Statistical Machine Translation with Word- and Sentence-Aligned Parallel Corpora

Chris Callison-Burch David Talbot Miles Osborne
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
callison-burch@ed.ac.uk

Abstract

The parameters of statistical translation models are typically estimated from sentence-aligned parallel corpora. We show that significant improvements in the alignment and translation quality of such models can be achieved by additionally including word-aligned data during training. Incorporating word-level alignments into the parameter estimation of the IBM models reduces alignment error rate and increases the Bleu score when compared to training the same models only on sentence-aligned data. On the Verbmobil data set, we attain a 38% reduction in the alignment error rate and a higher Bleu score with half as many training examples. We discuss how varying the ratio of word-aligned to sentence-aligned data affects the expected performance gain.

1 Introduction

Machine translation systems based on probabilistic translation models (Brown et al., 1993) are generally trained using *sentence-aligned* parallel corpora. For many language pairs these exist in abundant quantities. However for new domains or uncommon language pairs extensive parallel corpora are often hard to come by.

Two factors could increase the performance of statistical machine translation for new language pairs and domains: a reduction in the cost of creating new training data, and the development of more efficient methods for exploiting existing training data. Approaches such as harvesting parallel corpora from the web (Resnik and Smith, 2003) address the creation of data. We take the second, complementary approach. We address the problem of efficiently exploiting existing parallel corpora by adding explicit *word-level* alignments between a number of the sentence pairs in the training corpus. We modify the standard parameter estimation procedure for IBM Models and HMM variants so that they can exploit these additional word-level alignments. Our approach uses both word- and sentence-level alignments for training material.

In this paper we:

1. Describe how the parameter estimation framework of Brown et al. (1993) can be adapted to incorporate word-level alignments;
2. Report significant improvements in alignment error rate and translation quality when training on data with word-level alignments;
3. Demonstrate that the inclusion of word-level alignments is more effective than using a bilingual dictionary;
4. Show the importance of amplifying the contribution of word-aligned data during parameter estimation.

This paper shows that word-level alignments improve the parameter estimates for translation models, which in turn results in improved statistical translation for languages that do not have large sentence-aligned parallel corpora.

2 Parameter Estimation Using Sentence-Aligned Corpora

The task of statistical machine translation is to choose the source sentence, $\mathbf{e}$, that is the most probable translation of a given sentence, $\mathbf{f}$, in a foreign language. Rather than choosing $\mathbf{e}^*$ that directly maximizes $p(\mathbf{e}|\mathbf{f})$, Brown et al. (1993) apply Bayes' rule and select the source sentence:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} p(\mathbf{e})p(\mathbf{f}|\mathbf{e}). \quad (1)$$

In this equation $p(\mathbf{e})$ is a language model probability and is $p(\mathbf{f}|\mathbf{e})$ a translation model probability. A series of increasingly sophisticated translation models, referred to as the IBM Models, was defined in Brown et al. (1993).

The translation model, $p(\mathbf{f}|\mathbf{e})$ defined as a marginal probability obtained by summing over word-level alignments, $\mathbf{a}$, between the source and target sentences:

$$p(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} p(\mathbf{f}, \mathbf{a}|\mathbf{e}). \quad (2)$$

While word-level alignments are a crucial component of the IBM models, the model parameters are generally estimated from sentence-aligned parallel corpora without explicit word-level alignment information. The reason for this is that word-aligned parallel corpora do not generally exist. Consequently, word level alignments are treated as hidden variables. To estimate the values of these hidden variables, the expectation maximization (EM) framework for maximum likelihood estimation from incomplete data is used (Dempster et al., 1977).

