Unsupervised Morpheme Analysis Evaluation by a Comparison to a Linguistic Gold Standard – Morpho Challenge 2007

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

Abstract

This paper presents the evaluation of Morpho Challenge Competition 1 (linguistic gold standard). The Competition 2 (information retrieval) is described in a companion paper. In Morpho Challenge 2007, the objective 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 choice of a meaningful evaluation for the submitted morpheme analysis was not straight-forward, because in unsupervised morpheme analysis the morphemes can have arbitrary names. Two complementary ways were developed for the evaluation: *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. Data sets for Competition 1 were provided for four languages: Finnish, German, English, and Turkish and the participants were encouraged to apply their algorithm to all of them. The results show significant variance between the methods and languages, but the best methods seem to be useful in all tested languages and match quite well with the linguistic gold standard. The Morpho 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.

Morpho Challenge 2007 is a follow-up to our previous Morpho Challenge 2005 (Unsupervised Segmentation of Words into Morphemes) [7]. 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.

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 [8], and information retrieval [10].

The problem of how to arrange a meaningful evaluation of the unsupervised morpheme analysis algorithms is not straight-forward, because in unsupervised morpheme analysis the morphemes can be called by arbitrary names which are not likely to directly correspond to the linguistic morpheme definitions. In this challenge we solved this 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 used. In the first evaluation, called *Competition 1*, the proposed morpheme analyses were compared to a linguistic gold standard citecreutz04.tr by counting the matching morpheme-sharing word pairs. In this way we did not have to try to match the names of the morphemes directly, but only to measure if the proposed algorithm can find the correct word pairs that share common morphemes. The second evaluation called *Competition 2* involved performing information retrieval (IR) experiments using the data of the state of art CLEF evaluation, 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 presents the Competition 1 and the Competition 2 is described in a companion paper [6].

2 Task

The Morpho Challenge 2007 task was set to return the unsupervised morpheme analysis of every word form contained in a long word list supplied by the organizers for each test language. The participants were pointed to corpora in which the words occur, so that the algorithms may utilize information about word context.

In the Morpho Challenge 2005 the morphological segmentation evaluations were performed for three languages: Finnish, English, and Turkish. Now a data set and evaluation were provided for one new text language, German. To achieve the goal of designing language independent methods, the participants were encouraged to submit results in all these languages. Having the theme of unsupervised machine learning, the participants were required to describe any supervision or parameter optimization steps that were taken in the algorithms. The participants did not need to worry about which names to use for the morphemes they discovered, because the evaluation was performed just by the F-measure of matching accuracy of the morpheme-sharing word pairs.

3 Data sets

The first and foremost type of data files were the word lists. The words had been extracted from a text corpus, and each word in the list was preceded by its frequency in the corpus used. For instance, a subset of the supplied English word list looked like this:

```
1 barefoot's
2 barefooted
6699 feet
653 flies
2939 flying
1782 foot
64 footprints
```

The result files that the participants' task was to return, were lists containing exactly the same words as in the input, with morpheme analyses provided for each word. Submission for the above English words might have looked like this:

```
barefoot's BARE FOOT +GEN
barefooted BARE FOOT +PAST
feet FOOT +PL
flies FLY_N +PL, FLY_V +3SG
flying FLY_V +PCP1
foot FOOT
footprints FOOT PRINT +PL
```

The order in which the morpheme labels appeared after the word forms does not matter; e.g., "FOOT +PL" is equivalent to "+PL FOOT". As the learning is unsupervised, the labels are arbitrary: e.g., instead of using "FOOT" one might use "morpheme784" and instead of "+PL" one might use "morpheme2". However, intuitive labels are preferable, because it becomes easier for anyone to get an idea of the quality of the result by looking at it.

If a word has several interpretations, all interpretations can be supplied: e.g., the word "flies" may be the plural form of the noun "fly" (insect) or the third person singular present tense form of the verb "to fly".Thus the analysis could be as: "FLY_N +PL, FLY_V +3SG". The existence of alternative analyses makes the task challenging, and it was left to the participants to decide how much effort they put into this aspect of the task. In English, for instance, in order to get a perfect score, it would be necessary to distinguish the different functions of the ending "-s" (plural or person ending) as well as the different parts-of-speech of the stem "fly" (noun or verb). As the results will be evaluated against reference analyses (our so-called gold standard), the guiding principles used when constructing the gold standard will be explained in Section 4.

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).

