Linguistic Structure in Statistical Machine Translation

zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften

der Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Teresa Herrmann

aus Dresden

Tag der mündlichen Prüfung:	19. Juni 2015
Erster Gutachter:	Prof. Dr. Alexander Waibel
Zweiter Gutachter:	Dr. Stephan Vogel

Abstract

The first approaches to machine translation required linguists or language experts writing language-dependent translation rules by hand. In the last 20 years, statistical approaches applying machine learning techniques have become the state-of-the-art. They automatically learn translation rules from large parallel corpora of existing translations. This facilitates fast development of translation systems for new languages without the need for language experts writing rules. In phrase-based machine translation, new translations are constructed from phrases seen in previous translations during training. In most cases, this flat composition of phrases leads to good results, especially for translations between languages with similar sentence structure.

However, this strong suit of statistical translation exposes a weakness, if there are structural differences between the languages. German is one of the languages where structural changes are necessary during translation, which increases the difficulty for translation. A particular challenge is the position of the verb in the German sentence, which depends on various factors that need to be taken into consideration when translating from or into German. The knowledge about sentence structure could provide helpful information for the translation process.

This thesis investigates the influence of linguistic structure in statistical machine translation. The first part of this thesis addresses the particular challenge that differences in word order between languages present for translation. We develop a word reordering model based on the structural information inherent in phrase structure trees. This reordering model is included as a separate component in a phrase-based machine translation system. It consists of rules on how to change the word order when translating from one language into another. These rule are automatically learned

from a parallel corpus that was annotated with syntactic parse trees for the source language. These rules define the search space for possible reorderings of the source sentence during translation. In comparison to a part-of-speech-based reordering model alone, the combination of the two models achieves an improvement of the word order in the target language, which can be further increased when including a lexicalized reordering model. Hence, the different word reordering models operating on different linguistic levels (words, parts-of-speech and syntax) have shown complementary effects which can be increased further when combined.

Measuring translation quality with regard to word order is difficult with automatic evaluation metrics. We perform a manual evaluation of the reordering approach on three different data sets representing different genres. Although the amount of affected sentences varies between genres, the manual evaluation of the translation quality on German-English translation confirms consistent improvements on all three data sets. The improvements introduced by the syntactic reordering model consist of translations of words that were removed from the translation before, as well as improved positions of words and whole constituents in the translated sentence. As intended in the design of the syntactic reordering model, verbs are the most affected word category.

Furthermore, we analyzed the potential and limits of the syntax-based reordering approach with experiments regarding the performance of the syntax-based and the part-of-speech-based reordering approaches. Oracle experiments revealed the maximal improvement that can be achieved with the syntax-based reordering model. First, the upper bound was determined for the source sentence reordering approach in general. This was compared with the highest achievable performance of the syntax-based approach and its actual performance generated by the decoder. It can be shown that the syntax-based reordering rules improve the search space of available reordering possibilities. More correct reordering options are generated and when translating from German to English those are also chosen frequently for translation. For translation from English into German, there is potential

for improving the search among the reordering options to generate better translations. The experiments also indicate that the reordering model would benefit from additional rules to expand the search space further in order to approximate the word order even better.

In the second part of the thesis we address the issues of translating pronouns and improving the morphological agreement for morphologically rich languages in statistical machine translation. For disambiguating between all possible translation options for a word, context information is needed. Especially when morphological agreement between words such as pronouns and their antecedent, or subjects and verbs is required, the dependencies between words are also important.

A source discriminative word lexicon (SDWL) is developed for predicting the translation for individual source words in their respective contexts. The prediction is performed as a classification task, where the context words and dependency relations of the given source word are used as structural features of the source sentence to guide the translation prediction. The evaluation of the prediction accuracy shows improved translation prediction by the SDWL over a baseline classifier, especially for pronouns, subjects and verbs. The translation predictions for individual words are combined on the sentence level and used in N-best list re-ranking. According to automatic evaluation, the translation quality after re-ranking is improved.

Zusammenfassung

In den ursprünglichen Ansätzen zur maschinellen Übersetzung waren Linguisten oder Sprachexperten notwendig, die sprachspezifische Übersetzungsregeln per Hand schreiben mussten. In den letzten 20 Jahren sind statistische Ansätze zur maschinellen Übersetzung zum State-of-the-Art geworden. Dabei werden mit Hilfe von Verfahren aus dem Maschinellen Lernen Übersetzungsregeln automatisch aus großen Korpora bestehender Übersetzungen gelernt. Diese Vorgehensweise ermöglicht die schnelle Entwicklung von Übersetzungssystemen für neue Sprachen, ohne dass Expertenwissen notwendig ist. Das Zusammensetzen einer neuen Übersetzung aus Mehrwortphrasen im Training gesehener Übersetzungsblöcke führt in den meisten Fällen zu guten Übersetzungen, besonders wenn die Übersetzung zwischen Sprachen erfolgt, deren Satzbau einer ähnlichen Struktur folgt.

Wenn Unterschiede in der Sprachstruktur zwischen den Sprachen bestehen, stellt dies allerdings eine Schwäche dar. Deutsch ist eine der Sprachen, bei denen während der Übersetzung strukturelle Änderungen notwendig sind, die den Übersetzungsprozess erschweren. Eine besondere Schwierigkeit im Deutschen stellt die Stellung des Verbs im Satz dar, welche von verschiedenen Faktoren abhängt, die bei Übersetzungen ins oder aus dem Deutschen beachtet werden müssen. Die Kenntnis der sprachlichen Struktur des Satzes kann jedoch für die Übersetzung hilfreiche Informationen liefern.

In dieser Arbeit wird der Einfluss von linguistischer Struktur in der statistischen maschinellen Übersetzung untersucht. Unterschiede in der Wortstellung zwischen Sprachen stellen für die Übersetzung eine besondere Schwierigkeit dar. Im ersten Teil der Arbeit wird eine Komponente für die Modellierung von Wortumordnungen entwickelt, die auf linguistischen Strukturinformation aus Phrasenstrukturbäumen basiert. Die Regeln für die

Wortstellungsänderungen werden automatisch aus einem parallelen Korpus, dessen Quellsprachseite mit syntaktischen Phrasenstrukturbäumen annotiert wurde, gelernt. Die gelernten Regeln definieren den Suchraum der möglichen Umordnungen des Quellsatzes für mögliche Übersetzungen. Gegenüber einem Umordnungsmodell das nur auf Wortkategorien basiert, können durch die Kombination der beiden Modelle Verbesserungen in der Wortstellung erzielt werden, die noch gesteigert werden können durch das weitere Hinzufügen eines lexikalischen Umordnungsmodells. Es kann somit gezeigt werden, dass die verwendeten Modelle zur Modellierung der Wortstellung in der Übersetzung zueinander komplementäre Effekte haben, die sich zur weiteren Steigerung der Wirksamkeit kombinieren lassen.

Die Übersetzungsqualität in Bezug auf die Wortstellung ist mit automatischen Evaluationsmetriken schwer zu bewerten. Daher wird in der Arbeit eine manuelle Evaluation des Umordnungsansatzes auf drei verschiedenen Textgenres und Stilrichtungen durchgeführt. Obwohl die Anzahl der jeweils betroffenen und analysierten Sätze variiert, können in einer manuellen Evaluation der Übersetzungsqualität auf Satzebene konsistente Verbesserungen auf allen drei Datensätzen nachgewiesen werden. Die Verbesserungen, die das Syntaxmodell einbringt, bestehen aus Übersetzungen für Wörter, die vorher aus der Übersetzung entfernt wurden, sowie die verbesserte Position von einzelnen Wörtern und ganzen Satzkonstituenten im übersetzten Satz. Wie im Entwurf des syntaktischen Umordnungsmodells vorgesehen, sind Verben die hauptsächlich betroffene Wortkategorie.

Des weiteren wurden die Möglichkeiten und Grenzen des Ansatzes analysiert. Dazu wurden Experimente zur Performanz der auf Wortkategorien und Syntaxbäumen basierenden Umordnungsmodelle durchgeführt. Mit Hilfe von Orakelexperimenten wurde die maximal erreichbare Verbesserung durch das auf Syntaxbäumen basierende Modell untersucht. Es wurde die obere Grenze für den Umordnungsansatz durch Umstellung der Quellsprachwörter gemäß der korrekten Wortstellung in der Zielsprache ermittelt. Dieser Übersetzung, die die maximal erreichbare Übersetzungsqualität mit diesem

Ansatz darstellt, wurden zwei Hypothesen gegenübergestellt: die Übersetzung, die die beste erreichbare Wortstellung mit dem syntaxbasierten Umordnungsmodell verwendet, sowie die vom Decoder generierte Übersetzung. Es kann gezeigt werden, dass durch die syntaktischen Umordnungsregeln der Suchraum der verfügbaren Umordnungen verbessert werden kann. Mehr korrekte Umordnungen sind verfügbar und für Übersetzungen aus dem Deutschen werden diese auch häufig für die Übersetzung ausgewählt. Für Umordnungen des englischen Quellsatzes könnte durch Verbesserung der Suche eine bessere Übersetzungsleistung erreicht werden. Des weiteren haben die Experimente gezeigt, dass das Modell von zusätzlichen Regeln profitieren könnte, die den Suchraum der Umordnungen erweitern, sodass die bestmögliche Wortstellung noch besser approximiert werden kann.

In einem weiteren Teil der Arbeit wird das Problem der Übersetzung von Pronomina und die Verbesserung der morphologischen Kongruenz für Sprachen mit komplexer Morphologie in der statistischen maschinellen Übersetzung behandelt.

Um zwischen allen möglichen Übersetzungen eines Wortes zu disambiguieren, werden Kontextinformationen benötigt. Insbesondere wenn morphologische Kongruenz zwischen bestimmten Wörtern, wie zum Beispiel Pronomen und deren Antezedenten, oder zwischen Subjekt und Verb bestehen muss, sind die Abhängigkeiten zwischen Wörtern ebenfalls wichtig. In dieser Arbeit wird ein diskriminatives Wortlexikon-Modell für die Quellsprache (source discriminative word lexicon, SDWL) vorgestellt, das die Übersetzung für einzelne Quellwörter in ihren entsprechenden Kontexten vorhersagt. Die Vorhersage erfolgt als Klassifikation anhand struktureller Merkmale des Quellsatzes bestehend aus Kontextwörtern und Dependenzrelationen des zu übersetzenden Quellwortes. Eine Evaluation der Vorhersagegenauigkeit zeigt, dass das SDWL Übersetzungen vorhersagt, die im Besonderen für Pronomen, Subjekte und Verben deutliche Verbesserungen gegenüber der Baseline aufweisen. Diese Klassifikatorvorhersagen für einzelne Wörter werden auf Satzebene kombiniert und für das Reranking

der N-Bestenliste des Decoders verwendet. Eine automatische Evaluation zeigt, dass die Übersetzungsqualität dadurch verbessert werden kann.

Acknowledgements

First of all I want to express my gratitude towards my advisor *Prof. Dr. Alex Waibel* for providing me with the opportunity to perform this research and for fruitful discussions and support during the course of my studies. I would also like to thank *Dr. Stephan Vogel* for co-advising my thesis, his time and effort.

I am indebted to my colleagues at the Interactive Systems Lab, who provided a productive and enjoyable working environment. Our discussions, joint work and mutual support contributed considerably in accomplishing this work. My thanks go to my colleagues in the machine translation group: *Eunah Cho*, *Thanh-Le Ha*, *Silja Hildebrand*, *Mohammed Mediani*, *Jan Niehues*, *Isabel Slawik* and *Yuqi Zhang*. Since collaboration and support does not stop at boundaries of topics and tasks at out lab, I would also like to thank all the other current and past members: *Silke Dannenmaier*, *Sarah Fünfer*, *Jonas Gehring*, *Michael Heck*, *Klaus Joas*, *Kevin Kilgour*, *Narine Kokhlikyan*, *Florian Kraft*, *Bastian Krüger*, *Patricia Lichtblau*, *Mirjam Mäß*, *Christian Mohr*, *Markus Müller*, *Bao-Quoc Nguyen*, *Huy-Van Nguyen*, *Thai-Son Nguyen*, *Margit Rödder*, *Virginia Roth*, *Christian Saam*, *Rainer Saam*, *Maria Schmidt*, *Matthias Sperber*, *Sebastian Stüker*, *Yury Titov*, *Joshua Winebarger* and *Liang Guo Zhang*. Furthermore, sharing office and joint lunches with the research team of the former Mobile Technologies GmbH (*Christian Fügen*, *Thilo Köhler* and *Kay Rottmann*) led to interesting and out-of-the-box insights. I also include them in my thanks.

I am grateful for the funding I received from the Interactive Systems Lab and through the projects I was involved in (Quaero, EU-Bridge) and also for access to the infrastructure of the lab, which allowed me to perform the research presented in this thesis.

I also want to thank *Dr. Andreas Eisele*, who introduced me to machine translation during my Master's studies, an exciting field, which has not stopped fascinating me since.

Last but not least, I am thankful to my friends and family, and especially to *Bastian Karweg* for constant encouragement and support.

Contents

List of Figures

List of Tables

Glossary

ASR Automatic Speech Recognition

BART Beautiful Anaphora Resolution Toolkit, a toolkit for automatic anaphora and co-reference resolution

BLEU Bilingual Evaluation Understudy, the most widely used evaluation metric for machine translation (Papineni et al., 2002)

BOW Bag-of-Words, a representation of a sequence of words where encoding consists of the uniquely participating words without information on their original order

CBOW Continuous Bag-of-Words, learning algorithm implemented in word2vec

CRF Conditional Random Field, discriminative learning framework for sequence labeling

CS Computer science

DE German (language)

Dev Development (Set), data that is used to train the log-linear translation model

DWL Discriminative Word Lexicon

EBMT Example-Based Machine Translation, a corpus-based approach to machine translation

EM Expectation Maximization, estimation algorithm for parameters in statistical models

EN English (language)

EPPS European Parliament Plenary Sessions, the largest available parallel corpus for most European languages

FR French (language)

GIZA(++) Implementation of the IBM models

HMM Hidden Markov Model

IBM Models Word-based translation models, first approach to statistical machine translation

ID Numerical identifier

IWSLT International Workshop on Spoken Language Translation

KenLM language modeling toolkit (Heafield et al., 2013)

KIT Karlsruhe Institute of Technology

KOUS POS tag: subjunctive conjunction

LexRM Lexicalized reordering model, orientation-based reordering model used in phrase-based machine translation

LM Language Model

MERT Minimum Error Rate Training, most commonly used optimization method for phrase-based machine translation

MKCLS Word cluster algorithm

MT Machine Translation

NC News Commentary, parallel corpus of several European l anguages

NLP Natural Language Processing, field in the area of computer science and computational linguistics that concentrates on processing natural language data

NN POS tag: noun; neural network

NNP POS tag: proper noun

NP POS tag: noun phrase

GLOSSARY

OOV Out-of-Vocabulary (word)

PBMT Phrase-Based Machine Translation, statistical machine translation approach

POS Part of speech, grammatical classes describing the function of a word in a sentence

PP POS tag: prepositional phrase

PPER POS tag: personal pronoun

PTKNEG POS tag: negative particle

PTKVZ POS tag: separable verb prefix

RBMT Rule-Based Machine Translation, comprehensive term for linguistic approaches to machine translation based on hand-written rules

RIBES Rank-based Intuitive Bilingual Evaluation Score, evaluation metric for jointly measuring machine translation and word order quality

SDWL Source discriminative word lexicon

SLT Spoken Language Translation

SMT Statistical Machine Translation, comprehensive term for statistical approaches to machine translation

SOV Subject-object-verb, order of the main word categories in a sentence (e. g. in the Japanese language)

SRI Research Institute

SRILM SRI Language Model, most commonly used language modeling toolkit

SVO Subject-verb-object, order of the main word categories in a sentence (e. g. in the English language, partially in the German language)

TED Technology, Entertainment, Design; global conference where invited speakers give talks on various topics. The talks are transcribed and translated into many languages.

TER Translation Error Rate, evaluation metric for machine translation

Test Test (Set), data that is used to test a translation system

TM Translation Model

TO POS tag: preposition to

VAFIN POS tag: finite auxiliary verb

VBZ POS tag (Penn Treebank tag set): finite verb, 3rd person singular present

VFIN POS tag: finite full verb

VMFIN POS tag: finite modal verb

VP POS tag: verb phrase

VVIMP POS tag: imperative form of full verb

VVPP POS tag: past participle of full verb

WIT^3 Web inventory of transcribed and translated talks (Cettolo et al., 2012), multilingual collection of TED talks

WMT Workshop on Machine Translation

word2vec Framework to learn continuous representations for words (Mikolov et al., 2013)

WSD Word Sense Disambiguation, field in computational linguistics

1

Introduction

Since the rise of statistical machine translation in the 1990s (Brown et al., 1990, 1993), phrase-based machine translation is one of the state-of-the-art approaches that continues to prove competitive in international evaluations of machine translation.

Compared to early approaches, language experts writing translation rules by hand are no longer needed. Instead, the translation is composed of pieces of previous translations collected in huge bilingual corpora. The choices of words and their order in the target sentence is guided by the statistical probability of word sequences previously seen in texts written in the target language. This process of flat composition of phrases without knowledge of the linguistic structures of source and target language works reasonably well, especially for languages that share a similar structure.

It is exactly this strong suit that also exposes a weakness as soon as differences in the sentence structure occur. German is one of the languages where structural changes during translation render the translation process difficult. In fact, this is observable in translation benchmarks when looking at translation results of language pairs where German is involved. The average translation quality is a lot lower for translations between German and English than between French and English, for example. Translating into German is even more difficult as results from international evaluation campaigns show (Callison-Burch et al., 2012a; Cettolo et al., 2013). One particular difficulty is the position of the verb in the German sentence, which depends on various factors that need to be taken into consideration when translating from and into German. The linguistic structure of the sentence can provide useful information under these circumstances.

Many researchers resort to hierarchical and syntax-based approaches to machine translation in such a scenario where structure is deemed beneficial for the translation process. However these approaches introduce an additional complexity which makes the search more difficult and slows down the system. We chose to stay within the phrase-based paradigm and introduce the linguistic structure by developing dedicated models to be used within the framework of the phrase-based machine translation system.

In this thesis we investigate the influence of linguistic structure in statistical machine translation exemplified on the German-English language pair. In the first part of this thesis we deal with the differences in word order between languages. Even though German and English stem from the same Germanic language family, producing the correct word order in the target language when translating from one into the other is a challenge. In order to address this issue, linguistic structure of the source language sentences is exploited in the development of a reordering model that learns how to change the word order of the source language sentence in order to achieve the word order of the target language. Then the translation can be performed without the need for additional reordering. This model is compared and combined with other successful reordering approaches like part-of-speech-based reordering and lexicalized reordering. We assess the potential of the source reordering approach with oracle experiments and perform a manual evaluation to confirm the performance of the model from a human point of view on three genres.

A second line of research is dedicated to the problem of pronoun translation and the improvement of morphological agreement for morphologically rich target languages in statistical machine translation. For disambiguating between all possible translation options for a word, context information is needed. Especially when morphological agreement between words such as pronouns and their antecedent, or subjects and verbs is required, the dependencies between words are also important.

We develop a disambiguation model (source discriminative word lexicon) that learns to predict the translation for each source word in a given source sentence by performing a classification task. Local word context and dependency relations between source words are used to represent the source sentence for classification. These structural features are intended to guide the translation prediction for the individual source words in a way that better choices can be made between different translation options. The

translation predictions for individual words are combined to form a sentence score in order to reassess translation hypotheses in N-best list re-ranking.

1.1 Overview

This section gives an overview of the contents of the individual chapters of this thesis.

Chapter 1 provides an introduction into the topic of this thesis: linguistic structure in statistical machine translation.

Chapter 2 explains the background of the thesis, presenting an introduction into the different approaches to machine translation, both rule-based and statistical ones. However, special focus is placed on the description of the phrase-based approach to statistical machine translation, which is applied in this work. In addition, we briefly introduce selected linguistic concepts which are relevant in the scope of this thesis.

Chapter 3 describes the particular linguistic challenges met in statistical machine translation more in detail, presenting translation examples which show the difficulty a phrase-based machine translation system has when dealing with linguistic phenomena such as word reordering, morphological agreement and pronominal anaphora.

Chapter 4 presents an overview of other research related to the topics dealt with in this thesis.

Chapter 5 introduces first the different types of data that are used for training and testing the methods developed in this thesis. In addition, we present a description of the statistical machine translation system and the particular components which play a significant role with regard to the experiments described in the following.

Chapter 6 presents the development, experiments and evaluation of the syntactic tree-based word reordering model developed within this thesis. After presenting a motivation, we introduce the framework for the development of the syntactic tree-based reordering model. Then the training and application of the reordering model based on syntactic parse trees is described. It is tested on several language

pairs and data sets. It is further compared and combined with other successful reordering techniques. The second part of this chapter assesses the potential of the source reordering method. First, an upper bound is established for setting the current performance of the approach in context with its optimal performance, uncovering possibilities for further improvements. In the third part of this chapter, the tree-based reordering approach is evaluated manually on three different data sets. The general impact of the model is assessed and a sentence-by-sentence comparison is conducted. In addition, a fine-grained examination gives insights on the types of changes the model introduces and the affected word classes.

Chapter 7 presents a second contribution to improving linguistic aspects in statistical machine translation. A translation disambiguation model is developed to perform predictions for translations of individual source words. A source discriminative word lexicon model predicts the translation for a given source word. Context and dependency relations are used to represent structural information features for this classification task. The individual translation predictions are combined as sentence features for N-best list re-ranking of machine translation output. First, a targeted evaluation of pronouns, subjects and verbs measures prediction accuracy. Secondly, the translation quality is measured after N-best list re-ranking.

Chapter 8 summarizes the experiments using linguistic structure for improving statistical machine translation, draws conclusions and suggests research directions for future work.

Appendix A describes a related project on anaphora resolution in machine translation and presents its results for comparison with the translation disambiguation model in Chapter 7.

2

Background

This chapter presents the fundamentals of machine translation, introducing the rule-based and statistical approaches to machine translation and how the quality of machine translation can be evaluated. Afterwards follows a short introduction of the linguistic concepts relevant for this thesis.

2.1 Machine Translation

In this section we give a short overview of the different approaches that can be adopted when the translation of a text from one natural language into another language is to be performed automatically without the need for a human translator.

2.1.1 Rule-based Machine Translation

The first approaches to machine translation were rule-based. It consists of translation rules that form the core of the translation approach. These rules are written by language experts to model the translation process from one language into another. Rule-based machine translation systems can be distinguished based on the level of abstraction applied during the translation process. Figure 2.1 depicts the Vauquois triangle (Vauquois and Boitet, 1985), illustrating the levels at which translation can take place. At the bottom of the triangle reside the machine translation approaches performing **direct translation** from the source language into the target text. No linguistic abstraction is carried out. Instead the translation is done at the surface word level or with minimal morphological processing. The simplest form of this approach consists of stemming

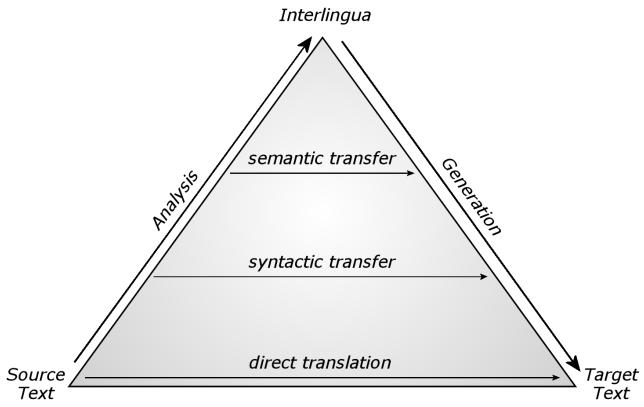

Figure 2.1: The Vauquois triangle

and dictionary look-up of individual words or multi-word phrases and performing a word-by-word translation. The rules consist of bilingual dictionary entries and possibly instructions for morphological processing.

Transfer-based machine translation approaches are characterized by a transfer component that performs the translation on a linguistic abstraction level such as syntax or semantics. An analysis component is applied to the surface words of the source language bringing them into the abstract representation of choice, conducting for example a morphological analysis and syntactic parsing to achieve a syntactic representation. Then transfer rules are applied to transform the source language representation into the corresponding representation of the target language. Finally, a generation component produces the target language sentence. The analysis and generation components are referred to as the grammar of the translation system. They encode the monolingual knowledge of word morphology and how to construct sentences from words. Typically, analysis and generation perform inverse operations, only that analysis operates on the source language and generation on the target language. The lexicon contains the rules that perform the mapping of words and abstract representations between the source and target language. The higher up in the triangle the transfer takes place, the more complex are analysis and generation, and the more abstract and language-independent the representation for transfer.

Interlingua Machine Translation is the approach that employs the highest degree of abstraction by operating in a so-called interlingua, which is in theory completely

language independent, so that no transfer is necessary any more. The target language sentence can be generated directly from the interlingua representation.

Common advantages of the rule-based approaches are their linguistic motivation. Language experts in source and target language explicitly model the human translation process in the particular languages. The disadvantage however, resides on the same matter, the necessity of a human expert to write the rules for each new language pair. The interlingua approach reduces the effort from quadratic to linear in the number of languages involved, but a true language independence is commonly regarded as not feasible for open-domain machine translation. Hence, interlingua approaches are typically only applied in well-defined scenarios e.g. as described in Levin et al. (1998) who present an interlingua for the translation of task-oriented dialogues in the travel domain for six languages.

2.1.2 Statistical Machine Translation

Statistical approaches to machine translation emerged in the 1990s (Brown et al., 1990, 1993). They make use of parallel corpora consisting of translations in order to learn statistical models on how to translate from one language into another and how to generate sentences in the target language. The original translation models modeled word-based translation, but they gave rise to a new era of statistical approaches to machine translation, which developed to be the current state-of-the-art. Among the most important are phrase-based machine translation, syntax-based machine translation and hierarchical machine translation, which will be described in the following sections.

These statistical approaches to machine translation have greatly reduced the human effort necessary for developing a machine translation system compared to the rule-based approaches presented above. However, this happened at the expense of the linguistic modeling, the arising issues of which we will discuss later on in this chapter.

In the presented descriptions of the statistical machine translation approaches and methods we have followed Koehn (2010), which may be referred to for more detailed information.

2.1.2.1 Word-based Translation Models

Even though word-based translation is not used as a stand-alone translation system any more, the original word-based models introduced in Brown et al. (1993) are still the

core of many state-of-the-art statistical translation systems. Hence, we will give a short overview of the word-based models. Brown et al. (1993) propose a cascaded model of five models, referred to as **IBM models**. Nowadays extended by a sixth model, they model the translation of words in a cascaded form with increasing complexity of each model. Provided a parallel bilingual corpus which is sentence-aligned, we want to learn a word correspondence, i.e. word alignment indicating which words are translations of each other. This problem, which is characterized by an interdependency of data and model, is addressed with the expectation-maximization (EM) algorithm, estimating the lexical probabilities and learning the alignment model in alternating steps until convergence. The result is called **IBM Model 1**, modeling the lexical correspondences on the word level without taking word reordering, nor the insertion and deletion of words into account. **IBM Model 2** improves the translation model in these regards additionally modeling absolute word positions by means of an alignment probability distribution. **IBM Model 3** introduces the concept of fertility, which allows one word in the source language to generate several words in the target language. This is for example the case in English-French translation, where the English negation *not* is realized in French by a construction of two words *ne ... pas* surrounding the negated word. The fertility value is then equal to the number of target words that a source word induces. **IBM Model 4** adds an alignment model with relative positions which encourage reordering of whole constituents and also introduces word classes for obtaining better probability estimates. **IBM Model 5** fixes deficiencies introduced in previous models which allowed multiple assignment of the same target position. In some systems a Hidden Markov Model, i.e. **HMM Model** (Vogel et al., 1996) is applied instead of IBM Model 2 in order to already use relative positions early on in training. Since only IBM Model 1 is guaranteed to find a global maximum and each of the following IBM models increases in complexity, the outputs of IBM Model 1 are typically used as initialization parameters for IBM Model 2 and those outputs again as initialization for IBM Model 3 and so forth.

2.1.2.2 Phrase-based Machine Translation

Phrase-based machine translation (Koehn et al., 2003) is one of the most important state-of-the-art statistical machine translation approaches, not least due to the availability of the open-source Moses toolkit (Koehn et al., 2007b). Phrase-based machine

translation builds upon the word-based models described above. Translating word sequences instead of single words at a time leads to better translations, since more context can be taken into account. For most words, there are no one-to-one translations. Many words need to be translated in context, especially multi-word expressions or idioms need to be handled as a unit, otherwise translation will not produce an acceptable result. Fertility and word deletions, as well as local reordering can be handled more effectively when using phrases as the smallest units for translation.

Translation Model The core of a phrase-based machine translation system is the phrase translation table. GIZA++, an implementation of the word-based models by Och and Ney (2003) is used in Moses to generate the word alignments and the word translation lexicon. Typically it is applied in both source-to-target and target-to-source direction. Then the heuristic combination of both alignments is extended by neighboring words to extract phrases that are consistent with the word alignment. A phrase pair is called consistent, if all source words within a phrase pair are aligned to target words within the phrase pair and all source words outside the phrase pair are aligned to target words outside the phrase pair. Phrase translation probabilities are computed from the relative frequencies of the phrases in the parallel training corpus. The translation model consisting of the phrase translation table is one of the core components of the phrase-based translation system.

