Improving alignment quality in statistical machine translation using context-dependent maximum entropy models*

Ismael García Varea
Dpto. de Informática
Univ. of Castilla-La Mancha
Campus Universitario s/n
02071 Albacete, Spain

Franz J. Och and **Hermann Ney**
Lehrstuhl für Inf. VI
RWTH Aachen
Ahornstr., 55
D-52056 Aachen, Germany

Francisco Casacuberta
Dpto. de Sist. Inf. y Comp.
Inst. Tecn. de Inf. (UPV)
Avda. de Los Naranjos, s/n
46071 Valencia, Spain

Abstract

Typically, statistical alignment models are based on single-word dependencies. These models do not include contextual information, which can lead to inadequate alignments. In this paper, we present an approach to include contextual dependencies in the statistical alignment model by using a refined lexicon model. Unlike previous work, we directly integrate this in the EM algorithm of statistical alignment models. Experimental results are given for the French-English Canadian Parliament Hansards task and the Verbmobil task. The evaluation is performed by comparing the obtained alignments with a manually annotated reference alignment.

1 Introduction

The performance of a statistical machine translation system depends directly on the quality of the lexicon and the alignment models used. So far, most of the statistical machine translation systems are based on single-word alignment models as described in (Brown et al., 1993). Typically, the lexicon models used in these systems do not include any linguistic or contextual information, which often yields inadequate alignments in pairs of sentences. In this paper, we present an approach to improve the quality of the word-to-word alignments for this family of statistical translation models by using a maximum entropy (ME) approach. We define a set of context-dependent ME lexicon models, which is directly integrated into a conventional EM training of statistical alignment models. Experimental results are given for the French-English Canadian Parliament Hansards corpus and the Verbmobil task. The evaluation is performed by comparing the obtained alignment with a manually annotated reference alignment.

The ME approach has been applied in natural language processing and machine translation to a variety of tasks. Berger et al. (1996) applies this approach to the so-called IBM Candide system to build context-dependent models, to compute automatic sentence splitting and to improve word reordering in translation. García-Varea et al. (2001) use ME models to reduce translation test perplexities and translation errors by means of a rescoring algorithm, which is applied to n-best translation hypotheses. Foster (2000) describes two methods for incorporating information about the relative position of bilingual word pairs into a ME translation model.

2 Statistical machine translation

The goal of the translation process in statistical machine translation can be formulated as follows: A source language string $\mathbf{f} = f_1^J = f_1 \dots f_J$ is to be translated into a target language string $\mathbf{e} = e_1^I = e_1 \dots e_I$. Every target string is regarded as a possible translation for the source language string with maximum a-posteriori probability $Pr(\mathbf{e}|\mathbf{f})$. According to Bayes' decision rule, we have to choose the target string that maximizes the product of both the target language model $Pr(\mathbf{e})$ and the string translation model $Pr(\mathbf{f}|\mathbf{e})$.

Alignment models to structure the translation model are introduced in (Brown et al., 1993). These alignment models are similar to the concept of Hidden Markov models (HMM) in speech recognition. The alignment mapping is $j \rightarrow i = a_j$ from source position j to target position $i = a_j$. In statistical alignment models, $Pr(\mathbf{f}, \mathbf{a}|\mathbf{e})$, the alignment $\mathbf{a}$ is introduced as a hidden variable.

* This work has been partially supported by Spanish CICYT under grant TIC2000-1599-C02-01

The translation probability $Pr(\mathbf{f}, \mathbf{a}|\mathbf{e})$ can be rewritten as follows:

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

Typically, the probability $Pr(f_j | f_1^{j-1}, a_1^j, e_1^I)$ is approximated to by a lexicon model $p(f_j | e_{a_j})$ by dropping the dependencies on f_1^{j-1}, a_1^{j-1}, and $e_1^I \backslash e_{a_j}$. Obviously, this simplification is not true for many natural language phenomena. The straightforward approach to include more dependencies in the lexicon model would be to add additional dependencies (e.g. $p(f_j | e_{a_j}, e_{a_{j-1}})$). This approach would yield a significant data sparseness problem.

