Reordering Metrics for Statistical Machine Translation

Alexandra Birch
School of Informatics
University of Edinburgh

Doctor of Philosophy
School of Informatics
University of Edinburgh
2011

Abstract

Natural languages display a great variety of different word orders, and one of the major challenges facing statistical machine translation is in modelling these differences. This thesis is motivated by a survey of 110 different language pairs drawn from the Europarl project, which shows that word order differences account for more variation in translation performance than any other factor. This wide ranging analysis provides compelling evidence for the importance of research into reordering.

There has already been a great deal of research into improving the quality of the word order in machine translation output. However, there has been very little analysis of how best to evaluate this research. Current machine translation metrics are largely focused on evaluating the words used in translations, and their ability to measure the quality of word order has not been demonstrated. In this thesis we introduce novel metrics for quantitatively evaluating reordering.

Our approach isolates the word order in translations by using word alignments. We reduce alignment information to permutations and apply standard distance metrics to compare the word order in the reference to that of the translation. We show that our metrics correlate more strongly with human judgements of word order quality than current machine translation metrics. We also show that a combined lexical and reordering metric, the LRscore, is useful for training translation model parameters. Humans prefer the output of models trained using the LRscore as the objective function, over those trained with the de facto standard translation metric, the BLEU score. The LRscore thus provides researchers with a reliable metric for evaluating the impact of their research on the quality of word order.

Acknowledgements

I would like to thank my supervisor Miles Osborne for his support and enthusiasm. His never-ending supply of ideas has been much appreciated. My second supervisor Philipp Koehn has also been instrumental in the completion of this thesis. He has provided the tools and infrastructure which has made it possible to run large scale translation experiments. He also trained the 110 translation models which are then analysed in Chapter 4 of this thesis. Chris Callison-Burch, Phil Blunsom, Barry Haddow and Vera Demberg have all given me valuable advice over the last few years and I shall always be grateful to them. Finally I would like to thank my partner Pedro, and my daughter Alba, for making these last few years the happiest and most rewarding time of my life.

Declaration

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

(*Alexandra Birch*
School of Informatics
University of Edinburgh)

To my mother, who taught me to love language.

Table of Contents

Chapter 1

Introduction

Machine translation can be viewed as consisting of two interrelated problems: predicting the words in the translation and deciding on their order. Although there is a large body of research aimed at improving the word order quality of machine translation systems, there has been surprisingly scant attention paid to how best to evaluate this research. In this thesis we present methods and metrics which allow researchers to understand the word ordering challenges facing them, and also to accurately evaluate the impact of their research.

The rest of this chapter proceeds as follows. First we describe the reasons why evaluating reordering is important. This leads to a discussion of the current evaluation methodology and why it fails to adequately measure reordering performance. We then introduce the general approach taken in this thesis and present the results of the most important experiments. Finally, we describe the main claims made in this thesis and we provide a summary of the work which follows.

1.1 Motivation

In this section we address the problem of reordering. We describe experiments which shows that the quality of the reordering in machine translations is poor, and that the amount of reordering present in a language pair is one of the most important factors in predicting the quality of the resulting translation. We then look at the existing automatic machine translation metrics and discuss why they are inadequate.

1.1.1 Reordering

Finding the correct order for translated words is a difficult problem because of the computational complexity involved. Searching all possible permutations of the words in a sentence of n words, requires $n!$ combinations. A complete search is intractable for all but the shortest sentences. Translation models apply reordering restrictions to the search, and only a small number of possible word orders are considered. The consequence of this is that translation models are able to perform small, local reorderings relatively well. However, large differences in word order are still problematic.

Human translation: to delay overhaul of America's antiquated spy network
Machine translation: could delay the old spy network system operation of the overhaul

Figure 1.1: A section of a Chinese sentence with two English translations, one produced by a human and the other by a machine translation system.

Some language pairs, such as Chinese-English, contain long distance word order differences. Figure 1.1 shows an example a section of a Chinese sentence with its human and machine translations. In the Chinese, the translation of the English noun "overhaul" appears at the end of the sentence while in the English human translation, the noun appears directly after the verb "delay". This is because the English prepositional phrase "of America's antiquated spy network" in Chinese, is in fact a modifier which occurs before the noun. We can see that the machine translation incorrectly follows the word order of the original Chinese, and it is therefore incomprehensible. Word order differences such as these are very frequent in Chinese-English, and consequently machine translation quality remains poor for this language pair.

1.1.2 Reordering in Parallel Corpora

There has been surprisingly little research done on analysing what word order differences exists in human translated corpora. Knowing the reordering characteristics of human translations is an important first step in successfully designing systems to model them.

In this thesis we present a novel method for analysing translated corpora (see Chapter 3 for details). Our method relies upon word alignments which connect words in the

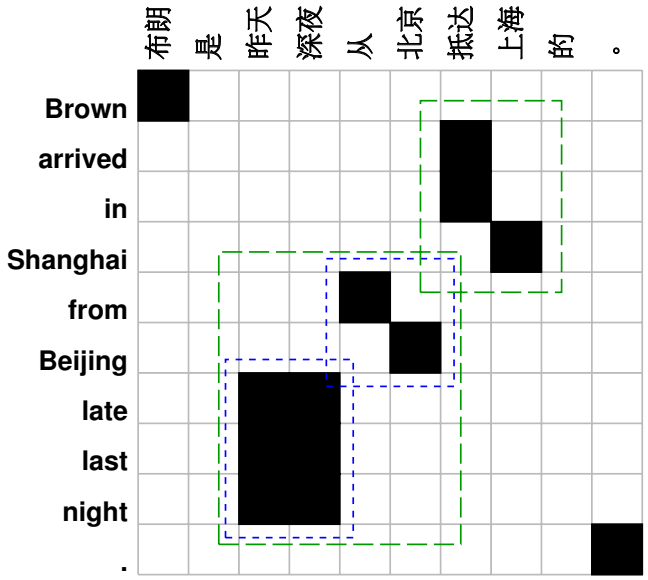


Figure 1.2: Sentence pair with Chinese source sentence and English target sentence. Word alignments are shown in black. Two reorderings are also shown, indicated with different dashed line styles.

source language to their translations in the target language. We define reorderings as two blocks of words which are adjacent in the source sentence and inverted in order in the translation. Figure 1.2 shows a word aligned sentence pair which has two reorderings. Each reordering consists of two blocks whose ordering in the target is inverted with respect to the source. The English target phrases “from Beijing” and “late last night” are a translation of Chinese source phrases which occur in the reverse order. As these are two relatively short phrases, and it represents a typical example of short distance or local reordering. These kinds of reorderings are easier for translation models to capture. The larger reordering involving “in Shanghai” and “from Beijing late last night” would be more of a challenge.

In order to determine how well our models are able to capture the reordering behaviour of the human translations, we collect statistics for parallel corpora by extracting the size and number of reorderings detected in the aligned sentences. We use the Chinese-English and Arabic-English language pairs because they are frequently studied in machine translation and they are important for commercial and strategic reasons.

Table 1.1 shows the average number of reorderings for each sentence pair. This table shows that for every hundred sentences for the Chinese-English language pair, there will be 29 word order differences which span more than 15 words, and 93 word

Language Pair	+15	+7
Chinese-English	0.29	0.93
Arabic-English	0.06	0.29

Table 1.1: Frequency of reorderings which affect more than 15 words or more than 7 words in the target language.

order differences which affect more than 7 words. These statistics show that large word order differences are very frequent in Chinese-English. Arabic-English, however, has many fewer long distance reorderings. It is common practice for statistical machine translation models to restrict the distance that words can be reordered to around seven positions. We can see that for Arabic-English this limit might work reasonably well, however, it will have a deleterious effect on Chinese-English models.

Apart from investigating the reordering behaviour of different models, we also apply our analysis of parallel corpora to determine the effect that the amount of reordering has on translation performance. In Chapter 4, we examine 110 language pairs of data from the European Parliament Proceedings. This wide-ranging study confirms the importance of reordering to the quality of machine translations. Variation in the amount of reordering accounts for 38% of the variation in performance and it has more influence than other factors, such as language relatedness and the morphological complexity of the source and target languages. We therefore demonstrate the importance of research aimed at improving the modelling of reordering, and the need for metrics to evaluate this research.

1.1.3 Machine Translation Metrics

We can only improve the reordering behaviour of translation models if we have reliable metrics for measuring the impact of our changes. There has recently been a great deal of interest in developing machine translation metrics. The Workshops on Statistical Machine Translation (Callison-Burch et al., 2007, 2008, 2009, 2010) and the NIST Metrics for Machine Translation 2008 Evaluation[1] have used human judgement data to compare a wide spectrum of metrics. Unfortunately there is no clear consensus about which is the best metric to use, as a variety of metrics perform well under different test conditions. Most importantly for this thesis, however, is that none of these metrics have been evaluated on how well they measure the quality of word order in translations.

[1] http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2008/

We select three commonly used metrics to highlight the problems that current machine translation metrics face with regard to measuring reordering performance: the BLEU score (Papineni et al., 2002); METEOR (Lavie and Agarwal, 2007); and TER (Snover et al., 2006). None of these metrics take the size of the word order differences into account. In Section 6.4 we demonstrate that these metrics are highly sensitive to the quality of the word choice, but that they are largely insensitive to the quality of the word order.

1.2 Permutation Distance Metrics

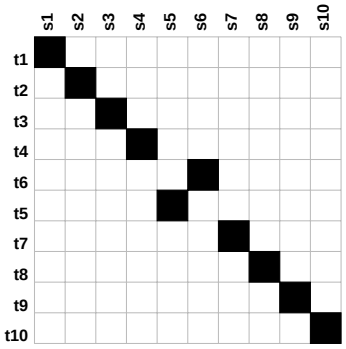
We have looked at the reasons why reordering is important, and we have mentioned some problems with current machine translation metrics. We now describe our approach to measuring the quality of word order in translations and we provide an example which illustrates the advantages of our approach.

1.2.1 Approach

In Chapter 5 we present a novel reordering metric which is able to isolate the effect of reordering by operating over alignments, not translations. Although we have already suggested a method for analysing reordering in parallel corpora, these methods cannot measure the similarity of the reference and translation alignments. We therefore suggest a different approach to measuring word order quality.

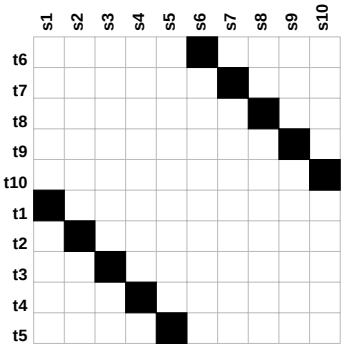
First, we convert alignments into permutations by iterating over the source words and extracting the relative order of their aligned target word. Figure 1.3 contains examples of source sentences (s1,...,s10) which are aligned to target sentences (t1,...,t10) which have different word orders. The resulting permutations are shown below the alignments. In example (a) there is a small word order difference, where only two words are swapped, and in example (b) there is a large word order difference, where the order of the two halves of the sentence has been swapped.

Formally permutations are defined as sets of ordered data and finding the distance between ordered sets is one of the fundamental problems of computer science. These distances have applications in many contexts such as statistics, coding theory, computing, DNA research and so on. In this thesis, we use distance metrics to compare two permutations: the permutation representing the source-reference alignment and the permutation representing the source-translation alignment.



(1 2 3 4 • 6 • 5 • 7 8 9 10)

Example (a)



(6 7 8 9 10 • 1 2 3 4 5)

Example (b)

Figure 1.3: Synthetic examples of two sentence pairs, showing their word alignment grids and below them, their permutations. Bullet points represent the non-sequential gaps in the permutation.

We use two different permutation distance metrics: the Hamming distance and Kendall's tau distance. These are well known distance metrics which are sensitive to the number of words which are out of order. The Hamming distance is an absolute measure of the amount of disorder between two permutations, and the Kendall's tau distance is a measure of the relative disorder. Kendall's tau distance is sensitive to how far words are out of order. As it is reasonable to suppose that humans are also sensitive to the size of reorderings, and not just to their number, we suggest that Kendall's tau is the more reliable metric. However, the Hamming distance is a simple and useful baseline metric.

Our metrics have a number of features which makes them well suited to measuring reordering in statistical machine translation:

- They measure the number of words which are out of order.
- They correlate with human judgements of reordering.
- The scores are meaningful at the sentence level. This allows researchers to analyse results at different levels of granularity and also makes it easier to inspect test results.
- They are language independent as they abstract away from the word choice in the translation. This is important for the metric to be widely useful.

- They are efficient to calculate. This means that they can be applied as an objective function for tuning the parameters of translation models.

Our approach to measuring reordering performance is quantitative. We are measuring the amount of the difference in word order. Humans are likely to also be sensitive to the kinds of words or phrases that are reordered. Taking this into account however, would require sophisticated linguistic knowledge which would be language dependent and introduce errors. Instead, we focus on simple intuitive measures.

1.2.2 Example

In order to highlight the problem of the current MT metrics, and to demonstrate the advantages of using permutation distance metrics, we refer to the two sentence pairs shown in Figure 1.3. We calculate the scores for these two sentences, the machine translation metrics and for the permutation distance metrics. The sentences are compared to a monotone reference sentence (t1, ... ,t10). Table 1.2 presents the results. In order to facilitate comparison, all metrics are transformed such that 0% represents the worst possible score, and 100% represents the best possible score.

Example	BLEU	METEOR	TER	Hamming	Kendall's tau
(a)	61.8	86.9	90.0	80.0	79.0
(b)	81.3	92.6	90.0	0.0	25.5

Table 1.2: Metric scores for examples in Figure 1.3 which are calculated by comparing the permutations to the monotone translation.

The example sentence pair in (a) represents a small reordering, and the sentence pair in (b) a large reordering. However, the machine translation metrics, such as BLEU, fail to recognise this. They are sensitive to breaks in order, but not to the actual amount of word order difference. The BLEU score detects three breaks in order in example (a), shown by the bullet points in the permutation, and assigns it a score of 61.8. There is only one break in example (b) and therefore more n-grams are matched and it consequently assigns a higher score of 81.3. METEOR counts the number of blocks that the translation is broken into, in order to align it with the source. (a) is aligned using four blocks and scores 86.9, whereas (b) is aligned using only two blocks and scores 92.6. TER counts the number of edits, allowing for block shifts. TER applies one

block shift for each example, resulting in an equal score of 90.0 for both sentences, thus demonstrating its insensitivity to the amount of reordering.

The reordering metrics correctly assign a lower score to (b) as they recognise the greater number of words affected by reordering. The Hamming distance detects two words out of order in (a), resulting in a score of 80.0, and all words are out of order in (b) and so it assigns the worst score of 0.0. Kendall's tau is the only metric which takes the distance words have moved into account. For the sentence in (a), Kendall's tau detects only one pair of words which are out of order, resulting a score of 79.0. For the sentence in (b) there are many more differences in order resulting in the lower score of 25.5. Even though (b) contains a large reordering, the words inside the two inverted blocks retain their relative order, and the Kendall's tau distance recognises this.

The examples in Figure 1.3 assume a perfect lexical match between the references and the translation. In real translation examples, the machine translation metrics are further hampered by lexical differences. Words which are not identical in the translation and the reference, are either considered to be breaks in order, or metrics attempt to align them using heuristics. Permutation distance metrics improve over current machine translation metrics, first by isolating the reordering component of the translation, and then by measuring the actual difference in word order.

1.2.3 Evaluation of Reordering Metrics

Using a rigorous evaluation methodology we demonstrate that permutation distance metrics are more appropriate than current metrics for measuring the quality of word order in translation.

Automatic metrics must be validated by human judgements. It is an open research question as to how best to utilise humans to evaluate translation. Most human evaluations are collected on the varied output of translation systems which makes it difficult to isolate the effect of reordering. We develop a novel human evaluation task which specifically measures reordering performance. This experiment shows that humans are able to distinguish between sentences with different levels of disorder. Furthermore, human judgements of reordering are shown to correlate strongly with permutation distance metrics.

Measuring word order differences in isolation is interesting, but for many circumstances a comprehensive metric is more appropriate. We present a novel metric, the Lexical Reordering score (LRscore), which combines these two important aspects of

machine translation quality. The LRscore is shown to correlate more strongly with human preference judgements than other machine translation metrics.

We also explore the ability of the LRscore to guide the reordering behaviour of translation models. We use the LRscore as an objective function during the tuning of the translation model parameters. We show that humans prefer the output of translation models trained with the LRscore over those trained with the BLEU score. We also show that when training with the LRscore, there is no discernible drop in performance with respect to the BLEU score.

1.3 Overview

The main claims defended in this thesis are the following:

- Reordering is an important factor in determining the overall performance of translation systems.
- Current machine translation metrics do not adequately measure reordering performance.
- Permutation distance metrics capture the quality of word order better than current machine translation metrics.

Current metrics are hampering progress of reseach in machine translation because they are not able to measure improvements in reordering performance reliably. Permutation distance metrics provide the solution by reflecting the true amount of word order difference between reference and translation sentences. Our metrics provide the key to the future development of the field.

1.3.1 Road Map of Thesis

This section contains a short summary the chapters in the rest of this thesis.

Chapter 2 provides background information about models of machine translation and how they deal with the reordering challenge. This is important for understanding the analysis of the reordering seen in the output of two different translation models presented in Chapter 3. Chapter 2 also provides a detailed discussion of the current approaches to evaluating machine translation output. Both human

evaluation campaigns and automatic evaluation metrics are discussed and their shortcomings regarding reordering are presented.

Chapter 3 proposes a method for extracting the reorderings seen in aligned parallel corpora. This results in a set of binary reorderings, where a block of contiguous words in the source is swapped in order in the target sentence. Using statistics about the distributions and sizes of the reorderings, we analyse the properties of two divergent language pairs, Chinese-English and Arabic-English. We then translate the source sides using two important translation models, the phrase-based model and a synchronous grammar-based model called the hierarchical model. We compare the reorderings seen in the output of the translation models to the human translated references and to each other. Part of this work has been published in Birch et al. (2008) and Birch et al. (2009).

Chapter 4 presents a survey of 110 different language pairs drawn from the Europarl project. By including so many language pairs, we are able to provide a "big-picture" view of the challenges facing machine translation. We start by extracting certain characteristics of the language pairs, such as the amount of reordering and a measure of language family relatedness. We train translation models and perform regression analysis, showing that reordering is the factor which correlates most strongly with translation performance. This extends sections of previous work published in Birch et al. (2008) and Koehn et al. (2009).

Chapter 5 proposes a method of evaluating reordering performance based on permutation distance metrics. We describe how permutations are extracted from alignments. We then describe two distance metrics, the Hamming distance, and Kendall's tau distance and how they are appropriate for comparing the word order seen in a reference sentence with the word order in a translation.

Chapter 6 evaluates the permutation distance metrics using three experiments. The first establishes that the metrics are able to distinguish human references from machine translations. The second proposes a novel human evaluation task which isolates reordering. We then extract the correlation of the permutation distance metrics and baseline metrics with human judgements of word order quality. Finally, we examine which aspects of a translation influence the current baseline machine translation metrics and show that they are largely insensitive to the quality of word order. Chapters 5 and 6 extend work published in Birch et al. (2010).

Chapter 7 presents a metric which combines lexical and reordering metrics in a simple, decomposable metric called the LRscore. We show how this metric is more meaningful and intuitive than current machine translation metrics and that it correlates better with human rank judgements of overall sentence quality. Preliminary results for this chapter have been published in Birch and Osborne (2010).

Chapter 8 demonstrates the usefulness of the LRscore. First, we apply the LRscore while training the parameters of our translation model to see whether information on reordering can help guide the translation model to produce better reorderings. We show that humans prefer the output of translation models trained with the LRscore over those trained with the BLEU score. Next, we present a set of experiments which show how using reordering metrics is more informative and more accurate than using other machine translation metrics when applying changes to the reordering behaviour of the model.

Chapter 9 summarises the main contributions made by this thesis, and gives an outlook on future work.

Chapter 2

Background

In this chapter we introduce statistical machine translation. We describe important work related to the reordering problem such as reordering models and reordering restrictions on the search. We then provide a detailed discussion of evaluation metrics for machine translation, focusing on their ability to measure reordering performance.

2.1 Statistical Machine Translation

Machine translation is a hard problem because of the highly complex, irregular and diverse nature of natural language. A principled approach to this problem is to use statistical methods to make optimum decisions given incomplete data. In statistical machine translation, we are given the source language sentence consisting of J words, $s_1^J = s_1 \cdots s_j \cdots s_J$, which is to be translated into the target language sentence, $t_1^I = t_1 \cdots t_i \cdots t_I$, consisting of I words. We must search for the highest probability sentence amongst all the possible target language sentences:

$$\hat{t}_1^I = \arg\max_{t_1^I} \{Pr(t_1^I|s_1^J)\} \tag{2.1}$$

The argmax operation denotes the search problem. We use Bayes rule to reformulate Equation 2.1:

$$\begin{aligned} \hat{t}_1^I &= \arg\max_{t_1^I} \frac{\{Pr(t_1^I) \cdot Pr(s_1^J|t_1^I)\}}{Pr(s_1^J)} \\ &= \arg\max_{t_1^I} \{Pr(t_1^I) \cdot Pr(s_1^J|t_1^I)\} \end{aligned} \tag{2.2}$$

The denominator $Pr(s_1^J)$ does not depend on t_1^I and it can therefore be ignored. This is known as the noisy channel approach and was suggested by Brown et al. (1990). The

noisy channel approach is commonly used in speech recognition and can be traced back to early information theory (Shannon and Weaver, 1948). It allows for an independent modelling of the target language model $Pr(t_1^I)$ and the translation model $Pr(s_1^J|t_1^I)$. The language model being a measure of how well formed the target sentence is and the translation model measures the likelihood of the target sentence being a translation of the source sentence.

The language model can be learned from large amounts of text in the target language and is usually based on n-gram frequencies. The translation model must be learned from parallel texts, or *bitexts*, where each sentence in one language is paired with a human translated sentence in the other language. The key to training a translation model is to use the idea of an *alignment*. Brown et al. (1990) define an alignment between a pair of strings as an object indicating for each word in the target string, the word in the source string from which it arose. The alignment is defined as a function $a : \{i \rightarrow j\}$. See Figure 2.1 for an example of a parallel sentence and its word alignment. For this sentence, "we" and "did" are aligned to "hemos" $a : \{1 \rightarrow 5, 2 \rightarrow 5\}$ and "not" to "no" $a : \{3 \rightarrow 4\}$ and so on.

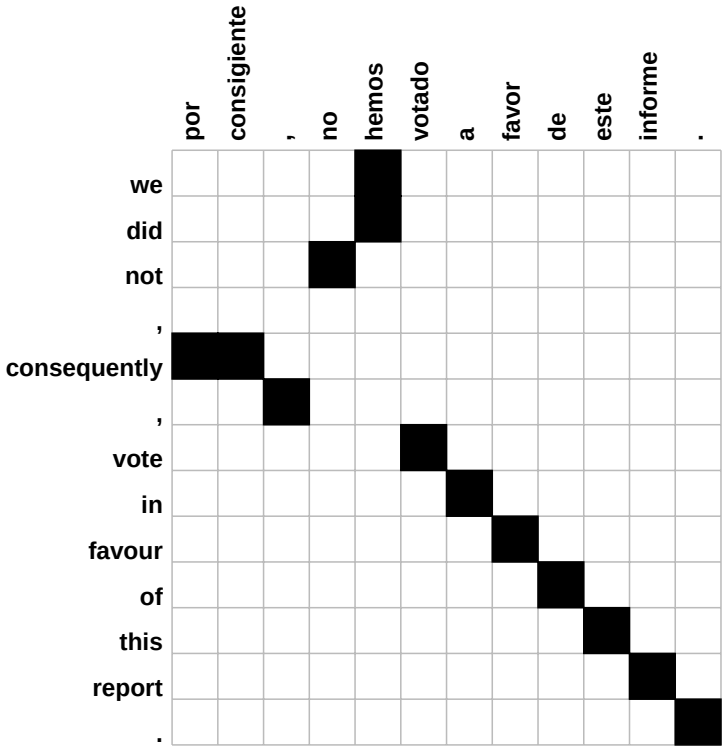


Figure 2.1: Spanish-English parallel sentence with word alignments marked in black squares.

Alignments are necessary for training translation models. The expectation maximisation (EM) algorithm (Dempster et al., 1977) is an iterative learning method that optimises parameters in situations where there is incomplete data. In this case the incomplete data is the hidden alignment information which is not directly available in the parallel corpus. EM calculates the probability of the translation model $Pr(s_1^J|t_1^I)$, by summing over all possible alignments, and then marginalising out the choice of alignment:

$$Pr(s_1^J|t_1^I) = \sum_{a_1^J \in \mathcal{A}} Pr(s_1^J, a_1^J|t_1^I)\} \tag{2.3}$$

where $\mathcal{A}$ is the set of all alignments over the sentence pair. The sum over all alignments is usually impossible to compute exactly and so it is necessary to restrict our EM training to considering only a small number of promising alignments that lie close to the most probable alignment, called the *Viterbi alignment*. The following equation defines the Viterbi alignment which depends on p_θ, the translation parameters of the current iteration of EM:

$$\hat{a}_1^J = \arg\max_{a_1^J} p_\theta(s_1^J, a_1^J|t_1^I) \tag{2.4}$$

The model defined so far operates over words, which is problematic when the relationship between the source and target words is not one-one. The alignment template translation model (Och et al., 1999) and others (Marcu and Wong, 2002; Koehn et al., 2003; Tillmann, 2003) advanced the state of the art by moving from using words as the basic unit of translation, to using *phrases*. Phrases in this context need not have any syntactic value and are simply sequences of words. They allow the translation models to learn local reorderings and idioms, and account naturally for the insertion and deletion of words in a local context. Performing EM with phrases is extremely expensive (Marcu and Wong, 2002; Birch et al., 2006) and so phrase pairs are extracted from sentence pairs where a Viterbi word alignment has been extracted. Phrase pairs are collected from an alignment by extracting all blocks which include aligned points, and are internally consistent. Consistent means that all the words within the phrase pair are only aligned to words within the phrase pair. Counts of the phrases are collected and used to calculate the probabilities of the phrase-based translation model $\phi(\bar{s}|\bar{t})$:

$$\phi(\bar{s}|\bar{t}) = \frac{count(\bar{s},\bar{t})}{count(\bar{s})} \tag{2.5}$$

The noisy channel approach struggles to include additional sources of knowledge in its probabilistic framework. The direct maximum entropy translation model was suggested by Och and Ney (2002) as an interesting alternative to the noisy channel approach. This log linear model can be easily extended by adding new feature functions which are each weighted separately. The log linear model directly models the posterior probability $Pr(t_1^I|s_1^J)$ and it allows us to use an arbitrary set of M feature functions $h(t_1^I, s_1^J)$ which are combined using optimised weights λ_m:

$$Pr(t_1^I|s_1^J) = \exp\left(\sum_{m=1}^{M} \lambda_m h_m(t_1^I, s_1^J)\right) \cdot \frac{1}{Z} \tag{2.6}$$

Z is the normalisation constant and as it is a sum over all possible t_1^I, it is not needed for the search, and thus we obtain the following decision rule:

$$\hat{t}_1^I = \arg\max_{t_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(t_1^I, s_1^J) \right\} \tag{2.7}$$

Equation 2.7 defines the decoding problem that the translation system must solve, and many alternative approaches have been suggested. Optimal search, such as A* search (Och et al., 2001) and integer programming (Germann et al., 2001), struggle to decode long sentences efficiently and greedy search algorithms can commit serious search errors (Germann et al., 2001). The most successful algorithms are based on breadth-first search with pruning (Tillmann and Ney, 2003; Och and Ney, 2004). This is called beam search. The beam search algorithm described by Koehn (2004a) generates multiple hypotheses which cover the target sentence from left to right. As the hypotheses grow, their probabilities are updated. Each new hypothesis extends the coverage of the source sentence. Hypotheses are placed in a stack with other hypotheses with the same number of source words covered. This allows for pruning and only the best n hypothesis are stored. The hypothesis with the highest probability that covers the source sentence is the output of the search. Figure 2.2 shows an example of hypothesis expansion for the Spanish sentence “Maria no daba una bofetada a la bruja verde”:

Equation 2.7 defines the translation model as a log-linear model, where $h_m(t_1^I, s_1^J)$ are the features and λ_m are the weights that balance the features. In this framework, the modelling problem amounts to developing suitable feature functions that capture the properties of the translation task, such as the probability of a target phrase given a source phrase $p(\bar{t}|\bar{s})$. The training problem amounts to obtaining suitable parameter

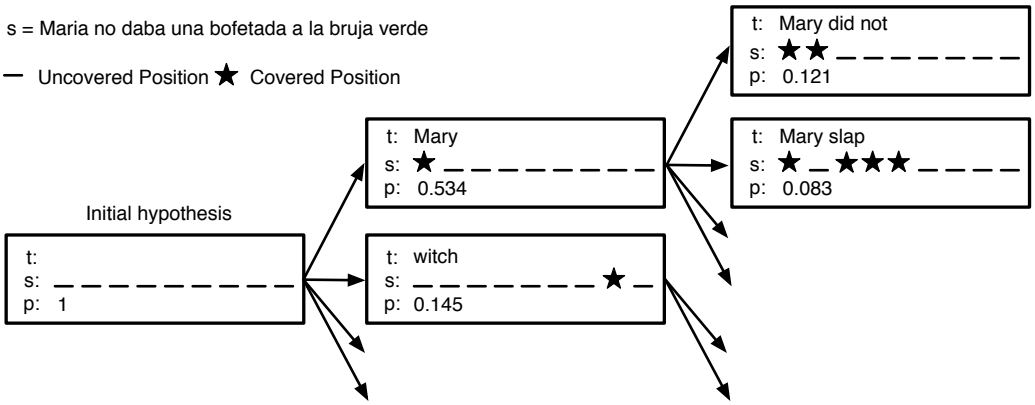


Figure 2.2: Hypothesis expansion in the beam search. Each expansion generates a new target word in the target string (t), marks the covered source words in the source bitvector (s) and calculates the updated probabilities in (p). From Koehn (2004)

values for λ_m. Och (2003) demonstrates that when setting these weights, the final evaluation metric should be taken into account. This is achieved by choosing the weights so as to maximise these scores given by the translation metric on a development set. In order to maximise the scores, a gradient-based optimisation technique cannot be used as the error surface is not smooth. Och also suggests applying an efficient algorithm to find good weights. This process is called Minimum Error Rate Training (MERT), and it is an essential part of the development of many machine translation systems.

We have thus presented an overview of the basic components of a translation system. This discussion has provided the context for the following section, which describes how word order differences are addressed in machine translation.

2.2 Reordering

Natural languages display a great variety of different word orders. The first researchers in statistical machine translation called this effect *distortion* but it is also known as *reordering* (Brown et al., 1990). Part of the reason that reordering is difficult to account for, is that different language pairs pose different reordering challenges. Many of the language pairs that have driven research, such as French-English, can on the whole be translated successfully when reordering is restricting to short, local movements. Other language pairs are more challenging because of long distance movement, or significant differences in syntactic structure. The basic Japanese word order is Subject-Object-Verb and long distance reordering is required in order to translate correctly into

English, which is a Subject-Verb-Object language. The computational complexity of exploring all possible word order differences means that machine translation is only able to allow a small number of orderings, normally those that are close to the monotone. Long distance reorderings are rarely considered, and even if they are, they would most likely be assigned a very low probability.

Trying to capture systematic differences in word order in a probabilistic framework is done using a *reordering model*. Many kinds of reordering models have been proposed, but most are relatively simple in order to restrict their impact on the size of the model, and on the efficiency of the search.

Apart from predicting movement of words in translation, restricting the kinds of movement allowed using *reordering restrictions* has also been widely studied. Here the restrictions are imposed primarily to improve the efficiency of the search. Good restrictions for reordering will allow plausible reorderings and discard large numbers of implausible reorderings. In practice, restricting the search often improves our chances of finding good word orderings.

Reordering models separate the ordering information from the translational probabilities. Syntax-based models merge the translation model and the reordering models in a synchronous grammar. Reordering models are weak and do not guide translation models to high probability hypotheses, whereas syntax-based model can succinctly encode long distance reorderings, by limiting the possible orderings to those seen in the training data.

The rest of this section will describe previous work done on reordering in statistical machine translation for different kinds of models. Deficiencies in current research and relevance to future work will be noted.

2.2.1 Reordering Models

Reordering in statistical machine translation was first proposed in the series of alignment models developed at IBM (Brown et al., 1993). These models are important because they introduced the fundamental concepts of statistical machine translation.

The simplest model, IBM Model 1, considers all possible alignments between words in the source and target sentences to be equally likely. This unrealistic assumption allows the model to search all possible alignments efficiently. Its parameters are then used to initialise the more complex alignment models. IBM Models 2 and 3 introduce distortion. Distortion probabilities are based on the absolute positions of the

source and the target words in their respective sentences, and the length of these sentences. These models do not generalise well since reordering will not often occur in the same way for the same word position over different sentences. This is especially true for longer sentences, where any estimates will not be realistic and will be affected by sparsity. Furthermore, this approach does not take into account the fact that words tend to move in blocks and not independently. IBM Models 4 and 5 replace absolute word positions with relative positions: the alignment of a word is dependent on the alignment of the previous word. The entire source and target vocabularies are reduced to a small number of classes for the purpose of estimating distortion parameters.

In the IBM Models, the addition of more sophisticated alignment models comes at the cost of greater complexity for the search algorithm. With Models 3, 4 and 5 certain optimisations can no longer be performed and therefore the search must be approximated.

Och and Ney (2003) analyse the different IBM alignment models. They show that first order dependencies in distortion are very important. The most successful alignment models combined the IBM Model 4 first order dependency in the source with the Hidden Markov alignment model described by Vogel et al. (1996) which has a first order dependency in the target. They also note that correct smoothing improves performance considerably.

Although the word-based translation models have been superseded by phrase-based and grammar-based models, they are still widely used together with EM to align large parallel corpora as they are relatively efficient. Corpora are aligned in both source-target and target-source directions, and the final alignment taken is calculated by combining their intersection and union using heuristics (Koehn et al., 2003).

2.2.2 Reordering Restrictions

Reordering restrictions on the search are necessary because even for the simplest form of statistical models like IBM Model 1, the decoding problem is NP-complete (Knight, 1999), which means that it is probably exponential in the length of the observed sentence. As Knight (1999) explained, this complexity is due to the combination of factors not present in other decoding problems: both overlapping bilingual dictionary entries and the word reordering problem. Efficiency considerations are therefore crucial.

Reordering restrictions for word-based decoders were introduced by Berger et al. (1996) and Wu (1995). The decoder presented in Berger et al. (1996) is based on the

A* search algorithm. Figure 2.3 represents a coverage vector over the source sentence where various different states are represented. Each position in the source sentence is marked as covered or uncovered. Berger defined a reordering constraint where the current target word being translated can only be generated from the last k uncovered source words. In Figure 2.3 k is equal to four and the source words which would be possible extensions to different hypotheses are marked with question marks. This constraint is sometimes called the IBM constraint, and it is commonly used today in phrase-based models. This means that there are $(k-1)$ words in the source that can be skipped or left untranslated until later in the sentence.

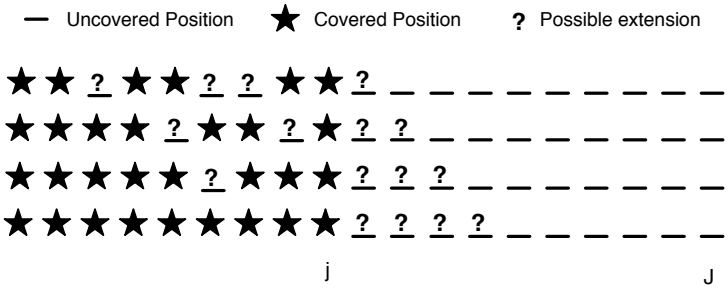


Figure 2.3: Illustration of IBM constraint: current word can be generated from last 4 words. From Zens and Ney (2003)

This models some distortion problems reasonably well, but not others. Using the German-English language pair, if we were translating into German, we could leave the verb untranslated until the end of the sentence, which is often required. However, in the other direction we would be unable to translate the verb in its correct place in English, if it was more than k positions further along in the German sentence. Zens and Ney (2003) show that these constraints allow for a polynomial time search.

Some phrase-based models apply an even stricter reordering constraint. The MOSES model specifies that the last word in a new phrase must translate the source word which occurs in a maximum of k (covered or uncovered) positions from the left-most untranslated word. Lopez (2009) compares a number of reordering restrictions and mentions their complexity.

Wu (1997) described another polynomial time algorithm which allows greater flexibility in ordering. He introduced the inversion transduction grammar (ITG), applying synchronous context free grammars (SCFG) to machine translation for the first time (see section 2.2.4 for further discussion on syntax-based models). An ITG is a grammar where each rule produces two streams of output, one for each language. ITG

allows the output to occur either in the same or in an inverted order. ITGs are reduced to normal form and a small example grammar is shown in Figure 2.4.

$$
\begin{array}{lcl}
\text{A} & \rightarrow & [\ \text{B C}\ ] \\
\text{A} & \rightarrow & <\ \text{B C}\ > \\
\text{B} & \rightarrow & \text{negro / black} \\
\text{C} & \rightarrow & \text{gato / cat}
\end{array}
$$

Figure 2.4: A small example of an inversion transduction grammar with a monotone rule [] and an inverted rule <>

The rule with the [] brackets indicates that the ordering within the two output streams is the same, whereas the rule with the <> brackets indicates an inverted order. In this grammar, [B C] would produce "gato negro / cat black", and < B C > would produce the correct output "gato negro / black cat". This is represented graphically in Figure 2.5.

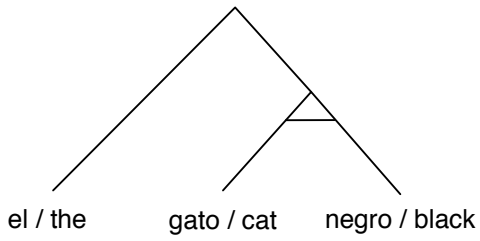


Figure 2.5: The graphical parse tree notation for ITG: the inverted rule <> is indicated with a horizontal line. The Spanish is read in the usual order, but for the English, the line means that the right subtree is read before the left.

An ITG allows the modelling of long distance dependencies, as a rule can cover a whole sentence. There are, however, many kinds of reordering that it cannot capture. ITG requires that the reordering occurs between two child nodes of the same parent and that the derivation trees between sentences are isomorphic. Two trees are isomorphic if the structures are identical, and only the order of the child nodes is allowed to vary. This restriction makes ITGs more efficient, but also less able to model some of the dependencies between languages.

Zens and Ney (2003) compare the decoding restrictions proposed by Berger et al. (1996) and Wu (1997). They attempt to investigate the *coverage* of the two types of

constraints using Viterbi word alignments. Every sentence is checked to see if the word alignment satisfies the constraints and the ratio of sentences that satisfy the constraints to the total number of sentences is referred to as coverage. The IBM constraints result in higher coverage than the ITG constraints for the French-English Canadian Hansard corpus, but the coverage was almost the same for the German-English Vermobil task. German-English has more long distance movement of words due to the verb final nature of German and therefore the ITG constraint is stronger for this corpus. This work is interesting because it presents an empirical comparison of the reordering capabilities of a finite-state based model and a context-free model. This is something we will expand upon in the thesis.

2.2.3 Phrase-Base Reordering

The shift from word-based statistical machine translation to phrase-based is largely motivated by the fact that bilingual phrase pairs, such as the alignment templates described by Och and Ney (2004), capture local reorderings. These have been shown to improve the quality of translations considerably. However, the ordering of phrases remains a challenge.

The phrase-based model described by Koehn et al. (2003) introduces a relative distortion model which is based on the assumption that monotone decoding is generally preferable. It is equivalent to summing over the distance (in the source language) between phrase pairs that are consecutive in the target language. The function is defined as $\sum_{i=0}^{I} abs(a_i - b_{i-1}) - 1$, where a_i denotes the start position of the source phrase that was translated into the ith target phrase, and b_{i-1} denotes the end position of the source phrase translated into the $(i-1)th$ target phrase. An example is given in Figure 2.6. For the target phrase "in my opinion" there is a monotone ordering with the start of the sentence, $a_1 = 1$, and $b_0 = 0$ and thus no distortion is detected. Distortion is detected for the phrase "current" where $a_3 = 6$ and $b_2 = 4$ and the difference is this $2 - 1$. The relative distortion model score for this example is equal to two.

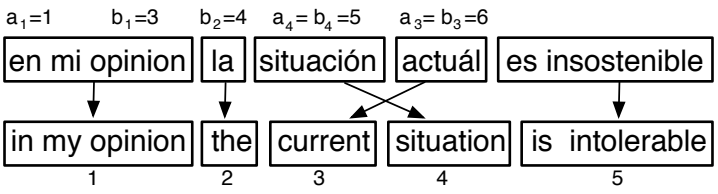


Figure 2.6: An example for the relative distortion model.

Reordering models can benefit from knowing which phrases are being reordered, and not just their relative distortion. Och et al. (2004) and Tillman (2004) suggest a lexicalised orientation model for phrases. Phrases pairs are assigned probabilities which relate to them having monotone, inverse or disjoint orders both with the phrase pairs that precede them in the sentence (backward direction) and with the phrase pairs that follow them (forward direction).

In Figure 2.7 we can see a word alignment from which phrases pairs are to be extracted. The orientations of the phrase pairs will correspond to the word alignments from which they have been extracted. This will mean that phrase pairs such as "situacion/situation" will learn a tendency for inverse orientation in the backward direction, and disjoint orientation in the forward direction.

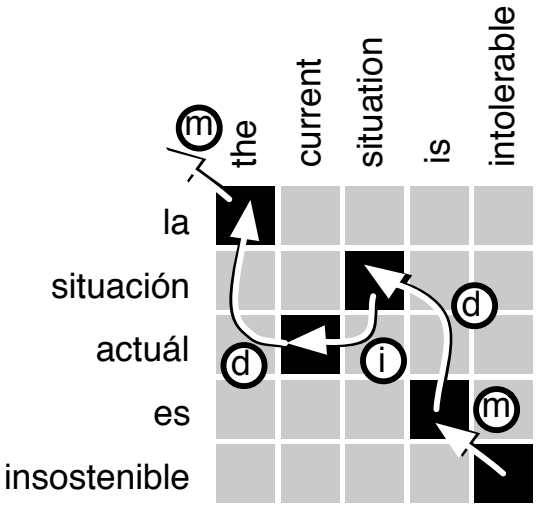


Figure 2.7: An example of lexicalised orientations with arrows indicating the backward direction, with three possible orientations (m)onotone, (i)nverse, and (d)isjoint.

Even if lexicalised reordering models have been successfully integrated into state-of-the-art phrase-based systems, they have some notable limitations. Thy have no ability to generalise, which is a problematic for unseen words, which are assigned small non-zero probabilities, and for phrase pairs that have only been seen a few times and have no reliable orientation statistics. Another important drawback is that the orientation information is limited to local decisions. The probabilities are assigned based on the ordering of phrase pairs which occur immediately before or after the current phrase pair, and therefore provide no information on longer distance reordering.

There have been a number of different papers proposing methods for extending

the range of reordering models for the phrase-based model. Kumar and Byrne (2005) suggest learning ordering information within a bigger window, although still limited to local distances. More recently Galley and Manning (2008) proposed a hierarchical reordering model. They used the shift-reduce parser extract reorderings from alignments, in a similar fashion to the way we extract reorderings from alignments in Chapter 3. This model allows longer distance reordering rules to be incorporated into a standard phrase-based system in an efficient manner.

These reordering models show some improvement over the basic phrase-based model, but they still do not significantly extend the ability of the model to capture reordering behaviour. Phrase-based systems still rely heavily on the language model to select among possible word order choices and reordering models have limited influence.

2.2.4 Syntax-Based Models

The essentially flat structure of phrase-based models means that they struggle to model the complex structural differences that can occur between languages. Synchronous grammars have been extensively investigated for their suitability to statistical machine translation. The main motivation for using a grammar based formalism is to capture long-range reorderings between source and target. Due to recursive sharing of subtrees among many derivations, we can search for hypotheses in polynomial time using dynamic programming algorithms (Melamed, 2003).

As introduced in Section 2.2.2, the ITG was the first synchronous context-free grammar to be proposed for statistical machine translation and it is a restricted case of syntax-directed grammars which are used in the theory of compilers Aho and Ullman (1969). ITG requires that the source and target sentences to be isomorphic which severely restricts the reorderings which are allowed between languages. Figure 2.8 shows the basic sentence structure (subject verb object or SVO) of a source English sentence. If we restrict ourselves to isomorphic trees in our synchronous grammar, we can only swap the order of child nodes. We can thus model the “SVO”, or the reordered “VOS” and “SOV” word order in the target. We cannot, however, model the “VSO” word order which is the canonical word order for Arabic.

Fox (2002) performed the first empirical study that showed that many common translation patterns fall outside the scope of the child reordering model. She found that even relatively similar languages such as English and French suffer from many struc-

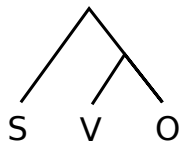


Figure 2.8: A basic English sentence with subject verb object (SVO) word order.

tural differences. For example, the "ne ... pas" construction wraps around a French verb so it will usually result in non isomorphic structure. Other languages are expected to be even more divergent in structure.

The difficulty in modelling non-isomorphic structures is a problem all synchronous grammars face. One way of dealing with this is to flatten the tree, giving more reordering possibilities amongst the larger number of child nodes as Yamada and Knight (2001) did. Another way of alleviating the non-isomorphism problem is to use synchronous grammars with richer expressive power, and whose rules apply to larger fragments of the tree. Eisner (2003) and Galley et al. (2004) use synchronous tree-substitution grammars which generate more tree relations than synchronous context-free grammars by using elementary structures beyond the scope of one-level context-free productions. These would be able to handle the reordering problem posed in Figure 2.8. Accounting for all non-isomorphic structure can be very difficult however, for instance, Galley et al. show that to cover all their Chinese-English sentence alignments, they would need extremely large tree fragments containing up to 43 nodes.

Syntax-based models are widely considered to have the right amount of structural information to model word order differences, but finding the balance between expressive power and efficiency is a serious challenge. More powerful grammar formalism are less efficient and often restricting the number of terminals and nonterminals allowed in each rule is necessary.

