FIT5190 Introduction to IT Research Methods

Assignment 3

A Review on Word Alignment Techniques

The Project Overview

Group 4

Ke Huishu 2819****

Guo Xuechun 2819****

29 May 2017

Southeast University - Monash University

Joint Graduate School (Suzhou)

Abstract

Word alignment has become an important technique used in Natural Language Processing. This paper is an overview of our future research. In this report, we give an introduction of the word alignment and relevant techniques firstly. Then, the objectives of our project are demonstrated. Our research aims to review the current outcomes, evaluate the models and predict the future tendency of word alignment. Thirdly, how to summarize the techniques and establish a standard are clarified. Finally, we give a brief conclusion of the review. Through this review, the readers can have a basic understanding about word alignment and its relevant techniques. Moreover, this review may provide a reference for the future works.

Keywords: Word Alignment, Natural Language Processing, Evaluation Rule, Machine Translation

1. Introduction

With the development of Natural Language Processing (NLP), there are more and more researches on corpus. As word alignment plays a crucial role in corpus processing, this has greatly stimulated the research on word alignment. Word alignment refers to identifying correspondences between the words/phases in source language and target language (Nguyen and Dinh, 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 general 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 (MT), lexicography and cross language information retrieval.

For word alignment, many approaches and models have been proposed to improve the performance and enhance the robustness between parallel sentences. Brown et al. (1993) brought up a series of five statistical models known as IBM 5-model. IBM 5-model is the most classical model of statistical approaches which depend on unknown parameters from training. However, these statistical approaches are sensible to training corpus and short of external knowledge resources such as syntactic information. There have been several methods proposed to integrate the additional information. For example, Quang-Hung and Anh-Cuong (2014) used bilingual syntactic patterns as external knowledge to narrow down the word alignment candidates. Although these statistical methods are widely used, it might still difficult to incorporate arbitrary external resources. There exist other approaches in word alignment field. The discriminative approaches were raised up that all knowledge is recognized as feature functions (Liu et al., 2010). Recently, many hybrid models and innovative models have been created. A hybrid model combining IBM model, word

entropy model and support information model was proposed to utilize advantages of each model (Chen et al., 2012). In addition, Legrand et al. (2016) proposed a neural network model to compute scores for word alignment candidates. Besides, Östling and Tiedemann (2016) developed a word alignment model using Markov Chain Monte Carlo (MCMC). There are also various evaluation rules such as Alignment Error Rate (AER), F-measure, Consistent phrase error rate (CPER) and Error-Sensitive Alignment Error Rate (ESAER) (Och, 2003; Fraser and Marcu, 2007; Ayan and Dorr, 2006; Huang et al., 2008).

As there are a series of researches on word alignment, a review should be conducted to summarize the current outcomes. Thus, our research aims to summary the models, establish evaluation rules and predict future work of word alignment.

2. Objectives

Through searching CNKI, ACM digital library and Monash library, we found there is no review about word alignment. Thus, our research aims to summary the classical and current model of word alignment. Meanwhile, we hope we can establish a uniform standard to analyze and evaluate the quality of word alignment. Through analysis and evaluation of various models, the advantages and disadvantages of each model can be found and the future tendency can be predicted. Therefore, we hope our review can give others a basic understanding of word alignments and relevant techniques, providing some experiences and research direction for the following studies at the same time.

3. Methodology

To achieve our objects, the process and some methods of our research are as follows.

First, we must search a large number of libraries to find the outcomes of word alignment. After searching, valuable research results should be selected and the other researches can be filtered out. Meanwhile, the different word alignment models are summarized from the aspects of time and basic principles. Then, we construct the framework of this review: summarizing the outcomes, constructing a uniform standard, evaluating different models and predicting future tendency.

Second, the three error types are defined as missing alignment connection, redundant alignment connection and wrong alignment connection (Huang et al., 2009). There also exist three difficulties in word alignment, namely incompatibility, various types and cross alignment. Thus, a series of variables are defined based on three difficulties and error types. We use S to represent source sentence and T to represent target sentence. s_i and t_j denote the i word in S and the j word in T. m and n

are the length of the S and T. $< s_i, t_j >$, is called an alignment link, represents a set of word correspondence in source sentence and target sentence. $d(p,q)$ is a distance function and denotes the distance between the p word and the q word. Meanwhile, α, β, γ denote the basic punishment factors.

