FIT5186 Introduction to IT Research Methods

Assignment 2

A Literature Review on Word Alignment Techniques

Group 4

Ke Huishu 2819****

Guo Xuechun 2819****

12 May 2017

SOUTHEAST UNIVERSITY - MONASH UNIVERSITY

JOINT GRADUATE SCHOOL (SUZHOU)

Abstract

Word alignment has become an important technique used in Natural Language Processing. The solution of word alignment can be basically classified to three approaches: the statistical-based approaches, the lexicon-based approaches and the hybrid approaches. This paper is to review the classical models based on this classification, including IBM models, ClassAlign model and a hybrid model. In this paper, we give an introduction of the word alignment firstly. Then talking about the scope and methods we used in the literature review. Thirdly, the detail principles and algorithms of these models are introduced. After introducing the classical models, we give a brief interpretation and conclusion of the review.

Key Words: Word Alignment, Natural Language Processing, IBM model, ClassAlign model

1. Introduction

With the development of the computer techniques, the bilingual parallel corpus has played an essential role of Natural Language Processing (NLP). The collection of the parallel texts is just the first step to build parallel corpus. The parallel corpus processing is the key procedure of building parallel corpus to trigger further research or application. There are various applications of parallel corpus, including the Machine Translation (MT),

cross-lingual text retrieval, dictionary construction etc.

For a parallel corpus, word alignment refers to identifying correspondences between the words/phrases in source language and in target language (Giang, 2012). According to the alignment granularity, the alignment level could be divided to section, paragraph, sentence, phrase and word. The difficulty of the alignment is proportional to the alignment granularity. Compared with paragraph and sentence alignment, word alignment provides more fine-grained bilingual translation information, which can provide knowledge support for machine translation, lexicography and cross language information retrieval. Therefore, word alignment takes a crucial part of parallel corpus processing. That's why we want to take a review of word alignment.

In order to reach a higher recall value and precision result, many researchers adopt various alignment algorithms and models to gain better performance. This review will look into the classical models of word alignment to know more about the fundamental knowledge in this field. The objective of this review is to extract meaningful information from previous research and organize these sources to put forward our own points and propose the tendency of future work.

2. Scope and Method

In this research, we focus on the word alignment and the relevant techniques. There are basically three kinds of approaches on word alignment: the lexicon-based approaches (Ker & Chang, 2002; Wang et al., 1999), the statistical-based approaches (Brown et al., 1993; Gale & Church, 1991), and the hybrid approaches (Wu et al., 2006; Chen et al., 2012). In the review, we select three representative models based on these approaches: (a) the classical IBM model that based on statistical approaches; (b) the ClassAlign model which based on lexicon approaches; (c) a hybrid model that combines the Support Information model, IBM 5-models and Word Entropy model.

All literatures are selected from Southeast University Library and Monash University Library. Meanwhile, all the papers have strong correlation with word alignment and relevant techniques.

3. Body of the Review

3.1 IBM model

The statistical-based approaches get the probabilities of the bilingual translation words through statistical training of the large-scale bilingual corpus. These probabilities could be the basis of alignment. IBM models are the classical alignment model based on statistical approaches.

IBM models identify the word alignment through indicating the corresponding words in parallel corpus. IBM models assign a probability to each of the possible word alignments. Then the most probable one are chosen as the result of word alignment (Brown et al, 1993). Specifically, the character $\mathbf{e}$ is used to present a string of English words. Every French string $\mathbf{f}$ is a possible translation of $\mathbf{e}$. The $\Pr(\mathbf{e}|\mathbf{f})$ is the probability of the translation relation. They choose the pair of words $(\mathbf{e},\mathbf{f})$ which makes $\Pr(\mathbf{e}|\mathbf{f})$ greatest. According to Bayes' theorem, They have

$$\Pr(\mathbf{e}|\mathbf{f}) = \frac{\Pr(\mathbf{e})\Pr(\mathbf{f}|\mathbf{e})}{\Pr(\mathbf{f})}$$

Since the $\Pr(\mathbf{f})$ is independent of $\mathbf{e}$, the $\mathbf{e}$ which makes the $\Pr(\mathbf{e})\Pr(\mathbf{f}|\mathbf{e})$ biggest should be find:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \Pr(\mathbf{e}) \Pr(\mathbf{f}|\mathbf{e})$$

