Appraise Evaluation Framework for Machine Translation

Christian Federmann
Microsoft Translator
One Microsoft Way
Redmond, WA 98052, USA
chrife@microsoft.com

Abstract

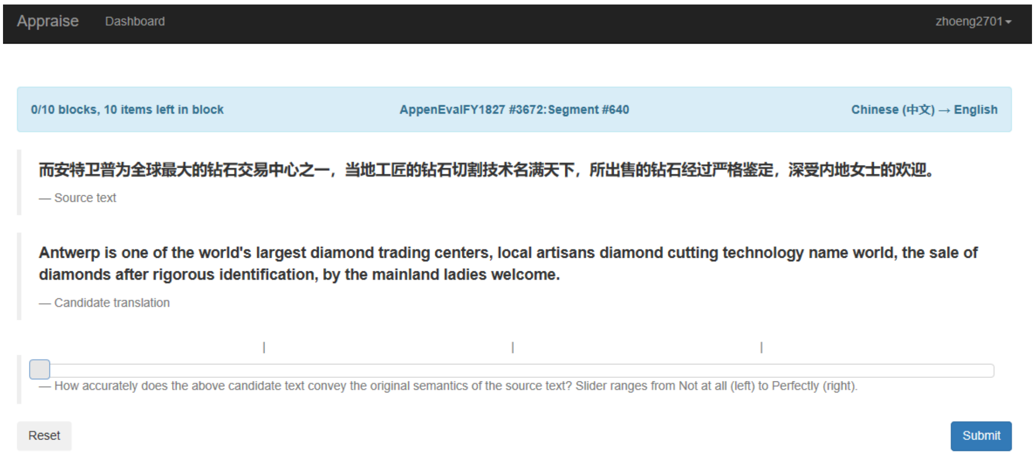
We present Appraise, an open-source framework for crowd-based annotation tasks, notably for evaluation of machine translation (MT) output. This is the software used to run the yearly evaluation campaigns for shared tasks at the WMT Conference on Machine Translation. It has also been used at IWSLT 2017 and, recently, to measure human parity for machine translation for Chinese to English news text. The demo will present the full end-to-end life cycle of an Appraise evaluation campaign, from task creation to annotation and interpretation of results.

1 Motivation

Human evaluation of machine translation is the ultimate measure of translation quality. However, due to data collection effort and annotation cost, many experiments and publications do not report results from human evaluation and rely on scores computed by automated metrics such as BLEU (Papineni et al., 2002) instead. We believe that machine translation researchers should be able to conduct manual annotation campaigns at scale, without having to re-implement the necessary infrastructure from scratch. Since 2007, development of the Appraise evaluation framework for machine translation has supported the research community, trying to bring more human evaluation into MT research.

2 Introduction

The Appraise framework has become a standard tool for machine translation evaluation. It is used for shared tasks at the yearly Conference on Machine Translation (WMT) (Bojar et al., 2017) and has been adopted at last year's IWSLT 2017 workshop (Cettolo et al., 2017). The Microsoft Translator team utilises the software for its internal quality monitoring. Figure 1 shows the annotation user interface.



0/10 blocks, 10 items left in block

AppenEvalFY1827 #3672:Segment #640

Chinese (中文) → English

而安特卫普为全球最大的钻石交易中心之一，当地工匠的钻石切割技术名满天下，所出售的钻石经过严格鉴定，深受内地女士的欢迎。

— Source text

Antwerp is one of the world's largest diamond trading centers, local artisans diamond cutting technology name world, the sale of diamonds after rigorous identification, by the mainland ladies welcome.

— Candidate translation

— How accurately does the above candidate text convey the original semantics of the source text? Slider ranges from Not at all (left) to Perfectly (right).

Figure 1: Screenshot of user interface for source-based direct assessment as implemented in Appraise.

Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 86–88
Santa Fe, New Mexico, USA, August 20-26, 2018.

