# ABSTRACT

Title of dissertation: Improving Statistical Machine Translation Using Comparable Corpora

Matthew Garvey Snover, Doctor of Philosophy, 2010

Dissertation directed by: Professor Bonnie Dorr
Department of Computer Science

With thousands of languages in the world, and the increasing speed and quantity of information being distributed across the world, automatic translation between languages by computers, *Machine Translation (MT)*, has become an increasingly important area of research. State-of-the-art MT systems rely not upon hand-crafted translation rules written by human experts, but rather on learned statistical models that translate a *source* language to a *target* language. These models are typically generated from large, parallel corpora containing copies of text in both the source and target languages. The co-occurrence of words across languages in parallel corpora allows the creation of translation rules that specify the probability of translating words or phrases from one language to the other. Monolingual corpora, containing text only in one language—primarily the target language—are not used to model the translation process, but are used to better model the structure of the target language. Unlike parallel data, which require expensive human translators to generate, monolingual data are cheap and widely available.

Similar topics and events to those in a source document that is being translated often occur in documents in a comparable monolingual corpus. In much the same way that a human translator would use world knowledge to aid translation, the MT system may be able to use these relevant documents from comparable corpora to guide translation by biasing the translation system to produce output more similar to the relevant documents. This thesis seeks to answer the following questions: (1) Is it possible to improve a modern, state-of-the-art translation system by biasing the MT output to be more similar to relevant passages from comparable monolingual text? (2) What level of similarity is necessary to exploit these techniques? (3) What is the nature of the relevant passages that are needed during the application of these techniques?

To answer these questions, this thesis describes a method for generating new translation rules from monolingual data specifically targeted for the document that is being translated. Rule generation leverages the existing translation system and topical overlap between the foreign source text and the monolingual text, and unlike regular translation rule generation does not require parallel text. For each source document to be translated, potentially comparable documents are selected from the monolingual data using cross-lingual information retrieval. By biasing the MT system towards the selected relevant documents and then measuring the similarity of the biased output to the relevant documents using Translation Edit Rate Plus (TERp), it is possible to identify sub-sentential regions of the source and comparable documents that are possible translations of each other. This process results in the

generation of new translation rules, where the source side is taken from the document to be translated and the target side is fluent target language text taken from the monolingual data. The use of these rules results in improvements over a state-of-the-art statistical translation system. These techniques are most effective when there is a high degree of similarity between the source and relevant passages—such as when they report on the same new stories—but some benefit, approximately half, can be achieved when the passages are only historically or topically related.

The discovery of the feasibility of improving MT by using comparable passages to bias MT output provides a basis for future investigation on problems of this type. Ultimately, the goal is to provide a framework within which translation rules may be generated without additional parallel corpora, thus allowing researchers to test longstanding hypotheses about machine translation in the face of scarce parallel resources.

# Improving Statistical Machine Translation Using Comparable Corpora

by

Matthew Garvey Snover

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2010

Advisory Committee:
Professor Bonnie Dorr, Chair/Advisor
Professor Carol Espy-Wilson
Professor William Gasarch
Professor Jimmy Lin
Professor Philip Resnik

For Gaja.

# Acknowledgments

I am deeply indebted to my advisor, Bonnie Dorr, who has guided me throughout my career. She provided me numerous opportunities to grow and learn. Without her patience and support, this thesis might never have been written. Richard Schwartz served as a second advisor to me during the course of my research. He was always willing and eager to look deep into any problems encountered, always with a push to keep the flow of research moving. He was an endless stream of ideas and inspiration.

Beginning in my first undergraduate cognitive philosophy classes through my Master's work and into the final days of my doctoral studies I have benefited from the guidance of many amazing teachers. It was Jesse Prinz who first pushed me towards the mysteries of language. Michael Brent taught me the beauty of probability and computation. The lessons I learned from him have remained with me and shape the ideals of my work. Philip Resnik asked hard questions which forced me to respond with better answers.

I am grateful for the guidance of my committee: Carol Epsy-Wilson, Bill Gasarch, Jimmy Lin, and Philip Resnik. This thesis is better for their questions and advice.

BBN Technologies provided me not only with funding and seemingly endless computational resources during my studies but also with the wisdom and experience of their outstanding MT research group, especially Rich Schwartz, John Makhoul,

Spyros Matsoukas, Jinxi Xu, Antti-Veikko Rosti, Jeff Ma, Mike Kayser, Libin Shen, and Bing Zhang.

One of the greatest resources and source of help to me have been my colleagues at the University of Maryland: Fazil Ayan, Jacob Devlin, Chris Dyer, Vladimir Eidelman, Nizar Habash, Adam Lopez, Nitin Madnani, Anton Rytting, Asad Sayeed, Nathaniel Waisbrot, and David Zajic.

David and Naomi Zajic provided me a home away from home when life pulled me away from Maryland. They opened their home to me and were more generous than I could ever have hoped. Nitin Madnani was my companion along the road to graduation. He has been a good friend and made me feel close to the University of Maryland even when I was far away.

My family's unconditional love and faith has supported me through the long years of graduate school. I am especially thankful to my parents Paul and Lydia Snover.

Throughout the researching and writing this thesis my dog, Ginger, has patiently sat by my side, channeling away my stress and forcing me to occasionally leave the computer behind for the relaxation of a good walk.

Finally, this thesis would not have been possible without the loving support of my wife, Gaja. No words of thanks or love will ever be enough.

This research was supported, in part, by BBN Technologies under the GALE Program, DARPA/IPTO Contract No. HR0011-06-C-0022. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor.

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| CLIR | Cross Lingual Information Retrieval |
| HTER | Human Mediated Translation Edit Rate |
| MT | Machine Translation |
| NIST | National Institute of Standards and Technology |
| NLP | Natural Language Processing |
| SMT | Statistical Machine Translation |
| TER | Translation Edit Rate |
| TERP | Translation Edit Rate Plus |
| WER | Word Error Rate |

# Chapter 1

## Introduction

With thousands of languages in the world, and the increasing speed and quantity of information being distributed across the world, automatic translation between languages by computers, *Machine Translation (MT)*, has become an increasingly important area of research. State-of-the-art MT systems rely not upon hand-crafted translation rules written by human experts, but rather on learned statistical models that translate a *source* language to a *target* language. These models are typically generated from large, parallel corpora containing copies of text in both the source and target languages. The co-occurrence of words across languages in parallel corpora allows the creation of translation rules that specify the probability of translating words or phrases from one language to the other. Monolingual corpora, containing text only in one language—primarily the target language—are not used to model the translation process, but are used to better model the structure of the target language. Unlike parallel data, which require expensive human translators to generate, monolingual data are cheap and widely available.

Similar topics and events to those in a source document that is being translated often occur in documents in a comparable monolingual corpus. In much the same way that a human translator would use world knowledge to aid translation, the MT system may be able to use these relevant documents from comparable corpora to

guide translation by biasing the translation system to produce output more similar to the relevant documents. This thesis seeks to answer the following questions: (1) Is it possible to improve a modern, state-of-the-art translation system by biasing the MT output to be more similar to relevant passages from comparable monolingual text? (2) What level of similarity is necessary to exploit these techniques? (3) What is the nature of the relevant passages that are needed during the application of these techniques?

The usefulness of this can be seen in the example shown in Figure 1.1, where a Chinese news story discusses the Hangzhou Water Treatment Development center. A relevant passage from several months earlier is found, and used to guide the translation. This fixes a mistake where "的 反 渗透 海水 淡化" was poorly translated as "the anti-infiltration of desalination", and causes it to be correctly translated as "The reverse osmosis desalinization".

The remainder of this chapter provides additional motivation for the questions raised in this thesis, outlines the thesis, and discusses the research contributions of this thesis.

## 1.1 Motivation

Statistical machine translation (SMT) systems rely upon mathematical models of language in order to translate sentences in the source language into sentences in the target language. The two core models of SMT systems are the language model and the translation model. The language model is estimated from monolingual text

**Source-Sentence**: 由 国家 海洋 局 杭州 水处理 技术 开发 中心 负责 开发 的 反 渗透 海水 淡化 技术 ， 是 一种 以 压力 为 驱动力 的 膜 分离 过程 ， 是 当今 国际 海水 淡化 研究 领域 的 热点.

**Correct-Translation:** The reverse osmosis sea water desalinization technology developed by the State Oceanic Administration 's Hangzhou Water Treatment Technology Development Center is a pressure-driven film separation process , a hot topic in current international sea water desalinization research.

**Original-Translation:** Hangzhou Water Treatment Technology Development Center of the State Oceanic Administration responsible for the development of the anti-infiltration of desalination technology is a process of pressure into a driving force membrane separation , and the current international hot spots in the field of seawater desalination research.

⇓

**Biased-Translation:** Hangzhou Water Treatment Technology Development Center of the State Oceanic Administration responsible for the development of the reverse osmosis desalination technology is a membrane separation process with the pressure as the driving force , the current international hot spots in the field of sea water desalination research.

**Relevant-Passage**: The reverse osmosis desalination demonstration project , which can treat 10,000 tons of sea water per day, was built by Hangzhou Development Center of Water Treatment of the State Oceanic Administration by utilizing membrane diffusion desalination technology.

Figure 1.1: Example of improving translation by biasing translation towards relevant passages.

in the target language and is used to calculate the likelihood that a string of words is a sentence in the target language. Stated succinctly, the job of the language model is to ensure that the SMT system's translation is fluent. The translation model estimates the probability that the given words in the target language are possible translations of the words in the source language, and generates the various translation hypotheses that the SMT system must decide between. Unlike the language model, the translation model is estimated using parallel or bi-text: text that exists

in two languages, normally with one side being a human translation of the other. Parallel text is a product of human translation and is therefore expensive to generate, especially in comparison to the monolingual text that is used to estimate the language model and can be found in vast quantities online.[1]

An alternative to creating parallel text by translation is to find *naturally* occurring parallel text. The distinction, in this case, is that with *natural* parallel text, the text in each of the two languages is created by native speakers, whereas with *created* parallel text, generally one side of the text is a translation by a non-native speaker of the language. In general, this leads to a lack of fluency in *created* parallel text that is not present in *natural* parallel text.

Finding naturally occurring parallel text is best illustrated by considering the task of translating news stories, one of the most studied genres for translation. Stories of international importance or interest are reported across the world in a multitude of languages. In most cases, these stories are not translations from another language, but are rather a retelling of the same events. Such stories are only partially parallel at the document, or story, level, and could not be used to estimate the translation model of an SMT system, which normally assumes sentence-level parallelism. Stories in the source and target language that report on the same events do contain many overlapping elements at the word and phrasal level, as well as possible statistical similarities. The stories in both languages are likely to contain the same person, places, and other entities in similar relationships or performing similar

[1]While parallel text can be found online, such as when a company presents its website in multiple languages, the data found in this way is very limited in size and suffers from issues of quality control (Resnik and Smith, 2003).

actions. For many of the phrases in the story from one language, there will exist a phrase in the other language's story that is actually a valid translation. Thus we can consider these stories parallel at a sub-sentential level.

This thesis explores several methods for utilizing relevant passages from comparable corpora to improve state-of-the-art machine translation. The focus is not on the task of building new training sets from non-parallel comparable corpora, but rather on extracting new, highly specific translation rules directly related to the source documents being translated—a process dubbed **translation-model adaptation**. In examining this, I also explore a similar method to aid the language model of the translation system—a process dubbed **language-model adaptation**. Both of these techniques seek to bias a statistical machine translation system to produce an output that is more similar to the relevant texts, while still being a valid translation of the source sentence.

This thesis seeks to answer the following questions:

1. Is it possible to improve a modern, state-of-the-art statistical translation system using language-model and translation-model biasing techniques that cause the translation output to be more similar to relevant passages from comparable monolingual text?
2. What level of sub-sentential parallelization is necessary to exploit such techniques?
3. What is the nature of the relevant passages that are needed in applying such techniques?

In response to these questions, I seek to validate the following hypotheses:

1. Improvements to the MT system are possible from both language-model and translation-model biasing techniques.
2. While little sub-sentential parallelization is necessary to exploit language-model adaptation, translation-model adaptation relies upon some level of sub-sentential parallelization consisting of a minimum of a few words of sub-sentential parallelization.
3. Those relevant documents that come from the same time period as the source document and cover the same story are the most useful and provide the greatest benefit for translation, although events that occur at different time periods can still be exploited to a lesser degree.

Answers to these questions allow MT researchers to further explore other important questions in MT by providing a new method to generate translation rules without the need for parallel corpora.

## 1.2 Outline of the Dissertation

Related work in using comparable monolingual corpora to improve machine translation is discussed in Chapter 2. Also discussed are the automatic machine translation evaluation metrics that are used to optimize MT systems and measure improvement over baseline systems.

Chapter 3 begins by defining the comparable corpora used in this research and details the cross-lingual information retrieval (CLIR) algorithm used to select the relevant passages. This is followed by a description of how language-model adap-

tation, combined with the CLIR approach to selecting relevant passages, improves translation quality over a baseline MT system. A novel translation-model adaptation approach is then introduced, wherein short phrasal translations are learned from a combination of source-language documents and relevant passages from comparable monolingual corpora. Although this technique over-generates translation rules, resulting in a large number of incorrect translation rules and a small number of correct translation rules, it can improve statistical machine translation systems where the translation system suffers from out-of-vocabulary words but still has enough coverage to filter out the incorrect translation rules generated. However, as the quality of the translation system improves, this method ceases to provides gains.

Improving upon the basic translation-model adaptation, Chapter 4 introduces a new method for selectively learning phrasal translation rules. Unlike the basic method which learns a large number of short translation rules, this method learns a very small number of translation rules that are generally much longer. The method described for learning these translation rules utilizes both the basic translation-model adaptation and the language-model adaptation methods described in Chapter 3, and also relies upon alignments produced by the TERP evaluation metric, introduced in Chapter 2. These new translation rules are shown to improve state-of-the-art statistical machine translation.

Chapter 5 fully details the new MT evaluation measure called TER-Plus, or TERP, that is used as an alignment tool in Chapter 4. This metric was found to be one of the best performing metrics in the NIST MetricsMATR 2008 Challenge, highly correlating with human judgments of translation quality (Przybocki et al.,

2008; Snover et al., 2009).

Finally, Chapter 6 contains conclusions and discusses future work.

## 1.3 Research Contributions

Through the research conducted in this thesis, I have made the following important research contributions:

- A method for selectively learning new phrasal translation rules without parallel corpora that improves state-of-the-art statistical machine translation. This method learns translations from the source documents to be translated and relevant passages from comparable monolingual text, by exploiting parallelization at a sub-sentential level. This selective translation rule learning relies upon language-model adaptation, basic translation-model adaptation and the use of TERp as an alignment tool.
- A new and simple method for translation-model adaptation using relevant texts from comparable corpora. Portions of this research have been previously published in Snover et al. (2008).
- Verification that language-model adaptation using relevant passages from comparable corpora can be used to improve state-of-the-art statistical machine translation. Portions of this research have been previously published in Snover et al. (2008).
- An automatic metric for machine translation evaluation, TER-Plus (TERp), which demonstrates a high level of correlation with human judgements of

quality—ranking at the top of automatic evaluation metrics at the NIST 2008 MetricsMATR challenge. TERP also provides a method to perform alignment between segments of English text, a feature used elsewhere in this thesis. Both this metric, and its predecessor TER, have been distributed to the NLP community where they have proved useful for both MT evaluation and alignment tasks. Portions of this research have been previously published in Snover et al. (2009) and Snover et al. (2010).

The task of improving state-of-the-art translation is difficult with only slow, gradual progress. Improving translation quality by learning new translation rules is a difficult task, generally reserved only for those languages where little or no parallel data is available. Learning new and useful translation rules in those situations where the MT system is already well developed and trained leaves little room for improvement, requiring new techniques. The work in this thesis leverages the already well trained translation system as well as the topical overlap present in real-world large data situations to learn highly specific but useful translation rules. These new translation rules can then be used to improve state-of-the-art translation quality.

The discovery of the feasibility of improving MT by using comparable passages to bias the output provides a basis for future investigation on problems of this type. Ultimately, the goal is to provide a framework within which translation rules may be generated without additional parallel corpora, thus allowing researchers to test longstanding hypotheses about machine translation in the face of scarce parallel resources.

# Chapter 2

# Related Work

This chapter is divided into two parts, covering previous work related to this thesis. Section 2.1 contains related work in the area of statistical machine translation, while Section 2.2 describes the evaluation metrics used to measure translation quality.

## 2.1 Previous work in Statistical Machine Translation

This section describes previous research investigating the use of comparable corpora to improve translation quality, as well as other relevant adaptation techniques.

### 2.1.1 Language Model Adaptation

The language model adaptation discussed in this thesis follows on Kim and Khudanpur (2003), Zhao et al. (2004), and Kim (2005). Kim (2005) used large amounts of comparable data to adapt language models on a document-by-document basis, while Zhao et al. (2004) used comparable data to perform sentence level adaptation of the language model. These adapted language models were shown to improve performance for both automatic speech recognition as well as machine translation. Kim and Khudanpur (2003) used cross-lingual information retrieval

(CLIR) to retrieve large numbers of comparable documents in the target language. A new language model was generated from these documents and then interpolated with the original language model. The interpolation weight used was selected as to the minimize perplexity of the interpolated model on a tuning set.

The technique used in this thesis for language model adaptation follows the same scheme as Kim and Khudanpur (2003), although there are a number of differences both in execution and in the use of this language model adaptation. The CLIR methods used in this thesis differ dramatically from those used by Kim and Khudanpur (2003), although the adaptation method is agnostic to this choice. Despite the fact that perplexity-based interpolation is a principled method, it is not necessarily ideal for machine translation, where the goal is not to reduce perplexity but to increase translation accuracy. For the MT system used in this thesis, a hand-chosen interpolation weight is shown to produce better than translations than those produced by techniques that use a minimizing weight. Most importantly, because Kim and Khudanpur (2003) used thousands of documents to build the comparable language model, a lesser biasing of the MT system was achieved. By using a smaller number of passages to generate the new language model, ranging from 300 down to a single passage, I create a much stronger biasing effect. Biasing to a very small number of passages can cause a large change in the output of the MT system, a result that is shown to be beneficial for the Selective Translation Model adaptation techniques described in Chapter 4.

This thesis takes language-model adaptation one step further, focusing on the learning of new translation rules from non-parallel comparable corpora. By

learning new translation rules, the capabilities of what translations are possible are extended.

### 2.1.2 Translation Model Weighting

A form of translation-model adaptation that used comparable out-of-domain parallel data was shown by Hildebrand et al. (2005) to yield significant gains over a baseline system. The translation model was adapted by selecting comparable sentences from parallel corpora for each of the sentences to be translated. This requires expensive parallel data whereas the techniques in this thesis adapt the translation model using much cheaper monolingual data. Adaptation using parallel data is much simpler as translation rules can be extracted using the standard rule-extraction techniques that used to train the generic translation-model–a method that is not possible when adaptation uses monolingual data. In addition to selecting out-of-domain data to adapt the translation model, comparable data selection techniques have been used to select and weight portions of the existing training data for the translation model to improve translation performance (Lu et al., 2007). These techniques extend the idea motivating language-model adaptation by applying it to the translation model, effectively re-weighting the parallel data in the same way that language-model adaptation re-weights the monolingual data. While Lu et al. (2007) has shown this to be beneficial for translation quality, this method, unlike the methods discussed in this thesis, does create any new translation rules, but only redistribute the probability mass associated with the existing translation

rules.

### 2.1.3 Using Statistical Co-ocurrence to Learn New Words

While the amount of parallel data available to train a statistical machine translation system is sharply limited, vast amounts of monolingual data are generally available, especially when translating to languages such as English. Yet monolingual data are generally only used to train the language model of the translation system. Previous work (Fung and Yee, 1998; Rapp, 1999) has sought to learn new translations for words by looking at comparable, but not parallel, corpora in multiple languages and analyzing the co-occurrence of words, resulting in the generation of new word-to-word translations. These methods have been shown to generate simple word-to-word translations but have not been used with modern phrasal translation systems or are not expected to be beneficial if a suitable amount of parallel data is already available.

Ji (2009) uses information extraction techniques to build entity relationship maps in multiple languages and then aligns these maps to find possible translations of named entities. This work addresses the co-occurrence of information across languages in a very different way from this thesis. Rather than looking for sub-sentential parallelization, Ji (2009) builds models of information separately in each language and then looks for parallelization between the entity-relationship models. It is unclear that these models can be used to improve translation quality, but as they approach the problem of topical-overlap in a different direction, it is possible this

method could be used in conjunction with the methods proposed in this thesis.

### 2.1.4 Mining Potentially Parallel Sentences

More recently, Resnik and Smith (2003) and Munteanu and Marcu (2005) have exploited monolingual data in both the source and target languages to find document or sentence pairs that appear to be parallel. These newly discovered bilingual data can then be used as additional training data for the translation system. Such methods generally have a very low yield leaving vast amounts of data that are only used for language modeling.

Abdul-Rauf and Schwenk (2009) seek to mine parallel sentences from non-parallel comparable corpora using CLIR—with WER, TER, and TERP as filters—to determine if sentences are parallel. Because Abdul-Rauf and Schwenk (2009) use TER and TERP, work developed in this thesis, they can be seen to built upon contributions we have already produced.

Moving beyond extracting only entire sentences that were potentially parallel, Munteanu and Marcu (2006) examined the extraction of elements of sentences from comparable corpora that were parallel at a sub-sentential level. The sub-sentential phrases were then used as additional training data for a phrasal translation system, although they proved less beneficial than the extracting entire sentences.

All of these methods seek to improve translation in more resource impoverished languages where parallel data are less available. Rather than improving translation directly they seek to build new parallel data, which can then be used to train

traditional translation models.

