A Feature-rich Supervised Word Alignment Model
for Phrase-based Statistical Machine Translation

Chooi-Ling Goh and Eiichiro Sumita

Language Translation Group

MASTAR Project

Knowledge Creating Communication Research Center

National Institute of Information and Communications Technology

3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, 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 build a phrase table, which is the core model in the decoding of 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 useful for handling sparse data and unseen words, such as POS tags, affixes, lemmas, etc., using generative models. A discriminative model such as conditional random fields (CRF) can solve this problem. We treat word alignment as a labelling problem, and 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 7% AER. Finally, we also prove that 2% higher BLEU score can be obtained with phrase-based SMT systems when our alignment models are used.

Keywords

Word alignment, statistical machine translation, machine learning, conditional random fields, sequential labelling.

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 of large sentence-aligned parallel corpora being present. In SMT systems, the system can be easily trained so long as parallel bilingual corpora exist for any language pair. However, while these corpora are typically sentence aligned, before

constructing the translation model, ones 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; phrase tables are necessary during decoding in the case of phrase-based SMTs (Koehn et al., 2003, Och and Ney, 2004).

Over the years, whether better word alignment leads to better translations has been a subject of dispute. Recently, Granchev et al. (2008) performed an extensive evaluation and showed that improvements in alignment accuracy would lead to improvements in machine translation. However, there still exists an agreement constraint between them. Therefore, a good word alignment model is still necessary.

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 large parallel corpora. Moreover, it functions very well with pairs comprising similar languages such as English and German; however, similar performances are not obtained when language pairs that are very different in their syntactic structures, such as English-Chinese and Japanese-English pairs, 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, it simultaneously creates more fake alignments (i.e., its precision is low).

A high recall definitely improves translation quality in the sense that the number of non-translated words is reduced, but low precision decreases the quality of translation. Therefore, a trade-off between recall and precision is very important for producing high-quality translations. In a phrase-based SMT system, a phrase table is generated after word alignment. Words that could not be aligned are freely attached to some phrases based on the context. A high recall and low precision in alignment will lead to fewer phrases being generated whereas a low recall and high precision will lead to more phrases being generated. High precision can be easily obtained only if high-accuracy links are generated. However, then the recall might be too low. The best situation would be a case wherein recall is improved and precision is maintained, and this is the aim of our study. In our research, we aim to train a model that can yield high precision with a comparable recall.

With the increase in numerous labeled data, recent research has investigated supervised or semi-supervised alignment (Ittycheriah and Roukos, 2005, Blunsom and Cohn, 2006, Fraser and Marcu, 2006, Wu et al., 2006, Moore, 2005, Taskar et al, 2005, Liu:et al., 2005). The current trend among researchers is to move from generative to 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 and assigned features. Our aim is to find an optimum set of features which are useful for alignment, such as Dice, relative sentence position, existence in a bilingual dictionary, part-of-speech tags, and word lemmas on inflectional languages. We can then create unigram and multi-gram features from them. Moreover, the words and POS tags in contexts are also used as features similar to that of a common sequential labeling problem. At the end, we also apply a heuristic model to further increase the recall. Our experiment was performed on a word-aligned corpus of 35K sentences between Chinese and English. The results have shown that SuperAlign has high accuracy especially in precision. Moreover, in the second part of our experiments, we have

also proven that good alignment results are useful in improving the translation quality in a phrase-based SMT.

