Incorporating Linguistic Information in Machine Translation Evaluation

Jason Adams

15 February 2008

Machine translation evaluation continues to be an active area of research. Completely automatic MT evaluation would require a system capable of human-level understanding of language, which would in turn allow us to create a human-level machine translation system, leading to a chicken and egg problem. We cannot create one without creating the other. We simplify the problem by using human generated reference translations. Current metrics evaluate MT system output by comparing machine generated sentences to these reference translations. Comparisons typically involve calculating measures of precision and recall between sentence pairs and producing sentence and/or system level assessments of quality. A recent trend has been to include more linguistic information in evaluation metrics. A growing number of papers show that by doing so, we can improve correlation with human judgments beyond BLEU (Papineni et al., 2002).

Liu and Gildea (2005) introduced the notion of incorporating parsing into machine translation evaluation through dependency-based headword chains. When a word is linked to its parent in a dependency parse, we have created a headword chain of length 2. Headword chains can be extended to arbitrary lengths (depending only on the sentences they are derived from) and compared in machine and human translations as a means of evaluating MT system quality. Liu and Gildea (2005) found that introducing syntactic features, such as the Headword Chain Metric (HWCM), improved sentence level correlation with human judgments over BLEU. Others have extended their work by using HWCM as a feature in a machine learning sytem (Albrecht and Hwa, 2007; Kuleska and Shieber, 2004). Owczarzak et al. (2007) brings in linguistic information by parsing the MT output and reference translations with a Lexical Functional Grammar (LFG) parser. Dependencies are extracted from the LFG parse and lexical variation is addressed using WordNet. They found that correlation with human judgments of fluency were higher than METEOR (Banerjee and Lavie, 2005) on Multiple-Translation Chinese data sets (parts 2 and 4), while METEOR outperformed them on adequacy. Another metric, Bllip, constructs dependency sets (a headword bichain) for the reference and machine translations using the Charniak parser (Pozar and Charniak, 2006). They compute the score using the number of matches and found this correlated with human judgments

better than BLEU.

The two representative papers are (Owczarzak et al., 2007) and (Albrecht and Hwa, 2007), but those may be revised.

References

Joshua S. Albrecht and Rebecca Hwa. A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 880–887, Prague, Czech Republic, 2007. ACM Press.

Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation measures for MT and/or Summarization*, 2005.

Alex Kuleska and Stuart M. Shieber. A Learning Approach to Improving Sentence-level MT Evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, 2004.

Ding Liu and Daniel Gildea. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 2005.

Karolina Owczarzak, Josef van Genabith, and Andy Way. Labelled Dependencies in Machine Translation Evaluation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 104–111, Prague, Czech Republic, 2007. ACM Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL '02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

Michael Pozar and Eugene Charniak. Bllip: An Improved Evaluation Metric for Machine Translation. Master's thesis, Brown University, 2006.