Karlsruhe Institute of Technology

Ambiguitiy Resolution with User Interaction

Study Thesis at the
Institute for Anthropomatics
Prof. Dr. Alex Waibel
Faculty for Computer Science
Karlsruhe Institute of Technology

from

Miriam Hundemer

Advisor:
Prof. Dr. Alex Waibel
Dipl.-Inform. Kay Rottmann

Date of Registration:	2012-08-27
Date of Submission:	2012-11-26

I declare that I have developed and written the enclosed Study Thesis completely by myself, and have not used sources or means without declaration in the text.

Karlsruhe, 2012-11-26

M. Andere

Contents

List of Tables

1. Introduction

Spoken language is one of the most important ways for human people to express themselves and communicate with each other. However, as soon as different languages are involved, several problems may occur that have to be solved in order to ensure a correct understanding of what has been said. One of these problems arises from the possibility that one word in a particular language may have different translations in another language depending on the context it is used in. For most human speakers this is not really a problem because they know in which context a word is used and therefore automatically choose the correct translation.

However, the correct translation of ambiguous words poses a serious problem to todays machine translation systems. Such translation systems could be used in many situations where people need to exchange information, and when there are few or no human translators available. This might be the case in disaster zones or during medical operations. In those situations people with different languages/linguistic backgrounds may want to exchange information quickly to be able to help each other.
But in order to really aid human communication it is important to resolve an ambiguous word in the correct way to preserve the intended meaning. Yet even in restricted domains such as medical environments there can be many ambiguous words whose disambiguation may be essential for understanding what has been said.

Therefore, solving the problem of finding ambiguous words and how to translate them correctly would greatly improve the quality of modern translation systems and help people to overcome linguistic barriers.

1.1 Goals

The aim of this thesis is to aid statistical machine translation by resolving ambiguous words with the help of a user.
In order to keep the system small and not create a need for additional resources only instruments should be used for disambiguation, which the baseline system already provided. This way there are less sources to get information from, but the system is going to be smaller and will be able to work even when there are no possibilities

to access additional data resources. Additionally, it will be easier to apply to other languages where only limited resources are available.

To aid the interaction with the user a Graphical User Interface (GUI) should be created, which presents the different possible translations for a word in a way that enables the user to choose the one he intends to use. It should be noted here that the user is considered to be monolingual.

Finally, a way has to be found to integrate the information provided by the user into the machine translation system, to make it available to the translation process.

1.2 Related Work

As natural language processing is a useful feature in many applications several approaches have been made to conquer this area of research. Some of these approaches will be presented in this section.

There is an approach from 1995 written by Blanchon et al. [BLKM95] to create an interactive disambiguation module for English. In this project ambiguities are divided into three different meta classes. Some of the classification rules may only be applied regarding an *ASR (Automatic Speech Recognition)* context but some can also be applied regarding only written language. An ambiguity is marked as *lexical ambiguity* when the analyser that is used cannot choose a word among homophones (as in *to* vs. *too* vs. *two*) or a syntactic class among homographs (as in *conduct* noun vs. verb). There is also a lexical ambiguity when there is no unique segmentation into words or terms ([right here] vs. [right] [here]).
A geometrical ambiguity is detected when there are several solutions with different geometries and no lexical ambiguities exist.
The third class is the labelling ambiguity where the analyser produces several solutions which have the same geometry but are no lexical ambiguities. Some further examples can be seen in table 1.1.

Geometrical Ambiguities	
Prepositional Attachment	Where can I catch a taxi **from** Kyoto station?
Adverbial Attachment	You can pay for it **right** on the bus.
Conjunction	Can I ask you to type in your name **and** the telephone number?
Labeling Ambiguities	
I want a reservation **for** the hotel.	

Table 1.1: Further Examples for Ambiguities

If any of these ambiguities are detected a question for the user is prepared. The questions are made of all the different meanings of the ambiguous word or sentence among which the user has to choose. The system contains pattern- and beam-matching mechanisms to identify ambiguities.

Another approach, from Yamaguchi et al. [YKI^{+}98], combines an automatic and an interactive method for disambiguation. In their system, the interactive method is only used when its calculated accuracy is higher than that of the automatic method.

The accuracy of the interactive mode is defined as the sum of all possibilities for the user to choose a translation while the accuracy of the automatic method corresponds to showing only one translation to the user. This is supposed to guarantee the highest possible overall accuracy because the user will not be confused by poor alternatives and does not need not be bothered if the automatic method would provide a better solution anyway.

In two more recent approaches from 2005 and 2006 additional resources such as *WordNet*[1] or *EuroWordNet*[2] are used to help creating monolingual questions for the user or to help translation.

In the first approach of Orasan et al. [OMC$^+$05] a word sense disambiguation (WSD) module is presented, which translates e-mails or other electronic documents which are redirected to a centrally based translation facility. In this project, EuroWordNet is used as a multilingual dictionary, but word senses had to be adjusted to maximize usefulness to their task. In order to reduce the number of ambiguous words several language processing filters are implemented.

There are, for example, *Part-of-speech tagging*, which assigns labels to words according to their grammatical category, *Named entity recognition* or *Multiword units identification*. *Part-of-speech tagging* allows for reducing the number of senses possible for one word by restricting to senses where the part of speech information of the word is the same.

Named entity recognition allows for the combination of words which refer to an entity with a special meaning such as *"Bill Gates"*. In this case the two words should not be translated separately or not at all.

Similarly, *mulitword units* should always remain combined and not be translated separately because their meaning would change and it even reduces the number of words to be translated. In the end [OMC$^+$05] showed that the language processing filters really helped to reduce the number of ambiguities but that the success also depends on the language pair and the combination of different filters.

The approach from Sammer et al. [SRS$^+$06] did not focus so much on reducing the amount of ambiguities but on creating a controlled language (CL) lexicon, presenting possible translations to the user and handling the choice that has been made. In the CL lexicon there is an entry for each distinct word sense of a term in the source language and is associated with the possible translations into the target language.