One approach has been particularly successful in demonstrating the benefit of using structure. The hierarchical phrase-based model (Chiang, 2005, 2007)) is based on the intuition that since phrases are good at learning the reordering of words, they can be used to learn the reordering of phrases too. Chiang defines a model based on *hierarchical phrases* which consist of words and place holders for subphrases. This model is formally a weighted synchronous context-free grammar but it is learned from a bitext without any syntactic annotations. The grammar consists of synchronous hierarchical

phrases where subphrases are marked by the single nonterminal symbol X. This allows the rules to act as both discontinuous phrases and as powerful reordering rules. Although this model is basically a lexicalised ITG, and its rules are limited to binary branches, it achieves performance comparable to state-of-the-art phrase-based models. The reason for this could be that even though it can potentially learn more powerful reordering rules, it is still able to retain the lexical dependencies that phrase-based systems retain. Another factor in the hierarchical model's success could also be its ability to cross linguistic phrase boundaries, making it more robust to rewording and loose translations. In Figure 2.9 we can see an example of a hierarchical phrase pair which is created by replacing subphrases with nonterminal symbols. The extraction process generates a large number of rules, as all possible subphrases are extracted. The rule $X \rightarrow (X_1$ duonianlai de $X_2 \parallel X_2$ over X_1 years) encodes a reordering over subphrases indicated by the relative order of the aligned non terminals X_1 and X_2.

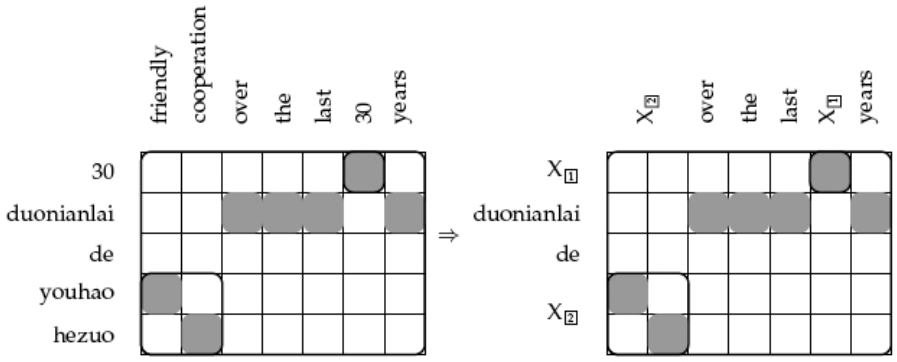


Figure 2.9: Creating a hierarchical phrase pair from word alignments. From Chiang (2007).

Syntax-based models are much more expressive than phrase-based models, but have generally lagged behind in terms of performance on large-scale evaluation campaigns. Part of the problem with synchronous grammar is that the size of the grammar becomes very large and this impacts on the space and time complexity of the decoding algorithm. This is exacerbated by including the language model scores. As soon as hypotheses with gaps on the target side can be created, all intermediate language model scores need to be stored until the final score can be calculated. In order to allow models to scale up to longer sentences and large corpora, aggressive pruning and reordering restrictions are necessary. The hierarchical model, for instance, allows rules to span a maximum of 10 source words. Rules are then glued together in a monotone fashion. This means that the model has a complexity which is linear with the length of

the sentence, and that the maximum reordering limit is of size 10.

Some syntactic models claim to implement global models of ordering. Chang and Toutanova (2007) present a reordering model which predicts the position of child nodes relative to their parent nodes using global features. Although their model is in theory capable of handling global features, in practice only local features are applied.

Syntax-based models are now competing with phrase-based models, as more efficient search algorithms and optimised pruning strategies such as those described in Huang and Chiang (2005) have allowed them to scale up to larger corpora.

2.2.5 Monolingual Reordering

Reordering makes decoding a computationally challenging task that cannot be performed exactly. If reordering is treated as a monolingual problem, allowing the decoding to be monotone, it has much less impact on efficiency. Monotone decoding translates words in the same order as they appear in the source language.

Xia and McCord (2004) propose a method to automatically acquire rewrite patterns that can be applied to any given input sentence so that the rewritten source and target sentences have similar word order. Apart from being able to perform an exact search, reordering of the source sentence uses linguistically motivated rules that a phrase-based model would not be able to incorporate. These linguistic rewrite rules allow for generalisation to unseen words. For instance the rewrite rule "Adj N $\Rightarrow$ N Adj" expresses the fact that in some languages the adjective precedes the noun and in others it follows the noun. These rewrite patterns are automatically extracted by parsing the source and target sides of the training parallel corpus. Their approach show a statistically-significant improvement over a phrase-based monotone decoder for French-English.

Collins et al. (2005) describe a system that is different from Xia and McCord (2004) in a few respects. They used only a handful of linguistically motivated transfer rules, rather than over 56,000 automatically learned context-free rules. They consider German-English which is more challenging than French-English and they were still able to show a significant improvement of the BLEU score over the normal phrase-based decoder with the usual reordering capabilities. However, the fact that these rules cannot be extracted automatically is a major drawback.

Rewriting the input or output sentence, whether using syntactic rules or heuristics, makes difficult decisions that can not be undone by the decoder. For this reason, reordering is often handled better during the search and as part of the optimisation

function.

However, a recent two-stage model proposed by Dyer and Resnik (2010) shows a great deal of promise. In this model the reordering step retains a large number of possible word order variations because it is encoded as a context-free forest. This lets the context free part handle mid-to-long range reordering, and lets the finite-state transducer handle local phrasal correspondences. Unlike SCFGs and phrase-based models, this model does not impose any distortion limits. Initial results are promising, but its ability to compete with state-of-the-art models is yet to be shown.

In this section we have examined research dealing with the challenges of reordering. We have looked at restrictions on the search, reordering models, and synchronous grammar models. We have seen that there is still no model which is able to perform long distance reorderings in a principled fashion, and that has lead to work which separates the ordering problem from the translation problem. We argue that part of the reason for lack of progress in modelling reordering is that most research is evaluated on the the BLEU score. This score is certainly useful for certain purposes, but we show that it is not a reliable metric of word order quality. In the next section we survey translation metrics which are currently in use, and look at some of their limitations.

2.3 Metrics for Machine Translation

Automatic metrics for evaluating the quality of machine translation are essential for researchers and developers working in the field. Automatic metrics produce scores for translations quickly and inexpensively, which means that they can be used to evaluate large amounts of data with minimal human effort. This makes them an essential tool for large-scale development of translation systems. One of their principal functions is allowing researchers to asses the impact of modifications to their systems, but they also play an important role in training the parameters of translation systems. Here, development data must be repeatedly translated and evaluated to assess the quality of the parameter settings.

Automatic metrics measure the similarity of system output with one or several gold standards. They produce a numeric score which is necessarily a simplification of the genuine differences that exist between references and translations. Automatic metrics cannot be considered to be a replacement for human judgement. In fact their usefulness can only be decided upon through correlation with human evaluations.

There is currently a great deal of interest in developing metrics, in part spurred

on by recent evaluation campaigns. The Workshop on Statistical Machine Translation (Callison-Burch et al., 2007, 2008, 2009, 2010) and the NIST Metrics for Machine Translation 2008 Evaluation (Przybocki et al., 2009) have collect human judgements of translations from different systems, and used them to evaluate a wide spectrum of translation metrics. With the proliferation of metrics it is not easy to know which one to use. Unfortunately the evaluation campaigns have not resulted in a consensus over which is the best metric, as there are many experimental conditions and a variety of metrics perform well under different conditions. In the rest of this section we describe different approaches to human and automatic evaluation of translation.

2.3.1 Human Evaluation

Automatic evaluation depends upon human evaluation, but even this is very difficult. Although there has been 60 years of research into machine translation, there is still no generally agreed upon methodology for humans to evaluate translations (Hutchins and Somers, 1992; Przybocki et al., 2009). The most obvious method of testing machine translation quality is by judging (a) its accuracy, or the amount by which the sentence contains the same information as the reference, and (b) its fluency, or the degree to which the sentence is easy to read and grammatical. These are somewhat orthogonal, as a sentence can be easy to read but distort the original message, and equally the sentence can be correct, but contain many disfluencies. They also overlap somewhat, as there is a point at which the sentence is so disfluent that it is no longer intelligible.

Until recently, this was the most widely adopted basis for evaluating machine translation. Humans were asked to assign values from two five-point scales representing fluency and adequacy. These scales were developed for the annual NIST Machine Translation Evaluation Workshop by the Linguistics Data Consortium (LDC, 2005). There are a number of problems with this approach, however, apart from the difficulty in separating quality into two scores. A more serious problem with this approach is the difficulty of assigning consistent scores across a number of different sentences. It appears that humans have been using these scores as a way of indicating preference of one translation over another. In other words, they use the scales as relative rather than absolute (Callison-Burch et al., 2007).

Due to these concerns with accuracy and fluency ratings, another evaluation task was proposed. Rather than having to assign each translation a value along an arbitrary scale, people simply compare different translations of a single sentence and rank

them. This type of human evaluation has been performed in the last four workshops on statistical machine translation.

Although it is useful to have a score or a rank for a particular sentence, especially for evaluating automatic metrics, these ratings are necessarily a simplification of the real differences between translations. Translations can contain a large number of different types of errors of varying severity. A simple approach to quantifying the differences is to count the number of edits a person has to make to correct the sentence. This is an extrinsic measure of sentence quality because it measures the effort needed to post edit machine translation output to make it acceptable. Translation for human post editing is one of the major applications of machine translation in industry (Allen, 2003; Simard et al., 2007).

Another approach is to categorise errors by different types of linguistic phenomenon, and by relative difficulty in fixing them. This is the approach taken by Vilar et al. (2006). This kind of fine grained evaluation would be particularly useful for system developers who need to guarantee a certain level of quality to end users of translation.

Human evaluation is essential for developers to determine how reliable their systems are. Its is also essential for determining the value of different automatic translation metrics. To our knowledge, so far there has been no human evaluation method which has been specifically designed and tested for measuring the quality of word order in translations.

2.3.2 Automatic metrics

The advent of the BLEU score (Papineni et al., 2002) had an enormous impact on the field of statistical machine translation. It was the first automated metric to demonstrate correlation with human judgements of quality. As such, BLEU provided a means for large scale evaluation and it quickly became the de facto standard metric for machine translation. Since BLEU was proposed, a number of other metrics have shown to correlate well with human judgements. We describe the most commonly used metrics below.

2.3.2.1 BLEU

The Bilingual Evaluation Understudy (BLEU) score is the de facto standard in machine translation evaluation. It measures how well a machine translation overlaps with multiple human translations using n-gram co-occurrence statistics. N-gram precision p_n is

computed for each n-gram length by summing over the matches for every hypothesis sentence S in the corpus C as follows:

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} \text{Count}_{clipped}}{\sum_{S \in C} \sum_{ngram \in S} \text{Count}}$$

Where $\text{Count}_{clipped}$ is the maximum number of n-grams co-occurring in a candidate translation and a reference translation, and Count is the number of n-grams in the candidate translation.

The BLEU score is measure of precision, and because recall is important but difficult to formulate over many references, a brevity penalty is used. This penalises translations which are too short. The brevity penalty is calculated as:

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

where c is the length of the corpus of hypothesis translations, and r is the reference corpus length. In the case of multiple references, the reference corpus length is most commonly set to the length of the reference corpus which is closest to the hypothesis corpus. However, some researchers use the length of the shortest reference corpus and a further alternative is to use the average length of the reference sentences.

Thus, the BLEU score is calculated as:

$$\text{Bleu} = BP * exp(\sum_{n=1}^{4} w_n log p_n)$$

In the standard application of the BLEU score $n = 4$ and the weights w_n are set to the uniform $\frac{1}{n}$. Shorter n-grams reflect the lexical coverage of the translation and word order is indirectly evaluated by the higher order n-grams. A BLEU score can range from 0 to 1, and a score of 1 is assigned when a hypothesis exactly matches one of the references, or contains all the n-grams that occur in a hypothesis.

Reference:	parliament launches action plan to reduce its carbon footprint .
Translation:	to reduce the parliament it plans to start its carbon footprint .

Table 2.1: A reference human translations and a machine translation

We will describe how the BLEU score would be calculated for the hypothesis translation show in Table 2.1. Counting each space separated token, we find that the translation has 7 out of 12 unigram matches, 4 out of 11 bigram matches, 2 out of 10

trigram matches and 1 out of 9 4-gram matches. This means that the precisions for the different n-grams, are as follows: $p_1 = 0.58$, $p_2 = 0.36$, $p_3 = 0.20$, and $p_4 = 0.11$. The length of the hypothesis is 12, and the reference translation is 10. This test case does therefore not incur a brevity penalty. The overall BLEU score would therefore be: $1 * exp(log 0.58 + log 0.36 + log 0.20 + log 0.11) = 26.2$. Normally the score would be calculated over an entire document and not over just one sentence.

There are some well known problems with the BLEU score. Not only does this method of measuring word order differences depend on there being words which exactly match the words in the reference, but it also does not reflect the order that matching n-grams occur in, or the distance that they have moved. The final score is the geometric mean the of n-gram precisions and a brevity penalty. This makes the score unreliable at a sentence level: if there are no matching 4-grams the BLEU score is zero. The BLEU score is really only appropriate for calculating document level scores, or scores for a collection of sentences.

There is a variation of BLEU called smoothed BLEU (Lin and Och, 2004a) which can be used to calculate BLEU on a sentence level. The numerator and denominator of the n-gram precisions for $n = (2,3,4)$ is incremented by 1. The sentence level scores cannot easily be compared with document level BLEU scores, but we are guaranteed a positive BLEU score unless no words match in which case even smoothed BLEU will return zero.

The BLEU score is efficient to calculate, and it requires no additional annotation. These considerations, as well as comparison with previous benchmarks encourages the continued use of the BLEU metric. BLEU has been shown to correlate well with human judgements of translation quality in many instances (Przybocki, 2004). However, BLEU has also been shown to systematically underestimate the quality of rule-based translation systems (Koehn and Monz, 2005) which are preferred by human judges. This is because BLEU does not address overall grammatical coherence, it is only operates at a local level. This might favour statistical systems which are good at producing n-grams, but bias it against rule-based systems which address global sentence structure.

Other issues have been identified by Callison-Burch et al. (2006a), and they argue that an improvement in BLEU score is neither necessary nor sufcient for achieving an actual improvement in translation quality. They point out that BLEU admits a huge amount of variation for identically scored hypotheses. Typically there are millions of permutations of a translation which can receive the same BLEU score, but all of

these orderings are clearly not equally good. This aspect of the BLEU score makes it particularly inappropriate for measuring word order performance.

2.3.2.2 METEOR

The Metric for Evaluation of Translation with Explicit ORdering (METEOR) (Banerjee and Lavie, 2005; Lavie and Denkowski, 2009) attempts to address some of the deficiencies of the BLEU score. METEOR does not require exact matching of words between the reference and the translation. It allows variability in word choice by matching stems and synonyms. It also includes a measure of recall, which is an improvement over the precision based approach of BLEU.

METEOR evaluates a translation by quantifying the number of words in the translation that are matched to a given reference translation. If more than one reference translation is available, the translation is scored against each reference independently, and the best scoring pair is selected.

Given a pair of strings to be compared, METEOR generates an alignment such that every word in each string maps to at most one word in the other string. The metric first matches all identical words, then all unmatched synonyms using WordNet, and then all unmatched identical stems. The alignment with the greatest number of matched items is selected, and if there is a tie it chooses the alignment with the least number of crossings. Based on the number of aligned unigrams found between the two strings (m), the number of unigrams in the translation (t) and the number of unigrams in the reference (r), precision $P = \frac{m}{t}$ and recall $R = \frac{m}{r}$ are calculated and combined in an harmonic mean:

$$F_{mean} = \frac{PR}{\alpha P + (1-\alpha)R} \tag{2.8}$$

It therefore has a reasonably sophisticated approach to detecting word correspondences. However, the way it handles word order differences is quite simplistic. It generates an ordering penalty for a hypothesis based solely on the number of chunks (ch) the translation needs to be broken into in order to align to the reference:

$$\text{Penalty} = \gamma(\frac{ch}{m})^{\theta} \tag{2.9}$$

The final score is as follows:

$$\text{METEOR} = (1 - \text{Penalty})F_{mean} \tag{2.10}$$

The lowest METEOR score of 0, would be assigned if there were no matching words found even after applying the modules such as stemming, and METEOR returns a score of 1 when there is a perfect match between the reference and the hypothesis.

We now return to the example which was used to describe how the BLEU score was calculated, shown in Table 2.1. The number of exact word matches between the reference and the translation is 7. The number of stemmed matches is 1, "plan", and there are no synonym matches. 8 words match for sentences where the translation is length 12 and the reference length 10. Precision is therefore $8/12 = 0.666$ and recall $8/10 = 0.8$. The basic parameter setting for α is 0.8, γ is 0.4 and θ is 2.5. F_{mean} is therefore equal to $\frac{0.666*0.8}{0.8*0.666+0.2*0.8} = 0.769$. The translation is broken into 4 chunks and so the fragmentation penalty is $0.4(\frac{4}{8})^{2.5} = 0.071$. The final METEOR score is therefore 0.715.

The recent workshops on machine translation show that METEOR correlates fairly well with human judgement when translating into English (Callison-Burch et al., 2010). One of the major problems with the METEOR score is that the search procedure is heuristic and likely to be error prone, especially as it relies upon stemming and synonym functions. Another problem with METEOR is that the handling of word order differences by counting chunks does not take into account the number of words affected by a reordering.

2.3.2.3 TER

The Translation Error Rate (TER) (Snover et al., 2006) score is an improvement of one of the original machine translation metrics, the Word Error Rate (WER) (Och et al., 1999). The WER was borrowed from speech recognition where it measures the number of insertions, deletions and substitutions required to transform the output sentence into the reference. Unfortunately WER is not as appropriate for evaluating machine translation, as it does not take reordering into account. This problem motivates the use of the Position-independent word Error Rate (PER) which does not penalise reorderings. This is also suboptimal, however, because word order differences should not be completely ignored. TER addresses these problems by allowing block movement of words within the hypothesis as a low cost edit, a cost of 1, the same as the cost for inserting, deleting or substituting a word.

When considering multiple references, the reference with which the hypothesis has the fewest number of edits is deemed the closest, and the number of edits is calculated relative to this reference. TER performs a greedy search as finding the optimal align-

ment is NP-hard. The score is calculated as follows:

$$\text{TER} = \frac{\text{Number Edits}}{\text{Ave Number Reference Words}} \tag{2.11}$$

TER will give a score of 0 for a hypothesis which is identical to the reference. When translations are very different to the hypothesis, the number of edits required to transform the hypothesis can be larger than the average number of reference words and thus TER can be greater than one. In practice TER values over one are uncommon, and usually only occur when there is a great difference in length between the reference and the hypothesis. In this case the number of inserts or deletes required can be greater than the number of words in the reference.

Looking again at the example in Table 2.1, TER calculates that there are two insertions, three substutions and one shift. This makes a total of six edits and because the reference is of length ten, the TER score is 0.6.

A major drawback of TER is that the block "shift" operation captures word order differences without taking the size of the block or the distance it has shifted into account. Another disadvantage of TER is that words are required to match exactly. TER-Plus (TERp) (Snover et al., 2008, 2009) addresses this problem by allowing for stem, synonym, and paraphrase substitutions. This flexibility hugely increases the search space, and in all likelihood, increases errors in aligning the translation and the reference. Even with these problems, both TER and TERp have shown good correlation with human judgements in recent evaluation campaigns, especially when translating out of English.

2.3.2.4 Other

Other metrics which demonstrate good correlation with human ratings combine simple and complex features such as semantic and dependency overlap. ULC (Giménez and Màrquez, 2008) is an arithmetic mean over other automatic metrics including METEOR, Rouge, measures of overlap between constituent parses, dependency parses, semantic roles, and discourse representations. Rich Textual Entailment (RTE) (Padó et al., 2009b) is a regression model over a features adapted from textual entailment systems. Another metric, RTE measure how closely syntactic and semantic structures are matched between references and translations. These more complex metrics are interesting, but are slow to run, language dependent and difficult to train. RTE took more than five days to run in the Metrics MATR workshop (Przybocki et al., 2009). Addi-

tionally, errors are introduced because of the need for multiple layers of processing. In this thesis we focus on shallow metrics which are more widely useful.

2.3.3 Evaluation of Automatic Metrics

The method for evaluating automatic metrics varies depending on whether we are correlating them with sentence (or segment) level human scores or if we wish to collate sentence level judgements into a document level score (typically ten or so sentences) or a system level score (the entire test set).

2.3.3.1 Sentence Level

The most widely used method to compute correlation between two metrics at the sentence level is the *Pearson correlation coefficient*. It is used as a measure of the strength of linear dependence between two sets of data points (x_i, y_i). The Pearsons correlation coefficient, or r_{xy} is equal to:

$$r_{xy} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y} \tag{2.12}$$

where $\overline{x}$, $\overline{y}$ are the sample means and s_x, s_y are the sample variances of the variables x and y.

The correlation coefficient ranges from -1 to 1. Values that are close to 1 or -1, mean that the linear relationship between x and y is very strong, whereas a value close to 0 implies that there is no linear correlation between the variables.

Thus an automatic evaluation metric with a higher absolute value for Pearsons correlation coefficient is making predictions that are more similar to the human judgements. The statistical significance of the correlation is calculated by using an asymptotic t-test approximation.

2.3.3.2 System Level

We typically measure the correlation of the automatic metrics with the human judgements of translation quality at the system level using Spearman's rank correlation coefficient ρ. When there are no ties ρ can be calculated using the equation:

$$\rho = 1 - \frac{6d_\rho}{n(n^2-1)}$$

where d_ρ is the difference between the rank for system i and n is the number of systems. ρ also ranges between -1 and 1.

2.3.4 Discussion

In this thesis we rely upon three metrics, the BLEU score, METEOR, and TER, as baseline metrics. These metrics have all performed well in the evaluation campaigns and they are widely used. They are all shallow metrics as no deep linguistic analysis is required. This is important as it makes them reasonably language independent and faster to compute, allowing them to be more widely useful and appropriate for training systems. METEOR does leverage stems and synonyms, but these modules are optional, and for languages where they are not available, exact match is used. We also choose them because they are representative of different types of automatic metrics. Przybocki et al. (2009) suggest that metrics can be placed in three different categories: n-gram metrics, edit distance metrics, and linguistic metrics. They state that BLEU, TER and METEOR are the representative examples of these three respective categories.

Although these metrics are widely used, we argue in this thesis that they are not appropriate for measuring the word order performance of translation systems. None of them take the size of the word order differences into account and none of them have been directly evaluated on a reordering task. Considering the fact that a large amount of the research in translation is dedicated to improving the quality of the word order, this is a surprising gap.

2.4 Summary

In this chapter we have briefly described the statistical machine translation task. We have looked in detail at the reordering component of different translation models and highlighted reordering models, reordering restrictions and reordering within syntax-based systems. We then investigated various strategies of human and machine evaluation of translations introducing the three baseline metrics we will use throughout the thesis BLEU, METEOR and TER. Finally we have looked at how to evaluate the automatic metrics. In the next chapter we examine how we can extract the reordering characteristics of parallel corpora, and we use this to compare human, phrase-based and syntax-based machine translation systems.

Chapter 3

Comparison of Reordering in Translation Models

3.1 Introduction

In the previous chapter, we presented a broad overview of research on reordering. This chapter provides an in-depth analysis of the reordering capabilities of two important translation models.

Phrase-based models (Koehn et al., 2003; Och and Ney, 2004) have been a major paradigm in statistical machine translation over the last seven years, showing state-of-the-art performance for many language pairs. They search all possible reorderings within a restricted window, and their output is guided by the language model and a lexicalised reordering model, both of which are local in scope. However, the lack of structure in phrase-based models makes it very difficult to model long distance movement of words between languages.

Synchronous context-free grammars can represent long-distance reordering without the exponential complexity which phrase-based models face. However, added modelling power comes with challenges such as the size of the grammar and spurious ambiguity. Some grammar-based models such as the hierarchical model (Chiang, 2005) and the syntactified target language phrases model (Marcu et al., 2006) have been preforming well in recent evaluation campaigns (NIST, 2009).

Exploring translation models with the aim of improve reordering performance has been the focus of much research in statistical machine translation. However, our understanding of the variation in reordering performance between phrase-based and synchronous grammar models has largely been limited to relative BLEU scores. Relying

on BLEU scores is problematic, as a very large number of orderings are give the same score (Callison-Burch et al., 2006a). There has been little direct research on empirically evaluating the reordering behaviour of different translation models.

This chapter proceeds as follows. Section 3.2 presents a novel method for characterising the word order differences found in parallel corpora. Reorderings are analysed quantitatively, by recording their frequency and their size. In Section 3.3 we describe the experimental setup, including the creation of a test set with known reordering properties. Section 3.4 presents the results of experiments where the performance of the two translation models are compared to each other, with the hierarchical model performing slightly better for language pairs with large amounts of reordering. However, both models are shown to produce largely monotone translations, failing to capture the reorderings seen in human translated corpora. Finally, in Section 3.5 we summarise our contributions and our findings.

3.2 Extracting Reorderings

Until now, we have used the term reordering quite loosely to mean a word order difference between the source and target language. In this section we will define reordering and describe the algorithm we use to extract reorderings.

Differences in word order can include ambiguous cases which are not in fact reorderings. Perhaps the essential characteristic of a reordering is that the order of two words, or sequences of words, must be swapped between the two languages. We argue that it is intuitive to define a reordering as the inversion of the relative ordering of two words between source and target languages. Figure 3.1 shows a reordering where the order of the source words “bruja verde” and the target words “green witch” have been swapped.

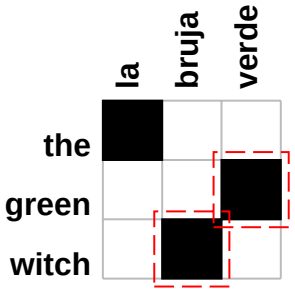
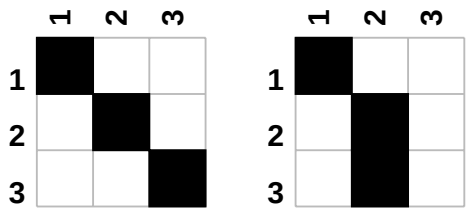


Figure 3.1: An example sentence with a reordering.



(a) (b)

Figure 3.2: Example sentences with no reordering.

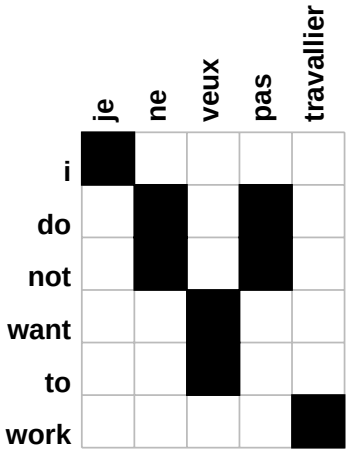


Figure 3.3: Example sentence with a discontinuous span and no reordering.

To support this intuition on the nature of reordering, it is helpful to consider cases where no reordering occurs. The most obvious case where no reorderings occur is when the sentence is translated in a monotone fashion. Figure 3.2 (a) shows us a simple example. Differences in the fertility of words also do not constitute reorderings, such as when a word is translated as two words or it is missed out in a translation, as shown in Figure 3.2 (b). Furthermore, in our opinion, no reordering exists for a more ambiguous case, when word is translated as more than one word, with a gap in between, i.e. a discontinuous alignment. An example of such a case is shown in Figure 3.3. Here, a word order difference exists, but there is no inversion in order of aligned words between the source and target.

Our definition of reordering follows this intuition, positioning it at the point where the difference in order is detected. We define reordering as a binary process occurring between two sequences of words that are adjacent in the source and are swapped in order in the target This definition agrees with the ITG constraint described in Sec-

tion 2.2.2. Under most conditions, our approach would essentially extract the spans defined by the inverse rules in an ITG grammar. Galley and Manning (2008) defined a method for extracting ITG reordering rules from word alignments using a shift-reduce parser which is quite similar to the method we use to extract reorderings. The main difference between these methods is in the manner in which they deal with reorderings which cannot be broken down into two inverted blocks, or *binarized*.

Our approach also has some similarities with the TER metric (Snover et al., 2006) which attempts to find the minimum number of edits to correct a hypothesis, and admits moving blocks of words. However TER relies upon a sequence of edits which transform the hypothesis at each application. Our method extracts a hierarchy of embedded reorderings from a fixed sentence pair without the confounding effect of the insert and delete actions of TER.

Wu (1997) discusses word order differences which cannot be modelled by two inverted blocks, of the kind we have defined. For sequences of length four, there are two out of a possible 24 orderings which are not binarizable, but for sequences of length 16, there are 20×10^{12} possible orderings of which only a very few are binarizable. Figure 3.4 shows the two interleaved reorderings of four words which broken down into two inverted blocks. Wu argues that these orderings are rare, however others have found these cases to be reasonably common (Zens and Ney, 2003).

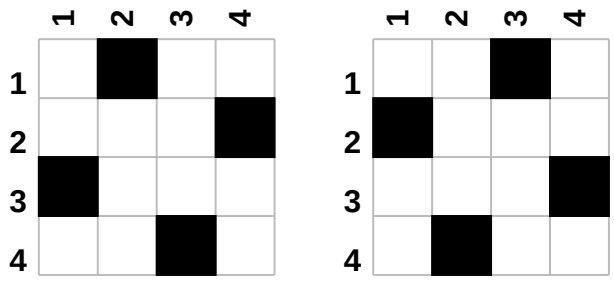


Figure 3.4: Example sentences with a reordering involving four interleaved words.

In our approach, we explicitly deal with the non-binarizable cases by extracting reorderings whenever adjacent source words are aligned to words in the target which are inverted in order. Our algorithm also differs from that of Galley and Manning by handling null alignments and discontinuous alignments, which occur frequently in human translated parallel texts.

The advantage of simplifying reorderings to binary inversions, is that any statistics thus collected can be used in a wide variety of translation models. Most models include

a concept of monotone or inverted orderings. The phrase-based models use lexicalised reordering models where the probabilities of monotone, inverted or disjoint orderings for a phrase pair are collected. Synchronous grammars perform binarization of rules in order to improve efficiency and the order of two non-terminals will either be monotone or inverted across languages. A recent study performed by Zhang et al. (2008) suggests that binary synchronous grammars are adequate for modelling translation.

3.2.1 Defining Concepts

For the purpose of extracting reorderings we must define exactly what a reordering is. We give here a strict definition and we use this for the experiments presented in this chapter and the following chapter, Chapter 4. In the rest of the thesis we will also use the term reordering to refer to the more general concept, where an inversion in word order has occurred, but it need not be an inversion between two blocks.

Before describing the extraction of reorderings, we need to define some concepts. We define a *block A* over an alignment grid as consisting of a source span, $A_{\bar{s}}$, which contains the positions from A_{smin} to A_{smax} and is aligned to a set of target words. The minimum and maximum positions (A_{tmin} and A_{tmax}) of the aligned target words mark the block's target span, $A_{\bar{t}}$. Figure 3.5 shows such a block.

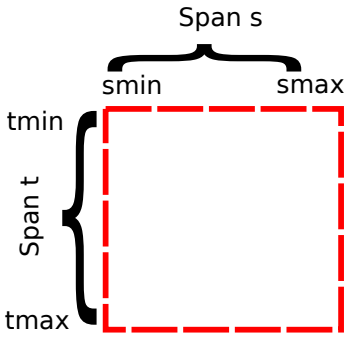


Figure 3.5: The dimensions of a block.

A reordering r consists of the two blocks rA and rB, which are adjacent in the source and where the relative order of the blocks in the source is reversed in the target. Figure 3.6 shows an example of a reordering with the two blocks. More formally:

$$rA_{\bar{s}} < rB_{\bar{s}}, \quad rA_{\bar{t}} > rB_{\bar{t}}, \quad rA_{smax} = rB_{smin} - 1$$

During the process of extracting reorderings, we rely upon the concept of consistency. A block is consistent if all the words which are inside the block are aligned to

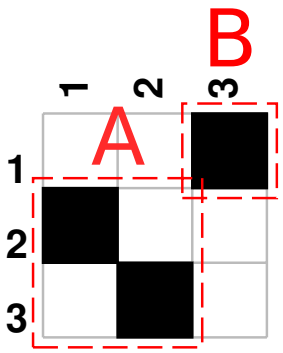


Figure 3.6: The definition of a reordering with two blocks A and B.

each other, and not to words outside the block. This concept is borrowed from work on phrase pair extraction from word alignments (Koehn et al., 2003). Figure 3.7 shows an example of an inconsistent block where target word two is aligned to a word which outside of the block.

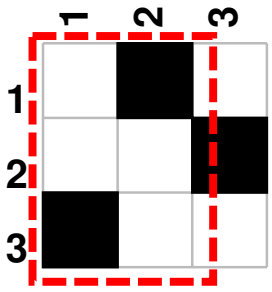


Figure 3.7: Example of an inconsistent block.

3.2.2 Extraction Algorithm

In this section we first describe the extraction algorithm which detects the existence of a reordering and determines the dimensions of the blocks involved. We then present some minor additions to the algorithm to handle null alignments and discontinuous alignments.

The algorithm proceeds as follows. We step through all the source words. We extract the positions of target words which are aligned to the current source word. We compare these target positions to the target words aligned to the previous source word. If they are inverted in order with respect to the source, a reordering has occurred.

The algorithm first sets the blocks to some initial positions and then grows them to their final dimensions. The initial position of block A is set to the previous source word and its aligned target words. Block B is set to the current source word and its aligned target words. Then we grow the blocks. When reorderings are embedded within each other the assumption is that they are right branching. This means that block A is grown to be as large as possible while block B is only grown the minimum necessary for the reordering as a whole to be consistent. This basic assumption is justified by the fact that English is considered to be a right branching language because the main verbs precede the direct objects. There are many languages which are left branching however, such as Japanese, and even English places prepositions and numerals before nouns. Ideally, the reordering would be constrained by a parse of the sentences as we do in other word (Birch et al., 2009).

Figure 3.8 shows the reorderings that are extracted from an alignment with embedded reorderings. Although (c) is not sanctioned by our algorithm, it could be the preferred reordering as determined by a parse tree. There is an extension of this work presented in Birch et al. (2009) which selects reorderings detected between child nodes in a parse tree. We do not pursue this method here as our experiments are quantitative and do not use the grammatical information that is thus extracted.

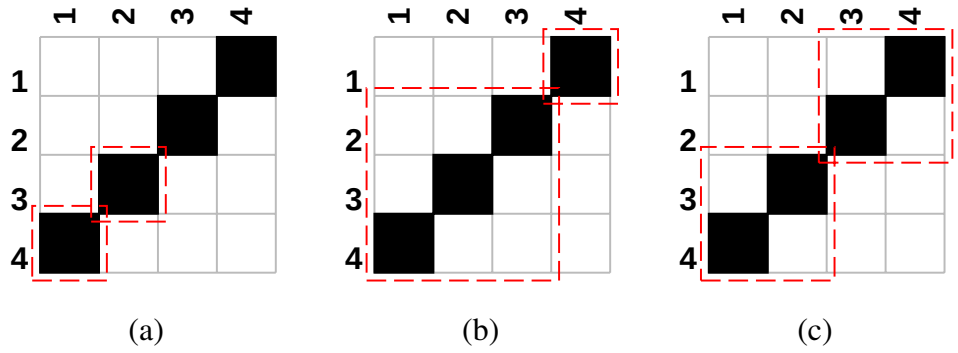


Figure 3.8: Examples of embedded reorderings. Examples (a) and (b) show two of the three reorderings supplied by our algorithm for this inverted alignment. Our algorithm would not produce the reorderings shown in (c).

From the initial dimensions of block A the algorithm attempts to grow block A from this point towards the source starting position. It extends the source span while the target span of A is greater than that of block B, and the new block A remains consistent. Finally, it extends block B towards the source end position, while the target span of B is less than that of A and the new reordering is inconsistent.

More formally, we define the algorithm as follows:

$J \leftarrow$ length of source sentence
for $s_{current} = 2$ to J **do**
  if $align(s_{current}) < align(s_{current} - 1)$ **then**
    B = getblock($s_{current}$,$s_{current}$)
    A = getblock($s_{current} - 1$,$s_{current} - 1$)
    $G = A$
    $x = s_{current} - 2$
    while $x >= 0$ **and** $G_{\bar{t}} > B_{\bar{t}}$ **do**
      G = getblock(x,$s_{current} - 1$)
      if G is consistent **then**
        $A = G$
      end if
      decrement x
    end while
    $G = B$
    $x = s_{current} + 1$
    while $x <= J$ **and** $A_{\bar{t}} > G_{\bar{t}}$ **and** reordering (A,G) is inconsistent **do**
      G = getblock($s_{current}$,x)
      if reordering (A,G) is consistent **then**
        $B = G$
      end if
      increment x
    end while
  end if
end for

See Figure 3.9 for an example of a sentence pair with two reorderings. The algorithm steps through the Chinese source words until it reaches the Chinese word for "from". It detects that the previous source word is aligned to a target word which precedes the current target word "from". A reordering is thus detected. The algorithm sets block A to "late" and block B to "from". It then continues to extend block A towards the start of the source sentence, while all the aligned target word positions are

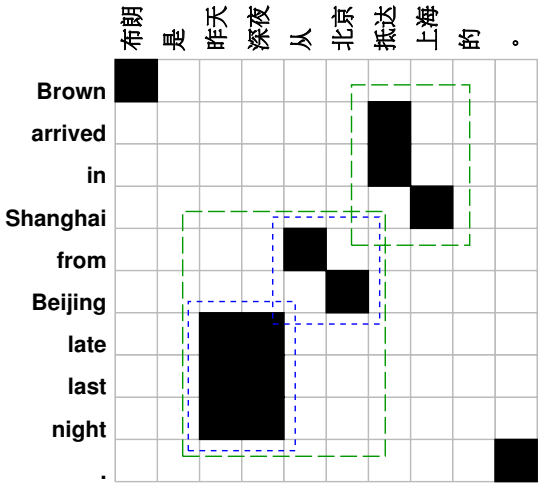


Figure 3.9: A sentence pair from the test corpus, with its alignment. Two reorderings are shown with two different dash styles.

greater than that of "from". It therefore stops when it reaches "Brown". The algorithm then tests whether the reordering is consistent and it discovers that the word "Beijing" is aligned to a word outside of the area of the reordering. It therefore grows block B towards the end of the source sentence. It stops once block B includes "Beijing". The next reordering is detected between "arrived in" and "Beijing", and the blocks are grown in a similar fashion. We can see that the algorithm attempts to grow A as large as possible, but it only grows B when the reordering is inconsistent. This algorithm has the worst case complexity of $O(\frac{n^2}{2})$ when the words in the target occur in the inverse order to the words in the source.

3.2.2.1 Null Alignments

In human translated text, null alignments will be relatively common. The way we deal with null alignments is to include them in the dimensions of the reordering if they occur between or inside the reordered blocks, but not if they occur on the outside of the blocks. This means that in the case of a word not being aligned to a target word, the next word is examined. Figure 3.10 provides three examples with null alignments which are included in the reordering blocks. Figure 3.9 includes two source words which are aligned to null and occur on the edge of reorderings and they are therefore not included.

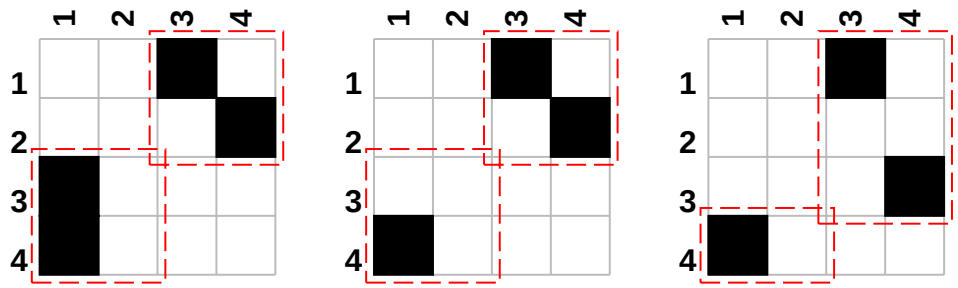


Figure 3.10: Examples of null alignments being included in reorderings.

3.2.2.2 Discontinuous Alignments

We have already discussed discontinuous alignments. We have argued that, although they represent a word order difference between the source and target, there is no inversion in word order, and therefore they are not considered to be reorderings. We do not want to extract discontinuous alignments as reorderings and we therefore handle them in the following manner. We identify the minimum consistent block that surrounds the discontinuous alignment, and we mark this area off. In Figure 3.11 we can see that the discontinuous alignment caused an area to blocked off in dark grey. No reorderings are extracted from within these blocks.

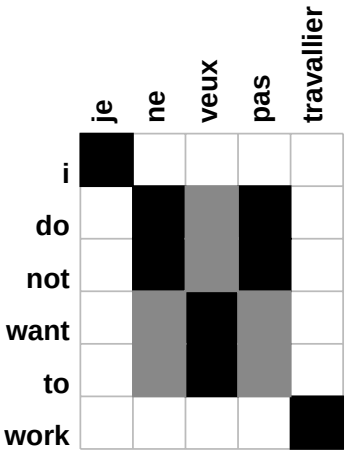


Figure 3.11: Example sentence with a discontinuous span and no reordering.

3.2.2.3 Non-binarizable Reorderings

Our approach to extracting reorderings relies upon the assumption that reorderings consist of two adjacent blocks which are inverted in order in the target. Figure 3.12 shows an example of a Chinese sentence with an interleaved reordering. It also shows the reorderings that our algorithm detects. There are two inversions and therefore two reorderings are extracted. The total source and target spans of these individual reorderings are very similar to the dimensions recorded if one interleaved reordering were to be extracted.

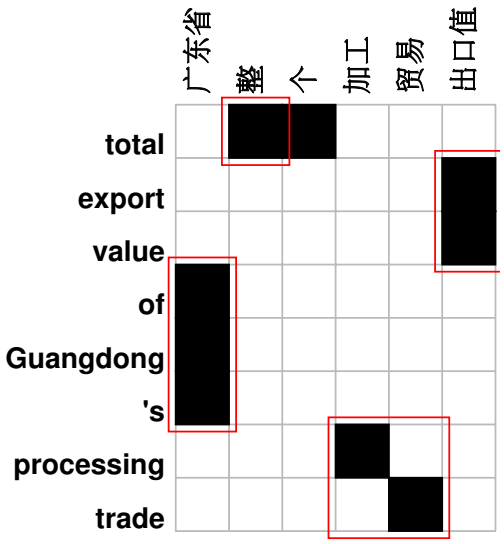


Figure 3.12: An example sentence where the reordering consists of 4 interleaving elements.

3.2.3 RQuantity

The reordering extraction technique allows us to analyse reorderings in corpora according to the distribution of reordering widths. In order to facilitate the comparison of different corpora, we combine statistics about individual reorderings into a sentence level metric which is then averaged over a corpus. This metric is defined using reordering widths over the target side, as the common language in the following experiments is the target language English.

We define RQuantity as follows:

$$RQuantity = \frac{\sum_{r \in R} |rA_{\bar{t}}| + |rB_{\bar{t}}|}{I}$$

where R is the set of reorderings for a sentence, I is the target sentence length, A and B are the two blocks involved in the reordering, and $|rA_{\bar{s}}|$ is the size or span of block A on the target side. RQuantity is thus the sum of the spans of all the reordering blocks on the target side, normalised by the length of the target sentence. The minimum RQuantity for a sentence would be 0. The maximum RQuantity occurs where the order of the sentence is completely inverted and the RQuantity is $\frac{\sum_{i=2}^{I} i}{I}$. The maximum RQuantity could potentially be greater than 1. For example, Figure 3.9 has an RQuantity of $\frac{3+2+5+3}{10} = 1.3$. RQuantity is not a true metric because it is not symmetric. Measuring the RQuantity from Chinese to English, could return different results than it would measuring from English to Chinese. The is partially because of the simplification assumptions needed to extract the reorderings, but also because we are only taking the length of sentence into account.

3.3 Experimental Design

We have presented our method for extracting reorderings from parallel corpora. We now apply this method to investigate what kind of reordering occurs in the output of two important state-of-the-art translation models. We aim to compare the reordering behaviour of the phrase-based model and the hierarchical model with each other and with the human translations. We also aim to determine whether the claim that the hierarchical model is better able to capture reordering is supported, and under what circumstances this is true.

3.3.1 GALE Data

Characterising the reordering present in different human generated parallel corpora is crucial to understanding the kinds of reordering we must model in our translations. In order to extract reorderings, word alignments are needed. The GALE project created an important and relevant resource which contains human annotated gold standard word alignments for a large number of Arabic-English (AR-EN) and Chinese-English (CH-EN) sentences[1]. A subset of these sentences come from the Arabic and Chinese treebanks, which provide gold standard parses of these sentences. Table 3.1 shows the number of sentences and the number of words in these corpora. In this chapter, we use the subset of the data with parsing information comprising of 3380 CH-EN and

[1] see LDC corpus LDC2006E93 version GALE-Y1Q4

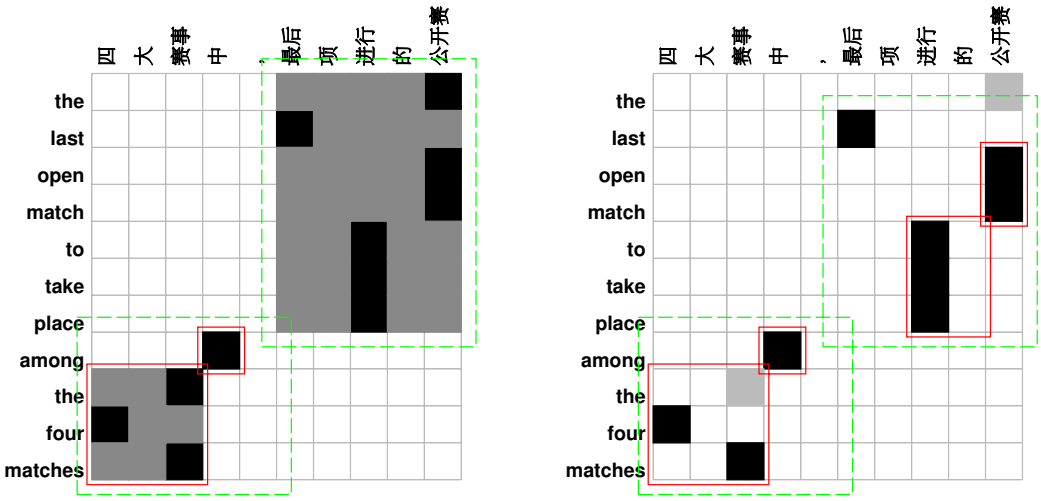
4337 AR-EN sentence pairs. The CH-EN corpus aligns English words with Chinese characters, and we apply the segmentation defined by the parse tree.

	Aligned	Aligned+Parsed
CH-EN		
Sentences	10,407	3,380
CH Words	236,634	84,408
EN Words	289,701	116,220
AR-EN		
Sentences	13,263	4,337
AR Words	277,744	140,091
EN Words	383,389	165,128

Table 3.1: GALE-Y1Q4 manually aligned corpus statistics.

