Toward hierarchical models for statistical machine translation of inflected languages

Sonja Nießen and **Hermann Ney**

Lehrstuhl für Informatik VI,
Computer Science Department
RWTH Aachen - University of Technology
D-52056 Aachen, Germany
{niessen,ney}@informatik.rwth-aachen.de

Abstract

In 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. Existing statistical systems for MT often treat different derivatives of the same lemma as if they were independent of each other. In this paper we argue that a better exploitation of the bilingual training data can be achieved by explicitly taking into account the interdependencies of the different derivatives. We do this along two directions: Usage of hierarchical lexicon models and the introduction of equivalence classes in order to ignore information not relevant for the translation task. The improvement of the translation results is demonstrated on a German-English corpus.

1 Introduction

The statistical approach to machine translation has become widely accepted in the last few years. It has been successfully applied to realistic tasks in various national and international research programs. However in many applications only small amounts of bilingual training data are available for the desired domain and language pair, and it is highly desirable to avoid at least parts of the costly data collection process.

Some recent publications have dealt with the problem of translation with scarce resources. (Brown et al., 1994) describe the use of dictionaries. (Al-Onaizan et al., 2000) report on an experiment of Tetun-to-English translation by different groups, including one using statistical machine translation. They 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 of the data.

In this paper, we take 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 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 languages like German.

We address the question of how to achieve a better exploitation of the resources for training the parameters for statistical machine translation by taking into account explicit knowledge about the languages under consideration. In our approach we introduce equivalence classes in order to ignore information not relevant to the translation process. We furthermore suggest the use of hierarchical lexicon models.

The paper is organized as follows. After reviewing the statistical approach to machine translation, we first explain our motivation for examining the morphological characteristics of an inflected language like German. We then describe the chosen output representation after the analysis and present our approach for exploiting the information from morpho-syntactic analysis. Experimental results on the German-English Verbmobil task are reported.

2 Statistical Machine Translation

The goal of the translation process in statistical machine translation can be formulated as follows: A source language string $f_1^J = f_1 \ldots f_J$ is to be translated into a target language string $e_1^I = e_1 \ldots e_I$. In the experiments reported in this paper, the source language is German and the target language is English. Every English string is considered as a possible translation for the input. If we assign a probability $Pr(e_1^I|f_1^J)$ to each pair of strings (e_1^I, f_1^J), then according to Bayes' decision rule, we have to choose the English string that maximizes the product of the English language model $Pr(e_1^I)$ and the string translation model $Pr(f_1^J|e_1^I)$.

Many existing systems for statistical machine translation (Wang and Waibel, 1997; Nießen et al., 1998; Och and Weber, 1998) make use of a special way of structuring the string translation model like proposed by (Brown et al., 1993): The correspondence between the words in the source and the target string is described by alignments which assign one target word position to each source word position. The lexicon probability $p(f|e)$ of a certain English word e is assumed to depend basically only on the source word f aligned to it.

The overall architecture of the statistical translation approach is depicted in Figure 1. In this figure we already anticipate the fact that we can transform the source strings in a certain manner.

3 Basic Considerations

The parameters of the statistical knowledge sources mentioned above are trained on bilingual corpora. In general, the resulting probabilistic lexica contain all word forms occurring in this training corpora as separate entries, not taking into account whether or not they are derivatives of the same lemma. Bearing in mind that 40% of the word forms have only been seen once in training (see Table 2), 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 lexica. As Table 2 shows, this problem is especially relevant for highly inflected languages like German: Texts in German contain many more different word forms than their English translations. The table also reveals that these words are often derived from a much smaller set of base forms ("lemmata"), and when we look at the number of different lemmata and the respective number of lemmata, for which there is only one occurrence in the training data, German and English texts are more resembling.

