A translation-based approach to the learning of the morphology of an under-resourced language

Tewodros Abebe
Addis Ababa University, IT–PhD
Addis Ababa, Ethiopia
wolaytta.boditti@gmail.com

Michael Gasser
Indiana University, School of Informatics
Bloomington, Indiana, USA
gasser@indiana.edu

Morphological analysis and generation are essential to many natural language processing tasks. There are now a number of tools for developing finite-state transducers (FSTs), which can be run either as analysers or as generators, for languages that are well studied and increasingly sophisticated algorithms for the automatic learning of morphology for languages with sufficient data. However, for most languages, there are neither sufficient linguistic resources nor sufficient data. One way of creating computational resources for languages that do not have many is to start with the resources that exist for other, closely related languages and then to learn differences based on the limited data available (Pretorius and Bosch, 2009). In this project, we apply this general idea to the problem of morphology learning and implement it for the specific case of the languages Wolaytta and Gofa.

Wolaytta and Gofa are members of the poorly researched Omotic family, spoken in southwestern Ethiopia. Wolaytta is the most spoken and best studied of the roughly 30 Omotic languages, while the closely related language Gofa has very few resources of any sort.

Given a target language (TL) whose morphology we would like to learn, our approach starts with a related, “source” language (SL) whose morphology is known. We assume the availability (or development as part of the project) of an FST for the SL. We also assume a small set of bilinguals who are literate in both SL and TL and can provide the system with word–word translation pairs.

The system begins with the assumption that SL and TL have identical morphology; a copy of the FST for the SL is the initial state for the TL. Given a translation pair, the basic idea is to attempt to translate the SL word, using the current state of the system’s TL knowledge, and compare the result with the correct TL translation. Small differences between the predicted and correct TL words lead to learning: the TL FST is modified in some way. Possible updates to the FST include modifications to the form of roots or affixes, to morphotactics (the sequence of potential affixes), and to alternation rules, which are responsible for the morphophonological changes that may take place at the boundaries between morphemes.

The project began in spring 2016 and will terminate in spring 2018. The first author has received funding for the research from the IT PhD program of Addis Ababa University. The expected outcomes of the project, all free, open-source, and available on GitHub under a GNU GPL3.0 license, are: (1) morphological analyser/generators (FSTs) for Wolaytta and Gofa, (2) a toolkit for the learning of FSTs for under-resourced languages based on the known morphology of a related language and a set of translation pairs.

Efforts so far have focused on developing an FST for Wolaytta, using the Helsinki Finite-State Transducer toolkit (Lindén et al., 2009), on collecting a data set of translation pairs from Wolaytta–Gofa bilinguals, and on solving the basic task of isolating roots and affixes when one or the other of these differs in a translation pair.

We are currently focusing on the more complex tasks of learning differences in the order or number of affixes and learning modified or new alternation rules. In both cases, it may be necessary to constrain the search space with phonological biases, for example, towards alternation rules that implement assimilation.

References

Lindén, Krister, Silfverberg, Miikka, and Pirinen, Tommi. 2009. HFST Tool for Morphology: an Efficient Open-Source Package for Construction of Morphological Analyzers. *SFCM 2009.*

Pretorius, L., and Bosch, S. 2009. Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology. In *Proceedings of the First Workshop on Language Technologies for African Languages* (pp. 96-103).