Language Model Another important component is the language model (LM). It models the fluency of the generated target sentence by estimating the probability of a sentence being a good representative of the target language. The most common type in machine translation are n-gram language models. An n-gram language model is a collection of statistics on the distribution of target language n-grams (sequences of n words) in large monolingual corpora in the target language. The statistics are based on maximum likelihood estimation from n-gram occurrences in this corpus. After training, the language model can provide a probability for individual words as well as complete sentences. For each word w_i in a sentence, the n-gram probability takes the history of $n-1$ previous words into account. The probability of a sentence s consisting of words $w_1, w_2, w_3 ... w_l$ is calculated as the product of the individual n-gram probabilities.

Equation 2.1 exemplifies this assuming a trigram language model.

$$\begin{aligned} p_{LM}(s) &= p_{LM}(w_1)p_{LM}(w_2|w_1)p_{LM}(w_3|w_1w_2)\dots p_{LM}(w_l|w_{l-2}w_{l-1}) \\ &= \prod_{i=1}^{l} p_{LM}(w_i|w_{i-2}w_{w-1}) \end{aligned} \tag{2.1}$$

The larger the corpus on which the language model is trained, the better the estimate that the language model can provide for a new sentence. However, a corpus can never be large enough to cover every conceivable n-gram in a language, especially for larger n. Therefore, smoothing and back-off techniques are applied to assign probability mass to unseen events. The SriLM toolkit (Stolcke, 2002) provides an implementation for language modeling commonly used in statistical machine translation systems. KenLM (Heafield et al., 2013) is another language modeling toolkit developed more recently for fast estimation of language models based on streaming algorithms.

Definition The fundamental definition of the machine translation task is already used in the word-based models. Equation 2.2 presents how the most probable translation $\hat{e}$ of a source sentence f is defined using the Bayes' rule, where $p(f|e)$ represents the translation probability of f into e, modeled by the translation model and $p(e)$ the probability of e being a good target language sentence, as modeled by the language model. The denominator $p(f)$ can be neglected, since it stays constant for each source sentence.

$$\begin{aligned} \hat{e} = argmax_e p(e|f) &= argmax_e \frac{p(f|e)p(e)}{p(f)} \\ &= argmax_e p(f|e)p(e) \end{aligned} \tag{2.2}$$

Log-Linear Model In order to allow for additional models to be included, the phrase-based statistical machine translation model is formulated as a log-linear model. Equation 2.3 gives the definition for a log-linear model, where n models h_i are combined with individual weights λ_i and the random variable x represents (e, f, start, end) for translation.

$$p(x) = exp \sum_{i=1}^{n} \lambda_i h_i(x) \tag{2.3}$$

The standard models used in a phrase-based system are translation model, language model, distance-based reordering model, word penalty and phrase penalty. Their weights are optimized on a set of development data, which should be similar to the actual test data for translation. Reference translations are necessary for the development data, so that the weights can be set through iterative search and comparison with the resulting translation and the reference. The model weights set during optimization, e.g. by minimum error rate training (MERT) (Och, 2003), are applied for the actual translation of the test data set. The idea being that for each translation task the importance of the individual models varies and their particular influence should be adapted to the task at hand.

Decoding The translation system obtains the translation for a given source sentence by performing a search for the best translation e that maximizes the objective function of one of the mathematical formulae presented above defining the machine translation problem. In order to find the best translation, the decoder searches the space of all possible translation hypotheses that are incrementally built during the translation process. This task has a complexity which is exponential in the length of the input sentence, so that an exhaustive search is computationally too expensive. Heuristic search methods are applied, such as beam search, in order to limit search errors and focus on the most promising translation hypotheses while introducing reordering restrictions. The search space for finding the best translation is built incrementally by expanding it with possible translation hypotheses from the phrase table and controlling its size by hypothesis recombination and pruning. In order to compare intermediate translations, the future costs for translating the rest of the sentence can be included. Finally, the translation hypothesis with the highest score is chosen as the best translation.

N-best List Re-ranking As an alternative to obtaining only the best translation for each input sentence, a list of the N best translation hypotheses with the scores computed during decoding can be produced as output by the translation system. Depending on the intended use or type of quality judgement, this N-best list may contain a better translation than the one that the decoder assigned the highest score to. Hence, re-assessing the hypotheses in the N-best list can lead to an improvement of the translation. This may be beneficial when a new model should be applied outside the log-linear

combination, assigning scores based on different aspects of translation quality or when combining the outputs of multiple machine translation systems.

2.1.2.3 Hierarchical Machine Translation

The hierarchical machine translation approach (Chiang, 2005) models the translation by assigning a sentence a hierarchical structure. It is operating with hierarchical phrases, which are phrases that contain phrases, generating a hierarchical structure of the sentence. The core phrases are extracted from the word alignment in the same way as in phrase-based machine translation. Additionally, the hierarchical phrases are extracted as larger phrases where alignment blocks qualifying as extractable phrases themselves are replaced by the non-terminal symbol X. Hence, the hierarchical phrases consist of both terminal symbols (words) and non-terminal symbols X (placeholder for phrases), which are indexed to avoid confusion if more than one non-terminal occurs. Both hierarchical as well as the traditional phrase translation pairs form the rules in the synchronous grammar. The rules are associated with scores, indicating the reliability of the rule. Typical scoring functions are based on the joint rule probability, rule application probability, direct and reverse translation probability and lexical translation probability. The decoding is framed as a parsing task, where source and/or target side hierarchical trees are constructed by applying the synchronous grammar rules. The translation then can be read off the leaves of the target side tree.

2.1.2.4 Syntax-based Machine Translation

There are also statistical approaches to machine translation that try to model the translation process based on actual linguistic information. In syntax-based machine translation (Yamada and Knight, 2001), a synchronous grammar is used to model the translation process in a similar fashion as in the hierarchical machine translation approach described above. However, actual syntactic parse trees are used to learn the synchronous grammar so that the hierarchical rules consist of terminal symbols (words) and various non-terminal symbols. In the syntax-based approaches the non-terminals are the syntactic categories in the parse tree (such as NP, i.e. noun phrase, VP, i.e. verb phrase, ...) instead of one single non-terminal symbol X in the hierarchical approach, which is not linguistically motivated. As for hierarchical machine translation, decoding is performed as parsing. A common approach is chart-parsing where the chart is used as

the organizing data structure. During parsing the chart is filled with chart entries covering continuous spans of the input sentence. The rules from the synchronous grammar are applied to build the sentence structure of source and target sentence simultaneously. Similar to decoding in phrase-based machine translation, recombination and pruning methods are applied to limit the search space. In order to allow the integration of a language model which further increases parsing complexity, cube pruning provides an algorithm for efficient computation while reducing the search error.

2.1.3 Evaluation of Machine Translation

The evaluation of the quality of machine translation output is a research task of its own, since there is not one correct translation for every sentence, but natural language provides many different ways of conveying the same meaning. Hence, countless methods have been proposed to evaluate machine translation quality, which can be divided into manual and automatic evaluation methods.

2.1.3.1 Manual Evaluation

Manual evaluation of machine translation quality is performed by human evaluators familiar with both source and target language, or at least the target language. They are not necessarily translators, interpreters or language experts, but rather non-specialists judging translation quality based on their own knowledge of the language. Depending on the setting for evaluation, the source text and/or a reference translation is available for assessing the translation quality. Often, several machine translation outputs are evaluated in comparison. A standard procedure for manual evaluation is to provide adequacy and fluency scales from 1 to 5 in order to judge translation adequacy and fluency in the target language separately (Callison-Burch et al., 2007). Another option is to indicate acceptability of particular sentence fragments (Callison-Burch et al., 2008) or measuring the human post-editing effort necessary to improve the machine translation output (Callison-Burch et al., 2009, 2010). In recent evaluation campaigns the human evaluation of choice was to apply a quality ranking of different machine translation outputs (Bojar et al., 2014a, 2013a). Alternatively, a task-specific evaluation can be chosen, e. g. an error classification (Vilar et al., 2006).

Even though a manual evaluation reflects the quality of the translation output best, especially with regard to the acceptability for a human user, in most situations a

manual evaluation is too expensive and time consuming. Typically it is applied for a very specific purpose focusing on particular issues and with a limitation of the amount of text which is evaluated. However, for the standard development cycle of a machine translation system, repeated evaluations are needed to guide the development process. Hence, automatic metrics for machine translation evaluation have become an accepted measure for assessing translation quality which is fast and generates comparable and reproducible results.

2.1.3.2 Automatic Evaluation

Automatic evaluation techniques for machine translation require the availability of a reference translation against which the machine translation output is compared. The most popular metric is the bilingual evaluation understudy (BLEU) by Papineni et al. (2002), which is used in many evaluations. It is next to the translation edit rate (TER) by Snover et al. (2006) the standard metric presented in publications describing research in machine translation. Even though the approaches of comparing n-gram overlaps in translation output and reference translation (i.e. BLEU) and counting insertions, deletions and shifts of words between translation and reference (i.e. TER) can be regarded as very simple, granted that they operate on the word level only. Many new, more complex metrics have been proposed postulating improved correlation with human judgements of translation quality, as presented in the metrics for machine translation task (Macháček and Bojar, 2013, 2014). It is carried out as a regular task of the workshop for machine translation (WMT) (Bojar et al., 2013b, 2014b; Callison-Burch et al., 2012b). However, none of the new metrics has replaced neither BLEU nor TER so far, due to their easy and straightforward application to all languages without the need for additional resources other than a reference translation. Additionally, in contrast to many other complex metrics, there is no need to adapt or optimize BLEU to the specific translation task. Hence, BLEU will also be the main evaluation metric used in this work, in alternation with both manual and automatic evaluations particularly directed at the investigated issues at hand.

2.2 Linguistic Concepts

This thesis focuses on the translation quality with regard to several linguistic phenomena. In the following we will give a short introduction into the linguistic concepts and terminology relevant for the rest of this work.

2.2.1 Words and Morphology

Words are the basic units of language. Even though in most western languages individual words are separated by blank spaces, this is not the case for all languages. The process of separating words and punctuation marks in order to provide individual tokens is called **tokenization**. Each word belongs to a grammatical category, defining the role it plays in the sentence. Nouns refer to objects, verbs to activities that objects or people can do, adjectives are modifiers for nouns, conjunctions are connectors for other words, word groups or sentences and so on. The grammatical category of a word is also called its **part-of-speech** (POS). A word can have multiple meanings which all have the same part-of-speech (*bank*, n. financial institution vs. *bank*, n. side of a river), but many words are also ambiguous regarding their part-of-speech (*can*, v. as in *I can cook* vs. *can*, n. as in *a can of tuna*). Disambiguation might be only possible when looking at the word in context. **Morphology** concerns the smallest units of language that carry meaning, called morphemes. A word consists of one or more morphemes that determine the meaning of the word. Morphemes can be divided into two groups: lexical morphemes and functional morphemes, where lexical morphemes form actual words as found in the dictionary. They can stand alone while functional morphemes need to combine with a lexical morpheme to form a word. Functional morphemes carry grammatical meaning and operate as indicators of grammatical properties, such as tense, count and person. They are realized as suffixes attachable to verbs in order to modify the respective grammatical property of the verb. For example, attaching the functional morpheme *-ed* in the English language to an English verb, e.g. the lexical morpheme *walk*, the resulting word is *walked*, the verb in past tense. Similarly, the functional morpheme *-s* changes the property of the verb into third person singular (*[he] walks*). In order to form a valid connection of subject and verb in a sentence, the verb needs to be finite, i.e. in a conjugated form by having a functional morpheme attached which sets the person, number and tense features. Those features need to match with those

of the noun or pronoun representing the subject. Such dependencies between words in a sentence that require congruency of morphological features is often referred to as **morphological agreement**. The most common forms of morphological agreement occur between subject and verb as described above and within noun phrases, which is discussed in the following.

The suffix *-s* mentioned above can also function as a count modifier for nouns, changing a singular noun into plural (*house* vs. *houses*). While English has limited morphological variation of words, other languages can express more varied grammatical properties through morphological variation. They are called **morphologically rich languages**. German for example falls into that category. In German there are three genders (masculine, feminine and neuter) and four cases (nominative, genitive, dative, accusative) that are realized by the respective functional morphemes attachable to German nouns. Nouns and pronouns have gender properties that are inherent in the word itself. The gender for attributive adjectives is adjusted when associated with a noun by attaching the respective functional morpheme as a suffix. Predicative adjectives are always used in the base form. Similarly, determiners are also declined for case and gender in accordance with the noun they belong to. Example 2.1 illustrates this process with noun phrases consisting of definite article, adjective and noun.

The analysis of morphological properties of words in a sentence or text as well as the generation of a target language sentence using a morphology component is standard in rule-based machine translation systems. However, most statistical machine translation systems do not apply particular handling of morphology in the standard configuration. However, explicit modeling of morphology could guide the translation process by uncovering dependencies between source words to resolve ambiguities, for example with regard to part-of-speech or grammatical properties. In addition, generating correct morphology in the target language is important in order to produce a grammatically correct target sentence, especially in morphologically rich languages.

2.2.2 Sentence Structure

Each language has their rules stating how a valid sentence can be constructed. In linguistics these rules on the compositionality of words into sentences is referred to as the **syntax** of a language. In English and German, for example, a sentence contains as main components a subject, a verb and zero or more objects. The number of objects

Adjective use	gender	case	Example
Predicative	masc.	nom.	*Der Schornstein ist blau. (The chimney is blue.)*
	fem.		*Die Tür ist blau. (The door is blue.)*
	neutr.		*Das Haus ist blau. (The house is blue.)*
Attributive	masc.	nom.	*der blau-e Schornstein (the blue chimney)*
	fem.		*die blau-e Tür (the blue door)*
	neuter		*das blau-e Haus (the blue house)*
	masc.	gen.	*des blau-en Schornstein-s (of the blue chimney)*
	fem.		*der blau-en Tür (of the blue door)*
	neuter		*des blau-en Haus-es (of the blue house)*
	masc.	dat.	*dem blau-en Schornstein (the blue chimney [indir. obj.])*
	fem.		*der blau-en Tür (the blue door [indir. obj.])*
	neuter		*dem blau-en Haus (the blue house [indir. obj.])*
	masc.	acc.	*den blau-en Schornstein (the blue chimney [dir. obj.])*
	fem.		*die blau-e Tür (the blue door [dir. obj.])*
	neuter		*das blau-e Haus (the blue house [dir. obj.])*

Example 2.1: Noun phrase agreement

is determined by the valency of the verb which needs to be complied with in order to obtain a valid sentence. The order in which subject, verb and object(s) are allowed to occur is fixed for a given language. In German and English the main word order type is subject – verb – object(s), also referred to as **SVO** order. Languages obeying this order are also called SVO languages. While this order is fixed for almost all sentences in English, in German this order only applies for main clauses. In German subordinate clauses, the SOV order is applied instead. Differences in word order is one of the main problems for statistical translation. Hence, developing models for word reordering in statistical machine translation is even a research direction of its own. Especially when languages belong to different word order groups, i.e. when translating from an SVO language into an SOV language, such as Japanese. In addition to simple word order categories based in the order of subject, verb and object(s), there are formal grammars describing and modeling the structure of sentences more in detail. For example, the **phrase structure grammar** is a grammar formalism which consists of a lexicon and a grammar component. The lexicon contains mappings of words to their parts-of-

speech and the grammar consists of rules on how to construct a sentence based on constituency and precedence principles. Applying the grammar rules to a sentence in order to generate a phrase structure tree (or **syntactic tree**) representing its structure is called **parsing**. During the parsing process, a constituency tree is built according to the rules in the grammar. Constituents are formed directly from the part-of-speech sequences of the words in the sentence or from higher order constituents until a full sentence (S) is constructed. Figure 2.2 shows a lexicon and grammar rules for a sample phrase structure grammar and the phrase structure tree generated by applying the grammar rules.

Lexicon:

Maria	NNP
Peter	NNP
gives	VBZ
the	DT
book	NN
to	TO

Grammar:

S	→ NP VP
NP	→ NNP
NP	→ DT NN
NP	→ NP PP
VP	→ VBZ NP
PP	→ TO NP

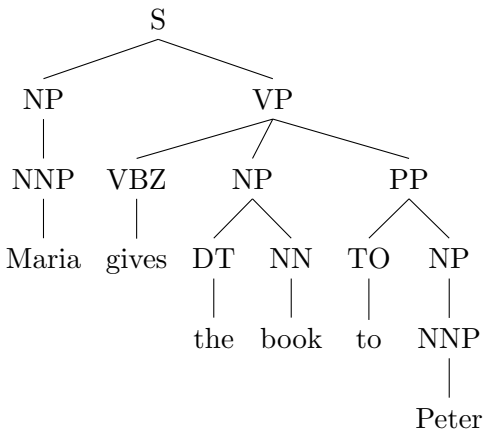

Figure 2.2: Example phrase structure grammar and tree

An alternative representation of sentence structure can be provided by a **dependency grammar**. In contrast to the phrase structure grammar which is guided by constituency, the dependency grammar is more motivated by the semantic relations between individual words. The verb represents the main semantic content of the sentence and forms the sentence's root. All other words are depending either directly on the root verb or on another word in the sentence which they form a close relation with. Figure 2.3 shows the dependency tree for the same sentence shown in Figure 2.2 in order to contrast the two different representations of sentence structure.

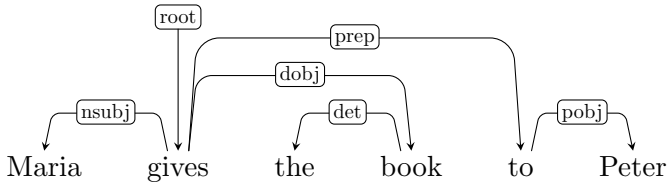

Figure 2.3: Example dependency tree

In this thesis we use both syntax trees based on phrase structure grammar as well as dependency trees for modeling sentence structure to improve the treatment of linguistic phenomena in statistical machine translation.

2.2.3 Discourse Phenomena

The previous sections have introduced linguistic concepts on the word and sentence level. Some linguistic phenomena may also go beyond the limits of sentence boundaries. An example for such phenomena are anaphora. **Anaphora** are words that refer to other words in the same or in a previous sentence. **Pronouns** are a particularly common type of anaphora which refer to a previously mentioned entity, called **antecedent**. The antecedent of a pronoun can be a person, thing or facts that were introduced earlier in the text or discourse. The connection between the pronoun and the antecedent is made obvious through congruence of person, number and gender features. That means a pronoun referring to a male person needs to have the following grammatical features: third person, singular and masculine. This grammatical agreement or congruency allows to disambiguate the connection. However, the mention of the antecedent could have happened several sentences ago, or it might even happen that several antecedents qualify according to their grammatical features. Humans have few problems keeping track of past mentions of possible antecedents and also use semantics and world knowledge for easily disambiguating between antecedent options. However, resolving **pronominal anaphora** automatically is a research field of its own in computational linguistics. During translation, an additional difficulty arises. A pronoun cannot be simply translated in isolation, but a pronoun needs to be chosen on the target side that exhibits the correct grammatical features coinciding with the translation of the antecedent. In this thesis we will introduce a method to treat pronominal anaphora in translation by including knowledge about context and sentence structure.

3

Linguistic Challenges

3.1 Reordering

When translating from German to English different word order is the most prominent problem. Especially the verb needs to be shifted over long distances in the sentence, since the position of the verb differs in German and English sentences. The finite verbs in the English language are generally located at the second position in the sentence. In German this is only the case in a main clause. In German subordinate clauses the verb is at the final position as shown in Example 3.1.

Source:	*..., nachdem ich eine Weile im Internet gesucht habe.*
Gloss:	`... after I a while in-the internet searched have.`
Translation:	*... as I have for some time on the Internet.*
Reference:	*... after browsing the web for a while.*

Example 3.1: Missing verbs in translation output

The example shows first the source sentence and an English gloss. The translation is produced by a phrase-based machine translation system applying a dedicated reordering model based on word categories, i. e. parts-of-speech. We can see that the translation is already partially correct. The auxiliary *habe/have* is shifted to the right position in the sentence. But the participle, which carries the main meaning of the sentence, gets lost during translation, rendering it unintelligible. The included reordering model does not realize that both the auxiliary and past participle are part of the verb and both

need to be shifted and translated as a whole.

The syntactic structure of the source sentence could help with this problem. To know which words form constituents in a syntactic parse tree would be useful information for a reordering model, so that these words would be shifted as a whole block. Abstracting from the word level to the constituent level also provides the advantage that even though reorderings are performed over long sentence spans, the rules consist of less reordering units (constituents of constituents or words) and can be learned more reliably.

Another challenging task during word reordering are verb prefixes which may be separated from the verb stem and placed at a distant position in the sentence. Example 3.2 shows a translation where the verb stem is translated in isolation, while ignoring the prefix. Since that verb stem also exists as a main verb, the translation of that verb on its own is technically correct. Furthermore, the prefix is dropped from the translation, so that no evidence of a mistranslation is left in the translated sentence.

Source:	*Die RPG Byty schlägt ihnen in den Schreiben eine Mieterhöhung von ca. 15 bis 38 Prozent vor.*
Gloss:	`The RPG Byty proposes-VFIN them in the letters a rent increase of ca. 15 to 38 percent proposes-PTKVZ`
Translation:	*The RPG Byty beats them in the letter, a rental increase of around 15 to 38 percent.*
Reference:	*RPG Byty proposes to increase rent by 15 to 38 percent in these letters.*

Example 3.2: Failed identification of verb prefix leads to wrong translations

If the connection between the verb and corresponding prefix were identified and both of them were moved together, the translation system had a better chance of generating a correct translation. Here, too, the sentence structure could serve as a basis for modeling the reordering better.

In order to address this problem, a reordering model based on syntactic parse trees is developed in this thesis. The syntactic trees encode information about sentence structure and the relationship between the words and constituents in the sentence. With this information, better reordering rules can be learned and the translation of verbs can be improved.

3.2 Translation of Pronominal Anaphora

Anaphora are another linguistic phenomenon that can lead to difficulties during translation. Although a quite common writing practice in order to avoid repetition of the same words, it is bound to produce ambiguity which is not easy to resolve by automatic means. Especially in the case of pronominal anaphora, where a pronoun is used to refer to a previously mentioned entity, called antecedent. Grammatical features, such as gender and number of antecedent and pronominal anaphora need to be congruent, but sometimes additional semantic or world knowledge is necessary to identify the anaphoric relationship, a process also called anaphora resolution.

During translation, the antecedent-anaphora relationship needs to be maintained grammatical. That means the pronoun cannot be translated in isolation, but in accordance with the translation of the antecedent. If the grammatical gender of the antecedent changes during translation, the grammatical gender of the translated pronoun needs to change accordingly. Especially when dealing with languages where gender in nouns is distributed dissimilarly, problems will arise. In English, masculine and feminine grammatical gender occurs only when referring to persons. Things are always neuter, and are referred to by the neuter pronoun *it*. In German, however, things may possess any of the three possible genders: masculine, feminine and neuter. As a consequence, the translation of an English neuter pronoun may result in a pronoun of either one of the three German genders, depending on the chosen German translation of the antecedent.

Example 3.3 shows a sentence, where the source pronoun *it* is translated as *sie* by the translation system. Since there is no antecedent in that sentence, it is only possible to know what *it* refers to in this context by the other words in the sentence, in this case the verb. Speaking about *sailing*, one can infer that it must be a boat or ship that is being referred to. The German translations would be neuter in both cases (*Boot*, n.; *Schiff*, n.), so that the pronoun to be chosen in this sentence should also be neuter and singular. The machine translation fails to make this admittedly very implicit connection and also chooses the wrong case.

Another problem in pronoun translation into German is shown in Example 3.4. Since German possessive pronouns behave similarly to determiners when they are used in an attributive way, they are subjected to declension and need to agree with the

Source:	*And I went sailing on <u>it</u> , and we did surveys throughout the southern South China sea and especially the Java Sea.*
Translation:	*Und ich ging auf <u>sie</u> segeln , und wir haben Umfragen in den südlichen Südchinesische Meer und vor allem die Java-See.*
Reference:	*Ich fuhr <u>darauf</u> mit und wir machten Erhebungen im ganzen südlichen Südchinesischen Meer und besonders in der Javasee.*

Example 3.3: Erroneous gender for pronoun

case and gender of their governing noun. In Example 3.4, the possessive pronoun *my* belongs to the governing noun *class* and in the German translation, those two words need to agree in case and gender. In this translation, the person and number of the pronoun in the baseline translation is correct (*mein-*), as well as the case (dative), but the case ending of *meinem* is masculine or neuter. However, to ensure agreement with the translation of the governing noun (*Klasse*), which is feminine, the feminine dative ending needs to be chosen (*meiner*). The translation system is not able to generate this correctly.

Source:	*I memorized in <u>my</u> anatomy class the origins and exertions of every muscle [...]*
Translation:	*Ich in <u>meinem</u> Anatomie der Klasse die Ursprünge und Strapazen eines jeden Muskel [...] auswendig [...]*
Reference:	*In <u>meiner</u> Anatomievorlesung lernte ich die Ursprünge und Ausläufer jedes Muskels [...]*

Example 3.4: Erroneous gender ending for pronoun

Example 3.5 presents the translation of another ambiguous pronoun. The English *that* is translated into the German conjunction *dass*. Although this could be correct in other instances, in this case *that* should have been translated into *die*, a relative pronoun referring to *Möglichkeiten*.

In order to facilitate the correct translation for pronouns, it is necessary to take more context or the structure of the sentence into account to identify the antecedent. It might even be necessary to go beyond sentence boundaries to find the noun being

Source:	*Somehow by ways <u>that</u> we don't quite understand, [...]*
Translation:	*Irgendwie durch Möglichkeiten, <u>dass</u> wir nicht ganz verstehen, [...]*
Reference:	*Auf irgend eine Art, <u>welche</u> wir noch nicht ganz verstehen, [...]*

Example 3.5: Failed disambiguation of ambiguous pronoun leads to wrong translation

referred to. In this thesis we explore the improval of pronoun translation by developing a model for translation prediction with context and dependency features.

3.3 Generating Morphological Agreement in the Target Language

When translating from a language with less morphological expressiveness to a morphologically rich language, the generation of morphological agreement of case, person, number and gender features in the target language poses a challenge for statistical machine translation. In English there are no case markers, but the role of a noun phrase in the sentence is solely determined by its position and/or by combination with a preposition. Hence, generating the correct case in German is difficult during English-to-German translation. Furthermore, German nouns belong to one of three genders: masculine, feminine, or neuter. This is an issue when a determiner or adjective is combined with the noun into a noun phrase. Then determiner and/or adjective and noun have to agree in case, number and gender, which means the appropriate word form needs to be chosen for each of them. Another level of complexity is added through the distinction into definite and indefinite articles as well as weak and strong adjectives, which follow different declension schemes. In English, there is no declension in the three mentioned word categories, except for pluralization, which is hence the only necessary morphological feature that requires agreement.

Example 3.6 shows an example where the English determiner *this* is translated into the correct German base word (*dies-*), but in the accusative instead of dative case form.

Source:	*[...] we can now write things in <u>this</u> code.*
Translation:	*[...], können wir jetzt die Dinge in <u>diesen</u> Code schreiben.*
Reference:	*dass wir, [...], selber Sachen in <u>diesem</u> Code schreiben können.*

Example 3.6: Failed case agreement between determiner and noun

A second type of agreement that needs to hold both in English and in German alike is the agreement between subject and verb. The difficulty here is not the underspecification in one of the involved languages. Both of them require that the subject and verb of a sentence agree in the person and number features. However, subject and verb might be separated by other words in the sentence and additionally necessary reordering could obfuscate the connection during translation. Or, as shown in the translation in Example 3.7, some English verbs have the same word form in singular and in plural, which can lead to wrong translations.

Source:	*There I think that the arts and film can perhaps fill the gap, and simulation.*
Translation:	*Ich glaube, dass die Kunst und Film kann vielleicht die Lücke füllen, und Simulation .*
Reference:	*Hier können, denke ich, die Kunst und der Film vielleicht die Lücke füllen, sowie Simulationen.*

Example 3.7: Failed case agreement between subject and verb

The structural features in the above mentioned model for translation prediction will also be of use for the modeling of agreement in the target language. The dependency features explicitly uncover the connection between subject and verb and can thus help to produce a translation with better subject-verb agreement. Similarly is the dependency relation between noun and modifiers, such as adjectives, determiners or possessive pronouns expected to support the modeling of noun-phrase-internal agreement.

4

Related Work

4.1 Word Reordering in Statistical Machine Translation

Word reordering has been addressed by many approaches in statistical machine translation systems. Already in the early days Wang and Waibel (1998) identified the problem in the word alignment models introduced by Brown et al. (1993) and suggested a structure-based alignment model to produce better alignments and therefore better translation of languages with different word order, like German and English.

In a state-of-the-art phrase-based machine translation system, the decoder processes the source sentence left to right, but allows changes in the order of source words while the translation hypothesis is generated. The window size for allowing changes in word order during translation can be set in the decoder according to the requirements of the language pair of translation. Many phrase-based systems, e.g. the open source machine translation system Moses (Koehn et al., 2007b) also include a lexicalized reordering model (Koehn et al., 2005; Tillmann, 2004) which provides additional reordering information for phrase pairs. It stores statistics on the orientation of adjacent phrase pairs on the lexical level. This reordering method affects the scoring of translation hypotheses but does not generate new reorderings. Another type of lexicalized reordering method is presented in Xiong et al. (2006).