$\Pr(\mathbf{e})$ is the prior language model probability. There are a series of five translation models to estimate the conditional probability $\Pr(\mathbf{f}|\mathbf{e})$ which is called the likelihood of the translation $(\mathbf{f},\mathbf{e})$. IBM models assume that all the connections for each French position are equally similar. Therefore, $\Pr(\mathbf{f}|\mathbf{e})$ is not affected by the order of the words in $\mathbf{e}$ and $\mathbf{f}$. In addition, the joint probability distribution $\Pr(\mathrm{F}=\mathbf{f},\mathrm{A}=\mathbf{a},\mathrm{E}=\mathbf{e})$ is used. F and E are the French string and English string, and the A is the alignment between them. Character l and m are used to represent the lengths of $\mathbf{e}$ and $\mathbf{f}$. Then, $\Pr(\mathbf{f}|\mathbf{e})$ could be described as

$$\Pr(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$$

If $\mathbf{e} = e_1^l \equiv e_1 e_2 \cdots e_l$ contains l words, $\mathbf{f} = f_1^m \equiv f_1 f_2 \cdots f_m$ has m words. The alignment a could be described as $a_1^m \equiv a_1 a_2 \cdots a_m$. $a_j = i$ mean that the word in position j of the French string is correspond to the word in position i of the English string. Then they can have:

$$\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \Pr(m \mid \mathbf{e}) \prod_{j=1}^{m} \Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \Pr(f_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathbf{e})$$

The algorithm based on the IBM models could be divided into three steps.

Firstly, they can compute the length of source language sentence according to the length of sentence in target language. Secondly, they should determine the position in target sentence that connect the first word in the source sentence. The position is chosen based on the knowledge of the target string and the length of the source string. Then, they choose the identity of the first word in the source string based on the position in the target string. Finally, going through the source string and repeating the first three steps based on the known knowledge, including the complete knowledge of the target string and previous decided alignments.

3.2 ClassAlign model

Unlike the statistical approaches, lexicon-based approaches take lexicon information as the necessary condition of word alignment. Ker and Chang (2002) proposed a Class Align model to solve the problem of word

alignment. As a promising alternative to statistical methods, ClassAlign could broaden wide coverage of alignments. There are three algorithms leading to ClassAlign word alignment models.

The first algorithm called DictAlign which attempts to obtain reliable connections between words in source language and target language. It is based on the dictionary translations of machine-readable dictionary (MRD). They defined translation pair as (S,T), s is a word in S while t represent its in-context translation in T. DT_s denotes the translations listed in dictionary translations for the headword s. DictAlign calculates the set $W_T = \{t \mid t \text{ is a word in T}\}$ and computes the similarity between each t and the DT_s relevant to S. The similarity could be described as

$$\mathrm{DT}Sim(s,t) = \max_{d \in DTs} \frac{2 \times |d \cap t|}{|d| + |t|}$$

Specifically, DictAlign aligns each word s in S with the in-context translations t in T based on DT_s.

Step 1: Remove all stop words in S to obtain a list of keywords, W_S.

Step 2: Lookup all possible words W_T of T in a dictionary.

Step 3: For each s in W_S, look up the root of s in a bilingual dictionary to obtain DT_S.

Step 4: For all $d \in DT_s$ and all $t \in W_T$, compute $\mathrm{DTSim}(d,t)$.

Step 5: For each $(s,t) \in W_s \times W_T$, compute $\mathrm{DTSim}(s,t)$.

Step 6: For each word s, produce a connection (s,t), if $\mathrm{DTSim}(s,t)$ is

maximized over $t \in W_T$ and $\text{DTSim}(s,t) > h_1$ where h_1 is a threshold.

Step 7: Compile the list of connections and denote the list as CONN.

The second algorithm called ClassAlign is to generalize the connections into lists of class-based rules. In-context translations

The third algorithm performs the actual word alignment based on acquired rules. For each pair of sentence (S,T), first, they tag each word in S with POS information and convert each word to the root form. Then, ANS is defined and initialized to an empty list. Third, running DictAlign on (S,T) to calculate a list of initial connections, *CONN.* Fourth, they look up the dictionary to obtain candidate words. Fifth, they repeat add the connection (s,t) to *ANS* and get the final result, *ANS*.

Through the above algorithm, the source sentences and target sentences can be organized and get the result of word alignment.

