Improving Statistical Machine Translation using Morpho-syntactic Information

Von der Fakultät für Mathematik, Informatik
und Naturwissenschaften der
Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades einer
Doktorin der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Diplom–Informatikerin
Sonja Nießen

aus Geilenkirchen

Berichter: Universitätsprofessor Dr.-Ing. Hermann Ney
Professor Dr. Enrique Vidal

Tag der mündlichen Prüfung: 2. Dezember 2002

Acknowledgments

This thesis is based on work carried out during my time as a research scientist at the Department for Computer Science at the University of Technology in Aachen, Germany.

First, I would like to express my gratitude to my advisor Professor Dr.-Ing. Hermann Ney, head of the Lehrstuhl für Informatik VI at the University of Technology in Aachen. His advice, his continuous interest, and his support made this thesis ultimately possible. I would also like to thank my second advisor Professor Dr. Enrique Vidal, from the Departamento de Sistemas Informáticos Y Computación at the Universidad Politecnica de Valencia, for his interest in this work and the valuable comments on the early drafts of this thesis.

Special thanks go to Gregor Leusch and Richard Zens for their valuable programming work. All the other people at the Lehrstuhl für Informatik VI are also deserving of my thanks for many fruitful discussions, and for the very good working atmosphere. Some of my colleagues became real friends joining me in my “ups” and helping me through the inevitable “downs”.

Thanks to the Nespole! consortium, listed on the project’s homepage [Nespole! 00], for making available part of the Nespole! data. Special thanks to Alon Lavie, Lori Levin, Stephan Vogel and Alex Waibel (in alphabetical order).

I am very grateful to my parents, Ingeborg Nießen and Karl-Heinz Nießen, who are always there for me and who teached me honesty and perseverance.

Finally, I would like to thank Torsten Bausch for his understanding, love and support.

Aachen, February 2003. Sonja Nießen

Abstract

In the framework of statistical machine translation, correspondences between the words in the source and the target language are learned from bilingual corpora, and often little or no linguistic knowledge is used to structure the underlying models. The work presented in this thesis is motivated by the well-known observation that training data typically does not sufficiently represent the range of phenomena in natural languages. In this thesis, various methods of incorporating morphological and syntactic information into systems for statistical machine translation are proposed and systematically assessed. The overall goal is to improve translation quality and to reduce the amount of parallel text necessary to train the model parameters. The development of the suggested methods is guided by the analysis of important causes of errors.

Large differences in word order between corresponding sentences are difficult to capture for automatic alignment algorithms. In this work, a range of sentence level restructuring transformations is introduced, which are motivated by knowledge about the sentence structure in the involved languages. These transformations aim at the assimilation of word orders in related sentences. A detailed analysis of the effect on the corpora and the translation quality reveals that their application results in better alignments and as a consequence in less noisy probabilistic lexica, broader applicability of multi-word phrase pairs and a better coverage of the language model.

Existing statistical systems for machine translation often treat different inflected forms of the same lemma as if they were independent of each other. A better exploitation of the bilingual training data can be achieved by explicitly taking into account the interdependencies of the related inflected forms. In this work a hierarchy of equivalence classes is defined on the basis of morphological and syntactic information beyond the surface forms. Features from those hierarchy levels are combined to form hierarchical lexicon models which can replace the standard probabilistic lexicon used in most statistical machine translation systems. The benefit from these combined models is twofold: Firstly, the lexical coverage is improved, because the translation of unseen word forms can be derived by considering information from lower levels in the hierarchy. Secondly, category ambiguity can be resolved, because syntactical context information is made locally accessible by means of annotation with morpho-syntactic tags.

Conventional bilingual dictionaries are often used as additional data to better train the model parameters. One of the disadvantages of these dictionaries as compared to full bilingual corpora is the fact that their entries typically contain no context to enable the distinction between the translations for different readings of a word. In this work a method for aligning corresponding readings in conventional dictionaries containing pairs of fully inflected word forms is proposed. The approach uses information deduced from one language side to resolve category ambiguity in the corresponding entry in the other language. The resulting disambiguated dictionaries are better suited for improving the quality of machine translation, especially if they are used in combination with the hierarchical lexicon models.

It is a costly and time consuming task to gather large texts and have them translated to form bilingual corpora suitable for training the model parameters for statistical machine translation. In this work the amount of bilingual data required to achieve an acceptable quality of machine translation is systematically investigated. All the methods presented in this thesis contribute to a better exploitation of the available bilingual data and thus to improving translation quality in frameworks with scarce resources.

The combination of the suggested methods results in substantial improvements on the Verbmobil task, the Nespole! task and the Zeres task, for German to English and English to German translation and for text input and on the output of a speech recognizer.

The second focus of this thesis is on evaluation of machine translation quality. A tool for the evaluation of translation quality which accounts for the specific requirements in a research environment is developed. Evaluation criteria which are more adequate than pure edit distance are defined. The measurement along these quality criteria is performed semi-automatically in a fast, convenient and consistent way using the tool and the corresponding graphical user interface. The quality criteria themselves are systematically assessed.

Zusammenfassung

Bei der statistischen maschinellen Übersetzung wird die Korrespondenz von Wörtern in der Quell- und der Zielsprache anhand von bilingualen Corpora gelernt, und häufig geht wenig oder gar kein linguistisches Wissen zur Strukturierung der zugrundeliegenden Modelle ein. Die hier dargestellte Arbeit ist motiviert durch die weithin bekannte Beobachtung, dass das Trainingsmaterial typischerweise die Bandbreite der Eigenheiten natürlicher Sprachen nicht ausreichend widerspiegelt. Es werden verschiedene Methoden zur Einbettung morphologischer und syntaktischer Information in statistische Übersetzungssysteme vorgestellt und systematisch getestet. Ziel ist allgemein die Verbesserung der Übersetzungsqualität und die Verringerung der zum Training der Modellparameter notwendigen Datenmenge. Die Entwicklung der vorgeschlagenen Methoden ist ausgerichtet an der Analyse vorherrschender Fehlerursachen.

Es ist schwierig für Alignierungsalgorithmen, größere Unterschiede in der Wortstellung zwischen einander entsprechenden Sätzen zu behandeln. In dieser Arbeit wird eine Reihe von Umordnungsoperationen auf Satzebene eingeführt, die auf Wissen über die Satzstruktur in den beteiligten Sprachen fußen. Zweck dieser Transformationen ist es, verwandte Sätze einander anzugleichen. Eine detaillierte Analyse der Auswirkung auf Corpora und Übersetzungsergebnisse lässt darauf schließen, dass ihre Anwendung zu besseren Wortalignments führt und folglich zu weniger verrauschten probabilistischen Lexika, zu breiterer Anwendbarkeit von Mehrwortphrasen und zu einer besseren Abdeckung durch das Zielsprachmodell.

Die existierenden statistischen Übersetzungssysteme betrachten verschiedene Wortformen des gleichen Lemmas als unabhängig voneinander. Das bilinguale Trainingsmaterial kann durch explizite Einbeziehung der wechselseitigen Abhängigkeiten verwandter Wortformen besser ausgeschöpft werden. In dieser Arbeit wird eine Hierarchie von Äquivalenzklassen definiert, die auf morphologischer und syntaktischer Information über die Oberflächenformen hinaus beruht. Durch die Kombination von Merkmalen aus den verschiedenen Hierarchieebenen werden hierarchische Lexikonmodelle gebildet, die die in den meisten statistischen Übersetzungssystemen üblichen probabilistischen Lexika ersetzen können. Diese kombinierten Modelle haben einen zweifachen Nutzen: Erstens verbessern sie die Vokabularabdeckung, da die Übersetzungen für ungesehene Wortformen aus Informationen hergeleitet werden können, die von tieferen Ebenen in der Hierarchie stammen. Zum Zweiten können Kategorie-Mehrdeutigkeiten aufgelöst werden, weil die Annotation mit morpho-syntaktischen Markierungen syntaktische Kontextinformation lokal zugreifbar macht.

Konventionelle bilinguale Lexika dienen häufig als zusätzliches Datenmaterial für das Training der Modellparameter. Gegenüber den parallelen Texten haben sie unter Anderem den Nachteil, dass ihre Einträge typischerweise keine Kontextinformation enthalten, anhand derer eine Unterscheidung zwischen den Übersetzungen der verschiedenen Lesarten einer Wortform möglich wäre. In dieser Arbeit wird eine Methode zur Zuordnung einander entsprechender Lesarten für konventionelle Lexika vorgestellt, die Paare von Wortvollformen enthalten. Dazu wird aus einer Lexikonseite abgeleitete Information zur Auflösung von Kategorie-Mehrdeutigkeit im entsprechenden Eintrag der anderen Lexikonseite verwendet. Die resultierenden disambiguierten Lexika eignen sich insbesondere in Kombination mit den hierarchischen Lexikonmodellen besser zur Verbesserung der Übersetzungsqualität.

Das Sammeln von Texten und ihre Übersetzung zum Zweck der Bereitstellung bilingualer Korpora für das Training der Modellparameter ist mühsam und teuer. In dieser Arbeit wird systematisch untersucht, wieviel paralleles Trainingsmaterial notwendig ist, um eine akzeptable Übersetzungsqualität zu erreichen. Alle hier dargestellten Methoden tragen zu einer verbesserten Ausschöpfung der bilingualen Daten und folglich zu einer Verbesserung der Übersetzungsqualität insbesondere im Fall von knappen Datenresourcen bei.

Durch die Kombination der vorgeschlagenen Methoden können erhebliche Verbesserungen nachgewiesen werden für die Aufgaben Verbmobil, Nespole! und Zeres, und zwar für die Übersetzung von Deutsch nach Englisch und umgekehrt, bei Texteingabe und bei Übersetzung der Ausgabe eines Spracherkenners.

Thema des zweiten Teils dieser Arbeit ist die Bewertung von Übersetzungen. Es wurde ein Werkzeug für diesen Zweck entwickelt, das den spezifischen Anforderungen im Forschungsumfeld Rechnung trägt. Des weiteren wurden Evaluationsmaße definiert, die angemessener sind als der einfache Editierabstand. Die Bewertung anhand dieser Kriterien erfolgt halbautomatisch auf schnelle, bequeme und konsistente Weise unter Verwendung des Werkzeugs und der zugehörigen graphischen Benutzeroberfläche. Die Qualitätsmaße selbst werden systematisch verglichen.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> ...the history of MT shows, to me at least, the truth of two (barely compatible) principles that could be put crudely as 'Virtually any theory, no matter how silly, can be the basis of some effective MT' and 'Successful MT systems rarely work with the theory they claim to.'
>
> (Yorick Wilks, 1989)

The statistical approach to machine translation has been justified by various successful comparative evaluations ever since its revival by the work of the famous IBM research group more than a decade ago. The considerable interest and more particularly the remarkable reluctance their earliest publications have produced might be due to the complete neglect of the linguistic approach dominating at that time and especially the fact that they consequently dispensed with linguistic analysis, at least in their earliest publications. Although also the IBM group finally made use of morphological and syntactic information to enhance translation quality [Brown & Della Pietra$^+$ 92, Berger & Brown$^+$ 96b], most of today's statistical machine translation (SMT) systems follow the tradition of considering only surface word forms and of not using linguistic knowledge about the structure of the involved languages.

The aim of this work is to incorporate information from morphological and syntactic analysis into SMT in order to better exploit high quality data for training model parameters and ultimately to improve translation quality. The procedure pursued to do so and proposed for future investigations along this line is the following: Analyze dominant sources of errors, like for instance difference in sentence structure, ambiguity, poor coverage of the vocabulary as is particularly caused by inflectional and compositional morphology, idiomatic expressions, etc. These are *real* problems in contrast to some other questions dealt with in linguistics, which are often of high theoretical but little practical interest. Then, identify knowledge sources with a potential to meet these problems and integrate them into a corpus based approach.

The methods for problem solving may be (a) largely language independent, like the suggestions in [Och & Tillmann$^+$ 99] or [Wang & Waibel 98] about phrase-level alignments or the work of [Wu 96] and [Yamada & Knight 01] who provide a framework for learning basic language pair specific differences in sentence structure; or (b) they can be language type specific, as is for instance the work in this thesis and in some of the related publications cited below which deal with characteristics of languages with rich inflectional and compositional morphology; or they can be (c) language specific, like for example the treatment of German prefix verbs described in this work.

After briefly reviewing the basic concepts of SMT, the remainder of this chapter describes the state of the art and related work as regards the incorporation of morphological and syntactic information into systems for natural language processing as well as the second focus of this thesis, namely the evaluation of machine translation quality in the framework of machine translation research.

1.1 Statistical machine translation

The goal of the translation process in statistical machine translation can be formulated as follows: Every target language string $e_1^I = e_1 \ldots e_I$ is assigned a probability $Pr(e_1^I)$ of being a valid word sequence in the target language and a probability $Pr(e_1^I|f_1^J)$ of being an admissible translation for the given source language string $f_1^J = f_1 \ldots f_J$. According to Bayes' decision rule, the optimal translation for f_1^J is the target string that maximizes the product of the target language model $Pr(e_1^I)$ and the string translation model $Pr(f_1^J|e_1^I)$.[1] Many existing systems for statistical machine translation [García-Varea & Casacuberta 01, Germann & Jahr$^+$ 01, Nießen & Vogel$^+$ 98, Och & Tillmann$^+$ 99] make use of a special way of structuring the string translation model [Brown & Della Pietra$^+$ 93b]: The correspondence between the words in the source and the target string is described by alignments that assign target word positions to each source word position. The probability of a certain target language word to occur in the target string is assumed to depend basically only on the source words aligned to it. The overall architecture of the statistical translation approach is depicted in Figure 1.1. This figure already anticipates the fact that the source strings will be transformed in a certain manner. If necessary the inverse of these transformations are also applied to the produced output strings. In Chapter 4 the applied transformations are explained in detail.

1.2 State of the art and related work

1.2.1 Incorporation of morphological and syntactic information

Although there has been a number of publications dealing with morphological and syntactic analysis in general and its application to machine translation in particular, there

[1] The notion of a "noisy channel model", which is at the basis of the argumentation for using the "inverted" probability $Pr(f_1^J|e_1^I)$ instead of $Pr(e_1^I|f_1^J)$ sometimes leads to notational confusion: [Wang 98] for example calls the language of the input to the noisy channel "source language" and the language of the output "target language", while in fact the output language of the channel is the language of the sentences presented as input to a translation system. To avoid confusion the notation in this work as well as in many other publications about statistical machine translation is rather task-oriented.

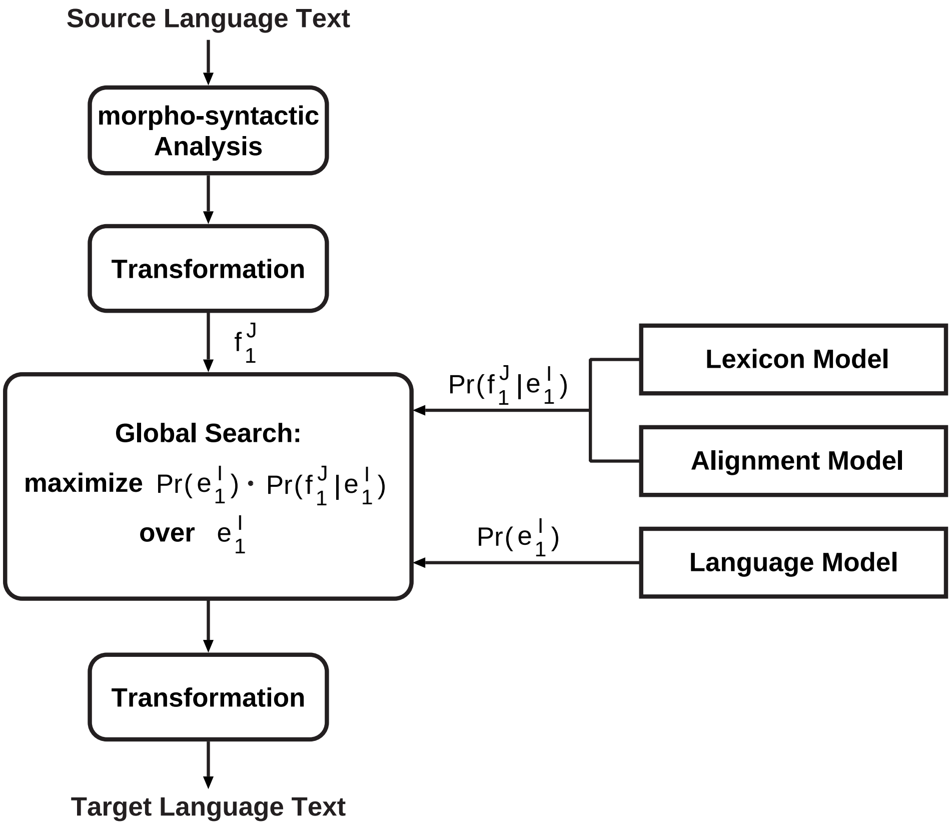


Figure 1.1: Architecture of the translation approach based on Bayes' decision rule.

have only been few which incorporate information from this analysis in the process of *statistical* machine translation.

Morphology:
Some publications have already dealt with the treatment of morphology in the framework of language modeling and speech recognition: [Kanevsky & Roukos$^+$ 97] propose a statistical language model for inflected languages. They decompose word forms into stems and affixes. [Maltese & Mancini 92] report that a linear interpolation of word-n-grams, POS-n-grams, and lemma-n-grams yields lower perplexity than pure word based models. [Larson & Willett$^+$ 00] apply a data-driven algorithm for decomposing compound words in compounding languages as well as for recombining phrases to enhance the pronunciation lexicon and the language model for large vocabulary speech recognition systems.

As regards machine translation, the treatment of morphology is part of the analysis and generation step in virtually any machine translation system based on the "classical" symbolic approach, at least when languages with some inflectional morphology (thus also English, for example) are involved and a more than experimental task is envisioned [Hutchins & Somers 92]. For this purpose the lexicon should contain base forms of words and the grammatical category, sub-categorization features and semantic information in order to enable a reduction of the size of the lexicon and in order to account for unknown word forms, that is word forms not present explicitly in the dictionary.

Virtually all of today's *statistical* machine translation (SMT) systems are based on or at least inspired by the pioneering work of a research group at the IBM Research Laboratories at Yorktown Heights, N.Y., in the late eighties and early nineties of the twentieth century [Brown & Cocke$^+$ 88, Brown & Cocke$^+$ 90, Brown & Della Pietra$^+$ 93b], and most of the current SMT research groups follow the strict interpretation of the data-driven approach in that they do not involve more than rather basic linguistic analysis or generation of sentences. The underlying (probabilistic) lexicon typically only contains pairs of full-forms. On the other hand, already [Brown & Della Pietra$^+$ 92] suggested to annotate word forms with morpho-syntactic information, but they did not perform any investigation on the effects.

Treatment of structural differences:
Many symbolic machine translation systems perform (more or less local) rearranging of the word order in the target language as a step of the generation process [Hutchins & Somers 92]. [Gamon & Reutter 97] and [Wolters 97] for example report on methods for treating German separable prefix verbs within the framework of symbolic natural language processing systems. For *statistical* machine translation systems it has proven beneficial to move the restructuring step to the other end, namely to apply (normally very local) reordering operations to the *source* language sentences, in order to make sentences in the two languages more similar to one another. For the language pair English and French, [Brown & Della Pietra$^+$ 92] have suggested some local word reordering transformations. They also introduce the notion of question inversion treatment (see Section 4.1.1). Unfortunately they did not report on experimental results revealing the effect of the reordering on the translation quality. [Wang 98] mentions some simple preprocessing operations, among other things splitting of compound words for which information about possible split points is available from a German dictionary within the Verbmobil project. [Och 02] uses pattern replacement tables based on regular expressions to perform preprocessing and postprocessing, e.g. splitting some German compound words which are potential sources of translation mistakes.

Translation with scarce resources:
Some recent publications have dealt with the problem of translation with scarce resources, like [Al-Onaizan & Germann$^+$ 00]. They report on an experiment of Tetun–to–English translation by different groups, including one using statistical machine translation. [Al-Onaizan & Germann$^+$ 00] assume the absence of linguistic knowledge sources such as morphological analyzers and dictionaries. Nevertheless, they found that human mind is very well capable of deriving dependencies such as morphology, cognates, proper names, spelling variations etc., and that this capability was finally at the basis of the better results produced by humans compared to corpus based machine translation. The additional information results from complex reasoning and it is not directly accessible from the full word form representation in the data.

This work takes a different point of view: Even if full bilingual training data is scarce, monolingual knowledge sources like morphological analyzers and data for training the target language model as well as conventional dictionaries (one word and its translation(s) per entry) may be available and of substantial usefulness for improving the performance of statistical translation systems. This is especially the case for highly inflected 'major' languages like German. The use of dictionaries to augment or replace parallel corpora has already been examined by [Brown & Della Pietra^{+} 93a] and [Koehn & Knight 01] for instance.

1.2.2 Evaluation of machine translation

Research in machine translation suffers from the lack of suitable, consistent, and easy-to-use criteria for the evaluation of the experimental results. The question of how the performances of different translation systems on a certain corpus can be compared or how the effects of small changes in the system prototypes can be judged in a fast and cheap way is still open. In the recent years, efforts in the field of the evaluation of translation quality have focused on measuring the suitability of a certain translation program as part of a distinct natural language processing task [White & Taylor 98, Sparck Jones & Galliers 96]. [ten Hacken 01] even postulates it to be an indicator for a paradigm change, that modern machine translation systems are evaluated according to the *success of communication*, as opposed to the traditional view reflected by [Nirenburg 87], who formulates *meaning preservation* as criterion for judging machine translation quality.

Evaluation methods, which are 'ideal' in the sense that they involve measuring the effect on the recipient as for instance described in [Tessiore & v. Hahn 00] should be applied at crucial points in the design and validation of machine translation systems, but they are too time-consuming and too expensive to help the daily work in the framework of machine translation research. When researchers compare the performances of different translation systems on a certain corpus or when they are interested in the effects of small changes in the system prototypes, they typically stick to inspecting individual sentence pairs and measuring meaning preservation. In order to do so, they have the choice between purely automatic measures and measures, which involve human judgment. In the following, some typical examples of these categories are listed.

Automatically computable:

Word Error Rate (WER): The edit distance $d(t, r)$ (minimal number of insertions, deletions and substitutions) between the produced translation t and one predefined reference translation r is calculated on the basis of a Levenshtein alignment [Levenshtein 65]. The edit distance has the main advantage to be automatically computable, and as a consequence, the results are inexpensive to get and reproducible, because the underlying data and the algorithm are always the same. The main disadvantage of the WER is the fact that it depends fundamentally on the choice of the sample translation. In machine translation this criterion is used e.g. in [Vidal 97], and [Tillmann & Vogel^{+} 97].

BLEU: [Papineni & Roukos^{+} 01] have proposed a method of automatic machine translation evaluation, which they call "BLEU", or "**Bil**ingual **E**valuation **U**nderstudy". It is based on the notion of modified n-gram precision, with $n \in \{1, \ldots, 4\}$: All candidate unigram, bigram, trigram and four-gram counts are collected and clipped against their corresponding maximum reference counts. The reference n-gram counts are calculated on a corpus of reference translations for each input sentence. These clipped candidate counts are summed and normalized by the total number of candidate n-grams. The geometric mean of the modified precision scores for a test corpus is calculated and multiplied by an exponential brevity penalty factor to penalize too short translations. [Papineni & Roukos$^+$ 01] state that their measure captures adequacy as well as fluency.

Involving human interaction:

Rate of not acceptable translations: The translations are scored by classification into a small number of quality classes, ranging from "perfect" to "absolutely wrong". In comparison to the WER, this criterion is more liable and conveys more information, but to perform the ranking is expensive, as it is not computed automatically but is the result of laborious evaluation by human experts. Besides, the results depend highly on the persons performing the evaluation and hence, the comparability of results is not guaranteed. Another disadvantage is the fact that the length of the sentences is not taken into account: The score of the translation of a long sentence has the same impact on the overall result as the score of the translation of a one-word sentence. The manual ranking is used e.g. in [Nießen & Vogel$^+$ 98].

The BLEU score is the only measure used in this thesis that measures *accuracy*: The others all measure *error rates*. Thus, high BLEU scores are better, and high WER figures, for example, are worse.

Chapter 2

Scientific Goals

> With some knowledge of PL/I, I am tempted to say that machine translation from Chinese to English is not only possible, it is also easy to program. . . . With more sophisticated linguistic data of both source and target languages, more efficient programming language, and bigger and faster computers, machine translation can be a reality in the near future.
>
> (Ching-Yi Dougherty, 1969)

The main objective of this thesis is the incorporation of morphological and syntactic information into statistical machine translation in order to improve the translation quality and to reduce the necessary amount of bilingual training data. The second focus is on evaluation of machine translation in a machine translation research environment.

Morpho-syntactic information for statistical machine translation

In the framework of statistical machine translation, correspondences between the words in the source and the target language are learned from bilingual corpora on the basis of so-called alignment models. Many of the statistical systems use little or no linguistic knowledge to structure the underlying models, but training data is often not large enough to sufficiently represent the range of phenomena in natural languages. It is thus reasonable to assume that statistical machine translation can take advantage of the explicit introduction of some knowledge about the languages under consideration. The methods of incorporating morpho-syntactic information described in this thesis are not orthogonal: they intersect and complement each other.

Treatment of structural differences:
Statistical alignment models are designed to capture the differences in word order in different languages. Although they are astonishingly successful in fulfilling this task, still the difference in word order is one of the main sources of errors in statistical machine translation. Thus it is promising to examine transformations which aim at "harmonizing" the word order in corresponding sentences on the basis of some linguistic knowledge about the sentence structure. The literature about statistical machine translation systems men-

tions the use of such restructuring operations, typically by means of simple replacement operations predefined by hand for a given set of words or phrases and designed to perform local rearrangements, e.g. in [Och 02] and [Wang 98]. To my knowledge, no systematic investigation on the effect of this kind of operations on the translation quality has yet been performed. In this work, a range of different types of restructuring operations is suggested, which allow for preprocessing even in the case of open vocabularies, that is when word forms not occurring in the training data are expected for the testing phase. Besides, some of the transformations are not restricted to local phenomena but operate on distant parts of the sentence. Unlike the aforementioned publications or also the publication of [Brown & Della Pietra$^+$ 92] about the pioneering IBM translation system, a detailed analysis of the effect on the corpora and the translation results is carried out. The work presented here is focused on aspects of restructuring for the language pair German and English, namely question inversion, separated verb prefixes, compound words, and multi-word phrases.

Hierarchical lexicon models:
Existing statistical systems for machine translation often treat different inflected forms of the same lemma as if they were independent of each other. A better exploitation of the bilingual training data can be achieved by explicitly taking into account the interdependencies of the related inflected forms. In this work a hierarchy of equivalence classes at different levels of abstraction is proposed. Features from those hierarchy levels are combined to form hierarchical lexicon models which can replace the standard probabilistic lexicon used in most statistical machine translation systems. Apart from the improved coverage, the proposed lexicon models enable the disambiguation of ambiguous word forms by means of annotation with morpho-syntactic tags.

Disambiguation of conventional bilingual dictionaries without context:
Conventional bilingual dictionaries are often used as additional data to better train the model parameters of statistical machine translation. One of the disadvantages of these dictionaries as compared to full bilingual corpora is the fact that their entries typically contain no context to enable the distinction between the translations for different readings of a word. In order to make dictionaries more valuable for natural language processing, some of the work presented in this thesis is devoted to resolving ambiguities using clues from corresponding entries on both language sides. The applicability of the resulting disambiguated dictionaries is not restricted to statistical machine translation: It is reasonable to expect that they would be useful for many other fields and approaches for natural language processing, for instance multi-lingual information retrieval and document classification.

Translation with scarce resources:
It is a costly and time consuming task to gather large texts and have them translated to form bilingual corpora suitable for training the model parameters for statistical machine translation. In this work the amount of bilingual data necessary to sufficiently cover the vocabulary expected in testing is investigated. One of the objectives of this thesis is to

introduce morphological knowledge in order to reduce this amount by enabling a better exploitation of the available language resources.

Evaluation of machine translation

Developers of machine translation systems are interested in the effects of small changes in the system prototypes, that is, they have a need for quick, frequent and inexpensive evaluations. Another important aspect of evaluation in a research environment is the fact that variants of system prototypes are frequently tested on few distinct sets of test sentences, and that the results often differ only in a small number of words.

A tool for evaluating translations:
A tool for the evaluation of translation quality which accounts for these requirements and these basic conditions is developed. It aims at facilitating the manual work of human evaluators and to help maintaining consistency. It also allows for the calculation of automatically computable measures as well as the estimation of quality scores.

Definition and comparison of evaluation measures:
The quality of machine translation can be measured along many different directions and with different granularities, and this measurement itself can be more or less costly and time consuming. One of the objectives of this work is firstly to define evaluation measures which are suitable for a machine translation research environment, and secondly to compare these different evaluation measures.

The presentation of the work is organized as follows: Chapter 3 describes the items of information provided by morpho-syntactic analysis and introduces a suitable representation of the analyzed corpus. In Chapter 4 the aforementioned restructuring operations are suggested as well as an algorithm for mapping unseen word forms to more abstract forms of the same lemma and automatic corrections of the translation output. The introduction of log-linear hierarchical lexicon models in Chapter 5 is prepared by a series of pre-studies. A method for disambiguating conventional dictionaries is proposed and finally, a procedure for combining all methods suggested in this thesis is presented which is especially well-suited for the case of scarce resources for training the model parameters. The second main focus of this work, the assessment of machine translation quality, is the topic of Chapter 6, which suggests a tool for semi-automatic evaluation. A range of evaluation measures is defined and compared. Experimental results are reported in Chapter 7. The Chapters 8 and 9 conclude the presentation with a discussion of the achievements of this work and an outlook on possible future directions.

The Appendices A and B review the task setups and the performance measures used in the experiments. Appendix C lists the definitions of symbols and acronyms and Appendix D describes a program for processing the output of morpho-syntactic analyzers.

Chapter 3

Representation of Morpho-syntactic Information

Every time I fire a linguist, my system's performance improves.

(Peter F. Brown, delivered statement)

A prerequisite for the methods for improving the quality of statistical machine translation described in this work is the availability of various kinds of morphological and syntactic information. The construction of analyzers which can provide this information is a demanding task on its own. Fortunately, there are high quality commercial constraint grammar parsers[1] available for lexical analysis and morphological and syntactic disambiguation of German and English. The information obtained from these analyzers is described in Section 3.2. A program called "`process_cg`" has been implemented to parse and process the output of these analyzers. Appendix D contains a description of its functionality. After reviewing some basic concepts of morphology, this chapter describes the output resulting from these tools and explains which parts of the analysis are used and how the output is represented for further processing.

3.1 Basic concepts of morphology

Linguists distinguish three different processes of word formation [Hausser 01]:

Inflection is a systematic variation adapting a word to the syntactic environment and marking it with a syntactic function. An example is the variation "goes" of the base form "go". In German there are about ten inflectional forms per base form.

Derivation is a variation using a prefix or a suffix attached to a free word form, like in "*un*comfortable" or "joy*ful*".

Compounding is the construction of a new word by appending two or more free word forms, like in the German "Tennisplatz" or the English "fisherman".

[1] For a description of the constraint grammar approach the reader is referred to [Karlsson 90].

Specifically the process of compounding leads to generally infinite vocabularies in languages like German: New word forms can be generated any time on demand and will be understood spontaneously. Verbs, nouns and adjectives and adverbs are subject to these three morphological processes which all result in continuous fluctuation in these word classes. For this reason these three parts of speech are called "*open word classes*" in contrast to the "*closed word classes*" of conjunctions, prepositions and articles.

3.2 Description of the analysis results

For obtaining the required morpho-syntactic information, commercial analyzers were applied. In detail, the following analyzers for German and English have been purchased from the Finnish software company *Lingsoft*: "gertwol" and "engtwol" for lexical analysis and "gercg" and "engcg" for morphological and syntactic disambiguation. Tables 3.1 and 3.2 give examples of the information provided by these tools:

- The base form, i.e. the word form that typically serves as a primary key to a lemma in a dictionary: For nouns, this is the singular nominative form, for verbs the present infinitive and for adjectives and adverbs the indefinite adverbial form. Of course the base form does not differ from the actual word form for words in the closed word classes (see Section 3.1). Sometimes, the base form is in lower case, whereas the original word form is in upper case, especially at the beginning of a sentence. As has been mentioned before, the corpora used for training are typically true–case–converted and consequently, case differences between inflected word form and assigned base form are very rare.

- The part of speech, e.g. *verb*, *noun*, *preposition*, *article* etc.

- The number, i.e. *singular* or *plural*.

Table 3.1: Sample analysis of a German sentence. Input: "Wir wollen nach dem Abendessen nach Essen aufbrechen." ("We want to start for Essen after dinner.")

Original	**Base Form**	**Tags**
Wir	wir	personal pronoun plural first nominative
wollen	wollen	verb indicative present plural first
nach	nach	preposition dative
dem	das	definite article singular dative neuter
Abendessen	Abend#essen	noun neuter singular dative
nach	nach	preposition dative
Essen	Essen Esse Essen Essen	noun name neuter singular dative noun feminine plural dative noun neuter plural dative noun neuter singular dative
aufbrechen	auf|brechen	verb separable infinitive

Table 3.2: Sample analysis of an English sentence.