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

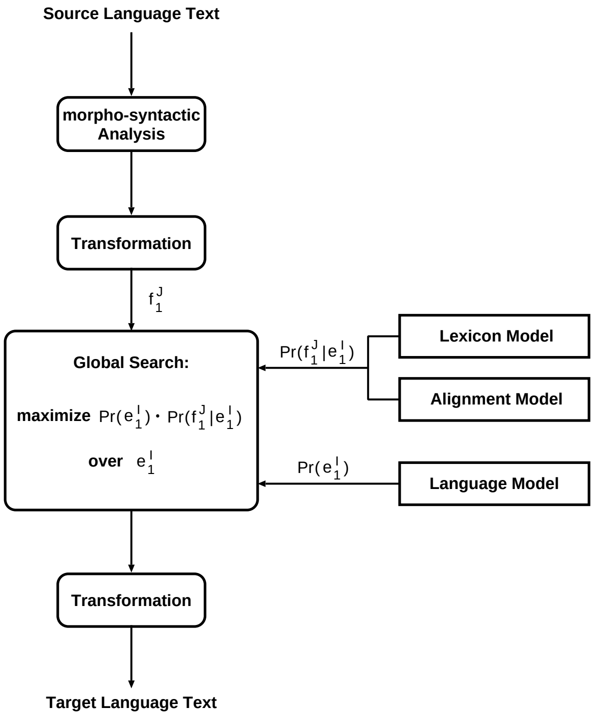

Figure 1: Architecture of the translation approach based on Bayes' decision rule.

because they are substantially different from full bilingual parallel corpora inasmuch 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 language is an interesting objective.

As a consequence of these considerations, we aim at taking into account the interdependencies between the different derivatives of the same base form.

4 Output Representation after Morpho-syntactic Analysis

We use GERCG, a constraint grammar parser for German for lexical analysis and morphological and syntactic disambiguation. For a description of the Constraint Grammar approach we refer the reader to (Karlsson, 1990). Figure 2 gives an example of the information provided by this tool.

```
Input:    Wir wollen nach dem Essen
nach Essen aufbrechen
"<*wir>"
    "wir"    * PRON PERS PL1 NOM
"<wollen>"
    "wollen" V IND PRÄS PL1
"<nach>"
    "nach"   pre PRÄP Dat
"<dem>"
    "das"    ART DEF SG DAT NEUTR
"<*essen>"
    "*essen" S NEUTR SG DAT
"<nach>"
    "nach"   pre PRÄP Dat
"<*essen>"
    "*essen" S EIGEN NEUTR SG DAT
    "*esse"  S FEM PL DAT
    "*essen" S NEUTR PL DAT
    "*essen" S NEUTR SG DAT
"<aufbrechen>"
    "aufbrechen" V INF
```

Figure 2: Sample analysis of a German sentence

A full word form is represented by the information provided by the morpho-syntactic analysis: From the interpretation "`gehen-V-IND-PRÄS-SG1`", i.e. the lemma plus part of speech plus the other tags the word form "`gehe`" can be restored. From Figure 2 we see that the tool can quite reliably disambiguate between different readings: It infers for instance 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 not from a verb, but from an adjective with the meaning "made of wool". In this sense, the information inherent to the original word forms is augmented by the disambiguating analyzer. This can be useful for deriving the correct translation of ambiguous words.

In the rare cases where the tools returned more than one reading, it is often possible to apply simple heuristics based on domain specific preference rules or to use a more general, non-ambiguous analysis.

The new representation of the corpus where full word forms are replaced by lemma plus morphological and syntactic tags makes it possible to gradually reduce the information: For example we can consider certain instances of words as equivalent. We have used this fact to better exploit the bilingual training data along two directions: Omitting unimportant information and using hierarchical translation models.

5 Equivalence classes of words with similar Translations

Inflected forms of words in the input language 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 bilingual German-English corpora, the German part contains many more different word forms than the English part (see Table 2). 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, because then, the resulting statistical translation lexica become smoother and the coverage is significantly improved. We construct these equivalence classes by omitting those informations from morpho-syntactic analysis, which are not relevant for the translation task.

The representation of the corpus like it is provided by the analyzing tools 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 find a method for automating this decision process. We found that the impact on the end result due to different choices of features to be ignored was not large enough to serve as reliable criterion. Instead, we could think of defining a likelihood criterion on a held-out corpus for this purpose. Another possibility is to assess the impact on the alignment quality after training, which can be evaluated automatically (Langlais et al., 1998; Och and Ney, 2000), but as we found that the alignment quality on the Verbmobil data is consistently very high, and extremely robust against manipulation of the training data, we abandoned this approach.

