# Overview of Morpho Challenge in CLEF 2007

Mikko Kurimo, Mathias Creutz, Ville Turunen
Adaptive Informatics Research Centre, Helsinki University of Technology
P.O.Box 5400, FIN-02015 TKK, Finland
`Mikko.Kurimo@tkk.fi`

**Abstract**

Morpho Challenge 2007 contained an evaluation of unsupervised morpheme analysis algorithms using information retrieval experiments utilizing data available in CLEF. The objective of the challenge was to design statistical machine learning algorithms that discover which morphemes (smallest individually meaningful units of language) words consist of. Ideally, these are basic vocabulary units suitable for different tasks, such as text understanding, machine translation, information retrieval, and statistical language modeling The evaluation of the submitted morpheme analysis was performed by two complementary ways: *Competition 1:* The proposed morpheme analyses were compared to a linguistic morpheme analysis gold standard by matching the morpheme-sharing word pairs. *Competition 2:* Information retrieval (IR) experiments were performed, where the words in the documents and queries were replaced by their proposed morpheme representations and the search was based on morphemes instead of words. This paper provides an overview of the IR evaluation. The IR evaluations were provided for Finnish, German, and English and participants were encouraged to apply their algorithm to all of them. The organizers performed the IR experiments using the queries, texts, and relevance judgments available in CLEF forum and morpheme analysis methods submitted by the challenge participants. The results show that the morpheme analysis has a significant effect in IR performance in all languages, and that the performance of the best unsupervised methods can be superior to the supervised reference methods. The challenge was part of the EU Network of Excellence PASCAL Challenge Program and organized in collaboration with CLEF.

## Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

## General Terms

Algorithms, Performance, Experimentation

## Keywords

Morphological analysis, Machine learning

# 1 Introduction

The scientific objectives of the Morpho Challenge 2007 were: to learn of the phenomena underlying word construction in natural languages, to advance machine learning methodology, and to discover approaches suitable for a wide range of languages. The suitability for a wide range of languages is becoming increasingly important, because language technology methods need to be quickly and as automatically as possible extended to new languages that have limited previous resources. That is why learning the morpheme analysis directly from large text corpora using unsupervised machine learning algorithms is such an attractive approach and a very relevant research topic today.

The problem of learning the morphemes directly from large text corpora using an unsupervised machine learning algorithm is clearly a difficult one. First the words should be somehow segmented into meaningful parts, and then these parts should be clustered in the abstract classes of morphemes that would be useful for modeling. It is also challenging to learn to generalize the analysis to rare words, because even the largest text corpora are very sparse, a significant portion of the words may occur only once. Many important words, for example proper names and their inflections or some forms of long compound words, may also not exist in the training material at all, and their analysis is often even more challenging. However, benefits for successful morpheme analysis, in addition to obtaining a set of basic vocabulary units for modeling, can be seen for many important tasks in language technology. The additional information included in the units can provide support for building more sophisticated language models, for example, in speech recognition [1], machine translation [10], and information retrieval [13].

The evaluation of the unsupervised morpheme analysis was in this challenge solved by developing two complementary evaluations, one including a comparison to linguistic morpheme analysis gold standard, and another including a practical real-world application where morpheme analysis might be useful. This paper presents an overview how the application-oriented evaluation called *Competition 2* was performed in the domain of finding useful index terms for information retrieval tasks in multiple languages using the queries, texts, and relevance judgments available in CLEF forum and morpheme analysis methods submitted by the challenge participants. The linguistic evaluation called *Competition 1* are described in [8] and Competition 2 in more detail in [7].

Traditionally, and especially in processing English texts, stemming algorithms have been used to reduce the different infected word forms into the common roots or stems for indexing. However, to achieve best results when ported to new languages the development of stemming algorithms requires a considerable amount of special development work. In many highly-inflecting, compounding, and agglutinative European languages the amount of different word forms is huge and the task of extracting the useful index terms becomes both more complex and more important.

The same IR tasks that were attempted using the Morpho Challenge participants' morpheme analysis, were also tested by a number of reference methods to see how the unsupervised morpheme analysis performed in comparison to them. These references included the organizers' public Morfessor Categories-Map [3] and Morfessor Baseline [2, 4], the Morfessor analysis improved by a hybrid method [12], grammatical morpheme analysis based on the linguistic gold standards [5], the traditional Porter stemming [11] of words and also by the words as such without any processing.