4 Gold standard morpheme analyses

The gold standard morpheme analyses are the correct grammatical morpheme analyses that were used as reference in the evaluation. The gold standard morpheme analyses were prepared in exactly the same format as that of the result file the participants were asked to submit. Because there are multiple correct analysis for some words, the alternative analyses are separated by a comma. See Table 1 for examples.

Table 1: Examples of gold standard morpheme analyses.

Language	Examples	
English	baby-sitters	baby_N sit_V er_s +PL
	indoctrinated	in_p doctrine_N ate_s +PAST
Finnish	linuxiin	linux_N +ILL
	makaronia	makaroni_N +PTV
German	choreographische	choreographie_N isch +ADJ-e
	zurueckzubehalten	zurueck_B zu be halt_V +INF
Turkish	kontrole	kontrol +DAT
	popUlerliGini	popUler +DER_lHg +POS2S +ACC, popUler +DER_lHg +POS3 +ACC3

The English and German gold standards are based on the CELEX data base[2]. The Finnish gold standard is based on the two-level morphology analyzer FINTWOL from Lingsoft[3], Inc. The Turkish gold-standard analyses have been obtained from a morphological parser developed at Bogazici University[4] [2, 5]; it is based on Oflazer's finite-state machines, with a number of changes.

The morphological analyses are *morpheme* analyses. This means that only grammatical categories that are realized as morphemes are included. For instance, for none of the languages there is a singular morpheme for nouns or a present-tense morpheme for verbs, because these grammatical categories do not alter or add anything to the word form. This is in contrast to, e.g., the plural form of a noun (house vs. house+s), or the past tense of verbs (help vs. help+ed, come vs. came).

The morpheme labels that correspond to inflectional (and sometimes also derivational) affixes have been marked with an initial plus sign (e.g., +PL, +PAST). This is due to a feature of the evaluation script: in addition to the overall performance statistics, evaluation measures are also computed separately for the labels starting with a plus sign and those without an initial plus sign. It is thus possible to make an approximate assessment of how accurately affixes are analyzed vs. non-affixes (mostly stems).

The morpheme labels that have not been marked as affixes (no initial plus sign) are typically stems. These labels consist of an intuitive string, usually followed by an underscore character (_) and a part-of-speech tag, e.g., "baby_N", "sit_V". In many cases, especially in English, the same morpheme can function as different parts-of-speech; e.g., the English word "force" can be a noun or a verb. In the majority of these cases, however, if there is only a difference in syntax (and not in meaning), the morpheme has been labeled as either a noun or a verb, throughout. For instance,

[1] http://corpora.informatik.uni-leipzig.de/
[2] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC96L14
[3] http://www.lingsoft.fi/
[4] http://www.boun.edu.tr/index_eng.html

the "original" part-of-speech of "force" is a noun, and consequently both noun and verb inflections of "force" contain the morpheme "force_N":

```
force force_N
force's force_N GEN
forced force_N +PAST
forces force_N +3SG, force_N +PL
forcing force_N +PCP1
```

Thus, there is not really a need for the participant's algorithm to distinguish between different meanings or syntactic roles of the discovered stem morphemes. However, in some rare cases, if the meanings of the different parts-of-speech do differ clearly, there are two variants, e.g., "train_N" (vehicle), "train_V" (to teach), "fly_N" (insect), "fly_V" (to move through the air). But again, if there are ambiguous meanings within the same part-of-speech, these are not marked in any way, e.g., "fan_N" (device for producing a current of air) vs. "fan_N" (admirer). This notation is a consequence of using CELEX and FINTWOL as the sources for our gold standards. We could have removed the part-of-speech tags, but we decided to leave them there, since they carry useful information without significantly making the task more difficult. There are no part-of-speech tags in the Turkish gold standard.

5 Participants and their submissions

Table 2: The submitted algorithms and reference methods

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 MAP"	The organizers	Helsinki Univ. Tech, FI
"Tepper"	Michael Tepper	Univ. Washington, USA
"Gold Standard"	The organizers	Helsinki Univ. Tech, FI

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 on all four test languages. All the submitted algorithms are listed in Table 2. In general, the submissions were all interesting and all of them met the exact specifications given and were able to get properly evaluated.

In addition to the competitors' 12 morpheme analysis algorithms, we evaluated a public baseline method called "Morfessor Categories-MAP" (or here just "Morfessor MAP" or "Morfessor", for short) developed by the organizers [3]. Naturally, the Morfessor competed outside the main competition and the results were included only as reference.