2.　Hand Aligned Corpus

First, since our model is a supervised model, we must prepare certain word-aligned data for the training. Our aim is to construct a corpus that will be aligned in such a way that it is suitable to be used for a SMT system.
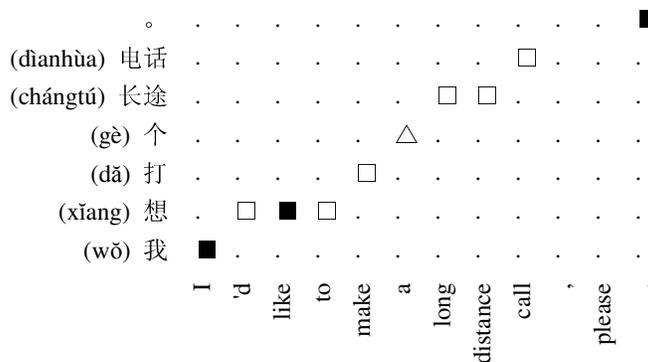
	I	'd	like	to	make	a	long	distance	call	,	please	.
。												■
(diànhùa) 电话									□			
(chángtú) 长途							□	□				
(gè) 个					△							
(dǎ) 打				□								
(xiǎng) 想		□	■	□								
(wǒ) 我	■											

Figure 1 Sample of alignment – diagonal alignment

	Tell	me	when	we	get	to	the	museum	.
。									
(shēng) 声	△								
(yì) 一	△								
(wǒ) 我		■							
(tōngzhī) 通知	■								
(qǐng) 请									
(shíhòu) 时候			■						
(de) 的									
(bówùguǎn) 博物馆							△	■	
(dào) 到					■	□			
(yào) 要									
(wǒmen) 我们				■					

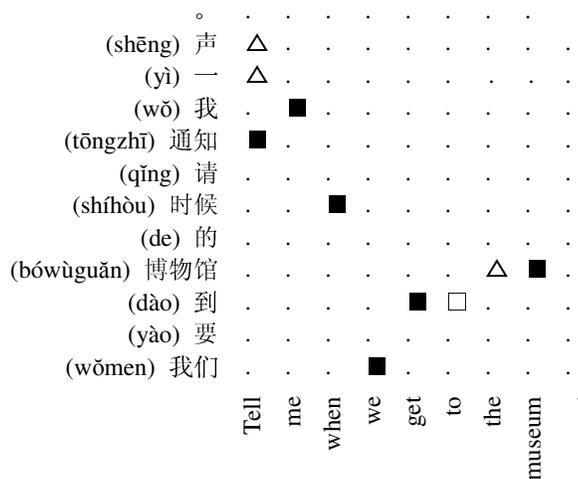Figure 2 Sample of alignment – scatter alignment

We have constructed a hand-aligned corpus with the co-operation of a research institute in China[1]. Four types of alignment links have been defined: strong, weak, pseudo,

---

[1]　Institute of Computing Technology, Chinese Academy of Sciences

and null, as proposed by Zhao et al. (2009). Strong links refer to words that are very good translations. Compound words and some possible alignments are represented by weak links. Both strong and weak links are considered to be genuine links. The alignments of functional words such as articles and prepositions are indicated using pseudo links. Finally, null links refer to words that do not align with any words. Figure 1 shows an example of an alignment; here, ■ represents a strong link, □ represents a weak link, and △ represents a pseudo link.

Chinese and English can be said to be two languages that have slightly different syntactic structures. Therefore, it is quite common for the translations between them to not align in a diagonal matrix space. Figure 2 shows an example in which the alignments are scattered across the matrix. Such alignments are difficult to resolve.

```
              ｡    .    .    .    .    ■
       (wèi) 味   .    .    △   □    .
       (hàn) 汗   .    .    △   ■    .
       (yǒu) 有   .    .    △   □    .
  (chènshān) 衬衫  .    ■    .    .    .
        (de) 的   □                   .
        (wǒ) 我   □    .    .    .    .
                 My   shirt  is  sweaty    .
```

Figure 3 Sample of alignment – phrase alignment

The manual alignment considers also alignment of phrases (Figure 3). For example, the phrase *is sweaty* is aligned with the phrase 有 汗 味. When we observe the annotations for each pair of words, we can find 3 types of links present in it. We will follow this alignment standard in our research for alignment and translation.

