Tuning Syntactically Enhanced Word Alignment for Statistical Machine Translation

Yanjun Ma[†] **Patrik Lambert**[‡] **Andy Way**[†‡]
[†]National Centre for Language Technology
[‡]Centre for Next Generation for Localisation
School of Computing, Dublin City University
Dublin 9, Ireland
{yma,plambert,away@computing.dcu.ie}

Abstract

We introduce a syntactically enhanced word alignment model that is more flexible than state-of-the-art generative word alignment models and can be tuned according to different end tasks. First of all, this model takes the advantages of both unsupervised and supervised word alignment approaches by obtaining anchor alignments from unsupervised generative models and seeding the anchor alignments into a supervised discriminative model. Second, this model offers the flexibility of tuning the alignment according to different optimisation criteria. Our experiments show that using our word alignment in a Phrase-Based Statistical Machine Translation system yields a 5.38% relative increase on IWSLT 2007 task in terms of BLEU score.

1 Introduction

Word alignment, which can be defined as a problem of determing word-level correspondences given a parallel corpus of aligned sentences, is a fundamental component in Phrase-Based Statistical Machine Translation (PB-SMT). The dominant approach to word alignment are generative models, including IBM models (Brown et al., 1993) and HMM models (Vogel et al., 1996; Deng and Byrne, 2006). While generative models trained in an unsupervised manner can produce high-quality alignments given a reasonable amount of training data, it is difficult to incorporate richer features into such models. On the other hand, discriminative models are more flexible to incorporate arbitrary features. However, these models need a certain amount of annotated word alignment data, which is often subject to criticism since the annotation of word alignment is a highly subjective task. Moreover, parameters optimised on manually annotated data are not necessarily optimal for MT tasks. Recent research attempts to combine the merits of both generative and discriminative models (Fraser and Marcu, 2007), or to tune a discriminative model according to MT metrics (Lambert et al., 2007).

In this paper, we introduce a simple yet flexible framework for word alignment. To take the advantage of the strength of generative models, we use these models to obtain a set of anchor alignments. We then incorporate syntactic features induced by the anchor alignments into a discriminative word alignment model. The syntactic features we used are syntactic dependencies. This decision is motivated by the fact that if words tend to be dependent on each other, so does the alignment (Ma et al., 2008). If we can first obtain a set of reliable anchor links, we could take advantage of the syntactic dependencies relating unaligned words to aligned anchor words to expand the alignment. Figure 1 gives an illustrative example. Note that the link (c_2, e_4) can be easily identified, but the link involving the fourth Chinese word (a function word denoting 'time') (c_4, e_4) is hard. In such cases, we can make use of the dependency relationship ('tclause') between c_2 and c_4 to help the alignment process.

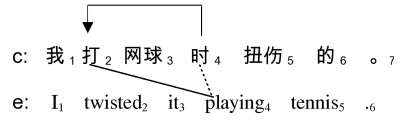

Figure 1: Dependencies for word alignment

Proceedings of the 13th Annual Conference of the EAMT, pages 250–257,
Barcelona, May 2009

Our experiments show that using our word alignment approach in a PB-SMT system can significantly improve the system over a strong baseline. The experiments also show that syntax is beneficial in word alignment. Given that the intrinsic quality of word alignment measured using F-score does not correlate well with PB-SMT performance measured using BLEU, we conducted experiments that can directly optimise the word alignments according to BLEU score. Experiments show that we can achieve higher performance using such an optimisation procedure and our word alignment approach is more flexible in a PB-SMT framework.

