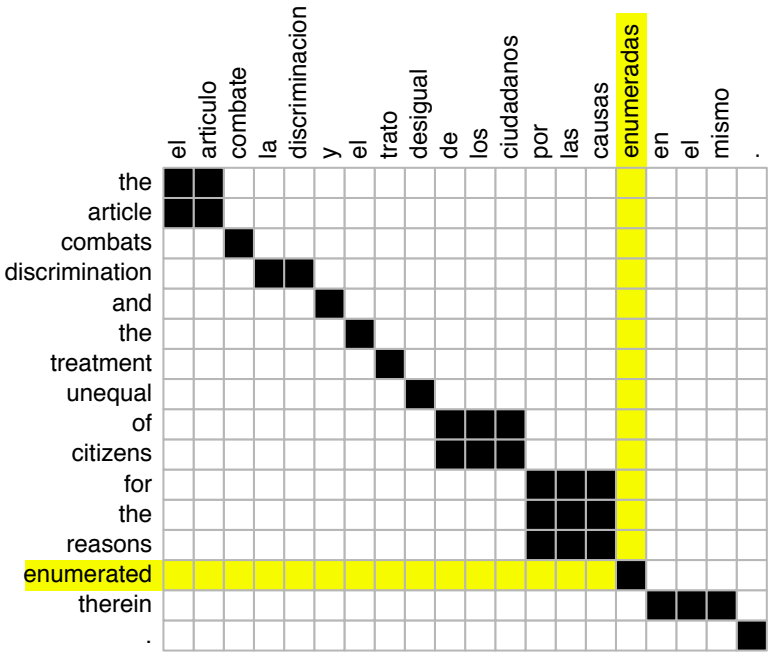


Figure 6.5: Pharaoh has a 'trace' option which reports which words in the source sentence give rise to which words in the machine translated output.

To specify the correspondences between the source sentence and the reference translations, we hired bilingual individuals to manually create word-level alignments. We implemented a graphical user interface, and specified a set of annotation guidelines that were similar to the Blinker project (Melamed, 1998). Figure 6.4 shows the alignment tool. The black squares indicate a correspondence between words. The annotators were also allowed to specify *probable* alignments for loose translations or larger phrase-to-phrase blocks. In order to make the annotators' job easier they were presented with the Viterbi word alignment predicted by the IBM Models, and edited that rather than starting from scratch. The average amount of time that it took for our annotators to create word alignments for a sentence pair was 3.5 minutes. While the creation of the word-level alignments was time consuming, it was a one-off preprocessing step. The data assembled during this stage could then be re-used for evaluating all of our different experimental conditions, and was therefore worth the effort.

To specify the correspondence between the machine translated output and the source sentence, we needed our machine translation system to report what words in the source were used to produce the different parts of its translation. Luckily, the Pharaoh decoder (Koehn, 2004) and the Moses decoder (Koehn et al., 2006) both provide a facility for doing this. For an input source sentence like the one given in Figure 6.3, the decoder

can produce a 'trace' of the output, which looks like

> The article |0 – 1| combats |2| discrimination |3 – 4| and |5| the |6| treatment |7| unequal |8| of citizens |9 – 11| for the reasons |12 – 14| enumerated |15| therein |16 – 18| . |19|

Each generated English phrase is now annotated with additional information, which indicates the indices of the Spanish words that gave rise to that English phrase. The trace allows us to extract correspondences between the source sentence and the translation, in the same way that the manual word-alignment did, as shown in Figure 6.5. Figure 6.6 shows the correspondences between the source sentence and the translations generated by different MT systems. The highlight portions show how we show the correspondences between the source phrase and the corresponding phrase in the MT output in Figure 6.3.

Note that Pharaoh only reports the correspondence between source words and the output translation at the level of granularity of the *phrases* that it selected, and is not necessarily as fine-grained as the word-level alignments that were manually created. In an ideal situation Pharaoh would produce a finer grained trace, which retained the word alignments between the phrases it uses. This would allow us to solicit judgments for very small units, or for larger chunks that spanned multiple units. However, for the evaluation that we conducted it was not an impairment. We were interested in soliciting judgments for source phrases that were previously untranslatable but which did have a translation after paraphrases. Therefore, we were interested in the particular phrases used by the decoder, so the correspondence that it reported was sufficient.

6.3.2 Reuse of judgments

In order to make the manual evaluation as quick and as painless as possible our evaluation software automatically re-used judgments if the translation of a source phrase for a given sentence was identical to a previous translation that had already been judged, or when it was identical to the corresponding segment in the reference human translation. This was partially inspired by the evaluation tool described by Niessen et al. (2000). They observed that one characteristic of MT research is that different versions of a translation system are tested *many times* on one distinct set of test sentences, and that often times the resulting translations differ only in a small number of words. Their tool facilitated fast manual evaluation of machine translation by using a database to store a record for an input sentence, which contained all its translations along with a

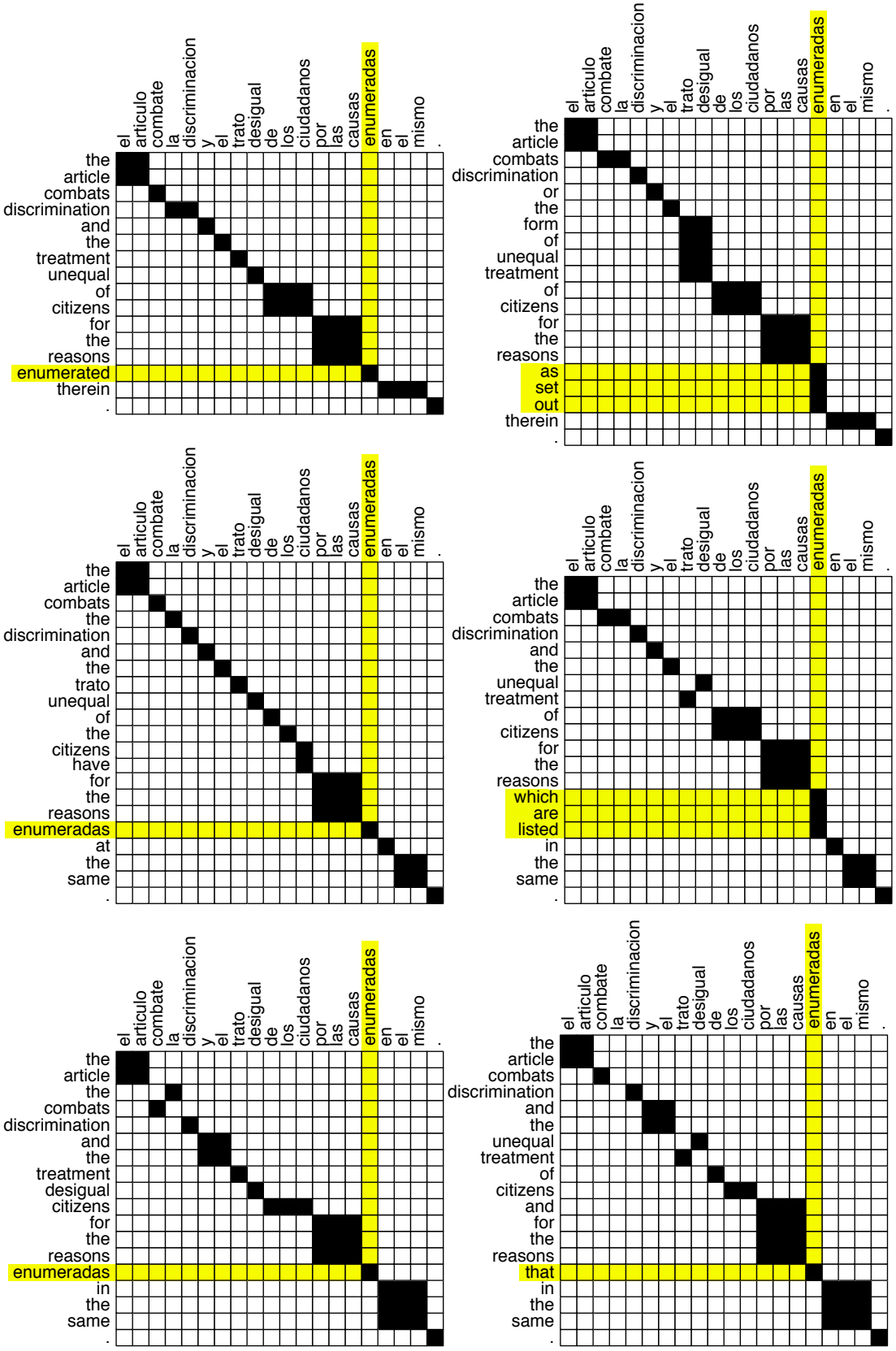


Figure 6.6: The ‘trace’ option can be applied to the translations produced by MT systems with different training conditions.

subjective sentence error rate (SSER) score for each translation. SSER is a ten point scale which range from 'nonsense' to 'perfect'. Storing scores in a database provided opportunities to automatically return the scores for translations which had already occurred, and to show judges the scores of previously judged translations if they differ from the new translation only by a few words. These reduced the number of judgments that had to be made, and helped to ensure that scores were assigned consistently over time.

We refine Niessen et al.'s methods by storing judgments about judgments for *subsentential units*. Rather than soliciting SSER scores about entire sentences, we ask judges to make a simpler yes/no judgment about whether the translation of a particular subphrase in the source sentence is correct. Decomposing the evaluation task into simpler judgments about smaller phrases gives several advantages over Niessen et al.'s use of SSER:

- Greater reuse of past judgments. Since the units in our database are smaller we get much greater re-use than Niessen et al. did by storing judgments for whole sentences.
- Simplification of the annotators' task. Asking about the translation of individual words or phrases is seemingly a simpler task than asking about the translation of whole sentences.
- The ability to define translation accuracy for a set of source phrases. This is described in the next section.

In the experiments described in the next chapter, we solicited 100 judgments for each system trained on each of the data sets described in Section 7.1.1. There were more than 3,000 items to be judged, but many of them were repeated. By caching past judgments in our database and only soliciting judgments for unique items, we sped the evaluation process considerably. The amount of re-use amounted to a semi-automation of the evaluation process. We believe that if judgments are retained over time, and built up over many evaluation cycles the amount of work involved in the manual evaluation is minimal, making it a potentially viable alternative to fully automatic evaluation.

6.3.3 Translation accuracy

In manual evaluations which solicit subjective judgments about entire sentences, as with Niessen et al.'s SSER or the LDC's adequacy and fluency scores, it is unclear

how to combine the scores. Can the scores be averaged across sentences, even if the sentences are different lengths? Since we solicit binary judgments of short phrases, we can combine our scores straightforwardly. We define translation accuracy for a particular system as the ratio of the number of translations that were judged to be correct to the total number of translations that were judged. We can further refine translation accuracy by restricting ourselves to judgments of a particular type of source phrase. For instance we could judge the translation accuracy of noun phrases or verb phrases in the source language, or we target specific improvements like word sense disambiguation and judge the accuracy of translation on polysemous words. In our experiments, we focused on source language phrases which were untranslatable prior to paraphrasing. By soliciting human judgments about whether our paraphrased translations were acceptable, we were able to get an indication of the accuracy of the newly-translated item.

It should be noted that the type of evaluation that we conducted is essentially focused on lexical choice, and that this is not the only aspect that determines translation quality. To judge other aspects of translation quality, like grammaticality, we would not only have to take into account word choice for particular phrases, but also that the composition of phrases leads to good word order, and that there were correct dependencies between words within the phrase and words outside of them (for things like agreement). If we had been investigating improvements to grammaticality instead of increasing coverage, then the focused manual evaluation would need to be formulated otherwise. However, evaluating lexical choice was well suited to the types of improvement that we were making to machine translation.

In the next chapter we describe the other aspects of our experimental design aside from those that pertain to evaluating translation quality. Section 7.1 outlines the data that we used to train our translation models and our paraphrase, and the different experimental conditions that we evaluated. Section 7.2 gives the results of our experiments.

Chapter 7

Translation Experiments

We designed a set of experiments to judge the extent to which paraphrasing can improve SMT. There are many factors to consider when designing such experiments. Not only do we have to choose an evaluation metric which is sensitive to our changes, we must also have appropriate conditions which highlight potential improvements and reveal problems. We attempted to ensure that our experimental setup was sensitive to potential improvements in translation quality. In particular we focused on the following elements of the experimental design:

- Since translation model coverage depends on the amount of available training data, we had several data conditions which used variously sized parallel corpora.
- Since a paraphrasing technique must be multilingual in order to be effectively applied to MT, we performed experiments in multiple languages.
- Since Bleu was potentially insensitive to our translation improvements, we also measured translation quality through a targeted manual evaluation.

The essence of our experiments was to train a baseline translation system for each of the training corpora, and to compare it against a *paraphrase system*. The paraphrase system's phrase table was expanded to include source language phrases that were untranslatable in the baseline system. The baseline and paraphrases systems were used to translate a set of held out test sentences, and the quality of their translations was analyzed. Since the baseline was a state-of-the-art phrase-based statistical machine translation system, it represented an extremely strong basis for comparison. Translation quality improvements therefore reflect a genuine advance in current technologies.

7.1 Experimental Design

The first half of this chapter is structured as follows: Section 7.1.1 describes data sets that were used in our experiments. Section 7.1.2 details the baseline SMT system and its behavior on unknown words and phrases. Section 7.1.3 describes the paraphrase system and how its phrase table was expanded to cover previously untranslatable words and phrases. Section 7.1.4 outlines the evaluation criteria that were used to evaluate our experiments. The results of our experiments are then presented in the second half of the chapter beginning in Section 7.2.

7.1.1 Data sets

In order to effectively apply a paraphrasing technique to machine translation it must be multilingual. Since we had already evaluated our paraphrasing technique on English, we choose two additional languages to apply it to. For these experiments we created paraphrases for Spanish and French, and applied them to the task of translating from from Spanish into English and from French into English. Our data requirements were as follows: We firstly needed data to train Spanish-English and French-English translation models. We additionally required data to create a Spanish paraphrase model, and data to create a French paraphrase model.

We drew data sets for both the translation models and for the paraphrase models from the publicly available Europarl multilingual parallel corpus (Koehn, 2005). We used the Spanish-English and French-English parallel corpora from Europarl to train our translation models. We created Spanish paraphrases using the Spanish-Danish, Spanish-Dutch, Spanish-Finnish, Spanish-French, Spanish-German, Spanish-Greek, Spanish-Italian, Spanish-Portuguese, and Spanish-Swedish parallel corpora. Crucially, we did not use any of the Spanish-English parallel corpus when training our paraphrase models. We created the French paraphrases in a similar fashion. The next two subsections give statistics about the size of the corpora used to train our translation models and our paraphrase models.

