# Exploiting Multi-Features for Word Alignment in Patent MT

**Li Zezhong**
Department of Computer Science
Ritsumeikan University
Japan
lizezhonglaile@163.com

**Hideto Ikeda**
Department of Computer Science
Ritsumeikan University
Japan
hikeda@is.ritsumei.ac.jp

**Nguyen Thanh Hung**
Department of Computer Science
Ritsumeikan University
Japan
hungnt@is.ritsumei.ac.jp

## Abstract

This paper presents a Maximum Entropy based word alignment in patent domain which incorporates various kinds of features. We expect that it  is capable of improving the quality of word alignment by deploying various features. Experiments show that our method can get a promising performance.

***Keywords:*** word alignment; Maximum Entropy; patent; machine translation

## 1 Introduction

Word alignment was first introduced as an intermediate result of statistical machine translation systems (Brown et al., 1993), but now not just limited into this, and has been shown a very broad applications.

In recent years, researchers have proposed a number of alternative alignment methods, which can be classified into two categories, generative vs. discriminative. The generative models (Brown et al., 1993) have got a great progress for learning translation  knowledge in a unsupervised way, but recently often criticized as its drawbacks that difficult to incorporate arbitrary features and one-to-many limitation. Thus the discriminative approach may be a promising direction.

The purpose of the present work is to build a specialized word alignment in patent domain. This year, we begin to take part in the NTCIR 9, our subtask is the Chinese-English Patent Translation Task. NII released a 1M Chinese-English sentences pairs. For each sentence, not only has a sentence ID but also has a document ID, from where we can get some document level information that may be beneficial for our word alignment. The following is an example:

*Proceedings of the 4th Workshop on Patent Translation, pages 86-90,*
*MT Summit XIII, Xiamen, China, September 2011*

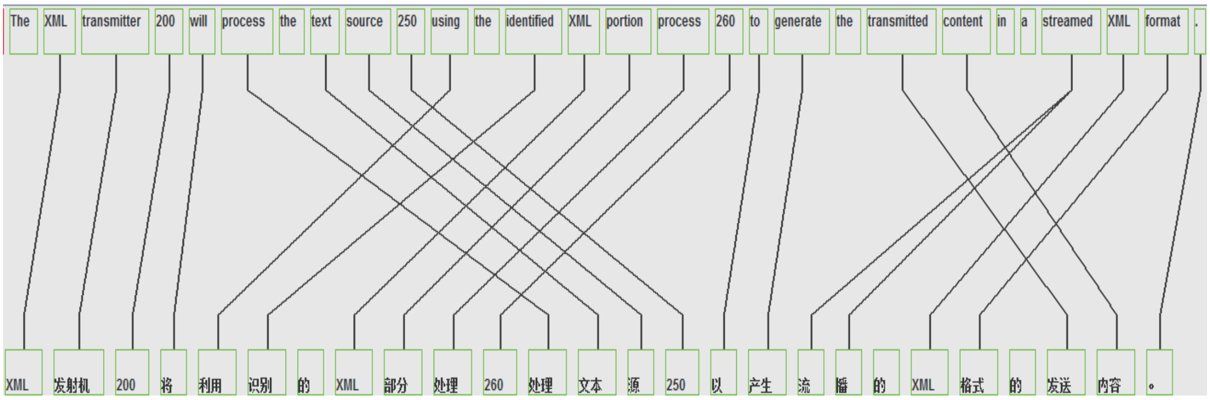

Fig. 1 word alignment for a sentence in patent domain

Fig. 1 shows an example of a Chinese-English sentence pair together with correct aligned phrase pairs. This sentence is extracted from the patent parallel corpus. At the first eye, it is easy to be observed that, the sentence in patent domain has a longer length than common sentence, and has more numbers and product names.

## 2 Maximum Entropy Model for Word Alignment

### 2.1 Maximum Entropy

Maximum Entropy is first proposed in (Jaynes, 1957), which can be expressed by the following famous formula:

$$p\ (y \mid x) = \frac{1}{Z_\lambda(x)} \exp\left[\sum_i \lambda_i f_i(x,y)\right]$$

In the above formula, $f_i$ is called feature function which denotes a kind of information that may be helpful for the final decision. In the framework of this log-linear model, arbitrary informative features can be deployed. Consider the task of word alignment, it can be viewed as a classification problem that to decide the existence of a link between a particular pair of words. To get high quality of alignment, it is essential to deploying kinds of features including lexical features, Part-of-Speech features, word frequency feature and so on, in this case, it is natural to take the Maximum Entropy model as a powerful tool for exploiting all the features and get the alignment finally.