To create this lexicon a dictionary from *UltraLingua*[3] and WordNet as machine readable dictionary for creating glosses for each entry were used. Here too, some changes were made to adapt WordNet to the task.

During the translation process the user is then confronted with possible translations for an ambiguous word and possibilities to leave the word untranslated as a proper noun or leave it unannotated. The user's choice is then integrated into the phrase translation table of the underlying SMT system, boosting the score of selected choices and lowering the scores of other translations. Experiments showed that this method of altering phrase-table entries according to human choices really was able to improve the translation accuracy of their SMT system.

[1] http://wordnet.princeton.edu/

[2] http://www.illc.uva.nl/EuroWordNet/

[3] http://www.ultralingua.com/products

1.3 Outline of the Thesis

In the past few paragraphs the goals of this research project were outlined and it was revealed how important spoken and written language are for human communication. There was also given a short overview of other research projects dealing with similar problems.

In chapter 2 a basic overview about phrase-based machine translation and evaluation is provided. There is also a short historical overview of machine translation.
Chapter 3 takes a closer look at the problems occuring when trying to resolve ambiguities. Problems of finding ambiguous words and choosing the right ones to present to the user are discussed here in detail.
Chapter 4 provides descriptions of the strategies used to solve the problems mentioned in chapter three. It contains approaches for phrase table pruning, choosing ambiguities and integration of the user choices into the machine translation system.
In the last two chapters the results of this research project are evaluated and there is an outlook on future development in section 6.

2. Fundamentals

In this chapter, a rough overview about the history of machine translation in general and Statistical Machine Translation as special approach will be provided. At the end of this chapter, the problem of evaluating Machine Translation methods and results will be detailed.

2.1 History of Machine Translation

The history of machine translation reaches back to the days of the first electronic computers. Great efforts were made in World War II to decode foreign language codes such as the German Enigma code. Therefore people are still talking about *decoding* a foreign language today. Researchers in those early days had great hopes and expectations on solving the problem of machine translation early but in 1966 the ALPAC report [Com66] changed minds in stating that there was no advantage in using machine translation systems and that funding and research should rather go into the improvement of human translation and linguistic research. After that, research went on mainly in commercial projects such as **Systran** or **Météo** for the translation of weather forecasts. It was due to the success of those systems that machine translation got more attractive again and research activity increased.

While the early machine translation systems were mostly using large bilingual dictionaries and hand-coded rules for fixing word reordering, in the 1980s and 1990s more research was done in the field of **interlingua-based** systems which represent meanings of a text independent of a specific language. Likewise, efforts were made on data driven methods such as **example-based translation** and later statistical machine translation first developed at IBM. Research in the field of statistical machine translation, went on during the 1990s and is still developed today. Reasons for the rise of statistical machine translation are for example the increase of computational power and data storage capacity and also the increased availability of bilingual corpora over the internet.
Today a lot of research is done on statistical machine translation in both academic and commercial research labs and large software companies such as IBM and Google.

2.2 Statistical Machine Translation

Statistical Machine Translation (SMT) is currently one promising approach to solving the problem of translating human speech with the help of computers. The underlying mathematical formula for SMT is the *Bayes rule*:

$$argmax_e\ p(e|f) = argmax_e\ p(f|e) \cdot p(e) \tag{2.1}$$

Given an foreign sentence f and an English sentence $\mathbf{e}$, it provides the best translation $argmax_e\ p(e|f)$. Referring to the sentences as *foreign* and *English* is based on the fact that in many early approaches the translations had been form a "foreign" language to English. In the Bayes rule, $p(f|e)$ stands for the translation model probability while $p(e)$ stands for the language model probability. The *translation model* (TM) describes the probability of translating a given sentence e into a foreign sentence f while the *language model* (LM) describes the probability of the word sequence e occurring in the target language in the first place.

A more recent approach being used in SMT is the so called *phrase-based MT*. In this approach the basic units can not only be single words but also so called *phrases*.
A *phrase* can be a single word, but can also cover whole sentences or anything in between. During the training process of a machine translation system word alignments between each sentence pair of a parallel corpus are created and phrase pairs are extracted that are consistent with this alignment. A word alignment is a bipartite graph which can be illustrated by the diagram in figure 2.1:

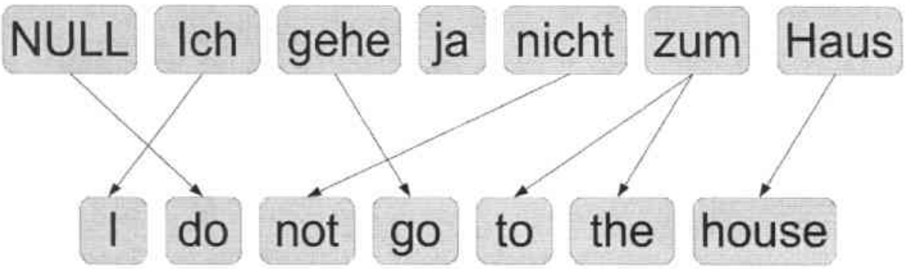


Figure 2.1: Example for a word alignment

In the further course of the training process a phrase translation table is created which contains all the extracted phrase pairs from the corpora used for training. Each entry of the phrase translation table also contains information about translation probabilities of the current phrase pair.
Usually the following scores are used: The *inverse phrase translation probability* $\varphi(f \mid e)$ discribes the probability of translating a phrase e, given in the target language, into a foreign phrase f. The second score is the *inverse lexical weighting* $p_w(f \mid e, a)$, which describes how well the words of a phrase pair f,e translate to each other, given an aligment a between the two. If there exists more than one alignment, the one with the highest lexical weight is used. The next two scores are the *phrase translation probability* and the *lexical weighting* which basically describe the same

probabilities but just the other way round. The fifth and last score ist the *phrase penalty* which penalizes the use of more phrases for a translation than necessary. The *phrase penalty* score is always set to the value 2.718 because it equates to e^1. In the log linear model this equates to a score of 1.0, which will later be optimized by the scaling factors.

