MORFESSOR AND HUTMEGS: UNSUPERVISED MORPHEME SEGMENTATION FOR HIGHLY-INFLECTING AND COMPOUNDING LANGUAGES

Mathias Creutz[1], **Krista Lagus**[1], **Krister Lindén**[1,2], **Sami Virpioja**[1]
[1] Helsinki University of Technology (Finland)
[2] University of Helsinki (Finland)

Abstract

In this work, we announce the *Morfessor* 1.0 software package, which is a program that takes as input a corpus of raw text and produces a segmentation of the word forms observed in the text. The segmentation obtained often resembles a linguistic morpheme segmentation. In addition, we briefly describe the *Hutmegs* package, also publicly available for research purposes. Hutmegs contains semi-automatically produced correct, or gold-standard, morpheme segmentations for a large number of Finnish and English word forms. One easy way for the reader to familiarize himself with our work is to test the *demonstration* program on our Internet site. The demo shows how Morfessor segments words that the user types in.

Keywords: unsupervised morpheme segmentation, morphology discovery and induction, language-independent, gold-standard, public resources, demo, Finnish, English

1. Introduction

In the theory of linguistic morphology, morphemes are considered to be the smallest meaning-bearing elements of language. Any word form can be expressed as a combination of morphemes, as for instance the following English words: 'arrange+ment+s, foot+print, mathematic+ian+'s, un+fail+ing+ly'.

It seems that automated morphological analysis would be beneficial for many natural language applications dealing with large vocabularies, such as speech recognition and machine translation. Many existing applications make use of words as vocabulary units. However, for highly-inflecting languages, e.g., Finnish, Turkish, and Estonian, this is infeasible, as the number of possible word forms is very high. The same applies (possibly less drastically) to compounding languages, e.g., German, Swedish, and Greek.

There exist morphological analyzers designed by experts for some languages (e.g., based on the two-level morphology methodology). However, expert knowledge and labor are expensive. Analyzers must be built separately for each language, and the analyzers must be updated on a continuous basis in order to cope with language change (mainly the emergence of new words and their inflections).

As an alternative to the hand-made systems there exist algorithms that work in an unsupervised manner and autonomously discover morpheme segmentations for the words in unannotated text corpora. *Morfessor* is a general model for the unsupervised induction of a simple morphology from raw text data. Morfessor has been designed to cope with languages having predominantly a concatenative morphology and where the number of morphemes per word can vary much and is not known in advance. This distinguishes Morfessor from resembling models, e.g., (Goldsmith 2001), which assume that words consist of one stem possibly followed by a suffix and possibly preceded by a prefix.

In this work, we present the publicly available Morfessor 1.0 software package. The program segments the word forms in its input into morpheme-like units (see Section 2). For evaluating the segmentation produced by Morfessor or some other segmentation algorithm, we provide the *Hutmegs* package. Hutmegs contains linguistic morpheme segmentations for a large number of Finnish and English word forms, as well as a set of tools for comparing the segmentation proposed by the splitting algorithm (e.g., Morfessor) to the correct segmentation of Hutmegs (see Section 3). The Morfessor and Hutmegs package are available on our Internet web page (`http://www.cis.hut.fi/projects/morpho`) together with an online demonstration program.

2. Morpheme segmentation with Morfessor

Morfessor is a general model framework for unsupervised morphology discovery. It takes as input an unannotated text corpus and produces a segmentation of every word in the corpus. We call the proposed segments *morphs*. The boundaries between the discovered morphs often coincide with linguistic morpheme boundaries.

The Morfessor method works in an unsupervised manner, which means that no linguistic knowledge is preprogrammed into it, except for some very general assumptions about model structure. By observing the language data alone Morfessor comes up with a model that captures regularities within the set of observed word forms. The underlying idea is to find the optimal *morph lexicon*, for producing a segmentation of the corpus, i.e., a vocabulary of morphs that is concise, and moreover gives a concise representation for the corpus. This objective corresponds to Occam's razor, which says that among equally performing models one should prefer the smallest one. A mathematical formulation can be obtained using the Minimum Description Length (MDL) principle or probabilistically using maximum a posteriori (MAP) estimation.

