Unsupervised Morpheme Analysis Evaluation by IR experiments – Morpho Challenge 2008

Mikko Kurimo and Ville Turunen
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 and results of Competition 2 (information retrieval experiments) in the Morpho Challenge 2008. Competition 1 (a comparison to linguistic gold standard) is described in a companion paper. In Morpho Challenge 2008 the goal was to search and evaluate unsupervised machine learning algorithms that provide morpheme analysis for words in different languages. The morpheme analysis can be important in several applications, where a large vocabulary is needed. Especially in morphologically complex languages, such as Finnish, Turkish and Arabic, the agglutination, inflection, and compounding easily produces millions of different word forms which is clearly too much for building an effective vocabulary and training probabilistic models for the relations between words. The benefits of successful morpheme analysis can be seen, for example, in speech recognition, information retrieval, and machine translation. In Morpho Challenge 2008 the morpheme analysis submitted by the Challenge participants were evaluated by performing information retrieval experiments, 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. The results indicate that the morpheme analysis has a significant effect in IR performance in all tested languages (Finnish, English and German). The best unsupervised and language-independent morpheme analysis methods can also rival the best language-dependent word normalization methods. 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; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Algorithms, Performance, Experimentation

Keywords

Morphological analysis, Machine learning

1 Introduction

The goal of the Morpho Challenge 2008 was to search and evaluate unsupervised machine learning algorithms in the task of morpheme analysis for words in different languages. The evaluation consisted of two parts: first a linguistic and then an application oriented performance analysis. The linguistic evaluation described in the companion paper [8], *Competition 1*, compared the suggested morpheme analyses to a linguistic morpheme analysis gold standard. The other evaluation *Competition 2*, described in this paper, carried out information retrieval (IR) experiments from CLEF, where the all the words in the queries and text corpus were replaced by the morpheme analyses of those words.

The Competition 2 IR tasks and corpora were the same as in our previous Morpho Challenge 2007 [6]. Additionally, there was an option to evaluate the IR performance using the morpheme analysis of word forms in their full text context. In our first Morpho Challenge 2005 [7], there were two speech recognition tasks instead of IR, and morpheme segmentations were utilized to train language models.

The morpheme analysis can be important in several applications, where a large vocabulary is needed. Especially in morphologically complex languages, such as Finnish, Turkish and Arabic, the agglutination, inflection, and compounding easily produces millions of different word forms which is clearly too much for building an effective vocabulary and training probabilistic models for the relations between words. The benefits of successful morpheme analysis can be seen, for example, in speech recognition [1, 7], information retrieval [12, 6] and machine translation [9, 11].

The same IR tasks that were attempted using the Morpho Challenge participants' morpheme analyses, were also tested by a number of reference methods to see how useful the unsupervised morpheme analysis could be. These references included the unsupervised baseline algorithms Morfessor Categories-Map [3] and Morfessor Baseline [2, 4], the rule-based grammatical morpheme analysis based on the linguistic gold standards [5], a commercial word normalization tool (TWOL) and traditional stemming approaches for different languages based on the Porter stemming [10]. The same IR statistics were also provided for words as such without any processing.

2 Task and Data in Competition 2

In Competition 2, the Morpho Challenge organizers performed IR experiments based on the morpheme analyses submitted by the participants for the given word lists. Two word lists in each language were provided for analysis, the first for Competition 1 and then another which included the same words plus the word forms that occurred in the IR tasks. For the IR experiments both the words in the documents and in the test queries were then replaced by their proposed morpheme representations and the search was based on morphemes instead of words. Three tasks were provided for three different languages: Finnish, German and English, and the participants were encouraged to use the same algorithms for all of them.

The data sets for testing the IR performance were exactly the same as in the previous Morpho Challenge 2007. In each language there were newspaper articles, test queries and the binary relevance judgments regarding to the queries. Because the organizers performed the IR experiments based on the morpheme analyses submitted by the participants, 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)[1], so that it was possible to use the full text corpora for preparing the morpheme analyses. In Morpho Challenge 2008, an option was also given for an IR performance evaluations using the morpheme analysis of word forms submitted in their full text context.

