Send Orders for Reprints to reprints@benthamscience.ae

Open Access

Co-occurrence Degree Based Word Alignment in Statistical Machine Translation

Chenggang Mi[1,2], Yating Yang[1,*], Lei Wang[1] and Xiao Li[1]

[1]*Xinjiang Technical Institute of Physics and Chemistry of Chinese Academy of Sciences, Urumqi, 830011, P.R.China*

[2]*University of Chinese Academy of Sciences, Beijing, 100049, P.R. China*

Abstract: To alleviate the data sparseness problem during word alignment, we propose a word alignment method based on word co-occurrence degree. In this paper, we propose a new method to get the statistical information from word co-occurrence. We combine the co-occurrence counts and the fuzzy co-occurrence weights as word co-occurrence degree. Fuzzy co-occurrence weights can be obtained by searching for fuzzy co-occurrence word pairs and computing differences of length between current word and other words in fuzzy co-occurrence word pairs. Experiments show that the quality of word alignment and the translation performance both improved.

Keywords: Co-occurrence Degree, Statistical Machine Translation, Word Alignment.

1. INTRODUCTION

Word alignment is the beginning stage of statistical machine translation, we extract phrase pairs or translation rules according to the result of word alignment, so the translation performance depends heavily on the quality of word alignment.

Traditional word alignment models like IBM model 1-5 and HMM all depend on the number of times the source-target word pairs appeared in a training set. However, language pairs are significantly different in morphology, only the number of times the word pairs appeared in training corpora cannot reflect relationships between source word and target word, for example: In Chinese-English parallel corpus, English word “father” is co-occurrence with “爸爸” in Chinese, There are also “their fathers” co-occurrence with “他们 的 爸爸”. Although English words “father” and “fathers” are two forms of “father”, we count their number of co-occurrence times independently according to current models. In this paper, we proposed a co-occurrence degree based word alignment model, which not only considers the number of times of the word pairs co-occurrence like current models, but also the co-occurrence information about other forms of a certain word.

2. RELATED WORK

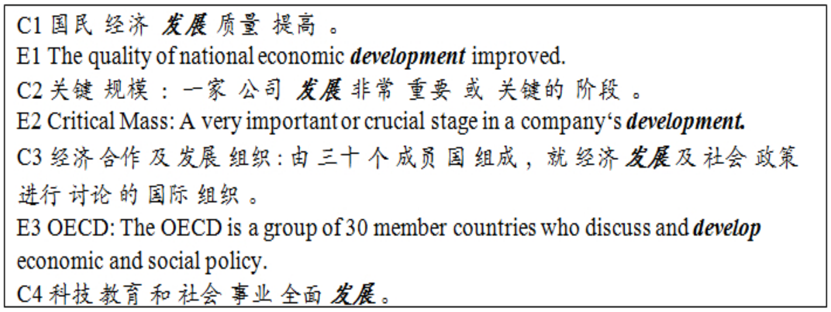
The original work of word alignment IBM-1 to IBM-5 were proposed by [1], which described a series of five statistical models of the translation process and gave algorithms for estimating the parameters of these models with a set of pairs of sentences that are translations of one another [2]. Presented the HMM-based word alignment model, which made the alignment probabilities dependent on the differences in the alignment positions rather than on the absolute positions.

Previous models mainly depended on co-occurrence counts of source and target words in parallel corpus, [3] presented a word alignment approach based on a combination of clues. Word alignment clues indicate associations between words and phrases, and they can be based on features such as frequency, part-of-speech, phrase type, and actual word form strings. Our work is similar with Tiedemann’s clue-based approach, but we are not only considering the associations of source and target word, but links between source words (which have the same word stem) are also combined into our model.

3. CO-OCCURRENCE DEGREE BASED WORD ALIGNMENT

3.1. Motivations

Sparse data is a general problem in natural language processing. In language model training, several smoothing algorithms have been proposed to relieve this problem. Also, in word alignment there exist data sparseness which affects the alignment performance [4]. Developed an optimization criterion based on a maximum-likelihood approach and described a clustering algorithm to determine bilingual word classes. Most current word alignment models based on the word co-occurrence information, which calculated by precise matching or counting the number of times word pair appeared in parallel texts [5-7]. However, some language pairs have remarkable differences in morphology or word-building, which makes the word co-occurrence counts unable to indicate the association between source words and target words


C1 国民 经济 ***发展*** 质量 提高 。

E1 The quality of national economic ***development*** improved.