7.1.1.1 Data for translation models

Since the problem of coverage in statistical machine translation depends in large part on the amount of data that is used to train the translation model, we extracted variously sized portions of Spanish-English and French-English parallel corpora from the

Spanish-English Training Corpora

Sentence Pairs	Spanish Words	English Words	Spanish Vocab	English Vocab
10,000	217,778	211,312	14,335	10,073
20,000	437,047	422,511	20,679	13,849
40,000	868,490	839,506	28,844	18,718
80,000	1,737,247	1,676,621	39,723	24,968
160,000	3,461,169	3,329,369	53,896	33,340
320,000	6,897,347	6,627,292	71,999	44,055

French-English Training Corpora

Sentence Pairs	French Words	English Words	French Vocab	English Vocab
10,000	230,462	203,675	13,049	10,006
20,000	460,213	404,401	18,196	13,630
40,000	917,133	806,984	25,051	18,420
80,000	1,832,336	1,612,403	33,649	24,709
160,000	3,643,936	3,202,861	44,601	32,999
320,000	7,249,043	6,388,281	58,199	43,438

Table 7.1: The size of the parallel corpora used to create the Spanish-English and French-English translation models

Europarl corpus. We trained translation models using each of the data sets listed in Table 7.1. We tested how effective paraphrasing was at improving translation quality for translation models trained from all of these sets. Because models trained from smaller amounts of training data are prone to coverage problems, the expectation was that translation quality will improve more for smaller training set, and that there was less potential for improving translation quality for the larger training sets.

7.1.1.2 Data for paraphrase models

We generated paraphrases for Spanish and French phrases that were unseen in the Spanish-English and French-English parallel corpora used to train the translation models. To train our paraphrase models we used all of the parallel corpora from Europarl aside from the Spanish-English and French-English corpora. To generate our Spanish paraphrases we used bitexts between Spanish and Danish, Dutch, Finnish, French,

Training Data for Spanish Paraphrases

Corpus	Sentence Pairs	Spanish Words
Spanish-Danish	621,580	12,896,581
Spanish-Dutch	746,128	15,919,006
Spanish-Finnish	697,416	15,263,785
Spanish-French	683,899	14,303,567
Spanish-German	703,286	16,114,427
Spanish-Greek	526,705	10,708,470
Spanish-Italian	703,286	15,010,437
Spanish-Portuguese	725,446	15,529,006
Spanish-Swedish	700,296	14,986,388
Totals	6,108,042	130,731,667

Training Data for French Paraphrases

Corpus	Sentence Pairs	French Words
French-Danish	713,843	16,068,205
French-Dutch	714,275	16,103,807
French-Finnish	659,074	14,940,748
French-German	699,149	15,837,749
French-Greek	466,064	10,433,920
French-Italian	647,525	14,973,400
French-Portuguese	693,949	15,673,798
French-Spanish	697,416	15,665,082
French-Swedish	656,803	14,802,257
Totals	5,948,098	134,498,966

Table 7.2: The size of the parallel corpora used to create the Spanish and French paraphrase models

German, Greek, Italian, Portuguese, and Swedish. To generated French paraphrases we used bitexts between French and Danish, Dutch, Finnish, German, Greek, Italian, Portuguese, Spanish, and Swedish. Table 7.2 gives the total amount of data that was used to train our paraphrase models. For the Spanish paraphrase model we had more than 130 million words worth of data between Spanish and other languages. For the French paraphrase model we had over 134 million words.

Table 7.3 shows how many of the Spanish and French phrases that occur in the training sets in Table 7.2 have paraphrases. We enumerated all unique phrases of various lengths and extracted paraphrases for them. For instance in the Spanish training corpora there were a total of 100,000 unique words, half of which could be paraphrased as another word or phrase. For both the Spanish and the French we see that as the original phrases get longer the proportion of them that can be paraphrased goes down. This is natural since they are less frequent and often match with foreign phrases that occur only once, which makes them impossible to paraphrase using our method. A large fraction of shorter phrases can be paraphrased, with more than 10% of 4-grams having paraphrases.

7.1.2 Baseline system

The baseline system that we used was a state-of-the-art phrase-based statistical machine translation model, identical to the one described by Koehn et al. (2005b). The model employes the log linear formulation given in Equation 2.11. The baseline model had a total of eight feature functions: a language model probability, a phrase translation probability, a reverse phrase translation probability, a lexical translation probability, a reverse lexical translation probability, a word penalty, a phrase penalty, and a distortion cost (detailed below). To set the weights for each of the feature functions we used a development set containing 500 sentence pairs that was disjoint from the training and test sets to perform minimum error rate training (Och, 2003). The objective function used in minimum error rate training was Bleu (Papineni et al., 2002). We trained a baseline model using each of the 12 training corpora given in Table 7.1. The parameters were optimized separately for each of them.

7.1.2.1 Software

We used the following software to train the models and produce the translations: Giza++ was used to train the IBM word alignment models (Och and Ney, 2003), the

Spanish Phrases in Training Corpora

Length	Number with paraphrases	Total phrases of this length	Ratio
1	51,497	102,215	.50
2	636,832	1,556,288	.41
3	1,424,162	5,177,193	.28
4	1,493,992	8,752,552	.17
5	1,088,152	10,659,152	.10
6	684,554	11,178,734	.06
7	423,811	10,981,884	.04
8	276,671	10,490,694	.03
9	193,924	9,892,640	.02
10	144,717	9,265,646	.02

French Phrases in Training Corpora

Length	Number with paraphrases	Total phrases of this length	Ratio
1	33,991	80,189	.42
2	429,796	1,306,284	.33
3	963,315	4,488,376	.21
4	1,004,219	7,729,360	.13
5	753,571	9,750,256	.08
6	491,504	10,606,654	.05
7	301,983	10,725,019	.03
8	183,396	10,448,055	.02
9	114,740	9,984,187	.01
10	75,416	9,438,390	.01

Table 7.3: The number phrases in the training sets given in Table 7.2 for which paraphrases can be extracted. The table gives the total number of phrases of each length, the number of those for which a non-identical paraphrase could be found, and the ratio that this represents.

SRI language modeling toolkit was used to train the language model (Stolcke, 2002), the Pharaoh beam-search decoder was used to produce the translations after all of the model parameters had been set (Koehn, 2004), and we used the scripts included with Pharaoh for performing minimum error rate training and for extracting phrase tables from word alignments. All the resources that we used are in the public domain in order to allow others researchers to recreate our experiments.

7.1.2.2 Feature functions

Here are the details for the eight feature functions in the model:

- The language model was fixed for all experiments. It was a trigram model trained on the English side of the full parallel corpus that used Kneser-Ney smoothing (Kneser and Ney, 1995). The choice of language model is not especially relevant for our experiments, since data available to train language models is more freely available than for translation models, and generally not affected by problems associated with coverage.

- The phrase translation probability feature functions assigned a value based on the probability of translating between the source language phrases (Spanish or French) and the corresponding English phrase. The phrase translation probabilities $p(\bar{e}|\bar{f})$ and $p(\bar{f}|\bar{e})$ were calculated using the maximum likelihood estimator given in Equation 2.7 by counting the co-occurrence of phrases which had been extracted from the word-aligned parallel corpora (as described in Section 2.2.2).

- The heuristics used to extract phrases are inexact, and occasionally align phrases erroneously. Because these events are infrequent and because the phrase translation probability is calculated using maximum likelihood estimation, $p(\bar{e}|\bar{f})$ and $p(\bar{f}|\bar{e})$ can be falsely high. It is common practice to offset these probabilities with lexical weight feature functions $lex(\bar{e}|\bar{f})$ and $lex(\bar{f}|\bar{e})$. The lexical weight is low if the words that comprise $\bar{f}$ are not good translations of the words in $\bar{e}$. The lexical weight feature functions were calculated as described by Koehn et al. (2003).

- The word and phrase penalty feature functions each add a constant factor (ω and π, respectively) for each word or phrase generated. The model prefers shorter translations when the weight of the word penalty feature function (ω) is positive,

aprobación

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
approval	0.64	0.78	0.13	0.44	2.718
discharge	0.17	0.09	0.63	0.18	2.718
passing	0.05	1.00	0.01	0.16	2.718
adoption	0.05	0.25	0.03	0.20	2.718

de

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
of	0.71	0.88	0.39	0.69	2.718
for	0.05	0.22	0.04	0.24	2.718
from	0.01	0.55	0.02	0.36	2.718
in	0.07	0.12	0.05	0.14	2.718
on	0.03	0.20	0.03	0.17	2.718

de la

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
of the	0.52	0.35	0.24	0.26	2.718
the	0.14	0.01	0.63	0.15	2.718
from the	0.03	0.38	0.01	0.13	2.718
of	0.05	0.00	0.39	0.05	2.718
in the	0.05	0.06	0.03	0.05	2.718

en

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
in	0.73	0.71	0.43	0.55	2.718
on	0.07	0.21	0.07	0.18	2.718
at	0.04	0.49	0.04	0.22	2.718
onto	0.01	0.78	0.02	0.35	2.718
to in	0.01	1.00	0.12	0.55	2.718

en favor

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
in favour	0.80	0.27	0.12	0.17	2.718
for	0.20	0.01	0.04	0.01	2.718

favor

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
favour	0.90	0.75	0.28	0.31	2.718
in favour	0.10	0.03	0.06	0.16	2.718

la

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
the	0.01	0.05	0.17	0.37	2.718
to the	0.01	0.12	0.02	0.18	2.718
's	0.01	0.01	0.06	0.37	2.718
of the	0.01	0.04	0.05	0.37	2.718
is the	0.01	0.33	0.01	0.37	2.718

la aprobación

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
the approval	0.57	0.66	0.08	0.16	2.718
the discharge	0.28	0.28	0.40	0.06	2.718
the passing	0.14	1.00	0.01	0.06	2.718

votar

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
vote	0.69	0.09	0.35	0.10	2.718
the vote	0.08	0.02	0.04	0.10	2.718
vote in favour	0.08	0.17	0.01	0.05	2.718
vote will be in favour	0.08	1.00	0.01	0.03	2.718

votar en

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
vote in	0.50	0.20	0.11	0.03	2.718
vote on	0.50	0.25	0.01	0.01	2.718

voto

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
vote	0.70	0.07	0.25	0.05	2.718
voting	0.10	0.10	0.12	0.08	2.718
favour	0.10	0.08	0.06	0.02	2.718
vote on this subject	0.10	1.00	0.01	0.01	2.718

voy a votar

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty
i shall vote	0.50	1.00	0.01	0.01	2.718

Table 7.4: Example phrase table entries for the baseline Spanish-English system trained on 10,000 sentence pairs

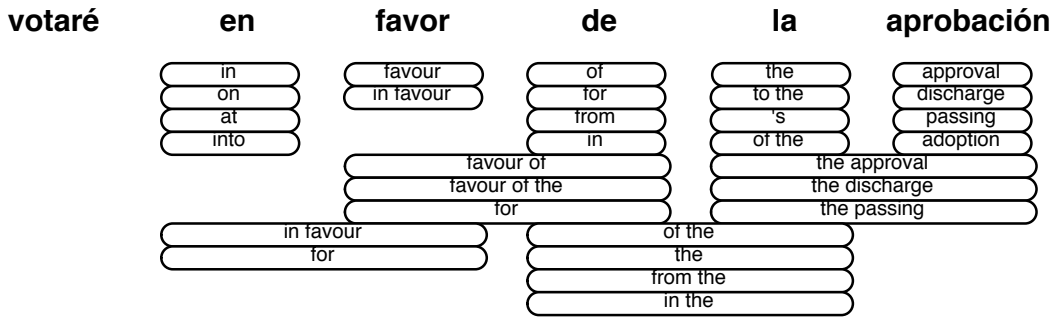


Figure 7.1: The decoder for the baseline system has translation options only for those words which have phrases that occur in the phrase table. In this case there are no translations for the source word *votaré*.

and longer translation when the weight is negative. The model prefers translations which are composed of a smaller number of long phrases when the weight of the phrase penalty feature function (π) is positive, and a greater number of short phrases when it is negative.

- The distortion cost adds a factor δ^n for phrase movements measured in a distance of n words. If the weight of the distortion feature function is positive, then translations which contain reordering are penalized exponentially with respect to the distance of the movement. Since this re-ordering model does not take the identity of the phrase into account, paraphrasing does not affect it at all. Lexicalized re-ordering models (Tillmann, 2004; Tillmann and Zhang, 2005; Koehn et al., 2005a) would require adaptation similar to the translation probability feature function.

7.1.2.3 Phrase Table

The baseline phrase table was created in the standard way by first assigning Viterbi word alignments for each sentence pair in the parallel corpus using the IBM Models, and then extracting phrase pairs from the word alignments (as described in Section 2.2.2). The phrase table contained these phrase pairs and their associated probabilities. Table 7.4 shows some of the entries that were contained in the phrase table for the baseline model which was trained on 10,000 Spanish-English sentence pairs. Section 7.2.2 discusses how much larger the French-English and Spanish-English phrase tables become after they are expanded using paraphrases.

