Improving Word Alignment Using Syntactic Dependencies

Yanjun Ma[1] **Sylwia Ozdowska**[1] **Yanli Sun**[2] **Andy Way**[1]
[1] School of Computing, Dublin City University, Dublin, Ireland
{yma,sozdowska,away}@computing.dcu.ie
[2] School of Applied Language and Intercultural Studies,
Dublin City University, Dublin, Ireland
yanli.sun2@mail.dcu.ie

Abstract

We introduce a word alignment framework that facilitates the incorporation of syntax encoded in bilingual dependency tree pairs. Our model consists of two sub-models: an anchor word alignment model which aims to find a set of high-precision anchor links and a syntax-enhanced word alignment model which focuses on aligning the remaining words relying on dependency information invoked by the acquired anchor links. We show that our syntax-enhanced word alignment approach leads to a 10.32% and 5.57% relative decrease in alignment error rate compared to a generative word alignment model and a *syntax-proof* discriminative word alignment model respectively. Furthermore, our approach is evaluated extrinsically using a phrase-based statistical machine translation system. The results show that SMT systems based on our word alignment approach tend to generate shorter outputs. Without length penalty, using our word alignments yields statistically significant improvement in Chinese–English machine translation in comparison with the baseline word alignment.

1 Introduction

Automatic word alignment can be defined as the problem of determining translational correspondences at word level given a parallel corpus of aligned sentences. Bilingual word alignment is a fundamental component of most approaches to statistical machine translation (SMT). Dominant approaches to word alignment can be classified into two main schools: generative and discriminative word alignment models.

Generative word alignment models, initially developed at IBM (Brown et al., 1993), and then augmented by an HMM-based model (Vogel et al., 1996), have provided powerful modeling capability for word alignment. However, it is very difficult to incorporate new features into these models. Discriminative word alignment models, based on discriminative training of a set of features (Liu et al., 2005; Moore, 2005), on the other hand, are more flexible to incorporate new features, and feature selection is essential to the performance of the system.

Syntactic annotation of bilingual corpora, which can be obtained more efficiently and accurately with the advances in monolingual language processing, is a potential information source for word alignment tasks. For example, Part-of-Speech (POS) tags of source and target words can be used to tackle the data sparseness problem in discriminative word alignment (Liu et al., 2005; Blunsom and Cohn, 2006). Shallow parsing has also been used to provide relevant information for alignment (Ren et al., 2007; Sun et al., 2000). Deeper syntax, e.g. phrase or dependency structures, has been shown useful in generative models (Wang and Zhou, 2004; Lopez and Resnik, 2005), heuristic-based models (Ayan et al., 2004; Ozdowska, 2004) and even for syntactically motivated models such as ITG (Wu, 1997; Cherry and Lin, 2006).

In this paper, we introduce an approach to improve word alignment by incorporating syntactic dependencies. Our approach is motivated by the fact that words tend to be dependent on each other. If

Proceedings of the Second ACL Workshop on Syntax and Structure in Statistical Translation (SSST-2), pages 69–77,
ACL-08: HLT, Columbus, Ohio, USA, June 2008.

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 illustrating example. Note that the link $(2, 4)$ can be easily identified, but the link involving the fourth Chinese word (a function word denoting 'time') $(4, 4)$ 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. Given such an observation, our model is composed of two related alignment models. The first one is an anchor alignment model which is used to find a set of anchor links; the other one is a syntax-enhanced alignment model aiming to process the words left unaligned after anchoring.

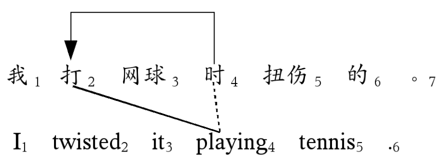

Figure 1: How syntactic dependencies can help word alignment: an example

The remainder of this paper is organized as follows. In Section 2, we introduce our syntax-enhanced discriminative word alignment approach. The feature functions used are described in Section 3. Experimental setting and results are presented in Section 4 and 5 respectively. In Section 6, we compare our approach with other related word alignment approaches. Section 7 concludes the paper and gives avenues for future work.

