LEVERAGING MULTIPLE LANGUAGES
TO IMPROVE STATISTICAL MT WORD ALIGNMENTS

Karim Filali and Jeff Bilmes

Departments of Computer Science & Engineering and Electrical Engineering
University of Washington
Seattle, WA 98195, USA
{karim@cs,bilmes@ee}.washington.edu

ABSTRACT

We present a new multilingual statistical MT word alignment model based on a simple extension of the IBM and HMM Models and a two-step alignment procedure. Preliminary results on a small hand-aligned subset of the Europarl corpus show a 7% relative improvement over a state of the art alignment model.

1. INTRODUCTION

The compilation of parallel bilingual corpora such as the *Canadian Parliamentary Proceedings (Hansards)* has undeniably contributed to the success of data-driven machine translation approaches. The recent availability of multilingual corpora (texts translated in three or more languages) such as *Europarl* [1] presents new possibilities for statistical models that exploit this potentially important source of information. One elusive goal of traditional machine translation has been to come up with an interlingual representation of language which would greatly simplify translation from any language to another. Data-driven approaches might take us closer to achieving that goal. In this paper we set out for a less ambitious project and try to answer the question

Can multilingual parallel translations help us learn better word alignment models than bilingual translations alone can?

It is not clear, a priori, that the answer would be affirmative. One could make the reasonable assumption that, when translating from language 1 to language 2, parallel sentences in language 3 do not add any information not already contained in corresponding sentences in language 1. It is however also clear that translations vary in quality and style and it might very well be that a judicious use of the parallel sentences can add to the robustness of our word alignments.

The central question this paper addresses is whether there is any benefit in using multilingual information even when

only bilingual translation is required. Another important question is how best to combine these parallel sources of information in a principled statistical framework. Here we opt for a simple modular solution that allows us to easily and efficiently extend current state of the art MT word alignment algorithms. This is especially important because, at least for some languages, parallel data are more and more abundant. The ability to scale up to large train set sizes is therefore very desirable.

There has been an increasing interest in exploiting multilingual resources in a variety of natural language applications. Two works are particularly relevant to ours. Och and Ney [2] train *separate* word alignment models for different target languages and a common source language. When inferring the optimal source language text given translations in the target languages, the translation models are combined to reduce search errors. We take a different tack and train a single joint alignment model that incorporates the information from several target languages at the same time. In [3], a trilingual parallel corpus is used to improve alignments between sentences (as opposed to words) in a text. While the two problems are different in terms of applicable alignment techniques, fundamentally they are both about how best to leverage parallel information that has a high degree of redundancy.

The paper is organized as follows: Section 2 briefly reviews the standard mathematics of bilingual word alignments following the presentation of Brown et al. [4]. Section 3 introduces our statistical multilingual word alignment model. We describe the experimental setup in section 4 and present our results in section 5.

2. MT AND WORD ALIGNMENTS

Suppose we want to translate a *target* string $\mathbf{f} = f_1^M = f_1 f_2 ... f_M$ to an *source*[1] string $\mathbf{e} = e_1^L = e_1 e_2 ... e_L$, where

---

[1] We follow the convention that we always translate from the target language to the source language according to the noisy channel naming convention. We think of the target string that we observe as a corrupted version of the source string we want to recover. This is also consistent with the

M and L are the lengths of the target and source sentences[2] respectively. In the standard statistical translation framework [5], the problem of finding the best source translation is modeled as recovering the most probable source string, $\hat{\mathbf{e}}$, which produces the target string after being sent through a noisy channel:

$$\hat{\mathbf{e}} = \underset{e_1^L}{\text{argmax}} \; P(f_1^M|e_1^L)P(e_1^L) \qquad (1)$$

where $P(f_1^M|e_1^L)$ is the learned translation model and $P(e_1^L)$ a language model.

Hidden alignment variables, a_1^M, are introduced to factorize $P(f_1^M|e_1^L)$ in terms of the translation probabilities of individual words.