3 EM training of simple alignment models (review)

In this section, we describe the training of the model parameters. Every model has a specific set of free parameters. For example, the parameters θ for Model 4 of (Brown et al., 1993), consist of alignment parameters $p_{alig}(\cdot)$ and fertility parameters $p_{fert}(\cdot)$ in addition to the lexicon parameters $p(f|e)$:

$$\theta = \left\{ \{p(f|e)\}, \{p_{alig}(\cdot)\}, \{p_{fert}(\cdot)\} \right\} \quad (2)$$

To train the model parameters θ, we pursue a maximum likelihood approach using a parallel training corpus consisting of S sentence pairs $\{(\mathbf{f}_s, \mathbf{e}_s) : s = 1, \ldots, S\}$:

$$\hat{\theta} = \arg\max_{\theta} \prod_{s=1}^{S} \sum_{\mathbf{a}} p_\theta(\mathbf{f}_s, \mathbf{a}|\mathbf{e}_s) \quad (3)$$

We do this by applying the EM algorithm (Baum, 1972). The different models are trained in succession on the same data, where the final parameter values of a simpler model serve as the starting point for a more complex model.

In the E-step, the lexicon parameter counts for one sentence pair $(\mathbf{e}, \mathbf{f})$ are calculated:

$$\begin{aligned} c(f|e; \mathbf{e}, \mathbf{f}) &= N(\mathbf{e}, \mathbf{f}) \cdot \sum_{\mathbf{a}} Pr(\mathbf{a}|\mathbf{e}, \mathbf{f}) \\ & \qquad \sum_{j} \delta(f, f_j) \delta(e, e_{a_j}) \end{aligned} \quad (4)$$

Here, $N(\mathbf{e}, \mathbf{f})$ is the training corpus count of the sentence pair $(\mathbf{f}, \mathbf{e})$.

In the M-step, we want to compute the lexicon parameters $\hat{p}(f|e)$ that maximize the likelihood of the training corpus. This results in the following re-estimation (Brown et al., 1993):

$$p(f|e) = \frac{\sum_s c(f|e; \mathbf{f}^{(s)}, \mathbf{e}^{(s)})}{\sum_{s,f} c(f|e; \mathbf{f}^{(s)}, \mathbf{e}^{(s)})} \quad (5)$$

Similarly, the alignment and fertility probabilities can be estimated for all other alignment models (Brown et al., 1993). When bootstrapping from a simpler model to a more complex model, the simpler model is used to weigh the alignments and the counts are accumulated for the parameters of the more complex model.

4 Maximum entropy modeling

Here, the role of ME is to build a stochastic model that efficiently takes a larger context into account. In the remainder of the paper, we shall use $p_e(f|x)$ to denote the probability that the ME model (which is associated to e) assigns to f in the context x. Actually, the context x refers to the dropped dependencies. Please note that the ME model must be distinguished by the basic lexicon model $p(f|e)$.

In the ME approach, we describe all properties that we deem to be useful by so-called feature functions $\phi_{e,k}(x, f), k = 1, \ldots, K_e$. For example, let us suppose we want to model the existence or absence of a specific word e'_k in the context of an English word e, which can be translated by f'_k. We can express this dependence using the following feature function:

$$\phi_{e,k}(x, f) = \begin{cases} 1 & \text{if } f = f'_k \text{ and } e'_k \in x \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Consequently the k-th feature for word e has associated the pair (e'_k, f'_k).

The ME principle suggests that the optimal parametric form of a model $p_e(f|x)$ taking into account the feature functions $\phi_{e,k}, k = 1, \ldots, K_e$ is given by:

$$p_e(f|x) = \frac{1}{Z_{\Lambda_e}(x)} \exp\left(\sum_{k=1}^{K_e} \lambda_{e,k}\phi_{e,k}(x, f)\right) \quad (7)$$

Here, $Z_{\Lambda_e}(x)$ is a normalization factor. The resulting model has an exponential form with free parameters $\Lambda_e \equiv \{\lambda_{e,k}, k = 1, \ldots, K_e\}$. The parameter values that maximize the likelihood for a given training corpus can be computed using the so-called GIS algorithm (generalized iterative scaling)(Darroch and Ratcliff, 1972) or its improved version IIS (Pietra et al., 1997; Berger et al., 1996).