Original	**Base Form**	**Tags**
`Do`	`do`	`verb present not-singular-third finite auxiliary`
`we`	`we`	`personal pronoun nominative plural first subject`
`have`	`have`	`verb infinitive not-finite main`
`to`	`to`	`infinitive marker`
`reserve`	`reserve`	`verb infinitive not-finite main`
`rooms`	`room`	`noun nominative plural object`

- The person, i.e. *first*, *second* or *third.*
- The case, i.e. *nominative*, *genitive*, *dative* or *accusative.*
- The gender, i.e. *masculine*, *feminine* or *neuter.*
- The tense, e.g. *present* or *past.*
- For verbs in the English sentence, whether they are *infinite* or *finite* and whether they are the *main* or an *auxiliary* verb of the sentence.
- For nouns and pronouns in the English sentence, whether they are *subject* or *object* of the sentence. This information can often not be provided unambiguously.
- For German words, especially nouns, whether they are compounds, that is words consisting of multiple components.
- For German verbs, whether it is a separable verb, that is a verb with a detachable prefix. Besides the ending position of the detachable prefix is indicated.
- "Weak" and "strong" split points between components are indicated with a '|' or a '#', respectively. Weak split points often mark derivations, like in the German "be|sprechen", whereas strong split points separate free word forms in compounds.
- Some frequent multi-word phrases are merged together with a separating '=' or '_', when they jointly fulfill a syntactic function: The phrase "irgend=etwas" ("anything") for example may form either an indefinite determiner or an indefinite pronoun. The phrase "in=order=to" is another example.
- Some additional information like the identification of a proper name etc.

3.3 Treatment of ambiguity

The examples in Tables 3.1 and 3.2 demonstrate the capability of the tools to disambiguate between different readings: For instance, they infer that the word "wollen" is a verb in the indicative present first person plural form. Without any context taken into account, "wollen" has other readings. It can even be interpreted as derived from an adjective with the meaning "made of wool". The inflected word forms on the German part of the Verbmobil (cf. Section A.1) corpus have on average 2.85 readings (1.86 for the English corpus), 58% of which can be eliminated by the syntactic analyzers on the basis of sentence context.

Common bilingual corpora normally contain full sentences which provide enough context information for ruling out all but one reading for an inflected word form. To reduce the remaining uncertainty, preference rules have been implemented. For instance, it is assumed that the corpus is correctly true–case–converted beforehand and as a consequence, non-noun readings of uppercase words are dropped. Besides, indicative verb readings are preferred to subjunctive or imperative. In addition, some simple domain specific heuristics are applied. The reading "plural of Esse" for the German word form "Essen" for instance is much less likely in the domain of appointment scheduling and travel arrangements than the readings "proper name of the town Essen" or the German equivalent of the English word "meal". As can be seen in Table 3.3, the reduction in the number of readings resulting from these preference rules is fairly small in the case of the Verbmobil corpus.

The remaining ambiguity often lies in those parts of the information which are not used or which are not relevant to the translation task anyway. An example for a frequent type of unresolved but also unimportant ambiguity is the distinction between different cases for some words in the German sentence: Relatively often, the analyzers cannot tell accusative from dative readings, but the case information is not essential for the translation task (see also Table 3.5). Section 3.5 describes a method for selecting morpho-syntactic tags considered relevant for the translation task, which results in a further reduction in the number of readings per word form to 1.06 for German and 1.01 for English. In these rare cases of ambiguity it is admissible to resort to the unambiguous parts of the readings, i.e. to drop all tags causing mixed interpretations. Table 3.3 summarizes the gradual resolution of ambiguity.

The analysis of conventional dictionaries poses some special problems, because they do not provide enough context to enable effective disambiguation. For handling this special situation dedicated methods have been implemented, which are presented in Section 5.3.

Table 3.3: Resolution of ambiguity on the Verbmobil corpus.

	#readings per word form	
disambiguation	German	English
none	2.85	1.86
by context	1.20	1.02
by preference	1.19	1.02
by selecting relevant tags	1.06	1.01
by resorting to unambiguous part	1.00	1.00

3.4 The lemma–tag representation

A full word form is represented by the information provided by the morpho-syntactic analysis: From the interpretation "`gehen verb indicative present first singular`", i.e. the base form plus part of speech plus the other tags, the word form "`gehe`" can be restored. It has already been mentioned that the analyzers can disambiguate between different readings on the basis of context information. In this sense, the information inherent in the original word forms is augmented by the disambiguating analyzer. This can be useful for choosing the correct translation of ambiguous words. Of course, these disambiguation clues result in an enlarged vocabulary. The vocabulary of the new representation of the German part of the Verbmobil corpus, for example, where full word forms are replaced by base form plus morphological and syntactic tags (in the following denoted by *lemma–tag representation*), is larger by a factor of 1.5 than the vocabulary of the original corpus. On the other hand, it is a characteristic feature of the lemma–tag representation that the information can gradually be accessed and finally be reduced: For example certain instances of words can be considered equivalent. This fact is used to better exploit the bilingual training data along two directions: Detecting and omitting unimportant information (see Section 3.5) and constructing hierarchical translation models (see Chapter 5). To summarize, the lemma–tag representation of a corpus has the following main advantages: It makes context information locally available and it allows for explicitly accessing information at different levels of abstraction.

3.5 Equivalence classes of words with similar translation

Inflected word forms in the input language often contain information that is not relevant for translation. This is especially true for the task of translating from a highly inflected language like German into English, for instance: In parallel German/English corpora, the German part contains many more distinct word forms than the English part (see for example Table A.1). It is useful for the process of statistical machine translation to define equivalence classes of word forms which tend to be translated by the same target language word: the resulting statistical translation lexicon becomes smoother and the coverage is considerably improved. Such equivalence classes are constructed by omitting those items of information from morpho-syntactic analysis, which are not relevant for translation.

The lemma–tag representation of the corpus helps to identify — and access — the unimportant information. The definition of relevant and unimportant information, respectively, depends on many factors like the involved languages, the translation direction and the choice of the models. Linguistic knowledge can provide information about which characteristics of an input sentence are crucial to the translation task and which can be ignored, but it is desirable to automate this decision process. One could think of defining a likelihood criterion for this purpose. Another possibility is to assess the impact on the alignment quality after training, which can be evaluated automatically [Ahrenberg & Merkel$^+$ 00, Och & Ney 00]. After some first experiments, this approach was abandoned because the alignment quality on the Verbmobil data for example is very robust against manipulation of the training data (see Table 3.4). This behavior does not sufficiently reflect the effect of the training corpus on the translation accuracy.

The approach finally chosen is to detect candidates for equivalence classes of words from the probabilistic lexicon trained for translation from German to English. For this purpose, those inflected forms of the same base form are inspected, which result in the same translation. For each set of tags, the algorithm counts how often an additional tag can be replaced by a certain other tag without effect on the translation. Table 3.5 lists some of the most frequently identified candidates to be ignored while translating: The gender of nouns is irrelevant for their translation (which is straightforward, as the gender of a noun is unambiguous) and the cases nominative, dative, accusative. For the genitive forms, the translation in English differs. For verbs the candidates number and person were found: the translation of the first person singular form of a verb, for example, is often the same as the translation of the third person plural form. Ignoring (dropping) those tags most often identified as irrelevant for translation results in building equivalence classes of words. Doing so results in a smaller vocabulary about 65.5% the size of the vocabulary of the full lemma–tag representation and about 99% the vocabulary size of the German part of the original Verbmobil corpus, for example.

The information described in this chapter is used along two directions to improve the quality of statistical machine translation and to better exploit the available bilingual resources. The next two chapters are dedicated to the description of these two approaches:

- Implementation of transformations performed on the input sentence before the actual translation process and thereafter on the output translations.
- Construction of hierarchical lexicon models which combine information on different levels of abstraction from the fully inflected word form to the base form.

Table 3.4: Effect of the size of the corpus for training on the alignment quality. Task: Verbmobil.

# sentences	precision	recall
58k	90%	88%
5k	84%	84%
1k	81%	82%

Table 3.5: Candidates for equivalence classes.

part of speech	**candidates**
noun	gender: masculine, feminine, neuter and case: nominative, dative, accusative
verb	number: singular, plural and person: 1,2,3
adjective	gender, case and number
number	case

Chapter 4

Integration of Morpho-syntactic Information via Preprocessing and Postprocessing

> Research in machine translation has developed traditional patterns which will clearly have to be broken if any real progress is to be made.
>
> (Martin Kay, 1996)

Figure 1.1 already anticipated the fact that the source language input strings may be transformed in a certain manner before the actual translation process and that also the output translations may undergo a postprocessing step. Similarly, such transformations can be applied to the corpora used for training the model parameters. Transforming the training corpora implies restarting the training procedure, but the algorithms for training and decoding themselves remain unchanged. The transformations to both the source and target language suggested in this chapter are motivated by insights about the nature and sources of typical errors of machine translation systems and are only possible on the basis of information from morpho-syntactic analysis. Thus, they provide a way of incorporating knowledge about the languages under consideration without making it necessary to adapt existing training and decoding procedures to new types of models.

Figure 4.1 depicts the overall process of using transformations in training and test. The training corpus consists of aligned source language sentences and target language sentences which both undergo a sentence level word restructuring step before being used for training the parameters of the translation model and the language model. The restructuring transformations are also applied to the input sentence during test, and the inverse of the transformations applied to the target language part of the training corpus are performed on the output of the translation process. In detail, restructuring entails: treatment of question inversion, treatment of separated verb prefixes, splitting of compound words, and merging of multi-word phrases, described in Section 4.1. Section 4.2 explains how unknown word forms occurring in the input sentence, that is word forms not seen in the training corpus, can be mapped to more abstract, known word forms in order to enable

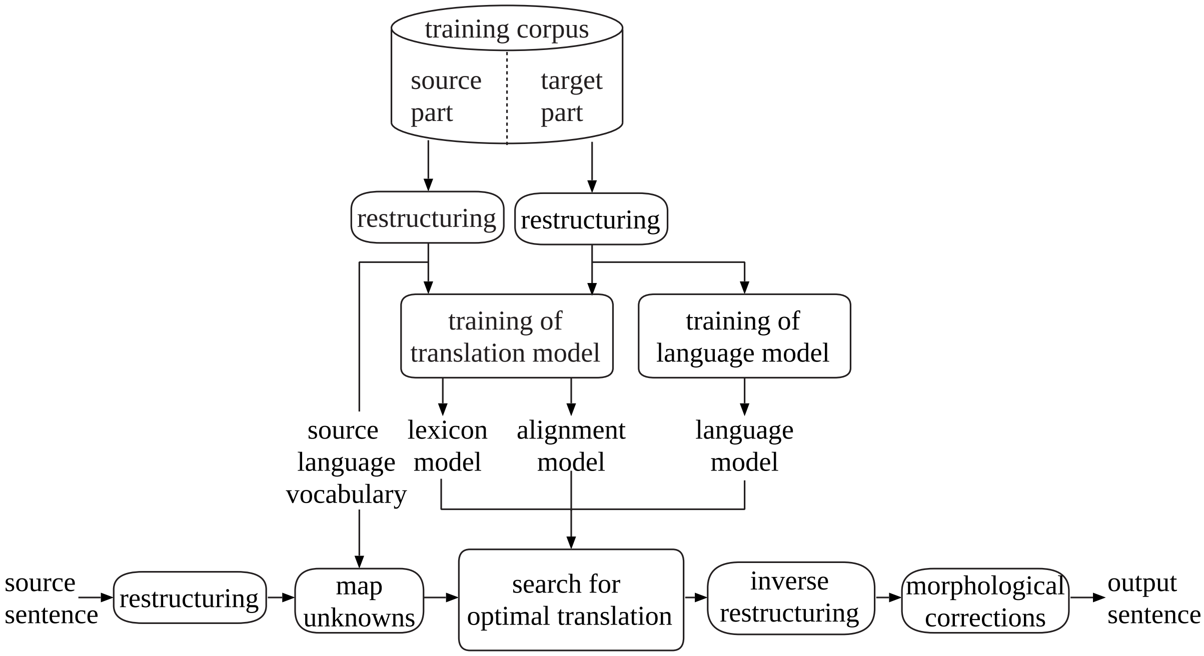


Figure 4.1: Training and test with transformations. "(inverse) restructuring", "map unknowns" and "morphological corrections" all require morpho-syntactic analysis of the transformed sentences.

at least an approximative translation. When translating from a less inflected language into a more inflected one, it can be beneficial to correct the translation output based on morpho-syntactic information. This will be the topic of Section 4.3.

4.1 Treatment of structural differences

As experiences with various translation tasks and language pairs show, difference in sentence structure is one of the main sources of errors in machine translation. This observation has been made previously, also for the language pair English and German, which has comparatively quite different word orders [Wang 98, Tillmann & Ney 00]. The reason is that in training the automatic alignment procedure tends to associate words in the source language sentence with words at *similar positions* in the target language sentence, while for language pairs like German and English the correct, 'ideal' alignment often contains long distance 'jumps'. Erroneous alignments on the bilingual training corpus result in noisy probabilistic lexica. It is thus promising to introduce transformations which aim at 'harmonizing' the word order in corresponding sentences.[1] The presentation in this chapter focuses on the following aspects:

[1] [Brown & Della Pietra$^+$ 92] formulate their view as to the validity of such operations as follows: "In some cases ... our simple rules will fail to apply where they should or will apply where they should not. While this is regrettable, we take a purely pragmatic attitude toward these errors: if the performance of the system improves when we use a transformation, then the transformation is good, otherwise it is bad."

Question Inversion: In many languages, the sentence structure of interrogative sentences differs from the structure in declarative sentences.

German verb prefixes: Some German verbs consist of a main part and a detachable prefix which can be shifted to the end of the clause.

Compound words: Compounds are words consisting of two or more relatively independent constituents. Typically they are translated into more than one word in English.

Multi-word phrases: Phrases are consecutive words repeatedly occurring as a fixed sequence in the corpus. Often they jointly fulfill a syntactic function in the sentence, and often they are translated into single words or identified phrases.

4.1.1 Question inversion

In German as well as in English and in many other languages, the sentence structure of interrogative sentences differs from the structure in declarative sentences in that the order of the subject and the corresponding finite verb is inverted. From the perspective of statistical translation, this behavior has some disadvantages: The algorithm for training the parameters of the target language model $Pr(e_1^I)$, which is typically a standard n-gram model, cannot deduce the probability of a word sequence in an interrogative sentence from the corresponding declarative form. For example, from the frequency of the sequence "you would have time" in the training corpus, the language model is not able to infer the probability of the sequence "would you have time". The same reasoning is valid for those statistical machine translation systems, which can learn the lexical translation probabilities of multi-word phrase pairs, like for instance the alignment template approach described in [Och & Tillmann$^+$ 99] and used as the translation system in the experiments: Without a special treatment of question inversion, such a system would not be able to learn the translation "ist es Dir recht" for "would you mind" from the bilingual sample "(you would mind"/"es ist Dir recht)".

The procedure for harmonizing the word order of questions with the word order in declarative sentences can best be understood by looking at the examples in Figure 4.2: The order of the subject (including the appendant articles, adjectives etc.) and the corresponding finite verb is inverted. In English questions supporting "do"s are removed. The application of the described preprocessing step on interrogative phrases in the bilingual training corpus implies the necessity of restoring the correct forms of the translations produced by the MT algorithm: In a postprocessing step the inverse restructuring operations are performed and in Yes/No-interrogatives the correct forms of the supporting "do" are inserted. This procedure was suggested by [Brown & Della Pietra$^+$ 92] for the language pair English and French, but they did not report on experimental results revealing the effect of the restructuring on the translation quality.

The reordering algorithm suggested here uses the information from syntactic analysis (see Tables 3.1 and 3.2), which helps to find the subject and the corresponding finite verb in an interrogative phrase. Because of the smaller variability regarding word order in English, this information is especially explicit and reliable for English. For the cases

may I take your order, sir?	→	I may take your order, sir?
do you know where the boutique is?	→	you know where the boutique is?
when would you have time for that?	→	when you would have time for that?
I mean, why should I reject it?	→	I mean, why I should reject it?
would you mind, if I come today?	→	you would mind, if I come today?
darf ich Ihre Bestellung aufnehmen?	→	ich darf Ihre Bestellung aufnehmen?
ich meine, warum sollte ich das tun?	→	ich meine, warum ich sollte das tun?
ist es OK, wenn ich heute komme?	→	es ist OK, wenn ich heute komme?
wie teuer ist ein Einzelzimmer?	→	wie teuer ein Einzelzimmer ist?

Figure 4.2: Examples for removing question inversion.

when subject and corresponding finite verb cannot unambiguously be identified from the analysis, some heuristics have been implemented which proved to be correct in most of the observed cases. Examples for the effect of question inversion treatment on the translation quality are given in Figure 4.3. In this figure as well as in the following ones, "⇓" denotes the translation process itself and "↓" and "⟶" indicate the effect of pre- or postprocessing.

4.1.2 Separated verb prefixes

Some verbs in German consist of a main part and a detachable prefix which can be shifted to the end of the clause, e.g. "losfahren" ("to leave") in the sentence "Ich fahre morgen los.". For the automatic alignment process it is often difficult to associate one English word with more than one word in the corresponding German sentence, namely the main

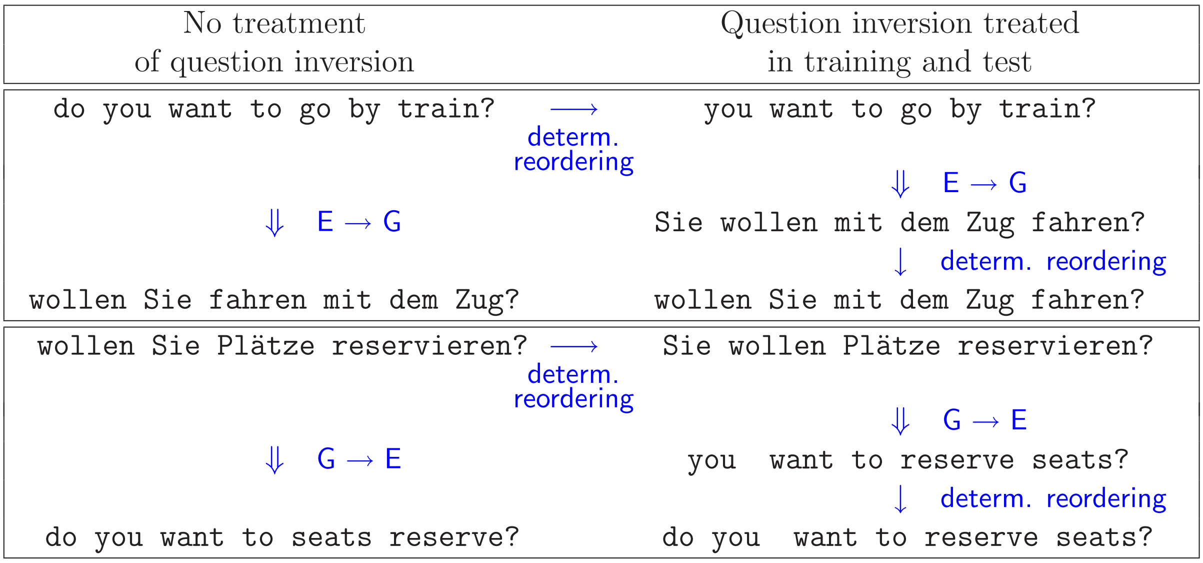

No treatment of question inversion
Question inversion treated in training and test
do you want to go by train?
⟶ determ. reordering
you want to go by train?
⇓ E → G
Sie wollen mit dem Zug fahren?
↓ determ. reordering
⇓ E → G
wollen Sie fahren mit dem Zug?
wollen Sie mit dem Zug fahren?
wollen Sie Plätze reservieren?
⟶ determ. reordering
Sie wollen Plätze reservieren?
⇓ G → E
you want to reserve seats?
↓ determ. reordering
⇓ G → E
do you want to seats reserve?
do you want to reserve seats?

Figure 4.3: Example for the effect of question inversion treatment on the translation quality.

part of the verb and the separated prefix. This difficulty is more serious in the (frequent) cases, where the distance between the positions of the main and the prefix part is large.

The procedure to solve the problem of separated prefixes is as follows: All separable word forms of verbs are extracted from the training corpus. The resulting list contains entries of the form `prefix|main`. For example, the entry "los|fahre" indicates that the prefix "los" can be detached from the word form "fahre". In all clauses containing a word matching a main part and a word matching the corresponding prefix part occurring at the end of the clause, the prefix is prepended to the beginning of the main part, as in "`Ich losfahre morgen`"[2]. This is carried out for the German part of the training corpus and, in the case that German is the source language, also for the input sentences in the testing phase. Examples for the effect on the translation quality are given in Figure 4.4.

The translation direction from English to German is more complicated than vice versa, because additional postprocessing of the German output sentence is needed in order to reconstruct the correct forms of the separable verbs. A language model rescoring approach is used to choose between different positions of the verb prefix, e.g. between "`Ich losfahre morgen`", "`Ich fahre los morgen`" and "`Ich fahre morgen los`". For this purpose, the trigram language model scores of the original sentence and the variants with moved prefixes are computed and the best scoring translation is chosen. Naturally, this language model is trained on the original, not transformed German part of the training corpus.

Figure 4.5 illustrates the results achieved with treating separated verb prefixes for the translation direction English to German. Note that there is an analogy between English and German in that for some separable German verbs there are correspondences in English which themselves consist of a main part and a particle, e.g. "pick up" for "abholen". Unlike the German separable verbs, these English phrases stay relatively close together, that is they are not separated by many words. As a consequence they can still be captured as a whole by a system like the alignment templates and thus it is not so important to treat them explicitly.

Treatment of infinitive markers inside German verbs

For German separable verbs, the infinitive marker "zu" does not precede the verb in the infinitive form, as it does normally in German and as does the English infinitive marker "to", like in "to be or not to be". Instead, the infinitive marker is inserted into the infinitive verb form between the main part and the prefix part, like in "los*zu*fahren". Consequently, the infinitive verb forms with and without corresponding infinitive marker form separate lexicon entries and thus aggravate the data sparseness problem. A straightforward complement of prepending detached prefixes to their corresponding main parts is to strip off the infinitive markers and move them to their usual position directly preceding the infinitive verb form. For example, "loszufahren" is replaced by "`zu losfahren`" in training and when translating from German into English. When testing is carried out for

[2] The result from restructuring is often not correct according to the grammatical constraints in the corresponding language. The sentences resulting from the restructuring operations are distinguished from the original sentences by using a different font.

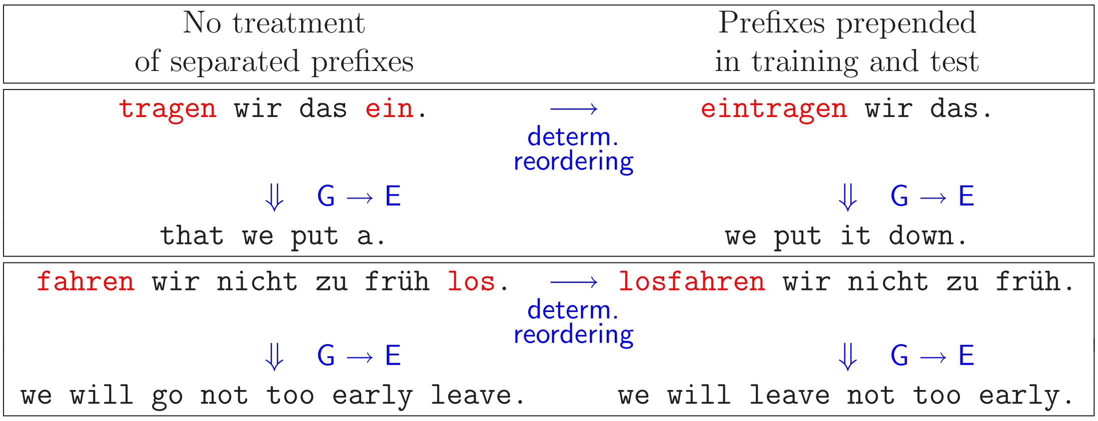

No treatment of separated prefixes
Prefixes prepended in training and test
tragen wir das ein.
determ. reordering
eintragen wir das.
G → E
G → E
that we put a.
we put it down.
fahren wir nicht zu früh los.
determ. reordering
losfahren wir nicht zu früh.
G → E
G → E
we will go not too early leave.
we will leave not too early.

Figure 4.4: Examples for the effect of prepending separated prefixes on the translation quality for the translation direction German to English.

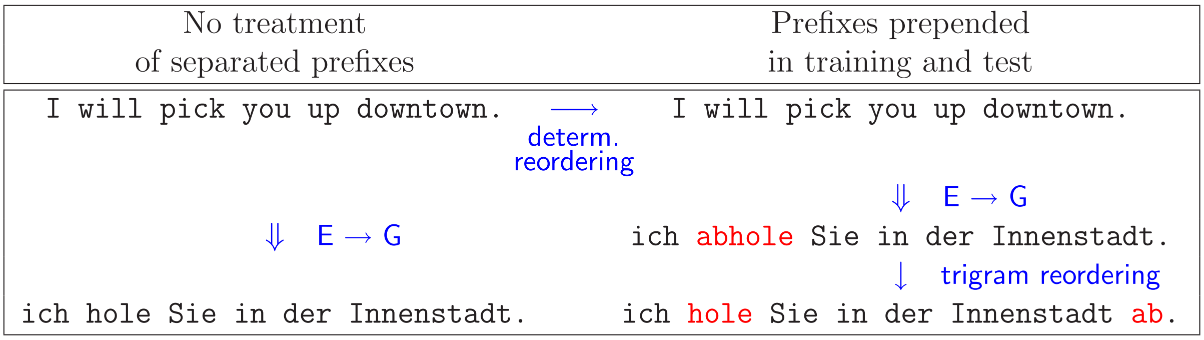

No treatment of separated prefixes
Prefixes prepended in training and test
I will pick you up downtown.
determ. reordering
I will pick you up downtown.
E → G
E → G
ich abhole Sie in der Innenstadt.
trigram reordering
ich hole Sie in der Innenstadt.
ich hole Sie in der Innenstadt ab.

Figure 4.5: Example for the effect of prepending separated prefixes on the translation quality for the translation direction English to German.

English to German translation, the inverse transformation is performed on the translation produced by the system trained on the transformed training corpus.

4.1.3 Compound splitting

Compound words pose special problems to the robustness of translation models because the word itself must be represented in the training data: the occurrence of each of the components is not enough. The word “Früchtetee” for example cannot be translated although its components “Früchte” and “Tee” appear in the training set of the Zeres task. Besides, even if the compound occurs in training, the training algorithm may not be capable of translating it properly as *two* words (in the mentioned case the words “fruit” and “tea”). Therefore the compound words are split into their components. As always, this transformation is performed before training as well as in test.

Figure 4.6 gives examples for the effect of splitting up compounds in the input language sentence. The compound “Tennisplätze” was translated correctly into “tennis courts” by the alignment templates system, which is in principle capable of translating phrases.

input	50 m entfernt befinden sich Tennisplätze und Segelschule.
baseline	50 m away to find tennis courts and Segelschule.
preprocessing	50 m away there are tennis courts and sailing school.

input	können Sie den Straßennamen bitte buchstabieren?
baseline	can you please spell the spell?
preprocessing	can you please spell the streets name?

Figure 4.6: Examples for the effect of compound splitting.

It is interesting to note that splitting up the word into its components does not harm the translation quality. The word “Segelschule” did not occur in training, unlike its components which after decomposition can correctly be translated into “sailing” and “school”. The word “Straßennamen” is an example of a word which cannot be translated correctly although it was seen in training. The reason is that the automatic alignment process had associated it with the word “spell” in training — obviously asking for the spelling of a street name is a frequent situation in the underlying domain. The components “Straßen” and “Namen” themselves are very frequent in the training corpus and can thus be translated by “streets” and “name”. Note that the splitting algorithm by default preserves the capitalization of the original word for all components. The components of a noun for instance adopt the upper case writing. This behavior of `process_cg` can be configured.

For some types of words, compound splitting is problematic. The reason is clear for compounds that represent proper names, like for instance the city name “Wuppertal”, which should not be translated as “Wupper valley”. Other difficulties arise with numbers and time expressions as they are very frequent in the Verbmobil task: Experiences show that these phrases can most robustly be translated with simple finite state transducers. Alternatively, [Wang 98] uses a special preprocessing step for German numbers, which transforms for instance the split version “`sechs` (six) `und` (and) `zwanzigsten` (twentieth)” of “sechsundzwanzigsten (twenty-sixth)” into “`sechste` (sixth) `und` (and) `zwanzig` (twenty)”. In order to make compound splitting (as well as the other operations on words) dependent on the type of the words, `process_cg` inspects the morpho-syntactic tags and can thus decide not to split numbers and identified proper names.

4.1.4 Multi-word phrases

Some recent publications deal with the automatic detection of multi-word phrases [Och & Weber 98, Tillmann & Ney 00]. These methods are very useful, but they have one drawback: they rely on sufficiently large training corpora, because they detect the phrases from automatically learned word alignments. In this section methods for detecting multi-word phrases are suggested, which merely require monolingual syntactic analyzers and a conventional electronic dictionary.

- Some multi-word phrases which jointly fulfill a syntactic function are provided by the analyzers. The phrase “irgend etwas” (“anything”) for example may form either an indefinite determiner or an indefinite pronoun. “`irgend=etwas`” is merged in order to form one single vocabulary entry. In the German part of the Verbmobil training corpus 26 different, non-idiomatic multi-word phrases are merged, while there are 318 phrases suggested for the English part. Some examples are listed in Table 4.1.

- In addition, syntactic information like the identification of infinitive markers, determiners, modifying adjectives (example: “*single* room”), pre-modifying adverbials (“*more* comfortable”), and pre-modifying nouns (“*account* number”) are used for detecting multi-word phrases. When applied to the English part of the Verbmobil training corpus these hints suggest 7 225 different phrases.

Altogether, 26 phrases for German and about 7 500 phrases for English are detected in this way. It is quite natural that there are more multi-word phrases found for English, as German unlike English uses compounding. But the experiments show that it is not advantageous to use all these phrases for English. Electronic dictionaries can be useful for detecting those phrases, which are important in a statistical machine translation context: A multi-word phrase is considered useful if it is translated into a single word or a distinct multi-word phrase (suggested in a similar way by syntactic analysis) in another language. For English, 290 phrases are chosen in this way. Section 5.3 shows how multi-word phrases help learning the disambiguation between different readings within dictionaries.

A few examples for the effect of merging multi-word phrases are listed in Figure 4.7. The phrases “flight schedule” and “hotel rooms” result from the second method (“flight” and “hotel” are pre-modifying nouns) and “alles klar” and “all right” are detected by the analyzers themselves. The baseline system translates “schedule” in the first example independently as a verb into “einplanen”. The second and third example show errors resulting from misalignments: Without merging phrases, the word pair “rooms” and “Einzelzimmer”(English: “single room(s)”) is often aligned in the training corpus, and both “alles” and “klar” are aligned to “right”.

Table 4.1: Examples of multi-word phrases provided by the analyzers.

Phrase	**Function in sentence**
‘als=ob’	subordinating conjunction
‘ein=bißchen’	indefinite pronoun singular
‘irgend=etwas’	indefinite determiner singular
‘vor=allem’	adverb
‘was=für’	interrogative determiner
‘a=little’	absolute adjective
‘as=soon=as’	preposition
‘in=order=to’	infinitive marker
‘lots=of’	determiner singular/plural
‘no=one’	pronoun nominative singular

input	I have a flight schedule.
baseline	ich habe einen Flug einplanen.
preprocessing	ich habe einen Flugplan.
input	I will reserve the hotel rooms.
baseline	ich buche die Einzelzimmer.
preprocessing	ich buche die Hotelzimmer.
input	alles klar?
baseline	right?
preprocessing	all right?

Figure 4.7: Examples for the effect of merging phrases.

4.2 Treatment of unseen word forms

For statistical machine translation systems it is difficult to handle words not seen in training. For unknown proper names it is normally correct to place the word unchanged into the translation. This section reports on work about the treatment of unknown words of other types. As already mentioned in Section 4.1.3 the splitting of compound words can reduce the number of unknown German words. In addition methods of replacing a fully inflected word form by a more abstract word form known from training have been examined. The translation of the simplified word form is generally not the precise translation of the original one, but sometimes the intended semantics is conveyed.

The mapping operations that can be applied to a word form not found in the vocabulary extracted from the training corpus are categorized as follows:

- Change lower case into upper case or vice versa, e.g.

“Vorweihnachtsfeiern”	→	“vorweihnachtsfeiern”
“Beschädigen”	→	“beschädigen”
“reden”	→	“Reden”

- Use the base form, e.g.

“Vorweihnachtsfeiern”	→	“Vorweihnachtsfeier”
“unternehmensübergreifende”	→	“unternehmensübergreifend”
“Reden”	→	“Rede”

- Split at “strong” split points, that is at the position between relatively independent, equally important components e.g.

“Vorweihnachtsfeier”	→	“Vorweihnachts Feier”
“unternehmensübergreifend”	→	“unternehmens übergreifend”

- Delete linking letters (“Fugenzeichen”), e.g.

"Vorweihnachts Feier"	→	"Vorweihnacht Feier"
"unternehmens übergreifend"	→	"unternehmen übergreifend"

- Split at "weak" split points, that is between the main part(s) and a particle, e.g.

