A Comparison of Alignment Models for Statistical Machine Translation

Franz Josef Och and **Hermann Ney**
Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen - University of Technology
D-52056 Aachen, Germany
{och,ney}@informatik.rwth-aachen.de

Abstract

In this paper, we present and compare various alignment models for statistical machine translation. We propose to measure the quality of an alignment model using the quality of the Viterbi alignment compared to a manually-produced alignment and describe a refined annotation scheme to produce suitable reference alignments. We also compare the impact of different alignment models on the translation quality of a statistical machine translation system.

1 Introduction

In statistical machine translation (SMT) it is necessary to model the translation probability $Pr(f_1^J|e_1^I)$. Here $f_1^J = \mathbf{f}$ denotes the (French) source and $e_1^I = \mathbf{e}$ denotes the (English) target string. Most SMT models (Brown et al., 1993; Vogel et al., 1996) try to model word-to-word correspondences between source and target words using an alignment mapping from source position j to target position $i = a_j$.

We can rewrite the probability $Pr(f_1^J|e_1^I)$ by introducing the 'hidden' alignments $a_1^J := a_1...a_j...a_J$ $(a_j \in \{0, \ldots, I\})$:

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

To allow for French words which do not directly correspond to any English word an artificial 'empty' word e_0 is added to the target sentence at position $i = 0$.

The different alignment models we present provide different decompositions of $Pr(f_1^J, a_1^J|e_1^I)$. An alignment $\hat{a}_1^J$ for which holds

$$\hat{a}_1^J = \arg\max_{a_1^J} Pr(f_1^J, a_1^J|e_1^I)$$

for a specific model is called Viterbi alignment of this model.

In this paper we will describe extensions to the Hidden-Markov alignment model from (Vogel et al., 1996) and compare these to Models 1 - 4 of (Brown et al., 1993). We propose to measure the quality of an alignment model using the quality of the Viterbi alignment compared to a manually-produced alignment. This has the advantage that once having produced a reference alignment, the evaluation itself can be performed automatically. In addition, it results in a very precise and reliable evaluation criterion which is well suited to assess various design decisions in modeling and training of statistical alignment models.

It is well known that manually performing a word alignment is a complicated and ambiguous task (Melamed, 1998). Therefore, to produce the reference alignment we use a refined annotation scheme which reduces the complications and ambiguities occurring in the manual construction of a word alignment. As we use the alignment models for machine translation purposes, we also evaluate the resulting translation quality of different models.

2 Alignment with HMM

In the Hidden-Markov alignment model we assume a first-order dependence for the alignments a_j and that the translation probability depends only on a_j and not on a_{j-1}:

$$Pr(f_j, a_j|f_1^{j-1}, a_1^{j-1}, e_1^I) = p(a_j|a_{j-1}, I)p(f_j|e_{a_j})$$

Later, we will describe a refinement with a dependence on $e_{a_{j-1}}$ in the alignment model. Putting everything together, we have the following basic HMM-based model:

$$p(f_1^J|e_1^I) = \sum_{a_1^J} \prod_{j=1}^{J} \left[p(a_j|a_{j-1}, I) \cdot p(f_j|e_{a_j})\right] \quad (1)$$

with the alignment probability $p(i|i', I)$ and the translation probability $p(f|e)$. To find a Viterbi alignment for the HMM-based model we resort to dynamic programming (Vogel et al., 1996).

The training of the HMM is done by the EM-algorithm. In the E-step the lexical and alignment counts for one sentence-pair $(\mathbf{f}, \mathbf{e})$ are calculated:

$$c(f|e;\mathbf{f},\mathbf{e}) = \sum_{\mathbf{a}} Pr(\mathbf{a}|\mathbf{f},\mathbf{e}) \sum_{i,j} \delta(f,f_j)\delta(e,e_i)$$
$$c(i|i',I;\mathbf{f},\mathbf{e}) = \sum_{\mathbf{a}} Pr(\mathbf{a}|\mathbf{f},\mathbf{e}) \sum_{j} \delta(i',a_{j-1})\delta(i,a_j)$$

In the M-step the lexicon and translation probabilities are:

$$\begin{aligned} p(f|e) &\propto \sum_s c(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)}) \\ p(i|i',I) &\propto \sum_s c(i|i',I;\mathbf{f}^{(s)},\mathbf{e}^{(s)}) \end{aligned}$$

To avoid the summation over all possible alignments $\mathbf{a}$, (Vogel et al., 1996) use the maximum approximation where only the Viterbi alignment path is used to collect counts. We used the Baum-Welch-algorithm (Baum, 1972) to train the model parameters in our experiments. Thereby it is possible to perform an efficient training using all alignments.