EM seeks to maximize the marginal log likelihood, $\log p(\mathbf{f}|\mathbf{e})$, indirectly by iteratively maximizing a bound on this term known as the *expected complete log likelihood*, $\langle \log p(\mathbf{f}, \mathbf{a}|\mathbf{e}) \rangle_{q(\mathbf{a})}$,[1]

$$\begin{aligned} \log p(\mathbf{f}|\mathbf{e}) &= \log \sum_{\mathbf{a}} p(\mathbf{f}, \mathbf{a}|\mathbf{e}) && (3) \\ &= \log \sum_{\mathbf{a}} q(\mathbf{a}) \frac{p(\mathbf{f}, \mathbf{a}|\mathbf{e})}{q(\mathbf{a})} && (4) \\ &\geq \sum_{\mathbf{a}} q(\mathbf{a}) \log \frac{p(\mathbf{f}, \mathbf{a}|\mathbf{e})}{q(\mathbf{a})} && (5) \\ &= \langle \log p(\mathbf{f}, \mathbf{a}|\mathbf{e}) \rangle_{q(\mathbf{a})} + H(q(\mathbf{a})) \end{aligned}$$

where the bound in (5) is given by Jensen's inequality. By choosing $q(\mathbf{a}) = p(\mathbf{a}|\mathbf{f}, \mathbf{e})$ this bound becomes an equality.

This maximization consists of two steps:

- E-step: calculate the posterior probability under the current model of every permissible alignment for each sentence pair in the sentence-aligned training corpus;
- M-step: maximize the expected log likelihood under this posterior distribution, $\langle \log p(\mathbf{f}, \mathbf{a}|\mathbf{e}) \rangle_{q(\mathbf{a})}$, with respect to the model's parameters.

While in standard maximum likelihood estimation events are counted directly to estimate parameter settings, in EM we effectively collect *fractional* counts of events (here permissible alignments weighted by their posterior probability), and use these to iteratively update the parameters.

Since only some of the permissible alignments make sense linguistically, we would like EM to use the posterior alignment probabilities calculated in the E-step to weight plausible alignments higher than the large number of bogus alignments which are included in the expected complete log likelihood. This in turn should encourage the parameter adjustments made in the M-step to converge to linguistically plausible values.

Since the number of permissible alignments for a sentence grows exponentially in the length of the sentences for the later IBM Models, a large number of *informative* example sentence pairs are required to distinguish between plausible and implausible alignments. Given sufficient data the distinction occurs because words which are mutual translations appear together more frequently in aligned sentences in the corpus.

Given the high number of model parameters and permissible alignments, however, huge amounts of data will be required to estimate reasonable translation models from sentence-aligned data alone.

3 Parameter Estimation Using Word- and Sentence-Aligned Corpora

As an alternative to collecting a huge amount of sentence-aligned training data, by annotating some of our sentence pairs with word-level alignments we can explicitly provide information to highlight plausible alignments and thereby help parameters converge upon reasonable settings with less training data.

Since word-alignments are inherent in the IBM translation models it is straightforward to incorporate this information into the parameter estimation procedure. For sentence pairs with explicit word-level alignments marked, fractional counts over all permissible alignments need not be collected. Instead, whole counts are collected for the single hand annotated alignment for each sentence pair which has been word-aligned. By doing this the expected complete log likelihood collapses to a single term, the *complete log likelihood* ($p(\mathbf{f}, \mathbf{a}|\mathbf{e})$), and the E-step is circumvented.

The parameter estimation procedure now involves maximizing the likelihood of data aligned only at the sentence level and also of data aligned at the word level. The mixed likelihood function, $\mathcal{M}$, combines the expected information contained in the sentence-aligned data with the complete information contained in the word-aligned data.

$$\mathcal{M} = \sum_{s=1}^{N_s} (1 - \lambda) \langle \log p(\mathbf{f}_s, \mathbf{a}_s|\mathbf{e}_s) \rangle_{q(\mathbf{a}_s)}$$

[1] Here $\langle \cdot \rangle_{q(\cdot)}$ denotes an expectation with respect to $q(\cdot)$.

$$+ \sum_{w=1}^{N_w} \lambda \log p(\mathbf{f}_w, \mathbf{a}_w | \mathbf{e}_w) \quad (6)$$

Here s and w index the N_s sentence-aligned sentences and N_w word-aligned sentences in our corpora respectively. Thus $\mathcal{M}$ combines the expected complete log likelihood and the complete log likelihood. In order to control the relative contributions of the sentence-aligned and word-aligned data in the parameter estimation procedure, we introduce a mixing weight λ that can take values between 0 and 1.

