Challenges of Cheap Resource Creation for Morphological Tagging

Jirka Hana
Charles University
Prague, Czech Republic
first.last@gmail.com

Anna Feldman
Montclair State University
Montclair, New Jersey, USA
first.last@montclair.edu

Abstract

We describe the challenges of resource creation for a resource-light system for morphological tagging of fusional languages (Feldman and Hana, 2010). The constraints on resources (time, expertise, and money) introduce challenges that are not present in development of morphological tools and corpora in the usual, resource intensive way.

1 Introduction

Morphological analysis, tagging and lemmatization are essential for many Natural Language Processing (NLP) applications of both practical and theoretical nature. Modern taggers and analyzers are very accurate. However, the standard way to create them for a particular language requires substantial amount of expertise, time and money. A tagger is usually trained on a large corpus (around 100,000+ words) annotated with the correct tags. Morphological analyzers usually rely on large manually created lexicons. For example, the Czech analyzer (Hajič, 2004) uses a lexicon with 300,000+ entries. As a result, most of the world languages and dialects have no realistic prospect for morphological taggers or analyzers created in this way.

We have been developing a method for creating morphological taggers and analyzers of fusional languages[1] without the need for large-scale knowledge- and labor-intensive resources (Hana et al., 2004; Hana et al., 2006; Feldman and Hana, 2010) for the target language. Instead, we rely on (i) resources available for a related language and (ii) a limited amount of high-impact, low-cost manually created resources. This greatly reduces cost, time requirements and the need for (language-specific) linguistic expertise.

The focus of our paper is on the creation of resources for the system we developed. Even though we have reduced the manual resource creation to the minimum, we have encountered a number of problems, including training language annotators, documenting the reasoning behind the tagset design and morphological paradigms for a specific language as well as creating support tools to facilitate and speed up the manual work. While these problems are analogous to those that arise with standard resource creation, the approach to their solution is often different as we discuss in the following sections.

2 Resource-light Morphology

The details of our system are provided in (Feldman and Hana, 2010). Our main assumption is that a model for the target language can be approximated by language models from one or more related source languages and that inclusion of a limited amount of high-impact and/or low-cost manual resources is greatly beneficial and desirable.

We use TnT (Brants, 2000), a second order Markov Model tagger. We approximate the target-language emissions by combining the emissions from the (modified) source language corpus with information from the output of our resource-light analyzer (Hana, 2008). The target-language transitions are approximated by the source language (Feldman and Hana, 2010).

3 Resource creation

In this section we address the problem of collection, selection and creation of resources needed by our system. The following resources must be available:

- a reference grammar book for information

[1] Fusional languages are languages in which several feature values are realized in one morpheme. For example Indo-European languages, including Czech, German, Romanian and Farsi, are predominantly fusional.

about paradigms and closed class words,

- a large amount of plain text for learning a lexicon, e.g. newspapers from the Internet,
- a large annotated training corpus of a related language,
- optionally, a dictionary (or a native speaker) to provide analyses of the most frequent words,
- a non-expert (not a linguist and not a native speaker) to create the resources listed below,
- limited access to a linguist (to make non-obvious decisions in the design of the resources),
- limited access to a native speaker (to annotate a development corpus, to answer a limited number of language specific questions).

and these resources must be created:

- a list of morphological paradigms,
- a list of closed class words with their analyses,
- optionally, a list of the most frequent forms,
- a small annotated development corpus.

For evaluation, an annotated test corpus must be also created. As this corpus is not part of the resource-light system per se, it can (and should) be as large as possible.

3.1 Restrictions

Since our goal is to create resources cheaply and fast, we intentionally limit (but not completely exclude) the inclusion of any linguist and of anybody knowing the target language. We also limit the time of training and encoding of the basic target-language linguistic information to a minimum.

3.2 Tagset

In traditional settings, a tagset is usually designed by a linguist, moreover a native speaker. The constraints of a resource-light system preclude both of these qualifications. Instead, we have standardized the process as much as possible to make it possible to have the tagset designed by a non-expert.