### 2.1.5 Self-Training

In addition to language model adaptation this thesis examines the modification of the translation model, adding additional translation rules that enable the translation of new words and phrases in both the source and target languages, as well as increasing the probability of existing translation rules. Translation adaptation using the translation system's own output, known as Self-Training (Ueffing, 2006) has previously shown gains by augmenting the translation model with additional translation rules. In that approach however, the translation model was augmented using parallel data, rather than comparable data, by interpolating a translation model trained using the system output with the original translation model.

Self-training does not seek to exploit comparable corpora, but rather seeks to adapt the translation model so that it produces output that is more similar to the best output it has generated in previous iterations. This provides a new method of building parallel training data. Rather than find it from non-parallel sources, the translation generates it directly by translating foreign texts in the target language and treating the high confidence translation as though they were parallel. The result is that target side of the translation is purely a product of machine translation, whereas the methods used in this thesis use natural and fluent text from the source and target languages to generate translation rules.

## 2.2 Previous Evaluation Metrics

Automatic evaluation is one of the major challenges in machine translation (MT), and is itself an active area of research, whereas it is often considered a solved and trivial problem in many other areas. For example, in speech recognition or most machine learning problems, for any given input $i$, there is a single correct output $o$. By contrast, in machine translation there is an effectively unlimited set of correct translations $o = \{o_1, ..., o_\infty\}$. It is impossible to even generate the full set of correct translations, so to serve as a proxy for $o$ when evaluating MT systems, several samples can be drawn from $o$ to generate $o'$ (typically consisting of 1 to 4 translations—dubbed *reference translations*) which then can represent the space of correct translations. This solution is obviously suboptimal, as there will always remain an additional correct translation that was not selected. Thus, if the MT system generates a translation that is not in $o'$, we cannot be sure that it is not a correct translation. Some studies of this sampling of correct translations have shown it to be responsible for over-estimating the error rate of current state-of-the-art SMT systems by approximately 30% (absolute).

The second challenge in automatic MT evaluation is created by the difficulty of the task, or, depending on one's viewpoint, the poor performance of current MT systems. State-of-the-art MT rarely translates any sentence of non-trivial length without some sort of error. To measure progress, it is not useful to simply mark a sentence as correctly or incorrectly translated, but rather it is necessary to give a partial score to indicate how close it is being a correct translation. Humans tasked

with measuring this have shown poor inter-annotator agreement, (Snover et al., 2006; Turian et al., 2003) indicating this to be a difficult task even for humans. When combined together, these two challenges result in a problem where given an $i$ source sentence, a hypothesized $h$ translation, and a small subset of the correct translations, $o'$, the task is to assign a score indicating how similar $h$ is to the correct translations in the unseen set $o$.

The most commonly used and accepted automatic metrics in machine translation, WER, TER, BLEU, and METEOR are presented below. An extended version of the TER metric—called TERP—is briefly described, as this new metric serves as a useful alignment tool in future chapters. A full description of the TERP metric is presented in Chapter 5.

### 2.2.1 Word Error Rate (WER)

One of the first automatic metrics used to evaluate automatic machine translation (MT) systems was Word Error Rate (WER) (Nießen et al., 2000), which remains the standard evaluation metric for Automatic Speech Recognition. WER is computed as the Levenshtein (Levenshtein, 1966) distance between the words of the system output and the words of the reference translation divided by the length of the reference translation. The Levenshtein distance is computed using dynamic programming to find the optimal alignment between the MT output and the reference translation, with each word in the MT output aligning to either 1 or 0 words in the reference translation, and vice versa. Those cases where a reference word is aligned

to nothing are labeled as *deletions*, whereas the alignment of a word from the MT output to nothing is an *insertion*. If a reference word matches the MT output word it is aligned to, this is marked as a *match*, and otherwise is a *substitution*. The WER is then the sums of the number of substitutions (SUB), insertions (INS), and deletions (DEL) divided by the number of words in the reference translation (N) as shown in equation 2.1.

$$\text{WER} = \frac{\text{SUB} + \text{INS} + \text{DEL}}{N} \tag{2.1}$$

WER deals only with a single reference translation, and is referred to as MWER (Multi-Reference WER) (Nießen et al., 2000) when used with multiple references, and is defined as the minimum of the WER scores between the MT output and each reference. In essence, MWER is the WER between the MT output and the closest reference translation. While this allows WER to be used with multiple references, the references are not combined in any fashion and are not truly exploited by the metric.

As mentioned earlier, MT differs from speech recognition in that there are many correct translations for any given foreign sentence. These correct translations differ not only in their word choice but also in the order in which the words occur. Because WER fails to adequately combine knowledge from multiple reference translations and also fails to model the reordering of words and phrases in translation, it is generally seen as inadequate for evaluation for machine translation. This has spurred new directions in machine translation evaluation, producing the metrics described below.

### 2.2.2 Translation Edit Rate (TER)

Translation Edit[1] Rate (Snover et al., 2006) (TER) addresses the issues described above by allowing block movement of words, called *shifts*. within the hypothesis. Shifting a phrase is assumed to have the same *edit cost* as inserting, deleting or substituting a word, regardless of the number of words being shifted. While a general solution to WER with block movements is NP-Complete (Lopresti and Tomkins, 1997), TER computes an approximate solution by using a greedy search to select the words to be shifted, as well as imposing additional constraints on these words. These constraints are intended to simulate the way in which a human editor might choose the words to shift. Other automatic metrics exist that have the same general formulation as TER but address the complexity of shifting in different ways, such as the CDER evaluation metric (Leusch et al., 2006).

The number of edits for TER is calculated in two phases. The number of insertions, deletions, and substitutions is calculated using dynamic programming. A greedy search is used to find the set of shifts, by repeatedly selecting the shift that most reduces the number of insertions, deletions and substitutions, until no more beneficial shifts remain. Note that a shift that reduces the number of insertions, deletions, substitutions by just one has no net reduction in cost, due to the cost of 1 for the shift itself. However, in this case, we still adopt the shift, because we find that the alignment is more correct subjectively and often results in slightly lower edit distance later on. Then dynamic programming is used to optimally calculate the

[1]The TER metric is also occasionally referred to as Translation **Error** Rate in the MT community based upon the abbreviation of WER for Word Error Rate.

remaining edit distance using a minimum-edit-distance (where insertions, deletions and substitutions all have cost 1). The pseudo-code for calculating the number of edits in TER is shown in Algorithm 1.

**Algorithm 1** Calculate Number of Edits in TER

```
input: HYPOTHESIS h
input: REFERENCES R
E ← ∞
for all r ∈ R do
  h' ← h
  e ← 0
  repeat
    Find shift, s, that most reduces min-edit-distance(h', r)
    if s reduces edit distance then
      h' ← apply s to h'
      e ← e + 1
    end if
  until No shifts that reduce edit distance remain
  e ← e+ min-edit-distance(h', r)
  if e < E then
    E ← e
  end if
end for
return E
```

The shifting constraints used by TER serve to better model the quality of translation as well as to reduce the model's computational complexity. Examining a larger set of shifts, or choosing them in a more optimal fashion might result in a lower TER score but it would not necessarily improve the ability of the measure to determine the quality of a translation. The constraints used by TER are as follows:

1. Shifts are selected by a greedy algorithm that chooses the shift that yields the largest reduction in WER between the reference and the hypothesis.
2. The sequence of words shifted in the hypothesis must *exactly match* the se-

quence of words in the reference that it will align with after the shift.

3. The words being shifted, and the matching reference words, must each contain at least one error, according to WER, before the shift occurs. This prevents the shifting of words that are already correctly matched.

When TER is used in the case of multiple references, it does not combine the references, but scores the hypothesis against each reference individually. The reference with which the hypothesis has the fewest number of edits is deemed the closet reference, and that number of edits is used as the numerator for calculating the TER score, as is done in MWER. Rather than use the number of the words in the closet reference as the denominator, TER uses the average number of words across all of the references. Thus the equation for the TER score, where SUB, INS, DEL and SHIFT are the number of substitutions, insertions, deletions and shifts, respectively, and $\bar{N}$ is the average number of reference words, is shown in equation 2.2.

$$\text{TER} = \frac{\text{SUB} + \text{INS} + \text{DEL} + \text{SHIFT}}{\bar{N}} \tag{2.2}$$

### 2.2.3 BLEU

BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002) is the current standard for automatic Machine Translation evaluation. The BLEU score of a system output is calculated by counting the number of n-grams, or word sequences, a maximum length of four words is common, in the system output that occur in the set of reference translations. BLEU is a precision-oriented metric in that it mea-

sures how much of the system output is correct, rather than measuring whether the references are fully reproduced in the system output. BLEU could be gamed by producing very short system outputs consisting only of highly confident n-grams, if it were not for the use of a brevity penalty which penalizes the BLEU score if the system output is shorter than the references.

$$p_n = \frac{\sum_{C \in \{Can\}} \sum_{n\text{-}gram \in C} Cnt_{clip}(n\text{-}gram)}{\sum_{C' \in \{Can\}} \sum_{n\text{-}gram' \in C'} Cnt(n\text{-}gram')} \tag{2.3}$$

$$\text{BP} = \begin{cases} 1, & \text{if } c > r; \\ e^{(1-r/c)}, & \text{if } c \leq r. \end{cases} \tag{2.4}$$

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.5}$$

Equation 2.3 shows the computation of the BLEU precision scores for n-grams of length $n$, where *Can* are the sentences in the test-corpus, *Cnt*(*n-gram*) is the number of times an n-gram occurs in a candidate, and $Cnt_{clip}$(*n-gram*) is the minimum of the unclipped count and the maximum number of times it occurs in a reference translation. Equation 2.4 shows the calculation of the BLEU brevity penalty, where $c$ is the length of the candidate translation, and $r$ is the length of the reference translation. These terms are combined, as shown in equation 2.5, to calculate the total BLEU score, where $N$ is typically 4, and $w_n$ is usually set to $1/N$.

Since its introduction, BLEU has become widespread in the machine translation community and is the most commonly reported evaluation metric. Several

shortcomings of the BLEU evaluation metric have been brought forth by the measure's critics (Callison-Burch et al., 2006; Lavie et al., 2004; Turian et al., 2003). One of the primary critiques of BLEU is absence of recall in its formulations. In addition, BLEU was designed for, and has been shown to work best when used on, large test corpora, such that the scores are averaged over many sentences. BLEU scores of individual sentences are not considered reliable. A number of new automatic evaluation measures for machine translation have been proposed in recent years to compensate for the perceived failings of the BLEU scoring measure. These measures all fundamentally deal with the notion of string matching between reference translations and hypothesized translations.

Despite these criticisms, BLEU remains the most commonly used automatic metric both for the optimization of system parameters and for final evaluation of the quality of an MT system. The use of the BLEU metric has driven development in the MT research community, and it is now the automatic evaluation metric against which all new metrics are compared.

### 2.2.4 METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee and Lavie, 2005) is an evaluation specifically designed to address several observed weaknesses in BLEU. METEOR is a recall-oriented metric, whereas BLEU is generally precision-oriented metric.[2] Unlike BLEU which only calculates precision, METEOR

[2]The brevity penalty in BLEU addresses this issue by penalizing short translation which BLEU would otherwise be unfairly biased towards. Without the brevity penalty, BLEU would be purely a precision-oriented metric.

calculates both precision and recall, and combines the two, as shown in equation 2.6, with a large bias towards recall, to calculate the harmonic mean.[3] In more recent work (Lavie and Agarwal, 2007), higher correlations with human judgments were obtained by optimizing the parameters of the harmonic mean for specific target languages.

$$F_{\text{mean}} = \frac{P \cdot R}{\alpha P + (1-\alpha)R} \tag{2.6}$$

METEOR uses several stages of word matching between the system output and the reference translations in order to align the two strings. The matching stages are as follows:

1. **Exact matching**. Strings which are identical in the reference and the hypothesis are aligned.
2. **Stem matching**. Stemming is performed, so that words with the same morphological root are aligned.
3. **Synonymy matching**. Words which are synonyms according to WordNet (Fellbaum, 1998) are aligned.

In each of these stages, only words that were not matched in previous stages are allowed to be matched. Only unigrams, single words, are compared for matches. Precision in METEOR is defined as number of matches divided by the number of words in the system output, and recall is defined as the number of matches divided by the number of words in the reference.

[3]The default parameters for the harmonic mean set $\alpha = 0.9$.

In addition to the $F_{\text{mean}}$, METEOR also uses a fragmentation penalty to bias the score against system outputs that have many short sequences of consecutive matches, called chunks. Fragmentation is calculated as the number of chunks divided by the number of unigram matches. The fragmentation is calculated as shown in equation 2.7, with default parameters of $\beta = 3.0$ and $\gamma = 0.5$.

$$Pen = \gamma \cdot frag^{\beta} \tag{2.7}$$

This fragmentation penalty causes METEOR to correctly penalize "word salad" MT output that would be allowed under the PER metric, and is an essential portion of the METEOR scoring metric. The final METEOR score is calculated as: $score = (1 - Pen) \cdot F_{\text{mean}}$.

Unlike BLEU, METEOR does not penalize longer answers and incorporates a level of linguistic knowledge in the form of its stem and synonym matching allowing it to identify equivalences between the MT output and the reference translation that would ignored by these earlier measures. METEOR lacks one of BLEU's key features however: the exploitation of multiple references, as METEOR cannot combine knowledge from multiple references into its score. The highly recall-based measure though can be exploited by the inclusion of additional highly likely words (such as "the" in English) in the MT output, giving higher scores to outputs with these additional padded words—although such behavior is not typically exhibited by modern machine translation systems.

### 2.2.5 Translation Edit Rate plus (TERP)

The TER-Plus, or TERP, evaluation metric (Snover et al., 2009) is an extension of the TER metric designed to improve the alignment between the hypothesis and reference and the resulting judgement of translation quality. In addition to aligning words in the hypothesis and reference if they are exact matches, TERP uses stemming and synonymy to allow matches between words. It also uses probabilistic phrasal substitutions to align phrases in the hypothesis and reference. These phrase substitutions are generated by considering possible paraphrases of the reference words. Matching techniques that use stems and synonyms (Banerjee and Lavie, 2005) as well as using paraphrases (Kauchak and Barzilay, 2006; Zhou et al., 2006) have been shown to be beneficial for automatic MT evaluation. Paraphrases have been shown to be additionally useful in expanding the number of references used for evaluation (Madnani et al., 2008) although they are not used in this fashion within TERP. The use of synonymy, stemming, and paraphrases allows TERP to better cope with the limited number of reference translations provided. TERP was one of the top metrics submitted to the NIST Metrics MATR 2008 challenge (Przybocki et al., 2008), having the highest average rank over all the test conditions (Snover et al., 2009).

Because the development of TERP constitutes a major contribution to the framework described in the remainder of this thesis, an entire chapter is dedicated to a detailed description of TERP, its implementation, and the results of comparison with other metrics in Metrics MATR 2008. (See Chapter 5.)

The research described above, in both statistical machine translation and machine translation evaluation, has provided a solid basis for the work provided in this thesis. The work described in this thesis builds upon and goes beyond these foundational pieces of research.

# Chapter 3

# Basic Adaptation

## 3.1 Defining and Selecting Relevant Documents

Statistical machine translation systems generally rely on bilingual data for translation-model training and on monolingual data for language-model training. When translating news stories, or other documents whose content is likely to have appeared in multiple languages, such systems fail to exploit the redundancy of information across texts of various languages. In particular, news stories of international import are likely to have been reported in both the target and source language. Even if the stories in each language are not exact translations of each other, they are likely to contain the same person, places, and other entities in similar relationships or performing similar actions. The stories in the target language can then be used to inform the translation of the source document. This need not apply only to documents where the same story is reported. With the widespread digital distribution of news, past stories on a given situation are very likely have been reported, and will contain information that may be able to inform translation of the source document. Even related stories on similar types of events could be exploited in such a way. These target language documents, while not translations of the source document, are similar enough to the source document to be considered *relevant* or *comparable* documents.

Comparable corpora are further defined in Section 3.1.1 while the probabilistic cross-lingual information retrieval method used for passage selection from comparable corpora is described in Section 3.1.2. It is through the methods described in these sections that passages relevant to the source document are selected.

### 3.1.1 Comparable Corpora

The methods for improving translation quality proposed in this thesis rely upon comparable corpora, that is, multiple corpora that cover the same general topics and events. Comparable documents occur because of the repetition of information across languages, and in the case of news data, on the fact that stories reported in one language are often reported in another language. In cases where no direct translation can be found for a source document, it is often possible to find documents in the target language that are on the same story, or even on a related story, either in subject matter or historically. Such documents can be classified as comparable to the original source document. Phrases within comparable documents are likely to be translations of phrases in the corresponding source documents, even if the documents themselves are not parallel.

Figure 3.1 shows an excerpt of the reference translation of an Arabic document, and Figure 3.2 shows a comparable passage.[1] In this case, the two new stories are not translations of each other and were not reported at the same time—the comparable passage being an older news story—but both discuss actress Angelina Jolie's visit

[1]This is the reference translation of an actual source document from the tuning set used in my experiments, and the first of a number of similar passages found by the comparable text selection system described in section 3.1.2. All examples in this thesis come from actual experimental results.

```
Cameras are flashing and reporters are following up, for Hollywood star
Angelina Jolie is finally talking to the public after a one-month stay
in India, but not as a movie star.  The Hollywood actress, goodwill
ambassador of the United Nations high commissioner for refugees, met with
the Indian minister of state for external affairs, Anand Sharma, here
today, Sunday, to discuss issues of refugees and children.  ...  Jolie,
accompanied by her five-year-old son, Maddox, visited the refugee camps
that are run by the Khalsa Diwan Society for social services and the high
commissioner for refugees Saturday afternoon after she arrived in Delhi.
Jolie has been in India since October 5th shooting the movie "A Mighty
Heart," which is based on the life of Wall Street Journal correspondent
Daniel Pearl, who was kidnapped and killed in Pakistan.  Jolie plays the
role of Pearl's wife, Mariane.
```

Figure 3.1: Excerpt of Example Reference Translation of an Arabic Source Document

to India. Many phrases and words are shared between the two, including: the name of the movie, the name and relationship of the actress' character, the name and age of her son and many others. Such a pairing is extremely comparable, although even less related document pairs could easily be considered comparable.

Parallel or bilingual documents are pairs of documents that are sentence-by-sentence translations of each other. There are a multitude of possible ways of translating a document from one language to another, but—depending on the task for which the translation is done—there are various degrees of translation accuracy. For some tasks, it is desirable to preserve the tone of a source document when translating it, whereas for other tasks it is sufficient to have a translation that merely preserves the bare facts of the source document while being fluent in the target language. In parallel data, a sentence in one language will translate into a sentence (or sometimes a pair of sentences) in the other language. The alignment of sentences

Actress Angelina Jolie hopped onto a crowded Mumbai commuter train Monday to film a scene for a movie about slain journalist Daniel Pearl, who lived and worked in India's financial and entertainment capital. Hollywood actor Dan Futterman portrays Pearl and Jolie plays his wife Mariane in the "A Mighty Heart" co-produced by Plan B, a production company founded by Brad Pitt and his ex-wife, actress Jennifer Aniston. Jolie and Pitt, accompanied by their three children -- Maddox, 5, 18-month-old Zahara and 5-month-old Shiloh Nouvel -- arrived in Mumbai on Saturday from the western Indian city Pune where they were shooting the movie for nearly a month. ...

Figure 3.2: Excerpt of Example Comparable Document

between the two languages is exploited by statistical machine translation systems to generate translation rules from one language to the other.

Comparable documents differ from parallel documents in several respects. No assumption is made that there is any correspondence between the sentences in the source document and the sentences in the comparable document. Sentences and facts in the source document will often be completely unaccounted for in the target document, and additional sentences and facts will be present in the target document. The exact definition of comparable document will often vary from task to task, but for the purposes of this work they must have certain properties in the common, namely the entities and places involved in the actions of the story, and the events that occur in the story. An ideal comparable document is about the same event and the same people, but is not parallel. Differences in some of the entities and places, as well as differences in the story, would not prevent the documents from being comparable. The document may still be considered comparable even if the story is unrelated but very similar. Commonly, a comparable document will not

be about the same story as that of the source document, but will be about an earlier, related story. Much of the information between the source and comparable document is repeated in these historically related stories, and the people, places and events described are often the same or very similar. Thus there are various degrees of comparability.

Execution of a Saudi After His Conviction for Murdering One of His Fellow Citizens Riyadh 2/7 (AFP) - The Saudi Interior Ministry announced in a report the implementation of the death penalty today, Tuesday, in the area of Medina (West) of a Saudi citizen convicted of murdering a fellow citizen. The report issued by the Saudi News Agency has revealed that Ghazi bin Ruwaydi bin Salih al-Jabiri (a Saudi National) blatantly murdered 'Ubayd bin 'Atiqallah bin 'Ubayd Al-Jabiri (a Saudi National) " by shooting him with a rifle, injuring him, and causing his death as a result of a dispute between them." The report added that the investigation of the perpetrator resulted in "leveling charges against him of committing his crime and referring him to the General Shari'a Court, which issued a statuary record confirming the validity of the charges brought against him and he was sentenced to death. The decision was authenticated by the Supreme Court and by the permanent Supreme Judicial Council." And this is the first execution to be announced in Saudi Arabia this current year. Saudi Arabia witnessed the execution of 83 individuals in 2005 in contrast to 53 in 2004 and 25 in 2003, according to statistics prepared by the France Press Agency based on official Saudi reports. In Saudi Arabia, individuals charged with rape, murder, apostasy, or armed robbery, as well as drug smuggling, are given the death penalty.

