UNIVERSITÄT
DES
SAARLANDES

Syntactic Word Reordering For Statistical Machine Translation

Maria Nãdejde

Master′s Thesis

Supervisors:
Prof. Dr. Dietrich Klakow
Prof. Dr. Hans Uszkoreit

October, 2011

Abstract

Statistical machine translation for the English-German language pair is prone to errors due to the structural differences between the two languages. Frequent translation errors that result from these differences are: missing words, in particular verbs, incorrect word order and wrong lexical choices. This work investigates two approaches to these problems that target different components of a statistical machine translation system.

The first approach is pre-reordering of the source sentence, such that its word order becomes similar to the word order of the target language. In a recent work, Tromble and Eisner (2009) proposed a framework for modeling the reordering of words as the Linear Ordering Problem. Following this framework, the present work investigates two state-of-the-art linear ordering algorithms as applied to the task of word reordering. Results of automatic evaluation indicate an improvement in translation quality, while error analysis shows that this approach reduces the number of missing verbs, incorrect words and improved word order.

The second approach is to improve the language model component by modeling the dependencies between words in adjacent sentences. Momtazi et al. (2010) proposed the Within-and-Across Sentence Boundary language model (ASB), which models this relation by estimating unigram probabilities based on the word distribution in previous sentences. This thesis presents results of integrating the ASB language model in the phrase-based statistical machine translation system Moses. Automatic evaluation indicates a small contribution from the language model, but further experiments and error analysis show that this model can improve word disambiguation, lexical choices and reduce the number of missing words.

Acknowledgements

I would like to thank my supervisors Prof. Klakow and Prof. Uszkoreit, and my advisor Sabine Hunsicker for their advices and their patience. I am also very grateful to Ivan Titov and Alexandre Klementiev for answering my questions about learning algorithms. I want to thank Tommaso Schiavinotto for providing the original implementation of the Memetic Algorithm and Marc Schulder for doing the error analysis.

I am most of all grateful to my parents for always supporting my decisions and helping me in times of distress. I am also grateful for all my friends that I've made in the past two years and for my friends from home that are missing me. I would also like to thank the wonderful people and professors that are part of the Erasmus Mundus Masters in Language and Communication Technologies.

Declaration

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Saarbrücken, 31 October 2011

Signature:

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Machine translation (MT) is the task of automatically translating text from a source language into a target language. This task is difficult since it needs to model the divergences between languages that can be both grammatical and idiosyncratic differences. For the English-German language pair major differences include word order and morphology. Specifically, while German has a flexible word order allowed by its rich morphology, English has a rather fixed word order and is a morphologically poor language. German allows the word order of *Subject-Verb-Object* to be both **SVO** and **OVS**, while in English it is always **SVO**. German also allows verbs to appear in different positions: in perfect tense the main verb appears at the end of the sentence and some verbs have separable particles that are placed at the end of the sentence. These differences are a cause of frequent translation errors like incorrect word order, missing words, wrong lexical choices and morphology. This thesis proposes two approaches to these problems.

One approach is to pre-reorder the source sentences such that their word order becomes similar the word order in the target language. As proposed by Tromble and Eisner (2009), word reordering is modeled as the Linear Ordering Problem and finding an optimal word order involves solving this problem. Two algorithms that solve this problem are applied in the present word reordering work: a Branch-and-Bound algorithm (Charon and Hudry 2006) which finds a global optimal solution and the Memetic algorithm (Schiavinotto and Stützle 2004) which finds local optimal solutions. A pre-reordering model is learned using the first algorithm and applied as a pre-processing step for MT. This approach is shown to reduce the number of missing verbs, incorrect words and improve word order. The present work will also show that the second algorithm is suitable for the task of word reordering and can be applied to efficiently reorder longer sentences. New dependency features will be introduced as an attempt to capture syntactic word reordering.

The second approach is to capture the dependencies between words in adjacent sentences. This has been done by integrating the Within-and-Across Sentence Boundary language model (Momtazi et al. 2010) in a phrase-based statistical MT system. The idea behind this language model is that words in previous sentences trigger words in the current sentence. The present work will show that this approach can improve word disambiguation, lexical choices and reduce the number of missing words in the machine translation output. It will also identify some difficulties in using this language model for machine translation and propose solutions.

The rest of the thesis is organized as follows. Chapter 2 gives an overview of statistical MT. Chapter 3 describes the pre-reordering model and the results of applying it as a pre-processing step to a phrase-based statistical MT system. Chapter 4 describes the Within-and-Across Sentence Boundary language model and the results of integrating it in a phrase-based statistical MT system. Finally chapter 5 draws some conclusions about the contributions of the two approaches.

CHAPTER 2

Statistical Machine Translation

The aim of this chapter is to offer an overview of the statistical approach to Machine Translation. Brown et al. (1990) formalized the word-based statistical translation system and introduced the notions of Translation Model and Language Model. In a later paper Brown et al. (1993) describe the efficient estimation of the more refined models known as the IBM models. The phrase-based models build upon the word-based models and are able to overcome some of the drawbacks of their predecessors.

This chapter is organized as follows. Section 2.1 introduces the word-based translation models. Section 2.2 describes the state-of-the-art phrase-based translation model. Section 2.3 gives an overview of n-gram language models. Finally section 2.4 discusses how machine translation can be evaluated.

2.1 Word-Based Models

In order to preserve the generality of the models and for brevity, the pair of languages considered for the translation task will be omitted. A string of words in a foreign language such as French will be denoted by **f** and the string corresponding to its translation in a target language such as English will be denoted by **e**. The random variable that describes the position of words in **f** is denoted by **a**. The subscript notation is used to refer to particular element, for example f_j refers to the *j-th* word in string **f**. The superscript notation is used to refer to a sequence up to a particular position, for example $a_1^m = a_1 a_2 ... a_m$ refers to the positions of the first m words. The random variable describing the length of **f** is denoted by l and the random variable describing the length of **e** is denoted by m.

Considering a pair of sentences **e** and **f** that are translations of each other, if one of the sentences is given, for example **f**, then there are two ways of thinking about how to recover the other sentence **e**. **e** can be the result of translating **f** or **e** is the sentence that originally produced **f** as its translation.

The intuition behind the statistical models proposed by Brown et al. (1990) and Brown et al. (1993) is that a string **f** in a foreign language is actually the product of a mental translation from a string **e**. Given a string **e** and a string **f**, the statistical translation model assigns a probability $P(f|e)$, that can be interpreted as the probability that a translator will produce string **f** as the translation of string **e**. This is an interpretation of the *Noisy Channel* model from information theory. With this idea in mind, the task of translating string **f** is equivalent to finding the string **ê** that initially produced the string **f**. The string **ê** is chosen to maximize the probability $P(e|f)$. Using the Bayes theorem this probability can be written as in Equation 2.1.1.

$$P(e|f) = \frac{P(e)P(f|e)}{P(f)} \tag{2.1.1}$$

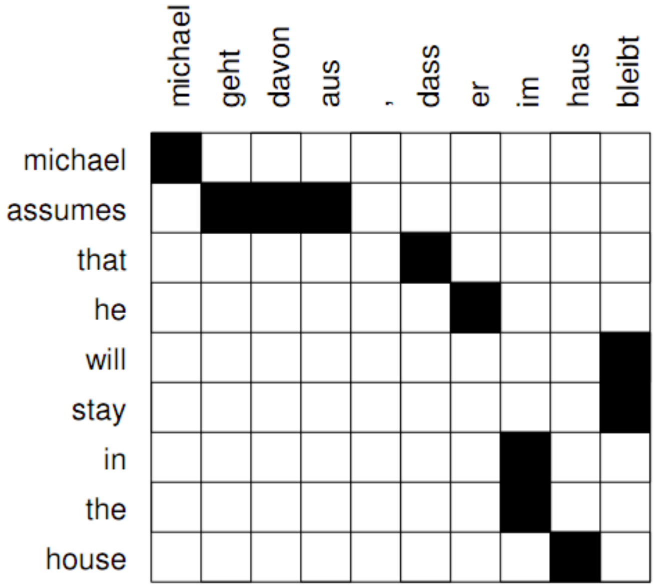
Since Equation 2.1.1 has to be maximized with respect to **ê** and the denominator is independent of this variable, the denominator can be dropped. Then the string **ê** is the solution to Equation 2.1.2.

$$\hat{e} = \operatorname*{argmax}_e Pr(e|f) = \operatorname*{argmax}_e Pr(f|e) \cdot Pr(e) \tag{2.1.2}$$

If we were to consider the probability P(e | f), then starting from a well-formed string f we would search through an immense space of well-formed and ill-formed strings e. Since the probability mass concentrated on the well-formed strings e would be too small we will not be able to find a good enough translation. By solving P(f | e) we are not concerned with how to probability mass is distributed over ill-formed or well-formed f strings, since the maximization is done with respect to e. That is the e maximizing P(f | e) for a given f is also maximizing c*P(f | e) if the probability mass is actually higher for that f.

According to Equation 2.1.2, the translation system consists of two components. The first component called *the language model* is model by $P(e)$ and the second component called *the translation model* is modeled by $P(f|e)$. The translation model is responsible for selecting an *adequate* string of words **ê** that could produce **f**, and the language model selects the strings **e** which have a $fluent$ word order in the target language. The estimation of the translation model from a parallel corpus is described in this chapter, while the language model will be described in a later chapter.

Given a pair of word strings **e** and **f** that are translations of each other, a word-based translation model tries to find a correspondence between words in **e** and words in **f**



	michael	geht	davon	aus	,	dass	er	im	haus	bleibt
michael	■									
assumes		■	■	■						
that						■				
he							■			
will										■
stay										■
in								■		
the								■		
house									■	

Figure 2.1: *Alignment matrix*. Alignments between words are indicated by the shaded *alignment points*. (Koehn 2009)

called *alignment*. The *alignment* indicates which words in **e** were responsible for producing some particular words in **f**. One way to represent word alignments is by an *alignment matrix* like the one shown in Figure 2.1. In the alignment matrix the alignments between words are indicated by the shaded *alignment points*. Adjacent alignment points indicate that a word in one string corresponds to several words in the other string. For example, in Figure 2.1 there is a so called *one-to-many* alignment between the word *assumes* in the English string and the group of words *geht davon aus* in the German string.

The total number of possible alignments for a pair of strings of lengths l and m is 2^{lm} since each of the l words can correspond to any of the m words and there are 2^{lm} ways of combining these correspondences to form an alignment. The conditional probability of string **f** having been produced by **e** can be written in terms of the possible alignments between the two strings as in Equation 2.1.3

$$P(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e}) \tag{2.1.3}$$

Depending of the complexity of the considered pair, a word in **e** can produce one or more words in **f**. The number of corresponding words in **f** is called the $fertility$ of **e**. The positions of the words matched from **e** and **f** can also be different depending on the word order of the considered languages. For example the order of adjective and nouns or the order of verb and object can change in the target language. The change of word positions in the target string is called *distortion* and can also be modeled within the translation model. In other words, to generate a string **f** and an alignment **a** from a string **e**, the length of **f** is chosen considering only the string **e**, then the positions of words in string **f** are chosen considering the length of string **f**, the string **e** and the previously selected positions and finally the actual words in string **f** are generated considering the position of word, the previously generated words, the length of **f** and the string **e**. The generative story is comprised in Equation 2.1.4 that estimates the joint likelihood of producing the string **f** and alignment **a** from a string **e**.

$$P(\mathbf{f},\mathbf{a}|\mathbf{e}) = P(m|\mathbf{e}) \prod_{j=1}^{m} P(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) P(f_j|a_1^{j}, f_1^{j-1}, m, \mathbf{e}) \tag{2.1.4}$$

The number of parameters in Equation 2.1.4 is too large to allow estimating all parameters independent of each other, therefore several assumptions are made about the parameters. The assumptions about the independence of parameters is what differentiates several word-based models, namely the **IBM models**, that are described in sections 2.1.1, 2.1.2 and 2.1.3.

2.1.1 IBM Model 1

IBM Model 1 assumes that m, the variable describing the length of string **f**, can take any value with equal probability, in other words $P(m|\mathbf{e}) = \epsilon$, where ϵ is a predetermined constant. The positions of the words generated for **f** are determined by conditioning the probability $P(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e})$ only on the variable l, the length of string **e**, making this probability equal to $(l+1)^{-1}$. The probability of generating the particular words of **f** will only depend on the words themselves and their positions, and this *translation probability* is denoted by $t(f_j|e_{a_j}) \equiv P(f_j|a_1^{j}, f_1^{j-1}, m, \mathbf{e})$. By making these assumptions Equation 2.1.4, that estimates the joint probability of generating string **f** with the alignment **a** given the string **e**, can be rewritten as Equation 2.1.6.

$$P(\mathbf{f},\mathbf{a}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.1.5}$$

Since the index of a_j ranges from 1 to m and the possible values can range from 0 to l, the likelihood of **(f | e)** can be written as:

$$P(\mathbf{f}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \dots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.1.6}$$

The *translation probabilities* $t(f|e)$ have to be estimated such that they maximize the likelihood of **(f | e)**. In order to estimate $t(f|e)$ Equation 2.1.6 is further mathematically transformed to allow an iterative algorithm to solve it. The iterative algorithm, called the *EM algorithm*, estimates at each step the number of times a word f is aligned to a word e in the given pair $(\mathbf{f}|\mathbf{e})$, based on the estimates of the previous step. Because the number of possible alignments is $(l+1)^m$ it would be inefficient to make these estimates, but with some mathematical observations (Brown et al. 1993) show it is possible to efficiently estimate these counts with a computational effort proportional to $l+m$.

2.1.2 IBM Model 2

While IBM Model 1 allowed words to be placed in any position of string **f** with equal probability, **IBM Model 2** improves by conditioning the probability $P(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e})$ on the index of the corresponding word in string **e** and on the lengths m and l of the two strings. Therefore the *alignment probability* of generating the *j-th* word in string **f** in a particular position a_j is:

$$a(a_j|j,m,l) \equiv P(a_j|a_1^{j-1}, f_1^{j-1}, m, l) \tag{2.1.7}$$

By introducing the *alignment probability* the likelihood of **(f | e)** becomes:

$$P(\mathbf{f}|\mathbf{e}) = \epsilon \sum_{a_1=0}^{l} \ldots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) a(a_j|j,m,l) \tag{2.1.8}$$

The parameters learned for IBM Model 1 are used to initialize the parameters of IBM Model 2 and compute the first estimated counts used in the iterative learning algorithm.

2.1.3 IBM Model 3-5

The **IBM Models 3,4,5** take into consideration that the correspondence between words is not always *one-to-one* and introduce in the generative story the $fertility$ of words in string **e**. Depending on the $fertility$ of a word in string **e**, several words are selected in string **f**, forming a *cept*, that will be modeled by a random variable τ. If the $fertility$ of a word is 0 that accounts for inserting words in the translated string that don't have a correspondent in the source string. The order in which the words of a *cept* appear in the generated string **f** is modeled by a random variable π. There are several possibilities for choosing τ and π to obtain the same string **f** with alignment **a**, more precisely $\prod_{i=0}^{l} \phi_i!$

since for every selection of words τ_i there are $\phi_i!$ ways of permuting them. Therefore the likelihood of a pair $\langle \mathbf{f}, \mathbf{a} \rangle$ can be written as:

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \sum_{(\tau,\pi)\in\langle \mathbf{f},\mathbf{a}\rangle} P(\tau, \pi|\mathbf{e}) \tag{2.1.9}$$

In the case of the **IBM Model 3** the assumptions made on the independence of parameters result in the following model parameters: *fertility probabilities* $n(\phi|e_i)$, *translation probabilities* $t(f|e_i)$ and *distortion probabilities* $d(j|i, m, l)$. An additional *NULL* word e_0 will account for words in string **f** that will not be explicitly translated because they don't have a direct correspondence to a word in string **e**. The fertility of the *NULL* word is ϕ_0 and words will be aligned to this word with probability p_0 or to some other word with probability p_1. The probability that ϕ_0 words out of a sequence of ϕ_1^l words will be aligned to the additional word is:

$$P(\phi_0|\phi_1^l, \mathbf{e}) = \begin{pmatrix} \phi_1 + ... + \phi_l \\ \phi_0 \end{pmatrix} p_0^{\phi_1+...+\phi_l-\phi_0} p_1^{\phi_0} \tag{2.1.10}$$

Then the likelihood of generating string **f** depending on string **e** becomes:

$$\begin{aligned} P(\mathbf{f}|\mathbf{e}) &= \sum_{a_1=0}^{l} ... \sum_{a_m=0}^{l} P(\mathbf{f}, \mathbf{a}|\mathbf{e}) \qquad (2.1.11) \\ &= \sum_{a_1=0}^{l} ... \sum_{a_m=0}^{l} \begin{pmatrix} m - \phi_0 \\ \phi_0 \end{pmatrix} p_0^{m-2\phi_0} p_1^{\phi_0} \prod_{i=1}^{l} \phi_i! n(\phi_i|e_i) \times \prod_{j=1}^{m} t(f_j|e_{a_j}) d(j|a_j, m, l) \end{aligned}$$

In the case of Equation 2.1.11 there is no mathematical transformation than can be applied in order to efficiently evaluate it, as in the case of Model 1 and Model 2. For this reason only the most probable alignments will be considered as summands of this equation. Initially the most probable alignment is selected and then small changes are applied to this alignment in order to reach other more probable alignments.

A problem with Model 3 is that the *distortion probabilities* don't take into account the positions that have been filled by previous words. That is why for some strings, several words can have the same position, while some positions remain empty. The model is called $deficient$ since it wastes probability mass on such strings.

Another aspect not considered in Model 3 is that words that form phrases tend to be displaced as a unit. In **IBM Model 4** such movement is modeled by selecting a *head* for each *cept* and writing the *distortion probability* in terms of two new parameters: one is the probability of assigning a position for the *head* and the second is the probability of

Model 1 - lexical translation model
Model 2 - absolute alignment model
Model 3 - fertility model
Model 4 - relative alignment model
Model 5 - fixes deficiency

Table 2.1: Summary of the contributions of IBM Models (Koehn 2009)

assigning a position for any other word of the *cept*. In order to obtain reliable estimates for these parameters, the words are grouped in *classes* and the parameters depend on word classes instead of individual words. The estimation of parameters of Model 4 also requires selecting a subset of alignments and makes use of estimated parameters of its predecessor, **Model 3**. Much like Model 3, Model 4 is also *deficient* and allows in addition for words to be positioned outside the limits of the string **f**.

