Morpho Challenge - Evaluation of algorithms for unsupervised learning of morphology in various tasks and languages

Mikko Kurimo, Sami Virpioja, Ville Turunen, Teemu Hirsimäki
Adaptive Informatics Research Centre
Helsinki University of Technology
FI-02015, TKK, Finland
Firstname.Lastname@tkk.fi

Abstract

After the release of the open source software implementation of Morfessor algorithm, a series of several open evaluations has been organized for unsupervised morpheme analysis and morpheme-based speech recognition and information retrieval. The unsupervised morpheme analysis is a particularly attractive approach for speech and language technology for the morphologically complex languages. When the amount of distinct word forms becomes prohibitive for the construction of a sufficient lexicon, it is important that the words can be segmented into smaller meaningful language modeling units. In this presentation we will demonstrate the results of the evaluations, the baseline systems built using the open source tools, and invite research groups to participate in the next evaluation where the task is to enhance statistical machine translation by morpheme analysis.

A proposal for a Type II Demo

1 Extended Abstract

1.1 The segmentation of words into morphemes

One of the fundamental tasks in natural language processing applications, such as large-vocabulary speech recognition (LVCSR), statistical machine translation (SMT) and information retrieval (IR), is the morphological analysis of words. It is particularly important for the morphologically complex languages, where the amount of different word forms is substantially increased by inflection, derivation and composition. The decomposition of words is required not only for understanding the sentence, but in many languages also for just representing the language by any tractable and trainable statistical model and lexicon. The manually composed rule-based morphological analyzers can solve these problems to some extent, but only a fraction of the existing languages have been covered so far, and for many the coverage of the relevant content is insufficient.

The objective of the Morpho Challenge[1] is to design and evaluate new unsupervised statistical machine learning algorithms that discover which morphemes (smallest individually meaningful units of language) words consist of. The goal is to discover basic vocabulary units suitable for different tasks, such as LVCSR, SMT and IR. In unsupervised learning the list of morphemes is not pre-specified for each language, but the optimal morpheme lexicon and morpheme analysis of all different word forms is statistically optimized from a large text corpus in a completely data-driven manner.

The evaluation of the morpheme analysis algorithms is performed both by a linguistic and an application oriented task. The analysis obtained for a long list of words is first compared to the linguistic gold standard representing a grammatically correct analysis by verifying that the morpheme-sharing word pairs are the correct ones (Kurimo et al., 2007). This is repeated in different languages and then the obtained decomposition of words is applied in state-of-the-art systems running various

[1] See http://www.cis.hut.fi/morphochallenge2009/

Proceedings of NAACL HLT 2009: Demonstrations, pages 13–16,
Boulder, Colorado, June 2009.

NLP applications. The suitability of the morphemes is verified by comparing the performance of the systems to each other and to systems using unprocessed words or conventional word processing algorithms like stemming or rule-based decompositions.

As a baseline method in all application, we have built systems by applying the Morfessor algorithm, which is an unsupervised word decomposition algorithm developed at our research group (Creutz and Lagus, 2002) and released as open source software implementation[2].

1.2 Morphemes in Information Retrieval

In information retrieval (IR) from text documents a typical task is to look for the most relevant documents for a given query. One of the key challenges is to reduce all the inflected word forms to a common root or stem for effective indexing. From the morpheme analysis point of view this task is to decompose all the words in the query and text documents and find out those common morphemes which form the most relevant links.

In Morpho Challenge the IR systems built using the unsupervised morpheme analysis algorithms are compared in state-of-the-art CLEF tasks in Finnish, German and English (Kurimo and Turunen, 2008) using the mean average precision metric. The results are also compared to those obtained by the grammatical morphemes as well as the stemming and word normalization methods conventionally used in IR.

1.3 Morphemes in Speech Recognition

In large-vocabulary continuous speech recognition (LVCSR) one key part of the process is the statistical language modeling which determines the prior probabilities of all the possible word sequences. An especially challenging task is to cover all the possible word forms with sufficient accuracy, because any out-of-vocabulary words will not only be never correctly recognized, but also severely degrade the modeling of the other nearby words. By decomposing the words into meaningful sub-word units, such as morphemes, large-vocabulary language models can be successfully built even for the most difficult agglutinative languages, like Finnish, Estonian and Turkish (Kurimo et al., 2006b).

