Improving Word Alignment Quality using Morpho-syntactic Information

Maja Popović and Hermann Ney
Lehrstuhl für Informatik VI - Computer Science Department
RWTH Aachen University
Ahornstrasse 55
52056 Aachen
Germany
{popovic,ney}@cs.rwth-aachen.de

Abstract

In this paper, we present an approach to include morpho-syntactic dependencies into the training of the statistical alignment models. Existing statistical translation systems usually treat different derivations of the same base form as they were independent of each other. We propose a method which explicitly takes into account such interdependencies during the EM training of the statistical alignment models. The evaluation is done by comparing the obtained Viterbi alignments with a manually annotated reference alignment. The improvements of the alignment quality compared to the, to our knowledge, best system are reported on the German-English Verbmobil corpus.

1 Introduction

In statistical machine translation, a translation model $Pr(f_1^J|e_1^I)$ describes the correspondences between the words in the source language sentence f_1^J and the words in the target language sentence e_1^I. Statistical alignment models are created by introducing a hidden variable a_1^J representing a mapping from the source word f_j into the target word e_{a_j}. So far, most of the statistical machine translation systems are based on the single-word alignment models as described in (Brown et al., 1993) as well as the Hidden Markov alignment model (Vogel et al., 1996). The lexicon models used in these systems typically do not include any linguistic or contextual information which often results in inadequate alignments between the sentence pairs.

In this work, we propose an approach to improve the quality of the statistical alignments by taking into account the interdependencies of different derivations of the words. We are getting use of the hierarchical representation of the statistical lexicon model as proposed in (Nießen and Ney, 2001) for the conventional EM training procedure. Experimental results are reported for the German-English Verbmobil corpus and the evaluation is done by comparing the obtained Viterbi alignments after the training of conventional models and models which are using morpho-syntactic information with a manually annotated reference alignment.

2 Related Work

The popular IBM models for statistical machine translation are described in (Brown et al., 1993) and the HMM-based alignment model was introduced in (Vogel et al., 1996). A good overview of all these models is given in (Och and Ney, 2003) where the model IBM-6 is also introduced as the log-linear interpolation of the other models.

Context dependencies have been introduced into the training of alignments in (Varea et al., 2002), but they do not take any linguistic information into account.

Some recent publications have proposed the use of morpho-syntactic knowledge for statistical machine translation, but mostly only for the preprocessing step whereas training procedure of the statistical models remains the same (e.g. (Nießen and Ney, 2001a)).

Incorporation of the morpho-syntactic knowlegde into statistical models has been dealt in (Nießen and Ney, 2001): hierarchical lexicon models containing base forms and set of morpho-syntactic tags are proposed for the translation from German into English. However, these lexicon models are not used for the training but have been created from the Viterbi alignment obtained after the usual training procedure.

The use of POS information for improving statistical alignment quality of the HMM-based model is described in (Toutanova et al., 2002). They introduce additional lexicon probability for POS tags in both languages, but actually are not going beyond full forms.

3 Statistical Alignment Models

The goal of statistical machine translation is to translate an input word sequence $f_1, \ldots, f_J$ in the source language into a target language word sequence $e_1, \ldots, e_I$. Given the source language sequence, we have to choose the target language sequence that maximises the product of the language model probability $Pr(e_1^I)$ and the translation model probability $Pr(f_1^J|e_1^I)$. The translation model describes the correspondence between the words in the source and the target sequence whereas the language model describes well-formedness of a produced target sequence. The translation model can be rewritten in the following way:

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

where a_1^J are called alignments and represent a mapping from the source word position j to the target word position $i = a_j$. Alignments are introduced into translation model as a hidden variable, similar to the concept of Hidden Markov Models (HMM) in speech recognition.

The translation probability $Pr(f_1^J, a_1^J|e_1^I)$ can be further rewritten as follows:

$$\begin{aligned} Pr(f_1^J, a_1^J|e_1^I) &= \prod_{j=1}^{J} Pr(f_j, a_j|f_1^{j-1}, a_1^{j-1}, e_1^I) \\ &= \prod_{j=1}^{J} Pr(a_j|f_1^{j-1}, a_1^{j-1}, e_1^I) \cdot \\ &\quad \cdot Pr(f_j|f_1^{j-1}, a_1^j, e_1^I) \end{aligned}$$

where $Pr(a_j|f_1^{j-1}, a_1^{j-1}, e_1^I)$ is called alignment probability and $Pr(f_j|f_1^{j-1}, a_1^j, e_1^I)$ is lexicon probability.

In all popular translation models IBM-1 to IBM-5 as well as in HMM translation model, the lexicon probability $Pr(f_j|f_1^{j-1}, a_1^j, e_1^I)$ is approximated with the simple single-word lexicon probability $p(f_j|e_{a_j})$ which takes into account only full forms of the words f_j and e_{a_j}. The difference between these models is based on the definition of alignment model $Pr(a_j|f_1^{j-1}, a_1^{j-1}, e_1^I)$. Detailed description of those models can be found in (Brown et al., 1993), (Vogel et al., 1996) and (Och and Ney, 2003).

