Guessers for Finite-State Transducer Lexicons

Krister Lindén

Department of General Linguistics, P.O. Box 9, FIN-00014 University of Helsinki
Krister.Linden@Helsinki.fi

Abstract. Language software applications encounter new words, e.g., acronyms, technical terminology, names or compounds of such words. In order to add new words to a lexicon, we need to indicate their inflectional paradigm. We present a new generally applicable method for creating an entry generator, i.e. a paradigm guesser, for finite-state transducer lexicons. As a guesser tends to produce numerous suggestions, it is important that the correct suggestions be among the first few candidates. We prove some formal properties of the method and evaluate it on Finnish, English and Swedish full-scale transducer lexicons. We use the open-source *Helsinki Finite-State Technology* [1] to create finite-state transducer lexicons from existing lexical resources and automatically derive guessers for unknown words. The method has a recall of 82-87 % and a precision of 71-76 % for the three test languages. The model needs no external corpus and can therefore serve as a baseline.

1 Introduction

New words and new usages of old words are constantly finding their way into daily language use. This is particularly prominent in rapidly developing domains such as biomedicine and technology. The new words are typically acronyms, technical terminology, loan words, names or compounds of such words. They are likely to be unknown by most hand-made morphological analyzers. In some applications, hand-made guessers are used for covering the low-frequency vocabulary or the strings are simply added as such.

Mikheev [2] and [16] noted that words unknown to the lexicon present a substantial problem to part-of-speech tagging and he presented a very effective supervised method for inducing a guesser from a lexicon and an independent training corpus. Oflazer & al. [3] presented an interactive method for learning morphologies and pointed out that an important issue in the wholesale acquisition of open-class items is that of determining which paradigm a given citation form belongs to.

Recently, unsupervised acquisition of morphologies from scratch has been studied as a general problem of morphology induction in order to automate the morphology building procedure. For overviews, see Wicentowski [4] and Goldsmith [5]. If we do not need a full analysis, but only wish to segment the words into morph-like units, we can use segmentation methods like Morfessor [6]. For a comparison of some recent successful segmentation methods, see the Morpho Challenge [7].

Although unsupervised methods have advantages for less-studied languages, for the well-established languages, we have access to fair amounts of lexical training ma-