2 Syntactically Enhanced Word Alignment Model

2.1 General Model

Given a source sentence $C = c_1^J$ that consists of J Chinese words $\{c_1, \cdots, c_J\}$ and target sentence $E = e_1^I$ which consists of I English words $\{e_1, \cdots, e_I\}$, we seek to find the optimal alignment $\hat{A}$ such that:

$$\hat{A} = \underset{A}{\operatorname{argmax}} P(A|c_1^J, e_1^I)$$

We use a model (1) that directly models the linkage between source and target words similarly to (Ittycheriah and Roukos, 2005). The Chinese-to-English word alignment $A_{C \to E} = \{i|a_j = i\}$ is modelled as shown in (1). We decompose this model into an emission model and a transition model (4). The emission model can be further decomposed into an anchor alignment model (2) and a syntactically enhanced model (3) by distinguishing the anchor alignment from the non-anchor alignment.

$$\begin{aligned} p(A|c_1^J, e_1^I) &= \prod_{j=0}^{J} p(a_j|c_1^J, e_1^I, a_1^{j-1}) && (1) \\ &= \frac{1}{Z} \cdot p_\epsilon(A_\Delta|c_1^J, e_1^I) \cdot && (2) \\ &\quad \prod_{j \in \bar{\Delta}} p(a_j|c_1^J, e_1^I, a_1^{j-1}, A_\Delta) && (3) \\ &\quad \prod_{j=1}^{J} p(a_j|a_{j-1}, A_\Delta) && (4) \end{aligned}$$

2.2 Emission Model

2.2.1 Anchor Word Alignment

The anchor alignment model $p_\epsilon(A_\Delta)$ aims to find a set of high-precision links. Various approaches can be used for this purpose.

We can use the asymmetric IBM models for bidirectional word alignment and derive the intersection. Using this approach, we can obtain a set of anchor alignments $A_\Delta = \{i|i \in \Delta\}$. Subsequently, the anchor model is estimated as follows:

$$p(a_j) = \begin{cases} \alpha & \text{if } a_j = i \text{ and } i \in \Delta, \\ \frac{1-\alpha}{I} & \text{otherwise.} \end{cases}$$

The parameter α can be optimised on the development set. In our experiments we set $\alpha = 0.9$.

2.2.2 Syntactically Enhanced Word Alignment

The syntactically enhanced model is used to model the alignment of the words left unaligned after anchoring. We directly model the linkage between source and target words using a discriminative word alignment framework where various features can be incorporated. Given a source word c_j and the target sentence e_1^I, we search for the alignment a_j such that:

$$\begin{aligned} \hat{a_j} &= \underset{a_j}{\operatorname{argmax}} \{p_{\lambda_1^M}(a_j|c_1^J, e_1^I, a_1^{j-1}, A_\Delta)\} && (5) \\ &= \underset{a_j}{\operatorname{argmax}} \{\sum_{m=1}^{M} \lambda_m h_m(c_1^J, e_1^I, a_1^j, A_\Delta, T_c, T_e)\} \end{aligned}$$

In this decision rule, we assume that a set of highly reliable anchor alignments A_Δ has been obtained, and T_c (resp. T_e) is used to denote the dependency structure for source (resp. target) language. In such a framework, various machine learning techniques can be used for parameter estimation. The feature functions we used are described in section 3.

2.3 Transition Model

Given the anchor alignment, the first-order transition probability model (4) can be defined as follows:

$$p(a_j|a_{j-1}, A_\Delta) = \begin{cases} 1.0 & \text{if} \in \Delta, \\ \hat{p}(a_j|a_{j-1}) & \text{otherwise.} \end{cases}$$

Such a definition implies that an anchor alignment is always believed to be a correct alignment, maximum likelihood estimates obtained on a gold-standard word alignment corpus are used when the current word f_j is not involved in an anchor alignment. The estimation of $\hat{p}(a_j|a_{j-1})$ is calculated following the homogeneous HMM model (Vogel et al., 1996). Under this model, we assume that the

probability depends only on the jump width $(i-i')$, in order to make the alignment parameters independent of absolute word positions. Using a set of non-negative parameters $\{c(i-i')\}$, the transition probability can be written in the form:

$$p(a_j|a_{j-1}, A_\Delta) = \frac{c(i-i')}{\sum_{i''=1}^{I} c(i''-i')}$$

We use the refined model which extends the HMM network with I empty words e_{I+1}^{2I} and adds parameter p_0 to account for the transition probability to empty words (Och and Ney, 2003).

If a zero-order dependence is assumed in a transition model, the emission models is the only information to guide the word alignment.

