Statistical Multi-Source Translation

Franz Josef Och and **Hermann Ney**

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

Abstract

We describe methods for translating a text given in multiple source languages into a single target language. The goal is to improve translation quality in applications where the ultimate goal is to translate the same document into many languages. We describe a statistical approach and two specific statistical models to deal with this problem. Our method is generally applicable as it is independent of specific models, languages or application domains. We evaluate the approach on a multilingual corpus covering all eleven official European Union languages that was collected automatically from the Internet. In various tests we show that these methods can significantly improve translation quality. As a side effect, we also compare the quality of statistical machine translation systems for many European languages in the same domain.

Keywords

statistical machine translation, multi-source translation, automated acquisition of parallel corpora

1 Introduction

In many applications for machine translation, it is necessary to translate a document into multiple languages. For example, in international organizations such as the European Union or the United Nations, all relevant documents must be translated into all official languages. Very often, the document is originally written in one language, and then translated into the other languages. If, for example, first is produced an English translation of a French document, then this translation should be used as additional knowledge source when producing a German translation. So far, existing machine translation technology is not able to make use of these additional knowledge sources.

From performing multi-source translation, we expect a better machine translation quality due to the following reasons:

- Better word sense disambiguation: Often ambiguities that need to be resolved between two languages do not exist between other languages.
- Better word reordering: A significant source of errors in statistical machine translation is the word reordering problem (Och et al., 1999). The word order between related languages is often very similar while the word order between distant languages might differ significantly. By using more source languages, we can expect that among the source languages there is one with a similar word order.
- Reduction of the need for explicit anaphora resolution: By having various translations of a pronoun in different languages the probability increases that it can be translated correctly without performing a full anaphora resolution.

Our method for performing multi-source translation fits nicely into the statistical approach and is relatively easy to implement. We are able to deduce a general statistical approach to multi-source translation. Our method is very general as it is independent of specific models, languages or application domains. Ultimately, the approach boils down to a multiplicative combination of various statistical translation models.

In principle, multi-source translation is not restricted to a statistical approach and it would be possible to pursue it also in a classical transfer-based approach. Yet, we believe that this would be significantly more complicated as already the development of transfer rules for single-source translation is nontrivial and requires experts.

The structure of the paper is as follows: Section 2 will describe the statistical approach to machine translation. Section 3 will describe a general statistical approach to multi-source translation and will introduce two specific methods for performing model combination. Section 4 will describe the collection of our large multilingual corpus with about two million words in the eleven official languages of the European Union. Section 5 will provide a discussion of the obtained results.

2 The Statistical Approach to Translation

Here, we will consider single-source translation. Multi-source translation will be described in Section 3.

The goal is the translation of a text given in some source language into a target language. We are given a source string $\mathbf{f} = f_1^J = f_1 ... f_j ... f_J$, which is to be translated into a target string $\mathbf{e} = e_1^I = e_1 ... e_i ... e_I$. In this paper, the term *word* always refers to a *full-form* word. Among all possible target strings, we will choose the string with the highest probability which is given by Bayes' decision rule (Brown et al., 1993):

$$\begin{aligned} \hat{e}_1^I &= \arg\max_{\mathbf{e}} \{Pr(\mathbf{e}|\mathbf{f})\} \\ &= \arg\max_{\mathbf{e}} \{Pr(\mathbf{e}) \cdot Pr(\mathbf{f}|\mathbf{e})\} \quad . \end{aligned}$$

$Pr(\mathbf{e})$ is the language model of the target language, whereas $Pr(\mathbf{f}|\mathbf{e})$ is the string translation model. The argmax operation denotes the search problem, i.e. the generation of the output sentence in the target language.

The overall architecture of the statistical translation approach is summarized in Fig. 1. In general, as shown in this figure, there may be additional transformations to make the translation task simpler for the algorithm. The transformations may range from the categorization of single words and word groups to more complex preprocessing steps that require some parsing of the source string. In this work, we only use tokenization, mapping words at the beginning of a sentence to their true case and categorization of numbers.

