Measuring Performance in Machine Translation Quality Estimation

Yvette Graham

Machine Translation Group
Dublin City University

Summary

Machine Translation (MT) aims to automatically learn how to translate text or speech from one natural language to another and provides one of the most challenging tasks in natural language processing. Given the vast number of possible system outputs, even at state-of-the-art performance, MT systems do not always produce perfect translations. Automatic estimation of the quality of MT output is known as MT Quality Estimation (QE) and is useful, for example, for flagging low quality output that requires post-editing prior to publication. MT QE provides its own set of challenges and evaluation of systems commonly takes the form of measurement of the degree of error that exists between QE system predictions and corresponding gold standard labels for a particular test set of translations. In this paper, we identify issues that arise during comparison of QE prediction score distributions and gold label distributions. We provide an analysis of methods of comparison and identify areas of concern with respect to widely used measures, such as the ability to gain by prediction of aggregate statistics specific to gold label distributions or by optimally conservative variance in prediction score distributions. As an alternative, we propose the use of the unit-free Pearson correlation, in addition to providing an appropriate method of significance testing improvements over a baseline. Components of quality estimation shared tasks are replicated to reveal substantially increased conclusivity in system rankings, including identification of outright winners of tasks. This paper received the best paper award at the 53rd Annual Meeting of the Association of Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJNLP 2015) [1].

Acknowledgement. This project has received funding from the European Union Horizon 2020 research and innovation programme under grant agreement 645452 (QT21) and the ADAPT Centre for Digital Content Technology at Dublin City University funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund.

References

1. Graham, Y.: Improving evaluation of machine translation quality estimation. In: 53rd Annual Meeting of the Association for Computational Linguistics and Seventh International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. pp. 1804–1813 (2015)