7.1.2.4 Behavior on unseen words and phrases

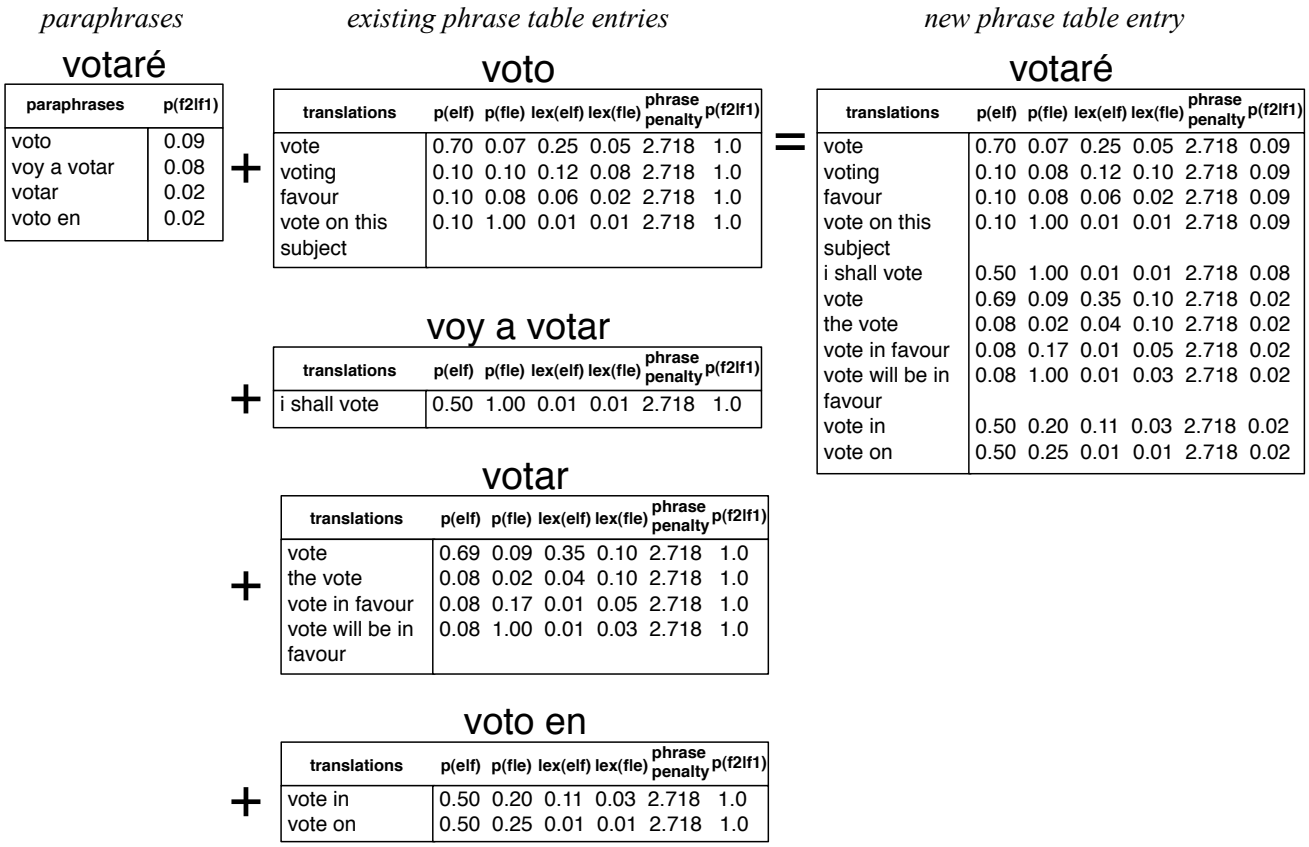
The decoder retrieves translations of each subphrase in an input sentence. It uses these as the translation options during its search for the best translation (as described in Section 2.2.3). Figure 7.1 shows the translation options for the Spanish sentence "*Votaré en favor de la aprobación.*" A word cannot be translate when it doesn't have any entries in the phrase table, as with *votaré*. The behavor of our baseline system was to reproduce the source word in the translated output. This is the default behavior for most systems, as noted in Section 5.2. When the baseline system encountered an unknown phrase, it attempts to translate each of its subphrases.

7.1.3 Paraphrase system

The paraphrase system differed from the baseline system in two ways: Its phrase table was expanded with paraphrases and it included a paraphrase probability feature function. We expanded each baseline phrase table by enumerating all words and phrases in the source language (French or Spanish) sentences in the test set and checking them against the baseline phrase table. For each word and phrase that was not in the baseline phrase table, we generated a list of its paraphrases. For each of the paraphrases of the unknown item, we checked whether it had any entries in the baseline phrase table. If the translations of one or more paraphrases were in the baseline phrase table we created a new entry for the unknown item with the translations of its paraphrases. The resulting phrase tables were used in the paraphrase systems. Each of the expanded phrase tables contained all of the entries from the baseline phrase tables, plus the additional entries created through paraphrasing.

7.1.3.1 Expanded phrase table

Figure 7.2 gives an example of how the phrase table for the paraphrase system was expanded to include an entry for the unknown source word *votaré*. Using the paraphrase model trained on the data listed in Table 7.2. The paraphrase model generates four potential paraphrases *voto*, *voy a votar*, *votar*, and *voto en*. These are present in the baseline phrase table that was trained on 10,000 sentence pairs (given in Table 7.4). Their translations and feature function values are combined into a new phrase table entry for *votaré*, as illustrated in Figure 7.2. This process can also be repeated for unknown phrases like *votaré en*.



votaré

paraphrases	p(f2lf1)
voto	0.09
voy a votar	0.08
votar	0.02
voto en	0.02

voto

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty	p(f2lf1)
vote	0.70	0.07	0.25	0.05	2.718	1.0
voting	0.10	0.10	0.12	0.08	2.718	1.0
favour	0.10	0.08	0.06	0.02	2.718	1.0
vote on this subject	0.10	1.00	0.01	0.01	2.718	1.0

voy a votar

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty	p(f2lf1)
i shall vote	0.50	1.00	0.01	0.01	2.718	1.0

votar

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty	p(f2lf1)
vote	0.69	0.09	0.35	0.10	2.718	1.0
the vote	0.08	0.02	0.04	0.10	2.718	1.0
vote in favour	0.08	0.17	0.01	0.05	2.718	1.0
vote will be in favour	0.08	1.00	0.01	0.03	2.718	1.0

voto en

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty	p(f2lf1)
vote in	0.50	0.20	0.11	0.03	2.718	1.0
vote on	0.50	0.25	0.01	0.01	2.718	1.0

votaré

translations	p(elf)	p(fle)	lex(elf)	lex(fle)	phrase penalty	p(f2lf1)
vote	0.70	0.07	0.25	0.05	2.718	0.09
voting	0.10	0.08	0.12	0.10	2.718	0.09
favour	0.10	0.08	0.06	0.02	2.718	0.09
vote on this subject	0.10	1.00	0.01	0.01	2.718	0.09
i shall vote	0.50	1.00	0.01	0.01	2.718	0.08
vote	0.69	0.09	0.35	0.10	2.718	0.02
the vote	0.08	0.02	0.04	0.10	2.718	0.02
vote in favour	0.08	0.17	0.01	0.05	2.718	0.02
vote will be in favour	0.08	1.00	0.01	0.03	2.718	0.02
vote in	0.50	0.20	0.11	0.03	2.718	0.02
vote on	0.50	0.25	0.01	0.01	2.718	0.02

Figure 7.2: A phrase table entry is added for *votaré* using the translations of its paraphrases. The feature function values of the paraphrases are also used, but offset by a paraphrase probability feature function since they may be inexact.

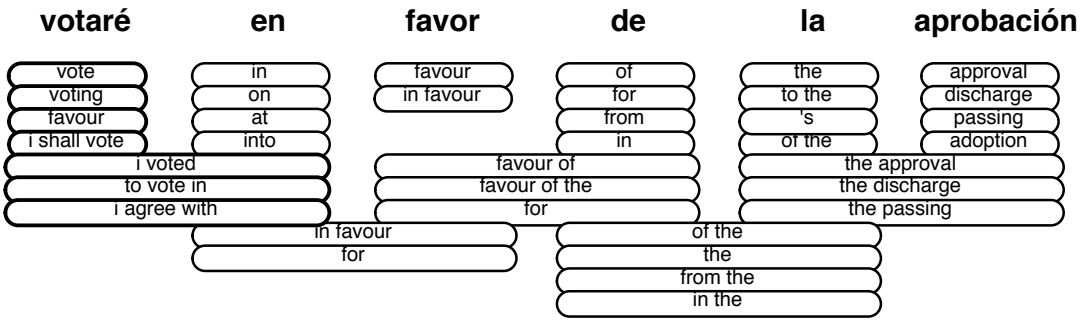


Figure 7.3: In the paraphrase system there are now translation options for *votaré* and and *votaré en* for which the decoder previously had no options.

7.1.3.2 Behavior on previously unseen words and phrases

The expanded phrase table of the paraphrase system results in different behavior for unknown words and phrases. Now the decoder has access to a wider range of translation options, as illustrated in Figure 7.3. For unknown words and phrases for which no paraphrases were found, or whose paraphrases did not occur in the baseline phrase table, the behavior of the paraphrase system is identical to the baseline system.

We did not generate paraphrases for names, numbers and foreign language words, since these items should not be translated. We manually created a list of the non-translating words from the test set and excluded them from being paraphrased.

7.1.3.3 Additional feature function

In addition to expanding the phrase table, we also augmented the paraphrase system by incorporating the paraphrase probability into an additional feature function that was not present in the baseline system, as described in Section 5.4.2. We calculated paraphrase probabilities using the definition given in Equation 3.6. This definition allowed us to assign improved paraphrase probabilities by calculating the probability using multiple parallel corpora. We omitted other improvements to the paraphrase probability described in Chapter 4, including word sense disambiguation and re-ranking paraphrases based on a language model probability. These were omitted simply as a matter of convenience and their inclusion might have resulted in further improvements to translation quality, beyond the results given in Chapter 7.2.

Just as we did in the baseline system, we performed minimum error rate training to set the weights of the nine feature functions (which consisted of the eight baseline feature functions plus the new one). The same development set that was used to set the

eight weights in the baseline system were used to set the nine weights in the paraphrase system.

Note that this additional feature function is not strictly necessary to address the problem of coverage. That is accomplished through the expansion of the phrase table. However, by integrating the paraphrase probability feature function, we are able to give the translation model additional information which it can use to choose the best translation. If a paraphrase had a very low probability, then it may not be a good choice to use its translations for the original phrase. The paraphrase probability feature function gives the model a means of assessing the relative goodness of the paraphrases. We experimented with the importance of the paraphrase probability by setting up a contrast model where the phrase table was expanded but this feature function was omitted. The results of this experiment are given in Section 7.2.1.

7.1.4 Evaluation criteria

We evaluated the efficacy of using paraphrases in three ways: by computing Bleu score, by measuring the increase in coverage when including paraphrases, and through a targeted manual evaluation to determine how many of the newly covered phrases were accurately translated. Here are the details for each of the three:

- The Bleu score was calculated using test sets containing 2,000 Spanish sentences and 2,000 French sentences, with a single reference translation into English for each sentence. The test sets were drawn from portions of the Europarl corpus that were disjoint from the training and development sets. They were previously used for a statistical machine translation shared task (Koehn and Monz, 2005).

- We measured coverage by enumerating all unique unigrams, bigrams, trigrams and 4-grams from the 2,000 sentence test sets, and calculating what percentage of those items had translations in the phrase tables created for each of the systems. By comparing the coverage of the baseline system against the coverage of the paraphrase system when their translation models were trained on the same parallel corpus, we could determine how much coverage had increased.

- For the targeted manual evaluation we created word-alignments for the first 150 Spanish-English sentence pairs in the test set, and for the first 250 French-English sentence pairs. We had monolingual judges assess the translation accuracy of parts of the MT output from the paraphrase system that were untrans-

latable in the baseline system. In doing so we were able to assess how often the newly covered phrases were accurately translated.

7.2 Results

Before giving summary statistics about translation quality we will first show that our proposed method does in fact result in improvements by presenting a number of example translations. Appendix B shows translations of Spanish sentences from the baseline and paraphrase systems for each of the six Spanish-English corpora. These example translations highlight cases where the baseline system reproduced Spanish words in its output because it failed to learn translations for them. In contrast the paraphrase system is frequently able to produce English output of these same words. For example, in the translations of the first sentence in Table B.1 the baseline system outputs the Spanish words *alerta*, *regreso*, *tentados* and *intergubernamentales*, and the paraphrase system translates them as *warning*, *return*, *temptation* and *intergovernmental*. All of these match words in the reference except for *temptation* which is rendered as *tempted* in the human translation. These improvements also apply to phrases. For instance, in the third example in Table B.2 the Spanish phrase *mejores prácticas* is translated as *practices in the best* by the baseline system and as *best practices* by the paraphrase system. Similarly, for the third example in Table B.3 the Spanish phrase *no podemos darnos el lujo de perder* is translated as *we cannot understand luxury of losing* by the baseline system and much more fluently as *we cannot afford to lose* by the paraphrase system.

While the translations presented in the tables suggest that quality has improved, one should never rely on a few examples as the sole evidence on improved translation quality since examples can be cherry-picked. Average system-wide metrics should also be used. Bleu can indicate whether a system's translations are getting closer to the reference translations when averaged over thousands of sentences. However, the examples given in Appendix B should make us think twice when interpreting Bleu scores, because many of the highlighted improvements do not exactly match their corresponding segments in the references. Table 7.5 shows examples where the baseline system's reproduction of the foreign text receives the same score as the paraphrase system's English translation. Because our system frequently does not match the single reference translation, Bleu may *underestimate* the actual improvements to translation quality which are made my our system. Nevertheless we report Bleu scores as a rough

REFERENCE	BASELINE	PARAPHRASE
tempted	tentados	temptation
I will vote	votaré	I shall vote
environmentally-friendly	repetuosos with the environment	ecological
to propose to you	proponerles	to suggest
initated	iniciados	started
presidencies	presidencias	presidency
to offer	to	to present
closer	reforzada	increased
examine	examinemos	look at
disagree	disentimos	do not agree
entrusted with the task	encomendado has the task	given the task
to remove	remover	to eliminate
finance	financiará	fund

Table 7.5: Examples of improvements over the baseline which are not fully recognized by Bleu because they fail to match the reference translation

indication of the trends in the behavior of our system, and use it to contrast different cases that we would not have the resources to evaluate manually.

7.2.1 Improved Bleu scores

We calculated Bleu scores over test sets consisting of 2,000 sentences. We take Bleu to be indicative of general trends in the behavior of the systems under different conditions, but do not take it as a definitive estimate of translation quality. We therefore evaluated several conditions using Bleu and later performed more targeted evaluations of translation quality. The conditions that we evaluated with Bleu were:

- The performance of the baseline system when its translation model was trained on various sized corpora
- The performance of the paraphrase system on the same data, when unknown words were paraphrased.
- The performance of the paraphrase system when unknown multi-word phrases were paraphrased.

	Spanish-English					
Corpus size	10k	20k	40k	80k	160k	320k
Baseline	22.6	25.0	26.5	26.5	28.7	**30.0**
Single word	23.1	25.2	26.6	**28.0**	**29.0**	**30.0**
Multi-word	**23.3**	**26.0**	**27.2**	**28.0**	28.8	29.7

Table 7.6: Bleu scores for the various sized Spanish-English training corpora, including baseline results without paraphrasing, results for only paraphrasing unknown words, and results for paraphrasing any unseen phrase. Corpus size is measured in sentences. Bold indicates best performance over all three conditions.