IBM Model 5 solves the deficiencies of Model 3 and Model 4 by only allowing words to be placed in positions that have not been already attributed to previous words. In order to estimate the parameters of Model 5 even less alignments are considered. The counts for Models 2,3,4 are computed considering only the alignments selected for Model 5 and used to estimate the parameters of this model.

The main limitation of the IBM Models is that they only allow *one-to-many* alignments. The *phrase-based model* discussed in the section 2.2 will build on the IBM models alignments to finally model *many-to-many* alignments. Koehn (2009) summarizes the contributions of each of the IBM models as in Table 2.1. In order to improve the alignments and to avoid the issue that a target word cannot be aligned with several source words, alignments are computed in both translation directions. The alignment points of the resulting alignments can be intersected or taken all together, resulting in better alignments. This process is called *symmetrization of word alignments* (Koehn 2009).

2.2 Phrase-Based Models

Phrase-based models are state-of-the art statistical translation models. An example of a statistical machine translation system that uses such a model is Moses (Koehn et al. 2007), the system on which the work presented in this thesis was built on. This section introduces the main concepts behind the phrase translation model and the distortion model that can capture how some phrases are reordered more often than others.

2.2.1 Phrase Translation Model

In general translations require *many-to-many* alignments between words. This means that a group of words in the source language should be translated by a group of words in the target language and there might not be a word-level correspondence between these groups. In *phrase-based models* the smallest unit used for translation is called a *phrase*, which is a group of words that will usually not form a linguistic constituent such as NP or VP. By translating phrases instead of words, the problem of word ambiguity is partially solved, and there is no need to directly model $fertility$ or arbitrary insertion and deletion of words.

The translation model considers the sentences to be split into phrases, and all ways to split the sentences are equally likely. The likelihood of generating string **f** given string **e** is then factored in terms of probabilities of generating a phrase $\bar{f}_i$ given a phrase $\bar{e}_i$:

$$P(\mathbf{f}|\mathbf{e}) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1}) \tag{2.2.1}$$

The distortion model accounts for the movement of words as a unit, assigning a probability to the length of the displacement for a given phrase $d(a_i - b_{i-1})$, where a_i and b_i are notations for the starting and ending position of the phrase generated by $\bar{e}_i$.

Koehn et al. (2003) investigate several methods for learning phrase translation pairs and find that their method of processing heuristically the word-based alignments gives the biggest improvement of the translation quality. The first step of the phrase extraction algorithm is the symmetrization of word alignments obtained with the IBM Models. Then phrase pairs are selected if they are *consistent with an alignment*, meaning that the words of the phrase pair have no alignment points with words not considered in the phrase pair. The phrase pairs can be extended by adding neighboring words that are unaligned.

After the phrase translation pairs have been extracted, the translation probabilities are estimated using relative counts as in Equation 2.2.2.

$$\phi(\bar{f}_i|\bar{e}_i) = \frac{count(\bar{f}_i, \bar{e}_i)}{\sum_{\bar{f}} count(\bar{f}_i|\bar{e}_i)} \tag{2.2.2}$$

An estimate for the quality of a phrase translation pair $p_w(\bar{f}_i|\bar{e}_i, a)$, called *lexical weight*, is computed based on how probable are the translations between the words of the pair. The lexical weight is introduced in the translation model and its importance is defined by the parameter λ:

$$P(\mathbf{f}|\mathbf{e}) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1}) p_w(\bar{f}_i|\bar{e}_i, a)^{\lambda} \tag{2.2.3}$$

2.2.2 Distortion Model

The distortion model accounts for movement of words as a unit. When the word order in the source language is very different from that in the target language, then the complexity of reordering the translated phrases increases. The distance based distortion model generally penalizes long distance movement of words. This becomes a problem when translating for example from German which allows verbs to appear at the end of the sentence. A more informed distortion model that is used with phrase-based SMT is the *lexicalized model*. The lexicalized model conditions reordering on the phrase translation pair and therefore captures how some phrases have to be reordered more often than others. This model will predict the *orientation* of a given phrase. To reduce the complexity of the model the possible orientations of a phrase are limited to *monotone* meaning the original order is preserved, and *swap* meaning the current phrase swaps positions with the previous phrase. The parameters of the model are estimated using maximum likelihood as in Equation 2.2.4.

$$P_o(\text{orientation}|\bar{f}_i, \bar{e}_i) = \frac{count(\text{orientation}, \bar{e}_i, \bar{f}_i)}{\sum_o count(o, \bar{e}_i, \bar{f}_i)} \tag{2.2.4}$$

2.3 Language Models

The role of the language model is to give an estimate of how probable a sequence of words is to appear in the target language. The language model helps the translation system with selecting words or phrases appropriate for the local context and with combining them in a sequence with better word order. The most common approach to language modeling is to estimate the probability of a word conditioned on a window of preceding words called the *history*. These types of language models are called *N-Gram Language Models*. In order to have a reliable estimate for the language model probabilities, the context of the language model is usually restricted to a few words. In this case the language model will not be able to capture long distance movement of words like in the case of German verbs with separable prefixes. Another restriction on the context of language models is that the probability of words is computed within the boundaries of a sentence. A language model will be presented in Chapter ASB, that can model the probability of words in a sentence conditioned on the words in previous sentences.

The main aspects of estimating *N-Gram Language Model* probabilities are discussed in the rest of this chapter.

2.3.1 N-Gram Language Models

An *N-Gram Language Model* expresses the probability of a sequence of words $w_1, ..., w_k$ as a *Markov chain*, a product of conditional probabilities that take into account the history of a word:

$$P(w_1, .., w_k) = P(w_1)P(w_2|w_1)...P(w_n|w_1...w_{k-1}) \tag{2.3.1}$$

The *Markov assumption* is made that the probability of a word is affected only by a few preceding words, therefore the history is limited to n words:

$$P(w_1^k) = \prod_{i=1}^{k} P(w_i|w_{i-n+1}^{i-1}) \tag{2.3.2}$$

where $w_1^k = w_1, .., w_k$ and the parameters of the model are the *n-gram* probabilities $P(w_i|w_{i-n+1}^{i-1})$. The n-gram probability is roughly an estimate of how often those words are encountered in a corpus. The *maximum likelihood estimate* of this probabilities results in the normalized count of that n-gram:

$$P(w_i|w_{i-n+1}^{i-1}) = \frac{count(w_{i-n+1}^{i})}{\sum_{w_i} count(w_{i-n+1}^{i})} \tag{2.3.3}$$

To give an example using the Europarl corpus and considering a 4-gram language model, the probability of P(Council | President of the) conditioned on a history of 3 words can be computed as:

$$P(Council|President\ of\ the) = \frac{count(President\ of\ the\ Council)}{\sum_{w_i} count(President\ of\ the\ w_i)} = \frac{2884}{25209} = 0.114 \tag{2.3.4}$$

But if the language model is required to give an estimate for *P(council | President of the)*, the count *c(President of the council)* turns out to be zero and then any product containing this term will be zero. In general an n-gram language model with a vocabulary of size V has to estimate $V^n - 1$ independent parameters. Many of these parameters will correspond to n-grams not seen in the training data that receive zero probability or to n-grams with low counts that lead to poor estimates . The issue of improving insufficient statistics is addressed in section 2.3.2 where *smoothing* is discussed.

Finally a measure is required for evaluating and comparing language models. Evaluating a language model as part of a system is too computationally expensive, therefore a measure called *perplexity* is used to estimate how well the language model recognizes some test data independently of the task it is used for. The perplexity of a model p is related to the cross-entropy of that model $H_p(T)$ on the test data T:

$$PP_p(T) = 2^{H_p(T)} = P(T)^{-1/|T|} \tag{2.3.5}$$

where $P(T)$ is the probability of the entire test set, computed as the product of the probabilities of all the sentences in the test set.

2.3.2 Smoothing

To avoid assigning zero or low probabilities to unseen events, the maximum likelihood estimates for the conditional probabilities are adjusted to make the distribution more uniform. By doing so, zero and low probabilities are increased and high probabilities are decreased. By applying this type of technique called *smoothing*, the estimations are improved and unseen n-grams receive some probability. Chen and Goodman (1996) make an extensive evaluation of different smoothing techniques and propose a version called *Modified Kneser-Ney smoothing*. The main ideas behind this technique are introduced in this section.

Linear interpolation

When there is insufficient data, higher-order n-gram models are poorly estimated. Therefore it proves useful to adjust these estimates by interpolating higher-order n-gram models with lower-order n-gram models that have better estimates. Smoothing by *interpolated estimation*, proposed by Jelinek and Mercer (1980), interpolates the n-order model with the $(n-1)$-order model:

$$P_{interp}(w_i|w_{i-n+1}^{i-1}) = \lambda_{w_{i=n+1}^{i-1}} P_{ML}(w_i|w_{i-n+1}^{i-1}) + (1 - \lambda_{w_{i=n+1}^{i-1}}) P_{interp}(w_i|w_{i-n+2}^{i-1}) \tag{2.3.6}$$

where P_{ML} is the fixed maximum likelihood estimate of the probability and $\lambda_{w_{i=n+1}^{i-1}}$ is estimated such that if P_{ML} is reliably estimated its weight will be higher. When the higher-order model has an unreliable estimate, the weight of the lower-order model will be higher and therefore the interpolated model will be falling back to the more reliable estimate.

Absolute discounting

The idea of *absolute discounting* is to take some probability mass from seen n-grams and distribute it to the unseen n-grams. The counts of seen n-grams are discounted by a factor $D \leq 1$ that will increase the weight of lower-order models:

$$P_{abs}(w_i|w_{i-n+1}^{i-1}) = \frac{\max\{count(w_{i-n+1}^{i}) - D, 0\}}{\sum_{w_i} count(w_{i-n+1}^{i})} + (1 - \lambda_{w_{i=n+1}^{i-1}}) P_{abs}(w_i|w_{i-n+2}^{i-1}) \quad (2.3.7)$$

with the weight of the lower-order n-gram estimated as:

$$1 - \lambda_{w_{i=n+1}^{i-1}} = \frac{D}{\sum_{w_i} count(w_{i-n+1}^{i})} N_{1+}(w_{i-n+1}^{i-1}\bullet) \quad (2.3.8)$$

where $N_{1+}(w_{i-n+1}^{i-1}\bullet)$ is the number of unique words that have the history $w_{i=n+1}^{i-1}$. When the count of an n-gram is zero, this model will simply *backoff* to the lower order model.

Kneser-Ney smoothing

The *Kneser-Ney smoothing* modifies Equation 2.3.7 by using a dedicated lower-order back-off distribution for unseen n-grams. The dedicated lower-order distribution accounts for the fact that some words are frequent but only appear in a particular context. Therefore the lower-order distribution for a word should be proportionate to $N_{1+}(\bullet w_{i-n+2}^{i})$, the number of possible histories for that particular word:

$$P_{KN}(w_i|w_{i-n+2}^{i-1}) = \frac{N_{1+}(\bullet w_{i-n+2}^{i})}{N_{1+}(\bullet w_{i-n+2}^{i-1}\bullet)} \quad (2.3.9)$$

The *interpolated* version of the *Kneser-Ney smoothing* is obtained by introducing P_{KN} as the lower-order model in Equation 2.3.7. If the lower-order distribution is used only for computing the probability of unseen n-grams, and does not contribute to the probability of seen n-grams than the model is called a *backoff* model. The modification that Chen and Goodman (1996) bring to the *Kneser-Ney smoothing* is that instead of one discounting parameter D, they use different discount parameters depending on the frequency of the n-gram:

$$D(c) = \begin{cases} 0 & \text{if count=0} \\ D_1 & \text{if count=1} \\ D_2 & \text{if count=2} \\ D_{3+} & \text{if count} \geq 3 \end{cases} \tag{2.3.10}$$

2.3.3 Class-Based Language Models

Another way to have more reliable estimates for n-grams, is to account for the fact that words can be grouped together, either by meaning, syntactic function or considering the similar context in which they appear. Brown et al. (1992) propose mapping the words in the vocabulary to C classes and estimate the parameters of the *n-gram class language model* as:

$$P(w_i|w_{i-n+1}^{i-1}) = P(w_i|c_i)P(c_i|c_{i-n+1}^{i-1}) \tag{2.3.11}$$

where w_i is mapped to class c_i. The first term is called the *emission probability* and the second term will be a product of *transition probability*. The maximum likelihood estimates of these terms are:

$$P(w_i|c_i) = \frac{count(w_i)}{count(c_i)} \tag{2.3.12}$$

$$P(c_i|c_{i-n+1}^{i-1}) = \frac{count(c_{i-n+1}^{i})}{\sum_{c_i} count(c_{i-n+1}^{i})} \tag{2.3.13}$$

Since $P(c_i) = count(c_i)/T$, where T is the total number of words occurring in the training corpus, and each word will be mapped to only one class then:

$$P(w_i) = P(w_i|c_i)P(c_i) = \frac{count(w_i)}{T} \tag{2.3.14}$$

By grouping words into word classes the number of independent parameters reduces to $C^n - 1 + V - C$, $V - C$ for the emission probabilities and $C^n - 1$ for the transition probabilities, making the estimates of these parameters more reliable. The algorithm proposed by Brown et al. (1992) for mapping words to classes does not make use of any linguistic information, instead it groups words according to the statistical similarity of the contexts they appear in and maximizes the likelihood of the language model. An implementation of this algorithm is provided with the SRILM toolkit Stolcke (2002) and an example of a word class obtained with this algorithm is given in Table 2.2.

judicial Catholic specialised bureaucratic decentralised presidential supervisory scientific academic capitalist centralised totalitarian Parliamentary liberal parliamentary communist socialist professional sovereign democratic vocational federal colonial creative specialist supranational

Table 2.2: Example of a word class obtained with SRILM toolkit from Europarl

Combining Language Models

In order to draw from the strengths of different language models one can combine them. For example *domain adaptation* involves combining models trained on data from different domains. To help with generalization and unseen n-grams, a word-based model can be combined with a class-based model. One way of combining models is again by *linear interpolation*:

$$P_{LI}(w_i|w_{i-n+1}^{i-1}) = \sum_j \lambda_j(w_{i-n+1}^{i-1}) P_j(w_i|w_{i-n+1}^{i-1}) \tag{2.3.15}$$

where the weights of each of the j-th language model are optimized by maximizing the interpolated probability on some *held-out* data.

Klakow (1998) proposes a method for combining language models called *log-linear interpolation (LLI)* and shows that it gives lower perplexities than *linear interpolation* on the tasks of adaptation and combining models of different context length. The *log-linear interpolation* of n-gram models is defined as:

$$P_{LLI}(w_i|w_{i-n+1}^{i-1}) = \frac{1}{Z_\lambda(w_{i-n+1}^{i-1})} \prod_i P_j(w_i|w_{i-n+1}^{i-1})^{\lambda_j} \tag{2.3.16}$$

where Z_λ is a normalization factor and λ_j are the model weights. The normalization factor is needed to make Equation 2.3.16 a probability, but it is expensive to compute since it sums over all parameters of the model:

$$Z_\lambda(w_{i-n+1}^{i-1}) = \sum_{w_i} \prod_j P_j(w_i|w_{i-n+1}^{i-1})^{\lambda_j} \tag{2.3.17}$$

Results in speech recognition have shown that the normalization factor can be left out without affecting much the result of the system. The normalization is needed though for optimizing the model weights, that is done by maximizing the log-likelihood on a held-out test set.

2.4 Evaluation

2.4.1 Automatic Evaluation

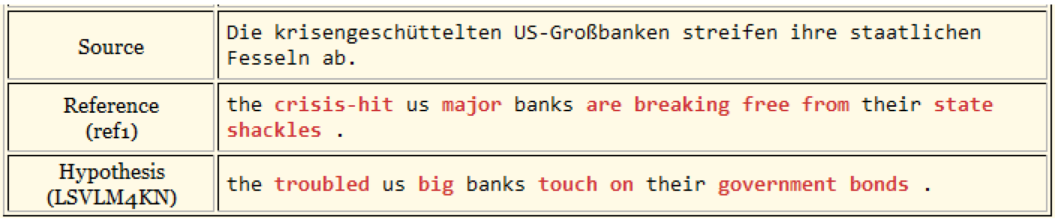
Automatic evaluation of the output of machine translation systems is essential for the development of these systems. Since evaluation is a bottle-neck in the development cycle there has been significant interest in designing an automatic metric that can evaluate both adequacy and fluency of a translation. The most widely used metric for machine translation is the BLEU metric (Papineni et al. 2002). Based on matching n-grams between a proposed translation and a reference translation, the BLEU metric computes a modified n-gram precision. The definition of the metric, considering matching n-grams of order up to 4, is given in Equation 4.5. The first term, called *brevity penalty*, accounts for missing words in the translation and reduces the score if the output is too short.

$$\text{BLEU-4} = \min\left(1, \frac{output - length}{reference - length}\right) \prod_{i=1}^{4} precision_i \tag{2.4.1}$$

In order to avoid 0 n-gram counts the score is usually computed for an entire test set, rather than for individual sentences. Although this metric is a good instrument for evaluating and comparing machine translation systems, it is not able to detect what types of changes in the output of these systems. It is also unable to deal with morphology, synonyms and paraphrases of the reference translation.

2.4.2 Manual Evaluation

Often when trying to improve a machine translation system some particular translation errors are targeted. An automatic metric, such as the BLEU metric, will be uninformative with respect to the types of changes that are of interest. For this reason a good practice is to do an error analysis of the translation output. This thesis follows the framework for error classification proposed by Vilar et al. (2006). The following classes of errors were considered: missing words, word order, incorrect words and punctuation. The *missing words* category accounts for words in the source sentence that are missing from the translation. This category is further split into *content words* and *filler words*. Content words are those words that by their absence would change the meaning of the sentence, such as nouns or verbs. Filler words refer to other words that make the sentence grammatically correct but do not alter the meaning of the sentence. The *word order* category is concerned, as its name suggests, with the order of words in the translation. It is also split in two categories that account for *word based* reordering and *phrase*



Source	Die krisengeschüttelten US-Großbanken streifen ihre staatlichen Fesseln ab.
Reference (ref1)	the crisis-hit us major banks are breaking free from their state shackles .
Hypothesis (LSVLM4KN)	the troubled us big banks touch on their government bonds .

