Comparative Study on "Sentence Rewriting for Semantic Parsing" and "Graph-Based Translation Via Graph Segmentation"

Tian, Ying

Published in:
Journal of Physics: Conference Series

Published: 01/01/2021

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Published version (DOI):
10.1088/1742-6596/1852/4/042021

Publication details:
Tian, Y. (2021). Comparative Study on "Sentence Rewriting for Semantic Parsing" and "Graph-Based Translation Via Graph Segmentation". *Journal of Physics: Conference Series*, *1852*(4), [042021]. https://doi.org/10.1088/1742-6596/1852/4/042021

Comparative Study on “Sentence Rewriting for Semantic Parsing” and “Graph-Based Translation Via Graph Segmentation”

Comparative Study on "Sentence Rewriting for Semantic Parsing" and "Graph-Based Translation Via Graph Segmentation"

Ying Tian*

City University of Hong Kong, Hong Kong, 999077, China

*Corresponding author e-mail: zora_tianying@cityu.edu.hk

Abstract. Machine translation is universally discussed by people worldwide nowadays along with the developments of computer science and artificial intelligence, however, sentence parsing and the translation errors between original language to target language are still difficult problems during the translation. This essay illustrates the main points of the two articles both of which concentrate on the methods to enhance the precision rate and accuracy of semantic parsing and translation, and makes a comparison between them. Even though these articles can increase the accuracy of translation, some of the backwards still exist and they are proposed to make the articles improved with their objectivity and precision.

Keywords: Comparative Study, Machine Translation, Semantic Parsing and Translation

1. Introduction

With the development of economic globalization and political multi-polarization, translation with high equality and accuracy is necessarily required for many international situations such as business conferences, academic forums, political negotiations, etc. where most of conferees are from different language backgrounds. The rapid improvement of technology and computer science have also make it possible and reliable for high precision machine translation. However, taking the various grammatical rules and syntactic structures into consideration, to make machine translation system fully understand the original language, semantic parse it and recombine the translated phrases to the target language are the most challenging parts for machine translation. Even though the phrase-based translation model, which changed the form of translation from translating word by word to phrase by phrase, proposed by Koehn et al. in 2003 [1] considered to be a milestone in the field of machine translation, it still could be improved to translate more accurately and precisely. Semantic parsing is regarded as a structural method and crucial stage of understanding the original language, however, tend to mismatch the vocabularies and phrases. Considered its importance, the method of improving its accuracy and precision is significantly meaningful.

The authors of the two articles proposed two different methods to enhance the accuracy of sentence parsing and translation with sentence rewriting and graph-based segmentation respectively. "Sentence Rewriting for Semantic Parsing" written by Chen, Sun, Han and An in 2016 proposed the method to

increase the accuracy of sentence parsing through rewrite the sentences in original languages and applied them into translation which proved to effectively improve the quality of machine translation by being compared with other methods [2]. Li, Way and Liu introduced another approach for translation called graph-based segmentation based on the former phrase-based method in "Graph-Based Translation Via Graph Segmentation" [3]. They also tested their method with its application in Chinese-English and German-English translations and then compared the accuracy rates of it with that of other translation methods, which proved with the significant improvement of their graph-based translation method.

2. Summaries of the Two Articles

2.1. Sentence Rewriting Method

Chen, Sun, Han and An illustrated the reasons for their creation of the sentence rewriting system that they had realized both the importance and difficulty of semantic parsing in their article "Sentence Rewriting for Semantic Parsing" [2]. The authors also explained the challenging part of semantic parsing which was cause by "the variety of natural language expressions", i.e. various sentences could be used to express the same meaning, which not only increased the degree of difficulty for decoding the different sentences with similar meaning, also decreased the correctness of sentence parsing and translation. Therefore, Chen and his teammates created two models to rewrite sentences, one was based on dictionary and the other one was based on template. In the dictionary-based for 1-N mismatch, they made fully use of the explanations of various words, especially nouns in the dictionary to reduce the mismatch translation of the words, for instance, they rewrote "*daughter*" as "*female child*" with the semantic structure "*child^female*". Template-based for N-1 mismatch was an algorithm for rewriting the complicated sentence into its simpler form following the grammatical structure of target language. It could be divided into three steps: firstly, the words to indicate specific places or time needed to be replaced by the symbol "*$y*"; secondly, the most appropriate candidate (with the highest probability) of the rewriting sentences would be chosen; and finally, the substituted word took place of "*$y*"and combined with the chosen sentence in second step. Taking the sentence "how many people live in Berlin" as example, the sentence firstly was substituted as "how many people live in *$y*" and in second step, the most appropriate candidate "what is the population of *$y*" was chosen and finally, the whole sentence was rewritten as "what is the population of Berlin" with the semantic structure "*population (Berlin)*". These two sentence-rewriting approaches were applied into semantic parsing and translation, and then their model was designed to learn from 65% questions of WEBQUESTIONS dataset [4] and then was tested with the other 35% questions. The model was also measured with another 50 mismatch problems to prove that it was effective to solve mismatch problems of vocabulary. The results of these tests were also compared with the same tests on another five translation systems such as QA based system created by Bast and Haussmann [5].