Chinese does not contain determiners and the annotation guidelines for the GALE data indicate that determiners in English are aligned to the head of the noun phrase. This creates a large number of discontinuous word alignments which result in blocked off areas, from which no reorderings are extracted. For a significant proportion of the sentences, these blocked off areas cover large areas of the sentence. We solved this problem by unaligning determiners in a preprocessing step where we POS tagged the English side of the corpus using the Stanford POS tagger (Toutanova and Manning, 2000). The results of this preprocessing can be seen in Figure 3.13. In (a) we see the original sentence with discontinuous alignments and two large blocked off areas in dark grey. In (b) we see the unaligned determiners are shown in light grey. By unaligning the determiners, we reduce, and often remove, areas of the sentence which are blocked off. More reorderings within these previously blocked off areas are now available for extraction. In Figure 3.13 we can see that a new reordering has been identified between “open match” and “to take place”.

We apply the reordering extraction algorithm to these corpora. Figure 3.14 shows the distribution of reorderings in the CH-EN and AR-EN corpora broken down by the total width of the target span of the reorderings. This figure clearly shows how different the distributions of reorderings are in the two language pairs. AR-EN has far fewer reorderings over the medium and long distances, but surprisingly, it has many more short distance reorderings. We define *short*, *medium* or *long distance* reorderings to mean that they have a reordering of width of between 2 to 4 words, 5 to 8 and



(a) (b)

Figure 3.13: An example sentence with discontinuous alignments (a) before and (b) after determiners are unaligned. The resulting reorderings are also indicated.

more than 8 words respectively. These definitions are somewhat arbitrary, but relate to reordering performance of the current translation models. Most translation models handle short distance reorderings relatively well. They can also sometimes correctly perform a reordering over a medium distance, but almost all long distance reorderings fail. We analysed the reorderings seen in the Chinese-English to see how many of them were binarizable. For 2990 of the 3380 Chinese-English sentences (88.46%), all the reorderings comply with the ITG assumption. This is a high percentage of sentences which do not contain non-binarizable reorderings.

We also investigate what kind of RQuantity values are returned for the corpora. Figure 3.15 and Figure 3.16 shows the RQuantity for CH-EN and AR-EN for sentences of different lengths. The size of the standard deviation is indicated by vertical lines. The CH-EN corpus displays about three times the amount of reordering than the AR-EN corpus. For CH-EN, the RQuantity increases with sentence length and for AR-EN, it remains constant. This seems to indicate that for longer CH-EN sentences there are larger reorderings, but this is not the case for AR-EN. RQuantity is low for very short sentences.

Al-Onaizan and Papineni (2006) propose an alternative method for comparing reordering in different parallel corpora. They take the reference sentence and reorder it according to the word order shown by the word alignments. They then measure how scrambled the sentence is by computing the BLEU score between the original reference

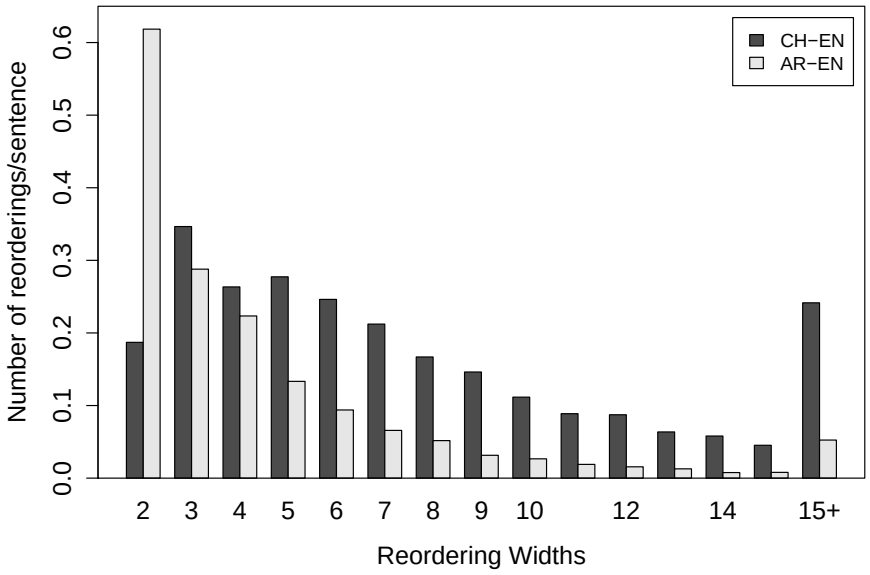


Figure 3.14: Comparison of reorderings of different widths for the CH-EN and AR-EN corpora.

sentence and the reordered reference. They show that Arabic-English is more monotone than Chinese-English because it reports a higher BLEU score. Their method is simple and provides some insight, unfortunately one BLEU score could account for a vast number of different possible orderings and is therefore not particularly informative.

3.3.2 Reordering Test Corpus

In order to determine what effect reordering has on translation, we extract a test corpus with specific reordering characteristics. We divide up the sentences into groups depending on the amount of reordering they display. By separating sentences with little or no reordering from sentences with a large amount of reordering, we can evaluate models on their treatment of sentences that we know contain large amounts of reordering.

To minimise the impact of sentence length, we select sentences with target sentence lengths of 20 to 39 words inclusive. In this range, the amount of reordering for different sentence lengths is relatively stable, as shown in Figures 3.15 and 3.15. We then split these sentences into four different sets. The first set consists of sentences

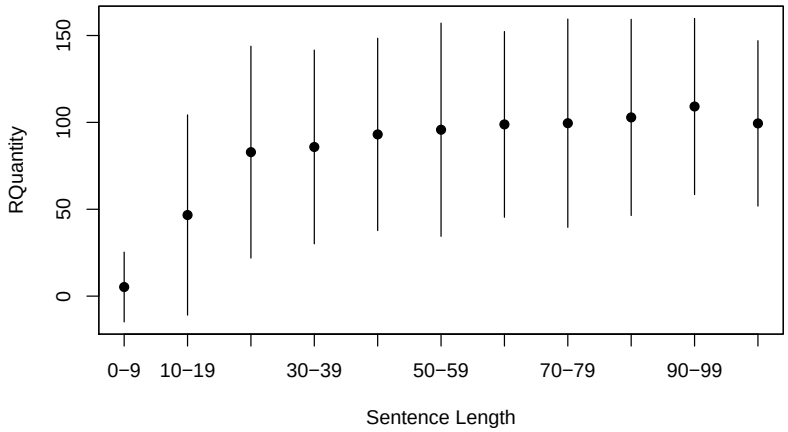


Figure 3.15: Average RQuantity, with standard deviation shown, for the CH-EN corpora for different English sentence lengths.

with no reordering. Some of these sentences have areas which are blocked off due to discontinuous alignments which can contain unextracted reorderings. On examination, the blocked off areas are small and most of these sentences do in fact have very little reordering.

We split the rest of the sentences into groups of equal size. They are divided into groups depending on their RQuantity and we end up with three sets of sentences: “low”, “medium” and “high”.

	None	Low	Medium	High
RQuantity				
CH-EN	0	0.39	0.82	1.51
AR-EN	0	0.10	0.25	0.57
Sentences				
CH-EN	105	367	367	367
AR-EN	293	379	379	379

Table 3.2: The RQuantity, and the number of sentences for each reordering test set.

Table 3.2 reports the RQuantity and the number of sentences for each of the four test sets. It is important to note that although we might name a set “low” or “high”, this is only relative to the other groups for the same language pair. The “high” AR-EN set, has a lower RQuantity than the “medium” CH-EN set.

Figure 3.17 shows distribution of the average number of reorderings per sentence

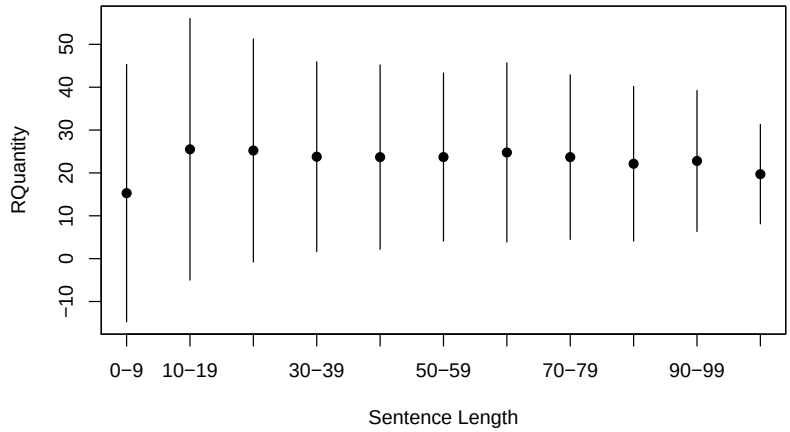


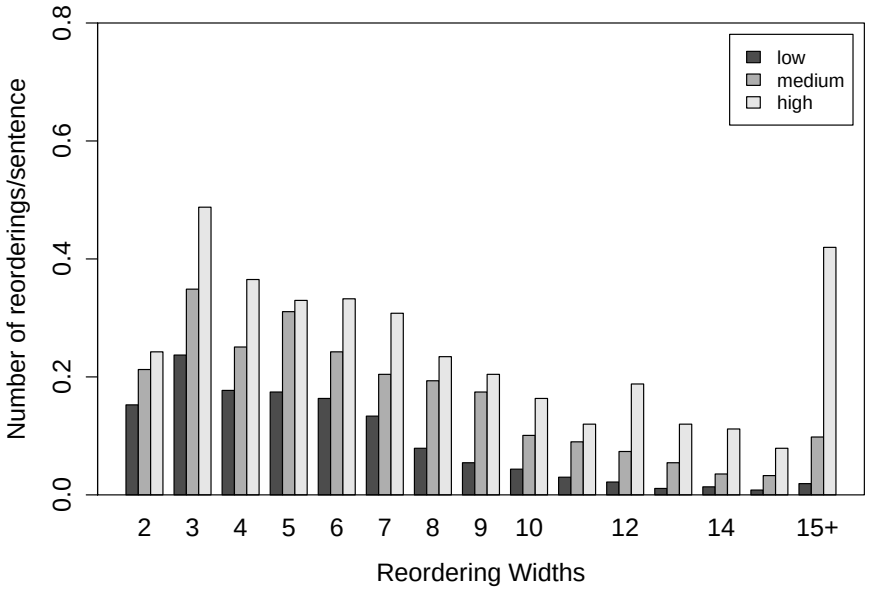
Figure 3.16: Average RQuantity, with standard deviation shown, for the AR-EN corpora for different English sentence lengths.

for each of the test sets, broken down by the total span of the reordering on the target side. As expected, we see more medium and long distance reorderings for Chinese to English than for Arabic to English. These graphs show that the reorderings in the higher RQuantity groups have more and longer reorderings.

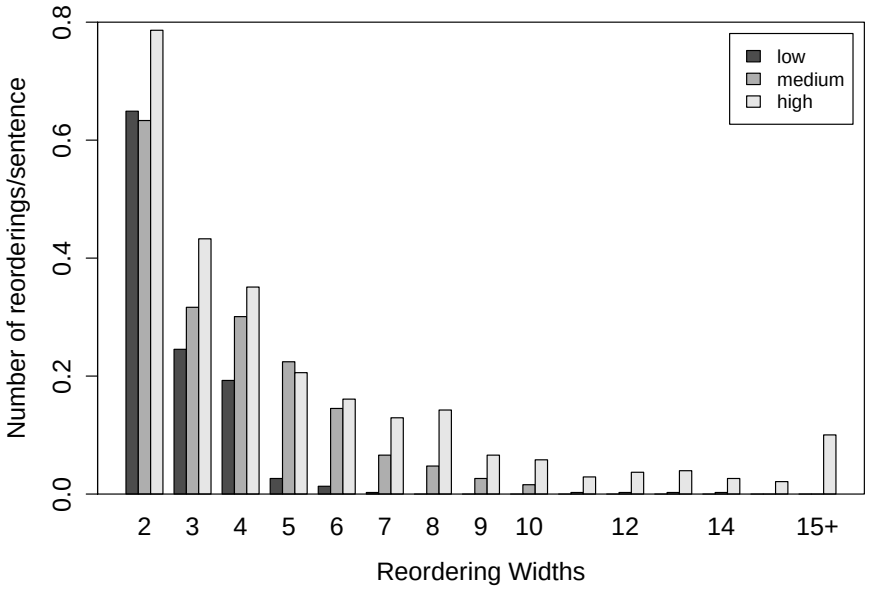
3.3.3 Translation Models

The following experiments use the MOSES implementation of the phrase-based model (Koehn et al., 2007), and the hierarchical model is an implementation of Hiero (Chiang, 2007) with all the default settings. For details please see Appendix A.

We trained both models on subsets of the NIST 2008 data sets, consisting mainly of news data. Table 3.3 reports the training corpora's characteristics: the number of sentences and the number of English and foreign words that they contain. We used a trigram SRILM language model, interpolated with kndiscount, on the entire English side (211M words) of the NIST 2008 Chinese-English training corpus. Although a higher order language model would have slightly increased translation performance across the board, it would not have changed the behaviour of the models with regard to the medium or longer distance reorderings, as even a stronger language model still operates at a local level. Minimum error rate training was performed on the 2002 NIST test for CH-EN, and the 2004 NIST test set for AR-EN.



(a)



(b)

Figure 3.17: Number of reorderings in the (a) CH-EN and (b) AR-EN test set plotted against the total width of the reorderings.

	CH-EN	AR-EN
Sentences	547K	1,069K
CH/AR words	10.2M	23.4M
English words	12.3M	26.9M

Table 3.3: NIST 2008 training data characteristics in thousands of sentences and millions of words.

3.3.4 Example Translation

Before analysing the results of the translation experiments, it is instructive to look at an example sentence pair.

Figure 3.18 shows the human annotated alignment of a Chinese sentence with its reference translation. Figures 3.19 and 3.20 show the translation model output of the phrase-based MOSES decoder, and the hierarchical HIERO decoder. The alignments are the actual alignments used by the translation models to construct the translation, and the reorderings extracted by our extraction algorithm are also shown. Reading the translations, it is clear that they are very poor. We can see that both models perform an essentially monotone translation of the source, even though the original human reference sentence contains a large number of word order differences relative to the source. Even if we take into account the fact that a certain amount of variation in word order is permissible in the translation, it seems perfectly clear that the lack of sensible reordering in the models is contributing to the poor quality of the output. The phrase-based model does perform some local reorderings, but these account more for lexical variation between the reference and the translation, than for differences in structure between the languages. The hierarchical decoder performs even fewer reorderings than the phrase-based model.

Figure 3.20 also shows some discontinuous alignments, which cause areas of the sentence to be blocked off. These represent rules in the grammar which have multiple terminal symbols, or words. These discontinuous alignments occur because the actual word alignments are not resolved. All words in the source side rule are aligned to all words in the target side rule. Although it would have been preferable to extract word alignments from these sentences, we can see here that the discontinuous alignments do not contain interesting reordering information. Even though we are potentially underestimating the number of reorderings seen in the output of the hierarchical model, the difference between the word order of the human translation and the Hiero translation

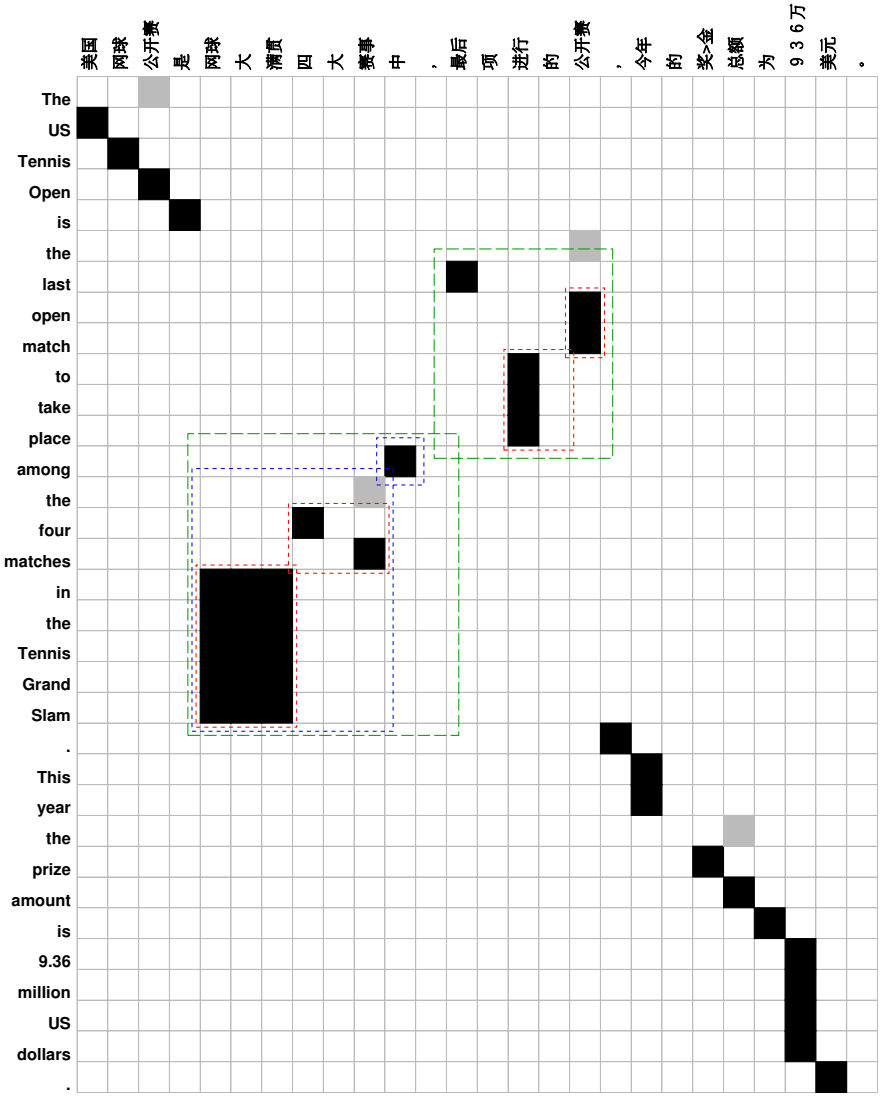


Figure 3.18: An example of a human translated Chinese-English sentence pair.

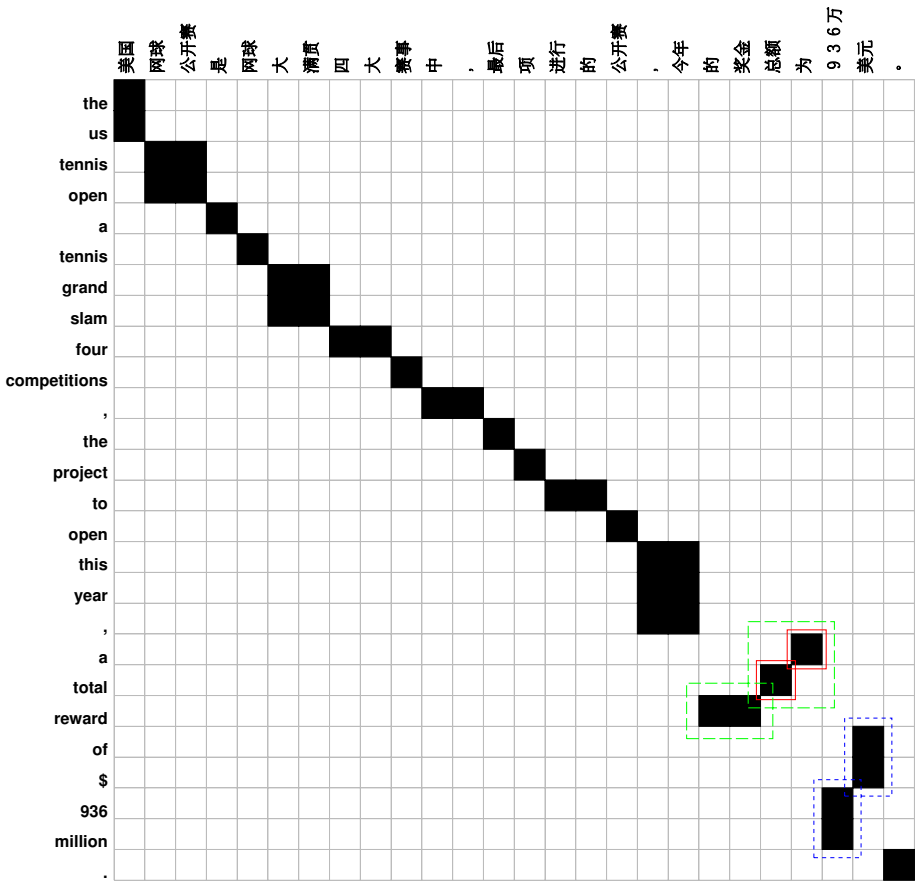


Figure 3.19: The phrase-based translation of the Chinese source in Figure 3.18.

is so marked, that we can confidently say that the model is under-performing.

This example highlights the fact that neither model comes close to modelling the reorderings seen in the human translated texts. Researchers motivate more powerful models, such as the hierarchical model, by claiming that they can model reordering better. However, these models are then evaluated using metrics which do not measure the word order directly.

3.3.5 Manual Analysis

In our experiments, we investigate whether or not reorderings which occur in the reference, also occur in the translations. The only way to verify the relevance of automatically detecting if reorderings are reproduced, is to use humans to evaluate how many reproduced and un-reproduced reorderings are correct.

We present human judges with the reference and the translation of 50 randomly

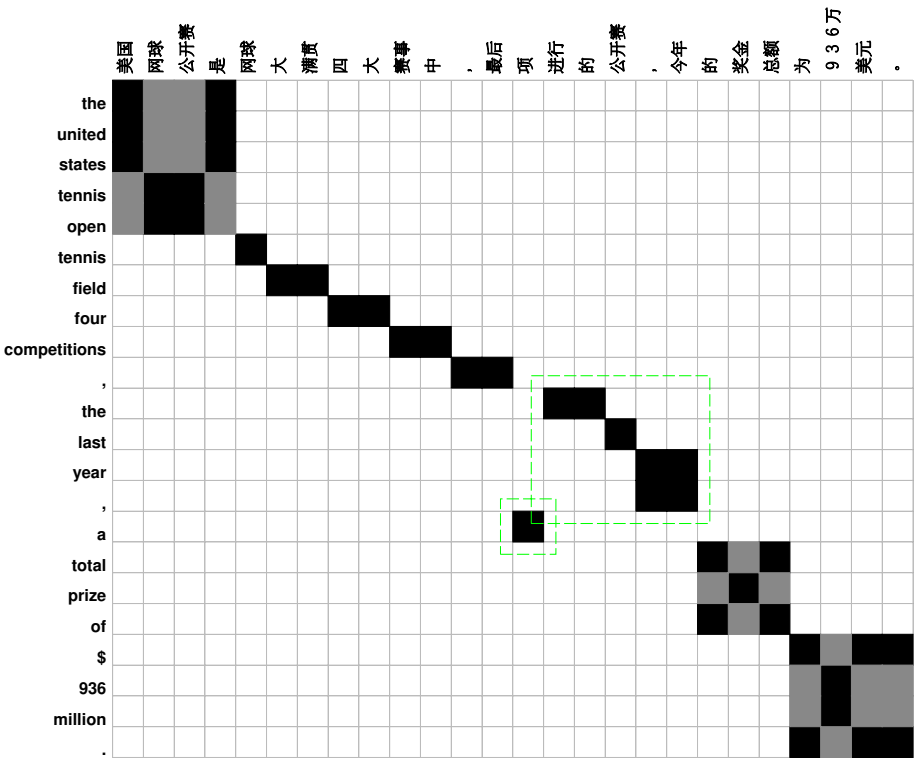


Figure 3.20: The hierarchical model translation of the Chinese source in Figure 3.18.

selected CH-EN sentences from the reordering test set. We mark the target ranges of the blocks that are involved in the particular reordering we are analysing, and ask the evaluator if the word order in the translation is "correct", "incorrect" or "not applicable". The judges were told to select the "not applicable" label when the translated words are so different from the reference, that their ordering is irrelevant.

There were three human evaluators who were approached personally. They were all fluent English speakers. They each judged 25 CH-EN reorderings which were reproduced and 25 CH-EN reorderings which were not reproduced. The 50 examples were presented to the evaluators in the same document which randomly permuted the reproduced and non-reproduced reorderings. In total 150 judgements were collected. No experimental software was used. Please see Appendix B for detailed instructions and an example test case.

3.4 Results

In the following experiments we explore ways of quantifying the differences between between the human translation and the machine translations, and between the phrase-based and hierarchical models.

3.4.1 Performance on Test Sets

The reordering test sets were created to see what effect reordering has on the performance of two translation systems. In this section we compare the translation output for the phrase-based and the hierarchical system using the standard machine translation metric, the BLEU score.

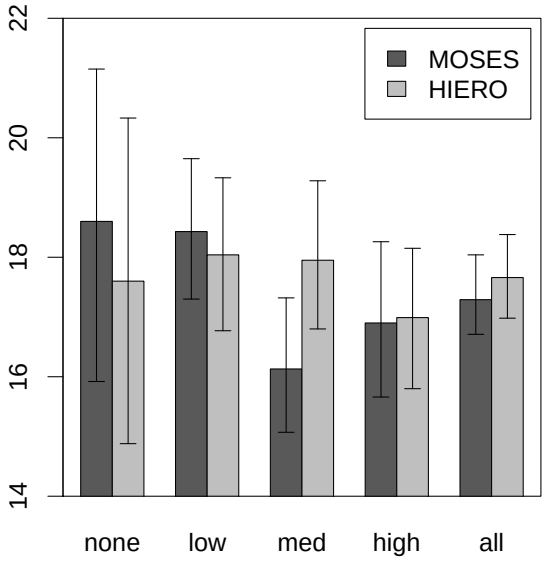


Figure 3.21: BLEU scores and 95% confidence intervals for the different CH-EN reordering test sets and the combination of all the groups for the two translation models.

Figure 3.21 and Figure 3.22 show the BLEU score results of the phrase-based model and the hierarchical model on the reordering test sets. The 95% confidence interval for the results is shown, and this was calculated using bootstrap resampling (Koehn, 2004b). We can see that the models display quite different behaviour for the test sets across the two language pairs.

The hierarchical model outperforms the phrase-based model when applied to the CH-EN language pair as a whole, but performs significantly worse on the AR-EN language pair. For the CH-EN test sets, the phrase-based model does a little better on

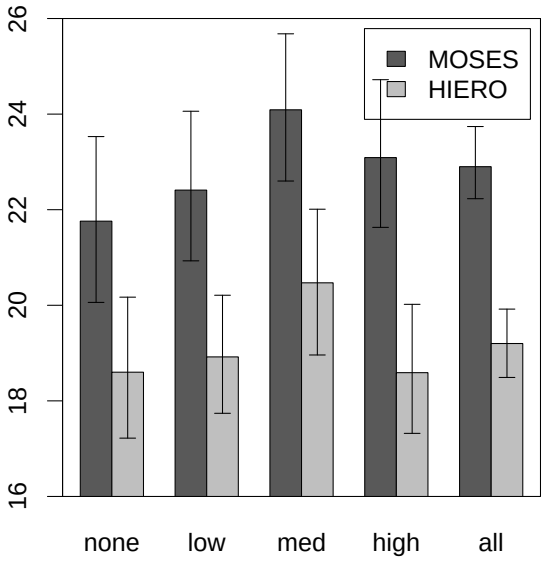


Figure 3.22: BLEU scores and 95% confidence intervals for the different AR-EN reordering test sets and the combination of all the groups for the two translation models.

the "none" and "low" test sets, but it performs worse on the "medium" test set. It seems that the phrase-based system is able to model the shorter distance reorderings, but the hierarchical model is able to model medium distance reorderings better. The fact that both model show equal performance on the "high" RQuantity test set suggests that the hierarchical model has no advantage over the phrase-based model when the reorderings are long enough and frequent enough. The performance of both systems on the "high" test set is surprisingly good, but this could also be due the fact that BLEU is unreliable at capturing reordering performance. This motivates analysing translations specifically for reordering as we do in the next section. In fact this thesis will propose a metric that takes into account both ordering and lexical variation, but making it easy to examine each component score in isolation.

For the AR-EN results (Figure 3.22), the phrase-based system has an advantage over the hierarchical system. This is because almost all the reorderings in the AR-EN test sets are reasonably short distance and the phrase-based system seems to handle these reorderings better than the hierarchical model. The phrase-based model considers all possible orderings within the distortion limit, whereas the hierarchical model requires evidence of a reordering occurring in the training corpus. These results indicate that the choice of translation model should be informed by the amount and type of reordering present in the language pair, and that more structured models are not

necessarily preferable.

Our results here compliments an empirical comparative study of MOSES and HIERO performed by Zollmann and Venugopal (2006). They tried to ascertain which is the stronger model under different reordering scenarios by varying distortion limits and the strength of language models. They show that the hierarchical models do slightly better for Chinese-English systems, but worse for Arabic-English. Although quite thorough, their study did not pick up the fact that even for Chinese-English, Moses performs better for sentences with low amounts of reordering and it performs as well for sentences with very large amounts of reordering.

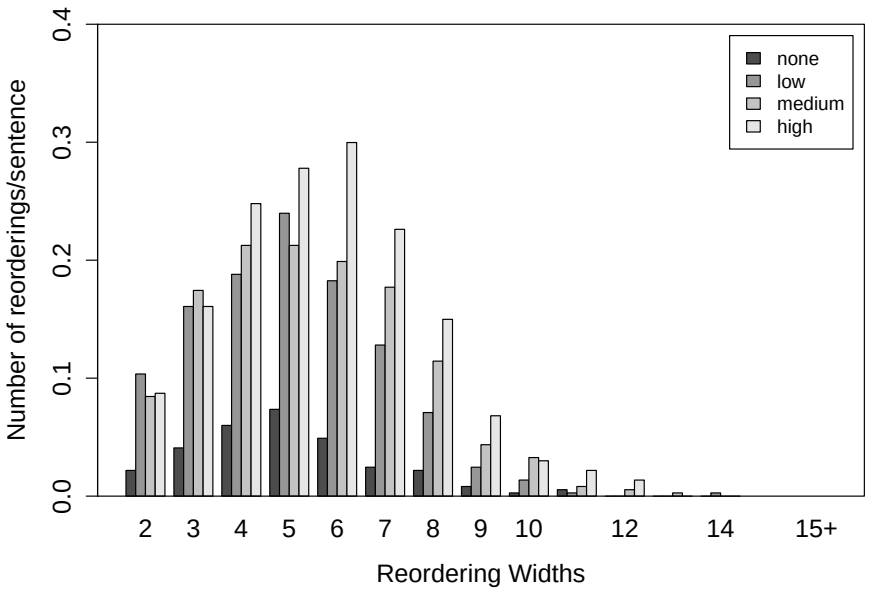
3.4.2 Reorderings in Translation

Reordering performance can only be partially revealed by the BLEU score, and so we perform a more detailed analysis. We use our extraction algorithm to extract the set of reorderings from the output of the translation models.

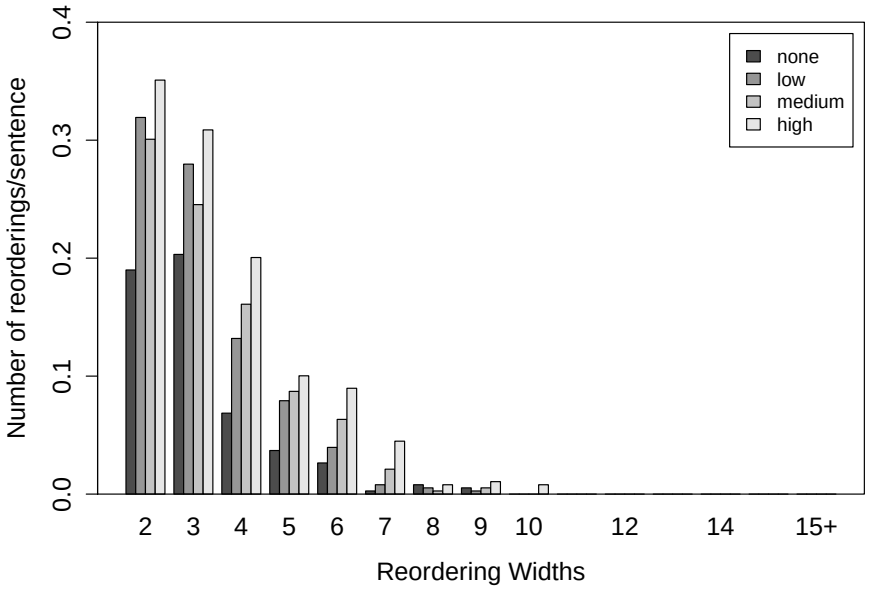
Figure 3.23 plots the frequency of the reorderings detected in the output of the phrase-based model, breaking down the analysis based on the total size of the reorderings on the target side. This graph is interesting when read in conjunction with Figure 3.17, which shows the reorderings that exist in the original reference sentence pairs.

The Moses translations have far fewer reorderings than the human reference translations. Those reorderings that do occur are predominantly short or medium length reorderings and almost no long distance reorderings occur.

Figure 3.24 shows the reorderings contained in the output of the hierarchical model. The results are very different to both the phrase-based model output (Figures 3.23) and to the original reference reordering distribution (Figures 3.17), there are many fewer reorderings. However, the BLEU score performance of this system is better than that of the phrase-based system for the “medium” test set. As we are missing some reorderings due to the discontinuous alignments contained within the hierarchical rules, the numbers here represent a lower-bound on the number of reorderings. Even so, it is clear that the hierarchical model has failed to capture the reordering behaviour of the human translated corpus.

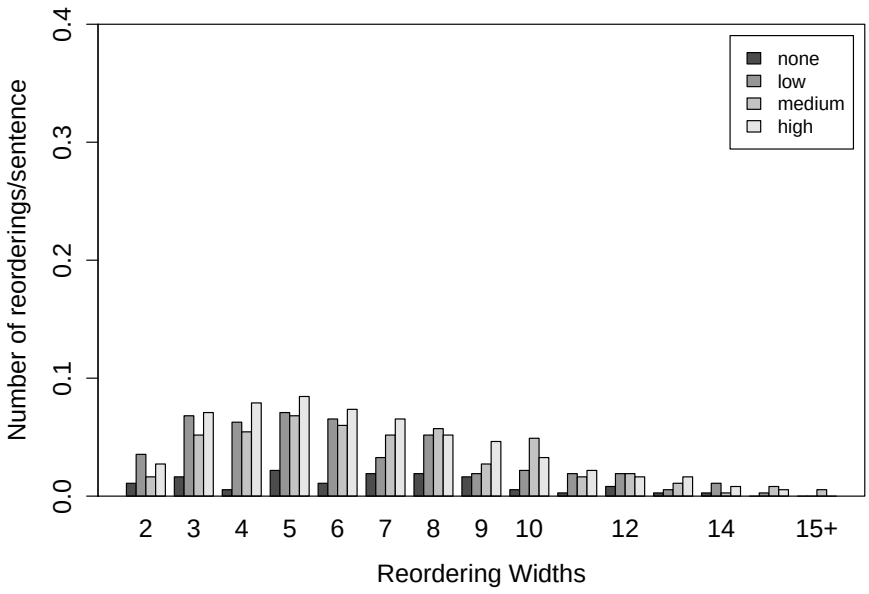


CH-EN

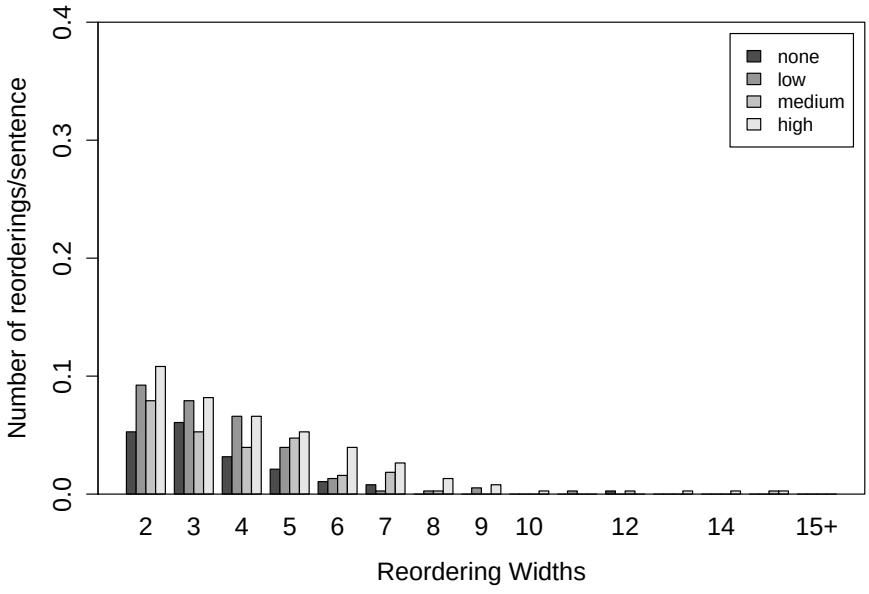


AR-EN

Figure 3.23: Reorderings in the MOSES translations, plotted against the total target width of the reorderings.



CH-EN



AR-EN

Figure 3.24: Reorderings in the Hiero translations, plotted against the total target width of the reorderings.

3.4.3 Reproducing Alignments

Although we can now quantify the amount of reordering occurring in translations, we still have no idea whether they are correct or spurious. One way to approach this question is to investigate whether the reorderings seen in the human reference are being reproduced in the machine translations.

We proceed as follows. Individual reorderings between the source and reference sentences in the test set are identified. We then test translations to see whether they contain the same reorderings as the reference. By doing so, we identify which reorderings are being reproduced by the different translation models.

If a reordering has been translated by one phrase pair, we say that the reordering has been reproduced because the reordering could exist inside the phrase. If the segmentation is slightly different, but a reordering occurred within the scope of the reference reordering, we also claim that it has been reproduced. The results are therefore an upper-bound on how many reorderings were actually reproduced.

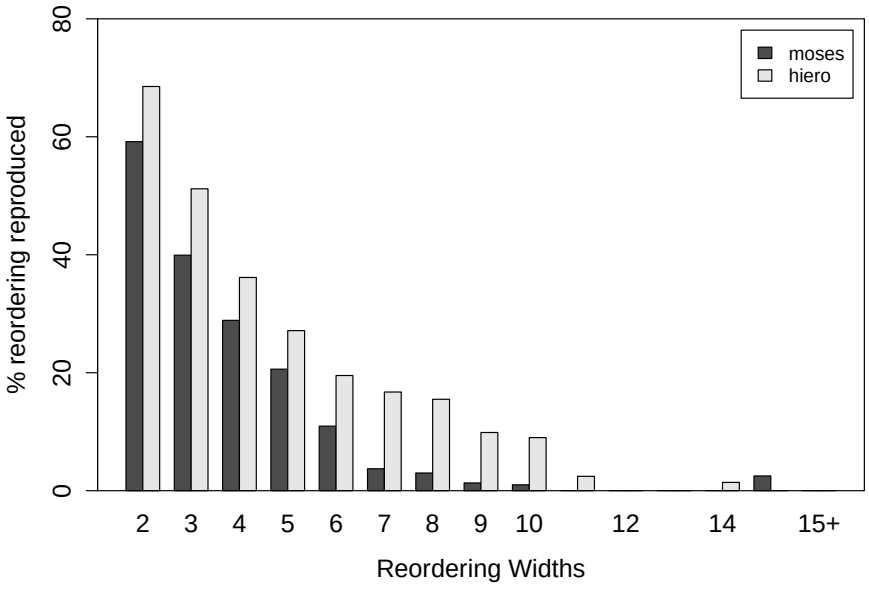


Figure 3.25: Percentage of reorderings reproduced by the phrase-based and hierarchical models for the combination of all the CH-EN reordering test sets. The data is shown relative to the length of the total target width of the reordering.

Figure 3.25 shows that the hierarchical model reproduces more reorderings of all widths than the phrase-based system, but especially for reorderings of width six to ten. This means that Hiero is performing better reordering than the phrase-based model.

Both systems retain very few reorderings however, and for distances of more than ten they reproduce practically none of the reorderings seen in the reference. As both models impose reordering restrictions, this is not surprising. Moses uses a distortion limit, and Hiero imposes a maximum source span for a rule, and in this thesis it is set to ten. As rules are then glued together in a monotone fashion, this means that no reorderings larger than ten are considered. Thus, any claims about the hierarchical model being able to perform long distance reorderings are clearly not supported.

3.4.4 Manual Analysis of Reproduced Alignments

We have established what reorderings have been reproduced. However we still need to determine whether reorderings which are reproduced are more likely to be correct. The translation model can compensate for not performing a reordering by using different lexical items. To judge the relevance of the evaluation performed in the previous section, Section 3.4.3, we perform a manual evaluation described in Section 3.3.5.

	Correct	Incorrect	NA	Total
Participant 1				
Reproduced	21	0	4	25
Not Reproduced	12	6	7	25
Participant 2				
Reproduced	21	0	4	25
Not Reproduced	11	10	4	25
Participant 3				
Reproduced	19	4	2	25
Not Reproduced	9	15	1	25
Total				
Reproduced	61	4	10	75
Not Reproduced	32	31	12	75

Table 3.4: Human evaluation of individual reorderings where they were either reproduced in the translation or they were not.

The results in Table 3.4 show that the reorderings which were reproduced are generally judged to be correct. If the reordering is not reproduced, then the evaluators divided their judgements evenly between the reordering being correct or incorrect. It

seems that the fact that a reordering is not reproduces does indicate that it is more likely to be incorrect. The cases where the reordering was reproduces, but it was judged to be incorrect could either be due to bad choice of words in the translation, or to human error. We used Fleiss' Kappa to measure the correlation between annotators. It expresses the extent to which the amount of agreement between raters is greater than what would be expected if all raters made their judgements randomly. In this case Fleiss' kappa is 0.357 which is considered to be a fair correlation.

3.5 Summary

This chapter provides a systematic analysis of reordering both in the original corpus, and in the output of two state-of-the-art translation models. In order to achieve this we present a novel method for extracting reorderings from parallel sentences.

This method of analysing reorderings is validated by detecting more and longer reorderings for the Chinese-English parallel corpus than for the Arabic-English corpus. More surprisingly, it shows that Arabic-English has more short distance reorderings than Chinese-English.

Finally, we show that the hierarchical model performs better than the phrase-based model in situations where there are many longer distance reorderings. However, we also show that the choice of translation model should be guided by the type of reorderings in the language pair, as the phrase-based model outperforms the hierarchical model where there a many short distance reorderings. Importantly, neither model is able to capture the reordering behaviour of the reference corpora adequately.

Chapter 4

Impact of Reordering on Translation Quality

4.1 Introduction

In the previous chapter, we proposed a method for analysing reordering in parallel corpora and used this to compare the performance of different translation models. In this chapter, we apply the same methods to a wide coverage study of the impact of reordering on translation performance.

The performance of machine translation systems varies greatly depending on the source and target languages involved. Knowing what characteristics of the language pair contribute to the variation in system performance is key to knowing what aspects of machine translation need to improve, and which have little impact. We are primarily interested in what impact reordering has on translation quality, but we also investigate two other factors: the morphological complexity and the language family similarity of the two languages. We wish to compare the importance of reordering as a factor in determining the performance of translation models, with other potential factors. Morphology and language relatedness were chosen, because they represent fundamental aspects of the challenge of translation. Morphological complexity makes it much more difficult to find the right words in translation, and lack of language relatedness would mean more divergence in language structure and lexical items.

We perform a survey of 110 different language pairs drawn from the Europarl project (Koehn, 2005). This contains parallel data for 11 official languages of the European Union and provides a rich variety of data for our experiments. Most research in machine translation only reports results on one or two languages pairs, by

analysing so many language pairs, we are able to provide a much wider perspective on the challenges facing machine translation.

The rest of the chapter proceeds as follows. In Section 4.2 we describe the Europarl corpus. Section 4.3 demonstrates the validity of using automatic alignments as the basis for extracting reorderings. Then the amount of reordering across the language pairs is investigated. Section 4.4 describes our approach to extracting the morphological complexity of a language and the language relatedness of a language pair. In Section 4.5, we describe the experimental design and, in particular, we describe our approach to regression analysis. In Section 4.6 we present the results of the analyses using the BLEU score as our dependent variable and the other factors as our independent variables. Finally, in Section 4.7 we discuss the contributions of this chapter.

4.2 Europarl

In order to analyse the influence of different language pair characteristics on translation performance, we need access to a large variety of comparable parallel corpora. A good data source for this is the Europarl Corpus (Koehn, 2005). Europarl Version 3 consists of a collection of the proceedings of the European Parliament, including the years from 1996 to 2006. It consists of up to 44 million words for each of the 11 official languages of the European Union. Table 4.1 lists the languages grouped in their language families.

Indo-European						**Non Indo-European**	
Germanic		Romance		Greek		Finno-Ugric	
Swedish	sv	French	fr	Greek	el	Finnish	fi
German	de	Portuguese	pt				
Dutch	nl	Italian	it				
Danish	da	Spanish	es				
English	en						

Table 4.1: Europarl languages and their abbreviations grouped together in their language families

In trying to determine the effect of properties of the languages involved in translation performance, it is important that other confounding factors are minimised. The Europarl corpus contains data from just one domain, and, with the exception of Greek,

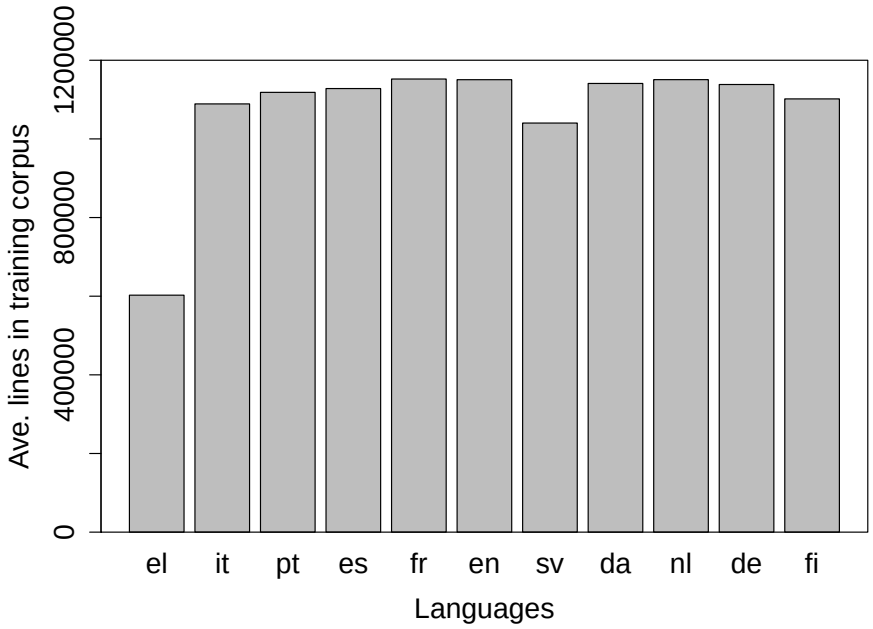


Figure 4.1: Average size of corpora for language pairs which include the language specified.

all the corpora have similar sizes.