C2 关键 规模 ： 一家 公司 ***发展*** 非常 重要 或 关键的 阶段 。

E2 Critical Mass: A very important or crucial stage in a company's ***development.***

C3 经济 合作 及 发展 组织：由 三十 个 成员 国 组成，就 经济 ***发展*** 及 社会 政策 进行 讨论 的 国际 组织 。

E3 OECD: The OECD is a group of 30 member countries who discuss and ***develop*** economic and social policy.

C4 科技 教育 和 社会 事业 全面 ***发展*** 。

Fig. (1). Chinese-English sentence pairs.

sufficiently. For example, in Chinese-English word alignment, some sentence pairs may appear like the following (Fig. **1**).

In these sentence pairs, source word "发展" and target word "develop" have appeared once, and source word "发展" and target word "development" appeared twice. The target word "develop" is a part of target word "development", their source words are same, but under the traditional word alignment model, co-occurrence counts of them are independent of each other. For the full use of given parallel corpus and to enhance the relationships between source word and corresponding target word, we proposed a co-occurrence degree based word alignment model, which not only considers the precise matching of source word and target word, but also the fuzzy matching of them. Accordingly, we summed the precise and fuzzy matching scores as the result of co-occurrence degree, and combine it into word alignment models.

3.2. Co-occurrence Degree

3.2.1 Definition of Co-Occurrence Degree

Co-occurrence degree of a word pair (source-target word pair) is the sum of the word pair co-occurrence counts and the score of fuzzy co-occurrence.

We get the co-occurrence counts by computing the number of times a certain source-target word pair appeared in parallel texts. However, there always exists data sparseness because of the limitation of corpora in statistical word alignment. We found that most words in our parallel texts have several forms especially mount languages, which build new words as follows:

$$\mathrm{affix}_1 + \cdots + \mathrm{affix}_k, (k>=0) \tag{1}$$

stem is the original form of one word, affix_i are affixes of the word stem, and represents different meanings. We compute the co-occurrence counts of one word pair in a traditional way; the fuzzy co-occurrence score calculated as a string similarity of the current word and words in corpus.

3.2.2. Co-occurrence Degree Calculation

Co-occurrence Counts

The well-known word alignment models IBM model 1 and IBM model 5 are based on the statistics of word co-occurrence counts. Most of the word co-occurrence information is collected during the training of IBM model 1. A word co-occurrence count is the number of times a certain word pair (source-target) appeared in training corpus. We denote co-occurrence counts of a word pair$\mathrm{WORD}_{\mathrm{src}}$ and $\mathrm{WORD}_{\mathrm{tgt}}$ $\mathrm{asco\_counts}(\mathrm{WORD}_{\mathrm{src}}, \mathrm{WORD}_{\mathrm{tgt}})$.

Fuzzy Co-occurrence Weights

We can compute the string similarity of source words (or target words), and take it as a factor of the score of fuzzy co-occurrence.

If current word pair is $< WORD_{src}, WORD_{tgt} >$, the traditional way to get the co-occurrence statics is just counting the number of times this word pair appeared in parallel corpus. According to our approach, we enhance the association between words in word pairs, and calculate the co-occurrence of other forms of current words (both source word and target word), and combine the score of fuzzy co-occurrence into word co-occurrence degree. The details of our approach are described as follows: we have two Chinese-English sentence pairs: C1-E1, and C2-E2:

C1:$c_{11}c_{12}c_{13} \ldots c_{1m}$

E1: $e_{11}e_{12}e_{13} \ldots e_{1l}$

C2:$c_{21}c_{22}c_{23} \ldots c_{2o}$

E2: $e_{21}e_{22}e_{23} \ldots e_{2n}$

Firstly, we stem English words, and keep both the word stem and affixes, and indicate an English word as

$$\mathrm{word}_e = \mathrm{stem} + \mathrm{affix}_0 + \mathrm{affix}_1 + \cdots + \mathrm{affix}_k \tag{2}$$

When we compute the score of fuzzy co-occurrence of a word pair <e , f>, we should follow three rules:

1) If we have the same Chinese word in two different Chinese sentences (**C1,C2**), we also have same English word in English sentences (**E1,E2**), we do nothing with these words;

2) If we have the same Chinese word in two different Chinese sentences (**C1,C2**), but we do not have the same English word, if the English word e is substring of English word e', then score of fuzzy co-occurrence <c, e>,

3) $\text{co\_fuzzy}(c,e) = \text{co\_fuzzy}(c,e) +,$ (3)

4) If we have the same Chinese word in two different Chinese sentences (**C1,C2**), but we do not have the same English word, if the English word e' is substring of English word e, then score of fuzzy co-occurrence<e,f>

$$\text{co\_fuzzy}(c,e) = \text{co\_fuzzy}(c,e) + (1\text{-}(\text{len}(e)\text{-len}(e'))/\text{len}(e')) \quad (4)$$

len(*str*) is the length of string *str*.

