Comparison of Different Modeling Units for Language Model Adaptation for Inflected Languages

Tanel Alumäe

Institute of Cybernetics at Tallinn University of Technology,
Akadeemia tee 21, Tallinn, 12618, Estonia
tanel.alumae@phon.ioc.ee

Abstract. This paper presents a language model adaptation framework for highly inflected languages that use sub-word units as basic units in a language model for large vocabulary speech recognition. The proposed adaptation method uses latent semantic analysis based information retrieval to find documents similar to a tiny adaptation corpus. The approach enables to use different language units for modeling document similarity. The method is tested on an Estonian broadcast news transcription task. We compare words, lemmas and morphemes as basic units for similarity modeling. We observe a drop in speech recognition error rate after building adapted language model for each news story. Morpheme-based adaptation is found to give significantly larger improvement than word and lemma-based adaptation.

1 Introduction

Language model adaptation is a task of building a language model (LM) for speech recognition that is better suited for the given domain than a general background model, given a small adaptation corpus. In recent years, *latent semantic analysis* (LSA) has been successfully used for integrating long-term semantic dependencies into statistical language models [1]. The LSA-based approach gradually adapts the background language model based on the recognized words by boosting the unigram probabilities of semantically related words, using co-occurrence analysis of words and documents.

However, this approach cannot be efficiently directly used for highly inflective and/or agglutinative languages, such as Estonian, Finnish, Turkish, Korean and many others. In such languages, each word-phrase can occur in a large number of inflected forms, depending on its syntactic and semantic role in the sentence. In addition, many such languages are also so-called compounding languages, i.e., compound words can be formed from shorter particles to express complex concepts as single words. The compound words can again occur in different inflections. As a result, the lexical variety of such languages is very high and it is not possible to achieve a good vocabulary coverage when using words as basic units for language modeling. In order to increase coverage, subword units,