Unsupervised Morpheme Analysis Evaluation by a Comparison to a Linguistic Gold Standard – Morpho Challenge 2008

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

Abstract

The goal of Morpho Challenge 2008 was to find and evaluate unsupervised algorithms that provide morpheme analyses for words in different languages. Especially in morphologically complex languages, such as Finnish, Turkish and Arabic, morpheme analysis is important for lexical modeling of words in speech recognition, information retrieval and machine translation. The evaluation in Morpho Challenge competitions consisted of both a linguistic and an application oriented performance analysis. This paper describes an evaluation where the competition entries were compared to a linguistic morpheme analysis gold standard. Because the morpheme labels in an unsupervised analysis can be arbitrary, the evaluation is based on matching the morpheme-sharing words between the proposed and the gold standard analyses. In addition to Finnish, Turkish, German and English evaluations performed in Morpho Challenge 2007, the competition this year had an additional evaluation in Arabic. The results in 2008 show that although the level of precision and recall varies substantially between the tasks in different languages, the best methods seem to manage all the tested languages quite well. 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 topic of the Morpho Challenge 2008 competition is to evaluate proposed unsupervised machine learning algorithms in the task of morpheme analysis for words in different languages. The Morpho Challenge evaluation consisted of both a linguistic and an application oriented performance analysis. The linguistic evaluation described in this paper, *Competition 1*, is based on a comparison of the suggested morpheme analysis to a linguistic morpheme analysis gold standard. The practical application oriented evaluation described in the companion paper [10], *Competition 2*, contained information retrieval (IR) experiments from CLEF, where the all the words in the queries and text corpus were replaced by their morpheme analyses.

The Morpho Challenge 2008 tasks and training corpora were the same as in our previous Morpho Challenge 2007 [8], except that it involved one additional morphologically complex language, Arabic. There was also an optional evaluation of the IR performance using the morpheme analysis of word forms in their full text context. The difference to our first Morpho Challenge 2005 [9] which focused on just the segmentation of words into morphologically meaningful units, was that the units should further be clustered into the abstract classes of morphemes. For example, this analysis should find the link between the word forms "foot" and "feet".

Especially in morphologically complex languages, such as Finnish, Turkish and Arabic, the morpheme analysis is important for lexical modeling of words in speech recognition [1, 9], information retrieval [13, 7] and machine translation [11, 12]. Due to the high level of agglutination, inflection, and compounding, there are millions of different word forms, which is clearly too much for building an effective vocabulary and training probabilistic models for the relations between words. There also exist carefully constructed linguistic tools for morphological analysis, but only for few languages. Even in these cases using statistical machine learning methods we may still discover interesting alternatives that may rival even the most sophisticated linguistically designed morphologies.

The scientific objectives of the Morpho Challenge competitions are: to learn about the word construction in natural languages, to advance machine learning methodology, and to discover approaches that are suitable for many languages. The portability to different languages is very important, because the language technology often needs to be quickly extended to various new languages for which there are limited amount of resources available. Unsupervised learning is then the most attractive approach for data analysis, because the majority of the available data is unannotated and human annotation work is expensive.

2 Task and Data in Competition 1

The task in the Morpho Challenge 2008 was to return the given list of words in each language extended by the morpheme analysis of each word form. The morpheme analyses should be obtained by an unsupervised learning algorithm that would preferably be as language independent as possible. In each language, the participants were pointed to a training corpus in which all the words occur (in a sentence), so that the algorithms may also utilize information about the word context. The tasks were the same as in the Morpho Challenge 2007 last year with the addition of one new language, Arabic.

The training corpora were the same as in the Morpho Challenge 2007, except for Arabic: 3 million sentences for English, Finnish and German, and 1 million sentences for Turkish in plain unannotated text files that were all downloadable from the Wortschatz collection[1] at the University of Leipzig (Germany). The corpora were specially preprocessed for the Morpho Challenge (tokenized, lower-cased, some conversion of character encodings).

The Arabic text data (135K sentences with 3.9M words) is the same as used by Habash and Sadat [6]. Because this text data is unfortunately not freely available, only a list of word forms was provided, so if the participants wanted to use typical word contexts in training their models in Arabic, they had to find their own text corpus. All words in the Arabic data were presented in Buckwalter transliteration[2]. In other languages the lists of word forms to be analyzed were extracted from the Wortschatz corpora and included all the different word forms existing there and their frequencies in the corpora. The total amount of word types were 2,206,719 (Finnish), 617,298 (Turkish), 1,266,159 (German), 384,903 (English), and 143,966 (Arabic).