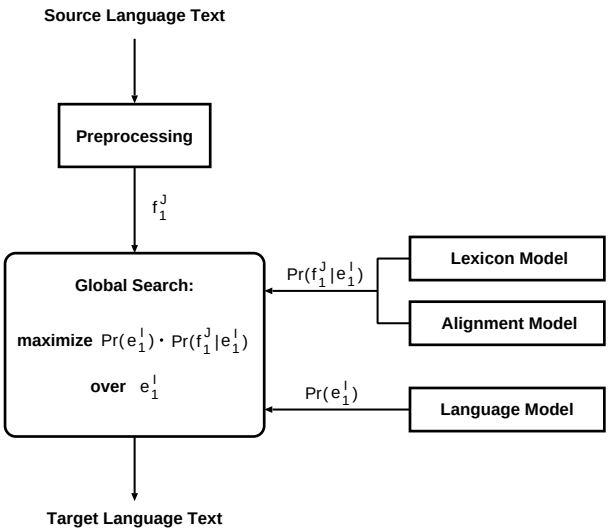

Figure 1: Architecture of the Translation Approach based on Bayes decision rule.

A key issue in modeling the string translation probability $Pr(f_1^J|e_1^I)$ is the question of how we define the correspondence between the words of the target sentence and the words of the source sentence. In typical cases, we can assume a sort of pairwise dependence by considering all word pairs (f_j, e_i) for a given sentence pair $[f_1^J; e_1^I]$. Typically, the model is further constrained by assigning each source word to *exactly one* target word. Models describing these types of dependencies are referred to as *alignment models* (Brown et al., 1993), (Vogel et al., 1996), (Och and Ney, 2000a).

Here, we use the alignment template system (Och et al., 1999). This system is based on a more general alignment model that allows also for phrase alignments. Alignment templates are pairs of phrases together with an alignment between the words within the phrases. Compared to the single-word based models the alignment templates take explicitly into account word context and local changes in word order. We observe that this approach typically produces better translations than the single-word based models. The alignment templates are automatically trained using a parallel training corpus. For more information about the alignment template approach see (Och et al., 1999).

A main advantage of the statistical approach to machine translation is its ability to learn from training data. Therefore, a statistical machine translation system can be bootstrapped very quickly given the availability of training data. For a performance comparison of the statistical approach to other machine translation approaches see (Ney et al., 2000).

3 Statistical Multi-Source Translation

The goal in multi-source translation is the translation of a text given in N source languages into a single target language. We are given N source strings $\mathbf{f}_1^N = \mathbf{f}_1, \ldots, \mathbf{f}_N$, which are to be translated into a target string $\mathbf{e}$. Among all possible target strings, we will choose the string with the highest probability:

$$\begin{aligned} \hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} \{Pr(\mathbf{e}|\mathbf{f}_1^N)\} \\ &= \arg\max_{\mathbf{e}} \{Pr(\mathbf{e}) \cdot Pr(\mathbf{f}_1^N|\mathbf{e})\} \quad . \end{aligned}$$

As in single-source translation $Pr(\mathbf{e})$ is the language model of the target language, whereas $Pr(\mathbf{f}_1^N|\mathbf{e})$ is the multi-source string translation model.

Combination method PROD

Now, we make the following assumption: Given the hypothesized target string $\mathbf{e}$, the source strings $\mathbf{f}_n$ are

considered statistically independent. Thus, we obtain:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \{p(\mathbf{e}) \cdot \prod_{n=1}^{N} p(\mathbf{f}_n|\mathbf{e})\} \quad .$$

In principle, we have to hypothesize all possible target strings to perform this maximization. Efficient search algorithms for doing this are described for example in (Och et al., 1999). Yet, the development of such a search algorithm suitable for multi-source translation is nontrivial. Therefore, as a first step, we use the approximation that for every language n the best translation $\mathbf{e}_n$ is computed by taking into account only the translation model for this language:

$$\mathbf{e}_n = \arg\max_{\mathbf{e}} \{p(\mathbf{e}) \cdot p(\mathbf{f}_n|\mathbf{e})\}, \quad n = 1, \ldots, N \quad .$$

To this purpose, we can use the standard single-source search algorithm (Och et al., 1999). In the search process for multi-source translation, we hypothesize then only these N different target sentences $\mathbf{e}_1, \ldots, \mathbf{e}_N$. Obviously, this is a severe restriction of the search space resulting in search errors. Hence, we expect that we will obtain better results if we implement a general search algorithm.

