Assigning an Inflectional Paradigm
using the Longest Matching Affix

Krister Lindén

Department of General Linguistics, P.O. Box 9, FIN-00014 University of Helsinki
Krister.Linden@Helsinki.fi

Abstract. Language software applications encounter new words, e.g., acronyms, technical terminology, loan words, names or compounds of such words. In English new compounds are formed as separate words or using a hyphen, but in other Germanic languages and in Fenno-Ugric languages there is no special indicator between the parts of a compound word. In order to add new words to a lexicon, we need to indicate the inflectional paradigm. In this article, a method for assigning the inflectional paradigm to new compound base forms based on the longest matching affix is evaluated. The method is 99.75 % accurate for Finnish compound base forms.

1 Introduction

New words and new usages of old words are constantly finding their way into daily language use. This is particularly prominent in rapidly developing domains such as biomedicine and technology. Humans deal with new words based on previous experience: we treat them by analogy to known words. The new words are typically acronyms, technical terminology, loan words, names or compounds of such words. They are likely to be unknown by most hand-made morphological analyzers. In some applications, hand-made guessers are used for covering this low-frequency vocabulary or the strings are simply added as such.

Unsupervised acquisition of morphologies from scratch has been studied as a general problem of morphology induction in order to automate the morphology building procedure. For overviews, see Wicentowski (2002) and Goldsmith (2007). If we do not need a full analysis, but only wish to segment the words into morph-like units, we can use segmentation methods like Morfessor (Creutz & al, 2005). For a comparison of some recent successful segmentation methods, see the Morpho Challenge (Kurimo & al, 2007).

Unsupervised methods have advantages for less-studied languages, but for the well-established languages, we have access to fair amounts of training material in the form of analyzes in context of more frequent words. In addition, there are a host of large but shallow hand-made morphological descriptions available, e.g., the Ispell collection of

dictionaries (Kuening, 2007) for spell-checking purposes, and many well-documented morphological analyzers are commercially available, e.g., from Lingsoft, Inc (2007).

Especially for Germanic and Fenno-Ugric languages, there are already large-vocabulary descriptions available and new words tend to be compounds of existing words. In English, the words are written separately or the junction is indicated with a hyphen, but in other Germanic languages and in the Fenno-Ugric languages, there is usually no word boundary indicator within the compounds.

We propose and evaluate a method for automatically classifying base forms of new compound words according to their inflectional paradigm. In Section 2, we describe the methodology. In Section 3, we present the training and test data for Finnish. In Section 4, we evaluate the model and show that it works with 99.75 % accuracy. In Section 5, we discuss the method and the test results, and propose an extension.

2 Methodology

Assume that we have a set of base forms that have been classified according to their inflectional paradigm. We also have another set of previously unseen compound words in base form for which we wish to determine their inflectional paradigm by analogy with the classified base forms. Also assume that the language mainly attaches its inflections to one end of the word, i.e., the language is an affixing language. We call the set of classified base forms an affix list. We then classify the compound base forms according to the paradigm of the longest matching affix from the affix list.

3 Data Sets

In order to test our method, we used the Finnish word list *Nykysuomen sanalista* (2007), which contains 94 110 words. Of these, approximately 43 000 are classified non-compound base forms and 51 000 are unclassified compound base forms.

The non-compound base forms consist of approximately 30 000 nouns, adjectives and verbs, which are classified into 76 different inflectional paradigms. In addition, the inflectional paradigms can be subdivided into 12 different stem change categories. There are also two broad categories for the remaining 13 000 non-compound words. All in all, there is a theoretical set of 914 classes. We use these non-compound words as our training material.

From the set of compound base forms, we removed a non-productive category, where both parts of the word inflect. These were 93 base forms. We also removed 525 words

that only occur in the plural form. The remaining 49 537 compound base forms, we used as our test material.

4 Experiments

We tested how well the classifier of compound base forms is able to predict the paradigm for a base form using the test data mentioned in Section 3. Of the 49 537 compound base forms only 122 received an incorrect classification using the longest matching suffix, i.e. 99.75 % were correctly classified. If we apply stricter criteria and demand that word boundaries are also correct, no more than 303 words had incorrect word boundaries, i.e. less than 0.6 % incorrect words boundaries

5 Discussion

In this section, we discuss the test results and give some final notes on the nature of Finnish morphology and the implementation of the method.