Figure 3.3: Sample Source Document Translation

Figure 3.3 displays the reference translation of a source document, and Figure 3.4 displays the passage, from the English Gigaword corpus and the FBIS corpus, deemed most relevant, by a comparable data selection algorithm, described below in Section 3.1.2. The two stories were published years apart but bear significant similarities, though they are clearly not translations of the same source document

```
A Saudi citizen convicted of shooting a compatriot to death was beheaded
by the sword in the Mecca region Monday , the interior ministry said.
Rajeh bin Ahmad bin Mohammad al-Yacoubi was found guilty of murdering
Awad bin Mohammad bin Ali al-Yacoubi after a row and sentenced to death,
said a ministry statement quoted by the official SPA news agency.  The
beheading took to 24 the number of executions announced in Saudi Arabia
this year, according to an AFP tally based on official statements.
Executions generally take place in public in the conservative kingdom
which applies a strict form of sharia , or Islamic law , imposing
the death penalty for murder, rape, apostasy, armed robbery and drug
trafficking.  At least 48 people were executed in Saudi Arabia in
2002.
```

Figure 3.4: Sample Comparable Document

and are not parallel. Both stories describe the execution of a Saudi citizen, and contain many of the same phrases. In this case, some overlap of people and places occurs in a similar event, but the story is not quite the same, e.g., the first one focuses on the beheading of Ghazi bin Ruwaydi bin Salih al-Jabiri whereas the second one focuses on the execution several years previously of Rajeh bin Ahmad bin Mohammad al-Yacoubi.

### 3.1.2 Selecting Relevant Passages Using CLIR

In the implementation developed for this thesis, relevant passages are selected for every source document from a large monolingual corpus in the target language. In practice, one could search the World-Wide-Web for documents that are relevant to a set of source documents, but this approach presents problems for ensuring the quality and formatting of the retrieved documents. The experiments in this thesis use relevant text selected from a collection of English news texts that are considered

comparable to the source documents. Because these texts are all fluent English, and of comparable genre to the test set, they are also used for training the standard language-model training.

The selection of comparable or relevant texts in one language, given a query in another language, is a problem that has been widely studied in the information retrieval community and cross-lingual information retrieval (CLIR) (Levow et al., 2005; Oard and Dorr, 1998). The implementation developed for this thesis uses CLIR to select a ranked list of documents in the target language (English). In the experiments described below, the source-language document that we wish to translate is viewed as a query in the CLIR framework; the result is a set of comparable target-language documents.

The CLIR problem can be framed probabilistically as: Given a query $Q$, find a document $D$ that maximizes the expression $\Pr(D \text{ is relevant}|Q)$. This expression can be expanded using Bayes' Law as shown in equation 3.1. The prior probability of a document being relevant can be viewed as uniform, and thus in this work, we assume $\Pr(D \text{ is relevant})$ is a constant.[2] The $\Pr(Q)$ is constant across all documents. Therefore finding a document to maximize $\Pr(D \text{ is relevant}|Q)$ is equivalent to finding a document that maximizes $\Pr(Q|D \text{ is relevant})$.

$$\Pr(D \text{ is relevant}|Q) = \frac{\Pr(D \text{ is relevant}) \Pr(Q|D \text{ is relevant})}{\Pr(Q)} \tag{3.1}$$

[2]In fact, it can be beneficial to use features of the document to estimate $\Pr(D \text{ is relevant})$ (Miller and Schwartz, 1998) but this has not explored in this work.

A method of calculating the probability of a query given a document was proposed by Xu et al. (2001)[3] and is shown in Equation 3.2. In this formulation, each foreign word, $f$, in the query is generated from the foreign vocabulary with probability $\alpha$ and from the English document with probability $1 - \alpha$, where $\alpha$ is a constant.[4] The probability of $f$ being generated by the general foreign vocabulary, $F$, is $\Pr(f|F) = \mathit{freq}(f, F)/|F|$, the frequency of the word $f$ in the vocabulary divided by the size of the vocabulary. The probability of the word being generated by the English document is the sum of the probabilities of it being generated by each English word, $e$, in the document which is the frequency of the English word in the document, $(\Pr(e|D) = \mathit{freq}(e, D)/|D|)$ multiplied by the probability of the translation of the English word to the foreign word, $\Pr(f|e)$.

$$\Pr(Q|D) = \prod_{f \in Q} (\alpha \Pr(f|F) + (1 - \alpha) \sum_{e} \Pr(e|D) \Pr(f|e)) \tag{3.2}$$

This formulation favors longer English documents over shorter ones. In addition, many documents cover multiple stories and topics. For the purposes of adaptation, shorter, fully relevant documents are preferred to longer, only partially relevant documents. Because of this, all of the documents in the monolingual data are divided into overlapping passages of approximately 300 words in length, with

[3]Xu et al. (2001) formulated this for the selection of foreign documents given an English query. We reverse this to select English documents given a foreign query.

[4]As in Xu et al. (2001), a value of 0.3 was used for $\alpha$.

divisions occurring only at sentence boundaries, so that no sentences are broken into two passages, ensuring that the final passages are fluent. The length of 300 was chosen as this was approximately the same length as the source documents.[5] Documents that were originally shorter than 300 words in length were not divided. These passages were used as the documents to be searched in the CLIR system.

For each source document, the CLIR system returns a ranked list of passages, of which the top $N$ are used, with the value of $N$ varying for different applications. These top $N$ passages are not guaranteed to be relevant and are often largely unrelated to the story or topic in the source document. The set of passages selected by the CLIR system is dubbed as the *bias text* to differentiate it from relevant text, as the adaptation methods use this text to *bias* the MT system so that its output will be more similar to the bias text.

While experiments have not been conducted using other CLIR systems, the adaptation methods presented in this thesis could be applied without modification using another CLIR system, as the adaptation method treats the CLIR system as a black box. The algorithm of Xu et al. (2001) is used without any significant modification, including the use of a stop word list for both the English and foreign texts. The parameters for $\Pr(f|F)$ and $\Pr(f|e)$ were estimated using the same parallel data on which our translation system was trained.

The bias texts selected by CLIR from the comparable data provide the basis for the adaptation techniques discussed in this thesis.

[5]Passage lengths of 100 or 200 words were also explored, but the use of these shorter passages proved less beneficial for later adaptation steps.

## 3.2 Model Adaptation

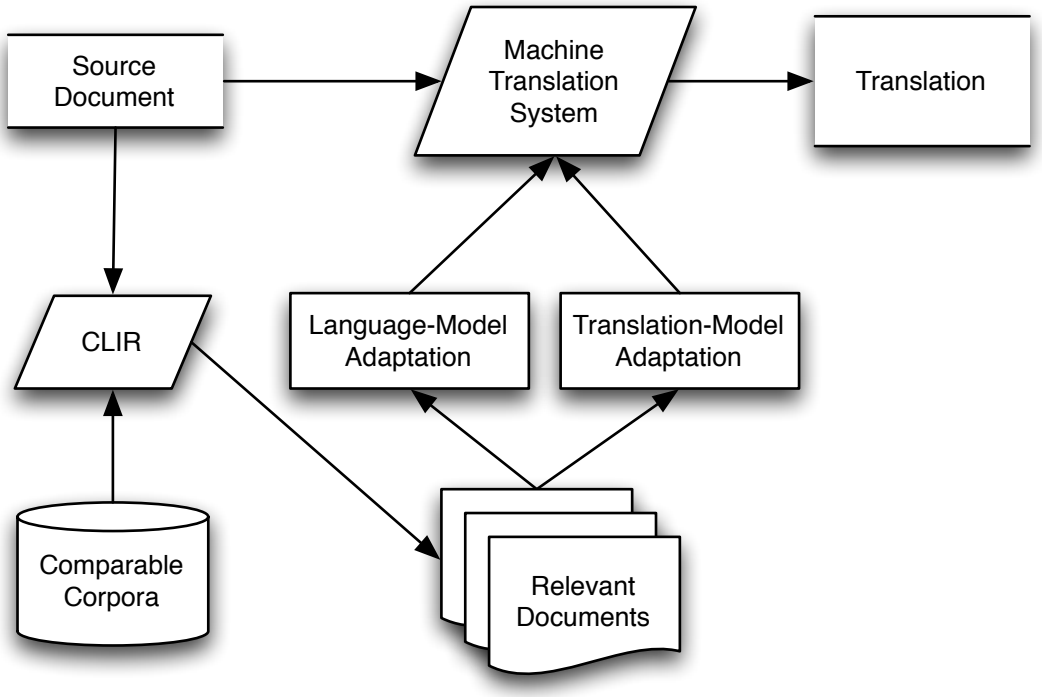



Figure 3.5: Flowchart illustrating the incorporation of language-model (LM) and translation-model (TM) adaptation into a statistical machine translation system.

We use the same bias text to adapt both the language model and the translation model. For language-model adaptation, we increase the probability of the word sequences in the bias text, and for translation-model adaptation we use additional phrasal translation rules. Figure 3.5 shows the integration of both adaptation methods into the statistical machine translation system. The adaptations can be done independently and while they can augment each other when used together, this is not required. It is not necessary to use the same number of passages for both forms of adaptation, although doing so makes it more likely both that the English side of the new translation rule will be assigned a high probability by the adapted

language model, and that the translation model produces the English text to which the language model has been adapted. Bias text that is used by one adaptation but not by the other will receive no special treatment by the other model. This could result in new translation rules that produce text to which the language assigns low probability, or it could result in the language model being able to assign a high probability to a good English translation that cannot be produced by the translation model due to a lack of necessary translation rules.

While both adaptation methods are integrated into a hierarchical translation model (Chiang, 2005), they are largely implementation independent. Language-model adaptation could be integrated into any statistical machine translation that uses a language model over words, while translation-model adaptation could be added to any statistical machine translation that can utilize phrasal translation rules.

### 3.2.1 Language-Model Adaptation

For every source document, we estimate a new language model, the *bias language model*, from the corresponding bias text. Since this bias text is short, the corresponding bias language model is small and specific, giving high probabilities to those phrases that occur in the bias text. The bias language model is interpolated with the *generic language model* that would otherwise be used for translation if no LM adaptation was used. The new bias language model is of the same order as the generic language model, so that if a trigram language model is used for the MT

decoding, then the biased language model will also be a trigram language model. The bias language model is created using the same settings as the generic language model. In our particular implementation however, the generic language model uses Kneser-Ney smoothing, while the biased language model uses Witten-Bell smoothing due to implementation limitations. In principle the biased language model can be smoothed in the same manner as the generic language model.

We interpolate the bias language model and the generic language model as shown in equation 3.3, where $\Pr_g$ and $\Pr_b$ are the probabilities from the generic language model and the bias language model, respectively. A constant interpolation weight, $\lambda$ is used to weight the two probabilities for all documents. While a value for $\lambda$ could be chosen that minimizes perplexity on a tuning set, in a similar fashion to Kim (2005), it is unclear that such a weight would be ideal when the interpolated language model is used as part of a statistical translation system. In practice we have observed that weights other than one that minimizes perplexity, typically a lower weight, can yield better translation results on the tuning set.

$$\Pr(e) = (1 - \lambda) \underset{g}{\Pr}(e) + \lambda \underset{b}{\Pr}(e) \tag{3.3}$$

The resulting interpolated language model is then used in place of the generic language model in the translation process, increasing the probability that the translation output will resemble the bias text. It is important to note that, unlike the translation-model adaptation described in section 3.2.2, no new information is added to the system with language-model adaptation. Because the bias text is extracted

from the same monolingual corpus that the generic language model was estimated from, all of the word sequences used for training the bias language model were also used for training the generic language model. Language-model adaptation only increases the weight of the portion of the language-model data that was selected as comparable.

### 3.2.2 Basic Translation-Model Adaptation

It is frequently the case in machine translation that unknown words or phrases are present in the source document, or that the known translations of source words are based on a very small number of occurrences in the training data. In other cases, translations may be known for individual words in the source document, but not for longer phrases. Translation-model adaptation seeks to generate new phrasal translation rules for these source words and phrases. The bias text for a source document may, if comparable, contain a number of English words and phrases that are the English side of these desired rules.

Because the source data and the bias text are not translations of each other and are not sentence aligned, conventional alignment tools, such as GIZA++ (Och and Ney, 2000), cannot be used to align the source and bias text. Because the passages in the bias text are not translations of the source document, it will always be the case that portions of the source document have no translation in the bias text, and portions of the bias text have no translation in the source document. In addition a phrase in one of these texts might have multiple, differing translations in

the other text.

Unlike language-model adaptation, the entirety of the bias text is not used for translation adaptation. We extract those phrases that occur in at least $M$ of the passages in the bias texts. A phrase is only counted once for every passage in which it occurs, so that repeated use of a phrase within a passage does not affect whether it used to generate new rules. Typically, passages selected by the CLIR tend to be very similar to each other if they are comparable to the source document and are very different from each other if they are not comparable to the source document. Phrases that are identical across passages are the ones that are most likely to be comparable, whereas a phrase or word that occurs in only one passage is likely to be present only by chance or in cases where the passage it is in is not comparable. Filtering the target phrases down to those that occur in multiple passages therefore serves not only to reduce the total number of rules, but also to filter out phrases from passages that are not comparable.

For each phrase in the source document we generate a new translation to each of the phrases selected from the bias text, and assign it a low uniform probability.[6] For each translation rule we also have a lexical translation probability that we estimate correctly from the trained word model. These new rules are then added to the phrase table of the existing translation model when translating the source document. Rather than adding probability to the existing generic rules, the new rules are marked as bias rules by the system and given their own feature weight. While

[6]A probability of 1/400 is arbitrarily used for the bias rules although it is then weighted by the bias translation rule weight.

the vast majority of the bias translation rules are incorrect translations, the translation system will naturally be biased against these incorrect rules. If the source side of a translation rule already has a number of observed translations, then the low probability of the new bias rule will cause it to not be selected by the translation system. If the new bias translation rules would produce garbled English, then the language model will be biased against the target side of those bias rules. When translation-model adaptation is combined with language-model adaptation, a natural pressure is exerted to use the bias rules for source phrases primarily when the resulting output would look more like the bias text.

Consider the example of bias translation-model adaptation presented in Figure 3.6, where the source text discusses a campaign to make the ancient city of Petra one of the new seven wonders of the world. The MT system has difficulties translating the Arabic word "البتراء" for "Petra" and so attempts to use a bias translation rule. Samples of sentences from the top 10 comparable documents are also presented in Figure 3.6. These are highly comparable sentences, also discussing the possibility of Petra and other sites becoming wonders of the world. Extracting just those phrases that occur multiple times in the bias text results in 1332 different phrases, of which only 17 contain the word Petra. These 17 English phrases are listed below:

**Source:** اعلنت هيئة تنشيط ال سياحة فى الاردن فى بيان ل ها امس عن اطلاق حملة وطنية فى عموم ال بلاد اعتبارا من غد ال اثنين تهدف الى حث ال ناس على ال تصويت ال اثرية ل تصبح احدى عجائب ال دنيا ال سبع ال جديدة البتراء ل صالح مدينة ال حسب تقرير وكالة ال صحافة ال فرنسي

**Correct-Translation:** In a statement yesterday, the Tourism Board in Jordan announced the launching of a national campaign all around the country starting tomorrow, Monday, with the goal of urging the people to vote in the favor of making the ancient city of Petra one of the new seven wonders of the world, according to a report by Agence France Presse.

**Sample-Comparable-Sentences:**

1. China 's Great Wall , the Stonehenge monoliths in England and the desert city of Petra in Jordan are among 21 candidate sites to be named the new seven wonders of the world , organizers said Tuesday.

2. LISBON -- The Great Wall of China , Rome 's Colosseum , and India 's Taj Mahal were among seven architectural marvels named the new wonders of the world yesterday. The other four winners , chosen by a global poll , were Peru 's Machu Picchu , Brazil 's Statue of Christ Redeemer , Jordan 's Petra , and Mexico 's Chichen Itza pyramid .

3. A privately funded organisation , the New 7 Wonders Foundation , has put forward a shortlist of 21 landmarks from across the globe. They include Rome 's Colosseum , Jordan 's ancient city of Petra , Britain 's Stonehenge and the Great Wall of China .

4. The winners in an internet and text - message contest that attracted 70 million were : the Chichen Itza pyramid in Mexico , the Christ Redeemer statue in Rio de Janeiro , the Great Wall of China, the Inca city of Machu Pichu in Peru , the Petra site in Jordan , the Colosseum in Rome and the Taj Mahal in India.

5. The proposed 21 sites include the Petra in Jordan , the Statue of Liberty in the United States , the Eiffel Tower in Paris , the Opera House in Sydney , Stonehenge Fort in Britain , Taj Mahal in India , Timbuktu Fort in Mali.

Figure 3.6: Example Arabic and English sentences for Basic Translation-Model Adaptation

(1) [petra ;] (2) [petra in] (3) [the petra] (4) [mahal and petra] (5) [petra site] (6) [petra ,] (7) [the petra site] (8) [petra] (9) [petra site in] (10) [city of petra] (11) [of petra] (12) [and petra] (13) [jordan 's petra] (14) [petra in jordan] (15) [, the petra] (16) [of petra in] (17) ['s petra]

There are 18 words in the source segment, resulting in approximately 51 Arabic different phrases. Taking the combination of all English phrases to all Arabic phrases yields 67,932 translation rules for this sentence. The MT system however generally discards those that do not have "البتراء" as the source side as there already translation rules that have a much higher probability for those words, eliminating the possibility of bias translation rule working in such a situation. This still leave 1332 translation rules, only 17 of which contain the word "Petra" . The language-model adaptation provides a strong filter in this case, eliminating those phrases that are unlikely according to the language model. The final rule that the system chooses and uses is show is (1). Although this is not perfectly correct, "Petra" might have been preferred according to the references, it is still a possible translation ("Petra" is sometimes referred to as "the Petra" in English, as shown in the comparable sentences.).

(1) البتراء ⇒ the Petra

The final resulting translation uses only this single bias translation rule and is shown below:

the tourism activation authority in jordan declared in a statement yesterday the release of the national campaign in all the country starting from tomorrow monday aimed at urging people to vote in favor of the ancient city of " the petra " to become one of the new seven wonders of the world , according to the report of the french press agency.

## 3.3 Experimental Results

The performance of our language- and translation-model adaptation approach was evaluated against a MT system baseline (described in Section 3.3.1) under two conditions, the details of which are presented in section 3.3.2. One condition involved a small amount of parallel training, such as one might find when translating a less commonly taught language (LCTL). The other condition involved the full amount of training available for Arabic-to-English translation. In the case of LCTLs we expect our translation model to have the most deficiencies and to be the most in need of additional translation rules. So, it is under such a condition we would expect the translation-model adaptation to be the most beneficial. We evaluate the system's performance under this condition in section 3.3.3. The effectiveness of this technique on state-of-the-art systems and its efficiency when used with a well trained generic translation model are presented in section 3.3.4.

### 3.3.1 Baseline Statistical Machine Translation System

The HierDec MT system (Shen et al., 2008) was used as a baseline MT system and as the foundation for evaluating the adaptation techniques. HierDec is a hierarchical translation system with string-to-dependency rules. Both Arabic-to-English and Chinese-to-English language pairs were examined. A trigram language model was used during decoding, and a 5-gram language model was used to re-score the n-best list after decoding.

All conditions were optimized using Bleu (Papineni et al., 2002) and evaluated

using both BLEU and Translation Edit Rate (TER) (Snover et al., 2006). BLEU is an accuracy measure, so higher values indicate better performance, while TER is an error metric, so lower values indicate better performance. Optimization was performed on a tuning set of newswire data, comprised of portions of MTEval 2004, MTEval 2005, and GALE 2007 newswire development data, a total of 48921 words of English in 1385 segments and 173 documents. Results were measured on the NIST MTEval 2006 Arabic Evaluation set, which was 55578 words of English in 1797 segments and 104 documents. Four reference translations were used for scoring each translation.

Parameter optimization was done using n-best optimization. The MT decoder was run on the tuning set generating an n-best list (where $n = 300$), on which all of the translation features (including bias rule weights) were optimized using Powell's (Powell, 1964) method. These new weights were then used to decode again, repeating the whole process, using a cumulative n-best list. This continued for several iterations until performance on the tuning set stablized. The resulting feature weights were used when decoding the test set. A similar, but simpler, method was used to determine the feature weights after 5-gram rescoring.

### 3.3.2 Implementation Details

Both language-model and translation-model adaptation were implemented on top of the HierDec system described in section 3.3.1. While generalized rules were generated from the parallel data, rules generated by the translation-model adapta-

tion were not generalized and were used only as phrasal rules. In addition to the features described in Shen et al. (2008), a new feature was added to the model for the bias rule weight, allowing the translation system to effectively tune the probability of the rules added by translation-model adaptation in order to improve performance on the tuning set.

Bias texts were selected from three monolingual corpora: the English Gigaword corpus (2,793,350,201 words), the FBIS corpus (28,465,936 words), and a collection of news archive data collected from the websites of various online, public news sites (828,435,409 words). All three corpora were also part of the generic language-model training data. Language-model adaptation on both the trigram and 5-gram language models used 10 comparable passages with an interpolation weight of 0.1. Translation-model adaptation used 10 comparable passages for the bias text and a value of 2 for $M$.

Each selected passage contains approximately 300 words, so in the case where 10 comparable passages were used to create a bias text, the resulting text was 3000 words long on average. The language models created using these bias texts were very specific giving large probability to n-gram sequences seen in those texts.

The construction of the bias texts increased the overall run-time of the translation system, although in practice this was a small expenditure. The most intensive portion was the initial indexing of the monolingual corpus, but this was only required once and could be reused for any subsequent test set that was evaluated. This index could then be quickly searched for comparable passages. When considering research environments, test sets were used repeatedly and bias texts only need to be built

once per set, making the building cost negligible. Otherwise, the time required to build the bias text was still small compared to the actual translation time.

