Morpho Challenge competition 2005-2010: Evaluations and results

Mikko Kurimo, Sami Virpioja, Ville Turunen, Krista Lagus
Adaptive Informatics Research Centre
Aalto University, Espoo, Finland
Firstname.Lastname@tkk.fi

Abstract

Morpho Challenge is an annual evaluation campaign for unsupervised morpheme analysis. In morpheme analysis, words are segmented into smaller meaningful units. This is an essential part in processing complex word forms in many large-scale natural language processing applications, such as speech recognition, information retrieval, and machine translation. The discovery of morphemes is particularly important for morphologically rich languages where inflection, derivation and composition can produce a huge amount of different word forms. Morpho Challenge aims at language-independent unsupervised learning algorithms that can discover useful morpheme-like units from raw text material. In this paper we define the challenge, review proposed algorithms, evaluations and results so far, and point out the questions that are still open.

1 Introduction

Many large-scale natural language processing (NLP) applications, such as speech recognition, information retrieval and machine translation, require that complex word forms are analyzed into smaller, meaningful units. The discovery of these units called morphemes is particularly important for morphologically rich languages where the inflection, derivation and composition makes it impossible to even list all the word forms that are used. Various tools have been developed for morpheme analysis of word forms, but they are mostly based on language-specific rules that are not easily ported to other languages. Recently, the performance of tools based on language-independent unsupervised learning from raw text material has improved significantly and rivaled the language-specific tools in many applications.

The unsupervised algorithms proposed so far in Morpho Challenge typically first generate various alternative morphemes for each word and then select the best ones based on relevant criteria. The statistical letter successor variation (LSV) analysis (Harris, 1955) and its variations are quite commonly used as generation methods. LSV is based on the observation that the segment borders between the sub-word units often co-occur with the peaks of variation for the next letter. One popular selection approach is to minimize a cost function that balances between the size of the corpus when coded by the morphemes and the size of the morpheme codebook needed. Selection criteria that produce results resembling the linguistic morpheme segmentation include, for example, the Minimum Description Length (MDL) principle and maximum a posteriori (MAP) probability optimization (de Marcken, 1996; Creutz and Lagus, 2005).

The Morpho Challenge competition was launched in 2005 to encourage the machine learning people, linguists and specialists in NLP applications to study this field and come together to compare their best algorithms against each other. The organizers selected evaluation tasks, data and metric and performed all the evaluations. Thus, participation was made easy for people who were not specialists in the chosen NLP applications. Participation was open to everybody with no charge. The competition became popular right from the beginning and has gained new participants every year.

Although not all the authors of relevant morpheme analysis algorithms have yet submitted their algorithms for this evaluation campaign, more than 50 algorithms have already been evaluated. After the first five years of Morpho Challenge, a lot has been learned on the various possible ways to solve the problem and how the different methods work in various NLP tasks. How-

Proceedings of the 11th Meeting of the ACL-SIGMORPHON, ACL 2010, pages 87–95,
Uppsala, Sweden, 15 July 2010.

ever, there are still open questions such as: how to find meaning for the obtained unsupervised morphemes, how to disambiguate among the alternative analyses of one word, and how to use context in the analysis. Another recently emerged question that is the special topic in 2010 competition is how to utilize small amounts of labeled data and semi-supervised learning to further improve the analysis.

2 Definition of the challenge

2.1 Morphemes and their evaluation

Generally, the morphemes are defined as the smallest meaningful units of language. Rather than trying to directly specify which units are meaningful, the Morpho Challenge aims at finding units that would be useful for various practical NLP applications. The goal is to find automatic methods that can discover suitable units using unsupervised learning directly on raw text data. The methods should also not be restricted to certain languages or include many language and application dependent parameters that needed to be hand tuned for each task separately. The following three goals have been defined as the main scientific objectives for the challenge: (1) To learn of the phenomena underlying word construction in natural languages. (2) To discover approaches suitable for a wide range of languages. (3) To advance machine learning methodology.

