

Word Sense Disambiguation and Machine Translation

Jason Adams

23 January 2008

It seems natural to assume that achieving correct word sense disambiguation (WSD) can help to correctly translate from one language to another. Until the middle part of this decade, little work had been done to confirm this assumption, especially with regards to statistical machine translation (SMT) systems. These systems typically rely on local context for lexical choice and so do some word sense disambiguation on their own (Carpuat and Wu, 2005a). A dedicated WSD system should be able to produce higher quality results which could then be introduced into an SMT system to improve lexical choice. Initial results by Carpuat and Wu (2005a) actually found this assumption to be incorrect. However, the same authors later showed that by integrating WSD into a phrase-based SMT system, BLEU scores never decreased and often increased (Carpuat and Wu, 2007a). Another approach by Chan et al. has shown improved results using WSD with a hierarchical phrase based SMT system.

Carpuat and Wu (2007a) propose a method by which WSD is incorporated directly into an SMT system and performs what they call "*multi-word phrasal* lexical disambiguation." The WSD system they use consists of an ensemble of four WSD models: naïve Bayes, maximum entropy, boosting, and Kernel PCA. The WSD component is added to the decoder in the Pharaoh system by generating a lexicon for every sentence. For more details, please see (Carpuat and Wu, 2007b). They report improved results on a variety of metrics including BLEU and METEOR for the NIST MT 2004 evaluation set.

The approach by Chan et al. involves integrating a WSD classifier into the phrase based SMT system, Hiero. The WSD classifier is built using an SVM and is independently evaluated on the SENSEVAL-3 Chinese lexical-sample task. The classifier is incorporated into Hiero and offers words to achieve higher probability translations during decoding. They report a statistically significant improvement of 0.63 BLEU points over their baseline (tested on the NIST MT 2003 evaluation set).

Further Reading

The two primary papers I hope to examine are (Carpuat and Wu, 2007a) and (Chan et al.). Further reading is given in the bibliography that is not

cited above.

References

- C. Cabezas and P. Resnik. Using WSD Techniques for Lexical Selection in Statistical Machine Translation. Technical report, Technical report, University of Maryland, 2005.
- M. Carpuat and D. Wu. Word sense disambiguation vs. statistical machine translation. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 387–394, 2005a.
- M. Carpuat and D. Wu. Evaluating the word sense disambiguation performance of statistical machine translation. *Second International Joint Conference on Natural Language Processing (IJCNLP-2005)*, 2005b.
- M. Carpuat and D. Wu. Improving Statistical Machine Translation Using Word Sense Disambiguation. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, 2007a.
- M. Carpuat and D. Wu. How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation, 2007b.
- Y.S. Chan, H.T. Ng, and D. Chiang. Word Sense Disambiguation Improves Statistical Machine Translation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40.
- M. Holmqvist. Memory-Based Learning of Word Translation. *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*, 2007.
- H. Li and C. Li. Word Translation Disambiguation Using Bilingual Bootstrapping. *Computational Linguistics*, 30(1):1–22, 2004.
- L. Specia. A Hybrid Model for Word Sense Disambiguation in English-Portuguese Machine Translation. *Proceedings of the 8th Research Colloquium of the UK Special-interest Group in Computational Linguistics*, pages 71–78, 2005.
- D. Vickrey, L. Biewald, M. Teyssier, and D. Koller. Word-sense disambiguation for machine translation. *HLT/EMNLP*, 2005.