With the tests and comparisons, the authors found that both of the two sentence-rewriting models they proposed could effectively reduce the vocabulary mismatch problems, which leaded to a significantly increase in the accuracy and precision of machine translation. But when put the two models into comparison, they concluded that the N-1 model (template-based model) is more effective than 1-N model (dictionary-based model), which may be caused by the fact that "N-1 mismatch problems are more frequently appears in the WEBQUESTIONS dataset" [2]. As for robustness of the semantic parsing, they concluded that template-based approach can attach specific information to the original sentence, which declined the inaccuracy of sentence parsing. Chen and his teammates also claimed that their rewriting patterns was effective to solve out of vocabulary (OOV) problems. The authors also carry error analysis into execution, and they summarized four reasons that could cause the error that "the label issue, n-ray predicate problem, temporal clause, and superlative case" [2]. Despite the errors might be made during the tests, the two sentence-rewriting approaches for semantic parsing and translation is significantly useful for enhancing the accuracy and precision of translation.

2.2. Graph Segmentation Method

Li, Way, and Liu specifically illustrated their graph-based segmentation method which is created based on phrase-based translation model in their article "Graph-Based Translation Via Graph Segmentation" in 2016. Li and his colleges generated the shortcomings of those existed translated systems such as phrase-based translation system which was lack of accuracy of translating discontinuous phrases, for example, "*ne…pas*" in French as "*not*" in English, and syntax-based translation approach which was restrained by hierarchical associations of the language use and grammatical structure, and etc. Therefore, Li, Way and Liu proposed a relatively new translation pattern — graph-based translation method by which the sentence in original language was decoded into various subgraphs could be divided into three main steps. First of all, the sentence of natural language was parsed into several subgraphs with their own dependency relations and grammatical relations also known as bigram relations were used to connect those subgraphs with each other according to semantic parsing rules. Secondly, those subgraphs were translated into phrases or strings in target language respectively by machine from left to right. Finally, the translated phrases and strings were combined to form the new sentence in the target language and the translation of the sentence was completed. The graph-based translation model was tested by the authors with its application in Chinese-English translation and German-English translation. Moreover, the same tests were executed with another five similar translation models, such as PBMT, Treelet, DTU, and etc. and all of the five translation patterns were evaluated by BLEU, METEOR, and TER algorithms which proved that the graph-based translation model was significantly effective for improving the accuracy and precision of machine translation, especially in terms of discontinuous phrases. According to the authors, in the graph-based translation pattern, the vocabularies were divided into fifty categories by *mkcls* which was learned from SMT [6] to generalize the lexicons more specific and accurate [3].

The results of the tests and comparisons from the three metrics indicated that GBMT (graph-based machine translation) could translate sentences much more accurately and precisely than PBMT (phrase-based machine translation). GBMT model performed especially well in handling long-distance relations i.e. graph-based machine translation pattern could decode discontinuous phrases more correctly than the other five models. Moreover, the authors also found that larger translation units were tending to be used in Chinese-English translation than German-English translation which also leaded to the higher accuracy and precision of German-English translation than Chinese-English translation. Three examples of the five various translation methods were listed by the authors to further illustrate the effectiveness of GBMT model in the article. The significant increase of translation accuracy and precision could be well completed by graph-based translation model proposed by Li and his colleges.

3. Comparisons between the Two Methods

3.1. Similarities

Even though the two articles were about different approaches for machine translation, they share some common effects essentially.

Firstly, both of the articles and their authors have recognized the backwards and disadvantages of the existed machine translation models. In "Sentence Rewriting for Semantic Parsing", Chen, Sun, Han, and An claimed that "the vocabulary mismatch problem between natural language and target ontology" was one of the main difficulties for sentence parsing and translation [2]. They also figured out that many translation models had upsetting accuracy on sentence parsing of the original language, which further influenced the precision of translation to target language, because of the low quality of dealing with vocabulary mismatch issues. The similar problem had been illustrated in "Graph-Based Translation Via Graph Segmentation" by Li, Way, and Liu, in which they maintained that numerous statistical machine translation models had problems in identifying discontinuous phrases in sentence parsing and segmentation, which leaded to misunderstanding of the original language and false translation and information attached to the target language [3]. The authors of these two articles paid

attention to those shortcomings of the translation models and proposed their models respectively, trying to improve the accuracy of sentence parsing and segmentation of the natural language and enhance the quality of machine translation.

