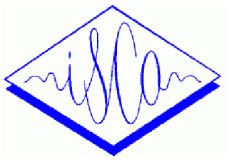
ISCA Archive
http://www.isca-speech.org/archive



International Workshop on Spoken Language Translation (IWSLT) 2006
Keihanna Science City, Kyoto, Japan
November 27-28, 2006

AER: Do we need to "improve" our alignments?

David Vilar, Maja Popović and Hermann Ney

Lehrstuhl für Informatik 6 – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{vilar,popovic,ney}@cs.rwth-aachen.de

Abstract

Currently most statistical machine translation systems make use of alignments as a first step in the process of training the actual translation models. Several researchers have investigated how to improve the alignment quality, with the (intuitive) assumption that better alignments increase the translation quality. In this paper we will investigate this assumption and show that this is not always the case.

1. Introduction

Alignments are a key concept to statistical machine translation. They represent the correspondence between the words of the source and target sentences. They were introduced in the mathematical context of [1] as a hidden variable and used in the framework of the EM Algorithm to estimate the lexicon probabilities and further parameters of the IBM-1 to IBM-5 translation models. Further development and research in statistical machine translation moved from the original single-word-based models to phrase-based-models, in order to better capture the context dependencies of the words in the translation process. The starting point for the training of these models was however the Viterbi alignment produced as a byproduct of the training of the original IBM models, that is, the alignment with the highest probability given the final parameter estimations. Most state-of-the-art machine translation systems, normally based on a phrase-based translation scheme or variations of it, make use of this Viterbi alignment as a first step in the training process [2, 3, 4]. Other translation approaches also benefit from the use of alignments [5].

It is then to expect that an increase in quality of the alignment should lead to an increase in translation quality. At least, it is expected that an improvement in the alignments does not hurt translation performance. In [6] the "Alignment Error Rate" (AER) is introduced as a measure of alignment quality. Given a reference alignment, consisting of a set S of "Sure", unambiguous alignment points and a set P of "Possible", ambiguous alignment points, with $S \subseteq P$, the AER of an alignment $A = \{(j, a_j)\}$ is defined to be

$$\mathrm{AER}(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} . \qquad (1)$$

This error rate is related to the well known F-measure, where the recall is computed using the sure alignments and the precision using the possible alignments. In the same paper, an exhaustive study of different alignment models is carried out.

Following this work, numerous new alignment methods or refinements to existing ones have appeared in the literature, which increase the alignment quality over the standard IBM models. However many of them do not report translation results, and the implicit assumption is made that the improvements on alignment quality will influence the translation process in a positive way.

In this paper we will present two counter-examples to this assumption, that is, we will present (review in one of the cases) two relatively simple refinements of the standard alignment process using the IBM models that actually *deteriorate* the alignment quality. However, they improve the translation performance. We will show this on two translation models, a phrase based system similar to the one used in [7] and a finite state transducer based system as presented in [8]. The key point is that these methods adapt the alignments to the translation models that will make further use of them.

2. Related Work

In [9] the authors conduct an experimental study on the correlation of AER as defined above and the actual translation performance. To our knowledge this is the first work that carries out such a detailed study. The conclusion of their work is that the alignment error rate is not a good measure for predicting translation performance. The main reason given is that AER does not penalize an unbalanced precision and recall. They propose to use the "standard" F-measure directly, defined as

$$\text{F-measure}(A, P, S, \alpha) = \frac{1}{\frac{\alpha}{\text{Precision}(A,P)} + \frac{1-\alpha}{\text{Recall}(A,S)}} , \qquad (2)$$

where, as is the case with alignment error rate, precision and recall are defined as

$$\text{Precision}(A, P) = \frac{|A \cap P|}{|A|} \qquad (3)$$

and

$$\text{Recall}(A, S) = \frac{|A \cap S|}{|S|} . \qquad (4)$$

Note the introduction of a new parameter α which controls the weighting of precision and recall. In their work, the authors find that the more appropriate value of α lies between 0.2 and 0.4, depending on the corpus. Furthermore, they discourage the use of possible alignments in the gold standard reference alignment.

Our goal in this paper is, on the one hand, to provide further empirical evidence that AER is not a suitable measure that can provide insight into the translation process. However, we also show that the proposed F-measure also does not necessarily help in this case. The main flaw found in both of these measures is that they do not take the structure of the translation model into account.