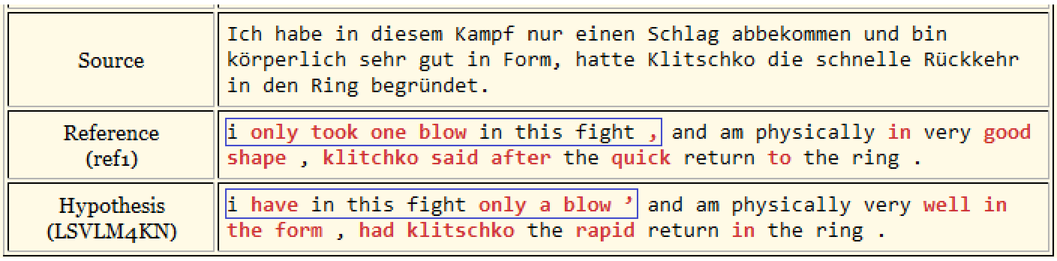
Figure 2.2: Example of a German into English translation with wrong lexical choices.

based reordering. The *incorrect words* category is the most comprehensive and distinguishes between translated words with a *wrong sense* or just a *wrong form*. Words with wrong sense are also split between words that have been incorrectly disambiguated and words that clearly represent a wrong lexical choice and disrupt the meaning of the sentence. In one of the error analysis presented in this thesis, a special attention was given to verbs. Therefore the following sub-categories were added: missing content words that are main verbs, missing content words that are auxiliary verbs, wrong word order of verbs, incorrect sense of a verb that fails to select its agent and patient and wrong verb tense.

2.5 Issues with SMT

Although phrase-based statistical machine translation systems represent the state-of-the-art, there are still many aspects that have to be improved. A major problem is caused by data sparsity which makes the models unable to generalize well to unseen data. A direct consequence is that the translation has wrong lexical choices or fails to correctly disambiguate the words in the given context, as can be seen in Figure 2.2.

The models are also unable to handle differences in word order between languages like German and English. The translation in this case mimics the word order of the source language and often important information is omitted, like the main verb. Figure 2.3 shows an example of such a faulty translation. Designing a better reordering model and allowing the language model to capture the dependencies between words in adjacent sentences and not only those in the current sentence are some possible improvements to the current model, that thesis investigates.



Source	Ich habe in diesem Kampf nur einen Schlag abbekommen und bin körperlich sehr gut in Form, hatte Klitschko die schnelle Rückkehr in den Ring begründet.
Reference (ref1)	i only took one blow in this fight , and am physically in very good shape , klitchko said after the quick return to the ring .
Hypothesis (LSVLM4KN)	i have in this fight only a blow ' and am physically very well in the form , had klitschko the rapid return in the ring .

Figure 2.3: Example of a German into English translation with wrong order highlighted and missing main verbs "took" and "said".

CHAPTER 3

Pre-Reordering Model

3.1 Pre-Reordering the source sentence

German and English have different word order which is problematic for statistical machine translation. Figure 3.1 a) shows an example of how a German sentence can be aligned to an English sentence. Because of the difference in word order, words in the source sentence are aligned to target words which have a different position. Problems such as poor estimates for probabilities of extracted translation pairs appear when the distance between the source position and the target position is too large. Another problem is that neither the local reordering within a phrase pair, the reordering model used as part of the statistical machine translation system nor the language model will be able to capture such a long range movement. For this reason the translations will have an incorrect word order that resembles the word order of the source sentence, and often verbs or other words will be omitted. By pre-reordering the source sentence is brought to a form that has a word order similar to that in the target language, as can be seen in Figure 3.1 b). This will allow a monotonic translation or at least reduce the amount of translated phrases that have to be reordered. Pre-reordering the entire parallel corpus would also increase the confidence of the alignments and improve the translation model.

3.2 Related Work

Collins et al. (2005) make some linguistic observations for German such as: finite verbs have a rigid second position in main clause and infinitive verbs come last in the sentence. Based on these observations they apply a sequence of rules to the parse tree structure of the source sentence to obtain a new word order. Rules such as moving

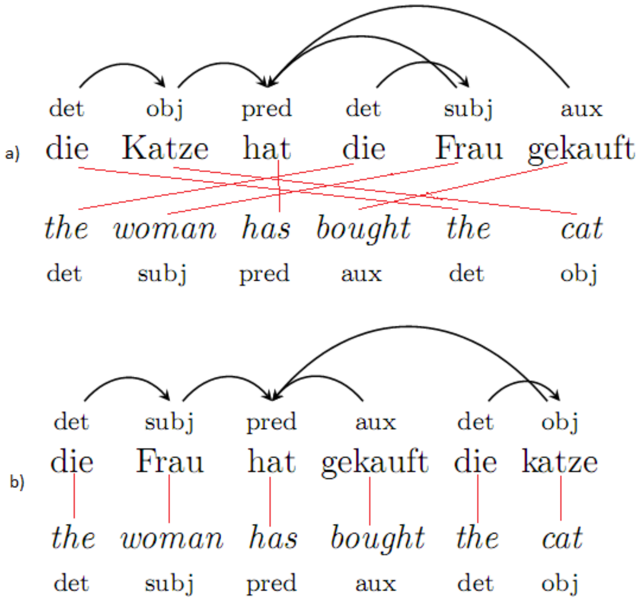

a)
det obj pred det subj aux
die Katze hat die Frau gekauft
the woman has bought the cat
det subj pred aux det obj
b)
det subj pred aux det obj
die Frau hat gekauft die katze
the woman has bought the cat
det subj pred aux det obj

Figure 3.1: a) Alignment between German and English sentences given the original word order. b) Alignment given the pre-reordered source sentence.

the infinitive verb to follow the finite verb or moving the subject to directly precede the head, give the source sentence a word order similar to English, the target language that was considered. This reordering approach is a pre-processing step used in both training and decoding phases of a PBSMT system.

In order to avoid hand-crafting rules based on linguistic observations or using examples of reordered sentences, Dyer and Resnik (2010) train a reordering component as a latent variable in a discriminative translation model using a parallel corpus. The translation model is split in a reordering model and a phrase transduction model. The reordering model takes the CFG representation of the source sentence and generates all permutations of the children of any node to produce the reordering forest. The phrase transduction model uses the reordering forest of the source sentence as a grammar to parse the finite-state transducer (FST) representation of the phrase-based translation model and generates a translation forest with all possible translations of the source sentence. The reordering of the source is considered to be a latent variable in the translation model and is learned such that it produces the best translation with respect to the likelihood of the parallel corpus. The proposed translation model is very appealing since it handles mid-to-long range reordering with the context-free representation and captures local correspondences with the phrase transduction model. It is also convenient to use if there are robust parsers available for the source language, like English, and less resources are available for the target language.

Tromble and Eisner (2009) model word reordering as the Linear Ordering Problem (LOP) and apply this model as a pre-processing step to reorder the source sentences. Since the LOP problem is NP-hard, the authors propose a solution that searches subsets of the permutation space Π_n, using a CKY-parsing-like algorithm. At each iteration the partial solution is improved by making a "local change". The local optimal permutation is searched in the neighborhood $N(\pi)$ of the current permutation π, where neighbors are defined as permutations that can be aligned with an Inversion Transduction Grammar (ITG). This model is particularly interesting because it captures the pairwise dependencies between words and therefore seems suitable for modeling syntactic word reordering. For this reasons the present work considers this model for pre-reordering of words. Section 3.3 describes the LOP and the algorithms applied in the present work to pre-reorder the source sentence for machine translation.

3.3 The Linear Ordering Problem

The Linear Ordering Problem is a well studied combinatorial optimization problem that was shown to be NP-hard. It is concerned with finding the optimal total order for a set of objects according to some, possibly inconsistent, pairwise preferences between these objects. Applications of this problem can be found in many fields like scheduling, economics or search engine combination. Tromble and Eisner (2009) proposed a framework for learning to reorder words in a sentence by modeling word ordering as the Linear Ordering Problem. Following this framework, the present work investigates two state-of-the-art linear ordering algorithms as applied to the task of word reordering. Section 3.3.1 describes the theory behind the Linear Ordering Problem. Section 3.3.2 introduces two algorithms that solve this problem, one approximate and the other exact. These algorithms are used to search for the best word order for a given sentence, as part of a learning algorithm. The learning framework is explained in section 3.4 and finally the results of applying the pre-reordering model to machine translation are discussed in section 3.5.

3.3.1 Description of the Problem

Martí and Reinelt (2011) give a definition of the Linear Ordering Problem over graphs. A complete directed graph $D_n = (V_n, A_n)$ is defined over the objects that need to be ordered, which for every pair of nodes i and j has an arc (i, j) and an arc (j, i). An *acyclic spanning tournament* in A_n is a subset of arcs that contains for every pair of nodes i and j either arc (i, j) or arc (j, i), and that doesn't form any directed cycle. If arc weights

c_{ij} are defined for every $i, j \in V_n$, then the linear ordering problem is to compute a acyclic spanning tournament for which $\sum_{(i,j)\in T} c_{ij}$ is maximized. A formulation that is closer to the way word reordering is usually addressed, in terms of permutations, is that used by Tromble and Eisner (2009) and Schiavinotto and Stützle (2004). Given an $n \times n$ matrix B, such that $B[i, j]$ is the score of placing i before j in a permutation of the n objects, the Linear Ordering Problem is to find the permutation that maximizes:

$$\hat{\pi} = \arg\max_{\pi\in\Pi_n} B(\pi) = \arg\max_{\pi\in\Pi_n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} B[\pi_i.\pi_j] \tag{3.3.1}$$

where Π_n is the set of all permutations with $|\Pi_n| = n!$. The score of the entire permutation is factorized as the sum of scores of the $\binom{n}{2}$ pairwise orderings. In the case of word reordering, the permutations are defined over words in a given sentence. The entries in matrix B will define the preferred order of any two words w_i, w_j in a sentence. For example if $B[w_i, w_j] > B[w_j, w_i]$ then word w_i should be placed before w_j in the reordered sentence. In the present work we considered these entries to have integer values and the matrix to be symmetric, meaning that $B[w_i, w_j] = -B[w_j, w_i]$. The scores of the pairwise orderings are computed based on the features of the considered pair:

$$B_w[i, j] = \theta \cdot \phi(w, i, j) \tag{3.3.2}$$

where ϕ is a vector of features of the entire sentence and θ is the feature weight vector. Learning a word reordering model will involve learning the weights of the features required for computing the preference matrix. In order for the learning algorithm to learn these weights, it will need to search for the optimal permutation given the current preference matrix. The next section introduces two algorithms that solve the Linear Ordering Problem and therefore find that optimal permutation.

3.3.2 Algorithms

Martí and Reinelt (2011) make an extensive survey of algorithms for the Linear Ordering Problem. According to the benchmarks they used, the algorithm that finds the largest number of optimal solutions and is most efficient is the *Memetic algorithm (MA)* proposed by Schiavinotto and Stützle (2004). *MA* is a hybrid meta-heuristic method that combines a genetic algorithm with local search. By comparing the *MA* with a classical genetic algorithm, Martí and Reinelt (2011) conclude that the inclusion of local search to improve solutions makes the *MA* a better algorithm. The *MA* starts with a *population of individuals*, obtained by generating a number of distinct random permu-

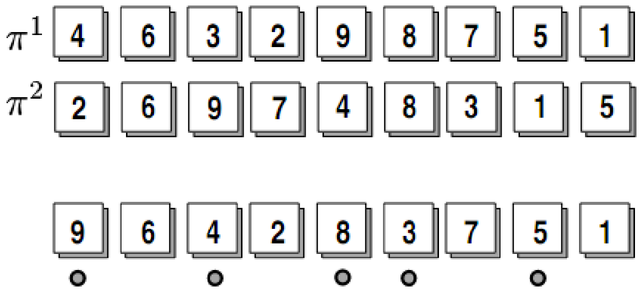


Figure 3.2: Order Based crossover operator. The reordered positions are marked by a circle. Schiavinotto and Stützle (2004)

tations and improving them with local search. The population is diversified at each iteration, by applying *crossover* and *mutation* operators to some of its individuals. The individuals to which these operators are applied to are selected randomly according to a uniform distribution. To generate a new individual, also called an *offspring*, the *crossover* operator combines parts of two existing individuals, and the *mutation* operator will interchange two positions within the same individual. The particular *crossover* operator that was used in word reordering experiments is the *order based (OB)* operator. The OB operator copies the first individual to the offspring, then selects some random positions and reorders them according to their order in the second individual. An example of how this operator works can be seen in Figure 3.2

In the next iteration, the new population will consists of the solutions obtained by applying local search to the new individuals and some of the best individuals from the previous population. There are several possible local search methods that can be used by the *Memetic algorithm*, one of which is based on the *first improvement interchange operator*. This method searches in the neighborhood of a permutation, N_x, that can be generated by interchange moves. The *interchange* operator maps a permutation to another permutation and is defined as:

$$interchange(\pi, i, j) = (..., \pi_{i-1}, \pi_j, \pi_{i+1}, ..., \pi_{j-1}, \pi_i, \pi_{j+1}, ...) \tag{3.3.3}$$

The size of the search space is therefore $|N_x| = n(n-1)/2$, but this particular search method will stop after it finds the first improving solution. Different types of crossover and mutation operators and local search methods are described in Schiavinotto and Stützle (2004), but were not applied to word reordering in the present work. Finally, if the population is not improved after some fixed number of iterations, the best individual is kept and a new population is initialized. The algorithm generates and improves the population for several iterations, until a stopping criteria is met and then the best individual is returned as the local optimal permutation. The *MA* proves to be very efficient for large problems, where other algorithms find fewer and not as good local optimal solutions, or are too slow. In the case of small to moderately sized problems, with n that can go up to 50, an exact algorithm will find a global optimal solution and can still be efficient. The *Branch-and-Bound (BB)* algorithm proposed by Charon and Hudry (2006) is an exact algorithm that was also evaluated by Martí and Reinelt (2011), who show that the algorithm efficiently solves moderately sized problems with $n = 40$. The *Branch-and-Bound* algorithm tries to split the problem into subproblems for which lower and upper bounds can be computed more efficiently, and that can be used to eliminate a large part of the solution space. These subproblems are relaxations of the initial problem and their solutions give an upper bound on the optimum objective function values of the original problem. If the relaxed problem is well chosen, then the search space will be significantly reduced, otherwise a branch-and-bound algorithm can end up enumerating most of the possible solutions. Charon and Hudry (2006) apply Lagrangian relaxation to the original problem in order to compute an upper bound for it. The Linear Ordering Problem is first formulated as an integer linear problem:

$$\begin{aligned} \max \sum_{(i,j)\in A_n} & c_{ij}x_{ij} \\ & x_{ij}+x_{jk}+x_{ki},\ for\ 1 \leq 2,\ for\ all\ distinct\ nodes\ i,\ j,\ k\ \in V_n, \\ & x_{ij}+x_{ji}=1,\ for\ 1 \leq i<j \leq n, \\ & x_{ij} \in 0,1,\ for\ 1 \leq i,j \leq n,\ i \neq j. \end{aligned} \tag{3.3.4}$$

If T denotes the set of all triples (r,s,t) such that $r < s$ and $r < t$ and $\mu \geq 0$ is a vector of Lagrange multipliers, then the Lagrangian relaxation becomes:

$$\begin{aligned} L(\mu) = & \max \sum_{(i,j)\in A_n} c_{ij}x_{ij} + \sum_{(r,s,t)\in T} (2 - x_{rs} - x_{st} - x_{tr})\mu_{rst} \\ & x_{ij}+x_{ji}=1,\ for\ 1 \leq i<j \leq n, \\ & x_{ij} \in 0,1,\ for\ 1 \leq i,j \leq n,\ i \neq j. \end{aligned} \tag{3.3.5}$$

The best bound given by the relaxed problem can be found by solving the *Lagrangian dual problem*:

$$\min_{\mu} L(\mu) \tag{3.3.6}$$

In the present work, the *Branch-and-Bound* algorithm was used without modifications and the *Memetic Algorithm* was subject to some restrictions that will be mentioned in section 3.3.1.

```
Averaged Perceptron
1. $w_0 \leftarrow 0$
2. for r=1 to R do
3.     for i=1 to N do
4.         $BP^i \leftarrow ORACLE(P^i)$
5.         $PP^i \leftarrow \text{argmax}_{PP \in GEN(P^i)} \phi(PP) \cdot w_j$
6.         if $PP^i \neq BP^i$ then
7.             $w_{j+1} \leftarrow w_j + L \cdot (\phi(BP^i) - \phi(PP^i))$
8. return $w = \sum_j w_j$
```

Figure 3.3: The Averaged Perceptron algorithm.

3.4 Learning Framework

3.4.1 Learning algorithm

As was mentioned in section 3.3, learning a reordering model involves learning the weights of the features used to compute the pairwise preference matrix of the Linear Ordering Problem. The algorithm that was used to learn the feature weights is the *Averaged Perceptron* (Freund and Schapire 1998, Collins 2002).

The Averaged Perceptron algorithm is an on-line learning-algorithm that was successfully applied to structured prediction tasks and is robust to approximate inference. The pseudo-code for the algorithm is given in Figure 3.3. As indicated by lines two and three, the algorithm runs repeatedly through the entire training data, and at each epoch it processes the training examples one-by-one. In the case of word reordering, a training example consists of all the word pairs in the sentence that has to be reordered, represented by their features. For each training example, the algorithm will predict which is the optimal permutation based on the features of the word pairs and the current weight vector. This step corresponds to line five where *PP* is the predicted permutation. If the predicted permutation is different from the true optimal permutation of that example, identified in line four as an *oracle* permutation, then the weight vector is updated. The features of a word pair that appeared in the predicted permutation, but not in the true optimal permutation, are decreased by an amount proportional to the *learning rate* and those of a word pair that was in the true optimal permutation, but not in the predicted permutation are increased by the same amount. These last actions correspond to lines six and seven in the pseudo-code. The final weight vector will be an average of all intermediate weight vectors, as computed on line eight, and together with the feature representation will represent the reordering model. If the training data is linearly separable then the algorithm will converge in a finite number of epochs.