It is important to stress that, in principle, we obtain one ME model for each target language word e. To avoid data sparseness problems for rarely seen words, we use only words that have been seen a certain number of times.

5 Contextual information and feature definition

Berger et al. (1996) use a window of 3 words to the left and 3 words to the right of the target word as contextual information. As in (García-Varea et al., 2001), in addition to a dependence on the words themselves, we also use a dependence on the word classes. We thereby, improve the generalization of the models and include some semantic and syntactic information. The word classes are computed automatically using the approach described in (Och, 1999).

Table 1 summarizes the feature functions that we use for a specific pair of aligned words (e_i, f_j): Category 1 features depend only on the source word f_j and the target word e_i. Categories 2 and 3 describe features that also depend on an additional word e' that appears one position to the left or to the right of e_i, respectively. The features of category 4 and 5 depend on an additional target word e' that appears in any position of the context x. Analogous features are defined using the word class associated to each word instead of the word identity.

To reduce the number of features, we perform a threshold-based feature selection. Any feature that occurs less than T times is not used. The aim of the feature selection is two-fold. Firstly, we obtain smaller models by using fewer features. Secondly, we hope to avoid overfitting on the training data. In addition, we use ME modeling for target words that are seen at least 150 times.

6 Training of refined alignment models

6.1 Basic/Dynamic approach

Using a ME lexicon model for a target word e, we have to train the model parameters $\Lambda_e \equiv \{\lambda_{e,k} : k = 1, \ldots, K_e\}$ instead of the parameters $\{p(f|e)\}$. We pursue the following approach. In the E-step, we perform a refined count collection for the lexicon parameters:

$$c(f|e, x; \mathbf{e}, \mathbf{f}) = N(\mathbf{e}, \mathbf{f}) \cdot \sum_{\mathbf{a}} Pr(\mathbf{a}|\mathbf{e}, \mathbf{f}) \sum_j \delta(f, f_j)\delta(e, e_{a_j})\delta(x, x_{j,a_j}) \quad (8)$$

Here, x_{j,a_j} should denote the ME context that surrounds f_j and e_{a_j}.

In the M-step, we want to compute the lexicon parameters that maximize the likelihood:

$$\hat{\Lambda}_e = \arg\max_{\Lambda_e} \prod_{f,x} c(f|e, x; \mathbf{e}, \mathbf{f}) \cdot \log p_e(f|x) \quad (9)$$

Hence, the refined lexicon counts $c(f|e, x; \mathbf{e}, \mathbf{f})$ are the weights of the set of training samples (f, e, x) which is used to train the ME model.

The re-estimation of the alignment and fertility probabilities does not change if we use a ME lexicon model.

Thus, we obtain the following steps of each iteration for the EM algorithm:

1. E-step:
   - Collect counts for alignment and fertility parameters.
   - Collect refined lexicon counts.
2. M-step:
   - Re-estimate alignment and fertility parameters.
   - Perform GIS training for lexicon parameters.

6.2 Simplification: Static approach

A simplification of the approach described above can be obtained in the following way:

Table 1: Meaning of different feature categories where $\square$ represents a specific target word (to be placed in $\bullet$) and $\diamond$ represents a specific source word, where k has associated the pair ($\square$,$\diamond$).

Category	$\phi_{e_i,k}(x, f_j) = 1$ if and only if ...
1	$f_j = \diamond$
2	$f_j = \diamond$ and $\square \in$ [ ][ ][$\bullet$][e_i][ ][ ][ ]
3	$f_j = \diamond$ and $\square \in$ [ ][ ][ ][e_i][$\bullet$][ ][ ]
4	$f_j = \diamond$ and $\square \in$ [$\bullet$][$\bullet$][$\bullet$][e_i][ ][ ][ ]
5	$f_j = \diamond$ and $\square \in$ [ ][ ][ ][e_i][$\bullet$][$\bullet$][$\bullet$]

First, perform a normal training of the EM algorithm. Then, after the final iteration, perform the ME training only once. Finally, a new EM training is performed where the lexicon parameters are fixed to the ME lexicon models obtained previously. This is why we call the basic approach the dynamic approach as well.

