Appraise: an Open-Source Toolkit for Manual Evaluation of MT Output

Christian Federmann

Language Technology Lab,
German Research Center for Artificial Intelligence,
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
cfedermann@dfki.de

Abstract. We describe Appraise, an open-source toolkit which can be used to do manual evaluation of Machine Translation output. Appraise allows to collect human judgments on translation output, implementing annotation tasks such as 1) translation quality checking, 2) ranking of translations, 3) error classification, and 4) manual post-editing. It uses an extensible format for import/export of data and can easily be adapted to new annotation tasks. The annotation tasks are explained in more detail in the paper. The current version of Appraise also includes automatic computation of inter-annotator agreement scores resulting in quick access to evaluation results. Appraise has successfully been used for a wide variety of research projects.

Keywords: Machine Translation, Evaluation, Applications

1 Introduction

Evaluation of Machine Translation (MT) output to assess translation quality is a difficult task. There exist automatic metrics such as BLEU [11] or Meteor [4] which are widely used in minimum error rate training [10] for tuning of MT systems and as evaluation metric for shared tasks such as, e.g., the Workshop on Statistical Machine Translation (WMT) [3]. The main problem in designing automatic quality metrics for MT is to achieve a high correlation with human judgments on the same translation output. While current metrics show promising performance in this respect, manual inspection and evaluation of MT results is still equally important as it allows for a more targeted and detailed analysis of the given translation output. The manual analysis of a machine translated text is, however, a time-consuming and laborious process; it involves training of annotators, requires detailed and concise annotation guidelines, and—last but not least—an annotation software that allows annotators to get their job done quickly and efficiently.

As we have mentioned before, the collection of manual judgments on machine translation output is a tedious task; this holds for simple tasks such as translation ranking but also for more complex challenges like word-level error analysis or post-editing of translation output. Annotators tend to lose focus after several

S. Wölfl (Ed.): Poster and Demo Track of the 35th German Conference on
Artificial Intelligence (KI-2012), pp. 130-134, 2012.

sentences, resulting in reduced intra-annotator agreement and increased annotation time. In our experience with manual evaluation campaigns it has shown that a well-designed annotation tool can help to overcome these issues.

In this paper, we describe *Appraise*, an open-source application toolkit that allows to perform manual evaluation of Machine Translation output. Appraise can be used to collect human judgments on translation output, implementing several annotation tasks. Development of the Appraise software package started back in 2009 as part of the EuroMatrixPlus project where the tool was used to quickly compare different sets of candidate translations from our hybrid machine translation engine to get an indication whether our system improved or degraded in terms of translation quality. A first version of Appraise was released and described by [5].

The remainder of this paper is structured as follows: Section 2 provides a brief description of the evaluation system before we highlight the different annotation tasks that have been implemented in Section 3. Finally, we describe several experiments where Appraise has proven useful (see Section 4) and give some concluding remarks in Section 5.

2 System Description

In a nutshell, Appraise is an open-source tool for manual evaluation of machine translation output. It allows to collect human judgments on given translation output, implementing annotation tasks such as (but not limited to):

- translation quality checking;
- ranking of translations;
- error classification;
- manual post-editing.

The software features an extensible XML import/output format[1] and can easily be adapted to new annotation tasks. The tool also includes code supporting the automatic computation of inter-annotator agreement scores, allowing quick access to evaluation results. We currently support computation of the following inter-annotator agreement scores:

- Krippendorff's α as described by [9];
- Fleiss' κ as published in [8];
- Bennett, Alpert, and Goldsteins S as defined in [1];
- Scott's π as introduced in [12].

Agreement computation relies on code from the NLTK project [2]. Additional agreement metrics can be added easily.

[1] An example of this XML format is available at GitHub: `https://raw.github.com/cfedermann/Appraise/master/examples/sample-ranking-task.xml`.

We have opened up Appraise development and released the source code on GitHub at `https://github.com/cfedermann/Appraise`. Anybody may fork the project and create an own version of the software. Due to the flexibility of the `git` source code management system, it is easy to re-integrate external changes into the master repository, allowing other developers to feed back bugfixes and new features, thus improving and extending the original software. Appraise is available under an open, BSD-style license.[2]

3 Annotation Tasks

We have developed several annotation tasks which are useful for the evaluation of machine translation output. All of these have been tested and used during the experiments described in Section 4. The following task types are available for the GitHub version of Appraise:

1. **Ranking** The annotator is shown 1) the source sentence and 2) several ($n \geq 2$) candidate translations. It is also possible to additionally present the reference translation. Wherever available, one sentence of left/right context is displayed to support the annotator during the ranking process.

   We also have implemented a special *3-way ranking task* which works for pairs of candidate translations and gives the annotator an intuitive interface for quick $A > B$, $A = B$, or $A < B$ classification.