Co-occurrence Degree

The co-occurrence degree of word alignment is the summation of word co-occurrence counts and the score of fuzzy co-occurrence. We be defined as

$$\begin{aligned} &co\_\deg ree(WORD_{src}, WORD_{tgt}) \\ &= co\_counts(WORD_{src}, WORD_{tgt}) \\ &+ co\_fuzzy(WORD_{src}, WORD_{tgt}) \end{aligned} \quad (5)$$

$\text{co\_degree}(\text{WORD}_{\text{src}}, \text{WORD}_{\text{tgt}})$ denotes the co-occurrence degree of source word WORD_{src} and target word WORD_{tgt}; $\text{co\_counts}(\text{WORD}_{\text{src}}, \text{WORD}_{\text{tgt}})$ denotes the co-occurrence counts of source word WORD_{src} and target word WORD_{tgt}; $\text{co\_fuzzy}(\text{WORD}_{\text{src}}, \text{WORD}_{\text{tgt}})$ denotes the score of fuzzy co-occurrence of source word WORD_{src} and target word WORD_{tgt}.

3.3. Co-occurrence Degree Based Word Alignment Model

The IBM model 1 is the simplest of the IBM models, which does not consider word order and one-to-many/many-to-one alignments. We can use model 1 for parameter estimations that are passed on to other IBM models. In this paper, we defined the co-occurrence degree of a word pair, and evaluation the quality of word alignment and translation performance, therefore, we describe details of IBM model 1 and combine the co-occurrence degree into it.

We can indicate IBM model 1 as follow formula:

$$p(e,a \mid f) = \frac{\varepsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j \mid f_{a(j)}) \quad (6)$$

l_f and l_e are the length of source and target sentence, respectively; a is the alignment function, a:j->i denotes the source word f_i is aligned to target word e_j; t(e|f) is the translation probability of source word f and target word e. At the beginning of IBM model 1 training, the t(e|f) was first defined as

$$t(e \mid f) = \frac{co\_counts(e,f)}{counts(f)} \quad (7)$$

$\text{co\_counts}(e,f)$ denotes the number of times source word f and target word e co-occur in parallel text; $\text{counts}(f)$ denotes number of times source word f appeared in corpus.

In this paper, we use the co-occurrence degree function $\text{co\_degree}(e,f)$ to replace the co-occurrence count $\text{tionco\_counts}(e,f)$. So we define the translation probability as

$$t'(e \mid f) = \frac{co\_\deg ree(e,f)}{counts(f)} \quad (8)$$

$\text{co\_degree}(e,f)$ denotes the co-occurrence degree of source word f and target word e, which is described in section 3.3.

4. EXPERIMENTS AND ANALYSIS

4.1. Set up

We base our co-occurrence degree based word alignment approach on word alignment corpus. For the test, the result of the word alignment‘s effects on translation performance are analyzed, we also use our word alignment model to train translation models in machine translation. In word alignment experiments, we use the GIZA++[1] and Berkeley aligner[2] as our baseline alignment systems and use default parameters of these tools in our experiments. The language models we used in machine translation experiments are three-grams. We take the widely used open-source machine translation system Moses [7, 8] as our baseline system in evaluation of translation performance.

We divide our experiments into two parts: Evaluation of precision of word alignment and Evaluation of the performance of machine translation. We described the details of experiments in section 4.2 and section 4.3.

4.2. Effecting on the Precision of Word Alignment

We report the performance of our different word alignment tools in terms of precision, recall, and alignment error rate (AER). We annotated 700 sentences with labels that distinguish between sure (labeled with ‘S’) and possible (labeled with ‘P’) alignments. We used 500 sentences as tuning set, and 200 sentences as a test set. The precision, recall and alignment error rate (AER) of word alignment are defined as

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

$$precision = \frac{|A \cap P|}{|A|}$$

$$AER = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \quad (9)$$

Where, S denotes the annotated set of sure alignments, P denotes the annotated set of possible alignments, and A denotes the set of alignments produced by the model under test. We take AER, which is derived from F-measure, as our primary evaluation metric.