6.3 Avoiding overfitting

ME modeling is maximum likelihood training for exponential models (Berger et al., 1996). As with other maximum likelihood methods, we have to deal with the problem of overfitting on the training data. To address this problem, we usually apply smoothing. We perform a linear interpolation of the baseline lexicon model with the ME lexicon model:

$$p'_e(f|x) = \lambda \cdot p_e(f|x) + (1 - \lambda) \cdot p(f|e) \qquad (10)$$

The interpolation parameter λ is optimized during training using held-out data. Hence, we choose the λ that maximizes the log-likelihood of the test data. The value of λ obtained in the results presented is 0.5.

Overfitting in the GIS training should also be avoided. Therefore, we stop the training if the change in training perplexity from one iteration to the next is below a certain threshold. This threshold is adjusted empirically by taking into account the perplexity on a test corpus.

6.4 Comparison of the different approaches

In this work, the type of features and contexts used are very similar to those used in (Berger et al., 1996) and (García-Varea et al., 2001). In these studies, the ME models were obtained after the normal training of the translation models. These models had no effect on the training of the statistical alignment models itself. Thus, only a refined lexicon model was obtained, but the fertility and alignment model were not changed. In this work, the ME models are used and/or trained within the EM training to obtain a better set of parameters. In this work, all the other models (namely alignment and fertility models) are also indirectly improved thanks to the refined context-dependent lexicon parameters.

The dynamic/basic approach gives us a more feasible parameter estimation than the static approach. In the dynamic approach, we do not know the Viterbi alignment of a given pair of sentences during EM training. This leads to the problem of constructing/extracting the corresponding training sample for the defined ME model training. To solve this problem, the set of all possible alignments for each sentence pair is considered.

Static training has the following advantages: the training time is faster because only one ME training has to be performed; a bootstrapping strategy of refinement could be applied. Hence, iterate the process of: "EM training $\rightarrow$ use the Viterbi alignment to train the ME models $\rightarrow$ repeat the EM training using the last ME models $\rightarrow$ ...", and so on.

On the other hand, dynamic training has the following advantages: a tight and feasible integration is provided; a refined set of ME models is obtained in each iteration of the EM algorithm; the set of p_e models considered is refined from one iteration to another in the same way as the parameters of the other models.

7 Evaluation methodology

We use the same annotation scheme for single-word-based alignments and a corresponding evaluation criterion as described in (Och and Ney, 2000). The annotation scheme explicitly allows for ambiguous alignments. The people

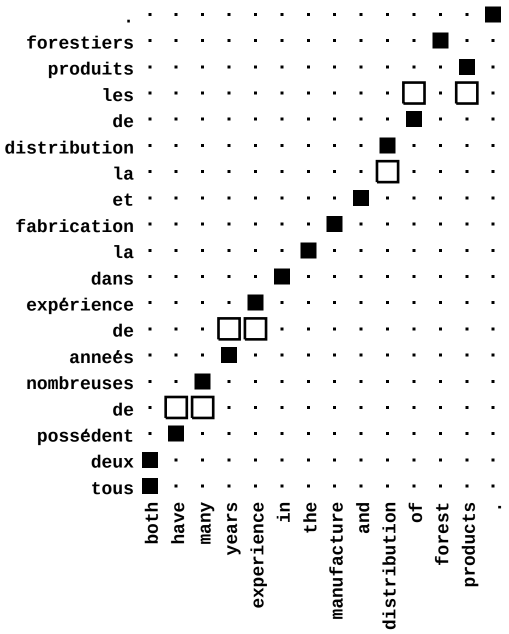

Figure 1: Example of a manual alignment with *S(ure)* (■)and *P(ossible)* (□) connections.

performing the annotation are asked to specify two different kinds of alignments: an *S(ure)* alignment, which is used for alignments that are unambiguous and a *P(ossible)* alignment, which is used for ambiguous alignments. The P label is used particularly to align words within idiomatic expressions, free translations, and missing function words ($S \subseteq P$).