## 2 Task

Morpho Challenge 2007 is a follow-up to our previous Morpho Challenge 2005 (Unsupervised Segmentation of Words into Morphemes) [9]. In Morpho Challenge 2005 the focus was in the segmentation of data into units that are useful for statistical modeling. The specific task for the competition was to design an unsupervised statistical machine learning algorithm that segments words into the smallest meaning-bearing units of language, morphemes. In addition to comparing the obtained morphemes to a linguistic "gold standard", their usefulness was evaluated by using them for training statistical language models for speech recognition.

In Morpho Challenge 2007 a more general focus was chosen to not only to segment words into smaller units, but also to perform *morpheme analysis* of the word forms in the data. For instance, the English words "boot, boots, foot, feet" might obtain the analyses "boot, boot + plural, foot, foot + plural", respectively. In linguistics, the concept of morpheme does not necessarily directly correspond to a particular word segment but to an abstract class. For some languages there exist carefully constructed linguistic tools for this kind of analysis, although not for many, but using statistical machine learning methods we may still discover interesting alternatives that may rival even the most careful linguistically designed morphologies.

## 3 Training data

The Morpho Challenge 2007 task, in practice, was to return the unsupervised morpheme analysis of every word form contained in a long word list supplied by the organizers for each test language [8]. The participants were pointed to corpora [8] in which the words occur, so that the algorithms may utilize information about word context. The text corpora where the word list were collected were obtained from the Wortschatz collection[1]. at the University of Leipzig (Germany). We used the plain text files (sentences.txt for each language); the corpus sizes are 3 million sentences for English, Finnish and German, and 1 million sentences for Turkish. For English, Finnish and Turkish we used preliminary corpora, which have not yet been released publicly at the Wortschatz site. The corpora were specially preprocessed for the Morpho Challenge (tokenized, lower-cased, some conversion of character encodings).

To achieve the goal of designing language independent methods, the participants were encouraged to submit results in all test languages. The information retrieval (IR) experiments were performed by the organizers based on the morpheme analyses submitted by the participants.

## 4 IR evaluation data

The data sets for testing the IR performance in each test language consisted of news paper articles as the source documents, test queries and the binary relevance judgments regarding to the queries. The organizers performed the IR experiments based on the morpheme analyses submitted by the participants, so it was not necessary for the participants to get these data sets. However, all the data was available for registered participants in the Cross-Language Evaluation Forum (CLEF)[2].

The source documents were news articles collected from different newspapers selected as follows:

- In Finnish: 55K documents from short articles in Aamulehti 1994-95, 50 test queries on specific news topics and 23K binary relevance assessments (CLEF 2004)
- In English: 170K documents from short articles in Los Angeles Times 1994 and Glasgow Herald 1995, 50 test queries on specific news topics and 20K binary relevance assessments (CLEF 2005).
- In German: 300K documents from short articles in Frankfurter Rundschau 1994, Der Spiegel 1994-95 and SDA German 1994-95, 60 test queries with 23K binary relevance assessments (CLEF 2003).

When performing the indexing and retrieval experiments for Competition 2, it turned out that the test data contained quite many new words in addition to those that were provided as training data for the Competition 1 [8]. Thus, the participants were offered a chance to improve the retrieval results of their morpheme analyses by providing them a list of the new words found in all test languages. The participants then had the choice to either run their algorithms to analyze as many of the new words as they could or liked, or to provide no extra analyses. No text data resources to find context for the new words were provided, but it was made possible to register to CLEF to use the text data available in there or any other data the participants could get.

## 5 Participants and their submissions

By the deadline in May, 2007, 6 research groups had submitted the segmentation results obtained by their algorithms. A total of 12 different algorithms were submitted, 8 of them ran experiments

[1] `http://corpora.informatik.uni-leipzig.de/`
[2] `http://www.clef-campaign.org/`

Table 1: The submitted algorithms.

| Algorithm | Authors | Affiliation |
|---|---|---|
| “Bernhard 1” | Delphine Bernhard | TIMC-IMAG, F |
| “Bernhard 2” | Delphine Bernhard | TIMC-IMAG, F |
| “Bordag 5” | Stefan Bordag | Univ. Leipzig, D |
| “Bordag 5a” | Stefan Bordag | Univ. Leipzig, D |
| “McNamee 3” | Paul McNamee and James Mayfield | JHU, USA |
| “McNamee 4” | Paul McNamee and James Mayfield | JHU, USA |
| “McNamee 5” | Paul McNamee and James Mayfield | JHU, USA |
| “Zeman ” | Daniel Zeman | Karlova Univ., CZ |
| “Monson Morfessor” | Christian Monson et al. | CMU, USA |
| “Monson ParaMor” | Christian Monson et al. | CMU, USA |
| “Monson ParaMor-Morfessor” | Christian Monson et al. | CMU, USA |
| “Pitler” | Emily Pitler and Samarth Keshava | Univ. Yale, USA |
| “Morfessor Categories-MAP” | The organizers | Helsinki Univ. Tech, FI |
| “Morfessor Baseline” | The organizers | Helsinki Univ. Tech, FI |
| “dummy” | The organizers | Helsinki Univ. Tech, FI |
| “grammatical” | The organizers | Helsinki Univ. Tech, FI |
| “Porter” | The organizers | Helsinki Univ. Tech, FI |
| “Tepper” | Michael Tepper | Univ. Washington, USA |