For the first epoch of training, the model is initialized with the log-odds of the parameters as in Tromble and Eisner (2009). If Φ_m^{swap} is the feature set of word pairs that swap their order according to the true optimal permutation, and $\Phi_m^{\sim swap}$ is the feature set of word pairs that keep their initial order, then the initial model is:

$$w_m = log\left(|\Phi_m^{swap}| + \frac{1}{2}\right) - log\left(|\Phi_m^{\sim swap}| + \frac{1}{2}\right) \qquad (3.4.1)$$

3.4.2 Oracle Reordering

The true optimal permutation, that the reordering model is trained to predict, is obtained heuristically from the word alignments. A word in the reordered source sentence will have the position that its corresponding target word has in the target sentence. If two source words are aligned to the same target word, then their initial relative order is preserved. If a source word is aligned to two target words, than the preferred position will be that of the right-most target word. This choice was made due to the right-branching character of both English and German, which can be exemplified by noun phrases, where a determinant will always precede the head noun, or prepositional phrase, where the preposition will precede the head noun. Source words that are not aligned to any target words will be grouped with the next aligned source word, and if there isn't such a word, then with the previous aligned source word. In contrast, Tromble and Eisner (2009) prefer to align source words to the left-most corresponding target word, and unaligned words will be grouped at the beginning of the sentence.

3.4.3 Features

The features that were used to compute the pairwise preference scores are global features based on the part-of-speech of words from the entire sentence. A feature $\phi(w, l, r)$ is a binary feature that fires when some properties of the word pair and of the sentence are present. These type of features were initially used for dependency parsing (Mcdonald et al. 2005) and then adapted for word reordering by (Tromble and Eisner 2009). The features are grouped in templates that consider, for example, the parts-of-speech of the word pair, of the word immediately preceding the left token and of the word immediately succeeding the right token. These templates are also extended with the binned distance between the right and left token of the considered word pair. Features based on dependency relations were also considered. Some features were also used by (Tromble 2009) to extend the reordering model but did not lead to an improvement in the machine translation output. In addition to those dependency features, the present work also considered the direction of attachment to the head and combined the dependency information with part of speech information. These further efforts to guide the reordering model with syntactic information have not improved the machine translation output, and possible reasons for this are discussed in section3.5. Lexical features were not considered since the number of features would become very large and the reordering model would become more domain dependent. All the considered feature templates are shown in Table 3.1 and Table 3.2 and the notation used to describe the features is explained next. t_l and t_r are the part-of-speech tag of the left token and respectively right token in the word pair, where l and r denote the index in the sentence. t_b is the part-of-speech of any word between the left and right tokens, $l < b < r$. All templates are also extended with the distance $r - l$, which is binned into 1,2,3,4, > 5 && <10 and > 10 . The dependency features are explained bellow:

1. *left parent* is the dependency relation between the left token and right token, when the right token is the head

2. *right parent* is the dependency relation between the left token and right token, when the left token is the head

3. *left sibling* is the dependency relation between the left token and its head, when the left token and right token have the same head

4. *right sibling* is the dependency relation between the right token and its head, when the left token and right token have the same head

5. *left dir* direction of attachment of the left token to its head

6. *right dir* direction of attachment of the right token to its head

7. $t_{left\ head}$ part-of-speech of the head of the left token. Activated only when the left and right tokens have different heads.

8. $t_{right\ head}$ part-of-speech of the head of the right token. Activated only when the left and right tokens have different heads.

These features were also extended with the binned distance $r - l$.

t_{l-1}	t_l	t_{l+1}	t_b	t_{r-1}	t_r	t_{r+1}
	●				●	
	●					
					●	
	●		●		●	
	●	●		●	●	
	●	●			●	
	●	●			●	●
	●				●	●
●	●				●	●
●	●				●	
●	●			●	●	
	●			●	●	

Table 3.1: Feature templates based on part-of-speech information.

left parent	*right parent*	*left sibling*	*right sibling*	*left dir*	*right dir*	$t_{left\ head}$	$t_{right\ head}$
●							
	●						
		●					
		●		●			
			●				
			●		●		
		●	●				
		●	●	●	●		
●						●	
	●						●
●	●					●	●
●				●		●	
	●				●		●
●	●			●	●	●	●

Table 3.2: Feature templates based on dependency information and part-of-speech information. The right side of the table contains features that were proposed in the present work in order to capture syntactic movement.

3.4.4 Distributed Training

One aspect of the Perceptron algorithm that has to be considered when applying it to structured prediction, where the inference algorithm can be very costly, is how the size of the corpus affects the training time. Since the algorithm processes the training examples one by one and inference is non-linear in the length of a training example, if the training corpus is very large then training time becomes unmanageable. (McDonald et al. 2010) analyze a distributed version of the Perceptron algorithm, based on *parameter mixing*. This approach splits the training corpus into smaller disjoint parts and trains models in parallel for each of them. The final model is a weighted mixture of the parameters of each model. The algorithm will converge on an individual subset of the training corpus if the data is linearly separable, but overall it might not separate the entire data even if the training corpus is separable. *Parameter mixing* was used in the present work, with all the models having an equal weight of one. The problem of the entire corpus not being separated by the final model was not an issue, since the algorithm was stopped before it converged on the subsets of the corpus because of efficiency reasons that will be further discussed in section 3.5.

3.5 Experiments

The first experiments described in this section were carried out in order to set the parameters of the learning algorithm. After explaining the choices made for these parameters, a monolingual evaluation of the reordering model is presented in terms of BLEU score, that measures how close the resulting reordered sentences are to the Oracle reordering. The reordering model is then applied as a pre-processing step in the machine translation pipeline. The machine translation output is evaluated with the BLEU metric and by making an error analysis.

3.5.1 Experimental Setup

The corpus used for training, tuning and testing of both the pre-reordering model and the translation models, was Europarl v6 (Koehn 2005). The pre-reordering model was applied as a pre-processing step of the English source sentences. The pre-reordered source sentences were matched with their corresponding German target sentences and used in a standard configuration of the Moses phrase-based SMT system. The standard configuration used a 4-gram language model trained with the SRILM toolkit on the target side of Europarl, word alignments trained with GIZA++ in both directions,

Data set	Sentences	Words - English	Words - German
training	940850	15376069	15071606
tuning	2000	32442	31968
test	2000	32822	32196

Table 3.3: Statistics of the data sets used in pre-reordering and translation experiments.

with the "grow-diag-final-and" heuristic, the lexicalized bidirectional msd reordering model and minimum error-rate training on the tuning set. The translation direction considered in the machine translation experiments was from English into German. It is interesting to apply word reordering for this translation direction since verbs in the target language have different position than in the source language. Although the word order is flexible in German there is a preferred word order especially of the subject and object as selected by the verb. To reduce the complexity of the pre-reordering model and therefore the training time, only source sentence between 5 and 25 words were used in the experiments. The sentences for the tuning and testing sets were chosen randomly from the Europal corpus. The remaining sentences were used for training the models. The data sets were chosen from the same domain to avoid possible model errors due to the mismatch between training and testing domains. It may be the case that such errors wouldn't occur anyway, since the features used for training the pre-reordering model were not lexicalized. Some statistics about the data are shown in Table 3.3. The training corpus was split randomly and evenly in 20 parts, resulting in smaller sets of approximately 47000 sentences each. The *parameter mixing* version of Perceptron was used because the training was distributed on several computers. The final weight vector was obtained by summing up the individual weight vectors, and applied to reorder the tuning and test sets. The *Oracle reordering* was extracted from word alignments trained with GIZA++ in both directions, with the "grow-diag-final-and" heuristic. To extract the features used to learn the pre-reordering model, the English side of the Europarl corpus was tagged with part-of-speech information using TreeTagger (Schmid 1994) and with dependency structure using an ensemble of parsing models (Surdeanu and Manning 2010) based on the MALT parser (Nivre and Nilsson 2004). The implementation of the BB algorithm was freely available online[1] and the implementation of the MA algorithm was made available by the author.

[1] http://www.enst.fr/ charon/tournament/median.html

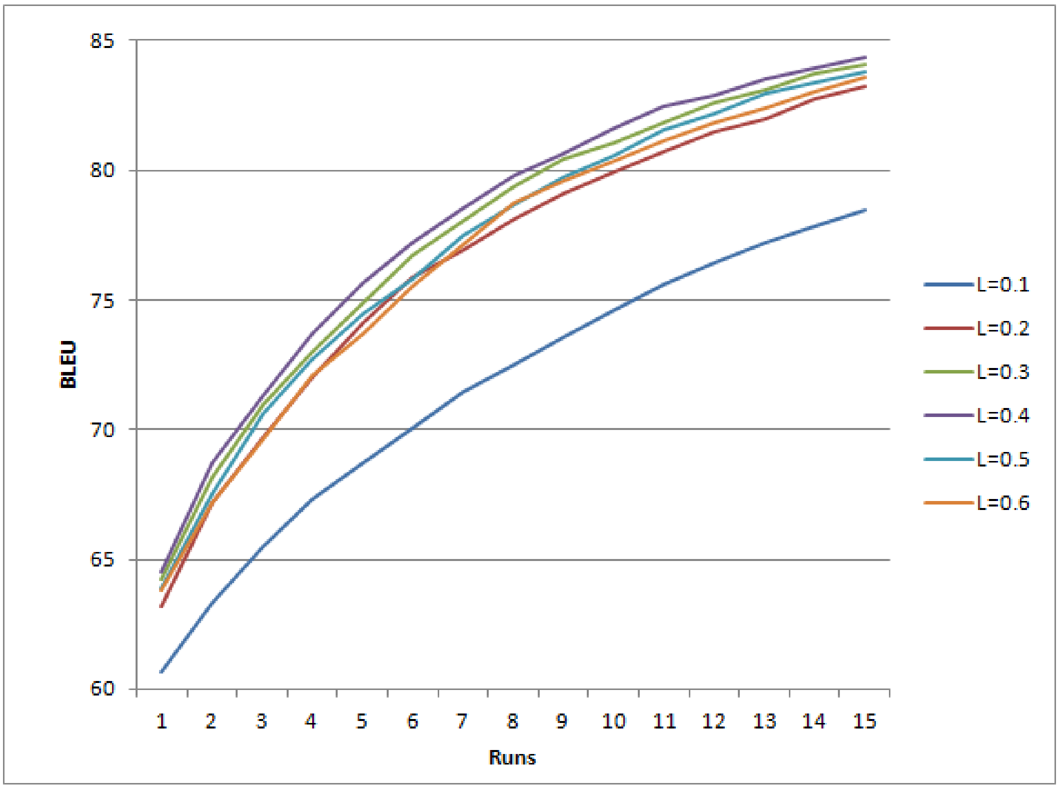


Figure 3.4: Learning curve for Perceptron with different learning rates. The search algorithm used is the Memetic algorithm and the training set has 2000 sentences.

3.5.2 Monolingual Evaluation

The results presented in this section motivate the selection of parameters for the learning algorithm. The Averaged Perceptron algorithm has several parameters that were optimized for the task at hand, taking into consideration the constraints on available resources. These parameters are: the size of the training corpus T, the number of epochs R, the learning rate L and the search algorithm. The models evaluated in this section were trained using only part-of-speech features.

The first experiments were conducted in order to choose the learning rate L. The learning rate affects how fast the algorithm convergences. A small L would lead to a slow convergence but possibly to a more accurate model, while a larger L would give a fast start but could make the model oscillate and finally also converge slowly. The value for this parameter was chosen as a trade-of between a fast improvement and a more accurate model, such that after a small number of epochs the model learned would result in higher monolingual BLEU scores. The learning curve of the Averaged Perceptron for different values of the learning rate is presented in Figure 3.4. The inference algorithm used is the Memetic algorithm.

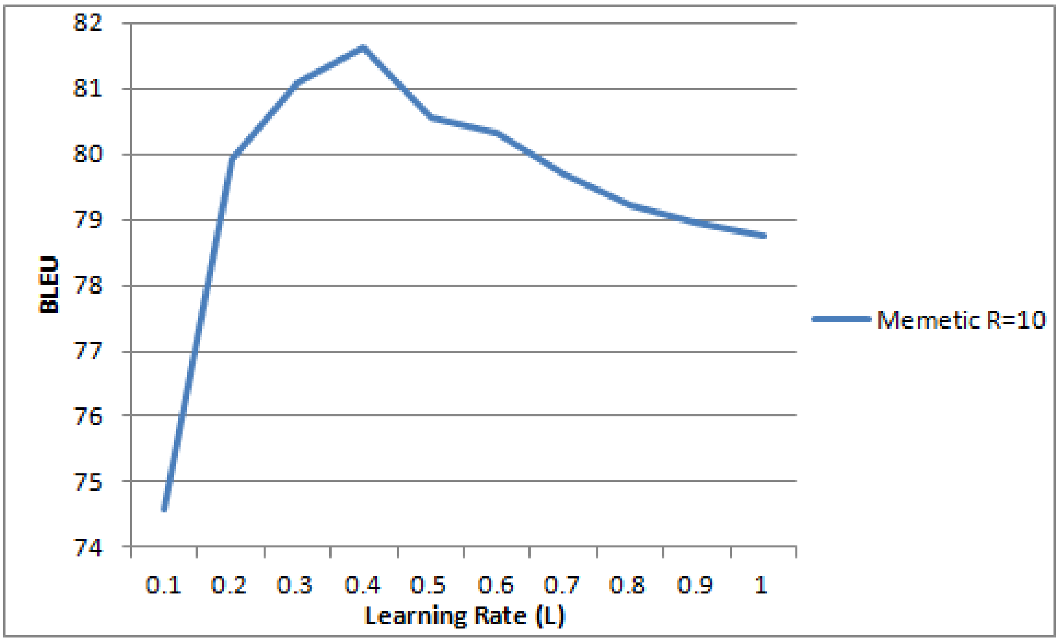


Figure 3.5: BLEU score after 10 runs for different learning rates. The search algorithm used is the Memetic algorithm and the training set has 2000 sentences.

Figure 3.4 shows that for learning rates of 0.3 and 0.4 the monolingual BLEU score increases faster, reaching after 10 epochs values of 81 and 81.5 respectively. Although the BLEU score is slightly lower when the learning rate is 0.3, this value was preferred in order to keep the models trained on subsets of the training corpus from varying too much. Further experiments will keep L fixed at 0.3. Figure 3.5 takes a closer look at how the learning parameters affect the final monolingual BLEU score reached after 10 epochs. It is evident that $L = 0.1$ is a value that is too small, and the model improves very slowly, while at the other end values larger than $L = 0.6$ make the model vary too much from one step to another, and the improvement is also slower.

The next experiment motivates the choice to limit the number of epochs R to 10, since the model does not improve much with additional passes through the training data. In Figure 3.6 the learning curve is shown for up to 20 epochs. In the first 10 epochs there is an increase of 17 BLEU points, from 64 to 81, while in the next 5 epochs the increase is of only 3 BLEU points and even less, 2 BLEU points, for the last 5 epochs. This shows that the learning curve flattens after 10 epochs and the relatively small gain in BLEU score does not justify doubling the training time to carry out the extra epochs.

Another decision that had to be made was which inference algorithm to use. The initial hypothesis was that the Memetic algorithm should be faster on large problems since it is an approximate algorithm, while the Branch and Bound algorithm would not make search errors but would be slower. It turns out that for sentences of up to 25 words, the

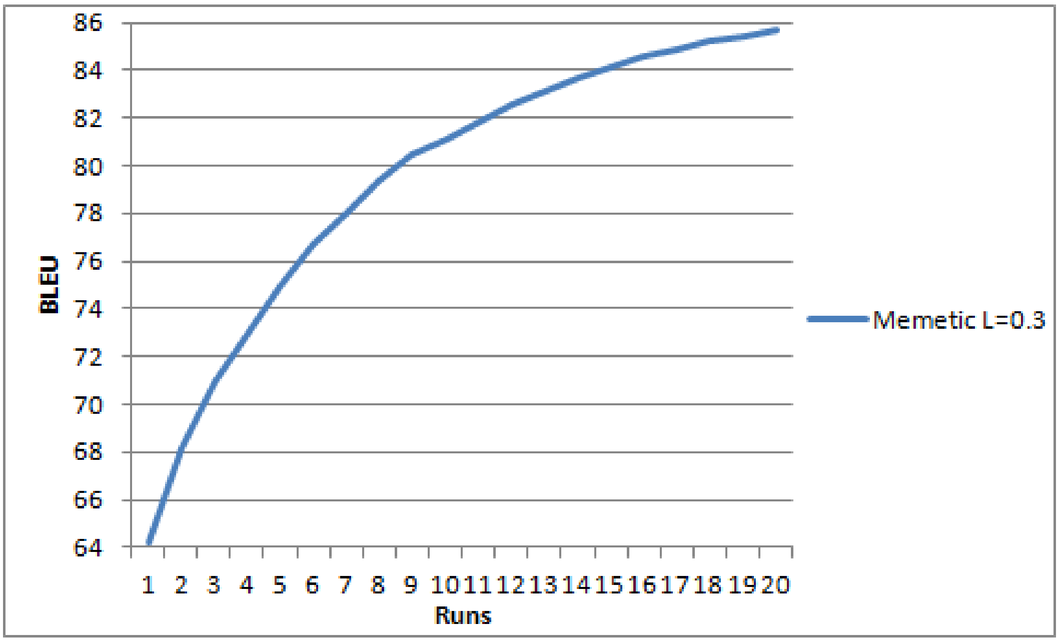


Figure 3.6: Convergence for L=0.3. The search algorithm used is the Memetic algorithm and the training set has 2000 sentences.

solutions found by the Memetic algorithm are as good as those found by the exact algorithm. This behavior can be seen in Figure 3.7 where the BLEU score after each run is almost identical. This result shows that the search space of the Memetic algorithm contains good solutions and even the global optimal solutions for this particular problem. This is especially interesting because the Memetic algorithm is affected by the choice of the local search procedure, and it appears that the *interchange operator* finds improving solutions rather quickly in the case of word reordering. It is also an indicator that the *oracle reordering* that the algorithm is trying to learn, is in fact just a few moves away from the original order. This result also strengthens the belief that for longer sentences the Memeric algorithm will not only be more efficient but also very accurate.