4.1.1 Preordering Approaches

A very popular approach is to detach the reordering from the decoding procedure and to perform the reordering on the source sentence before translation. Such preordering

approaches use linguistic information about the source and/or target language, such as parts-of-speech, dependency or constituency tree structures. They either apply handcrafted rules or automatically learn rules that change the order of the source sentence. Then monotone translation is performed.

In the first preordering approach, reordering rules for English-French translation are automatically learned from source and target language dependency trees (Xia and McCord, 2004). Since then many others adopted this method. In the beginning manually crafted reordering rules based on syntactic or dependency parse trees or part-of-speech tags were designed for particular languages (Collins et al., 2005; Habash, 2007; Popović and Ney, 2006; Wang et al., 2007). Later data-driven methods followed, learning reordering rules automatically based on part-of-speech tags (Niehues and Kolss, 2009; Rottmann and Vogel, 2007) or syntactic chunks or sequences (Crego and Habash, 2008; Elming, 2008; Zhang et al., 2007). Alternatively, word class information may be used to perform a translation of the original source sentence into a reordered source sentence (Costa-Jussà and Fonollosa, 2006). More recent work includes reordering rules learned from source and target side syntax trees (Khalilov et al., 2009), automatically learned reordering rules from IBM 1 alignments and source side dependency trees (Genzel, 2010) and using a classifier to predict source-sentence reordering (Lerner and Petrov, 2013). Du and Way (2010) perform classification of a particular construction in Chinese and learn corresponding reordering rules. In DeNero and Uszkoreit (2011) no parser is needed, but the sentence structure used for learning the reordering model is induced automatically from a parallel corpus. While some of the presented approaches perform a deterministic reordering of the source sentence, others store reordering variants in a word lattice leaving the selection of the reordering path to the decoder. Katz-Brown et al. (2011) address the problem from a completely different angle and train designated syntactic and dependency parsers to better concur with the word reordering task in a statistical machine translation system.

4.1.2 Syntax-based and Hierarchical Machine Translation

In contrast to phrase-based machine translation where no linguistic structure is taken into account and phrases are determined through co-occurrence and word alignment completely without linguistic motivation, there is another group of statistical approaches to machine translation. Syntax-based (Yamada and Knight, 2001) or syntax-augmented

(Zollmann and Venugopal, 2006) machine translation systems address the reordering problem by embedding syntactic analysis in the decoding process. They use syntactic trees of the source or target language and learn a synchronous grammar. Then they perform decoding as parsing. Hierarchical machine translation systems (Chiang, 2005) also use a synchronous grammar. However, instead of deriving the sentence structure from actual syntactic parses, a syntactic hierarchy is constructed, which is independent of linguistic categories. Nguyen and Vogel (2013) propose an extension to hierarchical machine translation by including reordering features from the phrase-based approach, namely the lexicalized reordering model and a distance cost. Galley and Manning (2008) present a hierarchical reordering model which extends the lexicalized reordering model (Tillmann, 2004) to hierarchical phrases and can be integrated into a phrase-based machine translation system.

Structural information such as syntactic or dependency parse trees can also be exploited in other ways in order to improve the word reordering problem: In Shen et al. (2004) and Och et al. (2004) syntactic information is used for re-ranking decoder output. Bach et al. (2009) use a reordering model based on dependency subtree movements and lexicalized reordering features. They apply it both during optimization and at decoding time.

4.1.3 Evaluating Word Order in Machine Translation

Related work regarding reordering metrics and reordering quality includes the first description of reorderings as permutations (Eisner and Tromble, 2006). Later, the use of permutation distance metrics to measure reordering quality (Birch et al., 2010) leveraged research into distance functions for ordered encodings. An approach to transform alignments into permutations (Birch, 2011) takes the particular characteristics of alignment functions into account. Another way of measuring reordering quality is using Kendall's τ distance (Kendall and Gibbons, 1990) in order to determine the distance between two reordering variants, e.g. the proposed reordering and a reference reordering. This metric only focuses on the order, so that is is completely independent of the actual translated words. The RIBES (Rankbased Intuitive Bilingual Evaluation Score) metric (Isozaki et al., 2010) combines Kendall's τ distance with precision and brevity penalty for jointly measuring reordering and translation quality. It is used as an alternative for BLEU (Papineni et al., 2002) for languages with distant word orders,

such as Japanese and English. Talbot et al. (2011) present a fuzzy reordering score which is based on the difference of the system's proposed reordering and a reference reordering in terms of jumps over chunks of words that are in the same order. Within their proposed framework, the reordering score provides a method to evaluate reordering quality for preordering approaches before deciding on a particular word order for translation.

4.1.4 Oracle Reordering

Oracle experiments have shown to be a valuable method for analyzing different aspects of machine translation. While an oracle BLEU score may serve for identifying translation errors in the phrase table (Wisniewski et al., 2010), another approach uses oracles for punctuation and segmentation prediction in speech translation (Cho et al., 2012). Efficient methods for finding the best translation hypothesis in a decoding lattice have been proposed (Sokolov et al., 2012). Furthermore, research on oracles regarding the reordering problem have been conducted. Dreyer et al. (2007) use linear programming to compare the best achievable BLEU scores when using different reordering constraints. Khalilov and Sima'an (2011) present a reordering method for translations from English to Spanish, Dutch and Chinese where deterministic reordering decisions are conditioned on source tree features and compared to several oracles.

4.1.5 Analysis of Reordering

Since the development cycles for machine translation systems are becoming shorter, automatic metrics are a popular method for measuring the quality of machine translation systems or their included models quickly and in a reproducible fashion (Lavie and Denkowski, 2009; Papineni et al., 2002; Snover et al., 2006). Since typical metrics for translation quality do not correlate well with reordering quality, explicitly measuring the reordering quality can provide insights on just this aspect (Birch, 2011). However, human judgment stays an important factor and is applied as an additional or even main decision criterion for translation quality in evaluation campaigns for machine translation systems (Bojar et al., 2013a; Federico et al., 2012). A classification scheme for human error analysis of machine translation is presented in Vilar et al. (2006). This scheme is also applied in a tool for performing manual error analysis for machine translation (Stymne, 2011), which allows choosing between error classification methods and

adding customized error classes. An extensive error analysis of different machine translation systems translating from English and Spanish to Catalán distinguishes linguistic error classes such as orthographic, lexical, morphological, semantic and syntactic errors (Farrús et al., 2012).

A framework towards an automatic error analysis directed in particular at different types of linguistic errors in machine translation is proposed in Popović and Ney (2011), also presenting a human error analysis as a reference for their automatic system. Another framework for semi-automatic error analysis makes use of manual and automatic annotations regarding several characteristics of input documents and connects them with system performance in order to identify features indicating system deficiencies (Kirchhoff et al., 2007).

4.2 Translation Disambiguation

The disambiguation of word senses as an individual task is closely related to the kind of disambiguation that has to be done when choosing a particular translation for an ambiguous word. Already Brown et al. (1991) have proposed an approach that performs statistical word sense disambiguation (WSD) by defining senses according to the different translations of a word. In their algorithm they exploit the word alignments and the context of a word to define a sense and improve translation when incorporating the sense disambiguation into the translation system.

Carpuat and Wu (2005) claim that contrary to common conception statistical machine translation is not good at performing word sense disambiguation and can benefit from an explicit modeling or integration of a word sense disambiguation component. Vickrey et al. (2005) cast word sense disambiguation as a word translation task for French-English translation using context features. Carpuat and Wu (2007) propose an integration of a word sense disambiguation approach in a phrase-based SMT system to perform multi-word lexical disambiguation for translation from Chinese to English. Chan et al. (2007) successfully integrate a word sense disambiguation component into a hierarchical phrase-based translation system. In addition to using the source context for disambiguation, Max et al. (2008) also use grammatical dependencies for a context-aware translation from English into French, a morphologically richer language.

Gimpel and Smith (2008) use context features including words, parts-of-speech and local syntactic structure for the prediction of phrase translations in Chinese-English and English-German translation with a phrase-based machine translation system. Specia et al. (2008) integrate word sense disambiguation and statistical machine translation with an N-best list re-ranking approach.

Apart from applying actual word sense disambiguation in machine translation, linguistic information, such as context words, dependencies or syntax can be integrated in machine translation as additional features in order to improve the translation quality, e.g. as done by Shen et al. (2009) and Haque et al. (2011).

Among the approaches that particularly model translation prediction as is done in this thesis, Mauser et al. (2009) predict the occurrence of a target word in a translated sentence given the source words using a discriminative approach. A similar approach is presented by Patry and Langlais (2009) using a multilayer perceptron. Tran et al. (2014) use a bilingual neural network to learn abstract word representations and features in order to predict word, stem and suffix translations for source words given the source context. Tamchyna et al. (2014) present a framework for training discriminative models on source context features and including the classifier predictions in the decoding process of phrase-based and hierarchical machine translation in Moses.

The representation can play an important role in prediction or classification tasks where many features are used. Gallant (1991) use a neural network to learn a context vector representation to be used for word sense disambiguation. They suggest that this type of representation is suitable to be used in various natural language processing (NLP) tasks, such as machine translation. The word2vec algorithm (Mikolov et al., 2013) became a quite popular way to learn word vector representations for natural language processing. Martinez Garcia et al. (2014) apply a semantic model built with the CBOW approach (Mikolov et al., 2013) to predict semantically related words in a bilingual setup and also integrate the semantic model in a statistical machine translation to translate ambiguous words.

4.2.1 Pronoun Resolution and Translation

Research on resolving co-referring expressions such as anaphora automatically as a stand-alone task is widely covered. Among the early rule-based approaches to co-reference resolution, Hobbs (1986) and Lappin and Leass (1994) are still used in current

work on pronoun translation (Le Nagard and Koehn, 2010). Qiu et al. (2004) present a reimplementation of Lappin and Leass (1994), providing a tool for benchmarking or to use in other natural language processing tasks. More recently, statistical approaches to co-reference and pronoun resolution have been presented, among them BART (Versley et al., 2008), based on the original algorithm introduced in Soon et al. (2001), as well as Stanford's co-reference system (Lee et al., 2011).

Research at the frontier of anaphora or co-reference resolution and machine translation includes approaches to multilingual resolution of co-reference (Harabagiu and Maiorano, 2000) and pronouns (Mitkov and Barbu, 2002) using parallel corpora. Others focus on the projection of co-references between languages by exploiting methods or resources from machine translation (de Souza and Orasan, 2011; Postolache et al., 2006; Rahman and Ng, 2012).

There is only limited research on modeling anaphora resolution for the translation of pronouns in a statistical machine translation system. Mitkov et al. (1995) were the first to integrate an anaphora resolution component within an MT system. The component is implemented by syntactic and semantic constraints and preferences within their unification-based framework designed for machine translation. Le Nagard and Koehn (2010) investigate the automatic translation of pronouns within a phrase-based statistical machine translation system. In their English-to-French experiments, they identify the antecedent of the English neuter pronouns *it* and *they* using two approaches to anaphora resolution. The pronouns are annotated with the gender of their antecedent's translation. Then a phrase-based machine translation system is trained on the annotated data. Hardmeier and Federico (2010) developed a word dependency model that makes use of anaphora-antecedent pairs obtained from an anaphora resolution tool. An additional model score is added for the probability of the translation of the pronoun given the translation of the antecedent. With this model, they can improve precision and recall of pronoun translation from English-to-German, but the improvements are not visible in BLEU. Guillou (2012) apply the approach by Le Nagard and Koehn (2010) to English-Czech pronoun translation, with similarly limited improvements as their predecessors. Pronoun translation is not only a problem for European language pairs. Taira et al. (2012) present a method to generate pronouns in the English translation in cases where the pronouns are omitted in the Japanese source language. Hardmeier et al. (2013) model the translation of the English third person

pronouns *he, she, it* and *they* into French using a neural network. Using automatic anaphora resolution output they perform a classification of the French translation into one of six classes, five of them being the French pronouns *ce, elle, elles, il, ils* or OTHER. Their neural network approach surpasses maximum entropy classification and can even be extended to perform latent anaphora resolution and translation prediction jointly, thus eliminating the need of an external anaphora resolution tool. After Novák et al. (2013b) perform an extensive analysis on the translation of the English pronoun *it* into Czech, Novák et al. (2013a) present an approach for modeling it within the deep syntax machine translation framework TectoMT (Žabokrtský et al., 2008). They perform classification of the pronoun into three classes triggering different treatment in the transfer and synthesis components of the tree-to-tree-based machine translation system. In Weiner (2014) another classification approach based on a discriminative word lexicon is applied to pronoun translation from English into German. The approach is described more in detail in Appendix A and will be part of the discussion in Chapter 7.

Popescu-Belis et al. (2012) annotated excerpts from the Europarl corpus (Koehn, 2005) with discourse-related annotations in English and French, providing a resource supporting further research on the automatic translation of pronouns and other discourse connectives.

4.2.2 Agreement in Statistical Machine Translation

When translating into a morphologically rich target language, the generation of correct word forms and agreement poses a challenge for statistical machine translation systems. Data sparsity issues are the main reason, since the system technically can only produce what it has seen in training. According to Birch et al. (2008), the success of machine translation depends to a great deal on the morphological complexity of the target language. Hence, alleviating the limitations of statistical systems by modeling target morphology in various ways is a popular direction of research. Minkov and Toutanova (2007) use morphological and syntactic resources for the prediction of inflected word forms in the target language. This prediction is integrated into phrase-based and syntactically informed machine translation systems investigating different integration strategies (Toutanova et al., 2008). Koehn and Hoang (2007) present a factored translation model treating word, lemma, part-of-speech and morphological

features as separate factors and performing morphological generation in a phrase-based machine translation system. This morphological generation model can translate previously unseen word forms. The factored model has been applied in Avramidis and Koehn (2008) for enriching the source language with linguistic information in order to address noun phrase and subject-verb agreement. Another application of the factored model is presented in Razavian and Vogel (2010). Instead of part-of-speech tags they use fixed-length suffixes in order to improve grammaticality of the translation output.

Mel'čuk and Wanner (2008) describe similar problems in a transfer-based machine translation system operating on a deep syntactic level when source and target language exhibit differences in morphological expressiveness. Cartoni (2009) presents a formalism for analyzing and generating neologisms in a transfer-based translation system.

Morphosyntactic processing of German is presented in Fraser (2009), where morphological splitting and stemming is performed for German as source language and a two-step processing is applied for German as target language. Translation is first performed into stemmed word forms from which then inflection is generated. Jeong et al. (2010) apply a discriminative lexicon model based on context, dependency and morphological features in a tree-to-string statistical machine translation system and report improvements on three morphologically rich target languages: Bulgarian, Czech and Korean.

In a string-to-tree machine translation system, Williams and Koehn (2011) model agreement by adding unification-based constraints for enforcing agreement within noun phrases and prepositional phrases as well as between the subject and verb of the sentence. An extension covering more phenomena is presented in Sennrich et al. (2014)

Conditional random fields (CRF) are a popular approach to sequence labeling. Clifton and Sarkar (2011) use them for the prediction of morphemes in post-processing after morpheme-based translation into Finnish. Green and DeNero (2012) propose an agreement model performing sequence scoring of morphosyntactic word classes with grammatical features. They apply CRFs for segmentation, tagging and scoring in their model. Two step translation in a similar fashion as proposed in Toutanova et al. (2008) is applied to English-German (Fraser et al., 2012) and English-French (Weller et al., 2013) translating first into non-inflected forms and using CRFs for predicting fully inflected forms afterwards. Another two-step translation is presented in Mareček et al.

(2011), where two translation systems are applied sequentially, first translation into simplified Czech with feature-enriched lemmas and then a second system monotonically translates simplified into fully inflected Czech. Kholy and Habash (2012) investigate different ways of translation into the morphologically rich target language Arabic. Surface form translation is compared against a two-step approach, first translating into enriched lemma and then generating or predicting fully inflected forms.

Operating on the language model is another common approach to deal with complex target morphology. Müller et al. (2012) combine a standard language model with a class-based language model based on morphological and shape features to reduce perplexity on 21 European languages. Bisazza and Monz (2014) perform a detailed evaluation of class-based approaches, comparing different kinds of classes, language model combination techniques and model forms. The representation of morphology in continuous space language models and its application in machine translation is investigated in Botha and Blunsom (2014).

Morphological preprocessing in the morphologically richer source language may consist of defining equivalence classes (Nießen and Ney, 2004) or simplification of the morphological variation and reduction to stems (Weller et al., 2013).

Another strategy to deal with morphology differences in the languages in translation is to augment the phrase table with synthetic phrases including predicted determiners (Tsvetkov et al., 2013) or morphological re-inflections of the target side of phrase pairs (Chahuneau et al., 2013).

5

Data and System

5.1 Data

In this thesis we examine linguistic phenomena of German and English which pose a challenge for automatic translation. Depending on the type of data, such phenomena can vary and result in an increased or reduced difficulty for the translation process. In order to present exhaustive results, the linguistic analyses and methods for dealing with particular linguistic phenomena presented in this thesis are applied to different text genres and domains. In the following we describe the different types of data that are used in the experiments of this thesis.

5.1.1 Text

We consider the genre of "text" to consist of well-written, grammatically correct text, such as News articles. Most of the available training data for natural language processing tasks can be assigned to this category of data. A characteristic of this kind of data is that it may consist of long sentences with embedded clauses, a rather formal style of writing and mostly describing events or third-party persons.

5.1.2 Speech

The data type "speech" on the other hand consists of spoken language presentations delivered for a particular audience. From a grammar point of view, speeches often differ from written text. Typically, spoken sentences are shorter and less complex, the used words are more common ones and the style is less formal. One of the biggest

problems in spoken language are disfluencies. Except for read or scripted speeches, a typical speech is characterized by the spontaneity of the spoken words. The speaker has a mental picture of what to say, but constructs the sentences on the fly, which results in nonverbal speech artifacts such as hesitations, filler words (*uh, uhm, hmm*), or stuttering. The speaker might abort and restart a sentence, because he changes his mind about how to formulate the sentence, or to correct a grammatical or content error. In other cases, he might even abandon the whole train of thoughts and start a sentence on a new topic without bringing the former one to a close. There is even a new dimension added in the genre of speech, which includes the "me and you" as well as the "here and now". The audience being the addressee of the speech as well as the speaker himself, the location and time at which he speaks may be referred to in the speech. This opens up a new possibility for ambiguities which make the translation even more challenging.

When translating speech with a statistical machine translation system, the typical procedure is to use an automatic speech recognition system and a statistical machine translation system in sequence. First the recorded or live stream of the speech is input to the speech recognition system and its output is then used as input to the machine translation system. Depending on the scenario, the machine translation system might have to deal with the correct speech transcript, which is manually written down by a human and serves as the reference for the speech recognition system or the actual output of an automatic speech recognition system, which possibly includes errors.

5.1.2.1 Manual Speech Transcripts

Using the manual speech transcripts for machine translation has the advantage that the machine translation performance can be measured independently of recognition errors. However, the main characteristics of speech as mentioned before are still present depending on the applied transcription method. More details on different transcription paradigms will follow later on.

5.1.2.2 Automatic Speech Recognition Output

If the actual output of a speech recognition system is used as input to the translation system, this obviously affects the translation performance. Possible recognition errors include omitted words or a word might be confused with another word with a similar

spelling. Homophones are another problem. They sound the same but are written differently, which can be confused by the recognition system. As a consequence, incorrect word boundaries may be determined, potentially leading to a series of wrong words. In addition, speech artifacts could be mistakenly recognized as words and sentence boundaries as predicted by the speech recognition system rarely correspond to full, grammatically correct sentences which the machine translation system is expecting. All of these kinds of recognition errors impede the translation in a way that improvements achieved in the machine translation system will be difficult to transfer to automatic recognition output and might be barely or not identifiable at all in the translation output.

In this thesis we will perform experiments on both text and speech data, but for speech data we choose the form of manual speech transcripts in most cases, in order to allow the measurement of the performance of the developed machine translation methods independently of speech recognition errors.

5.1.3 Domains

As mentioned above, data can be further distinguished by the domain it belongs to. In the following we will describe the three types of data used in the experiments in this thesis.

5.1.3.1 News Texts

Translation of news and news commentary texts is the main task of the Workshop on Machine Translation[1] (WMT). Started in 2006, it is carried out annually and has established a benchmark among its participants that come both from academic and industrial backgrounds. Every year, the organizers publish a new news data set for evaluating the quality of the participants' submitted translations. A typical data set contains several news articles, summing up in total to 2000 to 3000 sentences. The topics are various, ranging from economics to literature. Each data set comes with a human translation, which serves as reference for measuring translation quality of the machine-generated translation hypotheses. In the experiments in this thesis where translation of news data is performed, the translation system is developed using the

[1] e.g. `http://www.statmt.org/wmt15`

Data Set	Evaluation	Data Type	Name	Sentences
		Training	train	2079049
News	WMT 2012	Dev	newstest2010	2489
		Test	newstest2011	3003

Table 5.1: *News data*

data provided for the WMT 2012 evaluation campaign (Callison-Burch et al., 2012b). Table 5.1 shows an overview over the development and test data from the news domain.

5.1.3.2 TED Talks

TED talks are short presentations of up to 18 minutes on various topics held at the TED conference. TED originally stood for Technology, Entertainment and Design, but nowadays the topics are not restricted to any domain. Video, audio and subtitles for each TED talk are published on their website (`http://www.ted.com`). Since the original language of the talks is English, the TED Open Translation Project was created in order to make the content of the talks available to non-English-speaking users. Within this project, translations of the English talk subtitles are generated by volunteering TED users. Translators must follow TED's translation guidelines and their translations are submitted to review and approval by fellow TED translators.

TEDx are local, independently organized events in the spirit of the original TED conference, where talks are typically given in the local language. For those talks, both subtitles and translations are generated by TED users according to TED's transcription and translation guidelines. According to the guidelines, transcriptions of the talks follow the purpose of subtitles. Transcribers are advised to produce correct sentences, speaker's hesitations and speech artifacts such as "hm", "uhm" are not supposed to be included and obvious mistakes should be corrected, even though this should be indicated. Furthermore, subtitles are limited to a particular length that can be shown at once at the screen and sentences may be modified to fit the subtitling requirements and allow fluent reading and following the talk.

TED and TEDx subtitles and translations are collected and provided as parallel texts for research purposes as the Web Inventory of Transcribed and Translated Talks (WIT^3) (Cettolo et al., 2012). This collection is used in the annual evaluation campaign

Data Set	Evaluation	Data Type	Name	Talks	Sentences
TED	IWSLT 2013	Training	train	1064	158641
		Dev	dev2010	8	887
		Test	test2010	11	1565
	IWSLT 2014	Training	train	1361	171721
		Dev	test2011	16	1433
		Test	test2012	15	1700

Table 5.2: *TED data*

of the International Workshop for Spoken Language Translation (IWSLT) in the Automatic Speech Recognition (ASR), Machine Translation (MT) and Spoken Language Translation (SLT) tasks. Each year, portions of the WIT^3 corpus are provided for training and testing ASR, MT and SLT systems on different languages and translation directions.

For the experiments in this thesis that are operating on TED data, the training, development and test data from the IWSLT 2013 and 2014 evaluation campaigns (Cettolo et al., 2013, 2014) are used. Table 5.2 shows an overview over the TED data used in the TED translation systems.

5.1.3.3 University Lectures

The university lecture data consists of a collection of lectures on computer science and other subjects taught at the Karlsruhe Institute of Technology (KIT) (Stüker et al., 2012). Recordings of selected lectures are transcribed by research assistants according to detailed guidelines. These guidelines differ from the TED transcription guidelines in that the lecture transcriptions are intended as training and test data in automatic speech recognition and machine translation systems. Therefore, they have to meet particular requirements, necessary for research in those fields. As a consequence, the transcriptions need to be very close to the actual spoken words. Speech artifacts such as hesitations, stuttering, mumbled words, aborted and restarted words as well as sentences are annotated as they are spoken. This may result in ungrammatical text, which poses a difficulty for the statistical models of the translation system which are typically trained mostly on grammatically well-formed data.

In the experiments presented in this thesis we use a test set of seven lectures given by five different speakers, with topics covering computer science and history. The length

Data Set	Data Type	Lecture ID	Speaker ID	Length (in h:m:s)	Sentences
Lecture	Test	lect01	sp01	1:31:03	441
		lect02	sp01	1:04:08	398
		lect03	sp02	0:59:29	437
		lect04	sp03	0:46:53	348
		lect05	sp04	0:35:09	124
		lect06	sp05	0:36:17	251
		lect07	sp05	0:50:47	368
	total			4:52:43	1926

Table 5.3: *Lecture data*

of each lecture varies between 35 and 91 minutes. Table 5.3 shows statistics of the lecture data used as test data for the lecture translation task.

5.2 Phrase-based Machine Translation System

Throughout the experiments in this thesis we conduct translation experiments with a phrase-based machine translation system. Translations are generated using a beam search decoder originally developed at Carnegie Mellon University (Vogel, 2003) and continuously adapted at Karlsruhe Institute of Technology to incorporate new functionality. There are two alternatives for generating the word alignment underlying the translation table. On the one hand, we use the word alignment obtained from applying pgiza (Gao and Vogel, 2008), a parallel implementation of the standard GIZA++(Och and Ney, 2003). In some experiments, a discriminative word alignment (DWA) approach (Niehues and Vogel, 2008) is used. Phrases are extracted from the respective word alignment and the translation model is generated with the tools available in Moses (Koehn et al., 2007a). Language modeling is extended beyond standard target surface words to classes ranging from parts-of-speech to automatically generated word clusters using the MKCLS algorithm (Och, 1999). In addition, a bilingual language model is included. It poses an extension to the translation model by additional factors based on bilingual tokens. A language model is used to score the bilingual tokens which consist of source and target language words (Niehues et al., 2011).

The weights for all involved models are optimized by running 20 iterations of Minimum Error Rate Training (MERT) (Och, 2003). We apply a variant of the standard MERT as described in Venugopal et al. (2005). Optimization is done with respect to the BLEU metric (Papineni et al., 2002) on one reference translation, except for translation into French, where two references were available. Translation quality is also measured using the BLEU score.

5.2.1 Reordering Models

The translation system provides the possibility of using various techniques to model the changes of word order during translation. The models are shortly described in the following. Depending on the particular experiment, individual reordering models will be switched on or off in order to investigate interoperability of the respective models. This will be indicated accordingly in the description of the experiments.

Rule Type		Example Rule	
POS-based	Short-range	*VVIMP VMFIN PPER*	→ 2 1 0
	Long-range	*VAFIN * VVPP*	→ 0 2 1

Table 5.4: *Rule types*

5.2.1.1 Distance-based Reordering

This type of reordering model is applied at decoding time when processing the input sentence. While building up the translation incrementally from left to right, the decoder may delay the translation of words within a window of size *d*.

When one of the source reordering based on part-of-speech tags described below in Section 5.2.1.3 is applied, the reordering window is limited to a minimum (2). That means we typically allow reordering only by swapping adjacent words, if another dedicated reordering model is included.

5.2.1.2 Lexicalized Reordering Model

The lexicalized reordering model (Koehn et al., 2005; Tillmann, 2004) contains reordering probabilities for all phrases in the phrase table. Possible reordering orientation of a given phrase with respect to adjacent phrases are: monotone, swap and discontinuous. A phrase receives reordering probabilities for each of those orientations according to observed reordering instances in the training data. The lexicalized reordering model is part of the log-linear combination in the translation system and receives a model weight during optimization.

5.2.1.3 Part-of-Speech-based Reordering Model

We apply two approaches based on continuous and discontinuous sequences of parts-of-speech of the words in the sentence as described in Rottmann and Vogel (2007) and Niehues and Kolss (2009), respectively. By combining them, both short-range and long-range reordering phenomena between source and target language can be covered. We distinguish between short-range and long-range part-of-speech-based reordering rules. The part-of-speech tags are generated using the Tree Tagger (Schmid, 1994). Examples for each of the rule types are presented in Table 5.4.

Short-range Rules Short-range rules consist of a sequence of part-of-speech (POS) tags on the left hand side and an indexed representation of the target order of those POS tags on the right hand side of the rule. Each rule comes with an associated probability which is the relative frequency of the occurrence of this reordering in the training corpus.

Long-range Rules A long-range rule consists of a sequence of POS tags with placeholders on the left hand side. A placeholder can match arbitrary types and number of POS tags. The right hand side of the rule contains the reordered indices that indicate the new order of the components of the rule. The tags matched by the placeholder are assigned one index as a whole. Again, a probability is assigned to each rule.

Learning Reordering Rules For the training of the reordering rules a parallel corpus and a word alignment is required. In addition, POS tags are needed for the source side of the corpus for training the reordering rules. For each sentence in the training corpus we search for changes of word order between the source and target language sentence. When a crossing alignment indicates a different order of source and target language words, the alignment is monotonized and a rule is extracted that rearranges the source words in the order of the aligned target words. For more details refer to the descriptions of short-range and long-range POS-based rules (Niehues and Kolss, 2009; Rottmann and Vogel, 2007).