Figure 4.1 shows the average size of the corpora involved in the experiments. As there are 110 training corpora, we average the number of sentences in all parallel corpora in which each particular language is either the source or target language.

4.3 Reordering Characteristics

The overall quality of statistical machine translation has improved considerably over the last decade of intensive research, but some language pairs still result in very poor translations. Many researchers have postulated on the reasons why machine translation is hard. However, there has never been, to our knowledge, a systematic analysis of the effect of different characteristics of the language pairs on translation performance. Understanding where difficulties lie, allows researchers to focus their efforts on those aspects of translation that have the most impact on translation quality.

The basic challenges facing statistical machine translation were first outlined by Brown et al. (1993). The original IBM Models were broken down into separate translation and distortion models, thus recognising the importance of word order differences in modelling translation. Brown et al. (1993) also highlighted the importance

of modelling morphology, both for reducing sparse counts and improving parameter estimation and for the correct production of translated forms. We see these two factors, reordering and morphology, as fundamental to the quality of machine translation output, and we would like to quantify their impact on system performance. In this section, we measure the amount of reordering in a parallel corpora. We do this by adopting the reordering extraction approach described in the previous chapter, in Section 3.2. We first justify using automatic alignments, and then we describe the reordering characteristics of the Europarl corpus.

4.3.1 Automatic Alignments

The major difference between the treatment of reordering in this chapter and the previous one, (Chapter 3), is that gold standard word alignments are not available. Human annotated alignments are very expensive to create and only exist for a very small number of language pairs. We therefore need to rely upon automatic alignements. In order to justify using reordering data extracted from automatic alignments, we must show that they are similar to gold standard alignments.

4.3.1.1 Experimental Design

We compare reordering extracted from gold standard alignments and auatomatic alignments for the German-English language pair. We select German-English because it has a reasonably high expected level of reordering. We also have access to a manually aligned German-English corpus[1] which consists of the first 220 sentences of test data from the 2006 ACL Workshop on Machine Translation (WMT06) test set. This test set is from a held out portion of the Europarl corpus. The automatic alignments were extracted by appending the test set onto the German-English training corpus and aligning using GIZA++ and then applying the grow-final-diag algorithm.

4.3.1.2 Results

In order to use automatic alignments to extract reordering statistics, we need to show that reorderings from automatic alignments are comparable to those from manual alignments.

Table 4.2 shows the total amount of reordering for the manually and automatically aligned WMT06 test corpus and the automatically aligned Europarl training corpus.

[1] provided by Chris Callison-Burch

	RQuantity
Europarl, automatically aligned	0.62
WMT06 test, automatically aligned	0.65
WMT06 test, manually aligned	0.67

Table 4.2: The total amount of reordering for the different corpora.

The manually aligned test corpus has a slightly higher RQuantity of 0.67, and the automatically aligned test corpus has a slightly lower RQuantity of 0.65. But all these results are very similar.

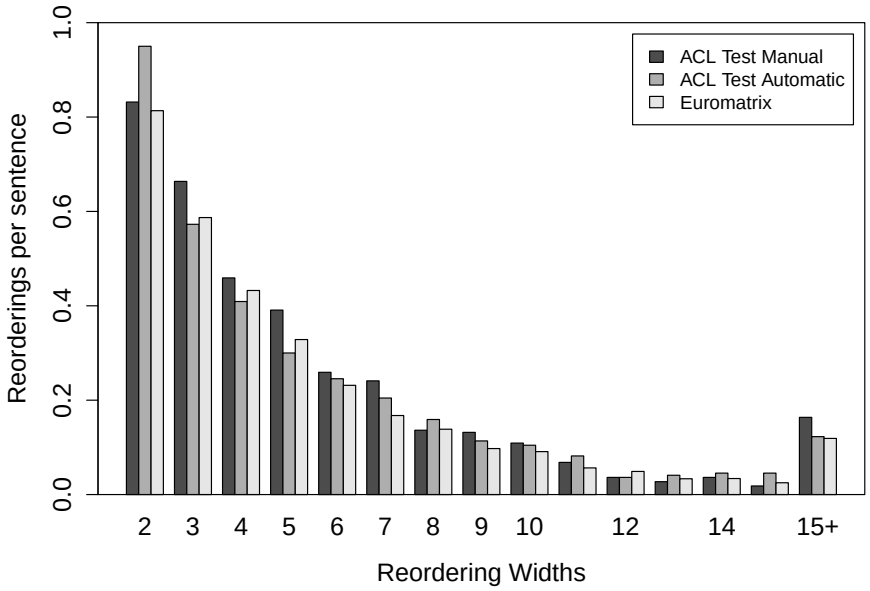


Figure 4.2: Average number of reorderings per sentence mapped against the total width of the reorderings for DE-EN.

Figure 4.2 shows the more detailed distributions of the reorderings for the three corpora. The corpora have very similar distributions with the automatically aligned test corpus showing slightly more reorderings of length two and the manually aligned corpus showing more reorderings of lengths greater than 15. These results provide evidence to support our use of automatic reorderings in lieu of manually annotated alignments. Firstly, they show that our WMT06 test corpus is very similar to the Europarl data, which means that conclusions that we reach using the WMT06 test corpus will hold for the Europarl data. Secondly, they show that the reordering characteris-

tics of the test corpus is very similar when extracted from automatic or from manual alignments.

Although we have shown that there are few differences between the manually and automatically aligned German-English corpus, there is no guarantee that this result extends to other corpora. Because German-English contains a reasonably large amount of reordering, it is likely to extend to more langauge pairs. However, there might exist a language pair whose alignments are very unsuited to the stochastic assumptions of the IBM or HMM alignment models. In any case, due to the number of language pairs involved in this study, we are obliged to rely upon automatic alignments.

4.3.2 Amount of reordering for the matrix

We extract RQuantity for the matrix of language pairs in the following manner. We randomly sampled a subset of 2000 sentences from each of the parallel training corpora. This is a large enough sample to accurately reflect the reordering characteristics of the whole Europarl training corpus. We then calculated the average RQuantity over the target side.

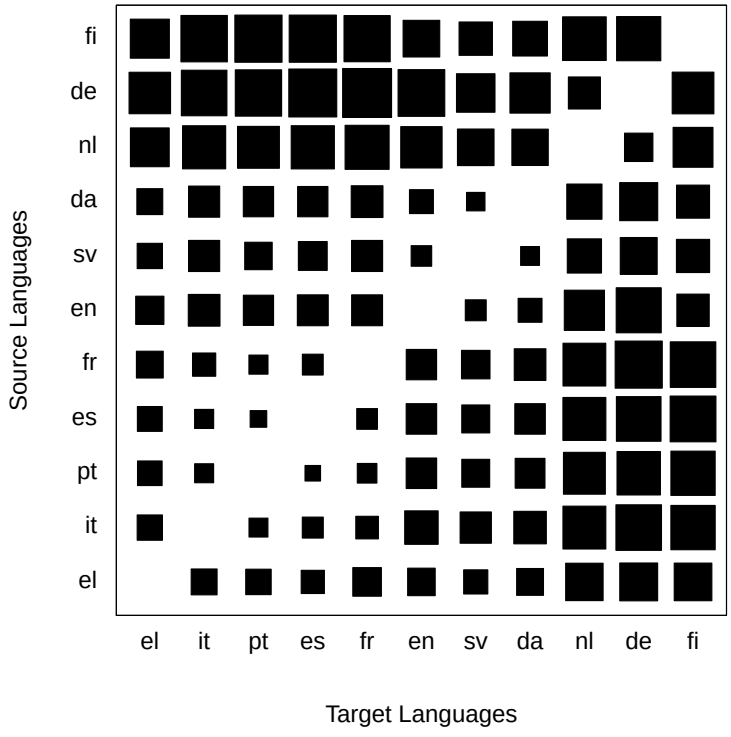


Figure 4.3: RQuantity for the matrix of langauge pairs

Figure 4.3 shows the RQuantity for each of the language pairs. The width of the

squares are proportional to the RQuantity. Note that the matrix is not quite symmetrical - reordering results differ depending on which language is chosen to measure the reordering span. The table of values for this Figure is provided in Appendix C.

Lowest RQuantity		Highest RQuantity	
pt-es	0.202	fr-de	0.613
es-pt	0.216	fi-pt	0.614
da-sv	0.240	fi-es	0.614
sv-da	0.245	de-es	0.624
it-pt	0.246	de-fr	0.637

Table 4.3: The language pairs with the lowest and highest amounts of reordering.

Table 4.3 shows a selection of results from the matrix, highlighting the lowest and the highest amounts of reordering. The lowest reordering scores are for languages in the same language group, like Portuguese-Spanish and Danish-Swedish, and the highest for languages from different groups, like German-French, and Finnish-Spanish.

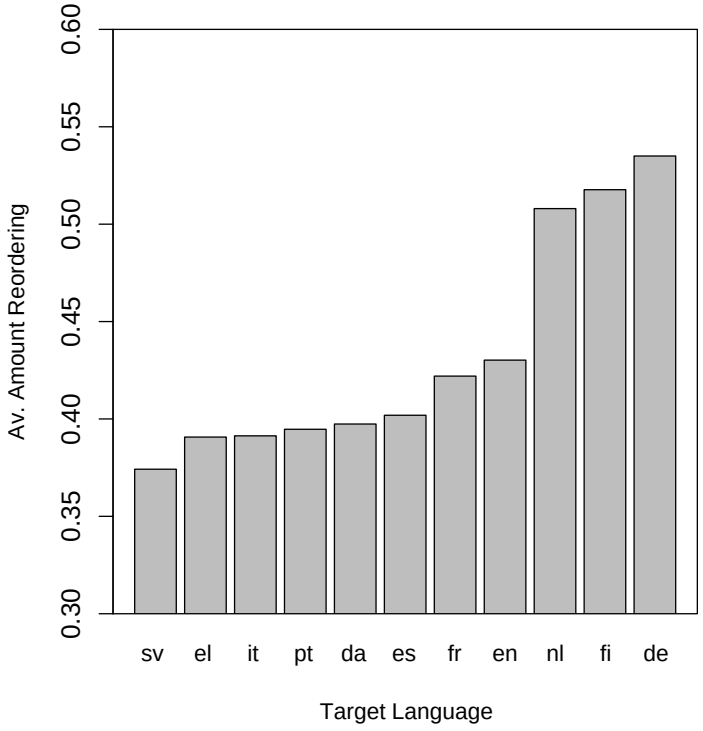


Figure 4.4: Average amount of reordering for each target language.

Figure 4.4 shows the average amount of reordering for each target language. German shows the largest amount of reordering overall. This is only partially explained

by the fact that it is unrelated to the biggest language group, the Romance languages. German also shows a relatively large amount of reordering for languages with which it is more closely related, such as Swedish, English and Danish.

Figure 4.5 shows the distribution of the reorderings for language pairs with small and large amounts of reordering. Here both short distance and long distance reorderings vary with the amount of reordering. Figure 4.6 shows a sample of reordering distributions where the target language is English. With the source language being French, there is a relatively large number of short distance reorderings, even though the total amount of reordering is quite small. This is because of the smaller number of medium and long distance reorderings as compared to the other two languages, Finnish and German. RQuantity is more sensitive to these larger reorderings and these are the ones that translation models struggle to capture. The graphs show researchers the reordering characteristics of language pairs. This allows them to choose appropriate language pairs for testing their improvements.

4.4 Other Characteristics

We would like to compare the impact of reordering with other important characteristics of translation. In this section we describe how to extract the morphological complexity and the language relatedness of a language pair.

4.4.1 Morphological Complexity

The morphological complexity of the languages involved in translation is widely recognised as one of the factors influencing translation performance. However, most statistical translation systems treat the various inflected forms of the same word as completely independent of one another. “cat” and “cats”, for example, are treated as unrelated words. This can result in sparse statistics and poorly estimated models, especially for languages with rich morphology. Furthermore, using the wrong form of a word may result in crucial differences in meaning that affect the quality of the translation.

Work on improving treatment of morphology has focused on either reducing word forms to lemmas to reduce sparsity (Goldwater and McClosky, 2005; Talbot and Osborne, 2006) or including morphological information in decoding (Dyer, 2007; Avramidis and Koehn, 2008). In this chapter we aim to discover the effect that different levels of morphological complexity has on translation.

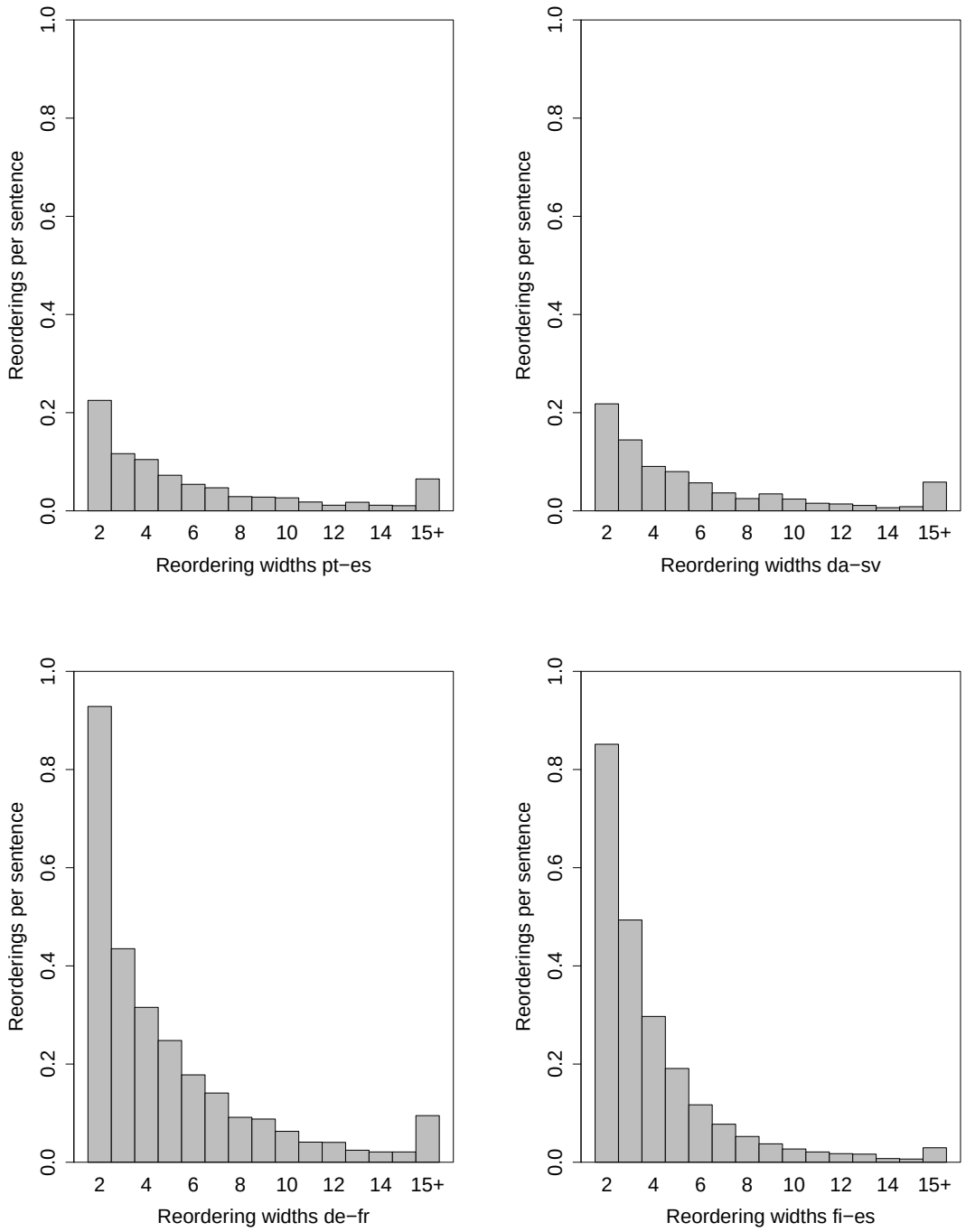


Figure 4.5: Distribution of reorderings in language pairs for cases with small amounts of reordering (Portuguese-Spanish and Danish-Swedish) and with large amounts of reordering (German-French and Finnish-Spanish). The reorderings are distributed according to the width of the reorderings on the source language side and are normalised by the number of sentences.

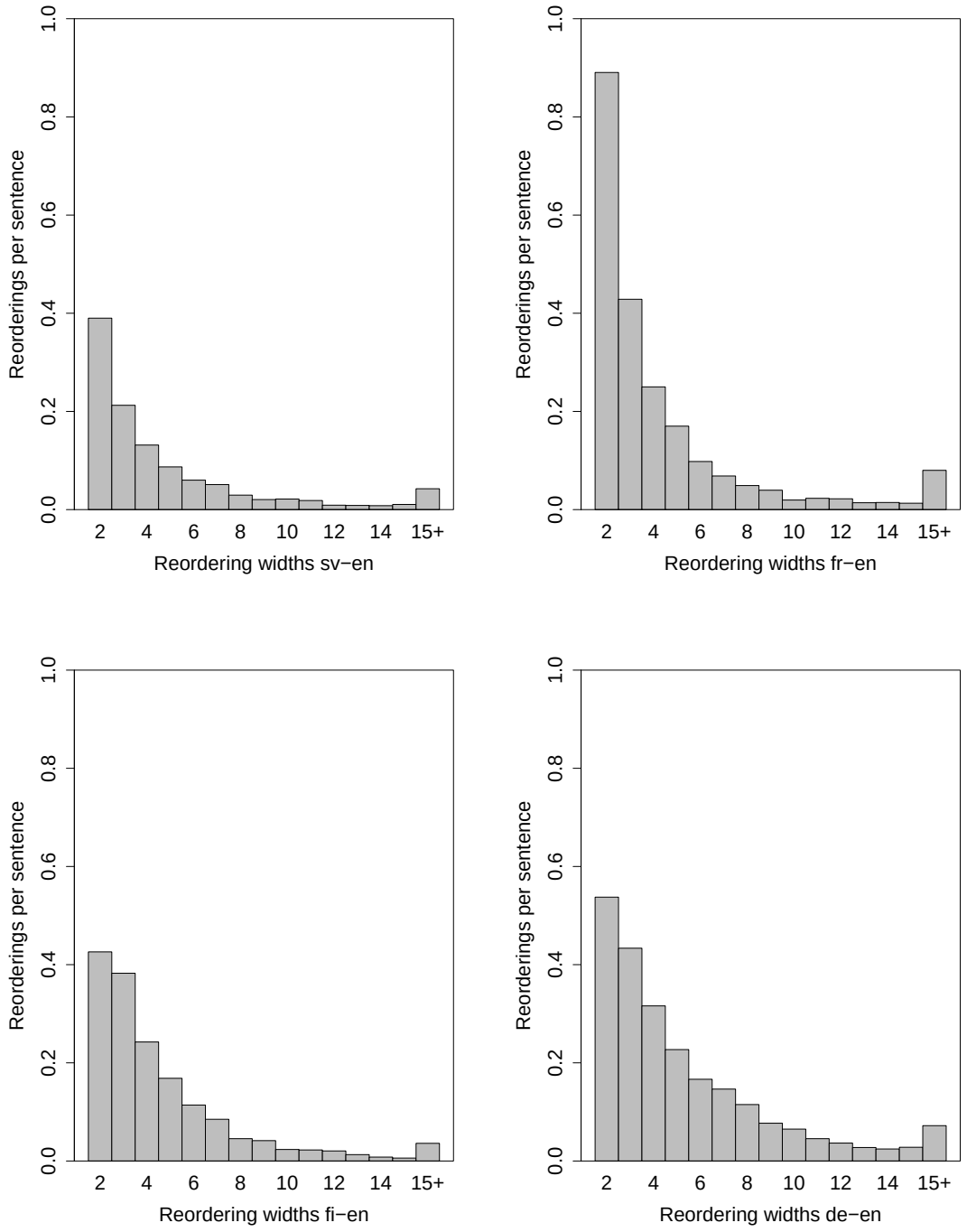


Figure 4.6: Distribution of reorderings for translation into English for cases with varying amounts of reordering from small to large (Swedish, French, Finnish and German). The reorderings are distributed according to the width of the reorderings on the source language side and are normalised by the number of sentences.

Some languages seem to be intuitively more complex than others, for instance Finnish appears more complex than English, but it is difficult to quantify this. One method of measuring complexity is by choosing a number of hand-picked, intuitive properties called *complexity indicators* (Bickel and Nichols, 2005) and then to count their occurrences. Examples of morphological complexity indicators could be the number of inflectional categories or morpheme types in a typical sentence. The major drawback of this method is finding a principled way of choosing which of the many possible linguistic properties should be included in the list of indicators.

A simple alternative employed by Koehn (2005) is to use vocabulary size as a measure of morphological complexity. Vocabulary size is strongly influenced by the number of words forms affected by number, case, tense etc. and its also affected by the number of agglutinations in the language. The complexity of the morphology of languages can therefore be approximated by examining vocabulary size.

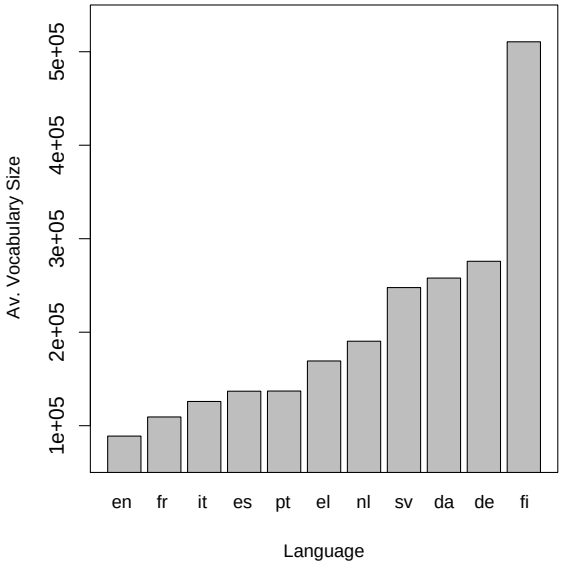


Figure 4.7: Average vocabulary size for each language.

Figure 4.7 shows the vocabulary size for our 11 languages. Each language pair has a slightly different parallel corpus, and so the size of the vocabularies for each language needs to be averaged. The size of the Finnish vocabulary is about six times larger (510,632 words) than the English vocabulary size (88,880 words). Finnish is an agglutinative language with fusional features and it has a highly productive morphology. It has been estimated that a Finnish noun can have more than 2000 different inflected and cliticized forms and verb morphology is even more complex (Laine et al., 1994).

An example of a noun with rich morphological information is shown in Figure 4.8.

Finnish word	tulu + i + ssu + ni + ko
Meaning	house + plural + inessive + possessive + clitic
English translation	in my houses ?

Figure 4.8: A Finnish noun broken down into its component parts. Example provided by Laine et al. (1994)

4.4.2 Language Relatedness

Comparative linguistics is field of linguistics which aims to determine the historical or phylogenetic relatedness of languages. Lexicostatistics is an approach to comparative linguistics that is appropriate for our purposes because it results in a quantitative measure of relatedness (Swadesh, 1955). It does this by comparing lists of lexical cognates.

The lexicostatistic percentages are extracted as follows. First, a list of universal culture-free meanings are generated. Words are then collected for these meanings for each language under consideration. We use the data from Dyen et al. (1992) who developed a list of 200 meanings for 84 Indo-European languages and calculated their lexicostatistics.

Cognacy decisions are then made by a trained linguist. For each pair of lists the cognacy of a form can be positive, negative or indeterminate. Finally, the lexicostatistic percentages is calculated. This percentage is related to the proportion of meanings for a particular language pair that are cognates, i.e. relative to the total without indeterminacy. Factors such as borrowing, tradition and taboo words can skew the results.

We show how the lexicostatistics are generated by using an example. A portion of the Dyen et al. (1992) data set is shown in Table 4.4. From this we could calculate the similarity of French, Italian and Spanish with each other as 100% because the two words are cognates. The Romance languages share one cognate with English, which means that the lexicostatistic percentage here would be 50%, and no cognates with the rest of the languages resulting in a score of 0%. We use these lexicostatistic percentages as our measure of language relatedness for the 55 bidirectional language pairs.

Figure 4.9 shows the symmetric matrix of language relatedness, where the width of the squares is proportional to the value of relatedness. The values range from Finnish

Language	“animal”	“black”
French	animal	noir
Italian	animale	nero
Spanish	animal	negro
English	animal	black
German	tier	schwarz
Swedish	djur	svart
Danish	dyr	sort
Dutch	dier	zwart

Table 4.4: A subset of the Dyen et al. (1992) cognate list.

to other languages, which is 0%, to Spanish-Portuguese, which is 87.4%. The table of actual values is provided in Appendix C.

Finnish is a Finno-Ugric language and does not form part of the Indo-European languages. The Dyen data does not include Finnish and we assume that it has 0% similarity with other languages. If one considers that English and Hindi are more closely related than English and Finnish, then this assumption seems justified. However, it is possible that the actual statistic might be higher than 0% because Finnish is likely to have borrowed words from neighbouring European languages.

4.5 Experimental Design

We analyse the performance of 110 different translation models drawn from the Europarl project. The purpose of doing so is to determine the impact of different language characteristics on translation quality.

The phrase-based model MOSES (Koehn et al., 2007) was used for the experiments with all the standard settings, including a lexicalised reordering model, and a 5-gram language model, trained on the target side of the corpora. Tests were run on the ACL WMT 2008 test set (Callison-Burch et al., 2008).

4.5.1 Evaluation of Translation Performance

We use the BLEU score to evaluate our systems. While BLEU scores are not strictly comparable across language pairs, they do give an indication of the quality of the translation. The translation setup is kept constant, with the only important difference

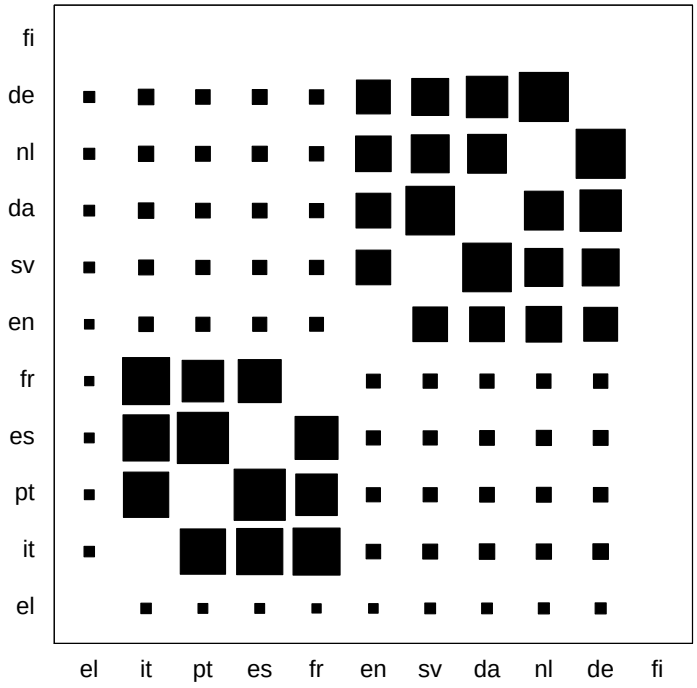


Figure 4.9: Lexicostatistic measure of langauge relatedness for the matrix

between systems being the language pair in question. This means that BLEU score differences should largely reflect the innate difficulty of translating the different language pairs.

Figure 4.10 shows the BLEU score results for the matrix. The table of values is provided in Appendix C. Comparing this figure to Figure 4.3 there seems to be a clear negative correlation between the amount of reordering and translation performance.

4.5.2 Regression Analysis

Linear regression We first perform simple linear regression in order to determine the relative strength of the relationship between the language characteristics and the quality of translation. In statistics, regression analysis helps us to understand how the value of the dependent variable changes when one of the independent variables is varied. We perform linear regression analyses using measures of morphological complexity, language relatedness and reordering amount as the independent variables. The dependent variable is the the BLEU score. We test how well the simple linear regression models explain the data using the r^2 test. r^2 is equal to the square of the Pearson's correlation coefficient between the

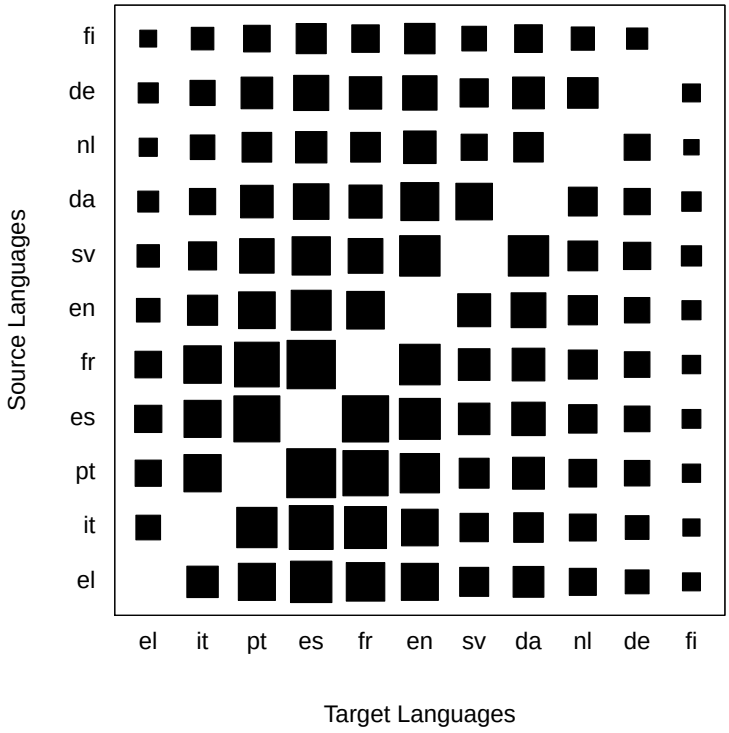


Figure 4.10: BLEU score performance of the different translation systems

observed and predicted data values. It is of interest because it provides a measure of how well future outcomes are likely to be predicted by the model. The two-tailed significance levels of coefficients are also given. We use a t-test to determine whether the coefficients for the independent variables are reliably different from zero. Significance results for the rest of the thesis are reported as follows: * means $p < 0.05$, ** means $p < 0.01$, and *** means $p < 0.001$.

Mixed-effects models We next investigate the effect of treating the source and target languages as random variables. We are interested in the experimental effects of reordering on languages in general, and not on their effect on a particular source or target language. Finnish, for instance, has a very high amount of reordering. The regression model should have the freedom to estimate a higher level of reordering for Finnish than the other languages. Allowing the model to incorporate different levels of reordering for different languages allows the final model to generalise better. In standard logistic regression analysis all features are assumed to be fixed effects, meaning that all possible values for these features are known, and each value may have an arbitrarily different effect on the outcome. However some features do not fit this pattern. Mixed-effects models are a gen-

eralisation of linear regression which allows for the inclusion of random effect, meaning that the features observed in the data are a random sample from a larger population (Pinheiro and Bates, 2009). Random effects are often the participants or items in an experiment, but in our case, they are languages used. We follow methods described by Baayen et al. (2008), who use linear mixed-effects models for the analysis of repeated measurement data. We fit our models using the lme4 package (Bates and Sarkar, 2007) of R (Team, 2009).

Model simplification The mixed-effects model can combine numerous fixed factors. We initially consider the maximal model which considers all factors and all their interactions. This results in a large model where many factors are insignificant. Following the principle of parsimony, we simplify the model using the Akaike Information Criterion (AIC) (Crawley, 2007). The AIC represents a trade-off between the fit of the model with the complexity, or degrees of freedom, of the model. At each step we test the least significant variable, seeing if removing it leads to a significant increase in deviance (or decrease in AIC) as compared to the current model. Significance is determined with a χ^2 test on the model with the variable and without it. Variables are removed if they do not significantly increase deviance. In this fashion we arrive at the minimal adequate model.

Collinearity The coefficients of the variables in the regression model have only limited usefulness as a measure of the impact of the explanatory variables in the model. One important factor to consider is that if the explanatory variables are highly correlated, then the values of the coefficients can be unstable. The model could attribute more importance to one or the other variable without changing the overall fit of the model. In our models, for instance, reordering and language similarity are likely to be correlated. We resolve this problem by residualising the effects with the correlated predictors in the model if there are high correlations between them ($>$0.2). Residualisation means to regress the collinear predictor against correlated predictors. Unfortunately this makes the effect sizes hard to interpret. The effect sizes now refer to the portion of the main effect that is not explained by the other correlated predictors.

Outlier Removal The final step is dealing with the problem of outliers. Outliers are data points that deviate markedly from the others in the sample, and thereby have an undue influence on the model. We detect outliers by examining residual values. The residual values of the regression model are the difference between the

observed values of the dependent variable and the values fitted by the model. We remove any points whose residual values are greater than 2 standard deviations from the mean of the distribution of residual values.

4.6 Results

4.6.1 Data Exploration

We start our experiments by investigating the relationship between each of the main explanatory variables and the BLEU score. We perform simple linear regressions with just one explanatory variable. We are particularly interested in the strength of the correlation of the effects with the BLEU score in isolation of each other, and seeing whether the assumptions of linear regression are valid.

Explanatory Variable	r^2	
Reordering Amount	0.391	***
Language Similarity	0.366	***
Target Vocabulary Size	0.387	***
Source Vocabulary Size	0.043	*
Corpus Size	0.059	*

Table 4.5: The goodness of fit of different simple linear regression models which use just one explanatory variable. The significance level represents the level of probability that the regression is appropriate.

Table 4.5 describes the amount of the variance of BLEU explained by the simple regression models with different explanatory variables. This table shows that reordering shows the highest correlation with the BLEU scores of all the explanatory variables. The reordering r^2 of 0.391 means that reordering can account for 39.1% of the variance of the BLEU scores. Language similarity and target vocabulary size account for slightly less variance than reordering does. Source vocabulary size and corpus size explain much less of the variance than the other variables.

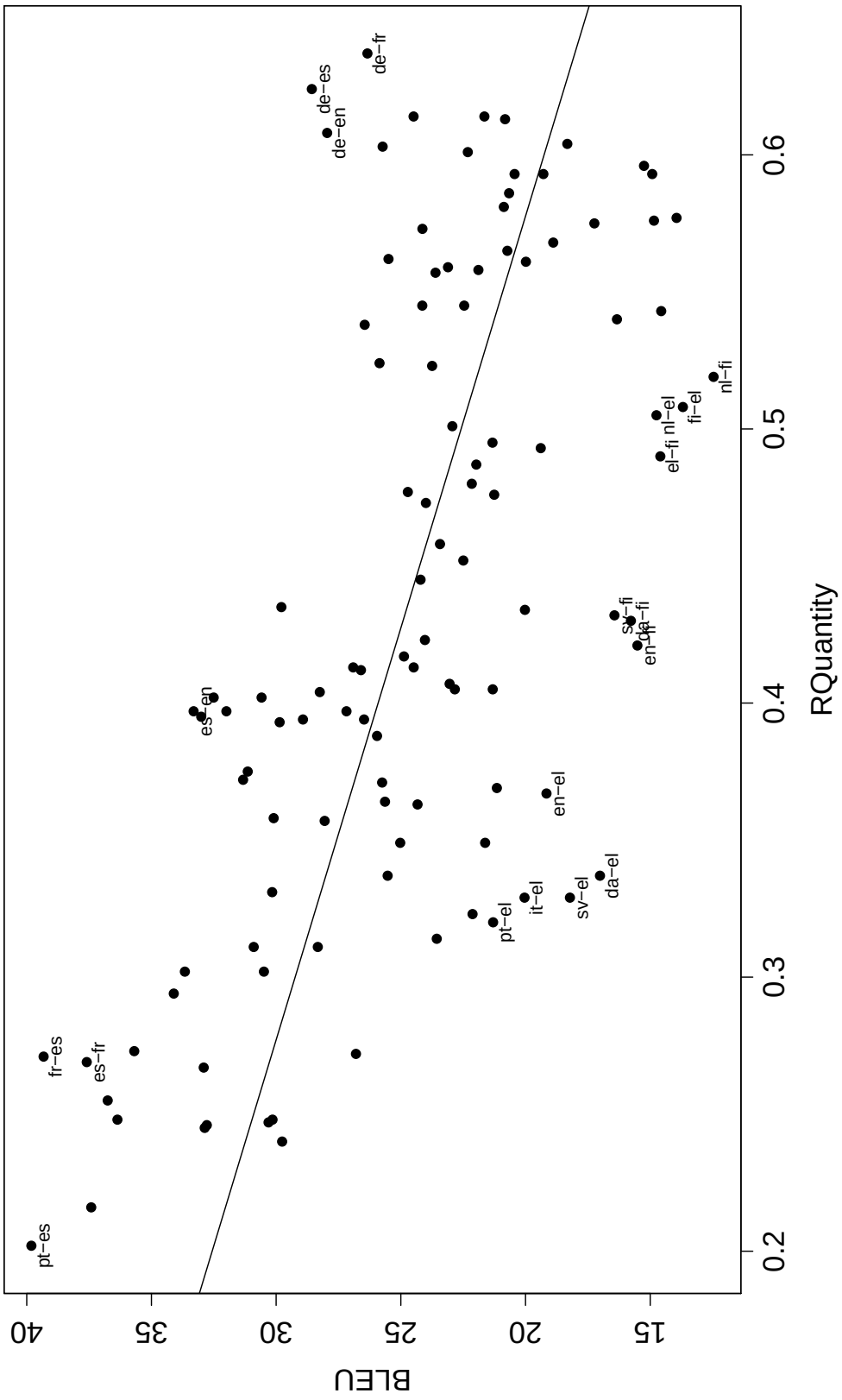


Figure 4.11: BLEU score vs. the amount of reordering, with the fitted model shown.

Figure 4.11 shows the simple regression model over the plot of BLEU scores against the amount of reordering. This graph clearly shows the impact that reordering has on performance. With more reordering, the performance of the translation model reduces. Data points with a large effect on the model are labelled for inspection. These are data points where the residuals are further than 1.5 standard deviations from the mean of the distribution of residual values. Data points with low levels of reordering and high BLEU scores tend to be language pairs where both languages are Romance languages. High BLEU scores with high levels of reordering tend to have German as the source language and a Romance language or English as the target.

Figure 4.12 shows the plot of the BLEU score and the other explanatory variables: source and target vocabulary size, corpus size and language similarity. Target vocabulary size and language similarity are much more important effects than source vocabulary size and corpus size, and their greater correlation with the BLEU score can be seen in the figure.

Explanatory Variable	Lang. Sim.	Target Vocab.	Source Vocab.	Corpus
Reordering Amount	-0.48	0.27	0.36	0.22
Language Similarity		-0.26	-0.26	0.31
Target Vocabulary Size			-0.09	0.12
Source Vocabulary Size				0.12

Table 4.6: Pearsons' correlation coefficient between predictors.

Table 4.6 shows the correlation of the effects with each other and many of the effects are relatively highly correlated with one another. Language similarity and reordering are particularly highly correlated with a Pearson's coefficient of -0.48, which is not unexpected. The further apart two languages are, the more their structure can diverge. The simple linear regression models are interesting because they allow us to gauge the intuitive impact of the different variables [2]. However, issues such as collinearity, outliers and random effects still need to be accounted for and they will be dealt with in the next experiment.

[2] We also fitted minimally adequate multiple regression model using normalised reordering, morphology and language relatedness as our independent variables. The r^2 of this model was 0.750 which means that together these factors explain most of the variability in translation performance.

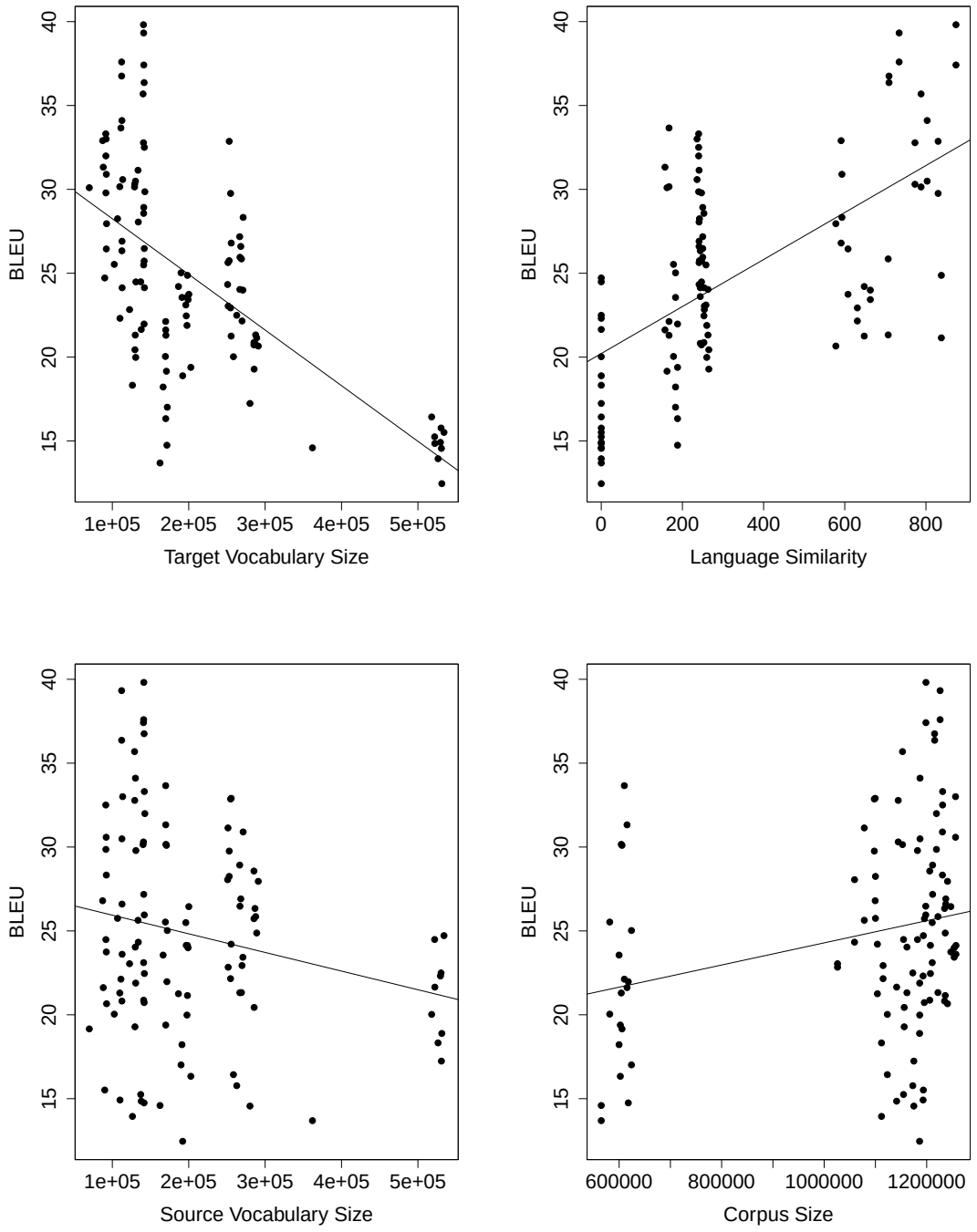


Figure 4.12: BLEU score vs. source and target vocabulary size, corpus size and language similarity. The fitted models are also shown.

4.6.2 Linear Mixed-Effects Model

We hypothesise that reordering is important to the performance of translation systems. The simple models are useful to explore the data available, but they do not conclusively demonstrate the unique contributions of the different effects. As described in Section 4.5, we fit a linear mixed-effects model to the data. In our experiments, we treat the source and target languages as random effects which means that our model contains an intercept for each. We also experimented with adding random slopes for source and target languages, but this failed to increase model fit and they were therefore discarded.

We start by standardising and centering all the data for the explanatory variables. As reordering and language similarity are highly correlated, we residualise them against each other. The initial maximal model includes all the interactions between the main effects: reordering, language similarity and target vocabulary and we add the corpus size, source vocabulary size and the square of reordering as additional explanatory variables. The reason for including the square of the reordering is that an analysis of the residual values versus predicted values of the simple linear regression model shows that it makes systematic errors when values are very small or very large. This means that the relationship between reordering and BLEU is not entirely linear.

Then we fit a minimal adequate model, by removing all terms which do not lead to a significant increase in the AIC of the model. During this procedure we discarded all source vocabulary, target vocabulary and corpus size terms. The source and target language random effects thus adequately account for differences in vocabulary and corpus size.

The final modification to the model is to remove data points with excessive influence on the model. Outliers are detected by taking points with residuals which are greater than 2 times the standard deviation of the distribution of residuals. There were four such points and removing them meant that the interaction term of reordering and language similarity, and the square of the reordering were no longer significant factors in the model.

Figure 4.7 reports the final fixed effects in the linear mixed effects model. The intercept of the model shows that the mean level of response or BLEU score would be 24.57. The model also shows the large negative impact that reordering has on performance where the coefficient is -3.45. This is the coefficient of the standardised, residualised reordering amount. Language similarity also has an important positive

Fixed Effects	Coefficient	Significance
Intercept	24.57	***
Reordering Amount	-3.45	***
Language Similarity	2.65	***

Table 4.7: Linear mixed model fixed effects coefficients and their significance.

impact on performance, with a coefficient of 2.65. This model conclusively proves the importance of reordering and language similarity in determining the success of the translation model. They are factors which contribute extra information above and beyond the knowledge of what source and target languages were used.

The conclusions that we draw in this chapter are only strictly relevant to the model for which this analysis has been performed, the phrase-based model. Models with different reordering capabilities, such as synchronous grammar-based models, might find that morphology contributes more to performance variability. However, in the previous chapter, Chapter 3, which addressed the reordering behaviour of different models, we demonstrated that reordering is still a big challenge for hierarchical models.

4.7 Summary

This chapter explores the amount and distribution of reordering seen across a wide variety of language pairs. Together with language similarity, reordering was seen to be a highly significant predictor of translation performance across the 110 language pairs that were examined.

During an initial exploration of the data, we investigated the simple linear relationship between the BLEU score and reordering, language similarity, source and target vocabulary size and corpus size. This exploration showed that reordering, language similarity and target vocabulary size each account for just over a third of the variation of the BLEU score. However, when applying linear mixed models with the source and target language as a random effect, then only reordering and language similarity still explain performance. Indeed, reordering has the largest coefficient, and therefore the greatest impact on performance. For this thesis we have thus demonstrated the importance of reordering in machine translation and this motivates further research on how best to measure translation quality.

