6.2 ***The Future of Sequence Modeling: A Peek into Modern Speech/Language Technology***

Session Abstract

Organizers: Jeff Bilmes, University of Washington and Hitoshi Isahara, NICT

Over the past 40 years, speech recognition technology has been at the forefront of statistical time-series modeling. Many of the methodologies that were developed for the problem of solving speech recognition, including the well-known hidden Markov model or HMM, have led to major improvements in terms of both speech recognition quality and computational demands. But these techniques have gone on to provide enormous benefits to other application areas as well such as biological sequence modeling (such as genomics and proteomics), economic modeling, and human activity and behavioral prediction. Here in 2008, however, while many application areas are still enamored of HMMs, speech recognition technology has evolved far beyond the HMM in many state-of-the-art systems. Given the success speech recognition technology has had in driving sequence modeling in the past, it is therefore likely that new speech technology can foretell some of the technological advances that will be successful in other application areas in the subsequent 40 years. In other words, we may take a peek into the future of sequence modeling by looking at today's modern speech recognition technology.

More than 50 years ago, research on natural language processing (NLP) started with so-called rule-based methodology, however, compilation of huge amount of grammar rules and dictionary entries are too difficult to develop practical systems. Then, trend of NLP research shifted to corpus-based or statistical systems. Thanks to the rapid improvement of computer power and data storage, nowadays we can utilize huge amount of actual linguistic data. Combining such linguistic resources and high quality language analyzer, we can extract useful linguistic information and develop practical systems for specific domain. Fusion of knowledge and example, or knowledge processing using linguistic resources, is one of the possibilities to develop high-performance NLP systems in the future. Based on this consideration, widely applicable and high-performance NLP technologies and linguistic resources are being developed. As for linguistic resources, we already compiled and published the Corpus of Spontaneous Japanese (CSJ), which is transcribed corpus of spontaneous speech and is used in various research on speech and text processing. R&D into NLP using language resources involves the development of fundamental NLP technologies to be utilized in speech processing and information retrieval.

In the first two talks of this session, we will hear from two young pioneering researchers in the area of speech recognition. In the first talk, Prof. Mark Hasegawa-Johnson will describe a language-independent speech information retrieval system, based on a large HMM with a few hundred thousand states; the states of the HMM are language-independent, but transition probabilities depend on the query being retrieved. He will argue that an HMM of this size provides a (usually) unique mapping, from a waveform snippet of arbitrary duration, into a

high-dimensional log-probability feature space.

Metadata in the training languages impose boundaries in feature space; by dividing at every well-supported boundary (using, e.g., a classification tree), we can design a set of sound units that cover most of the distinctions in most languages of the world. The second talk of this session will feature Prof. Karen Livescu describing new statistical models that may help us to deal with the challenge of pronunciation variation. In naturally spoken language, we often pronounce words in ways surprisingly different from their dictionary pronunciations, and this can dramatically reduce the accuracy of state-of-the-art speech recognizers. Prof. Livescu will describe new approaches, based on graphical models, that better account for this variation and that are also likely to have applications beyond speech recognition.

In the second half of this session, we will have two talks, each of which tackles real-world problems of speech recognition and web search engine for information retrieval on the internet. In the third talk, Prof. Tatsuya Kawahara will address automatic transcription of Japanese Diet meetings and classroom lectures. One of the biggest challenges encountered in such spontaneous speech is deviation from written language, including pronunciation variations and colloquial expressions. Prof. Kawahara will present a novel statistical scheme to approach the issue, and demonstrate transcription systems to be deployed in the real field. In the fourth talk, Mr Tatsumi Kobayashi will talk about how to incorporate a recommendation framework into the relevance-based traditional web search engines. He utilizes a collective intelligence technique to analyze user search behaviors, disambiguate query meanings and discover a co-related query group considering a trend and linguistic meanings. A proposed new recommendation framework is built on the top of deep analysis using a query log, user click behaviors and content analysis. This approach contributes an addition of trend sensitive contents into a robust nature of the current search engines, and helps users to find a valuable content effectively.