Applying Reordering Rules Before translation, a word graph (word lattice) is created for each sentence. First, the original source sentence is included as the monotone path and all edges are assigned a transition probability of 1. Then all matching reordering rules are applied and the resulting reordering variants of the sentence are stored in the word lattice. The edges of the reordered path are assigned transition probabilities according to the probability of the applied reordering rule. An edge branching from the monotone path receives the probability of the rule. The following edges in the reordered path are assigned a probability of 1. The edge on the monotone path where the branching takes place receives an update such that the probability of the applied rule is subtracted from the current transition probability of this edge. A minimum transition probability of 0.05 is kept for the monotone path, i.e. the original word order of the

sentence. Finally, the word lattice including all reordering variants is used as input to the decoder.

Judging Reordered Paths The probability of a given path in a reordering lattice is calculated as the product of the transition probabilities of the traversed edges. Since the transition probabilities are based on the occurrences of the reordering in the training data, higher scoring paths in the lattice should represent good reordering options for the given sentence. However, the final decision which reordering path to apply in translation is taken during decoding. Hence, the reordering lattice with its reordering paths and probabilities is included as an additional model in the log-linear model of the translation system. Its weight is set during optimization of the translation system together with the weights of the other models.

5.2.2 Discriminative Word Lexicon

The discriminative word lexicon (DWL) models the occurrence of individual words in the translation output. It consists of individual classifiers for each target word e_j in the translation of a given source sentence f. All source words f_j of the source sentence f are provided as features for the maximum entropy classifier, which then decides whether the current target word e'_j should occur in the translation or not (Mauser et al., 2009). Positive training examples for each classifier are compiled from all sentence pairs in the parallel training data where the target word e'_j occurs in the target sentence. Compared to the original DWL, Niehues and Waibel (2013) present an extension to the model changing the way negative training examples are generated. Instead of using all target sentences where e'_j does not occur, only sentences where e'_j is in the target vocabulary but not in the target sentences are used as negative examples. The target vocabulary of a sentence consists of all target side words of phrase pairs matching a source phrase in the source sentence of the training data.

Another difference to the original DWL is that source and target context is modeled by using new types of features. Source context and source word order is included by means of bag-of-source-n-gram features. They use one feature per n-gram up to the order of three and apply count filtering for bigrams and trigrams. Also, target context words are included in the set of features.

The sentence probability is calculated as the product of the individual word probabilities in the following way:

$$p(e|f) = \prod_{j=1}^{J} p(e_j|f) \tag{5.1}$$

In this definition, $p(e_j|f)$ is calculated using a maximum entropy classifier. In order to save time during decoding, the scores for all phrase pairs are precalculated.

5.2.3 N-Best List Re-Ranking

For some of the experiments in this thesis we perform N-best list re-ranking with the ListNet algorithm (Cao et al., 2007) on the 300 best translation hypotheses as described in Slawik et al. (2014). We use two data sets for training, one for validation and one for testing. For each translation hypothesis in the N-best lists, a set of scores is available from the translation system. This set of scores comprises several word-based, POS-based and cluster-based language model scores, translation model scores, scores from the reordering lattices as well as other models depending on the respective setup of the machine translation system. There are two possible ways to apply N-best list re-ranking. The first method is to use only the original set of scores from the translation system. Alternatively, additional scores can be included for new models that were not used in the original setup of the translation system. The re-ranking algorithm is used to learn new weights for the original and possibly new models in order to provide a better judgment for translation quality.

6

Syntactic Reordering

The linguistic challenges described in Chapter 3 show the need for a linguistically informed approach which can integrate knowledge about sentence structure. This chapter first presents the developed reordering model based on syntactic parse trees. The trees provide the information about the sentence structure in terms of the construction of words into constituents, constituents into bigger constituents and finally into a sentence. The reordering model consists of automatically learned reordering rules that determine how words of particular parts-of-speech and sequences of words in particular sentence constituents should be reordered in the source sentence in order to facilitate monotone translation. The reordering model is described in Section 6.1. Section 6.2 presents a method for combining it with other types of reordering models operating on different linguistic abstraction levels. Section 6.3 describes oracle experiments that investigate the current performance and potential of the tree-based reordering model as well as the source reordering approach in general. Both automatic and manual evaluations are performed. Section 6.4 shows the automatic evaluation of the tree-based reordering approach in German-to-English and German-to-French translation. Then the results of the oracle experiments on German-English and English-German translation are presented. Section 6.5 concludes this chapter with a detailed manual analysis of the tree-based reordering approach. Analyzing the overall impact, individual improvements and affected word categories in three different genres, the ability of the reordering model to generalize can be confirmed. The work presented in this chapter is based on the following publications. The the tree-based reordering model is introduced in Herrmann et al. (2013a) and Herrmann et al. (2013b) describes the oracle

experiments. The manual analysis is presented in Herrmann et al. (2014).

6.1 Source Reordering with Syntactic Parse Trees

The tree-based reordering model performs reordering on the source language side, learning how to rearrange the source words according to the correct word order of the target language. After the source words are reordered, monotone translation into the target language can be performed. The reordering model consists of reordering rules that operate on the syntactic level of the sentences of the source language. The rules are automatically learned and encourage word reordering motivated by the sentence structure. While the part-of-speech-based reordering rules proposed by Rottmann and Vogel (2007) and Niehues and Kolss (2009) are flat and perform the reordering on a sequence of words, the tree-based rules operate on subtrees in the syntactic parse tree of a complete sentence as shown in Figure 6.1. The subtree headed by a verb phrase (VP) with three child constituents (PTKNEG, NP and VVPP) is reordered by arranging the children in a new order.

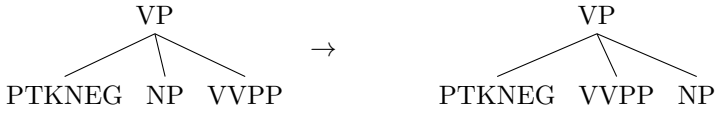

Figure 6.1: Example reordering based on subtrees

A syntactic parse tree contains both the word-level categories, i.e. parts-of-speech and higher order categories, i.e. constituents. In this way it provides information about the building blocks of a sentence that belong together and should not be taken apart by reordering. Consequently, the tree-based reordering operates both on the word level and on the constituent level to make use of all available information in the parse tree. It is able to handle long-range reorderings as well as short-range reorderings, depending on how many words the reordered constituents cover. If the constituents in the rule are on the word level and thus at the bottom of the tree, short-range reordering is performed. If the rule operates on a higher tree level, longer spans of the sentence are covered. The tree-based reordering rules should also be more stable and introduce less random word shuffling than the part-of-speech-based rules.

The reordering model consists of two parts. First the rule extraction is done in the training phase, where the rules are learned by searching the training corpus for non-monotonic alignments which indicate a reordering between source and target language. The application of the learned reordering rules to the input text takes place prior to translation.

6.1.1 Rule Extraction

As shown in Figure 6.1 we learn rules that reorder the children in a subtree of a syntactic parse tree for a sentence. Example 6.1 shows a reordering rule representing the tree reordering above. The first item in the rule is the head node H of the subtree and the rest represent the three children (indices 0 to 2). In the second part of the rule, the indices represent the new order in which the children of that subtree should be rearranged.

$$VP_H \ PTNEG_0 \ NP_1 \ VVPP_2 \rightarrow 0 \ 2 \ 1$$

Example 6.1: Tree-based reordering rule

Figure 6.2 presents an example for rule extraction: a sentence in its syntactic parse tree representation, the sentence in the target language and an automatically generated alignment. A reordering occurs between the constituents NP and $VVPP$.

In a first step the reordering rules have to be found. We extract the rules from a word aligned corpus where a syntactic parse tree is provided for each source side sentence. We traverse the tree top down and scan each subtree for instances of reordering, indicated by crossings of alignment links between source and target sentence. If there is a reordering, we extract a rule that rearranges the source side constituents according to the order of the corresponding words on the target side. Each constituent in a subtree comprises one or more words. For every source word f_i we define a_i as the set of indices of the target words e_j it is aligned to. We determine the lowest (min) and highest (max) alignment point for each constituent c_k and thus determine the range of the constituent on the target side. This can be formalized as $min(c_k) = min\{j | f_i \in c_k; j \in a_i\}$ and $max(c_k) = max\{j | f_i \in c_k; j \in a_i\}$. To illustrate the process, we have annotated the parse tree in Figure 6.2 with the alignment points (min-max) for each constituent.

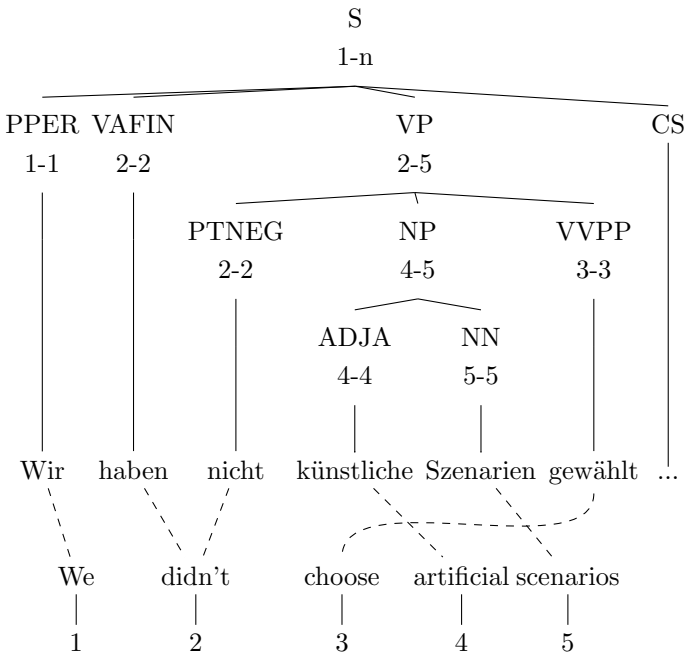

Figure 6.2: Example training sentence used to extract reordering rules

After defining the alignment range, we check for the following conditions in order to determine whether to extract a reordering rule.

1. all constituents have a non-empty range
2. source and target word order differ

First, for each subtree at least one word in each constituent needs to be aligned. Otherwise it is not possible to determine a conclusive order. Second, we check whether there is actually a reordering, i.e. the target language words are not in the same order as the constituents in the source language: $min(c_k) > min(c_{k+1})$ and $max(c_k) > max(c_{k+1})$.

Once we find a reordering rule to extract, we calculate the probability of this rule as the relative frequency. Hence, we divide the number of occurrences of this reordering in the training corpus by the number of total occurrences of this subtree in the corpus. We only store rules for reorderings that occur more than five times.

6.1.1.1 Partial Rules

The syntactic parse trees of German sentences are quite flat, i.e. a subtree usually has many children. When a rule is extracted, it always consists of the head of the subtree

and all its children. The application requires that the applicable rule matches the complete subtree: the head and all its children. However, most of the time only some of the children are actually involved in a reordering. There are also many different subtree variants that are quite similar. In verb phrases or noun phrases, for example, modifiers such as prepositional phrases or adverbial phrases can be added nearly arbitrarily. In order to generalize the tree-based reordering rules, we extend the rule extraction. We do not only extract the rules from the complete child sequence, but also from any continuous child sequence in a constituent. This way, we extract generalized rules which can be applied more often. Formally, for each subtree $h \rightarrow c_1^n = c_1 c_2 ... c_n$ that matches the constraints presented in Section 6.1.1, we modify the basic rule extraction such that $\forall l, m \; 1 \leq l < m \leq n : h \rightarrow c_l^m$. It could be argued that the partial rules might be not as reliable as the specific rules. In Section 6.4.1 we will show that such generalizations are meaningful and can have a positive effect on the translation quality.

6.1.2 Rule Application

During the training step all reordering rules are extracted from the parallel corpus. Prior to translation the rules are applied to the original source text, creating a word graph which is later used as input to the decoder. The word graph first includes only the source sentence in the original word order. Similar to the idea of graph grammars (Rozenberg, 1997), which have been successfully applied in many computer science tasks, e.g. in Reussner et al. (2005), we apply the reordering rules to the word graph of possible source word orders. In the following, we will refer to this word graph as word lattice or reordering lattice. Each rule is applied independently producing a reordering variant of that sentence. The rules may be applied recursively to already reordered paths. If more than one rule can be applied, all paths are added to the lattice unless the rules generate the same output. In this case only the rule with the highest probability is applied.

The edges in a word lattice for one sentence are assigned transition probabilities based on the rule probabilities. In the monotone path with original word order all transition probabilities are initially set to 1. In a reordered path the first branching transition is assigned the probability of the rule that generated the path. All other transition probabilities in this path are set to 1. Whenever a reordered path branches from the monotone path, the probability of the branching edge is subtracted from the

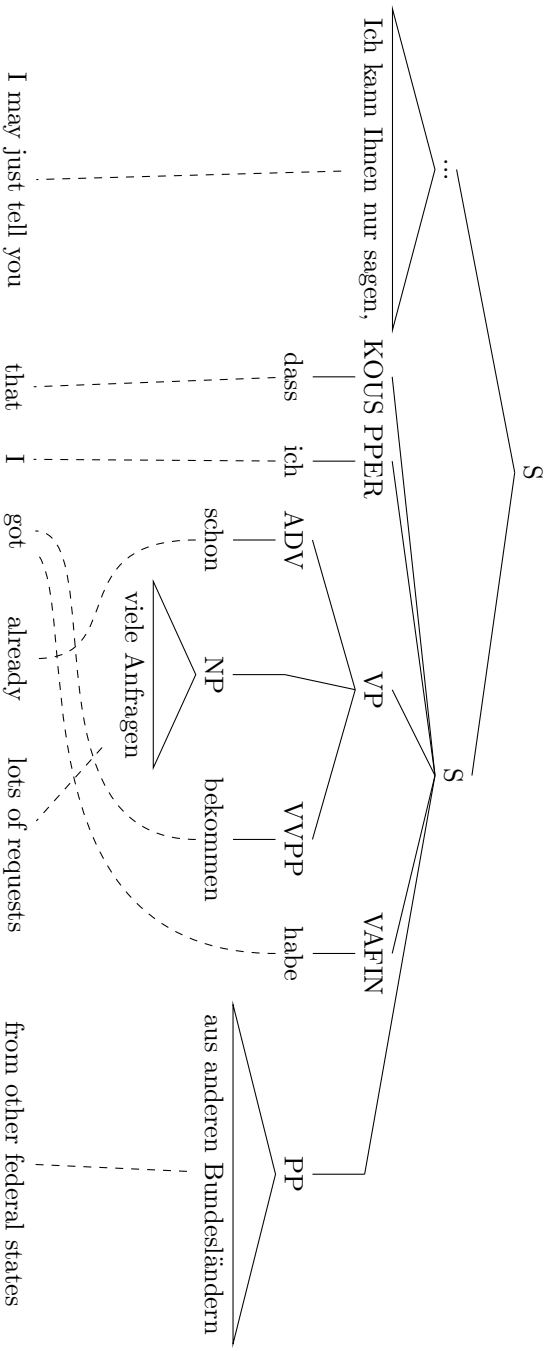

Figure 6.3: Example parse tree with separated verb particles

probability of the monotone edge. However, a minimum probability of 0.05 is reserved for the monotone edge in the path which represents the original word order. The score of the complete path is computed as the product of the transition probabilities. During decoding the best path is searched for by including the score for the current path weighted by the weight for the reordering model in the log-linear model of the translation system. In order to enable efficient decoding we limit the lattice size by only applying rules with a probability higher than 0.1. This threshold was determined empirically in initial experiments.

6.1.2.1 Recursive Rule Application

As mentioned above, the tree-based rules may be applied recursively. That means, after one rule is applied to the source sentence, a reordered path may be reordered again. The reason lies in the structure of the syntactic parse trees. Verbs and their particles are typically not located within the same subtree. Hence, they cannot be covered by one reordering rule. A separate rule is extracted for each subtree. Figure 6.3 demonstrates this in an example. The two parts that belong to the verb in this German sentence, namely *bekommen* and *habe*, are not located within the same constituent. The finite verb *habe* forms a constituent of its own and the participle *bekommen* forms part of the VP constituent. In English the finite verb and the participle need to be placed next to each other. In order to rearrange the source language words according to the target language word order, the following two reordering movements need to be performed: the finite verb *habe* needs to be placed before the VP constituent and the participle *bekommen* needs to be moved within the VP constituent to the first position. Only if both movements are performed, the correct word order can be generated.

However, the reordering model only considers one subtree at a time when extracting reordering rules. In the example sentence in Figure 6.3 two rules are learned, but if they are applied to the source sentence separately, they will end up in separate paths in the word lattice. The decoder then has to choose which path to translate: the one where the finite verb is placed before the VP constituent **or** the path where the participle is at the first position in the VP constituent.

In order to allow the correct reordering to be achieved in such cases, the rules may be applied recursively to the new paths created by our reordering rules. We use the same rules, but newly created paths are fed back into the queue of sentences to be

reordered. However, we only apply the rules to parts of the reordered sentence that are still in the original word order and restrict the recursion depth to 3 levels.

6.2 Combining Reordering Methods

We want to measure both the performance of the presented tree-based reordering model and compare its performance with state-of-the-art reordering models operating on different linguistic abstraction levels. This way, we hope to get a deeper insight into their individual strengths. By combining them we investigate whether their respective gains in translation quality overlap or complement each other. We address the word level using the lexicalized reordering, the morphosyntactic level by part-of-speech-based (POS-based) reordering and the constituent level by tree-based reordering.

6.2.1 POS-based and Tree-based Reordering Rules

For the combination of POS-based and tree-based reordering rules, we use POS-based reordering as described in Section 5.2.1.3. We apply both short-range reordering consisting of fixed POS sequences, and long-range reordering consisting of POS sequences with placeholders matching arbitrary embedded POS sequences. The general term POS-based reordering in our case typically comprises both types, short-range and long-range reordering. If only one rule type is used, it is indicated accordingly. The tree-based rules are trained separately as described above. First, the POS-based rules are applied to the monotone path of the source sentence and then the tree-based rules are applied independently, producing separate paths. Table 6.1 shows an overview of the three rule types used for combination.

Rule Type		Example Rule	
POS-based	Short-range	*VVIMP VMFIN PPER*	→ 2 1 0
	Long-range	*VAFIN * VVPP*	→ 0 2 1
Tree-based		*VP PTNEG NP VVPP*	→ 0 2 1

Table 6.1: *Rule types*

6.2.2 Reordering Rules and Lexicalized Reordering

As described in Section 6.1.2 we create word lattices that encode the reordering variants. The lexicalized reordering model (cf. Section 5.2.1.2) stores for each phrase pair the probabilities for possible reordering orientations at the incoming and outgoing phrase boundaries: monotone, swap and discontinuous. In order to apply the lexicalized reordering model on lattices the original position of each word is stored in the lattice. While the translation hypothesis is generated, the reordering orientation with respect to the original position of the words is checked at each phrase boundary. The probability for the respective orientation is included as an additional score in the log-linear model of the translation system.

6.3 Oracle Reordering

We want to assess the benefits of the source reordering approach and investigate how much it can help to improve the translation. For one, we want to determine lower and upper bounds for the translation quality that can be reached by this approach and to identify potential for further development. Furthermore, we want to assess the performance of the reordering model on two levels: The restriction of the search space of possible reorderings and the ranking of different reordering variants.
We designed oracle experiments that address the following questions:

- How good is the translation of the optimally reordered source sentence?
- How beneficial is the restriction of the search space through reordering lattices for translation quality?
- How accurate is the search for the best path in the reordering lattice?

In order to answer these questions, we compare the actual system performance against two different reordering oracles. The first oracle is the optimally reordered source sentence which presents the source words according to the target language word order. With this experiment we analyze the effectiveness of the preordering approach. By reordering the source sentence according to the target language word order we estimate an upper bound for translation quality using this strategy.

Then we investigate how the reordering lattices produced by the POS-based and tree-based reordering model restrict the search space for translation. Therefore, we compare the translation of the aforementioned oracle reordering with the translation of the oracle path. This is the path in the lattice that is closest to the oracle reordering of the source sentence. We perform this experiment for each of the different types of reordering rules.

In a third experiment we evaluate how good our models are at determining the best path in the lattice. In order to evaluate this aspect, we compare the translation of the oracle path with the actual translation where the path is chosen during decoding.

6.3.1 Optimally Reordered Sentence

In order to measure the oracle performance of the preordering approach, we use an optimally reordered sentence as input to the translation system and do not allow additional reordering during decoding. In order to create this oracle reordering for the source sentence, we make use of the word alignment between source sentence and reference translation. This alignment is generated by applying the alignment model trained during system development to the test data and its reference translation. After source and reference are aligned, we create a permutation of the source sentence (Birch et al., 2010).

In the permutation, words are generally assigned the position of the word they are aligned with. However, permutations are one-to-one alignments, while word alignments may also contain unaligned words, many-to-one alignments and one-to-many alignments. Therefore, some simplifying assumptions have to be made when transforming alignments to permutations (Birch, 2011): **unaligned source words** are aligned to the word after its predecessor or to the first word if it has no predecessor; **unaligned target words** are irrelevant to the source sentence order and are therefore ignored; for **many-to-one source-to-target alignments** the ordering is assumed to be monotone; in **one-to-many source-to-target alignments** the word is assumed to be aligned to the first target word. We will refer to this reordered source sentence as the **oracle reordering** of the input sentence.

6.3.2 Oracle Path

With our reordering model we generate many reordering variants by applying reordering rules to the source sentence and store these variants in a lattice. In order to know the upper bound of the restriction of the search space by the lattice we want to identify the best reordering variant in the reordering lattice. We define it as the path in the lattice which has the smallest distance to the oracle reordering as described above.

Among Hamming distance, Ulam's Distance and Kendall's τ distance, a version of Kendall's τ resulted to be the best distance metric, being the most reliable and correlating strongly with human fluency judgement (Birch et al., 2010). Hence, we calculate the Kendall's τ distance (Kendall and Gibbons, 1990) in order to find the path that is closest to the oracle reordering. The Kendall's τ distance is the minimum number of swaps between two adjacent symbols that transforms a permutation σ into another permutation π. This metric measures relative differences and takes both the number and the size of reorderings into account. We use the square root version (Birch, 2011) which corresponds closely with human perception of word order quality:

$$d(\pi, \sigma) = 1 - \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=i}^{n} x_{ij}}{Z}}$$

$$\text{where } x_{ij} = \begin{cases} 1 & \text{if } \pi(i) < \pi(j) \text{ and } \sigma(i) > \sigma(j) \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } Z = \frac{n \cdot (n-1)}{2}$$

If a path with the oracle reordering is in the lattice, this path is the closest path. However, if the oracle reordering is not in the lattice, several paths can have the smallest distance to the oracle reordering. Then we create lattices containing only the best paths and use these as input to the translation system.

Note that the best path or even the oracle reordering need not result in the best possible translation quality for two reasons. First, we rely on the alignment between source and reference for generating the oracle reordering. Errors in the alignment can introduce errors into the oracle reordering and the closest path. Another reason is that we generate an artificial word order which does not match the word order as seen in the training data. Therefore, we might not have well matching phrase pairs for generating the best possible translation.

6.4 Automatic Evaluation

In this section we perform automatic evaluations of the tree-based reordering model, the reordering model combinations and the oracle reordering experiments. The translation quality is measured using the automatic metric BLEU and reordering quality is presented according to Kendall's τ metric for measuring the distance between two reordering variants of a sentence.

6.4.1 Tree-based Reordering Model

The tree-based reordering model was tested on two language pairs, translating from German into English and from German to French. For both translation directions, we built systems using POS-based and tree-based reordering and show the impact of the individual models as well as their combination on the translation quality. For each system, two different setups were evaluated. First, with a distance-based reordering model only (noLexRM) and with an additional lexicalized reordering model (LexRM). The baseline system which uses no reordering rules at all allows a reordering window of 5 in the decoder for both setups. For all systems where reordering rules are applied, monotone translation is performed. Since the rules take over the main reordering effort, only monotone translation is necessary from the reordered word lattice input.

6.4.1.1 German-English

The results for German-to-English translation are presented in Table 6.2. In this experiment, we first compare the tree-based rules with and without recursive application, and the partial rules. Then the POS-based and tree-based reordering is combined as described in Section 6.2.1.

Compared to the baseline system using distance-based reordering only, 1.4 BLEU points can be gained by applying combined POS and tree-based reordering. The tree-based rules including partial rules and recursive application alone achieve already a better performance than the POS-based rules, but using them all in combination leads to an improvement of 0.4 BLEU points over the POS-based reordering alone. When lexicalized reordering is added, the relative improvements are similar: 1.1 BLEU points compared to the Baseline and 0.55 BLEU points over the POS-based reordering. We can therefore argue that the individual rule types of the rule-based reordering model as

System / Rule Type	noLexRM Dev	noLexRM Test	LexRM Dev	LexRM Test
Baseline (no Rules)	22.82	21.06	23.54	21.61
POS	24.33	21.98	24.42	22.15
Tree	24.01	21.92	24.24	22.01
Tree recursive	24.37	21.97	24.53	22.19
Tree recursive + partial	24.31	22.21	24.65	22.27
POS + Tree	24.57	22.21	24.91	22.47
POS + Tree recursive	24.61	22.39	24.81	22.45
POS + Tree recursive + partial	24.80	22.45	24.78	22.70

Table 6.2: *Tree-based reordering results: German-English*

well as the lexicalized reordering model each seem to address complementary reordering issues and can be combined successfully to obtain an even better translation quality.

We applied only tree rules with a probability of 0.1 and higher. Partial rules require a threshold of 0.4 to be applied, since they are less reliable. The recursive rule application is restricted to a maximum recursion depth of 3, such that a maximum of three rules is applied to a given subpath. This is meant to prevent the lattices from growing too large, which would increase decoding time severely. The values for the rule thresholds were set according to the results of initial experiments investigating the impact of the rule probabilities on the translation quality. Full rules and partial rules are not mixed during recursive application.

With the best system we performed a final experiment on the official testset of the WMT 2012 and achieved a score of 23.73 which is 0.4 BLEU points better than the best constrained submission.

6.4.1.2 German-French

The reordering model was also evaluated on German-French translation. For this language pair, similar improvements could be achieved by combining POS and tree-based reordering rules and applying a lexicalized reordering model in addition. Table 6.3 shows the results. Up to 0.7 BLEU points could be gained by adding tree rules and another 0.1 by lexicalized reordering.

System / Rule Type	noLexRM Dev	noLexRM Test	LexRM Dev	LexRM Test
POS	41.29	38.07	42.04	38.55
POS + Tree	41.94	38.47	42.44	38.57
POS + Tree recursive	42.35	38.66	42.80	38.71
POS + Tree recursive + partial	42.48	38.79	42.87	38.88

Table 6.3: *Tree-based reordering results: German-French*

6.4.1.3 Binarized Syntactic Trees

Since related work using syntactic parse trees in statistical machine translation for reordering purposes (Jiang et al., 2010) have reported an advantage of binarized parse trees over standard parse trees, we also produced binary tree rules. The Stanford parser (Rafferty and Manning, 2008) was used to generate the standard parse trees and to binarize them afterwards. However, binarizing our parse trees and working with binary rules led to decreased translation quality in our case. Even though the binary rules were tested with varying thresholds, the translation quality of the tree-based rules based on standard syntactic trees could not be reached. The BLEU scores were about 0.2 points lower. It seems that the flat hierarchical structure of standard parse trees enables our reordering model to learn the order of the constituents most effectively.

6.4.2 Oracle Reordering

In this section we present three experiments designed to address the three questions raised in in Section 6.3. First, we will analyze the potential of the source reordering approach. Afterwards, we investigate how the reordering lattices produced by our reordering model restrict the search space for translation. In a third experiment we compare the oracles with the actual performance of a system using the reordering lattices. This way we want to find out how good the models are at ranking different word orders.

6.4.2.1 Potential of Reordering the Source Sentence

When applying source reordering as a preprocessing step for translation, it is commonly assumed that arranging the source sentence according to target language word order should result in better translation quality. We want to question this assumption and investigate the benefits of the preordering approach in a first experiment that identifies

the lower and upper bounds of translation quality with respect to word order. We consider the lower bound of translation quality to be the performance that is obtained by translating the source sentence without allowing any additional reordering. Since the objective of the preordering approach is to obtain the source words in the order of the target language words, we regard the translation of the optimally reordered path to be the upper bound for translation quality. We generate the optimally reordered path using the reference translation and the alignment between source and reference as described in Section 6.3.1.

German-English Table 6.4 presents the results for the translation from German to English in two different domains: translation of News articles and TED talks. The difference between monotone translation and the translation of the oracle reordering is 5.2 and 6.2 BLEU points, for News and TED respectively. With a system using the lattice-based reordering approach in the standard way, applying both POS-based and tree-based rules, we achieve a performance that is approximately in the middle of that range. No oracle information is available. Instead, the decoder chooses the path with a particular reordered source sentence during translation.

Reordering Type	News	TED
Monotone	20.23	27.18
Lattice Reordering	22.45	30.87
Oracle	25.42	33.39

Table 6.4: *Oracle reordering: German-English*

English-German Table 6.5 shows the results for the reverse translation direction. We can see lower absolute BLEU scores, since translation into German is more difficult due to the highly inflective morphology of the German language. Compared to German-English translation, the difference between monotone and oracle translation is smaller, 2.9 and 4.6 BLEU points, for News and TED translation, respectively. Decoding with reordering lattices performs better than the monotone translation, but the gap towards the oracle translation is bigger. We infer that for English to German translation, there is even more potential for improvement through better reordering rules for generating the lattices.

