Morfessor in the Morpho Challenge

Mathias Creutz and **Krista Lagus**
Helsinki University of Technology, Adaptive Informatics Research Centre
P. O. Box 5400, FIN-02105 HUT, Finland
{mathias.creutz, krista.lagus}@hut.fi

Abstract

In this work, Morfessor, a morpheme segmentation model and algorithm developed by the organizers of the Morpho Challenge, is outlined and references are made to earlier work. Although Morfessor does not take part in the official Challenge competition, we report experimental results for the morpheme segmentation of English, Finnish and Turkish words. The obtained results are very good. Morfessor outperforms the other algorithms in the Finnish and Turkish tasks and comes second in the English task. In the Finnish speech recognition task, Morfessor achieves the lowest letter error rate.

1 Introduction

This paper briefly describes three consecutive steps in the development of a morpheme segmentation and simple morphology induction algorithm, called *Morfessor*. Morfessor has been developed by the organizers of the Morpho Challenge and was therefore excluded from the official competition. However, we believe that the performance of Morfessor in the Morpho Challenge task will be of interest to a broader audience than the current authors, especially since the obtained results are generally very good.

The readers should keep in mind that a comparison of Morfessor to its competitors is not entirely fair, since portions of the Finnish and English data sets used in the competition have been utilized during the development of the Morfessor model. It is thus probable that the model implementation to some degree reflects properties of these very data sets. Nevertheless, the data set of the third language, Turkish, is as new to the organizers as to the participants. No modifications to the tested versions of the Morfessor model have been made after the acquisition of the Turkish data.

In the following sections, some characteristics of the Morfessor model will be outlined and experimental results obtained in the morpheme segmentation as well as Finnish speech recognition task will be reported and discussed.

2 Characterization of the Morfessor model

Morfessor is an unsupervised method for the segmentation of words into morpheme-like units. The general idea behind the Morfessor model is to discover as compact a description of the data as possible. Substrings occurring frequently enough in several different word forms are proposed as *morphs* and the words are then represented as a concatenation of morphs, e.g., “hand, hand+s, left+hand+ed, hand+ful”.

An optimal balance is sought between compactness of the *morph lexicon* versus the compactness of the representation of the *corpus*. The morph lexicon is a list of all distinct morphs (e.g., “hand, s, left, ed, ful”) together with some stored properties of these morphs. The representation of the corpus can be seen as a sequence of pointers to entries in the morph lexicon; e.g. the word “left-handed” is represented as three pointers to morphs in the lexicon.

A very compact lexicon could consist of the individual letters of the language. However, this would result in a very expensive representation of the corpus, since every word would be broken down into as many morphs as the number of letters it contains. The opposite situation consists of having a short representation of the corpus (e.g., no words would be split into parts), but then the lexicon would necessarily be very large, since it would have to contain all distinct words that occur in the corpus. Thus, the optimal solution is usually a compromise between these two extremes.

Among others, de Marcken (1996), Brent (1999), Goldsmith (2001), and Creutz and Lagus (2002; 2003; 2004; 2005a; 2006) have

shown that the above type of model produces segmentations that resemble linguistic morpheme segmentations, when formulated mathematically in a probabilistic framework or equivalently using the Minimum Description Length (MDL) principle (Rissanen, 1989).

An alternative popular approach to the segmentation of words and phrases is based on the works by Zellig S. Harris (1955; 1967). For instance, Schone and Jurafsky (2000; 2001) make use of a Harrisian approach to suggest word stems and suffixes. In this approach, word or morpheme boundaries are proposed at locations where the predictability of the next letter in a letter sequence is low. Such a model does not use compactness of representation as an explicit optimization criterion. Other related work is described more thoroughly in our previous publications.

Next, the three tested versions of the Morfessor model will be described briefly. These versions are called *Morfessor Baseline*, *Morfessor Categories-ML*, and *Morfessor Categories-MAP*. The versions correspond to chronological development steps, starting with the simplest model and ending with the most complex one. For a discussion on how the early versions can be seen as special cases of the latest model, the reader is encouraged to consult (Creutz and Lagus, 2006). Note that the current paper merely presents the underlying ideas and characteristics of the Morfessor model; in order to find an exact mathematical formulation it is necessary to read our previous works.

2.1 Morfessor Baseline

The Morfessor Baseline algorithm was originally introduced in (Creutz and Lagus, 2002), where it was called the "Recursive MDL" method. Additionally, the Baseline algorithm is described in (Creutz and Lagus, 2005b; Hirsimäki et al., 2006). The implementing computer program is publicly available for download at `http://www.cis.hut.fi/projects/morpho/`.