Since the accuracy given by the two inference algorithms is almost the same, the only other criteria for selecting one or the other was robustness. The Branch and Bound algorithm proved to be more robust on this training sentences since it is an exact algorithm and the problems were relatively small. The Memetic algorithm keeps searching for improving solutions until the initial population has reached a fixed number of individuals. Because the local search it performs is based on some random choices, the Memetic algorithm is unable to detect when there are no more improving solutions and the search should stop. The Branch-and-Bound algorithm on the other hand, will have more information from the upper bounds of the objective function to determine if a better solution can be found. Although the size of the initial population of the Memetic algorithm was decreased for shorter sentences, which have fewer improving solutions,

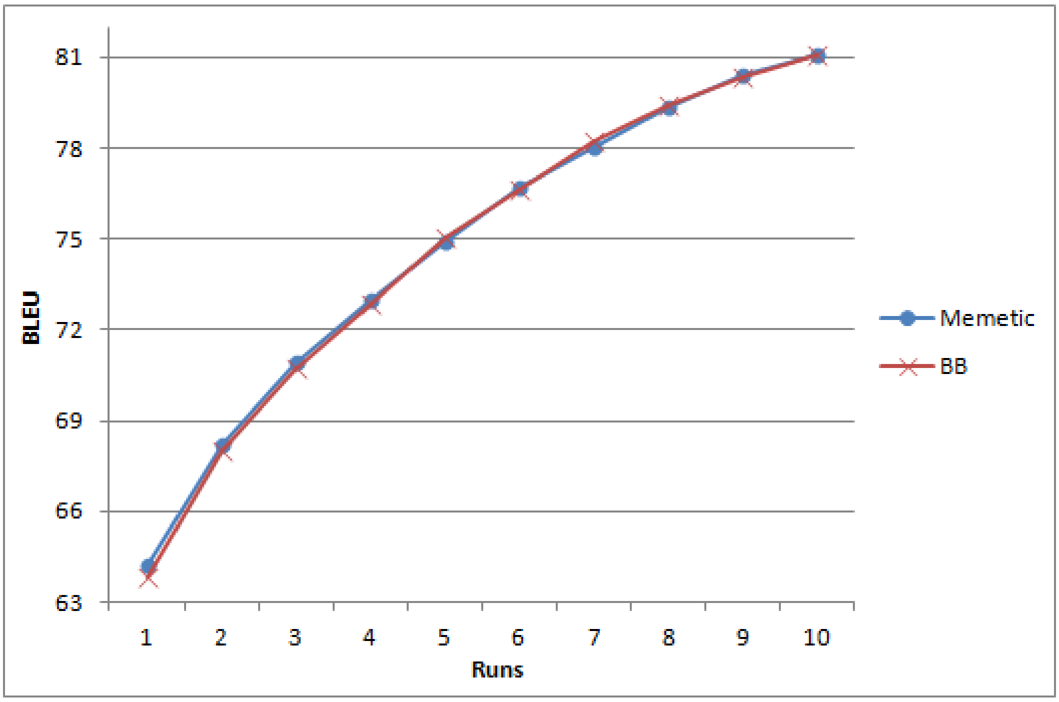


Figure 3.7: Comparison between learning curve when using Memetic algorithm and Branch and Bound algorithm as search algorithms, a learning rate of 0.3 and training set of 2000 sentences.

and time limits were set, without further tuning the algorithm was a bit slower than the Branch and Bound algorithm. For this reasons the more robust Branch and Bound algorithm was used in further experiments. Figure. 3.8 shows the learning curve of the pre-reordering model trained on 50000 sentences using the Branch-and-Bound algorithm. The BLEU scores and the improvement are of course reduced, since the training data is larger, but the asymptotic behavior is similar to that in previous experiments.

The following monolingual evaluations were conducted, after choosing all the parameters of the learning algorithm, in order to determine how well the pre-reording model is able to predict an optimal word reordering for unseen sentences. The first evaluation focused on how much training data is needed to improve the predicted reordering. Ideally, this evaluation is conducted by gradually adding more training data, retraining the model and applying it to a test set. Because retraining the models on more than 50000 sentences would be too expensive in terms of time, the true learning curve was approximated in the following way. To estimate the improvement due to training on 100000 sentences, two models trained on subsets of 50000 sentences were mixed and applied on the test set. To estimate the improvement due to training on 150000 sentences, three such models were mixed and so on. The final model used to approximate the improvement due to training on the entire data set was obtained by mixing

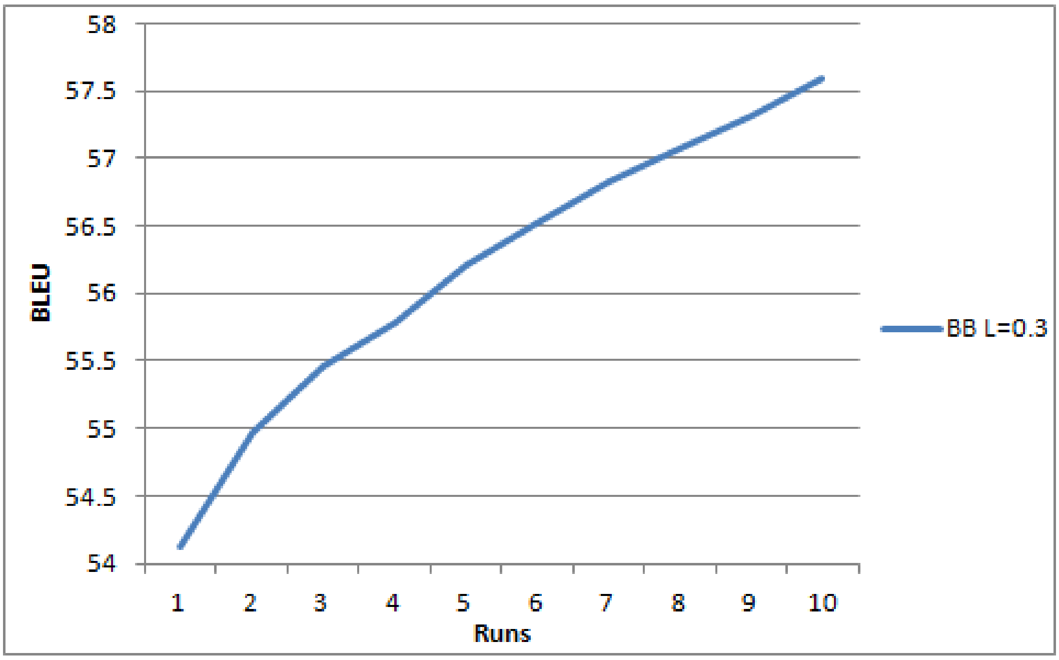


Figure 3.8: Learning curve when using Branch and Bound algorithm with learning rate 0.3 and a training set of 50000 sentences.

all models were trained on one of the 20 subsets of the training data. This approximated learning curve is shown in Figure 3.9. The learning curve starts to flatten around 500000 sentences and has a small drop at the end that is attributed to noise. Since the entire training data has to be reordered in order to be used for machine translation, it still makes sense to profit of a small improvement and train the model on the entire training set.

The final model was then applied to reorder the source side of the test set, and the resulting reordered sentences were evaluated in terms of BLEU score. This evaluation measures how close the predictions made by the pre-reordering model are to the desired Oracle reordering. Table 3.4 shows the comparison between the original order, the oracle reordering and the predicted reordering. The results indicate the original order has more n-grams in common with the oracle reordering than the predicted order has. This does not mean, though, that the original order is better. The BLEU score can only indicate the number of different n-grams but it is not informative about the type of differences from one order to another. As the machine translation experiments will show, exactly those small differences that were covered by the predicted reordering will improve the translation output. This monolingual evaluation also indicates that the oracle reordering is quite close to the original order and that the predicted order has not moved far from the original order, which could be improved upon.

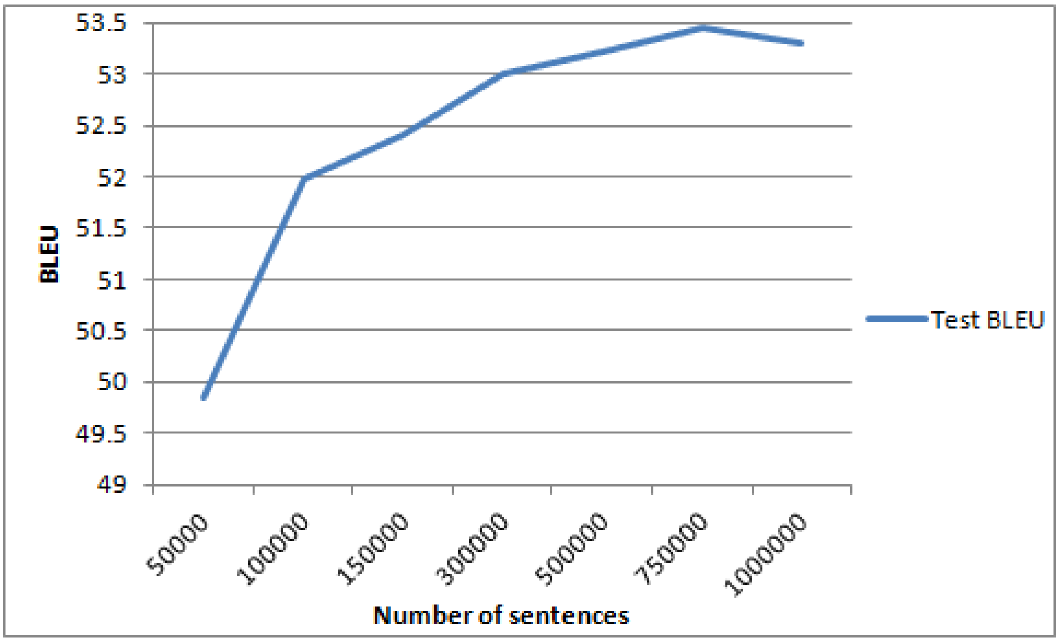


Figure 3.9: Evaluation of the pre-reordering model on the 2000 sentences test set used for translation experiments when more training sentences are gradually added. The learning curve is approximated by mixing an increasing number of models trained on subsets of 50000 sentences each.

Evaluated order	Reference order	2-grams	3-grams	4-grams	BLEU
Predict	Oracle	68.74	52.97	42.68	62.78
Original	Oracle	69.77	54.00	43.53	63.64
Predict	Original	87.72	81.25	76.08	85.81

Table 3.4: Monolingual evaluation of the reordered source sentences of the test set using BLEU scores. Each line compares two different word orders, that are either the orginal order, the oracle reordering or the reordering predicted by the final pre-reordering model.

Model	1-gram	2-gram	3-gram	4-gram	BLEU
Baseline	57.58	30.75	19.96	13.72	26.20
Oracle	60.35	38.44	27.47	20.22	33.69
BB[POS]	57.35	30.81	**20.22**	**14.05**	**26.62**
BB[POS+DEP]	56.96	30.34	19.64	13.54	26.04

Table 3.5: Evaluation in terms of BLEU scores of the machine translation output for the test set.

3.5.3 Machine Translation Evaluation

In section 3.5.2 a trade-off was made between efficiency and more accurate models resulting in the following values for the learning parameters: the learning rate was fixed to 0.3, the learning algorithm was stopped after 10 epochs, and the Branch and Bound algorithm was used to perform inference. The pre-reordering model, trained as described in sections 3.4 and 3.5.2, was applied to reorder the source side of the training, tuning and test sets used in machine translation experiments. The reordered source side of the training corpus was the result of the last epoch of training the pre-reordering model. The source side of the tuning and test sets were reordered using the final pre-reordering model. Table 3.5 shows the evaluation in terms of BLEU scores of the translation output given different word orders for the source sentences of the test set. The baseline was obtained by training, tuning and testing the statistical machine translation system using the original word order of the source sentences. An upper bound on how much can be gained by pre-reordering the source sentences was obtained by using the Oracle reordering of the source sentences of the training, tuning and test sets and running again the entire machine translation pipeline. To evaluate the contribution of the pre-reordering model the same was repeated using the reordering of the source sentences predicted by this model. Two pre-reordering models were considered, one trained using only part-of-speech features and the second trained using both part-of-speech and dependency features.

The results obtained when using the Oracle reordering show an improvement of more than 7 BLEU points, proving there is a lot to gain by pre-reordering the source sentence such that its word order resembles the word order in the target language. The pre-reordering model trained using part-of-speech features improved the baseline by 0.4 BLEU point, which leaves space for improving the model. The score break-down shows that the improvement comes from higher 3-gram and 4-gram scores, which indicates more fluent translations and therefore better word order. The pre-reordering

Model	BLEU
Baseline	25.51
Oracle	33.37
BB[POS]	25.65
BB[POS+DEP]	**25.93**

Table 3.6: Evaluation in terms of BLEU scores of the machine translation output for the tuning set.

model trained using both part of speech and dependency features didn't bring the expected results, the BLEU score being lower than that of the baseline. This result was disappointing since the improvement on the tuning set shown in Table 3.6 was encouraging. Further investigations have to be made about the reasons why the second model behaved worse on the testing set than the first model. A possible reason might be that the dependency parser was trained on data from a different domain than that of Europarl, resulting in less accurate dependency information. Another reason could be that only 11% of the head-dependent pairs in the entire training data swap their order according to the Oracle reordering. If the test set had few such pairs that should swap position, then the dependency features might have preferred the original word order. If the tuning set, on the other hand, had more pairs that should swap position according to the dependency features, that may have resulted in a better translation. Section 3.5.4 presents an error analysis of the machine translation output and tries to explain in more depth how pre-reordering improved the translation.

3.5.4 Error Analysis

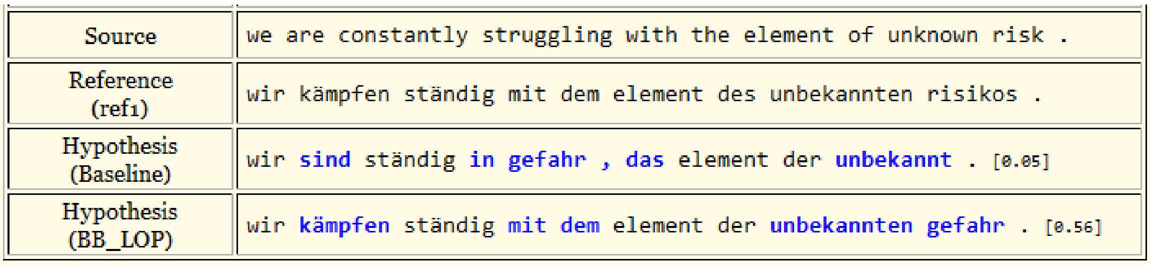
The error analysis takes a closer look at what improved in the machine translation output when the source sentence was pre-reordered with the model trained using part of speech features. Special attention was given to verbs, that were accounted for in separate categories than other words. One of the hypothesis made about how pre-reordering can improve translations was that it will reduce the number of missing verbs. As can be seen from the error analysis in Table 3.7 both the number of main verbs and auxiliary verbs were reduced when pre-reordering the source sentence. This contribution is indeed important since the verbs carry the most information in a sentence. Another hypothesis was that if the source and target sentences have similar word order, than the translation table will have better estimates for adequate translation pairs. Again we see a confirmation in the error analysis since the number of incorrect words

Error type	Error Sub-type	Baseline	BB[POS]
Missing Words		**22**	**6**
	content words - main verb	7	1
	content words - auxiliary verb	3	1
	content words - other	5	0
	filler words	7	4
Word order		**10**	**7**
	word based - verbs	3	2
	word based - other	2	2
	phrase based	5	3
Incorrect words		**41**	**22**
Wrong sense	wrong lexical choice	18	5
	incorrect disambiguation	7	6
	mixed agent/patient	3	0
Wrong form	verb tense	2	2
	concordance	13	10
Punctuation		**9**	**5**

Table 3.7: Error analysis.

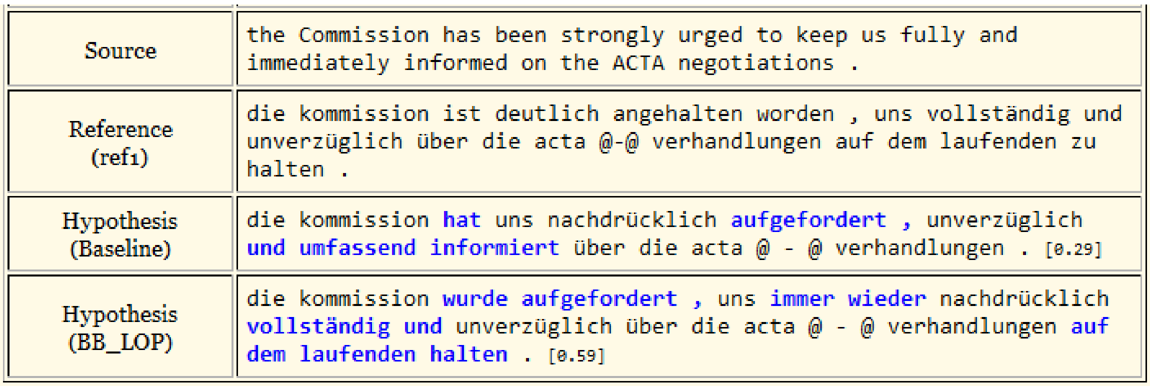
was reduced when pre-reordering the source sentence. An interesting change that was pointed out by the evaluator was that in some cases the verb is disambiguated such that it correctly selects its agent and patient. The error analysis also shows that word order was improved with this approach, but since some of the words that were missing in the baseline, appeared in the other translation but were misplaced, the numbers are only slightly smaller. Concordance remains an issue when translating into German due to the rich morphology of the language.

Finally we give some examples in order to compare the baseline translation with the translation resulted from the pre-reordering approach. In the example given in Figure 3.10 the pre-reordering approach translates a verb that was missing in the baseline translation and also correctly disambiguates another one. In the example given in Figure3.11 the pre-reordering approach generates a translation with better verb order and disambiguates the verb *wurde* such that it correctly selects its agent and patient.



Source	we are constantly struggling with the element of unknown risk .
Reference (ref1)	wir kämpfen ständig mit dem element des unbekannten risikos .
Hypothesis (Baseline)	wir **sind** ständig **in gefahr , das** element der **unbekannt** . [0.05]
Hypothesis (BB_LOP)	wir **kämpfen** ständig **mit dem** element der **unbekannten gefahr** . [0.56]

Figure 3.10: Example of how the pre-reordering approach gives a better translation by correctly translating a main verb that is missing in the baseline and correctly disambiguating another one.



Source	the Commission has been strongly urged to keep us fully and immediately informed on the ACTA negotiations .
Reference (ref1)	die kommission ist deutlich angehalten worden , uns vollständig und unverzüglich über die acta @-@ verhandlungen auf dem laufenden zu halten .
Hypothesis (Baseline)	die kommission **hat** uns nachdrücklich **aufgefordert ,** unverzüglich **und umfassend informiert** über die acta @ - @ verhandlungen . [0.29]
Hypothesis (BB_LOP)	die kommission **wurde aufgefordert ,** uns **immer wieder** nachdrücklich **vollständig und** unverzüglich über die acta @ - @ verhandlungen **auf dem laufenden halten** . [0.59]

Figure 3.11: Example of how the pre-reordering approach gives a better translation by generating a correct verb order and by correctly disambiguating the verb such that it properly selects its agent and patient.

