ARC '16

مؤتمر مؤسسة قطر
السنوي للبحوث
QATAR FOUNDATION
ANNUAL RESEARCH
CONFERENCE

Towards World-class
Research and Innovation

مؤسسة قطر
Qatar Foundation
لإطلاق قدرات الإنسان
Unlocking human potential

Information Communications Technology Pillar

http://dx.doi.org/10.5339/qfarc.2016.ICTOP2013

Annotation Guidelines and Framework for Arabic Machine Translation Post-Edited Corpus

Wajdi Zaghouani[1], Nizar Habash[2], Ossama Obeid[1], Behrang Mohit[1], Houda Bouamor[1], Kemal Oflazer[1]

[1]Carnegie Mellon University-Qatar, QA
[2]New York University Abu Dhabi, AE

Email: wajdiz@qatar.cmu.edu

1. Introduction

Machine translation (MT) became widely used by translation companies to reduce their costs and improve their speed. Therefore, the demand for quick and accurate machine translations is growing. Machine translation (MT) systems often produce incorrect output with many grammatical and lexical choice errors. Correcting machine-produced translation errors, or MT Post-Editing (PE) can be done automatically or manually.

The availability of annotated resources is required for such approaches. When it comes to the Arabic language, to the best of our knowledge, there is no MT manually post-edited corpora available to build such systems. Therefore, there is a clear need to build such valuable resources for the Arabic language. In this abstract, we present our guidelines and annotation procedure to create a human corrected MT corpus for the Modern Standard Arabic (MSA). The creation of any manually annotated corpus usually presents many challenges. In order to address these challenges, we created a comprehensive and simplified annotation guidelines which were used by a team of five annotators and one lead annotator. In order to ensure a high annotation agreement between the annotators, multiple training sessions were held and regular inter annotator agreement (IAA) measures were performed to check the annotation quality

2. Corpus

We collected a corpus of 100K of English news article taken from the collaborative journalism Wikinews website. Afterwards, the corpus collected was automatically translated from English to Arabic using the Google Translate API paid service.

Cite this article as: Zaghouani W, Habash N, Obeid O, Mohit B, Bouamor H, Oflazer K. (2016). Annotation Guidelines and Framework for Arabic Machine Translation Post-Edited Corpus. Qatar Foundation Annual Research Conference Proceedings 2016: ICTOP2013 http://dx.doi.org/10.5339/qfarc.2016.ICTOP2013.

3. Guidelines

In order to annotate the MT corpus, we use the general annotation correction guidelines we designed previously for L1 described in Zaghouani et al. (2014) and we add specific MT post-editing correction rules. In the general correction guidelines we place the errors to be corrected into seven categories: spelling, word choice, morphology, syntax, proper names, dialectal usage and punctuation. We refer to Zaghouani et al. (2014) for more details about these errors. In the MT post-editing guidelines, we provide the annotators with detailed annotation procedure and explain how to deal with borderline cases. We include many annotated examples to illustrate some specific cases of machine translation correction rules. Since there are equally-accurate alternative ways to edit the machine translation output, all being considered correct, some using fewer edits than others, we explained in the guidelines that the machine translated texts should be corrected with a minimum number of edits necessary to achieve an acceptable translation quality. However, correcting the accuracy errors and producing a semantically coherent text is more important than minimizing the number of edits and therefore, the annotators were asked to pay attention to the following three aspects: accuracy, fluency and style.

4. Annotation Pipeline

The annotation team consisted of a lead annotator and six annotators. The lead annotator is also the annotation workflow manager of this project. He frequently evaluate the quality of the annotation, monitor and report on the annotation progress. A clearly defined protocol is set, including a routine for the Post-editing annotation job assignment and the inter-annotator agreement evaluation. The lead annotators is also responsible of the corpus selection and normalization process beside the annotation of the gold standard to be used to compute the Inter-Annotator Agreement (IAA) portion of the corpus.

The annotation itself is done using an in house built web annotation framework built originally for the manual correction of errors in L1 and L2 texts (Obeid et al., 2013). This framework includes two major components: 1. The annotation management interface which is used to assist the lead annotator in the general work-flow process, it allows the user to upload, assign, monitor, evaluate and export annotation tasks. 2. The MT post-editing annotation interface is the actual annotation tool, which allows the annotators to do the manual correction of the MT Arabic output.

5. Evaluation

The low average WER of 4.92 obtained show a high agreement with the post-editing done in the first round between three annotators. The results obtained with the MT are comparable to those obtained with the L2 corpus, this can be explained by the difficult nature of both corpora and the multiple acceptable corrections for both.

6. Related Work

Large scale manually corrected MT corpora are not yet widely available due to the high cost related to building such resources. For the Arabic language, we cite the effort of Bouamor et al. (2014) who created a medium scale human judgment corpus of Arabic machine translation using the output of six MT systems and a total of 1892 sentences and 22k rankings. Our corpus is a part of the Qatar Arabic Language Bank (QALB) project, a large scale manually annotated annotation project (Zaghouani et al., 2014; Zaghouani et al., 2015). The project goal was to create an error corrected 2M-word corpus for online user comments on news websites, native speaker essays, non-native speaker essays and machine translation output.

7. Conclusion

We have presented in detail the methodology used to create a 100K word English to Arabic MT manually post-edited corpus, including the development of the guidelines as well as the annotation procedure and the quality control procedure using frequent inter-annotator measures. The created guidelines will be made publicly available and we look forward to distribute the post-edited corpus in a planned shared task on automatic error correction and getting feedback from the community on its usefulness as it was in the previous shared

tasks we organized for the L1 and L2 corpus (Mohit et al., 2014; Rozovskaya et al., 2015). We believe that this corpus will be valuable to advance research efforts in the machine translation area since manually annotated data is often needed by the MT community. We believe that our methodology for guideline development and annotation consistency checking can be applied in other projects and other languages as well.

8. Acknowledgement

This project is supported by the National Priority Research Program (NPRP grant 4-1058-1-168) of the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

9. References

Obeid, O., Zaghouani, W., Mohit, B., Habash, N., Oflazer,K., and Tomeh, N. (2013). A Web-based Annotation Framework For Large-Scale Text Correction. In The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations, Nagoya, Japan, October.

Mohit, B., Rozovskaya, A., Habash, N., Zaghouani, W., and Obeid, O. (2014). The first QALB shared task on automatic text correction for Arabic. ANLP 2014, page 39.

Rozovskaya Alla; Houda Bouamor; Nizar Habash; Wajdi Zaghouani; Ossama Obeid; Behrang Mohit. The Second QALB Shared Task on Automatic Text Correction for Arabic. In Proceedings of the ACL 2015 Workshop on Arabic Natural Language Processing (ANLP), Beijing, China, July 2015.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh,N., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014). Large scale Arabic error annotation: Guidelines and framework. In International Conference on Language Resources and Evaluation (LREC 2014).

Zaghouani, W., Habash, N., Bouamor, H., Rozovskaya, A., Mohit, B., Heider, A., and Oflazer, K. (2015). Correction annotation for non-native Arabic texts: Guidelines and corpus. Proceedings of The 9th Linguistic Annotation Workshop, pages 129–139.