Using the hidden alignment variables, $P(f_1^M|e_1^L)$ can be expressed as $\sum_{a_1^M} P(f_1^M, a_1^M|e_1^L)$ where a_1^M ranges over all possible alignments between e_1^L and f_1^M. These alignment variables also enforce the constraint that each French word aligns to at most one English word. Notation-wise, if a word in position j in the French string is connected to a word at position i in the English string, $a_j = i$, and if it is not connected to any English word, $a_j = 0$, in which case f_j is connected to the special **null word**, e_0. $P(f_1^M, a_1^M|e_1^L)$ can thus be written as

$$P(f_1^M, a_1^M|e_1^L) = \prod_{j=1}^{M} P(M|e_1^L)P(a_j|a_1^{j-1}, f_1^{j-1}, M, e_1^L) \cdot$$
$$P(f_j|a_1^j, f_1^{j-1}, M, e_1^L) \qquad (2)$$

Machine translation under the framework described above involves training the model using a parallel corpus of source and target sentence pairs; then finding optimal source translations for each target sentence in the test set using eqn 1.

A byproduct of the training step above is that we induce the values of the hidden alignment variables $\mathbf{a}$, for example during the expectation step of the Expectation Maximization procedure that is typically used. The Viterbi alignment is the assignment to the $\mathbf{a}$ variables that maximizes $P(f_1^M, a_1^M|e_1^L)$ and indicates which source word is the most likely to explain the presence at position j of each target word in the target sequence. Finding optimal alignments has received increased attention in the last few years because they have the potential to improve final translation quality and can also be used to build bilingual word or phrase translation tables that can be fed to other systems. For examples of different approaches to the alignment problem see [6, 7].

3. MULTILINGUAL ALIGNMENT TAG MODEL

Brown et al. [5] introduced *IBM Models* 1 through 5 by using different factorizations and simplifications of eqn. 2.

---

original paper on statistical MT [5]. We also follow the convention from the same paper to denote the target string by f_1^M, which stands for French and the source string by e_1^L (English).

[2]We use the word sentence here loosely to mean a unit of translation, typically a full sentence but possibly a fragment of a sentence, or several sentences.

These models have been widely used in the MT community. IBM Model4 and a few of its refinements has become the standard baseline for word alignments. IBM Models 1 and 2 make drastic independence assumptions; they have, however, been found to be useful for initializing higher models as well as features in log-linear alignment models for example. Models 3 and higher introduce the notion of *fertility*, which describes explicitly how many and which target words a source word can connect to. Vogel et al. [8] later introduced the *HMM alignment model* shown in eqn. 3. This model also makes a number of strong independence assumptions such as that the translation probability of the target word f_j depends on nothing else given the source word e_{a_j} it aligns to, but whereas models 1 and 2 assume alignment variables are independent of each other, the HMM model assumes that the way the current target word is aligned depends on how the preceding target word is aligned.

$$P(f_1^M, a_1^M|e_1^L) = P(M|L)\prod_{j=1}^{M} P(a_j|a_{j-1}, M, L) \cdot P(f_j|e_{a_j})$$
$$(3)$$

The HMM model is simpler mathematically and less expensive computationally than IBM models 3 to 5. At the same time the HMM model performance often approaches that of Model4 (among the state of the art alignment models). For both reasons, several lines of previous work that explored enriching alignment models with other sources of information, for instance, part-of-speech information [9], bilingual word clusters [10], or syntax-tree knowledge [11] have built their models on top of the HMM model only. In this paper, in addition to the HMM model we extend all IBM models up to model 4. However, for ease of exposition and to relate our work to previous work, we describe modifications to the HMM model. Extending the changes to the other models is straightforward.