The evaluation tasks, metrics and languages have been designed based on the scientific objectives of the challenge. It can not be directly verified how well an obtained analysis reflects the word construction in natural languages, but intuitively, the methods that split everything into letters or pre-specified letter n-grams, or leave the word forms unanalyzed, would not be very interesting solutions. An interesting thing that can be evaluated, however, is how close the obtained analysis is to the linguistic gold standard morphemes that can be obtained from CELEX or various language-dependent rule-based analyzers. The exact definition of the morphemes, tags, or features available in the gold standard to be utilized in the comparison should be decided and fixed for each language separately.

To verify that a proposed algorithm works in various languages would, ideally, require running the evaluations on a large number of languages that would be somehow representative of various important language families. However, the resources available for both computing and evaluating the analysis in various applications and languages are limited. The suggested and applicable compromise is to select morphologically rich languages where the morpheme analysis is most useful and those languages where interesting state-of-the-art evaluation tasks are available. By including German, Turkish, Finnish and Arabic, many interesting aspects of concatenative morphology have already been covered.

While the comparison against the linguistic gold standard morphemes is an interesting subgoal, the main interest in running the Morpho Challenge is to find out how useful the proposed morpheme analyses are for various practical NLP applications. Naturally, this is best evaluated by performing evaluations in several state-of-the-art application tasks. Due to the limitations of the resources, the applications have been selected based on the importance of the morpheme analysis for the application, on the availability of open state-of-the-art evaluation tasks, and on the effort needed to run the actual evaluations.

2.2 Unsupervised and semi-supervised learning

Unsupervised learning is the task of learning without labeled data. In the context of morphology discovery, it means learning without knowing where morpheme borders are, or which morphemes exist in which words. Unsupervised learning methods have many attractive features for morphological modeling, such as language-independence, independence of any particular linguistic theory, and easy portability to a new language.

Semi-supervised learning can be approached from two research directions, namely unsupervised and supervised learning. In an essentially unsupervised learning task there may exist some labeled (classified) data, or some known links between data items, which might be utilized by the (typically generative) learning algorithms. Turned around, an essentially supervised learning task, such as classification or prediction, may benefit also from unlabeled data which is typically more abundantly available.

In morphology modeling one might consider the former setup to be the case: the learning task is essentially that of unsupervised modeling, and morpheme labels can be thought of as known links

between various inflected word forms.

Until 2010 the Morpho Challenge has been defined only as an unsupervised learning task. However, since small samples of morphologically labeled data can be provided already for quite many languages, also the semi-supervised learning task has become of interest.

Moreover, while there exists a fair amount of research and now even books on semi-supervised learning (Zhu, 2005; Abney, 2007; Zhu, 2010), it has not been as widely studied for structured classification problems like sequence segmentation and labeling (cf. e.g. (Jiao et al., 2006)). The semi-supervised learning challenge introduced for Morpho Challenge 2010 can thus be viewed as an opportunity to strengthen research in both morphology modeling as well as in semi-supervised learning for sequence segmentation and labeling in general.

3 Review of Morpho Challenge competitions so far

3.1 Evaluation tasks, metrics, and languages

The evaluation tasks and languages selected for Morpho Challenge evaluations are shown in Figure 1. The languages where evaluations have been prepared are Finnish (FIN), Turkish (TUR), English (ENG), German (GER), and Arabic (ARA). First the morphemes are compared to linguistic gold standards in direct morpheme segmentation (2005) and full morpheme analysis (since 2007). The practical NLP application based evaluations are automatic speech recognition (ASR), information retrieval (IR) and statistical machine translation (SMT). Morphemes obtained by semi-supervised learning can be evaluated in parallel with the unsupervised morphemes. For IR, evaluation has also been extended for full sentences, where the morpheme analysis can based on context. The various suggested and tested evaluations are defined in this section.

year	new languages	new tasks
2005	FIN, TUR, ENG	segmentation, ASR
2007	GER	full analysis, IR
2008	ARA	context IR
2009	-	SMT
2010	-	semi-supervised

Table 1: The evolution of the evaluations. The acronyms are explained in section 3.1.