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

[1] `http://www.clef-campaign.org/`

- 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).

3 Participants and their submissions

Table 1: The submitted algorithms. "Comp 2" shows which were evaluated in Competition 2. "Only 1" means that only analyses of Competition 1 words were used in Competition 2.

Algorithm	Author	Affiliation	Comp 2
"Can (no wordlists)"	Burcu Can	Univ. York, UK	no
"Goodman (late submission)"	Sarah A. Goodman	Univ. Maryland, USA	no
"Kohonen"	Oskar Kohonen et al.	Helsinki Univ. Tech, FI	only 1
"McNamee five"	Paul McNamee	JHU, USA	yes
"McNamee four"	Paul McNamee	JHU, USA	yes
"McNamee lcn5"	Paul McNamee	JHU, USA	yes
"Monson Morfessor"	Christian Monson et al.	CMU, USA	yes
"Monson ParaMor"	Christian Monson et al.	CMU, USA	yes
"Monson ParaMor-Morfessor"	Christian Monson et al.	CMU, USA	yes
"Zeman 1"	Daniel Zeman	Karlova Univ., CZ	only 1
"Zeman 3"	Daniel Zeman	Karlova Univ., CZ	only 1

Four research groups submitted totally nine different algorithms by the deadline at the end of June, 2008 and one group after that. The algorithms and their authors are listed in Table 1. For more detailed analysis of the submissions, see [8].

In the IR task (Competition 2), totally nine algorithms were evaluated in all three languages. For six of those, the morpheme analyses were available for all the words in the IR text corpus. For the remaining three only those words were analyzed that existed in the text corpus for Competition 1 [8] and the others were indexed without analysis. In the Morpho Challenge 2007 [6] experiments were made to compare the IR performance with and without the analysis of these "new" words. The results indicated that in the Finnish task the extra analyses were helpful for almost all participants, but in the German and English task they did not seem to affect the results.

Unlike the others, the algorithms by McNamee were no real attempts to find morphemes, but rather focused directly on extracting substrings from words that would be suitable for IR.

4 Reference methods

In addition to the participating algorithms, a number of different reference methods were evaluated for the same tasks. The purpose of these methods was to provide views on the difficulty and various characteristics of these tasks and on the usefulness of the unsupervised morpheme analysis in the IR tasks.

1. *Morfessor Categories-Map*: The same Morfessor Categories-Map (or here just "catmap", for short) as described in Competition 1 [8] was used for the unsupervised morpheme analysis.

The stem vs. suffix tags were kept, but did not receive any special treatment in the indexing as we wanted to keep the IR evaluation as unsupervised as possible.

2. *Morfessor Baseline*: All the words were simply split into smaller pieces without any morpheme analysis. This means that the obtained subword units were directly used as index terms. This was performed using the Morfessor Baseline algorithm as in Morpho Challenge 2005 [7]. We expected that this would not be optimal for IR, but because the unsupervised morpheme analysis is such a difficult task, this simple method would probably do quite well.

3. *dummy*: No words were split nor any morpheme analysis provided except hyphens were replaced by spaces so that hyphenated words were indexed as separate words (changed from last year). This means that words were directly used as index terms as such without any stemming or tags. We expected that although the morpheme analysis should provide helpful information for IR, all the submissions would not probably be able to beat this brute force baseline. However, if some morpheme analysis method would consistently beat this baseline in all languages and task, it would mean that the method would probably be useful in a language and task independent way.

4. *grammatical*: The words were analyzed using the same gold standard analyses 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. For many words the gold standard analyses included several alternative interpretations that were all included in the indexing. However, we decided to also try the method adopted in the morpheme segmentation for Morpho Challenge 2005 [7] that only the first interpretation of each word is applied. This was here called "grammatical first" whereas the default was called "grammatical all". Words that were not in the gold standard segmentation were indexed as such. Because our gold standards are quite small, 60k (English) - 600k (Finnish), compared to the amount of words that the unsupervised methods can analyze, we did not expect "grammatical" to perform particularly well, even though it would probably capture some useful indexing features to beat the "dummy", at least.