2.2.1 Evaluation

The evaluation of a machine translation system is a very difficult task. One reason for this is that for one sentence, there is more than one possible answer which would be considered as a correct translation. Even human translators give different answers if you ask them to translate the same sentence. So in order to evaluate the output of a machine translation system, various metrics can be applied and different methods are used to measure the quality of a translation.
Although one may say that the output of a machine translation system is intended to be used by human people so the evaluation should also be done by humans, this approach is very time- and also money consuming, especially when it comes to very large corpora. Therefore automatic evaluation methods were invented to have a faster and less expensive way of telling if one machine translation system is better than another.

One such method is the *BLEU (Bilingual Evaluation Understudy)* metric [PRWZ02]. It is defined as:

$$BLEU = BP \cdot \sum_{n=1}^{N} w_n \log p_n \tag{2.2}$$

It scores the translation of a machine translation system according to *n-gram* matches compared to different reference translations provided by human translators. *n-grams* are words with a history of n-1 words.
In order to avoid matching the same word in a reference sentence more than once, a *modified n-gram precision* is used. This allows one n-gram to be matched only *max_ref_count* times where *max _ref_count* is the maximum number of times a word occurs in a reference sentence. There is also a penalty for candidate translations which are much too short compared to the references called the *brevity penalty* (BP).

3. Analysis

In human-human communication it is of significant importance to express oneself clearly so that the conversational partner understands what has been said and gets the right message. If people speak the same language, this is usually no problem because the conversational partners are able to uniquely identify the spoken words and know about the context they belong to. But if they do not share this linguistic similarity, communication can already get more difficult. If the context of the conversation is known, there might be a chance to guess what has been said; but if even the context is indistinct or may vary, it becomes really hard to identify the right meaning.

For SMT systems this translation task is even harder because they usually do not dispose of all the information a human speaker posesses. SMT systems usually have only limited knowledge of the domain a word belongs to. In the worst case, they know nothing about the current domain at all. This may happen if the domain the current source text belongs to is different from the domain of the training of the SMT system.

Because of this lack of knowledge it is sometimes difficult for SMT systems to identify ambiguities and especially to translate them correctly.

Therefore it is important to find strategies to aid Statistical Machine Translation to improve translation quality.

3.1 Ambiguities

As already defined in the introduction, one main goal of this research project is to aid the resolution of ambiguous words in order to improve translation quality. The first of various problems to be adressed here is to define what the word *ambiguous* is supposed to mean in a certain situation. Furthermore, how the words that would be ambiguous according to this definition are to be found in a text that also contains many other words that are not relevant for the task.

As can be seen in chapter 1.2, there are many different ways to address those questions. One way is to specify certain classes of ambiguities and applying special metrics to classify each word [BLKM95].Another method is to simply define every word with more than one meaning as ambiguous [SRS$^+$06].

As has been stated before only tools and sources of information should be used which are already included in the baseline system. This decision was made to keep the translation system as small as possible. That way it does not depend on external resources and can be used on any device the baseline systems runs on. Additionally, the amount of language pairs the system can be used for, only depends on the amount of parallel corpora available for a certain language pair. But for the identification of ambiguous words, the decision to keep the system as small as possible means a limitation of ways to gather information as there is no way to use external resources such as WordNet.

Still there has to be found a way to identify ambiguous words. Furthermore not all of the possible ambiguities found during the first attempt should be presented to a user in order to resolve the ambiguity. The list of possibly ambiguous words found in a first attempt may be very long and contain words that do not even carry much information. If a user had to resolve all those ambiguities he would soon be over-burdened and in the end not willing to use the translation system anymore, because it is too time-consuming and stressful.

Therefore, the number of ambiguities the user actually has to resolve needs to be reduced. In this thesis only important words should be presented to the user. Important words are meant to carry much information and are vital for the correct translation of a sentence. So a strategy has to be developed to seperate important, context carrying words from the words which do not carry any information.

As different platforms or devices the translation system runs on may make use of different user interfaces, there should be a simple interface providing easy access to the data which is to be presented to the user.

3.2 User Interface

In order to involve a user in the ambiguity resolving process, the identified ambiguities have to be made accessible. At this juncture it is assumed that the user is presumably monolingual. Therefore the intermediate results have to be prepared to allow for an easy understanding.

There should not be too many different possibilities for the user to choose from. The more possibilities there are for just one word, the more exhausting the translation task becomes for the user. He or she has to look at each translation possibility, compare it to all the other possibilities, and then decide if this is the one he (or she) actually intended to use.

This becomes even more stressful if the different possibilities are very similar to each other and only small nuances discriminate between the different meanings.

Therefore it is of vital importance that the user interface is concise and simple to use so that the user is able to focus on the translation task and not busy finding out how to use the interface he is being presented.

After the user made his choice, the decisions he made have to be integrated back into the translation system so that it can help improve the translation quality. All the possibilities the user did not choose should somehow be annotated too, so that the translation system learn the right translation for future translations and will hopefully not bother the user with unnecessary ambiguities again.

4. Design and Implementation

4.1 Preprocessing

As illustrated in the previous chapter, there are certain problems to overcome on the way of resolving an ambiguous word. This chapter presents one possible way to address those problems. At first, some preliminary steps have to be made on the parallel corpora that are used for developement and training of the MT system.
All of these preliminary steps are done using small Perl or Python scripts which edit the corpora or generate files that hold intermediate results.

4.1.1 Corpora Alignment

For the ambiguity resolution to work correctly it is vital that every sentence in the source language has its counterpart in the target language. Otherwise, there could be words or whole sentences that do not have a known alignment and it would be impossible to translate them during the later progress. Usually parallel corpora are used in SMT, so this should not pose a problem. However, it may happen that two texts are not completely parallel and the translation for some of them is missing. So one first step should be to get rid of sentences in the corpora that do not have a counterpart in the source or target language. In other words, the sentences whose counterpart is a blank line in the corresponding corpus have to be removed. That way it is certain that every source sentence in the corpus is aligned to a sentence in the target language.