Reordering Type	News	TED
Monotone	15.91	24.22
Lattice Reordering	16.34	24.95
Oracle	18.84	28.77

Table 6.5: *Oracle reordering: English-German*

From this experiment we can draw the conclusion that reordering the source sentences prior to translation indeed holds promising results. Our system using reordering lattices as translation input outperforms the monotone translation in all four translation tasks, and the oracle reordering shows that there is still potential for improvement through better reordering methods. In the following we will investigate how we can best address this potential by analyzing different aspects of the reordering approach in detail.

6.4.2.2 Lattice-based Restriction of the Search Space

In the previous experiment we have identified a gap between the actual performance of the system using reordering lattices and the oracle reordered translation. In our reordering approach we restrict the search space of possible reorderings by the reordering lattice. In this second experiment we want to investigate how much this restriction influences the drop in performance. Therefore, we evaluate how much better we could get, if the decoder found the best path in the given reordering lattices. As described in Section 6.3.2 we define the best path as the one that is closest to the oracle reordering, i.e. the optimally reordered sentence used in the previous experiment.

In order to compare the benefits of individual reordering rule types we apply all the different types of reordering rules and identify the oracle path within the lattices produced by those rules. Then we perform translation of the oracle path and compare the translation quality.

The tables in the following sections also include the scores for the monotone and oracle translation presented above. In addition, they show the translation results for systems using first short and long-range rules based on part-of-speech tags. Afterwards follow the tree-based rules, first the plain tree rules, then the tree-based rules with recursive rule application and the third tree rule option includes partial rules. The details on recursive rule application and partial rules are described in Sections 6.1.2.1 and 6.1.1.1. The three final systems combine all rule types.

German-English Table 6.6 shows the results for German-to-English translation. For each system using a different type of reordering we present translation quality and the size of the search space represented by the number of edges in the reordering lattice produced by the respective type of rule. As can be seen, the more complex the rule types for generating the reordering lattice, the more the search space increases. In the same way as the search space gets bigger, also the translation of the oracle path in that lattice gets better. The oracle path that is closest to the oracle reordering stems from the lattice produced by applying all rule types.

Reordering Type	News		TED	
	BLEU	Size	BLEU	Size
Monotone	20.23		27.18	
Short	21.37	193K	29.98	68K
Short + Long	21.41	255K	30.66	163K
Tree	21.88	140K	29.74	51K
Tree recursive	22.17	244K	30.11	81K
Tree recursive + partial	22.28	249K	30.22	82K
Short + Long + Tree	22.49	429K	30.97	182K
Short + Long + Tree recursive	22.64	534K	31.10	212K
Short + Long + Tree recursive+partial	22.65	538K	31.12	213K
Oracle	25.42		33.39	

Table 6.6: *Oracle path: German-English*

English-German Table 6.7 presents the same experiments for English-to-German translation. Again, the more complex rules and bigger search spaces lead to better oracle paths. Thus, we can confirm the findings in Section 6.4.1 namely that the different rule types produce complementary reordering possibilities which result in the best translation quality if combined in one lattice. We can also see that the translation of the best oracle path is still far from the oracle reordered translation. The lattices generated with the help of our reordering rules restrict the search space in a sensible way to allow for reorderings that are getting closer to the oracle reordered sentence. However, some reordering possibilities are still missing from our lattices. Therefore, research in the area of extending the search space by better rules seems to be promising.

Reordering Type	News		TED	
	BLEU	Size	BLEU	Size
Monotone	15.91		24.22	
Short	16.31	186K	25.83	76K
Short + Long	16.70	383K	25.99	170K
Tree	16.48	189K	25.31	71K
Tree recursive	16.60	726K	25.49	237K
Tree recursive + partial	16.60	727K	25.49	237K
Short + Long + Tree	17.00	496K	26.28	208K
Short + Long + Tree recursive	17.07	1M	26.38	373K
Short + Long + Tree recursive + partial	17.07	1M	26.38	373K
Oracle	18.84		28.77	

Table 6.7: *Oracle path: English-German*

6.4.2.3 Ranking different word orders

The experiments above revealed the best possible translation that can be produced by using the individual rule types and combinations thereof. Now we want to examine how well we actually perform in finding the best path in the lattices. Again, we tested on all the different rule types, but let the decoder find the best path for translation. It is worth mentioning that the decoder does not only utilize the scores of the reordering model described in Section 5.2.1.3 to find the path, but all the models in the log-linear model of the translation system contribute a score while constructing each translation hypothesis. For reference we include the BLEU scores achieved with the oracle paths from the previous experiment. In addition, we present the average distances between the decoder path used for translation and the optimally reordered sentence both for the decoder translation and for the translation of the oracle path. The distances are calculated using the Kendall's τ metric.

German-English We present the results for German-to-English translation in Tables 6.8 and 6.9. The differences between the oracle path scores and the actual performance of the system (**decoder path**) with the reordering lattices are very small. This means that the decoder is already quite good at finding the best path in the reordering lattice. To reach the translation quality of the oracle path, a further increase of 0.2 and 0.3 BLEU points would be possible for the News and the TED task, respectively.

Reordering Type	News			
	DecoderPath		OraclePath	
	BLEU	Distance	BLEU	Distance
Monotone			20.23	
Short	21.59	0.290	21.37	0.250
Long	21.35	0.286	21.41	0.259
Tree	21.78	0.286	21.88	0.250
Tree recursive	22.01	0.284	22.17	0.243
Tree recursive + partial	22.10	0.284	22.28	0.241
Short + Long + Tree	22.33	0.289	22.49	0.224
Short + Long + Tree recursive	22.44	0.288	22.64	0.220
Short + Long + Tree recursive + partial	22.45	0.288	22.65	0.220
Oracle			25.42	

Table 6.8: *Oracle vs. actual performance: German-English (News)*

Reordering Type	TED			
	DecoderPath		OraclePath	
	BLEU	Distance	BLEU	Distance
Monotone			27.18	
Short	30.00	0.179	29.98	0.124
Long	30.73	0.181	30.66	0.112
Tree	29.60	0.180	29.74	0.140
Tree recursive	29.88	0.179	30.11	0.135
Tree recursive + partial	29.96	0.179	30.22	0.133
Short + Long + Tree	30.82	0.182	30.97	0.106
Short + Long + Tree recursive	30.86	0.182	31.10	0.104
Short + Long + Tree recursive + partial	30.87	0.182	31.12	0.104
Oracle			33.39	

Table 6.9: *Oracle vs. actual performance: German-English (TED)*

The distances between decoder translation path and oracle reordering are shown in the column to the right of the decoder path, while the distances between the oracle path and the oracle reordering are shown in the column to the right of the scores reached by the oracle path translations. We can see that both the distances and the translation quality for the oracle path systems converge nicely for the News task. The closer the translation quality gets to the translation quality of the oracle reordering, the smaller also the reordering distance to the oracle reordering. In the TED task we also observe a good correspondence between translation quality and reordering distance for the oracle path results. The drop in BLEU score when using only tree rules is also obvious in the distance scores, which raise for those systems. For the decoder translation path, the distance to the oracle reordering seems to be not converging at all, it stays about the same both for News and TED translations.

English-German The results for English-to-German translation are presented in Tables 6.10 and 6.11. For this translation direction, the path in the reordering lattices chosen by the decoder is not very close to the optimal one yet. The decoder performance is 0.7 BLEU points worse than the translation of the oracle path for the best rule type of the News task. For the TED task, the difference between oracle path translation and decoder performance is even 1.4 BLEU points.

The distance scores show a similar behavior as observed for German-English translation. The distances from oracle path to oracle reordering get smaller as the translation quality increases. The distances from decoder translation path to oracle reordering do not converge. Compared to the German-English results, they vary even more. It is possible that this is due to the smaller differences in translation quality. In addition, outliers in the paths chosen by the decoder could cause the variations in the distance scores.

From these results we can draw the conclusion that some potential still lies in the reordering rules and therefore in the reordering lattices that the decoder is not yet able to make use of. The differences in the translation quality achieved by the decoder path and oracle path suggest that more complex scoring models for better judging reordering quality are needed, so that the decoder can make better decisions in choosing a reordering path from the lattice. As mentioned in Section 5.2.1.3, the score for a path in a reordering lattice is calculated from the probabilities of the reordering rules applied to generate this reordering. This seems to work reasonably well for German-English translation, where the path chosen by the decoder is quite close to the best path in the lattice. However, the results of English-German translation suggest that a better

Reordering Type	News			
	DecoderPath		OraclePath	
	BLEU	Distance	BLEU	Distance
Monotone			15.91	
Short	16.27	0.297	16.31	0.249
Long	16.31	0.311	16.70	0.236
Tree	16.21	0.306	16.48	0.252
Tree-rec	16.18	0.312	16.60	0.244
Tree-rec-partial	16.18	0.312	16.60	0.244
Short+Long+Tree	16.32	0.318	17.00	0.227
Short+Long+Tree-rec	16.34	0.321	17.07	0.222
Short+Long+Tree-rec-partial	16.34	0.321	17.07	0.222
Oracle			18.84	

Table 6.10: *Oracle vs. real: English-German (News)*

Reordering Type	TED			
	DecoderPath		OraclePath	
	BLEU	Distance	BLEU	Distance
Monotone			24.22	
Short	24.83	0.200	25.83	0.141
Long	24.87	0.214	25.99	0.129
Tree	24.47	0.206	25.31	0.163
Tree-rec	24.51	0.207	25.49	0.158
Tree-rec-partial	24.50	0.207	25.49	0.158
Short+Long+Tree	24.94	0.217	26.28	0.123
Short+Long+Tree-rec	24.95	0.218	26.38	0.120
Short+Long+Tree-rec-partial	24.95	0.218	26.38	0.120
Oracle			28.77	

Table 6.11: *Oracle vs. real: English-German (TED)*

scoring of reordering paths could help improve the translation and reordering quality for that translation direction.

6.5 Manual Analysis

The evaluation of word reordering models in machine translation is a difficult task. In the previous section we have used the common automatic metric BLEU for measuring translation quality and the Kendall's τ metric for measuring reordering quality of the developed tree-based reordering model. Such automatic evaluations are designed for quick and reproducible assessment of quality improvements. For system development purposes this is a very valuable service. However, they always compare against a reference translation and are therefore prone to underestimate improvements when the generated translation deviates from the reference. Hence, we perform a detailed analysis of the tree-based reordering approach applied in a German-to-English phrase-based machine translation system. We compare the translation outputs of two translation systems applying reordering rules based on parts-of-speech and syntax trees on a sentence-by-sentence basis. For each sentence-pair we examine the global translation performance and classify local changes in the translated sentences. This analysis is applied to three data sets representing different genres.

6.5.1 Analysis

We perform an analysis of two translation outputs, one using a reordering model based on only word-level information, i.e. parts-of-speech, and one using word-level and sentence structure information, i.e. syntactic parse trees. We assess the translation quality and determine the types of improvements and degradations introduced by the structure-aware tree-based reordering model. This way we investigate whether the structural information in the reordering model indeed produces translations with better sentence structure compared to a reordering model using only word-level information. We analyze four different aspects in our comparison of two translation outputs for three different data sets.

6.5.1.1 Data

The three data sets used in our analysis represent different genres. The first data set are news texts, which are written in formal style. They typically consist of grammatically correct, but longer and more complex sentences. The second data set consists of human transcripts of TED talks[1]. This type of presentations are practiced performances, so the speakers hardly make mistakes and spontaneous speech artifacts such as repetitions or stuttering are very rare. The transcripts are edited in subtitle style resulting in a

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

more written form of sentences. The third data set consists of human transcriptions of lectures and talks recorded at a university. Even though obvious spontaneous speech artifacts are removed from the data, no further editing is performed. Consequently, the style resembles more that of actual speech than it is the case with TED talks. The data sets are described in detail in Chapter 5.

By examining those three types of data, which exhibit different text characteristics and vary in their degree of grammaticality, complexity and spontaneity, we want to assess the impact of the tree-based reordering model more thoroughly and find out how it performs in these different environments. For each of the data sets we analyzed between 100 and 166 sentence pairs.

6.5.1.2 Impact of Trees depending on Genre

We first analyze how much the translations differ when using a word-level compared to a structure-aware reordering model. The word-level reordering model only includes reordering rules based on parts-of-speech, whereas the structure-aware model additionally includes the reordering rules based on syntactic parse trees. The rest of the translation system is identical and only the reordering model is changed to produce the two translations. Hence, there might be sentences which remain unchanged. The first aspect of our analysis therefore considers the amount of sentences affected by the change of the reordering model and how this impact varies across the data sets representing different genres.

6.5.1.3 Global Sentence Performance

Motivated by the sometimes inconclusive results when measuring joint reordering and translation quality with automatic metrics , the second part of the analysis is a manual evaluation of two translation outputs for each of the three data sets. The evaluation consists in a pairwise comparison of the translation quality of the two translations, one produced using the part-of-speech-based reordering model and the other one applying the tree-based reordering rules in addition. For the analysis one set of sentences is presented at a time, consisting of the source sentence and the two translations without revealing the system which generated each translation. The presentation of the translations takes place in random order to ensure anonymity. Then the overall better translation is chosen allowing ties. This assessment of the global translation performance on sentence basis is performed on all three genres.

6.5.1.4 Local Phenomena

As the third part of the analysis, the changes introduced by the tree-based reordering rules are examined more thoroughly. Each change in the translated sentence is classified according to the three steps presented in Table 6.12. First, we determine whether it represents an improvement or a degradation of the translation quality. Then further classification is performed, defining the role of the changed word(s) in the sentence, either by its part-of-speech, its constituent role or whether it globally affects the subject-verb-object (SVO) structure[1]. Then a more fine-grained distinction according to the type of the change is carried out. Since verbs are our special concern when translating between German and English, for verbs we distinguish between improved/degraded position, insertion, deletion, substitution by an improved/degraded verb form or a different word choice. For most other changes, we only discriminate between insertion/deletion and position changes.

change	role in sentence	type of change
improvement	verb	insertion
degradation	adverb	deletion
	adjective	position
	noun	substitution
	negation	- word choice
	preposition	- word form
	compound	
	PP	
	NP	
	SVO structure	
	...	

Table 6.12: *Classes in the classification scheme*

We provide statistics for total amounts of improvements and degradations introduced by the tree-based reordering model for the three genres and analyze which types of words or sentence parts are prominently affected by the model.

[1] Since we are analyzing English translation output, we expect an SVO sentence structure.

6.5.1.5 Local Changes and Global Translation Performance

In the last part of the analysis we examine the correlation between local changes and the global translation performance on the sentence basis for the individual data sets. We investigate how individual improvements and degradations affect overall sentence performance and whether conclusions about sentence quality can be drawn when certain changes are observed.

6.5.2 Results

In this section we present the results of the analysis. We translated three data sets by applying two versions of the reordering model within a phrase-based translation system as described in Chapter 5. The first system uses only POS-based reordering and the other one uses POS-based and tree-based reordering together.

In Section 6.5.2.1 we give the statistics of the different data sets. Section 6.5.2.2 describes how much the data sets were affected by the tree-based reordering model compared to the POS-based reordering model. In addition, we draw the connection to the translation quality measured with an automatic metric. Afterwards, we present the results of the pairwise comparison of translation quality, which was performed manually. The fine-grained analysis is presented in Section 6.5.2.4, showing first the number of improvements and degradations introduced by the tree-based reordering and then a more detailed examination of the types of changes. The final section presents the analysis of the correlation between local changes and global sentence performance.

6.5.2.1 Data Statistics

We used three different data sets for our analysis. Table 6.13 shows statistics on the data, which is described in detail in Chapter 5.

Data set	size	type
News	3003	text
TED	1565	speech
Lectures	2300	speech

Table 6.13: *Overview and statistics on the data sets*

6.5.2.2 Impact of Trees depending on Genre

As expected the translation outputs are quite similar, since the only difference between the systems is the addition of tree-based rules. However, there is a observable difference in the impact of the tree-based rules depending on the genre of the data sets.

Data set	size	POS	+Tree	
News	3003	21.98	22.45	+0.47
TED	1565	30.73	30.87	+0.14
Lectures	2300	25.64	25.65	+0.01

Table 6.14: *Translation accuracy (BLEU)*

The automatic assessment of translation quality using the BLEU score (Papineni et al., 2002) are presented in Table 6.14. It can be seen that only for the News data set a measurable difference between the translation quality can be achieved by adding the tree-based rules. For the translation of TED talks and lectures, the automatic score does not improve much or even stays practically the same. It is to be noted that this automatic measurement represents the translation accuracy on the translated document as a whole.

In order to get a deeper insight into this genre-dependent behavior, we analyzed the impact of the tree-based model on the sentence level. Table 6.15 shows for each of the three data sets the number of changed sentences due to the tree-based rules in relation to the total number of sentences in the translated document. For the News data, the translation of 75.5% of the sentences in the test set is changed due to the introduction of the tree-based rules. In contrast, the translation of speech data, i.e. the TED talks and university lectures, is a lot less affected by the tree-based rules. Only 16.3 and 22.5% of the sentences exhibit a changed translation.

Data set	size	different	%
News	3003	2267	75.5
TED	1565	255	16.3
Lectures	2300	518	22.5

Table 6.15: *Impact of tree model*

A reason for this difference between written text and speech data may be due to their different textual characteristics. Written text tends to contain more complex

sentences, which is the types of sentences where the tree-based reordering model can exert its strengths best. In spoken performances, overly complex sentences structures are typically avoided in order to facilitate comprehension on the part of the audience. Shorter and less complex sentences can be addressed well with the POS-based reordering rules, which explains why often word orders proposed by the tree-based model are not chosen for translation.

In order to confirm this assumption we examine different aspects of the data that could give an indication of the complexity of sentences. First of all, sentence length and the number of punctuation marks could be an indicator for complexity, since this increases parsing difficulty and could lead to erroneous parse trees.

Data set	sentence length (avg.)		# punctuation per sentence	
	all	subset	all	subset
News	20.83	23.29	4.8	5.1
TED	16.29	25.00	3.9	5.7
Lectures	19.30	27.01	4.3	4.8

Table 6.16: *Analysis of textual complexity*

Table 6.16 shows the two aspects mentioned above: average sentence length and number of punctuation marks per sentence both for the subsets of affected sentences and all sentences of the three data sets. As expected, the average amount of words per sentence as well as the number of punctuation marks is highest in the News data set. For the speech data sets, lectures contain longer sentences and more punctuation marks due to the specialized content in the university setting. TED talks are more general, popular talks directed at a broader audience where the appropriate presentation style consists of shorter, concise sentences. When considering only the subset of sentences affected by the tree-based rules, we can see that both the average sentence length and the number of punctuation marks increase for all data sets. This corresponds with our expectation that longer and complex sentences are explicitly targeted by the tree-based rules. For the subset, where the tree rules lead to different translations, the sentence length for the speech data is even longer than for text data. The reason might be that for the same sentence length, the structure of a written text is more complex than for a speech text. Therefore, the tree rules are already more important for shorter sentences.

These results may explain the difference in the proportion of affected sentences for the different data sets shown in Table 6.15. The differences in automatic translation scores between data sets will also be related to this finding. Since a lot fewer sentences are changed in the speech data sets, the tree rules' influence on the whole document is lower and therefore less noticeable in the BLEU score.

In order to evaluate the impact of the tree-based rules on the translation quality without the bias of unchanged sentences, we calculate the translation accuracy on a subset of the original data set consisting of the changed sentences only. Table 6.17 shows the automatic translation scores for these subsets.

Data set	size	POS	+Tree	
News	2267	21.38	21.87	+0.49
TED	255	27.10	27.51	+0.41
Lectures	518	23.53	23.60	+0.07

Table 6.17: *Translation accuracy on subsets (BLEU)*

These new scores show that for the TED data it was indeed the case that the lower number of affected sentences led to a underestimation of the impact of the tree-based reordering on the automatically measured translation quality. For the News data, the impact was already obvious, since the bigger part of the sentences were already affected by the tree-based model. Excluding the remaining sentences from the automatic scoring did not change the score much. We can therefore argue that the tree-based reordering affects the translation of the TED talks positively in a similar way as the News data, whenever the application of the tree-based reordering rules results in a changed translation. However, the automatic translation score for the translation of lectures shows not much of a difference compared to the previous results in Table 6.14.

After investigating the impact of the tree-based reordering model in various ways, we examine the changed translation hypotheses manually to find out whether the change introduced by the tree-based reordering resulted in a better translation.

6.5.2.3 Global Sentence Performance

From all the sentences which were translated differently due to the tree-based reordering, we extracted sentences from each of the data sets for manual analysis. Table 6.18 shows the exact amount of sentences analyzed for each data set. For TED and News data, the first 100 and 165 of the changed sentences of the document were chosen. For the lecture data, 166 sentences were chosen for analysis by taking an even amount from each of the individual lectures.

Data set	size
News	165
TED	100
Lectures	166

Table 6.18: *Amounts of manually analyzed data*

We analyzed the global sentence performance by comparing the two translation hypotheses created using only POS-based rules and using POS and tree-based reordering rules together. Table 6.19 shows the results. We can see that in 55-64% of the cases, the system using tree-based rules produced a better translation, while the translation using only POS-based reordering was considered the better translation for 24 to 28% of the sentences. There are more tree wins for the speech data sets than for the News data. However, the amount of POS wins is bigger for the speech data, while the amount of ties is lower. This might be both due to the above mentioned easier structure of speech sentences and the mismatch of training and test data for the parser.

Data set	Tree win	tie	POS win
News	55.8	19.4	24.9
TED	64.0	8.0	28.0
Lectures	60.8	12.7	26.5

Table 6.19: *Manual sentence-level analysis (%)*

In contrast to the automatic evaluation, which only indicates an improvement on the TED and News talks, the manual evaluation shows that the translation quality is improved on all three data sets when using the tree-based reordering approach.

6.5.2.4 Local Phenomena

The previous section presented an analysis of the global sentence performance, considering each translated sentence as a whole. Now we investigate the local phenomena more thoroughly, i.e. the individual changes of words and structure between the two translation hypotheses. We identify the changed regions in each sentence pair and determine for each of the changes introduced by the tree-based system, whether it improves the translation quality or degrades it.

Data set	++	%	- -	%	total	per sentence
News	119	65.0	64	35.0	183	1.11
TED	92	70.2	39	29.8	131	1.31
Lectures	159	70.4	67	29.6	226	1.36

Table 6.20: *Local phenomena*

Table 6.20 shows the amounts of improvements (++) and degradations (- -) among the total number of changes in all analyzed sentences of each data set. The News data set includes the lowest number of changes per sentence. More changes per sentence can be found in the two speech data sets. Consequently, even though much less sentences are affected by the tree-based model in the speech data sets (16% and 22% vs. 75% of the sentences, cf. Table 6.15), more changes are introduced per sentence in the affected sentences (1.3 in speech vs. 1.1 in text data).

	++		
	News	TED	Lectures
substitution	25.2	23.9	30.2
word choice	20.2	19.6	23.3
word form	5.0	4.3	6.9
position	30.3	43.5	42.1
insertion	44.5	32.6	27.0
deletion	0.0	0.0	0.6
total	100.0	100.0	100.0

Table 6.21: *Local phenomena - types of improvements (%)*

Tables 6.21 and 6.22 show what types of changes can be discerned in the improvements and degradations, respectively. We differentiate between substitutions, insertions

and deletions of words as well as position changes. Substitutions include different word choice and changed tense or other morphological changes to the word form.

	News	- - TED	Lectures
substitution	46.9	51.3	22.4
word choice	39.1	38.5	20.9
word form	7.8	12.8	1.5
position	34.4	25.6	43.3
insertion	0.0	7.7	0.0
deletion	18.8	12.8	34.3
total	100.0	100.0	100.0

Table 6.22: *Local phenomena - types of degradations (%)*

As can be seen, there is again a difference between the text and speech data sets. For the News data, nearly half of the improving changes (44%) are insertions of words, i.e. words appear in the translation that were not translated before. The rest of the changes are substitutions, i.e. different word choices (25%) and improved word positions (30%). For the two speech data sets, the biggest share of the improvements affect the position (43 and 42%), while insertions make up a smaller portion of the improvements. Deletions typically do not have a positive effect on the translation.

Analyzing the types of negative changes (Table 6.22) shows that for News and TED data the main source of degradations is word substitutions, i.e. different word choices or word forms that change the translation quality for the worse. For the lectures it is the changed positions and deleted words that make up most of the negative changes. This might be a reason for the low BLEU improvement on the lecture test set observed in Table 6.17.

Tables 6.23 and 6.24 show the types of changes according to word classes and sentence constituents. Changes in word form, position, insertions and deletions related to a word class are analyzed. Different word choices leading to a better or worse translation are not taken into account. It can be observed that throughout all data sets the most affected word classes are verbs and adverbs. Others are nouns and pronouns as well as prepositions. Regarding sentence structure, the position of whole prepositional phrases is one of the more prominently affected parts of the sentence.

		++	
	News	TED	Lectures
verb	49	53	81
adverb	9	6	11
pronoun	0	7	5
noun	7	1	2
compound	2	0	3
determiner	3	0	1
adjective	1	0	0
preposition	8	1	2
conjunction	2	1	4
negation	1	0	1
interjection	0	0	1
PP	9	4	8
NP	1	1	2
SVO structure	3	0	0
clause	0	0	1
	95	74	122
word choice	24	18	37
total	119	92	159

Table 6.23: *Local phenomena - word classes (improvements)*

The main word classes affected by degradations of translation quality are similar to the improved word classes, as Table 6.24 shows. Although fewer degradations are introduced by the tree-based reordering model, the changes still mainly affect verbs, adverbs, nouns, pronouns, prepositions and prepositional phrases. As mentioned before, the main types of degradations are degraded position and erroneously removed words.

	News	- - TED	Lectures
verb	14	9	18
adverb	2	0	8
pronoun	3	5	2
noun	3	0	4
compound	4	0	4
adjective	1	1	0
preposition	4	1	3
conjunction	0	2	3
interjection	0	0	1
PP	3	3	2
NP	3	0	5
SVO structure	1	1	1
clause	0	0	1
object	1	2	0
subject	0	0	1
	39	24	53
word choice	25	15	14
total	64	39	67

Table 6.24: *Local phenomena - word classes (degradations)*

6.5.2.5 Local Changes and Global Translation Performance

How are local changes correlated with the global translation performance? Table 6.25 shows how many of the positive changes in all word classes and in the verb class shown in Table 6.23 above were observed in a Tree win or POS win sentence. From these numbers we can draw the conclusion that between 90.8 and 96.2% of the improving changes in all classes result also in a globally improved translation quality. When we examine only the verbs, the tendency is similar. Between 83.7 and 95.1% of the verb-related improvements stem from a sentence produced by the tree-based reordering model and represent an improvement in translation quality over the sentence produced by the POS-based reordering model.

Table 6.26 shows the correlation between degradations and global sentence quality. We have already established that fewer negative changes than positive changes are introduced by the tree-based system. The previous table might indicate that a negative

		++	
	News	TED	Lectures
all classes			
Tree wins	90.8	94.6	96.2
POS wins	5.0	5.4	4.4
verbs			
Tree wins	83.7	92.5	95.1
POS wins	12.2	7.5	6.2

Table 6.25: *Local vs. global (improvements) (%)*

change should also correspond more likely with a worse translation quality of the output of the translation system using the tree-based reordering output, i.e. a POS win. When analyzing all word and constituent classes, the correlation between negative changes and POS wins is between 70.3 and 80.6%. For the verbs, the correspondence is a little higher, between 71.4 and 88.9%. However, the correlation is not as high as for positive changes with improved translation quality.

		- -	
	News	TED	Lectures
all classes			
Tree wins	17.2	20.5	19.4
POS wins	70.3	79.5	80.6
verbs			
Tree wins	14.3	11.1	16.7
POS wins	71.4	88.9	83.3

Table 6.26: *Local vs. global (degradations) (%)*

Hence, we can conclude that local improvements introduced by the tree-based model will most likely coincide with an overall better translation quality of that given sentence. Local degradations are not necessarily to correspond with a degraded translation quality of the whole sentence, although degradations in verbs have a more severe influence on the translation quality.

6.6 Translation Examples

This section shows examples for improved translations achieved by the tree-based reordering model. Example 6.2 shows how the translation of the challenging sentence presented in Chapter 3 is improved by adding the tree-based rules. We can see that using tree constituents in the reordering model indeed addresses the problem of verb particles and especially missing verb parts in German.

Source:	*..., nachdem ich eine Weile im Internet gesucht habe.*
Gloss:	`..., after I a while in-the Internet searched have.`
POS Reordering:	*... as I have for some time on the Internet.*
+Tree Reordering:	*... after I have looked for a while on the Internet.*
Reference:	*... after browsing the web for a while.*

Example 6.2: Recovering missing verbs in translation output

Example 6.3 shows that the tree-based rules can also address the problem of verb prefixes mentioned in Chapter 3. With the help of the tree-based reordering rules, it is possible to relocate the separated prefix of German verbs and find the correct translation. The verb *vorschlagen* consists of the main verb stem (VFIN) *schlagen* (here conjugated as *schlägt*) and the prefix (PTKVZ) *vor*. Depending on the verb form and sentence type, the prefix must be separated from the main verb and is located in a different part of the sentence. The two parts of the verb can also have individual meanings, *beats* is a correct translation for *schlagen* and *vor* as a preposition can be translated as *before, ago, in front of*. The translation of the verb stem were correct if it were the full verb. However, in this context, not recognizing the separated prefix and ignoring it in translation, corrupts the meaning of the sentence. With the help of the tree-based rules, the dependency between the main verb and its prefix is resolved and the correct translation can be produced.