Tables 3 - 6 show an example analysis and some statistics of each submission including the average amount of alternative analysis per word, the average amount of morphemes per analy-

Table 3: Statistics and example morpheme analyses in **Finnish**. #anal is the average amount of analysis per word (separated by a comma), #morph the average amount of morphemes per analysis (separated by a space), and lexicon the total amount of morpheme types.

Algorithm	Example word: linuxiin	#anal	#morph	lexicon
Bernhard 1	linux_B iin_S	1	3.16	87590
Bernhard 2	linux_B i_S in_S	1	3.89	87915
Bordag 5	linuxiin	1	2.84	517091
Bordag 5a	linuxia.linuxiin	1	2.84	514670
McNamee 3	xii	1	1	20063
McNamee 4	uxii	1	1	137629
McNamee 5	nuxii	1	1	435814
Zeman	linuxiin, linuxii n, linuxi in, linux iin	3.62	1.81	5434453
Morfessor MAP	linux +iin	1	2.94	217001
Gold Standard	linux_N +ILL	1.16	3.29	33754

Table 4: Statistics and example morpheme analyses in **Turkish**. #anal is the average amount of analysis per word (separated by a comma), #morph the average amount of morphemes per analysis (separated by a space), and lexicon the total amount of morpheme types.

Algorithm	Example word: popUlerliGini	#anal	#morph	lexicon
Bernhard 1	popUler_B liGini_S	1	2.48	86490
Bernhard 2	popUler_B liGini_S	1	2.73	87637
Bordag 5	popUlerliGini	1	2.24	219488
Bordag 5a	popUlerliGini	1	2.24	219864
McNamee 3	opU	1	1	19389
McNamee 4	pUle	1	1	102515
McNamee 5	Ulerl	1	1	226698
Zeman	popUlerliGin i, popUlerliGi ni	3.24	1.76	1205970
Morfessor MAP	pop +U +ler +liGini	1	2.64	114834
Tepper	popU lEr lWK W W	1	2.81	110682
Gold Standard	popUler +DER_lHg +POS2S +ACC, popUler +DER_lHg +POS3 +ACC3	1.99	3.36	21163

sis, and the total amount of morpheme types. The total amount of word types were 2,206,719 (Finnish), 617,298 (Turkish), 1,266,159 (German), and 384,903 (English). The Turkish word list was extracted in 1 million sentences, but the other lists from 3 million sentences per each language. In these word lists, the Gold Standard analysis were available for 650,169 (Finnish), 214,818 (Turkish), 125,641 (German), and 63,225 (English) words.

The algorthms by Bernhard, Bordag and Pitler were the same or improved versions from the previous Morpho Challenge [7]. Monson and Zeman were new participants who also provided several alternative analysis for most words. The most different approach was McNamee's algorithm, which did not attempt to provide a real morpheme analysis, but mainly to find a representative substring for each word type that would be likely to perform well in the IR evaluation (our Competition 2 [6]). Noteworthy in Tables 3 - 6 is also that the size of the morpheme lexicon varies a lot in different algorithms.

Table 5: Statistics and example morpheme analyses in **German**. #anal is the average amount of analysis per word (separated by a comma), #morph the average amount of morphemes per analysis (separated by a space), and lexicon the total amount of morpheme types.

Algorithm	Example word: zurueckzubehalten	#anal	#morph	lexicon
Bernhard 1	zurueckzu_P behalt_B en_S	1	3.12	56173
Bernhard 2	zurueckzu_P behalt_B e_S n_S	1	3.72	53497
Bordag 5	zu rueck zu be halt en	1	2.99	267680
Bordag 5a	zu rueck zu be ehalt.hale.halt.halte.helt en	1	2.99	266924
McNamee 3	kzu	1	1	16633
McNamee 4	kzub	1	1	130802
McNamee 5	kzube	1	1	434152
Zeman	zurueckzubehalten, zurueckzubehalte n, ...	4.15	1.81	4094228
Monson Paramor-M	+zurueck/PRE +zu/PRE +be/PRE halten/STM, zurueckzub +ehalten, zurueckzube +halten	2.91	2.20	1191842
Monson ParaMor	zurueckzub +ehalten, zurueckzube +halten	1.91	1.72	1001441
Monson Morfessor	+zurueck/PRE +zu/PRE +be/PRE halten/STM	1	3.10	166963
Morfessor MAP	zurueck zu be halten	1	3.06	172907
Gold Standard	zurueck_B zu be halt_V +INF	1.30	2.97	14298