2.4 Model Interpolation

We interpolate the general alignment model (1) as follows:

$$\begin{aligned} p(A|c_1^J, e_1^I) &= \frac{1}{Z} \cdot p_\epsilon(A_\Delta|c_1^J, e_1^I)^{1-\lambda} \cdot \\ &\prod_{j \in \bar{\Delta}} p(a_j|c_1^J, e_1^I, a_1^{j-1}, A_\Delta)^{1-\lambda} \cdot \\ &\prod_{j=1}^{J} p(a_j|a_{j-1}, A_\Delta)^{\lambda} \end{aligned}$$

We can use factor λ to weight the emission model and transition model probabilities so that the system can be optimised according to different objectives.

3 Feature Functions for Syntactically Enhanced Word Alignment

The various features used in our syntactically enhanced model can be classified into two groups: statistics-based features and syntactic features which are similar to those in (Ma et al., 2008)

3.1 Statistics-based Features

The statistics-based features we used include IBM model 1 score, Log-likelihood ratio (Dunning, 1993) and POS translation probability. We choose these features because they are empirically proven to be effective in word alignment tasks (Melamed, 2000; Liu et al., 2005; Moore, 2005).

3.2 Syntactic Features

The dependency relation R_e (resp. R_c) between two English (resp. Chinese) words e_i and $e_{i'}$ (resp. c_j and $c_{j'}$) in the dependency tree of the English sentence e_1^I (resp. Chinese sentence c_1^J) can be represented as a triple $<e_i, R_e, e_{i'}>$ (resp. $<c_j, R_c, e_{j'}>$). Given c_1^J, e_1^I and their syntactic dependency trees $T_{c_1^J}$, $T_{e_1^I}$, if e_i is aligned to c_j and $e_{i'}$ aligned to $c_{j'}$, according to the dependency correspondence assumption (Hwa et al., 2002), there exists a triple $<c_j, R_c, c_{j'}>$.

While we are not aiming to justify the feasibility of the dependency correspondence assumption by proving to what extent $R_e = R_c$ under the condition described above, we do believe that c_j and $c_{j'}$ are likely to be dependent on each other. Given the anchor alignment A_Δ, a candidate link (j, i) and the dependency trees, we can design four classes of feature functions.

3.2.1 Agreement features

The agreement features can be further classified into dependency agreement features and dependency label agreement features. Given a candidate link (j, i) and the anchor alignment A_Δ, the dependency agreement (DA) feature function is defined as follows:

$$h_{DA-1} = \begin{cases} 1 & \text{if } \exists <c_j, R_c, c_{j'}>, <e_i, R_e, e_{i'}> \\ & \text{and } (j', i') \in A_\Delta, \\ 0 & \text{otherwise.} \end{cases}$$

By changing the dependency direction between the words c_j and $c_{j'}$, we can derive another dependency agreement feature:

$$h_{DA-2} = \begin{cases} 1 & \text{if } \exists <c_{j'}, R_c, c_j>, <e_{i'}, R_e, e_i> \\ & \text{and } (j', i') \in A_\Delta, \\ 0 & \text{otherwise.} \end{cases}$$

We can define the dependency label agreement feature[1] as follows:

$$h_{DLA-1} = \begin{cases} 1 & \text{if } \exists <c_j, R_c, c_{j'}>, <e_i, R_e, e_{i'}> \\ & \text{and } (j', i') \in A_\Delta, R_c = R_e, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly we can obtain h_{DLA-2} by changing the dependency direction.

3.2.2 Source word dependency features

Given a candidate link (j, i) and anchor alignment A_Δ, source language dependency features are used to capture the dependency label between a

[1]Note that we used the same dependency parser for source and target language parsing.

source word c_j and a source anchor word $c_k \in \Delta$. For example, a feature function relating to dependency type 'PRD' can be defined as:

$$h_{src-1-PRD} = \begin{cases} 1 & \text{if } \exists < c_j, R_c, c_{j'} > \\ & \text{and } R_c = \text{'PRD'}, \\ 0 & \text{otherwise.} \end{cases}$$

By changing the direction we can obtain $h_{src-2-PRD}$.