We introduce what we term the **Alignment-tag Model**. Let g_1^M and s_1^L be two language sequences (German and Spanish for example) pre-aligned to the French and English strings f_1^M and e_1^L respectively. Precisely, what this means is that for each word f_i the corresponding tag g_i is the German word (possibly the null word) that aligns best to f_i according to a given alignment model. We can think of German as a noisy feature of our target language, French. We can then, using the probability chain rule in a similar fashion as in eqn. 2, write the probability $P(f_1^M, g_1^M, a_1^M|e_1^L, s_1^L)$ as

$$P(f_1^M, g_1^M, a_1^M|e_1^L, s_1^L) = \prod_{j=1}^{M} P(M|e_1^L, s_1^L) \cdot$$
$$P(a_j|a_1^{j-1}, f_1^{j-1}, g_1^{j-1}, M, e_1^L, s_1^L) \cdot$$
$$P(g_j|a_1^j, f_1^{j-1}, g_1^{j-1}, M, e_1^L, s_1^L) \cdot$$
$$P(f_j|a_1^j, f_1^{j-1}, g_1^j, M, e_1^L, s_1^L) \qquad (4)$$

Clearly if eqn. 2 is unwieldy, then eqn. 4 is even more so.

We introduce the following simplifying assumptions, again basing our exposition around extending the HMM model:

$$P(M|e_1^L, s_1^L) = P(M|L) \qquad (5)$$

$$P(a_j|a_1^{j-1}, f_1^{j-1}, g_1^{j-1}, M, e_1^L, s_1^L) = P(a_j|a_{j-1}, M, L) \qquad (6)$$

$$P(g_j|a_1^j, f_1^{j-1}, g_1^{j-1}, M, e_1^L, s_1^L) = P(g_j|s_{a_j}) \qquad (7)$$

$$P(f_j|a_1^j, f_1^{j-1}, g_1^j, M, e_1^L, s_1^L) = P(f_j|e_{a_j}) \qquad (8)$$

where in (5), we assume the length, M, of the $\mathbf{f}$ sequence (and thus also the length of the $\mathbf{g}$ tag sequence, which is pre-aligned to $\mathbf{f}$) depends on nothing else given the length, L of the $\mathbf{e}$ sequence (which is equal to the length of the $\mathbf{s}$ sequence). In (6), we assume that, given a_{j-1}, M, and L, the alignment variable a_j is independent of everything else. In (7), g_j is independent of everything else given the Spanish word at position a_j; and likewise in (8) for f_j given the English word at the *same* position a_j. Fig. 1 is a graphical representation of the independence assumptions described above.

We end up with eqn. 9, which has a similar form to the one describing the *POS-tagged HMM model* in [9].

$$P(f_1^M, g_1^M, a_1^M|e_1^L, s_1^L) = P(M|L) \prod_{j=1}^{M} P(a_j|a_{j-1}, M, L) \cdot$$
$$P(f_j|e_{a_j}) \cdot (P(g_j|s_{a_j}))^{\alpha} \qquad (9)$$

where $\alpha \in [0, 1]$ is a discount exponent we add to the model to decrease the importance of the $P(g_j|s_{a_j})$ factor. This is similar to what is done is speech recognition to control the contribution of the language model to the overall probability model. α gives us the option to correct for any imbalance that might exist between these two probability scores (as will be seen, we find that this helps in our results).

Note that compared to eqn. 3, the alignment variables a_1^M are now being used to align *both* French to English, and French's features (German) to English's features (Spanish).

As mentioned above, the model in eqn. 9 can easily be generalized to the other IBM models. In this paper, we use models 1, 2 and 4 in addition to the HMM model. We also use the *Refined Model 4* (4r), a further improvement over Model 4 [6]. This alignment procedure involves generating IBM Model 4 alignments in both directions (from the source language to the target and vice versa), then intersecting the two alignment sets and expanding them heuristically.