3.1.1 Comparisons to linguistic gold standard

The first Morpho Challenge in 2005 (Kurimo et al., 2006) considered unsupervised segmentation of words into morphemes. The evaluation was based on comparing the segmentation boundaries given by the competitor's algorithm to the boundaries obtained from a gold standard analysis.

From 2007 onwards, the task was changed to full morpheme analysis, that is, the algorithm should not only locate the surface forms (i.e., word segments) of the morphemes, but find also which surface forms are realizations (allomorphs) of the same underlying morpheme. This generalizes the task for finding more meaningful units than just the realizations of morphemes that may be just individual letters or even empty strings. In applications this is useful when it is important to identify which units carry the same meaning even if they have different realizations in different words.

As an unsupervised algorithm cannot find the morpheme labels that would equal to the labels in the gold standard, the evaluation has to be based on what word forms share the same morphemes. The evaluation procedure samples a large number of word pairs, such that both words in the pair have at least one morpheme in common, from both the proposed analysis and the gold standard. The first version of the method was applied in 2007 (Kurimo et al., 2008) and 2008 (Kurimo et al., 2009a), and minor modifications were done in 2009 (Kurimo et al., 2009b). However, the organizers have reported the evaluation results of the 2007 and 2008 submissions also with the new version, thus allowing a direct comparison between them. A summary of these results for English, Finnish, German and Turkish for the best algorithms is presented in Table 2. The evaluations in 2008 and 2009 were also performed on Arabic, but these results and not comparable, because the database and the gold standard was changed between the years. The exact annual results for all participants as well as the details of the evaluation in each year can be reviewed in the annual evaluation reports (Kurimo et al., 2006; Kurimo et al., 2008; Kurimo et al., 2009a; Kurimo et al., 2009b).

Already the linguistic evaluation of Morpho Challenge 2005 applied some principles that have been used thereafter: (1) The evaluation is based on a subset of the word forms given as training data. This not only makes the evaluation procedure lighter, but also allows changing the set when

English			
Method	P	R	F
2009			
Allomorfessor	68.98	56.82	62.31
Monson PMU	55.68	62.33	58.82
Lignos	83.49	45.00	58.48
2008			
Monson P+M	69.59	65.57	67.52
Monson ParaMor	63.32	51.96	57.08
Zeman 1	67.13	46.67	55.06
2007			
Monson P+M	70.09	67.38	**68.71**
Bernhard 2	67.42	65.11	66.24
Bernhard 1	75.61	57.87	65.56

Finnish			
Method	P	R	F
2009			
Monson PMU	47.89	50.98	49.39
Monson PMM	51.75	45.42	48.38
Spiegler PROMODES C	41.20	48.22	44.44
2008			
Monson P+M	65.21	50.43	**56.87**
Monson ParaMor	49.97	37.64	42.93
Monson Morfessor	79.76	24.95	38.02
2007			
Bernhard 2	63.92	44.48	52.45
Bernhard 1	78.11	29.39	42.71
Bordag 5a	72.45	27.21	39.56

German			
Method	P	R	F
2009			
Monson PMU	52.53	60.27	56.14
Monson PMM	51.07	57.79	54.22
Monson PM	50.81	47.68	49.20
2008			
Monson P+M	64.06	61.52	**62.76**
Monson Morfessor	70.73	38.82	50.13
Monson ParaMor	56.98	42.10	48.42
2007			
Monson P+M	69.96	55.42	61.85
Bernhard 2	54.02	60.77	57.20
Bernhard 1	66.82	42.48	51.94