"Vorweihnacht Feier"	→	"Vor Weihnacht Feier"
"unternehmen übergreifend"	→	"unternehmen über greifend"

- Split up phrases merged together in a preprocessing step (see Section 4.1.4).
- Split up phrases linked together with a dash, like "double-check" or "day-long".

If no transformation to the original word form results in a known word form, it is left unchanged. A sequence of these mapping operations can result in a sequence of word forms which are partly known from training. In general there are various ways of replacing a certain word form by more abstract word forms. The criteria for choosing among these possibilities are the following: Firstly, the resulting abstract word form and the original word form should have the same or at least a similar translation. To guarantee this, they should be as closely related as possible. Secondly, the resulting character sequence must be known from the training data. To ensure the first criterion, each mapping operation is associated with a penalty. These penalties have been empirically set as listed in Table 4.2. They can easily be changed by setting the values of the corresponding variables in `process_cg`. If the unknown word form is a compound word consisting of more than two components, each decomposition between either pair of components is counted as individual splitting operation.

When the original word form is not found in the vocabulary, the following algorithm finds an abstract form: For every possible combination of decompositions of the stem at the split points '`#`', '`=`', '`-`' and '`|`' and for every possible case-combination, using the original and the base form suffix (or word), and with and without linking letters: check, whether each part of this combination is exactly contained in the vocabulary. All admissible combinations are sorted according to their penalty and the best scoring one is passed as output — if there is more than one best scoring result, the choice is random.

Table 4.2: Empirically set penalties for mapping operations.

Mapping operation	**Penalty**	**Corresponding variable**
begin word	0	INITCOSTS
use different case	1	CASEMISMATCHCOSTS
use base form	2	STEMCOSTS
remove linking letter	1	SLASHCOSTS
split at '`#`' (strong split point)	3	HASHCOSTS
split at '`|`' (weak split point)	6	BARCOSTS
split at '`=`' (merged phrases)	3	EQCOSTS
split at '`-`'	3	DASHCOSTS

Given the penalties as in Table 4.2, Figure 4.8 illustrates the process of mapping German words not contained in the Zeres training corpus to more abstract word forms that occur in the corpus. “`Abhanden Kommen`”, “`Investitions Güter`” and “`fresken geschmückt`” are only intermediate steps — they cannot be used as final result, because the components “`Abhanden`”, “`Investitions`” and “`fresken`” are no meaningful German words. On the contrary, the intermediate steps “`Strudel Backen`” and “`Eis Sportverein`” might be the end results, if the respective components are found in the training corpus vocabulary.

4.3 Morphological corrections

Some translation errors which are typical of translating from a less inflected language like English into a more inflected one like German can be corrected after the translation process on the basis of morpho-syntactic information. The translation result is analyzed and represented as combination of base forms and morpho-syntactic tags as suggested in Section 3.4. Using this lemma–tag representation it is possible to identify groups of words belonging to the same concept, for instance noun phrases, and to detect mismatches within these groups in terms of case, number or gender. The values for these three features are extracted from the noun reading and taken over by the corresponding determiners and adjectives. From the corrected lemma–tag representation the presumably correct full word forms are generated. Examples for the effect of these corrections are given in Figure 4.9.

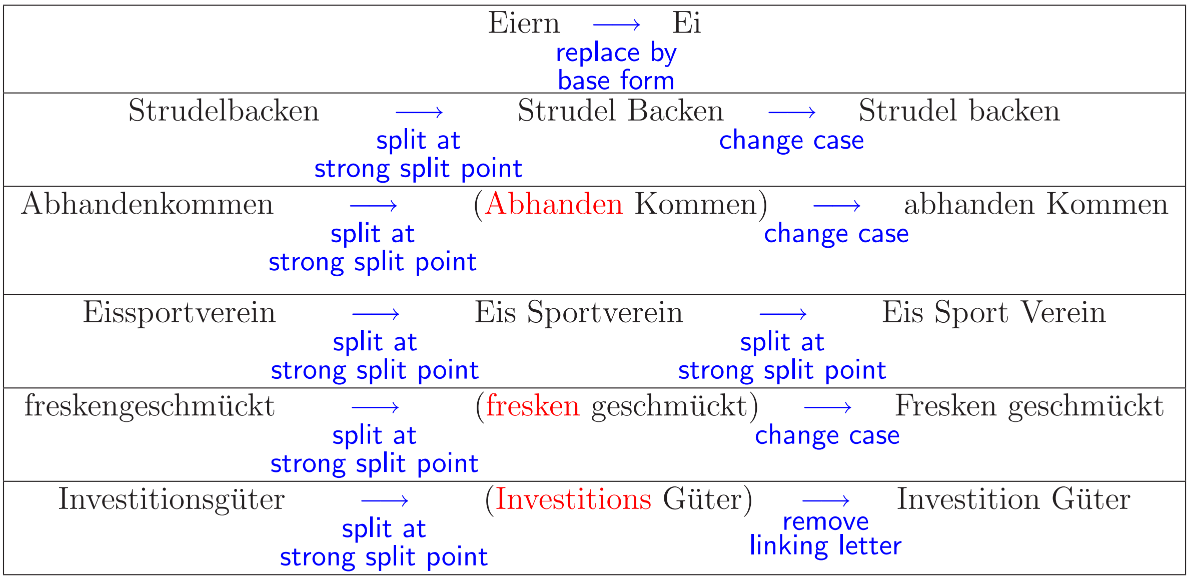

Eiern ⟶ Ei
replace by base form
Strudelbacken ⟶ Strudel Backen ⟶ Strudel backen
split at strong split point
change case
Abhandenkommen ⟶ (Abhanden Kommen) ⟶ abhanden Kommen
split at strong split point
change case
Eissportverein ⟶ Eis Sportverein ⟶ Eis Sport Verein
split at strong split point
split at strong split point
freskengeschmückt ⟶ (fresken geschmückt) ⟶ Fresken geschmückt
split at strong split point
change case
Investitionsgüter ⟶ (Investitions Güter) ⟶ Investition Güter
split at strong split point
remove linking letter

Figure 4.8: Examples for treatment of unseen word forms. These examples were taken from the Zeres task.

die Hotel liegt zentral.

↓ analysis

die-def.-art.-sg.-nom.-fem. Hotel-noun-neut.-sg.-nom.

↓ correction

die-def.-art.-sg.-nom.-neut. Hotel-noun-neut.-sg.-nom.

↓ generate words

das Hotel liegt zentral.

wir treffen uns *am Zugnummer* neunzehn.

↓ analysis

an-dem-prep.-def.-art.-sg.-dat.-masc. Zugnummer-noun-fem.-sg.-dat.

↓ correction

an-dem-prep.-def.-art.-sg.-dat.-fem. Zugnummer-noun-fem.-sg.-dat.

↓ generate words

wir treffen uns *an der Zugnummer* neunzehn.

Figure 4.9: Examples for morphological corrections. These examples were taken from the Verbmobil English to German test corpus.

Chapter 5

Hierarchical Lexicon Models for the Translation of Inflected Languages and Translation with Scarce Resources

Words, from the earliest times of which we have historical records, have been objects of superstitious awe.

(Bertrand Russel, 1940)

The parameters of the translation model are trained on a bilingual corpus. In general, the resulting probabilistic lexicon contains all word forms occurring in this training corpus as separate entries, not taking into account whether or not they are inflected forms of the same lemma. Bearing in mind that typically more than 40% of the word forms are only seen once in training (see for example Table A.1 and Table A.7), it is obvious that learning the correct translations is difficult for many words. Besides, new input sentences are expected to contain unknown word forms, for which no translation can be retrieved from the lexicon. This problem is especially relevant for highly inflected languages like German: Texts in German contain many more distinct word forms than their English translations. The tables in Appendix A also reveal that these words are often generated via inflection from a smaller set of base forms.

Another aspect is the fact that conventional dictionaries are often available in an electronic form for the considered language pair. Their usability for statistical machine translation is restricted because they are substantially different from full parallel corpora: The entries are often pairs of *base forms* that are translations of each other, whereas the corpora contain full sentences with inflected forms. To make the information taken from external dictionaries more useful for the translation of inflected languages is a relevant objective.

On the basis of these considerations it is straightforward to aim at taking into account the interdependencies between the inflected forms of the same base form. A first step toward this goal is the introduction of equivalence classes at various levels of abstraction starting with the inflected form and ending with the base form. As already explained in Section 3.4 the lemma–tag representation of the information from morpho-syntactic analysis makes it possible to gradually access information with different grades of abstraction. Consider, for example, the German verb form `"ankomme"`, which is the indicative present first person singular form of the lemma `"ankommen"` and which can be translated into English by `"arrive"`. The lemma–tag representation provides an 'observation tuple' consisting of

- the original full word form, e.g. "`ankomme`",
- morphological and syntactic tags (POS, tense, person, ..., case, ...)
  e.g. "`verb, indicative, present tense, 1st person singular`" and
- the base form, e.g. "`ankommen`".

In the following, $t_0^i = t_0, \ldots, t_i$ denotes the representation of a word where the base form t_0 and i additional tags are taken into account. For the example above, $t_0 =$ `"ankommen"`, $t_1 =$ `"verb"`, and so on. The hierarchy of equivalence classes $\mathcal{F}_0, \ldots, \mathcal{F}_n$ is as follows:

$$\begin{aligned}
\mathcal{F}_n = \mathcal{F}(t_0^n) &= \texttt{"ankommen verb indicative present singular 1"} \\
\mathcal{F}_{n-1} = \mathcal{F}(t_0^{n-1}) &= \texttt{"ankommen verb indicative present singular"} \\
\mathcal{F}_{n-2} = \mathcal{F}(t_0^{n-2}) &= \texttt{"ankommen verb indicative present"} \\
&\vdots \\
\mathcal{F}_0 = \mathcal{F}(t_0) &= \texttt{"ankommen"}\ .
\end{aligned}$$

n is the maximal number of morpho-syntactic tags. The mapping from the full lemma–tag representation back to inflected word forms is generally unambiguous, thus $\mathcal{F}_n$ contains only one element, namely `"ankomme"`. $\mathcal{F}_{n-1}$ contains the forms `"ankomme"`, `"ankommst"` and `"ankommt"`; in $\mathcal{F}_{n-2}$ the number (`singular` or `plural`) is ignored and so on. The largest equivalence class contains all inflected forms of the base form `"ankommen"`. The order of omitting tags can be defined in a natural way depending on the part of speech. In principle this decision can also be left to the ME training, when features for all possible sets of tags are defined, but this would cause the number of parameters to explode. As the experiments in this work have only been carried out with up to three levels of abstraction as defined in Section 5.2.1, the set of tags of the intermediate level is fixed and thus the priority of the tags needs not to be specified. The relation between this equivalence class hierarchy and the suggestions in Section 3.5 is clear: Choosing candidates for morpho-syntactic tags not relevant for translation amounts to fixing a level in the hierarchy. This is exactly what has been done to define the intermediate level in Section 5.2.1.

The methods described in the following sections make use of the aforementioned hierarchy definition to increasing extent, ranging from the translation of base forms in the first stage of a two-stage translation scheme as mentioned in Section 5.1.1, via using POS information for disambiguating a small set of frequent words (see Section 5.1.2) to the translation of equivalence classes instead of fully inflected word forms as in Section 5.1.3. Finally, the concept of *combining* information at different levels of abstraction is introduced, namely by means of linear interpolation in Section 5.1.4 and using log-linear models in Section 5.2.

5.1 Pre-studies

The methods described in this section are regarded as preparation for the log-linear models presented in Section 5.2, as the insights from these pre-studies guided the design of these models. Except for the investigation on two-stage translation introduced in Section 5.1.1 all of the approaches concerning inflectional morphology presented in this chapter apply to the *source language.*

5.1.1 Two-stage translation

Translating from a language with hardly any inflectional morphology like English into an inflected language like German raises additional difficulties, as the translation process has to add information to the word forms in the target language which cannot be directly inferred from the words in the input. The underlying idea of the approach proposed in this section and depicted in Figure 5.1 is to separate the choice of the word order from the choice of the correct inflected word form into two subsequent stages:

1. In a first stage, a translation system with good word reordering capability, e.g. the alignment templates system, is used to translate from normal English into pseudo-German where the sentences consist of sequences of *concepts* more abstract than

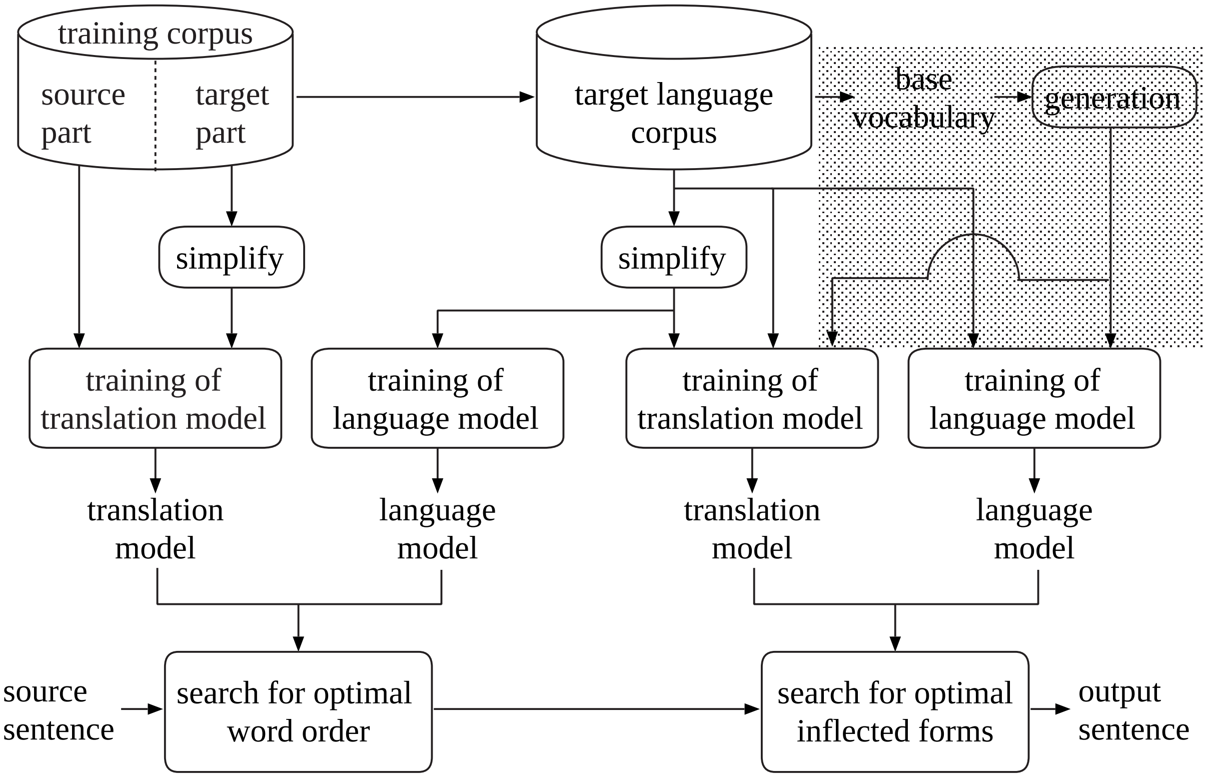


Figure 5.1: Training and test with two stages. "simplify" requires morpho-syntactic analysis of the transformed sentences and "generation" entails generating additional inflected forms of the base forms contained in the corpora.

fully inflected word forms. For this purpose the words in the German part of the training corpus are replaced by a simplified lemma–tag representation, for example base forms, and the translation model and the language model are trained using this transformed corpus. Alternatively a larger monolingual German corpus could be used for training the language model.

2. For the second stage the transformed German corpus (either the target language part of the bilingual corpus or else the larger monolingual corpus) serves as source language part of a bilingual corpus used for training a translation model, whereas the original German corpus is the target language part. Using the translation and language model probabilities learned in this way, the concepts produced in the first stage are replaced by fully inflected word forms using for example the single word system in a variant with strict monotonicity constraint.

The first stage fixes the *position* and the *lemma* of the words in the German translation of an English sentence, while the second stage generates the inflected word form. In first informal experiments for this pre-study at the beginning of the work for this thesis, 5 000 sentences of the Verbmobil corpus were used for training for the first stage, where the intermediate pseudo-German consisted of base forms. The full monolingual German corpus part was used for training for the second stage. The results on the Verbmobil Eval-2000 English to German task were disappointing: the translation quality was slightly worse than directly translating from English into fully inflected German using the alignment templates system. A control experiment, where the German reference translations were transformed into sequences of base forms and translated back into inflected word forms, proves that base forms do not contain enough information to properly reconstruct the correct forms: the lower bound for the translation errors committed in the second stage is 11% translation word error rate. In this thesis this approach has not been elaborated further, but there is a potential for improving the procedure:

- More information than pure base forms should be contained in the intermediate representation.
- The language model for the second stage should be more syntax-oriented than standard n-gram models.
- The shaded part in Figure 5.1 sketches the possibility of using a generation component. This module would be able to provide inflected forms of the base vocabulary, which are not contained in the training corpus.

5.1.2 Low-level disambiguation

Apart from the possibility of gradual information access, the second major advantage of the lemma–tag representation introduced in Section 3.4 is that it makes context information locally available: The analyzers use context to choose between ambiguous readings of words. This feature can be used in statistical machine translation to assist lexical choice in the target language, namely when the correct translation depends on the reading. In order to examine the potentials of disambiguating with the help of part of speech

(POS) information, a few frequent short words were examined that often cause errors in translation for the Verbmobil task. These words were annotated with their POS:

“aber” can be an adverb as in “Das ist aber nett.” or a conjunction as in “. . . , aber das ist nett.”.

“zu” can be an adverb as in “zu lang”, a preposition as in “zum (= zu dem) Bahnhof”, a separated verb prefix as in “Das trifft zu.”, or an infinitive marker as in “leicht zu lesen”.

“der”, “die” and “das” can be definite articles as in “das Hotel”, or pronouns as in “Das geht.”.

The difficulties due to these ambiguities are illustrated by the following examples: The sentence “Das würde mir gut passen.” is often translated by “*The* would suit me well.” instead of “*That* would suit me well.” and “Das war zu schnell.” is translated by “That was *to* fast.” instead of “That was *too* fast.”. Annotating these few words with their POS in training and in test yielded a relative improvement of the overall translation quality of almost 3% in terms of SSER from 20.3% to 19.7% in preliminary experiments carried out on the Verbmobil task with the alignment templates system. These results encourage the next logical step along this line, namely translating equivalence classes. For a definition of the SSER and the other measures used in the Tables 5.1 and 5.2 see Appendix B.

5.1.3 Translation of equivalence classes instead of fully inflected word forms

In this experiment, *all* words in the German source language part of the training corpus were replaced by their lemma–tag representation and the morpho-syntactic tags not considered relevant for the translation task (see Section 3.5) were dropped. In other words, the inflected word forms were represented by the “identifier” of an intermediate equivalence class $\mathcal{F}_i$ in the hierarchy exemplified on page 30. The same transformations were performed on the input sentences in the test set. This approach combines two aspects:

Specification via annotation with morpho-syntactic information, which makes context information locally accessible.

Abstraction by ignoring unimportant information, which results in a better coverage of the lexicon.

Table 5.1 shows the effect of introducing equivalence classes. The information from the morpho-syntactic analyzer is reduced by dropping unimportant information. All error measures could be decreased in comparison to using the original corpus with inflected word forms. For each word one single reading was chosen by applying some heuristics (see Section 3.3). For the normal training corpora, unlike conventional dictionaries, this is not critical because they contain predominantly full sentences which provide enough context for an efficient disambiguation. Meanwhile, methods for disambiguating conventional

Table 5.1: Effect of the introduction of equivalence classes. The original inflected word forms were used for the baseline experiment. Task: Verbmobil. Testing on 251 sentences ("`Test`"). System: Single word system.

	m-WER [%]	SSER [%]	ISER [%]
Inflected words	38.2	38.0	28.0
Equivalence classes	36.9	36.5	25.4

dictionaries have been implemented — they are documented in Section 5.3 of this thesis — but these experiments for equivalence classes were carried out for an earlier publication [Nießen & Ney 01b] using only bilingual corpora for estimating the model parameters.

The first example in Figure 5.2 demonstrates the effect of the disambiguating analyzer which identifies "Hotelzimmer" as singular on the basis of the context (the word itself can represent the plural form as well), and "das" as article in contrast to a pronoun. The second example shows the advantage of grouping words into equivalence classes: The training data does not contain the word "billigeres", but when generalizing over the gender and case information, a correct translation can be produced.

5.1.4 Linear combination

The previous sections more or less explicitly suggested to select certain levels in the hierarchy introduced on page 30 in order to replace the original inflected word forms. This section describes an approach to *combine* information at different levels of the hierarchy by means of linear interpolation.

Let $p(f|t_0^i, e)$ be the lexicon probability of a source language word f for a given partial reading t_0^i of f and a target language word e. Under the assumption that this probability does not depend on e this can be rewritten as $p(f|t_0^i, e) = p(f|t_0^i)$. For the sake of readability, the probability functions $p(f|t_0^i)$ are defined to yield zero for impossible interpretations t_0^i, that is when $f \notin \mathcal{F}(t_0^i)$. The inflected form is always assumed to be non-ambiguously derivable from the full lemma–tag representation t_0^n, that is $p(f|t_0^n) = 1$. In other words, the inflected form can non-ambiguously be derived from the full lemma–tag representation. $p(t_0^i|e)$ is the probability of the translation for e to belong to the equivalence class $\mathcal{F}_i$. The lexicon probability of a word f to be translated by e with respect to the level i can be defined by summing up over all possible readings of f:

$$p_i(f|e) = \sum_{t_0^i} p(f|t_0^i) \cdot p(t_0^i|e) \ . \tag{5.1}$$

ich reserviere das Hotelzimmer
⇓ G → E
I will reserve that hotel rooms

ich reserviere das Hotelzimmer
↓ specification
ich-pers.-pron.-sg.-1st-*nom.* reservieren-verb-ind.-pres.-sg.-*1st*
das-def.-art.-sg.-*acc.*-*neut.* Hotelzimmer-noun-*neut.*-sg.-*acc.*
↓ abstraction
ich-pers.-pron.-sg.-1st reservieren-verb-ind.-pres.-sg.
das-def.-art.-sg. Hotelzimmer-noun-sg.
⇓ G → E
I will reserve the hotel room

gibt es nichts billigeres?
⇓ G → E
there is do not UNKNOWN-billigeres?

gibt es nichts billigeres?
↓ specification
geben-verb-ind.-pres.-sg.-*3rd* es-pers.-pron.-sg.-3rd-*nom.*-*neut.*
nichts-indef.-det.-neg.-sg.-*acc.* billig-adj.-comp.-*sg.*-*nom./acc.*-*neut.*?
↓ abstraction
geben-verb-ind.-pres.-sg. es-pers.-pron.-sg.-3rd
nichts-indef.-det.-neg.-sg. billig-adj.-comp.?
⇓ G → E
there is nothing cheaper?

Figure 5.2: Examples for the effect of equivalence classes resulting from dropping morpho-syntactic tags not relevant for translation. First the translation using the original representation, then the new representation, its reduced form and the resulting translation.

The p_i can easily be combined by means of linear interpolation:

$$\begin{aligned} p(f|e) &= \lambda_0 p_0(f|e) + \ldots + \lambda_n p_n(f|e) \\ &= \sum_i \lambda_i \sum_{t_0^i} p(f|t_0^i) \cdot p(t_0^i|e) \\ &= \sum_{t_0^n} \sum_i \lambda_i \cdot p(f|t_0^i) \cdot p(t_0^i|e) \ , \end{aligned} \tag{5.2}$$

where the $n+1$ weight parameters λ_i fulfill the constraint $\sum_i \lambda_i = 1$. Results for the linear combination have been published in [Nießen & Ney 01b] for the case of only two levels, i.e. n in Equation (5.2) was set to 1. Thus, there was only one free parameter λ which was set to 0.5. $p(f|t_0)$ was modeled as a uniform distribution $\frac{1}{|\mathcal{F}(t_0)|}$ over all inflected forms with the base form t_0 occurring in the training data plus the base form itself, in case it is not contained. The process of lemmatization is unique in the majority of cases, and as a consequence, the sum in Equation (5.1) is not needed for a two-level lexicon combination of full word forms and base forms. Equation (5.2) then amounts to the following simpler definition of the lexicon probability:

$$p(f|e) \quad = \quad (1-\lambda) \cdot \frac{p(t_0|e)}{|\mathcal{F}(t_0)|} + \lambda \cdot p(t_0^n|e) \ . \tag{5.3}$$

The alignment on the training corpus was optimized using the original inflected word forms. From this alignment, the co-occurrence frequencies of aligned source and target “words” at corresponding positions were extracted to yield the lexicon probabilities in Equation (5.3). As the results summarized in Table 5.2 show, the combined lexicon outperforms the conventional one-level lexicon. Not surprisingly, the quality gain achieved by smoothing the lexicon is larger if the training procedure can take advantage of an additional conventional dictionary to learn translation pairs, because these dictionaries typically only contain base forms of words, whereas translations of fully inflected forms are needed in the test situation.

5.2 Log-linear combination

As there is a large overlap between the modeled events in the combined probabilistic models, it can be assumed that log-linear combination is better suited than linear interpolation. This is the topic of this section.

In modeling for statistical machine translation, a hidden variable a_1^J, denoting the hidden alignment between the words in source and target language, is usually introduced into the string translation probability:

$$Pr(f_1^J|e_1^I) \quad = \quad \sum_{a_1^J} Pr(f_1^J, a_1^J|e_1^I) = \sum_{a_1^J} Pr(a_1^J|e_1^I) \cdot Pr(f_1^J|a_1^J, e_1^I) \ . \tag{5.4}$$

In the following, $T_j = (t_0^n)_j$ denotes the lemma–tag representation of the jth word in the input sentence. The sequence T_1^J stands for the sequence of readings for the word sequence f_1^J, and can be introduced as a new hidden variable:

$$Pr(f_1^J|a_1^J, e_1^I) \quad = \quad \sum_{T_1^J} Pr(f_1^J, T_1^J|a_1^J, e_1^I) \ , \tag{5.5}$$

Table 5.2: Effect of two-level lexicon combination. As baseline serves the conventional one-level full form lexicon. Task: Verbmobil. Testing on 251 sentences (“Test”). System: Single word system.

	ext. dictionary	m-WER [%]	SSER [%]	ISER [%]
baseline	no	38.2	38.0	28.0
combined	no	38.3	37.6	27.2
baseline	yes	37.8	37.0	26.4
combined	yes	37.5	35.5	24.9

which can be decomposed into

$$Pr(f_1^J|a_1^J, e_1^I) \;=\; \sum_{T_1^J} \prod_{j=1}^{J} Pr(f_j, T_j | f_1^{j-1}, T_1^{j-1}, a_1^J, e_1^I) \; . \tag{5.6}$$

Furthermore, let $\mathcal{T}(f_j)$ denote the set of interpretations which are regarded valid readings for f_j by the morpho-syntactic analyzers on the basis of the whole sentence context f_1^J. Then we make the assumption that the probability functions defined above yield zero for all other readings, that is when $T_j \notin \mathcal{T}(f_j)$. Making the usual independence assumption, which states that the probability of the translation of words only depends on the identity of the words associated to each other by the word alignment, we get:

$$Pr(f_1^J|a_1^J, e_1^I) \;=\; \sum_{\substack{T_1^J \\ T_j \in \mathcal{T}(f_j)}} \prod_{j=1}^{J} p(f_j, T_j | e_{a_j}) \; . \tag{5.7}$$

As has been argued in Section 3.3, the number of readings $|\mathcal{T}(f_j)|$ per word form can be reduced to 1 for the tasks for which experimental results are reported here.

The elements in Equation (5.7) are the joint probabilities $p(f, T|e)$ of f and the readings T of f given the target language word e. The maximum entropy principle recommends to choose for p the distribution which preserves as much uncertainty as possible in terms of maximizing the entropy

$$H(p) = -\sum_x p(x) \log p(x) \; ,$$

while requiring p to satisfy constraints, which represent facts known from the data. These constraints are encoded on the basis of feature functions $h_m(x)$, and the expectation of each feature h_m over the model p is required to be equal to the observed expectation:

$$\sum_x p(x) h_m(x) = \sum_x \tilde{p}(x) h_m(x) \; ,$$

where $\tilde{p}$ is the empirical distribution in a training sample. The maximum entropy model can be shown to be unique and to have an exponential form involving a weighted sum over the feature functions h_m. In Equation (5.8), the notation t_0^n is used again for the lemma–tag representation of an input word (this was denoted by T in Equations (5.5)-(5.7) for notational simplicity):

$$p(f, T|e) = p_\Lambda(f, t_0^n|e) \;=\; \frac{\exp\left[\sum_m \lambda_m h_m(e, f, t_0^n)\right]}{\sum_{\tilde{f}, \tilde{t}_0^n} \exp\left[\sum_m \lambda_m h_m(e, \tilde{f}, \tilde{t}_0^n)\right]} \; , \tag{5.8}$$

where $\Lambda = \{\lambda_m\}$ is the set of model parameters with one weight λ_m for each feature function h_m. Furthermore, the maximum entropy model can be shown to be the maximum likelihood model in the class of exponential models given by Equation (5.8). Besides, the log-likelihood of the training corpus is concave in the model parameters Λ, and thus it is possible to implement converging iterative training procedures like described by [Darroch & Ratcliff 72] or [Della Pietra & Della Pietra+ 95]. For an introduction to maximum entropy modeling and the corresponding training procedures, the reader is referred to the corresponding literature, for instance [Berger & Brown+ 96a] or [Ratnaparkhi 97], which also contains proofs for some of the characteristics of maximum entropy models.

In the experiments presented in this thesis, the sum over the word forms $\tilde{f}$ and the readings $\tilde{t}_0^n$ in the denominator of Equation (5.8) is restricted to the readings of word forms having the same base form and partial reading as a word form f'' aligned at least once to e.

The new lexicon model $p_\Lambda(f, t_0^n|e)$ can now replace the usual lexicon model $p(f|e)$, compared to which it has the following main advantages:

1. The decomposition of the modeled events into feature functions allows for providing meaningful probabilities for word forms that have not occurred during training as long as the involved feature functions are well-defined. See also the argumentation on page 40 and the definition of first-level and second-level feature functions in Section 5.2.1.

2. Introducing the hidden variable $T = t_0^n$ and constraining the lexicon probability to be zero for interpretations considered non-valid readings of f (that is for $t_0^n \notin \mathcal{T}(f)$) amounts to making context information from the complete sentence f_1^J locally available: The sentence context was taken into account by the morpho-syntactic analyzer which chose the valid readings $\mathcal{T}(f)$.

5.2.1 Definition of feature functions

There are numerous possibilities of defining feature functions, as there are no constraints like the requirement that the components be disjoint and statistically independent or that different feature functions should have the same parametric form. Still it is necessary to restrict the number of parameters to be optimized in order to make the procedure of training them feasible. For the experiments reported in this thesis, the following types of feature functions have been defined on the basis of the lemma–tag representation (see Section 3.4):

1st-level: $m = \{L, \tilde{e}\}$, where L is the base form:

$$h^1_{L,\tilde{e}}(e, f, t_0^n) \;=\; \begin{cases} 1 & \text{if } e = \tilde{e} \text{ and } t_0 = L \text{ and } f \in \mathcal{F}(t_0^n) \quad (*) \\ 0 & \text{otherwise} \end{cases}$$

2nd-level: $m = \{T, L, \tilde{e}\}$, with subsets T of cardinality $\leq n$ of morpho-syntactic tags considered relevant (see Section 3.5 on page 15 for a description of the detection of relevant tags):

$$h^2_{T,L,\tilde{e}}(e, f, t_0^n) \;=\; \begin{cases} 1 & \text{if } (*) \text{ and } T \subseteq t_1^n \qquad\qquad (**) \\ 0 & \text{otherwise} \end{cases}$$

3rd-level: $m = \{F, T, L, \tilde{e}\}$, with the fully inflected original word form F:

$$h^3_{F,T,L,\tilde{e}}(e, f, t_0^n) \;=\; \begin{cases} 1 & \text{if } (**) \text{ and } F = f \\ 0 & \text{otherwise} \end{cases} \quad .$$

In terms of the hierarchy introduced on page 30, this means that information at three different levels in the hierarchy are combined. The subsets T of relevant tags mentioned above fix the intermediate level.[1] This choice of the types of features as well as the choice of the subsets T is reasonable but somewhat arbitrary. Alternatively one can think of defining a much more general set of features and applying some method of feature selection, as has been done for example by [Foster 00], who compared different methods for feature selection within the task of translation modeling for statistical machine translation.

Note that in contrast to Section 5.1.4, where there were only $n + 1$ parameters, the log-linear model introduced here uses one parameter per feature. For the Verbmobil task for example there are approximately 162 000 parameters: $47.8k$ for the first order features, $55.7k$ for the second order features and $58.5k$ for the third order features.

5.2.2 Training procedure

The overall process of training and testing with hierarchical lexicon models is depicted in Figure 5.3. The restructuring transformations presented in Chapter 4 can still be applied. This can even be advantageous, like for instance in the case of multi-word phrases, which jointly fulfill a syntactic function: Not merging them would raise the question of how to distribute the syntactic tags which have been associated with the whole phrase. Similarly, prepending detached verb prefixes prevents false interpretations of the prefix part. On the other hand, compound splitting cannot easily be applied here. The reasons are analogous to those mentioned with respect to the phrases.