3. Phrase-Based Translation

In this section we will briefly discuss the standard phrase based approach to machine translation, and we will pay special attention to the phrase extraction method. As usual, we will denote the (given) source sentence with $f_1^J = f_1 \dots f_J$, which is to be translated into a target language sentence $e_1^I = e_1 \dots e_I$.

The usual approach in most state-of-the-art translation systems models the translation probability directly using a log-linear model [10]:

$$p(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{\tilde{e}_1^I} \exp\left(\sum_{m=1}^M \lambda_m h_m(\tilde{e}_1^I, f_1^J)\right)}, \quad (5)$$

with a set of different models h_m, scaling factors λ_m and the denominator a normalization factor that can be ignored in the maximization process. The most important models in equation (5) normally are phrase-based models in both source-to-target and target-to-source directions.

In order to extract these phrase-based models, an alignment between the source and target training sentences is found by using the standard IBM models in both directions (source-to-target and target-to-source) and combining the two obtained alignments [6]. Given this alignment an extraction of contiguous phrases is carried out and their probabilities are computed by means of relative frequencies.

Let us examine this process of phrase extraction with more detail. Given a sentence pair with its corresponding alignment, we extract all phrases that fulfill the following restrictions:

1. all source words within the phrase are aligned only to target words within the phrase and
2. all target words within the phrase are aligned only to source words within the phrase.

More formally, the set of bilingual phrases consistent with a word alignment A is defined by

$$\mathcal{BP}(f_1^J, e_1^I, A) = \{(f_j^{j+m}, e_i^{i+n}) \mid \forall (i', j') \in A : j \le j' \le j+m \Leftrightarrow i \le i' \le i+n\}. \quad (6)$$

3.1. Alignment Adaptation

In the following example, we apply this phrase extraction algorithm to a German-English sentence pair taken from the Verbmobil corpus: "*wie sieht es irgendwann morgens am Dienstag , dem sechsten , aus ?*" – "*how about sometime in the morning on Tuesday the sixth ?*". The reference alignment for this sentence pair and the alignment found by GIZA++ [6] applying the IBM models can be seen in Figures 1(a) and 1(b), respectively. The automatically generated alignment perfectly matches the reference in this case. The German language has so called "separable verbs" ("trennbare Verben"), verbs that are formed from two parts, normally a main part and a short particle that determines the exact meaning. In the example in Figure 1 we have one such verb: "aus·sehen". The English expression "how about…?" corresponds to the German construction "wie sieht es…aus?", as reflected in the alignment with the link between "aus" and "about". We would like to extract phrases containing the pair "wie sieht es"–"how about", which is quite appropriate for the translation process. But, due to the link between "aus" and "about", the only phrases that we can extract containing this pair are the one shown in Figure 1(c) and the same including the question marks. Having such a long context, it is quite improbable that we could use one of these phrases in the translation process.

This is a clear example of a recurrent phenomenon that can be observed when looking into the alignments from German to English. A simple, "brute force" solution to this problem is to remove these distant points. For doing that, we simply compute for each alignment point the distance to the points in the previous and next non-empty columns. If both are above a given threshold (3 worked best in our case on a development corpus) the point is discarded from the alignment. Similarly, this is applied for the rows. The resulting alignment is shown in Figure 1(d). The point that links "aus" and "about" has been erased, and thus the desired phrase pair "wie sieht es"–"how about" can be extracted.

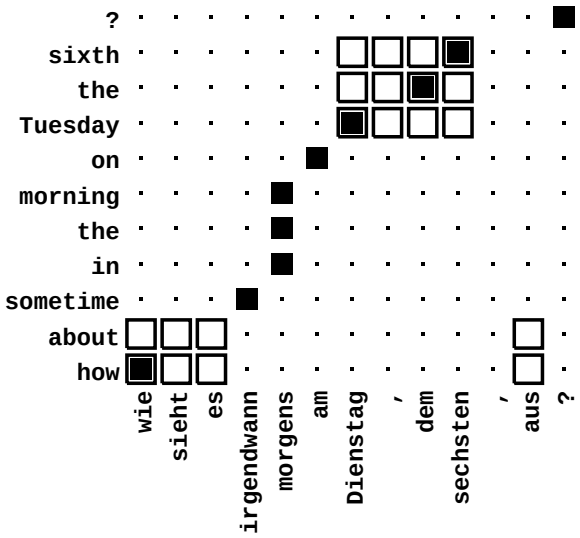
Note that in this case the alignment does not get worse as the link was marked as possible in the reference[1]. However it is expected that applying this method to the whole corpus will in fact increase the alignment error rate. A detailed analysis of the results is presented in section 5.