3.   Word Alignment with CRF

In SuperAlign, word alignment is treated as a sequential labeling problem. Each pair of words is assigned with some features and trained using a discriminative model, Conditional Random Fields. CRF has proven to be efficient in labeling sequential data (Lafferty et al., 2001). Moreover, it has been used for various natural language processing tasks such as morphological analysis, parsing, named entity recognition, information extraction, and text chunking.

A linear-chain CRF with parameters $\Lambda = \{\lambda_1, ..., \lambda_K\}$ defines a conditional probability for a label sequence $\mathbf{y} = y_1...y_T$ given an input sequence $\mathbf{x} = x_1...x_T$ to be:

$$P_\Lambda(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z_x} \exp\left(\sum_{t=1}^{T} \sum_k \lambda_k f_k(y_{t-1}, y_t, \mathbf{x}) \right)$$

where Z_x is the normalization factor that makes the probability of all state sequences sum to one; $f_k(y_{t-1}, y_t, \mathbf{x})$ is a feature function, and λ_k is a learned weight associated with feature f_k..

We use a public training tool CRF++[2], which is easy and fast, for training and decoding. For simplification, the following explanations assume that the source language is Chinese and the target language is English.

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.

3.2 Features

In order to train the CRF model, we must prepare a feature set. The features are chosen in a way 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 (Blunsom and Cohn, 2006, Ittycheriah and Roukos, 2005).

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 into one single word in the other language. A word in a source language can be translated into 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:

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

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

---

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

given in the equation above. For each translation e_i in E, if there 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, we are currently using it while we look for other choices.

3.2.3 Relative sentence position

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

where a_t is the position 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 not be helpful when the sentence structures are different, and the alignment is not 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[3], and the Chinese text is tagged with an in-house tagger that tags segmented text[4]. 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 sparseness of the translations.

3.2.5 Lemmatization

While English is an inflectional language, 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 sparseness, 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

---

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/Tree Tagger/
[4] In our case, the Chinese text must be pre-segmented as what we already have in our bilingual corpus.

introduce a new set of contextual features that allow our training 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.

3.2.7 Multi-gram features

In some previous work on sequential labeling problem, combination of different features showed significance improvement in the results. Here, we also try to combine features as either bigrams (using two features) or trigrams (using three features). The combination is normally done on features that are related to each other. In this case, the multi-gram features will become stronger than the single feature. The details of the combination will be given in the experiment section.

3.3 Heuristic model

$s1\ s2\ s3\ s4\ s5\ s6$　　　　　　$s1\ s2\ s3\ s4\ s5\ s6$

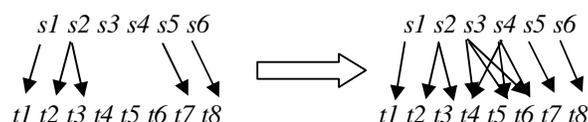$t1\ t2\ t3\ t4\ t5\ t6\ t7\ t8$　　　　$t1\ t2\ t3\ t4\ t5\ t6\ t7\ t8$

Figure 4 A heuristic model

A heuristic model is further used to increase the recall. Let's assume that sentence s is aligned with sentence t, and the word alignment output from SuperAlign is as shown at the left hand side of Figure 4. There exist some null links, ($s3\ s4$) and ($t4\ t5\ t6$), meaning words not aligned with any words. Then, by heuristic, the two words in s can be aligned with the three words in t. As a result, the alignment will be as shown at the right hand side of Figure 4. Empirically, we limit this kind of alignment to maximum 3 words at each side. The heuristic model is helpful when either the source words or the target words do not exist in the training corpus, the translation cannot be found in the dictionary or the POS tags are not tagged properly by the POS tagger.

4.　Experiments

To prove the effectiveness of our method, we use the Chinese-English hand-aligned Basic Traveler Expression Corpus (BTEC) (Kikui et al., 2006) 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[5] 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

---

[5]　http://www.slc.atr.jp/IWSLT2008/

training data.