3.1 The impact of word-level alignments

The impact of word-level alignments on parameter estimation is closely tied to the structure of the IBM Models. Since translation and word alignment parameters are shared between all sentences, the posterior alignment probability of a source-target word pair in the sentence-aligned section of the corpus that were aligned in the word-aligned section will tend to be relatively high.

In this way, the alignments from the word-aligned data effectively percolate through to the sentence-aligned data indirectly *constraining* the E-step of EM.

3.2 Weighting the contribution of word-aligned data

By incorporating λ, Equation 6 becomes an interpolation of the expected complete log likelihood provided by the sentence-aligned data and the complete log likelihood provided by word-aligned data.

The use of a weight to balance the contributions of unlabeled and labeled data in maximum likelihood estimation was proposed by Nigam et al. (2000). λ quantifies our relative confidence in the expected statistics and observed statistics estimated from the sentence- and word-aligned data respectively.

Standard maximum likelihood estimation (MLE) which weighs all training samples equally, corresponds to an implicit value of lambda equal to the proportion of word-aligned data in the whole of the training set: $\lambda = \frac{N_w}{N_w + N_s}$. However, having the total amount of sentence-aligned data be much larger than the amount of word-aligned data implies a value of λ close to zero. This means that $\mathcal{M}$ can be maximized while essentially ignoring the likelihood of the word-aligned data. Since we believe that the explicit word-alignment information will be highly effective in distinguishing plausible alignments in the corpus as a whole, we expect to see benefits by setting λ to amplify the contribution of the word-aligned data set particularly when this is a relatively small portion of the corpus.

4 Experimental Design

To perform our experiments with word-level alignements we modified GIZA++, an existing and freely available implementation of the IBM models and HMM variants (Och and Ney, 2003). Our modifications involved circumventing the E-step for sentences which had word-level alignments and incorporating these observed alignment statistics in the M-step. The observed and expected statistics were weighted accordingly by λ and $(1 - \lambda)$ respectively as were their contributions to the mixed log likelihood.

In order to measure the accuracy of the predictions that the statistical translation models make under our various experimental settings, we choose the alignment error rate (AER) metric, which is defined in Och and Ney (2003). We also investigated whether improved AER leads to improved translation quality. We used the alignments created during our AER experiments as the input to a phrase-based decoder. We translated a test set of 350 sentences, and used the Bleu metric (Papineni et al., 2001) to automatically evaluate machine translation quality.

We used the Verbmobil German-English parallel corpus as a source of training data because it has been used extensively in evaluating statistical translation and alignment accuracy. This data set comes with a manually word-aligned set of 350 sentences which we used as our test set.

Our experiments additionally required a very large set of word-aligned sentence pairs to be incorporated in the training set. Since previous work has shown that when training on the complete set of 34,000 sentence pairs an alignment error rate as low as 6% can be achieved for the Verbmobil data, we automatically generated a set of alignments for the entire training data set using the unmodified version of GIZA++. We wanted to use automatic alignments in lieu of actual hand alignments so that we would be able to perform experiments using large data sets. We ran a pilot experiment to test whether our automatic would produce similar results to manual alignments.