This n-best optimization method had subtle implications for translation-model adaptation. In the first iteration, few bias rules were used in decoding the 300-best, and those that were used frequently help, although the overall gain was small due to the small number of bias rules used. This caused the optimizer to greatly increase the weight of the bias rules, causing the decoder to overuse the bias rules in the next iteration causing a sharp decrease in translation quality. Several iterations were needed for the cumulative n-best to achieve sufficient diversity and size to assign a weight for the bias translation rules that resulted in an increase in performance over the baseline. Alternative optimization methods could likely circumvent this process. Language-model adaptation did not suffer from this phenomenon.

### 3.3.3 Less Commonly Taught Language Simulation

In order to better examine the nature of translation-model adaptation, a translation model that was trained on only 5 million words of parallel Arabic-English text was investigated. Limiting the translation-model training in this way simulated the problem of translating less commonly taught languages (LCTL) where less parallel text is available, a situation that is not the case for Arabic. Since the model was trained on less parallel data, it lacked a large number of translation rules, which was expected to be addressed by the translation-model adaptation. By working in an environment with a more deprived baseline translation model, we were giving the

translation-model adaptation more room to assist.

The experiments described below used a 5 million word Arabic parallel text corpus constructed from the LDC2004T18 and LDC2006E25 corpora. The full monolingual English data were used for the language model and for selection of comparable documents. Unless otherwise specified no language-model adaptation was used.

I first established an upper limit on the gain for translation-model adaptation, using the reference data to adapt the translation system. These reference data were considered to be extremely comparable, better than one could ever hope for with comparable-document selection. I first aligned this data using GIZA++ to the source data, simulating the ideal case where I could perfectly determine which source words translate to which comparable words. Because the translation-model adaptation system assigns uniform probability to all bias rules, I ignored the correct rule probabilities that we could extract from word alignment and assign uniform probability to all of the bias translation rules. As expected, this gave a large gain over the baseline.

I also examined limiting these new translation rules to those rules whose target side occurred in the top 100 passages selected by CLIR, thus minimizing the adaption to those rules that it theoretically could learn from the bias text. On average, 50% of the rules were removed by this filtering, resulting in a corresponding 50% decrease in the gain over the baseline. The results of these experiments and an unadapted baseline are shown in Table 3.1.

The fair translation-model adaptation system, however, does not align source

| Test Set | TM Adaptation | TER | BLEU |
|---|---|---|---|
| Tune | None | 49.84 | 40.80 |
| | Aligned Reference | 36.92 | 58.41 |
| | Overlapping Only | 41.79 | 51.38 |
| MT06 | None | 55.16 | 34.68 |
| | Aligned Reference | 45.17 | 52.16 |
| | Overlapping Only | 48.99 | 43.35 |

Table 3.1: LCTL Aligned Reference Adaptation Results

phrases to the correct bias text phrases in such a fashion, and instead aligns all source words to all target words. To investigate the effect of this over-production of rules, I again used the reference translations as if they were comparable data, but we ignored the alignments learned by GIZA++, and instead allowed all source phrases to translate to all English phrases in the reference text, with uniform probability. This still showed large gains in translation quality over the baseline, as measured by TER and BLEU. Again, I also examined limiting the text used for translation-model adaptation to those phrases that occurred in both the reference text and the top 100 comparable passages selected the CLIR system. While this decreased performance, the system still performs significantly better than the baseline, as shown in the following Table 3.2.

Applying translation-model and language-model adaptation fairly, using only bias text from the comparable data selection, yielded smaller gains on both the tuning and MT06 sets, as shown in Table 3.3. The combination of language-model and translation-model adaptation exceeded the gains that were achieved over the baseline by either method separately.

| Test Set | TM Adaptation | TER | BLEU |
|---|---|---|---|
| Tune | None | 49.84 | 40.80 |
| | Unaligned Ref. | 44.92 | 45.66 |
| | Overlapping Only | 48.08 | 43.13 |
| MT06 | None | 55.16 | 34.68 |
| | Unaligned Ref. | 52.54 | 39.90 |
| | Overlapping Only | 53.90 | 36.95 |

Table 3.2: LCTL Unaligned Reference Adaptation Results

| Test Set | Adaptation | TER | BLEU |
|---|---|---|---|
| Tune | None | 49.84 | 40.80 |
| | LM | 49.22 | 41.40 |
| | TM | 49.16 | 41.69 |
| | LM & TM | 48.88 | 42.44 |
| MT06 | None | 55.16 | 34.68 |
| | LM | 0.5559 | 34.90 |
| | TM | 55.45 | 34.78 |
| | LM & TM | 55.09 | 35.36 |

Table 3.3: LCTL Fair Adaptation Results

### 3.3.4 Full Parallel Training Results

While the simulation described in section 3.3.3 used only 5 million words of parallel training, 230 million words of parallel data from 18.5 million segments were used for training the full Arabic-to-English translation system. This parallel data included the LDC2007T08 "ISI Arabic-English Automatically Extracted Parallel Text" corpus (Munteanu and Marcu, 2007), which was created from monolingual corpora in English and Arabic using the algorithm described by Munteanu and Marcu (2005). This choice of corpus allowed the exploration of a more realistic

scenario, as the techniques used in that work are separate and independent from the adaptation methods described in this thesis.[7] Language-model adaptation and translation-model adaptation were applied both independently and jointly to the translation system, and the results were evaluated against an unadapted baseline, as shown in Table 3.4.

While gains from language-model adaptation were substantial on the tuning set, on the MT06 test set they were reduced to a 0.65% gain on BLEU and a negligible improvement in TER. The translation-model adaptation performed better with 1.37% improvement in BLEU and a 0.26% improvement in TER. This gain increases to a 2.07% improvement in BLEU and a 0.64% improvement in TER when language-model adaptation was used in conjunction with the translation-model adaptation, showing the importance of using both adaptation methods. While it could be expected that a more heavily trained translation model might not require the benefit of language and translation-model adaptation, a more substantial gain over the baseline could be seen when both forms of adaptation were used than in the case with less parallel training—a difference of 2.07% BLEU versus 0.68% BLEU. Improvements of 1% or more (absolute) in TER or BLEU are generally considered substantial.

Of the comparable passages selected by the CLIR system for the MT06 test set in the full training experiment, 16.3% were selected from the News Archive corpus, 81.2% were selected from the English GigaWord corpus, and 2.5% were selected

[7]The two methods are not directly comparable, and so we do not make any attempt to do so. Munteanu and Marcu (2005) creates new parallel corpora from two monolingual corpora. This new parallel data is generally applicable for training a translation model but does not target any particular test set. This adaptation method does not generate new parallel data, but creates a new, specific translation model for a test document that is being translated.

| Test Set | Adaptation | TER | BLEU |
|---|---|---|---|
| Tune | None | 43.39 | 46.61 |
| | LM | 42.27 | 48.57 |
| | TM | 43.51 | 46.57 |
| | LM & TM | 42.45 | 48.82 |
| MT06 | None | 51.46 | 38.52 |
| | LM | 51.40 | 39.17 |
| | TM | 51.20 | 39.89 |
| | LM & TM | 50.82 | 40.59 |

Table 3.4: Full Training Adaptation Results

from the FBIS corpus. A slightly different distribution was found for the Tuning set, where 17.8% of the passages were selected from the News Archive corpus, 77.1% were selected from the English GigaWord corpus, and 5.1% were selected from the FBIS corpus.

## 3.4 Limitations of Basic Translation-Model Adaptation

The basic translation-model adaptation technique suffers from several factors that limit its usefulness.

First, the number of bias rules added on a per-document basis was quite large and resulted in a large increase in the memory required to run the decoding step of the MT system. While the system was still runnable in the experiments described above, this memory requirement poses problems for future scaling up of the basic translation-model adaptation or for using this technique on computationally limited devices. This problem could be alleviated by pruning down the number of bias

translation rules, but doing so risks removing those rules that contain novel word translations. Distinguishing novel and good translation rules from garbage translation rules is very difficult as there as the prior probability of a bias being incorrect is very high, and the prior probability probability of a correct novel translation is very low. The use of surrounding context words can be used to distinguish the two cases though as will be shown in the next chapter.

Second, MT system tuning in the context of translation-model adaptation was found to be exceedingly difficult. When tuning an MT system, one does not begin from a set of completely neutral weights; instead one uses the best weights found in previous versions of the system, or even from other language pairs. While improvements in tuning methods might prove beneficial here, such as the MIRA optimization method (Chiang et al., 2009), the true problem is the lack of discriminative features used to distinguish the bias translation rules. Because of the uniform conditional probability of the translation rule, only the lexical probability is usable to distinguish the rule. This puts too much pressure on the language model and other components of the system to determine which bias rules to use and which to discard.

In the next chapter we explore solutions to both of these problems.

## 3.5 Improvements in Statistical Machine Translation

The research in this thesis was conducted over a period of several years on an actively improving state-of-the-art machine translation system: BBN Technologies'

HierDec translation system. The improvements in the system include the improvement of the tuning of discriminative features from Powell's method (Powell, 1964) to Expected-BLEU tuning (Devlin, 2009), increasing the number of discriminative features that can be tuned from a few dozen to thousands of features. These improvements to the system are reflected in the increasing quality of the baseline system in the next chapter.

The basic translation-model adaptation has shown benefits for Arabic translation, although it is difficult to use and does not benefit Chinese translation. This approach can be suitable for filling in gaps in the translation model where no suitable rules exist, it is generally limited by the overwhelming mass of incorrect translation rules.

# Chapter 4

## Selective Translation-Model Adaptation

In this chapter we present a refinement of the basic translation-model adaptation presented in the previous chapter. Unlike basic translation-model adaptation, *selective translation-model adaptation* does not exhaustively generate bias translation rules for all phrase pairings, but instead generates such rules only if the source words could generate a translation that is similar to the phrase in the relevant passage. This *selective translation-model adaptation* dramatically reduces the number of biased translation rules generated, allows the use of much longer translation rules, and generates biased translation rules that are much better translations. Selective translation-model adaptation provides gains to the translation system even in those cases when the basic translation-model adaptation ceases to be beneficial.

Selective translation-model adaptation is introduced in Section 4.1 with the procedure for selectively generating bias rules presented in Section 4.2. Selective translation-model adaptation can provide translations for new words as shown in Section 4.3. Experimental results showing the benefit of these bias translation rules are presented in Section 4.4.

## 4.1 Increasing Translation-Model Selectivity

Basic translation-model adaptation suffers from a vast over-production of low-quality translation rules. This can be addressed by attempting to filter out those rules that are poor translations by leveraging off of words that we already know how to translate. Only accepting biased translation rules where some of the words are already known translations is an unusable strategy for translation rules that are only a few words long. By filtering out poor translation rules however we can open up the possibility of generating much longer translation rules, which can contain a balance of unknown and known words. Even if all of the words in such a biased translation rule are known already, the rule introduces new phrases that may prove useful to the translation system. This chapter discusses such an approach to improving translation-model adaptation by increasing the selectiveness of the translation-model adaptation.

Selective translation-model adaptation takes the research in a new direction that utilizes my previous work in basic translation-model and language-model adaptation as well as my research into MT evaluation with TER-Plus (TERP). Translation-model adaptation is augmented by adding, for each source document to be translated, new bias translation rules to the set of existing rules. Cross-Lingual Information Extraction is used to find relevant passages from comparable monolingual (English) data. Previously, bias rules were generated from all possible combinations of the phrases from the source document with phrases from the relevant passages that occurred two or more times. Our new approach aligns each sentence from the

relevant passages with a translation of each source segment using the alignment capabilities of the TERp evaluation metric, described briefly in Section 2.2.5.

Using the previous techniques, bias rules are generated from regions of high overlap between the two segments, with more matching encouraged by a strong bias of the translation system towards the relevant passage. Initial results show that using these bias translation rules increases the IBM BLEU score of the Chinese Newswire Tune and Test sets by 1.0 and 0.8, respectively.

## 4.2 Selecting Bias Rules

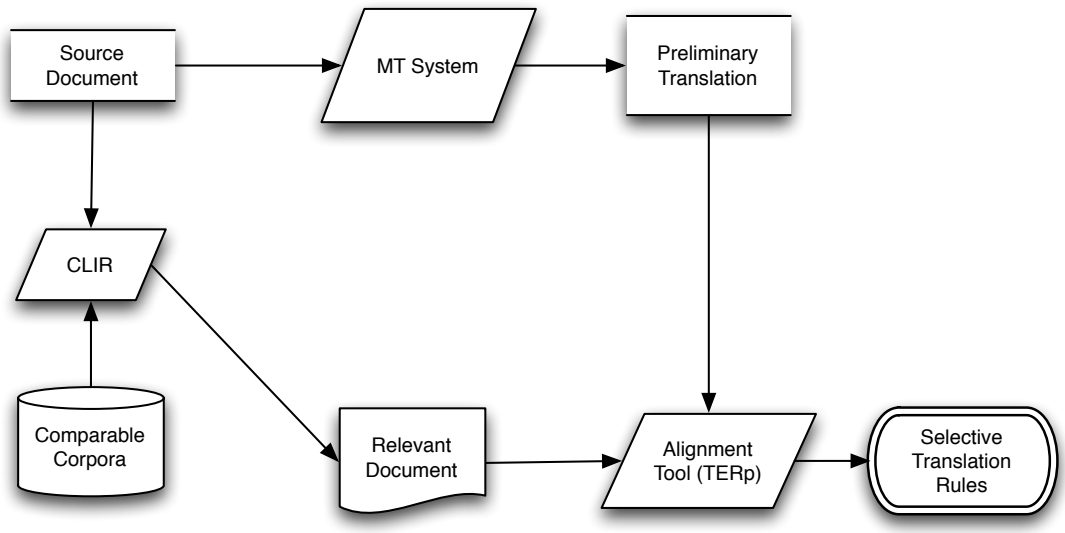



Figure 4.1: Flowchart illustrating the generation of bias rules using selective translation-model adaptation, without biasing the MT system.

The process of generating bias translation rules can be broken down into the following steps:

1. For every source document to be translated, $s \in S$, we select a ranked list

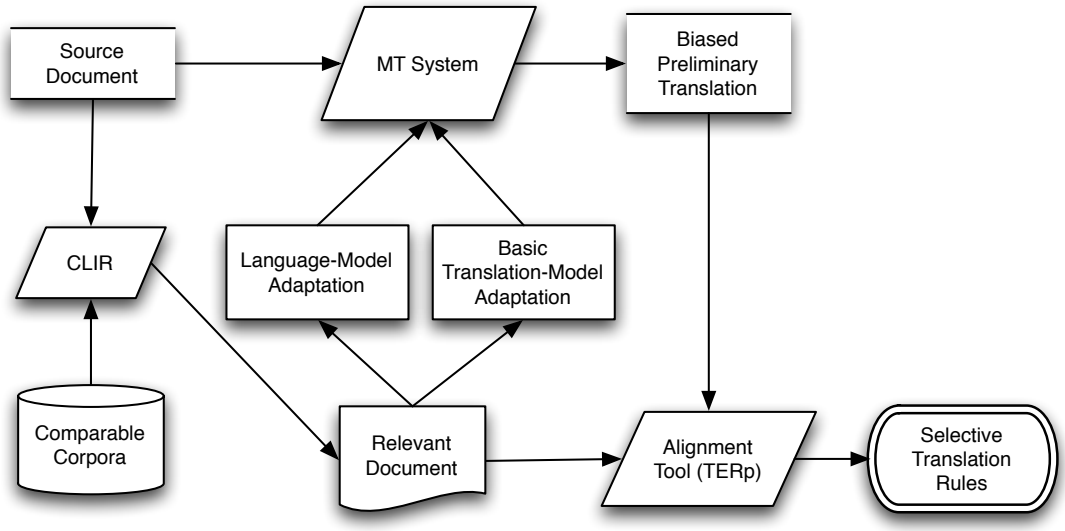



Figure 4.2: Flowchart illustrating the generation of bias rules using selective translation-model adaptation, using language-model and translation-model adaptation to generate a biased preliminary translation.

of relevant passages, $R = \{r_1, r_2, ..., r_n\}$, from the monolingual English data using the CLIR system. This is the same process described in Chapter 3.

2. For each of the top English passages selected $r_i \in R$, we generate a series of biased MT outputs, $B_i$. Each of these is an output of MT system that has been biased to generate an output that more strongly resembles the relevant passage $r_i$. Numerous biased MT outputs are generated with varying degrees of biasing for later contrastive steps. These biased MT outputs are generally much worse translations than the baseline system.

3. Each biased MT output in $B_i$ is then compared with the relevant passage $r_i$ using TERp to determine regions where there is a high degree of similarity. Recall from Section 2.2.5 that TERp extends TER in that it adds stemming, synonymy, paraphrasing, and other improvements. In addition, TERp outputs

a TER-style alignment between the two strings in addition to a score indicating the translation quality. If regions of the two passages that are similar enough are found, then we assume that the source words in $s$ that generated that region of the biased MT output and the region of the relevant passage $r_i$ may be translations of each other, and a phrasal translation rule is generated for the two with discriminative features based on the similarity of the alignment. These new translation rules are then added to the MT system as biased translation rules. The process of aligning the biased MT output to the relevant passage is detailed in section 4.2.1, while the discriminative features used with the new rules are discussed in section 4.2.2.

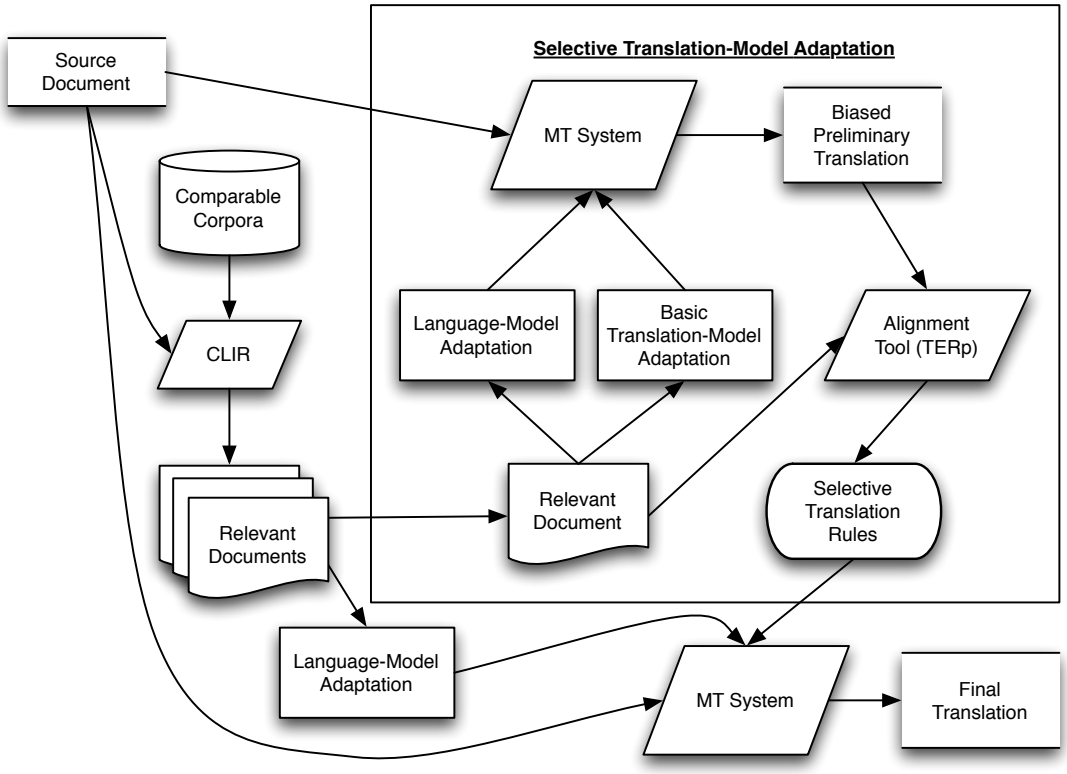



Figure 4.3: Flowchart illustrating the integration of selective translation-model adaptation into the machine translation pipeline.

A flowchart describing the process without biasing the MT output is shown in Figure 4.1, while the process with the biasing of the MT output is shown in Figure 4.2. These flowcharts only indicate the process by which the selective translation rules are generated. The integration of this process in the overall MT system is shown in Figure 4.3.

### 4.2.1 Aligning Biased MT Output to Relevant Passages

Bias translation rules are extracted from each pairing of source document relevant comparable passage according to the following procedure:

1. For each of the top English passages selected $r_i \in R$, we decode the source document $s$ with a strong bias to $r_i$. This strong bias is done using both language-model adaptation and basic translation-model adaptation to the single passage $r_i$. The language-model adaptation uses a high interpolation weight, such as 0.3, while the basic translation-model adaptation generates phrasal rules from all 1-2 word source phrasal to all 1-2 word target phrases. The basic translation-model adaptation is used to ensure the translation system could, in theory, generate the comparable text. Other than adding weights for the simple translation adaptation, the weights of MT system are unchanged. This causes the decoder to output a translation that is more similar to $r_i$, although it is not generally a better translation than the baseline system. We refer to this new translation in later steps as the *Biased MT Output*, or $b_i$.
2. For each of the sentences in $b_i$, we examine the alignment of the source to target

| | | Source | | | | |
|---|---|---|---|---|---|---|
| | | 精神 | 号 | 探测 | 车 | 是 |
| **Biased MT** | The | **X** | | | | |
| | Spirit | **X** | | | | |
| | Mars | | **B** | **B** | | |
| | rover | | | | **B** | **B** |

Figure 4.4: Example Alignment of Snippet of Biased MT Output to Snippet of Chinese Source. **B** and **X** indicate words are aligned using biased or regular translation rules, respectively.

words generated by our translation system, and select *Snippets* of the source segment and the biased MT output such that the source side snippet and the biased MT snippet are contiguous substrings of the source or biased MT sentences, and the source snippet was translated to the biased MT snippet. In particular, all of the target words generated by the words in the source snippet must be in the biased MT snippet and all of the source words that generated the words of the biased MT snippet must be in the source snippet. An example of this alignment is shown in Figure 4.4.