3.6 Discussion

This chapter proposed searching for an optimal word ordering using two state-of-the-art algorithms that solve the Linear Ordering Problem. One was the Branch-and-Bound algorithm which finds a global optimal solution to the problem and the other was the Memetic algorithm which finds local optimal solutions. Because it was more robust, the Branch-and-Bound algorithm was used to learn the final pre-reordering model that was applied as a pre-processing step in the machine translation pipeline. The automatic evaluation and error analysis of the machine translation output showed that pre-reordering the source sentences with the proposed model reduces the number of missing verbs and incorrect words, and improves word order. The monolingual evaluation showed that the Memetic algorithm is able to predict solutions that are very close to the global optimal solutions. Moreover, this was an indication that the optimal word order can be reached rather quickly by applying *interchange* moves to the original order. Since the Memetic algorithm is more efficient for larger problems it could be applied successfully to learn a pre-reordering model for longer sentences. The current reordering model could be used to initialize the parameters of the new model. Pre-reordering the entire Europarl corpus should improve translation output even more, since there is more word reordering required for longer sentences. The pre-reordering model could also be improved by splitting the training corpus in less parts in order to learn better feature weights. This could also result in more reliable dependency features which would contribute to syntactic word reordering.

CHAPTER 4

Across Sentence Boundary Language Model

The most widely used language models for machine translation are *n*-gram models. When a large training corpus is available, like the Google n-gram Corpus with more than 1 billion 5-grams, then *n* can take values of up to five or six. But more frequently there is either too little domain specific data and then data sparsity becomes a problem, or using a huge language model is not computationally or space efficient. Therefore the order of the language models is usually limited to three or four, making it harder to model long-distance dependencies between words. Another limitation for *n*-gram models is that the word contexts are taken only within sentence boundaries. The model proposed by Momtazi et al. (2010) tries to capture long range dependencies that cross sentence boundaries, by using words in previous sentences to estimate the probability of words in the current sentence. The model is presented in sections 4.1 and 4.2 and the results of integrating this language model with a phrase-based machine translation system are presented in sectionExpASB.

4.1 Across sentence boundary unigram model

The key component of the *Across Sentence Boundary Language Model (ASB)* proposed by Momtazi et al. (2010) is a unigram trigger model. The motivation behind the trigger model is that words in adjacent sentences are related, and therefore words in the current sentence are triggered by words in previous sentences. Arguments for this relation can be found in linguistic theories about discourse and text coherence, but modeling this relation statistically without any linguistic input is challenging. The trigger model estimates the unigram probabilities in the following way:

$$P(w|S_{-1}, S_{-2}, ...) = \sum_{u_i \in S_{-i}} P_{SentSent}(w|u_1, u_2...) \cdot f_{S_{-1}}(u_1) \cdot f_{S_{-2}}(u_2) \cdot ... \tag{4.1.1}$$

where S_{-i} are predecessor sentences, $f_{S_{-i}}(u_i)$ are the relative frequencies of the words in the predecessor sentences and $P_{SentSent}(w|u_1, u_2...)$ models the co-occurence of words in adjacent sentences. The improved unigram model will reduce perplexities of higher-order models, making an impact on the quality of the language model.

Although several predecessor sentences can be considered, for efficiency reasons only the immediately preceding sentence is used by the trigger model, making $P_{SentSent}$ a bigram model. The trigger model can also be trained on the words in the current sentence or on the current word itself. By combining trigger models trained in different ways, the unigram model is improved, which leads to better higher-order models.

4.2 Adapted Models

The trigger model described above is used as the unigram probability for an adapted word model. The adapted word model estimation is a result of fast marginal adaptation (Klakow 1998), a special case of log-linear interpolation:

$$P_{AdaptedWord}(w|h, S_{i-1}) = \frac{1}{Z_\lambda(h)} \left(\frac{P(w|S_{i-1})}{P(w)} \right)^{\lambda_u} P(w|h) \tag{4.2.1}$$

where h is the history, $P(w)$ is the usual unigram probability, $P(w|S_{i-1})$ it the across boundary emission probability and λ_u the interpolation weight.

An adapted class model is also defined by using the trigger model as the emission probability:

$$P_{AdaptedClass}(w|h, S_{i-1}) = P(w|c(w), S_{i-1})P(c(w)|c(h)) \tag{4.2.2}$$

where $P(w|c(w), S_{i-1})$ is the emission probability of the current words given its class and the words in the previous sentence.

The adapted word model and adapted class model are then combined using linear interpolation:

$$P_{AdaptedInterpolation}(w|h, S_{i-1}) = \alpha P_{AdaptedWord}(w|h, S_{i-1}) + (1 - \alpha) P_{AdaptedClass}(w|h, S_{i-1}) \tag{4.2.3}$$

corpus	sentences	words	vocabulary size
europarl v6	2032006	59411569	96801
newstest2009	2525	71489	9216
newstest2010	2489	68024	9148

Table 4.1: Statistics for the training, development and test sets used with the Across Sentence Boundary language model.

4.3 Experiments

4.3.1 Experimental Setup

The language model was trained on the English side of Europarl v6. The text used for tuning the parameters of the language model and for tuning the parameters of Moses was the 2009 news test set and the text used for evaluating the translation was the 2010 news test set, both provided in the Machine Translation Workshop. For the adapted class model 500 word classes were trained with the SRILM-toolkit using the entire corpus. In Table 4.1 some statistics are presented about the size of the corpus and the files used as development and test sets. The files were pre-processed by adding start-of-sentence and end-of-sentence symbols and by tokenizing and true-casing them. For the corpus used to train the trigger models, the end-of-sentence symbol was added to delimit the sentences used for update, but the start-of-sentence symbol was omitted since it would be uninformative for the trigger models.

4.3.2 Perplexity evaluation

As mentioned before, $perplexity$ is a measure of how well a language model recognizes some text. The parameters of the language models were tuned such that the perplexitiy will be minimized on the tuning set. The perplexity was again used to compare different language models on both the tuning and test sets.

The first hypothesis made about improving a language model was that improving the unigram model will lead to improved higher order models. Table 4.2 shows the perplexity of unigram models on the tuning and test sets. The baseline perplexity was computed for a normal unigram model. The second row shows the perplexity of a trigger model trained on the words in the current sentence and the third row the perplexity of a trigger model trained on the words in the previous sentence. The forth row shows the perplexity of a model that combines the previous two trigger models and a trigger model trained on the current word itself. Although the individual trigger mod-

Model	Tuning set	Test set
Baseline Unigram	2035.77	2279.35
ASB trigger trained on same sentence	1845.73	2087.7
ASB trigger trained on previous sentence	1928.01	2199.08
ASB trigger with combined training	1475.52	1634.13
Reduction	27.52%	28.31%

Table 4.2: Perplexities of the unigram models on the tuning and test set.

els reduce the perplexity only by a small percentage, the combination of these models yields a significant reduction of 27.52% on the tuning set and of 28.31% on the test set. Another aspect that needs to be considered is that the corpus used for training the language model is from a different domain than the tuning and test sets. For this reason the *out of vocabulary (OOV)* rate is of 2.0% for the tuning set and 2.3% for the test set, leading to a high value of the perplexity even when the combined model is used.

The combined across sentence boundary trigger model was then used as a unigram model in the adapted word model and as the emission probability in the class model. The perplexities of the adapted models on the tuning and test sets are presented in Table 4.3 and Table 4.4, respectively. The perplexities are shown for the tuning set since it was used for tuning the parameters of the language model, and in machine translation experiments for minimum error rate training. The results for the test set are shown in order to draw a conclusion later on about how changes in perplexity affects machine translation output. The first column shows the perplexities of the baseline models: a normal 4-gram model with Knesser-Ney smoothing, a class based 4-gram model and an interpolated model of the previous two. The perplexities of the across sentence boundary adapted models are shown in the second column. The adapted word model considered in these experiments uses a linear interpolation between the adapted unigram model and the normal unigram model instead of using a log-linear interpolation resulting from fast marginal adaptation. The reason for not evaluating the perplexity of the later model is that normalization is too expensive to compute for this model and without normalization the perplexity would be meaningless. The normalized model would be too slow to be used in machine translation experiments, therefore an unnormalized version was used and the result will be presented section 4.3.3. Another thing to notice is that the class-based models have quite high perplexities which leads to a smaller improvement of the interpolated models. This can be due to the fact that the data used to train the classes and the language model is from a different domain than the test set. Finally the last column shows the reduction in perplexity of

Model	Baseline	Across Sentence Boundary	reduction%
Word	463.884	332.835	28.25
Class	869.781	680.262	21.79
Interpolated	424.04	321.892	24.09

Table 4.3: Perplexities of the language models on the tuning set.

Model	Baseline	Across Sentence Boundary	reduction%
Word	496.454	355.47	28.40
Class	977.621	765.925	21.65
Interpolated	454.618	343.632	24.41

Table 4.4: Perplexities of the language models on the test set.

the adapted models with respect to the baseline models. The experiments resulted in a significant reduction in perplexity of the final interpolated across sentence boundary model compared to the baseline word model: 30.61% for the tuning set and 30.78% for the testing set. This result confirms the hypothesis that an improved unigram model reduces the perplexities of the higher-order models.

4.3.3 Autmatic Machine Translation Evaluation

The significant decrease in perplexity obtained by the adapted model means the language model is able to assign higher probabilities to valid text in the considered language. This result, however encouraging, does not guarantee a better performance of the machine translation system which incorporates this language model, since there is no known correlation between perplexity and an automatic metric for machine translation. The next experiments will therefore try to asses if the decrease in perplexity will lead to an increase in BLEU score.

The machine translation system was trained using several language models that were previously evaluated in terms of perplexity. The baseline system configurations used the baseline language models that have been already introduced. The other system configurations used adapted language models with the combined trigger unigram model obtained by log-linear interpolation without normalization (ASB) and with the combined unigram trigger model obtained by linear interpolation (ASB Interpolated). The results of the automatic evaluation using BLEU scores is shown in Table 4.5. The improvement between the results of the system using the normal baseline language model and the system using the interpolated adapted model, is not as high as expected con-

Model	Baseline	ASB	ASB Interpolated
Word	19.99	20.12	20.03
Class	18.59	18.55	18.56
Interpolated	20.03	**20.19**	20.11

Table 4.5: BLEU scores when translating with baseline word language model and with Across Sentence Boundary (ASB) models.

sidering the significant reduction in perplexity of the language models. An explanation for this small difference in BLEU score is that the language model is not used at its full potential. One aspect to be considered is that the sentences sent for update to the trigger model are inadequate sentences in the target language, making the trigger model inaccurate and therefore diminishes its contribution. The inadequacy of translated sentences for the German-English language pair is partially caused by the rich morphology of the source language, which causes data sparsity problems for the translation model. Another cause for inadequate translations is the different word orders of the two languages, that accounts for many missing verbs and other content words, as was explained in the chapter 3. The next experiments investigate possible problems with using the adapted language model for machine translation.

The first investigation is focused on how the quality of the sentences sent for update to trigger model affects the machine translation output. To be able to understand the value of valid text that is being sent for update to the language model, the reference translation sentences were given as update instead of the sentences that were being translated by the system. This leads to a 0.4 BLEU point improvement when using the adapted word language model, as can be seen in Table 4.6. As expected, better trigger words increase the contribution of the trigger unigram model, which increases the contribution of the entire language model and leads to more significant improvement of the machine translation output. The break-down of the score by different n-gram sizes shows that there is a larger contribution coming from the increased 1-gram and 2-gram counts. This would indicate that the larger context considered by the across sentence boundary model reduces the number of missing words and contributes to better lexical choice and word disambiguation for machine translation.

To further assess what would be an upper bound for the contribution of the adapted model if the training would provide the relevant statistics, the models were trained on the reference translations. The results in Table 4.7 show that the baseline language model could dramatically improve the machine translation output if provided with the proper statistics. Moreover the adapted word language model could improve this re-

Model	1-gram	2-gram	3-gram	4-gram	BLEU
Word Baseline	56.72	26.58	13.92	7.61	19.99
Word ASB	57.32	27.09	14.27	7.86	**20.42**

Table 4.6: BLEU scores when translating with word adapted language model with Oracle update.

Model	1-gram	2-gram	3-gram	4-gram	BLEU
Word Baseline	56.72	26.58	13.92	7.61	19.99
Word Baseline [REF]	75.33	58.17	49.54	42.03	45.44
Word ASB [REF]	76.05	59.02	50.35	42.7	46.16
Word ASB [REF] + ORACLE	**76.34**	**59.53**	**50.8**	**43.09**	**46.65**

Table 4.7: BLEU scores when translating with the following models trained on the reference set: baseline, adapted word and adapted word with Oracle update.

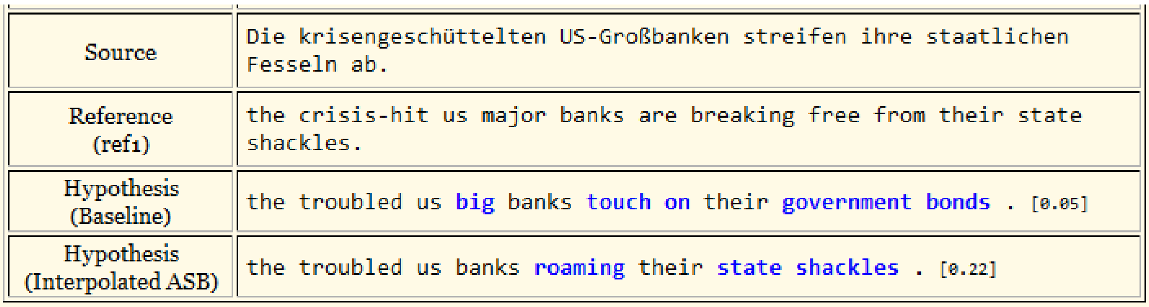
sult with an additional 0.7 BLEU point, and by using an adequate update for the trigger model, the increase would be of 1 BLEU point. The break down of the scores by n-gram sizes shows and improvement of the 1-gram score for the adapted model which can indicate better lexical choices. An even bigger improvement can be seen for the 2-gram and 3-gram scores, which suggests the translation has better word order. It is worth pointing out that the test set and the data used to train the language model are still from a different domain than the data used to train the translation model. Matching the two domains might further increase the contribution of the across sentence boundary model.

4.3.4 Error Analysis

The issue of how to automatically evaluate machine translation output has been widely discussed, and several arguments were raised against how well does the BLEU metric reflect the quality of the translation. Having only one reference translation is also a problem when using BLEU metric, since in natural language different surface realizations can bring across the same meaning. Therefore an obvious choice is to turn to a manual error analysis in order to determine how the adapted models affect the machine translation output. Since manual evaluation is a tedious job, the error analysis was conducted for 20 sentences with the hope that the general trend is captured in only these few sentences. The results of the error analysis are shown in Table 4.8. The error analysis indicates that the adapted models, especially the linear interpolation of the

Error type	Error Sub-type	Baseline	Word ASB	Interpolated ASB
Missing Words		**24**	**22**	**14**
	content words	18	16	6
	filler words	6	6	2
Word order		**18**	**18**	**25**
	word based	12	11	15
	phrase based	6	7	10
Incorrect words		**84**	**79**	**71**
Wrong sense	wrong lexical choice	49	47	42
	incorrect disambiguation	27	24	18
Wrong form	verb tense	5	5	7
	concordance	3	3	4
Punctuation		**12**	**7**	**6**

Table 4.8: Error analysis.

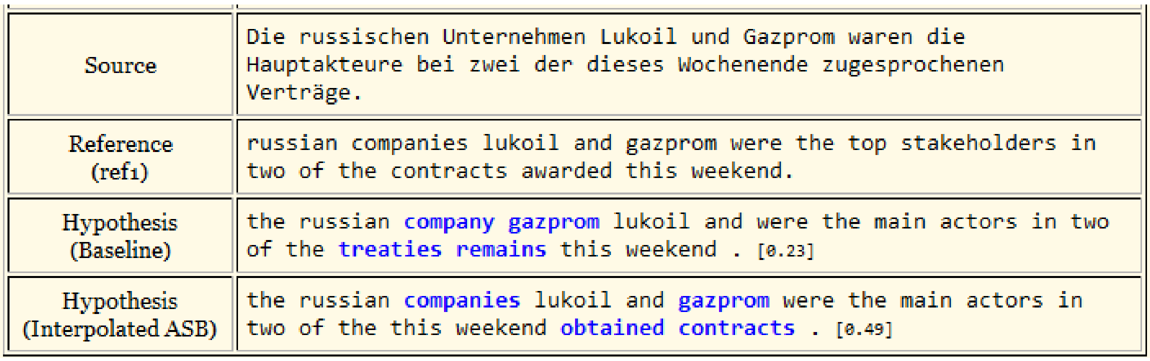


Source	Die krisengeschüttelten US-Großbanken streifen ihre staatlichen Fesseln ab.
Reference (ref1)	the crisis-hit us major banks are breaking free from their state shackles.
Hypothesis (Baseline)	the troubled us big banks touch on their government bonds . [0.05]
Hypothesis (Interpolated ASB)	the troubled us banks roaming their state shackles . [0.22]

Figure 4.1: Comparison between machine translation with baseline language model and with interpolated adapted language model.

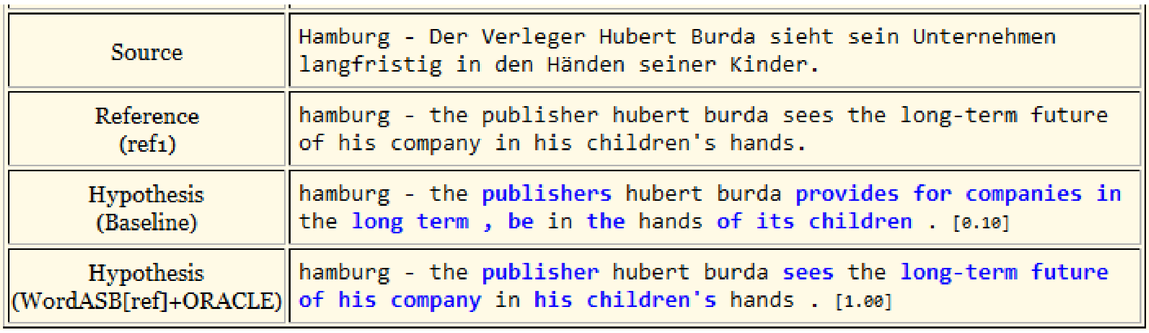
word and class-based adapted models, help reduce the number of missing words, incorrect lexical choices and incorrectly disambiguated words. The number of mistakes attributed to wrong word order is slightly increased because words that did not appear in the baseline translation, but were present in the other two translations, were sometimes misplaced.