on all four test languages. All the submitted algorithms are listed in Table 1. In addition to the competitors' 12 morpheme analysis algorithms, we evaluated a number of reference methods:

1. Public baseline methods called “Morfessor Baseline” and “Morfessor Categories-MAP” (or here just “Morfessor MAP”) developed by the organizers [3].

2. No words were split nor any morpheme analysis provided, “dummy”.

3. The words were analyzed using the gold standard in each language that were utilized as the “ground truth” in the Competition 1 [8]. Besides the stems and suffixes, the gold standard analyses typically consist of all kinds of grammatical tags which we decided to simply include as index terms, as well. “grammatical first” uses only the first interpretation of each word whereas “grammatical all” use all.

4. *Porter*: No real morpheme analysis was performed, but the words were stemmed by the Porter stemming, an option provided by the Lemur toolkit.

5. *Tepper*: A hybrid method developed by Michael Tepper [12] was utilized to improve the morpheme analysis reference obtained by our Morfessor Categories-MAP.

The outputs of the submitted algorithms are analyzed closer in [8]. From the IR point of view it is interesting to note that only Monson and Zeman decided to provide several alternative analysis for most words instead of just the most likely one. McNamee's algorithms did not attempt to provide a real morpheme analysis, but focused directly on finding a representative substring for each word type that would be likely to perform well in the IR evaluation.

# 6 Evaluation

In this evaluation, the organizers applied the analyses provided by the participants in information retrieval experiments. The words in the queries and source documents were replaced by the corresponding morpheme analyses provided by the participants, and the search was then based on morphemes instead of words.

The evaluation was performed using a state-of-the-art retrieval method (the latest version of the freely available LEMUR toolkit[3]. We utilized two standard retrieval method: Tfidf and Okapi term weighting. The Tfidf implementation in LEMUR applies term frequency weights for both query and document based on the BM25 weighting and the Euclidean dot-product as similarity measure. Okapi in LEMUR is an implementation of the BM25 retrieval function as described in [6].

The evaluation criterion was Uninterpolated Average Precision There were several different categories and the winner with the highest Average Precision was selected separately for each language and each category:

1. All morpheme analyses from the training data are used as index terms *"withoutnew"* vs. additionally using also the morpheme analyses for new words that existed in the IR data but not in the training data *"withnew"*.

2. Tfidf term weighting was utilized for all index terms without any stoplists vs. Okapi term weighting for all index terms excluding an automatic stoplist consisting of the most common terms (frequency threshold was 75,000 for Finnish and 150,000 for German and English). The stoplist was developed for the Okapi weighting, because otherwise Okapi weights were not suitable for the indexes that had many very common terms.

# 7 Results

The results of the information retrieval evaluations are shown in Table 2. Here we have selected only the best runs from each participant (in bold) and reference method. For the full results see [7]. Indexing is performed using Tfidf weighting for all morphemes (left) and Okapi weighting for all morphemes except the most common ones (stoplist) with frequency higher than 150,000 (right).

In the Finnish task, the highest average precision was obtained by the "Bernhard 2" algorithm, which was also won the Competition 1 [8]. The highest average precision 0.49 was obtained using the Okapi weighting and stoplist for both the originally submitted morpheme analysis (for Competition 1) and the morpheme analysis for the new words added for Competition 2. The "Bernhard 1" algorithm obtained the highest average precision 0.47 for the German task using the new words, Okapi and stoplist. For English, the highest average precision was obtained by the "Bernhard 2" algorithm, which was also won the Competition 1 [8]. As in Finnish and German, the highest average precision 0.39 was obtained with the new words and using the Okapi weighting and stoplist.

As expected, the "grammatical" reference method based on linguistic Gold Standard morpheme analysis [8] did not perform very well. However, with stoplist and Okapi term weighting it did achieve better results than the "dummy" method in all languages. In Finnish and English the performance was better than average, but quite poor in German. The "grammatical first" that utilized only the first of the alternative analysis in indexing was at least as good or better than the "grammatical all", which seems to indicate that the alternative analysis are not very useful here.