The Baseline method is a *context-independent* splitting algorithm. It is used as a baseline, or initialization, for the later *context-dependent* model versions (Categories-ML and Categories-MAP). In slightly simplified form, the optimization criterion utilized in Morfessor Baseline corresponds to the maximization of the following posterior probability:

$$P(\text{lexicon} \mid \text{corpus}) \propto P(\text{lexicon})P(\text{corpus} \mid \text{lexicon}) = \prod_{\text{letters } \alpha} P(\alpha) \cdot \prod_{\text{morphs } \mu} P(\mu). \quad (1)$$

The lexicon consists of all distinct morphs spelled out; this forms a long string of letters α. The probability of the lexicon is the product of the probability of each letter in this string. Analogously, the corpus is represented as a sequence of morphs, which corresponds to a particular segmentation of the words in the corpus. The probability of this segmentation equals the product of the probability of each morph token μ. Letter and morph probabilities are maximum likelihood estimates.

When segmentations produced by the Baseline method are compared to linguistic morpheme segmentations, the algorithm suffers from three types of fairly common errors: *undersegmentation* of frequent strings, *oversegmentation* of rare strings, and *morphotactic violations*. This follows from the fact that the most concise representation is obtained when any frequent string is stored as a whole in the lexicon (e.g., English "having, soldiers, states, seemed"), whereas infrequent strings are better coded in parts (e.g., "or+p+han, s+ed+it+ious, vol+can+o"). Morphotactic violations are a consequence of the context-independent nature of the model: For instance, the morphs "-s" and "-ed" are frequently occurring *suffixes* in the English language, but the algorithm occasionally suggests them in word-initial position as *prefixes* ("s+wing, ed+ward, s+urge+on").

2.2 Morfessor Categories-ML

Morfessor Categories-ML (Creutz and Lagus, 2004) introduces morph categories. The segmentation of the corpus is modeled using a Hidden Markov Model (HMM) with transition probabilities between categories and emission probabilities of morphs from categories (see Fig. 1). Three categories are used: *prefix*, *stem*, and *suffix* and an additional *non-morpheme* (or *noise*) category. Some distributional properties of the morphs in a proposed segmentation of the corpus are used for determining category-to-morph emission probabilities. A morph that is observed to precede a large number of different morphs is a likely prefix (e.g., English "re-, un-, mis-"); this is measured by *right perplexity* (Fig. 2a). Correspondingly, a morph that is observed to follow a large set of

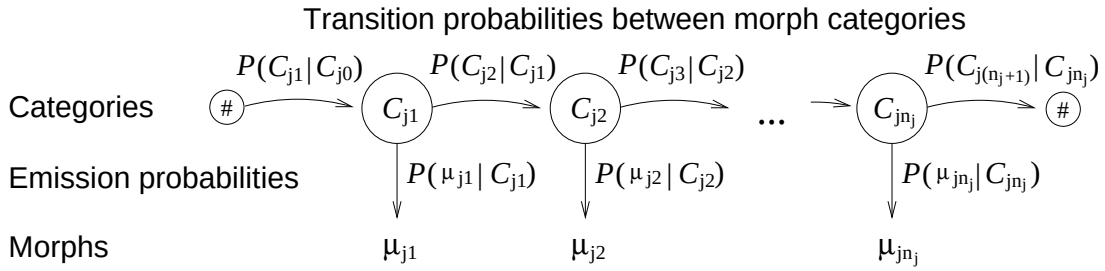


Figure 1: Hidden Markov model used in Categories-ML and Categories-MAP to compute $P(\text{corpus} \mid \text{lexicon})$. The picture shows the HMM representing one word in the corpus (the j^{th} word, which has been split into n_j morphs). The word consists of a sequence of morphs $\mu_{j\cdot}$ which are emitted from latent categories $C_{j\cdot}$. Note that the transition probabilities comprise transitions from and to a special word boundary category (#).

morphs is likely to be a suffix (e.g., "-s, -ed, -ing"); this is measured by *left-perplexity* (Fig. 2b). A morph that is not very short is likely to be a stem (e.g., "friend, hannibal, poison"); see Fig. 2c. A morph that is not an obvious prefix, stem, or suffix in the position it occurs may be an indication of an erroneous segmentation. Such morphs are tagged as noise (e.g., all morphs in the segmentation "vol+can+o").