3. For each of these snippet pairs, we align the biased MT snippet to each of the sentences in the comparable passage to determine whether there is a portion of the comparable sentence that is very similar to the biased MT snippet. This approach is based on the belief that if such a substring of the comparable sentence exists, then it is a likely translation of the source snippet, and might serve as a candidate for the target side of a bias translation rule with the source snippet as the source side. This alignment is further broken down into the following steps:

| Biased MT | Relevant: Mars | Relevant: Spirit | Relevant: rover |
|---|---|---|---|
| The | | | |
| Spirit | | **M** | |
| Mars | **M** | | |
| rover | | | **M** |

Figure 4.5: Example Alignment of Biased MT Output to Relevant Text. **M** indicate words are exact matches in TERP.

(a) We use the TERP evaluation metric to perform the alignment between the biased MT output and the comparable sentence, treating the biased MT output as the system output and the comparable sentence as the reference translation. An example of the TERP alignment between the biased MT output and the relevant text is shown in Figure 4.5.

(b) The comparable sentence is then divided into various possible substrings or snippets, such that all of the words in the snippet are contiguous, in an attempt to maximize the number of words in the comparable snippet that are matched in both the comparable and biased MT snippets. This results in a set of triples of source, biased MT, and comparable snippets. Each of these comparable snippets is then realigned to the biased MT snippet because the TERP alignment between the comparable and biased MT snippets will possibly have changed due to removing portions of the comparable sentence.

(c) These triples of snippets are then evaluated to determine if enough of the words in the biased MT snippet and the comparable snippet match each other, and if the TERP score of the alignment is low enough to warrant

keeping the triple (a low TERP score indicates better alignment). A cutoff of 0.75 is used to threshold the TERP score, although we additionally require that at least 25% of the biased MT words are matched and at least 50% of the comparable segment is matched. Additional filters are also used to prune down the number of snippet triples.

| | | **Source** | | | | |
|---|---|---|---|---|---|---|
| | | 精神 | 号 | 探测 | 车 | 是 |
| **Relevant** | Mars | | **B-M** | **B-M** | | |
| | Spirit | **X-M** | | | | |
| | rover | | | | **B-M** | **B-M** |

Figure 4.6: Example Alignment of Relevant Text to Chinese Source.

(d) For each resulting triple, a bias translation rule is generated between the source snippet and the comparable snippet. The biased MT snippet is not used directly in the resulting rule. An alignment between the source and comparable words is generated by following the path of the alignment from the source words to the biased MT words to the TERP alignment of biased MT words to the comparable words. An example of the projected alignment from the relevant snippet to the comparable text is shown in Figure 4.6.

The selective translation-model adaptation process has many free parameters that serve as heuristics to reduce the computational complexity, limit the number of bias rules generated and guide the adaptation process. A list of these parameters and the values used in the experiments for this thesis are presented in Table 4.1.

### 4.2.2 Selective Translation Rule Discriminative Features

For each bias MT rule, features are generated representing the quality of the rule using the TERP score, the alignment from the biased MT snippet to the comparable snippet, and the lexical translation probability of the source words to the comparable words. In addition, features are added for the CLIR score of the original bias passage. If multiple rules are generated for a source document with the same source and target side, these rules are combined into a single rule using the features from the rule with the lowest TERP score. An additional feature is used to indicate the number of these duplicates that were found. While the results described below utilize bias rules with about 65 bias rule features, only slightly worse results were obtained using only 7 of these features.

The bias translation rules are added to the MT system on a document-by-document basis, so that bias rules generated for one source document are not used when translating other source documents. These rules are not combined with the generic hierarchical translation rules that were learned from parallel rules, but are used as additional phrasal rules. The generic rules have feature values of 0 for all bias rule features, and the bias rules have feature values of 0 for all generic rule features. Using expected-BLEU tuning, the MT system is optimized to learn appropriate feature weights for both the generic and bias rule features.

## 4.3 Biased Translation Rules with New Words

Translation-model adaptation enables the MT system to learn translation rules for new phrases and new words. In most cases, the translation system has rules for translating all of the individual words but does not have translation rules for longer phrases. The selective translation-model adaptation adds such rules enabling long phrases to be translated together in a fluent manner. Selective translation model, like basic translation-model adaptation, can also add translations for new words that were not previously in the translation dictionary. In some cases the source words, or the desired target language words, were never seen before in training, and in other cases, all of the words have known translations, but a particular translation that is desired is missing.

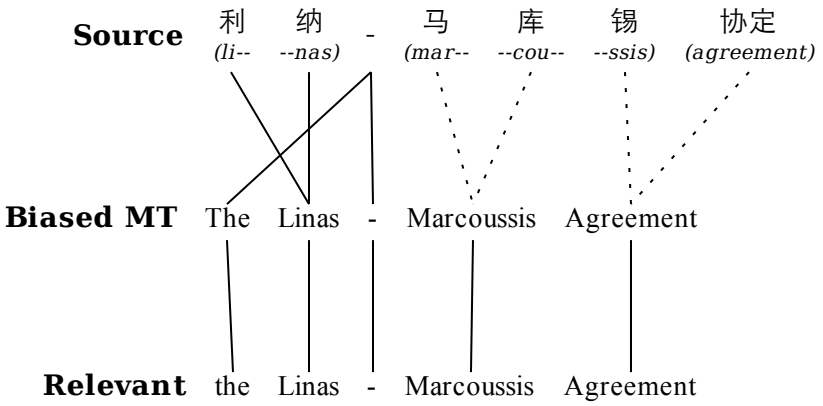



Figure 4.7: Alignment with new lexical translations using basic translation-model bias rules (dashed lines) in Chinese-to-English translation

There are two ways that selective translation-model adaptation allows the learning of new words. The first method is through the use of bias translation rules, from basic translation-model adaptation, to generate the words of the biased MT

output. An example of this type is shown Figure 4.7, where the dashed lines between the Source and the Biased MT text indicate that a biased translation rule, from basic translation-model adaptation, was used to translate the source word to the English word, and undashed lines indicate that a generic translation rule was used. In this example, the Chinese characters "马 库 锡" have been incorrectly segmented into three different words, as the MT system does not have the unsegmented word in its vocabulary. The baseline MT system incorrectly translates the Chinese as "Linas-Kumar-suk agreement," incorrectly translating "马 库 锡 (Marcousis)" as "Kumar - suk" (as well as incorrectly omitting "the"). Because no good translation was available from the MT system, the biased MT generation used bias rules to translate the three characters. In this case, although a correct translation was known for "协定 (agreement)," a bias translation rule was used so as to accommodate the third chinese character for "Marcousis," "." When TERP compares the biased MT output to the relevant passage that was used to bias the MT output, it finds an exact match, and therefore generates the bias translation rule (1).

(1) 利 纳 - 马 库 锡 协定 ⇒ the Linas-Marcoussis Agreement

It should be noted that the reference translation contains the word "Accord" rather than "Agreement," although "Agreement" is arguably an equally good translation, especially as the latter term is to used in the name of the peace accord in native English language news stories, as shown by its presence in the English monolingual text.

The second method that selective translation-model adaptation uses to learn new words does not rely on bias translation rules, but rather allows errors in the bias MT output that are then accepted by TERP. An example of new lexical translations using acceptance by TERP is shown in Figure 4.8. These new bias translation rules tend to be much longer as the errors in the biased MT output must be balanced by a number of correct matches for TERP to reach an acceptable translation error score; the example presented in Figure 4.8 has a TERP score, between the biased MT output and the relevant text snippet, of 0.364.

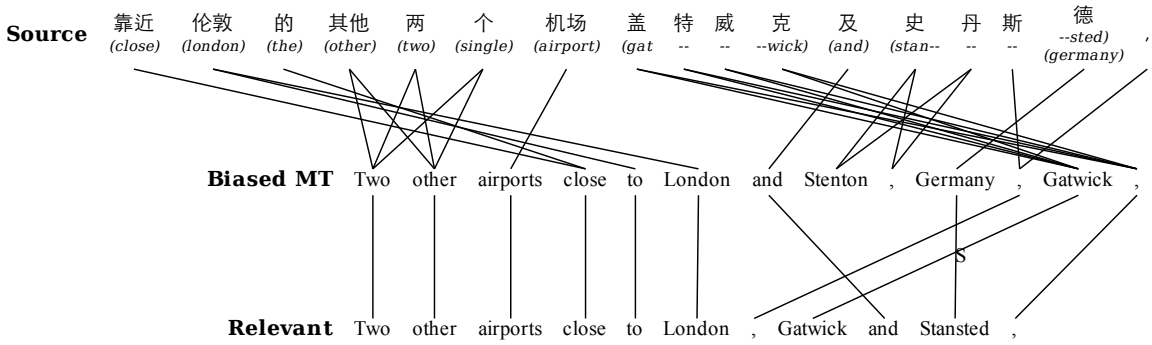
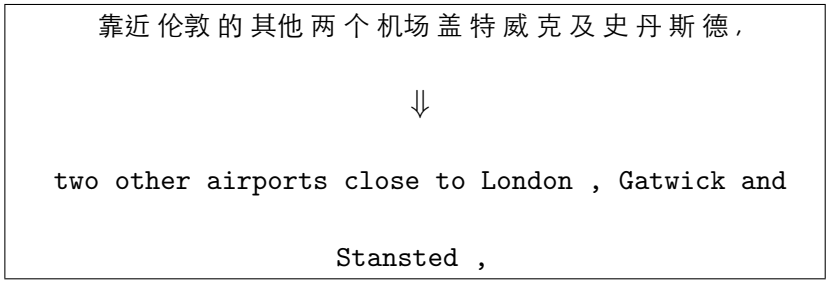



Figure 4.8: Alignment with new lexical translations using TERP edits in Chinese-to-English translation

In this example, the MT system does not know how to translate "史丹斯德" as "Stansted". Biased translation rules fail to generate the correct phrase as well, instead generating two words "Stenton" and "Germany." When the biased MT snippet is compared to the relevant passage there is a high degree of similarity between the two strings, despite the shifting of some words, the insertion of "Stenton" and a comma, as well as the substitution of "Germany" with "Stansted." Because the remaining words match, the rule is accepted, although the discriminative features

assigned to rule will indicate the lack of a perfect match. In this case, the new biased rule is a perfect translation and exactly matches the reference translation.

(2)




The two methods presented above allow the selective translation-model adaptation procedure to generate new translation rules that possess new word-to-word translations that were not originally seen in training.

## 4.4 Experimental Results

To evaluate the effectiveness of the translation-model adaptation, a baseline system that included language-model adaptation (using the top 100 passages from CLIR)—but no translation-model adaptation—was compared to the same system, with the addition of the selective translation-model adaptation technique described above. The systems were optimized to maximize BLEU on the Chinese Newswire Tune set and were tested on the Chinese Newswire Test set. BLEU and TER scores for these systems are shown in Table 4.2. While the BLEU score increased by 1 point on the Tune set, and 0.8 points on the Test set, the TER score remained relatively unchanged.

A total of 43,596 bias rules were learned for the two sets, although only 1048 of these rules, or 2.4%, were used in the 1-best answer. Of the 5,378 segments in

these sets, 905 segments, or 16.83%, used biased translation rules. These biased rules had target sides that were, on average, 7.73 words long, with 6% of rules being 20 or more words long.

## 4.5 Improvements to Bias Translation Rules Using Additional Features

This section provides an overview of the set of discriminative features used by the MT system to separate beneficial bias rules from those that are detrimental to translation, as well as other improvements to translation-model adaptation.

The previous section reported results where bias translation rules were generated using intermediate bias MT output that was generated by biasing source documents to each of 10 comparable passages. To produce these rules, a fixed bias language-model interpolation of 0.3 was used for the language-model adaptation and a weight of 1.5 was used for the basic translation-model rules. The basic translation-model rules all have uniform probability of $\frac{1}{400}$, so the weight of 1.5 in the log-linear model corresponds to applying an exponent of 1.5 to this probability. We take this approach a step further, exploring the use of other weights, and in particular using the output of all resulting translations as intermediate forms to aid in the extraction of bias translation-model rules. We used biased language-model interpolation weights of 0.0 (no language-model adaptation), 0.1, 0.3, and 0.5. We also used basic translation-model adaptation weights of 1.5 and 6.0, as well as disabling basic translation-model adaptation. Because the basic translation adaptation weight is

applied in a log-linear model to a rule probability of $\frac{1}{400}$, the higher weight of 6.0 corresponds to the use of fewer basic bias translation rules.

All combinations of these weights were used to generate the biased MT output with the exception of using the translation-model adaptation when no language-model biasing was used as translation-model adaptation without language-model adaptation is of limited benefit. This resulted in 10 permutations of weights used for biased MT output. New discriminative features were added to the decoder for the level of biasing used to generate the biased MT output, both in the language-model adaptation and the basic translation-model adaptation. Binary features for each level were used (such as a binary feature indicating if a language-model interpolation weight of 0.3 had been used), as well as real-valued features for the actual language-model interpolation weight and the basic translation-model adaptation weight. By examining the weights of these features, and the number of rules extracted from each of these conditions versus the number of rules used, we saw that the MT system strongly favored those bias rules that used the basic translation-model adaptation at a weight of 1.5, and disfavored the lack of use of language-model adaptation. It is inconclusive which interpolation weight for the language-model adaptation is most preferred, although the weight of 0.3 appears to be slightly preferred in the majority of cases.

The features of the bias rules were also expanded to include the CLIR score of the retrieved passage (as well as a number of functions of the score to reflect its rank and how it compared to other passages). Higher ranked passages were generally more useful to the translation-model system than lower ranked passages. The

results indicated that the vast majority of bias rules were obtained from the top 5 passages, with very few being generated by the 10th passage. We thus conclude that using more passages from the CLIR system will likely result in rapidly diminishing returns.

Another feature that was added was one corresponding to whether the comparable passage came from the same time period as the source document. For some comparable documents, only the year of publication was available, and for others a range of months was specified. For over 95% of the documents, however, we could determine the month that the comparable text was published. Because the comparable documents span a time period before, during, and after our tuning and test documents, we can determine for almost all cases whether the source document was published in the same month as our comparable document. Binary feature were added to indicate whether the comparable text was from the past, present or future of the source document, as well as integer-valued features indicating the distance in months in the future and past. Naturally, rules extracted from passages from the same month as the source document were preferred and given a higher weight, although a large number of rules were extracted and used from passages that originated in the past or future of the source document.

One of the primary features used previously was the number of TERP edits between the biased MT output and the comparable passage. This was expanded to include the number of contiguous edits of type, up to a length of 3, so that the decoder could discriminate on whether a long series of insertions or deletions had been present, or whether there was a long series of matches. This expanded to

include the number of edits for just words in a stop list, just upper-cased words, and words of certain part-of-speech categories.

In addition, we used the Charniak parser (Charniak and Johnson, 2005) was used to parse both the biased MT output and the selected comparable passages. Features were added to indicate whether the snippets selected were headed by a single constituent, and what the type of that constituent was. A binary feature was also added for whether the heads of the constituents of the two snippets are identical. Very few extracted rules are headed by a single constituent on either the biased MT output side or the comparable text, but those that are were assigned a higher weight and were used a disproportionately high amount of the time.

The bias rules used a total of 168 features, in comparison to our previously reported results, which utilized only 65 features. The automatic scores from both of these conditions, as well as those of the baseline system, are shown in Table 4.3. Note that the baseline system did not utilize translation-model adaptation, but did use language-model adaptation.

The gain over the baseline system on the Test validation set increased by 0.31 Bleu points and by 0.27 Ter points, for a total gain of 1.10 Bleu points and 0.29 Ter points.

## 4.6 Examining Date Overlap

The monolingual text from which relevant passages are drawn from an overlapping time period as the source documents to be translated. This reflects real-world

translation as we can always acquire new monolingual data that overlap in time with the source data that we wish to translate. Examining the issue of date overlap however allows us to approximate the issue of event similarity, bringing into question how similar the source- and target-language stories must be for the selective translation-model adaptation to prove effective. If we use date overlap as a surrogate for event similarity, then we can examine whether this translation-model adaptation is beneficial only when the source- and target-language news stories are on the same event or if the technique is beneficial when the events are only more distantly related.

Exact dates for the source and relevant documents are not available, although for each document we know the year and month it was published,[1] allowing us to determine if a source document and a relevant document came from the same time period. We can then divide the relevant passages into those documents that come from the same month as the source document and those that come from either before or after the source document. We can then generate bias translation rules from just those relevant passages in the top ten passages that meet either criteria.

The extent of the publication date overlap can be seen in Table 4.4, which shows the percentage of the top ten relevant passages selected by CLIR that overlap in publication date with their corresponding source document, as well as the percentage of bias rules that are generated for each condition. While only 26.52% of the selected passages overlap overlap with the source document publication date, these

[1]For a small number of documents in the monolingual data we know a range of few months in which is was published, although we do not know the exact month it was published. For the purposes for these experiments we assume that the source document was published in any of those months then the documents overlap in time period.

passages are responsible for 50.39% of the bias translation rules that are learned by system. This trend is reversed in the case of relevant passages that are older than the source document. More bias rules are therefore generated for those documents that overlap in date than are generated when the relevant document is from the past or future.

Table 4.5 shows the effect of date overlap on the selective-adaptation process. Identical language-model adaptation was used in all four experimental results, regardless of the date of the relevant documents. In the *No Overlap* condition rules were only generated from documents that did not come from the same time period as the source document, whereas only those that did come from the same time period were used in the *Date Overlap* condition. All documents were used to generate rules for the *All Dates* condition. The gains from using documents that did not overlap in date were much smaller than using those that did overlap in dates, although the gain from using both sets of documents slightly exceeded the gains from using the date overlapping documents alone. This indicates that while the selective translation-model adaptation works best when there is date overlap it provides some smaller gains when there is no date overlap between the source and relevant texts. Note the results in Table 4.5 differ from those in Table 4.3, due to differences in the baseline system over time.

Below I present sets of sentences from the source document, reference translation and relevant passage sentences that illustrate the pairing of sentences that result in bias rules that were used in translation. The portion of the source sentence and the relevant sentence that comprised the biased translation rule are underlined. The

examples are divided in those that overlapped in date, and those that did not.

### 4.6.1 Examples without Date Overlap

The following examples from Chinese-to-English translation do not overlap in publication date. All rules were used in the final output of the MT system.

(3)

**Source:** 前 反抗 军 " 民主 保卫 军 " 对此 举 表示 欢迎, 称赞 这 是 迈向 建立 新 武装部队 "重要 的 一步".

**Reference Translation:** This work is chiefly related to the ‘‘Forces for the Defense of Democracy,’’ the largest of the 6 rebellious military organizations that signed the truce.

**Relevant Sentence:** Presidents, including Pierre Buyoya of Burundi, will evaluate the ceasefire talks between the transitional government and the National Council for Defense of Democracy - Forces for Defense of Democracy (CNDD - FDD) and PALIPEHUTU - Forces for National Liberation (PALIPEHTU - FNL).

Example (3) shows a translation rule that learned a proper noun phrase, “Forces for Defense of Democracy” when it was used in a different and unrelated news story. This can be contrasted to the baseline system which produced “Army Defence of Democracy” for the same phrase. It should be noted that the bias translation rule is not perfect however as it lacks the “the” before “Defense,” which is used in the reference translation. While this results in a single edit difference for TER, it results in a much larger number of missed n-grams for BLEU. The bias rule is still strongly preferred by automatic metrics compared to the baseline.

(4)

**Source:** 由 国家 海洋 局 杭州 水处理 技术 开发 中心 负责 开发 的 反 渗透 海水 淡化 技术，是 一种 以 压力 为 驱动力 的 膜 分离 过程，是 当今 国际 海水 淡化 研究 领域 的 热点

**Reference Translation:** Responsible personnel from the center said that through technological improvement and equipment localization, the cost and power consumption of the reverse osmosis sea water desalinization project have both been slashed.

**Relevant Sentence:** The reverse osmosis desalination demonstration project , which can treat 10,000 tons of sea water per day, was built by Hangzhou Development Center of Water Treatment of the State Oceanic Administration by utilizing membrane diffusion desalination technology.

Example (4) is another case of a proper noun translation that occurs in a different news story. In this case, the baseline translation was "anti-infiltration and seawater desalination."

(5)

**Source:** 今年 march ， 陷入 亏损 的 福特 以 848000000 美元 出售 了 pag 旗下 的 阿 斯顿 马丁 ； 上月，它 又 表示 考虑 出售 路 虎 和 捷 豹.

**Reference Translation:** In March this year , a money-losing Ford sold PAG 's Aston Martin for US $848000000 ; last month , it also said it was considering selling Land Rover and Jaguar.

**Relevant Sentence:** The Ford Motor Co, which sold its Aston Martin sports car brand in March , would like to be rid of its Land Rover and Jaguar operations by the end of the summer , and Volvo by the time winter arrives.

In Example (5), both the source passage and relevant passage refer to the same events that occurred in the past even though the documents themselves were pub-

lished at different times. In addition the bias rule learned in this case is incorrect as it lacks the proper noun "PAG." While this is a single word that might be missed, it is an important content word and results in an incorrect translation. The baseline translation did not translate the Chinese phrase as single unit and kept "PAG" in the final translation, although it incorrectly stated the sale was to occur in the future and did not specify what was being sold.

In two of these three examples, date overlap correctly served as a proxy for story overlap, although it failed in the third case as both stories were reporting on prior events. Due to the size of the data involved and the non-trivial nature of the identifying whether two sentences report the same event, discrepancies of this type have not been empirically calculated.

### 4.6.2 Examples with Date Overlap

The following examples from Chinese-to-English translation illustrate cases where the publication date of the source document and the relevant passage do overlap.