Specific models presented by us can be seen as instances of the general Morfessor family. In this context we call the models as follows: Morfessor *Baseline* (Creutz and Lagus 2002), Morfessor *Baseline-Length* (Creutz and Lagus 2002), Morfessor *Categories-ML* (Creutz and Lagus 2004), and Morfessor *Categories-MAP* (Creutz and Lagus 2005a).

Table 1 shows example segmentations obtained by three of the models for the Finnish words 'megatähdeksi' ("[become a] megastar") and 'megatähdistä' ("from megastars") as well as the English words 'tyrannizes' and 'tyrannizing'. The algorithms produce different amounts of information: the Baseline and Baseline-Length methods only produce a segmentation of the words, whereas the category algorithms (Categories-ML and Categories-MAP) also indicate whether a segment functions as a prefix, stem, or suffix. Additionally, the morph lexicon learned by Categories-MAP contains hierarchical representations for some of its entries. These have been visualized using nested brackets.

Table 1: Examples of word segmentations learned by versions of Morfessor from a 16 million word Finnish corpus and a 12 million word English corpus. Proposed prefixes are <u>underlined</u>, stems are rendered in **bold-face**, and suffixes are *slanted*. Square brackets [ ] indicate higher-level entries in the hierarchical lexicon learned by Categories-MAP.

Baseline	Categories-ML	Categories-MAP
mega tähdeksi	<u>mega</u> **tähd** *e* *ksi*	[**mega** [**tähde** *ksi*]]
mega tähdistä	<u>mega</u> **tähd** *i* *stä*	[**mega** [**tähdi** *stä*]]
tyrann ize s	**tyrann** *ize* *s*	**tyrannize** *s*
tyrann izing	**tyrann** *izing*	**tyranni** **zing**

2.1. Software

The Morfessor 1.0 software package is publicly available on the Internet. The software consists of a Perl script and it is documented in a technical report (Creutz and Lagus 2005b). The Morfessor 1.0 package implements the Morfessor Baseline and Baseline-Length methods.

2.2. Internet demonstration

In addition to the downloadable software, there is a demonstration program on our Internet site. The user types in words of his own choice and the demo shows the analysis (segmentation) that Morfessor produces for these words. It is possible to select the model (Baseline or Categories-ML) and the data used for training the model. Small and large Finnish and English corpora are available.

3. Evaluation of the segmentation with Hutmegs

The Helsinki University of Technology Morphological Evaluation Gold Standard (Hutmegs) package contains fairly accurate morpheme segmentations for 1.4 million Finnish and 120 000 distinct English word forms. To produce these gold-standard segmentations for the words, we have processed the output of the two-level morphology analyzer FINTWOL (Koskenniemi 1983) and the contents of the English CELEX database (Baayen et al. 1995). For every word, an alignment between a surface (or allomorph) segmentation and a deep-level (or morpheme) segmentation has been obtained, as in the following examples:

```
megatähdeksi      mega:mega|PFX tähd^e:tähti|N ksi:TRA
megatähdistä      mega:mega|PFX tähd:tähti|N i:PL stä:ELA

tyrannizes        tyrann:tyrant|N iz^e:ize|s s:V+e3S
tyrannizing       tyrann:tyrant|N iz:ize|s ing:V+pe
```

For instance, the surface segmentation of the English word ‘tyrannizing’ is ‘tyrann + iz + ing’, which has the underlying deep-level representation ‘tyrant + ize + V+pe’. (Here, the label ‘V+pe’ corresponds to the present tense participle ‘ing’.) Additionally we know that ‘tyrant’ is a noun (N) and that ‘ize’ is a suffix (s).

There is also an option for so called “fuzzy” boundaries in the Hutmegs annotations (marked with ^). Fuzzy boundaries are applied in cases where it is inconvenient to define one exact transition point between two morphemes. For instance, in English, the final 'e' is dropped in some forms. Here we allow two correct segmentations, namely the traditional linguistic segmentation in 'tyrann + ize + s' as well as the alternative interpretation, where the 'e' is considered part of the following suffix: 'tyrann + iz + es'. (The latter can be compared to the form 'tyrann + iz + ing', where there is no 'e'.) The forms of the Finnish noun 'tähti' (star) behave in a similar way: In singular forms, we allow the final 'e' to belong to the stem or the ending: 'tähde + ksi' vs. 'tähd + eksi'; in plural there is no stem-final vowel and the segmentation is always 'tähd + i + stä'.