For the "Morfessor" references it is interesting to note that they always performed better than the "grammatical", which seems to suggest that the coverage of the analysis ("Morfessor" does not have any out-of-vocabulary words) is more important for IR than the grammatical correctness. In general, the old "Morfessor Baseline" seems to provide a very good baseline in all tested languages also for the IR tasks as it did for the language modeling and speech recognition in [9].

[3]`http://www.lemurproject.org/`

Table 2: The obtained average precision (AP%) in the information retrieval task for the best submitted segmentation for each participant and reference method.

**Finnish:**

| Tfidf weighting for all morphemes | | | Okapi weighting and a stoplist | | |
|---|---|---|---|---|---|
| METHOD | WORDLIST | AP% | METHOD | WORDLIST | AP% |
| Morfessor baseline | withnew | 0.4105 | **Bernhard** 2 | withnew | 0.4915 |
| **Bernhard** 1 | withoutnew | 0.4016 | Morfessor baseline | withnew | 0.4412 |
| grammatical first | withoutnew | 0.3995 | **Bordag** 5a | withnew | 0.4309 |
| **Bordag** 5 | withnew | 0.3831 | grammatical all | withoutnew | 0.4307 |
| **McNamee** 5 | withoutnew | 0.3646 | **McNamee** 5 | withnew | 0.3684 |
| Porter | withnew | 0.3566 | Porter | withnew | 0.3517 |
| dummy | withnew | 0.3559 | dummy | withnew | 0.3274 |
| **Zeman** | withoutnew | 0.2494 | **Zeman** | withoutnew | 0.2813 |

**German:**

| Tfidf weighting for all morphemes | | | Okapi weighting and a stoplist | | |
|---|---|---|---|---|---|
| METHOD | WORDLIST | AP% | METHOD | WORDLIST | AP% |
| Morfessor baseline | withnew | 0.3874 | **Bernhard** 1 | withnew | 0.4729 |
| **Bernhard** 1 | withoutnew | 0.3777 | **Monson** Morfessor | withnew | 0.4602 |
| Porter | withnew | 0.3725 | Morfessor MAP | withnew | 0.4571 |
| **Monson** Morfessor | withnew | 0.3520 | **Bordag** 5 | withnew | 0.4308 |
| dummy | withnew | 0.3496 | Porter | withnew | 0.3866 |
| **Bordag** 5a | withnew | 0.3496 | **McNamee** 5 | withoutnew | 0.3617 |
| **McNamee** 5 | withoutnew | 0.3442 | grammatical first | withoutnew | 0.3467 |
| grammatical first | withoutnew | 0.3223 | dummy | withnew | 0.3228 |
| **Zeman** | withoutnew | 0.2828 | **Zeman** | withoutnew | 0.2568 |

**English:**

| Tfidf weighting for all morphemes | | | Okapi weighting and a stoplist | | |
|---|---|---|---|---|---|
| METHOD | WORDLIST | AP% | METHOD | WORDLIST | AP% |
| Porter | withnew | 0.3052 | Porter | withnew | 0.4083 |
| **McNamee** 5 | withoutnew | 0.2888 | **Bernhard** 2 | withnew | 0.3943 |
| Morfessor baseline | withnew | 0.2863 | Morfessor baseline | withnew | 0.3882 |
| Tepper | withoutnew | 0.2784 | grammatical first | withoutnew | 0.3774 |
| dummy | withnew | 0.2783 | Tepper | withoutnew | 0.3728 |
| **Bernhard** 1 | withoutnew | 0.2781 | **Monson** Morfessor | withoutnew | 0.3721 |
| **Monson** Morfessor | withoutnew | 0.2676 | **Pitler** | withoutnew | 0.3652 |
| **Pitler** | withoutnew | 0.2666 | **McNamee** 4 | withoutnew | 0.3577 |
| grammatical all | withoutnew | 0.2619 | **Bordag** 5 | withoutnew | 0.3427 |
| **Zeman** | withoutnew | 0.2297 | dummy | withnew | 0.3123 |
| **Bordag** 5 | withoutnew | 0.2210 | **Zeman** | withoutnew | 0.2674 |

# 8 Discussions

The comparison of the results in the Tfidf and Okapi categories show that the Okapi with stoplist performed significantly better for all languages. We also run Tfidf with stoplist (the results not included here) which achieved results that were better than the plain Tfidf and only slightly inferior to Okapi with stoplist. However, we decided to rather report the original Tfidf, since we wanted to show what is the performance and the relative ranking of the methods without the stoplist.