	French-English					
Corpus size	10k	20k	40k	80k	160k	320k
Baseline	21.9	24.3	26.3	27.8	28.8	29.5
Single word	22.7	24.2	26.9	27.7	28.9	**29.8**
Multi-word	**23.7**	**25.1**	**27.1**	**28.5**	**29.1**	**29.8**

Table 7.7: Bleu scores for the various sized French-English training corpora, including baseline results without paraphrasing, results for only paraphrasing unknown words, and results for paraphrasing any unseen phrase. Corpus size is measured in sentences. Bold indicates best performance over all three conditions.

- The paraphrase system when the paraphrase probability was included as a feature function and when it was excluded.

Table 7.6 gives the Bleu scores for Spanish-English translation with baseline system, with unknown single words paraphrased, and for unknown multi-word phrases paraphrased. Table 7.7 gives the same for French-English translation. We were able to measure a translation improvement for all sizes of training corpora, under both the single word and multi-word conditions, except for the largest Spanish-English corpus. For the single word condition, it would have been surprising if we had seen a decrease in Bleu score. Because we are translating words that were previously untranslatable it would be unlikely that we could do any worse. In the worst case we would be replacing one word that did not occur in the reference translation with another, and thus have no effect on Bleu.

	Single word paraphrases			Multi-word paraphrases		
Feature Function	10k	20k	40k	10k	20k	40k
Translation Model	0.044	0.026	0.011	0.033	0.024	0.085
Lexical Weighting	0.027	0.018	0.001	0.027	0.031	-0.009
Reverse Translation Model	-0.003	0.033	0.014	0.047	0.142	0.071
Reverse Lexical Weighting	0.030	0.055	0.015	0.049	0.048	0.079
Phrase Penalty	-0.098	0.001	-0.010	-0.197	0.032	0.007
Paraphrase Probability	0.616	0.641	0.877	0.273	0.220	0.295
Distortion Cost	0.043	0.038	0.010	0.035	0.092	0.062
Language Model	0.092	0.078	0.024	0.097	0.124	0.137
Word Penalty	-0.048	-0.111	-0.039	-0.242	-0.286	-0.254

Table 7.8: The weights assigned to each of the feature functions after minimum error rate training. The paraphrase probability feature receives the highest value on all occasions

More interesting is the fact that by paraphrasing unseen multi-word units we get an increase in quality above and beyond the single word paraphrases. These multi-word units may not have been observed in the training data as a unit, but each of the component words may have been. In this case translating a paraphrase would not be guaranteed to received an improved or identical Bleu score, as in the single word case. Thus the improved Bleu score is notable.

The importance of the paraphrase probability feature function

In addition to expanding our phrase table by creating additional entries using paraphrasing, we incorporated a feature function into our model that was not present in the baseline system. We investigated the importance of the paraphrase probability feature function by examining the weight assigned to it in minimum error rate training (MERT), and by repeating the experiments summarized in Tables 7.6 and 7.7 and dropping the paraphrase probability feature function. For the latter, we built models which had expanded phrase tables, but which did not include the paraphrase probability feature function. We re-ran MERT, decoded the test sentences, and evaluated the resulting translations with Bleu.

Table 7.8 gives the feature weights assigned by MERT for three of the Spanish-English training corpora for both the single-word and the multi-word paraphrase con-

	Spanish-English					
Corpus size	10k	20k	40k	80k	160k	320k
Single word w/o ff	23.0	25.1	26.7	28.0	**29.0**	29.9
Multi-word w/o ff	20.6	22.6	21.9	24.0	25.4	27.5

Table 7.9: Bleu scores for the various sized Spanish-English training corpora, when the paraphrase feature function *is not* included. Bold indicates best performance over all three conditions.

	French-English					
Corpus size	10k	20k	40k	80k	160k	320k
Single word w/o ff	22.5	24.1	26.0	27.6	28.8	29.6
Multi-word w/o ff	19.7	22.1	24.3	25.6	26.0	28.1

Table 7.10: Bleu scores for the various sized French-English training corpora, when the paraphrase feature function *is not* included.

ditions. In all cases the feature function incorporating the paraphrase probability received the largest weight, indicating that it played a significant role in determining which translation was produced by the decoder. However, the weight alone is not sufficient evidence that the feature function is useful.

Tables 7.10 and 7.9 show definitively that the paraphrase probability into the model's feature functions plays a critical role. Without it, the multi-word paraphrases harm translation performance when compared to the baseline.

7.2.2 Increased coverage

In addition to calculating Bleu scores, we also calculated how much *coverage* had increased, since it is what we focused on with our paraphrase system. When only a very small parallel corpus is available for training, the baseline system learns translations for very few phrases in a test set. We measured how much coverage increased by recording how many of the unique phrases in the test set had translations in the translation model. Note by *unique phrases* we refer to types not tokens.

In the 2,000 sentences that comprise the Spanish portion of the Europarl test set there are 7,331 unique unigrams, 28,890 unique bigrams, 44,194 unique trigrams, and unique 48,259 4-grams. Table 7.11 gives the percentage of these which have transla-

Size	1-gram	2-gram	3-gram	4-gram
10k	48%	25%	10%	3%
20k	60%	35%	15%	6%
40k	71%	45%	22%	9%
80k	80%	55%	29%	12%
160k	86%	64%	37%	17%
320k	91%	71%	45%	22%

Table 7.11: The percent of the unique test set phrases which have translations in each of the Spanish-English training corpora prior to paraphrasing

Size	1-gram	2-gram	3-gram	4-gram
10k	90%	67%	37%	16%
20k	90%	69%	39%	17%
40k	91%	71%	41%	18%
80k	92%	73%	44%	20%
160k	92%	75%	46%	22%
320k	93%	77%	50%	25%

Table 7.12: The percent of the unique test set phrases which have translations in each of the Spanish-English training corpora after paraphrasing

tions in the baseline system's phrase table for each training corpus size. In contrast after expanding the phrase table using the translations of paraphrases, the coverage of the unique test set phrases goes up dramatically (shown in Table 7.12). For the training corpus with 10,000 sentence pairs and roughly 200,000 words of text in each language, the coverage goes up from less than 50% of the vocabulary items being covered to 90%. The coverage of unique 4-grams jumps from 3% to 16% – a level reached only after observing more than 100,000 sentence pairs, or roughly three million words of text, without using paraphrases.

7.2.3 Accuracy of translation

To measure the accuracy of the newly translated items we performed a manual evaluation. Our evaluation followed the methodology described in Section 6.3. We judged the translations of 100 words and phrases produced by the paraphrase system which

	Spanish-English					
Corpus size	10k	20k	40k	80k	160k	320k
Single word	48%	53%	57%	67%*	33%*	50%*
Multi-word	64%	65%	66%	71%	76%*	71%*

Table 7.13: Percent of time that the translation of a Spanish paraphrase was judged to retain the same meaning as the corresponding phrase in the gold standard. Starred items had fewer than 100 judgments and should not be taken as reliable estimates.

	French-English					
Corpus size	10k	20k	40k	80k	160k	320k
Single word	54%	49%	45%	50%	39%*	21%*
Multi-word	60%	67%	63%	58%	65%	42%*

Table 7.14: Percent of time that the translation of a French paraphrase was judged to retain the same meaning as the corresponding phrase in the gold standard. Starred items had fewer than 100 judgments and should not be taken as reliable estimates.

were untranslatable by the baseline system.[1] Tables 7.13 and 7.14 give the percentage of time that each of the translations of paraphrases were judged to have the same meaning as the corresponding phrase in the reference translation. In the case of the translations of single word paraphrases for the Spanish accuracy ranged from just below 50% to just below 70%. This number is impressive in light of the fact that none of those items are correctly translated in the baseline model, which simply inserts the foreign language word. As with the Bleu scores, the translations of multi-word paraphrases were judged to be more accurate than the translations of single word paraphrases.

In performing the manual evaluation we were additionally able to determine how often Bleu was capable of measuring an actual improvement in translation. For those items judged to have the same meaning as the gold standard phrases we could track how many would have contributed to a higher Bleu score (that is, which of them were exactly the same as the reference translation phrase, or had some words in common with the reference translation phrase). By counting how often a correct phrase would have contributed to an increased Bleu score, and how often it would fail to increase the

[1]Note that for the larger training corpora fewer than 100 paraphrases occurred in the set of word-aligned data that we created for the manual evaluation (as described in Section 6.3.1). We created word alignments for 150 Spanish-English sentence pairs and 250 French-English sentence pairs.

	Spanish-English					
Corpus size	10k	20k	40k	80k	160k	320k
Single word	88%	97%	93%	92%	95%	96%
Multi-word	87%	96%	94%	93%	91%	95%
Baseline	82%	89%	84%	84%	92%	96%

Table 7.15: Percent of time that the parts of the translations which were not paraphrased were judged to be accurately translated for the Spanish-English translations.

	French-English					
Corpus size	10k	20k	40k	80k	160k	320k
Single word	93%	92%	91%	91%	92%	94%
Multi-word	94%	91%	91%	89%	92%	94%
Baseline	90%	87%	88%	91%	92%	94%

Table 7.16: Percent of time that the parts of the translations which were not paraphrased were judged to be accurately translated for the French-English translations.

Bleu score we were able to determine with what frequency Bleu was sensitive to our improvements. We found that Bleu was insensitive to our translation improvements between 60-75% of the time, thus re-inforcing our belief that it is not an appropriate measure for translation improvements of this sort.

Accuracy of translation for non-paraphrased phrases

It is theoretically possible that the quality of the non-paraphrased segments got worse and went undetected, since our manual evaluation focused only on the paraphrased segments. Therefore, as a sanity check, we also performed an evaluation for portions of the translations which were not paraphrased prior to translation. We compared the accuracy of these segments against the accuracy of randomly selected segments from the baseline (where none of the phrases were paraphrased).

Tables 7.15 and 7.16 give the translation accuracy of segments from the baseline systems and of segments in the paraphrase systems which were not paraphrased. The paraphrase systems performed at least as well, or better than the baseline systems even for non-paraphrased segments. Thus we can definitively say that it produced better overall translations than the state-of-the-art baseline.

7.3 Discussion

As our experiments demonstrate paraphrases can be used to improve the quality of statistical machine translation addressing some of the problems associated with coverage. Whereas standard systems rely on having observed a particular word or phrase in the training set in order to produce a translation of it, we are no longer tied to having seen every word in advance. We can exploit knowledge that is external to the translation model and use that in the process of translation. This method is particularly pertinent to small data conditions, which are plagued by sparse data problems. In effect, paraphrases introduce some amount of *generalization* into statistical machine translation.

Our paraphrasing method is by no means the only technique which could be used to generate paraphrases to improve translation quality. However, it does have a number of features which make it particularly well-suited to the task. In particular our experiments show that its probabilistic formulations helps it to guide the search for the best translation when paraphrases are integrated.

In the next chapter we review the contributions of this thesis to paraphrasing and translation, and discuss future directions.

Chapter 8

Conclusions and Future Directions

Expressing ideas using other words is the crux of both paraphrasing and translation. They differ in that translation uses words in another language whereas paraphrasing uses words in a single language. Statistical models of translation have become commonplace due to the wide availability of bilingual corpora which pair sentences in one language with their equivalents in another language. Corpora containing pairs of equivalent sentences in the same language are comparatively rare, which has stymied the investigation of statistical models of paraphrasing. A number of research efforts have focused on drawing pairs of similar English sentences from comparable corpora, or on the miniscule amount of data available in multiple English translations of the same foreign text. In this thesis we introduce the powerful idea that paraphrases can be identified by pivoting through corresponding phrases in a foreign language. This obviates the need for corpora containing pairs of paraphrases. This allows us to use abundant bilingual parallel corpora to train statistical models of paraphrasing, and to draw on alignment techniques and other research in the statistical machine translation literature. One of the major contributions of this thesis is a probabilistic interpretation of paraphrasing, which falls naturally out of the fact that we employ the data and probabilities from statistical translation.

8.1 Conclusions

We have shown both empirically and through numerous illustrative examples that the quality of paraphrases extracted from parallel corpora is very high. We defined a baseline paraphrase probability based on phrase translation probabilities, and incrementally refined it to address factors that affect paraphrase quality. Refinements included the in-

tegration of multiple parallel corpora (over different languages) to reduce the effect of systematic misalignments in one language, word sense controls to partition polysemous words in training data into classes with the same meaning, and the addition of a language model to ensure more fluent output when a paraphrase is substituted into a new sentence. We developed a rigorous evaluation methodology for paraphrases, which involves substituting phrases with their paraphrases and having people judge whether the resulting sentences retain the meaning of the original and remain grammatical. Our baseline system produced paraphrases that met this strict definition of accuracy 50% of the time, and which had the correct meaning 65% of the time. Refinements increased the accuracy to 62%, with more than 70% of items having the correct meaning. Further experiments achieved an accuracy of 75% and a correct meaning 85% of the time with manual gold standard alignments, suggesting that our paraphrasing technique will improve alongside statistical alignment techniques.

In addition to showing that paraphrases can be extracted from the data that is normally used to train statistical translation systems, we have further shown that paraphrases can be used to improve the quality of statistical machine translation. Beyond its high accuracy, our paraphrasing technique is ideally suited for integration into phrase-based statistical machine translation for a number of other reasons. It is easily applied to many languages. It has a probabilistic formulation. It is capable of generating paraphrases for both words and phrases. A significant problem with current statistical translation systems is that they are slavishly tied to the words and phrase that occur in their training data. If a word does not occur in the data then most systems are unable to translate it. If a phrase does not occur in the training data then it is less likely to be translated correctly. This problem can be characterized as one of coverage. Our experiments have shown that coverage can be significantly increased by paraphrasing unknown words and phrases and using the translations of their paraphrases. For small data sets paraphrasing increases coverage to levels reached by the baseline approach only after ten times as much data has used. Our experiments measured the accuracy of newly translated items both through a human evaluation, and with the Bleu automatic evaluation metric. The human judgments indicated that the previously untranslatable items were correctly translated up to 70% of the time.

