# Semi-supervised learning of concatenative morphology

**Oskar Kohonen** and **Sami Virpioja** and **Krista Lagus**
Aalto University School of Science and Technology
Adaptive Informatics Research Centre
P.O. Box 15400, FI-00076 AALTO, Finland
{oskar.kohonen,sami.virpioja,krista.lagus}@tkk.fi

## Abstract

We consider morphology learning in a semi-supervised setting, where a small set of linguistic gold standard analyses is available. We extend Morfessor Baseline, which is a method for unsupervised morphological segmentation, to this task. We show that known linguistic segmentations can be exploited by adding them into the data likelihood function and optimizing separate weights for unlabeled and labeled data. Experiments on English and Finnish are presented with varying amount of labeled data. Results of the linguistic evaluation of Morpho Challenge improve rapidly already with small amounts of labeled data, surpassing the state-of-the-art unsupervised methods at 1000 labeled words for English and at 100 labeled words for Finnish.

## 1 Introduction

Morphological analysis is required in many natural language processing problems. Especially, in agglutinative and compounding languages, where each word form consists of a combination of stems and affixes, the number of unique word forms in a corpus is very large. This leads to problems in word-based statistical language modeling: Even with a large training corpus, many of the words encountered when applying the model did not occur in the training corpus, and thus there is no information available on how to process them. Using morphological units, such as stems and affixes, instead of complete word forms alleviates this problem. Unfortunately, for many languages morphological analysis tools either do not exist or they are not freely available. In many cases, the problems of availability also apply to morphologically annotated corpora, making supervised learning infeasible.

In consequence, there has been a need for approaches for morphological processing that would require little language-dependent resources. Due to this need, as well as the general interest in language acquisition and unsupervised language learning, the research on unsupervised learning of morphology has been active during the past ten years. Especially, methods that perform morphological segmentation have been studied extensively (Goldsmith, 2001; Creutz and Lagus, 2002; Monson et al., 2004; Bernhard, 2006; Dasgupta and Ng, 2007; Snyder and Barzilay, 2008b; Poon et al., 2009). These methods have shown to produce results that improve performance in several applications, such as speech recognition and information retrieval (Creutz et al., 2007; Kurimo et al., 2008).

While unsupervised methods often work quite well across different languages, it is difficult to avoid biases toward certain kinds of languages and analyses. For example, in isolating languages, the average amount of morphemes per word is low, whereas in synthetic languages the amount may be very high. Also, different applications may need a particular bias, for example, not analyzing frequent compound words as consisting of smaller parts could be beneficial in information retrieval. In many cases, even a small amount of labeled data can be used to adapt a method to a particular language and task. Methodologically, this is referred to as semi-supervised learning.

In semi-supervised learning, the learning system has access to both labeled and unlabeled data. Typically, the labeled data set is too small for supervised methods to be effective, but there is a large amount of unlabeled data available. There are many different approaches to this class of problems, as presented by Zhu (2005). One approach is to use generative models, which specify a join distribution over all variables in the model. They can be utilized both in unsupervised



*Proceedings of the 11th Meeting of the ACL-SIGMORPHON, ACL 2010*, pages 78–86,
Uppsala, Sweden, 15 July 2010. 

and supervised learning. In contrast, discriminative models only specify the conditional distribution between input data and labels, and therefore require labeled data. Both, however, can be extended to the semi-supervised case. For generative models, it is, in principle, very easy to use both labeled and unlabeled data. For unsupervised learning one can consider the labels as missing data and estimate their values using the Expectation Maximization (EM) algorithm (Dempster et al., 1977). In the semi-supervised case, some labels are available, and the rest are considered missing and estimated with EM.

In this paper, we extend the Morfessor Baseline method for the semi-supervised case. Morfessor (Creutz and Lagus, 2002; Creutz and Lagus, 2005; Creutz and Lagus, 2007, etc.) is one of the well-established methods for morphological segmentation. It applies a simple generative model. The basic idea, inspired by the Minimum Description Length principle (Rissanen, 1989), is to encode the words in the training data with a lexicon of morphs, that are segments of the words. The number of bits needed to encode both the morph lexicon and the data using the lexicon should be minimized. Morfessor does not limit the number of morphemes per word form, making it suitable for modeling a large variety of agglutinative languages irrespective of them being more isolating or synthetic. We show that the model can be trained in a similar fashion in the semi-supervised case as in the unsupervised case. However, with a large set of unlabeled data, the effect of the supervision on the results tends to be small. Thus, we add a discriminative weighting scheme, where a small set of word forms with gold standard analyzes are used for tuning the respective weights of the labeled and unlabeled data.