The identification of "noise' and likely erroneous segmentations makes it possible to apply some heuristics in order to partly remedy the shortcomings of Morfessor Baseline. Undersegmentation is reduced by forcing splits of redundant morphs in the lexicon. These morphs consist of other morphs that are also present in the lexicon (e.g., "seemed = seem+ed"). Some restrictions apply, such that splitting into noise morphs is prohibited. The opposite problem, oversegmentation, is alleviated by joining morphs tagged as noise with their neighbors (e.g, "vol+can+o" becomes "volcano"). Morphotactic violations are less likely to occur due to the context-sensitivity of the HMM model.

2.3 Morfessor Categories-MAP

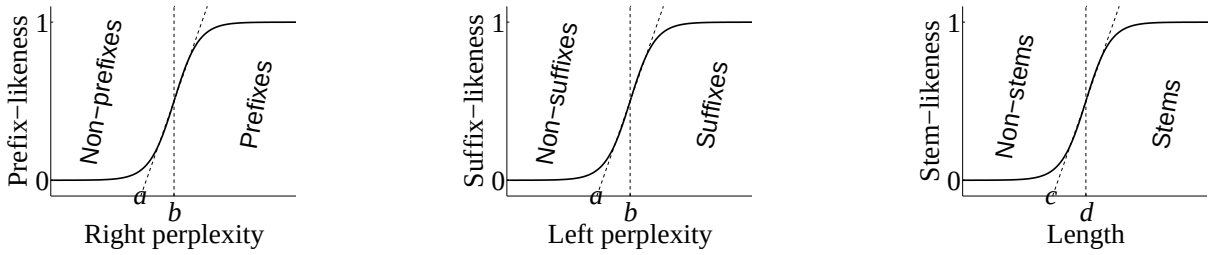
The Categories-MAP model version (Creutz and Lagus, 2005a) emerged in an attempt to reformulate Categories-ML in a more elegant fashion. In Categories-ML, the optimal segmentation of the corpus is sought through Maximum Likelihood (ML) re-estimation, whereas the complexity of the lexicon is controlled heuristically. In a Maximum a Posteriori (MAP) model, an explicit probability is calculated for both the lexicon and the representation of the corpus conditioned on the lexicon. Categories-MAP and the Baseline method are MAP models.

The most important new feature of the Categories-MAP model is that the lexicon may contain hierarchical entries. That is, a morph can either consist of a string of letters (as in the previous models) or of two submorphs, which can recursively consist of submorphs.

As was the case in the Baseline model, frequent strings typically end up as entries of their own in the lexicon (e.g, the English word "straightforwardness"). However, unlike in the Baseline model, these frequent strings now have a hierarchical representation; see Figure 3. In a morpheme segmentation task, the existence of this inner structure makes it possible to "expand" morphs into their submorphs, thereby avoiding undersegmentation. Since every morph at every level is tagged with its most likely category, it is possible to avoid *over*segmentation as well, since one can refrain from expanding nodes in the tree if the next level contains *non-morphemes*, i.e. "noise morphs". For instance, in Figure 3, the word "straightforwardness" is expanded into "straight+forward+ness". The morph "forward" is not expanded into its constituents "for+ward" (although this may have been appropriate), because "for" is tagged as a non-morpheme in the current context.

3 Morpheme Segmentation Experiments

In the following, some differences between the tested versions of Morfessor as well as the three tested languages are illustrated in the light of experimental results. The experiments were run on the datasets provided in the Challenge. The



(a) (b) (c)

Figure 2: Sketch of sigmoid functions (used in the Categories models), which express how the right and left perplexity as well as the length of a morph affect its tendency to function as a prefix, suffix, or stem. The parameters a, b, c, d determine the shape of the sigmoids. A probability distribution is obtained by first computing the probability that a morph μ belongs to *none* of the three categories. The probability of this so-called non-morpheme, or noise, category given the morph μ equals: $(1 - \textit{prefix-like}(\mu)) \cdot (1 - \textit{suffix-like}(\mu)) \cdot (1 - \textit{stem-like}(\mu))$. Then the remaining probability mass is distributed between prefix, stem and suffix proportionally to the prefix-, stem- and suffix-likeness values.

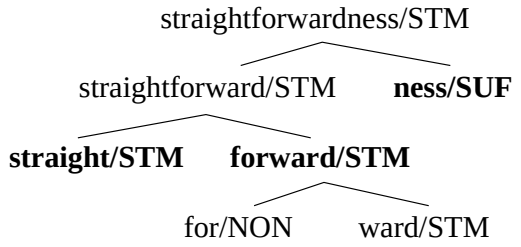


Figure 3: Hierarchical representation of the English word "straightforwardness" in the lexicon induced by Morfessor Categories-MAP. Each morph has been tagged with a category: stem (STM), suffix (SUF), or non-morpheme (NON). (No morph was tagged as a prefix in this example.) The finest resolution that does not contain non-morphemes is rendered using a bold-face font. This corresponds to the proposed morpheme segmentation.