We will refer to our family of models as **MxTy**, where $\mathbf{x} \in \{$'1', '2', 'hmm', '4'$\}$ is the alignment model used to obtain the alignments between $\mathbf{g}$ and $\mathbf{f}$ and between $\mathbf{s}$ and $\mathbf{e}$. $\mathbf{y} \in \{$'1', '2', 'hmm', '4', '4r'$\}$ is the final alignment model in which the output probabilities $P(f|e)$ have been replaced by $P(f|e) \cdot P(g|s)$ as is shown in the case $\mathbf{x} =$'hmm' in eqn. 9.
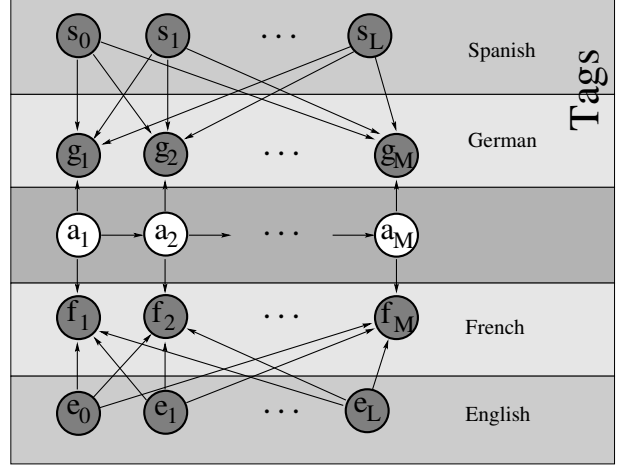
The full alignment-tag algorithm is shown below.

Fig. 1. **Alignment-tag Model**. *Each random variable g_j and s_i in the sequences $\mathbf{g}$ and $\mathbf{s}$ is assigned the word in the German and Spanish string resp. that best aligns with the value of the random variable f_j and e_i resp. We can therefore think of $\mathbf{g}$ and $\mathbf{s}$ as tag sequences for $\mathbf{f}$ and $\mathbf{e}$ resp. Each alignment variable a_j $(1 \leq j \leq M)$ decides both which word e_i $(0 \leq i \leq L)$ connects to f_j $(1 \leq j \leq M)$, and which word s_i $(0 \leq i \leq L)$ connects to g_j $(1 \leq j \leq M)$. The Naive Bayes-like assumption made by the model is clear from the fact $\mathbf{g}$ and $\mathbf{f}$ are marginally dependent of each other but independent conditioned on the alignment variables $\mathbf{a}$. Likewise for $\mathbf{s}$ and $\mathbf{e}$. For simplicity the figure does not show the dependence of the alignment variables on the source and target lengths L and M. The languages used in this work (German, Spanish, etc.) are also used to name the variables $\mathbf{g}$, $\mathbf{s}$, $\mathbf{f}$, and $\mathbf{e}$ but it should be noted that any languages could be used. In particular, we could use a common language (e.g., Finnish) both for $\mathbf{g}$ and $\mathbf{s}$.*

Model MxTy, $\quad \mathbf{x} \in \{\textbf{1,2,hmm,4}\} \quad \mathbf{y} \in \{\textbf{1,2,hmm,4,4r}\}$

Input: English, French, German, and Spanish sentences that are translations of each other.

Output: English to French word alignments.

1. Using Model $\mathbf{x}$, align Spanish to English, to generate Spanish tags for the English sentence.

2. Using Model $\mathbf{x}$, align German to French, to generate German tags for the French sentence.

3. Using Model $\mathbf{y}$ with augmented output probabilities, $P(f|e) \cdot P(g|s)$, generate English to French word alignments.

4. EXPERIMENTAL SETUP

We describe our experimental setup in particular the data processing steps we took to be able to use all the languages at our disposition. Any references to specific languages in this section and subsequent ones refer to the actual languages used as opposed to the placeholder languages used above for ease of exposition.

Our corpus is a subset of Europarl, which was compiled for statistical machine translation research by Koehn [1]. Specifically we use the corpus made available for the 2005 ACL workshop on machine translation [12]. Four parallel corpora were provided: French-English (EN-FR), Spanish-English (ES-EN), German-English (DE-EN), and Finnish-English (FI-EN). These corpora originate from the same European Parliament session transcript; however, because of variations in preprocessing, paragraph segmentation, and sentence alignment for each of the four pairs, the resulting set of English sentences was different for each of the four corpora. To resolve these differences, we aligned the English sentences from each corpus to each other and threw out all mismatches and their corresponding translations. We thus generated a single common set of English sentences and their corresponding translations in the four other languages. Using this procedure, the final number of sentences for each language was reduced from about 700k to 545379 sentences.