4. Tuple-Based Translation

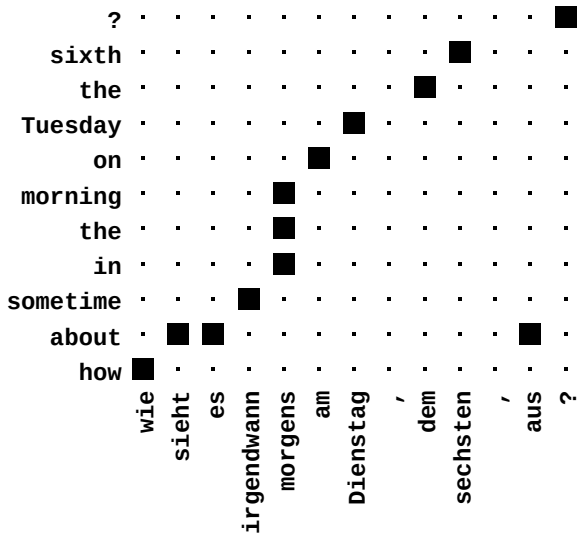
In this section we will briefly discuss an alternative translation model and present how to obtain alignments that better match the probabilistic model. A detailed description can be found in [8]. We will denote with $\tilde{e}_1^J$ a segmentation of a target sentence e_1^I into J phrases such that f_1^J and $\tilde{e}_1^J$ can be aligned to form bilingual tuples $(f_j, \tilde{e}_j)$.

We can then formulate the problem of finding the best

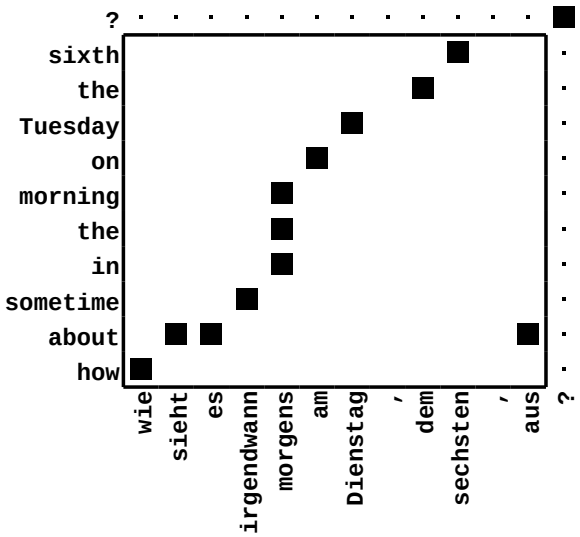
[1] Note however that, because of the simplicity of the algorithm, we have also removed the link between the question marks. This in fact affects the alignment quality.



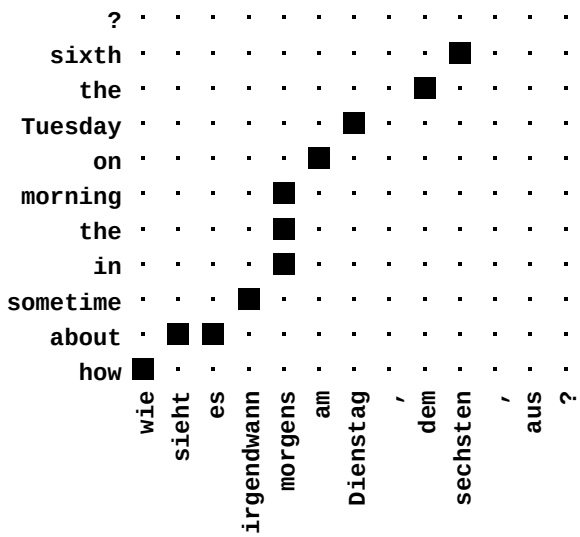
(a) Reference alignment. Sure points are marked as full boxes, Possible as empty boxes.



(b) Automatically found alignment.



(c) Phrase extraction with the generated alignment.



(d) Simplified alignment.