We divided our manual word alignments into training and test sets and compared the performance of models trained on human aligned data against models trained on automatically aligned data. A 100-fold cross validation showed that manual and automatic alignments produced AER results that were similar to each other to within 0.1%.[2]

[2]Note that we stripped out probable alignments from our

	Size of training corpus			
Model	.5k	2k	8k	16k
Model 1	29.64	24.66	22.64	21.68
HMM	18.74	15.63	12.39	12.04
Model 3	26.07	18.64	14.39	13.87
Model 4	20.59	16.05	12.63	12.17

Table 1: Alignment error rates for the various IBM Models trained with sentence-aligned data

	Size of training corpus			
Model	.5k	2k	8k	16k
Model 1	21.43	18.04	16.49	16.20
HMM	14.42	10.47	9.09	8.80
Model 3	20.56	13.25	10.82	10.51
Model 4	14.19	10.13	7.87	7.52

Table 2: Alignment error rates for the various IBM Models trained with word-aligned data

Having satisfied ourselves that automatic alignment were a sufficient stand-in for manual alignments, we performed our main experiments which fell into the following categories:

1. Verifying that the use of word-aligned data has an impact on the quality of alignments predicted by the IBM Models, and comparing the quality increase to that gained by using a bilingual dictionary in the estimation stage.

2. Evaluating whether improved parameter estimates of alignment quality lead to improved translation quality.

3. Experimenting with how increasing the ratio of word-aligned to sentence-aligned data affected the performance.

4. Experimenting with our λ parameter which allows us to weight the relative contributions of the word-aligned and sentence-aligned data, and relating it to the ratio experiments.

5. Showing that improvements to AER and translation quality held for another corpus.

5 Results

5.1 Improved alignment quality

As a staring point for comparison we trained GIZA++ using four different sized portions of the Verbmobil corpus. For each of those portions we output the most probable alignments of the testing data for Model 1, the HMM, Model 3, and Model 4,[3] and evaluated their AERs. Table 1 gives alignment error rates when training on 500, 2000, 8000, and 16000 sentence pairs from Verbmobil corpus without using any word-aligned training data.

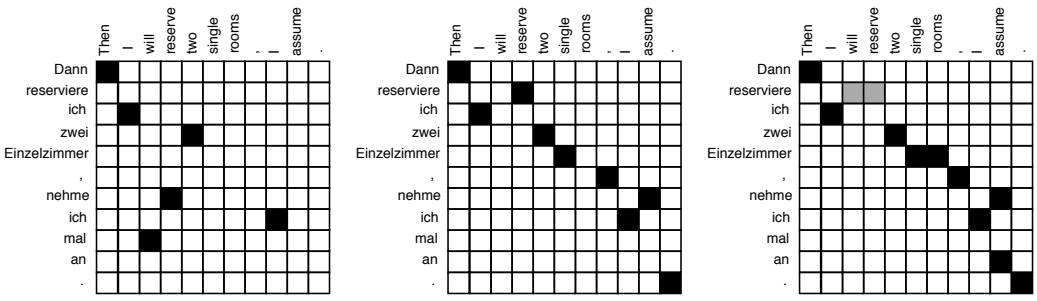
We obtained much better results when incorporating word-alignments with our mixed likelihood function. Table 2 shows the results for the different corpus sizes, when *all* of the sentence pairs have been word-aligned. The best performing model in the unmodified GIZA++ code was the HMM trained on 16,000 sentence pairs, which had an alignment error rate of 12.04%. In our modified code the best performing model was Model 4 trained on 16,000 sentence pairs (where all the sentence pairs are word-aligned) with an alignment error rate of 7.52%. The difference in the best performing models represents a 38% relative reduction in AER. Interestingly, we achieve a lower AER than the best performing unmodified models using a corpus that is one-eight the size of the sentence-aligned data.

Figure 1 show an example of the improved alignments that are achieved when using the word aligned data. The example alignments were held out sentence pairs that were aligned after training on 500 sentence pairs. The alignments produced when the training on word-aligned data are dramatically better than when training on sentence-aligned data.