[2]See http://www.cis.hut.fi/projects/morpho/

In Morpho Challenge the unsupervised morpheme algorithms have been compared by using the morphemes to train statistical language models and applying the models in state-of-the-art LVCSR tasks in Finnish and Turkish (Kurimo et al., 2006a). Benchmarks for the same tasks were obtained by models that utilize the grammatical morphemes as well as traditional word-based language models.

1.4 Morphemes in Machine Translation

The state-of-the-art statistical machine translation (SMT) systems are affected by the morphological variation of words at two different stages (Virpioja et al., 2007). In the first stage, the alignment of the source and target language words in a parallel training corpus and the training of the translation model can benefit from the decomposition of complex words into morphemes. This is particularly important when either the target or the source language, or both, are morphologically complex. The final stage where the target language text is generated, may also require morpheme-based models, because the large-vocabulary statistical language models are applied in the same way as in LVCSR.

In the on-going Morpho Challenge 2009 competition, the morpheme analysis algorithms are compared in SMT tasks, where the analysis is needed for the source language texts. The European Parliament parallel corpus (Koehn, 2005) is used in the evaluation. The source languages are Finnish and German and the target in both tasks is English. To obtain a state-of-the-art performance in the tasks the morpheme-based SMT will be combined with a word-based SMT using the Minimum Bayes Risk (MBR) interpolation of the N-best translation hypothesis of both systems (de Gispert et al., 2009).

1.5 Morpho Challenge 2009

As its predecessors, the Morpho Challenge 2009 competition is open to all and free of charge. The participants' are expected to use their unsupervised machine learning algorithms to analyze the word lists of different languages provided by the organizers and submit the results of their morpheme analysis. The organizers will then run the linguistic evaluations and build the IR and SMT systems and provide all the results and comparisons of the different systems. The participated algorithms and evaluation

results will be presented at the Morpho Challenge workshop that is currently planned to take place within the HLT-NAACL 2010 conference.

Acknowledgments

The Morpho Challenge competitions and workshops are part of the EU Network of Excellence PASCAL Challenge program and organized in collaboration with CLEF. We are grateful to Mathias Creutz, Ebru Arisoy, Stefan Bordag, Nizar Habash and Majdi Sawalha for contributions in processing the training data and creating the gold standards. The Academy of Finland has supported the work in the projects *Adaptive Informatics* and *New adaptive and learning methods in speech recognition*.

References

M. Creutz and K. Lagus. 2002. Unsupervised discovery of morphemes. In *Workshop on Morphological and Phonological Learning of ACL-02*.

A. de Gispert, S. Virpioja, M. Kurimo, and W. Byrne. 2009. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. Submitted to *HLT-NAACL*.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*.

M. Kurimo and V. Turunen. 2008. Unsupervised morpheme analysis evaluation by IR experiments – Morpho Challenge 2008. In *CLEF*.

M. Kurimo, M. Creutz, M. Varjokallio, E. Arisoy, and M. Saraclar. 2006a. Unsupervised segmentation of words into morphemes - Challenge 2005, an introduction and evaluation report. In *PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*.

M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar. 2006b. Unlimited vocabulary speech recognition for agglutinative languages. In *HLT-NAACL*.

M. Kurimo, M. Creutz, and M. Varjokallio. 2007. Morpho Challenge evaluation using a linguistic Gold Standard. In *CLEF*.

S. Virpioja, J. J. Väyrynen, M. Creutz, and M. Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *MT Summit XI*. Denmark.

2 Script outline for the demo presentation

In this demo we will present the achievements of the Morpho Challenge 2005-2008 competition in graphs and the baseline systems for various languages developed using the Morfessor algorithm for word decomposition, IR, LVCSR and SMT. The audience will also be welcome to try their own input for these baseline systems and view the results.

The script is presented below for a poster-style and try-it-yourself on laptop demo, but it will work well as a lecture-style show, too, if needed.

In the poster we illustrate the following points:

1. Basic characteristics of the unsupervised learning algorithms and morpheme analysis results in different languages (Finnish, Turkish, German, English, Arabic) as in Table 1, demo: *http://www.cis.hut.fi/projects/morpho/*.

2. The results of the evaluations against the linguistic gold standard morphemes in different languages, see e.g. Figure 1.

3. The results of the IR evaluations and comparisons to the performance of grammatical morphemes, word-based methods and stemming in different languages, see e.g. Figure 2.

4. The results of the LVCSR evaluations with comparisons to grammatical morphemes and word-based methods, see e.g. Figure 3.

5. The call for participation in the Morpho Challenge 2009 competition where the new evaluation task is using morphemes in SMT.