3.2.1 Positional Tagset

All languages we work with are morphologically rich. Naturally, such languages require a large number of tags to capture their morphological properties. An obvious way to make it manageable is to use a structured system. In such a system, a tag is a composition of tags each coming from a much smaller and simpler atomic tagset tagging a particular morpho-syntactic property (e.g. gender or tense). This system has many benefits, including the 1) relative easiness for a human annotator to remember individual positions rather than several thousands of atomic symbols; 2) systematic morphological description; 3) tag decomposability; and 4) systematic evaluation.

3.2.2 Tagset Design: Procedure

Instead of starting from scratch each time a tagset for a new language is created, we have provided an annotated tagset template. A particular tagset can deviate from this template, but only if there is a linguistic reason. The tagset template includes the following items:

- order of categories (POS, SubPOS, gender, animacy, number, case, ...) – not all might be present in that language; additional categories might be needed;
- values for each category (N – nouns, C – numerals, M – masculine);
- which categories we do not distinguish, even though we could (proper vs. common nouns);
- a fully worked out commented example (as mentioned above).

Such a template not only provides a general guidance, but also saves a lot of time, because many of rather arbitrary decisions involved in any tagset creation are done just once (e.g. symbols denoting basic POS categories, should numerals be included as separate POS, etc.). As stated, a tagset may deviate from such a template, but only if there is a specific reason for it.

3.3 Resources for the morphological analyzer

Our morphological analyzer relies on a small set of morphological paradigms and a list of closed class and/or most frequent words.

3.3.1 Morphological paradigms

For each target language, we create a list of morphological paradigms. We just encode basic facts about the target language morphology from a standard grammar textbook. On average, the basic morphology of highly inflected languages, such as Slavic languages, are captured in 70-80 paradigms. The choices on what to cover involve a balance between precision, coverage and effort.

3.3.2 A list of frequent forms

Entering a lexicon entry is very costly, both in terms of time and knowledge needed. While it is usually easy (for a native speaker) to assign a word to one of the major paradigm groups, it takes considerably more time to select the exact paradigm variant differing only in one or two forms (in fact, this may be even idiolect-dependent). For example, in Czech, it is easy to see that the word *atom* 'atom' does not decline according to the neuter paradigm *město* 'town', but it takes more time to decide to which of the hard masculine inanimate paradigms it belongs. On the other hand, entering possible analyses for individual word forms is usually very straightforward. Therefore, our system uses a list of manually provided analyses for the most common forms.

Note that the process of providing the list of forms is not completely manual – the correct analyses are selected from those suggested on the basis of the words' endings. This can be done relatively quickly by a native speaker or by a non-native speaker with the help of a basic grammar book and a dictionary.

3.4 Documentation

Since the main idea of the project is to create resources quickly for an arbitrarily selected fusional language, we cannot possibly create annotation and language encoding manuals for each language. So, we created a manual that explains the annotation and paradigm encoding procedure in general and describes the main attributes and possible values that a language consultant needs to consider when working on a specific language. The manual has five parts:

1. How to summarize the basic facts about the morphosyntax of a language;
2. How to create a tagset
3. How to encode morphosyntactic properties of the target language in paradigms;
4. How to create a list of closed class words.
5. Corpus annotation manual

The instructions are mostly language independent (with some bias toward Indo-European languages), but contain a lot of examples from languages we have processed so far. These include suggestions how to analyze personal pronouns, what to do with clitics or numerals.

3.5 Procedure

The resource creation procedure involves at least two people: a native speaker who can annotate a development corpus, and a non-native speaker who is responsible for the tagset design, morphological paradigms, and a list of closed class words or frequent forms. Below we describe our procedure in more detail.

3.5.1 Tagset and MA resources creation

We have realized that even though we do not need a native speaker, some understanding of at least basic morphological categories the language uses is helpful. So, based on our experience, it is better to hire a person who speaks (natively or not) a language with some features in common. For example, for Polish, somebody knowing Russian is ideal, but even somebody speaking German (it has genders and cases) is much better than a person speaking only English. In addition, a person who had created resources for one language performs much better on the next target language. Knowledge comes with practice.

The order of work is as follows:

1. The annotator is given basic training that usually includes the following: 1) brief explanation of the purpose of the project; 2) tagset design; 3) paradigm creation.
2. The annotator summarizes the basic facts about the morphosyntax of a language,
3. The first version of the tagset is created.
4. The list of paradigms and closed-class words is compiled. During this process, the tagset is further adjusted.

