Keynotes

Improving SMT by Model Filtering and Phrase Embedding

Chengqing Zong, National Laboratory of Pattern Recognition

Abstract

In phrase-based and hierarchical phrase-based statistical machine translation systems, translation performance is heavily dependent on the size and quality of the translation table. To meet the requirements of enabling real-time responses, some research has focused upon filtering (pruning) of the translation table. However, most existing filtering methods have been based on application of one or two constraints that act as hard rules, such as disallowing phrase-pairs with low translation probabilities. These approaches sometimes result in rigid constraints because they consider only a single factor instead of composite factors. In view of the above considerations, I will introduce a machine learning-based framework that integrates multiple features when pruning translation models. In addition, to improve the performance of phrase-based translation models, I will propose Bilingually-constrained Recursive Auto-encoders (BRAEs) for learning semantic phrase embeddings (compact vector representations for phrases), capable of distinguishing phrases with different semantic meanings. This method has been evaluated on two end-to-end SMT tasks and shows remarkable effectiveness on both tasks.

Bio

Chengqing Zong received his Ph.D. from the Institute of Computing Technology of the Chinese Academy of Sciences, in March, 1998. From May, 1998 to April, 2000 he worked as a post-doctoral researcher at the National Laboratory of Pattern Recognition (NLPR) in the Institute of Automation of the Chinese Academy of Sciences. He joined the NLPR in April, 2000, and is now a professor. In 1999 and 2001 he visited Japan's Advanced Telecommunications Research Institute International (ATR) as a guest researcher. From October, 2004 to February, 2005 he visited CLIPS-IMAG in France. His research interests include machine translation, natural language processing, and sentiment classification. He has authored a book and published more than 100 papers. He is a member of the International Committee on Computational Linguistics (ICCL) and the chair of Special Interest Group on Chinese Language Processing (SIGHAN) of the Association for Computational Linguistics (ACL). He is an associate editor of ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) and an editorial board member of IEEE Intelligent Systems, Machine Translation, Journal of Computer Science and Technology. He also served the

ACL-IJCNLP 2015 conference as a programming committee co-chair, and has helped organize many other conferences, including IJCAI, COLING, EMNLP and WWW etc., as a programming committee member or in other leadership positions.