Alignment error rate (AER) is a measurement parameter for alignment tasks proposed by (Och and Ney, 2003). AER is calculated based on the *Sure* and *Possible* links. However, according to a survey conducted by (Fraser and Marcu, 2007), AER does not correlate with the translation quality (BLEU score). The F-measure calculated by varying the trade-off between recall and precision has a better correlation. In their study, a constant *α* was used as the weight applied to recall (*α*) and precision (*1-α*). A value of less than 0.5 places more weight on recall and vice versa. Finding a good *α* setting is not straightforward and depends highly on the language pairs and the size of the corpus. Therefore, we only used the simple balanced F-measure, i.e., *α=0.5*, to evaluate the performance of our alignment models.

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∩S*, the correct alignments. In this case, we do not consider the different types of links.

$$precision \ = \frac{|\,A \bigcap S\,|}{|\,S\,|}$$

$$recall \ = \frac{|\,A \bigcap S\,|}{|\,A\,|}$$

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

Features	Precision (%)	Recall (%)	F-measure
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
All +160K Dice	89.56	80.42	84.74
All +heuristic	87.58	82.24	84.83

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 we used 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 (before and after)
  - 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. In our feature set, Dice and bilingual dictionary features are independent of our training corpus. Therefore, if we can obtain a larger bilingual sentence-aligned corpus, we can recalculate the Dice. The second last line of Table 1 shows that if we use the Dice calculated using a 160K sentence-aligned corpus, we can further increase the F-measure to 84.74. Currently, we do not have a larger bilingual dictionary but we are sure that a better and larger bilingual dictionary will definitely improve our model. The last line of Table 1 shows the results where the heuristic model is applied to the output of SuperAlign using 160K Dice. We can see that heuristic model helps to increase the recall and the balanced F-measure has been increased to 84.83, but the precision has dropped quite a lot. We will see the effects of this model in the following experiment on SMT.
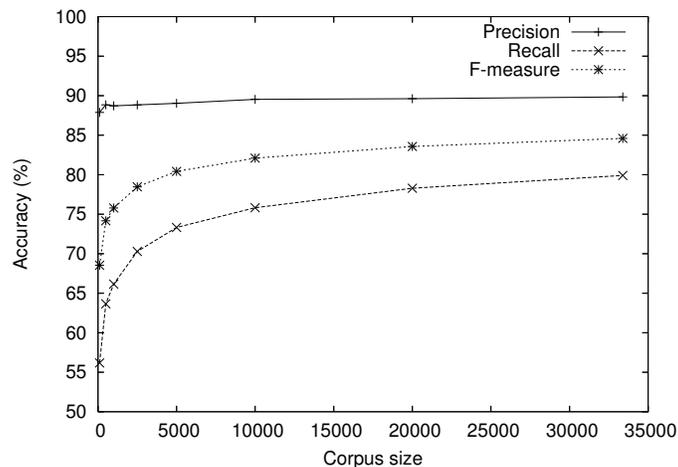

Figure 5 The accuracy of alignment versus size of training corpus

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 5 shows a graphical output of the accuracy versus the size of training corpus. All features are used, including multi-gram and context features. 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++ ($I^5H^53^34^3$) refined with the grow-diag-final-and method with SuperAlign.

Although AER does not correlate with translation quality, it is still a commonly used evaluation measurement for alignment tasks. Hence, we also calculate AER for comparisons with GIZA++. Since we do not annotate the corpus as defined for AER, we can only perform an estimation. We assume that our strong and weak links are equal to their *Sure* (S) link, and the pseudo link becomes their *Possible* (P) link. Hence, we define the equation as a measure of our AER:

$$AER = 1 - \frac{|A \bigcap S| + |A \bigcap P|}{|A| + |S|}$$

where A = system output, S = strong+weak links and P = strong+weak+pseudo links

Method	Precision (%)	Recall (%)	F-measure	AER (%)
CRF	89.84	79.91	84.59	11.83
--strong	93.03	89.28 (92.97)	91.11	-
--weak	71.49	63.90 (68.94)	67.48	-
--pseudo	69.10	47.31 (54.56)	56.17	-
CRF (5000)	89.04	73.32	80.42	15.05
CRF (1000)	88.72	66.15	75.79	18.92
GIZA++(all)	76.51	79.38	77.92	18.74
--strong	-	(94.33)	-	-
--weak	-	(69.39)	-	-
--pseudo	-	(47.60)	-	-
GIZA++(test)	62.05	67.23	64.54	32.78

Table 2 Comparison with GIZA++ alignment

Table 2 shows the results for each type of links and a comparison with GIZA++. Both SuperAlign and GIZA++ perform 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. The figures shown in brackets are recalls without concern to the types detected, since GIZA++ does not output link types. We can see that SuperAlign is better in detecting the pseudo links than GIZA++, which is the most difficult one.

In our experiment, we trained two GIZA++ models. The first model uses all 35k of the training data, including held-out and testing data. The second model uses only the testing data. The results show that the second model is much worse than the first. This also proves that GIZA++ requires a bigger training corpus in order to have a good performance.

In contrast, SuperAlign obtains AER 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 7% AER. 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 and AER 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 of 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 in-house 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.

The top of Table 3 (3a and 3b) 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 hand-aligned (HA) corpus produces less alignments points but with a larger phrase table, as shown in the row HA (swp). We also test the translation quality by excluding the pseudo links as shown in the row HA (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. In addition, by using only the strong links, HA (s), the translation performance deteriorated tremendously as the alignment links obtained are not sufficient to create the phrase table. So, we can conclude that at least strong and weak links must exist in order to produce a good translation model.

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. The middle section of Table 3 shows the experiment results. As predicted from the previous experiments, SuperAlign leads to better translation quality by approximately 2% on the BLEU score 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 three times 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 in the phrase-table creation phase[6].

Train corpus	Align method	IWSLT 2008 test data			IWSLT 2007 test data		
		bleu	meteor	(b+m)/2	bleu	meteor	(b+m)/2
BTEC hand 35K	GIZA++	0.4692	0.6016	0.5354	0.3038	0.5205	0.4121
	HA (swp)	0.4886	**0.6248**	0.5567	**0.3278**	**0.5427**	**0.4352**
	HA (sw)	**0.4995**	0.6162	**0.5578**	0.3107	0.5345	0.4226
	HA (s)	0.4442	0.5739	0.5090	0.2516	0.4887	0.3701
IWSLT 2008 20K	GIZA++	0.4042	0.5823	0.4932	0.2707	0.5063	0.3885
	SA (swp)	0.4325	**0.6049**	0.5187	0.2838	0.5187	0.4012
	SA (sw)	**0.4397**	0.6006	**0.5201**	0.2861	0.5199	0.4030
	SA (1K)	0.4199	0.5787	0.4993	0.2736	0.5086	0.3911
	SA (+h)	0.4315	0.5973	0.5144	**0.3031**	**0.5290**	**0.4160**
BTEC 160K	GIZA++	0.6083	0.6527	0.6305	0.4403	0.5942	0.5172
	SA (35K)	0.6043	0.6546	0.6294	0.4368	0.5918	0.5143
	SA (160K)	0.6077	0.6634	0.6355	0.4491	0.6002	0.5246
	SA (+h)	**0.6275**	**0.6762**	**0.6518**	**0.4681**	**0.6148**	**0.5414**

Table 3a Translation results obtained using different training corpora (HA: hand-aligned, SA: SuperAlign, +h: +heuristic, swp: strong/weak/pseudo links)

