Co-occurrence Degree Based Word Alignment: A Case Study on Uyghur-Chinese

Chenggang Mi[1, 2], Yating Yang[1], Xi Zhou[1], Xiao Li[1], Turghun Osman[1]

[1] Xinjiang Technical Institute of Physics & Chemistry of Chinese Academy of Sciences
Urumqi, Xinjiang 830011, China
[2] University of Chinese Academy of Sciences
Beijing, 100049, China
michenggang@gmail.com,{yangyt,zhouxi, xiaoli}@ms.xjb.ac.cn,
turghunjan@sina.com

Abstract. Most widely used word alignment models are based on word co-occurrence counts in parallel corpus. However, the data sparseness during training of the word alignment model makes word co-occurrence counts of Uyghur-Chinese parallel corpus cannot indicate associations between source and target words effectively. In this paper, we propose a Uyghur-Chinese word alignment method based on word co-occurrence degree to alleviate the data sparseness problem. Our approach 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 Uyghur word and other Uyghur words in fuzzy co-occurrence word pairs. Experiment shows that with the co-occurrence degree based word alignment model, the performance of Uyghur-Chinese word alignment result is outperform the baseline word alignment model, the quality of Uyghur-Chinese machine translation also improved.

Keywords: Uyghur-Chinese word alignment; co-occurrence degree; co-occurrence count; agglutinative language; data sparseness

1 Introduction

Statistical machine translation (SMT) [1] is one of the most popular machine translation frameworks where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual parallel corpora. The statistical machine translation technologies have been extended from word-based model to phrase-based model, and nowadays, as the advent of strong stochastic parsers, syntax-based models are also built [2]. Due to lack of syntax-annotated bilingual parallel corpora, syntax-based models are mainly used in research. Hierarchical phrase-based translation [3] combines the strengths of phrase-based and syntax-based translation, which take phrases as units for translation and synchronous context-free grammars as rules. Phrase-based and hierarchical-based models are often used in recent days. The translation model training, parameters tuning and decoding are all based on the output of word alignment model. To some extent, the performance of word alignment model affects the quality of statistical machine translation.

Research on machine translation between minority languages and Chinese like Uyghur-Chinese machine translation is still at preliminary stage. Because of low-resource and word forming distinctions between Uyghur and Chinese [4], we cannot get a desired translation performance with traditional word alignment models.

In this paper, we propose a co-occurrence degree based Uyghur-Chinese word alignment method to alleviate the data sparseness problem during models training, which combines co-occurrence counts and fuzzy co-occurrence weights to train word alignment models. Experiments show that our method outperforms traditional word alignment models both in Uyghur-Chinese word alignment and Uyghur-Chinese machine translation.

2 Related Research

The original work of bilingual word alignment IBM models 1-5 were proposed by Brown et.al, which described a series of five statistical models of the translation process and gave algorithms for estimating the parameters (Expectation-Maximum algorithm (EM) [5]) of these models given a set of pairs of sentences that were translations of each other. In 1996, Vogel presented the HMM-based word alignment model [6], which made the alignment probabilities dependent on the differences in the alignment positions rather than on the absolute positions. Liu et.al proposed a log-linear model for

word alignment [7], which treats all knowledge sources as feature functions; Liang's work focused on word alignment agreement [8].

For solving data sparseness during models training, Tiedemann et.al [9] proposed a clue-based word alignment method, which added several features like string similarities between source language words and target language words into the training of word alignment models, performance of word alignment model and quality of machine translation were both improved. The method proposed by Tiedemann et.al only performed well on cognate languages, but for language pairs like Uyghur-Chinese, the improvement of word alignment performance is not significantly.

The model proposed in this paper based previous works, we extend theirs methods in a situation that source language and target language are not cognate languages (Uyghur and Chinese) and the data sparseness problem is relatively serious. In our method, we replace the traditional word co-occurrence counts with word co-occurrence degree in word alignment model; the word co-occurrence degree consists of co-occurrence counts and fuzzy co-occurrence weights.

3 Uyghur-Chinese Word Alignment

Chinese is one part of Sino-Tibetan language family, and Uyghur belongs to Altaic language family. Because of differences among language families, Chinese and Uyghur have some significant distinctions in word forming and syntactic structure. In this part, we first introduce features of Uyghur language, and then we compare with Chinese and describe problems which exist in Uyghur-Chinese word alignment.