3.2.3 Target word dependency features

Target word dependency features can be defined in a similar way as source word dependency features.

3.2.4 Target anchor feature

The target anchor feature defines whether the target word e_i is an anchor word.

$$h_{src-1-PRD} = \begin{cases} 1 & \text{if } i \in a_\Delta, \\ 0 & \text{otherwise.} \end{cases}$$

4 Experimental Setting

4.1 Data

The experiments were carried out using the Chinese–English datasets provided within the IWSLT 2007 evaluation campaign (Fordyce, 2007). We tagged all the sentences in the training and devset3 using a maximum entropy-based POS tagger, namely MXPOST (Ratnaparkhi, 1996), trained on the Penn English and Chinese Treebanks. Both Chinese and English sentences are parsed using the Malt dependency parser (Nivre et al., 2007), which achieved 84% and 88% labelled attachment scores for Chinese and English (each has 11 dependency labels) respectively.

4.1.1 Word Alignment

We manually annotated word alignments on devset3. Following recent research in measuring word alignment quality for SMT purposes, we set all the word alignment links as sure (S) links (cf. (Fraser and Marcu, 2007)). IWSLT devset3 consists of 502 sentence pairs after cleaning. We used the first 300 sentence pairs for training, the following 50 sentence pairs as validation set and the last 152 sentence pairs for testing. The various statistics for the gold-standard corpus is listed in Table 1.

		Chinese	English
Train	Sentences	300	
	Running words	2,231	2,704
	Vocabulary size	636	709
	Sure links	2773	
Dev.	Sentences	50	
	Running words	445	451
	Vocabulary size	205	212
	Sure links	555	
Eval.	Sentences	152	
	Running words	1,107	1,149
	Vocabulary size	394	413
	Sure links	1400	

Table 1: Chinese–English word alignment gold-standard corpus statistics

4.1.2 Machine Translation

Training was performed using the default training set (39,952 sentence pairs), to which we added the set devset1 (506 sentence pairs) and devset2 (500 sentence pairs).[2] We used devset4 (489 sentence pairs, 7 references) to tune various parameters in the MT system and IWSLT 2007 test set (489 sentence pairs, 6 references) for testing. Detailed corpus statistics are shown in Table 2.

		Chinese	**English**
Train	Sentences	40,958	
	Running words	357,968	385,065
	Vocabulary size	11,362	9,718
Dev.	Sentences	489 (7 ref.)	
	Running words	5,717	46,904
	Vocabulary size	1,143	1,786
Eval.	Sentences	489 (7 ref.)/489 (6 ref.)	
	Running words	3,166	23,181
	Vocabulary size	862	1,339

Table 2: Corpus statistics IWSLT 2007 data set

4.2 Alignment Training and Decoding

In our experiments, we treated anchor alignment and syntactically enhanced alignment as separate processes in a pipeline. The anchor alignments are kept fixed so that the parameters in the syntactically enhanced model can be optimised.[3] We used the support vector machine (SVM) toolkit, SVM_light[4] to optimise the parameters in (5). Our model is constrained in such a way that each

[2]More specifically, we chose the first English reference from the 16 references and the Chinese sentence to construct new sentence pairs.

[3]Note that our anchor alignment does not achieve 100% precision. Since we performed precision-oriented alignment for the anchor alignment model, the errors in anchor alignment will not bring much noise into the syntactically enhanced model.

[4]http://svmlight.joachims.org/

source word can only be aligned to one target word.

In SVM training, we transform each possible link involving the words left unaligned after anchoring into an event. Positive examples (aligned pairs) are assigned the target value 1 and negative examples (unaligned pairs) assigned -1. Using this training data, we can build a regression model to estimate the reliability of alignment given a pair of words. The value of functional margin obtained by applying the regression model serves as the emission probability in our word alignment model.

For the first-order transition model, we estimate the transition probability on a gold-standard word alignment corpus in training. In decoding, the best alignment path is searched out using a Viterbi-style decoding algorithm. The interpolation factor λ can be optimised on development set. When a zero-order transition model (a uniform transition distribution) is used, we constrain the emission probability by a threshold t, which is set as the minimal reliability score for each link. Again, t can be optimised according to the development set.