Another difficulty in performing word alignment experiments using the Europarl corpus is that as of now no hand-aligned data exist for any pair of languages. We therefore hand-aligned 107 sentences from French to English, totaling about 5000 words[3]. Before we aligned enough data, we also we used as gold standard a 2000-sentence subset of automatic alignments provided by [12]. These alignments were generated using the refined IBM Model4 using all of the French-English corpus. The trends we observed on that test set were similar to those on the current test set.

For training our **Alignment-tag Model** we use training subsets of the full corpus ranging in size from 1k to 160k sentence pairs. For the baseline, we used train set sizes up to 700k, which corresponds to the largest bilingual corpus we had at our disposition *before* we reduced its size to make it suitable for our multilingual experiments as explained above. The maximum sentence length in the corpus is 40 words. The number of running words in the 700k corpus is about 15M words for French and 14M for English. In the 160k train set, the number of words is 3.6M for French, 2.3M for Finnish, and around 3.1M words for the remaining languages. Vocabulary size is 72k words for French, 55k for English, 88k for Spanish, 155k for German, and 290k for Finnish.

We used five EM training iterations each of Models 1,

2, HMM, 3, and 4. Each model being used to initialize the next one. We tried several different variations in the training schedule. The changes in performance were not significant for IBM Model4, the model we mostly care about.

For a given sentence pair (S_i, T_i), an alignment between S_i and T_i can be represented as the set of pairings $A_i = \{(j, a_j) | (j, a_j) \neq (0, 0)\}$, where j is the position in the source sentence and a_j the position in the target sentence. If one of j or a_j is zero resp., we call (j, a_j) a **null pairing** to refer to the fact a source or target word resp. is aligned to the null word.

It is common for human annotators to label pairings as either **sure** or **possible** to allow for loose translations or matching of idiomatic expressions. We follow this convention and define S_i as the set of sure pairings and P_i that of possible ones for the ith sentence pair.

Alignment Error Rate (AER) [13] is a well accepted measure of alignment quality and favors high precision on the *sure* set and high recall on the *possible* set. AER is defined as

$$AER = 1 - \frac{\sum_i^D (|A_i \cap S_i| + |A_i \cap P_i|)}{\sum_i^D (|S_i| + |A_i|)}$$

where A_i is the set of hypothesized alignments for sentence pair i, and S_i and P_i reference alignments defined as above. D is the number of sentence pairs in the test corpus.

Finally, we follow the convention of evaluating our alignments on the basis of both AER as defined above and the **nonnull alignment** AER. In the latter case, any null pairings in the reference or the hypothesized alignments are removed before the AER is computed. Typically the nonnull alignment AER is lower than the regular AER. Large differences between the two error rates can be informative of model assumptions relating to null alignments. Unless specified otherwise, we present regular AER results; our nonnull alignment results follow the same trend as regular ones.

5. RESULTS

The results in this section assume a scenario under which we translate from English to French. Using the notation introduced earlier, this means the **e** sequence is in French and the **f** is in English.

Figure 2 shows baseline alignment error rates using different models and with varying amounts of training data. The trends–the increasing but diminishing gains with larger training sizes; and the consistent advantage of the higher models over the training size range–are in line with what is frequently reported in the word alignment literature (see for example [6]).

Table 1 shows our best results obtained using the tag-alignment model in eqn. 9, where both the **g** and the **s** languages are Spanish, and where the tags were generated using IBM Model4. In order to balance the effect of the annotation model with the true source/target language model,
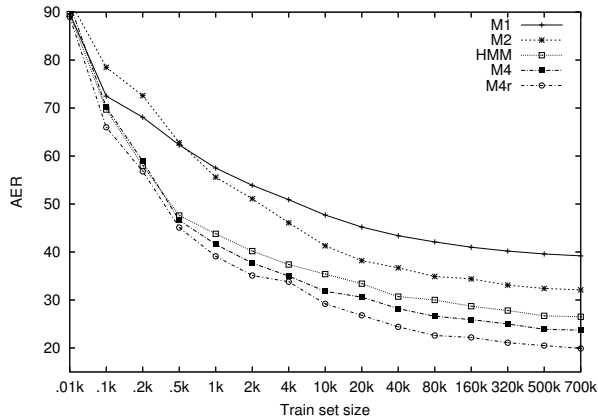
Fig. 2. *French-English %AER baselines for different training set sizes.*

we experimented with various weighting schemes on the annotation model. Specifically, we exponentiate the factor $P(g_j|s_{a_j})$ to $\alpha = 0.1$ in order to decrease the contribution of tag alignment probability relative to the word alignment probability.