4.1.2 High Frequency Words

The next step is to work through the corpora and search for *high frequency words*. These are often numbers, articles and conjuctions such as *and*, *or* or *the* that do not carry much information. For this experiment, high frequency words would be rather disturbing because the user should not be troubled translating such words. So it is best to filter them out beforehand. To achieve that, the occurences of all distinct words in the trainings corpus are counted to create a list, sorted by the number of occurences. Because there is no additional information available such as *part of speech tags* to exactly identify possible high frequency words, the 1000 most frequent words of the sorted list are defined as such. Those 1000 words will be used to narrow down the list of possibly ambiguous words that have to be resolved later.

4.1.3 Phrase Table Pruning

The phrase table represents the most important source of information for the ambiguity identifying process in this thesis. From its entries the possible ambiguities are extracted and stored in special files the preparation of which is described in next section. How the phrase table is prepared in order to extract ambiguous words will be the subject of this section.
To reduce the amount of phrase table entries to a reasonable size the question needs to be adressed which entries could or should be removed without loosing too much of information. Some of these techniques are implemented to achieve this goal.

- A first and rather simple approach is to prune all entries that consist of only numbers or punctuation marks because they will surely not carry any context at all. This step already removes quite a lot of entries which are not relevant for the task of resolving ambiguities. However the phrase table, most certainly, still constains too many entries to present to a user.
- To simplify the task a bit more, all phrases that cover more than one word are removed so that only one word phrases remain. That will shorten the phrase table to quite a reasonable size already, but there still remain a few steps of pruning that can be applied.
- So now, the before gathered high frequency words come into use. As has been stated before, high frequency words are not likely to carry much context or provide any useful information for the user. And even if they do, there would be enough examples so the translation systems would not need additional information from the user. So phrase table entries containing those words can safely be removed, further reducing the overall size of the phrase table. To enable a faster comparison between the current source or target phrase and all the before gathered high frequency words, the latter ones are converted into a Perl *hash*. That way, the look-up of a certain word is much faster than going through a list of all the high frequency words.
- Finally one last step can be applied to shorten the phrase table to its final size. If, after all the before mentioned pruning steps, there remains only one single entry in the phrase table for a particular word this one can be safely removed too. Because with only one remaining phrase table entry for that particular word, there also is only one possible translation the translation system is going to choose. Therefore that word can not be an ambiguity and the user should not be bothered with it.

Applying all the before mentioned steps creates a new, pruned phrase table to work with and gather ambiguities from.
However, during the later process of creating an ambiguity lexicon this pruned phrase table turned out to be still too large. There were still far too many possible ambiguities to be resolved within a reasonable time frame and through reasonable effort.

Pruning Step	Lines Remaining
Initial phrase table	1074665
With only one-word phrases	160931
Without high frequency words	67622
Without single remaining entries	48522

Table 4.1: Remaining Lines in Pruned Phrase Table

Therefore an additional list was created containing the high frequency words for the target language. This list was created in the same way as the previous one for the source language which was described in section 4.1.2. Now a new, pruned phrase table was created, but this time taking into account both high frequence sets, removing all one-word-phrases that contain either a high frequency word of the source or target language. Using the high frequency words for both the source and the target language finally helped to shorten the phrase table to a somewhat reasonable size.
Of course,all the before mentioned pruning steps can always introduce new problems because it can happen any time that vital information is pruned away. Problems that arose from the pruning steps and other limitations will be described in detail in the evaluation chapter 5. Table 4.1 illustrates the number of lines that remained in the phrase table after each pruning step. However, in the implementation the first and third pruning step were combined in one step.

4.1.4 Ambiguity Files

This step is the last one in the preprocessing part of the project. Here, a file is created containing all the words considered to be an ambiguity, due to all the heuristics that were already mentioned before and an additional one that will be described in this subsection. Together with the possible ambiguities all different possible translations belonging to it are stored in that file.
A first approach had been to present only the information contained in that file to the user and ask him to resolve the ambiguous words. But this task proofed not to be feasible. Situations did arise in which the user did not know how to resolve the current word because there was no information about the context the word was used in, which made a decision rather difficult. In other cases the user had to look really carefully at the different translation possibilites to make out differences and finally make the right choice. The whole resolving process turned out to be very exhausting. Therefore, an additional file, containing an extended three word context around the ambiguity and the corresponding phrases in the target language, is now created in this step. This file is later needed to aid the presentation of ambiguous words to the user.
But at first the two files have to be created. As has already been stated before, the pruned phrase table is the most important source of information that is needed for this task. Additionally, the developement set, which will be called dev set for the rest of this thesis, is needed again, to reduce the final number of ambiguities. As has been mentioned above there was another heuristic, designed to reduce to number of possible translations for one ambiguity. The heuristic is based on the *phrase table scores* of the distinct phrases.
As has been explained in section 2.2 there are five scores for each entry in the phrase

table. But because all those scores are not weighted, they can not be used directly as a basis for comparing different translations. The real phrase score, that would also be used by the translation system during the translation process, has to be calculated from the *phrase table scores* and their associated translation model weights λ_i using formular 4.1.

$$\begin{aligned} score\ (f,e) = -\lambda_1 \cdot log(\varphi(f \mid e)) - \lambda_2 \cdot log(lex(e|f)) - \\ \lambda_3 \cdot log(\varphi(e \mid f)) - \lambda_4 \cdot log(lex(f|e)) \end{aligned} \tag{4.1}$$

The weights λ_i are optimized during *Minimum Error Rate Training (MERT)* and used to prioritize the different scores. With the calculated phrase score at hand a word is considered ambiguous if there is more than one possible translation left after applying heuristic 4.2.

$$score\ (possible\ translation) < best \cdot beamsize \tag{4.2}$$

In this inequation *best* refers to the best score of all possible translations for one ambiguous word. *Beamsize* allows for varying the interval in which the scores of possible translations are supposed to reside. If *beamsize* is big, more possible translations are going to have a score that is smaller than *best · beamsize*. If it is set to a smaller value, the number of possible translations fullfilling inequation 4.2 will also be smaller.