Combination method MAX

We obtain an even simpler decision rule if we perform an additional approximation by replacing the product over all languages by a maximum operation over these languages:

$$\begin{aligned} \hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} \{p(\mathbf{e}) \cdot \max_{n} p(\mathbf{f}_n|\mathbf{e})\} \\ &= \arg\max_{\mathbf{e},n} \{p(\mathbf{e}) \cdot p(\mathbf{f}_n|\mathbf{e})\} \quad . \end{aligned}$$

In other words, we simply translate using any of the N source languages. Finally, we choose the translation that obtains the best score. For this combination method, it is not necessary to develop a specific search algorithm.

In order to take into account differences in the quality of various models we can introduce in both methods a scaling factor α_n for every source language: $p(\mathbf{f}_n|\mathbf{e}) \rightarrow p(\mathbf{f}_n|\mathbf{e})^{\alpha_n}$. We could adjust these values by optimizing the error rate on held out data. Informal experiments have shown that the optimal scaling factors do not deviate much from 1. Therefore, in the experiments in this paper we do not use scaling factors.

4 Data Collection

We know of no sentence aligned multilingual corpus that is suited to evaluate the methods described here. Therefore, we collected our own corpus from the Internet. We used the *Bulletin of the European Union* which exists in the 11 official languages of the European Union and which is available on the Internet. We performed the following steps to obtain a multilingual corpus:

1. We downloaded this corpus for all eleven languages in HTML format.

2. We performed an alignment on text level by file name matching.

3. We extracted the raw text from this corpus by extracting all text segments within HTML tags. Very often, these segments correspond to paragraphs. Thereby, we obtained a sequence of text segments for every text in every language.

4. We performed a segment alignment between two languages using a dynamic programming algorithm, which optimizes a length-based heuristic as in (Gale and Church, 1993). We performed this segment alignment for the ten language pairs we were interested in. Thereby, we obtained ten bilingual corpora aligned on paragraph level.

5. We performed a sentence alignment using similar heuristics as in the paragraph alignment. Thereby, we obtained ten bilingual corpora aligned on sentence level.

6. From the resulting bilingual corpora, we filtered all sentences which seem to have wrong alignments such as alignments of very long sentences with very short sentences or alignments which have a very low probability according to the Hidden Markov alignment model.

7. For all languages, we performed a very simple preprocessing. This includes tokenization, mapping of words at the beginning of a sentence to their true case and categorization of numbers.

Table 1 shows the corpus statistics of the collected training corpora. Due to the filtering of poor alignments the numbers for English differ with respect to the considered language pair up to 10 percent.

The vocabulary sizes differ considerably between the different languages. Languages like Finnish with a very rich morphology have a very large vocabulary (of full-form words) and languages like English have a very small vocabulary.

We extracted one test corpus by finding all (English) sentences between 10 and 14 words that are available in all corpora. These sentences were removed from all

Table 1: Training corpus statistics (fr=French, es=Spanish, pt=Portuguese, it=Italian, sv=Swedish, da=Danish, nl=Dutch, de=German, el=Greek, fi=Finnish, en=English).

Lang	Sentences	Words	Voc.
fr	117K	2.32M	50462
es	120K	2.32M	50949
pt	120K	2.30M	50216
it	120K	2.21M	54986
sv	125K	2.02M	72517
da	131K	2.21M	70713
nl	121K	2.30M	58550
de	139K	2.23M	73506
el	131K	2.28M	68811
fi	120K	1.61M	106159
en		˜2.1M	˜45K

Table 2: Test corpus statistics.

