The Review of Discourse-Based Machine Translation Evaluation

Yunzhen Zhang [1], En Guo [2], Hui Yu [1, a)], Weizhi Xu [1]

[1] *School of Management Science and Engineering, Shandong Normal University, Jinan, China.*
[2] *Shandong HI team Software Engineering College, Jinan, China.*

[a)] huiyu0117@163.com

Abstract. This article reviews the discourse-based machine translation evaluation metrics. According to different methods of using, machine translation evaluation based on discourse is divided into two categories: evaluation based on discourse structure and evaluation based on discourse features. For these two different categories, we introduce their representative works, and compare the advantages and disadvantages of each method. Finally, we summarize the evaluation metrics based on discourse, points out the problems to be paid attention to when evaluating, and predicts the future development trends.

Key words: Machine translation, Evaluation metrics, Discourse structure, Discourse features.

INTRODUCTION

Machine translation (MT) has benefited greatly from the development of automatic evaluation in the past decade [1]. To a certain extent, its progress is limited by the evaluation indicators being used. To date, most efforts to assess the quality of MT output have remained focused on the sentence level without paying enough attention to the consistency of sentences at the document level. This is reflected in the main MT evaluation metrics, such as BLEU [2], METEOR [3] and TER [4], which use the sentence-by-sentence method to score MT output. The evaluation result of any document is usually a simple average of its sentence score. The disadvantage of this sentence-based evaluation is the neglect of the entire discourse structure.

The accuracy of document-level MT output is particularly important for MT users because they are concerned with the overall meaning of the text rather than the grammatical correctness of each sentence [5]. First, if a text conveys its purpose and meaning of communication to the reader, the text is coherent. Second, the text needs to exist as a whole, rather than a series of independent sentences. The next part mainly introduces our classification. According to different methods of using, machine translation evaluation based on discourse is divided into two categories: (1) MT evaluation based on discourse structure. (2) MT evaluation based on discourse features. Then in the third part, we summarized the current situation and looked forward to the future research direction of machine translation evaluation.

MT EVALUATION BASED ON DISCOURSE STRUCTURE

MT Evaluation Based on Logical Semantic Structure

Guzman and Joty (2014) [6] believe that discourse information should be complementary to existing evaluation methods and should not be ignored, and that the discourse structure can be used to improve automated MT assessment. First, they defined two simple measure of similarity of discourse awareness. They use the all-subtree kernel to calculate the similarity between the parse trees according to Rhetorical Structure Theory (RST) [7]. Then, after extensive experiments on WMT12 and WMT11 data, it was found that all existing evaluation metrics can benefit from discourse-based indicators, whether at segment or system level.

Comelles et al. (2010) [8] proposed MT assessment metrics based on discourse representation theory, which takes into account the characteristics of common cause relations and discourse relations to assess the quality of MT output. Unfortunately, their metrics have no higher correlation with human quality judgement than standard sentence-level MT assessment metrics. However, one might think that the metric assessment of the shared task itself is biased because the evaluation of the human rating of the document level is approximated by averaging the average of human judgments on the interpretation of the chapter.

MT Evaluation Based on Co-Reference Structure

Hardmeier and Federico (2010) [9] studied the problem of pronoun anaphora translation by manually evaluating German-English SMT output. Then, SMT is provided with a word-dependency model, which can represent links between pairs of words in the same or different sentences. We use this model to integrate the output of a co-reference resolution system into the English-German SMT to improve the translation of anaphora pronouns.

Lida and Tokunaga (2012) [10] proposed a metric for assessing discourse coherence based on the output of a co-reference resolution model to reflect the author's opinion in using connective or associative relationships when writing coherent texts. In order to study the impact of the proposed metrics, they performed an empirical evaluation of paired-ranking tasks, using the NAIST text corpora as the target data set. The evaluation results show that the metric calculated using the output of the NP co-reference resolution obtains better accuracy than the entity grid model. In addition, the combination of metrics and entity grid models shows that the accuracy has increased by 7 points.

MT EVALUATION BASED ON DISCOURSE STRUCTURE

MT Evaluation Based on Cohesion

Billy and Kit (2012) [11] tried to solve the problem that most of the existing MT assessment indicators ignore the sentence connectivity in the document. The high correlation between cohesion use and the adequacy of translation also shows that the more lexical cohesive means are used, the better the quality of machine translation output [12]. Therefore, they used two ratios to capture this correlation. The experimental results confirm the effectiveness of this function in calculating the document-level quality of the MT output. The performance of the two-evaluation metrics, BLEU and TER, has been greatly improved in their correlation with human assessment by combining this document-level feature.