The test material may be slightly skewed in that it contains only well-established words of the Finnish language. In addition, the training material is guaranteed to have at least one matching suffix for the final part of each compound in the test material. In this sense, the experiment can be seen to give an indication of the upper bound for classification using longest suffix matching on new base forms.

When studying the few words for which the classification failed, we see that the most frequent misclassification (23/122) is based on the word *oppi 5B* vs. *soppi 7B*, as in *kasvatus/oppi, lujuus/oppi, perus/oppi.* Another fairly frequent mistake (16/122) is based on compounds being misclassified into the non-inflecting category *99*, e.g. *edustus/elin 33 vs. edustu/selin 99*. From this we see, that restricting the first part of the word to a complete word would have removed some of the mistakes. In order to require that the first part is a word in some form, we would have needed a fairly extensive list of possible prefixes consisting of word forms from a corpus and some morphological model.

We also observe that the misclassifications arise when the last few characters of the compound prefix together with the final part of the compound create a longer but viable word. This is not so common in Finnish due to a fairly restricted phonological structure of words, i.e., there are only a fairly restricted set of sounds at the end of words. Of these word final sounds, only *s* seems to interact in a more systematic way: e.g., *soppi* vs. *oppi*, *soikeus* vs. *oikeus*, *saukko* vs. *aukko, etc* are existing words. However, most often the two alternatives both belong to the same paradigm, so the fact that the word boundary is incorrect does not necessarily affect the classification. In other languages, the phonology is less strict and run-on words become a more prominent factor demanding a stricter control of the prefixing word forms.

The model for finding the classification of a base form was implemented with a cascade of weighted finite-state transducers. The cascade decomposes the base form into a prefix and a classified suffix. The prefix is weighted according to its length. To classify the compound base form, we use the decomposition with the smallest weight. Open Source tools for weighted finite-state transducers have been implemented by, e.g., Allauzen & al (to appear) and Lombardy & al (2004).

6. Conclusion

We have tested a simple but effective model for classifying base forms of new compounds by analogy with a set of non-compound base forms. The experiment assumed that the inflections are encoded as suffixes, but the idea is easily modified for prefixing languages. We tested the model on Finnish, which is a highly inflecting language with a considerable set of inflectional paradigms and stem change categories. Our model reached a recall of 100 % with an accuracy of 99.75 %. The error analysis indicates that the method is easily extendible to other languages. However, a prefixing mechanism may be required to deal with compound word boundaries that are a potential problem in languages with a less strict phonological structure.

References

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W. & Mohri, M. (to appear). 'OpenFst : A General and Efficient Weighted Finite-State Transducer Library', *Lecture Notes in Computer Science*.

Creutz, M., Lagus, K., Lindén, K. & Virpioja, S. 2005, 'Morfessor and Hutmegs : Unsupervised Morpheme Segmentation for Highly-Inflecting and Compounding Languages', *Proceedings of the Second Baltic Conference on Human Language Technologies*, Tallinn, Estonia.

Goldsmith, J. 2007, 'Morphological Analogy : Only a Beginning',
    Available at: http://hum.uchicago.edu/~jagoldsm/Papers/analogy.pdf

Kuenning, G. 2007, 'Dictionaries for International Ispell',
    Available at: http://www.lasr.cs.ucla.edu/geoff/ispell-dictionaries.html

Kurimo, M., Creutz, M. & Turunen, V. 2007, 'Overview of Morpho Challenge in CLEF 2007', *Working Notes of the CLEF 2007 Workshop*, pp. 19-21

Lingsoft, Inc. 2007, *Demos*
    Available at: http://www.lingsoft.fi/?doc_id=107&lang=en

Lombardy, S., Régis-Gianas,Y. & Sakarovitch, J. 2004. 'Introducing Vaucanson', *Theoretical Computer Science,* vol. 328, pp. 77 – 96.

Nykysuomen sanalista (computer file) 2007,
    Available at: http://kaino.kotus.fi/sanat/ nykysuomi/

Wicentowski, R. 2002, 'Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework', *PhD Thesis*, Baltimore, USA.