The paper is organized as follows: First, we discuss related work on semi-supervised learning. Then we describe the Morfessor Baseline model and the unsupervised algorithm, followed by our semi-supervised extension. Finally, we present experimental results for English and Finnish using the Morpho Challenge data sets (Kurimo et al., 2009).

### 1.1 Related work

There is surprisingly little work that consider improving the unsupervised models of morphology with small amounts of annotated data. In the related tasks that deal with sequential labeling (word segmentation, POS tagging, shallow parsing, named-entity recognition), semi-supervised learning is more common.

Snyder and Barzilay (2008a; 2008b) consider learning morphological segmentation with non-parametric Bayesian model from multilingual data. For multilingual settings, they extract 6 139 parallel short phrases from the Hebrew, Arabic, Aramaic and English bible. Using the aligned phrase pairs, the model can learn the segmentations for two languages at the same time. In one of the papers (2008a), they consider also semi-supervised scenarios, where annotated data is available either in only one language or both of the languages. However, the amount of annotated data is fixed to the half of the full data. This differs from our experimental setting, where the amount of unlabeled data is very large and the amount of labeled data relatively small.

Poon et al. (2009) apply a log-linear, undirected generative model for learning the morphology of Arabic and Hebrew. They report results for the same small data set as Snyder and Barzilay (2008a) in both unsupervised and semi-supervised settings. For the latter, they use somewhat smaller proportions of annotated data, varying from 25% to 100% of the total data, but the amount of unlabeled data is still very small. Results are reported also for a larger 120 000 word Arabic data set, but only for unsupervised learning.

A problem similar to morphological segmentation is word segmentation for the languages where orthography does not specify word boundaries. However, the amount of labeled data is usually large, and unlabeled data is just an additional source of information. Li and McCallum (2005) apply a semi-supervised approach to Chinese word segmentation where unlabeled data is utilized for forming word clusters, which are then used as features for a supervised classifier. Xu et al. (2008) adapt a Chinese word segmentation specifically to a machine translation task, by using the indirect supervision from a parallel corpus.

## 2 Method

We present an extension of the Morfessor Baseline method to the semi-supervised setting. Morfessor Baseline is based on a generative probabilistic model. It is a method for modeling concatenative morphology, where the morphs—i.e., the sur-

face forms of morphemes—of a word are its non-overlapping segments. The model parameters $\boldsymbol{\theta}$ encode a morph lexicon, which includes the properties of the morphs, such as their string representations. Each morph $m$ in the lexicon has a probability of occurring in a word, $P(M = m \mid \boldsymbol{\theta})$.[1] The probabilities are assumed to be independent. The model uses a prior $P(\boldsymbol{\theta})$, derived using the Minimum Description Length (MDL) principle, that controls the complexity of the model. Intuitively, the prior assigns higher probability to models that store fewer morphs, where a morph is considered stored if $P(M = m \mid \boldsymbol{\theta}) > 0$. During model learning, $\boldsymbol{\theta}$ is optimized to maximize the posterior probability:

$$\begin{aligned}\boldsymbol{\theta}^{\text{MAP}} &= \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\boldsymbol{D}_W) \\ &= \arg\max_{\boldsymbol{\theta}} \big\{P(\boldsymbol{\theta})P(\boldsymbol{D}_W|\boldsymbol{\theta})\big\}, \quad (1)\end{aligned}$$

where $\boldsymbol{D}_W$ includes the words in the training data. In this section, we first consider separately the likelihood $P(\boldsymbol{D}_W|\boldsymbol{\theta})$ and the prior $P(\boldsymbol{\theta})$ used in Morfessor Baseline. Then we describe the algorithms, first unsupervised and then semi-supervised, for finding optimal model parameters. Last, we shortly discuss the algorithm for segmenting new words after the model training.