To make the alignment parameters independent from absolute word positions we assume that the alignment probabilities $p(i|i',I)$ depend only on the jump width $(i-i')$. Using a set of non-negative parameters $\{c(i-i')\}$, we can write the alignment probabilities in the form:

$$p(i|i',I) = \frac{c(i-i')}{\sum_{i''=1}^{I} c(i''-i')} \quad . \tag{2}$$

This form ensures that for each word position i', $i' = 1, ..., I$, the alignment probabilities satisfy the normalization constraint.

Extension: refined alignment model

The count table $c(i-i')$ has only $2 \cdot I_{max} - 1$ entries. This might be suitable for small corpora, but for large corpora it is possible to make a more refined model of $Pr(a_j|f_1^{j-1}, a_1^{j-1}, e_1^I)$. Especially, we analyzed the effect of a dependence on $e_{a_{j-1}}$ or f_j. As a dependence on all English words would result in a huge number of alignment parameters we use as (Brown et al., 1993) equivalence classes G over the English and the French words. Here G is a mapping of words to classes. This mapping is trained automatically using a modification of the method described in (Kneser and Ney, 1991). We use 50 classes in our experiments. The most general form of alignment distribution that we consider in the HMM is $p(a_j - a_{j-1}|G(e_{a_j}), G(f_j), I)$.

Extension: empty word

In the original formulation of the HMM alignment model there is no 'empty' word which generates French words having no directly aligned English word. A direct inclusion of an empty word in the HMM model by adding an e_0 as in (Brown et al., 1993) is not possible if we want to model the jump distances $i - i'$, as the position $i = 0$ of the empty word is chosen arbitrarily. Therefore, to introduce the empty word we extend the HMM network by I empty words e_{I+1}^{2I}. The English word e_i has a corresponding empty word e_{i+I}. The position of the empty word encodes the previously visited English word.

We enforce the following constraints for the transitions in the HMM network ($i \leq I$, $i' \leq I$):

$$\begin{aligned} p(i+I|i',I) &= p_0^H \cdot \delta(i,i') \\ p(i+I|i'+I,I) &= p_0^H \cdot \delta(i,i') \\ p(i|i'+I,I) &= p(i|i',I) \end{aligned}$$

The parameter p_0^H is the probability of a transition to the empty word. In our experiments we set $p_0^H = 0.2$.

Smoothing

For a better estimation of infrequent events we introduce the following smoothing of alignment probabilities:

$$p'(a_j|a_{j-1},I) = \alpha \cdot \frac{1}{I} + (1-\alpha) \cdot p(a_j|a_{j-1},I)$$

In our experiments we use $\alpha = 0.4$.

3 Model 1 and Model 2

Replacing the dependence on a_{j-1} in the HMM alignment model by a dependence on j, we obtain a model which can be seen as a zero-order Hidden-Markov Model which is similar to Model 2 proposed by (Brown et al., 1993). Assuming a uniform alignment probability $p(i|j,I) = 1/I$, we obtain Model 1.

Assuming that the dominating factor in the alignment model of Model 2 is the distance relative to the diagonal line of the (j,i) plane the model $p(i|j,I)$ can be structured as follows (Vogel et al., 1996):

$$p(i|j,I) = \frac{r(i - j\frac{I}{J})}{\sum_{i'=1}^{I} r(i' - j\frac{I}{J})} \quad . \tag{3}$$

This model will be referred to as diagonal-oriented Model 2.

4 Model 3 and Model 4

Model: The fertility models of (Brown et al., 1993) explicitly model the probability $p(\phi|e)$ that the English word e_i is aligned to

$$\phi_i = \sum_j \delta(a_j, i)$$

French words.

Model 3 of (Brown et al., 1993) is a zero-order alignment model like Model 2 including in addition fertility parameters. Model 4 of (Brown et al., 1993) is also a first-order alignment model (along the source positions) like the HMM, but includes also fertilities. In Model 4 the alignment position j of an English word depends on the alignment position of the previous English word (with non-zero fertility) j'. It models a jump distance $j-j'$ (for consecutive English words) while in the HMM a jump distance $i-i'$ (for consecutive French words) is modeled. The full description of Model 4 (Brown et al., 1993) is rather complicated as there have to be considered the cases that English words have fertility larger than one and that English words have fertility zero.

For training of Model 3 and Model 4, we use an extension of the program GIZA (Al-Onaizan et al., 1999). Since there is no efficient way in these models to avoid the explicit summation over all alignments in the EM-algorithm, the counts are collected only over a subset of promising alignments. It is not known an efficient algorithm to compute the Viterbi alignment for the Models 3 and 4. Therefore, the Viterbi alignment is computed only approximately using the method described in (Brown et al., 1993). The models 1-4 are trained in succession with the final parameter values of one model serving as the starting point for the next.