Train corpus	Align method	IWSLT 2006 test data			# of align points	Size of phrase table	Total # of non-translated
		bleu	meteor	(b+m)/2			
BTEC hand 35K	GIZA++	0.1914	0.4380	0.3147	375,353	626,502	583
	HA (swp)	**0.2101**	**0.4631**	**0.3366**	369,587	661,104	497
	HA (sw)	0.1904	0.4531	0.3217	293,848	1,339,597	479
	HA (s)	0.1544	0.4163	0.2853	200,206	3,137,429	544
IWSLT 2008 20K	GIZA++	0.1614	0.4293	0.2953	212,869	357,237	791
	SA (swp)	0.1785	0.4425	0.3105	183,535	593,841	662
	SA (sw)	0.1762	0.4399	0.3080	151,545	964,829	650
	SA (1K)	0.1456	0.4210	0.2833	153,432	957,325	786
	SA (+h)	**0.1870**	**0.4475**	**0.3215**	195,220	465,926	628
BTEC 160K	GIZA++	0.1407	0.3968	0.2687	1,457,710	1,372,172	690
	SA (35K)	0.1505	0.4059	0.2782	921,457	4,692,876	674
	SA (160K)	0.1524	0.4088	0.2806	983,241	3,855,362	637
	SA (+h)	**0.1713**	**0.4214**	**0.2963**	1,053,634	3,280,926	626

Table 3b Translation results obtained using different training corpora (HA: hand-aligned, SA: SuperAlign, +h: +heuristic, swp: strong/weak/pseudo links)

---

[6] Refer to (Koehn et al., 2003) for phrase table creation.

SuperAlign also helps in reducing the non-translated words. The last column in Table 3 shows the total number of non-translated (unknown) words from all of the testing data. In other words, some of the words that have not been aligned with GIZA++ have been successfully aligned using SuperAlign. We will look at some examples in the section analysis later. Furthermore, if we increase the recall of alignment using the heuristic model, more words can be aligned correctly and the size of the phrase table is decreased. The translation quality is further improved and the number of non-translated words decreases as well.

As explained in the previous section, if we have a huge sentence-aligned corpus, we can re-calculate the Dice measure and re-train the CRF model using the new Dice. In Table 1, we have proven that Dice from a huge corpus can improve the accuracy of alignments further. In this experiment, we use the 160K BTEC corpus to train both GIZA++ and to calculate the Dice for SuperAlign. The SMT system is also trained using the same 160K corpus. The bottom section of Table 3 shows the experiment results. If we align the 160K corpus with the 35K-trained SuperAlign model, we obtain results that are somewhat equivalent to those obtained using GIZA++, whereas, if we align the 160K corpus with the 160K-trained SuperAlign model, we obtain slightly better results than those obtained using GIZA++. However, since the SuperAlign model generates less alignment points, the phrase table generated is approximately three times larger than the GIZA++ model. By adding more alignments with the heuristic model, we can further improve the translation quality. The heuristic model also reduced the number of non-translated words. However, the phrase table is still large compare to GIZA++; the resulting decoding time will be the main concern in future analysis.

4.3 Experiment on Japanese-English Translation

A similar experiment has also been carried out on a Japanese-English (J-E) language pair. We manually aligned 9,960 sentences from BTEC corpus using the same annotation as the Chinese-English (C-E) corpus. The Japanese language has fairly different syntactic structure from English as compared to Chinese. Therefore, the alignment is expected to be more difficult. There are a total of 120,981 links in the J-E corpus where 37.49% are strong links, 29.65% are weak links and 32.86% are pseudo links. The number of pseudo links is more than the C-E language pair, which makes the alignment more difficult. The Japanese lemmas and POS tags are obtained using ChaSen[7]. The Japanese-English dictionary is a commercial dictionary that contains about 2 millions entries[8]. The experimental results on alignment, using 1000 sentences as the test data, gave a precision of 82.05%, a recall of 72.41% and an F-measure of 76.93 points[9]. Overall, the performance of alignment is not as good as with the C-E language pair.

