

System	baseline	+reranking
BLEU	0.406	0.401
NIST	7.048	6.368
PER	0.424	0.417
TER	0.424	0.388
WER	0.500	0.481
METEOR	0.662	0.658
GTM	0.695	0.680
F1	0.699	0.709
PREC	0.708	0.745
RECL	0.690	0.676

Table 2: Official evaluation scores on eval09 obtained by primary Chinese-English baseline system vs. semi-supervised contrastive system, truecased version.

System	BLEU	PER	Meteor	NIST
case+punc	0.41	0.42	0.66	7.05
no case+punc	0.40	0.45	0.62	7.30

Table 3: Chinese-English translation results on the IWSLT09 eval set - subset of the official evaluation scores.

previous years, several techniques that were observed to increase MT performance (e.g. POS-based language models) did not show improvements on this years tasks, possibly due to the limitation of only using the BTEC training data. Our main goal this year was the integration of novel semi-supervised learning techniques, in particular semi-supervised ranking. Results from this method are inconclusive, improving some performance measures while decreasing others. This is in contrast to earlier experiments on two previous IWSLT data sets – further experiments need to be conducted to determine the reason for this discrepancy.

9. Acknowledgements

This work was partially funded by NSF grant IIS-0840461.

10. References

- [1] Bennett, K., Demiriz, A., and Maclin, R., “Exploiting Unlabeled Data in Ensemble Methods”, *Proceedings of SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.
- [2] Chang P-C., M. Gally and C. Manning, “Optimizing Chinese Word Segmentation for Machine Translation Performance”, *ACL 2008 Third Workshop on Statistical Machine Translation*, 2008.
- [3] Freund, Y., Iyer, R., Schapire, R., and Singer, Y., “An Efficient Boosting Algorithm for Combining Preferences”, *Journal of Machine Learning Research* 4, 2003.

System	BLEU	PER	Meteor	NIST
case+punc	0.48	0.35	0.72	6.85
no case+punc	0.48	0.38	0.69	6.93

Table 4: Arabic-English translation results on the IWSLT09 eval set - subset of the official evaluation scores.

- [4] Stolcke, A., “SRILM: An Extensible Language Modeling Toolkit”, *Proceedings of ICSLP*, 2002.
- [5] Habash, N. and O. Rambow, “Arabic Tokenization, Morphological Analysis, and Part-of-Speech Tagging in One Fell Swoop”, *Proceedings of ACL*, 2005
- [6] Byrne, W., and Deng, Y., “MTTK: An Alignment Toolkit for Statistical MACHine Translation”, *Proceedings of HTL/NAACL*, New York, 2006.
- [7] Koehn, P. and Och, F.J. and Marcu, D., “Statistical phrase-based translation”, *Proceedings of HLT/NAACL*, Edmonton, Canada, 2003.
- [8] Koehn, P. et al., “Moses: Open Source Toolkit for Statistical Machine Translation”, *Proceedings of ACL demo session*, Prague, 2007.
- [6] Och, F.J., and Ney, H., “A systematic comparison of various statistical alignment models”, *Computational Linguistics* 29(1), 19-52, 2003
- [9] Och, F.J., “Minimum Error Rate Training for Statistical Machine Translation”, *Proceedings of ACL*, Sapporo, Japan, 2003.
- [10] Ratnaparkhi, A., “A maximum entropy part-of-speech tagger”, in *Proc. of Empirical Methods in Natural Language Processing (EMNLP)*, 1996.
- [11] Zhang, B. and J.G. Kahn, *Evaluation of Decatur Text Normalizer for Language Model Training*, Technical report, University of Washington