Turkish			
Method	P	R	F
2009			
Monson PMM	48.07	60.39	53.53
Monson PMU	47.25	60.01	52.88
Monson PM	49.54	54.77	52.02
2008			
Monson P+M	66.78	57.97	**62.07**
Monson ParaMor	57.35	45.75	50.90
Monson Morfessor	77.36	33.47	46.73
2007			
Bordag 5a	81.06	23.51	36.45
Bordag 5	81.19	23.44	36.38
Zeman	77.48	22.71	35.13

Table 2: The summary of the best three submitted methods for years 2009, 2008 and 2007 using the linguistic evaluation of Morpho Challenge 2009. The complete results tables by the organizers are available from `http://www.cis.hut.fi/morphochallenge2009/`. The three columns numbers are precision (P), recall (R), and F-measure (F). The best F-measure for each language is in boldface, and the best result that is not based on a direct combination of two other methods is underlined.

the old one is considered to be "overlearned". (2) The frequency of the word form plays no role in evaluation; rare and common forms are equally likely to be selected, and have equal weight to the score. (3) The evaluation score is balanced F-measure, the harmonic mean of precision and recall. Precision measures how many of the choices made by the algorithm are matched in gold standard; recall measures how many of the choices in the gold standard are matched in the proposed analysis. (4) If the linguistic gold standard has several alternative analysis for one word, for full precision, it is enough that one of the alternatives is equivalent to the proposed analysis. The same holds the other way around for recall.

All of the principles can be also criticized. For example, evaluation based on the full set would provide more trustworthy estimates, and common word forms are more significant in any practical application. However, the third and the fourth principle have problems that can be considered to be more serious.

Balanced F-measure favors methods that are able to get near-to-equal precision and recall. As many algorithms can be tuned to give either more or less morphemes per word than in the default case, this encourages using developments sets to optimize the respective parameters. The winning methods in Challenge 2009—Monson's ParaMor-Morfessor Union (PMU) and ParaMor-Morfessor

Mimic (PMM) (Monson et al., 2009), and Allomorfessor (Virpioja and Kohonen, 2009)—did this, more or less explicitly.[1] Moreover, it can be argued that the precision would be more important than recall in many applications, or, more generally, that the optimal balance between precision and recall is application dependent. We see two solutions for this: Either the optimization for F-measure should be allowed with a public development set, which means moving towards semi-supervised direction, or precision-recall curves should be compared, which means more complex evaluations.

The fourth principle causes problems, if the evaluated algorithms are allowed to have alternative analyses for each word. If several alternative analyses are provided, the obtained precision is about the average over the individual analyses, but the recall is based on the best of the alternatives. This property have been exploited in Challenges 2007 and 2008 by combining the results of two algorithms as alternative analyses. The method, Monson's ParaMor+Morfessor (P+M) holds still the best position measured in F-measures in all languages. Combining even better-performing methods in a similar manner would increase the scores further. To fix this problem, either the evaluation metric should require matching number of alternative analyses to get the full points, or the symmetry of the precision and recall measures has to be removed.

Excluding the methods that combine the analyses of two other methods as alternative ones, we see that the best F-measure (underlined in Table 2) is held by Monson's ParaMor-Morfessor Mimic from 2009 (Monson et al., 2009) in Turkish and Bernhard's method 2 from 2007 (Bernhard, 2006) in all the other three languages. This means that except for Turkish, there is no improvement in the results over the three years. Furthermore, both of the methods are based purely on segmentation, and so are all the other top methods presented in Table 2 except for Bordag's methods (Bordag, 2006) and Allomorfessor (Virpioja and Kohonen, 2009).

3.1.2 Speech recognition

A key factor in the success of large-vocabulary continuous speech recognition is the system's ability to limit the search space using a statistical language model. The language model provides the probability of different recognition hypothesis by using a model of the co-occurence of its words and morphemes. A properly smoothed n-gram is the most conventional model. The n-gram should consist of modeling units that are suitable for the language, typically words or morphemes.