Using 160k train sentences and model M4T4r (table 1), the alignment error rate improves from 22.2% to 20.7% (and, not shown in the table, from 20.0% to 18.5% measured using the nonnull alignment AER) compared to the Refined IBM Model4. This constitutes a 7% relative improvement. Because it is reasonable to expect the amount of bilingual data to exceed that of multilingual data at any point of time, we also compare the performance of our model trained using 160k sentences to baselines trained with more data. Even when twice as much data is used to train the baseline, the alignment-tag model keeps a slight edge. With more than four times as much data, the baseline finally improves on the alignment-tag model by 0.8% absolute. This suggests that scalability should be an important consideration in designing any models for multilingual alignments.

We should also note that we found 4k to be the threshold above which we start seeing improvements from using the alignment-tag model over the baseline. With fewer than 4k sentences, the AER is high enough that the noise in the tags hurts us. This is of course very much expected and confirms the more accurate the tags the better the overall performance. This also suggests that an iterative procedure that cycles between improving the alignments between different language pairs, which in turn are used as for each other might work well.

As one would expect, Model4 tags help improve M4T1 and M4T2 (and M4Thmm to a lesser degree) models to a greater extent than they help M4T4 or M4T4r. In the latter case the tag quality is no better than the final alignment model while in the former cases, the tags are of better qual-

ity and help the models choose correct alignments. This effect is accentuated at higher values of α because the tag translation probabilities are trusted more. Table 2 shows the effect of varying α from 0.1 to 1.5 for different tag models.

TrainSet	Models				
	M4T1	**M4T2**	**M4Thmm**	**M4T4**	**M4T4r**
160k	37.6	31.2	27.0	24.9	20.7
	M1	**M2**	**HMM**	**M4**	**M4r**
B160k	41.0	34.4	28.7	25.9	22.2
B320k	40.2	33.1	27.8	25.0	21.1
B700k	39.2	32.1	26.5	23.7	19.9

Table 1. *Alignment-tag Model trained on 160k sentences. Spanish is used for both g and s. $\alpha = 0.1$. Rows B160k, B320k, and B700k show the baseline performance with 160k, 320k, and 700k training sets resp.*

α	**M4T1**	**M4T2**	**M4Thmm**	**M4T4**
0.1	37.6	31.2	27.0	**24.9**
0.3	34.7	29.5	**26.1**	**24.5**
0.5	**33.7**	**28.9**	**26.2**	25.1
0.8	**33.9**	30.0	28.2	26.4
1.5	36.6	33.2	32.0	31.3

Table 2. *%AER at different α values. Spanish is used for both g and s. The training set size is 160k.*

So far we have used Spanish as the tag language which we align to French and English. It is interesting to look at what effect using different languages as tags would have on alignment performance. Table 3 shows that English, Spanish, and French seem to help the most when used as tags, compared to German and Finnish. This is consistent with the reported performance of various translation systems on different language pairs. For example, in a recent machine translation shared task, Spanish-English and French-English translation were the highest scoring [12]. This suggests that the reason for the improvement seen is owing to the alignment accuracy for these languages rather than any inherent property of the languages themselves. In other words, we would expect similar improvements if we were able to align, say German to English, as accurately as Spanish to English.

We also observe that single language tags (with the notable exception of Finnish) help decrease AER the most. For example, (**g** = English, **s** = English) and (**g** = Spanish, **s** = Spanish) tags rank highest. We were surprised that (**s** = English, **g** = French) tags (i.e., English is used to tag French (**e**), and French is used to tag English (**f**)) were among the worst of the combinations shown. Note, also, that the M4T4 AER for the last three rows in table 3 is worse than the corresponding baseline.