2 Word Alignment Model

2.1 Notation

While in this paper we focus on Chinese–English, the method proposed is applicable to any language pair. The notation will assume Chinese–English word alignment and Chinese–English MT. Here we adopt a notation similar to (Brown et al., 1993). Given a Chinese sentence c_1^J consisting of J words $\{c_1, ..., c_J\}$ and an English sentence e_1^I consisting of I words $e_1, ..., e_I$, we define the alignment A between c_1^J and e_1^I as a subset of the Cartesian product of the word positions:

$$A \subseteq \{(j, i) : j = 1, ..., J; i = 1, ..., I\}$$

Our alignment representation is restricted so that each source word can only be aligned to one target word. The alignment A consists of associations $j \rightarrow i = a_j$ from a source position j to a target position $i = a_j$. The 'null' alignment $a_j = 0$ with the 'empty' word e_0 is used to account for source words that are not aligned to any target word.

We use A_Δ to denote a subset of A. The indices of the K source words involved in A_Δ are represented as Δ_1^K and the corresponding target indices for Δ_k are represented as a_{Δ_k}. The unaligned source words are represented as $\bar{\Delta}$.

2.2 General Model

Given a source sentence c_1^J and target sentence e_1^I, we seek to find the optimum alignment $\hat{A}$ such that:

$$\hat{A} = \underset{A}{\operatorname{argmax}} P(A|c_1^J, e_1^I) \tag{1}$$

We use a model (2) that directly models the linkage between source and target words similarly to (Ittycheriah and Roukos, 2005). We decompose this model into an anchor alignment model (3) and a syntax-enhanced model (4) by distinguishing the anchor alignment from the non-anchor alignment.

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

2.3 Anchor Alignment Model

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. In this paper we adopted the following two approaches.

2.3.1 Heuristics-based Approach

The problem of word alignment is regarded as a process of word linkage disambiguation, i.e. choosing the correct links between words from all competing hypothesis (Melamed, 2000; Deng and Gao, 2007).

We constrain the link probabilities in such a way that:

$$\forall i' \in \{1, ..., I\}, i' \neq i : \frac{p((j,i))}{p((j,i'))} > \epsilon_1 \qquad (5)$$

$$\forall j' \in \{1, ..., J\}, j' \neq j : \frac{p((j,i))}{p((j',i))} > \epsilon_2 \qquad (6)$$

Condition (5) implies that for the source word c_j, the link with the target word e_i is more probable (with reliability threshold ϵ_1) than the link with any other target word. Condition (6) guarantees that for the target word e_i, c_j is the only most probable (with threshold ϵ_2) source word to be linked to.

2.3.2 Intersected Generative Word Alignment Models

We can use the asymmetric IBM models for bidirectional word alignment and get the intersection.

2.4 Syntax-Enhanced Word Alignment Model

The syntax-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:

$$\begin{aligned}\hat{a_j} &= \underset{a_j}{\operatorname{argmax}}\{p_{\lambda_1^M}(a_j|c_1^J, e_1^I, a_1^{j-1}, A_\Delta)\} \qquad (7)\\ &= \underset{a_j}{\operatorname{argmax}}\{\textstyle\sum_{m=1}^M \lambda_m h_m(c_1^J, e_1^I, a_1^j, A_\Delta, T_c, T_e)\}\end{aligned}$$

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.

3 Feature Function for Syntax-Enhanced Model

The various features used in our syntax-enhanced model can be classified into three groups: statistics-based features, syntactic features and relative distortion features.