6 Evaluation

For each language, the morpheme analyses proposed by the participants' algorithm were compared against the linguistic gold standard. Since the task at hand involves unsupervised learning, it cannot be expected that the algorithm comes up with morpheme labels that exactly correspond to the ones designed by linguists. That is, no direct comparison will take place between labels as such (the labels in the proposed analyses vs. labels in the gold standard). What can be expected, however, is that two word forms that contain the same morpheme according to the participants' algorithm also have a morpheme in common according to the gold standard. For instance, in the English gold standard, the words "foot" and "feet" both contain the morpheme "foot_N". It is thus desirable that also the participants' algorithm discovers a morpheme that occurs in both these word forms (be it called "FOOT", "morpheme784", "foot" or something else).

In practice, the evaluation took place by randomly sampling a large number of word pairs, such that both words in the pair have at least one morpheme in common. The exact constitution of this set of word pairs was not revealed to the participants. In the evaluation, word frequency played no role. Thus, all word pairs were equally important, whether they were frequent or rare. The size of the randomly chosen set of word pairs set varied depending on the size of the word lists and Gold Standard given in the previous section: 200,000 (Finnish), 50,000 (Turkish), 50,000 (German), and 10,000 (English) word pairs.

As the evaluation measure, we applied *F-measure*, which is the harmonic mean of *Precision* and *Recall*:

$$\text{F-measure} = 1/(1/\text{Precision} + 1/\text{Recall})\,. \tag{1}$$

Precision is here calculated as follows: A number of word forms will be randomly sampled from the result file provided by the participants; for each morpheme in these words, another word containing the same morpheme will be chosen from the result file by random (if such a word exists). We thus obtain a number of word pairs such that in each pair at least one morpheme is shared between the words in the pair. These pairs will be compared to the gold standard; a point is given for each word pair that really has a morpheme in common according to the gold standard. The total number of points is then divided by the total number of word pairs.

Table 6: Statistics and example morpheme analyses in **English**. #anal is the average amount of analysis per word (separated by a comma), #morph the average amount of morphemes per analysis (separated by a space), and lexicon the total amount of morpheme types.

Algorithm	Example word: baby-sitters	#anal	#morph	lexicon
Bernhard 1	baby_P -_L sitt_B er_S s_S	1	2.61	55490
Bernhard 2	baby_P -_L sitt_B er_S s_S	1	2.90	52582
Bordag 5	baby sitters	1	1.97	190094
Bordag 5a	baby sitters	1	1.97	189568
McNamee 3	by-	1	1	15212
McNamee 4	aby-	1	1	98475
McNamee 5	y-sit	1	1	243578
Zeman	baby-sitter s, baby-sitt ers	3.18	1.74	905251
Monson Paramor-M	+baby-/PRE sitter/STM +s/SUF, bab +y, sit +ters, sitt +ers, sitte +rs, sitter +s	3.42	1.93	386257
Monson ParaMor	bab +y, sit +ters, sitt +ers, sitte +rs, sitter +s	2.42	1.88	233981
Monson Morfessor	+baby-/PRE sitter/STM +s/SUF	1	2.07	137973
Pitler	baby- sitt ers	1	1.57	211475
Morfessor MAP	baby - sitters	1	2.12	132086
Tepper	baby - sit ers	1	2.53	99937
Gold Standard	baby_N sit_V er_s +PL	1.10	2.13	16902

For instance, assume that the proposed analysis of the English word "abyss" is: "abys +s". Two word pairs are formed: Say that "abyss" happens to share the morpheme "abys" with the word "abysses"; we thus obtain the word pair "abyss - abysses". Also assume that "abyss" shares the morpheme "+s" with the word "mountains"; this produces the pair "abyss - mountains". Now, according to the gold standard the correct analyses of these words are: "abyss_N", "abyss_N +PL", "mountain_N +PL", respectively. The pair "abyss - abysses" is correct (common morpheme: "abyss+_N"), but the pair "abyss - mountain" is incorrect (no morpheme in common). Precision here is thus $1/2 = 50$

Recall is calculated analogously to precision: A number of word forms are randomly sampled from the gold standard file; for each morpheme in these words, another word containing the same morpheme will be chosen from the gold standard by random (if such a word exists). The word pairs are then compared to the analyses provided by the participants; a point is given for each sampled word pair that has a morpheme in common also in the analyses proposed by the participants' algorithm. The total number of points is then divided by the total number of sampled word pairs.

