Experiments in Cross-language Morphological Annotation Transfer

Anna Feldman, Jirka Hana, Chris Brew

Ohio State University
Department of Linguistics
Columbus, OH 43210-1298, USA

Abstract. Annotated corpora are valuable resources for NLP which are often costly to create. We introduce a method for transferring annotation from a morphologically annotated corpus of a source language to a target language. Our approach assumes only that an unannotated text corpus exists for the target language and a simple textbook which describes the basic morphological properties of that language is available. Our paper describes experiments with Polish, Czech, and Russian. However, the method is not tied in any way to these languages. In all the experiments we use the TnT tagger ([3]), a second-order Markov model. Our approach assumes that the information acquired about one language can be used for processing a related language. We have found out that even breathtakingly naive things (such as approximating the Russian transitions by Czech and/or Polish and approximating the Russian emissions by (manually/automatically derived) Czech cognates) can lead to a significant improvement of the tagger's performance.

1 Introduction

Genetically related languages posses a number of properties in common. For example, Czech and Russian are similar in many areas, including lexicon, morphology, and syntax (they have so-called free word-order). This paper explores the resemblances between Czech, Russian, and Polish, as well as exploits linguistic knowledge about these languages for automatic morpho-syntactic annotation without using parallel corpora or bilingual lexicons. Our experiments use these three languages; however, a broader goal of this work is to explore the general possibility of porting linguistic knowledge acquired in one language to another. This portability issue is especially relevant for minority languages with few resources.

Cross-language information transfer is not new; however, most of the existing work relies on parallel corpora (e.g. [7, 11, 12]) which are difficult to find, especially for lesser studied languages, including many Slavic languages. In our work, we explore a new avenue — We use a resource-rich language (e.g. Czech/Polish) to process a resource-poor genetically related language (e.g. Russian) without using a bilingual lexicon or a parallel corpus.

We tag Russian by combining information from a resource-light morphological analyzer ([5]) and information derived from Czech and Polish.