Morphosyntactic Annotation and Lemmatization Based on the Finite-State Dictionary of Wordformation Elements

Brian O'Donovan and Alexander Troussov

IBM Dublin Software Lab, Airways Ind. Est., Cloghran, Dublin 17, Ireland

Abstract: Dictionary-based methods in morphological analysis can provide accurate lemmatization and rich annotation, including part-of-speech, number, gender, etc. A morphological guesser can be used to process out-of-vocabulary words. Industrial text processing applications require high performance, which suggests the need to merge these two types of applications. In this paper we discuss the conversion of a pre-existing high coverage morphosyntactic lexicon into a deterministic finite-state device which: preserves accurate lemmatization and annotation for vocabulary words, allows acquisition and exploitation of implicit morphological knowledge from the dictionaries in the form of ending guessing rules to process out-of-vocabulary words, allows seamless integration of additional hand-crafted ending guessing rules.

INTRODUCTION

Dictionary-based methods in morphological analysis are able to provide accurate lemmatization and rich annotation, including part-of-speech, number, gender, etc. Morphological guesser can be used to process out-of-vocabulary words. Industrial applications used in automatic text processing require high performance, which suggests the need of merging these two types of applications. In this paper we discuss conversion of a pre-existing high coverage morphosyntactic lexicon into a deterministic finite-state device which:

- preserves accurate lemmatization and annotation for vocabulary words;
- allows acquisition and exploitation of implicit morphological knowledge from the dictionaries in the form of ending guessing rules to process out-of-vocabulary words (following the approach proposed by Jan Daciuk);
- allows seamless integration of additional hand-crafted ending guessing rules.

DICTIONARY

Morphological analysis, considered as the mapping of surface forms into normalized forms (lemmatization) with morphosyntactic annotation for surface forms (part-of-speech tagging), can be efficiently implemented for many languages as a simple lookup-mechanism in an exhaustive full-form morphosyntactic lexicon. This approach obviates the need for computationally expensive run-time morphological analysis. Different language models, including two-level morphology, can be used for the development of full-form lexicons.

Logically the full-form lexicon consists of lexical entries: keyword (surface form), associated glosses, which can include morphosyntactic information, morphotactics, and lemma (which can be represented by "cut-and-paste" code). Such dictionaries can be efficiently implemented as acyclic finite-state automata, also known as Directed Acyclic Word Graphs (DAWG). The input list is compiled into a letter tree, which is then minimized to reuse common postfixes. Glosses can be attached to the leaves (the terminals) of the letter tree. In this case two postfixes can be merged only if they lead to exactly the same information. Finite-state automata constructed this way are acyclic and deterministic (for each state and each character there can be only one or zero output links labeled by this character).

Among the advantages of using DAWGs for morphological analysis are excellent performance and their amenability for customization on the fly, including adding of new entries and modifying glosses.

DECOMPOUNDING AND CLITIC PROCESSING

The need for treating such linguistic phenomena as compounding and clitics shows the limitation of straightforward usage of full-form lexicon dictionary look-up method in crosslinguistic lexical analysis. Additional consideration is the need for keyword extraction in information retrieval, which is especially acute in so-called compounding languages like German, Dutch, Norwegian, Danish, Swedish and Greek, in which complex concepts are often expressed as single orthographic words.

The remedy, which allows us to stay inside the full-form lexicon paradigm, is that the recognition lexicon must be augmented with linguistically meaningful wordformation units. To allow for efficient interface between tokenization and the morphological analyzer one could place additional prefabricated units related to constituents of orthographic words in the same lexicon of surface forms, provided with constraints indicating their usage in morphological and lexical analysis. This assumes, that, for example, in case of the German word *Schule* all the alternants *Schule, Schul-* must be explicitly listed in the lexicon provided with constraints in the form of morphotactics needed to discriminate which of these forms must be used in particular case of decompounding. Specifically, the allomorph *Schul-* must be used in German compounds if followed by some other morpheme (e.g. *Schulrat*), while the morpheme *schule* is used in the final position of a compound (e.g. *Hochschule*).

The use of prefabricated units allows us to reduce the morphological analysis of out-of-vocabulary words as decomposition into non-overlapping non-gapped constituents satisfying constraints. For example, German *Hochschullehrer* can be analysed as *Hoch+schul+lehrer* or *Hochschul+lehrer*.

Such "mechanical decompounding" can be efficiently implemented based on the dictionary look-up in the full-form lexicon. The key observation is that one dictionary look-up in DAWG allows us to extract all prefixes of the input string together with their glosses presented in the lexicon. For example, one look-up for *Hochschullehrer* could produce all of the prefixes *Hoch, Hochschul, Hochschullehrer*. To continue segmentation, one must strip a found prefix and continue by recursion. Additional checks for validity of constraints can be easily fit into this scheme. For example, the recursion branch could be terminated if the prefix *schule* is encountered without reaching the end of the analysed form.

A commonly used alternative to this computational method, based on a single lexicon compiled as DAWG, is to use a non-deterministic finite-state device. However, this will eliminate the afore mentioned advantages of DAWGs. In addition, the simple programming logic used in our approach is not necessarily slower than backtracking in non-deterministic devices, but allows integration of additional constraints. For example, the final constituent of the compound *Hochschullehrer* is a noun, which requires that the compound must be also a noun. Another example: the verbalizing prefix *en-* allows for the correct part-of-speech annotation of English verb *encentre*.