We resorted to detecting candidates from the probabilistic lexica trained for translation from German to English. For this, we focussed on those derivatives of the same base form, which resulted in the same translation. For each set of tags, we counted how often an additional tag could be replaced by a certain other tag without effect on the translation. Table 1 gives some of the most frequently identified candidates to be ignored while translating: The gender of nouns is irrelevant for their translation (which is straightforward, because the gender is unambiguous for a certain noun) and the case, i.e. nominative, dative, accusative. For the genitive forms, the translation in English differs. For verbs we found the candidates number and person. That is, the translation of the first person singular form of a verb is often the same as the translation of the third person plural form, for example.

Table 1: Candidates for equivalence classes.

POS	candidates
noun	gender: MASK,FEM,NEUTR
	and case: NOM,DAT,AKK
verb	number: SG,PL
	and person: 1,2,3
adjective	gender, case and number
number	case

As a consequence, we dropped those tags, which were most often identified as irrelevant for translation from German to English.

6 Hierarchical Models

One way of taking into account the interdependencies of different derivatives of the same base form is to introduce equivalence classes c_i at various levels of abstraction starting with the inflected form and ending with the lemma.

Consider, for example, the German verb form f = `"ankomme"`, which is derived from the lemma `"ankommen"` and which can be translated into English by e = `"arrive"`. The hierarchy of equivalence classes is as follows:

$$
\begin{aligned}
c_n &= \texttt{"ankommen-V-IND-PRÄS-SG1"} \\
c_{n-1} &= \texttt{"ankommen-V-IND-PRÄS-SG"} \\
c_{n-2} &= \texttt{"ankommen-V-IND-PRÄS"} \\
&\vdots \\
c_0 &= \texttt{"ankommen"} .
\end{aligned}
$$

n is the maximal number of morpho-syntactic tags. c_{n-1} contains the forms `"ankomme"`, `"ankommst"` and `"ankommt"`; in c_{n-2} the number (`SG` or `PL`) is ignored and so on. The largest equivalence class contains all derivatives of the infinitive `"ankommen"`.

We can now define the lexicon probability of a word f to be translated by e with respect to the level i:

$$p_i(f|e) = \sum_{[t_0^i]} p(t_0^i|e) \cdot p(f|t_0^i, e) \; , \qquad (1)$$

where $t_0^i = t_0, \ldots, t_i$ is the representation of a word where the lemma t_0 and i additional tags are taken into account. For the example above, t_0 = `"ankommen"`, t_1 = `"V"`, and so on.

$p(f|t_0^i, e)$ is the probability of f for a given t_0^i. We make the assumption that this probability does not depend on e. $p(f|t_0^n)$ is always assumed to be 1. In other words, the inflected form can nonambiguously be derived from the full interpretation.

$p(t_0^i|e)$ is the probability of the translation for e to belong to the equivalence class c_i. The sum over $[t_0^i]$ amounts to summing up over all possible readings of f.[1]

We combine the p_i by means of linear interpolation:

$$p(f|e) = \lambda_0 p_0(f|e) + \ldots + \lambda_n p_n(f|e) \ . \quad (2)$$

7 Translation Experiments

Experiments were carried out on Verbmobil data, which consists of spontaneously spoken dialogs in the appointment scheduling domain (Wahlster, 1993). German source sentences are translated into English.

7.1 Treatment of Ambiguity

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, we have implemented preference rules. For instance, we assume that the corpus is correctly true-case-converted beforehand and as a consequence, we drop non-noun interpretations of uppercase words. Besides, we prefer indicative verb readings instead of subjunctive or imperative. For the remaining ambiguities, we resort to the unambiguous parts of the readings, i.e. we drop all tags causing mixed interpretations.

There are some special problems with the analysis of external lexica, which do not provide enough context to enable efficient disambiguation. We are currently implementing methods for handling this special situation.

It can be argued that it would be more elegant to leave the decision between different readings, for instance, to the overall decision process in search. We plan this integration for the future.

7.2 Performance Measures

We use the following evaluation criteria (Nießen et al., 2000):

- 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 error count is not increased, even if there are slight syntactical errors, which do not seriously deteriorate the intelligibility.