Figure 1: Example of a problematic alignment for the phrase-based system.

translation $\hat{e}_1^I$ of a source sentence f_1^J (here $\mathcal{A}$ denotes the set of all possible alignments):

$$\begin{aligned}
\hat{e}_1^I &= \operatorname*{argmax}_{e_1^I} Pr(f_1^J, e_1^I) \\
&= \operatorname*{argmax}_{\tilde{e}_1^J} \sum_{A \in \mathcal{A}} Pr(f_1^J, \tilde{e}_1^J, A) \\
&\cong \operatorname*{argmax}_{\tilde{e}_1^J} \max_{A \in \mathcal{A}} Pr(A) \cdot Pr(f_1^J, \tilde{e}_1^J | A) \\
&\cong \operatorname*{argmax}_{\tilde{e}_1^J} \max_{A \in \mathcal{A}} \prod_{f_j : j=1\ldots J} Pr(f_j, \tilde{e}_j | f_1^{j-1}, \tilde{e}_1^{j-1}, A) \\
&= \operatorname*{argmax}_{\tilde{e}_1^J} \max_{A \in \mathcal{A}} \prod_{f_j : j=1\ldots J} p(f_j, \tilde{e}_j | f_{j-m}^{j-1}, \tilde{e}_{j-m}^{j-1}, A) \,.
\end{aligned}$$

In other words: if we assume a uniform distribution for $Pr(A)$, the translation problem can be mapped to the problem of estimating an m-gram language model over a learned set of bilingual tuples $(f_j, \tilde{e}_j)$. In our case we represent this language model as a weighted finite state transducer, but this is not the only possibility [11].

Assume that the alignment is a function of the target words $A' : \{1, \ldots, I\} \rightarrow \{1, \ldots, J\}$, then the bilingual tuples $(f_j, \tilde{e}_j)$ can be inferred with e. g. the GIATI method of [4]. Each source word will be mapped to a target phrase of one or more words or an "empty" phrase ε. In particular, the source words which will remain non-aligned due to the alignment functionality restriction are paired with the empty phrase. However the alignments produced by the standard alignment generation procedure do not have this function-like property.

Furthermore, assuming that we could have such an alignment, when the function A' is not monotonic, the target language phrases $\tilde{e}$ can become very long. For example, given a completely non-monotonic alignment, all target words will be paired with the last aligned source word. All other source words form tuples with the empty phrase. Therefore, for language pairs with big differences in word order, probability estimates may be poor.

4.1. Alignment Adaptation

This problem can be solved by reordering either the source or the target training sentences (both in training and test phases) in a way such that alignments become monotonic for all sentences. In [8] a method is presented to obtain an alignment that fulfill both requirements. Here we will give an overview of it.

First, we estimate a cost matrix C for each sentence pair (f_1^J, e_1^I). The elements of this matrix c_{ij} are the local costs of aligning a source word f_j to a target word e_i. This cost matrix is estimated using the original IBM models, see [12] for more detail. For a given alignment $A \subseteq I \times J$, define the costs of this alignment, $c(A)$, as the sum of the local costs of all aligned word pairs:

$$c(A) = \sum_{(i,j) \in A} c_{ij} \tag{7}$$

		German	English
Train	Sentences	751 088	
	Words	15 256 793	16 052 269
	Vocabulary	195 291	65 889
Test	Sentences	2 000	
	Words	54 247	57 945

Table 1: Statistics of the Europarl corpus.

The goal is to find an alignment with the minimum costs which fulfills the given constraints.

In a first step, we require the alignment to be a function of *source* words A_1: $\{1, \ldots, J\} \rightarrow \{1, \ldots, I\}$ in order to uniquely define a reordering of the source sentence. This is easily computed from the cost matrix C as:

$$A_1(j) = \operatorname*{argmin}_i c_{ij} \,. \tag{8}$$

Non-aligned source words are not allowed. A_1 naturally defines a new order of the source words f_1^J.

In the second pass we extract the alignment that is a function of the target words for computing the corpus of bilingual tuples, and is also monotonic. This is computed as a minimum-cost alignment (using a "reordered" cost matrix) with a dynamic programming algorithm similar to the Levenshtein string edit distance algorithm. An example of this method is shown in Figure 2.