### 2.2 Features

In this section, we present several feature functions indicting links between bilingual words.

- **Lexical features**

Lexical model estimated from a bilingual corpus using the well-known IBM models, is one of the most important information for word alignment. For the simplicity of parameter estimation, it's better to have a binaryzation of feature. The lexicalq features can be classified into 3 features by the value of translation probability: well-translated, partial-translated and bad-translated

- **Cognate features**

Motivated by the fact that proper nouns are often share the same appearance for any language, there is a need to capture the feature that whether there is a cognate or same digital and Roman character. Two features are used: (1) whether e and f share exactly the same form; (2) whether there is a partial match.

- **Neighbor features**

A neighborhood of an alignment link $l(e_i, f_j)$, which is denoted by $N(i,j)$, contains 8 possible alignment links when the window size is 3. Neighbor features can be expressed as: whether a particular neighbor exits. Thus, this kind of features includes 8 specific features to indicate whether the link exists.

- **MWE feature**

A multiword expression (MWE) is a word sequence with relatively fixed structure representing special meanings. Here we redefine the MWE as "more than one source words that share the same target word as their translation". For example, there are many multiword expressions in the patent document, such as "轮轴"(wheel axles), "子 组合"(subcombination). Another unique characteristics for Chinese is that segmentation error frequently happens especially in a professional domain such as patent. The name of an invention or a kind of technology, often as an OOV and difficult to identified by the Chinese segmenter. For example, "硅氧烷"(siloxanes) is divided into "硅 氧 烷", "阻尼器"(attenuator) is divided into " 阻尼 器". Thus the word alignment for Chinese should consider the factor of segmentation, here we also view the mistakenly divided Chinese word as a MWE. Let $c(f_{j-1}f_j)$ denotes the frequency of 2-gram $f_{j-1}f_j$ in its document, it is prone to be a multiword when the frequency larger than a given threshold $\theta$.

$$f(y,x)=\begin{cases}1 & \text{if } (l(e_i,f_{j-1})=1 \text{ and } c(f_{j-1}f_j)>\theta) \text{ or } \\ & \quad (l(e_i,f_{j+1})=1 \text{ and } c(f_jf_{j+1})>\theta) \\ 0 & \text{otherwise}\end{cases}$$

This feature can be consider as a document level feature, and unique for the patent corpus. Although there are many criteria to measure the termhood of multiword, such as C-value and NC-value (Frantzi et al., 1998), when we restrict the scope as several sentences, the frequency of n-grams is very effective and informative for searching the multiword.

- **Rare word features**

A typical word alignment error was found that the tendency of rare word to have high fertilities, which usually causes many source language tokens to correspond to a single low-frequency target.Therefore, we need to discourage the link that from a rare word to a high frequency word. The feature is devised as follows: whether the link is from a local word to a high frequency word.

Another more explicit feature controls the fertility of rare word is activated when the rare one has links from not adjacent words in the source language, and this feature directly penalizes the high fertilities for rare words.

- **Part-of-Speech feature**

We encourage the link that has a high Part-of-Speech translation probability.

- **Link count features**

The word that now has a less number of links is prone to have another link. Link features include: (1) whether neither $e_i$ nor $f_j$ has an alignment; (2) whether just one of $e_i$ and $f_j$ has no alignment.

- **Distance penalty feature**

It seems intuitive that the likelihood of a target word aligns to a distant source word is much less. The link that jump a long distance will be penalized.

## 3 Word Alignment

We use GIS algorithm to train the λ of the log-linear models. After estimation of parameters, we adopt a similar way as introduced in (Ayan and Dorr, 2005) which focus on combining several alignments at the level of alignment links rather than at the sentence level for getting the alignment. But we put an emphasis on the order of adding links, this is very important because we have used some link features. We also follow the grow-diag-final way but make some differences. Extending points first from the intersection of both directions, and then extend alignment points from the union point set. The foundation behind this approach is that the intersection has a higher precision and a lower recall, while the union has a lower precision but higher recall. Different to the Och's heuristic method, we use a discriminative way to model the decision of extending links, which fully utilize kinds of informations. The algorithm runs as follows:

Step 1: Intersection $\mathrm{A} = e2f \cap f2e$ , union $\mathrm{B} = e2f \cup f2e$ , and set $C = B - A$ , where $e2f$ denotes the alignment from English to Chinese, $f2e$ is the reverse direction.