[1] The probability functions are defined to return zero for impossible interpretations of f.

7.3 Translation Results

The training set consists of 58 322 sentence pairs. Table 2 summarizes the characteristics of the training corpus used for training the parameters of Model 4 proposed in (Brown et al., 1993). Testing

Table 2: Corpus statistics: Verbmobil training. Singletons are types occurring only once in training.

	English	German
no. of running words	550 213	519 790
no. of word forms	4 670	7 940
no. of singletons	1 696	3 452
singletons [%]	36	43
no. of lemmata	3 875	3 476
no. of singletons	1 322	1 457

was carried out on 200 sentences not contained in the training data. For a detailed statistics see Table 3.

Table 3: Statistics of the Verbmobil test corpus for German-to-English translation. Unknowns are word forms not contained in the training corpus.

no. of sentences	200
no. of running words	2 055
no. of word forms	385
no. of unknown word forms	25

We used a translation system called "single-word based approach" described in (Tillmann and Ney, 2000) and compared to other approaches in (Ney et al., 2000).

7.3.1 Lexicon Combination

So far we have performed experiments with hierarchical lexica, where two levels are combined, i.e. n in Equation (2) is set to 1. λ_0 and λ_1 are set to 0.5 and $p(f|t_0)$ is modeled as a uniform distribution over all derivations of the lemma 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 (1) is not needed for a two-level lexicon combination of full word forms and lemmata.

As the results summarized in Table 4 show, the combined lexicon outperforms the conventional one-level lexicon. As expected, 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.

Examples taken from the test set are given in Figure 3. Smoothing the lexicon entries over the derivatives of the same lemma enables the translation of “sind” by “would” instead of “are”. The smoothed lexicon contains the translation “convenient” for any derivative of “bequem”. The comparative “more convenient” would be the completely correct translation.

7.3.2 Equivalence classes

As already mentioned, we resorted to choosing one single reading for each word by applying some heuristics (see Section 7.1). For the normal training corpora, unlike additional external dictionaries, this is not critical because they contain predominantly full sentences which provide enough context for an efficient disambiguation. Currently, we are working on the problem of analyzing the entries in conventional dictionaries, but for the time being, experiments for equivalence classes have been carried out using only bilingual corpora for estimating the model parameters.

Table 5 shows the effect of the introduction of equivalence classes. The information from the morpho-syntactic analyzer (stems plus tags like described in Section 4) is reduced by dropping unimportant information like described in Section 5. Both error metrics could be decreased in comparison to the usage of the original corpus with inflected word forms. A reduction of 3.3% of the information item semantic error rate shows that more of the intended meaning could be found in the produced translations.

Table 5: Effect of the introduction of equivalence classes. For the baseline we used the original inflected word forms.

	SSER [%]	ISER [%]
inflected words	37.4	26.8
equivalence classes	35.9	23.5

The first two examples in Figure 4 demonstrate 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 third example shows the advantage of grouping words in 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.

8 Conclusion and Future Work

We have presented methods for a better exploitation of the bilingual training data for statistical machine translation by explicitly taking into account the interdependencies of the different derivatives of the same base form. We suggest the usage of hierarchical models as well as an alternative representation of the data in combination with the identification and omission of information not relevant for the translation task.

First experiments prove their general applicability to realistic tasks such as spontaneously spoken dialogs. We expect the described methods to yield more improvement of the translation quality for cases where much smaller amounts of training data are available.

As there is a large overlap between the modeled events in the combined probabilistic models, we assume that log-linear combination would result in more improvement of the translation quality than the combination by linear interpolation does. We will investigate this in the future. We also plan to integrate the decision regarding the choice of readings into the search process.

Table 4: Effect of two-level lexicon combination. For the baseline we used the conventional one-level full form lexicon.

	ext. dictionary	SSER [%]	ISER [%]
baseline	yes	35.7	23.9
combined	yes	33.8	22.3
baseline	no	37.4	26.8
combined	no	36.9	25.8

input	`sind Sie mit einem Doppelzimmer einverstanden?`
baseline	`are you agree with a double room?`
combined lexica	`would you agree with a double room?`
input	`mit dem Zug ist es bequemer`
baseline	`by train it is UNKNOWN-bequemer`
combined lexica	`by train it is convenient`

Figure 3: Examples for the effect of the combined lexica.

