A Resource-Light Approach to Morpho-Syntactic Tagging

Anna Feldman* and Jirka Hana‡
(*Montclair State University, ‡Charles University)

Amsterdam: Rodopi (Language and computers: Studies in practical linguistics, volume 70), 2010, xiv+185 pp; hardbound, ISBN 978-90-420-2768-8, €40.00

Reviewed by
Christian Monson
Oregon Health & Science University

Anna Feldman and Jirka Hana had a problem. Wanting to extract Russian verb frames, they lacked a tool for the necessary first step: morphological analysis of Russian words, disambiguated for context. To avoid the significant overhead of building a contextualized morphological analyzer from scratch, Feldman and Hana wondered if an analyzer that was already available for Czech would perform adequately on Russian.

This book is the culmination of five years' research on projecting to a target language a contextualized morphological analyzer that was built for a separate source language, when both source and target belong to the same language family (Slavic, Romance, etc.). The authors succeed at building competitive morphological analysis systems for the target languages they consider (Russian, Catalan, and Portuguese), while expending a minimum of effort to construct specialized resources for these targets. At the culmination of their book, in Chapter 7, Feldman and Hana report a 6% absolute improvement, 79.7% vs. 73.5% labeling accuracy, when using a Czech morphological analyzer projected to Russian as opposed to training a statistical analyzer directly on a small sample (1,758 words) of hand-annotated Russian.

Unfortunately missing is a formal demonstration that hand-labeling 1,758 words with morphological analyses requires an equivalent human effort to projecting an analyzer from one language to another. The authors' final morphological projection incorporates a variety of improvements that require human intervention: from a hand-built morphological guesser on the target language side, to hand-defined rules that identify cognates between source and target languages and that render the syntactic structure of the source language more similar to the target's structure. Nowhere do the authors report the person-hours required to build each of these components and the reader is left to trust that constructing the projected systems takes as little time as is implied.

A word of warning to those with a linguistics background: The authors prefer the language of natural language processing (NLP) to standard linguistic terminology. As a prime example, the title of this book includes the phrase *morpho-syntactic tagging*, a term from NLP. Part-of-speech tagging, in languages with little inflectional morphology, such as English, involves assigning to each word one part-of-speech tag from a small set of 50 or so possible tags. For the more inflected Slavic and Romance languages considered in this book, the tag sets include as many as 4,000 tags, each marking a full suite of morpho-syntactic features, such as *tense*, *case*, or *number*. Thus, in linguistic terms, morpho-syntactic tagging is exactly contextually disambiguated morphological analysis.

Feldman and Hana's book substantiates a number of useful insights beyond the core observation that projecting morphology to a target language may be both adequate and cheaper than starting from scratch. The authors show, for example, that even in languages with significant inflectional morphology, such as Czech, the vast majority of

lemmas occur in very few surface forms. In one corpus of 620K Czech words, 93% of all lemmas appear in four or fewer forms. This implies that a hand-compiled list of the most common surface forms of a set of lemmas will likely cover a significant portion of the words found in running Czech text.

Another noteworthy result demonstrated in this book is that training an HMM tagger to identify coarser sets of morphological labels does not always improve performance. Extensive experiments in Section 7.8 show that the Russian projected tagger attains higher performance on individual features, such as *number* or *gender*, when trained to identify all features together than when taught features in isolation. The authors suggest that the presence of multiple features in the training data informs the Markov structure of their HMM tagger.

At times, Feldman and Hana forget that tagging, or rather morphological analysis, is not an end in itself. In Chapter 4 the authors rationalize their decision to base the Russian morphological tag set on that already defined for Czech because doing so makes the task of projecting a morphological analyzer from Czech to Russian "easier" (Section 4.4.1: Russian). But ease of projection is a specious argument for the design of a tag set. The morpho-syntactic categories that a tag set contains should be exactly those categories that the language marks and that are relevant for the downstream natural language application. And indeed, elsewhere in Chapter 4 the authors say as much (Sections 4.3.3 and 4.4.2: Catalan).

Backing up for an overview: After briefly motivating cross-language projection of contextually disambiguated morphological analyzers in Chapter 1, the next two chapters place this book within the field of natural language processing. Chapters 2 and 3 review, at a high level, prior work in sequential tagging and, respectively, prior work in resource-light approaches to various NLP tasks. Chapters 4 and 5, together with appendices, give clear descriptions of the linguistic and distributional properties of the languages, corpora, and tag sets (i.e., morpho-syntactic feature sets) that are used in the authors' experiments. Chapter 6 describes a process for building high-recall/low-precision morphological guessers for target languages; Chapter 7 reports projection results; and Chapter 8 proposes future research directions.

Feldman and Hana's book is the definitive reference for their encouraging research on projecting morphological analysis systems across languages. The advice and findings of this book are directly valuable to anyone modeling the morphology of less-studied languages, and more broadly, this work should influence solutions to a wide range of NLP problems for languages with scarce labeled training data.

Christian Monson is a Senior Research Scientist at Nuance Communications Inc, and was previously a Post-Doctoral Research Fellow at Oregon Health & Science University. His unsupervised morphology induction algorithm, named ParaMor, placed first in four of five language tracks at the Morpho Challenge 2009 competition. Monson's e-mail address is `christian.monson@gmail.com`.