2. **Quality Estimation** The annotator is given 1) the source sentence and 2) one candidate translation which has to be classified as *Acceptable*, *Can easily be fixed*, or *None of both*. We also show the reference sentence and again present left/right context if available. This task can be used to get a quick estimate on the *acceptability* of a set of translations.

3. **Error Classification** The annotator sees 1) the source (or target) sentence and 2) a candidate translation which has to be inspected wrt. errors that can be observed in the translation output. Error annotation is possible on the sentence level as well as for individual words. The annotator can choose to skip a translation marking it as containing *"too many errors"* and is able to differentiate between *"minor"* and *"severe"* errors in the annotation.

4. **Post-editing** The annotator is shown 1) the source sentence, with left/right context wherever available, and 2) one or several candidate translation. The task is defined as choosing the translation which is *"easiest to post-edit"* and then performing the post-editing operation on the selected translation.

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

4 Experiments

We have created Appraise to support research work on hybrid MT, especially during the *EuroMatrixPlus* project. We have also used Appraise in the *taraXÜ* project, conducting several large annotation campaigns involving professional translators and language service providers. In the *T4ME* project, we investigate how hybrid machine translation can be changed towards optimal selection from the given candidate translations. Part of the experimental setup is a shared task (ML4HMT) in which participants have to implement this optimal choice step. We use Appraise to assess the translation quality of the resulting systems. Appraise has also been used in research work related to the creation of standalone hybrid machine translation approaches. Finally, we use Appraise in the context of terminology translation for the financial domain in the *MONNET* project.

4.1 Results

As Appraise is a tool supporting evaluation it is difficult to point to individual results achieved through its usage. We were, however, able to find experimental proof that aforementioned automated evaluation metric Meteor correlates "best" with results from human judgement. This work has been described and published in [6]. Also, using Appraise, we were able to show that rule-based systems which performed worse than statistical MT systems according to automatic metrics were actually better in translation quality. This is described in our submission to last year's WMT shared task [7].

5 Conclusion and Outlook

We have described Appraise, an open-source tool for manual evaluation of MT output, implementing various annotation tasks such as error classification or post-editing. We also briefly reported on research projects in which different versions of the Appraise toolkit have been used, feeding back into and supporting the tool's development, eventually leading to its current version.

Maintenance and development efforts of the Appraise software package are ongoing. By publicly releasing the tool on GitHub, we hope to attract both new users and new developers to further extend and improve it. Future modifications will focus on new annotation tasks and a more accessible administration interface for large numbers of tasks. Last but not least, we intend to incorporate detailed visualisation of annotation results into Appraise.

Acknowledgments

The research work in this paper has been funded through the T4ME contract (grant agreement no.: 249119) under the Seventh Framework Programme for Research and Technological Development of the European Commission. We are grateful to the anonymous reviewers for their valuable feedback.

References

1. Bennett, E.M., Alpert, R., Goldstein, A.C.: Communications through limited-response questioning. Public Opinion Quarterly 18(3), 303–308 (1954)
2. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly, Beijing (2009), `http://www.nltk.org/book`
3. Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., Specia, L. (eds.): Proceedings of the Seventh Workshop on Statistical Machine Translation. Association for Computational Linguistics, Montréal, Canada (June 2012), `http://www.aclweb.org/anthology/W12-31`
4. Denkowski, M., Lavie, A.: Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 85–91. Association for Computational Linguistics, Edinburgh, Scotland (July 2011), `http://www.aclweb.org/anthology-new/W/W11/W11-2107`
5. Federmann, C.: Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valetta, Malta (May 2010), `http://www.lrec-conf.org/proceedings/lrec2010/pdf/197_Paper.pdf`
6. Federmann, C.: Results from the ml4hmt shared task on applying machine learning techniques to optimise the division of labour in hybrid machine translation. In: Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4. META-NET (11 2011)
7. Federmann, C., Hunsicker, S.: Stochastic parse tree selection for an existing rbmt system. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 351–357. Association for Computational Linguistics, Edinburgh, Scotland (July 2011), `http://www.aclweb.org/anthology/W11-2141`
8. Fleiss, J.: Measuring Nominal Scale Agreement among Many Raters. Psychological Bulletin 76(5), 378–382 (1971)
9. Krippendorff, K.: Reliability in Content Analysis. Some Common Misconceptions and Recommendations. Human Communication Research 30(3), 411–433 (2004)
10. Och, F.J.: Minimum error rate training in statistical machine translation. In: ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. pp. 160–167. Association for Computational Linguistics, Morristown, NJ, USA (2003)
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002), `http://acl.ldc.upenn.edu/P/P02/P02-1040.pdf`
12. Scott, W.A.: Reliability of Content Analysis: The Case of Nominal Scale Coding. The Public Opinion Quarterly 19(3), 321–325 (1955)