3.1 Statistics-based Features

3.1.1 IBM model 1 score

IBM model 1 is a position-independent word alignment model which is often used to bootstrap parameters for more complex models. Model 1 models the conditional distribution and uses a uniform distribution for the dependencies between source word positions and target word positions.

$$Pr(c_1^J, a_1^J|e_1^I) = \frac{p(J|I)}{(I+1)^J}\prod_{j=1}^J p(c_j|e_{a_j}) \qquad (8)$$

3.1.2 Log-likelihood ratio

The log-likelihood ratio statistic has been found to be accurate for modeling the associations between rare events (Dunning, 1993). It has also been successfully used to measure the associations between word pairs (Melamed, 2000; Moore, 2005). Given the following contingency table:

	c_j	$\neg c_j$
e_i	a	b
$\neg e_i$	c	d

the log-likelihood ratio can be defined as:

$$G^2(c_j, e_i) = -2log\frac{B(a|a+b,p_1)B(c|c+d,p_2)}{B(a|a+b,p)B(c|c+d,p)}$$

where $B(k|n,p) = \binom{n}{k}p^k(1-p)^{n-k}$ are binomial probabilities. The probability parameters can be obtained using maximum likelihood estimates:

$$p_1 = \frac{a}{a+b}, p_2 = \frac{c}{c+d} \qquad (9)$$

$$p = \frac{a+c}{a+b+c+d} \qquad (10)$$

3.1.3 POS translation probability

The POS tags can provide effective information for addressing the data sparseness problem using the lexical features (Liu et al., 2005; Blunsom and Cohn, 2006). The POS translation probability can be easily obtained using maximum likelihood estimation from an annotated corpus:

$$Pr(T_c|T_e) = \frac{COL(T_c, T_e)}{COF(T_e)} \qquad (11)$$

where T_c is a Chinese word's POS tag and T_e is an English word's POS tag. $COL(T_c, T_e)$ is the count of T_c and T_e being linked to each other in the corpus, and $COF(T_e)$ is the frequency of T_e in the corpus.

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:

$$h_{DA-1} = \begin{cases} 1 & \text{if } \exists <c_j, R_c, c_{j'}>, <e_i, R_e, e_{i'}> \\ & \text{and } (j', i') \in A_\Delta, \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

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

$$h_{DA-2} = \begin{cases} 1 & \text{if } \exists <c_{j'}, R_c, c_j>, <e_{i'}, R_e, e_i> \\ & \text{and } (j', i') \in A_\Delta, \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

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

$$h_{DLA-1} = \begin{cases} 1 & \text{if } \exists <c_j, R_c, c_{j'}>, <e_i, R_e, e_{i'}> \\ & \text{and } (j', i') \in A_\Delta, R_c = R_e, \\ 0 & \text{otherwise.} \end{cases} \quad (14)$$

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 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:

$$h_{src-1-PRD} = \begin{cases} 1 & \text{if } \exists <c_j, R_c, c_{j'}> \\ & \text{and } R_c = \text{'PRD'}, \\ 0 & \text{otherwise.} \end{cases} \quad (15)$$

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.

$$h_{src-1-PRD} = \begin{cases} 1 & \text{if } i \in a_\Delta, \\ 0 & \text{otherwise.} \end{cases} \quad (16)$$

3.3 Relative distortion feature

We can design features encoding the relative distortion information which can be used to evaluate a candidate link by computing its relative position change with respect to the anchor alignment. The relative position change of a candidate link $l = (j, i)$ is formally defined as follows:

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

$$D(l) = min(|d_L|, |d_R|) \quad (17)$$
$$d_L = (j - j_L) - (i - i_L) \quad (18)$$
$$d_R = (j - j_R) - (i - i_R) \quad (19)$$

where (i_L, j_L) is the leftmost anchor link of l, (i_R, j_R) is the rightmost anchor link of l. The less the relative position changes, the more likely the candidate link is. With a set of anchor alignments, we can obtain the distribution of the relative position changes from an annotated corpus using maximum likelihood estimation. In our experiments, we used the following four probabilities: $p(D = 0)$, $p(D = 1, 2)$, $p(D = 3, 4)$ and $p(D > 4)$.

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), extracted from the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2002). This multilingual speech corpus contains sentences similar to those that are usually found in phrase-books for tourists going abroad.

We tagged all the sentences in the training and devset3 using a maximum entropy-based POS tagger–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 respectively.

4.1.1 Word Alignment

We manually annotated word alignments on devset3. Since manual word alignment is an ambiguous task, we also explicitly allow for ambiguous alignments, i.e. the links are marked as sure (*S*) or possible (*P*) (Och and Ney, 2003). 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.