(6)

**Source:** 商业部 今天 提出 的 报告 将 对 美元 造成 压力 ，并 再度 呼吁 华府 在 美国 能源 独立 与 化解 美中 贸易 不平衡 上 应 有所 作为.

**Reference Translation:** The report by the Department of Commerce today will put pressure on the US dollar and renew calls for Washington to do something about the US energy independence and the trade imbalance with China.

**Relevant Sentence:** Wednesday's report is likely to put more pressure on the greenback and renew calls for Washington to do something about energy independence and the trade imbalance with China.

The source document and relevant passage in Example (6) were published only days apart, and largely overlap. The difference in date, from "today" to "Wednesday", and the naming of the report's author in the first half of the sentence prevent the two sentences from being parallel (the slang usage of "greenbacks" for "US dollar" is a potentially correct translation). The second half of the sentence is effectively parallel, and the resulting translation is correct. In this case the baseline translation is ", and once again called on Washington to resolve the U.S. energy independence and the Sino-US trade imbalance should do something.", which is only partially correct. The use of the long and highly fluent bias rule dramatically improves the translation quality in this case. Cases such as this are common when date overlap occurs.

(7)

**Source:** 负责 落实 推动 " 美国 访客 及 移民 身分 显示 技术 " ( us - visit ) 计画 的 国土 安全部 指出 , 大部分 属 欧洲 的 28 个 国家 的 公民 可 豁免 前述 程序 .

**Reference Translation:** The Department of Homeland Security , which implements and runs the ‘‘ United States Visitor and Immigrant Status Indicator Technology ’’ (US-VISIT ) , pointed out that citizens from 28 countries , mostly in Europe , are exempted from the aforesaid process.

**Relevant Sentence:** He announced that US-VISIT , for United States Visitor and Immigrant Status Indicator Tech nology , will be implemented at 115 American airports and 14 seaports, including LAX.

The two stories in Example (7) both discuss the "US-VISIT" technology in different contexts. The bias rule extracted correctly finds a translation of the Chinese to the English phrase. In the baseline system, this was translated using several individual translation rules resulting in a translation of "U.S. Visitor and Immigration Status display technology," which is an accurate translation, although it does not properly capture the acronym, replacing "indicator" with "display." This is an example of a very similar event happening at same time, using common phrases.

(8)

**Source:** 荷兰 说 , 欧盟 如何 为 这 项 禁运 画下 休止符 , 不 须 妄 加 揣测 .

**Reference Translation:** The Netherlands stated that how EU will end this embargo should not be wildly speculated.

**Relevant Sentence:** China has made clear it thinks human rights should not be linked to the arms embargo and has said it will make no concessions.

Example (8) presents a case where the relevant passage is not the same story as the source sentence, even though the relevant passage is from the same month as the source document. The resulting bias rule is also incorrect and results in an incorrect translation. This example shows how using date overlap as an approximation to story overlap can be incorrect, although in the majority of cases it may serve as an adequate proxy.

Selective translation-model adaptation addresses the limitations of basic translation-model adaptation by leveraging the MT system and the surrounding context to detect both phrasal and word translations that are novel and useful for the MT system. The analysis of date and topic overlap indicates that this technique is most beneficial when the relevant passages are from the same time as the source document, although the MT system can still benefit even when this similarity is reduced.

| Parameter Name | Value(s) | Description |
|---|---|---|
| Number of Relevant Passages | 10 | Specifies the number of passages used to generate selective translation-model adaptation rules. |
| Passage Length | 300 | Documents are pre-split into overlapping passages that are approximately 300 words long. |
| Minimum Source Phrase Length | 3 | This value was chosen to reduce the number of biased rules generated. |
| Minimum Target Phrase Length | 3 | This value was chosen to reduce the number of biased rules generated. |
| TERp Filter Level | 0.75 | Selective bias rules are only selected if the TERp score between the biased MT output and the relevant snippet are less than 0.75, where a score of 0 would indicate that there is a perfect match and a score of 1.0 or higher indicates that the pair are completely mis-matched. |
| LM Adaptation Interpolation Weights | 0, 0.1, 0.3, 0.5 | When generating the biased MT output, various interpolation weights are used for LM adaptation. |
| Basic Translation-Model Adaptation Levels | None, 1.5, 6.0 | When generating the biased MT output, various weights are used to weight the biased translation rules from basic-translation model adaptation. These weights are used in a log-linear model, so that a lower weight results in more biased translation rules being used. 'None' indicates that biased MT output was also generated without use of basic-translation model adaptation. |

Table 4.1: Free Parameters Used in Selective Translation-Model Adaptation

| Condition | Set | BLEU | TER |
|---|---|---|---|
| Baseline | Tune | 39.78 | 53.64 |
| | Test | 40.70 | 52.61 |
| 65 Features | Tune | 40.75 (+0.97) | 53.58 (-0.06) |
| | Test | 41.49 (+0.79) | 52.63 (+0.02) |

Table 4.2: Gains in Chinese Text Newswire Translation from Selective Translation-Model Adaptation

| Condition | Set | BLEU | TER |
|---|---|---|---|
| Baseline | Tune | 39.78 | 53.64 |
| | Test | 40.70 | 52.61 |
| 65 Features | Tune | 40.75 (+0.97) | 53.58 (-0.06) |
| | Test | 41.49 (+0.79) | 52.63 (+0.02) |
| 168 Features | Tune | 41.03 (+1.25) | 52.95 (-0.69) |
| | Test | 41.80 (+1.10) | 52.32 (-0.29) |

Table 4.3: Gains in Chinese-to-English translation from additional features in Translation-Model Adaptation

| Relevant Passage Date | Percent of Passages | Percent of Rules |
|---|---|---|
| Older than Source | 53.68% | 36.39% |
| Overlapping Date with Source[2] | 26.52% | 50.39% |
| Newer than Source | 18.80% | 17.54% |

Table 4.4: Percent of Passages and Rules that Overlap Source Documents Dates

| Condition | TER | BLEU | METEOR |
|---|---|---|---|
| | Test | | |
| Baseline | 54.69 | 35.96 | 60.36 |
| No Overlap | 54.65 (-0.04) | 36.26 (+0.30) | 60.34 (-0.02) |
| Date Overlap | 54.35 (-0.34) | 36.65 (+0.69) | 60.66 (+0.30) |
| All Dates | 54.34 (-0.35) | 36.84 (+0.88) | 60.75 (+0.39) |
| | Tune | | |
| Baseline | 53.92 | 37.06 | 60.63 |
| No Overlap | 53.76 (-0.16) | 36.86 (-0.20) | 60.62 (-0.01) |
| Date Overlap | 53.65 (-0.27) | 37.34 (+0.28) | 60.82 (+0.19) |
| All Dates | 53.65 (-0.27) | 37.33 (+0.27) | 60.75 (+0.12) |

Table 4.5: Effect of Date Overlap on Selective translation-model Adaptation for Chinese MT

# Chapter 5

## TER-Plus

In this chapter we introduce a novel automatic evaluation metric for machine translation, TER-Plus or TERP[1] (Snover et al., 2009, 2010) and demonstrate that it achieves higher correlation with human judgements of quality than other state-of-the-art evaluation metrics. The quality of TERP judgments is based upon the quality of the alignments it generates between English sentences. These alignments are essential to the selective translation model adaptation described in Chapter 4.

Section 5.1 provides the background behind the TERP measure. A detailed description of the design and implementation of TERP is presented in Section 5.2. To empirically test TERP, we measure its performance as a proxy for human judgements, as compared to other common metrics. The statistical nature of these comparisons is discussed in Section 5.3 with results presented in Section 5.4. As part of NIST's 2008 Metric MATR TERP was compared to a large number of potential evaluation metrics. An analysis of these results, showing TERP as one of the top performing metrics, is presented in Section 5.4.3. This is followed by a discussion of the benefit of each of the individual components of TERP in Section 5.5. Finally in Section 5.6 we discuss TERP as an alignment tool.

[1] TERP is named after the nickname—"terp"—of the University of Maryland, College Park, mascot: the diamondback terrapin.

## 5.1 Background

TERP is an automatic evaluation metric for machine translation (MT) that scores a translation (the *hypothesis*) of a foreign language text (the *source*) against a translation of the source text that was created by a human translator, which we refer to as a *reference* translation. The set of possible correct translations is very large, possibly infinite, and any one reference translation represents a single point in that space. Frequently, multiple reference translations—typically 4—are provided to give broader sampling of the space of correct translations. Automatic MT evaluation metrics compare the hypothesis against this set of reference translations and assign a score to the similarity, such that a better score is given when the hypothesis is more similar to the references.

TERP follows this methodology and builds upon an already existing evaluation metric, Translation Error Rate (TER) (Snover et al., 2006). In addition to assigning a score to a hypothesis, TER provides an alignment between the hypothesis and the reference, enabling it to be useful beyond general translation evaluation. While TER has been shown to correlate well with translation quality, it has several flaws: it only considers exact matches when measuring the similarity of the hypothesis and the reference, and it can only compute this measure of similarity against a single reference. The handicap of using a single reference can be addressed by constructing a lattice of reference translations—this technique has been used to combine the output of multiple translation systems (Rosti et al., 2007). TERP does not utilize this methodology and instead directly addresses the exact matching flaw

of TER.

In addition to aligning words in the hypothesis and reference if they are exact matches, TERP uses stemming and synonymy to allow matches between words. It also uses probabilistic phrasal substitutions to align phrases in the hypothesis and reference. These phrase substitutions are generated by considering possible paraphrases of the reference words. Matching using stems and synonyms (Banerjee and Lavie, 2005) as well as using paraphrases (Kauchak and Barzilay, 2006; Zhou et al., 2006) have been shown to be beneficial for automatic MT evaluation. Paraphrases have been shown to be additionally useful in expanding the number of references used for evaluation (Madnani et al., 2008) although they are not used in this fashion within TERP. The use of synonymy, stemming, and paraphrases allows TERP to better cope with the limited number of reference translations provided. TERP was one of the top metrics submitted to the NIST Metrics MATR 2008 challenge (Przybocki et al., 2008), having the highest average rank over all the test conditions (Snover et al., 2009).

## 5.2 The Design of Translation Edit Rate Plus (TERP)

TER-Plus extends the TER metric beyond the limitation of exact matches through the addition of three new types of edit operations, detailed in Section 5.2.1: stem matches, synonym matches, and phrase substitutions using automatically generated paraphrases. These changes allow a relaxing of the shifting constraints used in TER, which is discussed in Section 5.2.2. In addition, instead of all edit oper-

ations having a uniform edit cost of 1—as is the case in TER—the edit costs for TERP can be learned automatically in order to maximize correlation with human judgments. The details of this optimization are presented in Section 5.2.3.

### 5.2.1 Stem, Synonym, and Paraphrase Substitutions

In addition to the edit operations of TER—Matches, Insertions, Deletions, Substitutions and Shifts—TERP also uses three new edit operations: Stem Matches, Synonym Matches and Phrase Substitutions. Rather than treating all substitution operations as edits of cost 1, the cost of a substitution in TERP varies so that a lower cost is used if two words are synonyms (a Synonym Match), share the same stem (a Stem Match), or if two phrases are paraphrases of each other (a Phrase Substitution). The cost of these new edit types is set, along with the other edit costs, according to the type of human judgment for which TERP is optimized, as described in section 5.2.3.

TERP identifies stems and synonyms in the same manner as the METEOR metric (Banerjee and Lavie, 2005), where words are determined to share the same stem using the Porter stemming algorithm (Porter, 1980), and words are determined to be synonyms if they share the same synonym set according to WordNet (Fellbaum, 1998).

Phrase substitutions are identified by looking up—in a pre-computed *phrase table*—probabilistic paraphrases of phrases in the reference to phrases in the hypothesis. The paraphrases used in TERP are automatically extracted using the pivot-

based method (Bannard and Callison-Burch, 2005) with several additional filtering mechanisms to increase precision. The pivot-based method identifies paraphrases as English phrases that translate to the same foreign phrase in a bi-lingual phrase table. The corpus used for paraphrase extraction was an Arabic-English newswire bi-text containing a million sentences, resulting in a phrase table containing approximately 15 million paraphrase pairs. While an Arabic-English corpus was used to generate the paraphrases, the resulting phrase pairs are purely English paraphrases, and can be used when evaluating any translation into English regardless of the source language. It was previously shown that the choice of data for paraphrasing is not of vital importance to TERP's performance (Snover et al., 2009). A few examples of the extracted paraphrase pairs that were actually used by TERP in experiments described later are shown below:

brief $\Longleftrightarrow$ short
controversy over $\Longleftrightarrow$ polemic about
by using power $\Longleftrightarrow$ by force
response $\Longleftrightarrow$ reaction
agence presse $\Longleftrightarrow$ news agency
army roadblock $\Longleftrightarrow$ military barrier
think tank in $\Longleftrightarrow$ research center in
staff walked out $\Longleftrightarrow$ team withdrew
staged manner $\Longleftrightarrow$ gradually

Some paraphrases, such as *brief* and *short* are redundant with other edit types used by TERP such as synonym and stem matching.

A probability for each paraphrase pair is estimated as described in Bannard and Callison-Burch (2005). However, studies (Snover et al., 2009) of these paraphrase probabilities have shown that they are not always reliable indicators of the semantic relatedness of phrase pairs and further refinements of these probability estimates might prove valuable to TERP and other machine translation evaluation

metrics.

With the exception of the phrase substitutions, all of the edit operations used by TERP have fixed cost edits, i.e., the edit cost is the same regardless of the words in question. The cost of a phrase substitution is a function of the probability of the paraphrase and the number of edits needed to align the two phrases without the use of phrase substitutions. In effect, the probability of the paraphrase is used to determine how much to discount the alignment of the two phrases. For a phrasal substitution between a reference phrase $r$ and a hypothesis phrase $h$ where Pr is the the probability of paraphrasing $r$ as $h$, and edit$(r, h)$ is number of edits needed to align $r$ and $h$ without any phrasal substitutions, the edit cost is specified by three parameters, $w_1$, $w_2$, and $w_3$ as follows:

$$\text{cost}(r, h) = w_1 + \text{edit}(r, h)(w_2 \log(\text{Pr}) + w_3)$$

Only paraphrases specified in the input phrase table are considered for phrase substitutions. In addition, the total cost for a phrasal substitution is limited to values greater than or equal to 0, to ensure that the edit cost for substitution operations is always non-negative. The parameter $w_1$ allows a constant cost to be specified for all phrase substitutions, while parameters $w_2$ and $w_3$ adjust the discount applied to the edit cost of the two phrases.

### 5.2.2 Additional Differences From TER

In addition to the new edit operations, TERP differs from TER in several other ways. First, TERP is insensitive to casing information since we observe that penalizing for errors in capitalization lowers the correlation with human judgments of translation quality. Second, TERP is capped at 1.0. While the formula for TER allows it to exceed 1.0 if the number of edits exceed the number of words, such a score would be unfair since the hypothesis cannot be more than 100% wrong.

The shifting criteria in TERP have also been relaxed relative to TER, so that shifts are allowed if the words being shifted are: (i) exactly the same, (ii) synonyms, stems or paraphrases of the corresponding reference words, or (iii) any such combination. In addition, a set of stop words is used to constrain the shift operations such that common words ("the", "a" etc.) and punctuation can be shifted if and only if a non-stop word is also shifted. This reduces the number of shifts considered in the search and prevents any shifts that may not correspond with an increase in translation quality.

More relaxed shift constraints have been explored that allowed shifts even if some words did not match at all. We have empirically found this greatly increased the number of shifts considered, but also significantly decreased correlation with human judgment. The shift constraints imposed by TER and TERP serve not only to speed up the algorithm but also correspond to those block movement of words that correspond with increased translation quality.

### 5.2.3 TERP Edit Cost Optimization

While TER uses uniform edit costs—1 for all edits except matches—, we seek to improve TERP's correlation with human judgments by weighting different edit types more heavily than others, as some types of errors are more harmful to translation quality than others.

TERP uses a total of eight edit costs. However, the cost of an exact match is held fixed at 0 which leaves a total of seven edit costs that can be optimized. Since the paraphrase edit cost is represented by 3 parameters, this yields a total of 9 parameters that are varied during optimization. All parameters, except for the 3 phrasal substitution parameters, are also restricted to be positive. A hill-climbing search optimizes the parameters to maximize the correlation of human judgments with the TERP score. In this work, these correlations are measured at the sentence, or *segment*, level. However, optimization could also be performed to maximize document level correlation or any other measure of correlation with human judgments.

While TERP can be run using a fixed set of parameters, it can be beneficial to tune them depending on the properties of translation desired. Optimization of MT evaluation metrics towards specific human judgment types was previously investigated in a similar manner by Lita et al. (2005). Depending on whether the end goal is to maximize correlation with HTER, adequacy, or fluency, different sets of parameters may better reflect translation performance (Snover et al., 2009).

## 5.3 Statistical Analysis of MT Evaluation Metrics

Ideally, if the end-goal of a translation system is translation itself, rather than as input to another application, bilingual humans who are fluent in both the source and target language could compare numerous translations of a source document and quantify the relative quality of the translations. Employing sufficient speakers who are fluent, rather than just proficient, in both languages is costly, while the evaluation procedure itself is very time-consuming. Robustness and reliability of these human judgments is an additional and nontrivial issue.

The goal of an automatic MT evaluation metric is to act as a substitute or proxy for these costly and slow human evaluation of system outputs. Given a group of automatic metrics, selecting the metric which is the best proxy for human judgments is reduced to a statistical question of correlation of the quality judgments of the metric with the quality assessments of the human judges.

There are two primary roles for MT evaluation metrics: (1) to compare the outputs of various systems and determine which system produced a better translation, and the magnitude of that difference, and (2) to be used to optimize system parameters using Minimum Error Rate Training (MERT) or another optimization method. A third role for some MT evaluation metrics is the aligning two strings, particularly a reference sentence and a hypothesis sentence. This role is only applicable for those MT evaluation metrics that generate a string-to-string alignment such as WER, TER, and TERP, and is heavily exploited in this thesis. Such alignments can be used to not just determine the similarity of two strings but to detect

which portions of the two strings correspond to each other. This role for TERP will be further discussed in Section 5.6). Studies comparing evaluation metrics tend to focus on the first, although the second is arguably of even greater importance. Empirical studies of the first role are relatively straightforward, requiring translations of the same source data to be generated by the candidate systems which are then evaluated by human judges. Correlations of the human judgments with scores assigned by the automatic metrics can be used to statistically compare the metrics.

Examining the suitability of an evaluation metric for parameter optimization is far less straightforward, as the results vary depending on the optimization procedure used and the parameters that are being optimized. To conduct a study of which metric is better (irrelevant of any purely computational constraints) system parameters would have to be optimized for each metric. Each set of parameters would then be used to translate a held-out validation set of source documents. The resulting final output on the validation set would then be evaluated by human judges to determine which of the two sets of output is preferred. Studies of this kind have not, as of this time, been conducted on any wide scale, leaving this as a vital and yet largely unexplored region of research.

The remainder of this chapter focuses on TERP's suitability for the first role of evaluation metrics—that of acting as a proxy for human judges on the final output of varying systems, as this more closely reflects the use of TERP in Chapter 4. The question of parameter optimization is left for future research.

The two most common measurements of the correlation of human judgments with scores from automatic metric are Pearson and Spearman correlation coeffi-

cients. Because Spearman correlations can be viewed as a special case of Pearson correlations, I shall first discuss the calculation and significance of Pearson correlation coefficients. Both of these measures are used to evaluate TERP and other metrics in Section 5.4.

### 5.3.1 Pearson Correlation Coefficients

The Pearson correlation coefficient, $r$, when calculated over a sample of $N$ paired data points, $(X_i, Y_i)$, is defined in equation 5.1, where $\mu$ is the sample mean and $\sigma$ is the sample standard deviation. The correlation coefficient $r$ is not normally distributed, and ranges from -1 to 1, where a value of 0 indicates that $X$ and $Y$ are not correlated, and a value of 1 indicates that two variables are perfectly correlated. A value of -1 indicates that the two variables are perfectly inversely correlated. Metrics that are measures of accuracy, such as BLEU and METEOR, will be inversely correlated with metrics that measure error, such as WER, TER, and TERP.

$$r = \frac{1}{N-1} \sum_{i=1}^{N} \frac{X_i - \mu_X}{\sigma_X} \frac{Y_i - \mu_Y}{\sigma_Y} \tag{5.1}$$

Pearson confidence intervals can be used to determine if two correlations are significantly different, or if the differences in $r$ are due to sampling differences. Confidence intervals are calculated using the Fisher's $r$-to-$z$ transformation, consulting a $z$-table to find the upper and lower bounds of a confidence interval, and then converting the values back to $r$ scores. The transformation for $r$-to-$z$, as well as the reverse transformation, is computed used Equation 5.2, while the standard error of

$z$, $\sigma_z$, is computed computed using Equation 5.3. This is solely a function of the correlation coefficient, $r$, and the number of data points, $N$, where $N \geq 4$.

$$z_r = \frac{1}{2}(\ln(1+r) - \ln(1-r)) \tag{5.2}$$

$$\sigma_z = \frac{1}{\sqrt{N-3}} \tag{5.3}$$

As an example, if we have calculated a Pearson correlation coefficient of 0.85 over 8 samples, the 0.95 confidence interval after the described conversions is from 0.363 to 0.972, indicating that thee is a 95% probability that the true correlation is within that range. The confidence interval grows smaller as the value of $r$ and $N$ increase. If the value of $r$ was 0.99 with 8 samples, the 0.95 confidence interval would be much tighter with a range of 0.944 to 0.998. If instead of 8 samples, we had 80 samples with $r = 0.85$, the 0.95 confidence interval would be from 0.776 to 0.901.