The decoding is performed separately in two directions (Chinese-to-English and English-to-Chinese), and we then obtain the refined alignments as the final word alignment.

4.3 Baselines

4.3.1 Word Alignment

We used the GIZA++ implementation of IBM word alignment model 4 (Brown et al., 1993; Och and Ney, 2003) for word alignment, and the heuristics described in (Koehn et al., 2003) to derive the intersection and refined alignment.

4.3.2 Machine Translation

We use a standard log-linear PB-SMT model as a baseline: GIZA++ implementation of IBM word alignment model 4,[5] the refinement and phrase-extraction heuristics described in (Koehn et al., 2003), minimum-error-rate training (Och, 2003), a trigram language model with Kneser-Ney smoothing trained with SRILM (Stolcke, 2002) on the English side of the training data, and Moses (Koehn et al., 2007) to decode.

[5]More specifically, we performed 5 iterations of Model 1, 5 iterations of HMM, 3 iterations of Model 3, and 3 iterations of Model 4.

4.4 Evaluation

We evaluate the intrinsic quality of the predicted alignment A with Precision, Recall and the balanced F-score with $\alpha = 0.5$ (cf. (Fraser and Marcu, 2007)).

$$\text{Recall} = \frac{|A \cap S|}{|S|} \quad \text{Precision} = \frac{|A \cap S|}{|A|}$$

$$\text{F-score}(A, S, \alpha) = \frac{1}{\frac{\alpha}{Precision(A,S)} + \frac{1-\alpha}{Recall(A,S)}}$$

Research has shown that an increase in AER does not necessarily imply an improvement in translation quality (Liang et al., 2006) and vice-versa (Vilar et al., 2006). Hereafter, we use a Chinese–English MT task to extrinsically evaluate the quality of our word alignment. The translation output is measured using BLEU (Papineni et al., 2002).

5 Experiments

5.1 Word Alignment Results

We performed word alignment bidirectionally using our approach to obtain the refined alignments (Koehn et al., 2003) and compared our results with two strong baselines based on generative word alignment models. The results are shown in Table 3. We can see that both the syntactically enhanced model based on HMM intersection anchors and on IBM model 4 anchors achieved higher F-scores than the pure generative word alignment models. It is also can be seen that zero-order syntactic models are better in precision and first-order models are superior in recall. The best result achieved 2.99% relative increase in F-score compared to the baseline when we use IBM model 4 intersection to obtain the set of anchor alignments.

model	Precision	Recall	F-score
Model 1	65.98	70.64	68.23
+Syntax-zero-order	**80.71**	69.93	**74.93**
+Syntax-first-order	72.84	**73.36**	73.10
HMM refined	73.80	73.86	73.83
+Syntax-zero-order	**83.65**	70.14	76.30
+Syntax-first-order	77.17	**76.07**	**76.62**
Model 4 refined	75.87	78.14	76.99
+Syntax-zero-order	**84.59**	74.50	**79.29**
+Syntax-first-order	80.21	**77.57**	78.87

Table 3: The performance of our syntactically enhanced word alignment approach

5.2 Machine Translation Results

Table 4 shows the influence of our word alignment approach on MT quality.[6] From Table 4, we can see that our zero-order syntactically enhanced model based on Model 4 anchors achieved 1.82 absolute BLEU score (5.38% relative) improvement compared to its baseline counterpart on the test set, which is statistically significant ($p<0.002$) using approximate randomisation (Noreen, 1989) for significance testing. However, the first-order model suffers from overfitting problems, with a significant improvement on the development set and no improvement on the test set.

	dev	test
Baseline-Model4	24.13	33.85
+Syntax-zero-order	25.41	**35.67**
+Syntax-first-order	25.47	33.70

Table 4: Syntactically enhanced word alignment for PB-SMT optimised according to BLEU