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).[2] We used devset2 (506 sentence pairs, 16 references) to tune various parameters in the MT system and IWSLT 2007 test set (489 sentence pairs, 6 references) for testing.

4.2 Alignment Training and Search

In our experiments, we treated anchor alignment and syntax-enhanced alignment as separate processes in a pipeline. The anchor alignments are kept fixed so that the parameters in the syntax-enhanced model can be optimized.[3] We used the support vector machine (SVM) toolkit–SVM_light[4] to optimize the parameters in (7). Our model is constrained in such a way that each source word can only be aligned to one target word. Therefore, in training, we transform each possible link involving the words left unaligned after anchoring into an event. In testing, the source words are consumed in sequence and the target words serve as states. The SVM dual variable was used to measure the reliability of each candidate link and the alignment link for each word is made independently, which makes the alignment search much easier. A threshold t was set as the minimal reliability score for each link. t is optimized according to alignment error rate (21) on the validation set.

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 (Och and Ney, 2003) to derive the intersection and refined alignment.

4.3.2 Machine Translation

We use a standard log-linear phrase-based SMT (PB-SMT) model as a baseline: GIZA++ implementation of IBM word alignment model 4,[5] the refine-

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

[3] Note 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 syntax-enhanced model.

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

[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.

ment 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.

4.4 Evaluation

We evaluate the intrinsic quality of predicted alignment A with precision, recall and alignment error rate (AER). Slightly differently from (Och and Ney, 2003), we use possible alignments in computing recall.

$$recall = \frac{|A \cap P|}{|P|}, precision = \frac{|A \cap P|}{|A|} \quad (20)$$

$$AER(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \quad (21)$$

We also extrinsically measure the word alignment quality via a Chinese–English translation task. The translation output is measured using BLEU (Papineni et al., 2002).

5 Experimental Results

5.1 Word Alignment

We performed word alignment bidirectionally using our approach to obtain the union and compared our results with two strong baselines based on generative word alignment models. The results are shown in Table 1. We can see that both the syntax-enhanced model based on HMM intersection anchors (Syntax-HMM) and on IBM model 4 anchors (Syntax-Model 4) are better than the pure generative word alignment models. Our approach is superior in precision with a disadvantage in recall. The best result achieved 10.32% relative decrease in AER compared to the baseline when we use IBM model 4 intersection to obtain the set of anchor alignments.

model	precision	recall	f-score	AER
HMM refined	0.8043	0.7592	0.7811	0.2059
Syntax-HMM	0.8744	0.7304	0.7959	0.1845
Model 4 refined	0.7941	0.7987	0.7964	0.1929
Syntax-Model 4	0.8566	0.7685	0.8102	**0.1730**

Table 1: Comparing syntax-enhanced approach with generative word alignment

5.1.1 The Influence of Anchor Alignment Quality

As we can see in Table 2, our precision-oriented approach to acquire anchor alignments was accomplished quite well. All four different anchor alignment models achieved high precision. However, the recall differs dramatically, with model 4 achieving the highest recall and the heuristics-based approach receiving the lowest. To investigate the influence

anchor model	precision	recall	f-measure	AER
Heuristics	0.9774	0.4047	0.5724	0.3947
Model 1	0.9509	0.5011	0.6563	0.3157
HMM	0.9802	0.5327	0.6903	0.2809
Model 4	0.9777	0.5677	0.7179	0.2533

Table 2: Performance of anchor alignment

of the anchor alignment model, we first obtained the intersection of the words left unaligned after anchoring using each of the anchor alignment models. We evaluate the alignment of these words against the gold-standard alignments involving these words. The influence of anchor alignment on the performance of the syntax-enhanced model can be seen in Table 3. The performance of the syntax-enhanced model is closely related to that of the anchor alignment method. As can be seen from Table 2 and 3, HMM anchoring achieves the best precision and so does the syntax-enhanced alignment; IBM model 4 achieves the best recall and so does the syntax-enhanced alignment. Finally, the best alignment performances are obtained with IBM model 4 anchoring, with the difference in recall between HMM and IBM model 4 anchoring being more significant than the difference in precision.