5. *snowball*: No real morpheme analysis was performed, but the words were stemmed by stemming algorithms provided by Snowball libstemmer library. Porter stemming algorithm was used for English. Finnish and German stemmers were used for the other languages. Hyphenated words were first split to parts that were then stemmed separately. Stemming is expected to perform very well for English but not necessarily for the other languages because it is harder to find good stems.

6. *TWOL*: Two-level morphological analyzer TWOL from Lingsoft[2] Inc. was used to find the normalized forms of the words. These forms were then used as index terms. Some words may have several alternative normalized forms and two cases were studied similarly to the grammatical case. Either all alternatives were used (”all”) or only the first one (”first”). Compound words were split to parts. Words not recognized by the analyzer were indexed as such. German analyzer was not available for the organizers.

7. *best 2007*: This is the algorithm in each task that provided the highest average precision in Morpho Challenge 2007. The IR tasks in 2007 were identical to 2008, but because some numbers in the joint word frequency statistics provided for the participants differed slightly, the 2007 results may not be exactly comparable.

[2] http://www.lingsoft.fi/

5 Evaluation

The submitted morpheme analyses were evaluated by IR experiments in three different tasks: one in Finnish, one in German and one in English. It would have been interesting to evaluate also the performance in Turkish and Arabic, but unfortunately no IR tasks in these languages were available to the organizers. In the IR corpora the words were replaced by the provided morpheme analyses both in the text and the queries, and then the search was performed based on morphemes instead of full words. Any word without morpheme analysis was left un-replaced and indexed as it were just a single morpheme on its own.

Those participants who only provided morpheme analyses for words that exist in the text corpus for Competition 1 [8] had a slight disadvantage, because then the "new" words in the IR task were indexed and searched without splitting. However, the experiments in the Morpho Challenge 2007 [6] revealed that the extra analyses were helpful only in the Finnish task. In the German and English task they did not seem to affect the results.

In Morpho Challenge 2008 we provided the participants an option to use the full text corpora in order to get information and train models using the context in which the different words occur and, for the first time, also to submit morpheme analysis for words in their actual context. However, none of the participants dared to go for this even more challenging option.

In practice, the IR evaluation was performed using the latest version of the freely available LEMUR toolkit[3]. Okapi (BM25) term weighting was used for all index terms excluding an automatic stoplist. The automatic stoplist was separately determined for each morpheme analysis run by extracting the morphemes that have a collection frequency higher than 75000 (Finnish) or 150000 (German and English). The stoplist was used with the Okapi weighting, because in the previous Morpho Challenge [6] it was observed that the performance of indexes that have many very common terms was poor. The evaluation criterion was Uninterpolated Average Precision.

6 Results

Table 2 presents the IR evaluation results. The algorithms had been improved from the previous competition, and in all tasks there was a new winner. The highest average precision in the Finnish task was, slightly surprisingly, achieved by the character 4-gram approach "McNamee four" that was equal in performance to last year's winner, but clearly beat the other 2008 competitors.

In the English and German tasks the winner was "Monson Paramor+Morfessor" that also won the Competition 1 in all languages. The marginal to the best 2007 results was very tight, but clear to the other 2008 competitors. In both English and German tasks the "McNamee four" was second after Monson's algorithms.

The "Monson Paramor+Morfessor" which was built by combining the publicly available Morfessor algorithm and the "Monson Paramor" managed to improve both of them, except in the Finnish task, where it was very close to "Monson Morfessor". It is interesting to note that while being far behind Morfessor in both Finnish and German, the "Monson Paramor" does a very good job in English being close to the combined version "Monson Paramor+Morfessor".

The new rule-based reference method "TWOL" that was evaluated this year in the Finnish and English task, was unbeatable in Finnish and only narrowly beaten in English by the best unsupervised algorithm and the traditional "Snowball Porter" stemmer. In Finnish and German the "Snowball" stemmers did not perform very well and had clearly lower average precision than the best unsupervised algorithms and "TWOL". The performance of the "grammatical" references based on the linguistic gold standards were not very high, which is not surprising given that the gold standards are relatively small.