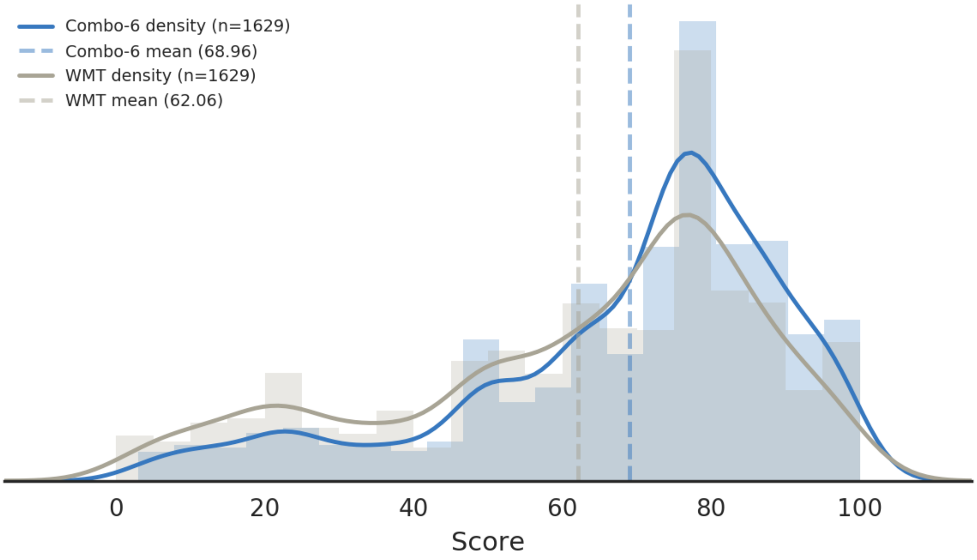


Figure 2: Visualisation of Appraise direct assessment results for Chinese to English news translation.

In 2018, Appraise was used as part of a research project which proved human parity for machine translation of Chinese to English news text (Awadalla et al., 2018), based on a large-scale evaluation campaign run using an Appraise system hosted on Azure. Figure 2 shows a graph visualising results from this work, comparing score distributions for the human parity system (COMBO-6) and the original WMT17 reference translation (WMT).

Our system demonstration provides an end-to-end overview on all aspects of a machine translation evaluation campaign within Appraise. We first describe input data, task creation, and campaign setup, including best practices regarding user and team management. Then, we show the annotation interface and discuss how annotator reliability is measured and monitored, allowing to detect spammers assigning random scores to candidate translations. We describe how statistical significance testing (Wilcoxon, 1945; Mann and Whitney, 1947; Riezler and Maxwell, 2005) can help to solve this problem. Lastly, we explain how final campaign results can be computed, extracted and visualised effectively, so that results are easily interpretable.

We also describe the annotation system's Python-based architecture and highlight implementation details as well as lessons learnt during ten years of human evaluation campaigns based on Appraise.

3 License

Appraise source code is available on GitHub[1] and is shared under a permissive license[2].

4 Conclusion

Our system demonstration explains the full end-to-end life cycle of an Appraise evaluation campaign. It gives an in-depth look into a decade of research on machine translation evaluation, including in-sights from several WMT campaigns as well as the evaluation part of Microsoft's recent human parity research breakthrough. This should lead to interesting discussions.

Acknowledgements

The author thanks the anonymous reviewers for their feedback. Appraise is developed incorporating feedback from the research community, participants of WMT and IWSLT shared tasks, and Microsoft. This support is crucial for the evolution of Appraise and, thus, much appreciated. Thank you!

[1] See https://github.com/cfedermann/Appraise/

[2] See https://github.com/cfedermann/Appraise/blob/master/appraise/LICENSE

References

Hany Hassan Awadalla, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. March.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, September. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuitho Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 Evaluation Campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)*, pages 2–12, Tokyo, Japan, December. IWSLT.

Christian Federmann. 2012. Appraise: An Open-Source Toolkit for Manual Evaluation of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35, September.

Christian Federmann. 2017. Appraise on Aure: A cloud-based, multi-purpose evaluation framework. In *Proceedings of the EAMT 2017: User Studies and Project/Product Descriptions*, page 32, Prague, Czech Republic, May. European Association for Machine Translation (EAMT).

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Stefan Riezler and John T Maxwell. 2005. On Some Pitfalls in Automatic Evaluation and significance Testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Stroudsburg, PA, USA, June. Association for Computational Linguistics.

Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.