Automatic Evaluation for Translation of English Learners—from the Perspective of systemic-functional linguistics

系统功能语言学视角下对英语学习者翻译质量自动评价方法的研究

Introduction

Translation is always accompanied by evaluation. It provides people with a feedback to make improvements. Human evaluation is a classical method while it is “extensive but expensive (Papineni et al., 2002).” With the development of machine translation, the demand for an automatic evaluation system has been increasing accordingly. Similarly, the growing number of English learners, especially translation learners, has also invoked the demand for such a system that can be used in online marking of large-scale examinations.

Our research starts with introducing several representative evaluation methods and how language knowledges and resources are applied to improve these methods. Then when turn to linguistic side to introduce a translation evaluation model under the perspective of systemic-functional linguistics that defines three “metafunctions” of language—ideational function, interpersonal function and textual function.

Under the theoretical framework of these metafunctions, we discuss a semantic based evaluation method which involves the concept AMR (Abstract Meaning Representation) and summarize how this method embodies equivalence of ideational function during translation. We would do some experiments to prove that this semantic-based method outperforms classical methods such as BLEU and NIST.

In the next part, we focus on other two metafunctions. First, we demonstrate that interpersonal factors and textual factors should be taken into consideration in translation evaluation. Since there are many examples of evaluating the interpersonal function manually while an automatic method has not been developed, we attempt to extract certain parameters affecting the interpersonal and textual function of the text for the future algorithm. We may propose our own evaluation method.

Literature Review

When we search with the keyword “automatic evaluation of translation” in google scholar, nearly all of the related work is about evaluating machine translation results. In fact, the study on machine translation which produces large amount of text generates the need for automatic evaluation.

The central idea of those classical evaluation methods is evaluating the similarity between reference translation and candidate translation. “The closer a machine translation is to a professional human translation, the better it is.” The BLEU algorithm is one of the representatives.

*The primary programming task for a BLEU implementer is to compare n-grams of the candidate with the n-grams of the reference translation and count the number of matches. These matches are position-independent. The more the matches, the better the candidate translation is. (Papineni et al., 2002)*

The method is able to give different scores to different n-grams. The score of unigrams reflects the accuracy of a translation while the score of 3-grams and 4-grams reflects the fluency of a translation.

However, is it necessary for the candidate translation to be strictly equivalent with the reference? Maybe not. As a result, researchers make improvements for the method by allowing more flexible matching criteria. M-BLEU uses stemming and Wordnet based word mapping modules. (Agarwal & Lavie, 2008), NIST is an algorithm that assigns higher weights to words that have fewer occurrences. (Doddington, 2002). METEOR focuses on word-level semantic similarity. It computes a score based on explicit word-to-word matches between the translation and a given reference translation. (Banerjee & Lavie, 2005)

Another track of improving involves the brevity penalty. In BLEU, if the candidate translation is shorter than the reference, the final score should be punished. The algorithm AMBER adds ten different kinds of penalties including continuity penalty, redundancy penalty, etc. (Chen & Kuhn, 2011)

All the previous methods are attempting to apply language knowledge to make improvement. However, those knowledges are scattered and the metrics mainly target at evaluation of machine translation.

It is noteworthy that BLEU remains the most commonly used evaluation metric for machine translation not only because of its speed of computing but also of the experimental result that models trained using BLEU obtain the highest scores from humans and even from other metrics (Cer et al., 2010). A possible explanation is that what computer does is optimization to get closer to the “standard answer.”

Nevertheless, it is not the case in human translation. Although these metrics can be adopted in automatic evaluation of human translations, they have unique features that affects their evaluation metric especially for the translation of English learners. Qin Ying (2018) proposes a hierarchal evaluation system which includes the overall score and specific errors in the text. It provides learners with error types and modification suggestions. Si Xianzhu (2004) proposes a translation quality assessment model from the perspective of systemic-functional linguistics which evaluates from three aspects-- ideational function, interpersonal function and textual function. Among them, ideational function is the most important.

In systemic functional linguistics, the concept “ideational function” is used to describe people’s experience in real world expressed through language (Halliday, 1994). In other words, to evaluate the equivalence of ideational function, we need to compare the similarity of the fundamental semantic structure in sentence level. Abstract Meaning Representation (AMR) is designed to complete whole-sentence semantic representations. It aims to abstract away from syntactic idiosyncrasies (Banarescu et al., 2013).