In Morpho Challenge state-of-the-art large-vocabulary speech recognizers have been built for evaluations in Finnish and Turkish (Kurimo et al., 2006). The various morpheme analysis algorithms have been compared by measuring the recognition accuracy with different language models each trained and optimized based on units from one of the algorithms. The best results were quite near to each other, but Bernhard (Bernhard, 2006) and Morfessor Categories MAP were at the top for both languages.

3.1.3 Information retrieval

In the information retrieval task, the algorithms were tested by using the morpheme segmentations for text retrieval. To return all relevant documents, it is important to match the words in the queries to the words in the documents irrespective of which word forms are used. Typically, a stemming algorithm or a morphological analyzer is used to reduce the inflected forms to their stem or base form. The problem with these methods is that specific rules need to be crafted for each language. However, these approaches were also tested for comparison purposes. The IR experiments were carried out by replacing the words in the corpora and queries by the suggested morpheme segmentations. Test corpora, queries and relevance assessments were provided by Cross-Language Evaluation Forum (CLEF) (Agirre et al., 2008).

To test the effect of the morpheme segmentation, the number of other variables will have to be minimized, which poses some challenges. For example, the term weighting method will affect the results and different morpheme analyzers may perform optimally with different weighting approaches. TFIDF and Okapi BM25 term weighting methods have been tested. In the 2007 Challenge, it was noted that Okapi BM25 suffers greatly if the corpus contains a lot of frequent terms. These terms are often introduced when the algorithms segment suffixes from stems. To overcome this problem, a method for automatically generating stop lists of frequent terms was intro-

[1] Allomorfessor was trained with a pruned data to obtain a higher recall, whereas ParaMor-Morfessor is explicitly optimized for F-measure with a separate Hungarian data set.

duced. Any term that occurs more times in the corpus than a certain threshold is added to the stop list and excluded from indexing. The method is quite simple, but it treats all morpheme analysis methods equally as it does not require the algorithm to tag which morphemes are stems and which are suffixes. The generated stoplists are also reasonable sized and the results are robust with respect to the stop list cutoff parameter. With a stop list, Okapi BM25 clearly outperformed TFIDF ranking method for all algorithms. However, the problem of choosing the term weighting approach that treats all algorithms in an optimal way remains open.

Another challenge is analyzing the results as it is hard to achieve statistically significant results with the limited number of queries (50-60) that were available. In fact, in each language 11-17 of the best algorithms belonged to the "top group", that is, had no statistically different result to the top performer of the language. To improve the significance of the results, the number of queries should be increased. This is a known problem in the field of IR. However, it is important to test the methods in a real life application and if an algorithm gives good results across languages, there is evidence that it is doing something useful.

Some conclusions can be drawn from the results. The language specific reference methods (Porter stemming for English, two-layer morphological analysis for Finnish and German) give the best results, but the best unsupervised algorithms are almost at par and the differences are not significant. For German and Finnish, the best unsupervised methods can also beat in a statistically significant way the baseline of not doing any segmentation or stemming. The best algorithms that performed well across languages are ParaMor (Monson et al., 2008), Bernhard (Bernhard, 2006), Morfessor Baseline, and McNamee (McNamee, 2008).

Comparing the results to the linguistic evaluation (section 3.1.1), it seems that methods that perform well at the IR task tend to have good precision in the linguistic task, with exceptions. Thus, in the IR task it seems important not to oversegment words. One exception is the method (McNamee, 2008) which simply splits the words into equal length letter n-grams. The method gives surprisingly good results in the IR task, given the simplicity, but suffers from low precision in the linguistic task.

3.1.4 Machine translation

In phrase-based statistical machine translation process there are two stages where morpheme analysis and segmentation of the words into meaningful sub-word units is needed. The first stage is the alignment of the parallel sentences in the source and target language for training the translation model. The second one is training a statistical language model for the production of fluent sentences in a morphologically rich target language.