Again, the alignment on the training corpus is trained using the original source language corpus containing inflected word forms. This alignment is then used to count the co-occurrences of the annotated "words" in the lemma–tag representation of the source language corpus with the words in the target language corpus. These event counts are used for the maximum entropy (ME) training of the model parameters Λ. Two different toolkits for ME training were used:

- `Yasmet`, a toolkit for conditional maximum entropy models [Och 01]. The optimization algorithm is generalized iterative scaling (GIS) [Darroch & Ratcliff 72].
- A maximum entropy toolkit implemented by E.S. Ristad[2], who uses the improved iterative scaling (IIS) algorithm [Della Pietra & Della Pietra$^+$ 95].

[1] Of course, there is not only one set of relevant tags, but at least one per part of speech. In order to keep the notation as simple as possible, this fact is not accounted for in the formulae and the textual descriptions.

[2] This toolkit was presented in a tutorial in conjunction with the ACL-EACL 1997 in Madrid. At that time it was freely available. Today, it can be purchased from [Ristad 01].

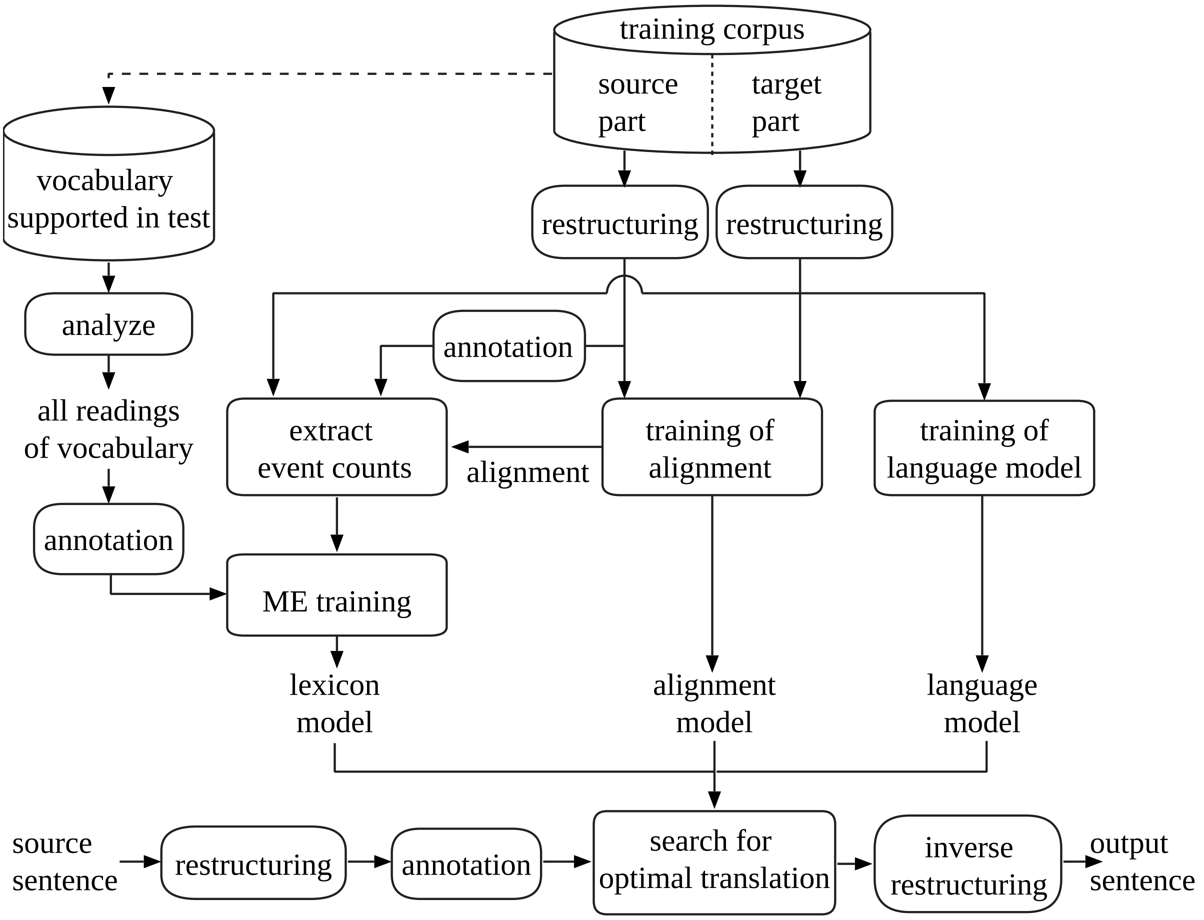


Figure 5.3: Training and test with hierarchical lexicon. "(inverse) restructuring", "analyze" and "annotation" all require morpho-syntactic analysis of the transformed sentences.

The translation results after training with `Yasmet` were consistently slightly better, but experiments on unrestricted bilingual training corpora could only be performed using Ristad's toolkit, because `Yasmet`'s memory management was not efficient enough yet[3].

The probability mass is distributed over (all readings of) the source language word forms to be supported for test (not necessarily restricted to those occurring during training). The only precondition is that the firing features for these unseen events are known. This "vocabulary supported in test", as it is named in Figure 5.3 can be a predefined closed vocabulary, as is the case in Verbmobil, where the output of a speech recognizer with limited output vocabulary is to be translated. In the easiest case it is identical to the vocabulary found in the source language part of the training corpus. The other extreme would be an extended vocabulary containing all automatically generated inflected forms of all base forms occurring in the training corpus. This vocabulary is annotated with morpho-syntactic tags, ideally under consideration of all possible readings of all word forms.

[3] June 2002

To enable the application of the hierarchical lexicon model the source language input sentences in test have to be analyzed and annotated with their lemma–tag representation before the actual translation process. So far, the sum over the readings in Equation (5.7) has been ignored, because applying the techniques for reducing the amount of ambiguity described in Section 3.3 and the disambiguated conventional dictionaries resulting from the approach presented later in this chapter in Section 5.3, there remains almost always only one reading per word form.

5.2.3 Application using alignment templates

Some additional considerations are necessary when the hierarchical lexicon model is applied using the alignment templates system [Och & Tillmann$^+$ 99]:

- Typically, when using the alignment templates system for translation, a direct translation model is applied, because this is computationally less expensive than the translation model in the Bayesian or source–channel formulation, and it has shown to yield comparable results [Och & Tillmann$^+$ 99]. The equivalent of Equation (5.8) then has the following form:

$$p_\Lambda(e|f,t_0^n) \;=\; \frac{\exp\left[\sum\limits_m \lambda_m h_m(e,f,t_0^n)\right]}{\sum\limits_{e'} \exp\left[\sum\limits_m \lambda_m h_m(e',f,t_0^n)\right]} \; . \qquad (5.9)$$

- The alignment templates system makes use of bilingual word classes to partition the source and target language vocabularies. These classes are learned with an automatic clustering algorithm on the basis of an alignment between the positions in the parallel sentences of the training corpus. This alignment results from the standard expectation–maximization (EM) translation model training [Dempster & Laird$^+$ 77, Och 99]. In connection with hierarchical lexicon models, the standard clustering algorithm to determine bilingual classes can be applied using the lemma–tag representation of the source language corpus, where the full word forms have been augmented by morpho-syntactic tags and the base form.

  Usually the only predefined requirement to be met by the clustering algorithm is the *number* of resulting bilingual classes, but in principle there is also the possibility to restrict the set of admissible ways of partitioning the vocabularies. The author of this thesis has experimented with the following setups:

  No restriction: The simplest approach is not to impose any restrictions.

  Word form restriction: All readings of one full word form must be placed into the same class.

  Base form restriction: All readings of all inflected forms of the same base form must belong to the same class.

- The alignment templates themselves are finally learned on the transformed corpus, using these bilingual classes and the alignment trained on the (possibly reordered) corpora in the original full word form representation.

5.3 Conventional dictionaries: Disambiguation without context

Conventional bilingual dictionaries are often used as an additional evidence to better train the model parameters of statistical machine translation. The notion "conventional dictionary" here denotes bilingual collections of word or phrase pairs predominantly collected "by hand", usually by lexicographers, as opposed to the probabilistic lexica, which in the framework of statistical machine translation are learned automatically from bilingual sentence aligned corpora. Apart from the theoretical problem of how to incorporate external dictionaries in a mathematically sound way into a statistical framework for machine translation [Brown & Della Pietra$^+$ 93a], there are also some pragmatic difficulties: one of the disadvantages of these conventional dictionaries as compared to full bilingual corpora is the fact that their entries typically contain single words or short phrases on each language side. Consequently, it is not possible to distinguish between the translations for different readings of a word. In normal bilingual corpora, the words can often be disambiguated by taking into account the sentence context in which they occur. For example, from the context in the sentence "Ich werde *die* Zimmer buchen.", it is possible to infer that 'Zimmer' in this sentence is plural and has to be translated by 'rooms' in English, whereas the correct translation of 'Zimmer' in the sentence "Ich hätte gerne *ein* Zimmer." is the singular form 'room'. The dictionary used by our research group for augmenting the bilingual data contains two entries for 'Zimmer': ('Zimmer'|'room') and ('Zimmer'|'rooms').

The idea of the approach described in this section is based on the observation that in many of the cases of ambiguous entries in dictionaries, the second part of the entry — that is the other language side, contains the information necessary to decide upon the interpretation. In some other cases, the same kind of ambiguity is present in both languages, and it would be possible and desirable to associate the (semantically) corresponding readings to each other. The method proposed here takes advantage of these facts in order to disambiguate dictionary entries. The author expects the resulting disambiguated dictionaries to be useful for natural language processing tasks other than statistical machine translation, like multi-lingual information retrieval and document classification.

The proposed procedure for disambiguating conventional dictionaries is sketched in Figure 5.4. Apart from the dictionary, in the following denoted with D, a bilingual corpus is required to learn tag sequence translation probabilities. As the word forms in this corpus, in the following denoted with C_1, do not have to match those in D, the only requirement for C_1 is that it must consist of the same language pair. C_1 is not necessarily the training corpus for the translation task in which the disambiguated version of D will be used. It does not even have to be taken from the same domain.

- A word alignment A_1 between the sentences in C_1 is trained with some automatic

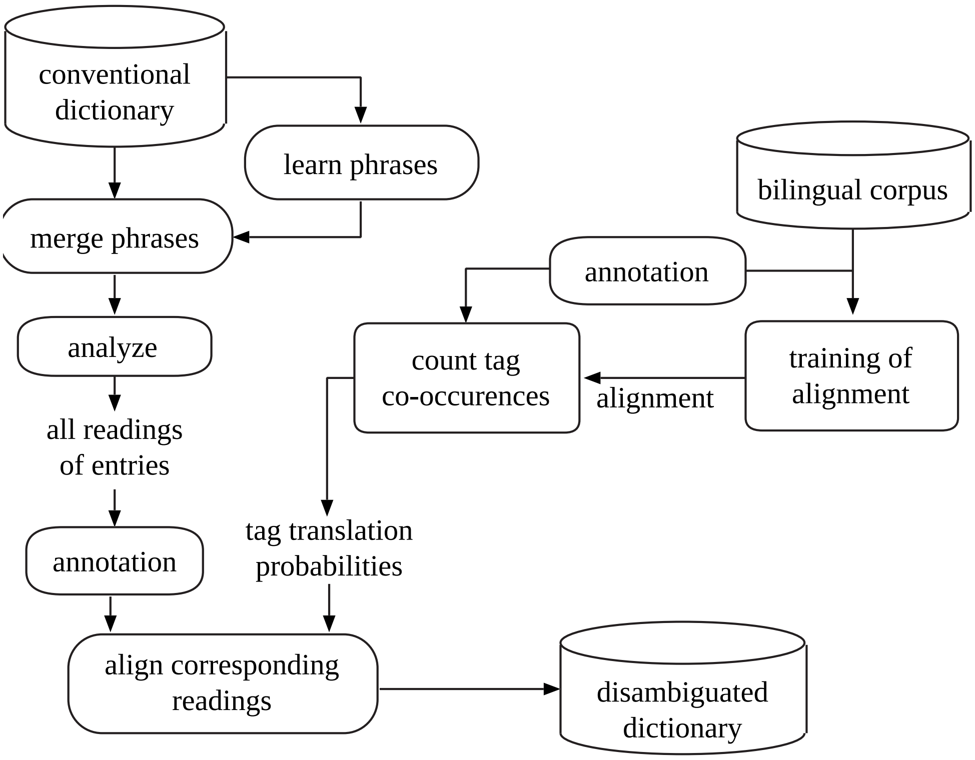


Figure 5.4: Disambiguation of conventional dictionaries.“learn phrases”, “analyze” and “annotation” require morpho-syntactic analysis of the transformed sentences.

alignment algorithm. Then the words in the bilingual corpus are replaced by a reduced form of their lemma–tag representation, where *only a subset of their morpho-syntactic tags* is retained — even the base form is dropped. The remaining subset of tags, in the following denoted with T_f for the source language and T_e for the target language, consists of tags considered relevant for the task of aligning corresponding readings. This is not necessarily the same set of tags considered relevant for the task of *translation*, which was used for example to fix the intermediate level for the log-linear lexicon combination on page 38. In the case of the Verbmobil corpus, the maximal length of a tag sequence is 5.

- The alignment A_1 is used to count the frequency of a certain tag sequence $\mathbf{t}_f$ in the source language to be associated with another tag sequence $\mathbf{t}_e$ in the target language and to compute the *tag sequence translation probabilities* $p(\mathbf{t}_f|\mathbf{t}_e)$ as relative frequencies. For the time being, these tag sequence translation probabilities associate readings of *words* in one language with readings of *words* in the other language: multi-word sequences are not accounted for.

- To alleviate this shortcoming it is possible and advisable to automatically detect and merge multi-word phrases. As mentioned in Section 4.1.4, the conventional bilingual dictionary itself can be used to learn and validate these phrases. This step is called “learn phrases” in Figure 5.4. The resulting multi-word phrases P_e for the target language and P_f for the source language are afterwards concatenated within D to form entries consisting of pairs of “units”.

- The next step is to analyze the word forms in D and generate all possible readings of all entries. It is also possible to ignore those readings that are considered unlikely for the task under consideration by applying the domain specific preference rules proposed in Section 3.3. The process of generating all readings includes replacing word forms by their lemma–tag representation, which is thereafter reduced by dropping all morpho-syntactic tags not contained in the tag sets T_f and T_e.

- Using the tag sequence translation probabilities $p(\mathbf{t}_f|\mathbf{t}_e)$, the readings in one language are aligned to readings in the other language. These alignments are applied to the full lemma–tag representation (not only tags in T_f and T_e) of the expanded dictionary containing one entry per reading of the original word forms. The highest ranking aligned readings according to $p(\mathbf{t}_f|\mathbf{t}_e)$ for each lemma are preserved.

The resulting disambiguated dictionary contains the entries (‘`Zimmer-noun-sg.`’|‘room’) and (‘`Zimmer-noun-pl.`’|‘rooms’) for the German word ‘Zimmer’. Note that this augmented dictionary, in the following denoted by $\hat{D}$, has more entries than D due to the step of generating all readings. The two entries (‘beabsichtigt’|‘intends’) and (‘beabsichtigt’|‘intended’) for example produce three new entries:
(‘`beabsichtigen-verb-ind.-pres.-sg.-3rd`’|‘intends’),
(‘`beabsichtigt-verb-past-part.`’|‘intended’) and
(‘`beabsichtigt-adjective-pos.`’|‘intended’).

5.4 Overall procedure for training with scarce resources

The motivation of the work on hierarchical lexicon models was the goal to take into account the interdependencies of related words, that is of inflected forms of the same base form. This is especially relevant when inflected languages like German are involved and when training data is sparse, as is often the case. In this situation many of the inflected word forms to account for in test do not occur during training. Sparse bilingual training data also makes additional conventional dictionaries especially important. Enriching them by aligning corresponding readings is in particular useful when they are used in conjunction with a hierarchical lexicon which can access the information necessary to distinguish readings via morpho-syntactic tags. The restructuring operations described in Chapter 4 also help coping with the data sparseness problem, because they make corresponding sentences more similar and thus reduce the perplexity of the corpora. This section proposes a procedure for combining the achievements of this thesis in order to improve the translation quality despite the sparseness of data. In other words, the combination of the suggested methods is expected to reduce the amount of full bilingual training data necessary to achieve reasonably good results with statistical machine translation systems. Figure 5.5 sketches the proposed procedure.

Three different bilingual corpora C_1, C_2, C_3, one monolingual target language corpus and a conventional bilingual dictionary D can contribute in various ways to the overall result. It is important to note here that C_1, C_2, C_3 can, but need not be distinct, and

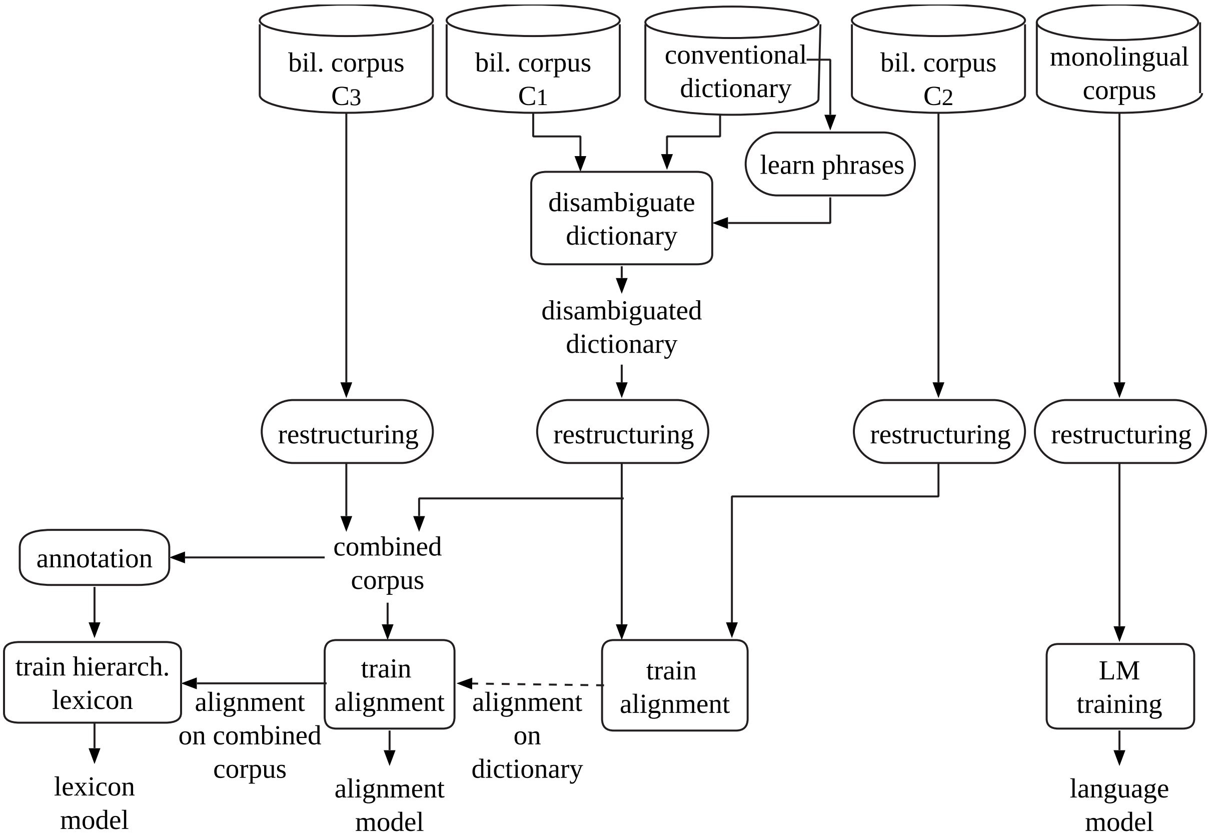


Figure 5.5: Training with scarce resources. "restructuring", "learn phrases" and "annotation" all require morpho-syntactic analysis of the transformed sentences.

the monolingual corpus can be identical to the target language part of C_3. Besides these corpora can be taken from different domains and C_1 and C_2 can be (very) small. C_3 is the only one of the three bilingual corpora that has to represent the domain and the vocabulary for which the translation system is built, and only the size of C_3 and the monolingual corpus have substantial effect on the translation quality. It is interesting to note though that a basic statistical machine translation system with an accuracy near 50% can be built *without any* domain specific bilingual corpus C_3, only on the basis of a disambiguated dictionary and the other methods suggested in this thesis, as Table 7.12 shows.

- In a first step, multi-word phrases are learned and validated on the dictionary D in the way described in Section 4.1.4. These multi-word phrases are concatenated in D. Then an alignment A_1 is trained on the first bilingual corpus C_1. On the basis of this alignment, the tag sequence translation probabilities are extracted which are needed to align corresponding readings in the dictionary as proposed in Section 5.3. The result of this step is an expanded and disambiguated dictionary $\hat{D}$. For this purpose, C_1 does not have to cover the vocabulary of D. Besides C_1 can be comparatively small given the limited number of tag sequence pairs $(\mathbf{t}_f|\mathbf{t}_e)$ for which translation probabilities must be provided: In the Verbmobil training corpus for example there are only 261 different German and 110 different English tag sequences.

- In the case that the domain specific corpus C_3 is very small it might be advantageous to determine the word alignment A_2 on $\hat{D}$ separately. As dictionaries consist of a limited number of pairs of words or at most pairs of short phrases, it is feasible to hand-align these entries. Another possibility is to train A_2 together with a different corpus C_2, which covers at least part of the words in the dictionary: C_2 and $\hat{D}$ can be considered as one bilingual corpus on which the standard training algorithm for word alignments is applied and A_2 is the dictionary portion of the resulting alignment. Experiments on Verbmobil show that the effect of this step is negligible or even slightly disadvantageous on this task: As can be concluded from Table 7.11, the task of associating words within conventional dictionaries seems to be so trivial that it can quite reliably be performed without any other knowledge source than the dictionary itself. On the other hand this conclusion might not be valid for other tasks.

- In the next step the third bilingual corpus C_3 and $\hat{D}$ are combined and a word alignment A_3 for both is trained. If applicable, the alignment A_2 on the dictionary may serve as an initial setting for the training of A_3. A_2 can alternatively replace the dictionary portion of A_3, if A_2 is considered more reliable. The resulting word alignment on C_3 and $\hat{D}$ is denoted with A_4. C_3 and $\hat{D}$ and A_4 are presented as input to the maximum entropy training of a hierarchical lexicon model as described in Section 5.2.2.

- The language model can be trained on a separate monolingual corpus. As monolingual data is much easier and cheaper to compile, this corpus might be (substantially) larger than the target language part of C_3.

The corpora and the dictionary can all be reordered before the actual training according to the suggestions in Chapter 4 in order to reduce their perplexity and to facilitate the task of establishing word alignments on them.

Chapter 6

Semi-automatic Evaluation of Machine Translation

> . . . , the sums that are being spent on MT . . . are large enough to make virtually inevitable the production of a second ALPAC report sometime in the next few years. . . . The report will be the more devastating for the fact that much of the money has in fact been spent frivolously, and much of the work has been incompetent, even by today's limited standards.
>
> (Martin Kay, 1986)

In this chapter a tool for the evaluation of translation quality is presented which is designed to meet the typical requirements in the framework of statistical machine translation (SMT) research. Evaluation criteria which are more adequate than pure edit distance are defined. Using the tool and the corresponding graphical user interface, the measurement along these quality criteria can be performed *semi-automatically* in a fast, convenient and above all consistent way.

One of the characteristics of SMT research is the fact that different prototypes of translation systems are tested *many times* on one distinct set of test sentences (for example for adjusting parameter settings or examining the effects of slight changes in system design). Sometimes the resulting translations differ only in a small number of words. The idea now is to store an input sentence together with all translations that have already been manually evaluated together with their scores in a database $\mathcal{DB}$. In addition, a suitable graphical user interface permits convenient manipulation of the database and provides means for calculating several kinds of statistics on it. This approach and the resulting evaluation tool [`EvalTrans`] provides the following opportunities:

- Define new types of quality criteria (see Sections 6.1.5 and 6.1.1).
- *Automatically* return the scores of translations that have already occurred at least once. Hence, consistency of quality judgments over time is guaranteed (see Section 6.1.4).
- *Extrapolate* scores for new translations by comparison with similar sentences in

$\mathcal{DB}$ (see Section 6.1.4). The advantage is among other things that costly human evaluations need only been done for those systems, for which this estimate yields promising results.

- Facilitate the evaluation of new translations that differ only slightly from previous ones (see Section 6.3.2). This makes evaluation more efficient and helps maintenance of consistency.

6.1 Evaluation measures

6.1.1 The multi-reference word error rate

An "enhanced" word error rate is computed as follows: a translation t is compared to each reference out of a set of references of the test sentence s and the edit distance of t and the most similar reference is used for the computation of the multi-reference word error rate. Let $\mathcal{R}(s)$ be the set of reference translations for s and $d(t, r)$ the edit distance between a translation candidate t and a reference $r \in \mathcal{R}(s)$. $d(t, \mathcal{R}(s))$ is the minimal edit distance of t compared to any reference of s:

$$d(t, \mathcal{R}(s)) = \min_{r \in \mathcal{R}(s)} d(t, r) \ .$$

The *multi-reference word error* (m-WER) of a set of translations $t_1^n = t_1 \dots t_n$ for a test corpus $s_1^n = s_1 \dots s_n$ can then be defined as follows:

$$\text{m-WER}(s_1^n, t_1^n) = \frac{\sum\limits_{i=1}^{n} d(t_i, \mathcal{R}(s_i))}{\sum\limits_{i=1}^{n} \frac{1}{|\mathcal{R}(s_i)|} \cdot \sum\limits_{r \in \mathcal{R}(s_i)} |r|} \ , \tag{6.1}$$

where $|r|$ is the length of the reference r and $|\mathcal{R}(s_i)|$ is the number of references for the i-th test sentence s_i. Note that the denominator consists of a sum over the *means* of all reference sentence lengths for an input language sentence. An alternative would be to define the set of the most similar references

$$\hat{r}(t, \mathcal{R}(s)) = \{r \in \mathcal{R}(s) \mid d(t, r) = d(t, \mathcal{R}(s))\}$$

and to replace the denominator of Definition (6.1) by the following expression:

$$\sum_{i=1}^{n} \frac{1}{|\hat{r}(t_i, \mathcal{R}(s_i))|} \cdot \sum_{r \in \hat{r}(t_i, \mathcal{R}(s_i))} |r| \ , \tag{6.2}$$

which amounts to summing up over the average lengths of the elements in the sets $\hat{r}(t_i, \mathcal{R}(s_i))$. Compared to the definition represented in Equation (6.1), this alternative normalization has the disadvantage of being dependent on the translation candidates themselves: As long candidates are more likely to have longer most similar references, the denominator tends to be larger for systems, which on average produce longer translations. The m-WER would then be biased towards such “wordy” systems.

The idea of computing the difference to more than one reference has been used before [Alshawi & Bangalore^{+} 98]. The advantage here is that the set of reference sentences comes for free as the database is enlarged: all translations of s in $\mathcal{DB}$ that have been judged “perfect” (score K, see Section 6.1.3) can be regarded as a reference for s. Besides, the new reference sentences produced by the translation systems under consideration are more adequate for the purpose of word-by-word comparison, because human translators tend to translate more or less freely, frequently resorting to synonyms and sentence restructuring.

6.1.2 The multi-reference word error rate with inversions

The usual (multi-reference) word error rate relies on the Levenshtein distance using the editing operations substitution, insertion and deletion. A translation candidate which differs from a reference only in that a phrase as a whole has been moved within the sentence, is assigned an often inadequately high error rate. This is the motivation for introducing inversion operations in addition to the other editing operations in order to penalize the inversion of phrases only by a constant cost.

Computing the minimum number of phrase movements necessary to transform one sentence into another causes exponential computation costs. For pragmatic reasons it is thus necessary to restrict the search to certain types of permutations of the sentence. In particular, overlapping and “inside-out” inversions seem unnatural. On the other hand, the recursive nature of languages calls for nested inversion operations, that is inversion within inverted sub-phrases. [Wu 95] has suggested these kinds of correspondences in his definition of a variant of bilingual context free grammars: an “inversion transduction grammar” (ITG) yields one single parse tree for a parallel sentence pair in two languages, while allowing inversions of corresponding sub-phrases across languages. Each ITG can be transformed into a normal form, which allows the following productions:

				applies to	
				in source language	in target language
lexical production:	C	$\rightarrow$	x/y	x	y
straight concatenation:	C	$\rightarrow$	$[AB]$	$A \cdot B$	$A \cdot B$
inverted concatenation:	C	$\rightarrow$	$\langle AB \rangle$	$A \cdot B$	$B \cdot A$

In the lexical productions, either x or y can be empty (ϵ). Parses can be represented by bracketing schemes. As an example, the following representation stands for a parse of the sentence pair (“Ich möchte Sie etwas fragen”/“I would like to ask you something”):

[[‘Ich’/‘I’ ‘möchte’/‘would like’ ϵ/‘to’] ⟨[‘Sie’/‘you’ ‘etwas’/‘something’] ‘fragen’/‘ask’⟩]

In the ITG formalism, a sentence as well as a sub-phrase itself is a composition of smaller sub-phrases. Two adjacent sub-phrases may be inverted. Movements across longer distances are achieved by first inverting larger fragments of the sentence. Within each sub-phrase, further inversions may take place, but crossings between two sub-phrases are not allowed. As a consequence, so-called “inside-out” matchings as sketched in Figure 6.1 for sentences of length 4 are excluded. [Wu 95] claims that this is actually a

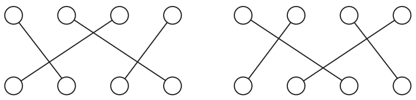


Figure 6.1: Alignments for sentences of length 4 not permitted in the ITG formalism.

benefit, because these types of crossings are very untypical between natural languages. The restrictions imposed by the ITG formalism enforce phrasal contiguity and above all achieve polynomial complexity for the parsing algorithm. The *multi-reference word error rate with inversions* (m-invWER) allows for the inversion of phrases in very much the same way as described above: The ITG formalism can be applied for establishing correspondences between candidate translations and references. A new edit distance is defined by introducing costs c for inversions in addition to the costs for substitutions, insertions and deletions:

$$\begin{array}{lllllll}
A & \rightarrow & [AA] & & c(\cdot) = 0 & \text{straight concatenation} & (1) \\
A & \rightarrow & \langle AA \rangle & & c(\cdot) = 1 & \text{inverted concatenation} & (2) \\
A & \rightarrow & x/x & & c(\cdot) = 0 & \text{match} & (3) \\
A & \rightarrow & x/y & (x \neq y) & c(\cdot) = 1 & \text{substitution} & (4) \\
A & \rightarrow & x/\epsilon & & c(\cdot) = 1 & \text{deletion} & (5) \\
A & \rightarrow & \epsilon/y & & c(\cdot) = 1 & \text{insertion} & (6)
\end{array}$$

Note that without rule (2), the minimum editing cost is the Levenshtein distance. The m-invWER is the minimal number of insertions, deletions, substitutions and inversions necessary to transform the evaluated translation candidates into the most similar among the reference translations, weighted by the length of the references, i.e. the only difference to the definition of the m-WER in Equation (6.1) is the definition of the edit distance.

Despite the rejection of certain types of permutations the computational effort for calculating the m-invWER is substantially higher than for the m-WER and implementing an efficient search algorithm including pruning techniques is essential [Leusch & Nießen 02].

6.1.3 The subjective sentence error rate

Automatic evaluation methods are very helpful for the daily work of MT research, but as yet, human inspection is considered more meaningful and reliable. Therefore, manual evaluation is performed at least from time to time and for the most promising system variants. The evaluation scheme is defined such that each translation t for an input sentence s is assigned a score $v(s,t)$ ranging from 0 points ("nonsense") to K points ("perfect"). A range from zero to ten in steps of one was chosen for our purposes. Table 6.1 gives an idea of how these scores should be interpreted. The *subjective sentence error rate* (SSER) of a set of translations t_1^n for a test corpus s_1^n can then be defined as follows:

$$\mathrm{SSER}(s_1^n, t_1^n) = 1 - \frac{1}{Kn} \sum_{i=1}^{n} v(s_i, t_i) \ . \tag{6.3}$$

This definition is based on the assumption that each individual score has the same weight, not taking the lengths of the scored sentence or the source language equivalent into ac-

Table 6.1: Definition of scores for human evaluation.

$0 \equiv$	nonsense.
$1 \equiv$	some aspects of contents are conveyed.
	...
$5 \equiv$	understandable with major syntactic errors.
	...
$9 \equiv$	OK. Only slight errors in register or style or minimal syntax errors.
K=10 $\equiv$	perfect.

count. Of course the sentence lengths are implicitly considered by the human evaluators in that small errors in long sentences are not penalized as much as in short sentences.