3.1. Experiment

Hutmegs also contains a number of Perl scripts and Makefiles for performing a quantitative evaluation of some suggested segmentation in relation to the desired segmentation in the Gold Standard. In Figure 1 we have plotted the results of an experiment, where the Morfessor Baseline model has been applied to different sized Finnish and English data.

Figure 1a shows how the *precision* and *recall* of the discovered morpheme boundaries develop, when the amount of data increases. Precision is the proportion of correctly discovered boundaries among all boundaries discovered by the algorithm. Recall is the proportion of correctly discovered boundaries among all correct boundaries. In order to get a comprehensive idea of the performance of a method, both measures must be taken into account.

A measure that combines precision and recall is the *F-measure*, which is the harmonic mean of the two and allows for a direct comparison of the goodness of segmentations. Figure 1b depicts the F-measure as a function of the data size.

Generally, the behavior of the Morfessor Baseline algorithm is such that precision increases as a function of the data size. That is, the proposed morpheme boundaries coincide more and more with morpheme boundaries in the Gold Standard. However, when the amount of data is very large, recall starts to decrease. That is, an increasing number of morpheme boundaries in the Gold Standard are missed by Morfessor Baseline.

3.2. Access

The Hutmegs package is a collection of files and documentation that are free to use for non-commercial purposes. However, to obtain the complete Finnish Gold Standard, a missing component must be licensed from Lingsoft, Inc. at an inexpensive price[1]. If the component is not purchased, the user will have access to all Hutmegs scripts and documentation, but only a sample Gold Standard containing the analyses of 700 Finnish word forms.

Likewise, the CELEX database is a prerequisite for accessing the complete English Gold Standard. Non-commercial licenses are available from the Linguistic Data Consortium[2]. The Hutmegs package provides sample segmentations for roughly 600 English word forms, which can be viewed without access to the CELEX database.

More detailed information about Hutmegs is available in a technical report (Creutz and Lindén 2004).

4. Conclusions

Currently, only the Baseline and Baseline-Length versions of Morfessor exist as public resources. The later models (Morfessor Categories-ML and Categories-MAP) may be

[1] URL: `http://www.lingsoft.fi`. Current price: 600 euros.
[2] URL: `http://www.ldc.upenn.edu/`. Current price: US$ 150.

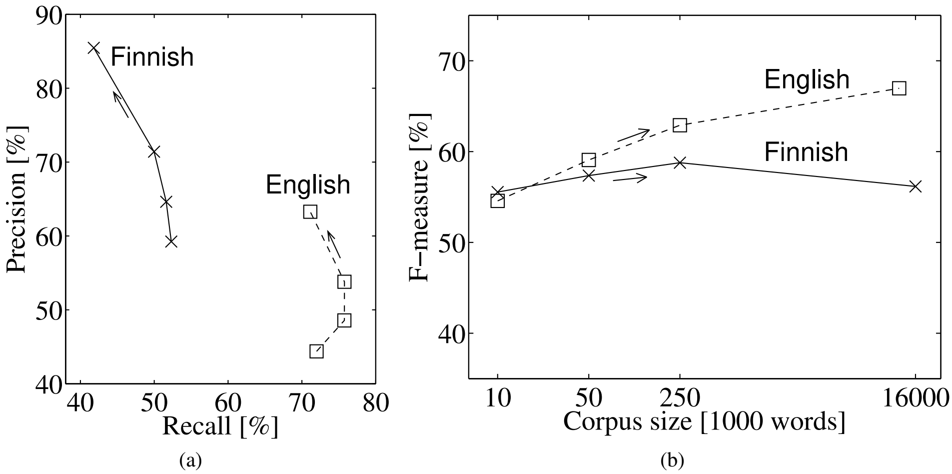

Figure 1: Performance of the Morfessor Baseline algorithm when evaluated against the Hutmegs Gold Standard on different sized corpora of Finnish and English text. In (a) the precision and recall of the discovered morpheme boundaries are shown. The points on the curves correspond to different data sizes and arrows indicate the direction of increasing data size. Precision measures the accuracy of the proposed splitting points, whereas recall describes the coverage of the splits. The most desirable area of the curve is the upper right corner, where both precision and recall are high. In (b) the corresponding F-measure values are shown as a function of the corpus size. The F-measure for Finnish is fairly constant across the corpus sizes, whereas the goodness of the English segmentation seems to improve when increasing the amount of data from 10 000 to 12 million words.

released in the future.