Sentences	1 302
English words	15 048
Trigram perplexity	179
Bigram perplexity	286

training corpora. Table 2 shows the test corpus statistics.

5 Results

Evaluation Criteria

In all experiments, we use the following two error criteria:

- WER (word error rate):
  The WER is computed as the minimum number of substitution, insertion and deletion operations that have to be performed to convert the generated string into the target string. This performance criterion is widely used in speech recognition.

- PER (position independent word error rate):
  A shortcoming of the WER is the fact that it requires a perfect word order. As a result, the word order of the automatically generated target sentence can be different from that of the given target sentence, but nevertheless acceptable so that the WER measure alone could be misleading. In order to overcome this problem, we introduce the position independent word error rate (PER) as additional measure. This measure compares the words in the two sentences *without* considering word order.

Table 3: Training corpus perplexity of HMM alignment model and translation results.

Lang	PP	WER	PER
fr	19.1	55.3	45.3
pt	21.3	58.9	48.2
es	18.4	59.2	47.6
it	24.3	59.5	48.8
sv	24.1	60.3	49.9
da	24.3	62.7	52.9
nl	17.6	64.3	51.7
de	31.7	66.9	54.2
el	31.7	72.4	53.0
fi	44.2	83.3	66.3

Both error rates are related to the post-editing effort that a human needs to invest to correct the machine translation output.

Single-Source Translation Results

For every bilingual corpus, we trained a single-word based alignment model (Och and Ney, 2000b), performed a word alignment and trained the alignment template system (Och et al., 1999). Thereby, we obtained ten translation systems from some language to English. Table 3 shows the training corpus perplexity (PP) and word error rate (WER) and position-independent word error rate (PER) of every translation system.

Looking at the Table 3, we make the following interesting observations:

- The error rates differ significantly for the different languages. The best translation quality is obtained for French (WER: 55.3%) and Portuguese (58.9%) and the worst translation quality is obtained with German (66.9%), Greek (72.4%) and Finnish (83.3%). Obviously, the languages with a very large vocabulary size, due to the rich morphology in these languages, result in a poor translation quality, which shows the necessity of morphological processing for these languages.

- The error rates correspond to training corpus perplexity. Very often, language pairs with a high translation model perplexity also result in a high WER (exception: Dutch).

Multi-Source Translation Results

Table 4 and Table 5 show the quality improvement in WER when combining the languages French, Span-

Table 4: Absolute improvements in WER combining two languages using method MAX compared with the best WER obtained by any of the two languages.

	fr	pt	es	it	sv	da	nl
fr	0.0	1.5	1.2	0.5	2.7	1.9	0.8
pt		0.0	2.2	2.1	4.0	3.4	1.3
es			0.0	2.4	3.9	2.6	1.7
it				0.0	3.5	3.2	1.6
sv					0.0	2.7	1.7
da						0.0	4.3
nl							0.0

Table 5: Absolute improvements in WER combining two languages using method PROD compared with the best WER obtained by any of the two languages.

	fr	pt	es	it	sv	da	nl
fr	0.0	0.8	0.1	0.4	1.0	0.8	-0.2
pt		0.0	2.6	2.1	2.6	2.8	-0.1
es			0.0	2.4	3.4	3.7	1.1
it				0.0	1.9	3.0	0.3
sv					0.0	1.8	0.5
da						0.0	1.5
nl							0.0

ish, Portuguese, Swedish, Danish, and Dutch using method MAX and using method PROD. These tables show the absolute improvement to the best word error rate obtained by any of the two languages. Using MAX, we observe an improvement in word error rate between 0.5 and 4.3 percent. Using PROD, the improvement is typically lower. Interestingly, the error rates almost never increase. This shows the robustness of the approach.

Table 6 shows the translation quality obtained using MAX when combining even more languages. We chose always the next language pair that yields the largest improvement. The additional improvement by using a third language is quite small. Translation quality does not improve when more than three languages are used.