In this experiment that number should neither be too big nor too small. Tests, calculating the total number of ambiguities that could be found with a specific value of *beamsize*, showed that beamsize 6 gives a pretty good median as can be seen in figure 4.1.

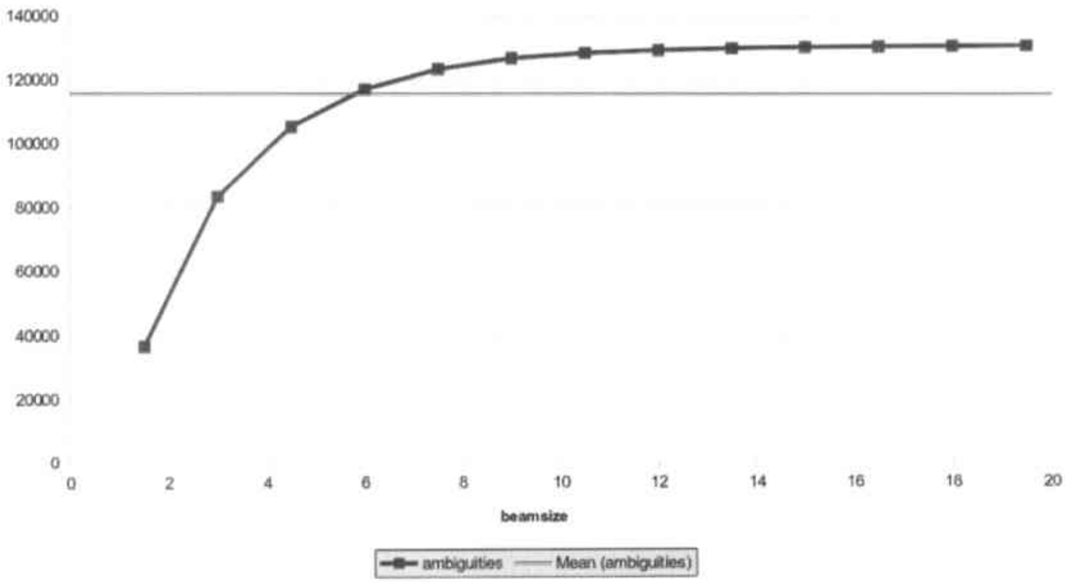


Figure 4.1: Total number of ambiguities for different values of *beamsize*

For values of *beamsize* that are bigger than 10 the total number of found ambiguities does not change except for a few. If the value becomes smaller than 4.5 the number gets considerably smaller. The red line in the diagramm marks the mean value.

Applying all the before mentioned steps and heuristics finally results in a file holding all found ambguities and all their possible translations.
An entry in this ambiguity file looks as follows:

```
ambiguity ||| translation 1 ||| translation 2 ||| ...
```

Finally the file containing the extended three-word context for each ambiguity and the corresponding translations has to be created. It is called *ambisen* for the rest of this thesis. The extended context is supposed to simplify the decision making process for the user giving him additional information about the ambiguous words and their possible translations. With the help of an alignment file the matching source sentence for every possible translation of an ambiguity is obtained and both are put into the *ambisen* file. The alignment file contains information about the alignments for every word in the source and target corpus in the form of `i-j`. The variables `i` and `j` describe the position of a word in a sentence which means that the notation `i-j` refers to the `i-th` source word beeing aligned to the `j-th` target word. An entry in the *ambisen* file would look like:

```
ambiguity ||| source sentence 1 *** target sentence 1 |||...
```

In order to keep the length of the source and target sentences reasonable and comparable, not whole sentences are used to make up the file but rather only an n-word context in front of and after the ambiguous word if there were enough words available. So if the ambiguity would be the *i-th* word in the sentence and a three-word context would be used, a source sentence would look like:

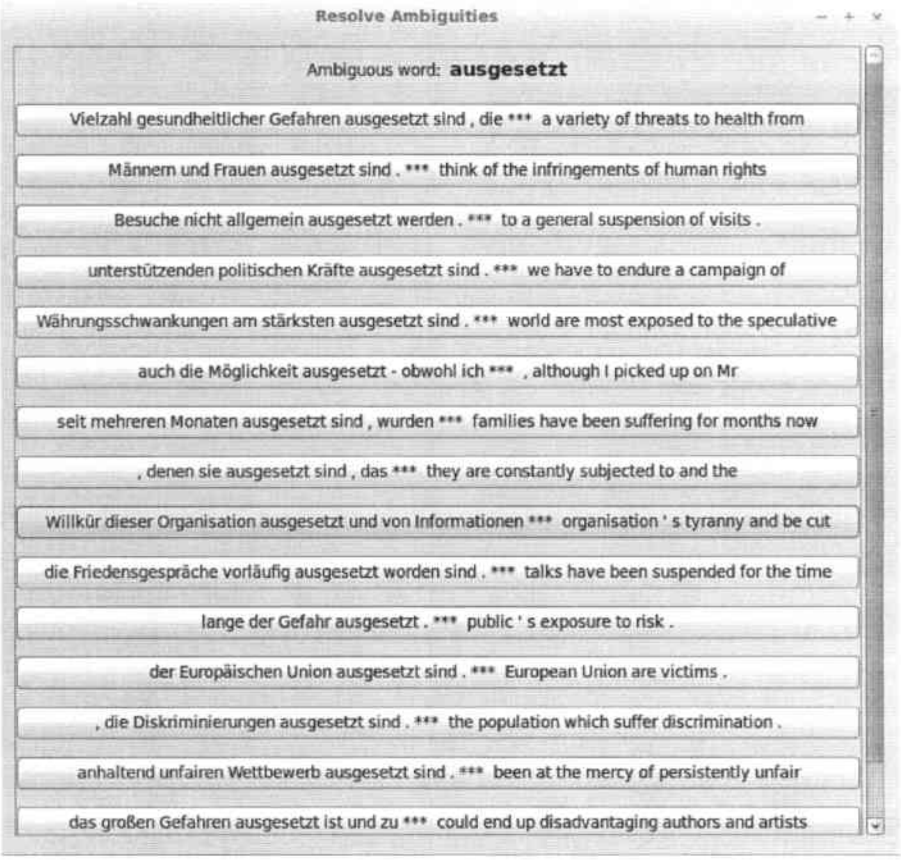
$\texttt{word}_{i-3}\ \texttt{word}_{i-2}\ \texttt{word}_{i-1}\ \ \texttt{ambiguity}\ \ \texttt{word}_{i+1}\ \texttt{word}_{i+2}\ \texttt{word}_{i+3}$