[1] `http://corpora.informatik.uni-leipzig.de/`
[2] `http://www.qamus.org/transliteration.htm`

The exact syntax of the word lists and the required output lists with the suggested morpheme analyses were explained previously in [8]. As the learning is unsupervised, the returned morpheme labels may be arbitrary: e.g., "foot", "morpheme42" or "+PL". The order in which the morpheme labels appear after the word forms does not matter. Several interpretations for the same word can also be supplied, and it was left to the participants to decide whether they would be useful in the task, or not.

3 Reference analysis

3.1 Linguistic Gold Standard

In Competition 1 the proposed unsupervised morpheme analyses were compared to the correct grammatical morpheme analyses called here the linguistic gold standard. The gold standard morpheme analyses were prepared in exactly the same format as the result file the participants were asked to submit, alternative analyses separated by commas. 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
Arabic	Algbn	gabon_POS:N Al+ +SG
	AlmtHdp	mut aHidap_POS:PN Al+ +SG,
		mut aHid_POS:AJ Al+ +SG

The gold standard reference analyses were the same as in the Morpho Challenge 2007 [8], except in Arabic. The Arabic gold standard analyses are based on the representation of lexeme and features used in the Aragen system (a wrapper using publicly available BAMA-1 databases) [5]. The first part of an analysis is a lexeme followed by a list of features. The original features were here modified to connect the POS label to the root of the word, e.g. "Algbn = gabon_POS:N Al+ +SG". In addition, the gender morphemes were removed (e.g. the German gold standard doesn't contain these either). This did not affect the ranking of the submissions, but made the evaluation resemble more the other tested languages.

In the word lists described in the previous section, the gold standard analyses were available for 650,169 (Finnish), 214,818 (Turkish), 125,641 (German), 63,225 (English), and 141,876 (Arabic) word types.

3.2 Morfessor

As baseline results for unsupervised morpheme analysis, the organizers provided morpheme analysis by a publicly available unsupervised algorithm called "Morfessor Categories-MAP" developed at Helsinki University of Technology [3] (or here "Morfessor catmap" or "Morfessor MAP", for short as in [8]). Analysis by the original Morfessor [2, 4] (or here "Morfessor baseline"), which provides only a surface-level segmentation, was also provided for reference.

4 Participants and their submissions

Table 2: The submitted algorithms. “Comp 1” shows which were evaluated in Competition 1.

Algorithm	Author	Affiliation	Comp 1
“Can (no wordlists)”	Burcu Can	Univ. York, UK	no
“Goodman (late submission)”	Sarah A. Goodman	Univ. Maryland, USA	yes
“Kohonen Allomorfessor”	Oskar Kohonen et al.	Helsinki Univ. Tech, FI	yes
“McNamee five”	Paul McNamee	JHU, USA	no
“McNamee four”	Paul McNamee	JHU, USA	no
“McNamee lcn5”	Paul McNamee	JHU, USA	no
“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	yes
“Zeman 3”	Daniel Zeman	Karlova Univ., CZ	yes

By the submission DL at the end of June, 2008, four research groups had submitted nine different algorithms which were then evaluated by the organizers. After the DL, more submissions were received from another author (Goodman), which were evaluated separately outside the Competition 1. One group (Can) decided not to submit the final wordlists that could be evaluated and one (McNamee) wanted only to participate in Competition 2. Thus, the final amount of evaluated algorithms was nine: six in Competition 1, one outside the competition, and two reference results by Morfessor. The algorithm submissions and their authors are listed in Table 2.

Some characteristics of morpheme analyses proposed by the unsupervised algorithms together with the gold standard analyses are briefly presented in Tables 3 and 4. The statistics of each submission include the average amount of alternative analyses per word, the average amount of morphemes per analysis, and the total amount of morpheme types. The “Allomorfessor” is an extension to the “Morfessor Baseline” that attempts to discover common baseforms for the different surface forms that are likely to represent the same morpheme. The “ParaMor” is another algorithm for segmenting words into morphemes which, after improvements from the previous Morpho Challenge, was submitted also as a combination with the publicly available “Morfessor CATMAP”. The “Zeman 1” is a resubmission from the previous Morpho Challenge which, after attempts to include a new treatment of prefix, was submitted as the “Zeman 3”. It is interesting to note that this year all the algorithms resulted in a very large lexicon, usually much larger than the reference methods did.