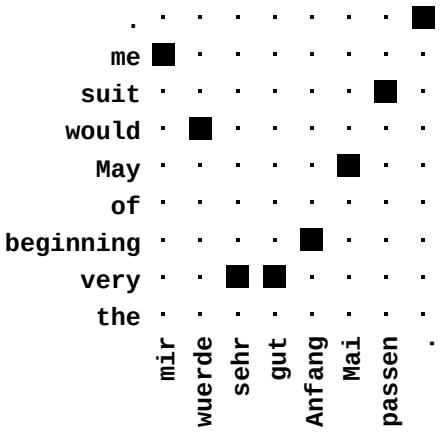
Because of the special constraints we require for this model, the alignment quality is expected to be relatively poor.

5. Experimental Results

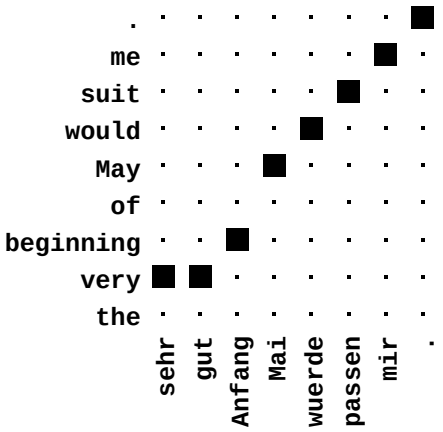
In this section we will analyze the impact the alignment methods described in Sections 3.1 and 4.1 have on both alignment and translation quality. For this, experiments will be reported on the Europarl corpus as used in the ACL 2005 Machine Translation Workshop Shared Task [13], for the German-English language pair. The corpus consists of the proceedings of the European Parliament, which are published on the web. Statistics are shown in Table 1. This corpus was chosen because of the different structure of the German and the English languages, that allows to better observe the effect of the alignments than for other language pairs, where the alignment is quasi-monotonic (e.g. English-Spanish).

In order to have a reference alignment, we randomly selected a subset of the training corpus, consisting of 508 sentences, and manually annotated the alignments. Contrary to the recommendation in [9], we used both sure and possible alignments, as the restriction of using only sure alignments is very restrictive and we feel that it does not completely reflect the correspondences between the two languages. This is especially true for "real-life" corpora[2], as the one we are

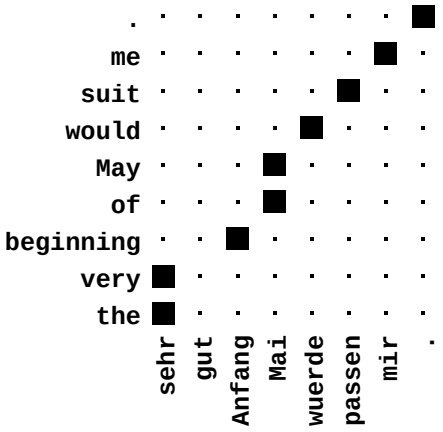
[2] In contrast to corpora created specifically for research purposes, where the translations are created specifically for one task and often are very literal translations.



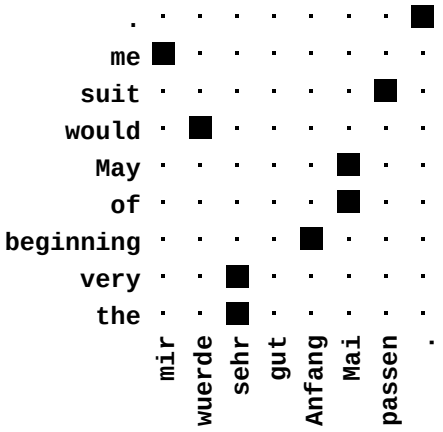
(a) Viterbi alignment.



(b) First pass alignment.



(c) Second pass alignment.



(d) Second pass alignment with original source sentence order.

Figure 2: Alignments for the tuple based model.

dealing with. In many cases the translator did not produce a one-to-one translation. Instead the same meaning is expressed in a way that is fits better in the structure of the target language. Possible alignments reflect better this process.

The results are shown in Table 2 (all systems were optimized for the BLEU score). It can be seen that both alignment transformation methods described before (entries "Phrases" and "Tuples" in the table) deteriorate the alignment quality both in terms of alignment error rate and F-measure. The error rate increases from 20.8% for the baseline to 24.2% in the case of the transformation for the phrase-based system and 26.4% for the alignments computed for the tuple-based one. The translation quality however measured by the BLEU score can be seen to improve[3] if we apply the alignment method with its corresponding translation system, slightly in the case of the phrase translation system. In the case of the tuple model, the absolute scores are significantly worse than for the phrase based model[4], but the effect of the alignment type is much more important in this case. Note also that applying the method that does not correspond to the system deteriorates the translation quality.