3.1 Features of Uyghur Language

Uyghur is an agglutinative language [10], which is a type of synthetic language with morphology that primarily uses agglutination: words are formed by joining phonetically unchangeable suffix morphemes to the stem. In agglutinative languages, each suffix is a bound morpheme for one unit of meaning, instead of morphological modifications with internal changes of the root of the word, or changes in stress or tone. The syntax structure of Uyghur is S (Subject)-O (Object)-V (Verb), which is significantly different with Chinese (S-V-O). We give some examples about the Uyghur words forming and syntax structure as follows:

Uyghur word forming examples:

سومكا + م → سومكام

我的包(My bag)： *stem* *suffix0*

كىتاب + ڭىز → كىتابىڭىز

您的书(Your book)： *stem* *suffix0*

Uyghur syntax structure examples:

مەن تاماق يېمەيمەن

我不吃饭。(I don't want to eat.) **Subject Object Verb**

3.2 Data Sparseness in Uyghur-Chinese Word Alignment.

Due to rare of Uyghur-Chinese parallel corpora, there exist data sparseness problems during training of Uyghur-Chinese word alignment models.

We train word alignment models based on bilingual parallel corpora. Compared with English-Chinese and English-French, Uyghur-Chinese parallel corpora are relatively rare. Additionally, due to the word forming of Uyghur, a stem in Uyghur may derive several Uyghur words; we cannot expect a Uyghur-Chinese dictionary or Uyghur-Chinese bilingual corpora can collect every word that a certain stem can forming. Therefore, data sparseness will occur during training of Uyghur-Chinese word alignment model, which affects the performance of word alignment model, even the quality of Uyghur-Chinese machine translation.

In this paper, we try to alleviate the data sparseness problem in Uyghur-Chinese word alignment models training.

4 Co-occurrence Degree Based Uyghur-Chinese Word Alignment

Most word alignment models like IBM Model 1-5, HMM are trained based on word co-occurrence information which is obtained by counting word co-occurrence in parallel texts. When a certain Uyghur-Chinese word pair appeared in Uyghur-Chinese parallel corpora, the co-occurrence count of this word pair increased. Due to shortage of Uyghur-Chinese parallel texts, data sparseness may occur during word alignment models training. For fully use of Uyghur-Chinese parallel texts, we propose a co-occurrence degree based word alignment method to replace traditional word co-occurrence counts based methods.

4.1 Word Co-occurrence Degree

Definition of Word Co-occurrence Degree

We obtain Uyghur-Chinese word co-occurrence degree by combine word co-occurrence counts and fuzzy co-occurrence weights. The word co-occurrence counts can be gotten as a common way-number of times a certain Uyghur-Chinese word appeared in corpora; we compute fuzzy co-occurrence weights as summing up length of Uyghur words that have the same stem, meanwhile there exist same Chinese word(s) in Chinese sentences.

Computation of Co-occurrence Degree in Uyghur-Chinese Word Alignment

The co-occurrence degree can be computed as:

$$Score_{co-degree} = Score_{co-counts} + Score_{co-fuzzy} \quad (1)$$

In (1), $Score_{co-counts}$ is the count of word co-occurrence, and $\mathrm{Score}_{\mathrm{co-fuzzy}}$ is the fuzzy co-occurrence weight.

Measure the Word Co-occurrence Counts.

As the traditional way, we simply get the word co-occurrence counts by counting number of times a certain Uyghur-Chinese word pair appeared in Uyghur-Chinese parallel corpora:

$$Score_{co-counts} = n \quad (2)$$

n is the number a certain word pair appeared in Uyghur-Chinese bilingual corpora.

Measure the Fuzzy Word Co-occurrence Weights.

Uyghur words are formed by joining phonetically unchangeable suffix morphemes to a certain stem. We compute the fuzzy word co-occurrence weights of a Uyghur-Chinese word pair based on bilingual corpora. In this paper, we suppose that if a word in current Uyghur sentence has the same stem with word in another Uyghur sentence, meanwhile, there are same Chinese words in Chinese sentences which aligned to above two Uyghur sentences; we consider these two Uyghur-Chinese word pairs are reference word aligned. These kinds of alignments are measured by fuzzy co-occurrence weights, which can be obtained as following two parts.