5 Evaluation

The evaluation of Competition 1 in Morpho Challenge 2008 was similar as in Morpho Challenge 2007 except that there was one new language, Arabic. The full description of the method to compare the submitted unsupervised morpheme analyses were to the linguistic gold standard analyses is in [8]. In the current paper we just remind the main points and obtained performance measures.

Because the morpheme analysis candidates are achieved by unsupervised learning, the morpheme labels can be arbitrary and different from the ones designed by linguists. The basis of the evaluation is, thus, to compare whether any two word forms that contain the same morpheme according to the participants' algorithm also has a morpheme in common according to the gold standard and vice versa. In practice, the evaluation is performed by randomly sampling a large number of morpheme sharing word pairs from the compared analyses. Then the *precision* is calculated as the proportion of morpheme sharing word pairs in the participant's sample that really has

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

Finnish	Example word: linuxiin	#a	#m	lexicon
Kohonen	linux iin	1	1.86	486096
Monson paramor	linux +iin	1	2.62	1123572
Monson morfessor	linux/STM +iin/SUF	1	2.83	223412
Monson p+m	linux/STM +iin/SUF, linux +iin	2	2.72	1359325
Zeman 1	linuxiin, linuxii n, linuxi in, linux iin	3.61	1.81	5379817
Zeman 3	linuxiin	1.21	1.62	1830751
Morfessor baseline	linux iin	1	2.21	149417
Morfessor catmap	linux +iin	1	2.94	217001
Gold Standard	linux_N +ILL	1.16	3.29	33754

Turkish	Example word: popUlerliGini	#a	#m	lexicon
Kohonen	popUler liGini	1	1.76	183297
Monson paramor	popUlerl +i +G +in +i	1	2.89	245737
Monson morfessor	pop/STM +U/SUF +ler/SUF +liGini/SUF	1	2.76	107431
Monson p+m	pop/STM +U/SUF +ler/SUF +liGini/SUF, popUlerl +i +G +in +i	2	2.83	354280
Zeman 1	popUlerliGin i, popUlerliGi ni	3.24	1.76	1205970
Zeman 3	popU lerliGi ni, popU lerliGin i, popU lerliGini, popUlerliGi ni, popUlerliGin i	1.14	1.52	501154
Morfessor baseline	popUler liGini	1	2.14	53473
Morfessor catmap	pop +U +ler +liGini	1	2.64	114834
Gold Standard	popUler +DER_lHg +POS2S +ACC, popUler +DER_lHg +POS3 +ACC3	1.99	3.36	21163

Arabic	Example word: AlmtHdp	#a	#m	lexicon
Monson paramor	AlmtHd +p	1	1.72	81978
Monson morfessor	+Al/PRE mtHd/STM +p/SUF	1	2.03	46526
Monson p+m	+Al/PRE mtHd/STM +p/SUF, AlmtHd +p	2	1.87	133309
Zeman 1	AlmtHdp, AlmtHd p, AlmtH dp	2.24	1.65	217232
Zeman 3	AlmtHdp	1.23	1.61	106378
Morfessor baseline	Al mtHdp	1	2.45	16735
Morfessor catmap	Al/PRE mtHd/STM p/SUF	1	2.04	46789
Gold Standard	mut aHidap_POS:PN Al+ +SG, mut aHid_POS:AJ Al+ +SG	1.78	3.39	43914

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

German	Example word: zurueckzubehalten	#a	#m	lexicon
Kohonen	zurueckzu behalten	1	1.83	334851
Monson paramor	zurueckzube +halten	1.25	1.65	908556
Monson morfessor	+zurueck/PRE +zu/PRE +be/PRE halten/STM	1	3.10	166963
Monson p+m	+zurueck/PRE +zu/PRE +be/PRE halten/STM, zurueckzube +halten	2.25	2.30	1094322
Zeman 1	zurueckzubehalten, zurueckzubehalte n, zurueckzubehalt en, zurueckzubehal ten, zurueckzubeha lten, zurueckzubeh alten, zurueckzube halten	4.11	1.80	4054397
Zeman 3	zurueckzubehalten	1.12	1.43	1053275
Morfessor baseline	zurueckzu behalten	1	2.30	90009
Morfessor catmap	zurueck zu be halten	1	3.06	172907
Gold Standard	zurueck_B zu be halt_V +INF	1.30	2.97	14298