Despite these marked improvements, the Bleu metric vastly underestimated the quality of our system. We analyzed Bleu's behavior, and showed that its poor model of allowable variation in translation means that it cannot be guaranteed to correspond to human judgments of translation quality. Bleu is incapable of correctly scoring trans-

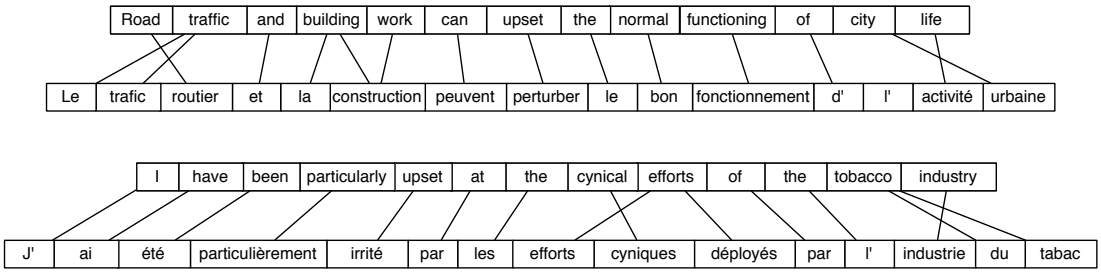


Figure 8.1: Current phrase-based approaches to statistical machine translation represent phrases as sequences of fully inflected words

lation improvements like ours, which frequently deviate from the reference translation but which nevertheless are correct translations. Its failures are by no means limited to our system. There is a huge range of possible improvements to translation quality that Bleu will be completely insensitive to. Because of this fact, and because Bleu is so prevalent in conference papers and research workshops, the field as a whole needs to reexamine its reliance on the metric.

8.2 Future directions

One of the reasons that statistical machine translation is improved when paraphrases are introduced is the fact that they introduce some measure of generalization. Current phrase-based models essentially *memorize* the translations of words and phrases from the training data, but are unable to *generalize* at all. Paraphrases allow them to learn the translations of words and phrases which are not present in the training data, by introducing *external knowledge*. However, there is a considerable amount of information *within* the training data that phrase-based statistical translation models fail to learn: they fail to learn simple linguistic facts like that a language's word order is subject-object-verb or that adjective-noun alternation occurs between languages. They are unable to use linguistic context to generate grammatical output (for instance, which uses the correct grammatical gender or case). These failures are largely due to the fact that phrase-based systems represent phrases as sequences of fully-inflected words, but are otherwise devoid of linguistic detail.

Instead of representing phrases only as sequences of words (as illustrated by Figure 8.1) it should be possible to introduce a more sophisticated representation for phrases. This is the idea of Factored Translation Models, which we began work on at a summer workshop at Johns Hopkins University (Koehn et al., 2006). Factored Translation

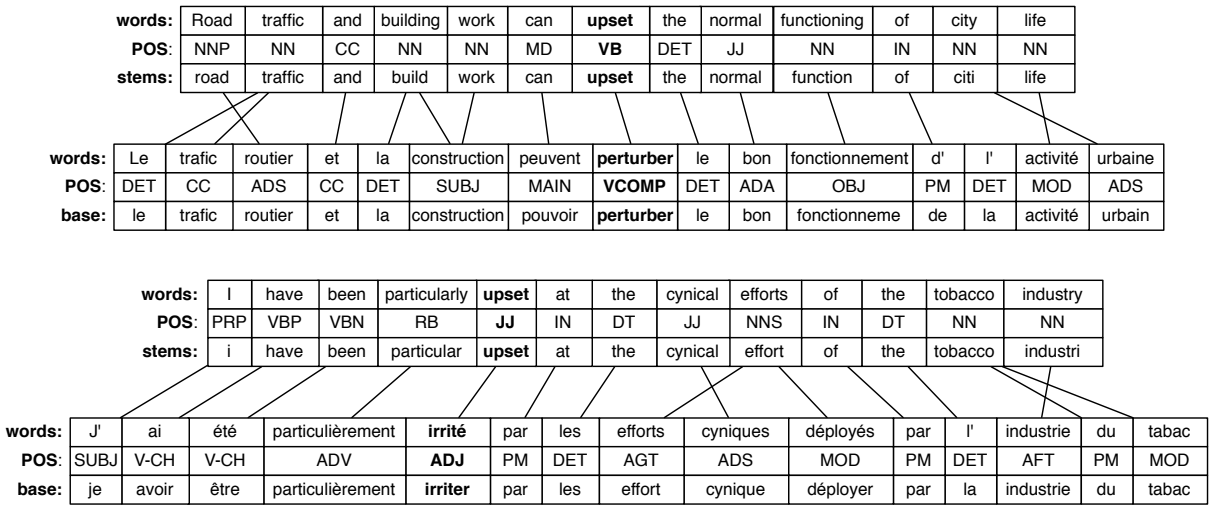


Figure 8.2: Factored Translation Models integrate multiple levels of information in the training data and models.

Models include multiple levels of information, as illustrated in Figure 8.2. The advantages of factored representations are that models can employ more sophisticated linguistic information. As a result they can draw generalizations from the training data, and can generate better translations. This has the potential to lead to improved coverage, more grammatical output, and better use of existing training data.

Consider the following example. If the only occurrences of *upset* were in the sentence pairs given in Figure 8.1, under current phrase-based models the phrase translation probability for the two French phrases would be

$$\begin{aligned} p(perturber|upset) &= 0.5 \\ p(irrité|upset) &= 0.5 \end{aligned}$$

Under these circumstances the French words *irrité* and *perturber* would be equiprobable and the translation model would have no mechanism for choosing between them. In Factored Translation Models, translation probabilities can be conditioned on more information than just words. For instance, by extracting phrases using a combination of factors we can calculate translation probabilities that are conditioned on both words and parts of speech:

$$p(\bar{f}_{words}|\bar{e}_{words},\bar{e}_{pos}) = \frac{count(\bar{f}_{words},\bar{e}_{words},\bar{e}_{pos})}{count(\bar{e}_{words},\bar{e}_{pos})} \tag{8.1}$$

Whereas in the conventional phrase-based models the two French translations of *upset* were equiprobable, we now have a way of distinguishing between them. We can now

correctly choose which French word to use if we know that the English word *upset* is a verb (VB) or an adjective (JJ):

$$\begin{aligned} p(\textit{perturber}|\textit{upset, VB}) &= 1 \\ p(\textit{perturber}|\textit{upset, JJ}) &= 0 \end{aligned}$$

$$\begin{aligned} p(\textit{irrité}|\textit{upset, VB}) &= 0 \\ p(\textit{irrité}|\textit{upset, JJ}) &= 1 \end{aligned}$$

The introduction of factors also allows us to model things we were unable to model in the standard phrase-based approaches to translation. For instance, we can now incorporate a translation model probability which operates over sequences of parts of speech, $p(\bar{f}_{pos}|\bar{e}_{pos})$. We can estimate these probabilities straightforwardly using techniques similar to the ones used for phrase extraction in current approaches to statistical machine translation. In addition to enumerating phrase-to-phrase correspondences using word alignments, we can also enumerate POS-to-POS correspondences, as illustrated in Figure 8.3. After enumerating all POS-to-POS correspondences for every sentence pair in the corpus, we can calculate $p(\bar{f}_{pos}|\bar{e}_{pos})$ using maximum likelihood estimation

$$p(\bar{f}_{pos}|\bar{e}_{pos}) = \frac{count(\bar{f}_{pos}, \bar{e}_{pos})}{count(\bar{e}_{pos})} \tag{8.2}$$

This allows us to capture linguistic facts within our probabilistic framework. For instance, the adjective-noun alternation that occurs between French and English would be captured because the model would assign probabilities such that

$$p(\text{NN ADJ}|\text{JJ NN}) > p(\text{ADJ NN}|\text{JJ NN})$$

Thus a simple linguistic generalization that current approaches cannot learn can be straightforwardly encoded in Factored Translation Models.

The more sophisticated representation of Factored Translation Models does not only open possibilities for improving translation quality. The addition of multiple factors can also be used to extract much more general paraphrases that we are currently able to. Without the use of other levels of representation, our paraphrasing technique is currently limited to learning only *lexical* or *phrasal* paraphrases. However, if the corpus were tagged with additional layers of information, then the same paraphrasing technique could potentially be applied to learn more sophisticated *structural* para-

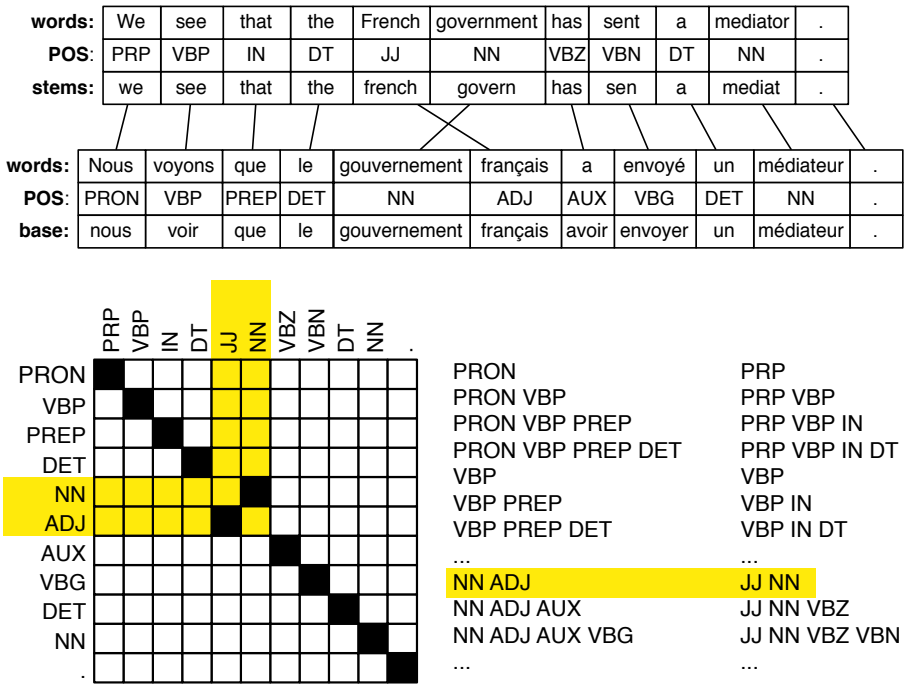


Figure 8.3: In factored models correspondences between part of speech tag sequences are enumerated in a similar fashion to phrase-to-phrase correspondences in standard models.

phrases as well, as illustrated in Figure 8.4. The addition of the part of speech information to the parallel corpus would allow us to not only learn the phrasal paraphrase which equates *the office of the president* with *the president's office*, but would also allow us to extract the general structural transformation for possessives in English $DT\ NN_1\ IN\ DT\ NN_2 = DT\ NN_2\ POS\ NN_1$. This methodology may allow us to discover other structural transformations such as passivization or dative shift. It could further point to other changes like nominalization of certain verbs, and so forth.

Multi-level models, such as Factored Translation Models, have the potential to have wide-ranging impact on all language technologies. Simultaneous modeling of different levels of representation – be they high level concepts such syntax, semantics and discourse, or lower level concepts such as phonemes, morphology and lemmas – are an extremely useful and natural way of describing language. In future work we will investigate a unified framework for the creation of multi-level models of language and translation. We aim to draw on all of the advantages of current phrase-based statistical machine translation – its data-driven, probabilistic framework, and its incorporation of various feature functions into a log-linear model – and extend it to so that it has the ability to generalize, better exploit limited training data, and produce more grammat-

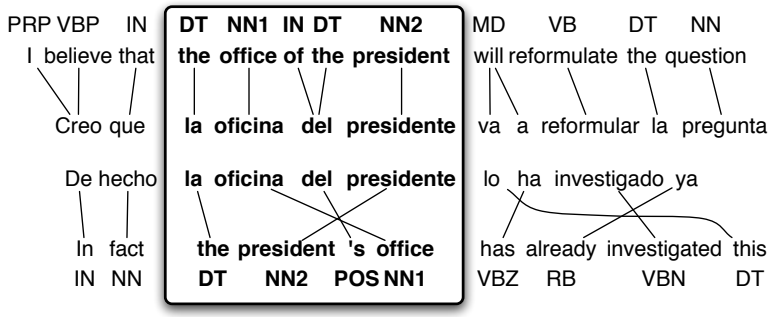


Figure 8.4: Applying our paraphrasing technique to texts with multiple levels of information will allow us to learn structural paraphrases such as $DT\ NN_1\ IN\ DT\ NN_2 \rightarrow ND\ NN_2\ POS\ NN_1$.

ical output text. We will investigate the application of multi-level models not only to translation, but also to other tasks including generation, paraphrasing, and the automatic evaluation of natural language technologies.

Appendix A

Example Paraphrases

This Appendix gives example paraphrases and paraphrase probabilities for 100 randomly selected phrases. The paraphrases were extracted from parallel corpora between English and Danish, Dutch, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish. Enumerating all unique phrases containing up to 5 words from the English section of the Europarl corpus yields approximately 25 million unique phrases. Using the method described in Chapter 3, it is possible to generate paraphrases for 6.7 million of these phrases, such that the paraphrase is different than the original phrase.

The phrases and paraphrases that are presented in this Appenix are constrained to be the same syntactic type, as suggested in Section 3.4. In order to identify the syntactic type of the phrases and their paraphrases, the English sentences in each of the parallel corpora were automatically parsed (Bikel, 2002), and the phrase extraction algorithm was modified to retain this information. Applying this constraint reduces the number of phrases for which we can extract paraphrases (since we are limited to those phrases which are valid syntactic constituents). The number of phrases for which we were able to extract paraphrases falls from 6.7 million to 644 thousand. These paraphrases are generally higher precision, but they come at the expense of recall.