Similarly, we also carried out an experiment on translation, using the 160K BTEC corpus as the training data, and 510 sentences as the test data. The setting is the same as the previous experiment on the C-E language pair. Table 4 shows the translation results. As we can see, although the alignment result is not as good as the C-E language pair, a similar gain (about 2%) on translation quality can still be obtained.

---

[7] http::/chasen-legacy.sourceforge.jp/
[8] This dictionary contains a lot of scientific and technical terms.
[9] The heuristic model does not work well on the alignment between Japanese-English language pair as their word order and syntactic structure are too different.

Align method	bleu	meteor	(b+m)/2	# of align points	Size of phrase table
GIZA++	0.6730	0.6935	0.6822	1,642,428	1,420,682
CRF	**0.6910**	**0.7177**	**0.7045**	1,261,928	3,643,445

Table 4 Translation results on Japanese-English

4.4 Analysis

In our experiment, we found that the SMT system using SuperAlign generates less non-translated words than GIZA++. In other words, some of the words actually exist in the training corpus but they could not be found during the decoding when using GIZA++. In the example below, the word 顺利 (all right) has not been translated using GIZA++ but has been successfully translated using SuperAlign.

- source: 不用 担心 。 一定 会 顺利 的 。
- GIZA++: do n't worry . will be #顺利 .
- SuperAlign: do n't worry . will be all right .

The translation to *all right* can be obtained from this sentence pair:

请 告诉 他们 我 已 顺利 下车 了 。
Please tell them I got off all right.

Figure 6 shows the output of the alignments performed by the two models. Symbol ⊙ represents the case in which both the models align the same words; ○, alignment by GIZA++ only; and ●, alignment by SuperAlign only.

Figure 6 Alignment outputs by GIZA++ and SuperAlign

Since Chinese is not an inflectional language, the past tense of the word *got* can be matched by the Chinese adverb 已 and the auxiliary word 了, which has been aligned correctly by GIZA++. However, the verb between them, 下车, has not been correctly

aligned by GIZA++, which is *got off*. Here, GIZA++ makes a mistake by aligning 顺利 with *got*. Hence, at the end, the translation pair 顺利 and *all right* cannot be extracted during the creation of the translation model. The possible phrases generated using SuperAlign are as follow:

顺利 ⇔ all right

顺利 下车 ⇔ got off all right

顺利 下车 了 ⇔ got off all right

已 顺利 下车 ⇔ got off all right

已 顺利 下车 了 ⇔ got off all right

whereas using GIZA++, we have only:

已 顺利 下车 了 ⇔ got off all right

As a result, the word 顺利 cannot be translated using GIZA++.

5.   Related Work

Our method is based on the concept proposed in (Blunsom and Cohn, 2006). They also trained a CRF model for inducing word alignment from sentence-aligned data. They have introduced more features than us; they have added the output of GIZA++ (models 1 and 4) as features. Moreover, due to the similarity between European languages, they have also introduced orthographic features (English-French and English-Romanian). However, their improvement on the alignment is not sufficient for improving the translation quality. In our method, the bilingual feature is not a true-false feature but a similarity measurement. Moreover, we have also proposed the use of the word lemma as a feature; it becomes useful for achieving word alignment between morphologically weak (Chinese) and strong (English) languages. Additionally, we have introduced contextual features that have helped in improving the results. As compared to (Blunsom and Cohn, 2006), our model is a fully supervised model where features from unsupervised model are not incorporated. Therefore, it is difficult to make a fair comparison to them.

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 research on discriminative word alignment models using hand-aligned training data. That research provided insights into the incorporation of more features, either lexically, syntactically, or statistically, to create a better model.

While almost all the previous studies have used the output of GIZA++ as part of the features, our model does not incorporate any features from GIZA++. 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 error links, and in phrase-based SMTs, such error links will cause problems in creating the translation table required during decoding. While our method can promise high precision, we only produce "good" align points, and it is up to the translation model creation phase to generate the possible translation phrases.

