A Feature-rich Supervised Word Alignment Model for Phrase-based SMT

Chooi-Ling Goh and Eiichiro Sumita
NICT, MASTAR Project, Language Translation Group
619-0288 Kyoto, Japan
{chooiling.goh,eiichiro.sumita}@nict.go.jp

Abstract

Word alignment plays an important role in statistical machine translation (SMT) systems. The output of word alignment can be used to decode new sentences. Most current SMT systems use GIZA++, a generative model, to automatically align words from sentence-aligned parallel corpora. GIZA++ works well when large sentence-aligned corpora are used. However, it is difficult to encode syntactic and lexical features such as POS tags, affixes, lemmas, etc., which are useful for handling sparse data and unseen words, using generative models. A discriminative model such as conditional random fields (CRF) can solve this problem. We treat word alignment as a labeling problem, and we encode the syntactic, lexical, and contextual features. Our experiments were conducted using a 35K Chinese-English hand-aligned corpus. Our model gives better word alignment results than GIZA++ by 6.7% F-measure. Finally, we also prove that 2% higher translation score can be obtained with phrase-based SMT systems when our alignment models are used.

1 Introduction

Current research has shown that statistical machine translation (SMT) systems generate better translations than other systems such as those using example-based and rule-based methods, especially in the case where large sentence-aligned parallel corpora are present. In SMT systems, the system can be easily trained so long as there exist parallel bilingual corpora for each language pair. However, while these corpora are typically sentence aligned, before constructing the translation model, one must automatically match the words with their translations; this is referred to as word alignment. The predicated word alignments are then used to build a phrase table which is necessary during decoding in the case of phrase-based SMTs (Koehn et al., 2003; Och and Ney, 2004).

Currently, generative models for word alignment, such as GIZA++ (Och and Ney, 2003), which is based on the IBM models (Brown et al., 1993), are widely used for SMT systems. GIZA++ gives good results when it is trained on a large parallel corpora. Moreover, it functions very well with pairs comprising similar languages such as English and German; however, similar performance is not obtained when language pairs that are very different in their syntactic structures, such as English-Chinese pair, are aligned. While GIZA++ does attempt to align most of the words between the sentences (few null alignments) and retains a high recall with alignment, simultaneously, it creates more fake alignments, i.e., its precision is low.

With the increase in available labeled data, recent research has investigated supervised or semi-supervised alignment (Blunsom and Cohn, 2006; Fraser and Marcu, 2006; Wu et al., 2006; Moore, 2005; Taskar et al., 2005; Liu et al., 2005) using discriminative models. Discriminative models allow the introduction of various features, either lexically, syntactically, or statistically during the training. Previous results have shown that discriminative models outperformed generative models in both precision and recall.

In this study, we apply a discriminative model, conditional random fields (CRF), to solve the word alignment problem. We name this model SuperAlign since it is a supervised model that is powerful (efficient) in learning the features. The alignment problem is treated as a labeling problem of a pair of words given some features such as Dice, relative sentence position, existence in a bilingual dictionary, part-of-speech tags, word lemmas on inflectional languages and contexts. Our experiment was performed on a word-aligned corpus of 35K sentences between Chinese and English. The results showed that SuperAlign has high accuracy which is useful in improving the translation quality in a phrase-based SMT.

2 Related Work

Our method is based on the concept proposed in (Blunsom and Cohn, 2006). They trained a CRF model for inducing word alignment from sentence-aligned data. They have introduced some linguistic features and incorporated the output of GIZA++ (models 1 and 4) as features. Due to the similarity between European languages, the also introduced orthographic features (English-French and English-Romanian). However, their improvement on the alignment is not sufficient for improving the translation quality.

There are a few more discriminative models (Moore, 2005; Taskar et al., 2005; Liu et al., 2005); these models share similar features, and they were the very first researches on discriminative word alignment models using hand-aligned training data. These researches provide some insights into the incorporation of more features, either lexically, syntactically, or statistically, to create a better model.

While most of the previous studies have used the output of GIZA++ as part of the features, we propose not to incorporate any features from it. This is because we do not want our model to work "like" GIZA++ since although GIZA++ gives high recall in alignment, its precision is not satisfactory. It generates many erroneous links, and in phrase-based SMTs, such error links will cause problems in creating the translation table required during decoding. We would like to have a model that can produce "good" align points, and during the translation model creation phase, only necessary phrases for translation are generated.