Xiong and Ben (2013) [13] proposed three different models to incorporate the three types of lexical cohesive devices, namely repeat words, synonyms/near synonyms and super-coordinate devices into SMT. These three models are the first attempts to successfully integrate lexical cohesion into document-level machine translation and achieve substantial improvements in the baseline. They integrated these three models into a hierarchical phrase based SMT system and conducted a series of experiments to verify their effectiveness. The experimental results show that all three models can basically improve the translation quality of the BLEU.

MT Evaluation Based on Coherence

Karamanis et al. (2004) [14] and Miltsakaki and Kukich (2000) [15] proposed a coherence assessment method that uses text-center transitions directly as the central theory does. Karamanis et al. defined an indicator based on the number of missing back-sight centers [16], where each center is a discourse entity that appears in the current discourse and is considered to be the most prominent discourse entity of the previous discourse. On the other hand, Miltsakaki focused on the relationship between the coherence of the text and the transitions of the center and revealed that the rough transitions of the center are related to the incoherence of the text.

Barzilay and Lapata (2005; 2008) [17-18] proposed an entity-based model to represent and evaluate local discourse coherence. The model is inspired by the central theory, which states that subsequent sentences in partially coherent text may continue to focus on the same entity as the previous sentence. Barzilay realized discourse entity transition in discourse theory by creating an entity grid model and demonstrated that their models can recognize coherence texts. Barzilay and Lee proposed a domain-dependent HMM model to capture topic transitions in the text, where topics are represented by hidden states and sentences are observed. The global coherence of the text can be summarized from the overall probability of the first sentence to the last sentence.

CONCLUSION

In recent years, discourse-based machine translation evaluation has received a lot of attention. According to the differences of the using discourse, we divide it into MT evaluation based on the discourse structure and MT evaluation based on the discourse feature. From the above introduction we can see that each method has its own advantages and disadvantages, and several directions for subsequent research: (1) We can further improve the shortcomings of the various methods described above to find better discourse-based machine translation evaluation methods. (2) We can conduct research on other aspects of the discourse. We can study other structures in the text structure, such as topic structure, function structure, event structure, etc. to create new machine translation evaluation method based on discourse structure.

ACKNOWLEDGMENTS

The work is supported by NSFC project (No. 61602284, No. 61602282 and No. 61602285), Shandong Natural Science Foundation, China (No. ZR2015FQ009 and No. ZR2016FP07), and Primary Research & Development Plan of Shandong Province (2017GGX10112).

REFERENCES

1. Deborah Coughlin. 2003. Correlating automated and human assessments of machine translation quality. In Proceedings of the Machine Translation Summit X, MT Summit'03, pages 23–27, New Orleans, LA, USA.
2. K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, 2002: 311-318.
3. Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan, June 2005: 65-72.
4. Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In Proceedings of association for machine translation in the Americas, 2006: 223-231.
5. Higgins, D., Burstein, J., Marcu, D., and Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL2004), pages 185-192.
6. Francisco Guzman, Shafiq Joty, Lluıs Marquez and Preslav Nakov. Using Discourse Structure Improves Machine Translation Evaluation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 687-698.
7. Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. Text, 8(3):243-281.
8. Elisabet Comelles, Jesus Gimenez, Lluıs Marquez, Irene Castellon, and Victoria Arranz. 2010. Document-level automatic MT evaluation based on discourse representations. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics- MATR, pages 333-338.
9. Christian Hardmei er and Marcello Federico. 2010.Modelling pronominal anaphora in statistical machine translation. In Proceedings of the International Workshop on Spoken Language Translation, pages 283 ‑ 289.
10. Ryu Iida, Takeno bu Tokunaga. A Metric for Evaluating Discourse Coherence based on Co-reference Resolution. In Proceedings of COLING 2012: Posters, pages 483-494.
11. Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL, pages 1060-1068.
12. Kazem Lotfipour-Saedi. 1997.Lexical cohesion and translation equivalence. Meta, 42(1):185-192.
13. D Xiong, Y Ding, M Zhang and C Tan. Lexical Chain Based Cohesion Models for Document-Level SMT. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1563-1573.
14. Karamanis, N., Poesio, M., Mellish, C., and Oberlander, J. (2004). Evaluating centering-based metrics of coherence using a reliably annotated corpus. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pages 391-398.

15. Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. Computational Linguistics, 21(2):203-226.
16. Miltsakaki, E. and Kukich, K. (2000). Automated evaluation of coherence in student essays. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000).
17. Barzilay, R. and Lapata, M. (2005). Modeling local coherence: An entity-based approach. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pages 141-148.
18. Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. Computational Linguistics, 34(1):1-34.