The examples given in the next 18 pages show phrases that were randomly drawn from the 644 thousand phrases for which the syntax-refined method was able to extract paraphrases. The original phrases are italicized, and their paraphrases are listed in the next column. The paraphrase probabilities are given in the final column. The paraphrase probability was calculated using Equation 3.7.

a completely different path	a completely different path	0.635
	the opposite direction	0.083
	an entirely different direction	0.052
	a completely different direction	0.052
	a different direction	0.028
	an apparently opposing direction	0.028
	a markedly different path	0.028
	a very different direction	0.024
	totally different lines	0.024
	quite a different direction	0.024
a conscientious effort	a conscientious effort	0.792
	a conscious attempt	0.125
a special prosecuting office	a special prosecuting office	0.684
	a european public prosecutor 's office	0.070
	a european prosecutor	0.053
	a european public prosecutor	0.053
	a european public ministry	0.035
	the european public prosecutor	0.035
	a european public prosecution office	0.035
	a european public prosecution service	0.018
	a european prosecution service	0.018
a speedy expansion	a speedy expansion	0.444
	the swift expansion	0.236
	rapid enlargement	0.153
	a rapid enlargement	0.139
	a rapid expansion	0.028

a tribunal	a tribunal	0.358
	a court	0.266
	the court	0.102
	court	0.05
	a court of law	0.05
	the courts	0.035
	the court of justice	0.018
	one court	0.016
	justice	0.014
	we	0.014
a well-known public fact	a well-known public fact	0.778
	a well known fact	0.111
	common knowledge	0.111
about cap reform	about cap reform	0.722
	on the reform of the cap	0.167
	on the reform of the common agricultural policy	0.056
	on cap reform	0.056
according to this article	according to this article	0.599
	according to that article	0.175
	under this rule	0.087
	under this article	0.083
	on the basis of this rule	0.056
advance that	advance that	0.458
	drive that forward	0.083
	take that forward	0.083
	move this	0.063
	allow that	0.063
	take this forward	0.063
	achieving this	0.063
	pursuing this	0.063
	carried through	0.063

all the fundamental rights	all the fundamental rights	0.8
	fundamental rights	0.171
	the fundamental rights	0.007
	basic rights	0.005
	human rights	0.002
	their basic rights	0.002
	the constitutional rights	0.002
	constitutional rights	0.002
	fundamental human rights	0.002
	citizens ' rights	0.002
an agreed timetable	an agreed timetable	0.857
	a timetable	0.095
	any set time-frame	0.048
any difficulty whatsoever	any difficulty whatsoever	0.873
	any difficulty	0.032
	no problems	0.032
	a problem	0.016
	no difficulties	0.016
	no problem	0.016
	no difficulty	0.016
any real voice	any real voice	0.9
	a say	0.1
are uniting europe	are uniting europe	0.75
	unite europe	0.25
both articles	both articles	0.875
	these two articles	0.125
brings these together	brings these together	0.938
	put all that together	0.063
cannot support this	cannot support this	0.767
	does not receive my support	0.2
	do not accept it	0.017
	cannot accept this	0.017

comparison to the present situation	comparison to the present situation	0.361
	relation to the current situation	0.333
	comparison with the current situation	0.194
	relation to the situation we have at present	0.111
considered this matter	considered this matter	0.897
	dealt with the subject	0.013
	has now tackled this subject	0.013
	looked at this matter	0.013
	has studied the matter	0.013
	discussed this issue	0.01
	considered this question	0.01
	considering this issue	0.008
	examined the issue	0.008
	take this into account	0.008
criticize	criticize	0.386
	criticise	0.382
	condemn	0.034
	blame	0.018
	denounce	0.014
	reproach	0.014
	censure	0.013
	attack	0.012
	be	0.009
	say	0.009
delivery of money	delivery of money	0.7
	payments	0.233
	payment	0.05
	payment appropriations	0.017
environmental decision-making	environmental decision-making	0.861
	decision-making on the environment	0.083
	environmental decision-taking	0.056

eu principles	eu principles	0.722
	the eu 's principle	0.139
	the very principles of the european union	0.083
	the community principle	0.056
for small and medium-sized producers	for small and medium-sized producers	0.863
	for small and medium-sized enterprises	0.061
	for smes	0.014
	for small and medium-sized businesses	0.012
	for small- and medium-sized enterprises	0.007
has never been implemented	has never been implemented	0.875
	has ever been applied	0.063
	was not done	0.063
has taken upon his shoulders	has taken upon his shoulders	0.775
	has committed itself	0.025
	have signed up to	0.025
	have committed themselves	0.025
	are taken up	0.025
	have agreed to	0.025
	has undertaken	0.025
	have entered into	0.025
	has entered into	0.025
	have committed themselves to	0.025
have to be controlled	have to be controlled	0.9
	has to be halted	0.1
healthily	healthily	0.833
	right	0.167
hundreds of millions of jobs	hundreds of millions of jobs	0.886
	hundreds of thousands of jobs	0.114

i can confirm	i can confirm	0.899
	i can confirm to you	0.033
	i can only echo	0.025
	i can assure you	0.022
	i said	0.013
	i can guarantee	0.008
	i can confirm	0.958
	i can confirm to you	0.042
i turn to you	i turn to you	0.870
	i am addressing these words to you	0.093
	i address you	0.037
in our own century	in our own century	0.867
	in our century	0.067
	in the twentieth century	0.067
in the recipient country	in the recipient country	0.72
	in the host country	0.15
	in the receiving state	0.07
	in the country of destination	0.02
	of the receiving country	0.02
	of the host country	0.02
into irresponsible hands	into irresponsible hands	0.667
	in irresponsible hands	0.167
	in the wrong hands	0.167
is a major misconception	is a major misconception	0.743
	is a serious misunderstanding	0.1
	would be a great mistake	0.071
	is a grave error	0.029
	is completely misguided	0.029
	is a complete misunderstanding	0.029
is a separate question	is a separate question	0.736
	is a separate issue	0.125
	is an additional question	0.063
	is another question	0.056
	is a different matter	0.021

is admissible	is admissible	0.541
	is acceptable	0.1
	is now admissible	0.018
	is admissibility	0.017
	be inadmissible	0.015
	is out of order	0.015
	was admissible	0.015
	is permitted	0.014
	is in order	0.014
	meet the conditions	0.009
is rather too complicated	is rather too complicated	0.806
	is complicated	0.082
	is too complex	0.071
	are complex	0.020
	is quite difficult	0.020
is required instead	is required	0.25
	is required instead	0.194
	are needed	0.179
	is needed	0.086
	required	0.083
	are required	0.061
	are necessary	0.061
	being called for	0.056
	needed	0.030
its three amendments	its three amendments	0.917
	the three proposed amendments	0.083

meet the target	meet the target	0.735
	achieve the targets	0.037
	achieve their goal	0.027
	achieve the goal	0.026
	achieve its objectives	0.026
	achieve the objectives	0.021
	reach the goal	0.016
	meet the objectives	0.016
	fulfill that objective	0.016
	reach the objectives	0.010
most traditional	most traditional	0.888
	more classic	0.063
	more traditional	0.05
must put in place	must put in place	0.7
	obliged to introduce	0.1
	are supposed to implement	0.1
	be done	0.1
my surname	my name	0.767
	my surname	0.198
	my own behalf	0.014
	my own name	0.01
	my attendance	0.003
	myself	0.002
	the minutes	0.002
	my behalf	0.002
	my group	0.002
need revitalising	need revitalising	0.867
	to rebuild	0.067
	is to be reconstructed	0.067

no boundaries	no boundaries	0.584
	no borders	0.161
	no limit	0.1
	no frontiers	0.069
	no bounds	0.032
	no limits	0.014
	no national borders	0.011
	no barriers	0.007
	any limits	0.004
	no end	0.004
non-eu states	third countries	0.517
	non-eu states	0.395
	non-eu countries	0.027
	other countries	0.017
	other states	0.015
	third states	0.008
	other third countries	0.006
	non-member states	0.004
	third world countries	0.003
	non-member countries	0.002
occur	occur	0.251
	happen	0.07
	arise	0.068
	take place	0.027
	exist	0.024
	happens	0.016
	occurs	0.011
	to happen	0.009
	prevent	0.007
	happened	0.007
of a package of reforms	of a package of reforms	0.74
	of the reform package	0.26
of banking and finance	of the banking and financial sector	0.5
	of banking and finance	0.5

of paragraph 18	of paragraph 18	0.733
	in section 18	0.067
	in paragraph 18	0.04
	at point 18	0.03
	to paragraph 18	0.02
	in point 18	0.01
of such sales	of such sales	0.917
	of these sales	0.083
of the cabinet	of the cabinet	0.786
	of cabinet	0.071
	of the federal cabinet	0.071
	of the council	0.048
	from the ministry	0.012
	of the minister	0.012
of the problems there are	of the problems there are	0.75
	there are problems	0.25
of voluntary organizations and foundations	of voluntary organizations and foundations	0.752
	of organizations and foundations	0.181
	of associations and foundations	0.011
on human subjects	on human subjects	0.381
	in people	0.083
	of human beings	0.083
	on human beings	0.081
	on humans	0.067
	on people	0.061
	to humans	0.061
	in man	0.047
	of people	0.033
	to people	0.033
our own heads	our own heads	0.867
	our minds	0.133

part of the agreement	part of the agreement	0.739
	part of the treaty	0.089
	the agreement	0.059
	a part of the agreement	0.018
	a condition of the agreement	0.018
	part of the overall settlement	0.011
	partners	0.011
	part of the settlement	0.009
	part of it	0.008
	the treaty	0.007
policy making	policy making	0.383
	political decisions	0.154
	the legislative process	0.139
	the political stage	0.048
	politics	0.040
	political attention	0.028
	decision-making	0.028
	the political scene	0.022
	the politicians	0.021
	policy decisions	0.020
quite obviously	clearly	0.32
	obviously	0.245
	quite obviously	0.166
	naturally	0.079
	quite clearly	0.044
	certainly	0.02
	very clearly	0.017
	apparently	0.007
	evidently	0.006
	indeed	0.006
really democratic	really democratic	0.825
	truly democratic	0.15
	thoroughly democratic	0.025

reprocessed	reprocessed	0.625
	processed	0.181
	made	0.063
	established	0.063
	incorporated	0.031
	included	0.019
	taken	0.006
	used	0.006
	fed	0.006
	taken	0.2
	used	0.2
	processed	0.2
	been included	0.2
	included	0.2
rescission of the contract	rescission of the contract	0.75
	the cancellation of the agreement	0.25
scots	scots	0.528
	scotland	0.293
	the scots	0.124
	the people of scotland	0.029
	scotsmen	0.026
serious faults	serious faults	0.522
	serious defects	0.153
	serious shortcomings	0.114
	serious deficiencies	0.082
	grave shortcomings	0.022
	significant deficiencies	0.015
	considerable shortcomings	0.015
	severe shortages	0.015
	shortcomings	0.015
	a lack	0.007

subjects	issues	0.303
	subjects	0.185
	matters	0.114
	questions	0.065
	areas	0.053
	points	0.037
	topics	0.031
	themes	0.020
	substances	0.018
	things	0.017
take that view	take that view	0.545
	agree	0.090
	think so	0.069
	agree with that	0.046
	share this view	0.032
	believe this	0.028
	shares this view point	0.016
	share this point of view	0.016
	shares this point of view	0.016
	shares that view	0.016
the appropriate adjustment	the appropriate adjustment	0.584
	the necessary adjustments	0.221
	the necessary amendment	0.071
	the necessary adjustment	0.049
	the necessary changes	0.036
	the necessary corrections	0.013
	the necessary amendments	0.013
	adjustments	0.013

the last issue	the last issue	0.282
	the last point	0.123
	my last point	0.098
	the final point	0.09
	my final point	0.069
	the last question	0.065
	the last item	0.019
	the final issue	0.017
	one final issue	0.017
	the final subject	0.013
the lessons	the lessons	0.494
	lessons	0.091
	the lesson	0.079
	a lesson	0.024
	experience	0.015
	its lesson	0.013
	the experience	0.012
	it	0.010
	we	0.008
	the example	0.007
the light of current circumstances	the light of current circumstances	0.9
	the light of current events	0.1
the one remaining hope	the only hope	0.290
	the one remaining hope	0.219
	their only hope	0.2
	our only hope	0.169
	the only real hope	0.121

the part of individual countries	the part of individual countries	0.5
	the individual member states	0.136
	the member states	0.127
	individual member states	0.042
	member states	0.017
	each member state	0.017
	the individual states	0.017
	the different member states	0.017
	one member state	0.017
	the national member states	0.008
the players	the players	0.385
	players	0.18
	operators	0.078
	the actors	0.048
	the parties	0.028
	those	0.023
	the operators	0.018
	all the players	0.011
	the stakeholders	0.01
	agents	0.01
the power of the union	the power of the union	0.697
	the responsibility of the union	0.064
	the competence of the eu	0.042
	the capacity of the union	0.042
	the european union 's remit	0.033
	the powers of the union	0.025
	the union 's ability	0.025
	eu competence	0.025
	the union 's scope	0.017
	the union 's capacity	0.008

the real choice	the real choice	0.83
	a genuine choice	0.038
	real choice	0.033
	the true choices	0.025
	the real election	0.017
	a real choice	0.015
	genuine choice	0.01
	the political options	0.008
	free choice	0.008
	a real election	0.005
the significant sums	the significant sums	0.778
	the substantial sums	0.222
the united kingdom conservative party	the united kingdom conservative party	0.544
	the british conservative party	0.345
	the conservative party in the united kingdom	0.063
	the british conservatives	0.045
	uk conservatives	0.004
the vast majority of researchers	the vast majority of researchers	0.917
	most researchers	0.083
the very best practice	best practice	0.401
	the very best practice	0.295
	the best practices	0.105
	the best practice	0.089
	best practices	0.087
	better practice	0.012
	the best possible practice	0.005
	best current practice	0.003
	good practices	0.003
these two budgets	these two budgets	0.806
	both these budgets	0.194
think in euros	think in euros	0.583
	thinking in euros	0.333
	think in euro terms	0.083

thirteen years ago	thirteen years ago	0.917
	13 years ago	0.068
	just 13 years ago	0.015
this french initiative	this french initiative	0.710
	the french initiative	0.212
	the french initiatives	0.044
	the text of the french initiative	0.022
	the french republic 's initiative	0.011
thousands of young men	thousands of young people	0.601
	thousands of young men	0.249
	hundreds of young people	0.064
	several thousand young people	0.029
	thousands of people	0.029
	thousands of young women	0.029
to be warmly welcomed	to be warmly welcomed	0.585
	to be welcomed	0.174
	very positive	0.056
	is very welcome	0.056
	to be very greatly welcomed	0.019
to previous presidencies	to previous presidencies	0.688
	to previous wars	0.125
	with the others	0.125
	from all those that went before	0.063
to solve the problem either	solving the problem	0.167
	to resolve the problem	0.096
	to solve this problem	0.088
	to address the problem	0.063
	to solve the problem either	0.063
	tackling the problem	0.033
	to answer the problem	0.033
	to solve that problem	0.033
	to resolve it	0.033
	tackling the issue	0.033

to the candidates themselves	to the candidates themselves	0.778
	by the applicant countries	0.111
	with the accession candidates	0.111
to the holding	to the holding	0.6
	to exploitation	0.083
	for exploitation	0.083
	of the company	0.067
	of the farm	0.033
	of the enterprise	0.033
	on the holding	0.033
	of the business	0.017
	in the business	0.017
	of any company structure	0.017
to the very limit	to the very limit	0.75
	to the limit	0.25
translation errors	translation errors	0.819
	translation error	0.152
	translation	0.029
ukraine and moldova	ukraine and moldova	0.833
	ukraine and moldavia	0.106
	the ukraine and moldova	0.061
very interesting things	very interesting things	0.917
	a lot that is of interest	0.083
voluntary organizations	voluntary organizations	0.441
	voluntary organisations	0.220
	non-governmental organisations	0.083
	ngos	0.047
	non-governmental organizations	0.028
	organisations	0.023
	associations	0.021
	the voluntary organisations	0.02
	the voluntary organizations	0.019
	organizations	0.016

wake up to this situation	wake up	0.278
	wake up to this situation	0.222
	frighten them	0.167
	worry	0.111
	to express concern	0.056
	happening	0.056
	express concern	0.056
	worry about	0.056
was suspended at 11.56 a.m.	was suspended at 11.56 a.m.	0.896
	was suspended at 11.55 a.m.	0.083
	was adjourned at 11.55 a.m.	0.021
we could describe it	we could describe it	0.75
	can be said	0.25
who we represent	who we represent	0.895
	that we represent	0.043
	we represent	0.037
	whom we represent	0.025
wish to clarify	wish to clarify	0.447
	want to make perfectly clear	0.167
	would like to ask	0.083
	would like to comment on	0.061
	would like to pick up	0.030
	should now like to comment on	0.030
	would like to mention	0.030
	would like to deal with	0.030
	would comment on	0.030
	should like to comment on	0.030