3 Word Alignment with CRF

In SuperAlign, word alignment is treated as a sequential labeling problem. Each pair of words is assigned some features and trained using a discriminative model, Conditional Random Fields (Lafferty et al., 2001). We use a public training tool CRF++[1], which is easy and fast, for training and decoding.

3.1 Sequence labeling

First, for each sentence pair, we build a list of word pairs $n \times m$ where n = # of Chinese words and m = # of English words. Our task is to label each pair of words into 4 categories: strong, weak, pseudo, or null. Strong links refer to words that are very good translations. Compound words and some possible alignments are represented by weak links. The alignments of functional words such as articles and prepositions are indicated using pseudo links.

3.2 Features

In order to train the CRF model, we must prepare a feature set. The features are chosen such that they will provide certain clues for the alignments. CRF allows the use of arbitrary and overlapping features. Hence, we are free to introduce any possible features such as syntactical, lexical, and contextual features.

3.2.1 Dice coefficient

The most useful feature is probably the Dice coefficient, which is an estimation of the closeness of two words. The word association is calculated using sentence aligned corpus.

$$Dice(e,f) = \frac{2 \times C_{EF}(e,f)}{C_E(e) + C_F(f)}$$

Here C_E and C_F represent the number of occurrences of the words e and f in the corpus while C_{EF} represents the number of co-occurrences. A high (low) value indicates that the word pair is closely (loosely) related to each other.

3.2.2 Bilingual dictionary

The second measurement parameter for the two words can be a bilingual dictionary. If the pair of words exists in the same entry in the dictionary, there is a high possibility that they can be aligned together. However, many words belonging to one language are not always translated to one single word in the other language. A word in a source language can be translated to a compound word in the other language and vice versa. This is especially true for translations between languages that are fairly different syntactically, such as, in our case, Chinese and English.

Therefore, the similarity between the two words is calculated as follows:

Bi-dic = $Sim(e,E)$ = Max($Sim(e,e_i) = \frac{1}{|e_i|}$ if $e \in e_i$ and $e_i \in E$ else 0)

Here, our source language is Chinese and the target language is English. Assume that the word pair that we consider for alignment is (c,e). Then, we search for the translation for c in the dictionary. There may exist multiple translations for c, i.e., E. We compare e and E as given in the equation above. For each translation e_i in E, if there

[1] http://crfpp.sourceforge.net/

is a one-to-one match, that is, if $e = e_i$, then the score is 1; else, the score is $1/N$ if word e exists in e_i where N is the number of words in the translation e_i; else, the score is 0. If the word e matches a few translations, we only take the maximum value. In this experiment, we use the LDC CEDICT dictionary, which contains 54,170 entries. It is not the ideal dictionary to use since the size is small, but, currently, we stick to it alone while looking for other choices.

3.2.3 Relative sentence position

$$Relpos = abs(\frac{a_t}{|e|} - \frac{t}{|f|})$$

where a_t is the postion of the aligned source word in e, and t is the position of the target word in f

The relative sentence position allows the model to learn the preferences for aligning words that are close to the alignment matrix diagonal. If two languages share similar grammar structures, this feature is useful. However, in the case of English and Chinese language pairs, this may be only of small assistance since the sentence structures mostly are different, and the alignment will not be placed on the diagonal. However, the phrase structures between them are sometimes fairly similar, and therefore, this feature might still be useful.

3.2.4 POS tags

In order to reduce the sparseness of the lexical words, part-of-speech tags for both languages are used as features. The English text is tagged with TreeTagger[2], and the Chinese text is tagged with an inhouse tagger that tags segmented text[3]. TreeTagger uses the Penn Treebank POS tagset while the Chinese tagger is trained using the Penn Chinese Treebank. Since both taggers share a similar tagset, we think that the POS tags can be matched to reduce the sparsity of the translations.

3.2.5 Lemmatization

While English is an inflectional lsanguage, Chinese words do not show any morphological changes. There are no conjugations in Chinese. Therefore, a word in present tense or past tense in English can be aligned to the same Chinese word. The tenses in Chinese are represented by some adverbs or are context-based. In order to reduce such sparsity, the English lemma is used. This is not necessary for Chinese since it is not an inflectional language. With the matching of inflectional words, this alignment can be enhanced even further. We also use the same English TreeTagger for their lemmas.