Source:	*Die RPG Byty schlägt ihnen in den Schreiben eine Mieterhöhung von ca. 15 bis 38 Prozent vor.*
Gloss:	`The RPG Byty proposes-VFIN them in the letters a rent increase of ca. 15 to 38 percent proposes-PTKVZ`
POS Reordering:	*The RPG Byty beats them in the letter, a rental increase of around 15 to 38 percent.*
+Tree Reordering:	*The RPG Byty proposes them in the letters a rental increase of around 15 to 38 percent.*
Reference:	*RPG Byty proposes to increase rent by 15 to 38 percent in these letters.*

Example 6.3: Reordering and successful translation of verb prefix

6.7 Conclusions

This chapter presented a reordering model making use of structural information provided by syntactic parse trees in order to produce better sentence structure in phrase-based machine translation output. We performed experiments and analyses on several languages and data sets and addressed the potential of the source reordering approach with oracle experiments. A manual evaluation investigated the changes introduced by the syntactic tree-based reordering model more in detail in a sentence-wise comparison of translation outputs on three data sets.

6.7.1 Tree-based Reordering Model

We have presented a reordering method based on syntactic tree constituents to model long-range reordering in phrase-based machine translation more reliably. We combined the reordering methods addressing different linguistic abstraction levels. Experiments on German-English and German-French translation showed that the best translation quality can be achieved by combining part-of-speech-based and tree-based rules. Adding a lexicalized reordering model increased the translation quality even further. In total we reached up to 0.7 BLEU points of improvement by adding tree-based and lexicalized reordering compared to only part-of-speech-based rules. Up to 1.1 BLEU points were gained over to a baseline system using a lexicalized reordering model and up to 1.4 BLEU points improvement was achieved when using a combination of tree-based, part-of-speech-based and lexicalized reordering over a baseline using no dedicated reordering model.

6.7.2 Oracle Reordering

In a second line of experiments we have analyzed the performance of the tree-based reordering model using oracle experiments. The experiments were conducted on German-to-English and English-to-German translation of News texts and TED talks.

The first set of experiments showed that source sentence reordering is a very promising approach. By translating an optimally reordered source sentence, an improvement of the translation performance by up to 6.2 BLEU points is possible.

This upper bound was compared to the oracle path in the reordering lattices encoding the reordering variants produced by different types of reordering rules. The results led to the conclusion that the restriction of the search space using the reordering lattices approximates the oracle reordering better when more complex and complementary reordering rules are used. However, the best oracle path and the oracle reordering are still far apart, leaving a lot of potential for discovering better reordering rules that approximate the oracle reordering even better. Both for German-English and English-German, a gap of 2.5 to 3.8 BLEU points remains until the best possible translation result can be reached. As a consequence, one direction of promising research is to expand the search space further to include reordering variants that better approximate the optimally reordered source sentence.

Comparing the decoder path with the oracle path showed that for German-to-English translation the path chosen by the decoder is quite close to the oracle path, both in terms of translation quality and reordering distance. The two paths are only 0.2 and 0.3 BLEU points apart. Hence, the current models used in the machine translation system are able to find almost the best source word order that exists in the search space. For English-to-German translation, however, finding the best path in the reordering lattice seems to be more difficult. A gap of 0.7 and 1.4 BLEU remains between the actual performance and the oracle path translation. We can conclude that at least for English-to-German translation a better ranking of the different reordering possibilities in the search space seems to hold a promising perspective for future research.

6.7.3 Manual Analysis

In addition to the automatic evaluation of the tree-based reordering approach, an in-depth analysis was performed for German-to-English translation. We examined the changes in the translation output introduced by the tree-based reordering rules compared to the part-of-speech-based reordering rules. We compared the results on three data sets which differ in genre and topic.

The findings of the detailed evaluation have shown that the tree-based reordering approach helps produce output of an improved translation quality on all three data sets. The impact of the tree-based reordering model is higher on data that consists of well structured, grammatically correct texts, while fewer sentences were affected for the two speech data sets. Taking only the affected sentences into account, the translation quality as measured with the automatic metric BLEU behaved similarly on the News and the TED data.

The manual evaluation of sentence-level translation quality confirmed consistent improvements by the tree-based reordering model throughout all three data sets. A similar behavior on the three data sets can also be reported for the local improvements in the sentence which include translations of words which were removed from the translation before as well as improved word and constituent positions in the translated sentence. As intended in the design of the tree-based reordering model, verbs are the main cause for local improvements. We observed a high correlation between local improvements in the sentence and an overall better sentence quality, while local degradations not necessarily lead to a worse translation on the sentence level.

7

Syntactic Structure for Translation Disambiguation

Ambiguity of words is a big challenge for all natural language processing tasks. Already within the same language, words can be ambiguous with regard to their part-of-speech (***can**, n. - **can**, v.*), word sense (***bank**, n., financial institution - **bank**, n., side of a river*) or what they are referring to in the given context (*The monkey eats the banana. **It** is brown.*). For translation, such ambiguities pose an additional difficulty. Unless the very same ambiguity exists in the target language, the ambiguity needs to be resolved in order to generate the correct translation. When translating into German, for example, depending on the correct part-of-speech, word sense and antecedent in the sentence, the translation for each of those examples is a different one.

The word(s) indicating which is the correct word sense or antecedent for an ambiguous word in a given context, could occur in a more distant part of the sentence. That means long-range dependencies need to be considered in order to generate the correct translation. We propose a discriminative framework for modeling these dependencies that allows utilizing any conceivable set of features for predicting the correct translation. We show the potential of this approach in detail on the third type of ambiguity mentioned above: The translation of pronouns, which is conditioned on the translation of the antecedent they refer to, since the pronoun in the target language needs to share the morphological properties of the antecedent in the target language.

An approach to explicitly performing anaphora resolution to uncover the pronoun-antecedent relationship for pronoun translation disambiguation was carried out in a related project described in Appendix A (Weiner, 2014). Their experiments motivated the present work, however the approach was adapted in the following ways: While

Weiner (2014) focus only third person pronouns, we include all pronouns and also take translations into other word categories into account. In order to allow for a more comprehensive exploration of the source discriminative word lexicon approach we apply it for translation disambiguation for all words and perform a separate evaluation of the performance on pronouns. We further evaluate it on another difficult agreement task, the agreement of subject and verb in a sentence.

The translation examples given in Chapter 3 have shown that a state-of-the-art machine translation system struggles with these particular kinds of linguistic requirements. Hence, we believe our approach can provide a comprehensive solution for many of these challenges where long-range dependencies have to be met in order to ensure congruency of linguistic features. In the remainder of this chapter we describe the setting, development and evaluation of a disambiguation model using structural features for translation prediction of two particular linguistic challenges, the translation of pronouns and the generation of morphological agreement in a morphologically rich target language. The work presented in the following is an extended version based on Herrmann et al. (2015).

7.1 Pronoun Translation

When translating pronouns, it is necessary to produce the correct pronoun-antecedent agreement in the translation. Number and gender of the generated pronoun need to agree with number and gender of the previously mentioned noun it refers to. That means that a pronoun cannot simply be translated in isolation, but the context of previously mentioned nouns needs to be taken into account. The referring noun can be located in the same sentence or in a previous sentence. Weiner (2014) performed an analysis showing that the location of the referring pronoun is dependent on the type of data. In News data, inter-sentence and intra-sentence anaphora occur in equal shares (Table A.1), while in TED data, the referring noun more often occurs in a previous sentence than within the same sentence (75% vs. 25% in Table A.1).

7.1.1 Analysis

We performed an analysis of how pronouns are translated for two translation directions, German-to-English and English-to-German, and two genres, TED talks and News texts. The analysis is based on an automatic word alignment between source text and human reference translation. For each of the involved languages, a set of pronouns was defined consisting of first, second and third person pronouns in nominative, genitive, dative

and accusative case for German (Eisenberg et al., 2005) and subjective, possessive, objective and reflexive pronouns for English (Crystal, 2003), in singular and plural. Tables 7.1 and 7.2 show an overview of those pronouns in the two languages.

Person	Number	Gender	Nom.	Gen.	Dat.	Acc.
1st	Singular	-	ich	mein, -e, -es, -er, -em, -en	mir	mich
2nd		-	du	dein, -e, -es, -er, -em, -en	dir	dich
3rd		Masculine	er	sein, -e, -es, -er, -em, -en	ihm	ihn
		Feminine	sie	ihr, -e, -es, -er, -em, -en	ihr	sie
		Neuter	es	sein, -e, -es, -er, -em , -en	ihm	es
1st	Plural	-	wir	unser, -e, -es, -er, -em, -en	uns	uns
2nd		-	ihr	euer, eure, -es, -er, -em, -en	euch	euch
3rd		Masculine				
		Feminine	sie	ihr, -e, -es, -er, -em, -en	ihnen	sie
		Neuter				

Table 7.1: *German pronouns*

Person	Number	Gender	Subj.	Poss.	Obj.	Refl.
1st	Singular	-	I	my, mine	me	myself
2nd		-	you	your, yours	you	yourself
3rd		Masculine	he	his	him	himself
		Feminine	she	her, hers	her	herself
		Neuter	it	its	it	itself
1st	Plural	-	we	our, ours	us	ourselves
2nd		-	you	your, yours	you	yourselves
3rd		Masculine				
		Feminine	they	their, theirs	them	themselves
		Neuter				

Table 7.2: *English pronouns*

The analysis of how these pronouns are translated into each other, was done in the following way: For all pronouns in the source text, the aligned words in the target text were extracted. If the aligned word is not in the set of target language pronouns, it was assigned to the class **other**. Tables 7.3(a) and 7.3(b) present the distribution

of pronoun-to-pronoun translations for English-German translation of News texts and TED talks and Tables 7.4(a) and 7.4(b) present the distribution of pronoun-to-pronoun translations for translation of News texts and TED talks for German-to-English translation. In the tables, the columns and rows of German possessive pronouns *mein, dein, sein, ihr, unser, euer* also subsume the respective declined word forms *meine, meines, meiner, ...* as shown in Table 7.1.

All four tables show an approximation of the expected distribution along the diagonal. However, some scattering can be observed which is due to ambiguous pronouns. On the English side there is the second person pronoun *you*, which can be both singular or plural. Similarly, the German *sein* can be both third person singular masculine and neuter and the very ambiguous pronoun *ihr* with its many morphological variations can represent the third person singular feminine, the third person plural genitive for all three genders and the second person plural nominative form. Since these cases cannot be distinguished if only the surface form of the pronouns is considered, we duplicate the respective rows in the table for the sake of completion. For example, the same row of the pronoun *sie* in Tables 7.4(a) and 7.4(b) occurs in two places, once as the third person singular feminine and again as third person plural. The ambiguities are clearly visible in the tables by clusters that deviate from the diagonal.

Another prominent deviation from the diagonal are the translations categorized as **other**. There is a remarkable amount of occurrences where the translation of a pronoun is not a pronoun in the target language. This subsumes null alignments where no target word is generated. Translations classified as **other** amount to 20 or even 50% of the target words, depending on translation direction and text genre. Although this might be partially due to errors in the automatically generated word alignment, the numbers are too high to be discarded as noise. Example 7.1 shows example sentences for German-English and English-German translation where a source pronoun is aligned to a word class other than pronoun or even unaligned in the target language sentence.

Source:	*[...] maybe even dancing with it.*
Reference:	*[...] und vielleicht sogar damit zu tanzen.*
Source:	*[...] sie zu vermeiden , noch sollten wir sie unter den Teppich kehren [...]*
Reference:	*[...] not [...] something we want to avoid or sweep under the rug [...]*

Example 7.1: Pronoun translation as **other**

Evidently, a good portion of pronouns is not translated as pronouns. This analysis confirms our decision not to restrict the prediction in the target language to pronouns in the translation prediction model, but to allow other translations as well.

	ich	mein	mir	mich	du	dein	dir	dich	er	sein	ihm	ihn	sie	ihr	es	sein	ihm	sich	wir	unser	uns	ihr	euer	euch	sie	ihr	ihnen	other
I	94	0	5	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
my	0	17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
mine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
me	0	0	9	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
myself	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
you	0	0	0	0	8	0	2	0	0	0	0	0	18	2	0	0	0	1	1	0	0	2	0	0	18	2	2	37
your	0	0	0	0	0	4	0	0	0	3	0	0	0	8	0	3	0	0	0	0	0	8	0	0	0	8	1	12
yours	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
yourself	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
he	1	0	0	0	0	0	0	0	253	2	9	1	1	0	5	2	9	0	0	0	0	0	0	0	1	0	0	50
his	0	0	0	0	0	0	0	0	9	156	4	1	0	2	0	156	4	0	0	0	0	2	0	0	0	2	0	41
him	0	0	0	0	0	0	0	0	6	0	12	18	1	3	0	0	12	0	0	0	0	3	0	0	1	3	1	12
himself	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	9
she	0	0	0	0	0	0	0	0	0	0	0	0	68	1	1	0	0	0	0	0	0	1	0	0	68	1	0	18
her	0	2	0	0	0	0	0	0	0	1	0	0	6	48	0	1	0	0	0	0	0	48	0	0	6	48	0	18
hers	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
herself	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3
it	0	0	1	0	0	0	0	0	18	0	0	5	54	1	163	0	0	8	3	0	0	1	0	0	54	1	0	216
its	0	0	0	0	0	0	0	0	0	78	0	0	0	46	1	78	0	4	0	0	0	46	0	0	0	46	0	66
itself	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	9
we	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86	0	10	0	0	0	0	0	0	19
our	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	27	0	0	0	0	0	0	0	5
ours	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
us	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	20	0	0	0	0	0	0	13
ourselves	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	5
you	0	0	0	0	8	0	2	0	0	0	0	0	18	2	0	0	0	1	1	0	0	2	0	0	18	2	2	37
your	0	0	0	0	0	4	0	0	0	3	0	0	0	8	0	3	0	0	0	0	0	8	0	0	0	8	1	12
yours	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
yourselves	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
they	0	0	0	0	0	0	0	0	6	0	0	1	165	4	4	0	0	5	1	0	0	4	0	0	165	4	7	81
their	0	0	0	0	0	0	0	0	0	2	0	0	3	124	0	2	0	1	0	0	0	124	0	0	3	124	0	58
theirs	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0
them	0	0	0	0	0	0	0	0	0	0	0	1	22	1	1	0	0	0	0	0	0	1	0	0	22	1	18	33
themselves	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	7	0	0	0	0	0	0	1	0	0	3

(a) News

	ich	mein	mir	mich	du	dein	dir	dich	er	sein	ihm	ihn	sie	ihr	es	sein	ihm	sich	wir	unser	uns	ihr	euer	euch	sie	ihr	ihnen	other
I	487	6	14	6	1	0	0	0	0	0	0	0	0	0	6	0	0	0	1	0	0	0	0	0	0	0	0	39
my	4	93	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
mine	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
me	7	3	28	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
myself	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
you	1	0	0	0	8	0	1	3	0	0	0	0	231	4	3	0	0	0	5	0	0	4	0	0	231	4	35	137
your	0	0	0	0	0	5	0	0	0	5	0	0	1	17	0	5	0	0	0	2	0	17	1	0	1	17	0	7
yours	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
yourself	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	5
he	0	0	0	0	0	0	0	0	94	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	7
his	0	0	0	0	0	0	0	0	3	25	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	3
him	0	0	0	0	0	0	0	0	1	0	2	14	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
himself	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
she	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
her	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0	0	2	0	0
hers	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
herself	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
it	0	0	0	0	0	0	0	0	20	1	2	16	65	1	209	1	2	0	0	0	0	1	0	0	65	1	0	194
its	0	0	0	0	0	0	0	0	0	4	0	0	0	11	0	4	0	0	0	0	0	11	0	0	0	11	0	8
itself	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	5
we	0	0	0	0	0	0	0	0	0	0	0	0	1	0	15	0	0	0	453	0	11	0	0	0	1	0	0	45
our	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	1	0	0	2	61	4	0	0	0	0	0	0	10
ours	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
us	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	2	44	0	0	0	0	0	0	7
ourselves	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	6	0	0	0	0	0	0	6
you	1	0	0	0	8	0	1	3	0	0	0	0	231	4	3	0	0	0	5	0	0	4	0	0	231	4	35	137
your	0	0	0	0	0	5	0	0	0	5	0	0	1	17	0	5	0	0	0	2	0	17	1	0	1	17	0	7
yours	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
yourselves	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
they	0	0	0	0	0	0	0	0	3	0	0	0	157	0	4	0	0	1	1	0	0	0	0	0	157	0	0	39
their	0	0	0	0	0	0	0	0	0	8	0	0	2	24	0	8	0	1	0	0	0	24	0	0	2	24	1	8
theirs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
them	0	0	0	0	0	0	0	0	0	0	0	0	19	1	0	0	0	0	0	0	0	1	0	0	19	1	20	26
themselves	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	6

(b) TED

Table 7.3: *Pronoun translation distribution: English-German*

	I	my	mine	me	myself	you	your	yours	yourself	he	his	him	himself	she	her	hers	herself	it	its	itself	we	our	ours	us	ourselves	you	your	yours	yourselves	they	their	theirs	them	themselves	OTHER
ich	0	2	0	10	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	545
mein	0	95	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	45
mir	0	2	2	26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36
mich	0	1	0	29	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21
du	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1
dein	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
dir	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
dich	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
er	0	0	0	0	0	0	0	0	0	95	3	1	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	9
sein	0	0	0	0	0	0	5	0	0	0	27	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	5	0	0	0	8	0	0	0	58
ihm	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ihn	0	0	0	0	0	0	0	0	0	1	0	14	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
sie	0	0	0	0	0	32	0	0	0	0	0	0	0	2	0	0	0	58	0	0	0	0	0	0	0	32	0	0	0	160	1	0	25	0	50
ihr	0	0	0	0	0	3	3	0	0	0	0	0	0	2	0	0	1	11	0	0	0	0	0	0	0	3	3	0	0	0	22	0	0	0	9
es	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	211	0	0	12	0	0	0	0	4	0	0	0	3	0	0	0	0	178	
sein	0	0	0	0	0	0	5	0	0	0	27	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	5	0	0	0	8	0	0	0	58
ihm	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
sich	0	0	0	0	0	4	0	0	1	0	0	0	2	0	0	0	1	1	4	0	0	0	0	0	4	0	0	0	1	0	0	1	6	82	
wir	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	448	2	0	18	1	4	0	0	0	0	0	0	0	0	0	46
unser	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	61	1	1	1	0	2	0	0	0	0	0	0	0	0	12
uns	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	8	3	0	44	7	1	0	0	0	0	0	0	0	0	0	26
ihr	0	0	0	0	0	3	3	0	0	0	0	0	0	2	0	0	1	11	0	0	0	0	0	0	3	3	0	0	0	22	0	0	0	9	
euer	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2
euch	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sie	0	0	0	0	0	32	0	0	0	0	0	0	0	2	0	0	0	58	0	0	0	0	0	0	0	32	0	0	0	160	1	0	25	0	50
ihr	0	0	0	0	0	3	3	0	0	0	0	0	0	2	0	0	1	11	0	0	0	0	0	0	3	3	0	0	0	22	0	0	0	9	
ihnen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	2

(a) News

	I	my	mine	me	myself	you	your	yours	yourself	he	his	him	himself	she	her	hers	herself	it	its	itself	we	our	ours	us	ourselves	you	your	yours	yourselves	they	their	theirs	them	themselves	OTHER
ich	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	58
mein	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
mir	0	1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
mich	0	0	0	6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
du	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	10
dein	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
dir	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
dich	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
er	0	0	0	0	0	0	0	0	0	145	6	5	0	1	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	72
sein	0	0	0	0	0	0	3	0	0	1	141	0	0	0	0	0	0	0	75	0	0	0	0	0	0	0	3	0	0	0	3	0	0	0	138
ihm	0	0	0	0	0	0	0	0	0	7	4	11	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
ihn	0	0	0	0	0	0	0	0	0	1	1	11	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	15
sie	0	0	0	0	0	1	0	0	0	1	0	2	0	43	4	0	1	36	0	0	0	0	0	0	0	1	0	0	0	121	2	0	13	0	144
ihr	0	0	0	0	0	1	0	0	0	0	2	1	0	2	43	0	0	2	44	0	0	0	0	0	0	1	0	0	0	5	112	1	0	0	113
es	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	101	1	0	2	0	0	0	0	0	0	0	0	5	0	0	0	0	193
sein	0	0	0	0	0	0	3	0	0	1	141	0	0	0	0	0	0	0	75	0	0	0	0	0	0	0	3	0	0	0	3	0	0	0	138
ihm	0	0	0	0	0	0	0	0	0	7	4	11	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
sich	0	0	0	0	0	1	0	0	1	0	1	0	9	0	0	0	1	6	3	4	0	0	0	0	1	0	0	0	5	1	0	0	7	307	
wir	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	50	0	0	2	0	0	0	0	0	0	0	0	0	0	23
unser	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	13
uns	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	15	0	0	0	0	0	0	0	0	0	0	21
ihr	0	0	0	0	0	1	0	0	0	0	2	1	0	2	43	0	0	2	44	0	0	0	0	0	0	1	0	0	0	5	112	1	0	0	113
euer	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
euch	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
sie	0	0	0	0	0	1	0	0	0	1	0	2	0	43	4	0	1	36	0	0	0	0	0	0	0	1	0	0	0	121	2	0	13	0	144
ihr	0	0	0	0	0	1	0	0	0	0	2	1	0	2	43	0	0	2	44	0	0	0	0	0	0	1	0	0	0	5	112	1	0	0	113
ihnen	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	11	0	13

(b) TED

Table 7.4: *Pronoun translation distribution: German-English*

7.2 Subject-Verb Agreement in Translation

Morphological agreement of the grammatical features of subject and verb in the sentence is a requirement for both German and English. Ensuring this type of agreement to be passed on during translation does not seem to be such a problem. For the most part, singular nouns tend to be translated as singular nouns, as do singular verbs. Hence, the agreement should be transferable by just individually translating subject and verb. However, the difficulty lies first in the dissimilarly distributed morphological features of the verbs in source and target language. Table 7.5 illustrates this by contrasting the conjugation of the English verb *live* with the German verb *leben*. On the German side, four out of the six grammatical persons (first, second, third person in singular and plural) correspond to a distinct surface form of the verb. On the English side however, only two distinct verb forms exist in the present tense. The third person singular has a separate surface form, while all other grammatical persons share the same form. In the past tense, only one surface form exists for all persons. Second, in case the subject is a pronoun, we encounter the aforementioned difficulties of pronoun translation. And third, as for pronoun-antecedent agreement, the distance between the involved parties might span several words within the sentence length such that the dependency cannot easily be established. A machine translation system additionally suffers from the limited context that can be taken into consideration during translation. Example 7.2 shows how long the distance between subject and verb can be in a sentence. For German, this distance can get very long, while in English subject and verb are mostly only separated by adverbs.

		German			English		
Person	Number	Subject	Verb		Subject	Verb	
			present	past		present	past
1st		ich	leb-e	**lebt-e**	I	**live**	**live-d**
2nd	Singular	du	leb-st	lebt-est	you	**live**	**live-d**
3rd		er, sie, es	**leb-t**	**lebt-e**	he, she, it	live-s	**live-d**
1st		wir	**leb-en**	**lebt-en**	we	**live**	**live-d**
2nd	Plural	ihr	**leb-t**	lebt-et	you	**live**	**live-d**
3rd		sie	**leb-en**	**lebt-en**	they	**live**	**live-d**

Table 7.5: *Verb conjugation in German and English*

German: *Und wir fanden heraus , dass es in der Tat einen Zusammenhang gab.*
English: *Well it obviously is not.*

Example 7.2: *Long-distance dependencies in German and English*

7.3 Source Discriminative Word Lexicon

We implement translation disambiguation as a prediction task. The prediction is motivated by the discriminative word lexicon (Niehues and Waibel, 2013). The discriminative word lexicon (DWL) operates on the target side and learns to predict for each target word whether it should occur in a given target sentence. The source discriminative word lexicon (SDWL) operates on the source side. For every source word a classifier is trained to predict its translation in the given sentence. We perform a multi-class classification task by identifying for every source word the 20 most frequent translations to be the classes we want to predict. We define the translations to be the aligned word(s) in the respective target sentence. All target language words that occur less often than the 20 most frequent words are assigned to one class, called **other**. Alignments to the NULL word on the target side are treated in the same way as if NULL were a word. Hence, NULL can form a class of its own if NULL alignments occur often enough to be part of the 20 classes, otherwise they are included in the **other** class. Then we train 20 classifiers for the source word and perform one-against-all classification. All sentence pairs where the source word occurs in the source sentence are selected as training examples for each of the 20 classifiers. The sentence pairs are divided into positive and negative training examples. Those sentence pairs, where the respective aligned word on the target side belongs to the current class, are positive examples. All other sentence pairs, where a different target word is aligned, represent negative training examples for that class. We use maximum entropy classification provided by the MegaM package[1] for training and applying the classifiers. The SDWL consists of a training and a prediction phase. The maximum entropy models for the individual classes of each source word are trained based on the given set of features extracted from the source sentence and the correct class of each training example. For the prediction, the test data is first separated into words. For each word the features are extracted from the source sentence it stems from. Then all the binary maximum entropy models for the individual classes are applied and each of them produces a prediction. The multi-class prediction output corresponds to the class with the highest prediction probability.

[1] `http://www.umiacs.umd.edu/~hal/megam/`

7.3.1 Structural Features

The training examples and test data for the classifiers are represented by a set of features and the class this example belongs to. We experiment with different types of features representing the structure of a sentence to varying degrees.

7.3.1.1 Bag-of-Words

A straight forward way to represent the source sentence for this classification task is to use the bag-of-words approach. The sentence is represented simply by the words it contains, however without information about their order and with every word only occurring once. This is the least structural informative feature which does not provide any knowledge about the sentence beyond the mere existence of the words in it.

7.3.1.2 Context

The context feature adds structural information about the preceding and succeeding words of the modeled source word in the sentence. In addition to the context words themselves, their position is encoded in the feature such that the same word occurring at a different position (relative to the source word in question) would result in a different feature. We include up to six context words, three on each side of the source word. Hence, this feature type provides structural information by means of sequential order within a limited context.

7.3.1.3 Dependency Relations

The feature contributing the most information about the sentence structure is based on the relations between the source sentence words in a dependency tree. In order to obtain the dependency relations, we extract a dependency tree from a constituency parse tree using the Stanford Parser (Klein and Manning, 2002, 2003). Then we include the dependency relations between the source word and its parent and children in the dependency tree as features. That means, we form a feature consisting of the governance relation (parent or child of the source word), the dependency relation type (from the set of dependency relations described in de Marneffe and Manning (2008) e. g., nsubj, dobj, vmod, ...) and the connected word itself. This type of feature allows to capture structure by means of semantic dependencies that can range over longer distances in the sentence, but are relevant due to the semantic connection to the current source word. An example for the features for the word *it* in a given sentence is presented in Example 7.3.

Sentence:		*Well it obviously is not.*
Features:	bag-of-words	not is it obviously well .
	context	−1_well +1_obviously +2_is
	dependency	dep_parent_nsubj_is

Example 7.3: *Representation of the source word "it" by the different features*

7.3.2 Feature Representation

We compare two methods of feature representation: word IDs and word vectors.

7.3.2.1 Word IDs

When representing words by word IDs, we use the source vocabulary size V_{source} as the dimension of the feature space, a word's ID in the vocabulary as a feature and we set the feature to 1 if it is used in the example. All other features are set to 0. For accommodating the context features (**context**), we extend the feature space such that $V_{context} = c * V_{source}$ where c equals the size of the context. Each position of a word in the context hence has its own range in the feature space, and words in different context positions can be distinguished accordingly. The features representing dependency relations (**dep**) are included in a similar fashion. Again, a new feature space is defined as $V_{dep} = d * V_{source}$ where d equals the size of the inventory of all dependency relations, where parent and child relations count separately. The feature types can be combined by simply concatenating the individual feature spaces. That means when all three types of features are used the size of the feature space amounts to $V_{source} + V_{context} + V_{dep}$. It is obvious, that with this strategy for design the feature space grows quite big, possibly leading to data sparseness problems. In order to reduce dimensions, the representation via word vectors seemed an appropriate measure.

7.3.2.2 Word Vectors

The word vectors for feature representation are generated using word2vec (Mikolov et al., 2013) with the number of dimensions set to 100. That means each word is represented by a 100-dimensional vector. However, it is not straight forward how multiple words should be expressed in this representation, so that the representation by word vectors is not applied for the bag-of-words features, but only for the context and dependency features. In case of the vector representation of the context features

(**contextVec**), each position in the context words receives its own range in the feature space. Hence, the size of the feature space equals to $V_{contextVec} = c * dim$, where c is the context size and dim the dimension of the vector representation. This amounts to a significant reduction compared to $V_{context}$ used in the representation method via word IDs. The feature space for dependency relations using word vectors (**depVec**) equals to $V_{depVec} = d * dim$ with d being the inventory of dependency relations. Compared to V_{dep}, this amounts again to a huge reduction. In addition to the **depVec** feature, further variants of the dependency feature are compared:

parentDepVec
: For this feature, only the dependency relation to the parent word is represented in vector representation.

parentWordVec
: This feature consists of the vector representation of the parent word and an additional binary feature that is 1 if the parent word is the root of the dependency tree.

parentWordVec+DepRel
: In addition to the **parentWordVec** feature, the dependency relation to the parent word is encoded as a vector.