### 2.1 Likelihood

The latent variable of the model, $\boldsymbol{Z} = (Z_1, \ldots, Z_{|\boldsymbol{D}_W|})$, contains the analyses of the words in the training data $\boldsymbol{D}_W$. An instance of a single analysis for the $j$:th word is a sequence of morphs, $z_j = (m_{j1}, \ldots, m_{j|z_j|})$. During training, each word $w_j$ is assumed to have only one possible analysis. Thus, instead of using the joint distribution $P(\boldsymbol{D}_W, \boldsymbol{Z} \mid \boldsymbol{\theta})$, we need to use the likelihood function only conditioned on the analyses of the observed words, $P(\boldsymbol{D}_W \mid \boldsymbol{Z}, \boldsymbol{\theta})$. The conditional likelihood is

$$\begin{aligned}&P(\boldsymbol{D}_W \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{\theta}) \\ &= \prod_{j=1}^{|\boldsymbol{D}_W|} P(W = w_j \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{\theta}) \\ &= \prod_{j=1}^{|\boldsymbol{D}_W|} \prod_{i=1}^{|z_j|} P(M = m_{ji} \mid \boldsymbol{\theta}), \quad (2)\end{aligned}$$

where $m_{ij}$ is the $i$:th morph in word $w_j$.

[1] We denote variables with uppercase letters and their instances with lowercase letters.

### 2.2 Priors

Morfessor applies Maximum A Posteriori (MAP) estimation, so priors for the model parameters need to be defined. The parameters $\boldsymbol{\theta}$ of the model are:

- Morph type count, or the size of the morph lexicon, $\mu \in \mathbb{Z}_+$
- Morph token count, or the number of morphs tokens in the observed data, $\nu \in \mathbb{Z}_+$
- Morph strings $(\sigma_1, \ldots, \sigma_\mu)$, $\sigma_i \in \Sigma^*$
- Morph counts $(\tau_1, \ldots, \tau_\mu)$, $\tau_i \in \{1, \ldots, \nu\}$, $\sum_i \tau_i = \nu$. Normalized with $\nu$, these give the probabilities of the morphs.

MDL-inspired and non-informative priors have been preferred. When using such priors, morph type count and morph token counts can be neglected when optimizing the model. The morph string prior is based on length distribution $P(L)$ and distribution $P(C)$ of characters over the character set $\Sigma$, both assumed to be known:

$$P(\sigma_i) = P(L = |\sigma_i|) \prod_{j=1}^{|\sigma_i|} P(C = \sigma_{ij}) \quad (3)$$

We use the implicit length prior (Creutz and Lagus, 2005), which is obtained by removing $P(L)$ and using end-of-word mark as an additional character in $P(C)$. For morph counts, the non-informative prior

$$P(\tau_1, \ldots, \tau_\mu) = 1/\binom{\nu - 1}{\mu - 1} \quad (4)$$

gives equal probability to each possible combination of the counts when $\mu$ and $\nu$ are known, as there are $\binom{\nu-1}{\mu-1}$ possible ways to choose $\mu$ positive integers that sum up to $\nu$.

### 2.3 Unsupervised learning

In principle, unsupervised learning can be performed by looking for the MAP estimate with the EM-algorithm. In the case of Morfessor Baseline, this is problematic, because the prior only assigns higher probability to lexicons where fewer morphs have nonzero probabilities. The EM-algorithm has the property that it will not assign a zero probability to any morph, that has a nonzero likelihood in the previous step, and this will hold for all morphs

that initially have a nonzero probability. In consequence, Morfessor Baseline instead uses a local search algorithm, which will assign zero probability to a large part of the potential morphs. This is memory-efficient, since only the morphs with nonzero probabilities need to be stored in memory. The training algorithm of Morfessor Baseline, described by Creutz and Lagus (2005), tries to minimize the cost function

$$L(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{D}_W) = -\ln P(\boldsymbol{\theta}) - \ln P(\boldsymbol{D}_W \mid \boldsymbol{z}, \boldsymbol{\theta}) \quad (5)$$

by testing local changes to $\boldsymbol{z}$, modifying the parameters according to each change, and selecting the best one. More specifically, one word is processed at a time, and the segmentation that minimizes the cost function with the optimal model parameters is selected:

$$z_j^{(t+1)} = \arg\min_{z_j} \left\{ \min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{z}^{(t)}, \boldsymbol{D}_W) \right\}. \quad (6)$$

Next, the parameters are updated:

$$\boldsymbol{\theta}^{(t+1)} = \arg\min_{\boldsymbol{\theta}} \left\{ L(\boldsymbol{\theta}, \boldsymbol{z}^{(t+1)}, \boldsymbol{D}_W) \right\}. \quad (7)$$

As neither of the steps can increase the cost function, this will converge to a local optimum. The initial parameters are obtained by adding all the words into the morph lexicon. Due to the context independence of the morphs within a word, the optimal analysis for a segment does not depend on in which context the segment appears. Thus, it is possible to encode $\boldsymbol{z}$ as a binary tree-like graph, where the words are the top nodes and morphs the leaf nodes. For each word, every possible split into two morphs is tested in addition to no split. If the word is split, the same test is applied recursively to its parts. See, e.g., Creutz and Lagus (2005) for more details and pseudo-code.