Chapter 5

Reordering Metrics

5.1 Introduction

In the preceding chapters, we have presented two important findings: translation models are still not close to modelling the reordering performance of human translators; and reordering is an important predictor of the quality of translation output. These findings motivate the need for both better models of reordering, and also better metrics to evaluate them. In this chapter we propose novel metrics of reordering which directly measure word order differences between human reference sentences and machine translations.

There is currently a great deal of research dealing with the problem of improving the reordering performance of translation systems. Reordering models, translation models, and search constraints have all been extensively investigated. However, this work is hampered by the fact that automatic machine translation metrics only measure word order quality indirectly.

We argue that it is important to evaluate reordering performance directly. Our approach relies upon the assumption that orderings which are close to the word order of the reference are going to be preferable to orderings which are very different and we present a method for doing this using *permutation distance metrics*. We first extract permutations from alignments, and then we apply standard distance metrics to compare the reference permutation and the translation permutation. These intuitive measures are sensitive to the size and frequency of reorderings. They are also efficient, language independent and they are meaningful at a sentence level. These properties make them desirable automatic machine translation metrics.

The rest of the chapter proceeds as follows. In Section 5.2 we define permutations

and describe how to convert alignments into ranked data. In Section 5.3 we present permutation distance metrics and discuss why they are appropriate. In Section 5.4 we explore their properties, with the help of an example, and contrast them with the machine translation metrics.

5.2 Permutations over Alignments

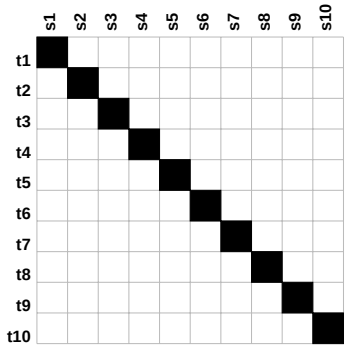
In machine translation the relative ordering of words in the source and target is encoded in alignments. A word alignment over a sentence pair allows us to transcribe the source word positions in the order of the aligned target words. This results in a permutation on which metrics for ordered encodings can be applied in order to measure and evaluate reorderings. Permutations have already been applied in machine translation. Eisner and Tromble (2006) present a reordering model which uses ordering costs to score possible permutations. Here, however, we use permutations in a novel fashion to evaluate reordering performance.

The ordering of the words in the target sentence can be seen as a permutation of the words in the source sentence. The source sentence s of length n consists of the word positions $s_0 \cdots s_i \cdots s_n$. Using an alignment function where a source word at position i is mapped to a target word at position j with the function $a : \{i \rightarrow j\}$, we can reorder the source word positions to reflect the order of the words in the target. This gives us a permutation.

A *permutation* is a bijective function from a set of natural numbers $1, 2, \cdots, n$ to itself. We name our permutations π and σ. The ith symbol of a permutation π is denoted as $\pi(i)$, and the inverse of the permutation π^{-1} is defined so that if $\pi(i) = j$ then $\pi^{-1}(j) = i$. The identity, or monotone, permutation id is the permutation for which $id(i) = i$ for all i. Figure 5.1 contains a number of alignments and their associated permutations. The permutations are calculated by iterating over the source words, and recording the relative order of the aligned target words.

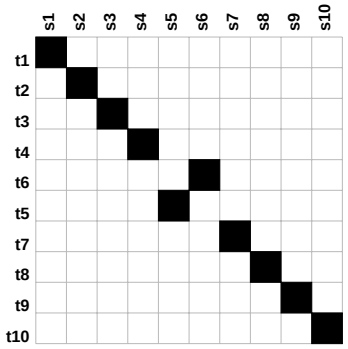
Permutations encode one-one relations, whereas alignments contain null alignments and one-many, many-one and many-many relations. We make some simplifying assumptions to allow us to work with permutations:

- **Unaligned source words**: Source words aligned to null ($a(i) \rightarrow null$) are assigned the target word position immediately after the target word position of the previous source word ($\pi(i) = \pi(i-1) + 1$). If the source word is the first word



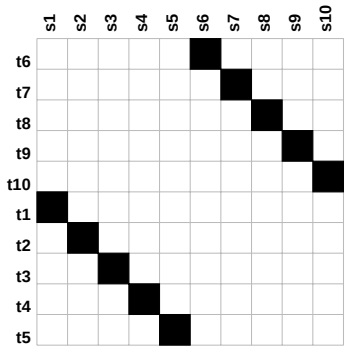
(1 2 3 4 5 6 7 8 9 10)

(a)



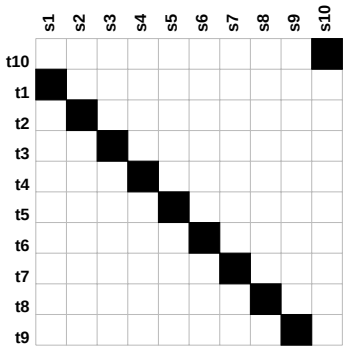
(1 2 3 4 • 6 • 5 • 7 8 9 10)

(b)



(6 7 8 9 10 • 1 2 3 4 5)

(c)

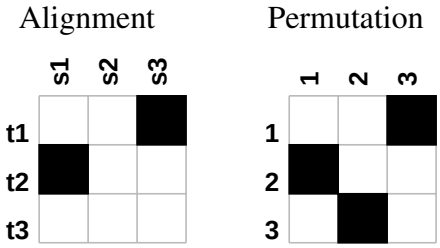


(2 3 4 5 6 7 8 9 10 • 1)

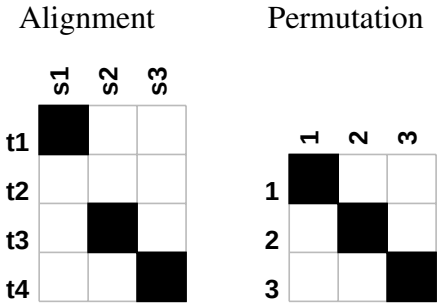
(d)

Figure 5.1: Synthetic examples of alignments and their permutations, where bullet points highlight non-sequential neighbours. (a) is a monotone translation, (b) is a translation with one short distance word order difference, (c) is a translation where the order of the two halves has been swapped, and (d) is a translation with a long distance reordering of the last source word.

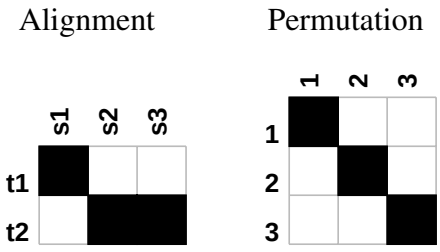
in the sentence, it is aligned to position 1. Below is an example of how an unaligned source word is assigned the position which follows the previous source word position.



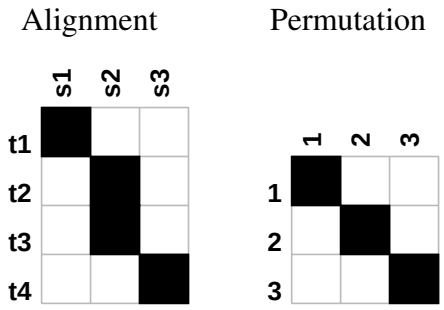
- **Unaligned target words**: These are ignored.



- **Many-to-one source to target alignment**: Where multiple source words are aligned to the same target word or phrase, the target ordering is assumed to be monotone.



- **One-to-many source to target alignment**: When one source word is aligned to multiple target words, the source word is assumed to be aligned to the first target word.



These simplifications are applied on the assumption that the default ordering is monotone, and this reflects the largely monotone ordering of translation output. Monotone orderings avoid introducing spurious reorderings which would occur if one linked an unaligned source word with, say, the first target position.

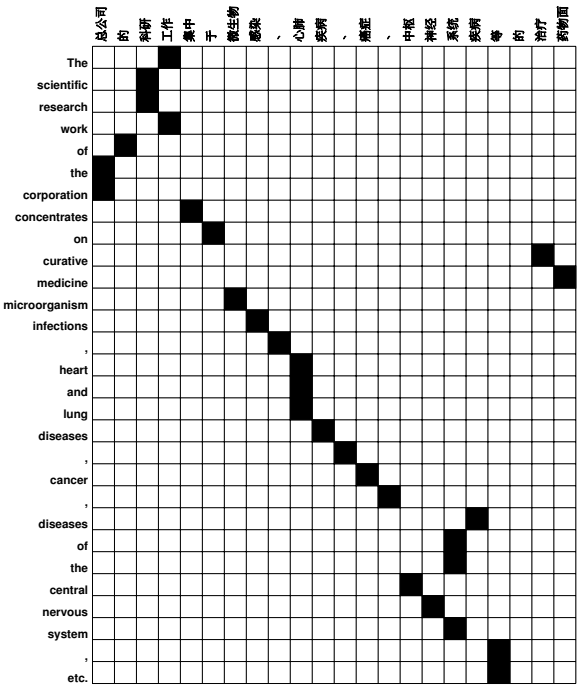
Although these simplification assumptions can result in significant changes to the original alignment, on the whole they are still able to capture the differences in order between the source and target language. Figure 5.2 presents an example of how the extraction process works with a non-trivial alignment. Although this sentence pair contains a complex alignment in (a), in (b) it shows how the simplification assumptions result in acceptable orderings.

In this section we have discussed the process of converting alignments into permutations. In the next section we describe metrics over these permutations which are intuitive and useful for machine translation.

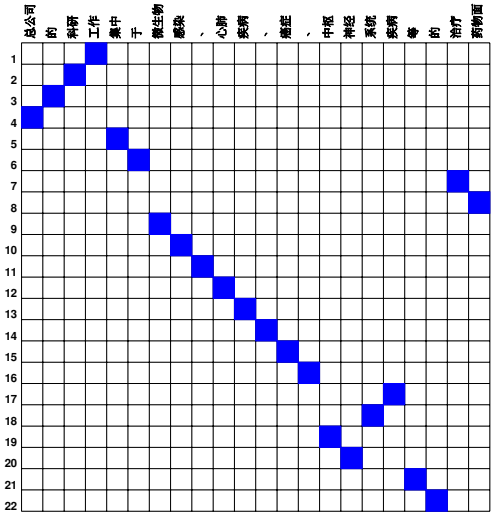
5.3 Permutation Distance Metrics

In human languages there is a certain amount of allowable variation in word order. It is difficult to judge automatically what is a good word order and what is not. However, we can be reasonably certain that the ordering of the reference sentence must be acceptable. We therefore compare the ordering of a translation, encoded in the permutation π, with that of the reference sentence, encoded in σ. The underlying assumption is that most reasonable word orderings should be fairly similar to the reference. This assumption is a necessary one. All automatic machine translation metrics assume that the translation should somehow be similar to the reference. We propose using *permutation distance metrics* to perform the comparison and calculate the difference between two sequences π and σ.

Permutation distance metrics have been used before in computational linguistics,



(a)



(b)

(4 • 3 • 2 • 1 • 5 6 • 9 10 11 12 13 14 15 16 • 19 20 • 18 • 17 • 21 22 • 7 8)

(c)

Figure 5.2: An aligned sentence (a) from the CH-EN Gale corpus together with its permutation (b) and (c), showing the source word positions ordered according to the aligned target words.

primarily for measuring the success of information ordering tasks. Ordering information is an essential step in applications such as text generation and summarisation. These tasks involve finding an acceptable ordering for items such as propositions (Karamanis, 2003), trees (Mellish et al., 1998) or sentences (Lapata, 2003). The success of these tasks is evaluated by comparing the ordering of the output to a gold standard ordering.

There are many different ways of measuring distance between two orderings, with different solutions originating in different domains (statistics, computer science, molecular biology, ...). Real numbered data leads to measures such as Euclidean distance and binary data to measures such as Hamming distance. But for ordered sets, there are many different options, and the best one depends on the task at hand. We choose two metrics which are widely used, efficient to calculate and capture the the number of elements which are out of order: the Hamming distance, and the Kendall's tau distance. See Deza and Huang (1998) for an in depth survey of metrics on permutations from a mathematical perspective.

We hypothesise that humans are sensitive to the number of words that are out of order in a sentence and both the metrics we use measure this. The Hamming distance is an absolute measure of the amount of disorder between two permutations, and the Kendall's tau distance is a measure of the relative disorder. Kendall's tau distance is also sensitive to how far words are out of order and this is something we would also like to capture, as it is reasonable to suppose that humans are sensitive to the size of reorderings as well as their frequency.

Our approach to measuring reordering performance is quantitative. We are measuring the amount of word order differences. Humans are also likely to be sensitive to the kinds of constituents that are reordered. Taking this into account however, would require sophisticated syntactic metrics of the kind discussed in the background chapter, in Section 2.3.2.4. The problem with these metrics is that they depend on rich source and target language information. This is particularly problematic if the model does not generate this information automatically. Extracting syntactic or semantic information can be difficult, especially when the quality of the translated sentence is poor. The advantage of using alignments is that they are an intrinsic part of the translation process, and can easily be produced along with the lexical tokens.

Another advantage of measuring reordering quality with distance metrics, is that the scores reported have an intrinsic meaning. However, an obvious disadvantage of this approach is reliable alignments are not available. If accuracy is paramount, test sets

with gold standard alignments can be used and translation models can output the actual word alignment used during translation. This approach was followed in Chapter 3. Alignments can also be generated automatically where gold standard alignments are not available. This approach was followed in Chapter 4.

We now describe the permutation distance metrics in more detail. Distance metrics decrease as the quality of translation increases, whereas many current machine translation metrics increase as the quality of translation increases. For ease of presentation, we would like all metrics to consistently increase with an increase in quality. We therefore subtract the distance metrics from one. Distance metrics are normalised to return distances between the values zero and one, although we report results as percentages. Comparing identical permutations thus return 0%, and completely inverted permutations return 100%.

5.3.1 Hamming Distance

The Hamming distance (Hamming, 1950) measures the number of disagreements between two permutations. The Hamming distance for permutations was proposed by Ronald (1998) and is also known as the *exact match distance*. It is defined as follows:

$$d_h(\pi,\sigma) = 1 - \frac{\sum_{i=1}^{n} x_i}{n}, x_i = \begin{cases} 0 \text{ if } \pi(i) = \sigma(i) \\ 1 \text{ otherwise} \end{cases}$$

where n is the length of the permutation. The Hamming distance will calculate the percentage of words in the translation which are in exactly the same order as in the reference sentence. The Hamming distance is the simplest permutation distance metric and is useful as a baseline. However, it has no concept of the relative ordering of words and this can lead to unintuitive scores. If all words are out of position by just one, the score will be zero. The Hamming distance is widely utilised in coding theory to measure the discrepancy between two binary sequences.

5.3.2 Kendall's Tau Distance

Kendall's tau distance is the minimum number of transpositions of two *adjacent* symbols necessary to transform one permutation into another (Kendall, 1938; Kendall and Gibbons, 1990). Kendall's tau seems particularly appropriate for measuring word order differences because it measures relative differences. It is sensitive to both the number and the size of the reorderings. Also, Kendall's tau distance is an intuitive measure of

strength of relationship between the permutations. It can be interpreted as a function of the probability of observing concordant and discordant pairs of elements (Kerridge, 1975). In other words it is the probability that two items are in the same order as opposed to in different orders, when comparing them between the two permutations π and σ.

For the case of translation, very few word order differences are completely inverted. Most word order differences are relatively small, and close to monotone. Because Kendall's tau is able to measure very large word order differences, this makes it rather insensitive to smaller reorderings. Therefore the range of values of Kendall's tau is too narrow for our purposes, with the majority of values bunched up close to 1. For this reason we take the square root of the standard metric and spread out the larger scores. This allows the metric to be more discerning of smaller word order differences and reflect more closely the human perception of word order quality. We show in experiments in Section 6.3.2.3 that the square root of Kendall's tau is more correlated with human judgements.

The Kendall's tau distance is thus defined as follows:

$$
\begin{aligned}
d_k(\pi,\sigma) &= 1-\sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}}{Z}} \\
\text{where } x_{ij} &= \begin{cases} 1 \text{ if } \pi(i)<\pi(j) \text{ and } \sigma(i)>\sigma(j) \\ 0 \text{ otherwise} \end{cases} \\
Z &= \frac{(n^2-n)}{2}
\end{aligned}
$$

Note that the distance metric range from 1, a perfect match, to 0 which indicates maximum disagreement. Normally a distance metric would use 0 to represent identical items, but we reverse the range so that our reordering metrics increase as the ordering matches the reference more closely. This makes it easier to compare results where our reordering distance metrics are presented next to machine translation metrics.

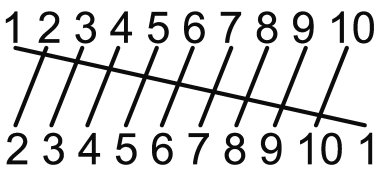


Figure 5.3: A visualisation of Kendall's tau distance

Figure 5.3 shows an example of two word orderings. The number of transpositions

can be calculated by counting the number of crossings. The example in Figure 5.3 shows nine crossings resulting in the following Kendall's tau distance:

$$d_k(\pi,\sigma) = 1 - \sqrt{\frac{9}{\frac{100-10}{2}}} = 1 - \sqrt{0.2} = 55.28\%$$

Without squaring, the score would be 80% instead of 55.28%. Considering the fact that there is a long distance reordering in this sentence, a score of 80% is perhaps too high.

In statistics, Kendall's tau rank correlation coefficient is a widely used non-parametric measures of association for two variables. Where the Kendall's tau distance metric counts the number of discordant pairs, the rank correlation measures the different between concordant and discordant pairs:

$$\tau = \frac{concordant - discordant}{Z}$$

As we are interested in the distance between two permutations, we use the distance metric formulation, but both measures essentially represent the same information.

In natural language processing research, Kendall's tau has been used as a means of estimating the distance between a system-generated and a human-generated gold-standard order for the sentence ordering task (Lapata, 2003, 2006). Kendall's tau has also been used in machine translation as a cost function in a reordering model (Eisner and Tromble, 2006). An MT metric called ROUGE-S (Lin and Och, 2004b) also measures the accuracy and precision of ordered pairs of words in the translation. This is similar to a Kendall's tau metric on lexical items. Our metric abstracts away from the words in the translation, and is a true measure of the word order similarity of a translation with a reference sentence.

In this thesis we have considered using two other distance metrics which measure relative ordering differences: the Ulam distance (Ulam, 1972) and Spearman's rank correlation (Diaconis and Graham, 1977). The Ulam distance between two permutations is the minimum number of single item movements required to transform one permutation into another. This metric does not take the distance a word is out of order into account and it did not correlate particularly strongly with human judgements. In statistics, Spearman's rank correlation is more widely used than Kendall's tau, and both metrics have the same sensitivity to detecting the existence of association. Spearman's

rank correlation has a number of disadvantages, however, such as the fact that it is a biased statistic (Kendall and Gibbons, 1990) which means that for smaller samples, it can often underestimate the strength of the correlation. Lapata (2006) argues that Kendall's tau distance is more appropriate for evaluating ordering tasks, and presents an overview of the differences.

5.4 Comparing Metric Properties

In the previous section, Section 5.3, we presented two permutation distance metrics for measuring reordering performance. These metrics have different properties and in this section we explore how appropriate they are for application to machine translation with the use of examples. We also compare them to commonly used machine translation metrics and to our previous metric, the RQuantity.

5.4.1 Baseline Metrics

For the rest of this thesis, we apply three metrics as our baselines: the BLEU score (Papineni et al., 2002); METEOR (Lavie and Agarwal, 2007); and TER (Snover et al., 2006). These metrics are described in detail in the background chapter, Chapter 2. BLEU and TER are shallow metrics, as they perform no deep linguistic analysis. Shallow metrics are of particular interest because the are reasonably language independent and fast to compute, and are therefore more generally applicable. METEOR uses optional stemming and synonym matching, but it is still fast to run and applicable to a variety of target languages. We select our baseline metrics because they are widely used and representative of the different kinds of metrics.

The BLEU score measures overlapping n-grams. METEOR is the harmonic mean of unigram precision and recall, and uses stemming and synonyms to allow for lexical variation. TER measures the number of edits required to change a system output into one of the references, and allows a block move edit. None of these metrics take the size of the word order differences into account and they all have parameters which are difficult to train, making the interpretation of the score more difficult. We adjust the TER metric by subtracting it from one in order for it to increase with an increase in translation quality, as all the other metrics do.

Example	Permutation
(a)	(1 2 3 4 5 6 7 8 9 10)
(b)	(1 2 3 4 • 6 • 5 • 7 8 9 10)
(c)	(6 7 8 9 10 • 1 2 3 4 5)
(d)	(2 3 4 5 6 7 8 9 10 • 1)
(e)	(2 • 1 • 4 • 3 • 6 • 5 • 8 • 7 • 10 • 9)
(f)	(4 • 3 • 2 • 1 • 5 6 • 9 10 11 12 13 14 15 16 • 19 20 • 18 • 17 • 21 22 • 7 8)
(g)	(4 • 3 • 2 • 1 • 5 6 7 8 9 10 11 12 13 14 15 16 • 19 20 • 18 • 17 • 21 22)

Table 5.1: Permutations representing a variety of characteristic reorderings. Most of these permutations correspond to the alignments shown in Figures 5.1. Example (f) correspods to the alignment in 5.2 and (g) is a new ordering which differs from (f) only in the positioning of items 7 and 8, shown with wavy underline.

Example	BLEU	METEOR	TER	d_h	d_k	d_k no sqrt
(a)	100.00	100.00	100.00	100.00	100.00	100.00
(b)	61.80	86.91	90.00	80.00	79.03	97.77
(c)	81.33	92.63	90.00	0.00	25.47	44.44
(d)	91.46	92.63	90.00	0.00	55.28	80.00
(e)	19.30	72.00	50.00	0.00	66.67	88.88
(f)	48.32	80.75	63.64	9.09	58.90	83.11
(g)	63.89	81.90	68.18	63.63	90.25	96.10

Table 5.2: Metric scores for permutations in the previous Table 5.1 calculated by comparing the disordered permutations to the monotone identity permutation (a).

5.4.2 Worked Examples

Table 5.1 contains a selection of permutations with a variety of characteristic reorderings. Previously we have shown permutations (a-d) together with their alignments in Figure 5.1 and (f) in Figure 5.2. The first five permutations are simple examples showing different ordering cases: (a) is a monotone ordering, (b) contains a small reordering where two words are swapped, (c) has a reordering where the two halves of the sentence are swapped, (d) is a reordering where the last source word is moved to the beginning of the target, and (e) is the case where there are many small word swaps. (f) is the real example sentence and (g) is a variation of that sentence ordering with one less long distance reordering. Table 5.2 presents the metric scores for the permutations

when compared to the monotone identity permutation.

We now discuss the scores which different metrics assign to the permutations and how they match our intuitions of what they should be measuring. We also compare distance metrics with each other and with other MT metrics. When calculating the MT metric scores, we assume that the words in the translation are in fact the numbers in the permutation. This means that all the "words" in the reference occur in the translation, just in a different order. Normally translations contain a great variety of words which do not match the reference, and so we are presenting the upper-bound of the metrics' performance.

Example (a) This permutation is identical to the monotone reference and so it has the highest scores for all metrics.

Example (b) When scoring (b) against the monotone, intuitively it should get a score very similar to (a) as it contains a very minor amount of disorder: just two words are swapped. As we have seen in the motivating examples in the introduction to this thesis, Section 1.1, BLEU, METEOR and TER fail to recognise that this is a small reordering and assign relatively poor scores to (b). In particular they score (b) with worse or equal scores than examples (c) and (d) which have much more reordering. All the reordering metrics correctly assign a high score to (b), much higher than examples (c) and (d). The Hamming and Kendall's tau distances are both reasonably sensitive to the small reordering. However, Kendall's tau with no square root gives a very high score to (b), one that could be problematic when trying to differentiate this permutation from a monotone permutation. This example illustrates the motivation for taking the square root of the standard Kendall's tau metric when applying it to reordering in machine translation.

Example (c) Example (c) is arguably the permutation with the most serious disorder as all words in the sentence have been moved by a long distance. The machine translation metrics give (c) a high score, not recognising the large amount of reordering present. The distance metrics are able to correctly measure the quantity of reordering, giving (c) the lowest (or joint lowest) score of all the permutations.

Example (d) This permutation is another case with a large amount of disorder, although most words have only moved by one position. Kendall's tau gives (d) a score which falls between the scores for (b) and (c), which is reasonable. The Hamming distance, however, measures absolute position and not relative posi-

tion. As all the elements are out of order it gives a score of zero, which is perhaps overly harsh. The Hamming distance should take into account that most of the relative word orders are the same as the reference, and only one word is out of order. The machine translation metrics give (d) a high score, failing to recognise the long distance reordering that has occurred. METEOR and TER give (d) the same score as they do to (c) which has, arguably, more disorder.

Example (e) In this permutation all the items are out of order, but only by a distance of 1. Only Kendall's tau distance is able to measure the real amount of reordering in this permutation. All other metrics penalise it heavily. Humans would probably give this kind of sentence a very low fluency score, but it is not unreasonable to suppose that they could still understand its meaning.

Example (f) This permutation is a non trivial example from the corpus. There is a lot of disorder, largely because of the long distance movement of target position (7 8) to the end of the sequence. The metrics can only be compared in relation to example (g).

Example (g) This permutation is the same as example (f), except (7 8) are no longer reordered to the end of the sequence. Unsurprisingly, all metrics score (g) better than (f). There is very little difference between the METEOR and the TER scores. BLEU and the distance metrics are able to easily distinguish the two permutations, giving (f) a much higher score than (g). Kendall's tau with no square root is much less sensitive to the difference than the Kendall's tau that we use d_k.

In real translation examples, there will be not only ordering differences, but also lexical differences to contend with. While the permutation distances are insensitive to lexical differences, the ability of MT metrics to detect word order differences are further hampered by differences in word choice. BLEU will consider every non-matching word to be a break, and so ordering differences will only be detected if they occur between words which are identical in the translation and the reference. METEOR will try to match synonyms and stems which leads to errors in the alignment. TER can account for differences in word order by using inserts and deletes. All commonly used MT metrics conflate the lexical and the ordering component of the measure, making it difficult to know what the actual reordering performance is.

5.4.3 Comparison with RQuantity

Until now we have used permutation distance metrics to measure the similarity of a permutation to another permutation, in order to evaluate the quality of the word order in a translation. By comparing a permutation to the identity permutation, we are also calculating the total amount of disorder in a sequence, as we do in the previous examples. This is very similar to the analysis we performed in Chapter 3, where we proposed a method for extracting reorderings from word aligned sentences. We then defined a metric for the amount of reordering in a sentence, the RQuantity.

The reason that we have proposed permutation distance metrics is that with the RQuantity we are unable to extract the distance between two word orderings where neither are monotone. Permutation distance metrics also handle non-binarizable reorderings naturally, as the orderings of interleaved items are taken into account.

Another difference with our reordering extraction method, is it takes both source and target dimensions into account, whereas with permutations, we are reducing all the target side properties to a simple ordering over source elements. The consequence of the two dimensional aspect of the algorithm, is that it was more sensitive to discontinuous word alignments than permutations are. With RQuantity, it was important to unalign determiners or large areas of the sentence could be blocked off and made unavailable for extracting reorderings.

5.5 Discussion

In this section we discuss a number of related approaches to measuring reordering and some considerations regarding our approach.

5.5.1 Related Work on Measuring Reordering

There have been a number of studies which have attempted to measure the complexity of the reorderings in sentences (Fox, 2002; Wellington et al., 2006; Galley et al., 2004). Much of this work has focused on the rank of the synchronous grammar rules or the size of the rules necessary to account for all reorderings seen the aligned sentence pairs in a corpus. These studies provide analyses which are tailored to a particular translation model and they are not widely applicable.

The permutation distance metrics return measures which are both generally useful, but they also handle all kinds of orderings, including the interleaved reorderings.

For example the typical non-binarizable reordering pattern of (1 • 3 • 4 • 2), see Figure 3.12, is easily compared to the monotone using the permutation distance metrics. Further analysis of permutations could lead to insight into the interleaved reorderings. A *permutation cycle* is a subset of a permutation whose elements trade places with one another. Cycles could be used to perform in depth analysis of more complex reordering patterns.

5.5.2 Permutations in Machine Translation

The ordering of words in a sentence can quite naturally be translated to a permutation. However, when you have a sentence pair with a complex many-to-many word alignment, a bijective permutation can fail to capture some of the real ordering dependencies between words. The fact that we only use the first alignment for a word, means that if subsequent alignments indicate that a reordering has occurred, we will have failed to identify this. In the case where there is a phrase with a gap such as "ne ... pas", we probably do not want to detect a reordering, as no inversion in order of the words has occurred. However, for the Chinese-English case where the determiner and the noun in English is aligned to the noun in the Chinese, because the determiner does not exist in Chinese, we would use only the alignment to the determiner, and the more important ordering of the noun is overlooked. Here it is possible that genuine differences in word order might be missed.

There has already been work which treats reordering as a permutation. Eisner and Tromble (2006); Tromble and Eisner (2009) propose a Linear Ordering Problem (LOP) model, which is capable of assigning a different score to every possible permutation of the source language sentence. It uses rich information about the source words and their relative positions to score different permutations. They describe ways to efficiently search over an exponentially large subset of sequence permutations using dynamic programming and apply this as a source reordering preprocessing step, before running the phrase-based model. Khalilov and Sima'an (2010) extend this work by introducing a tree-based reordering model which restricts the space of possible permutations by using tree contexts and limiting the permutations to data instances. Both these papers suggest an interesting approach to modelling reordering, however, in this thesis we focus on applying permutations as metrics, rather than using them to develop reordering models.

5.5.3 Reliance on Alignments

One of the drawbacks of the approach described in this chapter, is that we rely upon alignments and there is a scarcity of reliable gold standard alignments. Automatic alignments come with no guarantees about their accuracy or their lack of bias. This thesis presents a study which demonstrates that German-English gold standard alignments and German-English automatic alignments produce very similar reorderings (Section 4.3). The German-English language pair contains long distance reorderings so this is encouraging. However, this study cannot be reproduced on a large sample of language pairs due to shortages of gold standard alignments.

Although reliance on alignments is a drawback, almost all machine translation metrics have a similar problem. However, instead of source-target alignments, they create alignments directly between the translations and the references. These alignments are likely to be less accurate than alignments derived from standard alignment models, because metrics do not have the same resources to search for optimal alignments. The alignment models which generate the bilingual alignments are trained on large amounts of data and use highly refined search algorithms.

Our distance metrics could also be applied to the alignment between the translation and the reference sentence. These alignments could be generated using TER, for instance, and if TER alignments were shown to be reasonably accurate, this approach could work well. Even so, source-target alignments reflect the nature of the translation process and might be less ambiguous than alignments between two translations which were generated by two very different human and machine processes.

5.6 Summary

This chapter presents a method for measuring the quality of the word order in a translation. We compare the word order of the translation with a reference by using permutation distance metrics. We describe two different metrics: the Hamming distance, which measures absolute distance, and Kendall's tau distance, which measures relative distance. These are intuitive measures which are efficient, language independent and meaningful at a sentence level. These properties make them desirable machine translation metrics.

Using a variety of permutations representing reorderings with different properties, we compare the distance metrics to each other and to commonly used machine transla-

tion metrics. We show that the Hamming distance is an interesting baseline metric, but that it can be unreliable. It reports scores which are much lower than expected when a large number of words are shifted by only one or two positions.

Kendall's tau metric is indeed capturing the kind of information which we would like to see reflected in a reordering metric. Because it measures the size of large reorderings, it is insensitive to smaller reorderings. We take the square root of the standard Kendall's tau metric in order to create more variety in the scores. We show, using examples, that this leads to a metric which is sensitive to both smaller and larger reorderings.

The metrics proposed in this chapter work under the assumption that the number of words involved in a reordering and the distance that they move is relevant to the scores that word orders should receive. In the following chapter, Chapter 6, we will perform experiments to see whether this assumption is valid using experiments with human judges.

Chapter 6

Experiments with Reordering Metrics

6.1 Introduction

In the previous chapter, Chapter 5, we present novel reordering metrics which measure the quality of word order in translation. In this chapter, we show that these metrics correlate with human judgements of word order quality, and that current machine translation metrics are largely insensitive to the word order of the translation.

In order to establish the reliability of different metrics with regards to measuring reordering, it is necessary to collect human judgements on different word orders. We randomly permute test sentences in order to create several different orderings of each sentence, and we use these to extract human ratings. This allows us to control for sentence length, difficulty and domain. We then correlate metrics with human judgements to determine their ability to measure word order performance.

The rest of this chapter proceeds as follows. Section 6.2 presents an experiment which shows that our reordering metrics can distinguish human translations from machine translations. Section 6.3 presents a novel human evaluation task which isolates reordering and then correlates human judgements with current MT metrics and reordering metrics. Finally, in Section 6.4 we measure what percentage of variation of current MT metrics is due to lexical success and what percentage is due to reordering performance.

6.2 Distinguishing human and machine translations

Just as there are many different ways that the words in a sentence can be translated, there are also different ways to order them. A reordering metric must be able to dis-

tinguish between different word orderings. At the very least, they must be able to differentiate between the good orderings of human references and the often poor orderings of machine translations. The experiment we perform in this section is a sanity check. In a similar approach to Papineni et al. (2002), we verify that the permutation distance metrics return significantly better scores for human translations than for machine translations.

6.2.1 Experimental design

The goal of this experiment is to compare the metric scores for human translations with those of machine translations. We extract these scores on a standard test set which comprises of 1998 sentences from the GALE 2008 evaluation[1]. These are Chinese-English newswire sentences with four English reference sentences. We need a corpus with multiple references in order to extract scores comparing one reference to another.

We train a phrase-based model using MOSES (Koehn et al., 2007) on the full GALE 2008 Chinese-English training corpus. See Appendix A for details. With all the default options, we generate the translation output in English from the Chinese source sentence. We then word align the reference and the translated sentences to the Chinese source using the Berkeley word aligner (Liang et al., 2006) which was also trained on the full GALE 2008 training corpus. The Berkeley aligner has been shown to be more robust than using GIZA++ in situations where there are long sentences and sparse word counts (Koehn et al., 2008). This results in five sentence pairs and five alignments for each of the 1998 input test sentences. We then extract permutations for all alignments.

Each of the four references takes a turn as the gold standard. The metrics are applied, comparing the gold standard to the three other references and to the machine translation. We extract the set of scores for references and for translations and we compare them using a paired Wilcoxon signed-rank test (Wilcoxon, 1945). This test is a non-parametric statistical hypothesis test for the case of two paired samples. The Wilcoxon signed-rank test involves comparisons of differences between measurements and requires that the concepts “greater than”, “equal to” and “less than” are meaningful. The metrics scores consist of interval data and these concepts are thus applicable. The Wilcoxon signed-rank test is often used as an alternative to the paired Student’s t-test, when the distribution cannot be assumed to be normally distributed. For

[1] http://www.itl.nist.gov/iad/mig/tests/gale/2008/

a Wilcoxon signed-rank test the null hypothesis is that the central point would be expected to be zero, which indicates that there is no significant difference between the two variables.

We use the Kendall's tau distance and the Hamming distance as our two permutation distance metrics (See Section 5.3). They are scaled so that 0% is the worst score and 100% is the best possible score. We also compare them to the BLEU score and to extract meaningful BLEU scores at sentence level, we compute smoothed BLEU as described in Lin and Och (2004a).

6.2.2 Results

	Hamming	Kendall	BLEU
References	62.75	79.51	39.94
Translations	53.52	74.61	20.67

Table 6.1: The mean machine translations scores, compared to the mean scores for references.

We first report the metric scores which result from comparing the gold standard to another reference, or to the machine translation. Table 6.1 reports the mean metric scores. It shows that all metrics give higher scores to human references than to machine translations. BLEU shows the greatest difference in scores between translations and references. It is not surprising that BLEU is more sensitive to the differences between references and translations as it evaluates the words used as well as the word order, while the reordering metrics are guided purely by word order.

In order to visualise the behaviour of the two different sets of scores, Figure 6.1 contains the histogram plots of the distributions of metric scores. These plots show the percentage of sentences which give a certain range of scores. These plots also show that scores for the references are generally higher. Most noticeably, the number of sentences with maximum score of 100% has increased. Some references, especially for short sentences, are identical. Scores for translations are rarely 100%. The BLEU score almost never assigns 100% to a translation, which is probably desirable. The translations to which the Hamming and the Kendall's tau distance assign 100% to tend to be monotone translations which are compared to a monotone reference. The distributions of the scores also reveal interesting properties of the metrics. The Hamming distance scores are much more spread out than Kendall's tau and the BLEU score. Some of this

variation is due to the fact that small differences in order can lead to large differences in the Hamming distance score.

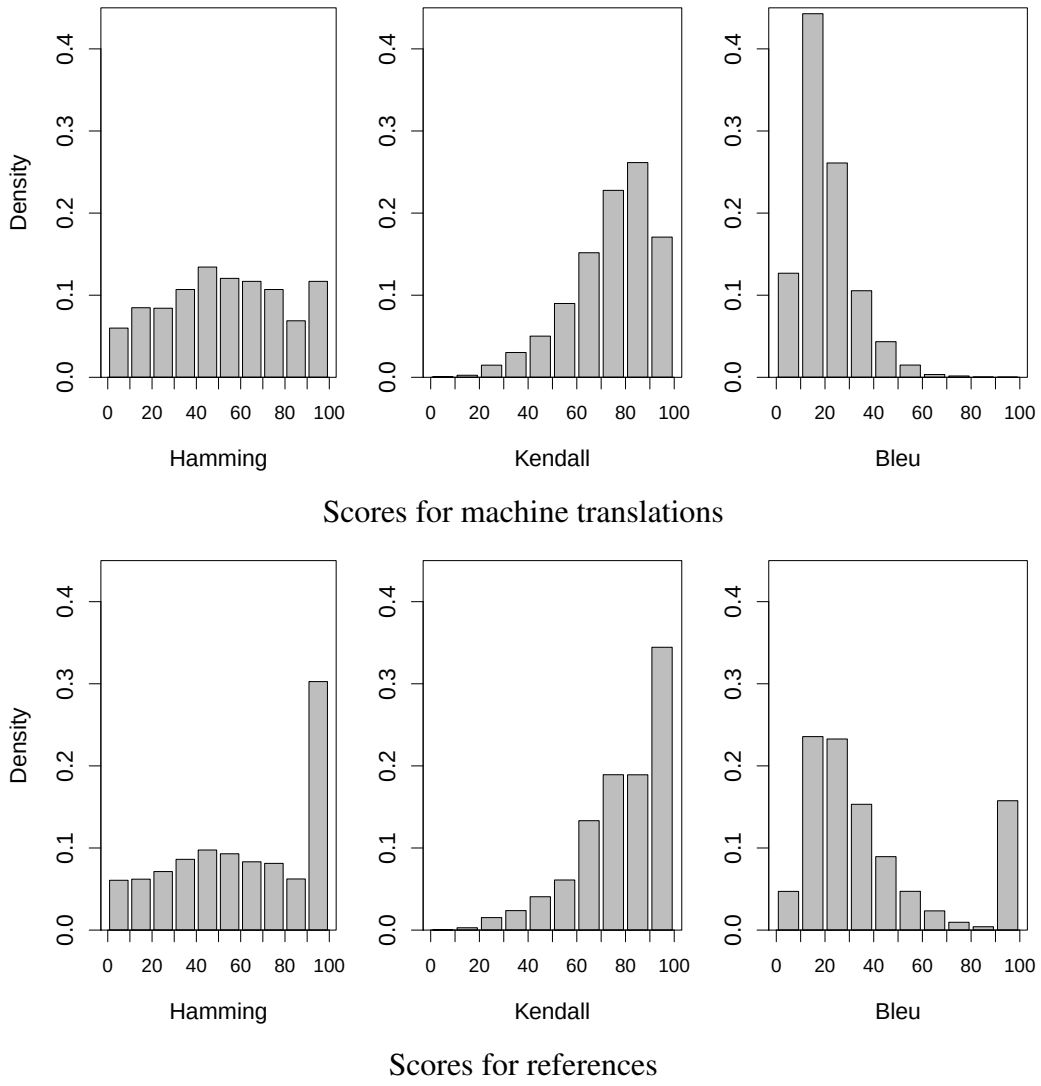


Figure 6.1: Distribution of sentence scores comparing a gold reference with either other references or with translations.

We want to show that the distance metrics are able to significantly distinguish the scores for the reference sentences from those of the machine translations. The significance is calculated by using the Wilcoxon signed-rank test. For each of the four gold references, we have metric scores which compare them to the three other references. We therefore have 12 sets of scores which are each compared to the set of scores for the machine translations. For each metric we perform 12 paired Wilcoxon signed-rank tests. For all metrics, for all tests, we can discard the null hypothesis that the scores for the references were not higher than the scores for the translations. The significance

levels of these experiments were all greater than 99.99%.

This experiment shows that the permutation distance metrics are capable of distinguishing between human and machine translation, and that these differences are in fact significant. In the next section, we explore the ability of reordering metrics to measure reordering performance.

6.3 Human evaluation of reordering

Machine translation metrics have been extensively applied to research on reordering. These metrics have been evaluated by comparing them to human judgements. However, these human judgements have not been shown to measure word order differences.

The most widely adopted methodology for humans evaluation of machine translation output is to assign values along a five-point scales for fluency and adequacy (LDC, 2005). Other popular human evaluation strategies have been to rank translations of a source sentence (Callison-Burch et al., 2007), or to perform post editing of the machine translation output (Callison-Burch et al., 2009; NIST, 2008). None of these human evaluation tasks have addressed the evaluation of reordering. The human scores are affected by lexical choice, sentence difficulty and sentence length. It is thus difficult to make any conclusions about the quality of word order in the translated sentences from these human evaluation experiments. Correlation with the judgements cannot be claimed to demonstrate that the metrics are able to measure the quality of word order.

In this section we design a novel experiment which isolates the effect of reordering on human judgements. We take a Chinese-English test set with human references and we artificially permute the English translation with different amounts of reordering. We thus control for all other confounding factors and we can say that this experiment does in fact specifically measure human judgements of word order quality. We then use this data to test the correlation of metrics with human judgements of reordering.

The rest of this section proceeds as follows. In Section 6.3.1 we describe our method for assembling a set of experimental materials and collecting human judgements. Then Section 6.3.2 reports the results of the experiments. We confirm that humans are able to reliably differentiate sentences with varying levels of reorderings and we show that permutation distance metrics do correlate with human judgements of reordering.

6.3.1 Experimental design

To assess whether automatic metrics correlate reliably with human evaluations, we need to design an experiment which gathers ratings on several different orderings of the same input. We then examine how well automatic metrics correlate with human judgements of reordering. A similar experiment for the information ordering task has been performed: humans judgements of comprehension for differently ordered sentences were collected (Lapata, 2006) and Kendall's tau was correlated with human judgements.

Data We use the Chinese-English parallel corpus that is provided by the GALE project[2] as it contains human annotated word alignments. Reorderings are extracted according to our reordering extraction algorithm, defined in Chapter 3. We calculate the amount of reordering, RQuantity, in a sentence by summing up the spans of the reorderings on the source sentence and normalised by the length of the source sentence. We randomly select 40 sentences which have a reasonably large amount of reordering (RQuantity > 1.3) and where the sentence length is between 10 and 40 words.

Baseline Metrics We compare the distance based metrics to three baseline metrics: BLEU, METEOR version 0.7, and TER version 0.7.25. These metrics are described in detail in Section 2.3.

Human Judgements During the study the participants were presented with a permuted sentence and asked to judge how fluent and comprehensible it was on a seven-point scale. We therefore collect data with a granularity which is informative without being unduly precise. The scalar scores do not assume a linear relationship between reordering amount and human fluency judgements. Ranking experiments would enforce a linear relationship, preventing us from distinguishing how much better or worse one reordering is from another.

Experimental Setup The study was conducted remotely over the Internet using Webexp [3] software. 28 unpaid volunteers were recruited by emailing the School of Informatics in the University of Edinburgh. They were all self-reported fluent speakers of English. Participants were instructed that some sentences would be perfectly understandable, and others would be scrambled and fairly incoherent.

[2] see LDC corpus LDC2006E93 version GALE-Y1Q4

[3] http://www.webexp.info/

They were shown examples of correctly and badly ordered sentences. Please see Appendix B for details of the instructions shown, and an example of one test case. From the set of test sentences we created five lists each consisting of different versions of the 40 sentences, following a Latin square design. Each pariticpant was randomly assigned one list which ensured that no user saw more than one ordering of the same sentence. In total $28 * 40 = 1120$ judgements were collected.

Extraction of Test Cases To asses whether our distance metrics reliably correlate with human judgements, we generate five different orderings of each reference sentence. This means that any preference shown by humans is based solely on word order differences. We start off by selecting the correctly ordered English reference sentence as our first test case. We then create the test case with the greatest amount of disorder. We do this by transforming the reference sentence so that the English words reflect the word order of the aligned Chinese sentence. We also generate three intermediate versions of the sentence. Each intermediate version falls into a bin with a different amount of reordering. The English word order of each intermediate version is the result of applying a random subset of the reorderings that were detected in the original Chinese-English sentence pair.

We choose to explore this particular space of possible word orders because it represents a wide range of humanly plausible reorderings. If we had explored the space of orderings that a translation system could produce, this would only represent a small and biased range of orderings. If, on the other hand, we had chosen to represent all possible word orderings, from inverted to monotone, this would represent a vast and totally implausible set of reorderings.

Illustrative Example Figure 6.2 shows an example sentence pair from the experiment. The original sentence pair is shown in (a). In (b) we see the shuffled test cases. In (c) we see the scores assigned by different metrics and the averaged human evaluations. The human results are averaged because we collect multiple ratings for each test case. Five test cases of this sentence were created and each participant only saw one of these cases. Bin Ref contains the original reference sentence and Bin 4 contains the reference reordered to reflect the word order of the aligned Chinese sentence. Test cases in Bins 1, 2, and 3 were created by applying a random subset of the Chinese-English reorderings shown in the alignment grid.

Let us take the test case in Bin 1. Here the scrambled version of the reference is created by applying a small reordering (with blocks that correspond to the English "in international competitions" and "before"). We do the same for the test case in Bin 3. Here we apply a large reordering (with blocks "has won many championships" and "in international competitions before").

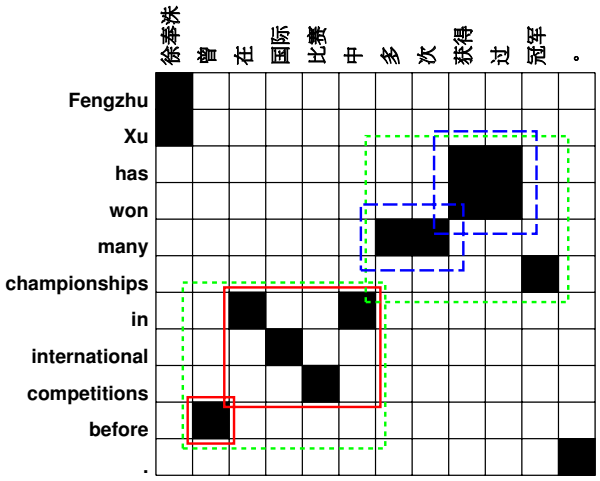
This example again demonstrates the problem with the current machine translation metrics. Humans give the test case in Bin 2 a score about one point higher than the test case in Bin 3, showing that they prefer the sentence with less reordering. Reordering metrics agree with humans and also give a higher score to Bin 2. BLEU, Meteor and TER however, give higher scores to Bin 3 because they are not sensitive to the size of the reordered chunks. Looking at Bin 4, we can see that it is completely garbled, but interestingly, the average humans judgement of comprehension is quite high. That might be because it is a short sentence and for simple sentences the meaning can be guessed at.