6.1.4 The extrapolated subjective sentence error rate

When a new set of translations for the test corpus s_1^n is generated, some of the pairs (s_i, t_i) have typically already been evaluated and their scores can be extracted from the database $\mathcal{DB}$. The remaining – really new – pairs are evaluated and added to $\mathcal{DB}$. Alternatively, the score for a new translation t_i can be extrapolated by comparison with other translations: Let $\mathcal{T}(s_i)$ be the set of evaluated translations for s_i stored in $\mathcal{DB}$. Provided that $\mathcal{T}(s_i) \neq \emptyset$, t_i is compared to all candidates in $\mathcal{T}(s_i)$ to calculate the minimum difference in terms of edit distance (0, if $(s_i, t_i) \in \mathcal{DB}$):

$$d(t_i, \mathcal{T}(s_i)) = \min_{t \in \mathcal{T}(s_i)} d(t_i, t) \tag{6.4}$$

and adopt the average score of the most similar candidates

$$\hat{t}(t_i, \mathcal{T}(s_i)) = \{t \in \mathcal{T}(s_i) \mid d(t_i, t) = d(t_i, \mathcal{T}(s_i))\} \tag{6.5}$$

to extrapolate the score of t_i:

$$\hat{v}(s_i, t_i, \mathcal{T}(s_i)) = \frac{1}{|\hat{t}(t_i, \mathcal{T}(s_i))|} \sum_{t \in \hat{t}(t_i, \mathcal{T}(s_i))} v(s_i, t) \ . \tag{6.6}$$

The extrapolated score is defined as follows:

$$\tilde{v}(s, t) = \begin{cases} v(s, t) & \text{if } (s, t) \in \mathcal{DB} \ , \\ \hat{v}(s, t, \mathcal{T}(s)) & \text{otherwise } \ . \end{cases} \tag{6.7}$$

The *extrapolated subjective sentence error rate* (eSSER) results from replacing $v(s_i, t_i)$ by $\tilde{v}(s_i, t_i)$ in Definition (6.3). The eSSER has been used during the development of the methods described in the preceding chapters of this thesis for intermediate assessments. For the final results presented in Chapter 7, all translation candidates have been evaluated, and the figures for the SSER in the corresponding tables are not extrapolated.

As an indicator of the accuracy of this extrapolation serves the *average normalized edit distance* $\bar{d}(t_1^n)$:

$$\bar{d}(t_1^n) = \frac{1}{n} \sum_{i=1}^{n} \frac{d(t_i, \mathcal{T}(s_i))}{|s_i|} \ , \tag{6.8}$$

which depends on the rate of new translations as well as on the degree of similarity of these new hypotheses to the other candidates in the database. The above definition is a consequence of the definition of the SSER, which takes the quality judgments of all sentences as equally important.

The approach for estimating the quality of translations described above relies on Levenshtein alignments to find the most similar sentences according to edit distance. [Vogel & Nießen$^+$ 00] have suggested a method for improving the score extrapolation using weighted edit distance with individual costs for insertions, deletions and substitutions. These weights are learned in an automatic training procedure. Three different levels with increasing number of free parameters have been tested:

1. One insertion score I, one deletion score D, and one substitution score S.
2. Individual costs $I(w)$, $D(w)$, and $S(w_1, w_2)$ for each word w or word pair w_1, w_2.
3. For each source language sentence s individual costs $I_s(w)$, $D_s(w)$, and $S_s(w_1, w_2)$.

The best results were obtained with the last setup.

6.1.5 The information item (semantic) error rate

It remains unclear how to rank long sentences consisting of correct and wrong parts. To overcome this shortcoming of the SSER the notion of "information items" is introduced. Each input sentence in the database is partitioned into segments representing the relevant items of information to be conveyed. Let $\mathcal{II}(s)$ be the set of information items for s. Then for each element of this set, a candidate translation t is assigned either "OK" or one out of a predefined set of error classes. For our purposes the following categories were chosen:

OK: The information is correctly conveyed and the translation is syntactically sound.

syntax: The information is correctly conveyed, but there are slight syntactic or stylistic errors, which do not seriously deteriorate the intelligibility.

missing: No part of the translation can be identified as translation of the information item, or a source language word has been inserted untranslated into the translation, for example because the word is not contained in the training corpus.

meaning: An ambiguous source language segment is translated wrongly. An example is the translation "I beat before." for "Ich schlage vor." ("I suggest.").

other: The information is not conveyed, but the categories "missing" and "meaning" are not adequate.

The *information item error rate* (IER) is the rate of information items not evaluated as "OK" for a set of translations t_1^n:

$$\text{IER}(s_1^n, t_1^n) = \frac{\sum\limits_{i=1}^{n} |\{ii | ii \in \mathcal{II}(s_i),\ ii \neq \text{“OK”}\}|}{\sum\limits_{i=1}^{n} |\mathcal{II}(s_i)|}\ , \tag{6.9}$$

For the *information item semantic error rate* (ISER), partial translations which are assigned the category "syntax" are considered correct:

$$\text{ISER}(s_1^n, t_1^n) = \frac{\sum\limits_{i=1}^{n} |\{ii | ii \in \mathcal{II}(s_i),\ ii \neq \text{“OK”},\ ii \neq \text{“syntax”}\}|}{\sum\limits_{i=1}^{n} |\mathcal{II}(s_i)|}\ . \tag{6.10}$$

6.1.6 Correlation of objective and subjective evaluation measures

For the comparison of the automatically computable ("objective") measures the manually assigned ("subjective") scores are considered as gold standard. Table 6.2 lists the correlations between the subjective quality score v and the objective measures (1) m-WER without inversions (see page 48), (2) m-WER with inversions (m-invWER) (see page 49) and (3) BLEU (see page 5). The correlations were computed using 40 hypotheses files with translation candidates from MT systems for the Verbmobil Eval-2000 German to English "`Test`" set comprising 251 source language sentences, and the corresponding evaluation database $\mathcal{DB}$. Three different types of comparisons were considered:

1. For each translation in $\mathcal{DB}$, correlate the automatically calculated scores and v.
2. For each hypothesis in the files, correlate the automatically calculated scores and v.
3. For each hypotheses file, correlate the automatically calculated scores and the SSER.

All three automatic measures correlate fairly well with the subjective sentence error rate for whole hypotheses files. On the contrary, when considering the scores for individual

Table 6.2: Correlation of automatic evaluation measures with manually assigned scores. Database: Verbmobil Eval-2000 German to English "`Test`".

measure	correlation with individual scores in DB	 hypotheses	 SSER for whole sets of hypotheses
m-WER	0.50	0.52	0.97
m-invWER	0.53	0.51	0.98
BLEU	0.33	0.41	0.98

sentences, both variants of the m-WER outperform the BLEU measure. A possible explanation might be the fact that BLEU is especially well-suited for measuring the syntactic well-formedness of translations, whereas human evaluators typically set great store by the preservation of meaning, at least within translation tasks like Verbmobil.

Figure 6.2 shows an effect observed on several evaluation databases when plotting the mean of the automatic scores for each category of the manually assigned scores: The BLEU score tends to overestimate the quality of very bad translations with a score, say below 3. Note that the ideal line for both variants of the m-WER in this picture would be along the graph of $1 - \frac{\text{score}}{K}$ and for BLEU along the graph of $\frac{\text{score}}{K}$. The fact that all curves in Figure 6.2 and especially the curve for the BLEU score are relatively flat in the range of medium quality translations corresponds to the observation that these measures discriminate rather well between translations of very different quality and also between very good and fairly good translations, but that small changes of the automatic measures in a medium range can be quite misleading (see for example the Tables 7.2, 7.6 and the first and the last line of Table 7.7 in the experimental part).

The comparison between the m-WER with and without inversions as revealed in Table 6.2 does not clearly answer the question which of both is better suited to substitute for the manual evaluation for purposes like large scale parameter tuning with thousands of translations to judge. Figure 6.3, which plots the objective scores versus the subjective

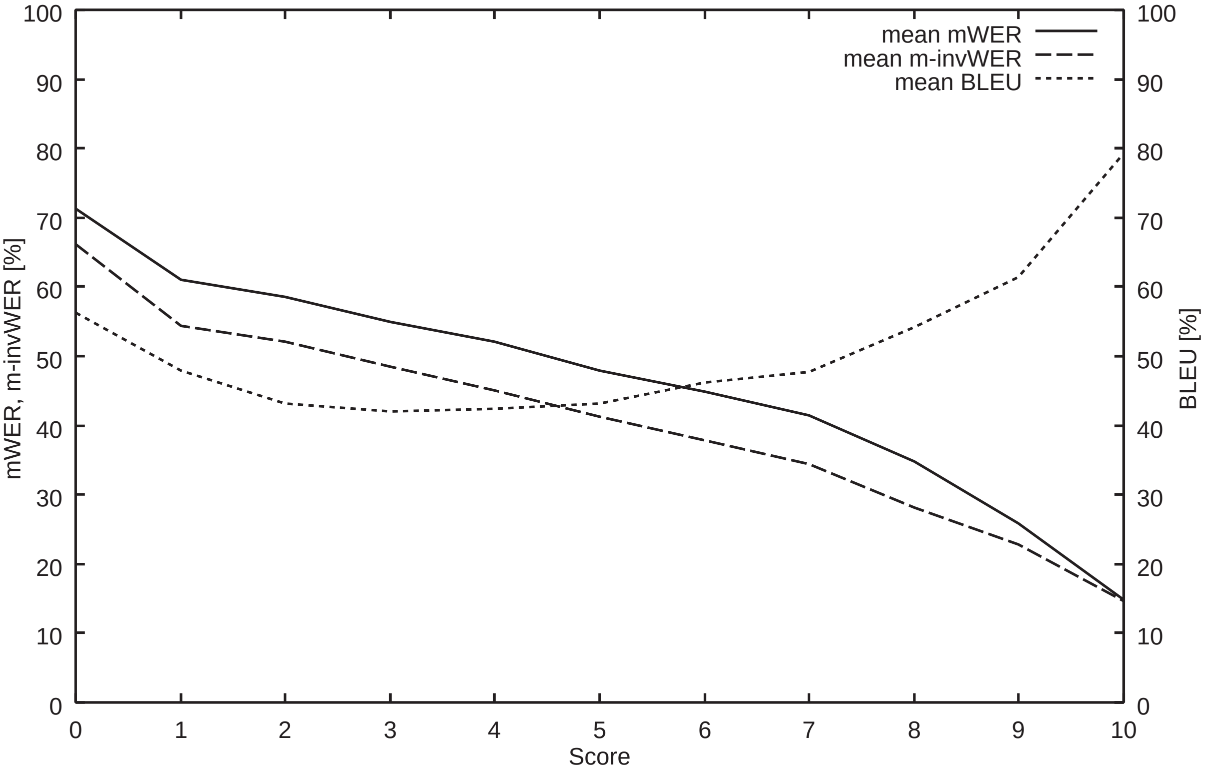


Figure 6.2: For each quality category, the means of the m-WER, the m-invWER and the BLEU score of the sentences in the database are displayed versus the manually assigned quality score. Database: Verbmobil Eval-2000 German to English "`Test`".

scores of the aforementioned 40 hypotheses files, also supports the conclusion drawn from the last column in Table 6.2, that all of the automatic measures seem to reflect equally well the subjective evaluation.

6.2 An XML format for evaluation databases

The manually judged translation candidates are stored in a database together with their scores. For our purposes, an XML format has been defined which is more flexible and less prone to mismatches than the format defined for the original BLEU measure as well as the format defined for the NIST variant of BLEU. An example of a source sentence in German, segmented into two information items, with two corresponding translations together with their evaluations is shown below.

```
<?xml version"1.0" encoding="iso-8859-1" ?><!DOCTYPE etdb SYSTEM "etdb.dtd">
<database>
 <version_id>$ Id: de_2000.etdb.xml,v 1.388 2002/05/06 11:25:54 schouten Exp$
 </version_id>
```

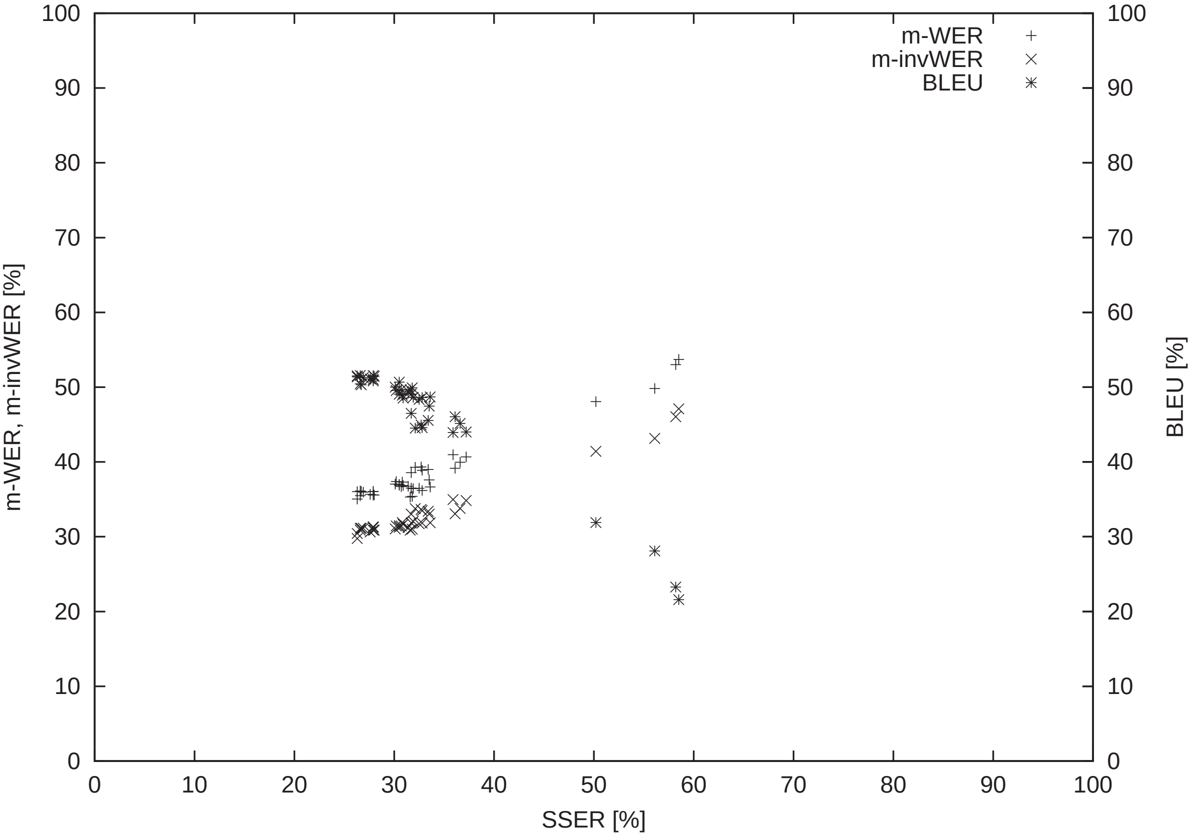


Figure 6.3: m-WER, m-invWER and BLEU score versus SSER of 40 hypotheses files with translation candidates for the Verbmobil Eval-2000 German to English “Test” set comprising 251 source language sentences.

```
 <source>
  <s_sent>alles klar. danke schoen.</s_sent>
  <ielist>
   <iedef id="0">alles klar.</iedef><iedef id="1">danke schoen.</iedef>
  </ielist>
  <targets>
   <tgt><t_sent>yes. thanks. fine.</t_sent>
    <eval val="6"/><comment>schoen translated by fine</comment></tgt>
   <tgt><t_sent>okay thanks.</t_sent>
    <eval val="10"/><ie id="0" val="ok"/><ie id="1" val="ok"/></tgt>
  </targets>
 </source>
</database>
```

The first line contains information about the character encoding and the second line points to the corresponding document type definition (DTD). The first entry of the "database" itself is the "version_id": Evaluation databases should be kept under version control. The version_id contains the information automatically updated by standard version control systems like *revision control system* (RCS) or *concurrent versions system* (CVS), namely the document source, the version number, time and date of the last update, the user who has performed the update (normally a person performing manual evaluations), and the document status. A database contains a structured list of "source" sentences, each of which has a list "ielist" of information item definitions "iedef" and a list of translations "targets". Each translation "tgt" can be assigned a quality score "eval" and an error class for each information item "ie". Comments about the quality judgments can be stored in the "comment" field.

6.3 The graphical user interface

A graphical user interface (GUI) facilitates the access to the database. Figure 6.4 gives an overview of the layout of the evaluation tool. The usual database manipulation operations are provided by the evaluation tool [EvalTrans]:

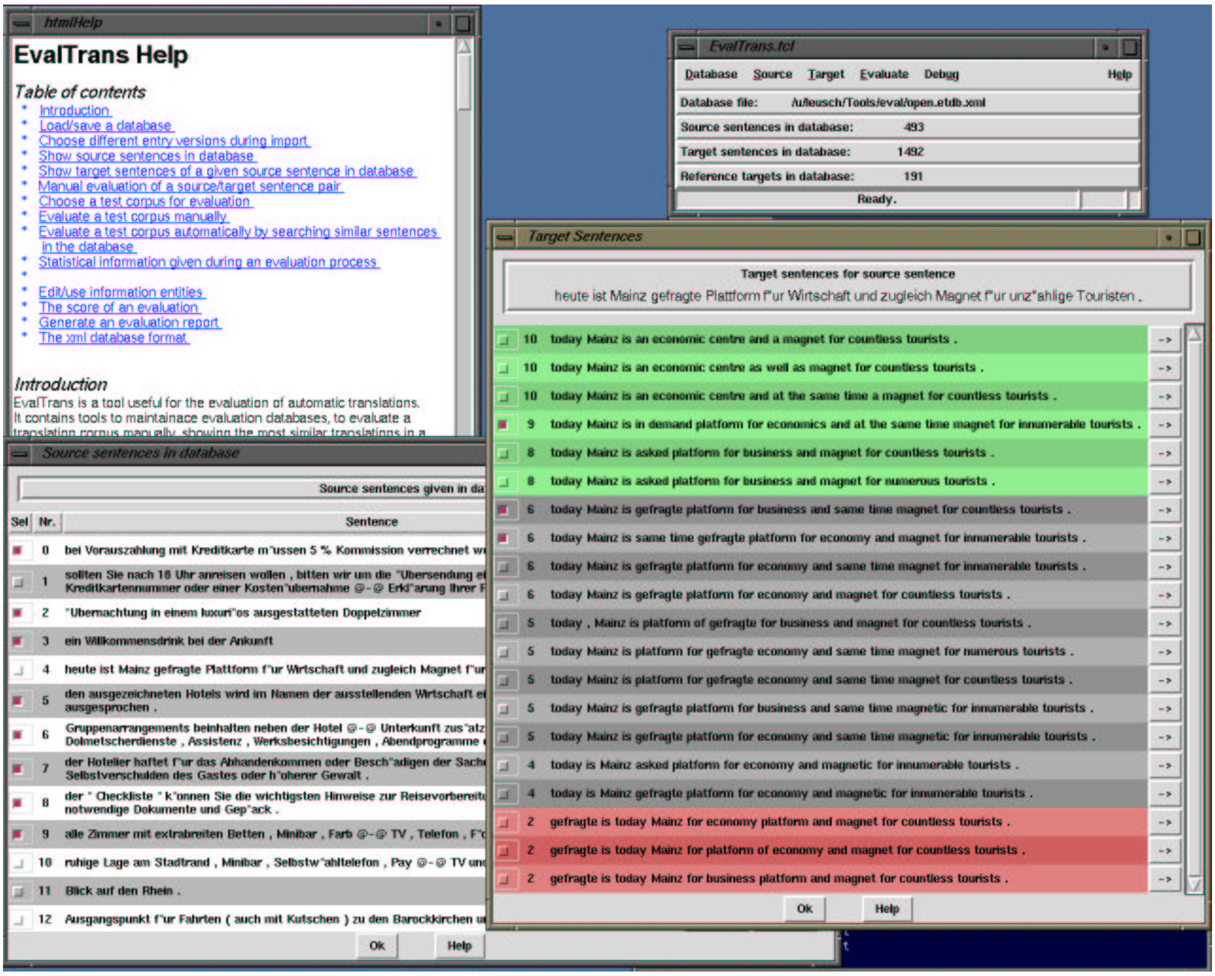
import from plain text. This feature is typically used to initialize an evaluation database with a set of source sentences and reference translations.

export into plain text and into the formats required by the BLEU evaluation script or the corresponding NIST variant hereof [Papineni & Roukos$^+$ 01]. Export causes loss of information, as other formats do not allow for information items or comments.

selection of source and target language sentences (via mouse click or automatically by comparison with sentences in a file).

deletion of source and target sentences.

sorting according to date of database entry, score or alphabetically.



htmlHelp

EvalTrans Help

Table of contents

- Introduction
- Load/save a database
- Choose different entry versions during import
- Show source sentences in database
- Show target sentences of a given source sentence in database
- Manual evaluation of a source/target sentence pair
- Choose a test corpus for evaluation
- Evaluate a test corpus manually
- Evaluate a test corpus automatically by searching similar sentences in the database
- Statistical information given during an evaluation process
-
- Edit/use information entities
- The score of an evaluation
- Generate an evaluation report
- The xml database format

Introduction

EvalTrans is a tool useful for the evaluation of automatic translations. It contains tools to maintainace evaluation databases, to evaluate a translation corpus manually, showing the most similar translations in a

Source sentences in database

Source sentences given in da

Sel	Nr.	Sentence
■	0	bei Vorauszahlung mit Kreditkarte m"ussen 5 % Kommission verrechnet w
□	1	sollten Sie nach 18 Uhr anreisen wollen , bitten wir um die "Ubersendung ei Kreditkartennummer oder einer Kosten"ubernahme @-@ Erkl"arung Ihrer F
■	2	"Ubernachtung in einem luxuri"os ausgestatteten Doppelzimmer
■	3	ein Willkommensdrink bei der Ankunft
□	4	heute ist Mainz gefragte Plattform f"ur Wirtschaft und zugleich Magnet f"ur
■	5	den ausgezeichneten Hotels wird im Namen der ausstellenden Wirtschaft ei ausgesprochen .
■	6	Gruppenarrangements beinhalten neben der Hotel @-@ Unterkunft zus"atz Dolmetscherdienste , Assistenz , Werksbesichtigungen , Abendprogramme
■	7	der Hotelier haftet f"ur das Abhandenkommen oder Besch"adigen der Sach Selbstverschulden des Gastes oder h"oherer Gewalt .
■	8	der " Checkliste " k"onnen Sie die wichtigsten Hinweise zur Reisevorbereitu notwendige Dokumente und Gep"ack .
■	9	alle Zimmer mit extrabreiten Betten , Minibar , Farb @-@ TV , Telefon , F"c
□	10	ruhige Lage am Stadtrand , Minibar , Selbstw"ahltelefon , Pay @-@ TV und
□	11	Blick auf den Rhein .
□	12	Ausgangspunkt f"ur Fahrten (auch mit Kutschen) zu den Barockkirchen u

Ok Help

EvalTrans.tcl

Database Source Target Evaluate Debug Help

Database file:	/u/leusch/Tools/eval/open.etdb.xml
Source sentences in database:	493
Target sentences in database:	1492
Reference targets in database:	191

Ready.

Target Sentences

Target sentences for source sentence

heute ist Mainz gefragte Plattform f"ur Wirtschaft und zugleich Magnet f"ur unz"ahlige Touristen .

□	10	today Mainz is an economic centre and a magnet for countless tourists .	->
□	10	today Mainz is an economic centre as well as magnet for countless tourists .	->
□	10	today Mainz is an economic centre and at the same time a magnet for countless tourists .	->
■	9	today Mainz is in demand platform for economics and at the same time magnet for innumerable tourists .	->
□	8	today Mainz is asked platform for business and magnet for countless tourists .	->
□	8	today Mainz is asked platform for business and magnet for numerous tourists .	->
■	6	today Mainz is gefragte platform for business and same time magnet for countless tourists .	->
■	6	today Mainz is same time gefragte platform for economy and magnet for innumerable tourists .	->
□	6	today Mainz is gefragte platform for economy and same time magnet for innumerable tourists .	->
□	6	today Mainz is gefragte platform for economy and magnet for countless tourists .	->
□	5	today , Mainz is platform of gefragte for business and magnet for countless tourists .	->
□	5	today Mainz is platform for gefragte economy and same time magnet for numerous tourists .	->
□	5	today Mainz is platform for gefragte economy and same time magnet for countless tourists .	->
□	5	today Mainz is gefragte platform for business and same time magnetic for innumerable tourists .	->
□	5	today Mainz is gefragte platform for economy and same time magnetic for innumerable tourists .	->
□	4	today is Mainz asked platform for economy and magnetic for innumerable tourists .	->
□	4	today is Mainz gefragte platform for economy and magnetic for innumerable tourists .	->
□	2	gefragte is today Mainz for economy platform and magnet for countless tourists .	->
□	2	gefragte is today Mainz for platform of economy and magnet for countless tourists .	->
□	2	gefragte is today Mainz for business platform and magnet for countless tourists .	->

Ok Help

Figure 6.4: Overview of the GUI layout.

search with regular expressions.

merging of databases. The user is warned when conflicts occur, for instance when the same source sentences has different information item definitions or the same target language sentence has different scores.

A small editor supports convenient definition of information items. The interface also contains a help system based on hypertext. The most important purpose of the GUI is on the one hand to display statistics about the status of the database and about a distinct set of candidate translations and on the other hand to facilitate the manual evaluation of new translations.

6.3.1 Displaying database statistics

The following major kinds of statistics can be displayed:

1. For a selected source sentence s, compute the average number of information items translated correctly by sentences in $\mathcal{T}(s)$, which conveys the "difficulty" of s. An

example is shown in Figure 6.5. A very high error rate for a certain information item often indicates word forms not seen in training.

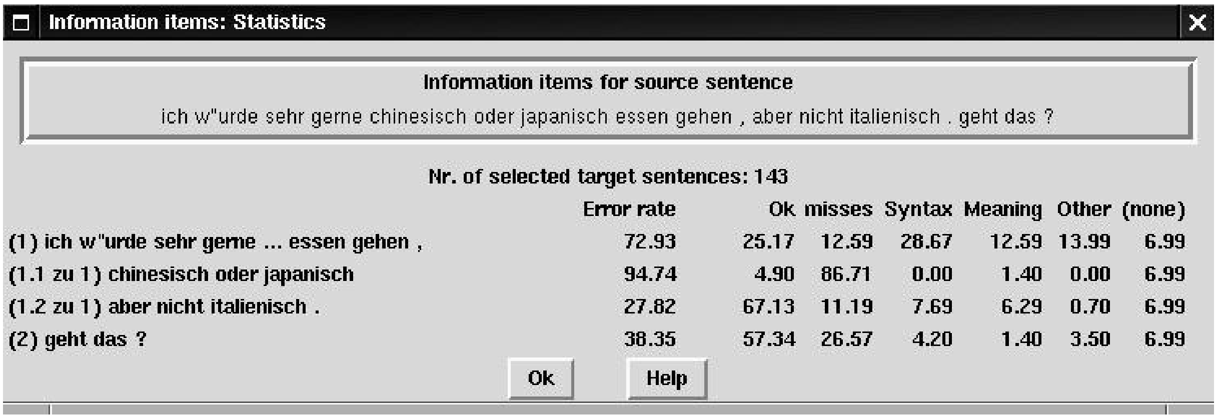
2. For any subset of all scored and stored target sentences, display the average (absolute) extrapolation error, an indicator for the reliability of quality estimates (see Section 6.4.2.1 and Figure 6.6).

3. Test the consistency of the automatically calculated measures (m-WER with and without inversions and the BLEU score) with the manually assigned scores and generate corresponding plot files (see Figure 6.7).

4. For a given set of pairs $(s_1, t_1) \ldots (s_n, t_n)$, the following operations are possible: Print the eSSER, the average normalized edit distance $\bar{d}(t_1^n)$, the IER and the ISER, the m-WER with and without inversions and the BLEU score. For all pairs (s_i, t_i), print the extrapolated score $\tilde{v}(s_i, t_i)$, the minimal edit distance $d(t_i, \mathcal{T}(s_i))$, the multi-reference edit distance, the BLEU score and, if (s_i, t_i) is already evaluated, the number of information items translated correctly. Report files containing all kinds of statistics can be generated. See Figure 6.8.

6.3.2 Manual evaluation of new translations

In the first place, the evaluation tool [`EvalTrans`] is designed for facilitating the work of manually judging evaluation quality. Figure 6.9 shows the corresponding interface. Those candidate translations in $\mathcal{DB}$ that are most similar to the sentence currently under evaluation are highlighted (so far, similarity is measured according to the standard definition of the edit distance). When moving the cursor over one of the candidates, all insertions, substitutions and deletions are marked in different colors and the classification of the information items is indicated. This results in speeding up the evaluation process and in improving evaluation consistency, as judgments can be made in comparison to other translations. The information items can be classified quickly by clicking on radio buttons for "OK" or one of the error classes.



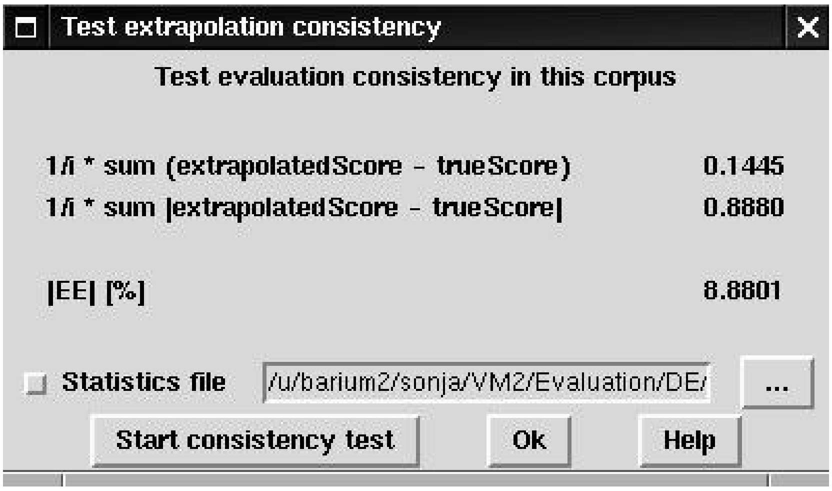
Information items for source sentence

ich w"urde sehr gerne chinesisch oder japanisch essen gehen , aber nicht italienisch . geht das ?

Nr. of selected target sentences: 143

	Error rate	Ok	misses	Syntax	Meaning	Other	(none)
(1) ich w"urde sehr gerne ... essen gehen ,	72.93	25.17	12.59	28.67	12.59	13.99	6.99
(1.1 zu 1) chinesisch oder japanisch	94.74	4.90	86.71	0.00	1.40	0.00	6.99
(1.2 zu 1) aber nicht italienisch .	27.82	67.13	11.19	7.69	6.29	0.70	6.99
(2) geht das ?	38.35	57.34	26.57	4.20	1.40	3.50	6.99

Figure 6.5: Statistics on information item error rate.



Test extrapolation consistency

Test evaluation consistency in this corpus

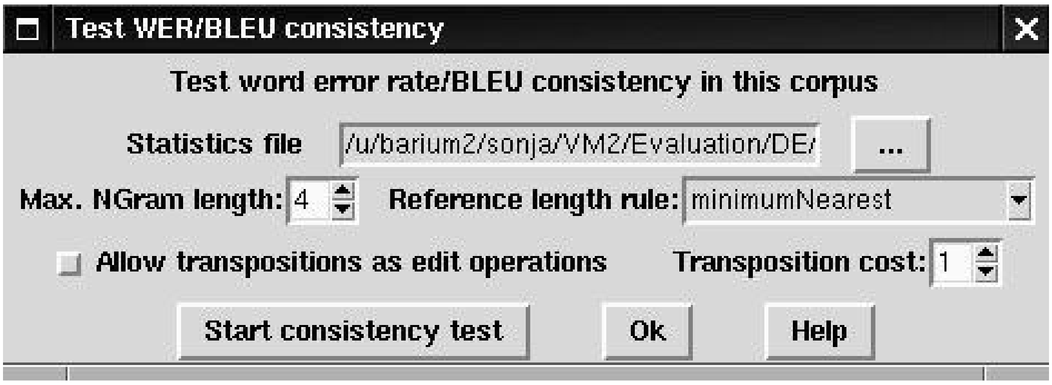
1/i * sum (extrapolatedScore - trueScore) 0.1445

1/i * sum |extrapolatedScore - trueScore| 0.8880

|EE| [%] 8.8801

□ Statistics file /u/barium2/sonja/VM2/Evaluation/DE/

Figure 6.6: The average (absolute) extrapolation error.