### 2.4 Semi-supervised learning

A straightforward way to do semi-supervised learning is to fix the analyses $\boldsymbol{z}$ for the labeled examples. Early experiments indicated that this has little effect on the results. The Morfessor Baseline model only contains local parameters for morphs, and relies on the bias given by its prior to guide the amount of segmentation. Therefore, it may not be well suited for semi-supervised learning. The labeled data affects only the morphs that are found in the labeled data, and even their analyses can be overwhelmed by a large amount of unsupervised data and the bias of the prior.

We suggest a fairly simple solution to this by introducing extra parameters that guide the more general behavior of the model. The amount of segmentation is mostly affected by the balance between the prior and the model. The Morfessor Baseline model has been developed to ensure this balance is sensible. However, the labeled data gives a strong source of information regarding the amount of segmentation preferred by the gold standard. We can utilize this information by introducing the weight $\alpha$ on the likelihood. To address the problem of labeled data being overwhelmed by the large amount of unlabeled data we introduce a second weight $\beta$ on the likelihood for the labeled data. These weights are optimized on a separate held-out set. Thus, instead of optimizing the MAP estimate, we minimize the following function:

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{D}_W, \boldsymbol{D}_{W \mapsto A}) = \\ & - \ln P(\boldsymbol{\theta}) \\ & - \alpha \times \ln P(\boldsymbol{D}_W \mid \boldsymbol{z}, \boldsymbol{\theta}) \\ & - \beta \times \ln P(\boldsymbol{D}_{W \mapsto A} \mid \boldsymbol{z}, \boldsymbol{\theta}) \end{aligned} \quad (8)$$

The labeled training set $\boldsymbol{D}_{W \mapsto A}$ may include alternative analyses for some of the words. Let $A(w_j) = \{a_{j1}, \ldots, a_{jk}\}$ be the set of known analyses for word $w_j$. Assuming the training samples are independent, and giving equal weight for each analysis, the likelihood of the labeled data would be

$$\begin{aligned} & P(\boldsymbol{D}_{W \mapsto A} \mid \boldsymbol{\theta}) \\ & = \prod_{j=1}^{|\boldsymbol{D}_{W \mapsto A}|} \prod_{a_{jk} \in A(w_j)} \prod_{i=1}^{|a_{jk}|} P(M = m_{jki} \mid \boldsymbol{\theta}). \end{aligned} \quad (9)$$

However, when the analyses of the words are fixed, the product over alternative analyses in $A$ is problematic, because the model cannot select several of them at the same time. A sum over $A(w_j)$:s would avoid this problem, but then the logarithm of the likelihood function becomes nontrivial (i.e., logarithm of sum of products) and too slow to calculate during the training. Instead, we use the hidden variable $\boldsymbol{Z}$ to select only one analysis also for the labeled samples, but now with the restriction that $Z_j \in A(w_j)$. The likelihood function for $\boldsymbol{D}_{W \mapsto A}$ is then equivalent to Equation 2. Because the recursive algorithm search assumes that a string is segmented in the same way irrespective of its context, the labeled data can still

get zero probabilities. In practice, zero probabilities in the labeled data likelihood are treated as very large, but not infinite, costs.

### 2.5 Segmenting new words

After training the model, a Viterbi-like algorithm can be applied to find the optimal segmentation of each word. As proposed by Virpioja and Kohonen (2009), also new morph types can be allowed by utilizing an approximate cost of adding them to the lexicon. As this enables reasonable results also when the training data is small, we use a similar technique. The cost is calculated from the decrease in the probabilities given in Equations 3 and 4 when a new morph is assumed to be in the lexicon.

## 3 Experiments

In the experiments, we compare six different variants of the Morfessor Baseline algorithm:

- **Unsupervised:** The classic, unsupervised Morfessor baseline.
- **Unsupervised + weighting:** A held-out set is used for adjusting the weight of the likelihood $\alpha$. When $\alpha = 1$ the method is equivalent to the unsupervised baseline. The main effect of adjusting $\alpha$ is to control how many segments per word the algorithm prefers. Higher $\alpha$ leads to fewer and lower $\alpha$ to more segments per word.
- **Supervised:** The semi-supervised method trained with only the labeled data.
- **Supervised + weighting:** As above, but the weight of the likelihood $\beta$ is optimized on the held-out set. The weight can only affect which segmentations are selected from the possible alternative segmentations in the labeled data.
- **Semi-supervised:** The semi-supervised method trained with both labeled and unlabeled data.
- **Semi-supervised + weighting:** As above, but the parameters $\alpha$ and $\beta$ are optimized using the the held-out set.