1) **Searching for Fuzzy Aligned Word Pairs**

Suppose we have three aligned Uyghur-Chinese sentence pairs: (**SENT-UYG1**, **SENT-CHN1**) and (**SENT-UYG2**, **SENT-CHN2**), words of these sentences distribute as follows:

SENT-UYG1:$W_{SentU11}, W_{SentU12}, W_{SentU13}, \dots, W_{SentU1(k-2)}, W_{SentU1(k-1)}, W_{SentU1k}$

SENT-CHN1:$W_{SentC11}, W_{SentC12}, W_{SentC13}, \dots, W_{SentC1(l-2)}, W_{SentC1(l-1)}, W_{SentC1l}$

SENT-UYG2: $W_{SentU21}, W_{SentU22}, W_{SentU23}, \dots, W_{SentU2(h-2)}, W_{SentU2(h-1)}, W_{SentU2h}$

SENT-CHN2: $W_{SentC21}, W_{SentC22}, W_{SentC23}, \dots, W_{SentC2(n-2)}, W_{SentC2(n-1)}, W_{SentC2n}$

SENT-UYG3:$W_{SentU31}, W_{SentU32}, W_{SentU33}, \dots, W_{SentU3(p-2)}, W_{SentU3(p-1)}, W_{SentU3p}$

SENT-CHN3: $W_{SentC31}, W_{SentC32}, W_{SentC33}, \dots, W_{SentC3(q-2)}, W_{SentC3(q-1)}, W_{SentC3q}$

$W_{SentXij}$ is the *j*th word in *i*th sentence (X: U for Uyghur sentence, C for Chinese sentence).*k, l ,h, n, p, q* are length of the 1st Uyghur sentence, length of 1st Chinese sentence, length of 2nd Uyghur sentence, length of 2nd Chinese sentence, length of 3rd Uyghur sentence and length of 3rd Chinese sentence, respectively. If a Uyghur word $W_{SentU1j}$ $(0 \le \mathrm{j} \le \mathrm{k}-1)$ in SENT-UYG1 have the same stem with a Uyghur word $W_{SentU2i}$ $(0 \le \mathrm{i} \le \mathrm{h}-1)$ in SENT-UYG2 and a Uyghur word $W_{SentU3r}(0 \le \mathrm{r} \le \mathrm{p}-1)$ in SENT-UYG3, $W_{SentU2i}$ and $W_{SentU3r}$ are same words; meanwhile, there exist a same Chinese word $W_{SentC1g}$ $(0 \le \mathrm{g} \le \mathrm{l}-1)$ in SENT-CHN1, SENT-CHN2 and SENT-CHN3, the word pair $<W_{SentU1i}, W_{SentC1g}>$ can be considered as fuzzy aligned in sentence pair SENT-UYG1 and SENT-CHN1.

2) **Computation of Fuzzy Co-occurrence Weights**

According to method described in **1)**, with the help of Uyghur-Chinese dictionary, we first collect all fuzzy co-occurrence pairs for current word pair. Then, we obtain the fuzzy co-occurrence weights of current word pair as combine differences of length between current Uyghur word and other Uyghur words in fuzzy aligned word pairs which obtain from **1)**:

$$Score_{co-fuzzy} = \sum_{i=1}^{k} \frac{|L_{uyg(cur)} - |L_{uyg(i)} - L_{uyg(cur)}||}{\max\{L_{uyg(i)}, L_{uyg(cur)}\}}, \quad 1 \le i \le k \tag{3}$$

We obtain the fuzzy co-occurrence word pair <*uyg(i), chn(cur)*> according to the method described in 1). *k* is the number of fuzzy co-occurrence word pairs, $L_{uyg(i)}$is the length of Uyghur word in *i*th word pair, $L_{uyg(cur)}$ is the length of current Uyghur word.

4.2 Combine the Word Co-occurrence Degree into Word Alignment Models

IBM models are traditionally trained based on word co-occurrence counts; IBM model 1 is the first and important model to collect lexical information for following models, which can be indicated as follows:

$$p(e,a|f) = \frac{\varepsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t\left(e_j \middle| f_{a(j)}\right), 1 \le j \le l_e. \quad (4)$$

l_f and l_e are the length of source language sentence (Uyghur) and target language sentence (Chinese), respectively; *a* is the word alignment function, *a: j->i* means source word f_i is align with target word e_j; *t(e|f)* is the translation probability of source word *f* and target word *e*. When training the IBM model 1, *t(e|f)* can be computed based on word co-occurrence counts:

$$t(e|f) = \frac{Score_{co-counts(e,f)}}{Score_{counts(f)}} \quad (5)$$

$Score_{co-counts(e,f)}$ is the word co-occurrence counts of source word f and target word **e** in bilingual parallel corpora, $Score_{counts(f)}$ is the number of times source word **f** appeared in corpora.In this paper, we replace *t(e|f)* (which based on Uyghur-Chinese word co-occurrence counts) with *t'(e|f)* (which based on Uyghur-Chinese word co-occurrence degree), and *t'(e|f)* can be computed as follows:

$$t'(e|f) = \frac{Score_{co-counts(e,f)} + Score_{co-fuzzy(e,f)}}{Score_{counts(f)}} \quad (6)$$