Test WER/BLEU consistency

Test word error rate/BLEU consistency in this corpus

Statistics file /u/barium2/sonja/VM2/Evaluation/DE/

Max. NGram length: 4

Reference length rule: minimumNearest

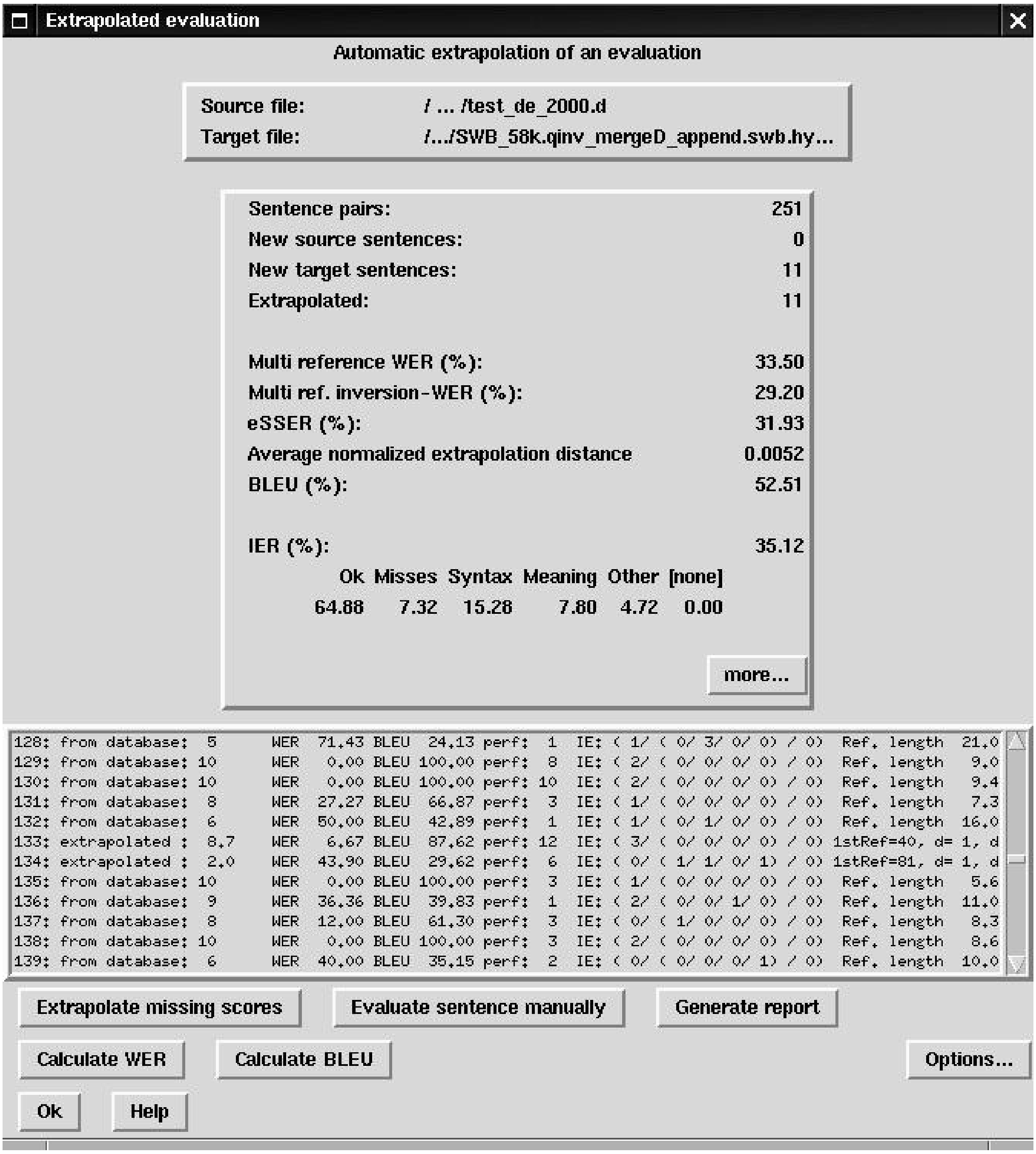
□ Allow transpositions as edit operations

Transposition cost: 1

Figure 6.7: Test the consistency of objective measures with manually assigned scores.

6.4 Assessment of the tool

Our research group constantly performs experiments to control the progress of the development of our translation systems. The evaluation tool has yet been used for the evaluation of results on various test sets for different tasks. Table 6.3 summarizes the statistics of the evaluation databases which have been used most often. The last database contains translations trained on the Zeres corpus with texts in the tourist domain. The corpus statistics are listed in Section A.2. The other lines correspond to the various test sets for Verbmobil with spontaneously spoken dialogs in the appointment scheduling domain (see Section A.1). The higher complexity of the Zeres task (increased vocabulary size, smaller amount of training data and less constrained domain) results in higher SSER. In Table 6.3, the column symbol "n" means "number of different source sentences". Note that n is often smaller than the number of sentences in the test corpus, because of duplicates. "T/n" stands for "average number of target sentences per source sentence" and R/n means "number of reference translations" (score K) per source sentence. The range of results for the Verbmobil German to English test sets Eval-2000 and Eval-147 is remarkably wide, the former because experiments using only a conventional dictionary


Extrapolated evaluation
Automatic extrapolation of an evaluation
more...
Extrapolate missing scores
Evaluate sentence manually
Generate report
Calculate WER
Calculate BLEU
Options...
Ok
Help

Source file: / ... /test_de_2000.d
Target file: /.../SWB_58k.qinv_mergeD_append.swb.hy...

Sentence pairs:	251
New source sentences:	0
New target sentences:	11
Extrapolated:	11
Multi reference WER (%):	33.50
Multi ref. inversion-WER (%):	29.20
eSSER (%):	31.93
Average normalized extrapolation distance	0.0052
BLEU (%):	52.51
IER (%):	35.12

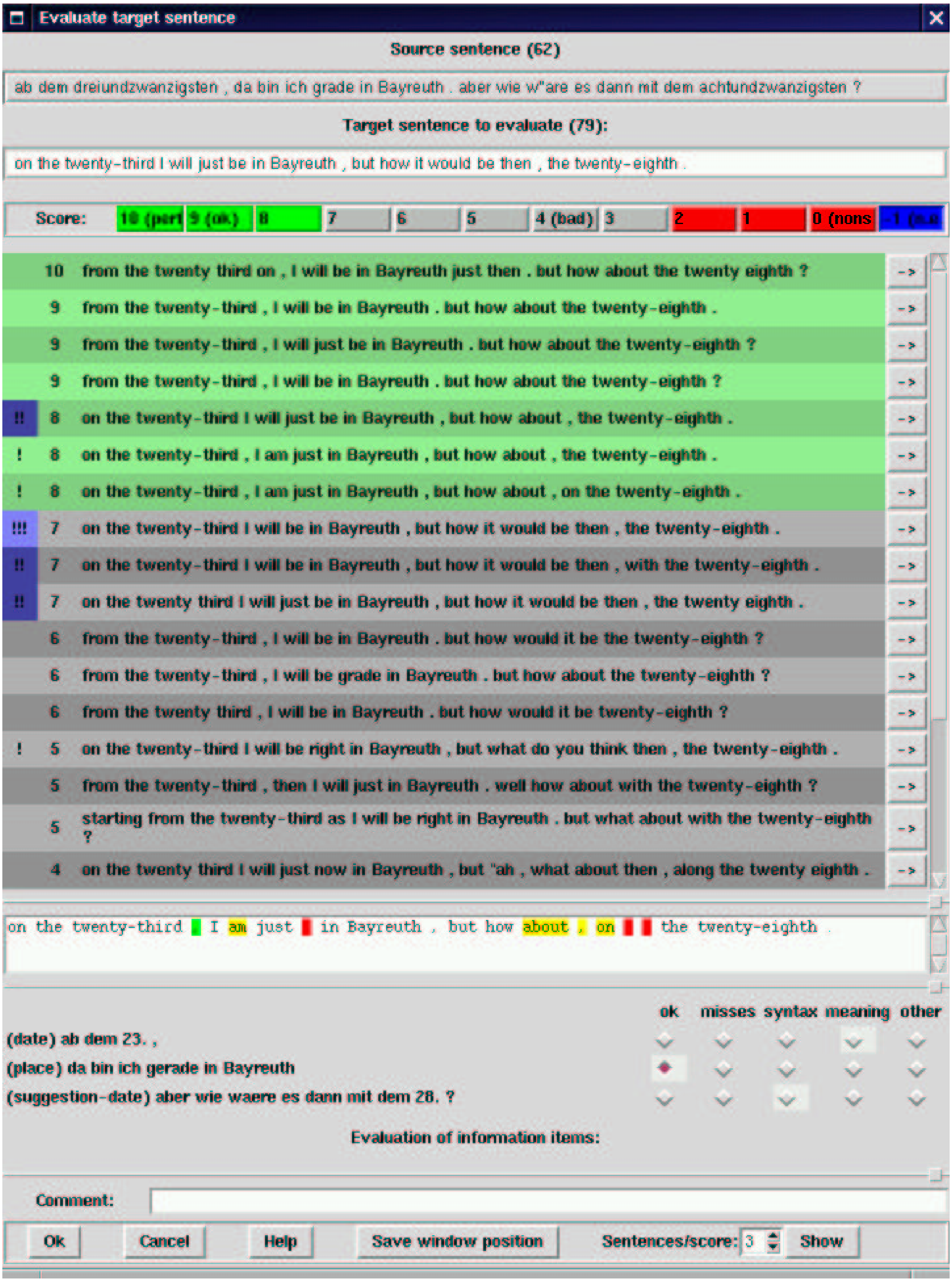
Ok	Misses	Syntax	Meaning	Other	[none]
64.88	7.32	15.28	7.80	4.72	0.00

```
128: from database:  5       WER  71.43 BLEU  24.13 perf:  1  IE: ( 1/ ( 0/ 3/ 0/ 0) / 0)  Ref. length  21.0
129: from database: 10       WER   0.00 BLEU 100.00 perf:  8  IE: ( 2/ ( 0/ 0/ 0/ 0) / 0)  Ref. length   9.0
130: from database: 10       WER   0.00 BLEU 100.00 perf: 10  IE: ( 2/ ( 0/ 0/ 0/ 0) / 0)  Ref. length   9.4
131: from database:  8       WER  27.27 BLEU  66.87 perf:  3  IE: ( 1/ ( 0/ 0/ 0/ 0) / 0)  Ref. length   7.3
132: from database:  6       WER  50.00 BLEU  42.89 perf:  1  IE: ( 1/ ( 0/ 1/ 0/ 0) / 0)  Ref. length  16.0
133: extrapolated :  8.7     WER   6.67 BLEU  87.62 perf: 12  IE: ( 3/ ( 0/ 0/ 0/ 0) / 0) 1stRef=40, d= 1, d
134: extrapolated :  2.0     WER  43.90 BLEU  29.62 perf:  6  IE: ( 0/ ( 1/ 1/ 0/ 1) / 0) 1stRef=81, d= 1, d
135: from database: 10       WER   0.00 BLEU 100.00 perf:  3  IE: ( 1/ ( 0/ 0/ 0/ 0) / 0)  Ref. length   5.6
136: from database:  9       WER  36.36 BLEU  39.83 perf:  1  IE: ( 2/ ( 0/ 0/ 1/ 0) / 0)  Ref. length  11.0
137: from database:  8       WER  12.00 BLEU  61.30 perf:  3  IE: ( 0/ ( 1/ 0/ 0/ 0) / 0)  Ref. length   8.3
138: from database: 10       WER   0.00 BLEU 100.00 perf:  3  IE: ( 2/ ( 0/ 0/ 0/ 0) / 0)  Ref. length   8.6
139: from database:  6       WER  40.00 BLEU  35.15 perf:  2  IE: ( 0/ ( 0/ 0/ 0/ 1) / 0)  Ref. length  10.0
```

Figure 6.8: Statistics for a sample set of candidate translations.

(see Section 7.3.2) have been performed on this set, and the latter because translations for speech input are contained in the corresponding database.

The Zeres tests are more difficult than the tests for Verbmobil. For this reason, and because less experiments have yet been run and thus less hypotheses have been evaluated for Zeres, the number of reference translations is small compared to an average of seven references for the Verbmobil Eval-147 test sentences. Figure 6.10 shows the development of the rate T/n of target sentences per source sentence and of the rate R/n of reference



Source sentence (62)

ab dem dreiundzwanzigsten , da bin ich grade in Bayreuth . aber wie w"are es dann mit dem achtundzwanzigsten ?

Target sentence to evaluate (79):

on the twenty-third I will just be in Bayreuth , but how it would be then , the twenty-eighth .

Score: 10 (perf | 9 (ok) | 8 | 7 | 6 | 5 | 4 (bad) | 3 | 2 | 1 | 0 (nons | -1 (n.e

	10	from the twenty third on , I will be in Bayreuth just then . but how about the twenty eighth ?	->
	9	from the twenty-third , I will be in Bayreuth . but how about the twenty-eighth .	->
	9	from the twenty-third , I will just be in Bayreuth . but how about the twenty-eighth ?	->
	9	from the twenty-third , I will be in Bayreuth . but how about the twenty-eighth ?	->
!!	8	on the twenty-third I will just be in Bayreuth , but how about , the twenty-eighth .	->
!	8	on the twenty-third , I am just in Bayreuth , but how about , the twenty-eighth .	->
!	8	on the twenty-third , I am just in Bayreuth , but how about , on the twenty-eighth .	->
!!!	7	on the twenty-third I will be in Bayreuth , but how it would be then , the twenty-eighth .	->
!!	7	on the twenty-third I will be in Bayreuth , but how it would be then , with the twenty-eighth .	->
!!	7	on the twenty third I will just be in Bayreuth , but how it would be then , the twenty eighth .	->
	6	from the twenty-third , I will be in Bayreuth . but how would it be the twenty-eighth ?	->
	6	from the twenty-third , I will be grade in Bayreuth . but how about the twenty-eighth ?	->
	6	from the twenty third , I will be in Bayreuth . but how would it be twenty-eighth ?	->
!	5	on the twenty-third I will be right in Bayreuth , but what do you think then , the twenty-eighth .	->
	5	from the twenty-third , then I will just in Bayreuth . well how about with the twenty-eighth ?	->
	5	starting from the twenty-third as I will be right in Bayreuth . but what about with the twenty-eighth ?	->
	4	on the twenty third I will just now in Bayreuth , but "ah , what about then , along the twenty eighth .	->

```
on the twenty-third 🟩 I am just 🟥 in Bayreuth , but how about , on 🟥 🟥 the twenty-eighth .
```

	ok	misses	syntax	meaning	other
(date) ab dem 23. ,	◇	◇	◇	◇	◇
(place) da bin ich gerade in Bayreuth	◆	◇	◇	◇	◇
(suggestion-date) aber wie waere es dann mit dem 28. ?	◇	◇	◇	◇	◇

Evaluation of information items:

Comment:

Sentences/score: 3

Figure 6.9: Manual evaluation of a new translation candidate.

Table 6.3: Statistics of evaluation databases.

database	range of SSER[%]	n	T/n	R/n
Verbmobil Eval-2000 English to German				
Test (first 100)	23–29	99	15.7	2.3
Verbmobil Eval-2000 German to English				
Test	26–60	247	64.4	2.8
Develop	26–60	270	21.6	2.1
Verbmobil Eval-147 German to English				
Speech and text input	17–41	144	59.1	6.9
Zeres: Test-Open (first 100)	48–59	100	19.9	1.4

sentences per source sentence on the Verbmobil Eval-147 database. On the x-axis, the respective database version of the Verbmobil Eval-147 evaluation database is shown (old versions can easily be retrieved, as the databases are under revision control).

6.4.1 Efficiency of manual evaluation

The human evaluators performing the manual evaluation of the experimental results are students from the university's languages department. Upon the installation of the tool

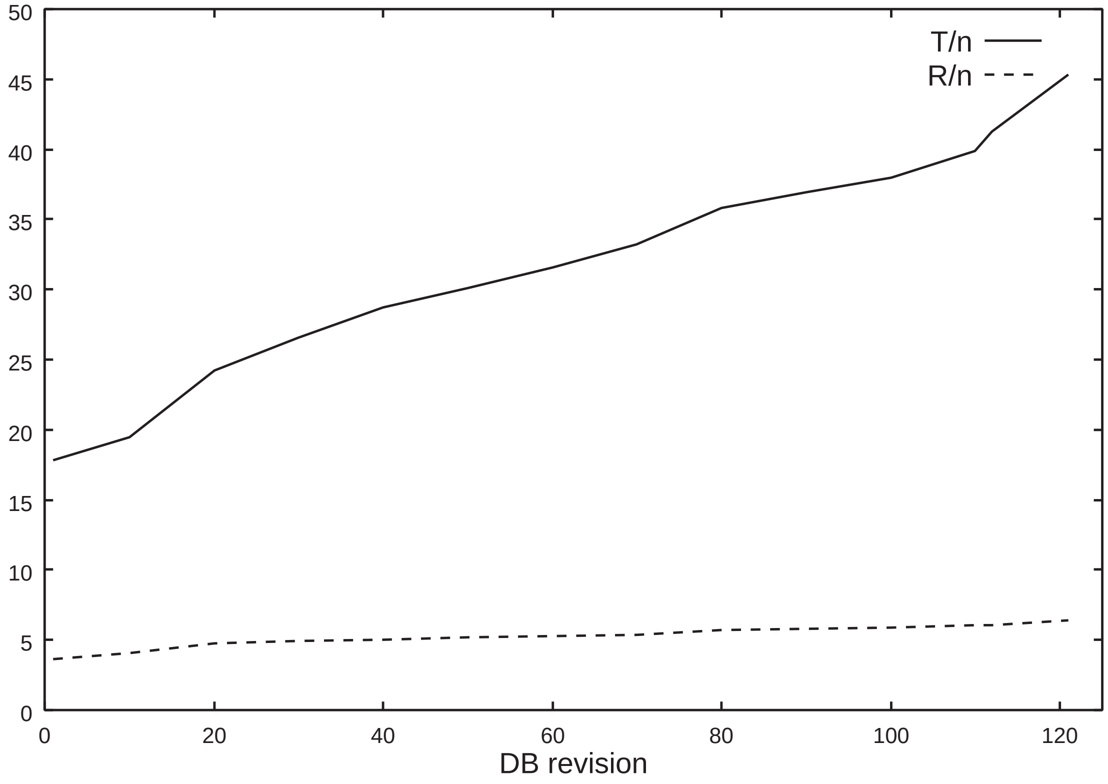


Figure 6.10: T/n and R/n versus revision number of $\mathcal{DB}$.

they reported a substantial facilitation for their work due to the graphical user interface. They also mentioned that judging the information items, though necessitating additional evaluation effort, helped getting a sense of the quality of the translation under consideration. Highlighting of the most similar translation candidates and also marking the respective difference in terms of substitutions, insertions and deletions in different colors (see section 6.3.2 and Figure 6.9) helps speeding up the evaluation process substantially. The evaluation of a new translation candidate takes approximately 30 to 60 seconds, depending on the length of the sentence, provided that the evaluators are already familiar with the *source* sentence from previous repeated evaluations on the same test set.

6.4.2 Reliability of quality extrapolation

The accuracy of the extrapolation of the SSER depends on many factors, like complexity of the translation task, variability of the evaluated translations, degree to which the database is filled, i.e. number of translations per source sentence, etc. The average normalized edit distance $\bar{d}(t_1^n)$ as defined in Equation (6.8) is a measure for the reliability of the eSSER for a set of new translations, whereas the method described in subsection 6.4.2.1 allows for the computation of the expected extrapolation error on translations yet to be produced.

6.4.2.1 Leaving-one-out validation

As a measure of the reliability of the extrapolation of scores for new translation candidates serves the *average absolute extrapolation error* $|EE|(\mathcal{DB})$:

$$|EE|(\mathcal{DB}) = \frac{1}{K \cdot |\mathcal{DB}|} \sum_{\substack{(s,t) \\ \in \mathcal{DB}}} |v(s,t) - \hat{v}(s,t,\mathcal{T}(s) \setminus \{t\})| \ ,$$

where $|\mathcal{DB}|$ is the number of pairs (s,t) contained in $\mathcal{DB}$ (normalization constant). This quantity conveys the following: For each target sentence t for a source sentence s, try to extrapolate the corresponding score from the *other* translation candidates (*leaving-one-out* scheme). The resulting estimate is compared to the real score of t. $|EE|(\mathcal{DB})$ gives the overall extrapolation error per sentence, i.e. a measurement of the reliability of the estimates for a distinct sentence. Note that the extrapolation process sometimes overestimates the quality of a translation, and sometimes the estimation is lower than the real score. It is for this reason that the eSSER on a set of n translation is more reliable than each extrapolated score of a distinct sentence t. In Table 6.4 the results of the leaving one out validation on the databases listed in Table 6.3 are summarized. [Vogel & Nießen$^+$ 00] report a substantial reduction of the $|EE|$ and a higher number of correctly extrapolated quality scores using weighted edit distance as discussed in Section 6.1.4. Figure 6.11 shows the development of the average absolute extrapolation error as the database is gradually filled. Again, the x-axis represents increasing revision numbers.

Table 6.4: Leaving-one-out validation on different databases.

database	$\|EE\|[\%]$
Verbmobil Eval-2000 English to German	
Test (first 100)	11.5
Verbmobil Eval-2000 German to English	
Test	9.9
Develop	8.9
Verbmobil Eval-147 German to English	10.4
Zeres Test-Open (first 100)	12.0

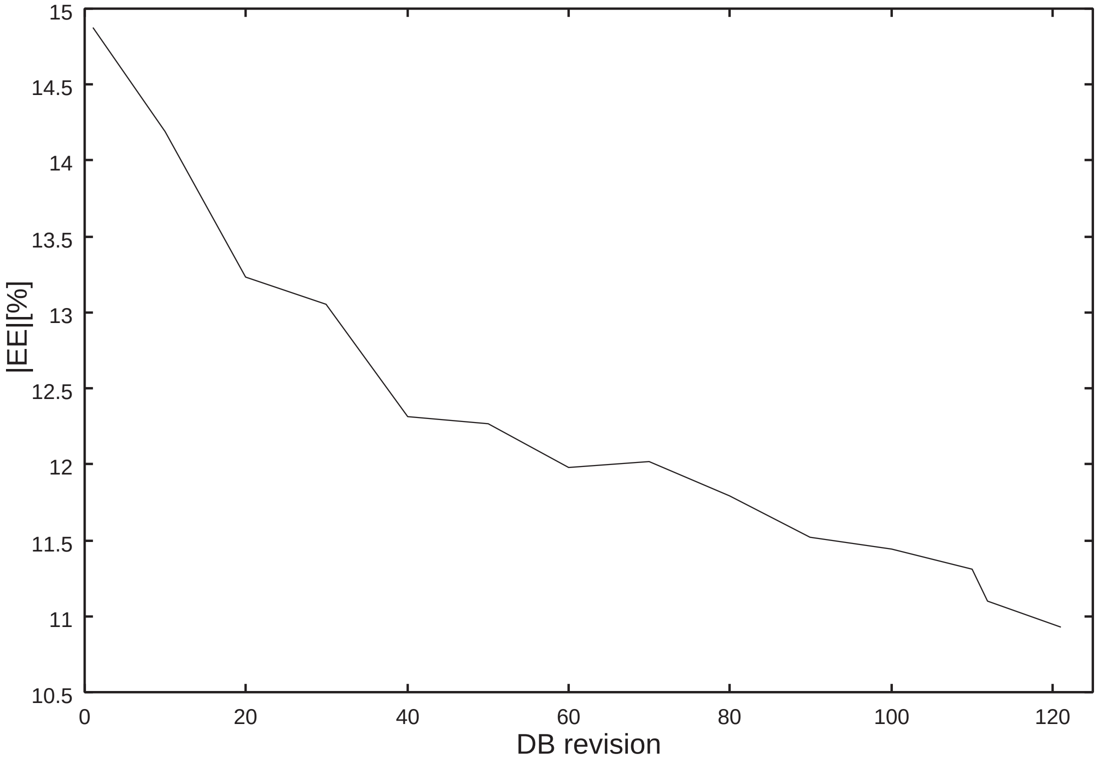


Figure 6.11: $|EE|(\mathcal{DB})$ versus revision number of $\mathcal{DB}$.

6.4.2.2 Example hypotheses files

For 26 sets of translations, the eSSER and the corresponding $\bar{d}(t_1^n)$ just before evaluation was stored and compared to the real SSER afterward (i.e. the extrapolation error |SSER - eSSER| was computed). The resulting diagram is shown in Figure 6.12. On the 26 files, the error |SSER - eSSER| was only 1.2% on average. Also for the estimation of the quality of entire sets of candidate translations, [Vogel & Nießen$^+$ 00] report a significantly smaller average error. On average 29.5% of the translation quality scores had to be estimated, i.e. were not yet present in the database. This means that the tool saved at least 70% of the evaluation effort for the evaluation of these 26 translation hypotheses files.

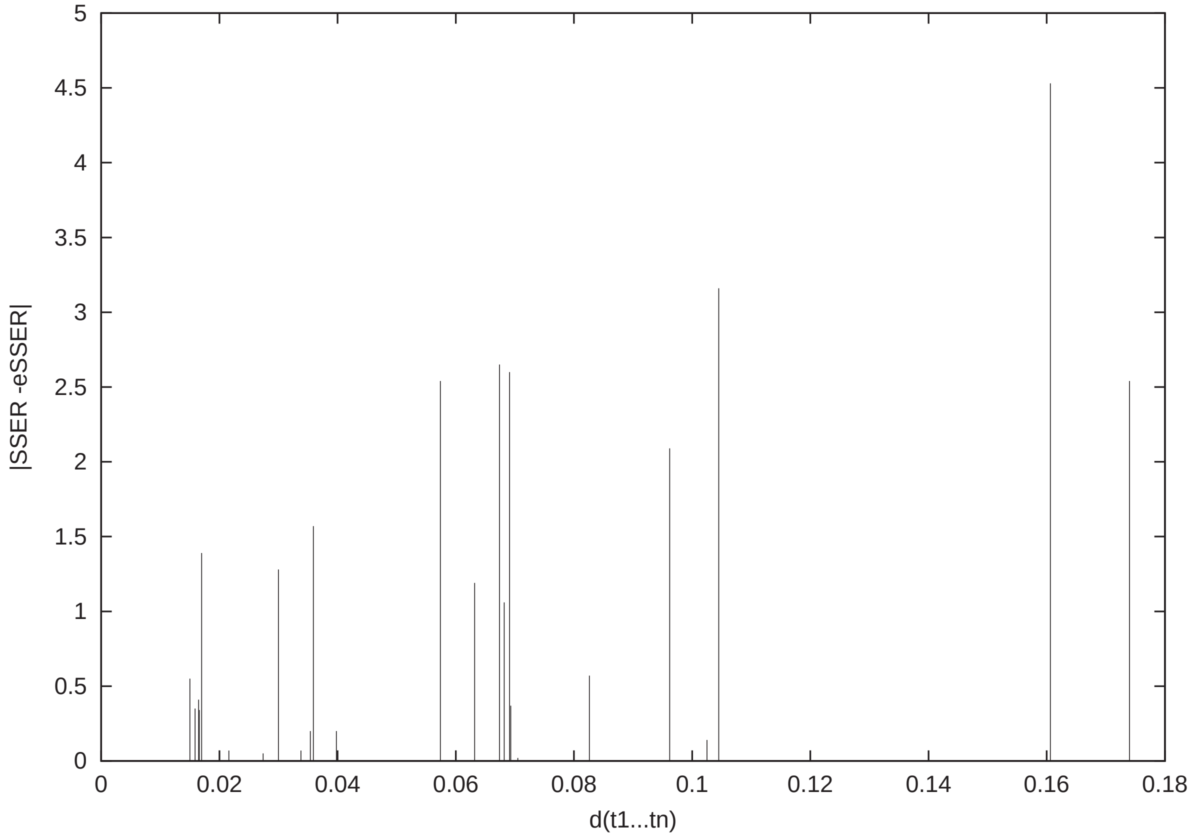


Figure 6.12: $\bar{d}(t_1^n)$ versus |SSER - eSSER|.

6.4.3 Consistency of results

The following experiment would convey information about the sensibility of the evaluation results against the so called "human factor", i.e. the question "how much would the SSER of a certain set of new candidates differ depending on which evaluator performs the evaluation and on his or her current mental constitution?": Randomly extract sentences with their scores from the database and make evaluators do the evaluation again. The resulting new score can be compared to the score formerly stored in the database. This experiment has not been performed so far.

Chapter 7

Experimental Results

> One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'
>
> (Warren Weaver, 1949)

Tests were carried out for the translation direction from German into English on Verbmobil data, on Nespole! data and on Zeres data and for the direction from English to German on Verbmobil data. As usual, the sentences from the test sets were not used for training. The corpus statistics are summarized in Appendix A. The training corpora were used for training the parameters of Model 4 as first proposed by [Brown & Della Pietra$^+$ 93b].

The performance measures are described in Appendix B. For most of the experiments, the 'objective' measures (m-WER,m-invWER and BLEU) and two 'subjective' measures (SSER and ISER) are specified. The reason for omitting the m-invWER, e.g. for the results for speech input, is the presence of very long sentences, which makes the calculation of the m-invWER computationally too costly. In almost all cases, the m-WER and the m-invWER yield the same rankings, anyway. When the results are discussed, the focus is on differences in terms of the SSER and the ISER, because the manually assigned quality scores are considered most liable and meaningful. All figures presented in this chapter for the subjective sentence error rate (SSER) are not extrapolated: the translations have been fully evaluated according to the manually assigned quality scores.

7.1 Description of the baseline setup

The translation system for most of the experiments was the alignment templates translation system described in [Och 02], which already has a reasonably good capability of performing word reordering on its own. The key elements of this system are the so-called *alignment templates*, which are pairs of source and target language phrases together with an alignment between the words within the phrases. The advantage of the alignment template approach compared to single word-based statistical translation models is that word context and local changes in word order are explicitly accounted for. The alignment template model refines the translation probability $Pr(f_1^J|e_1^I)$ by introducing two hidden variables z_1^K and a_1^K for the K alignment templates and the alignment of the alignment templates:

$$Pr(f_1^J|e_1^I) \quad = \quad \sum_{z_1^K, a_1^K} Pr(a_1^K|e_1^I) \cdot Pr(z_1^K|a_1^K, e_1^I) \cdot Pr(f_1^J|z_1^K, a_1^K, e_1^I)$$

Training the parameters for this system entails training of Model 4 parameters [Brown & Della Pietra$^+$ 93b] in both translation directions and combining the resulting alignments into one symmetrized alignment. From this symmetrized alignment, the lexicon probabilities as well as the alignment templates are extracted. For further details, the reader is referred to [Och & Tillmann$^+$ 99].

Compared to the results using the alignment templates system reported in [Och 02], there are some differences in the setups used there and in this work. These differences are categorized as follows:

Corpus format: The corpora for training originally were encoded in an ASCII format similar to TEX–style, where the German umlaut 'ä' for example is represented as '`"a`'. To enable parsing of the corpora by the analyzers described in Section 3.2, they were transformed into Iso format.

Hard-coded mapping: In [Och 02], some hard-coded replacement tables were used to pre- and post-process the data. These mapping operations were designed to split up some frequently occurring compound words, to merge some frequent multi-word phrases, to handle different notations of time expressions and to mark proper names. As most of these tasks are dealt with in a more systematic and automatic way by the methods described in this work, it is straightforward to abandon these replacement tables.

Weights for the dictionary: As regards the conventional dictionary, [Och 02] distinguishes between entries, which occur as aligned pair of phrases or words in the training corpus, and those which do not. The former are weighted with a factor of 10 as compared to pairings in the corpus, whereas the latter are weighted with a factor of 0.1 — that is, dictionary entries validated positively on the training corpus are considered 100 times more reliable than the others. As this distinction only makes sense when the corpus is sufficiently large and when the corpus size is kept fixed, and as this work contains an investigation on the impact of reducing the size of the training corpus, this distinction is abandoned.

Edited dictionary version: In this work, an edited version of the conventional dictionary is used, in which some errors have been corrected and some entries have been deleted and some inserted. The changes had different objectives:

1. to correct typographical errors, for example "cheeper" instead of "cheaper";
2. to eliminate entries which are valid in some contexts, but never in the considered domains, e.g. the translation "flight" for "Zug", which may originate from the interpretation "Zug ≡ Zug der Zugvögel ≡ flight of migrating birds";

3. to remove some entries which cannot be expected to occur in a typical conventional dictionary, like for example the translation "Backe" as translation for "Backe", which in the Verbmobil corpus occurs as proper name of a person, but which in general would be translated into "cheek" (in fact, the entry "cheek/Backe" is inserted into the edited dictionary);
4. to correct entries not representing exact translations, for example "discussion" instead of "discussions" as translation for "Gespräche";
5. to add translations for readings not yet accounted for, like for instance the entry "convenient" as translation for "bequemer".

The effects of these changes on the translation results are listed in Table 7.1. The baseline alignment templates setup used in this work corresponds to the first line in this table and the last line stands for the setup used in [Och 02]. The combination of restructuring transformations yielding the best results was also applied within the setup in [Och 02], except for the corpus format, which hardly affects the translation results. Table 7.5 reports the corresponding figures.

For the sake of completeness, experiments for restructuring and hierarchical lexicon models were also carried out using the single word based approach [Tillmann 02]. The corresponding results are summarized in the Tables 7.4 and 7.13.

