Haşim Sak

Thesis Supervisor: Assoc. Prof. Tunga Güngör
Thesis Co-Supervisor: Assoc. Prof. Murat Saraçlar

Integrating Morphology into Automatic Speech Recognition: Morpholexical and Discriminative Language Models for Turkish

Languages with agglutinative or inflectional morphology prove to be challenging for speech and language processing (SLP). The main problem with these languages is their relatively large vocabulary size, which may be infinite in some of these languages. The large vocabulary leads to a high number of out-of-vocabulary (OOV) words for SLP applications in which the vocabulary of the system is generally limited to a fixed size. Higher OOV rates decreases accuracy in these systems. Increasing the vocabulary size may degrade the time and space efficiency of the system. Besides, the reliability of parameter estimates in statistical models may deteriorate due to sparse data with increasing vocabulary sizes. These problems are more pronounced for Turkish with its extremely productive inflectional and derivational morphology and rather flexible word order in addition to its unlimited vocabulary.

In this thesis, we tackle with these challenges for Turkish in automatic speech recognition frame. First, we build the necessary tools and resources for Turkish, namely a finite-state morphological parser (computational lexicon), a perceptron-based morphological disambiguator, and a text corpus collected from web. Then using these tools, we build a generative *n*-gram model, *morpholexical language model*, where modeling units are lexical-grammatical morphemes instead of commonly used words or sub-words. The morpholexical language model is composed with the lexical transducer of the morphological parser to obtain a morphology-integrated search network (morpholexical search network), which effectively solves over-generation problem (generation of invalid word forms) of sub-lexical models. We also build a linear model trained discriminatively using morpholexical and morphosyntactic features to rerank *n*-best candidates obtained with the generative model. We apply the proposed models in broadcast news transcription task and give experimental results. The morpholexical model leads to an elegant morphology-integrated search network with unlimited vocabulary. Thus, it is highly effective in alleviating OOV problem and improves the word error rate (WER) over word and statistical sub-word models. The discriminative model with a rich set of morphological features and novel *n*-best-list edit features further improves the WER of the system. Finally, we present an algorithm for on-the-fly lattice rescoring with low-latency. The methodologies of this thesis can be applied to other morphologically rich languages and to other application areas, such as machine translation.

PUBLICATIONS

Journals

1) **Haşim Sak**, Tunga Güngör, and Murat Saraçlar: Resources for Turkish Morphological Processing. Language Resources and Evaluation, Vol. 45, No. 2, pp. 249–261, 2011. (SCIE-SSCI)
2) Ebru Arısoy, Doğan Can, Sıddıka Parlak, **Haşim Sak**, and Murat Saraçlar: Turkish Broadcast News Transcription and Retrieval. IEEE Transactions on Audio, Speech & Language Processing, Vol. 17, No. 5, pp. 874–883, 2009. (SCIE-SSCI)

Conferences

1) **Haşim Sak**, Murat Saraçlar, and Tunga Güngör: Discriminative Reranking of ASR Hypotheses with Morpholexical and N-best-list Features . IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), 2011, kabul edildi.
3) **Haşim Sak**, Murat Saraçlar, and Tunga Güngör: On-the-fly Lattice Rescoring for Real-time Automatic Speech Recognition. INTERSPEECH 2010, pp. 2450-2453, 2010.
4) **Haşim Sak**, Murat Saraçlar, and Tunga Güngör: Morphology-based and Sub-word Language Modeling for Turkish Speech Recognition. Acoustics Speech and Signal Processing (ICASSP), pp. 5402–5405, 2010.
5) **Haşim Sak**, Murat Saraçlar, and Tunga Güngör: Integrating Morphology into Automatic Speech Recognition. IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), 2009, pp. 354–358.
6) **Haşim Sak**, Tunga Güngör, and Murat Saraçlar. Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus. In GoTAL 2008, volume 5221 of LNCS, 2008, pages 417-427. Springer.
7) Tuncay Aksungurlu, Sıddıka Parlak, **Haşim Sak**, Murat Saraçlar. Turkçe Haber Programları için Dil Modelleme Yaklaşımlarının Karşılaştırılması. IEEE 16. Sinyal İşleme, İletişim ve Uygulamaları Konferansı (SİU). Didim, Türkiye, 2008.
8) Ebru Arısoy, **Haşim Sak**, and Murat Saraçlar. Language modeling for automatic Turkish broadcast news transcription. In Proceedings of Interspeech 2007 - Eurospeech, pp. 2381-2384, 2007.
9) **Haşim Sak**, Tunga Güngör, and Murat Saraçlar. Morphological disambiguation of Turkish text with perceptron algorithm. In CICLing 2007, volume LNCS 4394, pages 107-118, 2007.

Defense Jury Members

Assoc. Prof. Tunga Güngör	Bogazici University
Assoc. Prof. Murat Saraçlar	Bogazici University
Prof. Lale Akarun	Bogazici University
Prof. Fikret Gürgen	Bogazici University
Assist. Prof. Deniz Yüret	Koc University

Defense Date: 15.06.2011