Trends and challenges in language modeling for speech recognition and machine translation

Holger Schwenk, University of Le Mans

Abstract

Language models play an important role in large vocabulary continuous speech recognition (LVCSR) systems and statistical approaches to machine translation (SMT), in particular when modeling morphologically rich languages. Despite intensive research over more than 20 years, state-of-the-art LVCSR and SMT systems seem to use only one dominant approach: n-gram back-off language models. This talk first reviews the most important approaches to language modeling. I then discuss some of the recent trends and challenges for the future.

An interesting alternative to the back-off n-gram approach are the so-called continuous space methods. The basic idea is to perform the probability estimation in a continuous space. By these means better probability estimations of unseen word sequences can be expected. There is also a relative large body of works on adaptive language models. The adaptation can aim to tailor a language model to a particular task or domain, or it can be performed over time. Another very active research area are discriminative language models. Finally, I will review the challenges and benefits of language models trained an very large amounts of training material.