As for the word-based features, word vector features can be combined by concatenation of feature spaces.

7.3.3 Integration of SDWL Predictions

In order to integrate the individual translation predictions into a machine translation system we use the prediction probabilities for individual words to produce scores for whole sentences. The combination of individual translation predictions for words into a sentence score is explained in the following. These scores are then used in N-best list re-ranking as described in Chapter 5.

7.3.3.1 SDWL-based Re-ranking Scores

For each of the translation hypotheses in the N-best list, we generate a sentence score based on the translation predictions for the individual words in the sentence. We compare four methods to combine the individual word scores into a sentence score for a particular translation hypothesis.

Absolute number of predicted words in the sentence

All the words in the translation output produced by the translation system are compared against our predicted translations for the given source words. We use the alignment information from the phrase pairs used during decoding to know which target word is produced from which source word. If the predicted translation of the source word is the same as its translation in the hypothesis, we increase the sentence score, otherwise not. That means we count the number of word translations in the sentence that coincide with the predicted translations by the translation prediction model. If the translated word in the sentence is not one of the most frequent translations assigned to an individual class and the predicted translation is **other**, this is also counted as a match.

Relative number of predicted words in the sentence

As an alternative score we again count the number of words in the translation hypothesis that coincide with the predicted translation. This number of matches is then divided by the total number of target words generated by the source words according to the alignment.

Sum of prediction probabilities for the words in the sentence

The third type of score takes all words into account whether they coincide with the prediction or not. We do not just look up the prediction with the highest probability, but all the predictable words for a given source word and their prediction probabilities. Then we sum up the prediction probabilities of all the words that were used in the hypothesis.

Rank of the words in the sentence according to prediction rank

Instead of summing up the prediction probabilities of the words in the hypothesis, we sum up the ranks of the words according to their prediction probability. That means, the highest scoring predicted translation is equivalent to rank 0, the translation with the second highest prediction probability equals rank 1, and so forth. Consequently, if the hypothesis is composed only from words that are also the predicted translations, the sentence would get the score 0. The higher the sentence score, the more the hypothesis diverges from the translation predictions.

All these scores were both used individually and collectively as additional sentence scores for N-best list re-ranking, in order to find out which of them are most beneficial for judging translation quality.

7.4 Results

This section presents the results of the translation prediction model tested on English-to-German translation of TED talks. First, we will show that the prediction accuracy improves when applying the proposed set of structural features. In addition, the translation quality can be improved when using the translation predictions for N-best list re-ranking to find a better translation among the hypotheses in the N-best list of the translation system.

7.4.1 Translation Prediction

We compare the different features for representing the sentence and context for the translation prediction of individual source words described above. We measure the accuracy of the translation prediction achieved with each of the features and feature combinations. Table 7.6 presents an overview of the experiments. It shows the average prediction accuracy on all words in the data used for testing.

	Prediction Accuracy
Baseline	52.09
Bag-of-Words	53.29
Context (+/- 2 words)	58.74
ContextVec (+/- 2 words)	58.97
ContextVec (+/- 3 words)	57.48
Dep	56.07
DepVec	57.27
ParentDepVec	55.02
ParentWordVec	54.65
ParentWordVec+DepRel	55.20
ContextVec (+/-2) + DepVec	59.37

Table 7.6: *Translation prediction results: all words*

The baseline prediction is performed with a maximum likelihood classifier, which a priori chooses the most frequent class, without using any features at all. We can see that using the bag-of-words features consisting of the words contained in the source sentence already improves over the baseline prediction. When applying the more structurally informative features, both context and dependency features individually improve con-

siderably over the simple bag-of-words features. Among the context feature variants, the vector representation with 2 words of context in both directions performs best. For the dependency features, it is the vector representation using both parent and child relations, which leads to the best predictions. Combining the two best performing features **contextVec** and **depVec**, holds another small improvement leading to a prediction accuracy that is more than 7% higher than the baseline prediction, which corresponds to 14% relative improvement.

7.4.1.1 Pronoun Translation

In order to explicitly measure the accuracy of the translation prediction for pronouns, we selected the pronouns among the source words and measured the prediction accuracy of those words. Table 7.7 presents the prediction accuracy of the defined set of source language pronouns (Table 7.2). The pronouns achieve higher absolute numbers of translation accuracy. However, the improvements by the different types of features is comparable to the improvements on all words. The use of structural features led to an absolute and relative increase in prediction accuracy by more than 5% and 9%, respectively.

	Prediction Accuracy	
	all words	pronouns
Baseline	52.09	59.58
Bag-of-Words	53.29	60.03
ContextVec (+/- 2 words)	58.97	64.89
DepVec	57.27	63.12
ContextVec (+/-2) + DepVec	59.37	65.08

Table 7.7: *Translation prediction results: pronouns*

7.4.1.2 Subject-Verb Agreement

We also analyzed the accuracy of prediction features with respect to subject-verb agreement. For this purpose all word pairs connected by a subject relation were extracted from the dependency trees for the source sentences. All words posing as parents in such a dependency relation were taken to be possible verbs, and all children in a subject relation are considered as possible subjects. It has to be noted, though, that the subject and verb list can also contain words of other parts-of-speech, since relations such as

the one between nouns and adjectives can also be defined as a subjective relation in a dependency tree. However, manual inspection confirmed that apart from a few outliers it was indeed mostly words qualifying as subjects and verbs in the extracted list and we chose not to apply an additional manual filter. In order to produce comparable results, we measured the prediction accuracy of the words in the subject and verb lists in the same way as all words and pronouns in the results reported above. The results are presented in Table 7.8. It shows that the improvements of subjects and verbs are the highest, almost reaching 10% absolute and 20% relative improvement over the baseline prediction.

	Prediction Accuracy		
	all words	subjects	verbs
Baseline	52.09	46.81	46.71
ContextVec (+/-2) + DepVec	59.37	56.00	54.12

Table 7.8: *Translation prediction results: subjects and verbs*

7.4.2 N-Best List Re-ranking

The results of improved prediction accuracy of the SDWL model with structural informative features presented above are encouraging. Therefore, we want to use the predictions to judge the quality of a particular translation hypothesis in N-best list re-ranking. For the baseline, an N-best list re-ranking is performed, using the original sentence-based scores available from the translation system. Then we compare the four ways of generating an additional score for a given hypothesis based on the individual word translation predictions described above: absolute and relative number of predicted words in the hypothesis, sum of the prediction probabilities of the words chosen in the hypothesis and rank of the words in the hypothesis according to prediction probabilities.

Table 7.9 shows an overview over the results. Three of the methods to create the sentence score perform very similar, providing about 0.2 BLEU points of improvement. Only when using the prediction ranks of the words in the hypothesis, the translation quality is not increased. That means that the translation predictions can indeed serve as an indicator for translation quality when combined in one of the three proposed ways. By using the SDWL-based scores it is possible to select an even better hypothesis from the N-best list compared to using only the available scores from the translation system.

Translation System	BLEU
Baseline	24.04
SDWL: Abs	24.20
SDWL: Rel	24.22
SDWL: Sum	24.21
SDWL: Rank	23.98

Table 7.9: *N-best list re-ranking with prediction features: translation results*

7.4.3 Comparison with Weiner (2014)

Weiner (2014) applied the SDWL approach to the translation of third person pronouns. They investigated different types of features based on antecedents generated by an external anaphora resolution tool as well as reference antecedents determined manually to address pronoun translation. However, the antecedent features could not help to achieve an improvement over using simple bag-of-word features in the SDWL model. Furthermore, the SDWL model addressing pronoun translation did not surpass the baseline pronoun translation achieved by the statistical machine translation system in their case. The results of Weiner (2014) are reproduced in Tables A.6 and A.7 in Appendix A.

Our results presented above show that the SDWL approach is suitable for modeling translation prediction for all pronouns and even other word categories (Tables 7.7 and 7.6). We applied structural features modeling the context surrounding the given source word as well as dependency relations between the source word and other words in the sentence. This way an improvement of the prediction accuracy for pronouns and other words could be achieved over the baseline prediction, which applies a maximum likelihood classifier. In addition, by using the structural features modeling context and dependency relations we could improve over the SDWL with simple bag-of-words features, which is equivalent to the SDWL baseline in Weiner (2014) reproduced in Table A.7.

Since the SDWL did not give an advantage over using the pronoun translation already achieved in the translation system in their experiments, Weiner (2014) did not perform any further attempts to integrate it into the translation system. In contrast, the SDWL features presented here proved successful in N-best list re-ranking. The predicted translations for each source word were used to compute a sentence score for the translation hypotheses assessed in the re-ranking procedure. Using the hypotheses preferred by the SDWL features led to an improvement of the translation quality.

7.5 Translation Examples

We inspected the translation output after the N-best list re-ranking with prediction-based sentence scores and found that better translations with regard to pronouns and agreement were chosen compared to the baseline re-ranking. The challenging translation examples introduced in Chapter 3 could be improved with the presented method. The following examples show that the prediction model provides a better translation disambiguation both for pronouns and for satisfying agreement requirements.

The translation in Example 7.4 shows a translation of the pronoun *it*, which refers to a boat. This can only be inferred from the use of the verb *sailing*. The baseline translation system translates the English pronoun into the German *sie*, a feminine or plural pronoun. With the SDWL, the neuter translation *es* was chosen, generating the correct gender agreement with the implicit sailing boat, which is neuter in its German translation. However, the translation does not match with the translation in the reference. Hence, this is an example which would not affect the BLEU score, even though it is an improvement.

Source:	*And I went sailing on it , and we did surveys throughout the southern South China sea and especially the Java Sea.*
Translation:	*Und ich ging auf sie segeln , und wir haben Umfragen in den südlichen Südchinesische Meer und vor allem die Java-See.*
+SDWL-Model:	*Und ich ging es segeln , und wir haben Umfragen in der gesamten südlichen Südchinesische Meer und vor allem die Java-See.*
Reference:	*Ich fuhr darauf mit und wir machten Erhebungen im ganzen südlichen Südchinesischen Meer und besonders in der Javasee.*

Example 7.4: Correct gender for pronoun

Another improvement in pronoun translation is shown in Example 7.5. Here the person and number of the pronoun in the baseline translation is correct, and the right case is chosen. However, the gender is incorrect. It needs to agree with the connected noun *Klasse*, which is feminine. The SDWL generates the correct gender so that the grammatical agreement of the possessive pronoun and the noun holds in this noun phrase.

Example 7.6 shows that the translation prediction model also encourages morphological agreement between subject and verb. The information that the verb is actually

Source:	*I memorized in my anatomy class the origins and exertions of every muscle [...]*
Translation:	*Ich in meinem Anatomie der Klasse die Ursprünge und Strapazen eines jeden Muskel [...] auswendig [...]*
+SDWL-Model:	*Ich in meiner Klasse Anatomie die Ursprünge und Strapazen jeder Muskel [...] auswendig [...]*
Reference:	*In meiner Anatomievorlesung lernte ich die Ursprünge und Ausläufer jedes Muskels [...]*

Example 7.5: Correct gender ending for pronoun

in plural form is not encoded in the source language. The English verb *can* can be both singular and plural. Hence, producing a plural verb in the translation is not straight forward. Apparently, the structural features are able to capture the plural subject in the dependency feature and/or the plural indicator *and* in the context feature. As a result the translation hypothesis higher with the plural verb (*können*) achieves a higher rank in the N-best list and is chosen as the best translation.

Source:	*There I think that the arts and film can perhaps fill the gap, and simulation.*
Translation:	*Ich glaube, dass die Kunst und Film kann vielleicht die Lücke füllen, und Simulation.*
+SDWL-Model:	*Ich glaube, dass die Kunst und Film, vielleicht können die Lücke füllen, und Simulation.*
Reference:	*Hier können, denke ich, die Kunst und der Film vielleicht die Lücke füllen, sowie Simulationen.*

Example 7.6: Correct case agreement between subject and verb

The SDWL prediction model has also additional applications. It can provide disambiguation for types of difficult translations regarding pronouns and agreement, other than the ones this model was particularly tested on. In Example 7.7 the correct translation for the relative pronoun *that* can be chosen after applying the SDWL predictions in re-ranking. This is another example, where the translation and reference do not match and the BLEU score will not be affected positively even though the translation is improved.

Source:	*Somehow by ways that we don't quite understand, [...]*
Translation:	*Irgendwie durch Möglichkeiten, dass wir nicht ganz verstehen, [...]*
+SDWL-Model:	*Irgendwie von Wegen, die wir nicht ganz verstehen, ...*
Reference:	*Auf irgend eine Art, welche wir noch nicht ganz verstehen, [...]*

Example 7.7: Correct disambiguation of relative pronoun

The following example shows another improvement with regard to morphological agreement within a noun phrase. In Example 7.8 the SDWL prediction helps to choose the correct case for the translation of the noun phrase *this code*. In this sentence, the determiner *this* and the noun in the noun phrase both need to be used in the dative form, which is corrected when using the SDWL predictions in re-ranking.

Source:	*[...] we can now write things in this code.*
Translation:	*[...], können wir jetzt die Dinge in diesen Code schreiben.*
+SDWL-Model:	*..., wir können jetzt Dinge in diesem Code schreiben.*
Reference:	*dass wir, [...], selber Sachen in diesem Code schreiben können.*

Example 7.8: Correct case agreement between determiner and noun

7.6 Conclusion

We have presented a model for translation disambiguation using structural features in a classification task. The translation of a source word in a given sentence is predicted based on the classification into one of its 20 most frequent translation options. Structural features such as source context words and relations in the dependency tree of the source sentence allow to include knowledge about the sentence structure when modeling the prediction. The model is in particular aimed at improving challenging linguistic issues like the translation of pronouns and generating morphological agreement in the translated sentence.

The prediction results have shown that the accuracy of predicting a translation for individual source words increases considerably when including the context and dependency features. Representing the features by a word2vec word vector representation both reduces dimensions and increases prediction accuracy. Even though the context and dependency features contribute similar improvements individually, their combi-

nation provides the highest prediction accuracy. A separate inspection of pronouns, subjects and verbs confirms that these types of words were improved in particular by up to 10%.

The individual translation predictions for the source words in each sentence are combined into a sentence score used in N-best list re-ranking. Using the prediction scores in re-ranking improves the translation quality by 0.2 BLEU points. The translation obtained after the prediction-based re-ranking has shown to repair particular translation errors in pronoun translation and morphological agreement in the target sentence.

Directions for future work could be the investigation of features that include more semantic information such as the semantic distance between words. Furthermore, the current classification approach could be compared to other machine learning techniques such as neural networks which are able to model more implicit dependencies.

8

Conclusions

In this thesis we have investigated the influence of linguistic structure in statistical machine translation. We dealt with the differences in word order between languages and how to use linguistic structure from constituency trees to improve over a part-of-speech-based reordering model. A second line of research was dedicated to the problem of pronoun translation and the improvement of morphological agreement for morphologically rich target languages in statistical machine translation.

8.1 Syntactic Reordering

We developed a reordering model based on syntactic parse trees for targeting verb movements when translating from and into German. Since the location of the verb in a German sentence depends on various factors, we first focused on German as the source language for translation. The reordering experiments were performed on German-to-English and German-to-French translation. When comparing the tree-based reordering with POS-based reordering we observed that tree-based reordering can improve the translation quality over POS-based reordering. The best results are obtained when the tree-based and POS-based reordering models are combined. Further improvements were achieved when including a lexicalized reordering model in the machine translation system. Our results suggest that the different reordering methods have complementary reordering effects. Their individual improvements can be increased through combination.

8.1.1 Oracle Experiments with POS- and Tree-based Reordering

Next, we conducted performance experiments with the POS-based and tree-based reordering models. With the help of oracle experiments we analyzed the performance of the tree-based reordering on English-German and German-English translation of News texts and TED talks.

In order to establish an upper bound for translation quality, we translated an optimally reordered source sentence. The best oracle path in the lattices produced with the POS-based and tree-based reordering models bridges the gap half-way to the upper bound presented by the optimally reordered sentence. Therefore, a possible direction for future research is to develop techniques that predict reordering options that are currently not in the search space.

When examining the path that is actually chosen for translation by the decoder we found that the decoder path is quite close to the oracle path for German-English translation. When translating from English to German finding the translation path is more difficult. The experiments on both News and TED data showed that further improvement were possible with the presented reordering model, if better scoring methods provided a better discrimination between reordering options in the lattice.

8.1.2 Manual Analysis of the Tree-based Reordering Model

We also performed a comparative analysis of the POS-based and tree-based reordering models on three genres: News texts, TED talks and University lectures. We found that the impact of the tree-based model is higher for well structured, grammatically correct texts, while fewer sentences are affected in the two speech data sets when the tree-based reordering model is applied.

However, a manual evaluation of the translation quality on the sentence level confirmed consistent improvements throughout all three data sets. Around 100 sentences per data set were inspected manually, comparing the translation outputs after applying the POS-based reordering and the combined reordering model of POS-based and tree-based rules. In 55 to 64% of the cases, the system including tree-based rules produced a better translation, while only 24 to 28% of the sentences generated by the POS-based reordering rules were considered to be a better translation. That means that for 72 to 76% of the sentences, the tree-based reordering led to either an improvement of the translation quality or the translation quality stayed the same.

The improvements introduced by the tree-based reordering model include translations of words which were removed during the translation process before. In addition,

word and constituent positions in the translated sentence were improved. As intended in the design of the reordering model, verbs are the most frequently affected word class.

8.2 Linguistic Structure for Translation Disambiguation

The many ambiguities inherent to natural language make translation a difficult task. In this thesis we developed a model for translation disambiguation for the individual words in a source sentence. With features based on the sentence structure the translation disambiguation is modeled as a classification task where the classes are possible translations for the source language word. The output of the classification is the predicted translation within the given sentence and context. Two particularly difficult linguistic challenges are addressed with this translation prediction model: the translation of pronouns and the generation of morphological agreement in the target language.

8.2.1 Pronoun Translation

The translation of a pronoun depends on what its antecedent—the previously mentioned noun it refers to—is translated into. The pronoun in the target language then needs to exhibit a gender and number that is concordant with the morphological features of the antecedent's translation in the target language. In the translation prediction model, structural features such as context words and dependency relations in the sentence serve as a way to model this implicitly while learning how to translate a pronoun in a given sentence. The results show that pronoun translation is indeed influenced positively by the translation prediction model. The prediction accuracy for individual pronouns improved by 5% and the prediction accuracy for all words improved by 7% compared to a baseline classification. In addition, the translation quality of the translations chosen in N-best list re-ranking based on our predictions is improved by 0.2 BLEU points.

8.2.2 Morphological Agreement

Languages differ in terms of the explicitness of morphological features visible in a word's surface form. When translating into a language with rich morphology, generating the correct word form, e.g. in order to achieve morphological agreement, is a challenge for a machine translation system. Especially if the source language offers less morphological variation, the generation of correct morphological agreement without evidence from the source side often leads to ungrammatical or semantically wrong translation output.

The structural features used in the translation prediction model encode additional information that has shown to improve the accuracy for the translation prediction of subjects and verbs by up to 10%. In addition, morphological agreement is improved in the produced translations.

8.3 Summary

In this thesis two models were developed to improve a phrase-based machine translation system by incorporating information on the linguistic structure of the source language. The models were specifically targeted to three particular linguistic challenges presented by the English-German language pair. The differences in word order are addressed by a reordering model based on syntactic parse trees. The difficulties in pronoun translation and generating morphological agreement are addressed by a translation prediction model used in an N-best list re-ranking approach. In order to provide an overview of the contributions in this thesis, the methods were applied cumulatively to a strong baseline provided by a phrase-based machine translation system. Table 8.1 shows the development of the translation quality for translation from English to German on the translation of TED talks. The contributions from this thesis are highlighted in bold.

Translation System	BLEU
Baseline	23.47
+ Re-ranking	23.81
+ **Syntax-based Reordering**	**24.04**
+ **Structural Features for Prediction-based Re-ranking**	**24.22**

Table 8.1: *Thesis Overview: Translation Results*

8.4 Future Work

The topics covered in this thesis have been shown to be challenges which could be improved by the developed methods, but are far from being completely resolved. Translation quality is highly affected by word reordering and ambiguous words that require a linguistic dependency to hold in the target language after translation. Both of them should be investigated further in order to achieve translations that increasingly satisfy the linguistic constraints of natural languages.

The oracle experiments conducted in Chapter 6 revealed some directions of future work for the syntactic tree-based reordering approach. In general, it would be beneficial to extend the approach to learn additional rules that approximate the actual reordering that happens between the languages even better. Furthermore, the reordering approach could benefit from an improved scoring of the reordering options for a better differentiation between the suggested reordering options by the syntactic tree-based and part-of-speech-based reordering rules.

The approach for translation disambiguation presented in Chapter 7 provides a framework for straightforward extension and substitution of features. A future line of research could be to include features representing meaning such as the semantic distance between words. A comparison of different machine learning approaches for performing the classification task seems also promising. For example modeling the prediction with a neural network could enable a better modeling of the implicit dependencies between words in translation.

Appendices

Appendix A

Pronominal Anaphora in Machine Translation

In a related project, Weiner (2014) tried to improve pronoun translation in English-to-German translation, focusing on the third person pronouns *he, she*, and *it*. In that work, the automatic anaphora resolution tool JavaRAP (Qiu et al., 2004) was applied to obtain antecedent information for improving the translation of the pronouns. They first conducted a set of analyses, assessing the performance of the automatic anaphora resolution and locating where the antecedents occurred (Table A.1). Another part of the study investigated how well the pronoun translation of the machine translation system performs for the third person pronouns in general (Tables A.2 and A.3) and for the specific source–target pronoun pairs (Tables A.4 and A.5). These analyses showed that it is mainly the pronouns *it* and *its* that need special attention. They are only translated correctly in half of the cases.

Several approaches to improve the translation of pronouns are investigated. Tables A.6 and A.7 show an overview of the results. Two post-processing methods were compared. The first method substitutes pronouns in post-processing such that the gender and number agree with the antecedent identified in manual and automatic anaphora resolution. In the second method a hypothesis is chosen from the N-best list, such that pronoun features agree with the antecedent. Then two approaches using a discriminative word lexicon were applied. Different kinds of features were compared ranging from antecedent related features to previous nouns in the sentence. However, the discriminative word lexicon approaches could not improve over the baseline. Afterwards, the source discriminative word lexicon was applied with antecedent features. These results obtained in Weiner (2014) motivated the extended research on pronoun

translation including all person pronouns on source and target side when targeting the evaluation of pronouns performed in this thesis.

anaphora list	number of pairs	P	R	F1	intra pronouns	inter pronouns
news.a.manual	288				50.7	49.3
news.a.auto	368	0.40	0.51	0.44	63.3	36.7
ted.a.manual	170				24.1	75.9
ted.a.auto	176	0.47	0.48	0.47	54.5	45.5

Table A.1: *Anaphora statistics for News and TED*

source pronoun	occurrences	translated correctly
personal pronouns nominative		
he	49	100.0%
it	42	47.6%
she	10	90.0%
they	47	97.9%
personal pronouns objective		
her	6	100.0%
him	5	100.0%
them	11	100.0%
possessive pronouns		
his	21	100.0%
its	14	71.4%
their	44	88.6%

Table A.2: *Translations for News*

source pronoun	occurrences	translated correctly
personal pronouns nominative		
he	52	100.0%
it	36	47.2%
she	1	100.0%
they	28	100.0%
personal pronouns objective		
him	15	100.0%
them	3	100.0%
possessive pronouns		
his	14	100.0%
its	11	54.5%
their	1	100.0%

Table A.3: *Translations for TED*

source pronoun	target pronoun	how often	translated correctly
personal pronouns nominative			
he	er	100.0%	100.0%
it	er	19.0%	0.0%
	es	40.5%	76.5%
	ihn	7.1%	0.0%
	sie	33.3%	50.0%
she	er	10.0%	0.0%
	sie	90.0%	100.0%
they	es	2.1%	0.0%
	sie	97.9%	100.0%
personal pronouns objective			
her	ihr	16.7%	100.0%
	ihre	50.0%	100.0%
	ihrem	16.7%	100.0%
	ihren	16.7%	100.0%
him	ihm	60.0%	100.0%
	ihn	20.0%	100.0%
	seiner	20.0%	100.0%
them	sie	100.0%	100.0%
possessive pronouns			
his	sein	14.3%	100.0%
	seine	28.6%	100.0%
	seinem	9.5%	100.0%
	seinen	33.3%	100.0%
	seiner	14.3%	100.0%
its	ihrem	7.1%	100.0%
	ihren	7.1%	0.0%
	ihrer	7.1%	100.0%
	sein	14.3%	100.0%
	seine	28.6%	75.0%
	seinen	28.6%	75.0%
	seiner	7.1%	0.0%
	seinen	2.3%	0.0%
their	ihr	6.8%	100.0%
	ihre	45.5%	100.0%
	ihrem	13.6%	100.0%
	ihren	4.5%	100.0%
	ihrer	15.9%	100.0%
	ihres	2.3%	100.0%
	seine	4.5%	0.0%
	seinem	4.5%	0.0%
	seinen	2.3%	0.0%

Table A.4: *Translations for News*

source pronoun	target pronoun	how often	translated correctly
personal pronouns nominative			
he	er	100.0%	100.0%
it	er	22.2%	12.5%
	es	44.4%	100.0%
	ihn	5.6%	0.0%
	sie	27.8%	0.0%
she	sie	100.0%	100.0%
they	sie	100.0%	100.0%
personal pronouns objective			
him	ihm	26.7%	100.0%
	ihn	73.3%	100.0%
them	ihnen	33.3%	100.0%
	sie	66.7%	100.0%
possessive pronouns			
his	sein	14.3%	100.0%
	seine	42.9%	100.0%
	seinem	14.3%	100.0%
	seinen	7.1%	100.0%
	seiner	21.4%	100.0%
its	ihr	9.1%	0.0%
	ihre	9.1%	100.0%
	ihren	63.6%	57.1%
	seine	9.1%	0.0%
	seinen	9.1%	100.0%
their	ihren	100.0%	100.0%

Table A.5: *Translations for TED*

	he (49)	**she (10)**	**it (42)**	**they (47)**	**all (148)**
baseline translation	100.0	90.0	47.6	97.9	83.8
post-processing – correcting words					
corrected by pos.text (.a.manual.correctPair)	89.8	50.0	92.9	80.0	84.9
corrected by pos.pt (.a.manual.correctPair)	71.4	40.0	90.5	100.0	83.8
corrected by pos.text (.a.auto.correctPair)	91.8	60.0	73.8	97.9	86.5
corrected by pos.pt (.a.auto.correctPair)	91.8	60.0	73.8	97.9	86.5
post-processing – n-best					
corrected by pos.text (.a.manual.correctPair)	100.0	90.0	81.0	97.9	93.2
corrected by pos.pt (.a.manual.correctPair)	100.0	90.0	78.6	97.9	92.6
corrected (.a.auto.correctPair)	100.0	90.0	69.0	97.9	89.9
dwl words					
baseline	95.9	90.0	50.0	91.5	81.1
target antecedent pos2 (.a.manual.correctPair)	91.8	90.0	50.0	93.6	80.4
target antecedent pos2 (.a.auto.correctPair)	91.8	80.0	42.9	93.6	77.7
dwl ngrams					
baseline	87.8	90.0	40.5	89.4	75.0
previous nouns2 (.a.manual.correctPair)	91.8	90.0	23.8	89.4	71.6
previous nouns2 (.a.auto.correctPair)	91.8	90.0	23.8	89.4	71.6
sdwl words (SDWL-4c)					
baseline	100.0	90.0	38.5	97.9	81.2
target antecedent (.a.manual.correctPair)	100.0	90.0	30.8	97.9	79.0
target antecedent (.a.auto.correctPair)	100.0	100.0	31.0	100.0	80.4
sdwl ngrams (SDWL-4c)					
baseline	100.0	90.0	38.5	97.9	81.2
target antecedent (.a.manual.correctPair)	100.0	90.0	43.6	97.9	82.7
target antecedent (.a.auto.correctPair)	100.0	100.0	37.9	100.0	82.4

Table A.6: *Pronoun evaluation results for News (in %)*

	he (52)	**she (1)**	**it (36)**	**they (28)**	**all (117)**
baseline translation	100.0	100.0	47.1	100.0	83.7
post-processing – correcting words					
corrected by pos.text (.a.manual.correctPair)	96.2	0.0	94.4	78.6	90.6
corrected by pos.pt (.a.manual.correctPair)	96.2	0.0	88.8	78.6	88.9
corrected by pos.text (.a.auto.correctPair)	100.0	0.0	66.7	100.0	88.9
corrected by pos.pt (.a.auto.correctPair)	100.0	0.0	66.7	100.0	88.9
post-processing – n-best					
corrected by pos.text (.a.manual.correctPair)	100.0	100.0	77.7	89.3	90.6
corrected by pos.pt (.a.manual.correctPair)	100.0	100.0	72.2	89.3	88.9
corrected (.a.auto.correctPair)	100.0	100.0	52.8	100.0	85.5
dwl words					
baseline	96.2	100.0	47.2	89.3	79.5
target antecedent (.a.manual.correctPair)	100.0	100.0	47.2	92.9	82.0
target antecedent (.a.auto.correctPair)	100.0	100.0	44.4	92.9	81.2
dwl ngrams					
baseline	88.5	100.0	50.0	78.6	74.4
target antecedent pos (.a.manual.correctPair)	98.1	100.0	52.8	89.3	82.0
target antecedent pos (.a.auto.correctPair)	94.2	100.0	50.0	89.3	79.5
sdwl words (SDWL-4c)					
baseline	100.0	100.0	47.1	89.3	81.2
target antecedent (.a.manual.correctPair)	100.0	100.0	47.1	100.0	83.7
target antecedent (.a.auto.correctPair)	100.0	100.0	33.3	100.0	79.5
sdwl ngrams (SDWL-4c)					
baseline	100.0	100.0	47.1	97.9	83.7
target antecedent (.a.manual.correctPair)	100.0	100.0	47.1	92.9	82.0
target antecedent (.a.auto.correctPair)	100.0	100.0	33.3	100.0	79.5