Step 2: For each alignment point in $A$, get their neighbors $N = \{l_1, l_2 \dots l_n\}$ . For each link $l$, get its probability by using the log linear model. If $p(y(l) = 1 \mid x)$ is greater than $p(y(l) = 0 \mid x)$, delete $l$ from $C$ and add it to $A$, otherwise just delete it from $C$.

Step 3: If $C$ increases, goto Step 2.

Step 4: For the remained alignment points in $C$, compute its probability using the above formula and decide whether it should be added into $A$.

Step 5: Output $A$ as the final alignment.

## 4 Related Work

The literature contains numerous descriptions of discriminative approaches to word alignment, especially in the framework of maximum entropy, which can be classified into two categories, link modeling and alignment modeling. In our paper, we adopt the former approach, directly model a link, while some papers model the whole alignment (Liu et al., 2005). For the link modeling approach, some researchers suggest beam search algorithm for generating all the links form scratch at the level of the whole sentence (Ittycheriah and Roukos, 2005), one obvious drawback of this approach is the restraint of beam width, and the search space for alignment is very large. Another drawback is the difficulty of deploying contextual links information. Our method is most similar with the way introduced in (Ayan and Dorr, 2006) , which focuses on combining several alignments at the level of alignment links rather than at the sentence level, but we put an emphasis on the order of adding links, which is very important because we have used some link-dependent features.

## 5 Experimental Results

In this section, we present results of the word alignment for Chinese-English in domain of patent. The corpus includes 1M Chinese-English sentence pairs that released by NTCIR9. We annotated 300 word aligned sentence, and 250 for training, 50 for test. We adopt the standard AER as our evaluation metric, which is proposed by Och and Ney (2003), but here consider all the links as sure (Link includes sure link and possible link, which is depended on the degree of correspondence).

For comparison purposes, we list 4 alignments in the Table 1: (1) Intersection of both directions; (2) Union of both directions; (3) Och's grow-diag-final alignment; (4) Our alignment. As shown in the table, the intersection alignment has a higher precision, but just a lower recall, while the union alignment have a lower precision, but higher recall. The heuristic method gets a better performance than both the previous alignments. Finally, for our discriminative method, achieve a AER of 10.69%, and a relative AER reduction of 18.55% for the heuristic method, which proves the effectiveness of our method.

**Table 1: comparison of different alignments**

| type | P | R | AER |
|---|---|---|---|
| intersection | 97.35% | 72.94% | 16.60% |
| union | 78.78% | 93.06% | 14.68% |
| heuristic | 81.86% | 92.57% | 13.12% |
| discriminative | 88.30% | 90.35% | 10.69% |

## 6 Conclusions

In this paper, we present a maximum entropy model to improve the quality of word alignment. Our method is actually a hybrid method, the discriminative training is done on the basis of generative models, just requires a little annotated data, but has shown a significant improvements according to the AER.The final performance will be tested in our translation system.

## References

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.

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

Smith, Noah A. and Michael E. Jahr. 2000. Cairo: An Alignment Visualization Tool, in the proceedings of the Second International Conference on Language Resources and Evaluation(LREC 2000).

NTCIR Project: http://research.nii.ac.jp/ntcir/

E. T. Jaynes. 1957. Information Theory and Statistical Mechanics. Phys. Rev. 106(4): 620-630

Necip Fazil Ayan and Bonnie J. Dorr. 2006. A maximum entropy approach to combining word alignments. In Proceedings of the Human Language Technology Conference of the NAACL

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.

K. Frantzi, S. Ananiadou, and J. Tsujii, 1998. The C-Value/NC-Value Method of Automatic Recognition for Multi-word terms, In Proc. of the Second European Conference on Research and Advanced Technology for Digital Libraries.

Liu Yang, Liu, Qun, and Lin Shuoxun, 2005. Log-linear models for word alignment. In Proc. of ACL.