3.2.6 Contextual features

While GIZA++ enforces the competition for alignment between words, the outputs of Models 1 and 4 are used as features in (Blunsom and Cohn, 2006; Taskar et al., 2005) in order to bootstrap the training of the alignment. In our approach, we try not to use any features from GIZA++ since that will force our model to work like GIZA++. Therefore, we introduce a new set of contextual features that allow our learning to consider the competition between the adjacent words. Since our learning method is similar to a sequential labeling problem, the contexts can be the words and POS tags before and after current word pairs. Both Chinese and English contexts are added as the features.

4 Experiments

In this experiment, we use the Chinese-English hand-aligned Basic Traveler Expression Corpus (BTEC) for the training of CRF alignments. It consists of 35,384 sentence pairs with 369,587 links; of these links, 54.17% are strong links, 25.34% are weak links, and 20.49% are pseudo links.

Then, we use an IWSLT[4] evaluation campaign corpus to test the effectiveness of our alignment. The effects of CRF alignment on a phrase-based SMT system will be reported.

4.1 Experimental Results on Alignment

In the experiments on word alignment, we randomly chose a portion of 1000 sentence pairs as held out data and 999 sentence pairs as testing data. Finally, we retained 33K as the training data.

We measure the accuracy of the alignment using precision, recall, and F-measure, as given in the equations below; here, A represents the gold-standard alignments; S, the output alignments; and $A \cap S$, the correct alignments. In this case, we do not consider the different types of links.

$$precision = \frac{|A \cap S|}{|S|} \qquad recall = \frac{|A \cap S|}{|A|}$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}$$

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[3] In our case, the Chinese text must be pre-segmented as what we already have in our bilingual corpus.

Features	Prec (%)	Rec (%)	F-mea
All unigram	91.48	60.81	73.06
-sentence position	85.39	59.09	69.85
-Dice	88.19	49.43	63.35
-bilingual dictionary	90.89	57.00	70.07
-Chinese POS tags	91.33	61.10	73.22
-English lemma	91.12	60.89	73.00
-English POS tags	91.39	60.93	73.12
+context	90.37	63.46	74.56
All multi-gram	89.57	77.76	82.67
All multi-gram+context	89.84	79.91	84.59

Table 1: Comparison between features

Table 1 shows the results obtained when each feature is subtracted from the full model; we do this to find out which feature is useful for our task. Dice is the most useful feature, followed by relative sentence position and bilingual dictionary. POS tags and lemmatization do not improve the F-measure much (and they sometimes even deteriorate it) but they do improve precision. By adding contextual features, we further improve the accuracy. Thus far, all the features barring contextual features are unigram. We have also tried some bigram and trigram features, which gives us an incremental improvement. The combination of bigram and trigram features is determined using the held out data. The multi-gram features used by us are as follows:

- unigram features (C-word, E-word, relpos, Dice, Bi-dic, C-POS, E-lemma, E-POS)

- bigram features (C-word/E-lemma, C-word/C-POS, E-lemma/E-POS, C-POS/E-POS)

- trigram features (C-word/E-lemma/relpos, C-word/E-lemma/Dice, C-word/E-lemma/Bi-dic)

- contextual features (C-word-1, C-pos-1, E-lemma-1, E-pos-1, C-word+1, C-pos+1, E-lemma+1, E-pos+1, C-word-1/C-word, C-pos-1/C-pos, E-lemma-1/E-lemma, E-pos-1/E-pos, C-word/C-word+1, C-pos/C-pos+1, E-lemma/E-lemma+1, E-pos/E-pos+1)

Finally, by adding all the features together, we obtain the highest F-measure of 84.59 points.

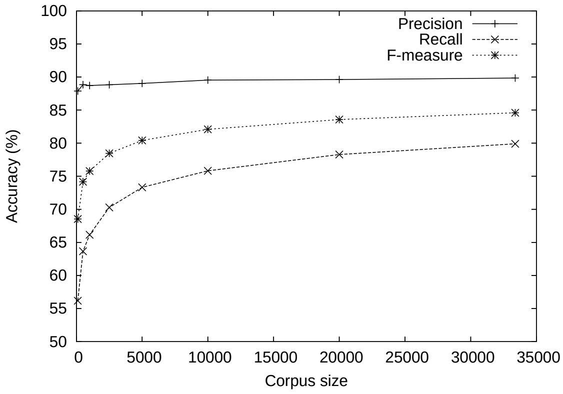

Figure 1: The accuracy of alignment versus size of training corpus

Method	Prec (%)	Rec (%)	F-mea
CRF (+context)	89.84	79.91	84.59
–strong	93.03	89.28	91.11
–weak	71.49	63.90	67.48
–pseudo	69.10	47.31	56.17
CRF (5000)	89.04	73.32	80.42
CRF (1000)	88.72	66.15	75.79
GIZA++(all)	76.51	79.38	77.92
GIZA++(test)	62.05	67.23	64.54