6.3.2 Results

The question which we address in these experiments is whether we can extract human judgements of reordering performance reliably, and then whether or not they correlate with a variety of translation metrics. In Section 6.3.2.1 we present the results of human judgements for test cases with varying amounts of reordering. In Section 6.3.2.2 we confirm that humans are able to reliably differentiate sentences with different levels of reorderings. Finally, in Section 6.3.2.3 we extract correlation statistics between automatic metrics and human judgements.

6.3.2.1 Human judgements of reordering scenarios

In order to develop automatic metrics of reordering performance we first need to establish that the amount of disorder in a sentence can be reliably detected by humans. In this experiment we examine the human judgements made on sentences with different levels of reordering. We analyse the fluency and comprehension judgements for different RQuantity bins. Table 6.2 presents the mean and standard deviation of the human judgements, for each of the RQuantity bins. We can see that humans are indeed sensitive to the amount of reordering. The higher the amount of reordering, the lower the fluency and comprehension scores are. Additionally, comprehension ratings



(a)

Bin	Sentence Test Cases with Permutations
Ref	Fengzhu Xu has won many championships in international competitions before .
	1 2 3 4 5 6 7 8 9 10 11
1	Fengzhu Xu has won many championships before in international competitions .
	1 2 3 4 5 6 • 10 • 7 8 9 • 11
2	Fengzhu Xu many has won championships before in international competitions .
	1 2 • 5 • 3 4 • 6 • 10 • 7 8 9 • 11
3	Fengzhu Xu in international competitions before has won many championships .
	1 2 • 7 8 9 10 • 3 4 5 6 • 11
4	Fengzhu Xu before in international competitions many has won championships .
	1 2 • 10 • 7 8 9 • 5 • 3 4 • 6 • 11

(b)

Bin	BLEU	METEOR	TER	d_h	d_k	Fluency	Comprehension
Ref	100.00	100.00	100.00	100.00	100.00	6.43	6.71
1	66.36	89.69	90.90	63.63	76.64	6.00	6.57
2	31.70	81.67	81.81	36.36	69.84	5.25	6.50
3	59.00	89.69	90.90	27.27	46.06	4.25	6.50
4	31.70	81.67	72.72	27.27	38.20	2.28	5.57

(c)

Figure 6.2: An example of a sentence pair used in the human evaluation campaign. The sentence pair in (a) is shown with the alignment and the reorderings, displayed with rectangles of different colours and line styles. Below the alignment in (b), the five differently ordered test versions of the reference sentence are displayed. Finally at the bottom in (c), a table with scores for the different test versions are presented, including metric scores and resulting average human judgements on fluency and comprehension.

are slightly higher and more variable than fluency ratings. This is because participants can sometimes understand a sentence even though it is somewhat scrambled. There is less variability amongst the best and worst bins, and slightly more variability in the intermediate bins. This is logical as it is easier to agree on an excellent or a terrible word order example than on an intermediate example.

Bins	Mean Fluency	Mean Comprehension
Ref	6.11 (1.16)	6.26 (1.21)
1	5.13 (1.80)	5.67 (1.64)
2	3.95 (1.65)	5.01 (1.69)
3	3.37 (1.59)	4.53 (1.72)
4	2.92 (1.42)	4.13 (1.64)

Table 6.2: The mean and standard deviation (in brackets) of human ratings for test items with different amounts of reordering, as shown in different bins. The Ref bin contains the reference sentences with RQuantity of 0. Bin 4 contains the most reordering and these test items have RQuantity of > 1.3.

We analysed the correspondence of human ratings with the RQuantity reordering bins, by performing an analysis of variance (ANOVA). Our ANOVA analysis had one factor, the reordering bin which can take one of 5 levels. The ANOVA showed that this factor was significant in both by-subject (F = 57.381, $p < 0.001$) and by-item (F = 24.49 , $p < 0.001$) analyses.

We use the Tukey's Honestly Significant Difference (HSD) test to determine if the ratings for sentences versions from different bins are all significantly different. Tukey's HSD compares all possible pairs of means and determines which means are significantly different from one another. It uses a similar distribution to the t-test, except that it corrects for the fact that the probability of making a type I error increases for multiple comparisons.

Tukey's HSD tests indicate that the ratings for sentences versions from different bins are all significantly different at the 99% level. The only exception is when comparing bins 2 and 3, where they are only significantly different at 95% level. We thus show that humans return low ratings of fluency and comprehension for sentences with large amounts of reordering, and conversely, that they return high ratings of fluency and comprehension for sentences with low amounts of reordering.

6.3.2.2 Reliability of human judgements

Not only must humans be able to detect differences in reordering, they must also show agreement with one another for the experiment to be useful. Inter-annotator agreement is also of interest because it acts as the upper-bound for agreement between human and automatic metrics.

To calculate inter-annotator agreement we use leave-one-out-resampling, which is a special case of n-fold cross-validation (Weiss and Kulikowski, 1991). For each set of judgements from a participant, we correlated their ratings with the averaged ratings of all the other subjects. We did this 28 times as we had 28 participants. In Table 6.3 we can see that the average human correlation quite high. This result contradicts previous research (Callison-Burch et al., 2007) which showed a low level of inter-annotator agreement for judgements on fluency and accuracy, which is a measure similar to comprehension. Callison-Burch et al. (2007) used the Kappa coefficient to calculate inter-annotator agreement, which is only applicable to categorical data. These judgements represent interval data and therefore the leave-one-out-resampling approach is more appropriate.

Another reason why we can rely upon the judgements, is that this experiment is controlled, as only the amount of word order differences vary. Previous human evaluations have rated machine translation output with confounding factors (such as word choice, sentence difficulty, sentence length, domain) all of which make the human evaluation task more unreliable.

	Fluency	Comprehension
Correlation	0.780 (0.112)	0.691 (0.106)

Table 6.3: The median and standard deviation (shown in brackets) inter-annotator agreement as calculated by leave-one-out-resampling with Pearsons correlation.

6.3.2.3 Correlation with permutation distance metrics

The ultimate goal of this experiment is to see if automatic metrics correlate with human judgements of the quality of word order in a sentence. We use correlation analysis to explore the linear relationship between human judgements and the metrics. This shows us if the metrics are indeed appropriate for evaluating reordering, and which metrics are best at capturing the reordering differences.

Metric	Fluency	Comprehension
BLEU	0.779	0.624
METEOR	0.802	0.638
TER	0.712	0.602
Hamming	**0.806**	**0.664**
Kendall's tau	0.795	0.657
Kendall's tau no sqrt	0.707	0.599

Table 6.4: The Pearsons correlation of metrics with human fluency and comprehension judgements averaged per test item.

In Table 6.4 we see the Pearson's correlation coefficients for the baseline and reordering metrics compared to the human fluency and comprehension ratings. As there are multiple human judgements collected for each item, the fluency and comprehension scores were averaged per test item. All the correlations are significant to the 99.9% level. We can see that in general correlation is strong, with the Hamming distance showing the highest correlation with human judgements for both fluency and comprehension. For fluency, the Hamming distance has a correlation coefficient of 0.802. This is slightly higher than the theoretical upper-bound, the inter-annotator agreement, which was 0.780. For comprehension, the Hamming distance has a correlation coefficient of 0.664 which is slightly lower than the inter-annotator agreement of 0.691. The strength of the correlation is generally lower for comprehension. This is explained by the fact that a sentence can be disfluent but can still be understood, which is hard to capture in an automatic metric. METEOR and Kendall's tau distance show correlations which are almost as strong as the Hamming distance.

We would like to test whether the correlation strengths are significantly different from another one. The two correlation coefficients are transformed with the Fisher Z-transform and the null hypothesis is that both samples of pairs show the same correlation strength. Performing significance tests between all pairs of correlation statistics for both fluency and comprehension, we find few significant differences. The only significant differences in correlation occur for fluency and are between the highest correlating metrics (METEOR, Hamming and Kendall's tau) with the lowest correlating metrics (TER and Kendall's tau with no square root).

This experiment uses sentences with perfect lexical overlap between the hypothesis and the reference, giving the baseline metrics, BLEU, METEOR and TER an unrealis-

tic advantage. When evaluating real translations, these metrics will be hampered by the fact that the words used in the translation are different to the reference. The reordering metrics are agnostic about the words used in the translation, as they abstract away from the words by using alignments. It is remarkable then, that the Hamming distance and the Kendall's tau distance show such strong correlation with human judgement in this experiment.

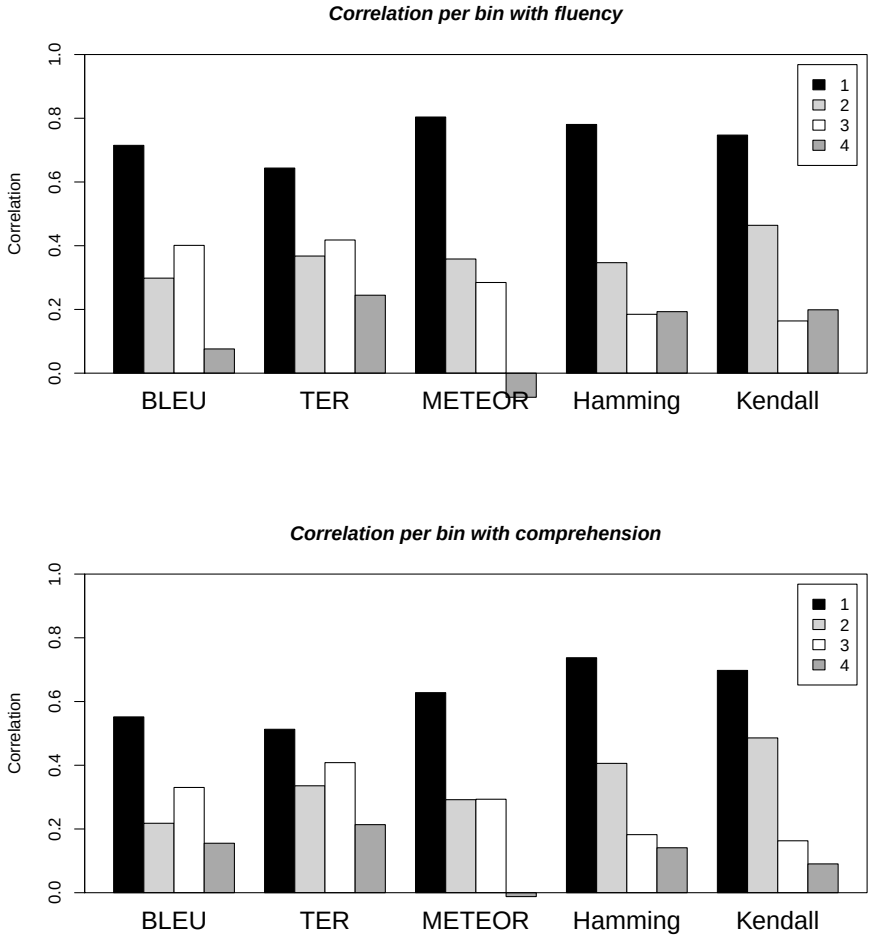


Figure 6.3: The Pearsons correlation of metrics with human fluency and comprehension judgements averaged per test item for each separate reordering bin.

In order to investigate the differences in correlation between metrics, we analyse the correlation of the metrics across the different reorderings bins. Figure 6.3 shows that the Hamming distance has the best overall correlation with human judgements of fluency and comprehension.

The strength of the Hamming distance is in fact somewhat surprising. It measures

the absolute amount of disorder in a sentence, but it does not measure how far out of order words are. The baseline metrics also do not measure how far out of order words are, and this makes us wonder if humans are perhaps more sensitive to the number of word order differences, than to the distance that they are moved. However, the difference between the correlation of the Hamming distance and the Kendall's tau distance is very small and not statistically significant. Long distance reordering will negatively affect the comprehension of a translation and translation models are not able to model them. Shorter distance word order differences would are more likely to be handled correctly by the translation models, either by reordering or by using different lexical items.

Kendall's tau distance is able to take the size of the reordering into account, which makes it more intuitive than either the baseline metrics or the Hamming distance. We also report the correlation coefficient of the Kendall's tau with no square root and this shows that it correlates much worse than the adjusted version which we suggest in the previous chapter, in Section 5.3.

METEOR also correlates very well with human judgement, but for large amounts of reordering it performs particularly badly, reaching negative correlation. TER does not correlate as well as METEOR but this seems only to affect the bin with the least amount of reordering. For the bin with the greatest amount of reordering, it performs better than all metrics, even the reordering metrics.

Figure 6.4 shows the relationship of current metrics to human judgements on fluency. All the plots show a group of points with the metric scores of 1. These are the reference sentences, which are assigned a variety of fluency scores by the participants. All metrics correlate very well with human judgements and it is not readily discernible where one metric is stronger than another. The difference between the plot of Kendall's tau and Kendall's tau without taking the square root, is that without the square root, the points are heavily clustered around the scores of 90-100% and when taking the square root, as we do for our Kendall's tau distance metric, the values are more spread.

6.4 Factors influencing machine translation metrics

In the previous experiments we used human judgements derived from an artificial experiment to evaluate the metrics. We have seen that under artificial test conditions where there is perfect lexical overlap between the reference and the translation, the machine translation metrics, BLEU, METEOR and TER, correlate reasonably well

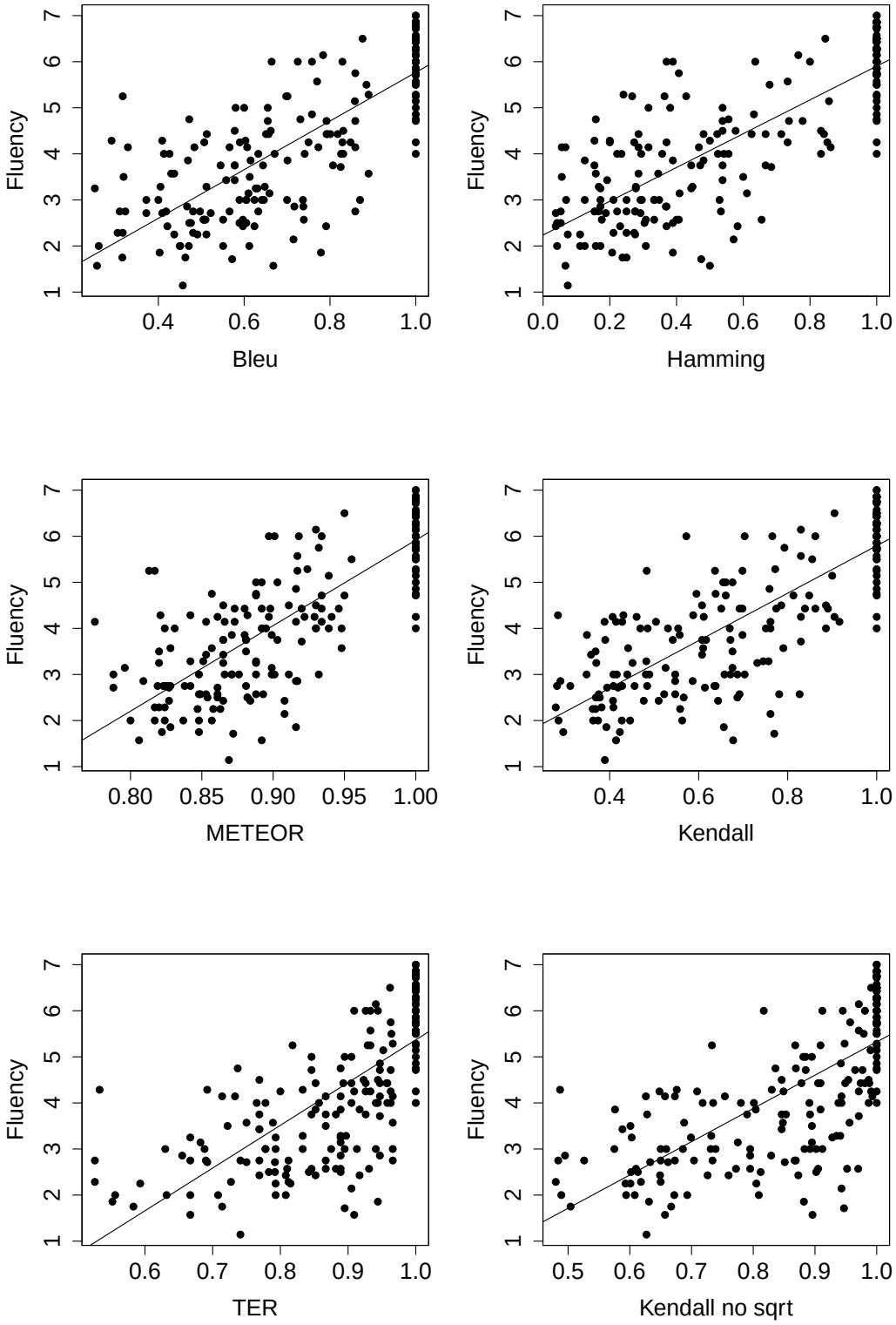


Figure 6.4: The averaged human fluency judgements for each sentence version compared to selected metrics.

with human judgements. However, these metrics are expected to perform much worse where there is lexical variation between reference and translation. We design an experiment to analyse what contribution lexical variation and word order performance have to the variability of the current machine translation metrics, under real test conditions.

6.4.1 Experimental design

We perform correlation analysis on our metric scores, comparing them with the amount of lexical overlap and the amount of reordering to see which factor affects them most. While the permutation distances are insensitive to lexical differences, the ability of MT metrics to detect word order differences is hampered by differences in word choice.

We used the 1-gram BLEU score, BLEU1, as our measure of lexical overlap. This is a precision score which takes into account multiple reference sentences and is defined as the number of matched words divided by the length of the translation. We have demonstrated that we are able to capture the reordering performance of sentences using the Kendalls tau distance, which measures relative word order and takes the size of reorderings into account. Multiple references are accounted for by measuring the distance to the reference with the closest word order.

The test data comprises of 1994 sentences from the GALE 2008 Chinese-English newswire test set which each have four English reference sentences, also used in Section 6.2. We also use the same translation model, training data and alignment model as described previously in Section 6.2.

6.4.2 Results

r			r^2		
Metric	BLEU1	Kendall's tau	Metric	BLEU1	Kendall's tau
BLEU	0.693	0.255	BLEU	0.481	0.065
METEOR	0.609	0.162	METEOR	0.371	0.026
TER	0.736	0.302	TER	0.543	0.091

Table 6.5: The Pearson's correlation r and the r^2 of lexical choice and reordering and current machine translation metrics. All regressions are significant to the 99.9% level

In this experiment we determine what influence lexical and reordering performance has on the MT metrics. Table 6.5 shows the Pearsons correlation of the metrics with

the BLEU1 score and the Kendall's tau distance metric. The results are all significant to the 99.9% level.

The correlation coefficients between the baseline metrics and the lexical metric are much larger than the correlations between the baseline metrics and the amount of reordering. The largest correlation of 0.736 is given by comparing TER and the lexical metric BLEU1. The largest correspondence between the amount of reordering, the Kendall's tau, and the MT metrics is also for the TER metric and it is 0.302. This is a weak, if significant, correlation. The results in this table show that the metrics are much more sensitive to the words used in translations than to their order.

Although the correlation coefficient r is a good indication of the strength of the relationship, taking it's square results in the r^2 which has an extremely useful interpretation. The r^2 allows us to describe the proportion of variability of the metric which is directly attributable to the variability in BLEU1 or Kendall's tau. The results for r^2 emphasise the fact that almost none of the variability in the metric scores can be attributed to reordering. The highest r^2 value for correspondence with Kendall's tau, is not even 10% for TER, and the lowest value is for METEOR, which is 2.6%.

Reordering seems to have a minimal effect on all of the metrics and we thus have evidence for one of the major claims made in this thesis. This can be visualised by looking at the plots in Figure 6.5 where the correlation between lexical overlap and machine translation metrics can clearly be seen, whereas the relationship between reordering and the metrics is minimal.

Finally, it is interesting to note that TER is more correlated with both lexical choice and reordering than the other two metrics. This indicates that TER is a more reliable measure of lexical and reordering success. METEOR is less correlated than the other two metrics, possibly due to errors introduced when matching stems and synonyms.

6.5 Summary

This chapter presents a number of experiments which justify using reordering metrics to evaluate word order quality in translations. First we show that our permutation distance metrics are able to distinguish between human and machine translations. Then we present a novel human evaluation experiment which specifically isolates the effect of word order differences. With this experiment we are able to show that humans are able to reliably discriminate between sentences with different levels of disorder. These judgements are then used to correlate with the MT and the reordering metrics. The

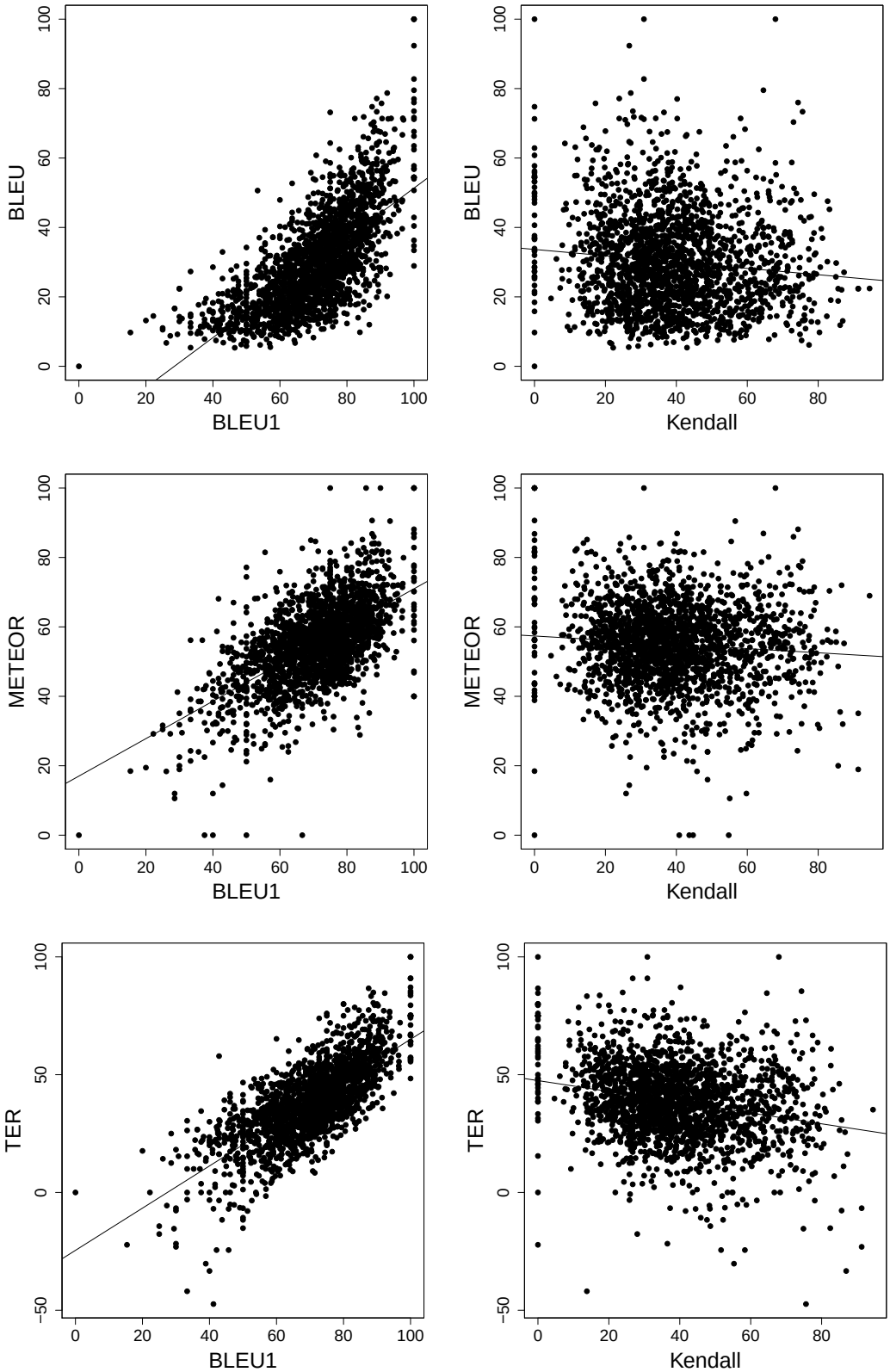


Figure 6.5: The lexical overlap and reordering amount plotted against MT metrics.

Hamming distance and the Kendall's tau distance are shown to correlate well with human judgements than the MT metrics, even in an artificial setting where the MT metrics have an unrealistic advantage.

Finally, we presented an experiment which shows that the current machine translation metrics are largely driven by the words used in the translation, and that they are quite insensitive to the order in which they appear.

Although reordering metrics which measure the quality of word order can be important for validating research aimed at improving reordering, they can never be considered a comprehensive metric as they only measure one aspect of translation quality. In the next chapter we propose a simple, intuitive combined lexical and reordering metric.

Chapter 7

LRscore: Combining Reordering and Lexical Metrics

7.1 Introduction

Research in machine translation has focused broadly on two main goals, improving word choice and improving word order in translation output and measuring the quality of these two aspects of translation is of fundamental importance. In the previous two chapters we proposed novel reordering metrics which we have shown correlate with human judgements of word order quality. We have also demonstrated that current metrics are relatively insensitive to word order quality. However, reordering metrics will always need to be used in conjunction with measures of the quality of word choice to be considered comprehensive metrics. In this chapter we present a novel metric, the Lexical Reordering score (LRscore), which explicitly combines a measure of lexical success with a reordering metric to provide a complete machine translation metric.

Apart from their inability to adequately measure reordering performance, a common criticism of current automatic MT metrics is that a particular score does not provide insight into quality (Przybocki et al., 2009) because they have no intrinsic significance. Ideally, the scores that the metrics report would be meaningful and stand on their own. For current MT metrics one can say that a higher score is better for accuracy metrics and a lower score is better for distance metrics, but it is very hard to extract any further insight. We argue that the LRscore is intuitive and meaningful because it is a simple, decomposable metric with only one parameter to train. The reordering component has an intrinsic significance. For the Hamming distance is represents absolute order and for Kendall's tau distance it represents relative order.

Ultimately, all automatic metrics need to be verified by correlating them with human judgements. We present experiments where human preference judgements are used to compare the LRscore with other existing metrics and we show that the LRscore is more consistent with humans judgements than other baseline metrics.

The rest of this chapter proceeds as follows. In Section 7.2 we start by describing the LRscore and its properties. In Section 7.3 we describe how to train the parameter of the metric using greedy hill-climbing. We also show that the LRscore is more consistent with human preference judgements than other commonly used MT metrics. In Section 7.4 we discuss the results and finally in Section 7.5 we summarise the main findings and contributions of the chapter.

7.2 LRscore

The main purpose of machine translation evaluation is to determine "to what extent the makers of a system have succeeded in mimicking the human translator" (Krauwer, 1993). Automatic evaluation assumes access to one or more reference translations created by humans. The task is to compare the system output with the references. However, unlike many natural language processing applications, machine translation has no unique "ground truth" as there are typically many possible correct translations. It is frequently impossible to judge automatically whether a translation is incorrect or simply unknown. It is even harder to judge how incorrect it is. Even so, automatic metrics are a necessary tool for developing machine translation systems. They allow developers to assess the impact of system modifications, and are critical for tuning statistical MT systems, for example in Minimum Error Rate Training (MERT).

There is a great deal of interest in developing automatic machine translation metrics. There have been a number of evaluation campaigns where metrics have been compared under different conditions, such as in the Workshops on Machine Translation (Callison-Burch et al., 2007, 2008, 2009) and the NIST Metrics for Machine Translation Challenge (MetricsMATR) (Przybocki et al., 2009). Although a large variety of metrics have been proposed, none of them specifically address the issue of reordering performance.

In this chapter we present the novel LRscore which includes a permutation distance metric which has been demonstrated to correlate strongly with human judgements of word order quality (see Section 6.3). It is a shallow metric which is quick to run and language independent. It is therefore an appealing metric for machine translation

researchers.

The LRscore is a linear interpolation of a reordering metric with a lexical metric, and each part of the score can be inspected independently if desired. Separating these two aspects of translation performance is somewhat simplistic, as word order affects word choice and vice-versa. However, decomposing a complex problem into two simpler, more manageable parts is an essential technique for solving scientific problems.

The LRscore is a weighted average of the reordering and lexical component and is defined as follows:

$$LRscore = \alpha * R + (1 - \alpha)L \tag{7.1}$$

The only weight present in the metric is α, which balances the contribution of the reordering metric, R, and the lexical metric, L. R is a permutation distance metric adjusted by the brevity penalty, and over a set of sentences S, it is calculated as follows:

$$R = \frac{\sum_{s \in S} d_s * BP_s}{|S|} \tag{7.2}$$

where d is the permutation distance score and BP is the brevity penalty. R is thus the average of the distance metrics, adjusted by the brevity penalty, over a set of sentences. In the following experiments d is either the Hamming distance or the Kendall's tau distance (see Section 5.3 for details).

The brevity penalty is calculated in the same manner as for the BLEU score:

$$BP = \begin{cases} 1 & \text{if } t > r \\ e^{1-r/t} & \text{if } t \leq r \end{cases} \tag{7.3}$$

where t is the length of the translation, and r is the length of the closest reference. If the reference sentence is slightly longer than the translation, then the brevity penalty will be a fraction somewhat smaller than 1. This has the effect of penalising translations that are shorter than the reference. The brevity penalty is necessary as the reordering metric provides the same score for a one word translation as it would a much longer monotone translation.

In these experiments, the lexical metric is the BLEU score, which is a product of the precisions of different n-gram lengths. We use two versions of the score: the 1-gram BLEU score, BLEU1, results in a lexical metric with no word order information; and the 4-gram BLEU score includes some measure of the local reordering success in the precision scores of the longer n-grams. BLEU is an important baseline, and improving on it by including a reordering metric is an interesting result.

Here we use the BLEU score, but the lexical component of the LRscore could be any metric which is meaningful for a particular target language. If a researcher was interested in morphologically rich languages, perhaps a metric which scores partially correct words would be more appropriate. We could, for example, apply METEOR which matches stems.

The LRscore returns both sentence level and system level scores. The only difference between the two is that the sentence level scores use smoothed BLEU (Lin and Och, 2004a), as BLEU is not stable at the sentence level.

The LRscore is not the first metric to be composed of a word choice and a word order component. Wong and Kit (2009, 2010) proposed the ATEC metric which also combines these two aspects of translation quality. ATEC is described as an F-measure which uses a matching function M to calculate precision and recall. M combines the number of matched words, weighted by their *tfidf* importance, with a measure of their position difference. The position difference score is the average difference of absolute and relative word positions and has no clear interpretation. ATEC also subtracts a score for unmatched words which undermines the interpretation of the supposed F-measure. The ATEC score is not intuitive nor easily decomposable. In fact it is more similar to METEOR than to the LRscore, because it mixes synonym and stem functionality with a reordering penalty.

7.3 Predicting Human Judgements

Even though the LRscore has many desirable properties, it must ultimately be judged on how well it correlates with with human judgements. This section explores how consistent the LRscore is with human judgements at the sentence and the system level.

In order to obtain optimal correlation with human judgement, the weight of the interpolation parameter must be set. We present experiments where we use a randomised hill climbing search for different language pairs, in order to train the LRscore.

Having to repeat this training for new language pairs requires access to human judgement data, which is not available for most test scenarios. We therefore investigate setting the parameter, based on the amount of reordering seen in the test set as a corpus with more reordering might require a higher weighting for the reordering component of the score. This is a novel approach to training a machine translation metric.

Language Pair	Pairwise Judgements	Ties
German-English	7,444	1,062
Spanish-English	4,808	702
French-English	7,772	1,504
Czech-English	3,150	899
English-German	7,351	788
English-Spanish	3,732	483
English-French	3,854	887
English-Czech	14,154	2,912
Total	52,265	9,237

Table 7.1: The number of human pairwise sentence rank judgements and the number of these judgements which were tied. They were collected in the 2009 Workshop on Machine Translation.

7.3.1 Experimental Design

Automatic metrics must be validated by correlating their scores with human judgements. We train the metric parameter to optimise consistency with human preference judgements across different language pairs and then we show that the LRscore is more consistent with humans than our baseline metrics.

7.3.1.1 Human Judgement Data

In the research community, there has recently been a lot of interest in developing automatic machine translation metrics and all metrics need to be validated by correlation with human judgements. However, the question of which is the best way of extracting human judgements, is still an open question. Various different human evaluation tasks have been evaluated for inter- and intra-annotator agreement, and ranking sentences was shown to be faster and more reliable than other human judgement tasks (Callison-Burch et al., 2007). Ranking has been chosen as the official determinant of translation quality for the 2009 Workshop on Machine Translation (Callison-Burch et al., 2009). We used human ranking data from this workshop to evaluate the LRscore.

Table 7.1 reports the number of pairwise ranking judgements for each language pair. The instructions provided to the annotators were: “Rank translations from Best to Worst relative to the other choices (ties are allowed).” Annotators were presented

Language Pair	No. Sentences	No. Words
German-English	1.58	41.15
Spanish-English	1.67	42.91
French-English	1.69	43.75
Czech-English	2.49	30.41

Table 7.2: The number of sentences and words (in millions) in the parallel corpora used for training the Berkeley alignment models.

with at most five translations at a time. Although there were more than five competing systems, there was no attempt to get a complete ordering over systems. The workshop organisers compiled a random selection and relied upon a reasonably large sample size to make the comparisons fair.

7.3.1.2 Alignments

Our reordering metric relies upon word alignments that are generated between source and reference sentences, and between source and translated sentences. In an ideal scenario, the translation system provides the actual alignments used to generate translations and the reference has gold standard human alignments. However, the human judgements have been collected for data which does not provide gold standard alignments, and we must resort to automatic alignments

We used version two of the Berkeley alignment model (Liang et al., 2006), with the posterior threshold set at 0.5. Our Spanish-, French- and German-English alignment models are trained using Europarl version 5 (Koehn, 2005). The Czech-English alignment model is trained on sections 0-2 of the Czech-English Parallel Corpus, version 0.9 (Bojar and Zabokrtsky, 2009). In Table 7.2 we can see the characteristics of the corpora used to train the alignment models.

7.3.1.3 Test Data

The metric scores are calculated for the test set from the 2009 workshop on machine translation. It comprises of 2525 sentences in English, French, German, Spanish and Czech. These sentences have been translated by different machine translation systems and the output submitted to the workshop. The system output along with human evaluations can be downloaded from the results section of the website of the Workshop on Machine Translation 2009. Participants used the training, development and test data

Language Pair	No. Systems	Kendall's tau
German-English	15	73.9
Spanish-English	9	80.5
French-English	14	80.4
Czech-English	3	81.1
English-German	11	73.9
English-Spanish	9	80.7
English-French	12	80.5
English-Czech	5	81.0

Table 7.3: The number of systems for which there are translations for each language pair. The average Kendall's tau reordering distance between the test and reference sentences is also reported.

provided by the workshop to train their particular translation system.

Table 7.3 reports the number of different translation systems which are provided by the workshop for download. The table also shows the amount of reordering that is present between the source and reference sentences. Remember that a higher score means less reordering. The amount of reordering for each language pair affects the importance of the reordering component of the score. The German-English language pairs have considerably more reordering than the other language pairs, because the Kendall's tau score is lower than for other language pairs.

7.3.1.4 System Level Correlation

Ultimately metrics are used to measure if one translation system is better than another. This is usually done over a test set consisting of a few thousand test sentences. When one score is reported for a whole test set, this is commonly called a system level score. It is useful to have a measure which can produce a meaningful system level score which correlates well with human judgements. System level correlations however suffer from having few data points and significant differences in metrics will be rare. Table 7.3 shows that the maximum number of data points is just 15 for German-English. We therefore use sentence level consistency, as described in the next section, as our main method of comparison.

To measure the correlation of the automatic metrics with the human judgements of translation quality at the system-level we use Spearmans rank correlation coefficient ρ.

We converted the raw scores assigned to each system into ranks. We follow Callison-Burch et al. (2009) in assigning an overall human ranking to each system based on the percent of time that their translations were judged to be better than or equal to the translations of any other system in the manual evaluation.

7.3.1.5 Sentence Level Consistency

Although we ultimately want a metric which outputs system level scores which correlate well with human judgements, a sentence level score is often more useful. Human judgements are not collected over whole collections of test sets, they are collected at the sentence level. There is a large amount of variability in human judgements between sentences and considering just one collective measure at the system level means that we loose a large amount of information. It is also interesting for researchers to have access to metrics which output sentence level scores in order to analyse translation output and determine the effect their changes have. One side effect, however, of looking at sentence level scores, is that shorter sentences are given the same importance as longer sentences. This is not necessarily undesirable, as the human judgements of shorter sentences are probably more reliable.

Utilizing sentence rank judgements is not as straightforward as using absolute scores of fluency and adequacy, for which correlation can be easily calculated. Lavie and Agarwal (2008) trained the parameters of the METEOR metric on rank data by calculating Spearman's rho correlation for the small number of rank judgements available for each sentence. They then take the average of the correlations across all sentences. This is an undesirable strategy because the correlations for small numbers of items are unreliable and correlation coefficients cannot simply be averaged as the correlation coefficient is not a linear function of the magnitude of the relation between the variables.

We therefore adopt the method used in the 2009 workshop on machine translation (Callison-Burch et al., 2009). We ascertain how consistent the automatic metrics are with human judgements by examining each pairwise comparison of translation output for single sentences by a particular judge. We then record whether or not the metrics are consistent with the human ranking (i.e. we counted cases where both the metric and the human judge agreed that one system is better than another). We divided this by the total number of pairwise comparisons to get a percentage which we call the consistency of a metric. There were many ties in the human data, but metrics rarely give the same score to two different translations. We therefore excluded pairs that the

human annotators ranked as ties.

It is important to be able to determine when a difference in consistency scores between two metrics represents a significant difference in their performance. Koehn (2004b) describes a method to compute statistical confidence intervals for automatic metrics using bootstrap resampling (Efron and Gong, 1983). Bootstrapping is a statistical technique for estimating the sampling distribution of a variable by sampling with replacement (i.e. allowing repetition of the values) from the original sample. The method has the practical advantage of being easy to implement and the theoretical advantage of not presupposing anything about the underlying distribution of the variable. A simple programming routine can calculate the estimators of the mean, variance, etc., of any random variable distribution. We use bootstrap resampling to estimate the 95%confidence intervals of the consistency of metrics with human judgements. If the intervals for different metrics do not overlap, we can say that one metric is significantly more consistent than another.

Given the consistency result of m, we would like to compute with a confidence q that the true consistency score lies in an interval $[a,b]$. We draw a test set from the space of all possible test cases, and we then calculate consistency. We do this for a large number test sets, and we sort the corresponding consistency scores. We drop the top 2.5% and the bottom 2.5% of the scores, and this leaves us with the remaining scores within an interval $[a,b]$. Our overall consistency score m is the mean of all the samples. The law of large numbers dictates, that with an increasingly large number of samples, the interval $[a,b]$ approaches the 95% confidence interval. We do not have access to the space of all possible test cases, and so we assume that estimating the confidence interval from a large number of test sets with n test cases drawn from a set of n test cases with replacement is as good as sampling n test cases from an infinite set of test cases.

7.3.1.6 Baseline Metrics

In order to evaluate the LRscore, it must be compared to our baseline metrics, BLEU, METEOR and TER. The BLEU score has five parameters, one for each n-gram, and one for the brevity penalty. These parameters are set to a default uniform value as is standard. When results are reported for system level scores, the BLEU score is used. When results are reported for sentence level scores, the smoothed BLEU score is used.

METEOR has 3 parameters which have been trained twice, once for human judgements of adequacy and fluency (Lavie and Agarwal, 2007) and once for human judge-

Metric Name	Reordering Metric	Lexical Metric
LR-HB1	Hamming	BLEU1
LR-HB4	Hamming	smoothed BLEU
LR-KB1	Kendall	BLEU1
LR-KB4	Kendall	smoothed BLEU

Table 7.4: The test conditions for the LRscore

ments of rank (Lavie and Agarwal, 2008). METEOR version 0.7 was used. The parameters optimised for adequacy and fluency have been used, these were applied separately for each of target languages. For English as the target language the exact match, porter stem and synonymy modules were used. For Czech as the target language the exact match module was used, and for the rest of the languages we used exact match and porter stem.

The other baseline metric used was TER version 0.7.25. As in previous chapters, we adapt TER by subtracting it from one, so that all metric increases mean an improvement in the translation. The TER metric has five parameters which have not been trained.

We test the LRscore with two reordering metrics, the Hamming distance and Kendall's tau distance. We also apply two lexical metrics, the 1-gram BLEU score, BLEU1, and the standard 4-gram BLEU score with uniform weight. See Table 7.4 for the breakdown of the LRscore variations.

7.3.1.7 Optimisation of Metrics

Automatic metrics of translation all have different components which are combined to form a complete metric. Training metric parameters is difficult as we rely upon human evaluation data, and many metrics either perform no optimisation of their parameters, or are optimised only once for a particular language pair and domain. Even the metrics which have been trained for a particular target language are not necessarily optimal for other language pairs or domains. For example, if the metric is trained on Arabic-English data, where there is little reordering, the word order component might receive a lower weight than it would for another language pair with more reordering.

Our first approach to optimising the LRscore is to train the parameter separately for each of the eight language pairs. We use greedy hill climbing in order to find the optimal setting. We optimise for sentence level consistency of the metric. As hill

climbing can end in a local minima, we perform 20 random restarts, and retain only the parameter value with the best consistency result. Random-restart hill climbing is a surprisingly effective algorithm in many cases. A reasonably good local maxima can often be found with a relatively small number of restarts (Russell et al., 1995).

The brevity penalty applies to both the reordering metric and the BLEU score. We do not set a parameter to regulate the impact of the brevity penalty, as we want to retain BLEU scores that are comparable with BLEU scores computed in published research.

7.3.1.8 Optimisation Across Language Pairs

There is very little human evaluation data available for training metrics and it is time consuming to train metric parameters for each new data set. It is therefore desirable to be able to set the metric parameters by simply calculating some characteristic of the language pair. The LRscore is simple and requires setting only one parameter which balances reordering and lexical metrics. It is logical to suppose that this parameter depends to a large degree on the importance of reordering in the language pair in question. A language pair with little or no reordering will have little use for a metric which measures this.

We propose a novel method for setting the metric parameter. First we train a language independent parameter which is then adjusted by the amount of reordering that exists in the test set. In order to apply the LRscore, the test set has to have been aligned to the reference sentences, and so extracting the amount of reordering with the LRscore is quick and simple. Researchers using our metric will thus be able to determine the reordering amount for each language pair and domain they wish to test with very little extra effort. The amount of reordering is calculated as the Kendall's tau distance between the source and the reference sentences as compared to dummy monotone sentences. The language independent parameter (θ) is adjusted by applying the reordering amount (d_k) as an exponent. This works in a similar way to the brevity penalty. With more reordering, the d_k becomes smaller. This leads to an increase of the final weight of α, which represents the percentage contribution of the reordering component in the LRscore:

$$\alpha = \theta^{d_k} \tag{7.4}$$

The language independent parameter θ is trained once, over multiple language pairs. This procedure optimises the average of the consistency results across the differ-

ent language pairs. The validity of this approach can be demonstrated by comparing the optimised consistency results obtained here, with those obtained when training is performed for each language pair individually. Once θ is trained on a particular set of language pairs, we then use it for new language pairs for which we have calculated the d_k. Thus, α can be set easily for any new language pair or domain.

7.3.2 Results

In the following experiments we aim to:

- Optimise the parameter the LRscore metric with respect to human judgements of rank.
- Compare the consistency of the LRscore with baseline metrics and show that the LRscore corresponds better with human judgement.
- Explore which combination of lexical and reordering components in the LRscore is more consistent with human judgements.
- Find the system level correlation of the LRscore with human judgements.
- Optimise the metric parameter using characteristics of the language pair instead of needing to train with human judgements for each test case.

7.3.2.1 Sentence Level Consistency

This experiment performs randomised hill-climbing for each of the language pairs in order to optimise the LRscore's sentence level consistency with human judgements. Once optimised, the consistency of the LRscore is compared with that of the baseline metrics.

Table 7.5 reports the optimal consistency of the LRscore and baseline metrics with human judgements for each language pair. The table also reports the results for the individual components of the LRscore in isolation.

The first thing to note in Table 7.5 is that, apart from Czech-English, the LRscore is the metric which is most consistent with human judgement. This is an important result which shows that combining lexical and reordering information makes for a stronger metric. The language pairs with the most reordering are the German-English and English-German pairs (as shown Table 7.3) and for these language pairs it seems that

Metric	de-en	es-en	fr-en	cz-en	en-de	en-es	en-fr	en-cz	ave
METEOR	58.6	58.3	58.3	**59.4**	52.6	55.7	61.23	55.6	57.5
TER	53.2	50.1	52.6	47.5	48.6	49.6	58.3	45.8	50.7
BLEU1	56.1	57.0	56.7	52.5	52.1	54.2	62.3	53.3	55.6
BLEU	58.7	55.5	57.7	57.2	54.1	56.7	**63.7**	53.1	57.1
Hamming	51.1	42.6	38.7	36.4	42.4	38.3	47.4	35.5	41.5
Kendall	53.2	44.6	40.5	42.1	45.6	39.2	49.2	37.3	44.0
LR-HB1	60.4	**60.6**	58.6	53.7	54.8	55.8	63.9	55.0	57.8
LR-HB4	60.5	58.9	**58.8**	57.7	55.0	**57.5**	**63.7**	55.1	58.4
LR-KB1	60.7	58.5	58.5	54.2	54.7	55.6	62.3	55.1	57.5
LR-KB4	**61.1**	59.9	58.6	58.9	**55.2**	57.4	**63.7**	**55.3**	**58.7**

Table 7.5: The percentage consistency between human judgements of rank and metrics. The LRscore variations (LR-*) are optimised for consistency for each language pair. A random baseline metric would get a 50% consistency score.