The reference alignment thus obtained may contain many-to-one and one-to-many relationships. Figure 1 shows an example of a manually aligned sentence with S and P labels.

The quality of an alignment $A = \{(j, a_j) | a_j > 0\}$ is then computed by appropriately redefined precision and recall measures:

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

and the following alignment error rate, which is derived from the well known F-measure:

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

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

The set of sentence pairs, for which the manual alignment is produced, is randomly selected from the training corpus. It should be emphasized that all the training is done in a completely unsupervised way, i.e. no manual alignments are used. From this point of view, there is no need to have a separate test corpus.

8 Experimental results

We show results on the Verbmobil task and the Hansards task. The Verbmobil task is a speech translation task in the domain of appointment scheduling, travel planning, and hotel reservation. The task is difficult because it consists of spontaneous speech and the syntactic structures of the sentences are less restricted and highly variable. The French-English Hansards task consists of the debates in the Canadian Parliament. This task has a very large vocabulary of more than 100,000 French words.

The corpus statistics are shown in Table 2. The number of running words and the vocabularies are based on full-form words including the punctuation marks. We produced smaller training corpora by randomly choosing 500, 8000 and 34000 sentences from the Verbmobil task and 500, 8000 and 128000 sentences from the Hansards task.

To train the context-dependent statistical alignment models, we extended the publicly available toolkit GIZA++ (Och and Ney, 2001). The training of the ME models was carried out using the YASMET toolkit (Och, 2002).

All the results shown in this paper were obtained using the static ME integration.

Table 3 and Table 4 show the alignment quality for different training sample sizes of the Hansards and Verbmobil tasks, respectively. These tables show the baseline AER for different training schemes and the corresponding values when the integration of the ME is done. The training scheme is defined in accordance with the number of iterations performed for each model (4^3 means 3 iterations of Model 4). In all the experiments, we started applying the ME models in the first iteration of Model 1.

The recall and precision results for the Hansards task with and without ME training are shown in Figures 2 and 3.

We observe that the alignment error rate im-

Table 2: Corpus characteristics.

		Verbmobil		Hansards	
		German	English	French	English
Train	Sentences	34446		1470K	
	Words	329625	343076	24.33M	22.16M
	Vocabulary	5936	3505	100269	78332

Table 3: AER [%] on Hansards task.

		Size of train corpus		
Training	Model	0.5K	8K	128K
1^5	1	48.0	35.1	29.2
	1+ME	47.7	32.7	22.5
$1^5 2^5$	2	46.0	29.2	21.9
	2+ME	44.7	28.0	19.0
$1^5 2^5 3^3$	3	43.2	27.3	20.8
	3+ME	42.5	26.4	17.2
$1^5 2^5 3^3 4^3$	4	41.8	24.9	17.4
	4+ME	41.5	24.3	14.1
$1^5 2^5 3^3 4^3 5^3$	5	41.5	24.8	16.2
	5+ME	41.5	24.5	14.3

Table 4: AER [%] on Verbmobil task.

		Size of train corpus		
Training	Model	0.5K	8K	34K
1^5	1	27.7	19.2	17.6
	1+ME	24.6	16.6	13.7
$1^5 2^5$	2	26.8	15.7	13.5
	2+ME	25.3	14.1	10.8
$1^5 2^5 3^3$	3	25.6	13.7	10.8
	3+ME	24.1	11.6	8.8
$1^5 2^5 3^3 4^3$	4	23.6	10.0	7.7
	4+ME	22.8	9.3	7.0
$1^5 2^5 3^3 4^3 5^3$	5	22.6	9.9	7.2
	5+ME	22.3	9.6	6.8

proves when using the context-dependent lexicon models. For the Verbmobil task, the improvements were smaller than for the Hansards task, which might be due to the fact that the baseline alignment quality was already very good. It can be seen that greater improvements were obtained for the simpler models.

As expected, ME training plays a more important role when larger sizes of the corpus are used. For the smallest corpora, the number of training events for the ME models is very low, so it is not possible to disambiguate some translations/alignments for different contexts. For larger sizes of the corpora, greater improvements are obtained. Therefore, we expect to obtain better improvements when using even larger corpora.