To test whether two correlations, $r_1$ and $r_2$, both calculated over a sample of size $N$ are significantly different, we calculate the confidence intervals, of probability $p$, as described and check to see if the sample correlation of $r_1$ is within the confidence interval of $r_2$, or vice versa—this is a symmetric relationship, even though the confidence intervals themselves are not symmetric. If $r_1$ is within the confidence interval of $r_2$ then the two are **not** significantly different with probability $\geq p$, while the difference, also with probability $\geq p$, is statistically significant if $r_1$ is not within $r_2$'s confidence interval. More exact tests can be computed if certain

other assumptions are true, although this is frequently not the case when comparing the correlations of evaluation metrics with human judgments.

To continue the previous example, consider three metrics, $m_1$, $m_2$ and $m_3$, whose correlations with human judgments on 8 data points are 0.85, 0.95 and 0.99 respectively. The 0.95 confidence intervals for these correlations are:

| Metric | lower bound | $r$ | upper bound |
|---|---|---|---|
| $m_1$ | 0.363 | 0.85 | 0.972 |
| $m_2$ | 0.743 | 0.95 | 0.991 |
| $m_3$ | 0.944 | 0.99 | 0.998 |

This reveals that the correlation with human judgments of $m_1$ is significantly worse from the correlation for $m_3$. The correlation of $m_2$ however is not significantly different from either $m_1$ or $m_3$. While the confidence intervals of $m_1$ and $m_3$ overlap, the $r$ value for the two metrics do not lie in the confidence interval of the other metric, so the difference between is significant.

### 5.3.2 Spearman Correlation Coefficients

Spearman correlations are rank correlations and ignore the degree of difference between data points and focus instead on how well the two distributions, $X$ and $Y$ agree on order of the data points in the sample. To calculate the Spearman correlation coefficient, the data points are first converted to ranks, with the highest value being given a rank of $N$ and the lowest value being given a rank of 1. For example, if the sample of scores $< (X_1, Y_1), (X_2, Y_2), ...(X_N, Y_N) >$ were $< (1.0, 3.0), (2.0, 4.0), (0.5, 6.8), (3.0, 5.6) >$, the ranks of the sample would be

$< (2,1),(3,2),(1,4),(4,3) >$. The Spearman correlation coefficient is then defined as the Pearson correlation coefficient of the ranked values.

While more exact methods, such as bootstrapping, can be used to calculate the confidence intervals of the Spearman correlation coefficients, reliable confidence intervals be obtained using the Fisher $r$-to-$z$ transformation method used for Pearson correlations.

Spearman correlations are frequently used in comparing evaluation metrics, but serious handicaps to their use exist. If ties exist in the ranks, so that if multiple samples have the same value under $X$ or $Y$, then the Spearman correlation can be inappropriate. While this may not occur when examining small samples, such as when only the total system level performance is examined, it is almost impossible to avoid as $N$ increases such as when correlation is computed at the segment or document level. In addition, the Spearman correlation ignores the degree of difference between values, which factors out an important component of an evaluation metric. For example, if a metric assigns the scores, $[0, 0.09, 0.10, 0.11, 0.9, 0.95]$ to a sample of size 6, the same ranks are generated as when the metric assigns the scores $[0, 0.2, 0.4, 0.6, 0.8, 1.0]$. In the first example, the second, third, and fourth scores are almost identical, while they are more widely spread out in the second example.

## 5.4 Evaluating TERP

This section empirically evaluates TERP by exploring the optimization of the edit costs and the correlation of TERP scores with human judgments of translation quality. TERP was also independently evaluated by NIST in the Metrics MATR 2008 Challenge, and results from that evaluation are presented below.

### 5.4.1 Optimization For Adequacy

In order to tune and test TERP, we used a portion of the Open MT-Eval 2006 evaluation set that had been annotated for adequacy (on a seven-point scale) and released by NIST as a development set for the Metrics MATR 2008 challenge (Przybocki et al., 2008). This set consists of the translation hypotheses from 8 Arabic-to-English MT systems for 25 documents, which in total consisted of 249 segments. For each segment, four reference translations were also provided. Optimization was done using 2-fold cross-validation. These optimized edit costs (and subsequent results) differ slightly from the formulation of TERP submitted to the Metrics MATR 2008 challenge, where tuning was done without cross-validation. Optimization requires small amounts of data but should be done rarely so that the metric can be held constant to aid in system development and comparison.

| **Match** | **Insert** | **Deletion** | **Substitution** | **Stem** |
|---|---|---|---|---|
| 0.0 | 0.20 | 0.97 | 1.04 | 0.10 |
| **Syn.** | **Shift** | **Phrase Substitution** | | |
| 0.10 | 0.27 | $w_1$: 0.0 | $w_2$: -0.12 | $w_3$: 0.19 |

Table 5.1: TERP Edit Costs Optimized for Adequacy

TERP parameters were then optimized to maximize segment level Pearson correlation with adequacy on the tuning set. The optimized edit costs, averaged between the two splits of the data, are shown in Table 5.1. Because segment level correlation places equal importance on all segments, this optimization over-tunes for short segments, as they have very minor effect at the document or system level. Optimization on length weighted segment level correlation would rectify this but would result in slightly worse segment level correlations.

### 5.4.2 Correlation Results

In our experiments, we compared TERP with METEOR (Banerjee and Lavie, 2005) (version 0.6 using the Exact, WordNet synonym, and Porter stemming match modules), TER (version 0.7.25), and the IBM version of BLEU (Papineni et al., 2002) with the default maximum $n$-gram size of 4 (BLEU). We also included a better correlating variant of BLEU with a maximum $n$-gram size of 2 (BLEU-2). TER and both versions of BLEU were run in case insensitive mode as this produces significantly higher correlations with human judgments, while METEOR is already case insensitive.

To evaluate the quality of an automatic metric, we examined the Pearson correlation of the automatic metric scores—at the segment, document and system level— with the human judgments of adequacy. Document and system level adequacy scores were calculated using the length weighted averages of the appropriate segment level scores.

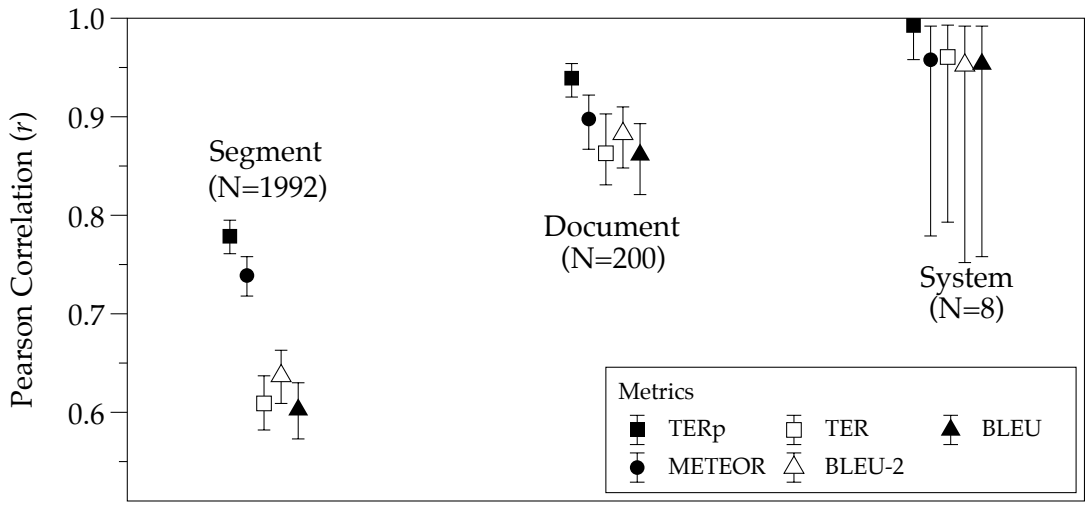



Figure 5.1: Metric correlations with adequacy on the Metrics MATR 2008 development set. Correlations are significantly different if the center point of one correlation does not lie within the confidence interval of the other correlation.

Pearson correlation results between the automatic metrics and human judgments of adequacy are shown in Figure 5.1. In order to determine the statistical significance of the differences in correlation between metrics, one can examine the confidence interval of the Pearson coefficient, $r$. If the correlation coefficient for a metric occurs within the 95% confidence interval of another metric, then the difference between the correlations of the metrics is not statistically significant.

TERP consistently outperformed all of the other metrics on the segment, document, and system level Pearson correlations, with all but one difference being statistically significant. While TERP had higher correlation than TER on the system level, the difference is not statistically significant—the differences with all other metrics are statistically significant. Of the other metrics, METEOR consistently had the highest Pearson correlation at the segment and document level.

### 5.4.3 NIST Metrics MATR 2008 Challenge

TERP was one of 39 automatic metrics evaluated in the 2008 NIST Metrics MATR Challenge. In order to evaluate the state of automatic MT evaluation, NIST tested metrics across a number of conditions across 8 test sets. These conditions included segment, document and system level correlations with human judgments of preference, fluency, adequacy and HTER. The test sets included translations from Arabic-to-English, Chinese-to-English, Farsi-to-English, Arabic-to-French, and English-to-French MT systems involved in NIST's MTEval 2008, the GALE (Olive, 2005) Phase 2 and Phrase 2.5 program, Transtac January and July 2007, and CESTA run 1 and run 2, covering multiple genres.

| **Match** | **Insert** | **Deletion** | **Substitution** | | **Stem** |
|---|---|---|---|---|---|
| 0.0 | 0.26 | 1.43 | 1.56 | | 0.0 |
| **Syn.** | **Shift** | **Phrase Substitution** | | | |
| 0.0 | 0.56 | $w_1$: -0.23 | $w_2$: -0.15 | $w_3$: 0.18 | $w_4$: -0.08 |

Table 5.2: Optimized TERP Edit Costs

The version of TERP described previously differs from the version submitted to the Metrics MATR challenge in two regards. First, the version of TERP submitted to this workshop was optimized as described in Section 5.4.1, except that 2-fold cross validation was not used but rather the development data was split into two portions, one for tuning and one for testing. Secondly, the formula for the cost of a phrase substitution contained another term, $w_4$ so that the cost for substituting a reference phrase $r$ with the hypothesis phrase $h$ is:

$$\text{cost}(r,h) = w_1 + \text{edit}(r,h)(w_2 \log(\Pr) + w_3 + w_4 \Pr)$$

The edit costs learned from the optimization performed are shown in Table 5.2. The development set upon which TERP was optimized was not part of the test sets evaluated in the challenge.

Due to the wealth of testing conditions, a simple overall view of the official MATR08 results released by NIST is difficult. To facilitate this analysis, we examined the average rank of each metric across all conditions, where the rank was determined by their Pearson and Spearman correlation with human judgments. To incorporate statistical significance, we calculated the 95% confidence interval for each correlation coefficient and found the highest and lowest rank from which the correlation coefficient was statistically indistinguishable, resulting in lower and upper bounds of the rank for each metric in each condition. The average lower bound, actual, and upper bound ranks (where a rank of 1 indicates the highest correlation) of the top metrics, as well as BLEU and TER, are shown in Figure 5.2, sorted by the average rank of the Pearson correlation. The same analysis for Spearman correlations is shown in Figure 5.3. Full descriptions of the other metrics,[2] the evaluation results, and the test set composition are available from NIST (Przybocki et al., 2008). The results in this analysis differ slightly

This analysis shows that TERP was consistently one of the top metrics across test conditions and had the highest average rank both in terms of Pearson and Spearman correlations. While this analysis is not comprehensive, it does give a general idea of the performance of all metrics by synthesizing the results into a

[2]System description of metrics are also distributed by AMTA: `http://www.amtaweb.org/AMTA2008.html`

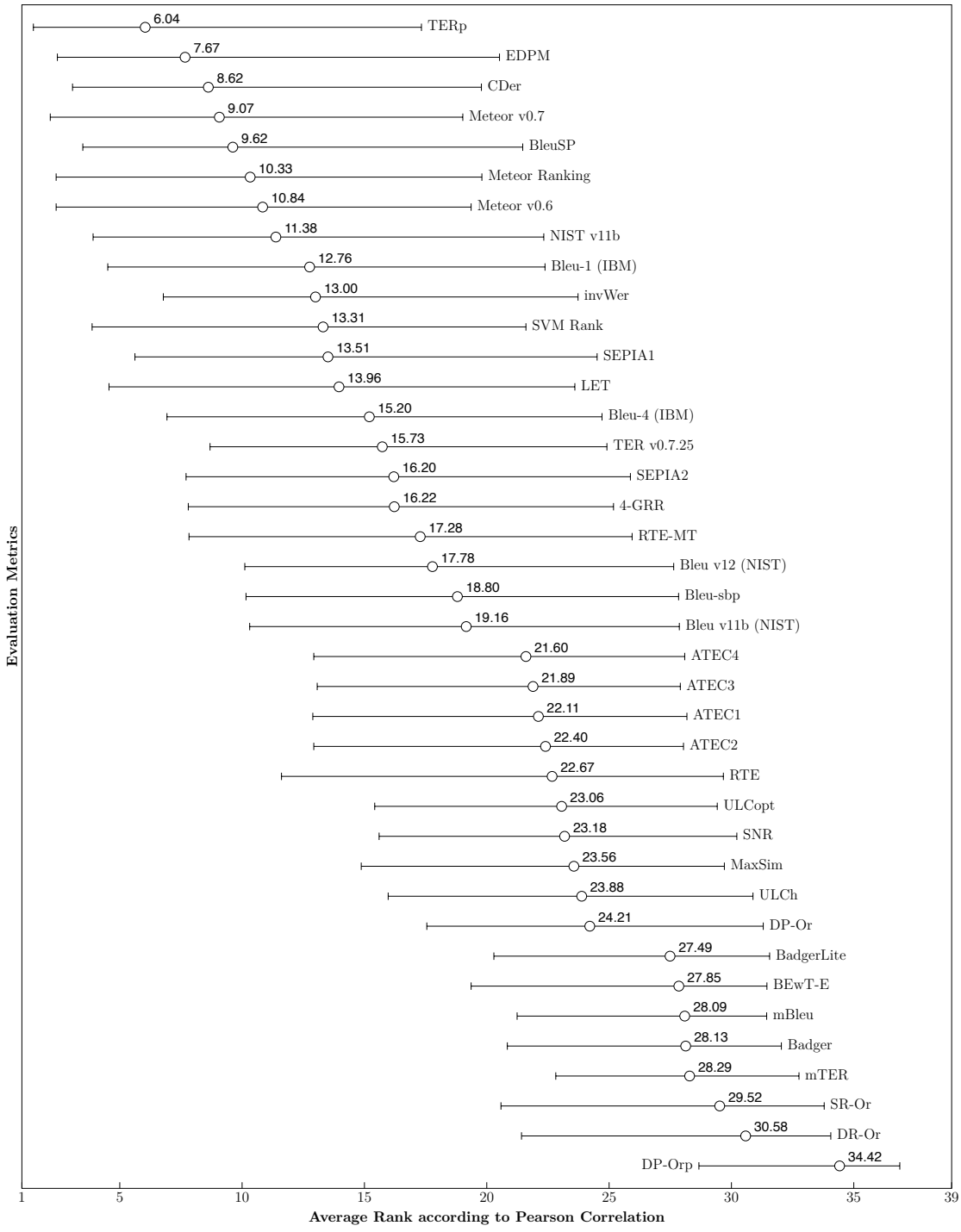



Figure 5.2: Average Metric Rank according to Pearson correlation in NIST Metrics MATR 2008 Official Results

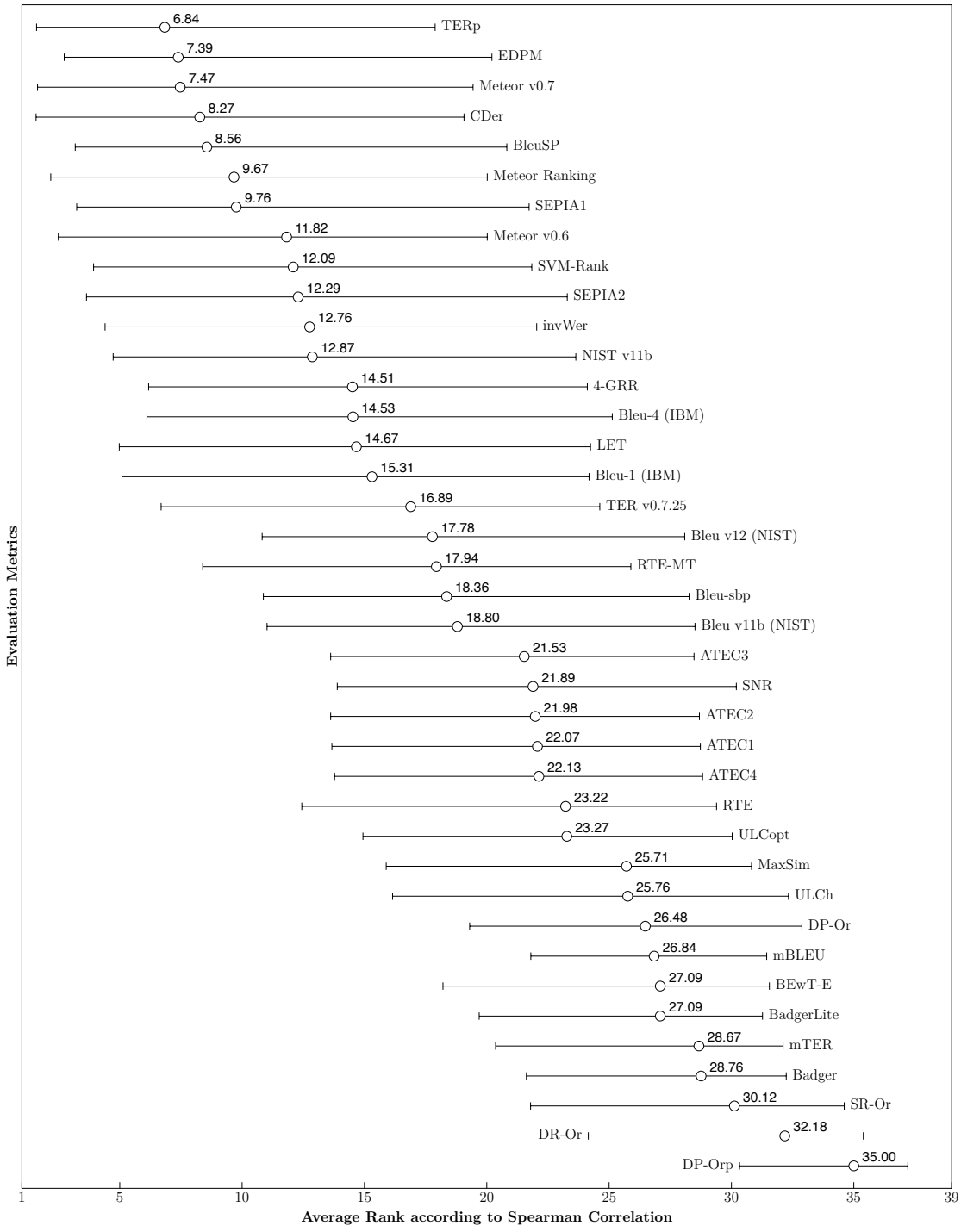



Figure 5.3: Average Metric Rank according to Spearman correlation in NIST Metrics MATR 2008 Official Results (Average Rank of 1 is highest rank)

single table. There are striking differences between the Spearman and Pearson correlations for other metrics, in particular the CDer metric (Leusch et al., 2006) had the second highest rank in Spearman correlations (after Terp), but was the sixth ranked metric according to the Pearson correlation. In several cases, Terp was not the best metric (if a metric was the best in all conditions, its average rank would be 1), although it performed well on average. In particular, Terp did significantly better than the Ter metric, indicating the benefit of the enhancements made to Ter.

## 5.5 Benefit of Individual TERp Features

In this section, we examine the benefit of each of the new features of Terp by individually adding each feature to Ter and measuring the correlation with the human judgments. Each condition was optimized as described in section 5.4.1. Figure 5.4 shows the Pearson correlations for each experimental condition along with the 95% confidence intervals.

The largest gain over Ter is through the addition of optimizable edit costs. This takes Ter from being a metric with balanced insertion and deletion costs to a recall-oriented metric which strongly penalizes deletion errors, while being forgiving of insertion errors. This single addition gives statistically significant improvements over Ter at the segment and document levels. This validates similar observations of the importance of recall noted by Lavie et al. (2004).