7.2 Results for preprocessing and postprocessing

7.2.1 Translation direction English to German

The results for translating from English to German are summarized in Table 7.2. The effect of treating question inversion is not clear for this translation direction: The slight reduction in the information item syntactic error rate (ISER) suggests that more of the intended information can be conveyed. From the increase in terms of subjective sentence error rate (SSER) on the other hand it can be concluded that this positive effect is compensated by a deteriorated syntax. This interpretation is enforced by the asymmetric behavior of the multiple reference word error rate (m-WER), which suggests an improve-

Table 7.1: Comparison of the baseline versus the results reported in [Och 02]. Task: Verbmobil. Testing on 251 sentences ("Test"). System: Alignment templates.

corpus format	preproc. via maps	dictionary version	dictionary weights	m-WER [%]	m-invWER [%]	BLEU [%]	SSER [%]	ISER [%]
Iso	**no**	**edited**	**no**	**33.8**	**28.4**	**53.8**	**31.4**	**16.4**
Iso	no	edited	yes	34.0	28.6	53.7	30.4	15.4
Iso	no	orig.	yes	34.3	29.0	53.5	30.8	16.3
Iso	yes	orig.	yes	34.6	28.7	53.4	30.1	13.2
Tex	**yes**	**orig.**	**yes**	**34.1**	**28.2**	**53.8**	**30.1**	**12.4**

Table 7.2: Effect of restructuring for the translation direction English to German. Task: Verbmobil. Testing on the first 100 sentences of "Test". System: Alignment templates.

	m-WER [%]	m-invWER [%]	BLEU [%]	SSER [%]	ISER [%]
baseline	30.2	25.1	59.2	26.3	14.6
treat question inversion	29.0	26.1	57.7	27.9	13.4
treat prefixes	27.9	23.4	59.1	23.7	12.7
+ morphological corrections	27.8	23.3	59.2	23.6	12.4
+ merge phrases in both languages	31.7	26.4	55.3	23.5	12.0

ment, and the BLEU score, which suggests a deterioration: As BLEU measures n-gram precision, its focus may be more on the syntax. An inspection of the translations shows that many of the translation errors are due to erroneous detection of the subject and the finite verb. In Section 4.1.1 it has already been argued that this is more difficult for German than for English because of the larger variability of the word order. A typical example is the incorrect translation "Acht ist Uhr in Ordnung?" for "Is eight o'clock OK?", where the cardinal number "acht" has been misclassified as the imperative second person singular form of the verb "achten". In contrast to question inversion treatment, treating separated verb prefixes improves the translation quality substantially. The further improvement by combining this method with merging multi-word phrases in both languages and with morphological corrections of the translations is rather small.

7.2.2 Translation direction German to English

Results on the Verbmobil task

In Table 7.3 the improvements achieved by various word restructuring techniques on the Verbmobil German to English translation task are listed. When translating German input sentences into English, both question inversion treatment and treatment of separated verb prefixes for themselves and even more the combination of both restructuring operations result in higher quality translations. The further improvement by merging multi-word phrases is comparatively small. Table 7.4 reports on the results achieved with the single word system in a variant with German to English reordering constraints.

Experiments for splitting compound words were not carried out on this task, because there was no reason to expect this operation to be beneficial on this task: As has been argued at the end of Section 4.1.3, compound splitting can be problematic for German proper names and cardinal and ordinal numbers, but these types of words are especially frequent in the Verbmobil corpus: More than half of the compound words belong to these categories. When identified proper names and numbers are excluded from compound splitting the effect of this restructuring operation on the training corpus is fairly small: only about 1% of the resulting tokens in the transformed corpus originate from compounds and the token–type ratio is only increased from 65.4 to 71.6. The number of singletons (words seen only once in training, an indicator for the degree to which the training sample

Table 7.3: Effect of restructuring for the translation direction German to English. Task: Verbmobil. Testing on 527 sentences ("`Test`" and "`Develop`"). System: Alignment templates.

	m-WER [%]	m-invWER [%]	BLEU [%]	SSER [%]	ISER [%]
baseline	34.1	27.6	53.7	30.2	14.1
treat prefixes	34.0	27.8	53.3	29.9	13.1
treat question inversion	32.5	26.4	56.4	27.6	13.6
+ merge German phrases	32.4	26.4	56.6	27.7	13.5
+ treat prefixes	32.6	26.7	56.2	27.1	13.1
+ treat inf. marker	32.4	26.6	56.3	26.9	13.3
+ merge English phrases	32.5	26.8	56.3	26.6	12.8

Table 7.4: Effect of restructuring for the translation direction German to English. Task: Verbmobil. Testing on 251 sentences ("`Test`"). "restructuring" entails: treatment of question inversion, separated verb prefixes and merging of phrases in both languages. System: Single word system.

	m-WER [%]	m-invWER [%]	BLEU [%]	SSER [%]	ISER [%]
baseline	35.2	30.7	50.1	33.5	19.2
restructuring	33.6	29.1	52.8	31.8	18.7

is representative) does not go down by more than 2.8% and the vocabulary size is only reduced by 7.6%, because the German compound words occurring in the corpus are very typical of the domain, like for instance "Doppelzimmer" (English: double room), "Hauptbahnhof" (main train station) or "Zugfahrplan" (train schedule), and they are contained in the corpus frequently.

The examples in Figure 7.1 illustrate the effect of the pre– and postprocessing for this translation direction. In the first example the joint effect of treating question inversion and separated verb prefixes turns a very bad translation into a perfect one: In the baseline the main verb part "fährt" and the detached prefix "ab" are translated individually: the first into "goes" and the second into "starting from", which are the highest ranking translations for the verb *parts* when they are treated as separated entries in the probabilistic lexicon. When restructuring has been active the reconstructed verb "abfährt" is correctly translated into "leaves" before treating question inversion replaces "leaves" by the structure "does . . . leave". From the second example it can be concluded that the translation pair "(kommen . . . an | arrive)" occurs sufficiently often in the training corpus to learn that "an" must not be translated individually (in the IBM alignment models there is the so-called empty word e_0 to account for these cases), but still the highest ranking translation for "kommen" is not "arrive" but "come". The third example shows that sometimes domain specific data can compensate for parts of the word order problems, as

input	`wann fährt der Zug genau ab?`
baseline	`when the train goes starting from right?`
restructuring	`when does the train leave exactly?`
input	`wir kommen um 12 Uhr mittags in Hannover an.`
baseline	`we will come at noon, in Hanover.`
restructuring	`we arrive at 12 o'clock in Hanover at noon.`
input	`dann schlage ich das Hotel Prinzenhof vor.`
baseline	`then I suggest the hotel Prinzenhof in front of.`
restructuring	`then I would suggest the hotel Prinzenhof.`
input	`dann müssen wir noch die Rückreise klären.`
baseline	`then we still have to clarify the return trip.`
restructuring	`then we still have to clear the return trip.`
input	`wir treffen uns am besten um 8 Uhr.`
baseline	`it would be best if we meet at 8 o'clock.`
restructuring	`we will meet on the best at 8 o'clock.`

Figure 7.1: Examples for the effect of restructuring on the translation quality.

the main verb part “schlage” is not translated by its generally most common translation “beat” but by “suggest”. Still, the detached prefix “vor” is translated individually by “in front of”.

As is often the case when the settings for statistical machine translation are changed, some of the Verbmobil test sentences are translated worse than before. The last two examples in Figure 7.1 give an idea of the newly introduced errors. They are typical in that the deterioration cannot be explained directly by the restructuring operations: the errors must be due to indirect effects on the alignments in training.

The restructuring methods yield a comparable improvement within the setup used in [Och 02]. The notation in Table 7.5 is as follows: “setup 1” is used in almost all experiments on Verbmobil reported in this work. It corresponds to the first line in Table 7.1. “setup 2” is used in [Och 02] and it corresponds to the last but one line in Table 7.1.

Results on speech: Some tests have also been carried out for speech translation. The test set characteristics are depicted in Table A.5. The translation performance results for translation of text and of the single-best hypothesis given by a speech recognizer (accuracy 69%) are given in Table 7.6. For both, text and speech input, the combination of treating separated prefixes and inserted infinitive markers and of merging phrases results in better translations. On speech input data it is difficult to achieve further improvements with question inversion treatment because the question marks are not directly available: the only clue for detecting interrogative sentences is prosodic markup.

The fact that such transformations are not only helpful on text input, but also on speech input is quite encouraging. As an example makes clear this cannot be taken for granted: The test sentence “Dann fahren wir dann los.” is recognized as “Dann fahren wir

Table 7.5: Results for restructuring with two setups. Setup 1 is the setup used as baseline in this work, and setup 2 corresponds to the setup used in [Och 02]. Task: Verbmobil. Testing on 251 sentences (“Test”). “restructuring” entails treatment of question inversion, separated verb prefixes and infinitive markers as well as merging of phrases in both languages. System: Alignment templates.

		m-WER [%]	m-invWER [%]	BLEU [%]	SSER [%]	ISER [%]
setup 1	baseline	33.8	28.4	53.8	31.4	16.4
	restructuring	32.7	27.8	55.8	26.6	13.8
setup 2	baseline	34.6	28.7	53.4	30.1	13.2
	restructuring	32.5	27.0	56.1	26.3	11.7

Table 7.6: Results for restructuring: Verbmobil Eval-147 text and speech input. System: Alignment templates.

		m-WER [%]	BLEU [%]	SSER [%]	ISER [%]
text	baseline	28.9	61.1	20.0	14.6
	treat prefixes + inf. marker + phrases	29.8	60.5	18.8	9.1
	+ treat question inversion	29.9	59.7	17.6	8.2
speech	baseline	50.7	39.1	40.8	38.9
	treat prefixes + inf. marker + phrases	50.9	37.9	39.9	33.0
	+ treat question inversion	49.9	39.5	40.5	30.4

dann uns.” and the fact that separable verbs do not occur in their separated form in the training data is unfavorable in this case. The figures suggest that the speech recognizer output still contains enough information for helpful preprocessing.

Results on the Zeres task

In contrast to the Verbmobil corpus the Zeres corpus lends itself to compound decomposition: The number of words seen only once in training can be reduced by 8.9%, the token–type ration can be increased from 8.6 to 12.3 and 6.8% of the tokens in the transformed corpus originate from compounds. Even more convincing is a reduction of the vocabulary size by 25%. Table 7.7 shows that the decomposition of compound words yields an improvement in the subjective sentence error rate of 6.6% and the treatment of unknown words improves the translation quality by an additional 1.1%. Treating separable verb prefixes in addition to splitting compounds and treating unknown words gives the best result with an improvement of 8.8% absolute compared to the baseline SSER and an improvement of 8.2% in terms of ISER. Question inversion treatment does not help on this task, it even deteriorates the results. Questions are far more important on a task like Verbmobil, where dialogs are translated.

Table 7.7: Results on Zeres. Testing on the first 100 sentences of "Test-Open". System: Alignment templates.

	m-WER [%]	BLEU [%]	SSER [%]	ISER [%]
baseline	58.2	26.0	55.0	42.8
+ map unknown word forms	57.6	27.2	51.5	41.8
split compounds	56.2	28.1	48.4	41.4
+ map unknown word forms	56.2	28.1	47.3	40.1
split compounds + treat prefixes	56.4	27.4	47.7	34.3
+ map unknown word forms	56.2	27.6	46.2	34.3
split compounds + treat prefixes + question inv.	58.1	24.4	53.1	41.4
+ map unknown word forms	58.1	24.5	52.4	42.1

7.3 Hierarchical lexicon models and translation with scarce resources

Experiments for hierarchical lexicon models were only carried out for the translation direction German to English and only on Verbmobil data. They have not been applied to the Zeres task because there the decomposition of compound words had a large positive impact on the translation quality. Until now, there is no possibility to use the hierarchical lexicon models in combination with compound word treatment, because it is not clear how the morpho-syntactic tags of a compound should be "distributed" over the components. Hierarchical lexicon models have not been applied to English to German translation yet, because for translating into a morphologically richer language a strong and more syntax oriented language model than pure n-gram models should be available in order to choose between different inflected word forms in the target language.

7.3.1 Conventional dictionaries: Disambiguation without context

Table 7.8 lists the results achieved with and without an additional conventional dictionary. The dictionary yields a relative improvement of about 4% in terms of subjective sentence error rate. In the Verbmobil training corpus, which is used for detecting the tag translation probabilities as described in Section 5.3, there are 261 different German tag sequences and 110 different English tag sequences. Only 1 199 of the 28 710 possible pairings of German and English tag sequences actually occur in the alignment. Table 7.9 lists the highest ranking pairings. Note that the tag sets for German and English are slightly different.

Sparse bilingual training data makes additional conventional dictionaries especially important. Table 7.10 reveals that disambiguating them as such helps improving the translation quality a little bit, but enriching them by aligning corresponding readings makes more sense when they are used together with a hierarchical lexicon which can access the information necessary to distinguish readings via morpho-syntactic tags.

Table 7.8: Impact of the conventional lexicon. Task: Verbmobil. Testing on 251 sentences ("Test"). System: Alignment templates.

additional dictionary available	m-WER [%]	m-invWER [%]	BLEU [%]	SSER [%]	ISER [%]
no	35.2	30.0	53.0	32.7	17.9
yes	33.8	28.4	53.8	31.4	16.4

Table 7.9: Highest ranking tag sequence pairs

Rank	**German**	**English**
1	adverb	adverb
2	personal pronoun	pronoun
3	noun singular	noun singular
4	preposition	preposition
5	definite article singular	determiner
6	coordinating conjunction	coordinating conjunction
7	cardinal number	cardinal number
8	ordinal number	ordinal number
9	verb indicative present 3rd singular	verb present 3rd singular
10	adjective positive	adjective absolute

7.3.2 Impact of the corpus size

It is a costly and time consuming task to compile large texts and have them translated to form bilingual corpora suitable for training the model parameters for statistical machine translation. As a consequence, it is an important question how much of this data is necessary to sufficiently cover the vocabulary expected in testing, and going further, to what extent the introduction of morphological knowledge sources can reduce this amount of necessary data. Figure 7.2 shows the relation between the size of a typical German corpus and the corresponding number of different full forms. At the size of 520 000 words, the size of the Verbmobil corpus used for training, this curve still has a large growth rate.

Table 7.10: Results for disambiguated lexicon. Task: Verbmobil. Testing on 527 sentences ("Test" and "Develop"). System: Alignment templates. Setup: 5k sentences for training; treatment of question inversion, separated verb prefixes and inserted infinitive markers; merging of multi-word phrases in both languages; standard full word form lexicon.

lexicon disambiguated	m-WER [%]	m-invWER [%]	BLEU [%]	SSER [%]	ISER [%]
no	34.7	28.7	52.1	33.6	15.2
yes	34.8	28.7	51.6	33.3	14.7

To investigate the impact of the size of the bilingual corpus available for training on the translation quality, three different setups for training the statistical lexicon on Verbmobil data have been defined:

1. Using the full training corpus as described in Table A.1, comprising 58 000 sentences;
2. restricting the corpus to only 5 000 sentences (≈ every 11th sentence);
3. using no bilingual training corpus at all (only a bilingual dictionary, cf. below).

The language model is always trained on the full English corpus. The argument for this is that monolingual corpora are always easier and less expensive to obtain than bilingual corpora. A conventional dictionary is used in all three setups to complement the bilingual corpus. In the last setup, the lexicon probabilities are trained exclusively on this dictionary — as always on the basis of automatically established word alignments. Generally the quality of the alignments decreases with the size of the training corpus. On the other hand, one can argue that the hand-crafting of word alignments on relatively small amounts of data is in principle feasible and in the case of conventional dictionaries, which contain pairings of words or short phrases, predominantly trivial. In order to better predict the benefit from performing this laborious hand-aligning, the hand-crafted alignment on the dictionary is simulated by adopting the alignment produced during the training in the first setup, where the whole bilingual corpus was available.

As Table 7.11 shows, the quality of translation drops significantly when the amount of bilingual data available during training is reduced: When restricting the training corpus to only 5 000 sentences, the SSER increases by about 7% and the ISER by about 3%. As could be expected, the translations produced by the system trained exclusively on a

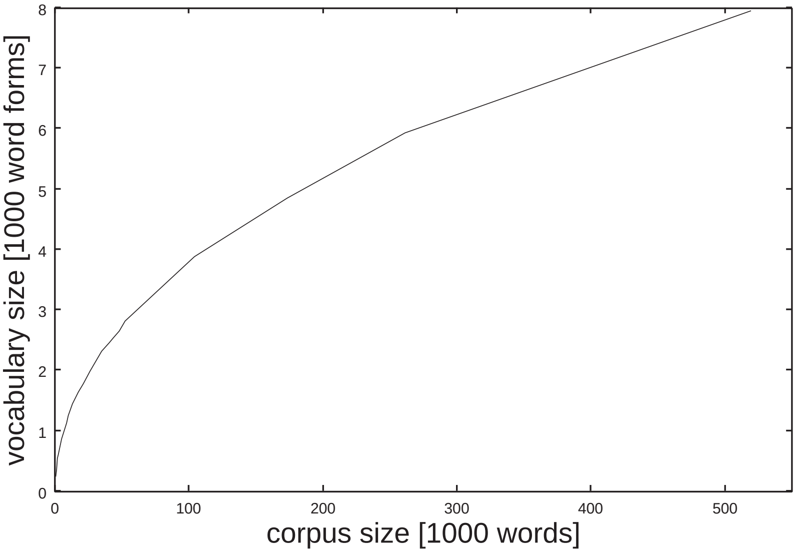


Figure 7.2: Impact of the corpus size (measured in number of running words in the corpus) on the vocabulary size (measured in number of different full form words found in the corpus) for the German part of the Verbmobil corpus.

Table 7.11: Impact of the size of the training corpus. Task: Verbmobil. Testing on 527 sentences (“Test” and “Develop”). The language model is trained on the full monolingual English corpus. System: Alignment templates.

#sentences for training	method for aligning	m-WER [%]	m-invWER [%]	BLEU [%]	SSER [%]	ISER [%]
58k	standard	34.1	27.6	53.7	30.2	14.1
5k	standard	38.0	31.4	47.4	37.3	17.4
0	standard	54.2	47.3	22.0	60.5	28.7
	simul. hand-aligning	53.6	46.2	23.3	60.4	29.8

conventional dictionary are very bad: The SSER jumps over 60%. Hand-aligning is not expected to improve the translation quality, as can be concluded from the result of the experiment with simulated hand-aligning.

7.3.3 Results for log-linear lexicon combination

Results on the Verbmobil task

As has already been pointed out in Chapter 5, the hierarchical lexicon is expected to be especially useful in the case that many of the inflected word forms to account for in test do not occur during training. To systematically investigate the model's generalization capability, they have been applied on the three different setups described in Section 7.3.2. The training procedure was the one proposed in Section 5.4, which includes restructuring transformations in training and test. Table 7.12 summarizes the improvement achieved for all three setups by the methods described in this thesis.

Training on 58k sentences plus conventional dictionary: Compared to the effect of restructuring already reported in Table 7.3 the additional improvement achieved with the hierarchical lexicon is relatively small in this setup. The combination of all methods results in a relative improvement in terms of subjective sentences error rate (SSER) of almost 13% and in terms of information item semantic error rate (ISER) of more than 16% as compared to the baseline.

Training on 5k sentences plus conventional dictionary: Restructuring improves the translation quality by 3.7% absolute. The benefit from the hierarchical lexicon is larger in this setup and the overall reduction in SSER is 5.5%. This is a relative improvement of almost 15.0%. The relative improvement in terms of ISER is even almost 22%. Note that by applying the methods proposed in this thesis the corpus for training can be reduced to less than 10% the original size while loosing only 1.6% in terms of SSER compared to the baseline when using the full corpus.

Training only on conventional dictionary: In this setup the impact of the hierarchical lexicon is clearly larger than the effect of the restructuring methods, because

Table 7.12: Results for hierarchical lexicon models and translation with scarce resources. "restructuring" entails treatment of question inversion, separated verb prefixes and infinitive markers as well as merging of phrases in both languages. A conventional dictionary is available in all three setups. For the experiments without any full bilingual corpora available for training, the hand-crafting of the word alignments on the dictionary is simulated. Task: Verbmobil. Testing on 527 sentences ("`Test`" and "`Develop`"). System: Alignment templates.

#sent. for training		m-WER [%]	m-invWER [%]	BLEU [%]	SSER [%]	ISER [%]
58k	baseline	34.1	27.6	53.7	30.2	14.1
	restructuring	32.5	26.8	56.3	26.6	12.8
	+ dict. disambiguated + hierarchical lexicon	31.8	25.8	57.1	26.3	11.8
5k	baseline	38.0	31.4	47.4	37.3	17.4
	restructuring	34.7	28.7	52.1	33.6	15.2
	+ dict. disambiguated + hierarchical lexicon	33.9	27.8	52.9	31.8	13.7
0	baseline	53.6	46.2	23.3	60.4	29.8
	restructuring	50.2	43.2	29.1	57.8	30.0
	+ dict. disambiguated + hierarchical lexicon	48.0	40.6	32.6	52.1	24.1

here the data sparseness problem is much more important than the word order problem. The overall relative reduction in terms of SSER is 13.7% and in terms of ISER 19.1%. An error rate of about 52% is still very bad, but it is close to what might be acceptable when only the gist of the translated document is needed, as is the case in the framework of document classification or multi-lingual information retrieval.

Results for hierarchical lexicon models with the single word system are presented in Table 7.13. The results on speech input using these models as summarized in Table 7.14 are especially interesting: In the case of sufficient training data, the success of hierarchical lexicon models relies primarily on their disambiguation capability. From the reduction in

Table 7.13: Results on Verbmobil Eval-2000 for hierarchical lexicon models with the single word system. Training on 58k sentences plus conventional dictionary. Testing on 527 sentences ("`Test`" and "`Develop`"). "restructuring" entails: treatment of question inversion and merging of phrases in both languages.

	m-WER [%]	m-invWER [%]	BLEU [%]	SSER [%]	ISER [%]
baseline	35.9	29.9	49.7	32.6	16.1
restructuring + hierarchical lexicon	34.3	28.3	52.8	29.1	14.3

Table 7.14: Results for hierarchical lexicon model: Verbmobil Eval-147 speech input. "restructuring" entails treatment of separated verb prefixes and infinitive markers as well as merging of phrases in both languages. The hierarchical lexicon model is used in connection with a disambiguated conventional dictionary. System: Alignment templates.

	m-WER	BLEU	SSER	ISER
	[%]	[%]	[%]	[%]
baseline	50.7	39.1	40.8	38.9
restructuring	50.9	37.9	39.9	33.0
+ hierarch. lexicon	50.5	37.9	39.2	29.2

ISER, the conclusion can be drawn that speech input contains enough well-formed information to enable correct disambiguation, which is encouraging though not self-evident.

Examples taken from the Verbmobil Eval-2000 test set are given in Figure 7.3. Smoothing the lexicon probabilities over the inflected forms of the same lemma enables the translation of "sind" by "would" instead of "are". The smoothed lexicon contains the translation "convenient" for any inflected form of "bequem". The comparative "more convenient" would be the completely correct translation. The last two examples demonstrate the effect of the disambiguating analyzer which on the basis of the sentence context identifies "Zimmer" as plural (it has been translated into the singular form "room" by the baseline system) and "das" as article to be translated by "the" instead of a pronoun, which would be translated as "that". The last example demonstrates that over-fitting on domain specific training can be problematic in some cases: Generally, "because" is a good translation for the co-ordinating conjunction "denn", but in the appointment scheduling domain, "denn" is often an adverb and it often occurs in the same sentence as "dann", like in "Wie wäre es denn dann?". The translation for this sentence is something like "How about then?". Because of the frequency of this domain specific language use the word

input	`sind Sie mit einem Doppelzimmer einverstanden?`
baseline	`are you agree with a double room?`
hierarch. lexicon	`would you agree with a double room?`
input	`mit dem Zug ist es bequemer.`
baseline	`by train it is UNKNOWN-bequemer.`
hierarch. lexicon	`by train it is convenient.`
input	`wir haben zwei Zimmer.`
baseline	`we have two room.`
hierarch. lexicon	`we have two rooms.`
input	`ich würde das Hilton vorschlagen denn es ist das beste.`
baseline	`I would suggest that Hilton then it is the best.`
hierarch. lexicon	`I would suggest the Hilton because it is the best.`

Figure 7.3: Examples for the effect of the hierarchical lexicon

form "denn" is often aligned to "then" in the training corpus. The hierarchical lexicon distinguishes the adverb reading and the conjunction reading, and the correct translation "because" is the highest ranking one for the conjunction.

Results on the Nespole! task

In the final phase of this work, I was provided with a small German–English corpus from the Nespole! project (Thanks to the consortium, listed on the Nespole! homepage [Nespole! 00]. Special thanks to Alon Lavie, Lori Levin, Stephan Vogel and Alex Waibel.). See Section A.3 for a description. From Table A.9 on page 94 it is obvious that this task is an example of very scarce training data, and it is thus interesting to test the performance of the methods proposed in this work on this task. The same conventional dictionary as for the experiments on Verbmobil data complemented the small bilingual training corpus. Besides, the (monolingual) English part of the Verbmobil corpus was used in addition to the English part of the Nespole! corpus for training the language model. Table 7.15 summarizes the results. Information items have not been defined for this test set. An overall relative improvement of 16.5% in the SSER can be achieved.

7.4 Summary

In this section the most important results presented in this chapter are summarized and discussed as to their statistical significance. The literature provides a range of techniques for testing the statistical significance of improvements. Such techniques involve the computation of confidence intervals for the expected number of errors on the basis of random samples. The difficulty in the case of natural language processing tasks in general and machine translation in particular lies in the definition of an error. Although any such definition is in a sense ad-hoc, the procedure proposed in this section is straightforward and allows some insight on the relevance of results.

Table 7.15: Results for hierarchical lexicon model: Nespole!. "restructuring" entails treatment of question inversion, separated verb prefixes and infinitive markers as well as merging of phrases in both languages. The same conventional dictionary as in the experiments for the Verbmobil data was used. The language model was trained on a combination of the English parts of the Nespole! corpus and the Verbmobil corpus. System: Alignment templates.

	m-WER [%]	m-invWER [%]	BLEU [%]	SSER [%]
baseline	50.2	45.1	31.6	41.1
restructuring	45.9	41.0	33.7	38.1
+ hierarch. lexicon	44.1	40.0	36.5	34.3

The *approximative 2-sample Gauß-test* requires two samples of size n_1 and n_2 for the methods 1 and 2 and the corresponding numbers of errors $\sum X_1$ and $\sum X_2$. The error counts are assumed to satisfy Binomial distributions. The *Counter-Hypothesis*

$$H_0 : \mu_1 \leq \mu_2$$

can be rejected with probability of a false reject below α, when a certain *test function* v is above the $(1-\alpha)$-quantile of the standard normal distribution. Rejecting H_0 supports the Hypothesis

$$H_1 : \mu_1 > \mu_2$$

that method 1 yields on average more errors than method 2. The test function is calculated as follows:

$$v = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(\sum X_1 + \sum X_2)(n_1+n_2-\sum X_1 - \sum X_2)}{(n_1+n_2)\cdot n_1 \cdot n_2}}} \ .$$

The m-WER (see Definition (6.1) on page 48) without the denominator is chosen for counting the errors (the quantities X_1 and X_2) within the words produced for the test sets by different systems. The number of words in the translations is then the sample size n_1 or n_2, respectively. The comparison is always carried out between the baseline result and the system yielding the best results according to subjective sentence error rate.

Verbmobil English to German: By applying the restructuring transformations, the SSER can be reduced by 10.3% relative from 26.3% to 23.5%.

On the basis of the word errors, a significance analysis cannot be successful for the best system variant, because due to the phrase merging the WER increases. Instead, a significance analysis has been performed for the second best system variant with SSER 23.6%, where prefix verb treatment has been combined with morphological corrections. For this variant, the m-WER decreases from 30.2% to 27.8%. According to this criterion, the counter-hypothesis which assumes that the new system is *not* better than the old one can be rejected to significance level 90%.

Verbmobil German to English: For all three setups discussed in Section 7.3.2, the combination of the methods proposed in this thesis yields a relative improvement in terms of subjective sentence error rate between 13% and 15% in terms of SSER and between 16% and 22% in terms of ISER. When the suggested methods are applied, the size of the training corpus can be reduced to less than 10% the size of the original corpus while loosing only 1.6% absolute in terms of SSER compared to the baseline result achieved using the full corpus.

The counter-hypothesis which assumes that the new, improved system is *not* better than the old one can be rejected on the basis of the word error counts to the significance level 99% for all three setups.

Zeres: Restructuring yields a relative reduction in SSER of 16% for this task. A significance analysis has not been performed, but the difference seems large enough to be meaningful.

Nespole!: Using a combination of all proposed methods the SSER could be reduced by 16.5% relative.

On the other hand, an m-WER improvement from 50.5% to 44.1% is achieved in this case. According to the word error counts, the counter-hypothesis which assumes that the new, improved system is *not* better than the old one can be rejected to the level 90%.

Chapter 8

Scientific Contributions

> There is little doubt that statistics-based techniques will be a feature of many future MT projects, although whether many will follow the exclusivity of the IBM team is uncertain. At the present time, the assumption is that linguistic data and methodology will remain at the centre of any practical MT system.
>
> (W. John Hutchins and Harold L. Somers, 1992)

Morpho-syntactic information for statistical machine translation

Various methods of incorporating morphological and syntactic information into systems for statistical machine translation have been proposed and systematically assessed. The overall goal was to improve translation quality and to reduce the amount of parallel text necessary to train the model parameters. The development of the suggested methods was guided by the analysis of important causes of errors. Substantial improvements on the Verbmobil task, the Nespole! task and the Zeres task, for German to English and English to German translation and for text input and speech input could be achieved.

Treatment of structural differences:
A range of sentence level restructuring transformations have been introduced which are motivated by knowledge about the sentence structure in the involved languages. These transformations aim at the assimilation of word orders in related sentences. In detail the suggested restructuring operations focus on the following aspects of structural difference: question inversion, German separable prefix verbs including inserted infinitive markers, compound words and multi-word phrases. A technique for detecting multi-word phrases using syntactical clues has been described, which uses conventional dictionaries for validating phrase candidates. The application of the suggested transformations results in better alignments and as a consequence in less noisy probabilistic lexica, broader applicability of multi-word phrase pairs and a better coverage of the language model.

Translation of word forms not seen in training:
In this work information from morphological analysis has been incorporated into statistical

systems for MT along two different lines in order to infer the translation of word forms not contained in the training data: (1) Mapping of unseen word forms to closely related, but more abstract known word forms and (2) using the information at different levels of abstraction as features in hierarchical lexicon models (see below). In addition, the decomposition of compound words results in a better coverage of the vocabulary.

Hierarchical lexicon models:
A hierarchy of equivalence classes has been defined on the basis of morphological and syntactic information beyond the surface forms. The study of the effect of using information from either degree of abstraction led to the construction of hierarchical lexicon models which combine different items of information in a log-linear way. The benefit from these combined models is twofold: Firstly, the lexical coverage is improved, because the translation of unseen word forms can be derived by considering information from lower levels in the hierarchy. Secondly, category ambiguity can be resolved, because syntactical context information is made locally accessible by means of annotation with morpho-syntactic tags. As a side-effect of the preparative work for setting up the underlying hierarchy of morpho-syntactic information, those pieces of information inherent in fully inflected word forms are detected, which are not relevant for the translation task.

Disambiguation of conventional bilingual dictionaries without context:
In this work a method for aligning corresponding readings in conventional dictionaries containing pairs of fully inflected word forms has been proposed. The approach uses information deduced from one language side to resolve category ambiguity in the corresponding entry in the other language. The resulting disambiguated dictionaries have proven to be better suited for improving the quality of machine translation, especially if they are used in combination with the hierarchical lexicon models.

Translation with scarce resources:
The amount of bilingual training data required to achieve an acceptable quality of machine translation has been systematically investigated. All the methods mentioned above contribute to a better exploitation of the available bilingual data and thus to improving translation quality in frameworks with scarce resources. Three setups for training the parameters of the statistical lexicon on Verbmobil data have been examined: (1) Using the full 58k sentences comprising bilingual training corpus; (2) restricting the corpus to 5k sentences and (3) using only a conventional dictionary. For each of these setups, a relative improvement in terms of subjective sentence error rate between 13 and 15 percent as compared to the baseline could be obtained using combinations of the methods described in this work. The amount of bilingual training data could be reduced to less than 10% of the original corpus, while loosing only 1.6% in subjective sentence error rate. A relative improvement of 16.5% in terms of subjective sentence error rate could also be achieved on the Nespole! task.

Evaluation of machine translation

A tool for the evaluation of translation quality which accounts for the specific requirements in a research environment has been developed. Evaluation criteria which are more adequate than pure edit distance have been defined. The measurement along these quality criteria is performed *semi-automatically* in a fast, convenient and above all consistent way using the tool and the corresponding graphical user interface. The quality criteria themselves have been systematically assessed. The software is registered in the Natural Language Software Registry of the Association of Computational Linguistics [ACL Registry] and freely available for non-commercial purposes [`EvalTrans`].

Chapter 9

Future Directions

> The practical upshot of all this is that if you stick a Babel fish in your ear you can instantly understand anything said to you in any form of language. The speech patterns you actually hear decode the brain-wave matrix which has been fed into your mind by your Babel fish.
>
> (Douglas Adams, “The Hitchhiker’s Guide to the Galaxy”)

For future work along the lines explored in this work, the development of a very general viewpoint on any kind of available source of information is desirable. It might be possible to investigate the usefulness of such varied types of data and knowledge sources as high quality bilingual sentence aligned corpora, comparable texts, dictionaries, monolingual data or even representations of linguistic language theory in a unified framework, no matter if the information is domain dependent or domain independent, specific for certain text types (like dialogs) or more generally applicable in a unified framework. For some languages, like German or English for instance, all of these kinds of information are available and can be generally useful. In these cases, data driven machine learning can be guided by linguistic knowledge. On the other hand real world data can validate the relevance of linguistic knowledge for the envisioned task. In other cases, not all are available and a framework for substituting one by the other is necessary. Again, the examination of methods for unifying the incorporation of different types of knowledge into more generalized frameworks is expected to be beneficial.