LR-KB4 is the best metric. This suggests that the Kendall's tau metric is more appropriate for language pairs with a reasonable amount of reordering. After the LRscore, METEOR shows the highest consistency, however for German-English and English-German, METEOR lags behind the BLEU score, suggesting that it is less appropriate for language pairs which contain a lot of reordering. The TER score shows the lowest consistency of all the complete metrics and it might be hampered by lack of tuning to the data set. The reordering metrics in isolation are clearly deficient. As reordering metrics were never intended to be used in isolation, their poor correlation is not a concern.

LR-HB1 is meant as a baseline metric, but it performs best for the Spanish-English language pair. This suggests that for this language pair, the longer n-grams are not important for human judgements of rank. However, for most other language pairs, using the full BLEU score does improve correspondence. Both LR-HB4 and LR-KB4 perform very well in this experiment, but LR-KB4 performs best for 5 language pairs, as opposed to LR-HB4 performing best for only 3 language pairs. Also, LR-KB4 performs best for the language pairs with the largest amount of reordering (those into and out of German) and we therefore select this as our preferred metric.

In order to judge the significance of these results, in Figure 7.1 we show 95% confidence interval for the consistency scores, extracted using bootstrap resampling. We

Metric	de-en	es-en	fr-en	cz-en	en-de	en-es	en-fr	en-cz
LR-HB1	36.44	22.42	19.45	06.60	18.63	14.67	21.34	52.13
LR-HB4	09.28	80.15	04.29	00.88	06.82	03.38	00.19	44.93
LR-KB1	47.98	29.21	05.94	07.44	22.67	16.90	51.49	59.56
LR-KB4	22.41	81.03	18.85	19.97	12.01	00.62	00.19	44.51

Table 7.6: The optimal parameter setting for each language pair and direction when trained with randomised hill climbing. This parameter refers to the percentage contribution of the reordering component in the linearly interpolated LRscore.

can see that the mean consistency for the LRscores is greater than that of the baseline metrics. However, the confidence interval of these results overlap significantly for METEOR and BLEU. This means that we cannot assert that the LRscore is significantly more consistent with human judgements than either BLEU or METEOR. Even without significance, the higher consistency of the LRscore and its ability to capture reordering make it an attractive choice for evaluating machine translation research.

In Table 7.6 we report the optimal parameter setting of α for the LRscore for each language pair and LRscore test condition. The parameter refers to the percentage contribution of the reordering component of the linearly interpolated LRscore. There is a great deal of variation between the settings. The largest setting of α is 81.03 for Spanish-English for LR-KB4 and the smallest is 00.19 for English-French LR-HB4. However, while training α we noticed that the range of values for consistency is quite narrow, varying from about 55% to the results seen in Table 7.5. We also noticed that these parameter setting are quite stable on random restarts. Looking at German-English, which has more reordering than other language pairs, it seems that the reordering component contributes more when the lexical component includes no notion of reordering, when the 1-gram BLEU score is used, as expected. When the 4-gram BLEU score is used, the reordering component is weighted less. Also, comparing the Hamming distance and the Kendall's tau distance metrics, it seems that for German-English, Kendall's tau is preferred. For the languages translating into English, the results are more mixed. Translating out of English, the contribution of reordering is slightly lower. For English-French and English-Spanish when using 4-gram BLEU, the contribution of reordering is close to zero. Perhaps languages with more morphology need to place a higher emphasis on the lexical component of the metric.

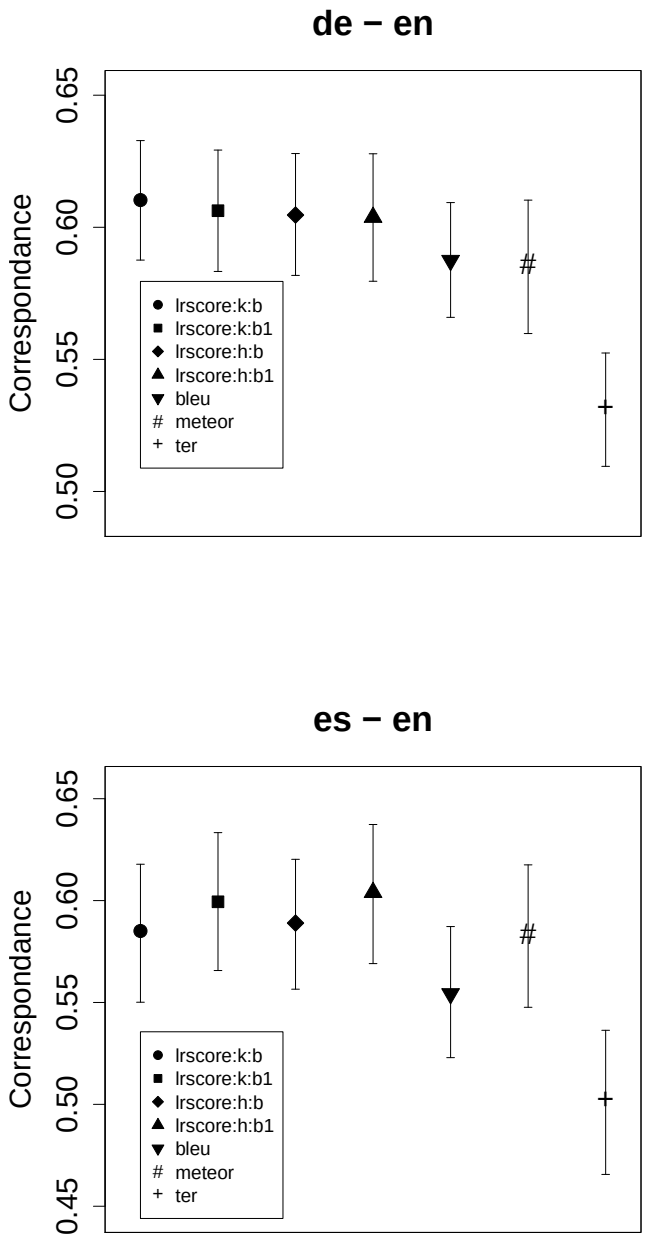


Figure 7.1: The mean consistency of metrics with their 95% confidence intervals extracted via bootstrap resampling.

Rank	Human	LR-HB4	LR-KB4	METEOR	TER	BLEU
1	rbmt2 (29)	google	google	google	umd	google
2	google (26)	uka	uka	uka	google	uka
3	rbmt3 (24)	umd	uedin	stuttgt	uka	umd
4	systran (22)	uedin	umd	uedin	rwth	uedin
5	uka (21)	stuttgt	stuttgt	systran	stuttgt	stuttgt
6	umd (21)	liu	liu	umd	uedin	liu
7	uedin (20)	rwth	systran	rbmt3	systran	rwth
8	rbmt4 (19)	systran	rwth	rbmt2	liu	systran
9	rbmt1 (14)	rbmt3	rbmt3	liu	rbmt3	rbmt3
10	stuttgt (14)	usaar	usaar	usaar	usaar	rbmt2
11	rwth (10)	rbmt2	rbmt4	rbmt1	rbmt2	usaar
12	usaar (9)	rbmt4	rbmt2	rwth	rbmt4	rbmt4
13	liu (8)	rbmt1	rbmt1	rbmt4	rbmt1	rbmt1
14	geneva (4)	geneva	geneva	geneva	geneva	geneva
15	jhu (3)	jhu	jhu	jhu	jhu	jhu

Table 7.7: The German-English translation systems ranked in order of preference for human judgements and for the automatic metrics. The human ranks are calculated by counting the number of sentences which are judged as best or tied as best for a particular machine translation system. This count is reported in brackets along with the human ranked systems.

7.3.2.2 System Level Correlation

The most common method of applying an MT metric is to compare the performance of two systems on a particular test set. This motivates the following experiment where system level correlations of metrics with human judgements are presented. In Table 7.7 we can see the ranking of the different German-English machine translation systems. The human ranks are based on the number of times that humans judged their translations to be better than or equal to the translations of any other system. These counts are shown in brackets next to the human ranks. The different automatic metrics are in broad agreement. They all disfavour the commercial rule based machine translation systems of "rbmt" and "systrans" which are highly regarded by humans. It seems that all automatic metrics struggle to mimic human preferences.

In Table 7.8 we report the system level Spearman's rho correlation between the hu-

man ranking and the metric ranking. The number of systems that are used to calculate the correlation were reported in Table 7.3.

Table 7.8 shows that the correlation of the LRscore metrics are comparable to the BLEU score correlation. Few of these correlations are statistically significant, because there are relatively few systems to be ranked, with the largest number being 15 systems for German-English. The Czech-English and English-Czech language pair only has three and five systems respectively. This makes is hard to gain any useful insight into the performance of the metrics. Furthermore, looking at the number of human judgements used to create the human ranking in Table 7.7, we see that the number of times that systems are judged as the best are quite small. This brings into question the value of this type of evaluation.

Even so, looking at the correlation data in Table 7.8, it seems that the LRscore correlates reasonably well with human data when compared to the BLEU score and with METEOR. In fact the average correlation of LR-HB4 and LR-KB4 is higher than that of the BLEU score, and only slightly lower than that of the METEOR metric. The results for the LRscore variations which use BLEU1, and therefore rely entirely on the reordering component of the metric for evaluating word order, are much better than those of BLEU1 and not much worse than BLEU, which shows that the reordering component is correctly contributing information on word order quality. In fact if you look at the reordering metrics in isolation, they seem correlate worse than all other metrics for French-English and Spanish-English, but they also correlate better than other metrics for English-German and English-Spanish. In fact for English-German, the reordering metrics are the only ones that are positively correlated with human judgements. This is likely to be partly due to the randomness of the small number of systems compared, but the reordering metrics could be contributing useful knowledge which is distinct from the information available to the other metrics.

7.3.2.3 Optimising across Language Pairs

It is time consuming and costly to optimise metric parameters, especially when there is no human evaluation data for a particular language pair or domain. We have proposed setting this parameter automatically based on the amount of reordering in the test set. This experiment aims to determining whether our approach is valid by comparing the consistency results obtained when optimising for each language pair, with the consistency results when optimising the language independent parameter θ over multiple language pairs.

Metric	de-en	es-en	fr-en	cz-en	en-de	en-es	en-fr	en-cz	ave
MET.	0.67	0.70*	0.66*	1.00	-0.25	0.23	0.88***	0.60	0.56
TER	0.44	0.41	0.50	1.00	-0.51	0.35	0.74**	0.10	0.38
BLEU1	0.46	0.40	0.58*	1.00	-0.45	0.38	0.87***	0.20	0.43
BLEU	0.49	0.60	0.65*	1.00	-0.29	0.37	0.86***	0.50	0.52
Hamming	0.45	0.33	0.11	-0.50	0.48	0.82*	0.78**	0.40	0.36
Kendall	0.25	0.05	-0.16	-0.50	0.78**	0.52	0.62*	0.70	0.28
LR-HB1	0.39	0.58	0.58*	1.00	-0.40	0.38	0.89***	0.60	0.50
LR-HB4	0.45	0.61	0.66*	1.00	-0.22	0.37	0.86***	0.70	0.55
LR-KB1	0.46	0.46	0.58*	1.00	-0.45	0.38	0.88***	0.70	0.50
LR-KB4	0.45	0.33	0.61*	1.00	-0.13	0.37	0.86***	0.70	0.52

Table 7.8: Spearman's rho correlation for system level evaluation of metrics with human judgements of the best or tied best translation.

Metric	θ	de-en $\alpha = \theta^{d_k}$
LR-HB1	13.32	22.54
LR-HB4	01.86	05.26
LR-KB1	28.20	39.24
LR-KB4	13.19	22.38

Table 7.9: The language independent parameter θ for each LRscore test condition, and the final parameter α for the German-English task after applying the reordering amount d_k of 0.739 to θ.

We perform a randomised hill climbing search for the best setting of θ. At each step instead of calculating the consistency for only one language pair, we calculate it for all language pairs and take the average. For this experiment, θ is adjusted for each language pair by applying as an exponent the Kendall's tau reordering amount shown in Table 7.3.

In Table 7.9 we can see the optimised language independent parameter θ for each LRscore setting. This is used to calculate α and α is then used to calculate the consistency of each metric for each language pair. Table 7.9 shows that the contribution of the reordering component is small for LR-HB4, but for the rest of the metrics, it is more important. The final parameter α is higher than the value in this table. We also provide the final α value for German-English, where the d_k of 0.739 was applied as an

Metric	de-en	es-en	fr-en	cz-en	en-de	en-es	en-fr	en-cz	ave
LR-HB1	59.7	60.0	58.6	53.2	54.6	55.5	63.7	54.5	57.5
LR-HB4	60.4	57.3	58.7	57.2	54.8	57.3	63.3	53.8	57.9
LR-KB1	60.4	59.7	57.9	54.0	54.1	54.7	63.4	54.9	57.5
LR-KB4	61.0	57.2	58.5	58.6	54.8	56.8	63.1	54.9	58.7

Table 7.10: The result of using a parameter setting based on language pair characteristics.

exponent. This Kendall's tau value was extracted from Table 7.3.

Table 7.10 reports the consistency for each language pair when using the language independent parameter. The average consistency is also reported, and this is value which is optimised. The results in this table should be compared with Table 7.5. This comparison shows that the consistency figures are only very slightly lower when training across language pairs. This leads us to conclude that we can reliably use the language independent parameter together with the amount of reordering in the test set to configure the LRscore for new language pairs and domains.

7.4 Discussion

In the previous experiments we have shown that the LRscore is consistent with human judgements of rank. We chose to use this human evaluation data because, as compared to accuracy and fluency judgements, it eliminates some confounding factors. As the person is simply comparing translations of the same source sentence, the original sentence length, sentence difficulty or sentence domain are kept constant. Even so, different translations will contain a variety of errors in both the words used in the translations and the word orderings. It is therefore not clear whether human preference judgements are indeed measuring the quality of the word order in the translation.

We have already established in a previous experiment (in Section 6.3) that we can reliably extract human judgements on word order and that permutation distance metrics are highly correlated with these human judgements. Although in this experiment we are evaluating the metrics on how well they correlate with human judgements, we can, in fact, also judge the human evaluation setup on how well they correlate with permutation distance metrics. The fact that the reordering metrics by themselves are not highly correlated with human rank judgements, see Table 7.5, indicates that these

human experiments are not especially sensitive to the quality of the word order. In fact, except for the case of German-English, the distance metrics agree with humans less than the random baseline would (50%). Although this human judgement data is not ideal, it is still the best we have available to evaluate the LRscore metrics of overall translation quality.

7.5 Summary

In this chapter, we present a novel metric called the LRscore. The main motivation for this metric is the fact that it measures the reordering quality of MT output by using permutation distance metrics. It is a simple, decomposable metric which interpolates the reordering component with a lexical component, the BLEU score.

This chapter demonstrates that the LRscore metric correlates better with human preference judgements of machine translation quality than other machine translation metrics. We show that combining two largely orthogonal information sources results in a superior combined metric.

We also demonstrate that the weight of the metric can be optimised on fairly small amounts of human judgement data when training each language pair individually. Furthermore, we present a novel approach to training a language independent parameter which is optimised across multiple language pairs. Combining the language independent parameter with a measure of the amount of reordering in the test set, displays correlation with human judgements which is comparable to that of training on each language pair. This makes it easy to tune the LRscore parameter without needing human judgements for each new language pair or domain.

In the next chapter we show that the LRscore is more sensitive to changes in reordering conditions than other baseline metrics. We also show that adding reordering to the objective function while training translation model parameters improves translation quality as judged by humans.

Chapter 8

Experiments with LRscore

8.1 Introduction

In the previous chapter, Chapter 7, we presented the LRscore. This metric is motivated by its ability to accurately measure reordering performance and the fact that the individual components of the score can be examined separately.

Automatic metrics are necessary for evaluating the quality of the output. However, an equally important function of automatic metrics is to provide an objective function for training the weights of the log linear translation model. In this chapter we apply the LRscore during minimum error rate training (MERT) (Och, 2003) in order reward the translation model for producing better reorderings. We show that humans prefer the output of translation models trained with the LRscore over those trained with the BLEU score. We also show that when training with the LRscore, there is no discernible drop in performance with respect to the BLEU score.

Another important characteristic of a good automatic metric is its ability to discriminate between systems of varying quality. The results must be sensitive enough to differentiate systems which are fairly close in quality. In this chapter we have designed a set of experiments which show that reordering metrics are more informative and more accurate than other machine translation metrics when conditions affecting reordering are varied.

The rest of this chapter proceeds as follows. In Section 8.2 we use the LRscore as the objective function during MERT training. Then, in Section 8.3, we describe experiments where we examine how sensitive metrics are at detecting changes in reordering conditions. Finally, in Section 8.4 we summarise the contributions and findings of the chapter.

8.2 Optimising Translation Models

The parameters of log linear translation models are commonly tuned using MERT. MERT searches for the parameter setting which maximises some objective function, typically an automatic translation metric such as BLEU, which is applied to the output of a translation model. The success of MERT therefore depends heavily on the evaluation metric, and the BLEU score is not particularly informative regarding the word order performance of the hypotheses. A model with optimised feature weights is likely to exhibit the properties that the metric rewards, but it will be blind to aspects of translation quality that are not captured by the metric. We apply the LRscore during MERT training in order to inject knowledge about reordering behaviour into the training process. If we are able to improve reordering, there could also be improvements in comprehension, grammaticality and the overall quality of the output.

Cer et al. (2010) explore how optimizing toward various automatic evaluation metrics (BLEU, METEOR, NIST, TER) affects the behaviour of the resulting model. They show that the although other metrics might correlate better with human judgements than the BLEU score, when used for training translation models, the BLEU score trained model is preferred by humans. They conclude that when using a metric to train a translation model, it can only be useful to the extent that the MT models structure and features allow it to take advantage of the metric. We therefore adopt the BLEU score as a strong baseline.

8.2.1 Experimental Design

We hypothesise that the LRscore is a good metric for training translation model weights. We test this hypothesis by evaluating the output of the tuned models, first with automatic metrics, and then by using human evaluation. We choose to run the experiment with the Chinese-English language pair as it contains a large amount of medium and long distance reorderings.

8.2.1.1 Experimental Conditions

We apply four variations of the LRscore as an objective function: BLEU1 and the complete BLEU score are used together with the Hamming distance and Kendall's tau distance. BLEU and BLEU1 are also applied on their own as baseline objective functions.

8.2.1.2 Data

It is very important that these experiments are as similar as possible to experiments that would be performed by researchers in the machine translation community. We therefore use the GALE 2008 Chinese-English data, which is a standard training set on which state-of-the-art models have been trained upon. We use the official test set of the 2006 NIST evaluation (1994 sentences). For the development test set, we used the evaluation set from the GALE 2008 evaluation (2010 sentences). Both development set and test set have four references. The translation model was built from 1.727M parallel sentences from the GALE 2008 training data.

8.2.1.3 Models

The MOSES phrase-based translation model was used, with a distortion limit of 6. See Appendix A for details. The SRILM language modelling toolkit (Stolcke, 2002) was used, with interpolated Kneser-Ney discounting to train three separate trigram language models. These were trained on the English side of parallel corpus, the AFP part of the Gigaword corpus, and the Xinhua part of the Gigaword corpus. For the final experiment we also added a 5-gram language model, trained on English side the parallel corpus. A lexicalised reordering model was used with the msd-bidirectional-fe option. The output was re-cased using a recaser trained as a monotone translation model.

The reordering metrics require alignments. Thus the development, test and translated sentences had to be aligned to the source. We did this using the Berkeley word alignment package version 1.1 (Liang et al., 2006), with the posterior probability set to being 0.5.

8.2.1.4 Baseline Metrics

We use the same baseline metrics as those described in Section 7.3.1.6.

8.2.1.5 LRscore parameter setting

We need to set the weight which balances the contribution of the lexical and the reordering component of the score. We use the language independent method described above in Section 7.3.2.3. We first extract the amount of reordering in the test set by calculating the Kendall's tau distance from the monotone. This value is 66.06% which is lower than any of the other language pairs seen so far, which means the translation

LR-HB1	LR-HB4	LR-KB1	LR-KB4
26.40	07.19	43.33	26.23

Table 8.1: The parameter setting representing the % impact of the reordering component for the different versions of the LRscore metric.

are further from the source ordering or that there is more reordering. We then calculate the optimal parameter setting by using the values from Table 7.10 for each of the four LRscore versions. We apply these adjusted parameters by using the reordering amount as a power exponent. Table 8.1 shows the final parameter settings we used in the following experiments. These parameters represent the percentage contribution of the reordering component of the LRscore metric.

8.2.1.6 Human Evaluation Setup

Human judgements of translation quality are necessary to determine whether humans prefer sentences from models trained with the BLEU score or with the LRscore. There have been some recent studies which have used the on-line micro-market, Amazons Mechanical Turk, to collect human annotations (Snow et al., 2008; Callison-Burch, 2009). While some of the data thus generated is very noisy, invalid responses are largely due to certain workers (Kittur et al., 2008). We use Mechanical Turk and we simulate expert-level quality by collecting multiple judgements, and eliminating workers who do not achieve a minimum level of performance on gold standard questions.

In previous human experiments, we recruited volunteers to evaluate translations on a web based interface. The advantage of Mechanical Turk is that a large amount of data can be collected from workers all over the world in a very short period of time and for relatively small amounts of money. This experiment was completed in one hour for a cost of about $30.

Our test data was generated by randomly selecting sentences from the test set for presentation to the judges. These sentences had to be between 15 and 30 words long. Shorter sentences were avoided as they tend to have uninteresting differences, and longer sentences may have many conflicting differences. We also eliminated sentences where the translation output was identical between the two systems. We selected 60 sentences for comparing BLEU with the LRscore using the Hamming distance (LR-HB4), and another 60 for comparing BLEU with the LRscore using Kendall's tau distance (LR-KB4). Workers were presented with randomly ordered test cases and

Reference	By providing free vocational skill training to the rural laborers, the city has removed 1,017 laborers out of the farmland for new jobs during the year.
Option A	through the rural labor force to free vocational skills training, as a whole, the transfer of 1,017 total labor force.
Option B	through the rural labor force to free vocational skills training, as a whole, the total labor force and 1,017.
Explanation	A contains 'the transfer of' which parallels the concept 'removed' that is present in the reference.

Table 8.2: An example of a gold test unit where Option A was labelled as correct.

completed as many examples as they wanted. Only one worker completed more than 30.

The instructions given to the workers were to read the reference sentence, and then to carefully compare the two translations. They should then select whether they preferred translation option A or translation option B, and only if there was no difference in quality should they select the final option "Don't Know". Option A and option B were randomly assigned either a translation from the BLEU score trained system or from the LRscore trained system. They were then given an example to clarify the instructions. Please see Appendix B for details.

Workers were screened to guarantee reasonable judgement quality. 20 sentence pairs were randomly selected from the 120 test units and annotated as gold standard questions. Workers who got less than 60% of these gold questions correct were disqualified and their judgements discarded.

After getting a gold question wrong, a worker is presented with the right answer and an explanation. This guides the worker on how to perform the task and motivates them to be more accurate. We used the Crowdflower[1] interface to Mechanical Turk, which implemented the gold functionality for us.

Table 8.2 shows as example of an annotated gold test unit. Option A was labelled as correct and 82% of the workers chose A as their preferred option. 6% chose B and 12% chose "Don't Know". Humans disagree on which translations they prefer, and so a relatively low threshold of 60% agreement was chosen. Users were able to express their disagreement with the gold standard annotations and one worker who had selected

[1] http://www.crowdflower.com

"Don't Know" objected to the classification of A being preferred by saying "Neither is even close to the meaning, or to being grammatically correct.". Even though experts can disagree on preference judgements, gold standard labels are necessary to weed out the substandard workers. There were 21 trusted workers who achieved an average accuracy of 91% on the gold. There were also 96 untrusted workers who averaged 29% accuracy on the gold and their judgements were discarded. Three judgements were collected from the trusted workers for each of the 120 test sentences. More than three judgements for the gold questions were collected, but only the first three were used so that all sentences are equally weighted.

8.2.2 Results

8.2.2.1 Automatic Metrics

In this experiment we demonstrate that the reordering metrics can be used as learning criterion in minimum error rate training to improve parameter estimation for machine translation.

Obj.Func. \ Metrics	BLEU1	BLEU	LR-HB1	LR-HB4	LR-KB1	LR-KB4	d_h	d_k	TER	METEOR
BLEU1	73.2	31.0	66.5	32.2	71.9	41.3	47.7	70.2	60.7	56.0
BLEU	73.4	31.1	66.0	32.1	71.5	41.1	45.4	69.1	60.7	55.6
LR-HB1	72.7	29.6	66.4	31.0	71.8	40.3	**48.7**	**70.6**	61.7	55.4
LR-HB4	**73.6**	**31.2**	**66.6**	**32.4**	**72.0**	**41.4**	47.1	70.0	**60.5**	**56.1**
LR-KB1	72.8	30.2	66.4	31.6	71.9	40.8	48.6	70.6	61.3	55.5
LR-KB4	73.3	30.1	65.5	31.1	71.3	40.2	43.8	68.7	61.2	55.6

Table 8.3: Rows represent systems trained with the objective functions indicated on the left. Columns represent the test results for these models for different metrics. The diagonal is expected to show the optimal results.

Table 8.3 reports the results of the MERT training with different objective functions. The lexical metrics BLEU1 and BLEU are used as objective functions in isolation, and also as part of the LRscore together with the Hamming distance (shown with prefix LR-H) and Kendall's tau distance (shown with prefix LR-K). The B1 suffix means BLEU1 has been used, and B4 means BLEU has been used. All the systems are trained using the objective functions and they are then evaluated using the automatic metrics reported in the columns. We test models using our different objective functions and we also apply distance metrics and the TER and METEOR scores. Tuning using reordering metrics resulted in very poor performance as would be expected as they are not complete metrics.

The first thing we note in Table 8.3 is that we would expect that the diagonal would report the highest scores, as MERT maximises the objective function on the development data set. This is not the case however. The best results, across the board, are reported for the LR-HB4 objective function which uses the Hamming distance. The only exception to this is that the reordering metrics report the highest scores when using the LR-HB1 objective function. This is an important result, even though the difference in scores is not large, as it shows that by training with the LRscore objective function, BLEU scores do not decrease. Although this is surprising, it can be explained by the fact that BLEU allows multiple solutions with the same score, and the LRscore allows us to select the one which has better reordering. The reordering metrics and the lexical metrics are orthogonal information sources, and combining them results in better performing systems. These results are reinforced in the next section where we show that humans also prefer the LRscore translations.

Another interesting finding reported in Table 8.3, is that there is very little difference between using BLEU1 and BLEU as the objective function. It seems that the higher order n-grams do not have a large impact on the performance of the trained models. This is surprising as higher order n-grams provide all of the BLEU score's ability to measure word order, and BLEU1 is a metric which only measures lexical success.

MERT does not find a global optimum, and it is possible that our training procedure found a poor local optimum. We therefore repeat MERT experiments two more times with different random starting points. Table 8.4 shows the outcome of three different MERT runs. Test scores are averaged and the standard deviation is shown in brackets. This table shows that the scores are relatively stable across different optimizations, as the standard deviations are quite small. METEOR changes the most between different

Metrics / Obj.Func.	BLEU	LR-HB4	LR-KB4	TER	METEOR
BLEU	**31.1 (0.0)**	32.1 (0.0)	41.0 (0.1)	60.7 (0.1)	55.5 (0.3)
LRHB4	**31.1 (0.2)**	**32.2 (0.1)**	**41.3 (0.1)**	60.6 (0.2)	55.7 (0.2)
LRKB4	31.0 (0.2)	**32.2 (0.2)**	41.2 (0.2)	**61.0 (0.5)**	**55.8 (0.4)**

Table 8.4: Average results and standard deviation (in brackets) of three different MERT runs for different objective functions.

MERT runs, and has a standard deviation of 0.4 percentage points. These results do not contradict the initial results reported in Table 8.3. When using the LRscore as an objective function, the other metrics' scores are not depressed. The best scores are now shared between the LRHB4 and the LRKB4 metrics.

Metrics / Obj.Func.	BLEU	LR-HB4	LR-KB4	TER	METEOR
BLEU	**32.2**	**33.2**	41.9	60.4	**55.9**
LRHB4	31.9	32.7	41.7	**60.9**	55.6
LRKB4	32.1	**33.2**	**42.0**	60.7	55.4

Table 8.5: Results for different objective functions with the addition of a large 5-gram language model.

The results in this chapter have been extracted from models using three trigram language models. Although these LM models improve local orderings, it is not anticipated that stronger language models change the findings of the experiment. Table 8.5 shows the results of an additional experiment where the models were trained and tested using a more powerful 5-gram language model as well as to the three trigram language models. We can see that all the scores improve in comparison to Table 8.4. The important result here is that there is still no notable drop in the BLEU score performance when training the model with the LRscore.

To better understand the impact of the different objective functions, Table 8.6 shows the translation model and reordering model weights that resulted from the MERT experiments shown in Table 8.3. When training the model with different objective functions, the only notable difference in the translation model weights is with the phrase penalty weight, where the LRscore leads to a much higher phrase penalty. A larger phrase penalty means that the model prefers translations which are composed

	Translation Model Weights					
Obj.Func.	p(f|e)	lex(f|e)	p(e|f)	lex(e|f)	ph.penalty	w.penalty
BLEU	0.027	0.062	0.061	0.029	0.035	-0.231
LR-HB4	0.043	0.056	0.041	0.015	0.085	-0.150
LR-KB4	0.040	0.063	0.056	0.024	0.097	-0.195

	Reordering Model Weights						
Obj.Func.	monof	swapf	discontf	monob	swapb	discontb	dist. cost
BLEU	0.006	0.012	0.049	0.189	0.085	0.023	0.104
LR-HB4	0.017	0.045	0.002	0.316	0.041	0.070	0.038
LR-KB4	0.022	0.047	0.023	0.213	0.040	0.046	0.048

Table 8.6: The weights of the models when training with different objective functions.

Obj. Funcs	Prefer LR	Prefer BLEU	Don't Know	Total
LR-KB4 vs. BLEU	96 (53.3%)	79 (43.9%)	5	180
LR-HB4 vs. BLEU	93 (51.7%)	79 (43.9%)	8	180
Total LR vs. BLEU	189 (52.5%)	158 (43.9%)	13	360

Table 8.7: The number of times human judges preferred the output of systems trained either with the LRscore or with the BLEU score, or were unable to choose.

of a smaller number of longer phrases. The reordering model weights are also quite mixed. The LRscore prefers the monotone orderings and the swap forward orderings. The BLEU score prefers the discontinuous forward ordering and the swap backwards ordering. These differences might not be very important, but the fact that the distortion cost is considerably lower is interesting. The LRscore trained models thus assign a lower cost to distortions.

8.2.2.2 Human Evaluation

Although it is interesting to consider the automatic metric scores and the model weights, any final conclusion on the impact of the metrics on training must use human evaluation of translation quality. We collect human preference judgements on the output of systems trained using the BLEU score and the LRscore. We thus aim to determine whether training with the LRscore leads to genuine improvements in translation quality. Table 8.7 presents the results of our human evaluation experiment. For both the

LR-KB4 vs. BLEU and the LR-HB4 vs. BLEU scenarios, humans show a greater preference for the output for systems trained with the LRscore. The difference in the number of times humans preferred the LRscore (189) vs the BLEU score (158) is quite large and it seems like reordering information genuinely improves the quality of the trained translation system.

The sign test can be used to determine whether the difference in preference is significant. The null hypothesis is that the probability of a human preferring the LRscore trained output is the same as that of preferring the BLEU trained output. The one-tailed alternative hypothesis is that humans prefer the LRscore output. If the null hypothesis is true, then there is only a probability of 0.048 that 189 out of 347 $(189+158)$ people will select the LRscore output. We therefore discard the null hypothesis and the human preference for the output of the LRscore trained system is significant to the 95% level.

In order to judge how reliable our judgements are we calculate the inter-annotator agreement. This is given by the Kappa coefficient (K):

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is the proportion of time that the workers agree, and P(E) is the proportion of time that they would agree by chance. Inter-annotator agreement was 64.28% and the expected agreement is 33.33%. The Kappa coefficient is therefore 0.464 which is considered to be a moderate level of agreement.

We expect that more substantial gains can be made in the future by using reordering metrics to train models which have more powerful reordering capabilities. A richer set of reordering features, and a model capable of longer distance reordering would better leverage metrics which reward good word orderings. Even though the phrase-based model struggles to model reordering, when analysing the output sentence, we found that output from the system trained with the LRscore tended to have better structure. In Table 8.8 we see a typical example. The word order of the sentence trained with BLEU is mangled, whereas the LR-KB4 model outputs a very clear translation which closely matches the reference. It also garners higher reordering and BLEU scores. The scores shown are calculated with all four references, not only the one reference that is shown.

Type	Sentence	Sm.BLEU	d_k
Reference	silicon valley is still a rich area in the united states. the average salary in the area was us $62,400 a year, which was 64% higher than the american average.	na.	na.
LR-KB4	silicon valley is still an affluent area of the united states, the regional labor with an average annual salary of 6.24 million us dollars, higher than the average level of 60 per cent.	34.6	78.2
BLEU	silicon valley is still in the united states in the region in an affluent area of the workforce, the average annual salary of 6.24 million us dollars, higher than the average level of 60 per cent	31.4	76.4

Table 8.8: A reference sentence is compared with output from models trained with BLEU and with the LR-KB4 lrscore.

8.3 Metric Sensitivity to Reordering Conditions

We have just demonstrated the value of the LRscore as an objective function for tuning the parameters of a translation model. In our final experiments, we demonstrate the LRscore's ability to evaluate research on different reordering conditions.

The following experiments vary factors which affect the reordering performance of the models. Distortion limits, lexicalised reordering models and language models are all examined. Although we know that these factors affect the word order of the output, it is not clear exactly what the effect is. Allowing some distortion is desirable, but how much does it improve translation and how much distortion should we allow? How does the lexicalised reordering model help translation? Does it encourage more reorderings, or fewer, but better chosen reorderings? These kinds of questions are very difficult to answer with current translation metrics. Using the LRscore and its individual score components, we gain insight into the effect that these conditions have on translation. The experiments in this chapter are aimed at supporting research into reordering.

As we are not sure of the actual effect of varying reordering conditions, metrics are not evaluated on their ability to measure a certain effect. Instead we evaluate the metrics based on their sensitivity to change. We aim to determine if they are able to detect differences in conditions reliably.

8.3.1 Experimental Design

The experiments performed in this section use the same experimental design as those described in the previous section. Please see Section 8.2.1 for details. Additionally, a small language model was trained on 100,000 lines of text from the English side of the GALE corpus. The BLEU score is used as our objective function so that results will be comparable with other work. Additionally, the individual n-gram precisions of the BLEU metric have been calculated. We report these scores as 1BLEU - 4BLEU. BLEU1 applies a brevity penalty but 1BLEU does not. When these precision scores are reported as a sentence level metric they are smoothed, and when they are reported as a document level metric they are not smoothed.

8.3.1.1 Experimental Conditions

We present experiments which explore the effect of varying the following reordering conditions:

- **Search Restrictions**

  As described in Section 2.2.2, reordering restrictions on the search for the best translation hypothesis are necessary in order to make decoding tractable. Although some reordering is undoubtedly desirable, when searching through a vast number of possible orderings, the number of search errors made by the decoder could grow. In practice a distortion limit of six is generally considered the best setting.

- **Lexicalised Reordering Model**

  Many phrase-based translation models apply a lexicalised reordering. This models the probability that a phrase is monotone, inverted or disjoint with respect to the preceding and following phrases (see Section 2.2.3 for details). The lexicalised reordering model generally improves translation quality as it provides more information for the decoder, however the effect of this model is limited in scope to local adjacency decisions.

- **Language Model**

> Language models are crucial for producing fluent translations. The effect of the language model on the quality of the output of the translation model is also local and limited to the n-gram length of the model. It assigns probabilities to consecutive segments of the translated sentences. Most MT models rely heavily on language model probabilities to influence the word order of the target sentence. The problem with relying on the language model is that it incorporates no knowledge of the source sentence. Over-reliance on the language model can lead to fluent but meaningless or confusing sentences.

There are other factors which influence word order such as the maximum phrase length, and the distance-based reordering model, which encodes the monotone assumption inherent in most translation models. However, since these factors are less important, we have not investigated them.

8.3.1.2 Statistical Significance

The main goal of this experiment is to test the sensitivity of the metrics to incremental changes. We test for the significance of the differences between two sets of sentence level metric values by using the Wilcoxon signed-rank test (Wilcoxon, 1945). As described in Section 6.2.2, this test is appropriate when the distribution cannot be assumed to be normally distributed. Our experimental results are mostly presented at the document level, but sentence level scores are used for significance testing.

8.3.2 Results

8.3.2.1 More Reordering

The distortion limit in the phrase-based models controls the amount of reordering that the translation model is allowed to perform. By increasing the distortion limit, we should initially see an improvement as the model is allowed to discover good sequences of target phrases. However beyond a certain limit, the model can be overwhelmed by the possible permutations and is not longer able to distinguish good orderings from poor ones. We increased the distortion limit from zero (forcing monotone translation) to twelve to see how this affects the metric scores. The word order of the hypothesis is guided by the language model and the default distance based reordering model which penalises reorderings. For this initial experiment, the lexicalised reordering model is not applied.

	dl0	dl3		dl6		dl9		dl12	
BLEU	26.8	26.3 (-0.5)	**	27.0 (0.7)		27.5 (0.5)	**	29.1 (1.6)	***
METEOR	54.7	54.6 (-0.1)		54.0 (-0.6)	***	55.1 (1.1)	***	55.6 (0.5)	*
TER	37.7	37.4 (-0.3)	**	36.0 (-1.4)	***	38.2 (2.2)	***	38.7 (0.5)	
1BLEU	69.8	69.7 (-0.1)		69.6 (-0.1)		71.0 (1.4)	***	72.8 (1.8)	***
2BLEU	36.4	35.9 (-0.5)	**	36.5 (0.6)	*	37.4 (0.9)	***	40.1 (2.7)	***
3BLEU	19.7	19.3 (-0.4)	**	20.0 (0.7)	***	20.4 (0.4)	**	22.6 (2.2)	***
4BLEU	10.8	10.5 (-0.3)	*	11.1 (0.6)	**	11.3 (0.2)		12.9 (1.6)	***
Hamming	69.4	69.9 (0.5)	***	72.5 (2.6)	***	73.7 (1.2)	***	74.5 (0.8)	*
Kendall	73.4	73.0 (-0.4)	***	71.6 (-1.4)	***	70.5 (-1.1)	***	71.5 (1.0)	***
LR-HB4	25.8	25.4 (-0.4)	***	25.3 (-0.1)		25.8 (0.5)	**	28.1 (2.3)	***
LR-KB4	36.5	36.1 (-0.4)	***	35.9 (-0.2)	*	36.2 (0.3)		36.6 (0.4)	**

Table 8.9: The document level metric scores for systems with different distortion limits, distortion limit 0 to 12, with no lexicalised reordering model. The difference between metric scores for successive distortion levels is shown between brackets. First distortion level 3 is compared to distortion level 0, then 6 to 3 etc. The significance of the difference is calculated by using the Wilcoxon Signed Rank Test.

In Table 8.9 we can see the metric scores for translation models with different distortion limits. The absolute scores are of interest, but we are more concerned with how these scores change as we increase the amount of reordering. We therefore present the differences in scores between the adjacent distortion limits in brackets, along with the significance of their difference.

The baseline metrics BLEU, METEOR and TER seem to give better scores for translations with larger distortion limits. Allowing reordering to occur is obviously beneficial, however some metrics, for example METEOR and TER, only show improvements over a distortion level of zero when the distortion limit reaches nine. The three baseline metrics show their best results for the maximum distortion level of 12. This is interesting because most reordering experiments set the limit to 6. We look at the permutation distance scores and broken down BLEU scores to gain insight into what is occurring.

The broken down BLEU scores show that small amounts of reordering, with a distortion level of three, slightly reduce the scores for all the n-gram BLEU scores. This is surprising and could be due to the fact that the Chinese-English language pair has a large proportion of longer distance reorderings (See Chapter 3) which cannot be covered by the distortion limit of three. Only with a distortion limits of six or more, is the language model able to find better matches for the longer ngrams 2BLEU, 3BLEU and 4BLEU. The purely lexical metric 1BLEU only seems to benefit with a distortion limit of nine or more.

Lexical metrics alone do not fully explain the differences between different distortion limits. The Kendall's tau distances show us that with increased reordering, the word order of the translation diverges more and more from the word order of the closest reference. It is only with a distortion limit of 12 that the scores improve a bit, even though they are still lower than when the distortion level is zero. The Hamming distance, however, shows improved scores for each increase in distortion. It seems that absolute order improves with more reordering, but not relative order.

There are slightly more correct bi-grams, tri-grams and 4-grams with larger amounts of reordering. The language model allows the translation model to find better local reorderings. However, this is offset by the overall increase in error in longer distance reorderings.

Combining lexical and reordering metrics into one score here seems less informative than looking at them separately. LRHB4 largely reflects the BLEU score performance and LRKB4 cancels the increases in BLEU with decreases in the Kendall tau's

	dl0	dl3	dl6	dl9	dl12	Ref.s
Hamming	81.8	78.9	65.2	58.5	59.4	41.9
Kendall	91.2	89.6	83.1	79.3	79.3	67.8

Table 8.10: The amount of reordering: distances calculated by comparing the word order in the translations and references to the monotone. Values closer to 100 are closer to monotone.

distance metric. The LRscore was designed so that the individual components of the score would be easy to examine.

Apart from judging the quality of the word order in the translations, we can also look at the quantity of reordering in the translations and indeed the references. We do this by calculating the distance to the monotone, and by doing this, we gain insight into the nature of the effect of the different test conditions on the translations.

In Table 8.10 we can see that the reordering metrics reflect the fact that with larger distortion limits, more reordering is performed as the score drops and the translations get further and further away from the monotone. This trend reverses at a distortion limit of 12, where the scores increases again slightly. The translations are not completely monotone with a distortion level of zero due to reorderings within phrase pairs and possible automatic word alignment effects. What is interesting to note here is that even though translations are far from monotone, they are still much closer to monotone than the reference. This means that we not only have to increase the quality of the reorderings to match human translations, but we have to increase their quantity. This insight is not available with the other translation metrics.

8.3.2.2 More Informed Reordering

Much research in reordering involves proposing better models of reordering. In this section we present an experiment where we apply an additional reordering model to the experiment in the previous section. The additional lexicalised reordering model provides adjacent phrase ordering information during the search for the best hypothesis. Typically applying lexicalised reordering models improves the translation quality, however, the effect of the lexicalised reordering model has usually been measured by indirect measures of word order performance, such as the BLEU score.

We argue that research in reordering requires explicit measures of success. Here we show that our permutation distance metrics are more sensitive and more reliable at

detecting the difference in word order performance than the MT metrics. We also show that part of the reason that there is better reordering, is that less incorrect reordering is occurring.

	dl0		dl3		dl6		dl9		dl12	
BLEU	25.9 (-0.9)	***	27.1 (0.8)	***	28.1 (1.1)	***	28.2 (0.7)	*	29.9 (0.8)	***
METEOR	53.6 (-1.1)	***	55.0 (0.4)	*	55.0 (1.0)	***	54.3 (-0.8)	***	55.7 (0.1)	
TER	36.8 (-0.9)	***	37.9 (0.5)	**	38.5 (2.5)	***	38.2 (0.0)		39.1 (0.4)	*
1BLEU	69.3 (-0.5)	*	70.1 (0.4)	*	70.8 (1.2)	***	70.3 (-0.7)	**	72.9 (0.1)	
2BLEU	35.7 (-0.7)	***	36.8 (0.9)	***	38.2 (1.7)	***	38.2 (0.8)	***	40.1 (0.6)	***
3BLEU	19.0 (-0.7)	***	20.1 (0.8)	***	21.2 (1.2)	***	21.2 (0.8)	***	22.6 (0.8)	***
4BLEU	10.2 (-0.6)	**	11.0 (0.5)	***	11.8 (0.7)	***	11.9 (0.6)	***	12.9 (0.7)	***
Hamming	69.7 (0.3)		69.9 (0.0)		70.7 (-1.8)	***	71.6 (-2.1)	***	74.9 (0.3)	*
Kendall	73.3 (-0.1)		73.1 (0.1)		72.8 (1.2)	***	72.1 (1.6)	***	71.1 (-0.4)	***
LR-HB4	25.1 (-0.7)	***	26.0 (0.6)	***	26.5 (1.2)	***	26.2 (0.4)	**	28.5 (0.4)	***
LR-KB4	35.9 (0.6)	***	36.6 (0.5)	***	37.0 (1.1)	***	36.7 (0.5)	***	36.9 (0.3)	

Table 8.11: The document level results for different metrics with different distortion limits where we have applied an additional lexicalised reordering model. The difference between models with and without (see Table 8.11) lexicalised reordering is shown in brackets. A positive number indicates that the addition of the reordering model improved the translation as expected. The significance of the difference is calculated by using the Wilcoxon Signed Rank Test.

Table 8.11 shows the metric scores for different distortion limits where we have applied a lexicalised reordering model. The important information in this table is how these results compare to the case without the lexicalised reordering model, shown in Table 8.9. To highlight the comparison, the differences between the scores in this table with lexicalised reordering and without lexicalised reordering, from Table 8.11, are shown in brackets. A positive score indicates an improvement in translations with the addition of the lexicalised reordering model. We also include the statistical significance of the difference over sentence level scores between the models with and without lexicalised reordering.

We can see that the addition of lexicalised reordering generally improves the quality of translations as judged by the metrics. The fact that scores go down slightly for the case of no distortion could be due to the fact that the lexicalised reordering model is causing non-optimal phrase pairs to be selected to try to fit the monotone ordering. For the case of the distortion limit of three, there is generally an improvement in scores, but it is not very significant. All metrics show very significant increases in scores with lexicalised reordering for the distortion limit of six, although it seems that absolute word order has deteriorated as measured by the Hamming distance. This amount of reordering seems to allow the model to make good use the ordering information from the reordering model. When the distortion limit reaches nine and twelve, the reordering metrics and the higher order n-gram metrics are the only metrics which are very sensitive to the effect of the reordering model. Overall the LRscores are slightly more sensitive to the effect of the reordering model across all distortion limits.

In Figure 8.1 we see a breakdown of the differences in scores between the lexicalised reordering and the non-lexicalised reordering scores for a distortion of level 6. These histograms group sentences into 20 groups according to their assigned scores. Firstly, it is interesting to see the distribution of scores assigned to the Chinese-English translations. But more relevant to this experiment, one can see the change in distribution with the addition of the lexicalised reordering metric. With lexicalised reordering, more sentences have higher scores for all metrics.