6.   Conclusion and Future Work

In this paper, we have introduced a supervised word alignment using a discriminative model, Conditional Random Fields. We treat the alignment as a sequential labeling problem and train the models to label each pair of words with a label that indicates the relations between the words in the sentence: strong, weak, pseudo, or null links. We have provided

the word pairs with useful features such as the Dice coefficient, relative position, similarities based on a bilingual dictionary, POS tags, and word lemmas. We have also defined the multi-gram features and the contextual features, that is, the words and POS tags around the current word pairs.

We trained the models using 35K sentences of Chinese-English hand-aligned corpus. Our experimental results show that SuperAlign achieved higher accuracy than an unsupervised generative model, GIZA++. SuperAlign achieved a 7% lower alignment error rate than GIZA++. SuperAlign always gives high precision no matter how small the training data is. Finally, we have also proven that the alignment output by SuperAlign improved the quality of translation in a phrase-based SMT system.

However, as compared to GIZA++, SuperAlign produced more null links. By applying the heuristic model, we have reduced some null links, but it is still not sufficient. In future research, we will try other methods to reduce the null links. Although the presence of null links does not affect the translation quality very much, they increase the size of the phrase table, thereby affecting the decoding time. Since our aligned corpus is annotated with 3 types of links, we may be able to make use of this information to create a better phrase table for translation. 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 present only on 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.

7. Acknowledgments

8. References

Blunsom, P. and Cohn, T., 2006, Discriminative word alignment with conditional random fields, In *Proceedings of COLING/ ACL*, pp.65–72.

Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., and Mercer, R.L., 1993, The mathematics of statistical machine translation: Parameter estimation, *Computational Linguistics*, vol.19, no.2, pp.263–311.

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

Fraser, A. and Marcu, D., 2007, Measuring word alignment quality for statistical machine translation, *Computational Linguistics, Squibs & Discussion*, vol.33, no.3, pp.293–303.

Granchev, K., Graça, J.V., and Taskar, B., 2008, Better alignments = better translations?, In *Proceedings of ACL*, pp.986–993.

Ittycheriah, A. and Roukos, S., 2005, A Maximum Entropy word aligner for Arabic-English Machine Translation, In *Proceedings of HLT/EMNLP*, pp.89–96.

Kikui, G., Yamamoto, S., Takezawa, T., and Sumita, E., 2006, Comparative study on corpora for speech translation, *IEEE Transaction on Audio, Speech and Language*, vol. 14(5), pp. 1674–1682.

Koehn, P., Och, F.J., and Marcu, D., 2003, Statistical phrase-based translation, In *Proceedings of HLT/NAACL*, pp.81–88.

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

Lafferty, J., McCallum, A., and Pereira, F., 2001, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In *Proceedings of ICML*, pp.282–289.

Liu, Y., Liu, Q., and Lin, S., 2005, Log-linear models for word alignment, In *Proceedings of ACL*, pp.459–466.

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

Och, F.J. and Ney, H., 2003, A systematic comparison of various statistical alignment models, *Computational Linguistics*, vol.29, no.1, pp.19–52.

Och, F.J. and Ney, H., 2004, The alignment template approach to statistical machine translation, *Computational Linguistics*, vol.30, no.4, pp.417–449.

Taskar, B., Lacoste-Julien, S., and Klein, D., 2005, A discriminative matching approach to word alignment, In *Proceedings of HLT/EMNLP*, pp.73–80.

Wu, H., Wang, H., and Liu, Z., 2006, Boosting statistical word alignment using labeled and unlabeled data, In *Proceedings of COLING/ACL Poster Sessio*n, pp.913–920.

Zhao, H., Liu, Q., Zhang, R., Lü, Y., Sumita, E., and Goh, C.L., 2009, A guideline for Chinese-English word alignment, *Journal of Chinese Information Processing*, vol.23, no.3, pp.65–87.