In the machine translation tasks used in the Morpho Challenge, the focus has so far been in the alignment problem. In the evaluation tasks introduced in 2009 the language-pairs were Finnish-English and German-English. To obtain state-of-the-art results, the evaluation consists of minimum Bayes risk (MBR) combination of two translation systems trained on the same data, one using words and the other morphemes as the basic modeling units (de Gispert et al., 2009). The various morpheme analysis algorithms are compared by measuring the translation performance for different two-model combinations where the word-based model is always the same, but the morpheme-based model is trained based on units from each of the algorithms in turns.

Because the machine translation evaluation has yet been tried only in 2009, it is difficult to draw conclusions about the results yet. However, the Morfessor Baseline algorithm seems to be particularly difficult to beat both in Finnish-German and German-English task. The differences between the best results are small, but the ranking in both tasks was the same: 1. Morfessor Baseline, 2. Allomorfessor, 3. The linguistic gold standard morphemes (Kurimo et al., 2009b).

3.2 Evaluated algorithms

This section attempts to describe very briefly some of the individual morpheme analysis algorithms that have been most successful in the evaluations.

Morfessor Baseline (Creutz and Lagus, 2002): This is a public baseline algorithm based on jointly minimizing the size of the morph codebook and the encoded size of the all the word forms using the minimum description length MDL cost function. The performance is above average for all evaluated tasks in most languages.

Allomorfessor (Kohonen et al., 2009; Virpioja and Kohonen, 2009): The development of this method was based on the observation that the

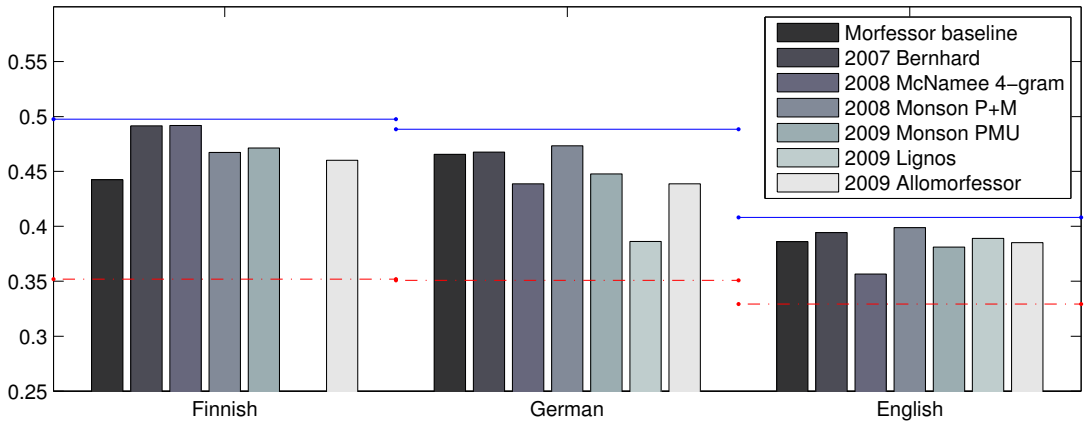

Figure 1: Mean Average Precision (MAP) values for some of the best algorithms over the years in the IR task. The upper horizontal line shows the "goal level" for each language, i.e. the performance of the best language specific reference method. The lower line shows the baseline reference of doing no stemming or analysis.

morph level surface forms of one morpheme are often very similar and the differences occur close to the morpheme boundary. Thus, the allomorphemes could be modeled by simple mutations. It has been implemented on top of the Morfessor Baseline using maximum a posteriori (MAP) optimization. This model slightly improves the performance in the linguistic evaluation in all languages (Kurimo et al., 2009b), but in IR and SMT there is no improvement yet.

Morfessor Categories MAP (Creutz and Lagus, 2005): In this method hidden Markov models are used to incorporate morphotactic categories for the Morfessor Baseline. The structure is optimized by MAP and yields slight improvements in the linguistic evaluation for most languages, but not for IR or SMT tasks.