A special problem in Model 3 and Model 4 concerns the deficiency of the model. This results in problems in re-estimation of the parameter which describes the fertility of the empty word. In normal EM-training, this parameter is steadily decreasing, producing too many alignments with the empty word. Therefore we set the probability for aligning a source word with the empty word at a suitably chosen constant value.

As in the HMM we easily can extend the dependencies in the alignment model of Model 4 easily using the word class of the previous English word $E = G(e_{i'})$, or the word class of the French word $F = G(f_j)$ (Brown et al., 1993).

5 Including a Manual Dictionary

We propose here a simple method to make use of a bilingual dictionary as an additional knowledge source in the training process by extending the training corpus with the dictionary entries. Thereby, the dictionary is used already in EM-training and can improve not only the alignment for words which are in the dictionary but indirectly also for other words. The additional sentences in the training corpus are weighted with a factor F_{lex} during the EM-training of the lexicon probabilities.

We assign the dictionary entries which really co-occur in the training corpus a high weight F_{lex} and the remaining entries a very low weight. In our experiments we use $F_{lex} = 10$ for the co-occurring dictionary entries which is equivalent to adding every dictionary entry ten times to the training corpus.

6 The Alignment Template System

The statistical machine-translation method described in (Och et al., 1999) is based on a word aligned training corpus and thereby makes use of single-word based alignment models. The key element of this approach are the *alignment templates* which are pairs of phrases together with an alignment between the words within the phrases. The advantage of the alignment template approach over word based statistical translation models is that word context and local re-orderings are explicitly taken into account. We typically observe that this approach 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).

7 Results

We present results on the Verbmobil Task which is a speech translation task in the domain of appointment scheduling, travel planning, and hotel reservation (Wahlster, 1993).

We measure the quality of the above mentioned alignment models with respect to *alignment quality* and *translation quality.*

To obtain a reference alignment for evaluating alignment quality, we manually aligned about 1.4 percent of our training corpus. We allowed the humans who performed the alignment to specify two different kinds of alignments: an S (sure) alignment which is used for alignments which are unambiguously and a P (possible) alignment which is used for alignments which might or might not exist. The P relation is used especially to align words within idiomatic expressions, free translations, and missing function words. It is guaranteed that $S \subseteq P$. Figure 1 shows an example of a manually aligned sentence with S and P relations. The human-annotated alignment does not prefer any translation direction and may therefore contain many-to-one and one-to-many relationships. The annotation has been performed by two annotators, producing sets S_1, P_1, S_2, P_2. The reference alignment is produced by forming the intersection of the sure alignments ($S = S_1 \cap S_2$) and the union of the possible alignments ($P = P_1 \cup P_2$).

The quality of an alignment $A = \{(j, a_j)\}$ is measured using the following alignment error rate:

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

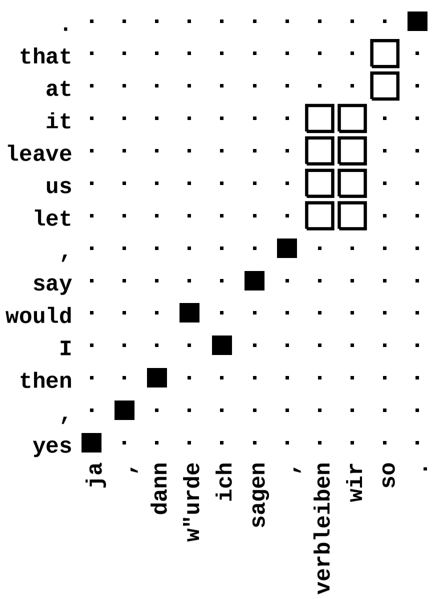

Figure 1: Example of a manually annotated alignment with *sure* (filled dots) and *possible* connections.

Obviously, if we compare the sure alignments of every single annotator with the reference alignment we obtain an AER of zero percent.

Table 1: Corpus characteristics for alignment quality experiments.

		German	English
Train	Sentences	34 446	
	Words	329 625	343 076
	Vocabulary	5 936	3 505
Dictionary	Entries	4 183	
	Words	4 533	5 324
Test	Sentences	354	
	Words	3 109	3 233

Table 1 shows the characteristics of training and test corpus used in the alignment quality experiments. The test corpus for these experiments (not for the translation experiments) is part of the training corpus.