We contrasted these improvements with the improvements that are to be had from incorporating a *bilingual dictionary* into the estimation process. For this experiment we allowed a bilingual dictionary to constrain which words can act as translations of each other during the initial estimates of translation probabilities (as described in Och and Ney (2003)). As can be seen in Table 3, using a dictionary reduces the AER when compared to using GIZA++ without a dictionary, but not as dramatically as integrating the word-alignments. We further tried combining a dictionary with our word-alignments but found that

manually produced alignments. Probable alignments are large blocks of words which the annotator was uncertain of how to align. The many possible word-to-word translations implied by the manual alignments led to lower results than with the automatic alignments, which contained fewer word-to-word translation possibilities.

[3] We used the default training schemes for GIZA++, and left model smoothing parameters at their default settings.



(a) Sentence-aligned (b) Word-aligned (c) Reference

Figure 1: Example alignments using sentence-aligned training data (a), using word-aligned data (b), and a reference manual alignment (c)

	Size of training corpus			
Model	.5k	2k	8k	16k
Model 1	23.56	20.75	18.69	18.37
HMM	15.71	12.15	9.91	10.13
Model 3	22.11	16.93	13.78	12.33
Model 4	17.07	13.60	11.49	10.77

Table 3: The improved alignment error rates when using a dictionary instead of word-aligned data to constrain word translations

	Sentence-aligned		Word-aligned	
Size	AER	Bleu	AER	Bleu
500	20.59	0.211	14.19	0.233
2000	16.05	0.247	10.13	0.260
8000	12.63	0.265	7.87	0.278
16000	12.17	0.270	7.52	0.282

Table 4: Improved AER leads to improved translation quality

the dictionary results in only very minimal improvements over using word-alignments alone.

5.2 Improved translation quality

The fact that using word-aligned data in estimating the parameters for machine translation leads to better alignments is predictable. A more significant result is whether it leads to improved translation quality. In order to test that our improved parameter estimates lead to better translation quality, we used a state-of-the-art phrase-based decoder to translate a held out set of German sentences into English. The phrase-based decoder extracts phrases from the word alignments produced by GIZA++, and computes translation probabilities based on the

Ratio	AER when λ = Standard MLE	when λ = .9
0.1	11.73	9.40
0.2	10.89	8.66
0.3	10.23	8.13
0.5	8.65	8.19
0.7	8.29	8.03
0.9	7.78	7.78

Table 5: The effect of weighting word-aligned data more heavily that its proportion in the training data (corpus size 16000 sentence pairs)

frequency of one phrase being aligned with another (Koehn et al., 2003). We trained a language model using the 34,000 English sentences from the training set.

Table 4 shows that using word-aligned data leads to better translation quality than using sentence-aligned data. Particularly, significantly less data is needed to achieve a high Bleu score when using word alignments. Training on a corpus of 8,000 sentence pairs *with* word alignments results in a higher Bleu score than when training on a corpus of 16,000 sentence pairs *without* word alignments.

5.3 Weighting the word-aligned data

We have seen that using training data consisting of entirely word-aligned sentence pairs leads to better alignment accuracy and translation quality. However, because manually word-aligning sentence pairs costs more than just using sentence-aligned data, it is unlikely that we will ever want to label an entire corpus. Instead we will likely have a relatively small portion of the corpus word aligned. We want to be sure that this small amount of data

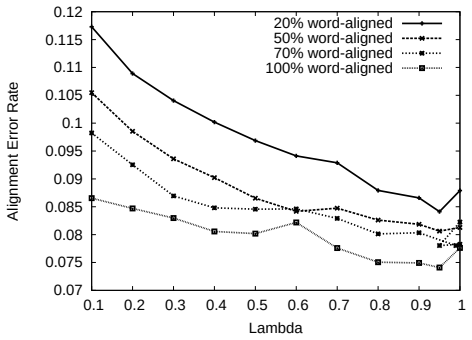


Figure 2: The effect on AER of varying λ for a training corpus of 16K sentence pairs with various proportions of word-alignments