Bernhard (Bernhard, 2006): This has been one of the best performing algorithms in Finnish, English and German linguistic evaluation and in IR (Kurimo et al., 2008). First a list of the most likely prefixes and suffixes is extracted and alternative segmentations are generated for the word forms. Then the best ones are selected based on cost functions that favour most frequent analysis and some basic morphotactics.

Bordag (Bordag, 2006): This method applies iterative LSV and clustering of morphs into morphemes. The performance in the linguistic evaluation is quite well for Turkish and decent for Finnish (Kurimo et al., 2008).

ParaMor (Monson et al., 2008): This method applies an unsupervised model for inflection rules and suffixation for the stems by building linguistically motivated paradigms. It has obtained one of the top performances for all languages when combined with the Morfessor Baseline (Kurimo et al., 2009a). Various combination methods have been tested: union, weighted probabilistic average and proposing both the analyses (Monson et al., 2009).

Lignos (Lignos et al., 2009): This method is based on the observation that the derivation of the inflected forms can be modeled as transformations. The best transformations can be found by optimizing the simplicity and frequency. This method performs much better in English than in the other languages (Kurimo et al., 2009b).

Promodes (Spiegler et al., 2009): This method presents a probabilistic generative model that applies LSV and combines multiple analysis using a committee. It seems to generate a large amount of short morphemes, which is difficult for many of the practical applications. However, it obtained the best performance for the linguistic evaluation in Arabic 2009 (Kurimo et al., 2009b), but did not survive as well in other languages, and particularly not in the IR application.

4 Open questions and challenges

Although more than 50 algorithms have already been tested in the Morpho Challenge evaluations and many lessons have been learned from the results and discussions, many challenges are still open and untouched. In fact, the attempts to solve the problem have perhaps produced even more open questions than there were in the beginning.

The main new and open challenges are described in this section.

What is the best analysis algorithm? Some of the suggested algorithms have produced good test results and some even in several tasks and languages, such as Bernhard (Bernhard, 2006), Monson ParaMor+Morfessor (Monson et al., 2008) and Allomorfessor (Virpioja and Kohonen, 2009). However, none of the methods perform really well in all the evaluation tasks and languages and their mutual performance differences are often rather small, even though the morphemes and the algorithmic principles are totally different. Thus, no dominant morpheme analysis algorithm have been found. Furthermore, reaching the performance level that rivals, or even sometimes dominates, the rule-based and language-dependent reference methods does not mean that the solutions are sufficient. Often the limited coverage or unsuitable level of details in the analysis for the task in the reference methods just indicates that they are not sufficient either and better solutions are needed. Another observation which complicates the finding and determination of the best algorithm is that in some tasks, such as statistical language models for speech recognition, very different algorithms can reach the same performance, because advanced modelling methods can compensate for unsuitable morpheme analysis.

What is the meaning of the morphemes? In some of the fundamental applications of morpheme analysis, such as text understanding, morpheme segmentation alone is only part of the solution. Even more important is to find the meaning for the obtained morphemes. The extension of the segmentation of words into smaller units to identification of the units that correspond to the same morpheme is a step taken to this direction, but the question of the meaning of the morpheme is still open. However, in the unsupervised way of learning, solutions to this may be so tightly tied to the applications that much more complex evaluations would be needed.

How to evaluate the alternative analyses? It is clear that when a word form is separated from the sentence context where it was used, the morpheme analysis easily becomes ambiguous. In the Morpho Challenge evaluations this has been taken into account by allowing multiple alternative analyses. However, in some evaluations, for example, in the measurement of the recall of the gold standard morphemes, this leads to unwanted results and may favour methods that always provide a large number of alternative analysis.

How to improve the analysis using context? A natural way to disambiguate the analysis involves taking the sentence context into account. Some of the Morpho Challenge evaluations, for example, the information retrieval, allow this option when the source texts and queries are given. However, this has not been widely tried yet by the participants, probably because of the increased computational complexity of the modelling task.