Finally some examples of translations are given to compare the output of the machine translation system when using the baseline language model and when using the adapted models. The iBLEU tool was provided to the evaluator to help visualize the translations (Madnani 2011).



Source	Die russischen Unternehmen Lukoil und Gazprom waren die Hauptakteure bei zwei der dieses Wochenende zugesprochenen Verträge.
Reference (ref1)	russian companies lukoil and gazprom were the top stakeholders in two of the contracts awarded this weekend.
Hypothesis (Baseline)	the russian company gazprom lukoil and were the main actors in two of the treaties remains this weekend . [0.23]
Hypothesis (Interpolated ASB)	the russian companies lukoil and gazprom were the main actors in two of the this weekend obtained contracts . [0.49]

Figure 4.2: Comparison between machine translation with baseline language model and with interpolated adapted language model.



Source	Hamburg - Der Verleger Hubert Burda sieht sein Unternehmen langfristig in den Händen seiner Kinder.
Reference (ref1)	hamburg - the publisher hubert burda sees the long-term future of his company in his children's hands.
Hypothesis (Baseline)	hamburg - the publishers hubert burda provides for companies in the long term , be in the hands of its children . [0.10]
Hypothesis (WordASB[ref]+ORACLE)	hamburg - the publisher hubert burda sees the long-term future of his company in his children's hands . [1.00]

Figure 4.3: Comparison between machine translation with baseline language model and with word adapted language model trained on reference set with Oracle update.

4.4 Discussion

This chapter proposed using a novel language model for statistical machine translation, that captures the relation between words in adjacent sentences. The error analysis showed that the adapted Across Sentence Boundary language model improves word disambiguation, lexical choices and reduces the number of missing words in the machine translation output. The automatic evaluation in terms of BLEU score showed only a small improvement, but further investigations indicated that using more adequate trigger words for the adapted model can improve the BLEU score significantly. One solution to avoid using inadequate trigger words that are translated by the machine translation system would be to use trigger words from the source sentences. The adapted model would therefore be more robust if it uses a combination of trigger unigram models trained on both the source and target sentences. Moreover, using only trigger words from the source sentences would allow the machine translation system to process sentences in parallel. This would be possible since the source sentences are known in advance and the system wouldn't have to first translate the previous two sentences in order to send the trigger words to the language model. In addition it would be possible to use trigger words from sentences that come after the sentence that is being processed. The trigger words are also affected by the different domains of the training and test data. The adapted model could have a greater impact if the two domains were matched.

CHAPTER 5

Conclusions

The focus of this thesis was dealing with frequent translation errors for the English-German language pair: incorrect word order, missing words, especially verbs, and wrong lexical choices. The present work approached these problems through two components of a statistical machine translation system: a pre-reordering component for English into German translation and the language model for German into English translation.

The first approach was to pre-reorder the source sentences to bring them to a form that has a word order similar to that of the target language. Word reordering was modeled as the Linear Ordering Problem and two algorithms were considered for solving this problem. One contribution of the present work was to successfully apply a Branch-and-Bound algorithm to learn a pre-reordering model. This algorithm was robust and efficient when applied to word reordering of sentences that had up to 25 words. Since sentences used in machine translation experiments can have up to 80 words, this algorithm might be too slow when reordering longer sentences because it finds a global optimal solution. Automatic evaluation indicated that using the pre-reordering model to pre-process the source sentences improved the machine translation output. This was further confirmed by an error analysis which showed that this approach reduced the number of missing verbs, incorrect words and improved word order. The features that were considered by the pre-reordering model were unlexicalized features based on part-of-speech information. This allows the model to generalize well to unseen data and reduces the problems caused by mismatching training and test domains. New dependency features were also proposed for learning a model that captures syntactic word reordering. Although these features didn't seem to improve word reordering for shorter sentences they might have a greater impact for longer sentences. This can be true because for shorter sentences only 11% of the head-dependent word pairs swap their order according to the oracle reordering, while for longer sentences more than

20% of these pairs swap oder. The present work also showed that the Memetic algorithm, which finds local optimal solutions to the Linear Ordering Problem, is suitable for the task of word reordering and could pre-reorder longer sentences more efficiently. This algorithm is able to find good local optimal solutions and even global optimal solutions because it seems that the optimal word order can be reached rather quickly from the original order by applying *interchange* moves. Therefore a direction for future work would be learning a pre-reordering model for longer sentences using the Memetic algorithm for inference and improved dependency features.

The second approach was to integrate a novel Within-and-Across Sentence Boundary language model in a phrase-based statistical machine translation system. This language model captures the dependencies between words in adjacent sentences by modeling how words in previous sentences trigger words in the current sentence. Another contribution of this thesis was to show how the ASB language model can help machine translation and what some of the hindrances of applying it to this task are. Error analysis indicated that this approach can improve word disambiguation, lexical choices and reduce the number of missing words in the machine translation output. A problem of using the ASB model for machine translation is that inadequate trigger words diminish the contribution of the model. Since the trigger words were taken from the previous two sentences translated by the system and the translations for the German-English language pair have many missing words and wrong lexical choices, the ASB model was less accurate. One proposal for making the language model more robust as part of a machine translation system was to train it with trigger words from both the source and target sentences. If the trigger words were taken only from the source sentence this would also allow the translation system to process words in parallel. Moreover, this would be interesting because it would allow modeling the dependency between words in the current sentence and the following sentences. Another proposal for enhancing the contribution of the trigger words was to match the domains of the training and testing data.

The pre-reordering model was applied to bring the source sentences to the more flexible word ordering of German. The translation output might be improved more when translating from German into English since the pre-reordering model would be learning a fixed word order. The Across Sentence Boundary language model could also help when translating into German, since concordance is a major issue which can be improved by modeling how words trigger each other. In order to draw on the benefits of each approach the next step would be applying them together to machine translation for both translation directions.

References

Roy Tromble and Jason Eisner. Learning linear ordering problems for better translation. 2009.

Saeedeh Momtazi, Friedrich Faubel, and Dietrich Klakow. Within and across sentence boundary language model. 2010.

Irène Charon and Olivier Hudry. A branch-and-bound algorithm to solve the linear ordering problem for weighted tournaments. *Discrete Appl. Math.*, 154, 2006.

Tommaso Schiavinotto and Thomas Stützle. The linear ordering problem: instances, search space analysis and algorithms. *Journal of Mathematical Modelling and Algorithms*, 2004.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Comput. Linguist.*, 1990.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 1993.

P. Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 2007.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, 2003.

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, 1996.

E Jelinek and R. L Mercer. Interpolated estimation of markov source parameters from sparse data. In *Proceedings, Workshop on Pattern Recognition in Practice*, 1980.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based n-gram models of natural language. *Comput. Linguist.*, 1992.

A. Stolcke. Srilm – an extensible language modeling toolkit. proc. intl. conf. on spoken language processing. 2002.

Dietrich Klakow. Log-linear interpolation of language models. In *ICSLP*, 1998.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. Error Analysis of Machine Translation Output. In *International Conference on Language Resources and Evaluation*, 2006.

Michael Collins, Philipp Koehn, and Ivona Kučerová. Clause restructuring for statistical machine translation. 2005.

Chris Dyer and Philip Resnik. Context-free reordering, finite-state translation. 2010.

R. Martí and G. Reinelt. *The Linear Ordering Problem: Exact and Heuristic Methods in Combinatorial Optimization*. Applied Mathematical Sciences. Springer, 2011.

Yoav Freund and Robert E. Schapire. Large margin classification using the perceptron algorithm. In *Machine Learning*, 1998.

Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *EMNLP*, 2002.

Ryan Mcdonald, Koby Crammer, and Fernando Pereira. Spanning tree methods for discriminative training of dependency parsers. 2005.

Roy Wesley Tromble. *Search and learning for the linear ordering problem with an application to machine translation*. PhD thesis, 2009.

Ryan McDonald, Keith Hall, and Gideon Mann. Distributed training strategies for the structured perceptron. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation, 2005.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees, 1994.

Mihai Surdeanu and Christopher D. Manning. Ensemble models for dependency parsing: Cheap and good? In *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference (NAACL-2010)*, Los Angeles, CA, June 2010.

Joakim Nivre and Jens Nilsson. Memory-based dependency parsing. 2004.

Nitin Madnani. ibleu: Interactively debugging & scoring statistical machine translation systems. In *Proceedings of the Fifth IEEE International Conference on Semantic Computing*, 2011.

CHAPTER 6

Appendicies

6.1 Error analysis for chapter 3

6.1.1 Evaluated source sentences

<seg id="1324"> since December , the Treaty of Lisbon commits the EU to join the Convention on Human Rights . </seg>
<seg id="1297"> the Commission has been strongly urged to keep us fully and immediately informed on the ACTA negotiations . </seg>
<seg id="990"> the proposed regulation would guarantee reliable , transparent and comparable results . </seg>
<seg id="161"> we are constantly struggling with the element of unknown risk . </seg>
<seg id="587"> nevertheless , our goal is in sight and we are moving slowly but surely towards it . </seg>
<seg id="683"> the European Council âĂŹ s decision seems to me to be a wise one . </seg>
<seg id="1132"> this dialogue is principally pursued through the heads of the EU missions in the countries concerned . </seg>
<seg id="1687"> the report calls on the budgetary authority to make a functioning budget available for direct Community measures for 1999 . </seg>
<seg id="1794"> on procedure , I would like to thank those colleagues in the Committee on Budget Control who have contributed towards this whole discharge process . </seg>
<seg id="1830"> I would like to highlight the existing cooperation between the energy and environmental sectors . </seg>
<seg id="1893"> representatives of the EU ′s member states will meet in Brussels in mid @-@ June to review a common policy towards Cuba . </seg>
<seg id="127"> finally , I would urge Parliament not to lose sight of the need for the development of an AIDS and malaria vaccine . </seg>
<seg id="221"> it is difficult to prophesy when the Commission will be able to table its proposals . </seg>
<seg id="462"> Turkey must interpret this challenge as a unique opportunity on its course to-

wards Europe . </seg>
<seg id="483"> the problems in the various less favoured zones can be solved through good levels of cooperation as part of the Union 's cohesion policy . </seg>
<seg id="578"> this will create synergies and will facilitate the sharing of data and working practices , in order to achieve a better understanding ... </seg>
<seg id="902"> fuel prices have hit the fisheries sector particularly badly . </seg>
<seg id="1329"> ethnic , religious , particularly Christian versus Muslim , tribal , cultural and economic tensions appear to be endemic in Nigeria . </seg>
<seg id="1286"> it is 10 o 'clock in the evening and we are staying for nothing if we do not receive a speaking slot . </seg>
<seg id="708"> what we are actually left with is a glass full of words and a hand full of nothing . </seg>

6.1.2 Evaluated reference sentences

<seg id="1324"> seit Dezember verpflichtet der Vertrag von Lissabon die EU , der Konvention zum Schutze der Menschenrechte und Grundfreiheiten beizutreten . </seg>
<seg id="1297"> die Kommission ist deutlich angehalten worden , uns vollständig und unverzüglich über die ACTA @-@ Verhandlungen auf dem Laufenden zu halten . </seg>
<seg id="990"> die vorgeschlagene Verordnung würde verlässliche , transparente und vergleichbare Ergebnisse garantieren . </seg>
<seg id="161"> wir kämpfen ständig mit dem Element des unbekannten Risikos . </seg>
<seg id="587"> dennoch ist unser Ziel in Sicht und wir bewegen uns langsam , aber sicher darauf zu . </seg>
<seg id="683"> die vom Europäischen Rat gefällte Entscheidung halte ich für sehr klug . </seg>
<seg id="1132"> dieser Dialog wird hauptsächlich durch die Leiter der EU @-@ Missionen in den betreffenden Ländern verfolgt . </seg>
<seg id="1687"> im Bericht wird die Haushaltsbehörde aufgefordert , für 1999 ein funktionierendes Budget für direkte GemeinschaftsmaÃ§nahmen bereitzustellen . </seg>
<seg id="1794"> bezüglich des Verfahrens möchte ich jenen Kollegen im AusschuÃ§ für Haushaltskontrolle danken , die zu diesem ganzen Entlastungsverfahren beigetragen haben . </seg>
<seg id="1830"> ich möchte die bestehende Zusammenarbeit zwischen dem Energie- und dem Umweltsektor hervorheben . </seg>
<seg id="1893"> Vertreter der EU @-@ Mitgliedsstaaten werden sich Mitte Juni in Brüssel treffen , um eine gemeinsame Politik gegenüber Kuba zu prüfen . </seg>
<seg id="127"> abschlieÃ§end möchte ich das Parlament dringend bitten , nicht den Blick für die Notwendigkeit der Entwicklung eines Impfstoffs gegen Aids und Malaria zu verlieren . </seg>
<seg id="221"> es ist schwer einzuschätzen , wann die Kommission ihre Vorschläge vorlegen kann . </seg>
<seg id="462"> die Türkei muss diese Herausforderung als eine einzigartige Gelegenheit für

ihren Kurs in Richtung Europa begreifen . </seg>
<seg id="483"> die Probleme der verschiedenen benachteiligten Regionen lassen sich durch eine gute Zusammenarbeit im Rahmen der Kohäsionspolitik der Union lösen . </seg>
<seg id="578"> dies wird zu Synergien führen und die gemeinsame Nutzung von Daten und Arbeitspraktiken erleichtern , um ein besseres Verständnis ... </seg>
<seg id="902"> die Kraftstoffpreise haben den Fischereisektor besonders schlimm getroffen . </seg>
<seg id="1329"> ethnische , religiöse , insbesondere Christen gegen Muslime , stammesbezogene , kulturelle und wirtschaftliche Spannungen scheinen in Nigeria endemischer Natur zu sein . </seg>
<seg id="1286"> es ist 22 Uhr und wir sitzen hier umsonst , wenn wir keine Redezeit bekommen . </seg>
<seg id="708"> was uns bleibt , ist ein Glas voller Worte und eine Handvoll Nichts . </seg>

6.1.3 Evaluated baseline sentences

<seg id="1324"> seit Dezember , der Vertrag von Lissabon verpflichtet sich die EU für den Beitritt zur Konvention zum Schutze der Menschenrechte und Grundfreiheiten . </seg>
<seg id="1297"> die Kommission hat uns nachdrücklich aufgefordert , unverzüglich und umfassend informiert über die ACTA @-@ Verhandlungen . </seg>
<seg id="990"> die vorgeschlagene Verordnung eine verlässliche und transparente und vergleichbare Ergebnisse . </seg>
<seg id="161"> wir sind ständig in Gefahr , das Element der unbekannt . </seg>
<seg id="587"> unser Ziel ist jedoch in Sicht , und wir werden langsam aber sicher zu bewegen . </seg>
<seg id="683"> der Europäische Rat die Entscheidung scheint mir vernünftig zu sein . </seg>
<seg id="1132"> dieser Dialog ist vor allem über die Köpfe der EU @-@ Missionen in den betreffenden Ländern . </seg>
<seg id="1687"> der Bericht fordert die Haushaltsbehörde eine Funktion für direkte MaÃ§nahmen der Gemeinschaft im Haushalt für das Jahr 1999 . </seg>
<seg id="1794"> zur Geschäftsordnung , ich danke den Kolleginnen und Kollegen im Ausschuss für Haushaltskontrolle , die dazu beigetragen haben , das gesamte Entlastungsverfahren . </seg>
<seg id="1830"> besonders hervorheben möchte ich die bestehende Zusammenarbeit zwischen den Sektoren Energie und Umwelt . </seg>
<seg id="1893"> die Vertreter der EU @-@ Mitgliedsstaaten wird Mitte Juni in Brüssel zu überprüfen , um eine gemeinsame Politik gegenüber Kuba . </seg>
<seg id="127"> abschlieÃ§end möchte ich das Parlament auffordern , nicht aus den Augen zu verlieren , die für die Entwicklung eines Impfstoffs AIDS und Malaria . </seg>
<seg id="221"> es ist schwer voraussagen , wenn die Kommission in der Lage sein werden , ihre Vorschläge zu unterbreiten . </seg>

<seg id="462"> die Türkei muÃ§ als eine einzigartige Gelegenheit , sich dieser Herausforderung auf ihrem Weg in Richtung Europa . </seg>
<seg id="483"> die Probleme in den benachteiligten Gebieten gelöst werden kann , durch eine gute Zusammenarbeit im Rahmen der Kohäsionspolitik der Union . </seg>
<seg id="578"> das schafft Synergieeffekte und erleichtert die gemeinsame Nutzung von Daten und die Verfahren , um ein besseres Verständnis zu erreichen ... </seg>
<seg id="902"> die von der Fischerei besonders schwer getroffen . </seg>
<seg id="1329"> ethnische , religiöse , insbesondere der christlichen und muslimischen , denen , die kulturelle und wirtschaftliche Spannungen , die in Nigeria . </seg>
<seg id="1286"> handelt es sich um 10 Uhr abends nichts weiter , und wir sind , wenn wir nicht die Redezeit erhalten . </seg>
<seg id="708"> das , was wir eigentlich noch ein Glas voller Worte und eine Hand voll von nichts . </seg>