For words that have several alternative analyses, as well as for word pairs that have more than one morpheme in common, the normalization of the points is carried out in order not to give these words considerably more weight in the evaluation than "less complex" words. The words are normalized by the number of alternative analyses and the word pairs by the number of matching morphemes. Details of the evaluation can be studied directly from the evaluation script[5] that was provided before the competition to let the participants evaluate their morpheme analysis relative to the gold standard samples provided in the Morpho Challenge.

7 Results

The precision, recall and F-measure percentages obtained in the evaluation for all the test languages are shown in Tables 7 - 10. The reference results that are given below each table were:

[5]The evaluation script can be downloaded from `http://www.cis.hut.fi/morphochallenge2007/`

Table 7: The submitted unsupervised morpheme analysis compared to the Gold Standard in **Finnish** (Competition 1).

METHOD	PRECISION	RECALL	F-MEASURE
Bernhard 2	59.65%	40.44%	48.20%
Bernhard 1	75.99%	25.01%	37.63%
Bordag 5a	71.32%	24.40%	36.36%
Bordag 5	71.72%	23.61%	35.52%
Zeman	58.84%	20.92%	30.87%
McNamee 3	45.53%	8.56%	14.41%
McNamee 4	68.09%	5.68%	10.49%
McNamee 5	86.69%	3.35%	6.45%
Morfessor MAP	76.83%	27.54%	40.55%

Table 8: The submitted unsupervised morpheme analysis compared to the Gold Standard in **Turkish** (Competition 1).

METHOD	PRECISION	RECALL	F-MEASURE
Zeman -	65.81%	18.79%	29.23%
Bordag 5a	81.31%	17.58%	28.91%
Bordag 5	81.44%	17.45%	28.75%
Bernhard 2	73.69%	14.80%	24.65%
Bernhard 1	78.22%	10.93%	19.18%
McNamee 3	65.00%	10.83%	18.57%
McNamee 4	85.49%	6.59%	12.24%
McNamee 5	94.80%	3.31%	6.39%
Morfessor MAP	76.36%	24.50%	37.10%
Tepper	70.34%	42.95%	53.34%

- *Morfessor Categories-Map*: The same Morfessor Categories-Map as described in Morpho Challenge 2005 [4] was used for the unsupervised morpheme analysis. Each morpheme was also automatically labeled as prefix, stem or suffix by the algorithm.
- *Tepper*: A hybrid method developed by Michael Tepper [9] was utilized to improve the morpheme analysis reference obtained by our Morfessor Categories-MAP.

For the Finnish task the winner (measured by F-measure) was the algorithm "Bernhard 2". It did not reach a particularly high precision, but the recall and the F-measure were clearly superior. It was also the only algorithm that won the "Morfessor MAP" reference.

For the Turkish task the competition was much tighter. The winner was "Zeman", but "Bordag 5a" and "Bordag 5" were very close. The "Morfessor MAP" and "Tepper" reference methods was clearly better than any of the competitors, but all the algorithms (except "Tepper") seem to have had problems with the Turkish task, because the scores were lower than for other languages. This is interesting, because in the morpheme segmentation task (Competition 1) of the previous Morpho Challenge [7] the corresponding Turkish task was not more difficult than the others.

The "Monson Paramor-Morfessor" algorithm reached the highest score in the German task, but the "Bernhard 2" who again had the highest recall as in Finnish was quite close. The "Bordag 5a" and "Bordag 5" were not very far, either, and managed to beat the "Morfessor MAP" reference.

For English, the "Bernhard 2" and "Bernhard 1" algorithms were the clear winners, but also "Pitler" and "Monson Paramor-Morfessor" and "Monson ParaMor" were able to beat the "Morfessor MAP" and some even the "Tepper" reference.

Table 9: The submitted unsupervised morpheme analysis compared to the Gold Standard in **German** (Competition 1).