Table 2 shows the alignment quality of different alignment models. Here the alignment models of HMM and Model 4 do not include a dependence on word classes. We conclude that more sophisticated alignment models are crucial for good alignment quality. Consistently, the use of a first-order alignment model, modeling an empty word and fertilities result in better alignments. Interestingly, the simpler HMM alignment model outperforms Model 3 which shows the importance of first-order alignment models. The best performance is achieved with Model 4. The improvement by using a dictionary is small compared to the effect of using better alignment models. We see a significant difference in alignment quality if we exchange source and target languages. This is due to the restriction in all alignment models that a source language word can be aligned to at most one target language word. If German is source language the frequently occurring German word compounds, cannot be aligned correctly, as they typically correspond to two or more English words.

Table 3 shows the effect of including a dependence on word classes in the alignment model of HMM or Model 4. By using word classes the results can be improved by 0.9% when using the HMM and by 0.5% when using Model 4.

Table 3: Effect of including a dependence on word classes in the alignment model.

	AER [%]	
Dependencies	HMM	Model 4
no	8.0	6.5
source	7.5	6.0
target	7.1	6.1
source + target	7.6	6.1

For the translation experiments we used a different training and an independent test corpus (Table 4).

Table 4: Corpus characteristics for translation quality experiments.

		German	English
Train	Sentences	58 332	
	Words	519 523	549 921
	Vocabulary	7 940	4 673
Test	Sentences	147	
	Words	1 968	2 173
	PP (trigram LM)	(40.3)	28.8

For the evaluation of the translation quality we used the automatically computable Word Error Rate (WER) and the Subjective Sentence Error Rate (SSER) (Nießen et al., 2000). The WER corresponds to the edit distance between the produced translation and one predefined reference translation. To obtain the SSER the translations are classified by human experts into a small number of quality classes ranging from “perfect” to “absolutely wrong”. In comparison to the WER, this criterion is more meaningful, but it is also very expensive to measure. The translations are produced by the alignment template system mentioned in the previous section.

Table 2: Alignment error rate (AER [%]) of different alignment models for the translations directions English into German (German words have fertilities) and German into English.

	English → German			German → English		
Dictionary	no		yes	no		yes
Empty Word	no	yes	yes	no	yes	yes
Model 1	17.8	16.9	16.0	22.9	21.7	20.3
Model 2	12.8	12.5	11.7	17.5	17.1	15.7
Model 2(diag)	11.8	10.5	9.8	16.4	15.1	13.3
Model 3	10.5	9.3	8.5	15.7	14.5	12.1
HMM	10.5	9.2	8.0	14.1	12.9	11.5
Model 4	9.0	7.8	6.5	14.0	12.5	10.8

Table 5: Effect of different alignment models on translation quality.

Alignment Model in Training	WER[%]	SSER[%]
Model 1	49.8	22.2
HMM	47.7	19.3
Model 4	48.6	16.8

The results are shown in Table 5. We see a clear improvement in translation quality as measured by SSER whereas WER is more or less the same for all models. The improvement is due to better lexicons and better alignment templates extracted from the resulting alignments.

8 Conclusion

We have evaluated various statistical alignment models by comparing the Viterbi alignment of the model with a human-made alignment. We have shown that by using more sophisticated models the quality of the alignments improves significantly. Further improvements in producing better alignments are expected from using the HMM alignment model to bootstrap the fertility models, from making use of cognates, and from statistical alignment models that are based on word groups rather than single words.

Acknowledgment

This article has been partially supported as part of the Verbmobil project (contract number 01 IV 701 T4) by the German Federal Ministry of Education, Science, Research and Technology.

References

Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, I. D. Melamed, F. J. Och, D. Purdy, N. A. Smith, and D. Yarowsky. 1999. Statistical machine translation, final report, JHU workshop. `http://www.clsp.jhu.edu/ws99/projects/mt/final_report/mt-final-report.ps`.

L.E. Baum. 1972. An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. *Inequalities*, 3:1–8.

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.

R. Kneser and H. Ney. 1991. Forming Word Classes by Statistical Clustering for Statistical Language Modelling. In *1. Quantitative Linguistics Conf.*

I. D. Melamed. 1998. Manual annotation of translational equivalence: The Blinker project. Technical Report 98-07, IRCS.

S. Nießen, F. J. Och, G. Leusch, and H. Ney. 2000. An evaluation tool for machine translation: Fast evaluation for mt research. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 39–45, Athens, Greece, May June.

F. J. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *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, USA, 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.

W. Wahlster. 1993. Verbmobil: Translation of face-to-face dialogs. In *Proc. of the MT Summit IV*, pages 127–135, Kobe, Japan.