Apart from these somewhat philosophical considerations, the following concrete suggestions for future refinements and investigations follow directly from the experiences made during this work:

Implementation of tools for morpho-syntactic analysis and generation: The morphological and syntactic information which is at the basis of the methods proposed in this work has yet been taken from analyses performed using commercial software. Implementing own tools for morpho-syntactic analysis would have various advantages: (1) To become independent of commercial software is favorable due to obvious considerations; (2) the software could be designed to be more flexible in order to add new features; (3) the implementation could be platform independent and library based for direct integration into demonstration prototypes; (4) the methodology could easily be applied to new languages with similar morphological and structural schemata; (5) moreover the algorithms themselves could be designed to be data driven thus be integrated in a more straightforward way into a machine learning approach to natural language processing; (6) it would be easier to integrate the analysis more tightly into the overall decision process in translation. In the course of such tighter integration, it would be possible (if necessary in tasks other than the relatively unambiguous texts considered in this work) to perform the weighted sum over the interpretations as in Equation (5.7) on page 37.

Note that the implementation of such tools is not trivial. An example is morphological analysis which (not only) because of allomorphy phenomena, like in "*Bücher*", the plural of "*Buch*", or "*win*" versus "*won*", does not merely amount to segmenting word forms into free morphemes and affixes.

A unified and refined model for compounds, separable verbs and phrases:
A more unified treatment of compound and non-compound words would be possible by learning alignments and lexical probabilities for both, the individual components *and* for their combination. The decision about which of them provides more useful information can be left to the overall decision process in translation. This also includes the decision about the often ambiguous segmentation of compound words, like with 'Wachtraum', which can be segmented into 'Wach Traum' ('day-dream') or into 'Wacht Raum' ('guard-room'): In most cases one of the segmentations has a significantly higher a-priory probability at least within one domain (is the document about 'dreams' or about 'guards'?). This procedure can also be applied to languages like English, which in contrast to German performs compounding mostly via juxtaposition: the components are still separated via blanks or dashes. Compounding does not cause such severe problems here as it does in German, as shown by the rate of singletons in both languages, but still it can be beneficial to consider compounds in English for closer inspection in a way similar to the one suggested above and less strict than pure phrase merging as proposed in this thesis. An analogous reasoning is valid for separable prefix verbs and their equivalents in English, namely verbs with particles or prepositions, like 'go back' or 'put down', although these phenomena are more difficult to capture because the considered groups of verbs are non-contiguous and modeling them not only affects the lexicon but also the alignment.

Translation from English into German:
The two stage translation as suggested in Section 5.1.1 could be elaborated further using more than only the base forms as output language representation in the first stage. A grammar based language model in combination with a generation component (see below) could be used to generate the correct inflected forms in the target language in the second stage. In addition to the schema depicted in Figure 5.1, annotation of the *input* language words with morpho-syntactic information can be expected to be beneficial also in the case of English as the source language: English has a high rate of category ambiguity, as unlike German, it does not distinguish verbs and adjectives on the one hand and nouns on the other hand via lower case or upper case writing.

Generation of translations for unseen word forms on demand:
In this work methods for inferring translations for unseen word forms in the *source* language have been suggested. Generally these methods amount to generating translations which are often less specific than the input. As an example, the word form "bequemer" (the comparative form of "convenient") can now be translated by the more general word "convenient", but there is no means yet to generate the comparative "*more* convenient". It can be imagined to implement a (data driven) system for generating, e.g. the word form "billig*eres*" from the information "billig in comparative singular nominative neuter form" which is the translation of "cheap in comparative form" in a context where the noun modified by "billig" is singular nominative neuter. This will be especially useful when the target language has more inflectional morphology than the source language like for instance in the case of English to German translation. Another interesting application is the translation of a concept language like in interlingua approaches or the translation of concepts representing languages of completely different nature than English or German, like for instance is the case for transcriptions of sign languages [Bauer & Nießen^{+} 99] into word forms in spoken languages.

Appendix A

Corpora

This annex summarizes information on the different corpora used in this work.

A.1 Verbmobil

Verbmobil was a project for automatic speech–to–speech translation of spontaneous speech. It was a joint initiative of information technology companies, universities, and research centers, and was funded by the German Federal Ministry for Education, Science, Research and Technology (BMBF). Verbmobil assists dialogs in the domain of appointment negotiation (phase I), travel planning and hotel reservation (phase II) for the languages German, English and Japanese. Details can be found in [Wahlster 00]. The Lehrstuhl für Informatik VI at the University of Technology Aachen contributed a statistical system for translating the output of a recognizer to the Verbmobil project. A detailed description of the system is given in [Ney & Nießen$^+$ 00] and in [Och 02].

Verbmobil training corpus: Table A.1 summarizes the characteristics of the corpus used for training the parameters of Model 4 [Brown & Della Pietra$^+$ 93b]. The original corpus was encoded in ASCII with a TEX-like notation for German umlauts. In the experiments described in this work, the corpus encoding was changed to Iso.

Conventional dictionary: A conventional dictionary complements the training corpus (see Table A.2 for the statistics).

The official vocabularies: The vocabulary in Verbmobil was considered closed: there are official lists of word forms, which can be produced by the speech recognizers. Such lists exist for both, German and English. See Table A.3.

Verbmobil German to English test corpora: Table A.4 lists the characteristics of the two test sets "`Test`" and "`Develop`" taken from the end–to–end evaluation in Verbmobil, the development part being meant to tune system parameters on a held out corpus different from the training as well as the test corpus. As no parameters are optimized on the development set for the methods described in this work, most

Table A.1: Corpus statistics: Verbmobil training. Singletons are types occurring only once in training.

	English	German
no. of sentences	58 073	
no. of running word forms	549 921	519 523
no. of running word forms w/o punctuation	453 612	418 974
no. of word forms	4 673	7 940
no. of singleton word forms	1 698	3 453
no. of base forms	3 639	6 063
no. of singleton base forms	1 236	2 546

Table A.2: Conventional dictionary used to complement the Verbmobil training corpus. The dictionary used in this work is slightly different from the one used in [Och 02].

	English	German
no. of entries	10 498	
no. of running word forms	15 305	12 784
no. of word forms	5 161	7 021
no. of base forms	3 666	5 479

of the experiments were carried out on a joint set containing both test sets. Some experiments were also carried out on a different test set, named “`Eval-147`”, for which the single-best hypotheses given by a speech recognizer (accuracy 69%) were available. Table A.5 provides the corresponding characteristics for the transcriptions (“`Transcription`”) and the single-best hypotheses (“`Single-Best`”).

Verbmobil English to German test corpus: The Verbmobil test set characteristics for translation direction English into German are summarized in Table A.6.

A.2 Zeres

The Zeres corpus, collected within the EuTrans project [Amengual & Benedí$^+$ 96], consists of different types of German–English texts belonging to the tourism domain: hotel web pages, brochures and business correspondence. Table A.7 summarizes the corpus

Table A.3: The official vocabularies in Verbmobil.

	English	German
no. of word forms	6 871	10 157
no. of base forms	3 268	6 667

Table A.4: Tests sets Verbmobil Eval-2000 for German to English translation (“`Test`” and “`Develop`”).

	`Test`	`Develop`
no. of sentences	251	276
no. of running word forms in German part	2 628	3 159
no. of word forms in German part	429	434
trigram LM perplexity of reference translation	30.5	28.1

Table A.5: Test set Verbmobil Eval-147 for German to English translation (“`Transcription`” and “`Single-Best`”).

	`Transcription`	`Single-Best`
no. of sentences	147	
no. of running word forms in German part	1 968	1 933
no. of word forms in German part	415	397
trigram LM perplexity of reference translation	28.1	

statistics of the Zeres training set for the original corpus and Table A.8 provides the corresponding figures for the open vocabulary test set. The first 100 test sentences contain 78 words never seen in training (549 on the whole open vocabulary test corpus).

A.3 Nespole!

Nespole! is a research project running from January 2000 to June 2002. It aims at providing multi-model support for negotiation [Nespole! 00, Lavie & Langley$^+$ 01]. Table A.9 summarizes the corpus statistics of the Nespole! training set. Table A.10 provides the corresponding figures for the test set used in this work.

Table A.6: Test set Verbmobil Eval-2000 for English to German translation (“`Test`”).

no. of sentences	248
no. of running word forms in German part	3 040
no. of word forms in German part	355
trigram LM perplexity of reference translation	54.6

Table A.7: Corpus statistics: Zeres training.

	English	German
no. of sentences	27 025	
no. of running word forms	561 804	498 420
no. of running word forms w/o punctuation	495 649	432 201
no. of word forms	33 905	58 271
no. of singleton word forms	16 003	34 359
no. of base forms	23 802	44 775
no. of singleton base forms	11 041	25 523

Table A.8: Corpus statistics: Zeres open vocabulary test set (“`Test-Open`”).

no. of sentences	493
no. of running word forms in German part	8 037
no. of word forms in German part	3 176
trigram LM perplexity of reference translation	240.3

Table A.9: Corpus statistics: Nespole! training.

	German	English
no. of sentences	3 182	
no. of *distinct* sentences	1 767	1 758
no. of running word forms	14 992	15 568
no. of running word forms w/o punctuation	11 672	12 461
no. of word forms	1 363	1 034
no. of singleton word forms	641	403
no. of base forms	1 072	870
no. of singleton base forms	461	326

Table A.10: Corpus statistics: Nespole! test.

no. of sentences	70
no. of running word forms in German part	456
no. of word forms in German part	180
trigram LM perplexity of reference translation	76.9

Appendix B

Performance Measures

The following evaluation criteria were used in the experiments:

m-WER (multi-reference word error rate):
For each test sentence there is a set of reference translations. For each translation hypothesis, the edit distance (number of substitutions, deletions and insertions) to the most similar reference is calculated.

m-invWER (multi-reference word error rate with inversions):
Similar to the m-WER: the usual Levenshtein distance is replaced by a different edit distance, which allows for an additional editing operation, namely the inversion of sub-phrases.

BLEU (**Bil**ingual **E**valuation **U**nderstudy):
This score has been proposed by [Papineni & Roukos$^+$ 01]. It is based on the notion of modified n-gram precision, with $n \in \{1, \ldots, 4\}$: All candidate unigram, bigram, trigram and four-gram counts are collected and clipped against their corresponding maximum reference counts. The reference n-gram counts are calculated on a corpus of reference translations for each input sentence. The clipped candidate counts are summed and normalized by the total number of candidate n-grams. The geometric mean of the modified precision scores for a test corpus is calculated and multiplied by an exponential brevity penalty factor to penalize too short translations. BLEU is an accuracy measure, while the others are error measures.

SSER (subjective sentence error rate):
Each translated sentence is judged by a human examiner according to an error scale from 0.0 (semantically and syntactically correct) to 1.0 (completely wrong).

ISER (information item semantic error rate):
The test sentences are segmented into information items; for each of them, the translation candidates are assigned either "OK" or an error class. If the intended information is conveyed, the translation of an information item is considered correct, even if there are slight syntactic errors, which do not seriously deteriorate the intelligibility.

Appendix C

Symbols and Acronyms

C.1 Mathematical symbols

$\mathcal{DB}$	evaluation database
$\|\mathbf{EE}\|(\mathcal{DB})$	average absolute extrapolation error
$\mathbf{K}$	highest ranking quality class
$\mathcal{II}(s)$	information items for test sentence s
$\mathcal{R}(\mathbf{s})$	references for test sentence s in the evaluation database
$\mathcal{T}(\mathbf{s})$	translations for test sentence s in the evaluation database
$\mathbf{d}(\mathbf{t}, \mathbf{r})$	edit distance: minimal number of insertions, deletions and substitutions necessary to transform t into r
$\bar{\mathbf{d}}(\mathbf{t}_\mathbf{1}^\mathbf{n})$	average normalized edit distance
$\mathbf{ii}$	information item
$\mathbf{t_0}$	tag at position zero in lemma–tag representation: base form
$\mathcal{F}_\mathbf{i} = \mathcal{F}(\mathbf{t}_\mathbf{0}^\mathbf{i})$	equivalence class of all words having a morpho-syntactic reading partly represented by the base form t_0 and the morpho-syntactic tags $t_1, \ldots, t_i$
$\mathbf{v}(\mathbf{s}, \mathbf{t})$	manually assigned quality score
$\hat{\mathbf{v}}(\mathbf{s}, \mathbf{t}, \mathcal{T}(\mathbf{s}))$	estimate of the score on the basis of the translations in $\mathcal{DB}$

C.2 Acronyms

BLEU	**bil**ingual **e**valuation **u**nderstudy
DTD	**d**ocument **t**ype **d**efinition
EM	**e**xpectation **m**aximization
eSSER	**e**xtrapolated **s**ubjective **s**entence **e**rror **r**ate
GIS	**g**eneralized **i**terative **s**caling
GUI	**g**raphical **u**ser **i**nterface
IIS	**i**mproved **i**terative **s**caling
IER	**i**nformation item **e**rror **r**ate
ISER	**i**nformation item **s**emantic **e**rror **r**ate
ITG	**i**nversion **t**ransduction **g**rammar
LM	**l**anguage **m**odel
m-WER	**m**ulti-reference **w**ord **e**rror **r**ate
m-invWER	**m**ulti-reference **w**ord **e**rror **r**ate with **inv**ersions
ME	**m**aximum **e**ntropy
MT	**m**achine **t**ranslation
POS	**p**art **o**f **s**peech
PP	**p**er**p**lexity
SMT	**s**tatistical **m**achine **t**ranslation
SSER	**s**ubjective **s**entence **e**rror **r**ate
WER	**w**ord **e**rror **r**ate

Appendix D

Implementation

`process_cg` is a program for parsing and processing the output of the morpho-syntactic analyzers used in this work. Its functionality can be categorized as follows.

Filtering:

- Apply preference rules based on the examination of syntactic tags or base forms (see Section 3.3).
- Ignore certain tags considered not important (see Section 3.5).
- Restrict the number of interpretations per word to a limited number, e.g. one single interpretation. There are several possibilities of imposing this restriction, for instance
  - use only the first interpretations (after application of preference rules for sorting the interpretations).
  - resort to the unanimous parts of all interpretations, dropping tags on which different interpretations disagree.
- Discard all but one of identical interpretations (which may occur after the application of some of the other filtering procedures).
- Remove all results which are identical to the original word form. This can be used to avoid trivial entries in replacement tables (see "Output format" below).

Restructuring:

- Treatment of question inversion, that is removing inversion in training and input in test, and restoring question inversion on the output translation in test (see Section 4.1.1).
- Treatment of separated verb prefixes, that is prepending separated verb prefixes to their main part (see Section 4.1.2). This functionality is only used during the

automatic extraction of separable German verbs on the training corpus. The actual transformation of training and test corpus is performed using pattern replacement tables.

- Splitting up compound words (See Section 4.1.3). Bookkeeping of the performed splits is possible to enable subsequent re-merging of compounds. Using this feature, it is also possible to associate word alignments resulting from training on the transformed corpora to the original word forms. The splitting operation can be restricted to word forms not belonging to certain categories, for instance numbers and proper names. Besides, it is possible to explicitly list compound parts which are excluded from splitting. Optionally, the components of upper case compounds can be written in upper case or lower case.
- Merging of multi-word phrases (See Section 4.1.4). Some of the resulting phrases are only used for *detecting phrase candidates* and the actual transformation is performed using replacement tables.

A similar kind of bookkeeping as for compound splitting is also possible for the sentence level reordering operations (as yet: treatment of question inversion and separated German verb prefixes).

Mapping of unseen word forms: See Section 4.2 for a detailed description.

Output format:

- Original fully inflected word form, base form, morpho-syntactic tags and any combination of them (see for example Section 3.4).
- Normal corpus format, format equivalent to the output format of the analyzers, replacement tables.

Other:

- Labeling of identified proper names and numbers.
- Special treatment of punctuation marks and end-of-turn markers[1].
- Branching and conditional application of the other features and functionalities.

[1] As end-of-turn marker serves the "fragment"-symbol of the analyzers, which is explicitly appended to the end of each turn before it is passed to analysis.

Bibliography

[ACL Registry] Natural Language Software Registry of the Association of Computational Linguistics. Hosted at DFKI in Saarbrücken. `http://registry.dfki.de/`.

[Ahrenberg & Merkel$^+$ 00] L. Ahrenberg, M. Merkel, A. Sågvall-Hein, J. Tiedemann: Evaluation of Word Alignment Systems. In *Proc. LREC 2000: The 2nd Int. Conf. on Language Resources and Evaluation*, pp. 1255–1261, Athens, Greece, May 2000.

[Al-Onaizan & Germann$^+$ 00] Y. Al-Onaizan, U. Germann, U. Hermjakob, K. Knight, P. Koehn, D. Marcu, K. Yamada: Translating with Scarce Resources. In *Proc. of the Seventeenth National Conference on Artificial Intelligence (AAAI)*, pp. 672–678, Austin, TX, Aug. 2000.

[Alshawi & Bangalore$^+$ 98] H. Alshawi, S. Bangalore, S. Douglas: Automatic Acquisition of Hierarchical Transduction Models for Machine Translation. In *Proc. COLING-ACL 1998: The 36th Annual Meeting of the Association for Computational Linguistics and the 17th Int. Conf. on Computational Linguistics*, pp. 41–47, Montréal, P.Q., Canada, Aug. 1998.

[Amengual & Benedí$^+$ 96] J.C. Amengual, J.M. Benedí, A. Castaño, A. Marzal, F. Prat, E. Vidal, J.M. Vilar, C. Delogu, A. di Carlo, H. Ney, S. Vogel: Example-Based Understanding and Translation Systems (EuTrans): Final Report, Part I. Deliverable of ESPRIT project No. 20268, Oct. 1996.

[Bauer & Nießen$^+$ 99] B. Bauer, S. Nießen, H. Hienz: Towards an Automatic Sign Language Translation System. In *Proceedings of the International Workshop on Physicality and Tangibility in Interaction: Towards New Paradigms for Interaction Beyond the Desktop*, 6 pages, Siena, Italy, Oct. 1999.

[Berger & Brown$^+$ 96a] A.L. Berger, P.F. Brown, S.A. Della Pietra, V.J. Della Pietra: A maximum entropy approach to natural language processing. *Computational Linguistics*, Vol. 22, No. 1, pp. 39–72, March 1996.

[Berger & Brown$^+$ 96b] A.L. Berger, P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, J.R. Gillett, A.S. Kehler: Language Translation Apparatus and Method of using Context-based Translation Models. United States Patent, Patent Number 5510981, April 1996.

[Brown & Cocke$^+$ 88] P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, R.L. Mercer, P.S. Roossin: A Statistical Approach to Language Translation. In

Proc. COLING 1988: The 12th Int. Conf. on Computational Linguistics, pp. 71–76, Budapest, Hungary, Aug. 1988.

[Brown & Cocke$^+$ 90] P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, P.S. Roossin: A Statistical Approach to Machine Translation. *Computational Linguistics*, Vol. 16, No. 2, pp. 79–85, 1990.

[Brown & Della Pietra$^+$ 92] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, J.D. Lafferty, R.L. Mercer: Analysis, Statistical Transfer, and Synthesis in Machine Translation. In *Proc. TMI 1992: 4th Int. Conf. on Theoretical and Methodological Issues in MT*, pp. 83–100, Montréal, P.Q., Canada, June 1992.

[Brown & Della Pietra$^+$ 93a] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, M.J. Goldsmith: But Dictionaries are Data Too. In *Proc. ARPA Human Language Technology Workshop '93*, pp. 202–205, Princeton, NJ, March 1993.

[Brown & Della Pietra$^+$ 93b] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, R.L. Mercer: Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[Darroch & Ratcliff 72] J.N. Darroch, D. Ratcliff: Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, Vol. 43, pp. 1470–1480, 1972.

[Della Pietra & Della Pietra$^+$ 95] S.A. Della Pietra, V.J. Della Pietra, J.D. Lafferty: Inducing features of random fields. Technical Report CMU-CS-95-144, Carnegie Mellon University, Pittsburgh, PA, 38 pages, May 1995.

[Dempster & Laird$^+$ 77] A.P. Dempster, N.M. Laird, D.B. Rubin: Maximum likelihood from incomplete data via the EM algorithm. *Journal Royal Statistical Society*, Vol. 39, No. 1, pp. 1–38, 1977.

[`EvalTrans`] EvalTrans, a tool for semi-automatic evaluation of machine translation. Authors: S. Nießen and G. Leusch. Published in [Nießen & Och$^+$ 00] and registered in [ACL Registry]. The tool is freely available; consult the website: `http://www-i6.Informatik.RWTH-Aachen.DE/~niessen/Evaluation/`.

[Foster 00] G. Foster: A maximum entropy/minimum divergence translation model. In *Proc. ACL 2000: The 38th Annual Meeting of the Association for Computational Linguistics*, pp. 37–44, Hong Kong, China, Oct. 2000.

[Gamon & Reutter 97] M. Gamon, T. Reutter: The Analysis of German Separable Prefix Verbs in the Microsoft Natural Language Processing System. Technical report, Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA, 7 pages, Sept. 1997. `http://www.research.microsoft.com/scripts/pubdb/pubsasp.asp?RecordID=16`.

[García-Varea & Casacuberta 01] I. García-Varea, F. Casacuberta: Search Algorithms for Statistical Machine Translation based on Dynamic Programming and Pruning Techniques. In *Proc. MT Summit VIII*, pp. 115–120, Santiago de Compostela, Spain, Sept. 2001.

[Germann & Jahr$^+$ 01] U. Germann, M. Jahr, K. Knight, D. Marcu, K. Yamada: Fast Decoding and Optimal Decoding for Machine Translation. In *Proc. ACL-EACL 2001: The 39th Annual Meeting of the Association for Computational Linguistics - joint with EACL 2001*, pp. 228–235, Toulouse, France, July 2001.

[Hausser 01] R. Hausser: *Foundations of Computational Linguistics, Human-Computer Communication in Natural Language*, chapter 13. Springer, 2nd edition, 2001.

[Hutchins & Somers 92] W.J. Hutchins, H.L. Somers: *An Introduction to Machine Translation.* Academic Press, 1992.

[Kanevsky & Roukos$^+$ 97] D. Kanevsky, S. Roukos, J. Sedivy: Statistical language model for inflected languages. United States Patent, Patent Number 5835888, 1997.

[Karlsson 90] F. Karlsson: Constraint Grammar as a Framework for Parsing Running Text. In *Proc. COLING 1990: The 13th Int. Conf. on Computational Linguistics*, Vol. 3, pp. 168–173, Helsinki, Finland, Aug. 1990.

[Koehn & Knight 01] P. Koehn, K. Knight: Knowledge Sources for Word-Level Translation Models. In L. Lee, D. Harman, editors, *Proc. EMNLP 2001: Conf. on Empirical Methods in Natural Language Processing*, pp. 27–35, Pittsburgh, PA, June 2001. SIGDAT.

[Larson & Willett$^+$ 00] M. Larson, D. Willett, J. Köhler, G. Rigoll: Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches. In *Proc. ICSLP 2000: 6th Int. Conf. on Spoken Language Processing*, Vol. 3, pp. 945–948, Beijing, China, Feb. 2000.

[Lavie & Langley$^+$ 01] A. Lavie, C. Langley, A. Waibel, F. Pianesi, G. Lazzari, P. Coletti, L. Taddei, F. Balducci: Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-commerce Applications. In J. Allan, editor, *Proc. HLT 2001: 1st Int. Conf. on Human Language Technology Research*, pp. 31–39, San Diego, CA, March 2001. Morgan Kaufmann Publishers.

[Leusch & Nießen 02] G. Leusch, S. Nießen: Algorithm for computing the multi-reference word error rate with inversions. Unpublished ongoing work, 2002.

[Levenshtein 65] V.I. Levenshtein: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Doklady Akademii nauk SSSR (in Russian)*, Vol. 163, No. 4, pp. 845–848, 1965. Also in Cybernetics and Control Theory, 10(8), pp 707-710, 1966.

[Maltese & Mancini 92] G. Maltese, F. Mancini: An automatic technique to include grammatical and morphological information in a trigram-based statistical language model. In *Proc. ICASSP 1992: Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 157–160, San Francisco, CA, March 1992.

[Nespole! 00] NEgotiating through SPOken Language in e-commerce. Project homepage, 2000. `http://nespole.itc.it/`.

[Ney & Nießen$^+$ 00] H. Ney, S. Nießen, F.J. Och, H. Sawaf, C. Tillmann, S. Vogel: Algorithms for Statistical Translation of Spoken Language. *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 1, pp. 24–36, Jan. 2000.

[Nießen & Ney 00] S. Nießen, H. Ney: Improving SMT Quality with morpho-syntactic Analysis. In *Proc. COLING 2000: The 18th Int. Conf. on Computational Linguistics*, pp. 1081–1085, Saarbrücken, Germany, July 2000.

[Nießen & Ney 01a] S. Nießen, H. Ney: Morpho-syntactic Analysis for Reordering in Statistical Machine Translation. In *Proc. MT Summit VIII*, pp. 247–252, Santiago de Compostela, Spain, Sept. 2001.

[Nießen & Ney 01b] S. Nießen, H. Ney: Toward Hierarchical Models for Statistical Machine Translation of Inflected Languages. In *39th Annual Meeting of the Association for Computational Linguistics - joint with EACL 2001: Proc. Workshop on Data Driven Machine Translation*, pp. 47–54, Toulouse, France, July 2001.

[Nießen & Och$^+$ 00] S. Nießen, F.J. Och, G. Leusch, H. Ney: An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proc. LREC 2000: The 2nd Int. Conf. on Language Resources and Evaluation*, pp. 39–45, Athens, Greece, May 2000.

[Nießen & Vogel$^+$ 98] S. Nießen, S. Vogel, H. Ney, C. Tillmann: A DP based Search Algorithm for Statistical Machine Translation. In *Proc. COLING-ACL 1998: The 36th Annual Meeting of the Association for Computational Linguistics and the 17th Int. Conf. on Computational Linguistics*, pp. 960–967, Montréal, P.Q., Canada, Aug. 1998.

[Nirenburg 87] S. Nirenburg: Knowledge and Choices in Machine Translation. In S. Nirenburg, editor, *Machine Translation: Theoretical and Methodological Issues*, pp. 1–21. Cambridge University Press, Cambridge, 1987.

[Och 99] F.J. Och: An Efficient Method for Determining Bilingual Word Classes. In *Proc. EACL 1999: The 9th Conf. of the Europ. Chapter of the Association for Computational Linguistics*, pp. 71–76, Bergen, Norway, June 1999.

[Och 01] F.J. Och: YASMET: Yet Another Maximum Entropy Toolkit, 2001. `http://www-i6.Informatik.RWTH-Aachen.DE/~och/software/YASMET.html`.

[Och 02] F.J. Och: *Machine Translation: From Single-Word Models to Alignment Templates.* Ph.D. thesis, Computer Science Department, RWTH - University of Technology, Aachen, Germany, 2002. Submitted, but not yet published.

[Och & Ney 00] F.J. Och, H. Ney: Improved Statistical Alignment Models. In *Proc. ACL 2000: The 38th Annual Meeting of the Association for Computational Linguistics*, pp. 440–447, Hongkong, China, Oct. 2000.

[Och & Tillmann$^+$ 99] F.J. Och, C. Tillmann, H. Ney: Improved Alignment Models for Statistical Machine Translation. In *Proc. EMNLP 1999: Conf. on Empirical Methods in Natural Language Processing*, pp. 20–28, University of Maryland, College Park, MD, June 1999.

[Och & Weber 98] F.J. Och, H. Weber: Improving Statistical Natural Language Translation with Categories and Rules. In *Proc. COLING-ACL 1998: The 36th Annual Meeting of the Association for Computational Linguistics and the 17th Int. Conf. on Computational Linguistics*, pp. 985–989, Montréal, P.Q., Canada, Aug. 1998.

[Papineni & Roukos$^+$ 01] K. Papineni, S. Roukos, T. Ward, W.J. Zhu: Bleu: A Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, 10 pages, Sept. 2001.

[Ratnaparkhi 97] A. Ratnaparkhi: A Simple Introduction to Maximum Entropy Models for Natural Language Processing. Technical Report 97–08, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA, 10 pages, May 1997.

[Ristad 01] E.S. Ristad: Predictive Modeling Toolkit (PMT), 2001. `http://www.mnemonic.com/pmt/`.

[Sparck Jones & Galliers 96] K. Sparck Jones, J.R. Galliers: *Evaluating Natural Language Processing Systems: An Analysis and Review.* Lecture notes in computer science. Springer-Verlag, 1996.

[ten Hacken 01] P. ten Hacken: Has There Been a Revolution in Machine Translation? *Machine Translation*, Vol. 16, No. 1, pp. 1–19, 2001.

[Tessiore & v. Hahn 00] L. Tessiore, W. v. Hahn: Functional validation of a machine interpretation system: Verbmobil. In [Wahlster 00], pp. 611–631.

[Tillmann 02] C. Tillmann: *Word Re-Ordering and Dynamic Programming based Search Algorithm for Statistical Machine Translation.* Ph.D. thesis, Computer Science Department, RWTH - University of Technology, Aachen, Germany, 2002.

[Tillmann & Ney 00] C. Tillmann, H. Ney: Word re-ordering and DP-based Search in Statistical Machine Translation. In *Proc. COLING 2000: The 18th Int. Conf. on Computational Linguistics*, pp. 850–856, Saarbrücken, Germany, Aug. 2000.

[Tillmann & Vogel$^+$ 97] C. Tillmann, S. Vogel, H. Ney, H. Sawaf, A. Zubiaga: Accelerated DP based Search for Statistical Translation. In *Proc. of the 5th European Conference on Speech Communication and Technology*, pp. 2667–2670, Rhodes, Greece, Sept. 1997.

[Vidal 97] E. Vidal: Finite-State Speech-to-Speech Translation. In *Proc. ICASSP 1997: Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 111–114, Munich, Germany, April 1997.

[Vogel & Nießen$^+$ 00] S. Vogel, S. Nießen, H. Ney: Automatic Extrapolation of Human Assessment of Translation Quality. In *2nd International Conference on Language Resources and Evaluation: Proc. of the Workshop on Evaluation of Machine Translation*, pp. 35–39, Athens, Greece, May 2000.

[Wahlster 00] W. Wahlster, editor: *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer Verlag, Berlin, Germany, 2000.

[Wang 98] Y.Y. Wang: *Grammar Inference and Statistical Machine Translation.* Ph.D. thesis, School of Computer Science, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 1998.

[Wang & Waibel 98] Y.Y. Wang, A. Waibel: Modeling with Structures in Statistical Machine Translation. In *Proc. COLING-ACL 1998: The 36th Annual Meeting of the Association for Computational Linguistics and the 17th Int. Conf. on Computational Linguistics*, pp. 1357–1363, Montréal, P.Q., Canada, Aug. 1998.

[White & Taylor 98] J.S. White, K.B. Taylor: A Task-Oriented Evaluation Metric for Machine Translation. In *Proc. First International Conference on Language Resources and Evaluation*, pp. 21–25, Granada, Spain, May 1998.

[Wolters 97] M. Wolters: Compositional Semantics of German Prefix Verbs. In *Proc. ACL-EACL 1997: The 35th Annual Meeting of the Association for Computational Linguistics - joint with EACL 1997*, pp. 525–527, Madrid, Spain, July 1997.

[Wu 95] D. Wu: Grammarless Extraction of Phrasal Translation Examples from Parallel Texts. In *Proc. TMI 1995: 6th Int. Conf. on Theoretical and Methodological Issues in MT*, Vol. 2, pp. 354–372, Leuven, Belgium, July 1995.

[Wu 96] D. Wu: A Polynomial-Time Algorithm for Statistical Machine Translation. In *Proc. ACL 1996: The 34th Annual Meeting of the Association for Computational Linguistics*, pp. 152–158, Santa Cruz, CA, June 1996.

[Yamada & Knight 01] K. Yamada, K. Knight: A Syntax-based Statistical Translation Model. In *Proc. ACL-EACL 2001: The 39th Annual Meeting of the Association for Computational Linguistics - joint with EACL 2001*, pp. 523–530, Toulouse, France, July 2001.

Lebenslauf

Angaben zur Person:

Sonja Nießen
geboren am 8. August 1969
Geburtsort: Geilenkirchen, Deutschland

Schul- und Berusausbildung:

1975 – 1988: Grundschule in Birgden und Gymnasium in Geilenkirchen

1988 – 1991: Ausbildung zur Mathematisch-technischen Assistentin am Lehr- und Forschungsgebiet für Prozessdatenverarbeitung und Prozessführung an der RWTH Aachen

Studium:

10/1991 – 7/1997: Informatikstudium an der RWTH Aachen, Deutschland und der EPFL Lausanne, Schweiz
Vertiefungsfach: Sprachverarbeitung und Mustererkennung
Abschluss als Diplom-Informatikerin mit Auszeichnung

Arbeitstätigkeiten:

9/1991 – 1/1993: Mathematisch-technische Assistentin im Lehr- und Forschungsgebiet für Prozessdatenverarbeitung und Prozessführung, RWTH Aachen

2/1993 – 12/1993: Studentische Hilfskraft am Lehrstuhl für Informationssysteme und Datenbanken, RWTH Aachen

1/1994 – 7/1997: Mathematisch-technische Assistentin am Lehrstuhl für Sprachverarbeitung und Mustererkennung, RWTH Aachen

7/1997 – 8/2002: Wissenschaftliche Angestellte am Lehrstuhl für Sprachverarbeitung und Mustererkennung, RWTH Aachen