Table 2: Comparison with GIZA++ alignment

Obtaining a hand-aligned training corpus is not an easy task. It is both resource- and time-consuming. Since our method requires a training corpus, we would also like to determine the amount of training data that is necessary for a reasonable result. Figure 1 shows a graphical output of the accuracy versus the size of training corpus. The increment of accuracy becomes slower after 10,000 training sentences. Hence, we can conclude that perhaps approximately 10,000 sentence pairs is sufficient to train the CRF alignment model for any new language pair.

Next, we would like to compare the accuracy obtained by using GIZA++ ($1^5H^53^34^3$) refined with the grow-diag-final-and method with SuperAlign. Table 2 shows the results for each type of links and a comparison with GIZA++. SuperAlign performs very well as far as labeling strong links is concerned since they are the easiest links to detect. Its performance is good for weak links but not very satisfactory for pseudo links. As explained earlier, pseudo links are mostly functional words that are not direct translations of each other. They highly depend on the context for determining the alignments. In other words, ambiguity is high since a word can be linked to different words depending on the context. Hence, the accuracy of alignment of pseudo links is low.

In our experiment, we have trained two GIZA++ models. The first model uses all 35k

[4] http://www.slc.atr.jp/IWSLT2008/

training data, including held-out and testing data. The second model uses only the testing data. The results show that the results of the second model is much worse than the first. This also proves that GIZA++ requires a bigger training corpus in order to have good performance.

In contrast, SuperAlign obtains F-measure that are equivalent to GIZA++ (trained with 35k) even when it is trained using only 1000 sentence pairs. When the full training data was used, SuperAlign outperformed GIZA++ by approximately 6.7% F-measure. The biggest advantage of SuperAlign was the precision gained. GIZA++ has good recall but the precision was relatively low. SuperAlign can always guarantee high precision even with a small set of training data. However, with only 1000 sentence pairs, the recall is quite low as compared to GIZA++, although the results for F-measure are equivalent. However, with 5000 sentence pairs, SuperAlign becomes better than GIZA++ by a large margin. In the following section, we will see how the precision and recall of alignments affect the translation quality.

4.2 Experimental Results on Translation

The first experiment is to test whether the hand-aligned corpus is really helpful in improving the translation quality in phrase-based statistical machine translations. We use the 35K hand-aligned corpus as the training corpus for the phrase-based SMTs. Moses (Koehn et al., 2007) is used as the training toolkit, and the decoder is an inhouse standard phrase-based decoder, CleopATRa. During the training, the refined method that begins from intersection and then increases to the neighbouring alignments (option grow-diag-final-and) is used to combine the output of GIZA++ in both directions. We directly replaced the output of these two steps when training Moses with the hand-aligned output. The development data (IWSLT 2005 test data) used for the optimization with a minimum error rate trainer (MERT) is identical for all our experiments. The testing data is obtained from IWSLT 2008, 2007, and 2006 testing data.

Table 3 shows the results of translations using the hand-aligned corpus as the training data. The results are measured using the BLEU score, which is a geometric mean of n-gram precision with respect to N reference translations, and METEOR, which calculates unigram overlaps between translations and reference texts using various levels of matches (exact, stem, synonym). The average of the two measures is used as the evaluation metric. In general, we obtain better scores than GIZA++ (by around 2%). However, while GIZA++ leads to more alignment points and the phrase table is smaller, our aligned corpus produces less alignments points but with a larger phrase table, as shown in the row BTEC (swp). We also test the translation quality by excluding the pseudo links as shown in the row BTEC (sw). The difference between the two models is not sufficiently clear to tell whether the pseudo links are useful in building the phrase table. However, since using all the links leads to a smaller phrase table, which, in turn, is faster during decoding, we conclude that the alignment of pseudo links is helpful in reducing the size of the phrase table but not in improving the quality of the translation.