6.1.4 Evaluated sentences translated using the pre-reordering approach

<seg id="1324"> seit Dezember , der Vertrag von Lissabon verpflichtet die EU die Menschenrechtskonvention beizutreten . </seg>
<seg id="1297"> die Kommission wurde aufgefordert , uns immer wieder nachdrücklich vollständig und unverzüglich über die ACTA @-@ Verhandlungen auf dem Laufenden halten . </seg>
<seg id="990"> die vorgeschlagene Verordnung würde verlässliche , transparente und vergleichbare Ergebnisse garantieren . </seg>
<seg id="161"> wir kämpfen ständig mit dem Element der unbekannten Gefahr . </seg>
<seg id="587"> dennoch ist unser Ziel in Sicht , und wir bewegen uns langsam aber sicher zu sein . </seg>
<seg id="683"> der Beschluss des Europäischen Rates halte ich für sehr klug . </seg>
<seg id="1132"> dieser Dialog ist hauptsächlich durch die Leiter der EU @-@ Missionen in den betreffenden Ländern verfolgt wird . </seg>
</seg>
<seg id="1687"> der Bericht fordert die Haushaltsbehörde verfügbar zu machen , ein funktionierendes Budget für 1999 für direkte MaÃ§nahmen der Gemeinschaft . </seg>
<seg id="1794"> zum Verfahren möchte ich den Kollegen im Ausschuss für Haushaltskontrolle danken , die sich auf das gesamte Entlastungsverfahren beigetragen haben . </seg>
<seg id="1830"> ich möchte die bestehende Zusammenarbeit zwischen der Energie- und Umweltpolitik Sektoren- hervorheben . </seg>
<seg id="1893"> die Vertreter der Mitgliedstaaten der EU in Brüssel zusammentreten wird Mitte Juni eine gemeinsame Politik gegenüber Kuba zu revidieren . </seg>
<seg id="127"> abschlieÃ§end möchte ich das Parlament dringend bitten , nicht aus den Augen , die Notwendigkeit der Entwicklung eines Impfstoffs AIDS und Malaria verliert . </seg>
<seg id="221"> es ist schwer voraussagen , wann wird die Kommission ihre Vorschläge vor-

legen können . </seg>
<seg id="462"> die Türkei muss diese Herausforderung als eine einzigartige Gelegenheit , auf seinem Weg in Richtung Europa zu interpretieren . </seg>
<seg id="483"> die Probleme in den verschiedenen benachteiligten Gebiete kann durch eine gute Zusammenarbeit im Rahmen der Politik des Zusammenhalts der Union gelöst werden . </seg>
<seg id="578"> dies wird , Synergien zu schaffen und die gemeinsame Nutzung von Daten und Arbeitspraktiken erleichtern , um ein besseres Verständnis ... </seg>
<seg id="902"> Kraftstoffpreise haben den Fischereisektor besonders schwer getroffen . </seg>
<seg id="1329"> ethnische , religiöse , insbesondere der Christen und Muslime , Stammesansprüche , kulturelle und wirtschaftliche Spannungen scheinen in Nigeria weit verbreitet werden . </seg>
<seg id="1286"> es ist 22 Uhr und wir nichts weiter tun , wenn wir nicht eine Zeitnische Sprachfluss erhalten . </seg>
<seg id="708"> was wir eigentlich noch bleibt , ist ein Glas voller Worte und eine Hand voll von nichts . </seg>

6.2 Error analysis for chapter 4

6.2.1 Evaluated source sentences

<seg id="12"> Mit Rücksicht darauf, dass die überstunden der dringlichste und brennendste Punkt hinsichtlich Forderungen und Beschwerden der Gewerkschafter waren, die vorige Woche die Produktion eingestellt hatten, hat die Geschäftsführung gestern nachmittag mit sofortiger Wirkung über die Aufhebung sämtlicher überstunden für den Monat Dezember entschieden, führte Vanek auf.</seg>
<seg id="24"> Ich habe in diesem Kampf nur einen Schlag abbekommen und bin körperlich sehr gut in Form, hatte Klitschko die schnelle Rückkehr in den Ring begründet.</seg>
<seg id="21"> Zerstörung und Besetzung sind völlig unakzeptable Mittel, sagte sie in einer Aktuellen Stunde am Donnerstag im Wiesbadener Landtag.</seg>
<seg id="21"> Niemand will das Jahresende in den Bergen ohne Schnee verbringen.</seg>
<seg id="14"> Diejenigen, die leistungsstarke Kinder untersuchen, sagen, diese haben oft eine natürliche Affinität für die Schule und einen ihnen innewohnenden Antrieb zum Erfolg.</seg>
<seg id="19"> Das Bundesprogramm, das Gracielas College-Tour finanziert hatte, ist ein nützliches Beispiel.</seg>
<seg id="6"> Bewandert in der komplizierten Geschichte und Politik des Walfangs, beschreibt er die lange Tradition der Arktiserforscher, die sich selbst bei der Jagd auf den Eisbären bewiesen haben.</seg>
<seg id="15"> Die russischen Unternehmen Lukoil und Gazprom waren die Hauptakteure bei zwei der dieses Wochenende zugesprochenen Verträge.</seg>

<seg id="20"> Der fünfte Kandidat, Vitali Banba, behauptet, weder die jetzige Regierung noch die Opposition zu unterstützen.</seg>
<seg id="2"> Die krisengeschüttelten US-GroÃ§banken streifen ihre staatlichen Fesseln ab.</seg>
<seg id="15"> Zur linken des Präsidenten zeigt sich der demokratische Vertreter Dennis Kucinich beunruhigt darüber, wie Barack Obama das Zurückgreifen auf die Streitkräfte verteidigt.</seg>
<seg id="121"> Traum ist ein anderer Zustand, beschreibt er sein Erleben.</seg>
<seg id="26"> Da die verwandte Linie der riesigen Sauropoden später ebenfalls Luftsäcke und Leichtbau-Knochen besaÃ§, hatte wahrscheinlich schon der gemeinsame Vorfahre beider Gruppen diese nützliche Innovation hervorgebracht.</seg>
<seg id="20"> Dies ist ganz klar ein Spiel, woraus sich eine neue wirtschaftliche Vorherrschaft entwickeln wird, so Ulate, der auch als regionaler Klimawandelberater für Conservation International in Mexiko und Mittelamerika tätig ist.</seg>
<seg id="31"> Neben Lob gab es aber auch Kritik an den Beschlüssen.</seg>
<seg id="12"> Frau Merkel muss klären, ob eine Strategie des gezielten Tötens Bestandteil der Afghanistan-Politik der Bundesregierung ist - und, ob Kanzleramt, Bundeswehr und Nachrichtendienst diese neue Strategie gebilligt haben, forderten Jürgen Trittin und Grünen-Kollegin Renate Künast.</seg>
<seg id="17"> Wir, das abchasische Volk, sind dankbar, hat er bei einer kleinen Feier gesagt, als Anspielung auf die Unterstützung Moskaus, insbesondere bezüglich der Entscheidung, Abchasiens Selbständigkeit anzuerkennen.</seg>
<seg id="18"> "Die Türkei kann diese (kurdische, Anm. d. Redaktion) Frage nicht anhand des Verbotes einer Partei klären", erklärte er der Presse.</seg>
<seg id="7"> Die Länder verpflichteten sich dazu, gemeinsame Strukturvorgaben für Bachelor- und Masterstudiengänge flexibler zu gestalten.</seg>
<seg id="27"> Brenders voraussichtlicher Nachfolger Peter Frey hatte sich dem Verwaltungsrat mit einem klugen Schachzug empfohlen - indem er das Gremium unmittelbar nach der Brender-Abwahl heftig dafür kritisierte.</seg>

6.2.2 Evaluated reference sentences

<seg id="12"> Considering that overtime was the most pressing and thorny question for the unions, as well as for the protesting employees who brought production to such a grinding halt last week, the management decided yesterday afternoon to immediately cancel all overtime for December, said Vanek.</seg>
<seg id="24"> I only took one blow in this fight, and am physically in very good shape, Klitchko said after the quick return to the ring.</seg>
<seg id="21"> Destruction and occupation are absolutely inacceptable means, the said in a topical session on Thursday in the Wiesbaden parliament.</seg>
<seg id="21"> Nobody wants to spend the end of the year in the mountains when there’s no snow.</seg>
<seg id="14"> Those who study high achievers say they often have a natural affinity for school

and an innate drive to succeed.</seg>
<seg id="19"> The federal program that funded Graciela's college tour is a useful example.</seg>
<seg id="6"> Well-versed in the complicated history and politics of whaling, he describes the long tradition of Arctic explorers who proved themselves by taking on the white bear.</seg>
<seg id="15"> Russian companies Lukoil and Gazprom were the top stakeholders in two of the contracts awarded this weekend.</seg>
<seg id="20"> The fifth candidate, Vitali Bganba, asserts not being in support of the current Government or the opposition.</seg>
<seg id="2"> The crisis-hit US major banks are breaking free from their state shackles.</seg>
<seg id="15"> To the president's left, the democratic representative Dennis Kucinich was alarmed to see Barack Obama justify resorting to force.</seg>
<seg id="121"> Dream is a different condition, he describes his experience.</seg>
<seg id="26"> As the related line of the huge sauropods later also had air bags and lightweight bones, the joint ancestor of these two groups will probably have come up with this useful innovation.</seg>
<seg id="20"> This is clearly a game where a new economic hegemony is being developed, said Ulate, who also serves as the regional Mexico and Central America climate change adviser for Conservation International.</seg>
<seg id="31"> On top of praise, there was also criticism for the decisions.</seg>
<seg id="12"> Merkel must clarify whether a strategy of targeted killing is part of the federal government's Afghanistan policy, and whether the Chancellor's Office, German army and intelligence service have approved this new strategy, demanded Jürgen Trittin and his Greens colleague Renate Künast.</seg>
<seg id="17"> We, the Abkhazian people, are thankful, he said during a small ceremony, alluding to the support from Moscow, and particularly to the decision to recognise the independence of Abkhazia.</seg>
<seg id="18"> Turkey cannot resolve this [Kurdish] issue by banning a party, he told the Press.</seg>
<seg id="7"> The states undertake to organise more flexibly joint structural requirements for bachelor and masters courses.</seg>
<seg id="27"> Brender's expected successor, Peter Frey, had recommended himself to the board of directors with a clever move - by harshly criticising the committee immediately after it voted out Brender.</seg>

6.2.3 Evaluated baseline sentences

<seg id="12"> In view of the fact that the overtime the most urgent and most burning issue in terms of demands and complaints of the trade unionists were the last week had yesterday afternoon, the management with immediate effect on the lifting of all overtime for the month of December, Vanek decided on.</seg>
<seg id="24"> I have in this fight only a blow âĂŹ and am physically very well in the form, had Klitschko the rapid return in the ring.</seg>

<seg id="21"> Destruction and occupation are completely unacceptable resources, she said in a recent hour on Thursday in Wiesbadener project.</seg>
<seg id="21"> Nobody wants the end of the year in the mountains without snow.</seg>
<seg id="14"> Those powerful children, say that this often have a natural affinity for school and an inherent momentum to success.</seg>
<seg id="19"> The Bundesprogramm, the Gracielas college Tour funded is a useful example.</seg>
<seg id="6"> Familiar with in the complicated history and policy of whaling, it presents the long tradition of Arktiserforscher, which even in the hunting of the polar bears.</seg>
<seg id="15"> The Russian company Gazprom Lukoil and were the main actors in two of the treaties remains this weekend.</seg>
<seg id="20"> The fifth candidate, Vitaly Banba, claiming that neither the present government to support the opposition.</seg>
<seg id="2"> The troubled US big banks touch on their government bonds.</seg>
<seg id="15"> To the left-wing of the President is the democratic representative Dennis Kucinich concerned, like Barack Obama resorting to the armed forces.</seg>
<seg id="121"> Dream is a different state of affairs, it presents its experience.</seg>
<seg id="26"> Since the related and foremost the enormous Sauropoden later also Luftsäcke and lighter vehicles-bones, was probably already the joint Vorfahre of both groups this useful innovation.</seg>
<seg id="20"> It is clear that this is a game, hence a new economic domination will develop, so Ulate, also as a regional Klimawandelberater for Conservation International in Mexico and Central America is operating.</seg>
<seg id="31"> Alongside praise, but also criticism of the decisions taken.</seg>
<seg id="12"> Mrs Merkel must clarify whether a strategy of targeted killing part of Afghanistan-policy of the Federal Government is - and whether Chancellery Bundeswehr and intelligence, this new strategy approved, have called for, Jürgen Trittin and Green colleague, Renate Künast.</seg>
<seg id="17"> We, the Abkhaz people, are grateful, he has said in a small celebration, as a reference to the support of Moscow, in particular with regard to the decision to recognise Abkhazia self-employment.</seg>
<seg id="18"> 'Turkey can this (Kurdish, Anm. that drafting) question is not on the basis of the ban on a party, he said "clarify the press.</seg>
<seg id="7"> The countries committed themselves to common Strukturvorgaben for Bachelor-and degrees more flexible.</seg>
<seg id="27"> Brenders voraussichtlicher successor Peter Frey had the Management Board with a wise move, recommended by the body immediately after the Brender-be lifted strongly criticised for it.</seg>

6.2.4 Evaluated sentences translated with the word adapted language model

<seg id="12"> In view of the fact that the overtime the most urgent and most burning issue in terms of demands and complaints of the trade unionists were the last week had yesterday afternoon, the management with immediate effect on the lifting of all overtime for the month of December, led Vanek decided on.</seg>
<seg id="24"> I have in this struggle only a blow âĂŹ and am physically very well in the form, had Klitschko the rapid return in the ring.</seg>
<seg id="21"> Destruction and occupation are completely unacceptable resources, she said in a recent hour on Thursday in Wiesbadener project.</seg>
<seg id="21"> Nobody wants the end of the year in the mountains without snow.</seg>
<seg id="14"> Those who are powerful children, say that this often have a natural affinity for school and an inherent momentum to success.</seg>
<seg id="19"> The Bundesprogramm, the Gracielas college Tour funded is a useful example.</seg>
<seg id="6"> Familiar with in the complicated history and policy of whaling, it presents the long tradition of Arktiserforscher, which even in the hunting of the polar bears.</seg>
<seg id="15"> The Russian company Gazprom Lukoil and were the main actors in two of the contracts are this weekend.</seg>
<seg id="20"> The fifth candidate, Vitaly Banba, claiming that neither the present government to support the opposition.</seg>
<seg id="2"> The troubled US big banks touch on their government bonds.</seg>
<seg id="15"> To the left-wing of the President is the democratic representative Dennis Kucinich concerned, like Barack Obama resorting to the armed forces.</seg>
<seg id="121"> Dream is a different state of affairs, it presents his experience.</seg>
<seg id="26"> Since the related and foremost the enormous Sauropoden later also Luftsäcke and lighter vehicles-bones, was probably already the common Vorfahre of both groups this useful innovation.</seg>
<seg id="20"> It is clear that this is a game, hence a new economic domination, will develop Ulate, also as a regional Klimawandelberater for Conservation International in Mexico and Central America is operating.</seg>
<seg id="31"> Alongside praise, but also criticism of the decisions taken.</seg>
<seg id="12"> Mrs Merkel must clarify whether a strategy of targeted killing part of Afghanistan-policy of the Federal German Government is - and whether Chancellery, Bundeswehr and intelligence have approved this new strategy, demanded Jürgen Trittin and Green colleague, Renate Künast.</seg>
<seg id="17"> We, the Abkhazian people, are grateful, he has said in a small celebration, as a reference to the support of Moscow, in particular with regard to the decision to recognise Abkhazia independence.</seg>
<seg id="18"> 'Turkey can this (Kurdish, Anm. that drafting) issue does not clarify the basis of the ban on a party', he told the press.</seg>
<seg id="7"> The countries committed themselves to common Strukturvorgaben for Bachelor-

and Masters Courses more flexible.</seg>
<seg id="27"> Brenders voraussichtlicher successor Peter Frey had the board of directors with a wise move, recommended by the body immediately after the Brender-be lifted strongly criticised for it.</seg>

6.2.5 Evaluated sentences translated with the interpolated adapted language model

<seg id="12"> In view of the fact that the overtime of the most pressing and most burning point in terms of demands and complaints of the trade unionists were, which had ceased production last week, the management with immediate effect yesterday afternoon on the lifting of all overtime for the month of December, led Vanek decided on.</seg>
<seg id="24"> I have only a blow in this fight obtain and am physically very well in the form, had Klitschko the rapid return in the ring-founded.</seg>
<seg id="21"> Destruction and occupation are totally unacceptable means, she said in a recent hour on Thursday in Wiesbadener project.</seg>
<seg id="21"> Nobody wants the end of the year without snow spend in the mountains.</seg>
<seg id="14"> Those who are powerful children investigate say that this often have a natural affinity for the school and a inherent impetus to success.</seg>
<seg id="19"> The Bundesprogramm Gracielas college tour, which had financed, is a useful example.</seg>
<seg id="6"> Familiar with in the complicated history and politics of whaling, he describes the long tradition of Arktiserforscher themselves in the hunting of the polar bears have demonstrated.</seg>
<seg id="15"> The Russian companies Lukoil and Gazprom were the main actors in two of the this weekend obtained contracts.</seg>
<seg id="20"> The fifth candidate, Vitaly Banba, claiming that neither the current government or the opposition to support.</seg>
<seg id="2"> The troubled US banks roaming their state shackles.</seg>
<seg id="15"> To the left of the President shows the democratic representatives Dennis Kucinich concerned, as Barack Obama resorting to the armed forces.</seg>
<seg id="121"> Dream is a different situation, he describes his experience.</seg>
<seg id="26"> Since the related line of the huge Sauropoden also later Luftsäcke and lighter vehicles-bones possessed, was probably already the joint Vorfahre of both groups this useful innovation.</seg>
<seg id="20"> This is clearly a game, which will develop a new economic domination, so Ulate, also as a regional Klimawandelberater for Conservation International in Mexico and Central America is operating.</seg>
<seg id="31"> Alongside praise, though, there was also criticism of the decisions.</seg>
<seg id="12"> Mrs Merkel must clarify whether a strategy of targeted killing part of Afghanistan-policy of the Federal German Government is - and whether Chancellery, Bundeswehr and intel-

ligence have approved this new strategy, demanded Jürgen Trittin and Green colleague, Renate Künast.</seg>
<seg id="17"> We, the Abkhazian people, are grateful, he was in a small celebration said, as a reference to the support of Moscow, in particular with regard to the decision to recognise Abkhazia independence.</seg>
<seg id="18"> 'Turkey can this (Kurdish, Anm. that drafting) issue does not clarify the basis of the ban on a party', he told the press.</seg>
<seg id="7"> The countries committed themselves to, common Strukturvorgaben for Bachelor- and Courses more flexible.</seg>
<seg id="27"> Brenders voraussichtlicher successor Peter Frey the Administrative Board had with a wise move recommended by the body - immediately after the Brender-dismissal strongly criticised for it.</seg>