All variations are evaluated using the linguistic gold standard evaluation of Morpho Challenge 2009. For supervised and semi-supervised methods, the amount of labeled data is varied between 100 and 10 000 words, whereas the held-out set has 500 gold standard analyzes. To obtain precision-recall curves, we calculated weighted F0.5 and F2 scores in addition to the normal F1 score. The parameters $\alpha$ and $\beta$ were optimized also for those.

### 3.1 Data and evaluation

We used the English and Finnish data sets from Competition 1 of Morpho Challenge 2009 (Kurimo et al., 2009). Both are extracted from a three million sentence corpora. For English, there were $62\,185\,728$ word tokens and $384\,903$ word types. For Finnish, there were $36\,207\,308$ word tokens and $2\,206\,719$ word types. The complexity of Finnish morphology is indicated by the almost ten times larger number of word types than in English, while the number of word tokens is smaller.

We applied also the evaluation method of the Morpho Challenge 2009: The results of the morphological segmentation were compared to a linguistic gold standard analysis. Precision measures whether the words that share morphemes in the proposed analysis have common morphemes also in the gold standard, and recall measures the opposite. The final score to optimize was F-measure, i.e, the harmonic mean of the precision and recall.[2] In addition to the unweighted F1 score, we have applied F2 and F0.5 scores, which give more weight to recall and precision, respectively.

Finnish gold standards are based on FINTWOL morphological analyzer from Lingsoft, Inc., that applies the two-level model by Koskenniemi (1983). English gold standards are from the CELEX English database. The final test sets are the same as in Morpho Challenge, based on $10\,000$ English word forms and $200\,000$ Finnish word forms. The test sets are divided into ten parts for calculating deviations and statistical significances. For parameter tuning, we applied a small held-out set containing $500$ word forms that were not included in the test set.

For supervised and semi-supervised training, we created sets of five different sizes: $100$, $300$, $1\,000$, $3\,000$, and $10\,000$. They did not contain any of the word forms in the final test set, but were otherwise randomly selected from the words for

[2] Both the data sets and evaluation scripts are available from the Morpho Challenge 2009 web page: `http://www.cis.hut.fi/morphochallenge2009/`

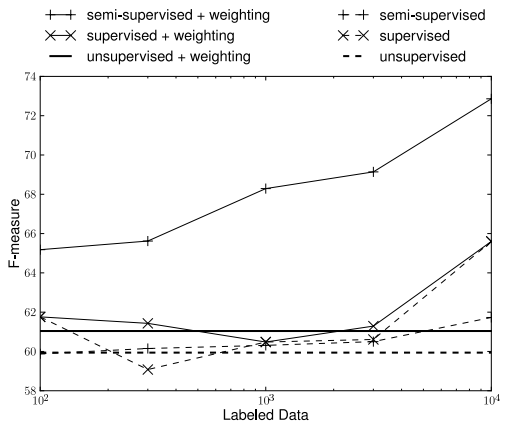


Figure 1: The F-measure for English as a function of the number of labeled training samples.

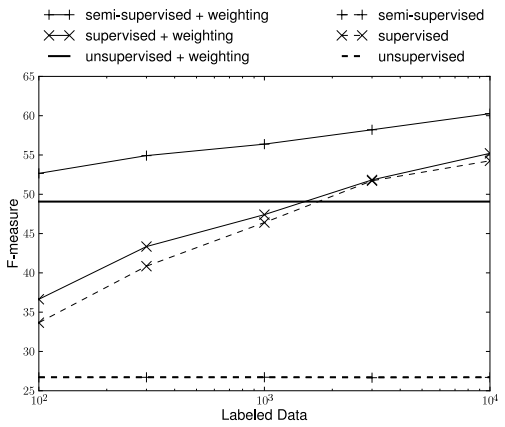


Figure 2: The F-measure for Finnish as a function of the number of labeled training samples. The *semi-supervised* and *unsupervised* lines overlap.

which the gold standard analyses were available. In order to use them for training Morfessor, the morpheme analyses were converted to segmentations using the Hutmegs package by Creutz and Lindén (2004).