The other three features of Terp—stemming, synonymy, and paraphrases—

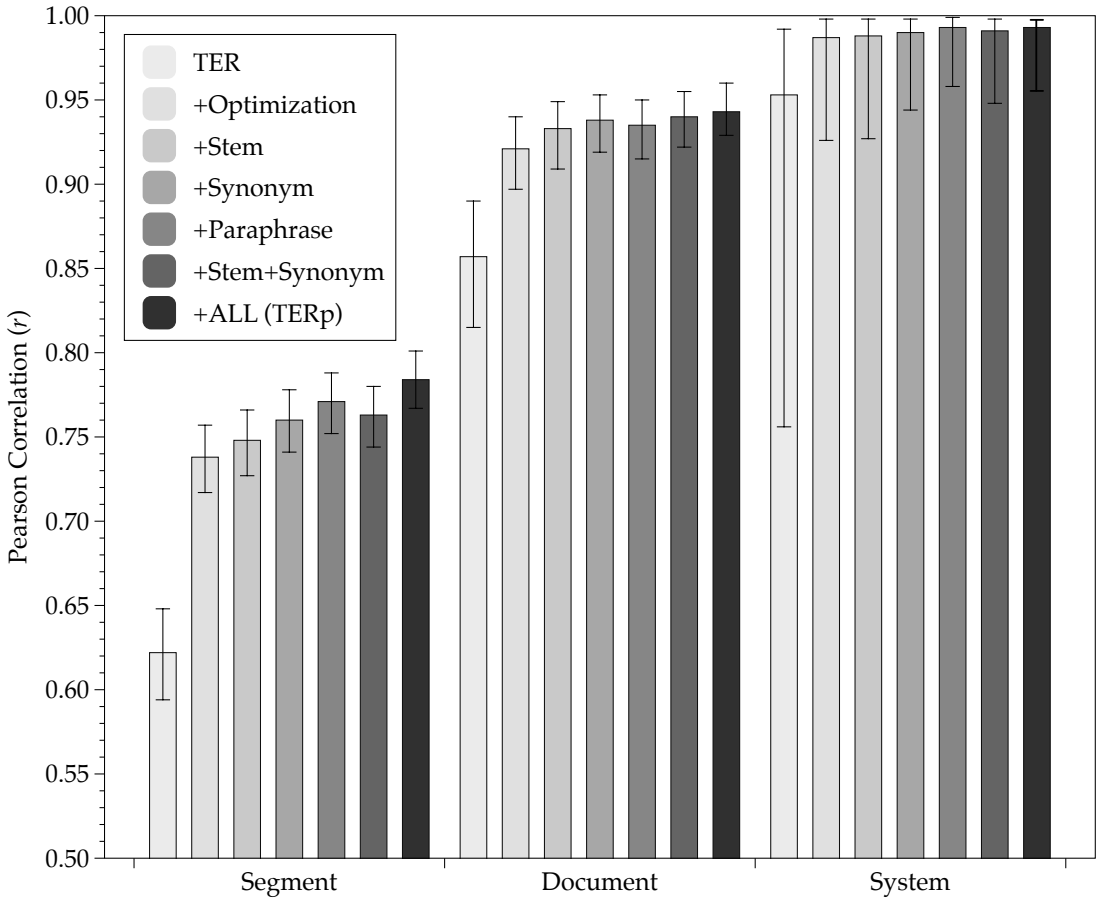



Figure 5.4: Pearson Correlation of TERp with Selective Features.

are added on top of the optimized TER condition since optimization is required to determine the edit costs for the new features. The addition of each of these features increases correlations over the optimized edit costs at all levels, with statistically significant gains at the segment level for the addition of synonymy or paraphrasing. The addition of paraphrasing gives the largest overall gains in correlation after optimization and is more beneficial than stemming and synonymy combined. A large percentage of synonym and stem matches are already captured in the paraphrase set and, therefore, the combination of all three features yields only a small gain over paraphrasing alone.

The TERP framework and software also provides for separate word classes with individual edit costs, so that the edit costs of various sets of words can be increased or decreased. For example, the cost of deleting content words could be set higher than that of deleting function words. It is difficult to set such costs manually as it is not clear how these phenomena are treated by human annotators of translation quality, although these costs could be determined by automatic optimization.

## 5.6 TERp Alignment

In addition to providing a score indicating the quality of a translation, TERP generates an alignment between the hypothesis and reference, indicating which words are correct, incorrect, misplaced, or similar to the reference translation. While the quality of this alignment is limited by the similarity of the reference to the hypothesis it can be beneficial in diagnosing error types in MT systems, or as a general sentence to sentence alignment tool, as is done in Chapter 4.

Several examples of TERP alignments are shown in Figure 5.5. Within each example, the first line is a snippet of the reference translation, the second line is the original hypothesis, and the third line is the hypothesis after all shifts have been performed. Words in **bold** are shifted, while square brackets are used to indicate other edit types: *P* for phrase substitutions, *T* for stem matches, *Y* for synonym matches, *D* for deletions, and *I* for insertions.

These alignments allow TERP to provide quality judgments on translations and to serve as a diagnostic tool for evaluating particular types of translation errors. In

R : ... [a number of]$_{\mathbf{D}}$ leaders expressed their opposition to [participating in]$_{\mathbf{P}}$ the government ...
H : ... the leaders expressed their opposition to the government **take part in** ...
H': ... [the]$_{\mathbf{I}}$ leaders expressed their opposition to [**take part in**]$_{\mathbf{P}}$ the government ...

5.27 (6)

R : ... [he]$_{\mathbf{D}}$ [went on to say]$_{\mathbf{P}}$ , "we also discussed how [to galvanize]$_{\mathbf{D}}$ the ... "
H : ... continued , "we also discussed how the activation of ... "
H': ... [continued]$_{\mathbf{P}}$ , "we also discussed how the [activation of]$_{\mathbf{I}}$ ... "

6.48 (8)

R : ... [but]$_{\mathbf{S1}}$ we [have]$_{\mathbf{Y1}}$ [Palestinian]$_{\mathbf{T}}$ [ ]$_{\mathbf{S2}}$ Arab [or]$_{\mathbf{D}}$ Islamic [alternatives]$_{\mathbf{Y2}}$ .
H : ... and we now possess an **Islamic** or the Palestinians and Arab options.
H': ...[and]$_{\mathbf{S1}}$ we now [possess]$_{\mathbf{Y1}}$ [an or the]$_{\mathbf{I}}$ [Palestinians]$_{\mathbf{T}}$ [and]$_{\mathbf{S2}}$ Arab **Islamic** [options]$_{\mathbf{Y2}}$ .

6.14 (10)

Figure 5.5: Examples of TERP Alignment Output. In each example, **R**, **H** and **H'** denote the reference, the original hypothesis and the hypothesis after shifting respectively. Shifted words are **bold**ed and other edits are in [brackets]. Number of edits shown: TERP (TER).

addition, TERP may also be used as a general-purpose string alignment tool, e.g., aligning multiple system outputs to each other for MT system combination (Rosti et al., 2007), a task for which TERP may be even better suited.

TERP extends the TER metric using stems, synonyms, and paraphrases and optimized weights to improve the alignment and relative scores of edits. Experimental results show that TERP has significant gain in correlation with human judgments over baseline MT evaluation metrics. Evaluation can be targeted towards specific types of human judgments, yielding different edit costs for TERP. This allows TERP to be optimized for use when a specific focus on MT output is required.

# Chapter 6

# Conclusion

This thesis explores the exploitation of monolingual comparable corpora to improve statistical machine translation. A natural duplication of information across languages due to the independent reporting of events and topics is utilized to find sub-sentential regions of monolingual text that are parallel to phrases in the source document to be translated. These regions can be detected by leveraging the existing power of the translation system, cross-lingual information retrieval and target-to-target language string alignment, using the TERP evaluation metric. New, document-specific, translation rules can then be hypothesized between the source words and the monolingual text.

The benefit of using these translation rules can be seen by the improvement according to standard automatic evaluation metrics and by qualitatively examining the use of these rules. In some cases, the rules are beneficial only in that they allow long fluent phrases to be used in place of many shorter generic translation rules. In other cases, these selective translation rules generate new words that were not previously in the lexicon of the MT system, or they generate new word-to-word translations that were absent from the MT system's training.

Rather than focus on resource impoverished languages, these techniques apply to resource-heavy languages, improving on state-of-the-art MT systems for well-

studied languages. While translation-model adaptation may be limited in benefit in these less commonly taught languages, it may still be applicable. Future research should investigate how these techniques that where developed for resource-heavy languages function when applied to more resource-poor languages.

## 6.1 Research Contributions

The contributions of this dissertation include the following:

- A method for selectively learning new phrasal translation rules without parallel corpora that improves state-of-the-art statistical machine translation. This method learns translations from the source documents to be translated and relevant passages from comparable monolingual text, by exploiting parallelization at a sub-sentential level. This selective translation rule learning relies upon language-model adaptation, basic translation-model adaptation and the use of TERp as an alignment tool.
- A new and simple method for translation-model adaptation using relevant texts from comparable corpora. Portions of this research have been previously published in Snover et al. (2008).
- Verification that language-model adaptation using relevant passages from comparable corpora can be used to improve state-of-the-art statistical machine translation. Portions of this research have been previously published in Snover et al. (2008).
- An automatic metric for machine translation evaluation, TER-Plus (TERp),

which demonstrates a high level of correlation with human judgements of quality—ranking at the top of automatic evaluation metrics at the NIST 2008 MetricsMATR challenge. TERP also provides a method to perform alignment between segments of English text, a feature used elsewhere in this thesis. Both this metric, and its predecessor TER, have been distributed to the NLP community where they have proved useful for both MT evaluation and alignment tasks. Portions of this research have been previously published in Snover et al. (2009) and Snover et al. (2010).

In this thesis I sought out to answer the following questions:

1. Is it possible to improve a modern, state-of-the-art statistical translation system using language-model and translation-model biasing techniques that cause the translation output to be more similar to relevant passages from comparable monolingual text?
2. What level of sub-sentential parallelization is necessary to exploit such techniques?
3. What is the nature of the relevant passages that are needed in applying such techniques?

I found is it is possible to improve a modern, state-of-the-art statistical translation system using language-model and translation-model adaptation. Standard automatic metrics and qualitative analysis of translation results show that the use of bias translation rules improves the translation quality of a state-of-the-art translation system. Such gains are the product of a small number of new translation rules.

I found that translation-model adaptation is most beneficial when there is a significant amount of sub-sentential parallelization, with the greatest benefits occurring when large amounts of the source sentence is sub-sententially parallel to the relevant monolingual text. These techniques still show some benefit when large parallelization is not present, although only shorter translation rules can be used in this case.

The most useful relevant passages for this technique are found when stories are repeated, typically within a short time period of the source document. The exclusive use of such passages provides almost as much benefit to the MT system as using more distantly related passages. Restricting the use of relevant passages that overlap in date with the source document does not remove the benefit of the adaptation techniques, but only cuts the gain in half.

In this thesis, I validated the following hypotheses:

1. *Improvements to the MT system are possible from both language-model and translation-model biasing techniques.* I found that by using these techniques in modern state-of-the-art MT system, I improve the quality of the translation.
2. *While little sub-sentential parallelization is necessary to exploit language-model adaptation, translation-model adaptation relies upon some level of sub-sentential parallelization consisting of a minimum of a few words of sub-sentential parallelization.* Translation-model adaptation only improved translation quality if there was some sub-sentential parallelization between the source and the relevant passage, be it a two or three words or a half a sentence.

3. *Those relevant documents that come from the same time period as the source document and cover the same story are the most useful and provide the greatest benefit for translation, although events that occur at different time periods can still be exploited to a lesser degree.* By limiting bias translation rules to those from texts that overlapped in date with the source document, I found the highest level of gain—a small additional gain was found by allowing relevant texts from different time periods. On their own, however, I found that relevant texts from different time periods were still useful, providing half of the gain as using the date-overlapping texts.

## 6.2 Future Work

The research presented in this thesis has a number of obvious extensions. Because the TERP alignment method is essential to the selective translation-model adaptation technique, improvements made to the alignment capabilities of the evaluation metric would likely have a direct effect on translation-model adaptation. These improvements would likely have additional innate benefit to MT evaluation. Future work in this direction is discussed in Section 6.2.1.

### 6.2.1 TERP

As a core element in selective translation-model adaptation, improvements to TERP could naturally lead to more reliable estimates on the similarity of the biased MT output and the relevant passages, enabling additional discriminative features

and enabling the system to generate a greater number of new biased translation rules. These improvements to TERP would also serve to improve its usability as an evaluation metric.

The paraphrases used by TERP are one of its most useful features; however, these paraphrases, being automatically generated using the pivot-based-method, contain many false paraphrases that cause TERP to be overly generous in alignment and evaluation. Improvements to paraphrasing that increase the precision of the paraphrases could dramatically increase the accuracy of the TERP metric. By examining the intersection of paraphrases from multiple pivot languages, it may be possible to find a high confidence set of paraphrases that could increase the usefulness of paraphrases to TERP for both alignment and evaluation purposes.

One of the key problems with the TERP metric is that it treats all words as being equally important. There exists a sharp contrast between the importance of content words/phrases and purely grammatical words/phrases. By identifying content words and phrases, and weighting them more heavily, TERP could ensure that the deletion of a proper noun is a much more serious error than the dropping of a determiner. Since unknown words in MT are dominated by proper nouns this would be especially important for the alignment of biased MT output to relevant text. By identifying cases where the only differences between the two strings are non-content words, TERP could more reliably identify regions where source words are truly sub-sententially parallel to words in a relevant passage. Such improvements to TERP could be achieved by incorporating part-of-speech tagging or named entity recognition to TERP, and separating edit types for various categories of words.

In addition, TERP is a tunable metric and could be tuned directly for the task of aligning the biased MT output to the relevant passages. It could also be tuned to provide scores that are more reflective of the quality of that alignment. Such tuning could be done either using human judgments of rule quality or by attempting to match rule scores, learning from discriminative training in the MT decoder itself.

Unfortunately, at present, the TERP metric relies upon English-only resources, limiting its use to the English language. Additional research and possible resource creation would be necessary to adapt TERP to function equally well in other languages. Full adaptation of TERP to other languages requires lemma matching resources, a list of synonyms (the English WordNet that is currently used by TERP was laboriously created by human experts), and paraphrases, in addition to data upon which to tune the TERP parameters. Of the first three resources, the paraphrases are the most useful in English, and can be automatically generated if parallel data exists between the new language and any other language using pivot-based paraphrasing techniques. An implementation of TERP that uses paraphrases but lacks stem matching or synonym matching would be inferior to the English version of TERP, but might suffice for use with selective translation-model adaptation.

### 6.2.2 Selective Translation-Model Adaptation

The work presented in this thesis is focused on the problem of translating news stories as these are high-priority genres in machine translation and are the most

likely place for the translation-model adaptation techniques described in this thesis to work. There is nothing intrinsic to these techniques that limits them to news stories, however. Translation-model adaptation should be beneficial whenever the source document to be translated contains information that is likely to be repeated in other languages. This would likely extend to the genre of broadcast news, where speech recognition is first used to transcribe speech in the source language before translating into the target language. Outside of the news domain, it is unclear how far this technique would extend in practice. To fully explore this, one would need to acquire large comparable corpora in these other genres. Assuming that comparable corpora can be obtained, the technique would seem to be best suited to current event driven data, where parallel text does not already exist. Much like news stories, blogs and product reviews tend to repeat information in many sources and across languages. Selective translation-model adaptation might be especially applicable in these genres as it can be used to find translations of new names and technical terms that are not currently in the translation lexicon of the MT system.

Extending selective translation-model adaptation beyond English requires several components. First, the Terp alignment tool must be adapted to the new target language, as discussed above. Second, an existing MT system must already be available in for the new source and target language pair—selective translation-model cannot replace an MT system or a lack of parallel data, but can possibly augment an MT system that has scarce resources. The CLIR system would also need to be adapted to the new language pair, although doing so is trivial given a existing statistical MT system for the language pair. There is nothing inherent in

the selective translation-model adaptation process that is English specific or that requires modification for other language pairs. Only the underlying tools that selective translation-model adaptation is built upon require adjustment when moving to new language pairs.

A major speed limitation to the selective translation-model adaptation technique is that several preliminary translations are required of the source text before bias rules are generated. In many cases, no bias rules are generated for a given passage. Such 'false' hits could be detected earlier by examining gloss lexical translations to detect a low likelihood of sub-sentential parallelization. By filtering in this manner, it would become practical to consider much larger sets of parallel documents.

The major limitation in using the selective translation rules is that the rules generated are completely phrasal, while the translation system is of a hierarchical nature. Because of this, the bias translation rules can be difficult to integrate into the hypotheses of the MT system. Generalizing the bias rules to be more hierarchical could allow them to be more easily used, and could be used to generalize away from regions of poor alignment in Terp alignment. Regions that do not align could be abstracted to a non-terminal so that the hierarchical system could attempt to use them without requiring an exact match against the relevant comparable text passage. This would allow two matching portions of the relevant text to be used, with the intervening region being abstracted as a non-terminal, resulting in a larger set of useful rules.

The work presented in this thesis presents many possibilities for the usefulness

of selective translation-model adaptation in real-world translation problems. Expanding the size of the monolingual corpus by orders of magnitude, such as using data from the World-Wide-Wide, might allow a much larger amount of relevant text to be retrieved. These more relevant texts would allow translation-model adaptation to be more heavily used, extending its applicability and value. It is in such real-world deployment that selective translation-model adaptation holds the most promise.

# Bibliography


Sadaf Abdul-Rauf and Holger Schwenk. Exploiting comparable corpora with ter and terp. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 46–54, Singapore, August 2009. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaulation Measures for MT and/or Summarization*, pages 228–231, 2005.

Colin Bannard and Chris Callison-Burch. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 597–604, Ann Arbor, Michigan, June 2005.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-Evaluating the Role of Bleu in Machine Translation Research. In *Proceedings of EACL-2006*, 2006.

E. Charniak and M. Johnson. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, page 180. Association for Computational Linguistics, 2005.

David Chiang. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of ACL*, pages 263–270, 2005.

David Chiang, Kevin Knight, and Wei Wang. 11,001 new features for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 218–226. Association for Computational Linguistics, 2009.

Jacob Devlin. Lexical features for statistical machine translation. Master's thesis, University of Maryland, December 2009.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998. http://www.cogsci.princeton.edu/~wn [2000, September 7].

Pascale Fung and Lo Yuen Yee. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of COLING-ACL98*, pages 414–420, August 1998.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval. In *Proceedings of EAMT 2005*, Budapest, Hungary, May 2005.

Heng Ji. Mining name translations from comparable corpora by creating bilingual information networks. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 34–37, Singapore, August 2009. Association for Computational Linguistics.

David Kauchak and Regina Barzilay. Paraphrasing for Automatic Evaluation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 455–462, 2006.

Woosung Kim. *Language Model Adaptation for Automatic Speech Recognition and Statistical Machine Translation*. PhD thesis, The Johns Hopkins University, Baltimore, MD, 2005.

Woosung Kim and Sanjeev Khudanpur. Cross-Lingual Lexical Triggers in Statistical Language Modeling. In *2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pages 17–24, July 2003.

Alon Lavie and Abhaya Agarwal. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgements. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, 2007.

Alon Lavie, Kenji Sagae, and Shyamsudar Jayaraman. The Significance of Recall in Automatic Metrics for MT Evaluation. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas*, pages 134–143, 2004.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. CDER: Efficient MT Evaluation Using Block Movements. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics)*, pages 241–248, 2006.

V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10:707–710, 1966.

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. Dictionary-based cross-language retrieval. *Information Processing and Management*, 41:523–547, 2005.

Lucian Vlad Lita, Monica Rogati, and Alon Lavie. BLANC: Learning Evaluation Metrics for MT. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 740–747, Vancouver, BC, October 2005.

Daniel Lopresti and Andrew Tomkins. Block edit models for approximate string matching. *Theoretical Computer Science*, 181(1):159–179, July 1997.

Yajuan Lu, Jin Huang, and Qun Liu. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 343–350, 2007.

Nitin Madnani, Philip Resnik, Bonnie J. Dorr, and Richard Schwartz. Are Multiple Reference Translations Necessary? Investigating the Value of Paraphrased Reference Translations in Parameter Optimization. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, pages 143–152, October 2008.

T. Leek Miller and Richard Schwartz. BBN at TREC7: Using Hidden Markov Models for Information Retrieval. In *TREC 1998*, pages 80–89, Gaithersburg, MD, 1998.

Dragos Stefan Munteanu and Daniel Marcu. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31: 477–504, 2005.

Dragos Stefan Munteanu and Daniel Marcu. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia, July 2006. Association for Computational Linguistics.

Dragos Stefan Munteanu and Daniel Marcu. Isi arabic-english automatically extracted parallel text. Linguistic Data Consortium, Philadelphia, 2007.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, pages 39–45, 2000.

Douglas W. Oard and Bonnie J. Dorr. Evaluating Cross-Language Text Retrieval Effectiveness. In Gregory Grefenstette, editor, *Cross-Language Information Retrieval*, pages 151–161. Kluwer Academic Publishers, Boston, MA, 1998.

F. J. Och and H. Ney. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Conference of the Association for Computational Linguistics*, pages 440–447, Hongkong, China, 2000.

Joseph Olive. *Global Autonomous Language Exploitation (GALE)*. DARPA/IPTO Proposer Information Pamphlet, 2005.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Traslation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Martin F. Porter. An algorithm for suffic stripping. *Program*, 14(3):130–137, 1980.

M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162, 1964.

Mark Przybocki, Kay Peterson, and Sébastian Bronsart. Official results of the NIST 2008 "Metrics for MAchine TRanslation" Challenge (MetricsMATR08). http://nist.gov/speech/tests/metricsmatr/2008/results/, October 2008.

Reinhard Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 519–526, 1999.

Philip Resnik and Noah Smith. The Web as a Parallel Corpus. *Computational Linguistics*, 29:349–380, 2003.

Antti-Veikko Rosti, Spyros Matsoukas, and Richard Schwartz. Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 312–319, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Libin Shen, Jinxi Xu, and Ralph Weischedel. A New String-to-Dependency Machine Translation Algorithm with a Target Dependency Language Model. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, June 2008.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006.

Matthew Snover, Bonnie Dorr, and Richard Schwartz. Language and Translation Model Adaptation using Comparable Corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 857–866, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece, March 2009. Association for Computational Linguistics.

Matthew G. Snover, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. *Machine Translation (to appear)*, 2010.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. Evaluation of Machine Translation and its Evaluation. In *Proceedings of MT Summit IX*, 2003.

Nicola Ueffing. Using Monolingual Source-Language to Improve MT Performance. In *Proceedings of IWSLT 2006*, 2006.

Jinxi Xu, Ralph Weischedel, and Chanh Nguyen. Evaluating a Probabilistic Model for Cross-lingual Information Retrieval. In *Proceedings of SIGIR 2001 Conference*, pages 105–110, 2001.

Bing Zhao, Matthias Eck, and Stephan Vogel. Language Model Adaptation for Statistical Machine Translation via Structured Query Models. In *Proceedings of Coling 2004*, pages 411–417, Geneva, Switzerland, Aug 23–Aug 27 2004. COLING.

Liang Zhou, Chon-Yew Lin, and Eduard Hovy. Re-evaluating Machine Translation Results with Paraphrase Support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 77–84, 2006.