Acknowledgement. This work was partly supported by the German Federal Ministry of Education, Science, Research and Technology under the Contract Number 01 IV 701 T4 (VERBMOBIL).

References

Yaser Al-Onaizan, Ulrich Germann, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Daniel Marcu, and Kenji Yamada. 2000. Translating with scarce resources. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI)*, pages 672–678, Austin, Texas, August.

P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. Mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

P. F. Brown, S. A. Della Pietra, and M. J. Della Pietra, V. J. a nd Goldsmith. 1994. But dictionaries are data too. In *Proc. ARPA Human Language Technology Workshop '93*, pages 202–205, Princeton, NJ, March. distributed as *Human Language Technology* by San Mateo, CA: Morgan Kaufmann Publishers.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *Proc. COLING 1990: The 13th Int. Conf. on Computational Linguistics*, volume 3, pages 168–173, Helsinki, Finmark.

Philippe Langlais, Michel Simard, and Jean Véronis. 1998. Methods and practical issues in evaluating alignment techniques. In *Proc. COLING-ACL 1998: The 36th Annual Meeting of the Association for Computational Linguistics and the 17th Int. Conf. on Computational Linguistics*, pages 711–717, Montréal, P.Q., Canada, August.

Hermann Ney, Sonja Nießen, Franz Josef Och, Hassan Sawaf, Christoph Tillmann, and Stephan Vogel. 2000. Algorithms for statistical translation of spoken language. *IEEE Transactions on Speech and Audio Processing*, 8(1):24–36, January.

Sonja Nießen, Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1998. 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*, pages 960–967, Montréal, P.Q., Canada, August.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proc. LREC 2000: The 2nd Int. Conf. on Language Resources and Evaluation*, pages 39–45, Athens, Greece, May.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proc. ACL 2000: The 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hongkong, China, October.

Franz Josef Och and Hans Weber. 1998. 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 Com-*

ich reserviere das Hotelzimmer
⇓
I will reserve that hotel rooms

ich-PRON-PERS-SG1-*NOM* reservieren-V-IND-PRÄS-SG*1*
das-ART-DEF-SG-*AKK-NEUTR* Hotelzimmer-S-*NEUTR*-SG-*AKK*
↓
ich-PRON-PERS-SG1 reservieren-V-IND-PRÄS-SG das-ART-DEF-SG Hotelzimmer-S-SG
⇓
I will reserve the hotel room

fliegen wir?
⇓
we flying?

fliegen-V-IND-PRÄS-PL*1* wir-PRON-PERS-PL*1-NOM*?
↓
fliegen-V-IND-PRÄS-PL wir-PRON-PERS-PL?
⇓
do we fly?

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

geben-V-IND-PRÄS-SG*3* es-PRON-PERS-SG3-*NOM-NEUTR*
nichts-DET-INDEF-NEG-SG-*AKK* billig-A-KOMP-*SG-NOM/AKK-NEUTR*?
↓
geben-V-IND-PRÄS-SG es-PRON-PERS-SG3 nichts-DET-INDEF-NEG-SG billig-A-KOMP?
⇓
there is nothing cheaper?

Figure 4: 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.

putational Linguistics, pages 985–989, Montréal, P.Q., Canada, August.

Christoph Tillmann and Hermann Ney. 2000. Word re-ordering and DP-based search in statistical machine translation. In *Proc. COLING 2000: The 18th Int. Conf. on Computational Linguistics*, pages 850–856, Saarbrücken, Germany, August.

Wolfgang Wahlster. 1993. Verbmobil: Translation of Face-to-Face Dialogs. In *Proceedings of the MT Summit IV*, pages 127–135, Kobe, Japan.

Ye-Yi Wang and Alex Waibel. 1997. Decoding algorithm in statistical translation. In *Proc. ACL-EACL 1997: The 35th Annual Meeting of the Association for Computational Linguistics - joint with EACL 1997*, pages 366–372, Madrid, Spain, July.