### 3.2 Results

Figure 1 shows a comparison of the unsupervised, supervised and semi-supervised Morfessor Baseline for English. It can be seen that optimizing the likelihood weight $\alpha$ alone does not improve much over the unsupervised case, implying that the Morfessor Baseline is well suited for English morphology. Without weighting of the likelihood function, semi-supervised training improves the results somewhat, but it outperforms weighted unsupervised model only barely. With weighting, however, semi-supervised training improves the results significantly already for only 100 labeled training samples. For comparison, in Morpho Challenges (Kurimo et al., 2009), the unsupervised Morfessor Baseline and Morfessor Categories-MAP by Creutz and Lagus (2007) have achieved F-measures of 59.84% and 50.50%, respectively, and the all time best unsupervised result by a method that does not provide alternative analyses for words is 66.24%, obtained by Bernhard (2008).[3] This best unsupervised result is surpassed by the semi-supervised algorithm at 1000 labeled samples.

As shown in Figure 1, the supervised method obtains inconsistent scores for English with the smallest training data sizes. The supervised algorithm only knows the morphs in the training set, and therefore is crucially dependent on the Viterbi segmentation algorithm for analyzing new data. Thus, overfitting to some small data sets is not surprising. At 10 000 labeled training samples it clearly outperforms the unsupervised algorithm. The improvement obtained from tuning the weight $\beta$ in the supervised case is small.

Figure 2 shows the corresponding results for Finnish. The optimization of the likelihood weight gives a large improvement to the F-measure already in the unsupervised case. This is mainly because the standard unsupervised Morfessor Baseline method does not, on average, segment words into as many segments as would be appropriate for Finnish. Without weighting, the semi-supervised method does not improve over the unsupervised one: The unlabeled training data is so much larger that the labeled data has no real effect.

For Finnish, the unsupervised Morfessor Baseline and Categories-MAP obtain F-measures of 26.75% and 44.61%, respectively (Kurimo et al., 2009). The all time best for an unsupervised method is 52.45% by Bernhard (2008). With optimized likelihood weights, the semi-supervised Morfessor Baseline achieves higher F-measures with only 100 labeled training samples. Furthermore, the largest improvement for the semi-supervised method is achieved already from 1000 labeled training samples. Unlike English, the supervised method is quite a lot worse than the unsupervised one for small training data. This is natural because of the more complex morphology

[3] Better results (68.71%) have been achieved by Monson et al. (2008), but as they were obtained by combining of two systems as alternative analyses, the comparison is not as meaningful.

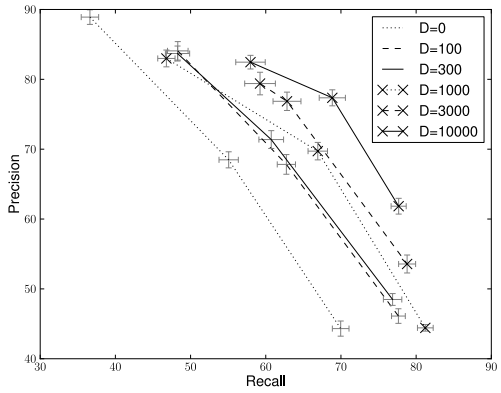


Figure 3: Precision-recall graph for English with varying amount of labeled training data. Parameters $\alpha$ and $\beta$ have been optimized for three different measures: F0.5, F1 and F2 on the held-out set. Precision and recall values are from the final test set, error bars indicate one standard deviation.

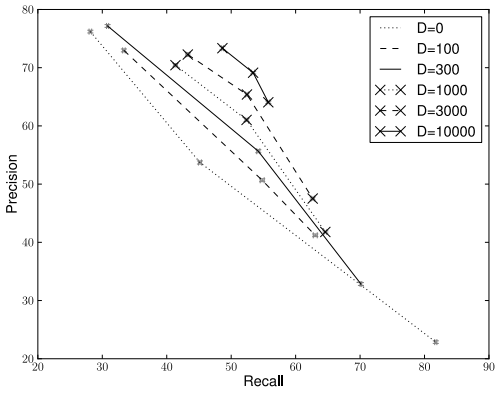


Figure 4: Precision-recall graph for Finnish with varying amount of labeled training data. Parameters $\alpha$ and $\beta$ have been optimized for three different measures: F0.5, F1 and F2 on the held-out set. Precision and recall values are from the final test set, error bars indicate one standard deviation, which here is very small.

in Finnish; good results are not achieved just by knowing the few most common suffixes.