3.5.2 Corpus annotation

The annotators do not annotate from scratch. We first run our morphological analyzer on the selected corpus; the annotators then disambiguate the output. We have created a support tool (`http://ufal.mff.cuni.cz/~hana/law.html`) that displays the word to be annotated, its context, the lemma and possible tags suggested by the morphological analyzer. There is an option to insert a new lemma and a new tag if none of the suggested items is suitable. The tags are displayed together with their natural language translation.

4 Case studies

Our case studies include Russian via Czech, Russian via Polish, Russian via Czech and Polish, Portuguese via Spanish, and Catalan via Spanish.

We use these languages to test our hypotheses and we do not suggest that morphological tagging of these languages should be designed in the way we do. Actually, high precision systems that use manually created resources already exist for these languages. The main reason for working with them is that we can easily evaluate our system on existing corpora.

We experimented with the direct transfer of transition probabilities, cognates, modifying transitions to make them more target-like, training a battery of subtaggers and combining the results (Reference omitted). Our best result on Russian is 81.3% precision (on the full 15-slot tag, on all POSs), and 92.2% (on the detailed POS). We have also noticed that the most difficult categories are nouns and adjectives. If we improve on these individual categories, we will improve significantly the overall result. The precision of our model on Catalan is 87.1% and 91.1% on the full tag and SubPOS, respectively. The Portuguese performance is comparable as well.

The resources our experiments have relied upon include the following:

1. Russian
    - Tagset, paradigms, word-list: speaker of Czech and linguist, some knowledge of Russian
    - Dev corpus: a native speaker & linguist
2. Catalan
    - Tagset: modified existing tagset (designed by native speaking linguists)
    - paradigms, word-list: linguist speaking Russian and English
    - Dev corpus: a native speaking linguists
3. Portuguese
    - Tagset: modified Spanish tagset (designed by native speaking linguists) by us
    - paradigms, word-list: a native speaking linguist
    - Dev corpus: a native speaking linguist
4. Romanian
    - Tagset, paradigms, word-list: designed by a non-linguist, speaker of English
    - Dev corpus – a native speaker

Naturally, we cannot expect the tagging accuracy to be 100%. There are many factors that contribute to the performance of the model:

1. target language morphosyntactic complexity,
2. source-language–target-language proximity,
3. quality of the paradigms,
4. quality of the cognate pairs (that are used for approximating emissions),
5. time spent on language analysis,
6. expertise of language consultants,
7. supporting tools.

5 Summary

We have described challenges of resource creation for resource-light morphological tagging. These include creating clear guidelines for tagset design that can be reusable for an arbitrarily selected language; precise formatting instructions; providing basic linguistic training with the emphasis on morphosyntactic properties of fusional languages; creating an annotation support tool; and giving timely and constructive feedback on intermediate results.

6 Acknowledgement

The development of the tagset was supported by the GAČR grant P406/10/P328 and by the U.S. NSF grant # 0916280.

References

Thorsten Brants. 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of 6th Applied Natural Language Processing Conference and North American chapter of the Association for Computational Linguistics annual meeting (ANLP-NAACL)*, pages 224–231.

Anna Feldman and Jirka Hana. 2010. *A Resource-light Approach to Morpho-syntactic Tagging*, volume 70 of *Language and Computers: Studies in Practical Linguistics*. Rodopi, Amsterdam/New York.

Jan Hajič. 2004. *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Karolinum, Charles University Press, Prague, Czech Republic.

Jirka Hana, Anna Feldman, and Chris Brew. 2004. A Resource-light Approach to Russian Morphology: Tagging Russian Using Czech Resources. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 222–229, Barcelona, Spain.

Jirka Hana, Anna Feldman, Luiz Amaral, and Chris Brew. 2006. Tagging Portuguese with a Spanish Tagger Using Cognates. In *Proceedings of the Workshop on Cross-language Knowledge Induction hosted in conjunction with the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 33–40, Trento, Italy.

Jirka Hana. 2008. Knowledge- and labor-light morphological analysis. *OSUWPL*, 58:52–84.