With the preparation of these two files the identification of ambiguous words in the corpus is finished. Having the *ambisen* file for additional information they can be presented to the user in the following step.

4.2 Presentation to the User

In section 4.1 the steps needed to identify ambiguous words were described.
This section describes the resolution process where the identified ambiguities are presented to the user. Additionally the phrase table is annotated to save the users choice and enable a new training considering them.

The next step is to present the identified ambiguities to the user in order to get a resolution. For this purpose the two files created before (like described in section 4.1.4) are browsed to find ambiguous words and their corresponding example sentences. This is done using a Perl script named `present_ambis.pl` that, for each match, calls another script called `gtkwindow.pl`. The latter implements a GUI (Graphical User Interface) that displays the current ambiguity and a selection of example sentences the ambiguous word was found in. Figure 4.2 illustrates an example of the window shown to the user:



Ambiguous word: **ausgesetzt**

Vielzahl gesundheitlicher Gefahren ausgesetzt sind , die *** a variety of threats to health from

Männern und Frauen ausgesetzt sind . *** think of the infringements of human rights

Besuche nicht allgemein ausgesetzt werden . *** to a general suspension of visits .

unterstützenden politischen Kräfte ausgesetzt sind . *** we have to endure a campaign of

Währungsschwankungen am stärksten ausgesetzt sind . *** world are most exposed to the speculative

auch die Möglichkeit ausgesetzt - obwohl ich *** , although I picked up on Mr

seit mehreren Monaten ausgesetzt sind , wurden *** families have been suffering for months now

, denen sie ausgesetzt sind , das *** they are constantly subjected to and the

Willkür dieser Organisation ausgesetzt und von Informationen *** organisation ' s tyranny and be cut

die Friedensgespräche vorläufig ausgesetzt worden sind . *** talks have been suspended for the time

lange der Gefahr ausgesetzt . *** public ' s exposure to risk .

der Europäischen Union ausgesetzt sind . *** European Union are victims .

, die Diskriminierungen ausgesetzt sind . *** the population which suffer discrimination .

anhaltend unfairen Wettbewerb ausgesetzt sind . *** been at the mercy of persistently unfair

das großen Gefahren ausgesetzt ist und zu *** could end up disadvantaging authors and artists

Figure 4.2: GUI illustrating ambiguities

The interface between the two modules `present_ambis.pl` and `gtkwindow.pl` is implemented as a JSON object.
JSON[1] is a lightweight data interchange format. It is text-based and language-independant but uses some conventions familiar from languages of the C-family. Because of its simple structure it is easy for humans to read and write. The integration into the perl scripts is done using the **CPAN** [2] module JSON.
The modular approach allows for the use of arbitrary executables that provide a graphical user interface. In this research project **Gtk2-Perl** has been chosen as a reference implementation.
Using a GUI it is easier to arrange the multiple choices clearly, and the user can just use a simple click to choose the best translation. After the user chose his favorite translation by clicking on the corresponding sentence pair, the target word corresponding to that sentence pair is transferred back to the calling script `present_ambis.pl`, where it is stored until every ambiguity has been be resolved.

Resolving the last ambiguity automatically initiates the annotation of the phrase table. A sixth score that mirrors the users choice is added to the five existing ones. The annotation is not done on the pruned phrase table that was created carrying out the steps of section 4.1.3, but on the initial phrase table from the beginning.

[1] (JavaScript Object Notation), http://www.json.org/

[2] (Comprehensive Perl Archive Network), http://www.cpan.org/

In this phrase table each entry is annotated with a sixth score. There are basically three rules how an entry can be adapted according to the users choices:

- If the source phrase of the current phrase table entry contains a word that has been identified as ambiguous and the target phrase of that entry contains the translation the user chose for this word, then the entry is annotated with a score of 1.0. This value is defined as a score indicating a phrase pair that represents a good translation of each other.
- If the source phrase contains a word that has been identified as ambiguous but the target phrase does not contain the translation the user chose but another one, the additional sixth score is set to 0.1. This value is supposed to represent phrase pairs that do not match as well.
- The phrase table entries that do not contain any ambiguous words at all are also annotated with a score of 1.0 because they are not considered in this research project and should not be ranked badly.

In this thesis, no deeper analysis has been conducted to clarify wether the values 1.0 and 0.1 are optimal weights or not. This exposes an area of possible future work.

At last the `moses.ini` file has to be adapted to the sixth score, too. This file is used by the translation system for decoding and contains information such as where language model files and the phrase table can be found. It also contains the number of scores the translation system has been trained with, so in order to adapt the system to the new sixth score the `moses.ini` file has to be changed. Additionally, a sixth weight has to be added to the `[weight-t]` section of the `moses.ini` file. In that section the weights for the translation model 4.1.4 are stored.
After finishing the adaptation of the system to the sixth score it can now be retrained to learn a better translation for the ambiguous word through the choices of the user.

5. Evaluation

Identifying and resolving the ambiguous words in a given text, several problems and difficulties occured. Some of them were already expected at the beginning of this project and in fact proved to be of significance during the later progress. Others were not foreseen and only became apparent during the experiments. Together with the results of the approach described in the previous chapter they will be presented in the following paragraphs.