How to effectively apply semi-supervised learning? In semi-supervised learning, a small set of labeled data in the form of gold standard analysis for the word forms are provided. This data can be used for improving the unsupervised solutions based on unlabeled data in several ways: (1) The labeled data is used for tuning some learning parameters, followed by an unsupervised learning process for the unlabeled data. (2) The labeled morphemes are used as an ideal starting point to bootstrap the learning on the unlabeled words (self-training). (3) Using the EM algorithm for estimating a generative model, the unlabeled cases can be treated as missing data.

The best and most practical way of using the partly labeled data will be determined in future when the semi-supervised task has been evaluated in the future Morpho Challenge evaluations. For the first time this task will be evaluated in the ongoing Morpho Challenge 2010.

Acknowledgments

We are grateful to the University of Leipzig, University of Leeds, Computational Linguistics Group at University of Haifa, Stefan Bordag, Ebru Arisoy, Nizar Habash, Majdi Sawalha, Eric Atwell, and Mathias Creutz for making the data and gold standards in various languages available to the Challenge. This work was supported by the Academy of Finland in the project *Adaptive Informatics*, the graduate schools in Language Technology and Computational Methods of Information Technology, in part by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022, and in part by the IST Programme of the European Community, under the FP7 project EMIME (213845) and PASCAL Network of Excellence.

References

Steven Abney. 2007. *Semisupervised Learning for Computational Linguistics*. Chapman and Hall/CRC.

Eneko Agirre, Giorgio M. Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters. 2008. CLEF 2008: Ad hoc track overview. In *Working Notes for the CLEF 2008 Workshop*.

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

Stefan Bordag. 2006. Two-step approach to unsupervised morpheme segmentation. In *Proc. of the PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*, Venice, Italy. PASCAL European Network of Excellence.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proc. SIGPHON/ACL'02*, pages 21–30.

Mathias Creutz and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *Proc. AKRR'05*, pages 106–113.

Adria de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. 2009. Minimum bayes risk combination of translation hypothesis from alternative morphological decompositions. In *Proc. NAACL'09*, pages 73-76.

C. G. de Marcken. 1996. *Unsupervised Language Acquisition*. Ph.D. thesis, MIT.

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222. Reprinted 1970 in Papers in Structural and Transformational Linguistics, Reidel Publishing Company, Dordrecht, Holland.

Feng Jiao, Shaojun Wang, Chi-Hoon Lee, Russell Greiner, and Dale Schuurmans. 2006. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *Proc. ACL'06*, pages 209–216.

Oskar Kohonen, Sami Virpioja, and Mikaela Klami. 2009. Allomorfessor: Towards unsupervised morpheme analysis. In *Evaluating systems for Multilingual and MultiModal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Revised Selected Papers*, Lecture Notes in Computer Science , Vol. 5706. Springer.

Mikko Kurimo, Mathias Creutz, and Krista Lagus. 2006. Unsupervised segmentation of words into morphemes - challenge 2005, an introduction and evaluation report. In *Proc. PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*, Venice, Italy. PASCAL European Network of Excellence.

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

Mikko Kurimo, Ville Turunen, and Matti Varjokallio. 2009a. Overview of Morpho Challenge 2008. In *Evaluating systems for Multilingual and MultiModal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Revised Selected Papers*, Lecture Notes in Computer Science , Vol. 5706. Springer.

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

Constantine Lignos, Erwin Chan, Mitchell P. Marcus, and Charles Yang. 2009. A rule-based unsupervised morphology learning framework. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece.

Paul McNamee. 2008. Retrieval experiments at morpho challenge 2008. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, September.

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

Christian Monson, Kristy Hollingshead, and Brian Roard. 2009. Probabilistic paraMor. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, September.

Sebastian Spiegler, Bruno Golenia, and Peter Flach. 2009. PROMODES: A probabilistic generative model for word decomposition. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, September.

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

Xiaojin Zhu. 2005. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.

Xiaojin Zhu. 2010. Semi-supervised learning. In *Encyclopedia of Machine Learning*. To appear.