anchor model	precision	recall	f-score	AER
Heuristics	0.4505	0.3270	0.3790	0.6210
Model 1	0.5538	0.3894	0.4573	0.5427
HMM	0.5932	0.3611	0.4489	0.5511
Model 4	0.5660	0.4216	0.4832	0.5168

Table 3: Influence of anchor alignment in syntax-enhanced model

5.1.2 The Influence of Syntactic Dependencies on Word Alignment

The influence of incorporating syntactic dependencies into the word alignment process is shown in Table 4. Syntax plays a positive role in all different anchor alignment configurations. The influence grows proportionally to the strength of the anchor alignment model. With the Model 4 intersection used as the set of anchor alignments, adding syntactic dependency features into the syntax-enhanced alignment model yields a 5.57% relative decrease in AER.

model	precision	recall	f-score	AER
Heuristics				
no syntax	0.8362	0.6751	0.7470	0.2302
w. syntax	0.8376	0.6894	0.7563	0.2240
Model 1				
no syntax	0.8759	0.6902	0.7720	0.2045
w. syntax	0.8542	0.7160	0.7790	0.2011
HMM				
no syntax	0.8655	0.7168	0.7841	0.1952
w. syntax	0.8744	0.7304	0.7959	0.1845
Model 4				
no syntax	0.8697	0.7340	0.7961	0.1832
w. syntax	0.8566	0.7685	0.8102	**0.1730**

Table 4: Influence of syntactic dependencies on word alignment

5.1.3 Contribution of Different Feature Classes

We interpret the contribution of each feature in terms of feature weights in SVM model training. The weights for the most discriminative features in each feature class in Chinese–English word alignment (using HMM intersection as anchor alignment) are shown in Table 5. As we can see, all statistics-based features are informative. Two target dependency features are informative: ‘PRD’ denoting ‘predicative’ dependency, and ‘AMOD’ denoting ‘adjective/adverb modifier’ dependency.

	weight
Model 1 Score	0.1416
POS	0.0540
Log-likelihood Ratio	0.0856
relative distortion	0.0606
DA-1	0.0227
DLA-2	0.0927
tgt-1-PRD	0.0961
tgt-2-AMOD	0.0621

Table 5: Weights of some informative features

5.2 Machine Translation

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 used a Chinese–English MT task to extrinsically evaluate the quality of our word alignment.

Table 6 shows the influence of our word alignment approach on MT quality.[6] On development set, we achieved statistically significant improvement using both our syntax-enhanced models—Syntax-HMM ($p<0.002$) and Syntax-Model 4 ($p<0.008$). On the test set, we observed that the MT output based on our alignment model tends to be shorter than the reference translations and the BLEU score is considerably penalized. If we ignore the length penalty (‘BP’ in Table 6) in significance testing, the improvement on test set is also statistically significant: $p<0.04$ for both Syntax-HMM and Syntax-Model 4. However, an indepth manual analysis needs to be carried out in order to determine the exact nature of the shorter sentences derived.

	dev. set	test set
Baseline	0.5412	0.3510 (BP=0.96)
Syntax-HMM	0.6015	0.3409 (BP=0.86)
Syntax-Model 4	0.5834	0.3585 (BP=0.91)

Table 6: The Influence of Word Alignment on MT

6 Comparison with Previous Work

Our syntax-enhanced model is a discriminative word alignment model. Certain generative word alignment models (e.g. HMM or IBM 4) also take the first-order dependencies into account. However, long distance dependencies between words are hard to incorporate into these models because of the explosive number of parameters. On the other hand, like existing discriminative models, our approach uses a set of informative features based on co-occurrence statistics, e.g. log-likelihood ratio and DICE score. The advantage of our approach is the mechanism by which syntactic features may be incorporated.