Limiting possible sources of information to resources already contained in the baseline system probably caused most of the problems. Due to the lack of additional information, words might, for example, be considered ambiguous by the system that would probably never be identified as such in a face-to-face conversation between humans. As a consequence the user has to resolve words that would normally not pose a problem.
For example, words that are in fact only conjugations of one another might be presented as completely different ambiguities. In such cases the user has to deal with words that share the same root word and most of the time also the same translation possibilities. In the end the resolution of conjugations does oftentimes not provide much useful information and is unnecessary and additionally time-consuming. Table 5.1 provides an example.

Another problem arising from a lack of information is the identification of proper names or names of a person. Those should, of course, not be translated or be used

Ambiguouity	**Shared Translations**			**Additional ones**
rasch	swiftly, swift, quick, rapid, rapidly	promptly	-	speedily, speeding, urgent, delay, briefly
rasche			speedy	sudden, trialogue, promt
raschen		-		reality, pace, introduction

Table 5.1: Conjugations

Ambiguous Word	Translations from the SMT system	Correct Translation(s)
sechs	fabric, seven, informed	six
Nein	nipping, Rossa, contrary, opposed, instead	no, nope
Regie	aegis, remit	administration, state direction
Potential	offers, bet, engine	potential
Mitgliedstaats	payments, contracting	member state

Table 5.2: Bad alingments for a word

as a translation. But with no way to identify them, sometimes names happend to be the only translation probability available for a word:

```
Abgeordneter ||| Celli ||| Ford ||| Gallagher ||| Graefe ...
```

Ambiguities for which no good translations were provided by the translation system posed another difficulty. The user has to take his time reading every possible translation only to find that none of them is suitable. The reason for that is not only sticking to a baseline configuration, of course, but also bad alignments produced from the translation system in the first place. Some examples for this kind of problem can be found in table 5.2.

The limitation to only one-word phrases made a first approach more simple than it would have been if also two or even n-word phrases would have been considered during the identifying steps 4.1.3 and 4.1.4. But some drawbacks do arise from this decision, especially regarding the German language which was one part of the language pair used in this research project. In the German language some words are in fact combinations of other words. Those are the so called *compounds*. Languages that do not have compounds tend to split them up and just translate each word of the compound as if it was a single word. With the limitation to one-word phrases it is not possible to cover all the words that would normally belong to a correct translation of a compound. The user can most of the time only choose one of the words as a translation. This chosen word is later, in the phrase table annotation step 4.2 rated higher than the other words that also belong to a correct translation. In the worst case this might even decrease the overall result of the translation system. Table 5.3 illustrates this problem.

But despite all problems described above, the modified translation system proved to have a slightly improved BLEU score compared to the baseline system which is illustrated in table 5.5. Examples for improved translation results due to resolved ambiguities can be found in table 5.7. In the translations of some sentences even better language model results can be seen. Table 5.6 shows some examples. This already is a promising result. With even more sophisticated modifications the translation results would probably become even better in the future. There are many possible modifications that might be able to improve the system presented in this thesis. Some of them will be described in the following section 5.1.

Ambiguous Word	Translations from the SMT system	Correct Translation
Strukturfonds	Structural, Funds	structural funds
Wahlkampf	electoral, campaign	election campaign
Plenarsitzung	session, turned, plenary	plenary session
Krisensituation	emergency, crises, situations	crisis/critical situation
Gleichbehandlung	treatment, treating, equality	equal treatment
langfristig	term, entrenched	long-term

Table 5.3: Compounds, Splitted Translations

Text	#Sentences	#Words: German - English
Europarl De-En	1.9M	44.5M - 47.8M
Training	200K	5.2M - 5.5M
Tuning	1K	26K - 28K
Test	1K	24K - 26K

Table 5.4: Corpus Information

Moses [KHB+07][1] is used as a translation system in this thesis. It is an open-source toolkit for Statistical Machine Translation and was created to provide an easy access for researchers to a fully featured machine translation system.
There are a variety of tools included in Moses for training and tuning. One of these tools is the **GIZA++** Toolkit [ON03][2] which is used for computing word alignments in this thesis. The use of efficient data structures allows to exploit large data resources even with limited hardware. The parallel texts that are used for training, tuning and testing derive form the **Europarl Corpus** [Koe05][3]. The europarl corpus is a parallel corpus extracted from the proceedings of the European Parliament and includes 21 European languages. Although there is lots of parallel data available, only a smaller set of parallel sentences is used in this thesis to keep the problem feasible. As a test language pair German-English is used. Table 5.4 displays some information about the corpus.

Baseline System	**24.75**
Modified System	**24.90**

Table 5.5: BLEU scores for the baseline and modified translations systems

5.1 Future Work

The previous section already showed that there are several problems that have to be addressed during an ambiguity solving process. However, the translation system

[1] http://www.statmt.org/moses/

[2] http://www.statmt.org/moses/giza/GIZA++.html

[3] http://www.statmt.org/europarl/

introduced in this thesis presented improved results nonetheless.
In this section several approaches and ideas will be discussed that could further improve the translation results.

One idea is to cancel the restriction of having a translation system that is as simple as possible. If it would be possible to include additional sources such as POS tagging or identification of root words, this could improve the system a lot. A POS tagger could, for example, simplify the task of deciding which words should be presented to the user. The set of words could be restricted to only those word classes that are supposed to carry the most information. Choosing important words would provide a much finer granularity. Additionally, a stemmer could be used to identify words that share the same root and are just conjunctions of each other. As can bee seen in table 5.1 conjunctions often share similar translations. If the user already translated one of the conjunctions, it would be possible to use the same translation for the others as well. This way the amount of ambiguities that have to be resolved by the user could further be reduced.

Another modification addresses the problem of compounds and other words the translation of which is more than one word or phrases consisting of more than one word but translating to only one. Using only one-word phrases it is impossible to treat those words right because the user is always able to choose only one of the words. If there was a possibility to identify compounds and other phrases that do not fit into a one-word phrase this could also improve translation results.