The Porter stemming that is a standard word preprocessing tool in IR remained unbeaten (by a narrow margin) in our evaluations in English, but in German and especially in Finnish, the unsupervised morpheme analysis methods clearly dominated the evaluation. There might exist better stemming algorithms for those languages, but because of the more complex morphology, their development might not be an easy task.

As future work in this field it should be relatively straight-forward to evaluate the unsupervised morpheme analysis in several other interesting languages, because it is not bounded to only those languages where rule-based grammatical analysis can be performed. It would also be interesting to try to combine the rival analysis to produce something better.

# 9 Conclusions

The objective of Morpho Challenge 2007 was to design a statistical machine learning algorithm that discovers which morphemes (smallest individually meaningful units of language) words consist of. Ideally, these are basic vocabulary units suitable for different tasks, such as text understanding, machine translation, information retrieval, and statistical language modeling. The current challenge was a successful follow-up to our previous Morpho Challenge 2005 (Unsupervised Segmentation of Words into Morphemes). This time the task was more general in that instead of looking for an explicit segmentation of words, the focus was in the morpheme analysis of the word forms in the data.

The scientific goals of this challenge were to learn of the phenomena underlying word construction in natural languages, to discover approaches suitable for a wide range of languages and to advance machine learning methodology. The analysis and evaluation of the submitted machine learning algorithm for unsupervised morpheme analysis showed that these goals were quite nicely met. There were several novel unsupervised methods that achieved good results in several test languages, both with respect to finding meaningful morphemes and useful units for information retrieval. The IR results also revealed that the morpheme analysis has a significant effect in IR performance in all languages, and that the performance of the best unsupervised methods can be superior to the supervised reference methods.

# Acknowledgments

We thank all the participants for their submissions and enthusiasm. We owe great thanks as well to the organizers of the PASCAL Challenge Program and CLEF who helped us organize this challenge and the challenge workshop. Especially, we would like to thank Carol Peters from CLEF for helping us to get Morpho Challenge in CLEF 2007 and organize a great workshop there. Our work was supported by the Academy of Finland in the projects *Adaptive Informatics* and *New adaptive and learning methods in speech recognition*. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. We acknowledge that access rights to data and other materials are restricted due to other commitments.

## References


[1] Jeff A. Bilmes and Katrin Kirchhoff. Factored language models and generalized parallel backoff. In *Proceedings of the Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 4–6, Edmonton, Canada, 2003.

[2] Mathias Creutz and Krista Lagus. Unsupervised discovery of morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, pages 21–30, 2002.

[3] Mathias Creutz and Krista Lagus. 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)*, pages 106–113, Espoo, Finland, 2005.

[4] Mathias Creutz and Krista Lagus. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology, 2005. URL: http://www.cis.hut.fi/projects/morpho/.

[5] Mathias Creutz and Krister Linden. Morpheme segmentation gold standards for finnish and english. Technical Report A77, Publications in Computer and Information Science, Helsinki University of Technology, 2004. URL: http://www.cis.hut.fi/projects/morpho/.

[6] S. Robertson et al. Okapi at TREC-3. In *Proceedings of the Third Text Retrieval Conference (TREC-3)*, pages 109–126, 1994.

[7] Mikko Kurimo, Mathias Creutz, and Ville Turunen. Unsupervised morpheme analysis evaluation by IR experiments – Morpho Challenge 2007. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 2007.

[8] Mikko Kurimo, Mathias Creutz, and Matti Varjokallio. Unsupervised morpheme analysis evaluation by a comparison to a linguistic Gold Standard – Morpho Challenge 2007. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 2007.

[9] Mikko Kurimo, Mathias Creutz, Matti Varjokallio, Ebru Arisoy, and Murat Saraclar. Unsupervised segmentation of words into morphemes - Challenge 2005, an introduction and evaluation report. In *PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*, Venice, Italy, 2006.

[10] Y.-S. Lee. Morphological analysis for statistical machine translation. In *Proceedings of the Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Boston, MA, USA, 2004.

[11] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.

[12] Michael Tepper. *A Hybrid Approach to the Induction of Underlying Morphology*. PhD thesis, University of Washington, 2007.

[13] Y.L. Zieman and H.L. Bleich. Conceptual mapping of user's queries to medical subject headings. In *Proceedings of the 1997 American Medical Informatics Association (AMIA) Annual Fall Symposium*, October 1997.