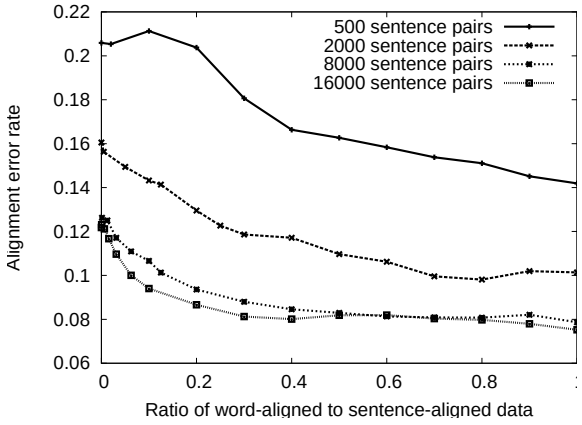


Figure 3: The effect on AER of varying the ratio of word-aligned to sentence-aligned data

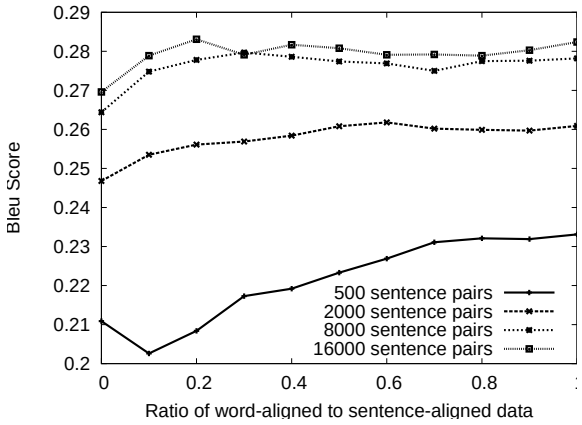


Figure 4: The effect on Bleu of varying the ratio of word-aligned to sentence-aligned data

labeled with word alignments does not get overwhelmed by a larger amount of unlabeled data. Thus we introduced the λ weight into our mixed likelihood function.

Table 5 compares the natural setting of λ (where it is proportional to the amount of labeled data in the corpus) to a value that amplifies the contribution of the word-aligned data. Figure 2 shows a variety of values for λ. It shows as λ increases AER decreases. Placing nearly all the weight onto the word-aligned data seems to be most effective.[4] Note this did not vary the training data size – only the relative contributions between sentence- and word-aligned training material.

5.4 Ratio of word- to sentence-aligned data

We also varied the ratio of word-aligned to sentence-aligned data, and evaluated the AER and Bleu scores, and assigned high value to λ (= 0.9).

Figure 3 shows how AER improves as more word-aligned data is added. Each curve on the graph represents a corpus size and shows its reduction in error rate as more word-aligned data is added. For example, the bottom curve shows the performance of a corpus of 16,000 sentence pairs which starts with an AER of just over 12% with no word-aligned training data and decreases to an AER of 7.5% when all 16,000 sentence pairs are word-aligned. This curve essentially levels off after 30% of the data is word-aligned. This shows that a small amount of word-aligned data is very useful, and if we wanted to achieve a low AER, we would only have to label 4,800 examples with their word alignments rather than the entire corpus.

[4] At $\lambda = 1$ (not shown in Figure 2) the data that is only sentence-aligned is ignored, and the AER is therefore higher.

Figure 4 shows how the Bleu score improves as more word-aligned data is added. This graph also reinforces the fact that a small amount of word-aligned data is useful. A corpus of 8,000 sentence pairs with only 800 of them labeled with word alignments achieves a higher Bleu score than a corpus of 16,000 sentence pairs with no word alignments.

5.5 Evaluation using a larger training corpus

We additionally tested whether incorporating word-level alignments into the estimation improved results for a larger corpus. We repeated our experiments using the Canadian Hansards French-English parallel corpus. Figure 6 gives a summary of the improvements in AER and Bleu score for that corpus, when testing on a held out set of 484 hand aligned sentences.