Figures 3 and 4 show precision-recall graphs of the performance of the semi-supervised method for English and Finnish. The parameters $\alpha$ and $\beta$ have been optimized for three differently weighted F-measures (F0.5, F1, and F2) on the held-out set. The weight tells how much recall is emphasized; F1 is the symmetric F-measure that emphasizes precision and recall alike. The graphs show that the more there are labeled training data, the more constrained the model parameters are: With many labeled examples, the model cannot be forced to achieve high precision or recall only. The phenomenon is more evident in the Finnish data (Figure 3), where the same amount of words contains more information (morphemes) than in the English data. Table 1 shows the F0.5, F1 and F2 measures numerically.

Table 2 shows the values for the F1-optimal weights $\alpha$ and $\beta$ that were chosen for different amounts of labeled data using the held-out set. As even the largest labeled sets are much smaller than the unlabeled training set, it is natural that $\beta \gg \alpha$. The small optimal $\alpha$ for Finnish explains why the difference between unsupervised unweighted and weighted versions in Figure 2 was so large. Generally, the more there is labeled data, the smaller $\beta$ is needed. A possible increase in overall likelihood cost is compensated by a smaller $\alpha$. Finnish with 100 labeled words is an exception; probably a very high $\beta$ would end in overlearning of the small set words at the cost of overall performance.

## 4 Discussion

The method developed in this paper is a straightforward extension of Morfessor Baseline. In the semi-supervised setting, it should be possible to develop a generative model that would not require any discriminative reweighting, but could learn, e.g., the amount of segmentation from the labeled data. Moreover, it would be possible to learn the morpheme labels instead of just the segmentation into morphs, either within the current model or as a separate step after the segmentation. We made initial experiment with a trivial context-free labeling: A mapping between the segments and morpheme labels was extracted from the labeled training data. If some label did not have a corresponding segment, it was appended to the previous label. E.g., if the labels for "found" are "find_V +PAST", "found" was mapped to both labels. After segmentation, each segment in the test data was replaced by the most common label or label sequence whenever such was available. The results using training data with 1 000 and 10 000 labeled samples are shown in Table 3. Although precisions decrease somewhat, recalls improve considerably, and significant gains in F-measure are obtained. A more advanced, context-sensitive labeling should perform much better.

| English | | | |
|---|---|---|---|
| *labeled data* | *F0.5* | *F1* | *F2* |
| 0 | 69.16 | 61.05 | 62.70 |
| 100 | 73.23 | 65.18 | 68.30 |
| 300 | 72.98 | 65.63 | 68.81 |
| 1000 | 71.86 | 68.29 | 69.68 |
| 3000 | 74.34 | 69.13 | 72.01 |
| 10000 | 76.04 | 72.85 | 73.89 |
| **Finnish** | | | |
| *labeled data* | *F0.5* | *F1* | *F2* |
| 0 | 56.81 | 49.07 | 53.95 |
| 100 | 58.96 | 52.66 | 57.01 |
| 300 | 59.33 | 54.92 | 57.16 |
| 1000 | 61.75 | 56.38 | 58.24 |
| 3000 | 63.72 | 58.21 | 58.90 |
| 10000 | 66.58 | 60.26 | 57.24 |

Table 1: The F0.5, F1 and F2 measures for the *semi-supervised + weighting* method.

| | **English** | | **Finnish** | |
|---|---|---|---|---|
| *labeled data* | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ |
| 0 | 0.75 | - | 0.01 | - |
| 100 | 0.75 | 750 | 0.01 | 500 |
| 300 | 1 | 500 | 0.005 | 5000 |
| 1000 | 1 | 500 | 0.05 | 2500 |
| 3000 | 1.75 | 350 | 0.1 | 1000 |
| 10000 | 1.75 | 175 | 0.1 | 500 |

Table 2: The values for the weights $\alpha$ and $\beta$ that the semisupervised algorithm chose for different amounts of labeled data when optimizing F1-measure.

The semi-supervised extension could easily be applied to the other versions and extensions of Morfessor, such as Morfessor Categories-MAP (Creutz and Lagus, 2007) and Allomorfessor (Virpioja and Kohonen, 2009). Especially the modeling of allomorphy might benefit from even small amounts of labeled data, because those allomorphs that are hardest to find (affixes, stems with irregular orthographic changes) are often more common than the easy cases, and thus likely to be found even from a small labeled data set.