Figures 3 and 4 show example translations for the phrase-based system and for the tuple-based system, respectively. As can be expected from the little difference in the evaluation measures, the differences in the phrase-based system are small when going from comparing the two alignment methods. However the examples show clearly the effect we presented in Section 3.1. In the first example the auxiliary verb "wird" is used to build the future tense of the main verb. In the baseline case the system is not able to find this and leaves the present tense, whereas in with the alignment adaptation the future tense is correctly translated. A similar effect can be seen in the second example, where the German passive construction does not allow to translate the verb correctly. In the case of the tuple-based system the improvement is more evident, as can be seen in Figure 4.

6. Discussion

Having arrived at this point, there are some open questions that should be discussed. The main one is of course whether the AER (or the F-measure as proposed in [9]) is an adequate measure of alignment quality. Actually we think it is. It is based on the precision, recall and F-measures that are widely used in the pattern recognition community (among others) and have proved to be quite useful. And in fact, when looking at the alignments, a human can see a good correlation between a lower alignment error rate and the quality of the alignments.

We think that the main problem lies in the "inconsistency" between the statistical models used in the alignment

[3]Other performance measures do not show this behavior, but this is mainly due to the fact that the systems were optimized with respect to BLEU.

[4]This can probably be explained by the lack of the combination of models, as happens in the case of the phrase-based translation system.

Original	Es wird ein ganzes Kapitel über Wissenschaft, Gesellschaft und Bürger geben.
Baseline	It is a chapter on science, society and citizens.
Phrases	It will be a whole chapter on science, society and citizens.
Reference	There will be an entire chapter on science, society and the citizens.

Original	Das reicht nicht aus, die gesamte Strategie muss stärker auf die Bürger und Bürgerinnen ausgerichtet werden.
Baseline	That is not enough, the whole strategy must be more closely to the citizens of Europe.
Phrases	That is not enough, the whole strategy must focus more on the citizens of Europe.
Reference	It is not enough; the whole strategy needs to be geared more to the citizens.

Figure 3: Example translations for the phrase based system.

Original	Litauen verfügt über ein beträchtliches Potential für ein langfristiges Wirtschaftswachstum.
Baseline	Has a considerable potential for a long-term Lithuania, although economic growth.
Tuples	Lithuania has a considerable potential for a long-term economic growth.
Reference	Lithuania has considerable potential for long-term economic growth.

Original	Gleichzeitig müssen berechtigte Interessen der Arbeitnehmer berücksichtigt werden.
Baseline	We must justified interests of employees.
Tuples	At the same time legitimate interests of employees must be taken into account.
Reference	At the same time, the workers' legitimate interests need to be considered.

Figure 4: Example translations for the tuple based system.

System	Alignment	AER[%]	F[%]	BLEU[%]	NIST	WER[%]	PER[%]
Phrase Based	Baseline	20.8	77.5	24.6	6.62	66.4	47.3
	Phrases	24.2	71.8	24.8	6.56	66.7	47.9
	Tuples	26.4	73.6	24.5	6.65	66.0	47.0
Tuple Based	Baseline	20.8	77.5	18.2	5.54	68.1	52.4
	Phrases	24.2	71.8	14.8	3.16	69.6	58.6
	Tuples	26.4	73.6	19.4	6.30	68.8	50.0

Table 2: Alignment and translation results for the different translation and alignment methods.

procedure and the models used later in the translation process. If we had perfect statistical translation models that could generate a completely correct translation given a perfect alignment, it could perfectly be that a direct relation between alignment quality and translation quality would exist. However we do not have such perfect models and the training procedure can be "confused" when it finds structures it does not expect, although they may be completely correct. Therefore it can be of advantage to sacrifice some alignment quality in order to better guide the training process and have more robust estimations.

A recent work [14] actually presents a new measure called "consistent phrase error rate" which tries to extend the AER to the concept of phrases. The authors show how this measure correlates better with translation performance, but it is however to much oriented to a phrase-based system and we expect it to perform poorly for other translation approaches[5].