After observing the common alignment errors, we plan to include more discriminanting features that would provide greater improvements. We also expect improvements by performing a refined modeling of the rare/infrequent words, which are currently not taken into account by the ME models.

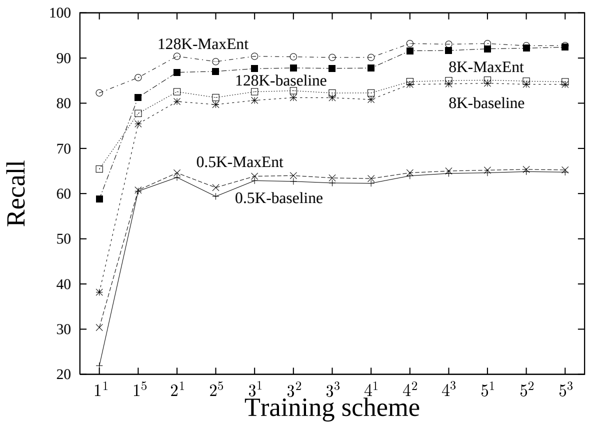

Figure 2: Recall [%] results for Hansards task for different corpus sizes.

9 Conclusions

In this paper, we show an efficient and straightforward integration of ME context-dependent models within a maximum likelihood training of statistical translation models.

We evaluate the quality of the alignments obtained with this new training scheme comparing the results with the baseline results. As can

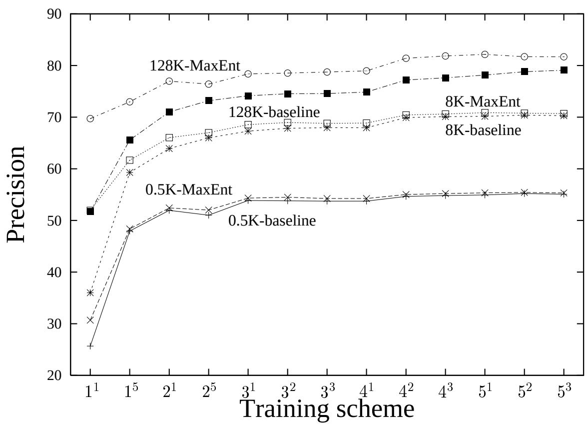

Figure 3: Precision [%] results for Hansards task for different corpus sizes.

be seen in Section 8, we obtain better alignment quality using the context-dependent lexicon model.

In the future, we plan to include more features in the ME model, such us dependencies with other source and target words, POS tags and syntactic constituents. We also plan to design ME alignment and fertility models. This will allow for an easy integration of more dependencies, such as second-order alignment models without running into the problem of an unmanageable number of alignment parameters. We have just started to perform experiments for a very distant pair of languages as is Chinese-English with very promising results.

References

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

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72, March.

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.

J.N. Darroch and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43:95–144.

George Foster. 2000. Incorporating position information into a maximum entropy/minimum divergence translation model. In *Proc. of CoNNL-2000 and LLL-2000*, pages 37–52, Lisbon, Portugal.

Ismael García-Varea, Franz J. Och, Hermann Ney, and Francisco Casacuberta. 2001. Refined lexicon models for statistical machine translation usign a maximum entropy approach. In *Proc. of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 204–211, Toulouse, France, July.

Franz J. Och and Hermann Ney. 2000. 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.

Franz J. Och and Hermann Ney. 2001. Giza++: Training of statistical translation models. `http://www-i6.Informatik.RWTH--Aachen.DE/~och/software/GIZA++.html`.

Franz J. Och. 1999. An efficient method for determining bilingual word classes. In *EACL '99: Ninth Conf. of the Europ. Chapter of the Association for Computational Linguistics*, pages 71–76, Bergen, Norway, June.

Franz J. Och. 2002. Yet another small maxent toolkit: Yasmet. `http://www-i6.Informatik.RWTH--Aachen.DE/~och/software/YASMET.html`.

Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. 1997. Inducing features in random fields. *IEEE Trans. on Pattern Analysis and Machine Inteligence*, 19(4):380–393, July.