Third, to evaluate result of different models, the alignment error results are conducted. For different error types, the error of each alignment link is calculated through distance and basic punishment factors. For example, if there is a wrong alignment connection in $< s_i, t_j >$, the error rate is calculated: $er(s_i, t_j) = \gamma \times [d(i, j)]^2 / (m+n)^2$. The overall sentence error rate is calculated as: $SER(s,t) = \sum_{0<i\leq m, 0<j\leq n} er(s_i, t_j)$. After a series of calculation, the sentence error rate can be qualified.

Fourth is the process of selecting corpus. As there are three challenges that must be solved, the test set chosen for different models is also very crucial. Concerning the three problems, the different weights should be set to distinguish. Meanwhile, the test set should be considered carefully. We should select three test sets at least, the three test sets have emphasis on incompatibility, various types and cross alignment. If an appropriate corpus cannot be found, we may have to construct a suitable test set by ourselves.

Fifth, according to the above step, the different models are measured. In each test set, we utilize each model to get the result of alignment links. Then, the results are compared with right links to calculate the overall sentence error rate and variance of sentence error rate.

Through the above five step, the quality of each word alignment model can be measured. After the measurement, the advantages and disadvantages of different models also can be analyzed according to the result. Moreover, these models can be classified based on the result. The models that only base a single approach are called single-based word alignment model. On the contrary, the models that utilize more than one approach are defined as multi-based word alignment. Based on the analysis, it is possible for us to find whether the single-based models are better or the multi-based models are better. If the single-based models are better, we should find the best one and discussed its performance on the three difficulties. If the multi-based models are better, how to combine those single approaches should also be predicted through the results.

Therefore, a uniform evaluation rule is conducted and various word alignment models can be measured by the rule. However, there are also some limitations of our method. For the process of calculating sentence error rate, it may not perform best to measure the quality of word alignment models. The detail of the algorithm needs further research. In addition, the weights of the three difficulties are not discussed now. The definition of weights may need experts or more experiments.

4. Novelty

The novelty of our research is that we not only decide to summarize the outcomes of word alignment but also establish an evaluation rule for different models. Instead of measuring different models based on previous researches, we hope to define a uniform standard to find the advantages and disadvantages of each model. Furthermore, the future tendency can also be predicted according to the analysis and compare.

5. Conclusion and Significance

In summary, our research mainly focuses on reviewing the outcomes of word alignment and evaluating the different word alignment models. We firstly summarize the classical and recent models of word alignment. Then, we establish our own uniform standard to measure and analyze these models. Utilizing the standard, different models will be analyzed and their limitations are also be clarified. Finally, we take advantage of these analyses to predict the future tendency of word alignment. Hence, the readers can have a basic understanding about word alignment and its relevant techniques. Moreover, this review may provide a reference for the future researches.

References

Ayan, N. F., & Dorr, B. J. (2006). Going beyond AER: an extensive analysis of word alignments and their impact on MT. *International Conference on Computational Linguistics and the, Meeting of the Association for Computational Linguistics*, 9-16.

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, 3*, 99-103.

Fraser, A., & Marcu, D. (2007). Measuring Word Alignment Quality for Statistical Machine Translation. *Computational Linguistics, 33*(3), 293-303.

Huang, S., Ning, X. I., Zhao, Y., Dai, X., & Chen, J. (2009). An error-sensitive evaluation metric for word alignment. *Journal of Chinese Information*

Processing.

Legrand, J., Auli, M., & Collobert, R. (2016). Neural network-based word alignment through score aggregation. 66-73.

Liu, Y., Liu, Q., & Lin, S. (2010). Discriminative word alignment by linear modeling. *Computational Linguistics,36*(3), 303-339.

Nguyen, G. T., & Dinh, D. (2012). Improving English-Vietnamese Word Alignment Using Translation Model. *IEEE Rivf International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future*, 1-4.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. *Meeting on Association for Computational Linguistics, 32*, 160-167.

Östling, R., & Tiedemann, J. (2016). Efficient word alignment with markov chain monte carlo. *Prague Bulletin of Mathematical Linguistics,106*(1), 125-146.

Quang-Hung, L. E., & Anh-Cuong, L. E. (2014). *Syntactic Pattern Based Word Alignment for Statistical Machine Translation*. IGI Global.