4 Hierarchical Representation of the Lexicon Model

Typically, the statistical lexicon model is based only on the full forms of the words and does not have any information about the fact that some different full forms are actually derivations of the same base form. For highly inflected languages like German this might cause problems because the coverage of the lexicon might be low since the token/type ratio for German is typically much lower than for English (e.g. for Verbmobil: English 99.4, German 56.3).

To take these interdependencies into account, we use the hierarchical representation of the statistical lexicon model as proposed in (Nießen and Ney, 2001). A constraint grammar parser GERCG for lexical analysis and morphological and syntactic disambiguation for German language is used to obtain morpho-syntactic information. For each German word, this tool provides its base form and the sequence of morpho-syntactic tags, and this information is then added into the original corpus. For example, the German word "gehe" (go), a verb in the indicative mood and present tense which is derived from the base form "gehen" is annotated as "gehe#gehen-V-IND-PRES#gehen".

This new representation of the corpus where full word forms are enriched with its base forms and tags enables gradual accessing of information with different levels of abstraction. Consider for example the above mentioned German word "gehe" which can be translated into the English word "go". Another derivation of the same base form "gehen" is "gehst" which also can be translated by "go". Existing statistical translation models cannot handle the fact that "gehe" and "gehst" are derivatives of the same base form and both can be translated into the same English word "go", whereas the hierarchical representation makes it possible to take such interdependencies into account.

5 EM Training

5.1 Standard EM training (review)

In this section, we will briefly review the standard EM algorithm for the training of the lexicon model.

In the E-step the lexical counts are collected over all sentences in the corpus:

$$C(f, e) = \sum_s \sum_{\mathbf{a}} p(\mathbf{a}|\mathbf{f}^s, \mathbf{e}^s) \sum_{i,j} \delta(f, f_{js}) \delta(e, e_{is})$$

In the M-step the lexicon probabilities are calculated:

$$p(f|e) = \frac{C(f,e)}{\sum_{\tilde{f}} C(\tilde{f},e)}$$

The procedure is similar for the other model parameters, i.e. alignment and fertility probabilities.

For models IBM-1, IBM-2 and HMM, an efficient computation of the sum over all alignments is possible. For the other models, the sum is approximated using an appropriately defined neighbourhood of the Viterbi alignment (see (Och and Ney, 2003) for details).

5.2 EM training using hierarchical counts

In this section we describe the EM training of the lexicon model using so-called hierarchical counts which are collected from the hierarchicaly annotated corpus.

In the E-step the following types of counts are collected:

- full form counts:

$$C(f,e) = \sum_s \sum_{\mathbf{a}} p(\mathbf{a}|\mathbf{f}^s,\mathbf{e}^s) \cdot \cdot \sum_{i,j} \delta(f,f_{js})\delta(e,e_{is})$$

  where f is the full form of the word, e.g. "gehe";

- base form+tag counts:

$$C(fbt,e) = \sum_s \sum_{\mathbf{a}} p(\mathbf{a}|\mathbf{f}^s,\mathbf{e}^s) \cdot \cdot \sum_{i,j} \delta(fbt,fbt_{js})\delta(e,e_{is})$$

  where fbt represents the base form of the word f with sequence of corresponding tags, e.g. "gehen-V-IND-PRES";

- base form counts:

$$C(fb,e) = \sum_s \sum_{\mathbf{a}} p(\mathbf{a}|\mathbf{f}^s,\mathbf{e}^s) \cdot \cdot \sum_{i,j} \delta(fb,fb_{js})\delta(e,e_{is})$$

  where fb is the base form of the word f, e.g. "gehen".

For each full form, refined hierarchical counts are obtained in the following way:

$$C_{hier}(f,e) = C(f,e) + C(fbt,e) + C(fb,e)$$

and the M-step is then performed using hierarchical counts:

$$p(f|e) = \frac{C_{hier}(f,e)}{\sum_{\tilde{f}} C_{hier}(\tilde{f},e)}$$

The training procedure for the other model parameters remains unchanged.

6 Experimental Results

We performed our experiments on the Verbmobil corpus. The Verbmobil task (W. Wahlster, editor, 2000) is a speech translation task in the domain of appointment scheduling, travel planning and hotel reservation. The corpus statistics is shown in Table 1. The number of sure and possible alignments in the manual reference is given as well. We also used a small training corpus consisting of only 500 sentences randomly chosen from the main corpus.

We carried out the training scheme $1^4H^5 3^3 4^3 6^5$ using the toolkit GIZA++. The scheme is defined according to the number of iterations for each model. For example, 4^3 means three iterations of the model IBM-4. We trained the IBM-1 and HMM model using hierarchical lexicon counts, and the parameters of the other models were also indirectly improved thanks to the refined parameters of the initial models.