Secondly, the purpose of the two articles are the same, i.e. both of the articles focused on the methods to enhance the accuracy and precision of machine translation, and the various approaches proposed in the two articles preformed well and effectively with high increases in correctness of the translation. Both of the new models created by the authors haven been tested several times and a large amount of comparisons were made between the new models and some of other famous and representative translation models by the authors respectively. According to Chen and his teammates, sentence-rewriting model was compared with "semantic parsing based systems [4][7][8], information extraction based systems [9], machine translation based systems [10], etc." with the results turned out that to rewrite the sentence in the original language is quite useful for improving the accuracy and correctness of the machine translation to translate into target language. Graph-based translation was also proved to be significantly useful with the comparison with "PBMT, Treelet, DTU, and etc." under the executions of BLEU, METEOR, and TER, and resulted "up to +1.5/+0.5 BLEU, +0.3/+0.2 METEOR, and -0.8/-0.4 TER improvements on Chinese-English translation and German-English translation respectively" according to Li, Way, and Liu [3]. Both of the new models for sentence parsing and translation are proved to be effective according to the tests.

3.2. Differences

Despite the similarities of the two new translation models, it does exist some differences between them inevitably.

To begin with, the start points of the two approaches are various. Chen, Sun, Han, and An started from sentence parsing while on the other hand, Li, Way, and Liu from the point of sentence segmentation. In the article "Sentence Rewriting for Semantic Parsing", the dictionary-based and template-based methods mainly took the grammatical functions of various words and phrases into consideration, by which the rewritten sentences could be more similar to the grammatical structure of the target language. By contrast, graph-based method, by which the sentence in original language could be divided into several graphs with bigram relations among them, in the article "Graph-Based Translation Via Graph Segmentation" focused mainly on the correctness of sentence segmentation, especially the segmentation on discontinuous phrases.

Secondly, the two models were established based on different former theories despite the fact that they were both related to machine translation. Sentence rewriting pattern was established on the basis of knowledge and sentence parsing. The rewriting of the sentences was based on the grammatical rules and syntactic structures with which the structure of the sentence tended to be more similar to the target language and the difficulty of decoding the original sentence was reduced and the correctness of decoding increased at the same time. Unlike sentence rewriting pattern, graph-based model was supported by phrase-based translation model [1], with which the bigram relations of the lexicons and phrases in the sentence of original language became more specific and clearer for machine to translate. As a result, the accuracy and precision of the translation were also improved.

Moreover, the executive patterns of the two models and the experiments designed for testing the significance of improvements of the two methods are various. The sentence-rewriting model needed to be trained before being tested, i.e. to solve the vocabulary mismatch problems with sentence rewriting method, many of the mismatch cases needed to be used to help the model "learn" from. Whereas, graph-based model could be used directly to execute the sentence and transformed it into several graphs. Owning to the variations of the models' execution, the measures to evaluate their effectiveness were designed differently, one needed to be trained while the other was not. Furthermore, graph-based translation was also designed to make a comparison between translation in different languages, with Chinese-English and German-English translating respectively.

4. Conclusion

To draw a conclusion, both of the sentence rewriting models and the graph-based translation model are established to effectively improve the correctness and precision of machine translation. Even though the two models achieve the results through different algorithms, they both focus the sentence parsing and segmentation of the original language. Both models have paid special attention to the phrases parsing, especially the discontinuous ones, and both of them are proved to be significantly effective in enhancing the machine translation accuracy. It would be better for machine translation to combine the two models, i.e. rewrite the sentences before they were separated into various subgraphs, without any doubt.

References

[1] Koehn, P., & Och, F. J., & Marcu. D. (2003). Statistical Phrase-Based translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, July, 48-54.

[2] Chen, Bo., & Sun, Le., & Han, Xianpei., & An, Bo. (2016). Sentence Rewriting for Semantic Parsing. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 7-12, Aug., 97-107.

[3] Li, Liangyou., & Way, Andy., & Liu, Qun. (2016). Graph-Based Translation Via Graph Segmentation. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 7-12, Aug., 97-107.

[4] Berant, J., & Chou, A., & Frostig, R. & Liang, P.. (2013). Semantic parsing on Freebase from question-answer pairs. Proceeding of the 2013 Conference on Empirical Methods in Natural Language Processing, Oct., 1533-1544.

[5] Bast, H., & Haussmann, E.. (2015). More accurate question answering on freebase. Proceedings of the 24th ACM International on Conference on information and Knowledge Maagement, 19-20, Oct., 1431-1440.

[6] Cherry, C.. (2013). Improved Reordering for Phrase-Based Translation using Sparse Features. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June, 22-31.

[7] Berant, J., & Liang, P.. (2014). Semantic parsing via paraphrasing. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1, June, 1415-1425.

[8] Berant, J., & Liang, P.. (2015). Imitation learning of agenda-based semantic parsers. Translations of the Association for Computational Linguistics, 3, 545-558.

[9] Yao, Xuchen, & Durme, B.V.. (2014). Information extraction over structured data: Question answering with freebase. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1, June, 956-966.

[10] Bao, Junwei., & Duan, Nan., & Zhou, Ming., & Zhao, Tiejun. (2014). Knowledge-based question answering as machine translation. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1, June, 967-976.