

Abstract

In open-domain language exploitation applications, a wide variety of topics with swift topic shifts has to be captured. Consequently, it is crucial to rapidly adapt all language components of a spoken language system. This thesis addresses unsupervised topic adaptation in both monolingual and crosslingual settings. For automatic speech recognition we rapidly adapt a language model on a source language. For statistical machine translation, we adapt a language model of a target language, a translation lexicon and a phrase table using a source text.

For monolingual adaptation, we propose latent Dirichlet-Tree allocation for Bayesian latent semantic analysis. Our model enables rapid incremental language model adaptation via caching the fractional topic counts of word hypotheses decoded from previous speech utterances. Latent Dirichlet-Tree allocation models topic correlation in a tree-based hierarchy and thus addresses the model initialization issue. To address the “bag-of-words” assumption in latent semantic analysis, we extend our approach to N-gram latent Dirichlet-Tree allocation. We investigate a fractional Kneser-Ney smoothing approach to handle fractional counts for topic models. The algorithm produces a more compact model compared to the Witten-Bell smoothing. Using multi-stage language model adaptation via N-gram latent Dirichlet-Tree allocation, we achieve significant reduction in speech recognition errors using our large-scale GALE systems on two different languages: Mandarin and Arabic. For end-to-end translation on speech inputs, applying topic adaptation on automatic speech recognition is beneficial to translation performance.

For crosslingual adaptation, we propose bilingual latent semantic analysis for statistical machine translation. A key feature of bilingual latent semantic analysis is a one-to-one topic correspondence between models of a source and a target language. Since topical information is language independent, our model enables transfer of a topic distribution inferred from a source text to a target language for crosslingual adaptation. Our approach has two advantages: first, it can be applied before translation, and thus has immediate impact on translation. Secondly, it does not rely on a translation output for adaptation, and therefore does not suffer from translation errors. Together with N-gram latent Dirichlet-Tree allocation on a target language, we achieve significant improvement in translation performance using our large-scale GALE systems for text translation.

A limitation of bilingual latent semantic analysis is the requirement of parallel corpora that are relative expensive to collect. We propose a semi-supervised approach to incorporate non-parallel documents into model training. We achieve improvement in crosslingual language model adaptation performance, especially when bilingual resources are deficient.