5 Experiments

5.1 Set up

We use GIZA++[1] which implements IBM models and HMM as the baseline in word alignment experiments and evaluates word alignment results by P_r (Recall), P_p(Precision) and *AER* (Alignment Error Rate). *A* indicates a set of word alignment results and *S* is a set of sure alignments in reference alignments. In the traditional way, the

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

computation of *AER* requires gold alignments annotated as "sure" or "possible", in this paper, we don't distinguish them. Therefore, we can compute AER as:

$$P_r = \frac{|A \cap S|}{|S|}, \quad P_p = \frac{|A \cap S|}{|A|}, \quad AER = 1 - \frac{2P_r P_p}{P_r + P_p} \tag{7}$$

We extract 200 sentence pairs from CWMT 2013 Uyghur-Chinese corpora (which collected from news reports and government documents) as the word alignment validate set, and annotated alignment associations by hands. For Uyghur-Chinese machine translation experiments, we use the CWMT 2013 corpora as training set and tune set, and select 1500 sentence pairs as the test set, Table 1.

Table 1. Statistics of corpora used in Uyghur-Chinese machine translation

Corpora	Size(pair)
Training set	109,895
Tune set	700
Test set	1,500

We use Moses[2] [11] as machine translation system and SRILM[3] [12] as language modeling tool. The results of Uyghur-Chinese machine translation are evaluated by BLEU [13].

5.2 Experiments

Uyghur-Chinese Word Alignment Experiments

We use the GIZA++ as the baseline in word alignment experiments. Uyghur-Chinese sentence pairs in word alignment validate set were preprocessed by methods described in 5.3.1. Then, we search for fuzzy aligned word pairs in bilingual corpus, and obtain co-occurrence degree as introduce in 4.1. The co-occurrence counts are replaced with co-occurrence degree in GIZA++. We take the co-occurrence degree as input in training of IBM Model 1. The performance of Uyghur-Chinese word alignment is evaluated by Recall, Precision and AER, respectively.

Uyghur-Chinese Machine Translation Experiments

[2] http://www.statmt.org/moses/

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

In Uyghur-Chinese machine translation experiments, we take the results by co-occurrence counts based word alignment model and co-occurrence degree based word alignment model as inputs of model training of Moses, respectively. Parameters of machine translation tools are set as follows: the language model is 5-gram; the maximum length of phrases (rules) is 11. We evaluate the quality of Uyghur-Chinese machine translation by the script multi-bleu.perl which included in Moses.

5.3 Analysis of Results

Table 2 and Table 3 are the experiment results of Uyghur-Chinese word alignment and Uyghur-Chinese machine translation, respectively.

Table 2. Evaluation on Uyghur-Chinese word alignment. GBaseline is the baseline (GIZA++) for word alignment experiments; GStemmer is the baseline (GIZA++) with an Uyghur Stemmer; and GCo-degree is the co-occurrence degree-based word alignment model which is described in this paper.

Evaluation（%）	Word Alignment Models		
	GBaseline	GStemmer	GCo-degree
Recall (R)	86.32	**87.69(+1.37)**	87.45(+1.13)
Precision (P)	80.40	82.33(+1.93)	**82.79(+2.39)**
AER	16.75	15.07(-1.68)	**14.94(-1.81)**

Through comparing with three different word alignment methods (in Table 2), recall (R) of the stem based method (GStemmer) achieved highest improvement (1.37%), which may because the Uyghur words stemming reduce the data sparseness, to some extent; but its improvement of precision (P) (1.93%) cannot outperform co-occurrence degree based method (GCo-degree) (2.39%), one possible reason is that the stem based method missing some important information of Uyghur words. The AER of co-occurrence degree based method (GCo-degree) achieved lowest among three methods (14.94%). Notice that the decrease of AER between the stem based method (GStemmer) and co-occurrence degree based method (GCo-degree) is not very significantly ((-1.68)-(-1.81) = 0.13), which means two methods both enhancing associations between source words (Uyghur words) and target words (Chinese words) that related with each other.