Even without labeling, it will be interesting to see how well the semi-supervised morphology learning works in applications such as information retrieval. Compared to unsupervised learning, we obtained much higher recall for reasonably good levels of precision, which should be beneficial to most applications.

| | **Segmented** | **Labeled** |
|---|---|---|
| **English,** $D = 1\,000$ | | |
| Precision | 69.72% | 69.30% |
| Recall | 66.92% | 72.21% |
| F-measure | 68.29% | 70.72% |
| **English,** $D = 10\,000$ | | |
| Precision | 77.35% | 77.07% |
| Recall | 68.85% | 77.78% |
| F-measure | 72.86% | 77.42% |
| **Finnish,** $D = 1\,000$ | | |
| Precision | 61.03% | 58.96% |
| Recall | 52.38% | 66.55% |
| F-measure | 56.38% | 62.53% |
| **Finnish,** $D = 10\,000$ | | |
| Precision | 69.14% | 66.90% |
| Recall | 53.40% | 74.08% |
| F-measure | 60.26% | 70.31% |

Table 3: Results of a simple morph labeling after segmentation with semi-supervised Morfessor.

## 5 Conclusions

We have evaluated an extension of the Morfessor Baseline method to semi-supervised morphological segmentation. Even with our simple method, the scores improve far beyond the best unsupervised results. Moreover, already one hundred known segmentations give significant gain over the unsupervised method even with the optimized data likelihood weight.

## Acknowledgments

This work was funded by Academy of Finland and Graduate School of Language Technology in Finland. We thank Mikko Kurimo and Tiina Lindh-Knuutila for comments on the manuscript, and Nokia foundation for financial support.

## References

Delphine Bernhard. 2006. Unsupervised morphological segmentation based on segment predictability and word segments alignment. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*, Venice, Italy. PASCAL European Network of Excellence.

Delphine Bernhard. 2008. Simple morpheme labelling in unsupervised morpheme analysis. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the CLEF*, volume 5152 of *Lecture Notes in Computer Science*, pages 873–880. Springer Berlin / Heidelberg.

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

Mathias Creutz and Krista Lagus. 2005. 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. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1), January.

Mathias Creutz and Krister Lindén. 2004. Morpheme segmentation gold standards for Finnish and English. Technical Report A77, Publications in Computer and Information Science, Helsinki University of Technology.

Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing*, 5(1):1–29.

Sajib Dasgupta and Vincent Ng. 2007. High-performance, language-independent morphological segmentation. In *the annual conference of the North American Chapter of the ACL (NAACL-HLT)*.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–189.

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

Mikko Kurimo, Mathias Creutz, and Matti Varjokallio. 2008. Morpho Challenge evaluation using a linguistic Gold Standard. In *Advances in Multilingual and MultiModal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, Lecture Notes in Computer Science , Vol. 5152, pages 864–873. Springer.

Mikko Kurimo, Sami Virpioja, Ville T. Turunen, Graeme W. Blackwood, and William Byrne. 2009. Overview and results of Morpho Challenge 2009. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, September.

Wei Li and Andrew McCallum. 2005. Semi-supervised sequence modeling with syntactic topic models. In *AAAI'05: Proceedings of the 20th national conference on Artificial intelligence*, pages 813–818. AAAI Press.

Christian Monson, Alon Lavie, Jaime Carbonell, and Lori Levin. 2004. Unsupervised induction of natural language morphology inflection classes. In *Proceedings of the Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*.

Christian Monson, Jaime Carbonell, Alon Lavie, and Lori Levin. 2008. ParaMor: Finding paradigms across morphology. In *Advances in Multilingual and MultiModal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, Lecture Notes in Computer Science , Vol. 5152. Springer.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 209–217. Association for Computational Linguistics.

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

Benjamin Snyder and Regina Barzilay. 2008a. Cross-lingual propagation for morphological analysis. In *AAAI'08: Proceedings of the 23rd national conference on Artificial intelligence*, pages 848–854. AAAI Press.

Benjamin Snyder and Regina Barzilay. 2008b. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745, Columbus, Ohio, June. Association for Computational Linguistics.

Sami Virpioja and Oskar Kohonen. 2009. Unsupervised morpheme analysis with Allomorfessor. In *Working notes for the CLEF 2009 Workshop*, Corfu, Greece.

Jia Xu, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. 2008. Bayesian semi-supervised chinese word segmentation for statistical machine translation. In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, pages 1017–1024, Morristown, NJ, USA. Association for Computational Linguistics.

Xiaojin Zhu. 2005. *Semi-supervised Learning with Graphs*. Ph.D. thesis, CMU. Chapter 11, Semi-supervised learning literature survey (updated online version).