Table 7 shows the translation quality obtained using PROD when combining even more languages. Here, we observe that the additional improvement by using more languages is still large. Using more than two languages, the combination method PROD yields better results than MAX. In the end, we obtain a WER improvement of 6.5% using six source languages instead of French alone.

Table 8 shows some of the examples where a combination of French and Spanish yields an improvement.

Table 6: Language combination using method MAX.

languages	WER	PER
fr	55.3	45.3
fr+sv	52.6	43.7
fr+sv+es	**52.0**	**43.2**
fr+sv+es+pt	52.3	43.6
fr+sv+es+pt+it	52.7	44.0
fr+sv+es+pt+it+da	52.5	43.9

Table 7: Language combination using method PROD.

languages	WER	PER
fr	55.3	45.3
fr+sv	54.3	44.5
fr+sv+es	51.0	41.4
fr+sv+es+pt	50.2	40.2
fr+sv+es+pt+it	49.8	39.8
fr+sv+es+pt+it+da	**48.8**	**39.1**

6 Conclusions

We have described methods for translating a text given in multiple source languages into a single target language. We have described the general statistical approach to this problem and have developed two specific statistical models: PROD and MAX. We have evaluated the approach on a multilingual corpus collected automatically from the Internet.

For a large number of language combinations we have been able to obtain significant improvements. The combination method MAX seems to be better suited for the combination of two languages while PROD yields better results if three or more languages are combined. Using PROD, we have been able to improve word error rate when translating into English from 55.3 percent using French as source language to 48.8 percent using five additional source languages.

Currently the search method for combination method PROD produces many search error as the number of considered hypotheses is extremely restricted. Therefore, we expect significant improvement from using a general search algorithm tuned for this problem. Thus, we will also be able to produce sentences that no single-source translation system produces.

The large discrepancies between the translation quality obtained with various languages seem to be mainly due to the sparse data problem resulting from the rich morphology in these languages. Therefore, a

Table 8: Combination examples ('+' denotes chosen translation).

Source: fr	L'existence de limites financières et sa justification;
Source: es	La existencia de límites financieros y su justificación;
Translation: fr	The existence of limit financial and its justification;
Translation: es +	The existence of financial limits and their justification;
Source: fr	Présentation des perspectives financières dans le cadre de l'élargissement.
Source: es	Presentación de las perspectivas financieras en el contexto de la ampliación.
Translation: fr	Presentation of the financial perspective in the framework of enlargement.
Translation: es +	Presentation of the financial perspective in the context of enlargement.
Source: fr	La Bosnie-et-Herzégovine est désormais acceptée comme une nation.
Source: es	Se reconoce a Bosnia y Herzegovina como un Estado nacional.
Translation: fr +	Bosnia and Herzegovina is now accepted as a nation.
Translation: es	Welcomed to Bosnia and Herzegovina as a State national.

systematic handling of morphology using preprocessing and postprocessing (see (Nießen and Ney, 2000)) in these languages would result in a comparable translation quality in all 10 source languages. A combination should lead to an additional significant improvement. Further improvements are expected by performing a finer combination of different languages not on a complete sentence level but on a phrase level.

References

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

W. A. Gale and K. W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–90.

H. Ney, F. J. Och, and S. Vogel. 2000. Statistical translation of spoken dialogues in the verbmobil system. In *Workshop on Multi-Lingual Speech Communication*, pages 69–74, Kyoto, Japan, October.

S. Nießen and H. Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *COLING '00: The 18th Int. Conf. on Computational Linguistics*, pages 1081–1085, Saarbrücken, Germany, July.

F. J. Och and H. Ney. 2000a. A comparison of alignment models for statistical machine translation. In *COLING '00: The 18th Int. Conf. on Computational Linguistics*, pages 1086–1090, Saarbrücken, Germany, August.

F. J. Och and H. Ney. 2000b. Improved statistical alignment models. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 440–447, Hongkong, October.

F. J. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, August.