Next, we will test the SuperAlign model on a real run. In this experiment, we use the IWSLT 2008 training corpus (20K) for the training of the phrase-based SMT model. The development data and testing data are the same as in the previous experiment. Table 4 shows the experiment results. As predicted from the previous experiments, SuperAlign leads to better translation quality by approximately 2% accuracy for all testing datasets. The experiment also showed that 1000 training sentence pairs for SuperAlign can give results equivalent to those obtained using GIZA++. However, since the recall is low when 1000 training pairs are used, the phrase table becomes approximately thrice than that when GIZA++ is used. Here, we can also conclude that precision plays an important role in creating the translation model. If we can ensure that only correct links are produced in the alignment phase, then the null links can be accounted for by the phrase-table creation phase.

SuperAlign also helps in reducing the non-translated words in the source. The last column in Table 4 shows the total number of non-translated (unknown) words from all the testing data. In other words, some of the words that have not been aligned with GIZA++ have been successfully aligned using SuperAlign.

5 Conclusion and Future Work

In this paper, we have introduced a supervised word alignment using a discriminative model, Conditional Random Fields. We trained the models using 35K sentences of hand-aligned corpus. Our experimental results show that SuperAlign achieved higher accuracy than an unsupervised

	2008			2007			2006			# of	size of	Total # of
	bleu	meteor	b+m/2	bleu	meteor	b+m/2	bleu	meteor	b+m/2	align points	phrase table	non-translated
GIZA++	0.4716	0.6064	0.5390	0.3075	0.5231	0.4153	0.1837	0.4335	0.3086	375,353	626,502	583
BTEC (swp)	0.4890	0.6309	0.5599	0.3332	0.5450	0.4391	0.2036	0.4630	0.3333	369,587	661,104	497
BTEC (sw)	0.4996	0.6221	0.5608	0.3129	0.5378	0.4253	0.1867	0.4524	0.3195	293,848	1,339,597	479

Table 3: Translation results obtained trained with 35K hand-aligned BTEC corpus

	2008			2007			2006			# of	size of	Total # of
	bleu	meteor	b+m/2	bleu	meteor	b+m/2	bleu	meteor	b+m/2	align points	phrase table	non-translated
GIZA++	0.4042	0.5823	0.4932	0.2707	0.5063	0.3885	0.1614	0.4293	0.2953	212,869	357,237	791
CRF (swp)	0.4325	0.6049	0.5187	0.2838	0.5187	0.4012	0.1785	0.4425	0.3105	183,535	593,841	662
CRF (sw)	0.4397	0.6006	0.5201	0.2861	0.5199	0.4030	0.1762	0.4399	0.3080	151,545	964,829	650
CRF (1000)	0.4199	0.5787	0.4993	0.2736	0.5086	0.3911	0.1456	0.4210	0.2833	153,432	957,325	786

Table 4: Translation results obtained trained with IWSLT 2008 training set

generative model, GIZA++ by 6.7% F-measure. SuperAlign always gives high precision no matter how small the training data is. Finally, we also proved that the alignment output by SuperAlign improved the quality of translation in a phrase-based SMT system.

However, compared to GIZA++, SuperAlign produced more null links. In future research, we will try to obtain methods to reduce the null links. Although the presence of null links does not affect the translation quality too much, they increase the size of the phrase table, thereby affecting the decoding time. Further, we would also like to apply SuperAlign on different language pairs to prove that our hypothesis works for any language pair. Our current corpus BTEC is an oral corpus in which the sentences are short and drawn from the travel domain. We will try our method on a corpus in a different domain in which the sentence length is longer and the sentence structure is more complicated.

Acknowledgments: This work is partly supported by the Grant-in-Aid for Scientific Research (C) and the Special Coordination Funds for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with conditional random fields. In *Proceedings of COLING/ACL*, pages 65–72.

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

Philip Koehn et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL Demo and Poster Sessions*, pages 177–180.

Alexander Fraser and Daniel Marcu. 2006. Semi-supervised training for statistical word alignment. In *Proceedings of COLING/ACL*, pages 769–776.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*, pages 81–88.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289.

Yang Liu, Qun Liu, and Shouxun Lin. 2005. Log-linear models for word alignment. In *Proceedings of ACL*, pages 459–466.

Robert C. Moore. 2005. A discriminative framework for bilingual word alignment. In *Proceedings of HLT/EMNLP*, pages 81–88.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Ben Taskar, Simon Lacoste-Julien, and Dan Klein. 2005. A discriminative matching approach to word alignment. In *Proceedings of HLT/EMNLP*, pages 73–80.

Hua Wu, Haifeng Wang, and Zhanyi Liu. 2006. Boosting statistical word alignment using labeled and unlabeled data. In *Proceedings of COLING/ACL Poster Session*, pages 913–920.