Appendix B

Example Translations

This Appendix gives a number of examples which illustrate the types of improvements that we get by integrating paraphrases into statistical machine translation. The tables show example translations produced by the baseline system and by the paraphrase system when their translation models are trained on various sized parallel corpora. The translation models were trained on corpora containing 10,000, 20,000, 40,000, 80,000, 160,000 and 320,000 sentence pairs (as described in Section 7.1). In addition to the MT output we provide the source sentences and reference translations.

The bold text is meant to highlight regions where the translations produced by the paraphrase system represent improvement in translation quality over the baseline system. In some cases a particular source word is untranslated in the baseline, but is translated by the paraphrase system. For instance, in the first example in Table B.1 the Spanish word *altera* is left untranslated by the baseline system, but the paraphrase system produces the English translation *warning*, which matches the reference translation.

In some cases neither the baseline system nor the paraphrase system manage to translate a word. For instance, in the same example as above, the Spanish word *ven* is left untranslated by both systems. Since the training data for the translation model was so small, none of the paraphrases of *ven* had translations, thus the paraphrase system performed similarly to the baseline system. We do not highlight these instances, since we intended the bold text to be indicative of improved translations.

SOURCE	REFERENCE	BASELINE SYSTEM	PARAPHRASE SYSTEM
estoy de acuerdo con su señal de **alerta** contra el **regreso** , al que algunos se ven **tentados** , a los métodos **intergubernamentales** .	i agree with his **warnings** against a **return** to **intergovernmental** methods , which some are **tempted** by .	i agree with the sign of **alerta** against the **regreso** , to which some are ven **tentados** the methods **intergubernamentales** .	i agree with the sign of **warning** against the **return** to which some are ven **temptation** to the **intergovernmental** methods .
votaré en favor de la aprobación del proyecto de reglamento .	**i will vote** to approve the draft regulation .	**votaré** in favour of the approval of the draft regulation .	**i shall vote** in favour of the approval of the draft regulation .
estos autobuses no sólo son más baratos y versátiles **internacionalmente** , sino también **más respetuosos con el medio ambiente** porque **utilizan menos combustible por pasajero** .	such buses are not only cheaper and **internationally** deployable , they are also **more environmentally-friendly** because they **use less fuel per passenger** .	not only are these autobuses more baratos and versátiles **internacionalmente** , but also **more respetuosos with the environment** because **less fuel used by pasajero** .	these people not only are more and versátiles **international** , but also **more ecological** because **used less fuel per passenger** .
por tanto , querría **proponerles** que el año próximo el parlamento no presente un informe general .	that is why i should like **to propose to you** that from next year we in parliament no longer present a general report .	therefore , i would like **proponerles** that next year parliament not produce a general report .	therefore , i would like **to suggest** that next year parliament not produce a general report .
considero que sobre la base de los trabajos **iniciados** por las anteriores **presidencias** , el estará en condiciones de **presentar** un balance preciso del proceso de adhesi'on .	i feel that on the basis of the work **initiated** by previous **presidencies** , he will be in a position **to offer** a quite precise overview of the accession process .	i think on the basis of the work **iniciados** by the previous **presidencias** , he will be able **to** a specific figures of the process of accession .	i think on the basis of the work **started** by the previous **presidency** , he will be able **to present** a course must be of the process of accession .

Table B.1: Example translations from the baseline and paraphrase systems when trained on a Spanish-English corpus with 10,000 sentence pairs

SOURCE	REFERENCE	BASELINE SYSTEM	PARAPHRASE SYSTEM
somos muchos los que queremos una federación de **estados-nación** .	there are many of us who want a federation of **nation states** .	many people are that we want a federation of **estados-nación** .	many of which we want a federation of **national states** .
quisiera que se empezara por esta **cooperación reforzada** para poner algunos ejemplos de la nueva **potencialidad** europea .	I would like to begin this **closer cooperation** so that we have some examples of the new european **potential** .	i would like to empezara for this **cooperation reforzada** to bring some examples of the new european **potencialidad** .	i would like to let for **increased cooperation** in order to bring some examples of the new european **potential** .
también pide que se establezcan valores de referencia para difundir las **mejores prácticas** en toda la ue .	he also calls for benchmarking to spread **best practices** across the eu .	it also calls for reference values and **practices in the best** we can help to spread throughout the eu .	that is also called and values of reference for we can help to spread the **best practices** throughout the eu .
lo que no significa que **dispondremos** del tiempo y de los medios necesarios para tratar cada una de ellas .	this does not mean that **we shall have** the time and resources to deal with each of them .	this does not mean that **dispondremos** time and resources needed to deal with each one of them .	this does not mean that **we have** the time and resources needed to deal with each one of them .
examinemos de nuevo los **flujos comerciales que existen actualmente** entre la unin europea y los pases de europa central y oriental .	**let us examine** the **trade flows that currently exist** between the european union and the central and eastern european countries .	**examinemos** once again that **there are currently flujos trade** between the european union and the countries of central and eastern europe .	**look at** new **trade that currently exist** between the european union and the countries of central and eastern europe .

Table B.2: Example translations from the baseline and paraphrase systems when trained on a Spanish-English corpus with 20,000 sentence pairs

SOURCE	REFERENCE	BASELINE SYSTEM	PARAPHRASE SYSTEM
sin embargo , hay **aspectos y cuestiones de envergadura** de los que **disentimos** .	however , there are **aspects and questions of significance** that **we disagree on** .	however , there are **issues and issues of content** of the **disentimos** .	however , there are **aspects and content of the issues** that **we do not agree** .
se me ha encomendado la misión de ser ponente en relación con estonia .	**i have been entrusted with the task** of acting as rapporteur where estonia is concerned .	**i encomendado has the task** of being rapporteur in connection with estonia .	**i was given the task** of being rapporteur in connection with estonia .
no podemos darnos el lujo de perder ocasiones como las que hubo a comienzos de los años 90 .	**we cannot afford to lose** more of the momentum that existed at the beginning of the nineties .	**we cannot understand luxury of losing** occasions as there was at the beginning of the 90 years .	**we cannot afford to lose** occasions as there was at the beginning of the 90 years .
la democracia no es solamente una cuestión de **mayorías** sino sobre todo de una codecisión equilibrada de las minorías .	democracy is not just about **majorities** , more than anything it is about minorities being given commensurate codecision powers .	democracy is not just a question of **mayorías** but above all a codecision balanced of minorities .	democracy is not just a question of **majority** but above all a codecision balanced of minorities .
no nos podemos dejar cegar por los **porcentajes** de derecho comunitario recogidos en la legislación nacional .	we **should not be blinded** by the **percentage** figures of community law transposed into national legislation .	we **can make cegar** by the **porcentajes** of community law incorporated into national legislation .	we **cannot be reduced** by the **percentage** of community law incorporated into national legislation .

Table B.3: Example translations from the baseline and paraphrase systems when trained on a Spanish-English corpus with 40,000 sentence pairs

SOURCE	REFERENCE	BASELINE SYSTEM	PARAPHRASE SYSTEM
tiene que ser un acuerdo que **impulse** la cooperacin pesquera y el sector pesquero marroquí , pero que **satisfaga** tambin plenamente los intereses pesqueros de las flotas europeas .	it must be an agreement which **provides an impetus** for fisheries cooperation and the moroccan fishing industry , but which also fully **satisfies** the fisheries interests of the european fleets .	it has to be a cooperation agreement **impulse** fisheries and the moroccan fisheries sector , but also **satisfaga** fully the interests of the european fisheries fleets .	it has to be an agreement which **pursue** cooperation fishing and the moroccan fisheries sector , but we also fully **satisfy** the fishing interests of the european fleets .
francia , por su lado , posee departamentos de **ultramar** en el **océano índico** y el **caribe** .	france , on the other hand , owns **overseas** departments in the **indian ocean** and the **caribbean** .	france , for its part , has **ultramar** departments of the **océano índico** and the **caribe** .	france , for its part , has the departments in the **indian ocean** and the **caribbean** .
desearía concluir , simplemente , **mencionando** tres problemas que han surgido .	i would like to conclude simply by **mentioning** three problems which have come up .	i would like to conclude , simply , **mencionando** three problems that have emerged .	i would like to conclude , this is simply **mentioning** three problems which have arisen .
hay que tratar de **remover** todas las causas de aprovechamiento existentes **buscando** y persiguiendo a quien se ha aprovechado , **ha abusado** de las poblaciones que se hallan en estado de necesidad .	we should try **to remove** all the existing causes of exploitation , **identifying** and bringing to justice those who **have abused** peoples which are in a state of need .	we must deal with **remover** all the causes of existing **buscando** exploitation and to pursue to anyone who has used this opportunity , has **has abusado** of people who have chosen a state of need .	we must try **to eliminate** all the causes of exploitation and pursuing existing **seeking** to anyone who **has abused** people who are in a state of need .
la auténtica sorpresa de este presupuesto se refiere a las acciones exteriores donde **domina el egoísmo** .	the real surprise in this budget relates to external actions , where **selfishness is the rule** .	the real surprise this budget relates to the external action **domina** where the **egoísmo** .	the real surprise this budget relates to the external actions , where the **dominant selfishness** .

Table B.4: Example translations from the baseline and paraphrase systems when trained on a Spanish-English corpus with 80,000 sentence pairs

SOURCE	REFERENCE	BASELINE SYSTEM	PARAPHRASE SYSTEM
basta con que miremos a nuestros **televisores** y veamos las dificultades actuales en el oriente medio para que lo recordemos todos los días .	we only have to look at our **television sets** to see the difficulties in the middle east to be reminded of that each and every day .	it is enough to look at our **televisores** and see the present difficulties in the middle east for what it every day .	it is enough to look to our **television** and see the present difficulties in the middle east to remind you that every day .
pero he **captado** perfectamente que más allá del aspecto lingüístico , su pregunta esconde otra preocupación .	but i **understand** that , behind the linguistic aspect , your question expresses another concern .	but i **captado** well beyond the linguistic aspect , your question hides another concern .	but i **understand** perfectly well that beyond the linguistic aspect , your question hides another concern .
eso demuestra que la comisión **no permaneció inactiva** y que llevó a efecto las observaciones que ustedes han formulado .	this shows that the commission **has not been idle** and that it is acting on the comments that you have made .	this demonstrates that the commission **permaneció not been idle** and led to effect the remarks that you have raised .	this demonstrates that the commission **has not been idle** and that led to effect the remarks that you have put forward .
nos referimos en particular al nuevo proyecto piloto que **financiará** acciones de información en materia de lucha contra la pederastia .	we refer in particular to the new pilot scheme **to finance** information measures in the fight against paedophilia .	we are referring in particular the new pilot project which **financiará** information actions to combat paedophilia .	we are referring in particular to the new pilot project **to fund** actions of information in the fight against paedophilia .
en lo referente al procedimiento **disciplinario** no unas reglas claras se deben aplicar dichas reglas . y tampoco la cuestión de la **externalización** está **definitivamente aclarada** .	as for **disciplinary** procedures , no clear rules have been formulated so far as to how such rules would be enforced , nor has the question of **external agencies** been **conclusively resolved** .	with regard to the procedure **disciplinario** not yet clear rules on how they should apply these rules , and also the question of the **externalización** is **definitely aclarada** .	as far as the **disciplinary** procedure but so far not clear rules on how these rules must apply either , nor the question of the **delegation** is **quite clear** .

Table B.5: Example translations from the baseline and paraphrase systems when trained on a Spanish-English corpus with 160,000 sentence pairs

SOURCE	REFERENCE	BASELINE SYSTEM	PARAPHRASE SYSTEM
resulta que en los últimos tiempos ese parlamento ha sido víctima de algunos **equívocos** e incluso de algunas críticas .	the central american parliament appears to have recently been the subject of certain **misunderstandings** and even criticism .	it is that in recent times that parliament has been the victim of some **equívocos** and even some criticisms .	it is that in recent times that parliament has been the victim of some **misunderstandings** and even some criticisms .
la política económica y social de la ue ha conducido a la reducción de **la capacidad adquisitiva** de los trabajadores .	the economic and social policy of the eu has resulted in the reduced **purchasing power** of workers .	the economic and social policy of the eu has led to the reduction in **capacity adquisitiva** of workers .	the economic and social policy of the eu has led to the reduction of **buying power** of workers .
la lucha contra la exclusión social queda **relegada** al rango de la asistencia .	the fight against social exclusion is **relegated** to the level of assistance .	the fight against social exclusion is **relegada** the rank of assistance .	the fight against social exclusion is **relegated** to the status of the assistance .
los interlocutores sociales se encuentran en la posición más favorable para proponer **soluciones practicables** .	the social partners are in the best place to work out **viable solutions** .	the social partners are in the more favourable position to propose **solutions practicables** .	the social partners are in the more favourable position to propose **practical solutions** .
en lo relativo a los descriptores , como “bajo contenido en alquitrán” , “suave” y “light” , todavía me sigue pareciendo que son **engañosos** , y por eso estoy en contra de las enmiendas .	i still feel that messages such as ‘low-tar’ , ‘mild’ and ‘light’ are **misleading** , so i will , in fact , oppose the amendments .	with regard to the descriptores , as “low tar content in” , “soft” and “light” , there is still that are **engañosos** , and i am therefore against the amendments .	with regard to the descriptores , as ‘low tar content in’ , ‘mild’ and ‘light’ , there is still are **misleading** , and that is why i am against the amendments .