*For example, the sentences “he described her as a genius”, “his description of her: genius”, and “she was a genius, according to his description” are all assigned the same AMR. (Banarescu et al., 2013)*

In the official guideline of AMR (Banarescu et al., 2013), the author raises a variety of concepts to represent all sentences in a certain language. And we notice that these concepts in AMR are related with the concept “process” which embodies transitivity in systemic-functional linguistics. We will analyze this relationship elaborately in our research.

AMR can be implemented in many fields such as natural language understanding and generating, while it can also play an important role in automatic evaluation of translation. The *smatch* metric is developed originally to assess inter-annotator agreement on AMR, as well as automatic AMR parsing accuracy. It enables us to compare semantic similarity between two texts which change into the candidate translation and reference translation in our research background.

According to Si Xianzhu (2005), other two metafunctions ought to be considered in translation evaluation. Since language is defined as a communicative behavior, the interpersonal function should not be lost throughout translation. Halliday (1994) holds that the interpersonal function is reflected by the mood system in the text. Yet, no current evaluation metrics would score this aspect separately mainly because most of them are designed for machine translation evaluation whose demand is only an overall score while this aspect becomes crucial in human translation. For instance, “walk” and “strut” have similar meaning, while they transmit different emotions.

Methodology

The study will start with a summary of current evaluation metric as the literature view part has mentioned. Then we draw conceptions of systemic-functional linguistics into discussion. We will analyze specific phenomena in AMR and figure out their relationship with verbal process which completes the ideational function.

Then we will implement an experiment to compare the performance of a semantic based metric and classical evaluation metrics. The performance is estimated by its correlation with the score of human evaluation. Our datasets are from LDC2006T04 and LDC2003T17. Each dataset includes 100 original text and every original text has 4 reference texts and 9 candidate texts with human evaluation score.

We will take interpersonal function into our evaluation. The dataset is selected from PACCEL. We will implement sentiment analysis to the reference and candidate text and figure out what can be applied in our evaluation metric.

Discourse analysis is the ultimate problem in natural language processing, which means that at present it is fairly difficult for computer to implement. We will raise the examples of discourse analysis and attempt to extract certain fixed parameters as auxiliary resources for the future algorithm.

Anticipated Results and Possible Explanations

Besides the analysis of current metrics from the perspective of systemic-functional linguistics, we expect to establish a demo system that automatically evaluates a translation of English learners. The evaluation includes an overall score which is computed by a semantic based metric and an independent score that reflects the deviation degree of interpersonal function. It should provide learners with error types and modification suggestions from the fundamental spelling error to deviation of textual function.

References

Agarwal, A. & A. Lavie. 2008. Meteor, m-bleu and m-ter: evaluation metrics for high correlation with human rankings of machine translation output [A]. Proceedings of the Third Workshop on Statistical Machine Translation [C]. Association for Computational Linguistics. 155-118

Banarescu, L., Bonial, C., Cai, S., et al. 2013. Abstract meaning representation for sembanking [A]. Proeedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse [C]. Association for Computational Linguistics. 178-186.

Banarescu, L., Bonial, C., Cai, S., et al. 2014. Abstract meaning representation (AMR) 1.2 specification. In parsing on freebase from question-answer pairs [A]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing [C]. Association for Computational Linguistics. 1533-1544.

Banerjee, S. & A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgements [A]. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and Summarization [C]. Association for Computational Linguistics. 65-72.

Cai, S. &K. Knight. 2013. Smatch: an evaluation metric for semantic feature structures [A]. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics [C]. Association for Computational Linguistics. 748-752.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics [A]. Proceedings of the Second International Conference on Human Language Technology Research [C].

Halliday, M. A. K. & C. M. Matthiessen. 2008. An Introduction to Functional Grammar [M]. Beijing: Foreign Language Teaching and Research Press.

Kishore Papineni., Salim Roukos., et al. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation [A]. Proceedings of the 40th Annual Meeting on ACL [C]. Association for Computational Linguistics. 311-318.

司显柱. 2004. 论功能语言学视角的翻译质量评估模式研究 [J]. 外语教学(4):45-50

秦颖. 2018. 翻译质量评价的标准与自动评价方法研究. 外语教学与研究出版社