Morfessor Baseline algorithm is entirely unsupervised and does not require that any parameters be set. The Categories algorithms have one parameter (the perplexity threshold b in Fig. 2) that needs to be set to an appropriate value for optimal performance. This parameter value was optimized separately for each language on the small development sets (model segmentations) provided.[1]

[1]A fixed (dataset-independent) scheme works fine for the other parameters in Fig. 2: $a = 10/b, c = 2, d = 3.5$. This is good, since the amount of necessary supervision should be kept to a minimum.

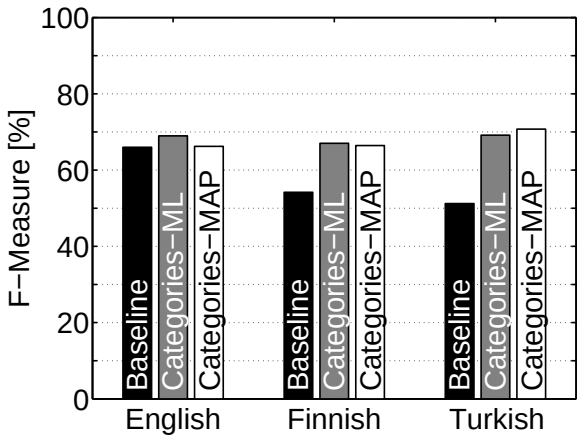


Figure 4: F-measures computed for the placement of morpheme boundaries in relation to linguistic morpheme segmentations, obtained by the three different versions of Morfessor on the three test languages.

3.1 Results

The morpheme segmentation task of the competition is won by the participant achieving the highest *F-measure* of correctly placed morpheme boundaries. Figure 4 shows the F-measures of the three Morfessor methods on the three tested languages. The F-measure is the harmonic mean of *precision* and *recall*. The precisions and recalls obtained by Morfessor are displayed in Figures 5 and 6, respectively.

The results show that there are different tendencies for the English data, on the one hand, and the

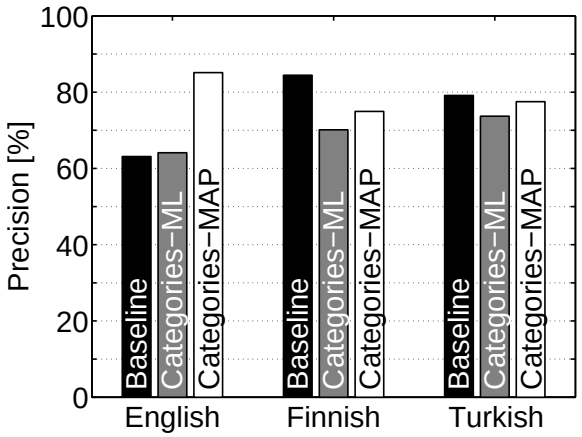


Figure 5: Precision of the three Morfessor methods on the three languages tested.

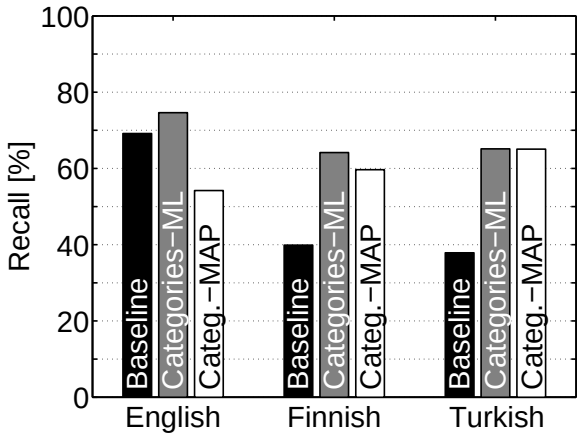


Figure 6: Recall of the three Morfessor methods on the three languages tested.

Finnish and Turkish data, on the other hand. For Finnish and Turkish, the context-dependent Categories models produce clear improvements over the context-independent Baseline splitting algorithm (with F-measures 10 – 20 points higher; Fig 4). For English, the improvement is minor, but on the other hand the Baseline here attains a considerably higher level than for Finnish and Turkish. The best F-measure obtained by Morfessor for all three languages is at the same level, around 70 %.

The precision and recall plots in Figures 5 and 6 provide more detailed information. For English, even though the F-measures of all three algorithms are approximately equal, the produced segmentations are very different. Categories-MAP has a significantly higher precision than the other model versions (and correspondingly a lower recall). For Finnish and Turkish, the Categories models display a great improvement of recall in relation to the Baseline method. This comes at the expense of lower precision, which is observed for Finnish and to a lesser degree on the Turkish data.