But one can take a step further. The alignment concept was first introduced as a hidden variable for the training of the single-word based models. Let them remain hidden then. When switching to phrase-based models the given data is assumed to be not only the training sentence pairs, but also the alignment between then. This alignment, however, is computed by using the parameter estimations of models that are only a (raw) approximation of the true parameter distributions, but are treated as an additional sure knowledge source. If we could retain the spirit of the alignments as a hidden variable in the spirit of the EM algorithm and include them in the training procedure of the more advanced models, the adaptation discussed in the preceding paragraph would be included automatically. First steps in this direction have already been undertaken [15, 16].

7. Conclusions

In this paper we have shown that the improvement in alignment quality does not always imply an improvement in translation quality. We have shown techniques to generate alignments that are better adapted to the characteristics of the translation models that will later make use of this information. Although the error rate of these transformed alignments was larger than the baseline, the translation quality actually improved. We have shown this effect on two different approaches for the modeling of the translation probability.

Seeing the outcome of the experiments presented in this paper, one clear conclusion can be drawn: future work on alignment (at least if oriented to machine translation) should always report results on translation quality.

[5]We want to thank the reviewers for pointing out this work

8. Acknowledgments

This work has been funded by the integrated project TC-STAR – Technology and Corpora for Speech-to-Speech Translation – (IST-2002-FP6-506738). We would also like to thank our colleagues who helped producing the reference alignments.

9. References

[1] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, June 1993.

[2] R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, and H. Ney, "The RWTH phrase-based statistical machine translation system," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Pittsburgh, PA, October 2005, pp. 155–162.

[3] D. Chiang, "A hierarchical phrase-based model for statistical machine translation," in *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI, June 2005, pp. 263–270.

[4] F. Casacuberta and E. Vidal, "Machine Translation with Inferred Stochastic Finite-State Transducers," *Computational Linguistics*, vol. 30, no. 2, pp. 205–225, 2004.

[5] J. M. Vilar, "Improve the learning of subsequential transducers by using alignments and dictionaries," in *Grammatical Inference: Algorithms and Applications*, ser. Lecture Notes in Artificial Intelligence, A. de Oliveira, Ed., vol. 1891. Lisbon, Portugal: Springer-Verlag, Sept. 2000, pp. 298–311.

[6] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, March 2003.

[7] D. Vilar, E. Matusov, S. Hasan, R. Zens, and H. Ney, "Statistical Machine Translation of European Parliamentary Speeches," in *Proceedings of MT Summit X*. Phuket, Thailand: Asia-Pacific Association for Machine Translation (AAMT), September 2005, pp. 259–266.

[8] S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney, "Novel reordering approaches in phrase-based statistical machine translation," in *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, Ann Arbor, Michigan, June 2005, pp. 167–174.

[9] A. Fraser and D. Marcu, "Measuring word alignment quality for statistical machine translation," ISI-University of Southern California, Tech. Rep., May 2006.

[10] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, July 2002, pp. 295–302.

[11] J. Mariño, R. Banchs, P. Lambert, M. Ruiz, J. Crego, and J. Fonollosa, "Bilingual N-gram Statistical Machine Translation," in *Proceedings of MT Summit X*. Phuket, Thailand: Asia-Pacific Association for Machine Translation (AAMT), September 2005, pp. 275–282.

[12] E. Matusov, R. Zens, and H. Ney, "Symmetric word alignments for statistical machine translation," in *COLING '04: The 20th Int. Conf. on Computational Linguistics*, Geneva, Switzerland, August 2004, pp. 219–225.

[13] P. Koehn and C. Monz, "Shared task: Statistical machine translation between european languages," in *Proceedings of the ACL 2005 Workshop on Parallel Text*, Ann Arbor, MI, USA, June 2005, pp. 119–124.

[14] N. F. Ayan and B. J. Dorr, "Going Beyond AER: An Extensive Analysis of Word Alignments and Their Impact on MT," in *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL'2006)*, Sydney, Australia, July 2006, pp. 9–16.

[15] A. Venugopal, S. Vogel, and A. Waibel, "Effective phrase translation extraction from alignment models," in *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 319–326.

[16] D. Marcu and W. Wong, "A phrase-based, joint probability model for statistical machine translation," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Philadelphia, PA, July 2002, pp. 133–139.