5.2.1 Different Optimisation Criteria

The parameter t (threshold) for zero-order models can be optimised with either F-score (OFscore) obtained on a gold-standard word alignment corpus, or BLEU score (OBLEU) on a development set of an MT system as the objective. Similarly for first-order models, parameters λ and p_0 can be optimised according to these two criteria. Given that we have a very limited number of parameters to optimise (just two, i.e. $t_{c\to e}$ for Chinese–English and $t_{e\to c}$ for English–Chinese in the zero-order model, and three parameters, i.e. $\lambda_{c\to e}$, $\lambda_{e\to c}$ and p_0 in the first-order model), we used a simple greedy search algorithm by search a predefined set of possible parameter settings. For example, we tried different value combinations from the set $\{-1.7, -1.6, \cdots, 0.0\}$ for $t_{c\to e}$ and for $t_{c\to e}$. Table 5 shows the results according to different optimisation criteria using Model 4 intersected alignments as anchors.

For the zero-order model, the best parameter set is $t_{c\to e} = -1.0$ and $t_{e\to c} = -0.6$ according to F-score; however, according to BLEU, the best parameters are $t_{c\to e} = -0.8$ and $t_{e\to c} = -0.9$. From Table 5, we can see that the BLEU score obtained when word alignment is optimised according to F-score is slightly inferior (not statistically significant) to that when optimised according to BLEU. The search graph of optimisation according to BLEU is shown in Figure 2. The different optimisation criteria do not have much impact on the F-score. For the first-order model, the best parameter setting is $\lambda_{c\to e} = 0.2$, $\lambda_{e\to c} = 0.2$ and $p_0 = 0.6$ according to F-score. However, according to BLEU, it is $\lambda_{c\to e} = 0.9$, $\lambda_{e\to c} = 0.3$ and $p_0 = 0.8$. From Table 5, we can observe that parameters optimised according to BLEU suffer from overfitting. The word alignment optimised according to F-score not only yields a higher F-score, but also achieves better performance on the test set when used in a PB-SMT system.

[6]Note that the only difference between our MT system and the baseline PB-SMT system is the word alignment component.

		BLEU		F-score	
		dev	test	dev	test
Zero-order	OFscore	24.74	35.21	*77.49*	79.23
	OBLEU	25.41	**35.67**	76.98	*79.25*
First-order	OFscore	23.75	34.32	76.41	78.87
	OBLEU	**25.47**	33.70	70.75	72.33

Table 5: Optimising syntactically enhanced word alignment for PB-SMT

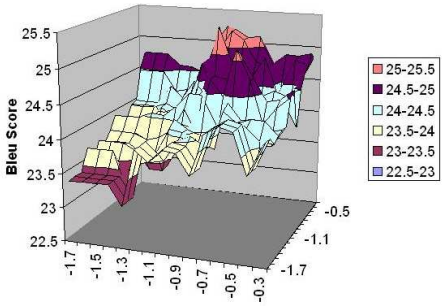

Figure 2: Search graph obtained when optimising BLEU

5.2.2 Phrase Extraction

To further investigate the impact of our word alignment on SMT, we compared the extracted phrase table using our word alignment against the baseline phrase table. Figure 3 shows the size of the phrase tables when the system use different word alignment. We observed that using the zero-order syntactically enhanced word alignment tends to extract fewer phrase pairs (more word alignment links) when optimised according to BLEU. As an exception, the first-order word alignment which suffered from overfitting extracted far more phrase pairs (fewer word alignment links) when optimised according to BLEU. All syntactically enhanced word alignments lead to larger phrase tables.

Figure 3: A comparison of the number of phrase pairs

5.2.3 Scaling Up

To test the scalability of our approach, we added in a further 130K sentence pairs from HIT corpus provided for IWSLT 2008 evaluation campaign. We re-use the parameters obtained from the IWSLT 2007 corpus in these experiments. Table 6 shows the results. For the zero-order syntactically enhanced model optimised according to BLEU, we observed an increase of 1.69 absolute BLEU scores over the baseline on the development set; on the test set, however, no improvement was achieved. For the first-order model, given the parameters we obtained on IWSLT 2007 data set by optimising BLEU suffered from overfitting, the consequence can also be seen on the experiments using the larger data set. From these results, we can see the limitation of the optimisation process and a more informative objective function is needed to achieve better performance.