PROCESSING OF OUT-OF-VOCABULARY WORDS BASED ON ENDING SEGMENTS ONLY

When out-of-vocabulary words can't be segmented into valid sequences of wordformation constituents, the lemmatization and annotation must be taken based on

some found constituents. Our morphological analyser tries to identify the unknown form by computing its potential linguistic characteristics including its canonical form(s) from its "head". This notion was introduced in morphology by Williams (1981) to account for the fact that a complex word shares most, if not all, properties with one of its constituents. The head of a word is typically either the rightmost or the leftmost morpheme of a word. In most commercially important languages it is primarily suffixes and the rightmost members of compounds which determine the properties (e.g. part of speech) of morphologically complex words.

The validity of this approach is based on the so-called Righthand Head Rule (see DiSciullo and Williams 1987). The generalization of this rule has problems, including the following systematic exceptions: category-changing prefixation in Romance and Germanic languages (English *en-*), regular compounding in Romance (French *essuie-glace* (wipe-windshield, windshield-wiper), circumfixation in Germanic languages. In this paper we focus only on the computationally efficient usage of the righthand part of unknown orthographic words based on the dictionary of wordformation elements.

For illustration purposes mainly, we provide several examples of "endstrings" for English. These endstrings are not necessarily limited by morphological boundaries. Most of the examples of out-of-vocabulary words are borrowed from www.rdues.liv.ac.uk/newwords.shtml.

Suffixes which are derived historically from Latin or Greek (such as *-ation*, *-ition*, *-ission*, etc.) are especially important for processing technical and scientific texts.This will allow the processing of such words as *Clerkenwellisation, contemporisation, fixturisation.* Words like *allgations* (probably, misspelled *allegations*) and *optoelctronic* (probably, misspelled *optoelectronic*) can be reliably annotated based on their Latinate endstrings. Neologism *Canadianisms* ("Some say it's English but listen out for those tell-tale Canadianisms") can be identified as abstract noun with the lemma *Canadianism*. For specific domains, general English suffixes must be augmented, for example with "biomedical" suffixes such as *-ase*, *-in* , *-yl* or *–ergic* (concatenation of basic suffixes *-erg-* and *-ic*).

Proper names are typically capitalized, but morphological analysis is useful to avoid reliance on this brittle feature only (titles, which are of special importance for information retrieval, are often written in all capital letters).

Endstrings like *-abad* (Astrabad, Leninabad, Ashkhabad), *-polis (*Argiroupolis), *-stan* (Afghanistan, Tatarstan, Tadzhikistan, Turkestan) indicate names of locations. Endstrings like *-shvili* and *-gusson* indicate family names.

As an example of languages other than English: the Spanish suffix *-mente* always indicates an adverb, and the suffix *-ción* always indicates a feminine singular noun.

The crosslinguistic application of the endstring method requires linguistic expertise. Even if some generic rules can be easily expressed in the form of the endstrings, practical implementation encounters severe problems with exceptional words. The solution proposed in this paper is to fully merge handcrafted endstrings with the full-form lexicon. The possibility of such a syncretism hinges on the fact, that DAWG representation of full-form lexicons is amenable for unsupervised extraction of morphological knowledge exactly in the form of ending segments (see Jan Daciuk 1998). As we mentioned earlier, full-form lexicons can be created based on different language models, and morphological knowledge used for the creation of language model is not explicitly presented in the result full-form lexicon.

IMPLEMENTATION OF ENDING SEGMENTS METHOD
COMBINING EXPLICIT AND IMPLICIT MORPHOLOGICAL KNOWLEDGE

By automatic analysis of e.g. English dictionary it is possible to discover that most of the words ending with *–ism* are nouns with plural form *–isms*. One can postulate, for example, that all out-of-vocabulary words ending with *–ism* are abstract nouns. Dictionary look-up can handle concrete nouns like *organism* or *prism*, and their prefixed forms like *microorganism* and *pentaprism*, while postulated knowledge will classify *Lamarkianism* and *vegetarianism* as abstract nouns.

The dictionaries under consideration are lists of words and wordformation elements provided with glosses: morphosyntactic annotation, morphotactic constraints, cut-and-paste code for lemmatization. This full-form lexicon is comprised from three sources:

- valid orthographic words;
- wordformation elements used for segmentation of solid compounds;
- endstrings used exclusively for "guessing".

Such a dictionary compiled into DAWG can be used for dictionary look-up and for word segmentation. The same dictionary can be used also for "guessing", but to facilitate the search of endstrings in an input word, another DAWG based on the same data can be created. In this finite-state device, which is called the reversed dictionary, each string representing a dictionary word is stored in the reversed sequence. For example, in a conventional dictionary the word *tables* is stored as the string "tables" while in the reversed dictionary it is stored as "selbat". To look-up a word in a reversed dictionary one must correspondingly reverse the order of the characters in the word.

The reversed lexicon can be used in combination with a conventional lexicon, or used as a stand-alone lemmatizer and word level tagger.

- Reversed dictionary entries are compiled into DAWG;
- Original glosses are attached to final states, which correspond to beginnings of words;
- The DAWG is analyzed to compute "probable" glosses for some intermediate nodes;
- To reduce the size DAWG can be pruned: if from a given state all paths lead to the same set of glosses, than all states between that state and the annotations can be removed with all their transitions.
- Run-time algorithm reverses analysed surface form and the consecutive characters are looked up in the DAWG; decision about gloss assignment based on all glosses collected during matching or on the longest match.

LITERATURE

Daciuk, J. *Incremental construction of finite-state automata and transducers, and their use in the natural language processing*. Rozprawa doktorska ETI. 1998.

Di Sciullo, A.-M. and Williams, E. *On the Definition of Word*. Cambridge, MA: MIT Press. 1987.

Williams, E. 1983. On the notions 'lexically related' and 'head of a word'. *Linguistic Inquiry,* 12:245-274, (1981).