Table B.6: Example translations from the baseline and paraphrase systems when trained on a Spanish-English corpus with 320,000 sentence pairs

Bibliography

Satanjeev Banerjee and Alon Lavie (2005). Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, Michigan.

Colin Bannard and Chris Callison-Burch (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan.

Regina Barzilay (2003). *Information Fusion for Mutlidocument Summarization: Paraphrasing and Generation*. PhD thesis, Columbia University, New York.

Regina Barzilay and Lillian Lee (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (HLT/NAACL-2003)*, Edmonton, Alberta.

Regina Barzilay and Kathleen McKeown (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*, Toulouse, France.

Dan Bikel (2002). Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of Second International Conference on Human Language Technology Research (HLT-02)*, San Diego, California.

Alexandra Birch, Chris Callison-Burch, and Miles Osborne (2006). Constraining the phrase-based, joint probability statistical translation model. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2006)*, Cambridge, Massachusetts.

Phil Blunsom and Trevor Cohn (2006). Discriminative word alignment with conditional random fields. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL-CoLing-2006)*, Sydney, Australia.

Peter Brown, John Cocke, Stephen Della Pietra, Vincent Della Pietra, Frederick Jelinek, Robert Mercer, and Paul Poossin (1988). A statistical approach to language translation. In *12th International Conference on Computational Linguistics (CoLing-1988)*.

Peter Brown, John Cocke, Stephen Della Pietra, Vincent Della Pietra, Frederick Jelinek, Robert Mercer, and Paul Poossin (1990). A statistical approach to language translation. *Computational Linguistics*, 16(2).

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer (1991). A statistical approach to sense disambiguation in machine translation. In *Workshop on Human Language Technology*, pages 146–151.

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer (1993). The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Colin Bannard, and Josh Schroeder (2005). Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder (2007). (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne (2006a). Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (HLT/NAACL-2006)*, New York, New York.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn (2006b). Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European*

Chapter of the Association for Computational Linguistics (EACL-2006), Trento, Italy.

Chris Callison-Burch, David Talbot, and Miles Osborne (2004). Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, Barcelona, Spain.

Michael Carl and Andy Way (2003). *Recent Advances in Example-Based Machine Translation*. Springer.

Colin Cherry and Dekang Lin (2003). A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, Sapporo, Japan.

David Chiang (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Philip Clarkson and Roni Rosenfeld (1997). Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings ESCA Eurospeech*.

Deborah Coughlin (2003). Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit IX*, New Orleans, Louisiana.

Ido Dagan and Alon Itai (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Tiphaine Dalmas (2007). *Information Fusion for Automated Question Answering*. PhD thesis, University of Edinburgh, Scotland.

Adrià de Gispert, Josè B. Marino, and Josep M. Crego (2005). Improving statistical machine translation by classifying and generalizing inflected verb forms. In *Proceedings of 9th European Conference on Speech Communication and Technology*.

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

Mona Diab (2000). An unsupervised method for word sense tagging using parallel corpora: A preliminary investigation. In *Proceedings of Special Interest Group in Lexical Semantics (SIGLEX) Workshop*.

Mona Diab and Philip Resnik (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, Pennsylvania.

George Doddington (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Human Language Technology: Notebook Proceedings*, pages 128–132, San Diego, California.

Bill Dolan and Chris Brockett (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of 3rd International Workshop on Paraphrasing*.

Bill Dolan, Chris Quirk, and Chris Brockett (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*.

Mark Dras (1997). Representing paraphrases using synchronous tree adjoining grammars. In *35tg Annual Meeting of the Association for Computational Linguistics (ACL-1997)*, Madrid, Spain.

Mark Dras (1999a). A meta-level grammar: Redefining synchronous TAGs for translation and paraphrase. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 98–104, Hong Kong.

Mark Dras (1999b). *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. PhD thesis, Macquarie University, Australia.

Pablo Ariel Duboue and Jennifer Chu-Carroll (2006). Answering the question you wished they had asked: The impact of paraphrasing for question answering. In *Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (HLT/NAACL-2006)*, New York, New York.

Helge Dyvik (1998). Translations as semantic mirrors. In *Workshop on Multilinguality and the Lexicon*, pages 24–44.

Alexander Fraser and Daniel Marcu (2006). Semi-supervised training for statistical word alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL-CoLing-2006)*, Sydney, Australia.

William Gale and Kenneth Church (1993). A program for aligning sentences in bilingual corpora. *Compuatational Linguistics*, 19(1):75–90.

Sharon Goldwater and David McClosky (2005). Improving statistical MT through morphological analysis. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP-2005)*, Vancouver, British Columbia., Canada.

Zellig Harris (1954). Distributional structure. *Word*, 10(2-3):146–162.

Eduard Hovy and Deepak Ravichandran (2003). Holy and unholy grails. Panel Discussion at MT Summit IX.

Ali Ibrahim, Boris Katz, and Jimmy Lin (2003). Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the Second International Workshop on Paraphrasing (ACL 2003)*.

Nancy Ide (2000). Cross language sense determination: Can it work? *Computers and the Humanities: Sepcail Issue on SENSEVAL*, 34:15–48.

Lidija Iordanskaja, Richard Kittredge, and Alain Polgére (1991). Lexical selection and paraphrase in a meaning text generation model. In Cécile L. Paris, William R. Swartout, and William C. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*. Kluwer Academic.

Abraham Ittycheriah and Salim Roukos (2005). A maximum entropy word aligner for arabic-english machine translation. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP-2005)*, Vancouver, British Columbia., Canada.

Hiroshi Kanayama (2003). Paraphrasing rules for automatic evaluation of translation into japanese. In *Proceedings of the Second International Workshop on Paraphrasing (ACL 2003)*, Sapporo, Japan.

David Kauchak and Regina Barzilay (2006). Paraphrasing for automatic evaluation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, Sydney, Australia.

Katrin Kirchhoff, Mei Yang, and Kevin Duh (2006). Machine translation of parliamentary proceedings using morpho-syntactic knowledge. In *Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation*.

Reinhard Kneser and Hermann Ney (1995). Improved smoothing for mgram language modeling. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*.

Philipp Koehn (2004). Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the 6th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, Washington DC.

Philipp Koehn (2005). A parallel corpus for statistical machine translation. In *Proceedings of MT-Summit*, Phuket, Thailand.

Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, David Talbot, and Michael White (2005a). Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of International Workshop on Spoken Language Translation*.

Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, David Talbot, and Michael White (2005b). Edinburgh system description for the 2005 NIST MT evaluation. In *Proceedings of the NIST 2005 Machine Translation Evaluation Workshop*.

Philipp Koehn, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Alexandra Constantin, Brooke Cowan, Chris Dyer, Marcello Federico, Evan Herbst, Hieu Hoang, Christine Moran, Wade Shen, and Richard Zens (2006). Factored translation models. CLSP Summer Workshop Final Report WS-2006, Johns Hopkins University.

Philipp Koehn and Kevin Knight (2003). Empirical methods for compound splitting. In *10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2003)*, Budapest, Hungary.

Philipp Koehn and Christof Monz (2005). Shared task: Statistical machine translation between European languages. In *Proceedings of ACL 2005 Workshop on Parallel Text Translation*, Ann Arbor, Michigan.

Philipp Koehn and Christof Monz (2006). Manual and automatic evaluation of machine translation between European languages. In *Proceedings of NAACL 2006 Workshop on Statistical Machine Translation*, New York, New York.

Philipp Koehn, Franz Josef Och, and Daniel Marcu (2003). Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (HLT/NAACL-2003)*, Edmonton, Alberta.

LDC (2005). Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5.

Audrey Lee and Mark Przybocki (2005). NIST 2005 machine translation evaluation official results. Official release of automatic evaluation scores for all submissions.

Vladimir I. Levenshtein (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Report*, 10(8):707–710.

Dekang Lin (1993). Parsing without over generation. In *31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio.

Dekang Lin and Patrick Pantel (2001). Discovery of inference rules from text. *Natural Language Engineering*, 7(3):343–360.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie Dorr (2007). Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Daniel Marcu and William Wong (2002). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, Philadelphia, Pennsylvania.

Kathleen R. McKeown (1979). Paraphrasing using given and new information in a question-answer system. In *17th Annual Meeting of the Association for Computational Linguistics*, La Jolla, California.

Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman (2002). Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of the Human Language Technology Conference*.

Dan Melamed, Ryan Green, and Jospeh P. Turian (2003). Precision and recall of machine translation. In *Proceedings of the Human Language Technology Conference*

of the North American chapter of the Association for Computational Linguistics (HLT/NAACL-2003), Edmonton, Alberta.

I. Dan Melamed (1998). Manual annotation of translational equivalence: The blinker project. Cognitive Science Technical Report 98/07, University of Pennsylvania.

Marie Meteer and Varda Shaked (1988). Strategies for effective paraphrasing. In *12th International Conference on Computational Linguistics (CoLing-1988)*, pages 431–436.

George A. Miller (1990). Wordnet: An on-line lexical database. *Special Issue of the International Journal of Lexicography*, 3(4).

Robert C. Moore (2004). Improving IBM word alignment model 1. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, pages 518–525, Barcelona, Spain.

Robert C. Moore (2005). A discriminative framework for bilingual word alignment. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP-2005)*, Vancouver, British Columbia., Canada.

Robert C. Moore, Wen-Tau Yih, and Andreas Bode (2006). Improved discriminative bilingual word alignment. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, Sydney, Australia.

Dragos Munteanu and Daniel Marcu (2005). Improving machine translation performance by exploiting comparable corpora. *Computational Linguistics*, 31(4):477–504.

Dragos Stefan Munteanu and Daniel Marcu (2006). Extracting parallel sub-sentential fragments from comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL-CoLing-2006)*, Sydney, Australia.

Makoto Nagao (1981). A framework of a mechanical translation between japanese and english by analogy principle. In A. Elithorn and R. Banerji, editors, *Artificial and Human Intelligence: edited review papers presented at the international NATO Symposium*, pages 173–180.

Sonja Niessen and Hermann Ney (2004). Statistical machine translation with scarce resources using morpho-syntatic analysis. *Computational Linguistics*, 30(2):181–204.

Sonja Niessen, Franz Josef Och, Gregor Leusch, and Hermann Ney (2000). An evaluation tool for machine translation: Fast evaluation for mt research. In *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece.

NIST and LDC (2007). Post editing guidelines for gale machine translation evaluation. Guidelines developed by the National Institute of Standards and Technology (NIST), and the Linguistic Data Consortium (LDC).

Doug Oard, David Doermann, Bonnie Dorr, Daqing He, Phillip Resnik, William Byrne, Sanjeeve Khudanpur, David Yarowsky, Anton Leuski, Philipp Koehn, and Kevin Knight (2003). Desperately seeking Cebuano. In *Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (HLT/NAACL-2003)*, Edmonton, Alberta.

Franz Josef Och (2003). Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, Sapporo, Japan.

Franz Josef Och and Hermann Ney (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, Pennsylvania.

Franz Josef Och and Hermann Ney (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och and Hermann Ney (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Karolina Owczarzak, Declan Groves, Josef Van Genabith, and Andy Way (2006). Contextual bitext-derived paraphrases in automatic mt evaluation. In *Proceedings of the SMT Workshop at HLT-NAACL*.

Bo Pang, Kevin Knight, and Daniel Marcu (2003). Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings*

of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (HLT/NAACL-2003), Edmonton, Alberta.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, Pennsylvania.

Mark Przybocki (2004). NIST 2004 machine translation evaluation results. Confidential e-mail to workshop participants.

Chris Quirk, Chris Brockett, and William Dolan (2004). Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, Barcelona, Spain.

Deepak Ravichandran and Eduard Hovy (2002). Learning sufrace text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, Pennsylvania.

Philip Resnik and Noah Smith (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Philip Resnik and David Yarowksy (1999). Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Satoshi Sato and Makoto Nagao (1990). Toward memory-based translation. In *Papers presented to the 13th International Conference on Computational Linguistics (CoLing-1990)*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2006)*, Cambridge, Massachusetts.

Harold Somers (1999). Review article: Example-based machine translation. *Machine Translation*, 14(2):113–157.

Andreas Stolcke (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado.

David Talbot and Miles Osborne (2006). Modeling lexical redundancy for machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL-CoLing-2006)*, Sydney, Australia.

Ben Taskar, Simon Lacoste-Julien, and Dan Klein. (2005). A discriminative matching approach to word alignment. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP-2005)*, Vancouver, British Columbia., Canada.

Henry Thompson (1991). Automatic evaluation of translation quality: Outline of methodology and report on pilot experiment. In *(ISSCO) Proceedings of the Evaluators Forum*, pages 215–223, Geneva, Switzerland.

Christoph Tillmann (2003). A projection extension algorithm for statistical machine translation. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*, Sapporo, Japan.

Christoph Tillmann (2004). A unigram orientation model for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (HLT/NAACL-2004)*, Boston, Massachusetts.

Christoph Tillmann and Tong Zhang (2005). A localized prediction model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan.

Stephan Vogel, Hermann Ney, and Christoph Tillmann (1996). HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics (Coling-1996)*.

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, and Alex Waibel (2003). The CMU statistical machine translation system. In *Proceedings of MT Summit IX*, New Orleans, Louisiana.

Dekai Wu and Pascale Fung (2005). Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP-2005)*, Jeju Island, Korea.

Mei Yang and Katrin Kirchhoff (2006). Phrase-based backoff models for machine translation of highly inflected languages. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy.

Ying Zhang and Stephan Vogel (2004). Measuring confidence intervals for the machine translation evaluation metrics. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2004)*.

Ying Zhang, Stephan Vogel, and Alex Waibel (2004). Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of Proceedings of Language Resources and Evaluation (LREC-2004)*, Lisbon, Portugal.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy (2006). Re-evaluating machine translation results with paraphrase support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, Sydney, Australia.