METHOD	PRECISION	RECALL	F-MEASURE
Monson Paramor-Morfessor	51.45%	55.55%	53.42%
Bernhard 2	49.08%	57.35%	52.89%
Bordag 5a	60.45%	41.57%	49.27%
Bordag 5	60.71%	40.58%	48.64%
Monson Morfessor	67.16%	36.83%	47.57%
Bernhard 1	63.20%	37.69%	47.22%
Monson ParaMor	59.05%	32.81%	42.19%
Zeman -	52.79%	28.46%	36.98%
McNamee 3	45.78%	9.28%	15.43%
McNamee 4	75.62%	6.67%	12.26%
McNamee 5	90.92%	4.21%	8.04%
Morfessor MAP	67.56%	36.92%	47.75%

8 Discussions

The significance of the differences in F-measure was analyzed for all algorithm pairs in all evaluations. The analysis was performed by splitting the data into several partitions and computing the results for each partition separately. The statistical significance of the differences between the participants' algorithms was computed by the Wilcoxon's Signed-Rank test for comparison of the results in the independent partitions. The results show that almost all differences were statistical significant, only the following pairs were not:

- In Finnish (Table 7): -
- In Turkish (Table 8): "Zeman" and "Bordag 5a", "Bordag 5a" and "Bordag"
- In German (Table 9): "Monson Morfessor" and "Bernhard 1"
- In English (Table 10): "Bernhard 2" and "Bernhard 1", "Monson Paramor-Morfessor" and "Monson ParaMor", "Monson Morfessor" and Zeman", "Bordag 5a" and "Bordag"

This result was not surprising since the random word pair samples were quite large and all these result pairs that were not significantly different gave very similar F-measures (less than 0.5 percentage units away).

By looking at the precision and recall results we see that the "McNamee 5" algorithm, who had clearly the highest precision in all languages, suffered from a very low precision and was not thus competitive in F-measure. However, McNamee's algorithms were not real attempts to provide good morpheme analysis, but mainly to find a representative substring for each word type that would be likely to perform well in the IR evaluation (our Competition 2 [6]). This is in line with our assumption that the precision evaluation could be closer to the IR task, because it measures the portion of matches from a chosen word to other words that agree with the grammatic analysis. This is related to what the most basic form of IR also does: to look for matches between the query word and the words in each document. The recall, however, may not be as relevant to IR, because it measures the portion of grammatically matching morphemes that are found by the algorithm. By looking at the Gold Standards (Table 1) we see that many of the grammatical morphemes (such as +PL and +PAST) are very common and may not be very relevant in IR and an algorithm like the "McNamee 5" would probably ignore them.

The future work in unsupervised morpheme analysis should develop further the clustering of contextually similar units for morphemes that would match better with the grammatical morphemes and thus, improve the recall. Most of the submitted algorithms probably did not take the

Table 10: The submitted unsupervised morpheme analysis compared to the Gold Standard in **English** (Competition 1).

METHOD	PRECISION	RECALL	F-MEASURE
Bernhard 2	61.63%	60.01%	60.81%
Bernhard 1	72.05%	52.47%	60.72%
Pitler	74.73%	40.62%	52.63%
Monson Paramor-Morfessor	41.58%	65.08%	50.74%
Monson ParaMor	48.46%	52.95%	50.61%
Monson Morfessor	77.22%	33.95%	47.16%
Zeman	52.98%	42.07%	46.90%
Bordag 5a	59.69%	32.12%	41.77%
Bordag 5	59.80%	31.50%	41.27%
McNamee 3	43.47%	17.55%	25.01%
McNamee 4	75.96%	13.67%	23.17%
McNamee 5	92.34%	7.38%	13.67%
Morfessor MAP	82.17%	33.08%	47.17%
Tepper	69.23%	52.60%	59.78%

provided possibility to utilize the sentence context for analyzing the words and finding the morphemes. Although this may not be as important for success in IR than improving the precision, it may provide useful additional information for some keywords.

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.

12 different segmentation algorithms from 6 research groups were submitted and evaluated. The evaluations included 4 different languages: Finnish, Turkish, German and English. The algorithms and results were presented in Morpho Challenge Workshop, arranged in connection with other CLEF 2007 Workshop, September 19-21, 2007. Morpho Challenge 2007 was part of the EU Network of Excellence PASCAL Challenge Program and organized in collaboration with CLEF.

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. We are most grateful to the University of Leipzig for making the training data resources available to the Challenge, and in particular we thank Stefan Bordag for his kind assistance. We are indebted to Ebru Arisoy for making the Turkish gold standard available to us. We thank also Krista Lagus for comments of the manuscript. 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] Ozlem Cetinoglu. Prolog based natural language processing infrastructure for Turkish. M.Sc. thesis, Bogazici University, istanbul, Turkey, 2000.

[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. Morfessor in the Morpho Challenge. In *PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*, Venice, Italy, 2006.

[5] Helin Dutagaci. Statistical language models for large vocabulary continuous speech recognition of Turkish. M.Sc. thesis, Bogazici University, istanbul, Turkey, 2002.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.