		dev	test
Baseline-Model4		27.05	**35.65**
Syntax-zero-order	OFscore	26.93	35.35
	OBLEU	**28.74**	35.47
Syntax-first-order	OFscore	27.05	35.16
	OBLEU	28.17	34.95

Table 6: Scaling up syntactically enhanced word alignment for PB-SMT

6 Comparison with Previous Work

(Fraser and Marcu, 2007) proposed a semi-supervised model that can take advantage of both generative and discriminative models. However, in their model word alignment is still a standalone component in PB-SMT and cannot be tuned for PB-SMT performance. (Lambert et al., 2007) attempted to tune a discriminative word alignment model directly with MT in mind. Our work investigates the tuning of word alignment that takes advantage of both generative and discriminative word alignment models. (Ma et al., 2008) proposed a similar word alignment framework; however, their word alignment was only tuned according to AER and the improvement for PB-SMT system was not statistically significant. We show that by tuning word alignment according to PB-SMT performance, we can achieve significantly better results.

7 Conclusions and Future Work

In this paper, we proposed a flexible syntactically enhanced word alignment model that can be tuned according to different end tasks. This model takes the advantages of both unsupervised and supervised word alignment approaches by obtaining anchor alignments from unsupervised generative models and seeding the anchor alignments into a supervised discriminative model. This model offers the flexibility of tuning the alignment according to different optimisation criteria.

Our model is superior to generative word alignment models in terms of both intrinsic and extrinsic quality. We observed a 2.99% relative increase in F-score compared to the best baseline system. Using our word alignment in a PB-SMT system yields a 5.38% relative increase in BLEU score.

In the future, we first plan to conduct an in-depth investigation regarding what type of word alignments are beneficial to MT. We also plan to refine the optimisation criteria to avoid overfitting problems. Finally, we will conduct experiments in other domains and on other language pairs.

Acknowledgement

This work is supported by Science Foundation Ireland (O5/IN/1732 and 07/CE/I1142) and the Irish Centre for High-End Computing.[7] We would like to thank the reviewers for their insightful comments.

References

Brown, Peter F., 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.

Deng, Yonggang and William Byrne. 2006. MTTK: An alignment toolkit for statistical machine transla-

[7] http://www.ichec.ie/

tion. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 265–268, New York City, NY.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Fordyce, Cameron Shaw. 2007. Overview of the IWSLT 2007 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12, Trento, Italy.

Fraser, Alexander and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 392–399, Philadelphia, PA.

Ittycheriah, Abraham and Salim Roukos. 2005. A maximum entropy word aligner for Arabic-English machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 89–96, Vancouver, British Columbia, Canada.

Koehn, Philipp, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 48–54, Edmonton, Canada.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Lambert, Patrik, Rafael E. Banchs, and Josep M. Crego. 2007. Discriminative alignment training without annotated data for machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 85–88, Rochester, NY.

Liang, Percy, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 104–111, New York, NY.

Liu, Yang, Qun Liu, and Shouxun Lin. 2005. Log-linear models for word alignment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 459–466, Ann Arbor, MI.

Ma, Yanjun, Sylwia Ozdowska, Yanli Sun, and Andy Way. 2008. Improving word alignment using syntactic dependencies. In *Proceedings of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, pages 69–77, Columbus, OH.

Melamed, I. Dan. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

Moore, Robert C. 2005. A discriminative framework for bilingual word alignment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 81–88, Vancouver, BC, Canada.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Ervin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Noreen, Eric W. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley-Interscience, New York, NY.

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

Och, Franz. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Ratnaparkhi, Adwait. 1996. A maximum entropy model for part-of-speech tagging. In Brill, Eric and Kenneth Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142, Somerset, NJ.

Stolcke, Andrea. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, Denver, CO.

Vilar, David, Maja Popovic, and Hermann Ney. 2006. AER: Do we need to "improve" our alignments? In *Proceedings of the International Workshop on Spoken Language Translation*, pages 205–212, Kyoto, Japan.

Vogel, Stefan, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 836–841, Copenhagen, Denmark.