One source of information that has not been used in this thesis at all is the language model. Including the information of the language model could help to reduce the amount of possible translations that have to be presented to a user. Those translation possibilities which would never be part of an actual translation because of a really low language model score could be filtered out beforehand and the user would not have to waste his time sorting them out.

Finally there are some modifications that concern mostly the presentation process itself. Some things could probably be done to make the ambiguity resolving task more concise and intuitive. If, for example, the possible translations that are presented to the user would be clustered according to the context they are used in, the user would be able to find his favourite translation much faster. It would be possible to search for the right context first and then choose one of the provided translation possibilities without even looking at the ones belonging to other contexts. For example, the German word *Hahn* might be used in an agricultural or in an industrial context. In the first one, *Hahn* refers to the animal, the second one to a tap. Seprating these two contexts would allow the user for choosing his preferred translation much faster.
It might happen that none of the translation possibilities available correspond to the conception of the user. This could be because the user is trying to translate a word in a context that has not been seen in training or due to bad alignments provided by the translation system 5.2. In such cases it would be an improvement to provide a way for the user not to choose any of the possible translations at all. In case of bad

Source Sentence	bei einer Sache kann ich allerdings der Kollegin Korhola nicht zustimmen : die Konsultation der Öffentlichkeit wird nur möglich sein , wenn man den Begriff " Öffentlichkeit " nicht zu vage definiert.
Baseline Translation	`in a matter of Mrs Korhola , however , I could not agree , the consultation of the public will only be possible if the term " public ' not too vague defined.`
New Translation	`in a matter , however , I can do not agree with Mrs Korhola , the consultation of the public will only be possible if we do not the concept of ' public too vague .`
Source Sentence	wir werden der Herausforderung nur dann gewachsen sein , wenn wir wirklich alles Erdenkliche tun , um die Krise zu bewältigen .
Baseline Translation	`we will be up to the challenge only then , if we really do all we can , in order to manage the crisis.`
New Translation	`we will only be up to the challenge , if we really do all we can , in order to manage the crisis.`
Source Sentence	das ist nicht nur für das irische Volk problematisch : es ist auch ein Problem für die Einwohner Großbritanniens und der Europäischen Union überhaupt.
Baseline Translation	`this is not only for the Irish people of a problem , it is also a problem for the inhabitants of the United Kingdom and the European Union.`
New Translation	`this is not only a problem for the Irish people : it is also a problem for the inhabitants of the United Kingdom and the European Union.`

Table 5.6: Improved Language Model

alignments all suggested translation could be annotated with a low score if the user decided to choose none of them. If there exists a better translation that happend to be pruned away before, its score could be improved towards the score of the others. However, an even better solution would be to provide the user with the possibility to place his own translation if none of the suggested are fitting. That way the user could actively help to improve translation quality by providing new possible translations for some of the words.

Source Sentence	. . . hat die britische Regierung einer Erweiterung des Betriebs dort zugestimmt .
Baseline Translation	`...the British Government has an extension of the operation there .`
New Translation	`...the British Government has accepted a enlargement of the operation.`
Source Sentence	mein Vorschlag an die Mitgliedstaaten daher : ersetzen Sie die unnötige bürokratische Kombination von Verwaltungs- und Regelungsverfahren durch eine Kombination aus Verwaltungs- und Beratungsverfahren .
Baseline Translation	`my proposal to the Member States therefore : you have the unnecessary ...`
New Translation	`my proposal to the Member States : you , therefore , to replace the unnecessary...`
Source Sentence	die internationalen Finanzaktivitäten auf den Devisenmärkten haben sich nämlich auf rund 1.200 Milliarden Dollar täglich verringert , und das liegt an der Schaffung der Währungsunion . . .
Baseline Translation	`the international Finanzaktivitäten to the Devisenmärkten , namely to approximately 1.200 billion dollars a day , and it is the result of the creation of the monetary union ...`
New Translation	`the international Finanzaktivitäten on the Devisenmärkten have to around 1.200 billion dollars a day , and it is reduced to the creation of the monetary union ...`

Table 5.7: Improved Translation Results

6. Summary and Conclusions

The goal of this study thesis has been to aid statistical machine translation by resolving ambiguities in interaction with a human user.
The previous chapters have presented several heuristics and approaches to achieve this goal. These were implemented and integrated in a existing machine translation system. With the help of a user the ambiguities of a text were resolved using a GUI that was also implemented in the context of this thesis.
The interface between the GUI and the data processing module has been kept simple to provide the possibility of changing the GUI.

With all implemented changes it is possible to achieve an improvement in BLEU score compared to the baseline system that has been used. These results could be improved even more with the implementation of the changes presented in chapter 5.1.

Bibliography

[BLKM95] Hervé Blanchon, Kyung-Ho Loken-Kim, and Tsuyoshi Morimoto. An interactive disambiguation module for english natural language utterances. In *Proceedings of NLPRS*, 1995.

[Com66] National Research Council (U.S.). Automatic Language Processing Advisory Committee. *Language and machines: computers in translation and linguistics; a report.* Publication (National Research Council (U.S.))). National Academy of Sciences, National Research Council, 1966.

[KHB$^+$07] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[Koe05] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT.

[OMC$^+$05] Constantin Orasan, Ted Marshall, Robert Clark, Le An Ha, and Ruslan Mitkov. Building a wsd module within an mt system to enable interactive resolution in the user's source language. In *Proceedings of EAMT*, 2005.

[ON03] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[SRS$^+$06] Marcus Sammer, Kobi Reiter, Stephen Soderland, Katrin Kirchhoff, and Oren Etzioni. Ambiguity reduction for machine translation: Human-computer collaboration. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas*, 2006.

[YKI+98] Masaya Yamaguchi, Takeyuki Kojima, Nobuo Inui, Yoshiyuki Kotani, and Hirohiko Nisimura. Combination of an automatic and an interactive disambiguation method. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 1423–1427, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.