3.3 The hybrid model

Chen et al. (2012) proposed a hybrid model for word alignment. This model is a combination of the previous model. In their study, they integrated the Support Information model, IBM 5-models and Word

Entropy model. IBM 5-models bring high recall value but low precision and Support Information model is exactly opposite. Meanwhile, the affected noises from other words can be efficiently reduced through Word Entropy model. Therefore, the three models are combined organically to form a new model.

In their model, they consider Chinese as source language and English as target language. Describing sentence pairing S as: $S = CS <=> ES$. CS represents Chinese sentences while ES represents English Sentence. The following is the main step of their algorithm.

First, the large scale parallel corpus should be preprocessed to ensure it is segmented and lowercased.

Second, they take advantage of Support Information model to achieve the first-time word alignment and word filtering. In this process, the Minimum Intersection model and the Minimum Difference model of Support Information model are used. For the Minimum Intersection model, they defined two pair of parallel sentences $ES_i <=> CS_i$ and $ES_j <=> CS_j$ to constitute a combination. If there are N combinations satisfying the following conditions:

$$w_e = ES_i \cap ES_j \qquad w_c = CS_i \cap CS_j$$

Then they defined $MinSup_{com}(w_e <=> w_c) = N$. For the Minimum Difference

model, if there are M combinations satisfy the conditions:

$$SET(E) = ES_i \cap ES_j \qquad SET(C) = CS_i \cap CS_j$$

$$w_e = ES_i - SET(E) \qquad w_c = CS_i - SET(C)$$

They defined $MinSup_{diff}(w_e <=> w_c) = M$. Thus, the value of N and M are support degree. From these results, they select a suitable threshold to get the first-time result. This result has high precision value but low recall value.

Third, after the second step, there may exist words that cannot find the target word. So, GIZA++ tool would be used to get candidate word for those words. However, a word usually has more than one candidate aligned words after this step. The result need further process in the following step.

Fourth, as there are more than one candidate words for one source word, they set a threshold and utilize Word Entropy to limit the probabilities of candidate words. After many experiments, the threshold is set to be 0.1 to get better performance.

4. Interpretation and conclusion

Word alignment is an important factor of Natural Language Processing and Machine Translation. It also plays a crucial role in many other applications. In this review, we select three respective models for word alignment. IBM model is the most classical model and still widely used now. ClassAlign is based on lexicon-approaches and has different principles with statistical-

based approaches. The hybrid model takes advantage of different model and attempts to get higher recall value and precision.

The traditional signal-based approaches all have their own limitations. For statistical-based approaches, words with high frequency would like to get high rate of precision. However, it cannot cover some less frequent words, thus leading to incomplete or incorrect alignments. Referring to IBM model, it could get the co-occurrence information, but it has a restriction that a source word is aligned to only one target word. However, one-to-multi alignment is usually happened in word alignment. Therefore, there are some defects and shortcomings in IBM model. For lexicon-based model, it has a strong dependency on lexicon and its calculation will largely increase with the increasing lexicon. Meanwhile, it lacks the compatibility and the model cannot be transplanted from original languages to other languages easily. In terms of hybrid model, it can take advantage of other models and have a better performance on word alignment.

Though there are some researches on word alignment techniques, the most widely used in practice is still IBM model. There still exists enough space for the development of the word alignment. It deserves deeper research and wider application in Nowadays. In conclusion, these papers are not only demonstrating the classical models in word alignment, but also teaching us

the mechanism and thought behind each model. We will extract meaningful sources from these literatures to explore new points.

References

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics, 19*(2), 263-311.

Chen, L., Xu, J., & Zhang, Y. (2012). A hybrid model for word alignment with bilingual corpus. *IEEE International Conference on Computer Science and Automation Engineering*, Vol. 3, 99-103.

Gale, W. A., & Church, K. W. (1991). A program for aligning sentences in bilingual corpora. *Meeting on Association for Computational Linguistics*, Vol. 19, 177-184.

Giang, T., & Dien, D. (2012). Improving English-Vietnames word alignment using translation model. *IEEE Rivf International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future*, 1-4.

Ker, S. J., & Chang, J. S. (2002). A class-based approach to word alignment. *Computational Linguistics, 23*(2), 313-343.

Wang, B., Liu, Q., & Zhang, X. (1999). An Automatic Chinese-English Word Alignment System. *International Conference on Multimodal Interaction.*

Wu, H., Wang, H., & Liu, Z. (2006). Boosting statistical word alignment using labeled and unlabeled data. *Coling/acl on Main Conference Poster Sessions*, Association for Computational Linguistics, 913-920.