We would like to know whether the amount of reordering has changed. Table 8.12 shows the difference in word order between the translation with the reordering model and the monotone. We show the difference in amount of reordering in brackets for the case without the reordering model. Many differences are positive which means that less reordering is occurring. We would expect that when we introduce the reordering model, the translation model learns that it can place more trust in reorderings, and

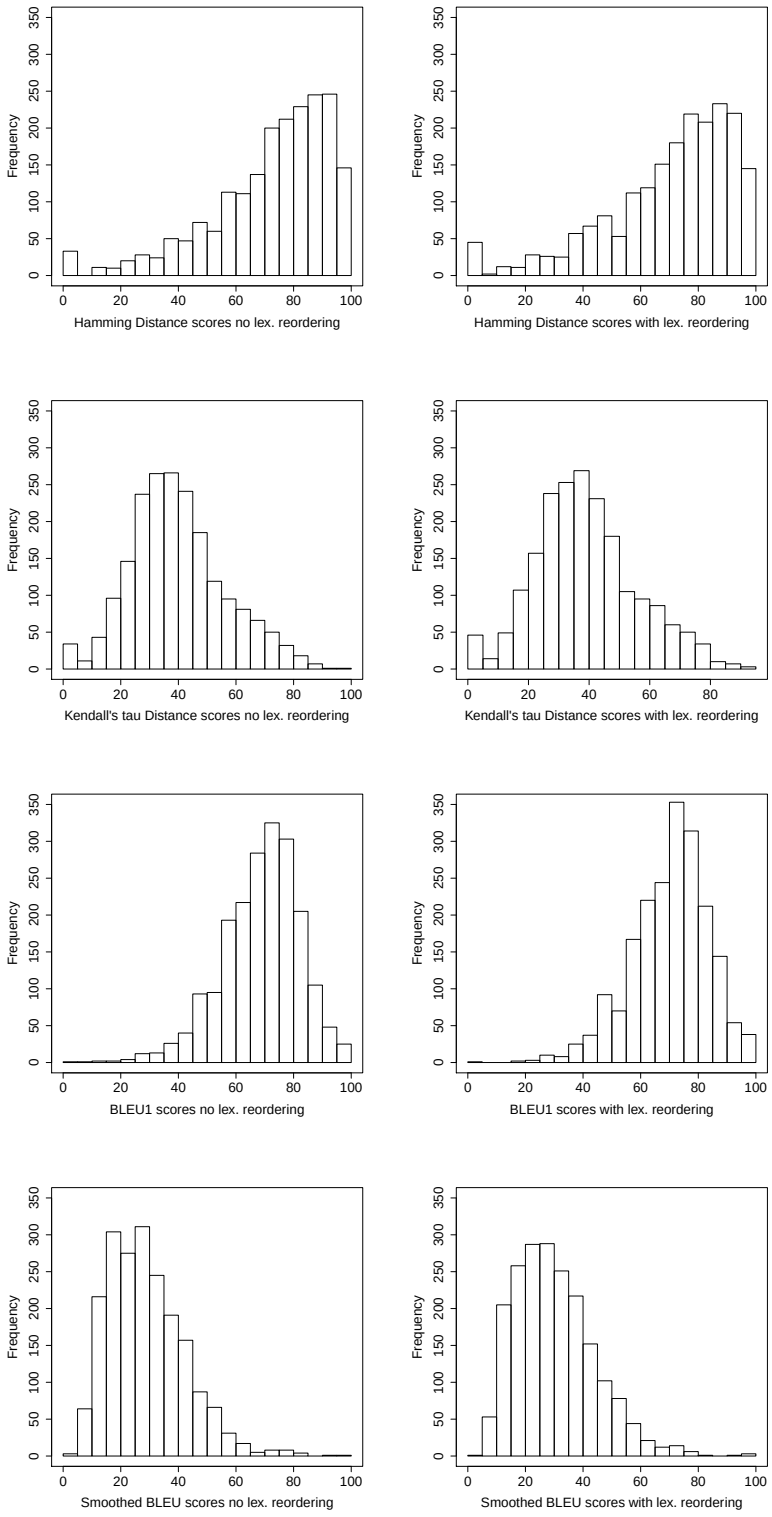


Figure 8.1: Comparing the distribution of scores for the baseline metrics, with and without the lexicalised reordering model, where the distortion level is 6.

	dl0	dl3	dl6	dl9	dl12	Ref.s
Hamming	81.4 (-0.4)	79.5 (0.6)	68.8 (3.6)	61.9 (3.4)	53.0 (-6.4)	41.9
Kendall	90.2 (-1.0)	89.4 (-0.2)	84.9 (1.8)	80.3 (1.0)	75.8 (-3.5)	67.8

Table 8.12: The amount of reordering occurring in output from models with lexicalised reordering. 100 means monotone ordering in the output. The differences in amounts of reordering with and without lexicalised reordering (see Table 8.10) are shown in brackets. A positive number indicates that the addition of the reordering model decreased the total amount of reordering in the translations, i.e. that the sentences with lexicalised reordering are closer to the monotone.

therefore and perform more of them. Instead, it is actually performing less reordering. However, when the distortion limit is zero or 12, then more reordering does occur.

As we can see in this section, BLEU, METEOR and TER are sensitive to both more reordering and more informed reordering. However, they do not reveal what causes the difference between two test conditions. Using reordering metrics, combined with lexical metrics, one can see exactly what changes in the output. We can detect improvements in word order. We can detect increases or decreases in the amount of reordering. We can also see if lexical choice improves. Using this method of analysis, we can provide strong support for claims that, for example, a new reordering model is improving the word order of translations.

8.3.2.3 Language Modelling

The language model is one of the largest contributors to the word order of a translation. We investigate what the effect is of applying language models of different quality and whether metrics are able to measure this.

In Table 8.13 we see the results of two systems where one system applies a very small trigram language model, and the other applies three large trigram language models, the ones used by the preceding experiments. Both models use a distortion limit of six and a lexicalised reordering model.

The baseline metrics improve considerably with higher quality language models. The greatest increase of 4.2 is reported for the BLEU score. The reordering metrics are largely unaffected by the large improvement in the language model. Kendalls tau improves by 0.6 but the Hamming distance even goes down slightly by -0.6. This is explained by the fact that they are not sensitive to the improved lexical choice that

	Small LM	Large LM	Difference
BLEU	23.9	28.1	(4.2) ***
METEOR	51.9	55.0	(3.1) ***
TER	34.9	38.5	(3.6) ***
1BLEU	68.1	70.8	(2.7) ***
2BLEU	33.7	38.1	(4.4) ***
3BLEU	17.1	21.1	(4.0) ***
4BLEU	9.0	11.8	(2.8) ***
Hamming	71.3	70.7	(-0.6) ***
Kendall	72.2	72.8	(0.6) ***
LR-HB4	23.3	26.5	(3.2) ***
LR-KB4	34.4	37.0	(2.6) ***

Table 8.13: The document level metric scores for systems with different sized language models.

the language model provides. The language model also, however, affects local word orderings. The Kendalls tau could benefit slightly from this because it is sensitive to relative word order, but the Hamming distance sees no benefit in absolute word order with the larger language models.

The effect of a larger language model on the broken down BLEU scores is more revealing. We can see that the language model improves the longer n-gram scores the most. We can see that 2BLEU gets the greatest increase of 4.4 points. Even 4BLEU improves more than 1BLEU. This shows us that the impact of the language model does not just improve lexical choice, which would have been demonstrated by a larger 1BLEU increase. The most important effect of the language model is to improve local ordering.

Local orderings are important to the quality of translation, however, they cannot adequately account for the large number of longer distance reorderings seen in the Chinese-English language pair. The MT metrics, and BLEU in particular, are most sensitive to improvements in local reorderings. The only way to get an idea of how the reordering has changed in the sentence as a whole is to use the reordering metrics.

8.3.3 Related Work

There have been very few studies which isolate the impact of design decisions on translation quality. Zollmann et al. (2008) perform a study where they vary the distortion limit of a phrase-based model and compare it to the hierarchical translation model. This work shows the persistent, although small, advantage of SCFG approaches. The problem with this study is that it assumes that the BLEU score is able to measure differences in reordering. In our experiments, we apply reordering metrics and a combination of baseline metrics (BLEU, METEOR and TER) to determine how appropriate the metrics are for this kind of research. We also examine the quantity of distortion, by comparing word order to the monotone. In this way we can see how much reordering is occurring and gain a deeper insight into the effect of the changes.

8.4 Summary

In this chapter we explore the usefulness of the LRscore metric. First we examine the effect of using the LRscore as an objective function while training translation model parameters. As a trained model is likely to exhibit the properties that the metric rewards, the goal was to improve the reordering behaviour of the model. We show that when training a phrase-based translation model with the LRscore, the model retains its performance as measured by the baseline metrics, in particular the BLEU score.

In order to determine whether the LRscore leads to real improvements in translation quality, we designed an experiment using human judges. We show that humans prefer the output of models trained with the LRscore, and thus confirm the value of the permutation distance metrics.

These experiments use the MOSES phrase-based decoder which is very limited in its ability to model long distance reordering. Apart from the restrictions on search, it also only applies local models of ordering. More powerful translation models, such as syntax-based models, which allow for longer distance reordering and have more structured models to guide their word order choice, would benefit even more from using the LRscore while tuning.

Tuning translation models is important, but researchers also need metrics which are sensitive to changes in reordering conditions for evaluating their research. We present experiments which demonstrate that reordering metrics are superior to current metrics both because of their sensitivity to changes in reordering conditions and because of the

insights reordering metrics can provide.

BLEU is heavily influenced by local word orderings, and it is thus sensitive to factors such as language models. However, it has little ability to capture long distance improvements in reordering. Our reordering metrics can measure both global and local reorderings and they can also measure either absolute order or relative word orderings. They can also be used to measure how much reordering is occurring which leads to new insight into the effect of lexicalised reordering models and language models.

It is important to note, however, that reordering metrics are not as sensitive as the broken down BLEU metrics to small, local reordering improvements. These improvements might be important for readability, but user comprehension is unlikely to be as affected by local reorderings as by larger reorderings which affect the structure of the sentence. We therefore conclude that the best approach is to apply the LRscore and to examine the individual lexical and a reordering scores that it provides. Looking at the breakdown of the score components will allow researchers to better judge the impact of a change to reordering conditions.

In the next and final chapter, we review the contributions of this thesis and we discuss future directions for research.

Chapter 9

Conclusion and Future Work

In this thesis we have introduced methods and metrics for quantitatively analysing reordering in parallel corpora.

The main claims defended in this thesis are:

- We have shown that reordering is an important factor in determining the performance of translation systems. We performed a regression analysis of translation systems over 110 language pairs which showed that the amount of reordering in a parallel corpus affects translation performance more than morphological complexity and language similarity. This wide ranging analysis provides strong evidence for the importance of research into reordering.

- We have shown that current machine translation metrics do not adequately measure reordering performance. We have described the limitations of the approaches that three commonly used shallow metrics take to measuring the quality of word order. Using examples, we demonstrate their failure to measure the amount of difference in word order between references and translations. Finally, we perform an experiment where metric scores are correlated with measures of lexical and reordering quality. Metric scores were very strongly correlated with lexical measures, and only slightly correlated with measures of reordering quality. This shows that current machine translation metrics are primarily responding to differences in the words used in translation, and that they are largely insensitive to word order quality.

- A large part of this thesis is dedicated to demonstrating that permutation distance metrics capture the quality of word order better than current machine translation metrics. We start by describing the properties of the distance metrics and their

advantages with respect to the current metrics. Of primary importance is the fact that they measure the number of words which are out of order. We design a novel human evaluation which isolates the effect of word order differences on fluency and comprehension judgements. We show that permutation distance metrics correlate more strongly with these judgements than the current machine translation metrics, even under conditions which favour the current machine translation metrics (perfect lexical overlap). We also show that the simple combined metric, the LRscore, correlates better with human preference judgements, of the overall quality of sentences. Finally, we show that the LRscore improves the quality of translation models when used as the objective function while training model parameters. Humans prefer the output of models trained using the LRscore over models trained using the BLEU score.

This thesis has contributed to our understanding of the challenges involved in modelling reordering. We have highlighted how poorly our current state-of-the-art models are performing and we have also shown that there is a great range of different distributions of reorderings amongst European languages and Chinese-English and Arabic-English. These findings allow researchers to select appropriate language pairs in order to test their theories, and to choose reasonable model parameters for those languages. If, for instance, someone makes a claim about improving long distance reordering, a language with a large number long distance reorderings can be selected.

The most significant contribution that this thesis makes to the field of statistical machine translation, however, is that it provides tools for measuring reordering performance. The permutation distance metrics, in particular the Kendall's tau distance, provide reliable, accurate measures of the amount of relative disorder between the translation and the reference sentences. Both small and large word order differences are detected and reported. These metrics have been rigorously evaluated and have shown to correlate well with human judgements of word order quality, and when combined with lexical metrics, with human judgements of overall quality. The code for these metrics is available as a standalone metric which has been distributed to various researchers on request. The code which incorporates the distance metrics and the LRscore is included as part of the open-source code base of the MOSES project and this promotes the diffusion and impact of the research described in this thesis. The code for the metric, independent of the MOSES optimisation module, is also available [1].

[1] see http://homepages.inf.ed.ac.uk/abmayne/

9.1 Contributions

A list of the major contributions of the thesis follows:

Methods for analysing corpora.

By defining a reordering as a pair of inverted blocks over the word alignment grid, as extracted by our reordering extraction algorithm, we are able to collect useful statistics over parallel sentences and corpora.

Comparison of human, phrase-based and hierarchical reorderings

We analyse the output of two different state-of-the-art translation models and show that neither of them are capable of capturing the great majority of the reorderings that exist in the reference sentences.

Analysis of reordering and its impact across many language pairs.

We show the great variety of reordering characteristics in different language pairs, highlighting languages that are particularly problematic, such as German-English and Chinese-English. Common parameter settings in state-of-the-art translation models have been shown to be inadequate, such as the distortion limit of 6 for the phrase-based model. We also demonstrate that the amount of reordering is the biggest factor influencing the performance of translation models.

Creation of a human evaluation which isolates reordering performance.

We create a human evaluation task which specifically isolates word order by artificially permuting a reference sentence with different amounts of disorder. This allows us to evaluate metrics on their correlation with human judgements of word order quality, something which no other human evaluation has been demonstrated to achieve.

Demonstrating that humans are sensitive to different amounts of reordering.

We show that humans can reliably distinguish sentences with different levels of reordering. This provides further confirmation that metrics should also measure this.

Definition of Permutation Distance Metrics for evaluating reordering.

We define novel reordering metrics based on permutation distance metrics which use word alignments to measure the quality of the word order in isolation from the actual words used in the translation.

Showing that reordering metrics correlate with human judgements of reordering.

Permutation distance metrics are strongly correlated with human judgements of word order quality.

Showing that machine translation metrics are largely insensitive to reordering.

We show that current machine translation metrics primarily measure the success of word choice and that they are largely insensitive to word order differences.

Definition of LRscore.

We present a complete machine translation metric which combines lexical and reordering metrics. We show that the LRscore correlates better with human preference judgements than baseline metrics.

Integration of the LRscore into the training of the translation model parameters.

When using the LRscore as the objective function for tuning the translation model parameters, translation quality improves. This is shown by the fact that humans prefer the output of models trained with the LRscore, over the output from the model trained with the BLEU score.

9.2 Discussion

Although our metrics are clearly better at measuring reordering performance than previous machine translation metrics, an obvious concern is the fact that two sets of word alignments are required: one for the source-reference sentence, and one for the source-translation. This need not be a major obstacle, however. Gold standard alignments are scarce, but if accuracy is paramount, a test set with manually annotated alignments could be selected. Also, the translation systems can output the word alignments that were used to generate the translation. This approach was followed in Chapter 3. Unfortunately gold standard alignments are often not available. Alignments can also be automatically generated using the alignment model that aligns the training data. This approach was followed in Chapter 4 where 110 translation models were analysed.

Apart from alignments, another issue to consider with regard to our method, is that we rely upon the assumption that word orderings should be close to reference. This is a strong assumption and might hold true for target languages with strict constraints on word order, but for languages with freer word order, such as Russian, it is not clear

that we will be successful. Essentially, we are aiming to capture the grammaticality and even more importantly, the accuracy of the translation, and word order is only a small part of this equation. It seems that, certainly for Chinese-English, a large part of the reordering problem exists at the clause level. Differences in ordering at this level lead to problems understanding how the parts of the sentence fit together.

There is some scope to believe that if the ordering of clauses in the translation was similar to the ordering in the reference, that these sentences are more likely be more comprehensible. Word order is not everything, however, and even if the order is correct, the linking words might not be. A more sophisticated metric, which could analyse the relationships between clauses, would arguably be a better reflection of the quality of the translation. Metrics such as textual entailment metrics (Padó et al., 2009a), which measures argument structure overlap, have already been proposed. Unfortunately, textual entailment metrics are slow and complex, sometimes more complex than the translation models themselves. Our approach is simple and efficient and will therefore be useful even in the event of a more knowledgeable metric becoming widely adopted.

I will make one final comment on the relative merits of different shallow metrics. The BLEU score is surprisingly good at measuring small, incremental improvements in the ordering capabilities of a translation system. If it matches even a few more n-grams between source and target sentences in a document, it will report improved scores. The problem with BLEU is that it is insensitive to large differences in word order, and it is therefore inappropriate to use it to compare systems which are very different. This is the same conclusion that was reached by Callison-Burch et al. (2006b). So if you wished to improve the lexicalised reordering model of the phrase-based paradigm, then perhaps the BLEU score is adequate. However, if you are proposing a new translation model, which, for example, solves the long distance ordering of the verb final phenomenon of German, BLEU would not be the best choice of metric. Under these circumstances, you would need to use the LRscore.

9.3 Future Directions

Our work on reordering is clearly an improvement on previous approaches. However, there is still more work to be done. A major drawback of the permutation distance metrics is that they rely upon a simplified representation of the alignment function. We could work around some of the assumptions if we adapt the metrics to partially

ranked data (Critchlow, 1985; Fagin et al., 2003), which are able to represent null and many-one alignments. However, metrics for partially ranked data require unintuitive extensions to handle ties. We could also abandon permutations and simply compare the aligned target word indexes using rank correlations, which would measure the strength of association between the two arrays. This approach would still be incapable of handling many-one alignments.

Developing distance metrics directly over the alignment grid would avoid this problem. Measuring the similarity of graphs is important for machine learning applications in diverse areas such as molecular biology, telecommunications, and social network analysis. This algorithmic problem has therefore received extensive attention. Graph kernels have been proposed as a theoretically sound and promising approach to the problem of graph comparison (Borgwardt, 2007), and can be efficient for graphs which are not excessively large, such as those found in sentence alignments.

Apart from improvements to the reordering metrics themselves, another important avenue of research is that of applying these metrics to translation models which actually model long distance reordering in an efficient manner. The phrase-based model was used for much of the research in this thesis, and the benefit of applying the reordering metric for training and evaluation of this model is limited to potential that the phrase-based model has for improvement. A syntax-based model which incorporates a strong model of reordering would potentially benefit more from having its weights trained using a reordering metric.

A large body of research into the problem of reordering has been evaluated using metrics which have been shown to be insensitive to the quality of word order. This thesis provides metrics which allow researchers to reliably evaluate their work. We believe that these metrics will be the key to driving significant improvements in the field.

Appendix A

Experimental Design: Models

This appendix describes the details of the translation and alignment models used for experiments in the thesis.

A.1 MOSES

The MOSES translation system is one of the most widely used open-source machine translation projects. It has an extensive homepage[1]. This thesis used the phrase-based model which was initially the only decoder included in the project. It can be downloaded from the Sourceforge code repository[2]. In the following sections we describe the implementation details of MOSES for the experimental sections of the thesis. Any differences from these settings are clearly discussed in the experimental design sections of the chapters containing translation experiments. No factors were used with Moses in this thesis.

Distortion Limit	6
Drop Unknown	Yes
NBest list	100
Alignment symmetrization	grow-diag-final
Lexicalised Reordering	msd-bidirectional-fe

Table A.1: MOSES settings

We extracted phrases as in Koehn et al. (2003) by running GIZA++ in both directions and merging alignments with the grow-diag-final heuristic. This instance of Moses contained 14 real-valued features:

- 1 language model feature
- 4 translation model features as described in Koehn et al. (2007)
- Phrase Penalty
- Word Penalty
- 6 lexicalised reordering features as described in Koehn et al. (2005)
- Distortion penalty

[1] http://www.statmt.org/moses/
[2] http://sourceforge.net/projects/mosesdecoder/

All settings of Moses are used irrespective of the language pair involved. The only changes which are not specified in the experimental design sections, are the exact versions of MOSES used in the experiments.

Chapter	Version
Chapter 3	2008-02-20
Chapter 4	2007-12-04
Chapter 6	2008-09-23
Chapter 8	2009-11-25

Table A.2: MOSES versions used in different experiments

In Table A.2 we can see the dates that the MOSES model was downloaded from the source control system, corresponding to different versions of the source code. For the MERT experiments in Chapter 8, we developed a novel version of the MERT code, in order to use the LRscore as the objective function. This is freely availables in the Sourceforge repository in the "mert-mtm5" branch.

A.2 HIERO

For the grammmar-based model experiments in Chapter 3, the Hiero hierarchical phrase-based decoder was used. The code was kindly provided by David Chiang.

Version	2006-05-02 Version 1.0
Minimum hole length	2
Maximum rule length	5
Maximum phrase length	10
Number unaligned words at edges	0
Beam threshold	15
Stack limit for chart pruning	30
Drop Unknown	Yes
NBest list	100

Table A.3: HIERO settings

This instance of Hiero contained 6 real-valued features:

- 1 language model feature

- 4 translation model features as described in Chiang (2005)
- Phrase Penalty

A.3 Berkeley Aligner

For some experiments, we use the Berkeley Aligner (Liang et al., 2006) to word align parallel sentence pairs, instead of GIZA++ aligner. The Berkeley aligner has been shown to be more robust than using GIZA++ in situations where there are long sentences and sparse word counts (Koehn et al., 2008). This software can be obtained from Google code[3]. In Chapter 6 and Chapter 7, we used version 2 and in Chapter 8, we used version 1.1. In both cases, the posterior threshold set at 0.5.

[3] http://code.google.com/p/berkeleyaligner/

Appendix B

Experimental Design: Instructions for Human Experiments

This appendix includes the experimental instructions and an example of the materials that were shown to judges in the experiments involving humans.

B.1 Reproduced Reorderings: Section 3.3.5

In Figure B.1 we can see the instructions given to human judges for the experiment described in Section 3.3.5.

You will be presented with a reference translation and a machine translation of the same unseen source sentence. They will each have two sequences of words, show with different underline styles. Please compare the ordering of these two sequences of words between the reference and the machine translation, and judge whether their order with respect to each other in the translation is “Correct” or “Incorrect”. Please select “Not Applicable” only when the translated words are so different from the reference that their ordering is irrelevant.

Figure B.1: The instructions given for manual evaluation.

Figure B.2 shows an example of a test case presented to the workers. The spans of the reordering are marked with different underline styles. In this example the ordering of the underlined phrases in the reference and the translation are different, showing that in this case the reordering was not reproduced.

Reference
corporations which drain contamination in huai river drainage area tributaries must implement a treatment deadline, and by the end of 1997, at the latest, stop draining contamination into tributaries.
Translation
for all to the huai river basin river pollution enterprises, to conduct a management at the end of 1997, the latest organisation to stop pollution.

Figure B.2: An example test case from the manual evaluation of word order task. The ordering under consideration is marked with double and wavy underlines.

B.2 Human Sensitivity to Reordering: Section 6.3

In Figure B.3 we can see the first page of instructions given to human judges for the experiment described in Section 6.3. Figure B.5 describes the experimental proceedure for the human judges. Figure B.4 shows the examples given to human judges to clarify the instructions. Figure B.6 shows an example of a test case presented to the workers.

Thanks for taking part in this experiment!
Please only participate if you consider your level of English to be fluent.
Please read through the instructions below before starting.

In this experiment you are asked to judge how fluent and how comprehensible sentences are on a scale of 1 to 7.

Fluency refers to whether the sentences are grammatical and well-formed in English.

- If the sentence is grammatical, then you should rate the sentence high in terms of fluency.
- If the sentences are something like word salad, then you should give the sentence a low number.

Comprehension refers to how understandable the sentences are.

- If the sentence is almost impossible to understand, then you should give it a low number.
- If the sentence is readily understandable, coherent and doesn't require any effort on the reader's part, then you should give it a high number.
- If a sentence is ungrammatical, but with effort you can make out what it means, then you should give it a medium to high score.

Try to use a wide range of numbers and to distinguish as many degrees of acceptability as possible.

Figure B.3: Basic instructions given for manual evaluation.

Suppose you were given the following sentence:

He had achieved complete victory in nine games with Chinese Go players before .

Then, you may rate it high in terms of fluency (e.g., 6 or 7) as the sentence is well-formed and grammatical. It would be also given a high comprehension score (e.g., 6 or 7) as it makes sense.
Now, take the following example:

This war , in including from Germany . those European countries from a total died of four million people

This sentence is much harder to read than the previous example. It contains grammatical errors and the individual words do not make sense together. So you would rate this sentence low in terms of fluency (e.g., 1 or 2). This sentence also lacks coherence. It is very difficult to figure out how the different parts of the sentence fit together. Overall the sentence is not comprehensible and would receive a low score (e.g., 1 or 2).

Awarding ceremony the was at The Philippines Cultural Center solemnly held .

This sentence is not grammatical but its meaning is reasonalby clear. This would get a low fluency score (e.g., 1 or 2), and a higher comprehension score (e.g., 5 or 6).

Figure B.4: Examples given for instructions in manual evaluation.

Procedure

When you start the experiment below you will be asked to enter your personal details. Next, you will be presented with 40 sentences to evaluate in the manner described above. Once you have completed your rating, click the button at the bottom of that page to advance to the next sentence.

Things to remember:

- Full-screen your web browser before starting the experiment.
- Keep this page open so that you can refer back to it if you are at all unsure of how to rate a sentence.
- Higher numbers represent a positive opinion of the sentence and lower numbers a negative one.
- Do not spend too long analysing the sentences; you should be able to rate them once you have read them for the first time.
- There is no right or wrong answer, so use your own judgement when rating each sentence.

Figure B.5: Proceedure given for manual evaluation.

This is the the Holland Trade Promotion Association
has established in China first representative office
that.

How fluent is the sentence?
1 2 3 4 5 6 7

How comprehensible is the sentence?
1 2 3 4 5 6 7

Figure B.6: Example of test case for manual evaluation.

B.3 Human Preference for LRscore output: Section 8.2.1.6

In Figure B.7 we can see a screen shot of the instructions. Figure B.8 shows an example of a test case presented to the workers.

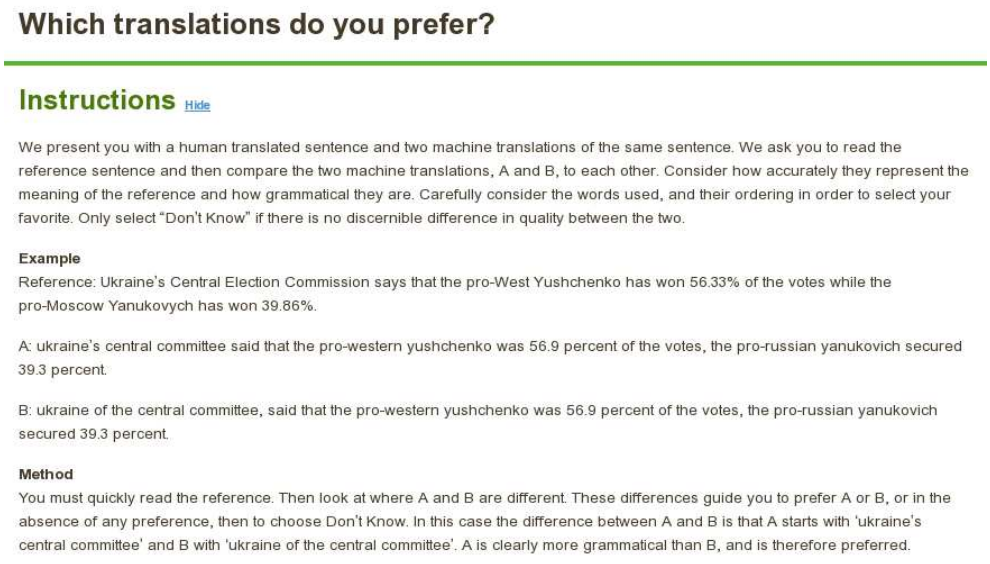


Figure B.7: The instructions shown to workers on Amazon's Mechanical Turk.

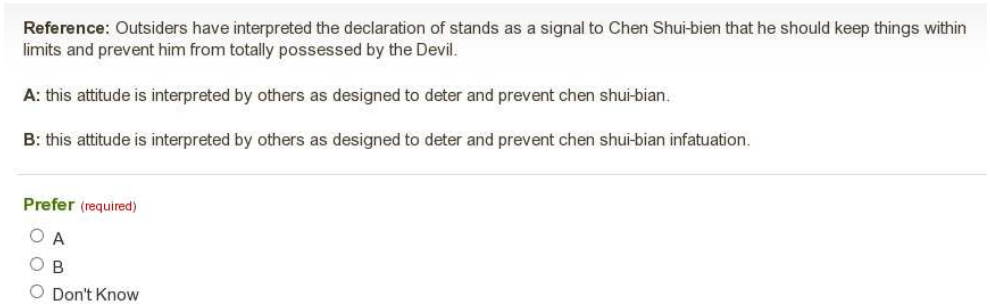


Figure B.8: An example of the judgements solicited from the workers on Amazon's Mechanical Turk.

Appendix C

Europarl Matrices

This appendix presents the values of the characteristics of different language pairs for the Europarl experiments reported with graphics in Chapter 4.

	el	it	pt	es	fr	en	sv	da	nl	de	fi
el	-	178	167	167	157	162	183	183	188	188	0
it	178	-	773	788	803	247	254	263	260	265	0
pt	167	773	-	874	709	240	241	250	253	247	0
es	167	788	874	-	734	240	241	250	258	253	0
fr	157	803	709	734	-	236	242	241	244	244	0
en	162	247	240	240	236	-	591	593	608	578	0
sv	183	254	241	241	242	591	-	830	648	631	0
da	183	263	250	250	241	593	830	-	663	707	0
nl	188	260	253	258	244	608	648	663	-	838	0
de	188	265	247	253	244	578	631	707	838	-	0
fi	0	0	0	0	0	0	0	0	0	0	-

Table C.1: The language similarity for Europarl matrix of languages.The rows represent the source language and the columns the target language.

	el	it	pt	es	fr	en	sv	da	nl	de	fi
el	-	0.337	0.331	0.302	0.372	0.358	0.314	0.349	0.487	0.493	0.490
it	0.329	-	0.246	0.273	0.294	0.435	0.407	0.423	0.558	0.593	0.577
pt	0.320	0.247	-	0.202	0.255	0.397	0.363	0.388	0.545	0.565	0.576
es	0.323	0.248	0.216	-	0.269	0.397	0.364	0.397	0.559	0.581	0.596
fr	0.349	0.302	0.248	0.271	-	0.395	0.371	0.412	0.557	0.613	0.593
en	0.367	0.413	0.393	0.402	0.402	-	0.272	0.311	0.523	0.586	0.421
sv	0.329	0.405	0.357	0.375	0.404	0.267	-	0.245	0.445	0.480	0.432
da	0.337	0.405	0.394	0.394	0.413	0.311	0.240	-	0.458	0.495	0.430
nl	0.505	0.561	0.545	0.562	0.573	0.538	0.476	0.473	-	0.369	0.519
de	0.540	0.593	0.603	0.624	0.637	0.608	0.501	0.524	0.417	-	0.543
fi	0.508	0.604	0.614	0.614	0.601	0.477	0.434	0.452	0.568	0.575	-

Table C.2: The reordering quantity RQuantity for Europarl matrix of languages. The rows represent the source language and the columns the target language.

	el	it	pt	es	fr	en	sv	da	nl	de	fi
el	-	25.53	30.16	33.66	31.33	30.10	23.56	25.03	21.98	19.39	14.60
it	20.04	-	32.78	35.69	34.11	29.79	23.05	24.04	21.89	19.29	13.95
pt	21.30	30.31	-	39.82	36.76	32.00	24.33	25.96	22.46	20.73	14.85
es	22.13	30.15	37.42	-	37.60	33.31	25.64	27.18	23.11	20.87	15.25
fr	21.62	30.49	36.37	39.33	-	33.01	25.75	26.60	23.61	20.82	14.92
en	19.16	24.49	29.86	32.51	30.59	-	26.80	28.33	23.75	20.66	15.52
sv	18.22	22.84	28.06	31.14	28.25	32.91	-	32.87	24.21	22.16	16.44
da	17.02	21.32	26.48	28.93	26.91	30.90	29.76	-	23.43	21.33	15.78
nl	14.75	19.99	24.14	25.50	24.14	26.45	21.26	24.00	-	21.15	12.46
de	16.34	20.44	25.73	28.57	26.34	27.96	22.94	25.86	24.88	-	14.56
fi	13.69	18.33	21.65	24.49	22.32	24.72	20.03	22.50	18.89	17.24	-

Table C.3: The Bleu scores for Europarl matrix of languages.The rows represent the source language and the columns the target language.

Bibliography

Aho, A. V. and Ullman, J. D. (1969). Syntax directed translations and the pushdown assembler. *Computer and System Sciences*, 3(1):37–56.

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

Allen, J. (2003). Post-editing. *Computers and Translation: a Translators Guide*, pages 297–317.

Avramidis, E. and Koehn, P. (2008). Enriching morphologically poor languages for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 763–770, Columbus, Ohio. Association for Computational Linguistics.

Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

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

Bates, D. and Sarkar, D. (2007). *lme4: Linear mixed-effects models using S4 classes*.

Berger, A., Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., Kehler, A. S., and Mercer, R. L. (1996). Language translation apparatus and method of using context-based translation models. *United States Patent*, Patent Number 5,510,981.

Bickel, B. and Nichols, J. (2005). *The World Atlas of Language Structures*, chapter Inflectional synthesis of the verb. Oxford University Press.

Birch, A., Blunsom, P., and Osborne, M. (2009). A Quantitative Analysis of Reordering Phenomena. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 197–205, Athens, Greece. Association for Computational Linguistics.

Birch, A., Blunsom, P., and Osborne, M. (2010). Metrics for MT Evaluation: Evaluating Reordering. *Machine Translation*.

Birch, A., Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Constraining the Phrase-Based, Joint Probability Statistical Translation Model. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 154–157, New York City. Association for Computational Linguistics.

Birch, A. and Osborne, M. (2010). Lrscore for evaluating lexical and reordering quality in mt. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 327–332, Uppsala, Sweden. Association for Computational Linguistics.

Birch, A., Osborne, M., and Koehn, P. (2008). Predicting Success in Machine Translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics.

Bojar, O. and Zabokrtsky, Z. (2009). CzEng0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92. in print.

Borgwardt, K. M. (2007). *Graph Kernels*. PhD thesis, LMU München.

Brown, P. F., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 286–295. Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and Zaidan, O. (2010). Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006a). Re-evaluating the role of Bleu in machine translation research. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006b). Re-evaluation the role of bleu in machine translation research. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Cer, D., Manning, C. D., and Jurafsky, D. (2010). The best lexical metric for phrase-based statistical mt system optimization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 555–563, Los Angeles, California. Association for Computational Linguistics.

Chang, P.-C. and Toutanova, K. (2007). A discriminative syntactic word order model for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 9–16, Prague, Czech Republic. Association for Computational Linguistics.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 263–270, Ann Arbor, Michigan.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics (to appear)*, 33(2).

Collins, M., Koehn, P., and Kucerova, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 531–540, Ann Arbor, Michigan.

Crawley, M. (2007). *The R book*. John Wiley & Sons Inc.

Critchlow, D. (1985). *Metric methods for analysing partially ranked data*, volume 34. Springer.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–22.

Deza, M. and Huang, T. (1998). Metrics on permutations, a survey. *Journal of Combinatorics, Information and System Sciences*, 23:173–185.

Diaconis, P. and Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *Royal Statistical Society Series B*, 32(24):262–268.

Dyen, I., Kruskal, J., and Black, P. (1992). An indoeuropean classification, a lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5).

Dyer, C. (2007). The 'noisier channel': Translation from morphologically complex languages. In *Proceedings on the Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Dyer, C. and Resnik, P. (2010). Context-free reordering, finite-state translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 858–866, Los Angeles, California. Association for Computational Linguistics.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48.

Eisner, J. (2003). Learning non-isomorphic tree mappings for machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 205—-208, Sapporo, Japan.

Eisner, J. and Tromble, R. W. (2006). Local search with very large-scale neighborhoods for optimal permutations in machine translation. In *Proceedings of the HLT-NAACL Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Processing*, pages 57–75, New York.

Fagin, R., Kumar, R., and Sivakumar, D. (2003). Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17:134–160.

Fox, H. J. (2002). Phrasal cohesion and statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 304–311, Philadelphia, USA.

Galley, M., Hopkins, M., Knight, K., and Marcu, D. (2004). What's in a translation rule? In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*, pages 273–280, Boston, USA.

Galley, M. and Manning, C. D. (2008). A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 847–855, Honolulu, Hawaii. Association for Computational Linguistics.

Germann, U., Jahr, M., Knight, K., Marcu, D., and Yamada, K. (2001). Fast decoding and optimal decoding for machine translation. In *Meeting of the Association for Computational Linguistics*, pages 228–235, Toulouse, France.

Giménez, J. and Màrquez, L. (2008). A smorgasbord of features for automatic MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198. Association for Computational Linguistics.

Goldwater, S. and McClosky, D. (2005). Improving statistical MT through morphological analysis. In *Proceedings of Empirical Methods in Natural Language Processing*.

Hamming, R. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, 26(2):147–160.

Huang, L. and Chiang, D. (2005). Better k-best parsing. In *Proceedings of the Workshop on Parsing Technologies*, pages 53–64, Vancouver, Canada. Association for Computational Linguistics.

Hutchins, W. and Somers, H. (1992). *An introduction to machine translation*. Academic Press New York.

Karamanis, N. (2003). *Entity Coherence for Descriptive Text Structuring*. PhD thesis, University of Edinburgh.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, 30:81–89.

Kendall, M. and Gibbons, J. D. (1990). *Rank Correlation Methods*. Oxford University Press, New York.

Kerridge, D. (1975). The interpretation of rank correlations. *Applied Statistics*, 2:257–258.

Khalilov, M. and Sima'an, K. (2010). A discriminative syntactic model for source permutation via tree transduction. In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*, pages 92–100, Beijing, China. Coling 2010 Organizing Committee.

Kittur, A., Chi, E., and Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 453–456. ACM.

Knight, K. (1999). Squibs and discussions: Decoding complexity in word-replacement translation models. In *Computational Linguistics*, volume 25, pages 607–615. MIT Press.

Koehn, P. (2004a). Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the Association for Machine Translation in the Americas*, pages 115–124.

Koehn, P. (2004b). Statistical significance tests for machine translation evaluation. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT-Summit*.

Koehn, P., Axelrod, A., Birch, A., Callison-Burch, C., Osborne, M., and Talbot, D. (2005). Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *International Workshop on Spoken Language Translation*.

Koehn, P., Birch, A., and Steinberger, R. (2009). 462 Machine Translation Systems for Europe. In *Machine Translation Summit XII*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics Companion Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P. and Monz, C. (2005). Shared task: Statistical machine translation between European languages. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 119–124. Association for Computational Linguistics.

Koehn, P., Och, F., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*, pages 127–133, Edmonton, Canada. Association for Computational Linguistics.

Koehn, P., Schroeder, J., and Osborne, M. (2008). Edinburgh University System Description for the 2008 NIST Machine Translation Evaluation. *NIST MT Evaluation Meeting*.

Krauwer, S. (1993). Evaluation of MT systems: a programmatic view. *Machine Translation*, 8(1):59–66.

Kumar, S. and Byrne, W. (2005). Local phrase reordering models for statistical machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 161–168, Vancouver, Canada. Association for Computational Linguistics.

Laine, M., Niemi, J., Koivuselkä-Sallinen, P., Ahlsén, E., and Hyönä, J. (1994). A neurolinguistic analysis of morphological deficits in a Finnish-Swedish bilingual aphasic. *Clinical Linguistics & Phonetics*, 8(3):177–200.

Lapata, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. *Computational Linguistics*, 29(2):263–317.

Lapata, M. (2006). Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

Lavie, A. and Agarwal, A. (2007). Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Workshop on Statistical Machine Translation at the Meeting of the Association for Computational Linguistics (ACL-2007)*, pages 228–231.

Lavie, A. and Agarwal, A. (2008). Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Workshop on Statistical Machine Translation at the Meeting of the Association for Computational Linguistics (ACL-2008)*.

Lavie, A. and Denkowski, M. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine Translation*.

LDC (2005). Linguistic data annotation specification: Assessment of fluency and adequacy in translations. revision 1.5.

Liang, P., Taskar, B., and Klein, D. (2006). Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York City, USA. Association for Computational Linguistics.

Lin, C.-Y. and Och, F. (2004a). Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the conference on Computational Linguistics*, page 501.

Lin, C.-Y. and Och, F. J. (2004b). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 605–612, Barcelona, Spain.

Lopez, A. (2009). Translation as weighted deduction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 532–540. Association for Computational Linguistics.

Marcu, D., Wang, W., Echihabi, A., and Knight, K. (2006). SPMT: Statistical machine translation with syntactified target language phrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 44–52, Sydney, Australia.

Marcu, D. and Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 133–139, Morristown, USA.

Melamed, I. D. (2003). Multitext grammars and synchronous parsers. In *Proceedings of the Human Language Technology Conference and the North American Association for Computational Linguistics*, pages 158–165, Edmonton, Canada.

Mellish, C., Knott, A., and Oberlander, J. (1998). Experiments Using Stochastic Search For Text Planning. In *In Proceedings of International Conference on Natural Language Generation*, pages 98–107.

NIST (2008). Evaluation plan for gale go/no-go phase 3.

NIST (2009). Open machine translation 2009 evaluation (mt09). http://www.itl.nist.gov/iad/mig//tests/mt/2009/.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., and Radev, D. (2004). A smorgasbord of features for statistical machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 161–168, Boston, USA. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 295–302, Philadelphia, USA.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):9–51.

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–450.

Och, F. J., Tillmann, C., and Ney, H. (1999). Improved alignment models for statistical machine translation. In *Proceedings of the Joint Workshop on Empirical Methods in NLP and Very Large Corpora*, pages 20–28, Maryland, USA.

Och, F. J., Ueffing, N., and Ney, H. (2001). An efficient A* search algorithm for statistical machine translation. In *Data-Driven Machine Translation Workshop*, pages 55–62, Toulouse, France.

Padó, S., Galley, M., Jurafsky, D., and Manning, C. (2009a). Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 297–305. Association for Computational Linguistics.

Padó, S., Galley, M., Manning, C. D., and Jurafsky, D. (2009b). Textual entailment features for machine translation evaluation. In *the EACL Workshop on Machine Translation (WMT)*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

Pinheiro, J. and Bates, D. (2009). *Mixed-effects models in S and S-PLUS*. Springer Verlag.

Przybocki, M. (2004). NIST 2004 machine translation evaluation results. Confidential e-mail to workshop participants.

Przybocki, M., Peterson, K., Bronsart, S., and Sanders, G. (2009). The nist 2008 metrics for machine translation challengeoverview, methodology, metrics, and results. *Machine Translation*, 23(2):71–103.

Ronald, S. (1998). More distance functions for order-based encodings. In *the IEEE Conference on Evolutionary Computation*, pages 558–563.

Russell, S., Norvig, P., Canny, J., Malik, J., and Edwards, D. (1995). *Artificial intelligence: a modern approach*. Prentice hall Englewood Cliffs, NJ.

Shannon, C. and Weaver, W. (1948). The mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Simard, M., Ueffing, N., Isabelle, P., and Kuhn, R. (2007). Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 203–206. Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *AMTA*.

Snover, M., Madnani, N., Dorr, B., and Schwartz, R. (2008). Terp system description. In *EMNLP: Metric MATR Workshop*.

Snover, M., Madnani, N., Dorr, B., and Schwartz, R. (2009). Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of Spoken Language Processing*, pages 901–904.

Swadesh, M. (1955). Lexicostatistic dating of prehistoric ethnic contacts. In *Proceedings American Philosophical Society*, volume 96, pages 452–463.

Talbot, D. and Osborne, M. (2006). Modelling lexical redundancy for machine translation. In *Proceedings of the Association of Computational Linguistics*, Sydney, Australia.

Team, R. D. C. (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Tillman, C. (2004). A unigram orientation model for statistical machine translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*, pages 101–104, Boston, USA. Association for Computational Linguistics.

Tillmann, C. (2003). A projection extension algorithm for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1–8, Sapporo, Japan.

Tillmann, C. and Ney, H. (2003). Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133.

Toutanova, K. and Manning, C. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics.

Tromble, R. and Eisner, J. (2009). Learning linear ordering problems for better translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1007–1016, Singapore. Association for Computational Linguistics.

Ulam, S. (1972). Some ideas and prospects in biomathematics. *Annual Review of Biophysics and Bioengineering*, pages 277–292.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error analysis of machine translation output. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 697–702, Genoa, Italy.

Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of International Conference On Computational Linguistics*, pages 836–841, Copenhagen, Denmark.

Weiss, S. and Kulikowski, C. (1991). *Computer systems that learn*. Morgan Kaufmann Publishers San Hatteo (CA) USA.

Wellington, B., Waxmonsky, S., and Melamed, I. D. (2006). Empirical lower bounds on the complexity of translational equivalence. In *Proceedings of the International Conference on Computational Linguistics and of the Association for Computational Linguistics*, pages 977–984, Sydney, Australia.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Wong, B. and Kit, C. (2009). ATEC: automatic evaluation of machine translation via word choice and word order. *Machine Translation*, pages 1–15.

Wong, B. and Kit, C. (2010). The parameter-optimized atec metric for mt evaluation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 360–364, Uppsala, Sweden. Association for Computational Linguistics.

Wu, D. (1995). Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1328–1334, Montreal, Canada.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Xia, F. and McCord, M. (2004). Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of International Conference On Computational Linguistics*, pages 508–514, Geneva, Switzerland.

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the Association for Computational Linguistics*, pages 523–530, Toulouse, France.

Zens, R. and Ney, H. (2003). A comparative study on reordering constraints in statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 144–151, Sapporo, Japan.

Zhang, H., Gildea, D., and Chiang, D. (2008). Extracting Synchronous Grammar Rules From Word-Level Alignments in Linear Time. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, pages 1081–1088.

Zollmann, A. and Venugopal, A. (2006). Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, New York City, USA. Association for Computational Linguistics.

Zollmann, A., Venugopal, A., Och, F., and Ponte, J. (2008). A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical mt. In *Proceedings of International Conference On Computational Linguistics*.