[1] http://code.google.com/p/giza-pp/

[2] http://code.google.com/p/berkeleyaligner/

Table 1. Evaluation of word alignment quality.

	GIZA++			Berkeley Aligner		
	BASE	**STEM**	**COD**	**BASE**	**STEM**	**COD**
Precision	77.40	78.96	**79.13**	78.31	78.62	**78.74**
Recall	72.30	**75.51**	74.32	71.03	**74.91**	73.50
AER	22.78	**21.15**	21.29	21.86	**21.43**	21.56

Table 2. Evaluation of translation performance.

	Moses + GIZA++			Moses + Berkeley Aligner		
	BASE	**STEM**	**CO Degree**	**BASE**	**STEM**	**CO Degree**
BLEU	33.10	34.12	**34.16**	32.93	33.06	**33.40**

In this section, we report the quality of different word alignment toolkits on Chinese-English parallel corpus. These corpora includes 200 sentence pairs which are annotated with sure (labeled by 'S') or possible (labeled by 'P') alignments. The results of these word alignment models are described in Table **1** (base: use default setting of word alignment tool; stemming: we stem the words in parallel texts before word aligning; CO degree: the co-occurrence degree based word alignment model).

Table **1** shows the results of different word alignment models. Because of word stemming, the word alignment models STEM achieved highest word alignment recalls and lowest AERs, that is because affixes of words has been stemmed, a lot of information was also missed. The performance of CO degree model is slightly worse than STEM. However, the CO degree models get the highest word alignment precisions (GIZA++: 79.13; Berkeley Aligner: 78.74).

4.3. Effecting on the Performance of Machine Translation

We investigate the translation performance on Chinese-English corpus. We use Moses as the baseline machine translation system, and we test our approach on the phrase-based translation model. Our training corpus contains 0.5M sentence pairs from LDC dataset. We train a 3-gram language model on the training data using SRI Language Toolkit[3]. Our tuning set contains 700 sentence pairs which are selected from LDC. We test our models on MT05 test set. The comparison of translation performance with different word alignment approaches are show in Table **2**.

Table **2** list results show the translation performance evaluation. With our co-occurrence degree based word alignment model (CO degree), we achieved highest BLEU in different translation systems (Moses + GIZA++: 34.16 and Berkeley Aligner: 33.40). The CO degree word alignment models also get the highest precisions. Which means the precision of word alignment is a key factor to the translation performance.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a co-occurrence degree based word alignment model, which combine co-occurrence counts and fuzzy co-occurrence scores and co-occurrence degree. Our method makes full use of parallel corpus, and alleviates the data sparseness during word alignment. Experiments show that with our approach, the precision of word alignment and the translation performance, both improved. In our future work, we will combine more language features into the computation of word co-occurrence degree.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA06030400), the West Light Foundation of Chinese Academy of Sciences (Grant No. LHXZ201301 XBBS201216), the Xinjiang High-Tech Industrialization Project (Grant No. 201412101) and Young Creative Sci-Tech Talents Cultivation Project of Xinjiang Uyghur Autonomous Region (Grant No. 2013731021).

REFERENCES

[1] P. E. Brown, S. A. D. Pietra, D. J. V. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation", *Comput. Linguist.*, vol. 19, pp. 263-311, 1993.

[2] S. Vogel, H. Ney., and C. Tillmann, "Hmm-based word alignment in statistical translation", In: *Proceedings of the 16th conference on Computational linguistics*, 1996, pp. 836-841.

[3] T. Jörg, "Combining clues for word alignment", In: *Proceedings of the tenth conference on European chapter of the Association for*

[3] http://www.speech.sri.com/projects/srilm/download.html

Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 339-346.

[4] F. J. Och, "An Efficient Method for Determining Bilingual Word Classes", In: *Proceedings of the European Chapter of the Association for Computational Linguistics*, 1999, pp. 71-76.

[5] F. J. Och and H. Ney, "Improved statistical alignment models", In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, pp. 440-447.

[6] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models", *Comput. Linguistic.*, vol. 29, pp:19-51, 2003.

[7] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation", In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003, pp. 48–54.

[8] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, Ri Zens, C. Dyer, O.Bojar, A. Constantin, E. Herbst.. Moses, "Open Source Toolkit for Statistical Machine Translation", Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, 2007.

Received: September 22, 2014 Revised: November 03, 2014 Accepted: November 06, 2014

© Mi *et al.*; Licensee *Bentham Open.*

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.