Ebru ARISOY
Thesis Supervisor: Assist. Prof. Murat Saraclar

STATISTICAL AND DISCRIMINATIVE LANGUAGE MODELING FOR TURKISH LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Turkish, being an agglutinative language with rich morphology, presents challenges for Large Vocabulary Continuous Speech Recognition (LVCSR) systems. First, the agglutinative nature of Turkish leads to a high number of Out-of-Vocabulary (OOV) words which in turn lower Automatic Speech Recognition (ASR) accuracy. Second, Turkish has a relatively free word order that leads to non-robust language model estimates.

These challenges have been mostly handled by using meaningful segmentations of words, called sub-lexical units, in language modeling. However, a shortcoming of sub-lexical units is over-generation which needs to be dealt with for higher accuracies. This dissertation aims to address the challenges of Turkish in LVCSR. Grammatical and statistical sub-lexical units for language modeling are investigated and they yield substantial improvements over the word language models. Our novel approach inspired by dynamic vocabulary adaptation mostly recovers the errors caused by over-generation and further improves the accuracy of sub-lexical units. Additionally, discriminative language models (DLMs) with linguistically and statistically motivated features are utilized. DLM outperforms the conventional approaches, partly due to the improved parameter estimates with discriminative training and partly due to integrating the complex language characteristics of Turkish into language modeling.

The significance of this dissertation lies in being a comparative study of several sub-lexical units on the same LVCSR system, addressing the over-generation problem of sub-lexical units and extending sub-lexical-based generative language modeling of Turkish to discriminative language modeling. These approaches can be easily extended to other morphologically rich languages that suffer from similar problems.

PUBLICATIONS

Journals

1) **E. Arısoy,** M. Saraclar, B.Roark and I. Shafran, "Discriminative Language Modeling with Linguistic and Statistically Derived Features", *IEEE Transactions on Audio, Speech and Language Processing*, 2011 (accepted for publication).
2) **E. Arısoy**, D. Can, S. Parlak, H. Sak, M. Saraçlar, "Turkish Broadcast News Transcription and Retrieval", *IEEE Transactions on Audio, Speech and Language Processing*, 17(5):874-883, July 2009.
3) **E. Arısoy** and M. Saraçlar, "Dynamic Vocabulary Adaptation for LVCSR of Turkish", *IEEE Transactions on Audio, Speech and Language Processing*, 17(1):163-173, January 2009.
4) M. Creutz, T. Hirsimaki, M. Kurimo, A. Puurula, J. Pylkkonen, V. Siivola, M. Varjokallio, **E. Arısoy**, M. Saraçlar, and A. Stolcke, "Morph-Based Speech Recognition and Modeling of Out-of-Vocabulary Words Across Languages", *ACM Transactions on Speech and Language Processing*, 5.1 Article 3, December 2007.

Book Chapters

1) **E. Arısoy**, M. Kurimo, M. Saraçlar, T. Hirsimaki, J. Pylkkonen, T. Alumae, H. Sak, "Statistical Language Modeling for Automatic Speech Recognition of Agglutinative Languages", Speech Recognition, Technologies and Applications. Book edited by: France Mihelic and Janez Zibert. ISBN 978-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria.

International Conference Papers

1) **E. Arısoy**, M. Saraçlar, B. Roark and I. Shafran, "Syntactic and Sub-lexical Features for Turkish Discriminative Language Models", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.
2) **E. Arısoy**, T. Pellegrini, M. Saraçlar and L. Lamel, "Enhanced Morfessor Algorithm with Phonetic Features: Application to Turkish", in *Proceedings of the International Conference on Speech and Computer (SPECOM)*, St. Petersburg, Russia, 2009.
3) **E. Arısoy**, B. Roark, I. Shafran, M. Saraçar, "Discriminative N-gram Language Modeling for Turkish", in *Proceedings of Interspeech*, Brisbane, Australia, 2008.
4) **E. Arısoy**, H. Sak and M. Saraçlar, "Language Modeling for Automatic Turkish Broadcast News Transcription", in *Proceedings of Interspeech – Eurospeech*, Antwerp, Belgium, 2007.
5) M. Creutz, T. Hirsimaki, M. Kurimo, A. Puurula, J. Pylkkonen, V. Siivola, M. Varjokallio, **E. Arısoy**, M. Saraçlar, and A. Stolcke, "Analysis of Morph-Based Speech Recognition and the Modeling of Out-of-Vocabulary Words Across Languages", in *Proceedings of Human Language Technologies / The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Rochester, USA, 2007.
6) **E. Arısoy** and M. Saraçlar, "Lattice extension and rescoring based approaches for LVCSR of Turkish", in *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP)*, Pittsburgh, PA, USA, 2006.
7) M. Kurimo, M. Creutz, M. Varjokallio, **E. Arısoy**, and M. Saraçlar, "Unsupervised segmentation of words into morphemes – Morpho Challenge 2005: Applications to automatic speech recognition", in Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP), Pittsburgh, PA, USA, 2006.
8) M. Kurimo, M. Creutz, M. Varjokallio, **E. Arısoy** and M. Saraçlar, "Unsupervised segmentation of words into morphemes: An Introduction and Evaluation Report", *PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*, 2006.
9) M. Kurimo, A. Puurula, **E. Arısoy**, V. Siivola, T. Hirsimaki, J. Pylkkonen, T. Alumae, and M. Saraçlar, "Unlimited vocabulary speech recognition for agglutinative languages", in Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL), New York, 2006.

National Conference Papers

1) **15) E. Arısoy** and M. Saraçlar, "Türkçe GDSKT için Konuşma Tanıma Hatalarının Analizi", in *Proceedings of the IEEE 17. Sinyal İşleme ve İletişim Uygulamaları Konferansı (SİU),* Side, Antalya, Turkey, 2009.
2) **16)** E. Arısoy and M. Saraçlar, "Türkçe Haber Programları için Konuşma Tanıma", in Proceedings of the IEEE 15. Sinyal İşleme ve İletişim Uygulamaları Konferansı (SİU), Eskişehir, Turkey, 2007.

3) **17)** İ. Uzun, **E. Arısoy**, R. Edizkan and M. Saraçlar, "Dağıtık Yapıda Türkçe Sürekli Konuşma Tanıma Sisteminde Seyrek Paket Kayıplarının Analizi ve Telafisi", in Proceedings of the IEEE 15. Sinyal İşleme ve İletişim Uygulamaları Konferansı (SİU), Eskişehir, Turkey, 2007.
4) **18) E. Arısoy** and M. Saraçlar, "Geniş Dağarcıklı Konuşma Tanıma Sistemleri için Örünün Yeniden Değerlendirilmesi Tabanlı Dil Modellemesi Yaklaşımları", in Proceedings of the IEEE 14. Sinyal İşleme ve İletişim Uygulamaları Konferansı (SİU), Antalya, Turkey, 2006.

Awards from the Thesis

1) Bogazici University Research Fund (BAP) Ph.D. Thesis Award, 2010.
2) Serhat Ozyar Young Scientist Honor Award, 2010.

Defense Jury Members

Asst. Prof. Murat Saraclar	Bogazici University
Prof. Ethem Alpaydın	Bogazici University
Prof. Levent M. Arslan	Bogazici University
Asst. Prof. Hakan Erdogan	Sabanci University
Assoc. Prof. Mikko Kurimo	Helsinki University of Technology
Prof. Bulent Sankur	Bogazici University

Defense Date: 17.12.2009