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

Some previous research also tried to make use of syntax in word alignment. (Wang and Zhou, 2004) investigated the benefit of monolingual parsing for alignment. They learned a generalized word association measure (crosslingual word similarities) based on monolingual dependency structures and improved alignment performances over IBM model 2 and certain heuristic-based models. (Cherry and Lin, 2006) used dependency structures as soft constraints to improve word alignment in an ITG framework. Compared to these models, our approach directly takes advantage of dependency relations as they are transformed into feature functions incorporated into a discriminative word alignment framework.

7 Conclusion and Future Work

In this paper, we proposed a model that can facilitate the incorporation of syntax into word alignment and measured the combination of a set of syntactic features. Experimental results have shown that syntax is useful in word alignment and especially effective in improving the recall. We have also observed that in our word alignment framework, the two submodels are closely related and the quality of the anchor alignment model plays an important role in the system performance.

The promising results will lead us to improve our model in the following aspects. First, the two submodels in our approach are two separate processes performed in pipeline. We plan to jointly optimize the two models in one go. Second, some of our experiments used complex IBM models, e.g. IBM Model 4, to obtain anchor alignment. We plan to boostrap the alignment using simple heuristics without relying on complex IBM models. Third, the alignment searching process assumed the alignment link for each word is made independently. A feasible markovian assumption will be tested for searching. Fourth, a comparison with traditional discriminative word alignment models is also necessary to justify the merits of our approach. Finally, we also plan to adapt our approach to larger data sets and more language pairs.

Acknowledgments

This work is supported by Science Foundation Ireland (grant number OS/IN/1732). Prof. Rebecca Hwa from University of Pittsburgh and Dr. Yang Liu from the Institute of Computing Technology, Chinese Academy of Sciences, are kindly acknowledged for providing us with their word alignment guidelines. We would also like to thank the anonymous reviewers for their insightful comments.

References

Necip Fazil Ayan, Bonnie Dorr, and Nizar Habash. 2004. Multi-align: Combining linguistic and statistical techniques to improve alignments for adaptable mt. In *Proceedings of the 6th Conference of the AMTA (AMTA-2004)*, pages 17–26, Washington DC.

Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with conditional random fields. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 65–72, Sydney, Australia.

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

Colin Cherry and Dekang Lin. 2006. Soft syntactic constraints for word alignment through discriminative training. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 105–112, Sydney, Australia.

Yonggang Deng and Yuqing Gao. 2007. Guiding statistical word alignment models with prior knowledge. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 1–8, Prague, Czech Republic.

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

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

Rebecca Hwa, 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.

Abraham Ittycheriah 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.

Philipp Koehn, 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.

Philipp Koehn, 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.

Percy Liang, 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.

Yang Liu, 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.

Adam Lopez and Philip Resnik. 2005. Improved HMM alignment models for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 83–86, Ann Arbor, Michigan, June.

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

Robert C. Moore. 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, British Columbia, Canada.

Joakim Nivre, 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.

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

Franz Och. 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.

Sylwia Ozdowska. 2004. Identifying correspondences between words: an approach based on a bilingual syntactic analysis of French/English parallel corpora. In *Proceedings of the COLING'04 Workshop on Multilingual Linguistic Resources*, pages 49–56, Geneva, Switzerland.

Kishore Papineni, 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.

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

Dengjun Ren, Hua Wu, and Haifeng Wang. 2007. Improving statistical word alignment with various clues. In *Machine Translation Summit XI*, pages 391–397, Copenhagen, Denmark.

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

Le Sun, Youbing Jin, Lin Du, and Yufang Sun. 2000. Word alignment of English-Chinese bilingual corpus based on chunks. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical Methods in Natural Language Processing and very large corpora*, pages 110–116.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of Third International Conference on Language Resources and Evaluation 2002*, pages 147–152, Las Palmas, Spain.

David Vilar, 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.

Stefan Vogel, 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.

Wei Wang and Ming Zhou. 2004. Improving word alignment models using structured monolingual corpora. In Dekang Lin and Dekai Wu, editors, *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 198–205, Barcelona, Spain.

Dekai Wu. 1997. Stochastic Inversion Transduction Grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.