The algorithms by Kohonen and Zeman that did not have morpheme analyses for all the words in the IR corpora were left behind Monson and McNamee. This may partly be due to those words that were not split in the morphemes, but as the importance of the analysis of those relatively rare words has not generally been very large in the previous tests, the performance gap may also be due to the morpheme analyses the algorithms provide.

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

Table 2: The obtained average precision (AP%) in the three different IR tasks. The Competition 2 participants are shown in bold and the various reference methods in normal font. (a) the IR tasks are the same as in Morpho Challenge 2007, but because some values in the word frequency statistics provided for the participants differed slightly, the 2007 results may not be exactly comparable. (b) some participants provided morpheme analyses only for words that existed also in the text corpus for Competition 1 [8].

Finnish IR task	AP%	English IR task	AP%
TWOL first	0.4976	snowball porter	0.4081
McNamee four	0.4918	**Monson Paramor+Morfessor**	0.3989
best 2007 Bernhard 2	0.4915a	TWOL first	0.3957
TWOL all	0.4845	best 2007 Bernhard 2	0.3943a
Monson Morfessor	0.4679	**Monson Paramor**	0.3928
Monson Paramor+Morfessor	0.4673	TWOL all	0.3922
McNamee five	0.4515	Morfessor baseline	0.3861
Morfessor catmap	0.4441	grammatical first	0.3734
Morfessor baseline	0.4425	Morfessor catmap	0.3713
grammatical first	0.4312	**Monson Morfessor**	0.3637
snowball finnish	0.4275	**McNamee five**	0.3630
grammatical all	0.4090	**McNamee four**	0.3566
Monson Paramor	0.3965	**McNamee lcn5**	0.3563
McNamee lcn5	0.3688	grammatical all	0.3542
Kohonen	0.3548b	**Kohonen**	0.3342b
dummy	0.3519	dummy	0.3293
Zeman 3	0.3282b	**Zeman 3**	0.3125b
Zeman 1	0.2627b	**Zeman 1**	0.2631b

German IR task	AP%
Monson Paramor+Morfessor	0.4734
best 2007 Bernhard 1	0.4729a
Monson Morfessor	0.4671
Morfessor baseline	0.4656
Morfessor catmap	0.4642
McNamee four	0.4388
McNamee five	0.4331
snowball german	0.3865
Kohonen	0.3671b
Monson Paramor	0.3631
dummy	0.3509
grammatical first	0.3353
McNamee lcn5	0.3276
Zeman 3	0.3206b
grammatical all	0.3014
Zeman 1	0.2343b

7 Discussions and Conclusions

The Morpho Challenge 2008 was a successful follow-up to our previous Morpho Challenge 2005 and 2007. Since the main tasks were unchanged the participants of the previous challenges were able to compare improvements of their algorithms and the new participants and those who missed the previous deadlines were able to try more established benchmark tasks. The new task which allowed full text context to be used in the unsupervised morpheme analysis was not yet attempted by anyone. However, as it seems like a natural way to improve the models, it may be included in the next Morpho Challenge as well, giving participants more time to develop the new kinds of models and learning algorithms needed.

As future work there remains the need to develop better methods to combine the different existing algorithms and to cluster the different surface forms produced by the morphemes. This might also somewhat improve the relatively low recall that several algorithms suffered in the Competition 1 [8]. New IR tasks should also be included and languages like Arabic which pose new kinds of morphological problems. To better serve the goal of producing a general purpose morpheme-based vocabulary that would be useful for several applications where large vocabulary is needed, we should also target new evaluation applications, e.g. in machine translation, text understanding and speech recognition.

Acknowledgments

We thank all the participants for their submissions and enthusiasm. We owe great thanks as well to the PASCAL Challenge Program and CLEF who helped us organize this challenge and the challenge workshop. 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] 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] Mikko Kurimo and Matti Varjokallio. Unsupervised morpheme analysis evaluation by a comparison to a linguistic Gold Standard – Morpho Challenge 2008. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[9] 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.

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

[11] Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark, 2007.

[12] 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.