[Interspeech 2010 tutorial]

Foundations of Statistical Machine Translation: Past, Present and Future

Taro Watanabe (National Institute of Information and Communications Technology)

Summary:

Statistical Machine Translation (SMT) was first introduced almost two decades ago by applying the concept of source-channel modeling, a theoretical background shared with Speech Recognition. SMT has attracted large interest from research community and its performance has drastically increased with regards to both quality and speed. One technique that has made a major contribution is tree-based SMT, which can incorporate syntactic knowledge empowered by efficient algorithms to handle such complex models. Recent advances in Machine Learning techniques together with efficient methods to handle large data sets have also resulted in significant improvements. This tutorial will serve as a basic introduction to SMT and summarize the twenty-year effort of improving SMT, but mainly concentrate on a few selected topics covering theoretical views and practical aspects.

First, we will review the theoretical backgrounds behind SMT and introduce three basic concepts, models, training, and search. Especially, we will detail the word alignment model and the phrase-based models, discuss generative and discriminative training, and present search algorithms based on Dynamic Programming.

Second, we will explore various tree-based SMT approaches with many alternative configurations, such as string-to-string or string-to-tree approaches, which can be cast as a synchronous-CFG formalism or a tree-transducer formalism. By incorporating syntactic knowledge into translation models, we can capture long distance reordering necessary to perform the reordering usually observed in syntactically divergent language pairs. These tree-based formalisms can be represented as a hypergraph framework, a natural extension of finite state models operating on strings to trees. We will cover various algorithms that operate over the hypergraph framework such as k-best extraction and the lazy feature function evaluation by cube pruning.

Third, we will present current hot topics for SMT in selected areas, data structures for efficient handling of large data, such as randomized data structures and succinct data structures, and Machine Learning techniques, such as large margin approaches and Bayesian approaches. We will also cover efficient algorithms for training or decoding, such as MBR decoding, by exploiting compact representations of many alternative translations, such as lattice or hypergraph data structures.

Biography:

Taro Watanabe received the B.E. and M.E. degrees in information science from Kyoto Univ., Kyoto, Japan in 1994 and 1997, respectively, and obtained the Master of Science degree in language and information technologies from the School of Computer Science, Carnegie Mellon University in 2000. In 2004, he received the Ph.D. in informatics from Kyoto Univ., Kyoto, Japan. Dr. Watanabe is an expert researcher of National Institute of Information and Communications Technology. His research interests include natural language processing, machine learning and statistical machine translation.