Table A.7: *Pronoun evaluation results for TED (in %)*

References

Eleftherios Avramidis and Philipp Koehn. Enriching Morphologically Poor Languages for Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ACL-HLT 2008, Columbus, OH, USA, 2008. 37

Nguyen Bach, Qin Gao, and Stephan Vogel. Source-side Dependency Tree Reordering Models with Subtree Movements and Constraints. In *Proceedings of the Twelfth Machine Translation Summit*, MT Summit XII, Ottawa, Canada, 2009. 31

Alexandra Birch. *Reordering Metrics for Statistical Machine Translation*. PhD thesis, University of Edinburgh, Edinburgh, Scotland, 2011. 31, 32, 60, 61

Alexandra Birch, Miles Osborne, and Philipp Koehn. Predicting Success in Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2008, Honolulu, HI, USA, 2008. 36

Alexandra Birch, Miles Osborne, and Phil Blunsom. Metrics for MT Evaluation: Evaluating Reordering. *Machine Translation*, 24(1), 2010. 31, 60, 61

Arianna Bisazza and Christof Monz. Class-Based Language Modeling for Translating into Morphologically Rich Languages. In *Proceedings of the International Conference on Computational Linguistics: Technical Papers*, Coling 2014, Dublin, Ireland, 2014. 38

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Hervé Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, WMT 2014, Baltimore, MD, USA, 2014a. 13

REFERENCES

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, WMT 2013, Sofia, Bulgaria, 2013a. 13, 32

Ondřej Bojar and Christian Buck and Chris Callison-Burch and Barry Haddow and Philipp Koehn and Christof Monz and Matt Post and Hervé Saint-Amand and Radu Soricut and Lucia Specia. *Proceedings of the Eighth Workshop on Statistical Machine Translation*. WMT 2013. Association for Computational Linguistics, Sofia, Bulgaria, 2013b. 14

Ondřej Bojar and Christian Buck and Christian Federmann and Barry Haddow and Philipp Koehn and Christof Monz and Matt Post and Lucia Specia. *Proceedings of the Ninth Workshop on Statistical Machine Translation*. WMT 2014. Association for Computational Linguistics, Baltimore, MD, USA, 2014b. 14

Jan A. Botha and Phil Blunsom. Compositional Morphology for Word Representations and Language Modelling. In *Proceedings of the International Conference on Machine Learning*, ICML 2014, Beijing, China, 2014. 38

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85, June 1990. 1, 7

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. Word-sense Disambiguation Using Statistical Methods. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, ACL 1991, Berkeley, CA, USA, 1991. 33

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993. 1, 7, 8, 29

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, WMT 2007, Prague, Czech Republic, 2007. 13

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further Meta-Evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, WMT 2008, Columbus, OH, USA, 2008. 13

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, WMT 2009, Athens, Greece, 2009. 13

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT 2010, Uppsala, Sweden, 2010. 13

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT 2012, Montréal, Canada, 2012a. 1

Chris Callison-Burch and Philipp Koehn and Christof Monz and Matt Post and Radu Soricut and Lucia Specia. *Proceedings of the Seventh Workshop on Statistical Machine Translation*. WMT 2012. Association for Computational Linguistics, Montréal, Canada, 2012b. 14, 42

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to Rank: From Pairwise Approach to Listwise Approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML 2007, Corvalis, OR, USA, 2007. 49

Marine Carpuat and Dekai Wu. Evaluating the word sense disambiguation performance of statistical machine translation. In *Proceedings of the Second International Joint Conference on Natural Language Processing*, IJCNLP 2005, Jeju, Republic of Korea, 2005. 33

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL 2007, Prague, Czech Republic, 2007. 33

Bruno Cartoni. Lexical Morphology in Machine Translation: A Feasibility Study. In *Proceedings of the Twelfth Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2009, Athens, Greece, 2009. 37

Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, EAMT 2012, Trento, Italy, 2012. xii, 42

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. Report on the 10th IWSLT Evaluation Campaign. In *Proceedings of the Tenth International Workshop on Spoken Language Translation*, IWSLT 2013, Heidelberg, Germany, 2013. 1, 43

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014. In *Proceedings of the Eleventh International Workshop on Spoken Language Translation*, IWSLT 2014, Lake Tahoe, CA, USA, 2014. 43

Victor Chahuneau, Eva Schlinger, Noah A. Smith, and Chris Dyer. Translating into Morphologically Rich Languages with Synthetic Phrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2013, Seattle, WA, USA, 2013. 38

Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL 2007, Prague, Czech Republic, 2007. 33

David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL 2005, Ann Arbor, MI, USA, 2005. 12, 31

Eunah Cho, Jan Niehues, and Alex Waibel. Segmentation and Punctuation Prediction in Speech Language Translation Using a Monolingual Translation System. In *Proceedings of the Ninth International Workshop on Spoken Language Translation*, IWSLT 2012, Hong Kong, China, 2012. 32

Ann Clifton and Anoop Sarkar. Combining Morpheme-based Machine Translation with Post-processing Morpheme Prediction. In *Proceedings of the Annual Meeting*

of the Association for Computational Linguistics: Human Language Technologies, ACL-HLT 2011, Portland, OR, USA, 2011. 37

Michael Collins, Philipp Koehn, and Ivona Kučerová. Clause Restructuring for Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL 2005, Ann Arbor, MI, USA, 2005. 30

Marta R. Costa-Jussà and José A. R. Fonollosa. Statistical Machine Reordering. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, EMNLP 2006, Sydney, Australia, 2006. 30

Josep M. Crego and Nizar Habash. Using Shallow Syntax Information to Improve Word Alignment and Reordering for SMT. In *Proceedings of the Third Workshop on Statistical Machine Translation*, WMT 2008, Columbus, OH, USA, 2008. 30

David Crystal. *The Cambridge Encyclopedia of the English Language*. Cambridge University Press, 2003. ISBN 9780521530330. 93

Marie-Catherine de Marneffe and Christopher D. Manning. Stanford typed dependencies manual. Technical report, Stanford University, Stanford, CA, USA, 2008. 99

José Guilherme Camargo de Souza and Constantin Orasan. Can Projected Chains in Parallel Corpora Help Coreference Resolution? In *DAARC*, volume 7099 of *Lecture Notes in Computer Science*. 2011. 35

John DeNero and Jakob Uszkoreit. Inducing Sentence Structure from Parallel Corpora for Reordering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2011, Edinburgh, Scotland, 2011. 30

Markus Dreyer, Keith Hall, and Sanjeev Khudanpur. Comparing Reordering Constraints for SMT Using Efficient BLEU Oracle Computation. In *Proceedings of the Workshop on Syntax and Structure in Statistical Translation*, SSST 2007, Rochester, NY, USA, 2007. 32

Jinhua Du and Andy Way. A Discriminative Latent Variable-Based "DE" Classifier for Chinese-English SMT. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Coling 2010, Beijing, China, 2010. 30

Peter Eisenberg, Jörg Peters, Peter Gallmann, Catherine Fabricus-Hansen, Damaris Nübling, Irmhild Barz, Thomas A. Fritz, and Reinhard Fiehler. *Duden, Grammatik*

der deutschen Gegenwartssprache, volume 4. Dudenverlag, Mannheim, Germany, 7., völlig neu erarbeitete und erweiterte Auflage edition, 2005. ISBN 3-411-04047-5. 93

Jason Eisner and Roy W. Tromble. Local Search with Very Large-Scale Neighborhoods for Optimal Permutations in Machine Translation. In *Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Processing*, New York City, NY, USA, 2006. 31

Jakob Elming. Syntactic Reordering Integrated with Phrase-Based SMT. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Coling 2008, Manchester, England, 2008. 30

Mireia Farrús, Marta R. Costa-Jussà, and Maja Popovic. Study and correlation analysis of linguistic, perceptual, and automatic machine translation evaluations. *JASIST*, 63(1):174–184, 2012. 33

Marcello Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker. Overview of the IWSLT 2012 evaluation campaign. In *Proceedings of the Ninth International Workshop on Spoken Language Translation*, IWSLT 2012, Hong Kong, China, 2012. 32

Alexander Fraser. Experiments in Morphosyntactic Processing for Translating to and from German. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, WMT 2009, Athens, Greece, 2009. 37

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. Modeling Inflection and Word-Formation in SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2012, Avignon, France, 2012. 37

S.I. Gallant. A practical approach for representing context and for performing word sense disambiguation using neural networks. In *International Joint Conference on Neural Networks*, volume ii of *IJCNN 91*, Seattle, WA, USA, July 1991. 34

Michel Galley and Christopher D. Manning. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2008, Honolulu, HI, USA, 2008. 31

Qin Gao and Stephan Vogel. Parallel Implementations of Word Alignment Tool. In *Proceedings of the Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, Columbus, OH, USA, 2008. 45

Dmitriy Genzel. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the International Conference on Computational Linguistics*, Coling 2010, Beijing, China, 2010. 30

Kevin Gimpel and Noah A. Smith. Rich Source-side Context for Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, WMT 2008, Columbus, OH, USA, 2008. 34

Spence Green and John DeNero. A Class-based Agreement Model for Generating Accurately Inflected Translations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL 2012, Jeju, South Korea, 2012. 37

Liane Guillou. Improving Pronoun Translation for Statistical Machine Translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2012, Avignon, France, 2012. 35

Nizar Habash. Syntactic preprocessing for statistical machine translation. In *Proceedings of the Eleventh Machine Translation Summit*, MT Summit XI, 2007. 30

Rejwanul Haque, Sudip Kumar Naskar, Antal van den Bosch, and Andy Way. Integrating source-language context into phrase-based statistical machine translation. *Machine Translation*, 25(3):239–285, 2011. 34

Sanda M. Harabagiu and Steven J. Maiorano. Multilingual Coreference Resolution. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, Seattle, WA, USA, 2000. 35

Christian Hardmeier and Marcello Federico. Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of the Seventh International Workshop on Spoken Language Translation*, IWSLT 2010, Paris, France, 2010. 35

Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. Latent Anaphora Resolution for Cross-Lingual Pronoun Prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2013, Seattle, WA, USA, 2013. 35

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL 2013, Sofia, Bulgaria, 2013. xi, 10

Teresa Herrmann, Jan Niehues, and Alex Waibel. Combining Word Reordering Methods on different Linguistic Abstraction Levels for Statistical Machine Translation. In *Proceedings of Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST 2013, Altanta, GA, USA, 2013a. 51

Teresa Herrmann, Jochen Weiner, Jan Niehues, and Alex Waibel. Analyzing the Potential of Source Sentence Reordering in Statistical Machine Translation. In *Proceedings of the Tenth International Workshop on Spoken Language Translation*, IWSLT 2013, Heidelberg, Germany, 2013b. 51

Teresa Herrmann, Jan Niehues, and Alex Waibel. Manual Analysis of Structurally Informed Reordering in German-English Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC 2014, Reykjavik, Iceland, 2014. 52

Teresa Herrmann, Jan Niehues, and Alex Waibel. Source Discriminative Word Lexicon for Translation Disambiguation. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, IWSLT 2015, Da Nang, Vietnam, 2015. 92

Jerry R. Hobbs. Resolving Pronoun References. In *Readings in Natural Language Processing*, pages 339–352. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1986. 34

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2010, Cambridge, MA, USA, 2010. 31

Minwoo Jeong, Kristina Toutanova, Hisami Suzuki, and Chris Quirk. A Discriminative Lexicon Model for Complex Morphology. In *Proceedings of the Conference of the Association for Machine Translation in the Americas*, AMTA 2010, Denver, CO, USA, 2010. 37

Jie Jiang, Jinhua Du, and Andy Way. Improved phrase-based SMT with syntactic reordering patterns learned from lattice scoring. In *Proceedings of the Conference of the Association for Machine Translation in the Americas*, AMTA 2010, Denver, Colorado, USA, 2010. 64

Jason Katz-Brown, Slav Petrov, Ryan McDonald, Franz Och, David Talbot, Hiroshi Ichikawa, Masakazu Seno, and Hideto Kazawa. Training a parser for machine trans-

lation reordering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2011, Edinburgh, Scotland, 2011. 30

Maurice Kendall and Jean D. Gibbons. *Rank Correlation Methods*. A Charles Griffin Title, 5 edition, September 1990. ISBN 0195208374. 31, 61

Maxim Khalilov and Khalil Sima'an. Context-Sensitive Syntactic Source-Reordering by Statistical Transduction. In *Proceedings of the International Joint Conference on Natural Language Processing*, IJCNLP 2011, Chiang Mai, Thailand, 2011. 32

Maxim Khalilov, José A.R. Fonollosa, and Mark Dras. A new subtree-transfer approach to syntax-based reordering for statistical machine translation. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, EAMT 2009, Barcelona, Spain, 2009. 30

Ahmed El Kholy and Nizar Habash. Translate, Predict or Generate: Modeling Rich Morphology in Statistical Machine Translation. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, EAMT 2012, Trento, Italy, 2012. 38

Katrin Kirchhoff, Owen Rambow, Nizar Habash, and Mona Diab. Semi-Automatic Error Analysis for Large-Scale Statistical Machine Translation Systems. In *Proceedings of the Eleventh Machine Translation Summit*, MT Summit XI, 2007. 33

Dan Klein and Christopher D. Manning. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Proceedings of the Conference on Neural Information Processing Systems*, NIPS 2002, Vancouver, Canada, 2002. 99

Dan Klein and Christopher D. Manning. Accurate Unlexicalized Parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL 2003, Sapporo, Japan, 2003. 99

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, MT Summit X, 2005. 36

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, New York City, New York, USA, 1st edition, 2010. ISBN 0521874157, 9780521874151. 7

Philipp Koehn and Hieu Hoang. Factored translation models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL 2007, Prague, Czech Republic, 2007. 36

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology*, NAACL-HLT 2003, Edmonton, Canada, 2003. 8

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, IWSLT 2005, Pittsburgh, PA, USA, 2005. 29, 46

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Demonstration Session*, ACL 2007, Prague, Czech Republic, 2007a. 45

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL 2007, Prague, Czech Republic, 2007b. 8, 29

Shalom Lappin and Herbert J. Leass. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535–561, December 1994. 34, 35

Alon Lavie and Michael J. Denkowski. The Meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115, 2009. 32

Ronan Le Nagard and Philipp Koehn. Aiding Pronoun Translation with Co-reference Resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT 2010, Uppsala, Sweden, 2010. 35

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the Conference on Natural Language Learning*, CoNLL 2011, Portland, OR, USA, 2011. 35

Uri Lerner and Slav Petrov. Source-Side Classifier Preordering for Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2013, Seattle, WA, USA, 2013. 30

Lori S Levin, Donna Gates, Alon Lavie, and Alex Waibel. An interlingua based on domain actions for machine translation of task-oriented dialogues. In *Proceedings of the Fifth International Conference on Spoken Language Processing*, ICSLP 98, Sydney, Australia, 1998. 7

Matouš Macháček and Ondřej Bojar. Results of the WMT13 Metrics Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, WMT 2013, Sofia, Bulgaria, 2013. 14

Matouš Macháček and Ondřej Bojar. Results of the WMT14 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, WMT 2014, Baltimore, MD, USA, 2014. 14

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. Two-step translation with grammatical post-processing. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT 2011, Edinburgh, Scotland, 2011. 37

Eva Martinez Garcia, Jörg Tiedemann, Cristina España Bonet, and Lluís Màrquez. Word's Vector Representations meet Machine Translation. In *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST 2014, Doha, Qatar, 2014. 34

Arne Mauser, Saša Hasan, and Hermann Ney. Extending Statistical Machine Translation with Discriminative and Trigger-based Lexicon Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2009, Suntec, Singapore, 2009. 34, 48

Aurélien Max, Rafik Makhloufik, and Philippe Langlais. Explorations in Using Grammatical Dependencies for Contextual Phrase Translation Disambiguation. In *Proceedings of the Twelfth Conference of the European Association for Machine Translation*, EAMT 2008, Hamburg, Germany, 2008. Poster. 33

Igor Mel'čuk and Leo Wanner. Morphological mismatches in machine translation. *Machine Translation*, 22(3):101–152, September 2008. 37

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781, 2013. xii, 34, 100

Einat Minkov and Kristina Toutanova. Generating complex morphology for machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL 2007, Prague, Czech Republic, 2007. 36

Ruslan Mitkov and Catalina Barbu. Using bilingual corpora to improve pronoun resolution. *Languages in Contrast*, 4(2):201–211, September 2002. 35

Ruslan Mitkov, Sung kwon Choi R, and All Sharp. Anaphora resolution in Machine Translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, 1995. 35

Thomas Müller, Hinrich Schütze, and Helmut Schmid. A Comparative Investigation of Morphological Language Modeling for the Languages of the European Union. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT 2012, Montreal, Canada, 2012. 38

ThuyLinh Nguyen and Stephan Vogel. Integrating Phrase-based Reordering Features into a Chart-based Decoder for Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL 2013, Sofia, Bulgaria, 2013. 31

Jan Niehues and Muntsin Kolss. A POS-Based Model for Long-Range Reorderings in SMT. In *Proceedings of the Workshop on Statistical Machine Translation*, WMT 2009, Athens, Greece, 2009. 30, 46, 47, 52

Jan Niehues and Stephan Vogel. Discriminative Word Alignment via Alignment Matrix Modeling. In *Proceedings of the Third Workshop on Statistical Machine Translation*, WMT 2008, Columbus, OH, USA, 2008. 45

Jan Niehues and Alex Waibel. An MT Error-driven Discriminative Word Lexicon using Sentence Structure Features. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, WMT 2013, Sofia, Bulgaria, 2013. 48, 98

Jan Niehues, Teresa Herrmann, Stephan Vogel, and Alex Waibel. Wider Context by Using Bilingual Language Models in Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT 2011, Edinburgh, Scotland, 2011. 45

Sonja Nießen and Hermann Ney. Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Computational Linguistics*, 30(2):181–204, 2004. 38

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. Translation of "It" in a Deep Syntax Framework. In *Proceedings of the Annual Meeting of the Association for*

Computational Linguistics, Workshop on Discourse in Machine Translation, Sofia, Bulgaria, 2013a. 36

Michal Novák, Zdeněk Žabokrtský, and Anna Nedoluzhko. Two Case Studies on Translating Pronouns in a Deep Syntax Framework. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, IJCNLP 2013, Nagoya, Japan, 2013b. 36

Franz Josef Och. An efficient method for determining bilingual word classes. In *Proceedings of the Ninth Conference on European chapter of the Association for Computational Linguistics*, EACL 1999, Bergen, Norway, 1999. 45

Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, ACL 2003, Sapporo, Japan, 2003. 11, 45

Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003. 9, 45

Franz Josef Och, Daniel Gildea, Sanjeev P. Khudanpur, Anoop Sarkar, Kenji Yamada, Alexander Fraser, Shankar Kumar, Libin Shen, David A. Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir R. Radev. A Smorgasboard of Features for Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference and the 5th Meeting of the North American Chapter of the Association for Computational Linguistics*, HLT-NAACL 2004, Boston, MA, USA, 2004. 31

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176 (W0109-022), IBM Research Division, T. J. Watson Research Center, 2002. xi, 14, 31, 32, 45, 77

Alexandre Patry and Philippe Langlais. Prediction of Words in Statistical Machine Translation using a Multilayer Perceptron. In *Proceedings of the Twelfth Machine Translation Summit*, MT Summit XII, 2009. 34

Andrei Popescu-Belis, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni, and Sandrine Zufferey. Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, LREC 2012, Istanbul, Turkey, 2012. 36

Maja Popović and Hermann Ney. POS-based Word Reorderings for Statistical Machine Translation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, LREC 2006, Genoa, Italy, 2006. 30

Maja Popović and Hermann Ney. Towards Automatic Error Analysis of Machine Translation Output. *Computational Linguistics*, 37(4):657–688, 2011. 33

Oana Postolache, Dan Cristea, and Constantin Orasan. Tranferring coreference chains through word alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, LREC 2006, Genoa, Italy, 2006. 35

Long Qiu, Min yen Kan, and Tat seng Chua. A public reference implementation of the rap anaphora resolution algorithm. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, LREC 2004, Lisbon, Portugal, 2004. 35, 119

Anna N. Rafferty and Christopher D. Manning. Parsing three German treebanks: lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*, Columbus, Ohio, USA, 2008. 64

Altaf Rahman and Vincent Ng. Translation-Based Projection for Multilingual Coreference Resolution. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT 2012, Montréal, Canada, 2012. 35

Narges Sharif Razavian and Stephan Vogel. Fixed Length Word Suffix for Factored Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL 2010, Uppsala, Sweden, 2010. 37

Ralf H. Reussner, Jens Happe, and Annegret Habel. Modelling Parametric Contracts and the State Space of Composite Components by Graph Grammars. In *Proceedings of the 8th International Conference, Held As Part of the Joint European Conference on Theory and Practice of Software Conference on Fundamental Approaches to Software Engineering*, FASE 2005, Edinburgh, Scotland, 2005. 55

Kay Rottmann and Stephan Vogel. Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model. In *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation*, TMI 2007, Skövde, Sweden, 2007. 30, 46, 47, 52

Grzegorz Rozenberg. *Handbook of Graph Grammars and Computing by Graph Transformation: Volume I. Foundations*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 1997. ISBN 98-102288-48. 55

Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, England, 1994. 46

Rico Sennrich, Philip Williams, and Matthias Huck. A tree does not make a well-formed sentence: Improving syntactic string-to-tree statistical machine translation with more linguistic knowledge. *Computer Speech & Language*, 2014. 37

Libin Shen, Anoop Sarkar, and Franz Och. Discriminative Reranking for Machine Translation. In *Proceedings of the Human Language Technology Conference and the 5th Meeting of the North American Chapter of the Association for Computational Linguistics*, HLT-NAACL 2004, Boston, MA, USA, 2004. 31

Libin Shen, Jinxi Xu, Bing Zhang, Spyros Matsoukas, and Ralph Weischedel. Effective Use of Linguistic and Contextual Information for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2009, Suntec, Singapore, 2009. 34

Isabel Slawik, Mohammed Mediani, Jan Niehues, Yuqi Zhang, Eunah Cho, Teresa Herrmann, Thanh-Le Ha, and Alex Waibel. The KIT Translation Systems for IWSLT 2014. In *Proceedings of the Eleventh International Workshop on Spoken Language Translation*, IWSLT 2014, Lake Tahoe, CA, USA, 2014. 49

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, AMTA 2006, Cambridge, MA, USA, 2006. 14, 32

Artem Sokolov, Guillaume Wisniewski, and François Yvon. Computing lattice BLEU oracle scores for machine translation. In *Proceedings of 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2012, Avignon, France, 2012. 32

Wee Meng Soon, Daniel Chung Yong Lim, and Hwee Tou Ng. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27 (4):521–544, 2001. 35

Lucia Specia, Baskaran Sankaran, and Maria Graças Volpe Nunes. n-Best Reranking for the Efficient Integration of Word Sense Disambiguation and Statistical Machine Translation. *Lecture Notes in Computer Science*, 4919:399–410, 2008. 34

Andreas Stolcke. SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, ICSLP 2002, Denver, CO, USA, 2002. 10

Sebastian Stüker, Florian Kraft, Christian Mohr, Teresa Herrmann, Eunah Cho, and Alex Waibel. The KIT Lecture Corpus for Speech Translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, LREC 2012, Istanbul, Turkey, 2012. 43

Sara Stymne. Blast: A Tool for Error Analysis of Machine Translation Output. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies Conference, System Demonstrations*, ACL-HLT 2011, Portland, OR, USA, 2011. 32

Hirotoshi Taira, Katsuhito Sudoh, and Masaaki Nagata. Zero Pronoun Resolution Can Improve the Quality of J-E Translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST 2012, Jeju, Republic of Korea, 2012. 35

David Talbot, Hideto Kazawa, Hiroshi Ichikawa, Jason Katz-Brown, Masakazu Seno, and Franz J. Och. A Lightweight Evaluation Framework for Machine Translation Reordering. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT 2011, Edinburgh, Scotland, 2011. 32

Aleš Tamchyna, Fabienne Braune, Alexander M. Fraser, Marine Carpuat, Hal Daumé III, and Chris Quirk. Integrating a Discriminative Classifier into Phrase-based and Hierarchical Decoding. *Prague Bulletin of Mathematical Linguistics*, 101:29–42, 2014. 34

Christoph Tillmann. A unigram orientation model for statistical machine translation. In *Proceedings of the Human Language Technology conference / North American Chapter of the Association for Computational Linguistics Annual Meeting*, HLT-NAACL 2004, Boston, MA, USA, 2004. 29, 31, 46

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. Applying Morphology Generation Models to Machine Translation. In *Proceedings of the Annual Meeting of the*

Association for Computational Linguistics: Human Language Technologies, ACL-HLT 2008, Columbus, OH, USA, 2008. 36, 37

Ke M. Tran, Arianna Bisazza, and Christof Monz. Word Translation Prediction for Morphologically Rich Languages with Bilingual Neural Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2014, Doha, Qatar, 2014. 34

Yulia Tsvetkov, Chris Dyer, Lori Levin, and Archna Bhatia. Generating English Determiners in Phrase-Based Translation with Synthetic Translation Options. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, WMT 2013, Sofia, Bulgaria, 2013. 38

Bernard Vauquois and Christian Boitet. Automated Translation at Grenoble University. *Computational Linguistics*, pages 28–36, 1985. 5

Ashish Venugopal, Andreas Zollman, and Alex Waibel. Training and Evaluation Error Minimization Rules for Statistical Machine Translation. In *Proceedings of the Workshop on Data-drive Machine Translation and Beyond*, WPT 2005, Ann Arbor, MI, USA, 2005. 45

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. BART: A Modular Toolkit for Coreference Resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, ACL-HLT 2008, Columbus, OH, USA, 2008. 35

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. Word-sense Disambiguation for Machine Translation. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT-EMNLP 2005, Vancouver, Canada, 2005. 33

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. Error Analysis of Machine Translation Output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, LREC 2006, Genoa, Italy, 2006. 13, 32

Stephan Vogel. SMT Decoder Dissected: Word Reordering. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China, 2003. 45

Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based Word Alignment in Statistical Translation. In *Proceedings of the International Conference on Computational Linguistics*, Coling 1996, Copenhagen, Denmark, 1996. 8

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, WMT 2008, Columbus, OH, USA, 2008. 36

Chao Wang, Michael Collins, and Philipp Koehn. Chinese Syntactic Reordering for Statistical Machine Translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, EMNLP 2007, Prague, Czech Republic, 2007. 30

Ye-Yi Wang and Alex Waibel. Modeling with Structures in Statistical Machine translation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, ACL-Coling 1998, Montréal, Canada, August 1998. 29

Jochen Weiner. Pronominal Anaphora in Machine Translation. Master's thesis, Karlsruhe Institute of Technology, Karlsruhe, Germany, 2014. v, 36, 91, 92, 106, 119

Marion Weller, Max Kisselew, Svetlana Smekalova, Alexander Fraser, Helmut Schmid, Nadir Durrani, Hassan Sajjad, and Richárd Farkas. Munich-Edinburgh-Stuttgart Submissions at WMT13: Morphological and Syntactic Processing for SMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, WMT 2013, Sofia, Bulgaria, 2013. 37, 38

Philip Williams and Philipp Koehn. Agreement Constraints for Statistical Machine Translation into German. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT 2011, Edinburgh, Scotland, 2011. 37

Guillaume Wisniewski, Alexandre Allauzen, and François Yvon. Assessing phrase-based translation models with oracle decoding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2010, Cambridge, MA, USA, 2010. 32

Fei Xia and Michael McCord. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the International Conference on Computational Linguistics*, Coling 2004, Geneva, Switzerland, 2004. 30

Deyi Xiong, Qun Liu, and Shouxun Lin. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. In *Proceedings of the Annual Meeting of*

the Association for Computational Linguistics, ACL 2006, Sydney, Australia, 2006. 29

Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, ACL 2001, Toulouse, France, 2001. 12, 30

Yuqi Zhang, Richard Zens, and Hermann Ney. Chunk-Level Reordering of Source Language Sentences with Automatically Learned Rules for Statistical Machine Translation. In *Proceedings of the Workshop on Syntax and Structure in Statistical Translation*, SSST 2007, Rochester, NY, USA, 2007. 30

Andreas Zollmann and Ashish Venugopal. Syntax augmented machine translation via chart parsing. In *Proceedings of the Workshop on Statistical Machine Translation*, WMT 2006, New York City, NY, USA, 2006. 31