Table 3. Evaluation on Uyghur-Chinese machine translation. GBaseline is the baseline (GIZA++) for word alignment experiments; GStemmer is the baseline (GIZA++) with an Uyghur Stemmer; and GCo-degree is the co-occurrence degree-based word alignment model which is described in this paper. (**PB** is short for the Phrase-Based Translation Model, and **HPB** is short for the Hierarchical Phrase-Based Translation Model)

TM(BLEU)	Word Alignment Models		
	GBaseline	GStemmer	GCo-degree
PB(test set)	38.63	39.72(+1.09)	39.75(+1.12)
HPB(test set)	39.00	40.57(+1.57)	**40.59(+1.59)**
PB (tune set)	31.69	33.40(+1.71)	33.60(+1.91)
HPB(tune set)	32.43	34.78(+2.35)	**35.00(+2.57)**

In Table 3, we compare performances of Uyghur-Chinese machine translation under different word alignment methods in test set and tune set. The stem based method (GStemmer) and the co-occurrence degree based method (GCo-degree) are both outperform the baseline (GBaseline) in Uyghur-Chinese machine translation. And the performance of co-occurrence degree based method (GCo-degree) achieved higher than stem based method (GStemmer), the most important reason is that the stem based method missing some information in Uyghur, and quality of stem based machine translation also rely on the performance of stemmers. Because of local reordering and generalization abilities, hierarchical phrase-based models outperform phrase based models. Although there are some different ideas about relationship between AER and BLEU in statistical machine translation, we validate some researchers' opinions that the BLEU of Uyghur-Chinese machine translation is related with the precision of Uyghur-Chinese word alignment: with the increase of precision of word alignment by our method, the improvement of Uyghur-Chinese machine translation performance increases correspondingly. This may because the precision of word alignment partly decide the alignment of translated words in Uyghur-Chinese bilingual corpora.

6 Conclusion and Future Work

In this paper, we propose a word co-occurrence degree based method for Uyghur-Chinese word alignment in SMT, which is different from traditional

co-occurrence counts based word alignment methods. We obtain the word co-occurrence degree by combine word co-occurrence counts and fuzzy co-occurrence weights. Experimental results show that with the method we present in this article, data sparseness in Uyghur-Chinese word alignment is alleviated effectively; comparing with stem based word alignment method, our approach maintain the integrity of Uyghur words. Performance of co-occurrence degree based word alignment model is significantly outperforming the word co-occurrence counts based method and stem based word alignment method; quality of Uyghur-Chinese machine translation also improved by our method. For future work, we will further investigate relationships between Chinese word segmentation and Uyghur-Chinese word alignment; we also plan to test our approach in other domains and on other language pairs.

Acknowledgements. This work is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA06030400), 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. Brown, P. E., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L.: The mathematics of statistical machine translation: Parameter estimation. Computational linguistics 19(2), 263-311 (1993)
2. Kenji, Y., and Kevin, K.: A syntax-based statistical translation model. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 523-530 (2001)
3. David, C.: Hierarchical Phrase-Based Translation. Computational Linguistics 33(2), 201-228 (2007)
4. Gulila, A., Mijit A.: Research on Uyghur Word Segmentation. Journal of Chinese Information Processing 18(6), 61-65 (2004)
5. Dempster, A., Laird, N. and Rubin: Maximum-likelihood from incomplete data via the EM algorithm. Journal of The Royal Statistical Society, Series B 39(1), 1-38 (1977)

6. Vogel, S., Ney, H., and Tillmann, C.: Hmm-based word alignment in statistical translation. In: Proceedings of the 16th conference on Computational linguistics, pages 836–841. Association for Computational Linguistics (1996)
7. Yang, L., Qun, L., and Shouxun, L.: Log-linear Models for Word Alignment. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, USA, pp. 459-466. Association for Computational Linguistics (June 2005)
8. Percy, L., Dan, K., and Michael, J.: Agreement-Based Learning. In: Proceedings of Advances in Neural Information Processing Systems (2008)
9. Jörg, T.: Combining clues for word alignment. In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1, Vol. 1, Stroudsburg, PA, USA, 339-346. Association for Computational Linguistics(2003)
10. Aykiz, K., Kaysar, K., Turgun, I.: Morphological Analysis of Uyghur Noun for Natural Language Information Processing. Journal of Chinese Information Processing 20(3), 43-48 (2006)
11. Philipp, K., Hieu, Hoang., Alexandra, B., Chris, C.B., Marcello, F., Nicola, B., Brooke, C., Wade, S, Christine, M., Richard, Zens., Chris, D., Ondrej, B., Alexandra, C., Evan, H.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of ACL, demonstration session, Prague, Czech Republic. Association for Computational Linguistics (2007)
12. Andreas, S.: SRILM -- an extensible language modeling toolkit. In: Proceedings of ICSLP, Vol. 2, pp. 901-904 (2002)
13. Kishore, P., Salim, R., Todd, W., and Weijing Z.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of ACL, Philadelphia, USA, 311-318 (2002)