		German	English
Train	Sentences	34446	
	Words	329625	343076
	Vocabulary	5936	3505
	Singletons	2600	1305
Test	Sentences	354	
	Words	3233	3109
	S relations	2559	
	P relations	4596	

Table 1: Corpus statistics for Verbmobil task

6.1 Evaluation Method

We use the evaluation criterion described in (Och and Ney, 2000). The obtained word alignment is compared to a reference alignment produced by human experts. The annotation scheme explicitly takes into account the ambiguity of the word alignment. The unambiguous alignments are annotated as sure alignments (S) and the ambiguous ones as possible alignments (P). The set of possible alignments P is used especially for idiomatic expressions, free translations and missing function words. The set S is subset of the set P ($S \subseteq P$).

The quality of an alignment A is computed as appropriately redefined precision and recall measures. Additionally, we use the alignment error rate (AER) which is derived from the well-known F-measure.

$$\text{recall} = \frac{|A \cap S|}{|S|}, \quad \text{precision} = \frac{|A \cap P|}{|A|}$$

$$\text{AER} = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

Thus, a recall error can only occur if a S(ure) alignment is not found and a precision error can only occur if a found alignment is not even P(ossible).

6.2 Alignment Quality Results

Table 2 shows the alignment quality for the two corpus sizes of the Verbmobil task. Results are presented for the Viterbi alignments from both translation directions (German→English and English→German) as well as for combination of those two alignments.

The table shows the baseline AER for different training schemes and the corresponding AER when the hierarchical counts are used. We see that there is a consistent decrease in AER for all training schemes, especially for the small training corpus. It can be also seen that greater improvements are yielded for the simpler models.

corpus size = 0.5k				
Training	Model	$D \rightarrow E$	$E \rightarrow D$	combined
1^4	ibm1	27.5	33.4	22.7
	+hier	24.8	30.3	20.5
1^4H^5	hmm	18.8	24.0	16.9
	+hier	16.9	21.5	14.8
$1^4H^53^3$	ibm3	18.4	22.8	17.0
	+hier	16.7	22.1	15.5
$1^4H^53^34^3$	ibm4	16.9	21.5	16.2
	+hier	15.8	20.7	14.9
$1^4H^53^34^36^5$	ibm6	16.7	21.1	15.9
	+hier	15.6	20.9	14.8

corpus size = 34k				
Training	Model	$D \rightarrow E$	$E \rightarrow D$	combined
1^4	ibm1	17.6	24.1	14.1
	+hier	16.8	21.8	13.7
1^4H^5	hmm	8.9	14.9	7.9
	+hier	8.4	13.7	7.3
$1^4H^53^3$	ibm3	8.4	12.8	7.7
	+hier	8.2	12.7	7.4
$1^4H^53^34^3$	ibm4	6.3	10.9	6.0
	+hier	6.1	10.8	5.7
$1^4H^53^34^36^5$	ibm6	5.7	10.0	5.5
	+hier	5.5	9.7	5.0

Table 2: AER [%] for Verbmobil corpus for the baseline system (name of the model) and the system using hierarchical method (+hier)

7 Conclusions

In this work we have presented an approach for including morpho-syntactic knowledge into a maximum likelihood training of statistical translation models. As can be seen in Section 5, going beyond full forms during the training by taking into account the interdependencies of the different derivations of the same base form results in the improvements of the alignment quality, especially for the small training corpus. We assume that the method can be very effective for cases where only small amount of data is available. We also expect further improvements by performing a special modelling for the rare words.

We are planning to investigate possibilities of improving the alignment quality for different language pairs using different types of morpho-syntactic information, like for example to use word stems and suffixes for morphologicaly rich languages where some parts of the words have to be aligned to the whole English words (e.g. Spanish verbs, Finnish in general, etc.) We are also planning to use the refined alignments for the translation process.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311

Ismael García Varea, Franz Josef Och, Hermann Ney and Francisco Casacuberta. 2002. Improving alignment quality in statistical machine translation using context-dependent maximum entropy models. In *Proc. of the 19th International Conference on Computational Linguistics (COLING)*, pages 1051–1057, Taipei, Taiwan, August.

Sonja Nießen and Hermann Ney. 2001a. Morpho-syntactic analysis for reordering in statistical machine translation. In *Proc. MT Summit VIII*, pages 247–252, Santiago de Compostela, Galicia, Spain, September.

Sonja Nießen and Hermann Ney. 2001. Toward hierarchical models for statistical machine translation of inflected languages. In *39th Annual Meeting of the Assoc. for Computational Linguistics - joint with EACL 2001: Proc. Workshop on Data-Driven Machine Translation*, pages 47–54, Toulouse, France, July.

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

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

Kristina Toutanova, H. Tolga Ilhan and Christopher D. Manning. 2002. Extensions to HMM-based statistical word alignment models. In *Proc. Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 87–94, Philadelphia, PA, July.

Stephan Vogel, Hermann Ney and Cristoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proc. of the 16th International Conference on Computational Linguistics (COLING)* , pages 836–841, Copenhagen, Denmark, August.

W. Wahlster, editor 2000. *Verbmobil: Foundations of speech-to-speech translations.* Springer Verlag, Berlin, Germany, July.