The segmentations in the Hutmegs Gold Standard have been designed for evaluating the accuracy of an unsupervised morphology-discovery algorithm. However, the given morpheme segmentations can also be used for other purposes. For instance, one can estimate n-gram language models from a corpus, where the words have been split into morphemes according to the Gold Standard. Such a language model can be utilized in unlimited-vocabulary continuous speech recognition; see e.g., (Siivola et al. 2003).

In conclusion, by supplying public benchmarking resources, we wish to contribute to the promotion of research in the fascinating field of unsupervised morphology discovery and morpheme segmentation.

References

Baayen, R. Harald; Piepenbrock, Richard; Gulikers, Léon 1995. The CELEX lexical database (CD-ROM). University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium. `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC96L14`

Creutz, M.; Lagus, K. 2002. Unsupervised discovery of morphemes. In: *Proc. Workshop on Morphological and Phonological Learning of ACL'02*, Philadelphia, Pennsylvania, USA. 21–30

Creutz, Mathias; Lagus, Krista 2004. Induction of a simple morphology for highly-inflecting languages. In: *Proc. 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, Barcelona. 43–51

Creutz, Mathias; Lagus, Krista 2005a. Inducing the morphological lexicon of a natural language from unannotated text. In: *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*. (submitted for review)

Creutz, Mathias; Lagus, Krista 2005b. Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor. Technical Report A81: Publications in Computer and Information Science, Helsinki University of Technology

Creutz, Mathias; Lindén, Krister 2004. Morpheme Segmentation Gold Standards for Finnish and English. Technical Report A77: Publications in Computer and Information Science, Helsinki University of Technology

Goldsmith, John 2001. Unsupervised learning of the morphology of a natural language. In: *Computational Linguistics* 27(2), 153–198

Koskenniemi, K. 1983. *Ph.D. thesis*: Two-level morphology: A general computational model for word-form recognition and production, University of Helsinki

Siivola, V.; Hirsimäki, T.; Creutz, M.; Kurimo, M. 2003. Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner. In: *Proc. Eurospeech'03*, Geneva, Switzerland. 2293–2296

MATHIAS CREUTZ is a post-graduate researcher at the Neural Networks Research Centre, Helsinki University of Technology (HUT). He received his M. Sc. degree at HUT in 2000 and is now working on his Doctoral thesis. His research interests concern the unsupervised induction of a morphology of natural languages, the automatic segmentation of words, and the application of subword-unit-based language models in unlimited vocabulary speech recognition. E-mail: mathias.creutz@hut.fi

KRISTA LAGUS is a teaching research scientist at the Neural Networks Research Centre, Helsinki University of Technology. She received her Ph.D. at HUT in 2000, dealing with text mining using the WEBSOM neural network method. Her research interests concern using machine learning methods for modeling the emergence of representations of language and cognition in artificial systems. She has taught courses on statistical language modeling and related topics at HUT. She has published 6 articles in international journals, several book chapters and over 30 conference articles. E-mail: krista.lagus@hut.fi

KRISTER LINDÉN is a postgraduate researcher at the Department of General Linguistics, Helsinki. He received his M.Sc. at the University of Helsinki. His research interests concern word sense disambiguation and word sense discovery and their application to machine translation, speech recognition and cross-lingual information retrieval. E-mail: krister.linden@helsinki.fi

SAMI VIRPIOJA is an undergraduate researcher at the Neural Networks Research Centre, Helsinki University of Technology. He is currently working on his Master's thesis concerning natural language modeling. E-mail: sami.virpioja@hut.fi