In order to better understand the differences observed in the results for the different languages, the output at various stages of the segmentation process has been studied for each of the Morfessor model variants. No obvious explanation has been found other than the difference in the morphological structures of the languages. Finnish and Turkish are predominantly agglutinative languages, in which words are formed through the concatenation of morphemes. The type/token ratio is high, i.e., the number of different word forms encountered in a piece of running text is relatively high. By contrast, word forming in English involves fewer morphemes. The type/token ratio is lower, and the proportion of frequently occurring word forms is higher.

In the Finnish and Turkish segmentation task, Morfessor outperforms all algorithms proposed by the participants of the Morpho Challenge; compare the following F-measures for Finnish: 67.0 % (Morfessor Categories-ML) vs. 64.7 % (best participant), and for Turkish: 70.7 % (Morfessor Categories-MAP) vs. 65.3 % (best participant). In the English segmentation task, Morfessor comes second: 69.0 % (Morfessor Categories-ML) vs. 76.8 % (best participant).

4 Finnish Speech Recognition Experiments

N-gram language models have been estimated from the segmentations produced by the three Morfessor models on the Finnish data. The language models have been used in speech recognition experiments, and results are shown in Table 1. The evaluation of the language models alone (cross-entropy on a held-out data set) suggests that the Categories models are better than Morfessor Baseline, since their cross-entropy is lower. The cross-entropies do not, however, correlate with the actual speech recognition results. Categories-MAP obtains the lowest letter error rate (LER) – 1.30 % of the recognized letters are incorrect in comparison with the reference transcript – which is also lower than the letter error rate achieved by any participant of the Challenge (best result:

Table 1: Results from the Finnish speech recognition experiments: cross-entropy (log-perplexity) of the language models (H), letter error rate (LER) and word error rate (WER).

Method	H [bits]	LER [%]	WER [%]
Baseline	13.59	1.31	9.84
Categ.-ML	13.53	1.32	10.18
Categ.-MAP	13.53	1.30	10.05

1.32 %). Nevertheless, the word error rate (WER) of Categories-MAP is higher than that of Morfessor Baseline and the WER:s of three participants. This suggests that the letter errors made by Categories-MAP are spread over a larger number of words, which increases WER, whereas the other methods have a concentration of errors on a smaller set of words.

5 Conclusions

In the morpheme segmentation task, the current versions of Morfessor attain an F-measure value of about 70 % for all three tested languages. For English, a language with "poorer" morphology and less morpheme boundaries to discover, the simple Baseline method seems to almost reach to this level. The characteristically agglutinative languages Finnish and Turkish, which have "richer" morphology and a larger number of morpheme boundaries to be detected, require more complex models (the context-sensitive Categories model) to perform on the same level. It is particularly encouraging to see that Morfessor performs so well in the Turkish segmentation task, since Turkish data was never used in the development of the model.

References

M. R. Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proc. Workshop on Morphological and Phonological Learning of ACL'02*, pages 21–30, Philadelphia, Pennsylvania, USA.

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

Mathias Creutz and Krista Lagus. 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)*.

Mathias Creutz and Krista Lagus. 2005b. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology.

Mathias Creutz and Krista Lagus. 2006. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*. (Accepted for publication).

Mathias Creutz. 2003. Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proc. ACL'03*, pages 280–287, Sapporo, Japan.

C. G. de Marcken. 1996. *Unsupervised Language Acquisition*. Ph.D. thesis, MIT.

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

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222. Reprinted 1970 in *Papers in Structural and Transformational Linguistics*, Reidel Publishing Company, Dordrecht, Holland.

Zellig S. Harris. 1967. Morpheme boundaries within words: Report on a computer test. *Transformations and Discourse Analysis Papers*, 73. Reprinted 1970 in *Papers in Structural and Transformational Linguistics*, Reidel Publishing Company, Dordrecht, Holland.

Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen. 2006. Unlimited vocabulary speech recognition with morph language models applied to finnish. *Computer Speech and Language*. (In press).

Jorma Rissanen. 1989. *Stochastic Complexity in Statistical Inquiry*, volume 15. World Scientific Series in Computer Science, Singapore.

P. Schone and D. Jurafsky. 2000. Knowledge-free induction of morphology using Latent Semantic Analysis. In *Proc. CoNLL-2000 & LLL-2000*, pages 67–72.

P. Schone and D. Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *Proc. NAACL-2001*.