On the whole, alignment error rates are higher and Bleu scores are considerably lower for the Hansards corpus. This is probably due to the dif-

	Sentence-aligned		Word-aligned	
Size	AER	Bleu	AER	Bleu
500	33.65	0.054	25.73	0.064
2000	25.97	0.087	18.57	0.100
8000	19.00	0.115	14.57	0.120
16000	16.59	0.126	13.55	0.128

Table 6: Summary results for AER and translation quality experiments on Hansards data

ferences in the corpora. Whereas the Verbmobil corpus has a small vocabulary (<10,000 per language), the Hansards has ten times that many vocabulary items and has a much longer average sentence length. This made it more difficult for us to create a simulated set of hand alignments; we measured the AER of our simulated alignments at 11.3% (which compares to 6.5% for our simulated alignments for the Verbmobil corpus).

Nevertheless, the trend of decreased AER and increased Bleu score still holds. For each size of training corpus we tested we found better results using the word-aligned data.

6 Related Work

Och and Ney (2003) is the most extensive analysis to date of how many different factors contribute towards improved alignments error rates, but the inclusion of word-alignments is not considered. Och and Ney do not give any direct analysis of how improved word alignments accuracy contributes toward better translation quality as we do here.

Mihalcea and Pedersen (2003) described a shared task where the goal was to achieve the best AER. A number of different methods were tried, but none of them used word-level alignments. Since the best performing system used an unmodified version of Giza++, we would expected that our modifed version would show enhanced performance. Naturally this would need to be tested in future work.

Melamed (1998) describes the process of manually creating a large set of word-level alignments of sentences in a parallel text.

Nigam et al. (2000) described the use of weight to balance the respective contributions of labeled and unlabeled data to a mixed likelihood function. Corduneanu (2002) provides a detailed discussion of the instability of maximum likelihood solutions estimated from a mixture of labeled and unlabeled data.

7 Discussion and Future Work

In this paper we show with the appropriate modification of EM significant improvement gains can be had through labeling word alignments in a bilingual corpus. Because of this significantly less data is required to achieve a low alignment error rate or high Bleu score. This holds even when using noisy word alignments such as our automatically created set.

One should take our research into account when trying to efficiently create a statistical machine translation system for a language pair for which a parallel corpus is not available. Germann (2001) describes the cost of building a Tamil-English parallel corpus from scratch, and finds that using professional translations is prohibitively high. In our experience it is quicker to manually word-align translated sentence pairs than to translate a sentence, and word-level alignment can be done by someone who might not be fluent enough to produce translations. It might therefore be possible to achieve a higher performance at a fraction of the cost by hiring a non-professional produce word-alignments after a limited set of sentences have been translated.

We plan to investigate whether it is feasible to use *active learning* to select which examples will be most useful when aligned at the word-level. Section 5.4 shows that word-aligning a fraction of sentence pairs in a training corpus, rather than the entire training corpus can still yield most of the benefits described in this paper. One would hope that by selectively sampling which sentences are to be manually word-aligned we would achieve nearly the same performance as word-aligning the entire corpus.

Acknowledgements

The authors would like to thank Franz Och, Hermann Ney, and Richard Zens for providing the Verbmobil data, and Linear B for providing its phrase-based decoder. This research was supported by an MRC Priority Area Studentship to the School of Informatics, University of Edinburgh.

References

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

Adrian Corduneanu. 2002. Stable mixing of complete and incomplete information. Master's thesis, Massachusetts Institute of Technology, February.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, Nov.

Ulrich Germann. 2001. Building a statistical machine translation system from scratch: How much bang for the buck can we expect? In *ACL 2001 Workshop on Data-Driven Machine Translation*, Toulouse, France, July 7.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the HLT/NAACL*.

I. Dan Melamed. 1998. Manual annotation of translational equivalence: The blinker project. Cognitive Science Technical Report 98/07, University of Pennsylvania.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In Rada Mihalcea and Ted Pedersen, editors, *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts*.

Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. IBM Research Report RC22176(W0109-022), IBM.

Philip Resnik and Noah Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380, September.