English	Example word: baby-sitters	#a	#m	lexicon
Kohonen	baby- sitters	1	1.62	180813
Monson paramor	bab +y, sitt +er +s	1.27	1.75	252997
Monson morfessor	+baby-/PRE sitter/STM +s/SUF	1	2.07	137973
Monson p+m	+baby-/PRE sitter/STM +s/SUF, bab +y, sitt +er +s	2.27	1.89	378364
Zeman 1	baby-sitter s, baby-sitt ers	3.18	1.74	905251
Zeman 3	baby-sitt ers, baby-sitter s	1.08	1.37	319982
Morfessor baseline	baby- sitters	1	2.32	40293
Morfessor catmap	baby - sitters	1	2.12	132086
Gold Standard	baby_N sit_V er_s +PL	1.10	2.13	16902

a morpheme in common according to the gold standard. Correspondingly, the *recall* is calculated as the proportion of morpheme sharing word pairs in the gold standard sample that also exist in the participant's submission. The sample size in different languages varied depending on the size of the word lists and gold standard: 200,000 (Finnish), 50,000 (Turkish), 50,000 (German), 10,000 (English), and 20,000 (Arabic) word pairs.

The *F-measure*, which is the harmonic mean of *Precision* and *Recall*, was selected as the final evaluation measure:

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

6 Results

Table 5: The submitted unsupervised morpheme analyses compared to the gold standard in **Finnish**, **Turkish** and **Arabic** (Competition 1). The Competition 2 participants are shown in bold and the various reference methods in normal font.

Finnish	PRECISION	RECALL	F-MEASURE
Monson p+m	49.76%	47.25%	48.47%
reference Morfessor catmap	76.83%	27.54%	40.55%
Monson paramor	46.40%	34.44%	39.53%
best 2007 Bernhard 1	75.99%	25.01%	37.63%
Monson morfessor	77.40%	21.52%	33.68%
Zeman 1	58.51%	20.47%	30.33%
reference Morfessor baseline	88.12%	12.01%	21.16%
Goodman methodB.deduped	62.19%	7.71%	13.71%
Kohonen allomorfessor	92.55%	6.89%	12.82%
Zeman 3	72.41%	3.42%	6.54%

Turkish	PRECISION	RECALL	F-MEASURE
Monson p+m	51.88%	52.10%	51.99%
Monson paramor	56.67%	39.42%	46.50%
Monson morfessor	73.92%	26.06%	38.53%
reference Morfessor catmap	76.36%	24.50%	37.10%
Zeman 1	65.81%	18.79%	29.23%
best 2007 Zeman	65.81%	18.79%	29.23%
reference Morfessor baseline	89.20%	11.32%	20.08%
Goodman pruned	69.96%	8.42%	15.04%
Kohonen allomorfessor	93.25%	6.15%	11.53%
Zeman 3	73.30%	3.01%	5.79%

Arabic	PRECISION	RECALL	F-MEASURE
Monson p+m	79.77%	27.47%	40.87%
reference Morfessor baseline	78.16%	23.74%	36.41%
reference Morfessor catmap	90.17%	20.97%	34.03%
Monson morfessor	90.35%	20.95%	34.01%
Zeman 1	77.24%	12.73%	21.86%
Monson paramor	78.58%	8.52%	15.37%
Zeman 3	89.62%	5.18%	9.79%

The results of the linguistic evaluation are presented in Tables 5 and 6. The tasks in Competition 1 were the same as in Morpho Challenge 2007, so it is possible to directly compare the improvements made over the previous algorithms. However, direct comparisons between the evaluation measures in different languages are not valid, because the corpora and gold standards are

Table 6: The submitted unsupervised morpheme analyses compared to the gold standard in **German** and **English** (Competition 1). The Competition 2 participants are shown in bold and the various reference methods in normal font.