	M4T1	M4T2	M4Thmm	M4T4
EN-EN	**37.4**	31.2	**27.1**	**24.8**
ES-ES	**37.6**	31.2	**27.0**	**24.9**
FR-FR	**37.9**	**30.1**	**26.9**	25.8
ES-EN	40.2	33.2	28.4	25.7
ES-DE	39.7	33.2	28.6	26.2
FI-FI	39.3	33.5	28.8	26.3
EN-FR	39.1	32.9	27.8	26.3

Table 3. *Tag language effect comparison using 160k sentences and $\alpha = 0.1$. Five languages are used, English (EN), Spanish (ES), French (FR), Finnish (FI), and German (DE). ES-DE, for example, means Spanish is used for* **g** *and German for* **s***.*

6. CONCLUSION

We have presented a statistical word alignment model that exploits information from parallel translations in more than two languages. We show a 7% decrease in alignment error rate relative to a state of the art alignment algorithm. Our model is a simple and efficient extension of the IBM and HMM models. The model makes a strong assumption of independence between words and their alignment-tags. While the effect of this assumption on alignment quality is not clear, a possible improvement of the algorithm might come from modeling the correlation between words and their tags. The challenge, however, will remain to do so efficiently to be able to scale to today's increasingly large corpora.

Also, given our two-step word alignment approach, a natural extension is an iterative procedure whereby after producing improved word alignments using the alignment-tag model we go back and generate new tags and new word alignments.

We believe the alignment-tag model can be a good starting point for more sophisticated multilingual alignment models, and ultimately full translation models. Most of the current statistical translation engines are based on generating word alignments as an intermediary step for learning higher-level translation units (e.g. phrases, trees, etc). An important future direction of research is to investigate whether our gains in multilingual alignment quality carry over and improve learning of phrase translation probabilities, for example. We should also note that multilingual information might help at different stages in a translation system. In [2], it was shown to help in the decoding stage. We have shown it to help improve word alignments and we think it can also help in the selection of better candidate phrases compared to a bilingual only system.

7. REFERENCES

[1] P. Koehn, "Europarl: A multilingual corpus for evaluation of machine translation," http://www.isi.edu/koehn/publications/europarl, 2002.

[2] F.J. Och and H. Ney, "Statistical multi-source translation," in *Proceedings of the MT Summit VIII*, September 2001, pp. 253–258.

[3] M. Simard, "Text-translation alignment: Three languages are better than two," in *Proceedings of EMNLP/VLC-99, College Park, MD*, 1999.

[4] 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, 1993.

[5] P. F. Brown, J. Cocke, S. A. Della Piettra, V. J. Della Piettra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, June 1990.

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

[7] J. Martin, R. Mihalcea, and T. Pedersen, "Word alignment for languages with scarce resources," in *Proc. ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, Michigan, June 2005, pp. 65–74, ACL.

[8] S. Vogel, H. Ney, and C. Tillmann, "HMM-based word alignment in statistical translation," in *Proceedings of the 16th conference on Computational linguistics*, Morristown, NJ, USA, 1996, pp. 836–841, ACL.

[9] K. Toutanova, H.T. Ilhan, and C.D. Manning, "Extensions to HMM-based statistical word align-ment models," in *Proc. Conf. on Empirical Methods for Natural Language Processing*, 2002, pp. 87–94.

[10] B. Zhao, E. P. Xing, and A. Waibel, "Bilingual word spectral clustering for statistical machine translation," in *Proc. ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, Michigan, June 2005, pp. 25–32, ACL.

[11] A. Lopez and P. Resnik, "Improved HMM alignment models for languages with scarce resources," in *Proc. ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, Michigan, June 2005, pp. 83–86, ACL.

[12] P. Koehn and C. Monz, "Shared task: Statistical machine translation between European languages," in *Proc. ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, Michigan, June 2005, pp. 119–124, ACL.

[13] F. J. Och and H. Ney, "Improved statistical alignment models," Hongkong, China, Oct 2000, pp. 440–447.