German	PRECISION	RECALL	F-MEASURE
Monson p+m	49.53%	59.51%	54.06%
best 2007 Monson p+m	51.45%	55.55%	53.42%
reference Morfessor catmap	67.56%	36.92%	47.75%
Monson morfessor	67.16%	36.83%	47.57%
Monson paramor	53.42%	38.15%	44.51%
Zeman 1	53.12%	28.37%	36.98%
reference Morfessor baseline	80.23%	19.22%	31.01%
Goodman methodB.deduped	54.53%	12.70%	20.60%
Kohonen allomorfessor	87.92%	7.44%	13.71%
Zeman 3	72.27%	7.15%	13.01%

English	PRECISION	RECALL	F-MEASURE
best 2007 Bernhard 2	61.63%	60.01%	60.81%
Monson p+m	50.64%	63.30%	56.26%
reference Morfessor baseline	71.93%	43.27%	54.04%
Monson paramor	58.50%	48.10%	52.79%
reference Morfessor catmap	82.17%	33.08%	47.17%
Monson morfessor	77.22%	33.95%	47.16%
Zeman 1	52.98%	42.07%	46.90%
Goodman methodB.deduped	66.19%	16.51%	26.43%
Kohonen allomorfessor	83.39%	13.43%	23.13%
Zeman 3	76.92%	8.47%	15.27%

different. In all tasks except the English one, improvements were made in 2008 and the best obtained F-measure was now higher. As clearly seen in Tables 5 and 6, this is mainly due to the improved version of "Monson paramor+morfessor" that dominated all tasks. The difference is especially clear in the recall statistics where the performance of the "Monson paramor+morfessor" is superior. Behind Monson's algorithms, the "Zeman 1" that is a re-submission from last year, was better than the rest of the algorithms, which all suffered from a very low recall. It is worth noting that the "Kohonen allomorfessor" algorithm achieved clearly the highest precision of all algorithms in all tasks, but due to the low recall, or undersegmentation, it got rather low F-measure values.

From the Competition 1 in Morpho Challenge 2007 [8], only the winner "best 2007" in each task was chosen in Tables 5 and 6 for reference. The "Monson paramor+morfessor" was able to clearly beat the publicly available reference methods "Morfessor baseline" and "Morfessor catmap" in all tasks. It is interesting to note that the "Morfessor baseline", which is the original simpler Morfessor version and only attempts to split words into morphemes without any further analysis, actually beats the more sophisticated "Morfessor catmap", as well as "Monson morfessor" and "Zeman 1", in English and Arabic. Otherwise, the ranking between the different 2008 algorithms remains the same in all tasks.

7 Discussions and Conclusions

The Morpho Challenge 2008 was a successful follow-up to our previous Morpho Challenges 2005 and 2007. Since the main tasks were unchanged, the participants of the previous challenges were able to track improvements of their algorithms. It also gave a possibility for the new participants and those who missed the previous deadlines to try more established benchmark tasks. This year the evaluation was performed also in Arabic, and despite the relatively small wordlist and the disability to distribute a relevant text corpus, this task was again successful in finding significant differences between the submitted algorithms.

The significance of the differences in F-measure was analyzed for all algorithm pairs in all tasks using the t-test. The analysis was performed by splitting the data into several partitions and comparing the results in each independent partition separately. The results of the tests show that all differences were statistically significant, except “Zeman 1” vs “Morfessor catmap” in the English task.

As already noted in the previous section, the ranking of the algorithms would have been very different, if only the precision measure was utilized. Some of the methods, especially “Kohonen allomorfessor” undersegmented the word forms heavily, which produced high precision but low recall. However, because it is difficult to estimate the relative weight of precision against recall in different applications, it remains for the application based evaluations in different tasks to show which algorithms are most useful. Many of the grammatical morphemes (such as +PL and +PAST in Table 1) are very common and may not be very relevant in IR, for example, compared to recognizing the right stem.

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. 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 are most grateful to the Nizar Habash from the University of Columbia for his kind assistance and making the Arabic word frequency list and reference analyses available to the Challenge. 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] Nizar Habash. Large scale lexeme based arabic morphological generation. In *Proceedings of Traitement Automatique du Langage Naturel (TALN-04)*, Fez, Morocco, 2004.

[6] Nizar Habash and Fatiha Sadat. Arabic preprocessing schemes 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)*, New York, USA, 2006.

[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] Mikko Kurimo and Ville Turunen. Unsupervised morpheme analysis evaluation by IR experiments – Morpho Challenge 2008. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[11] 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.

[12] 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.

[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.