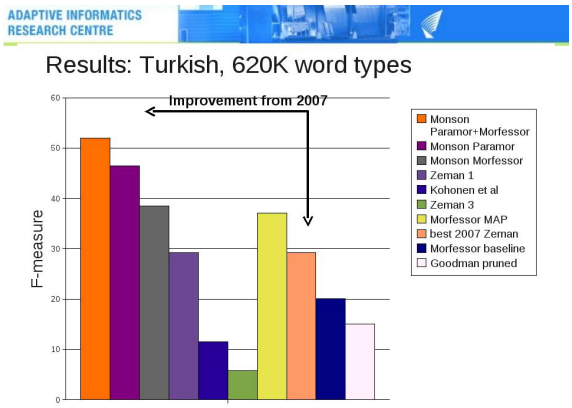

Figure 1: F-measures for the Turkish morpheme analysis.

The laptop is used to demonstrate the baseline systems we have recently developed for different tasks that are all based on unsupervised morphemes:

Example word	Morfessor analysis	Gold Standard
Finnish: linuxiin	linux +iin	linux_N +ILL
Turkish: popUlerliGini	pop +U +ler +liGini	popUler +DER_lHg +POS2S +ACC, popUler +DER_lHg +POS3 +ACC3
Arabic: AlmtHdp	Al+ mtHd +p	mut aHidap_POS:PN Al+ +SG, mut aHid_POS:AJ Al+ +SG
German: zurueckzubehalten	zurueck+ zu+ be+ halten	zurueck_B zu be halt_V +INF
English: baby-sitters	baby-+ sitter +s	baby_N sit_V er_s +PL

Table 1: Morpheme analysis examples in different languages.

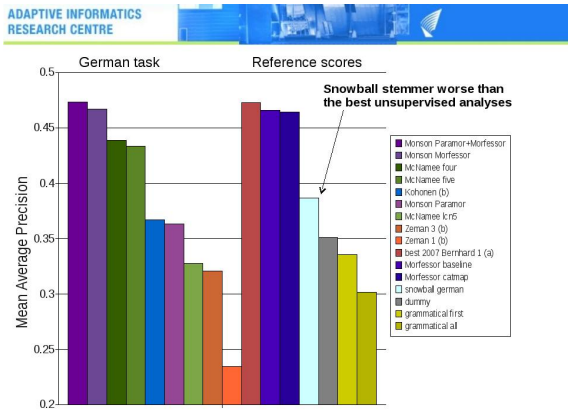

Figure 2: Precision performances for the German IR.

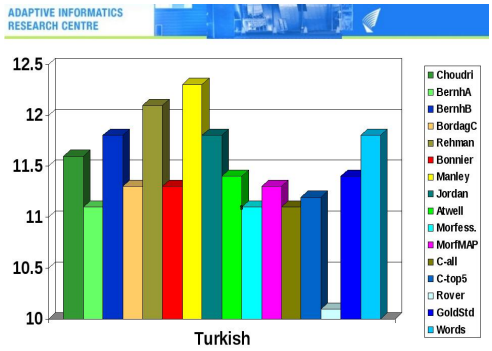

Figure 3: LVCSR error rates for the Turkish task.

1. Online LVCSR system for highly agglutinative languages, see e.g. screenshot in Figure 4.

2. Online IR system for highly agglutinative languages.

3. Online SMT system where the source language is a highly agglutinative language, see e.g. screenshot in Figure 5.

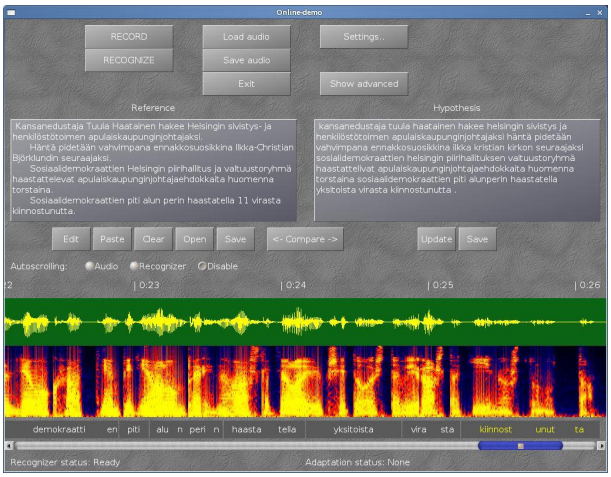

Figure 4: Screenshot of the morpheme-based speech recognizer in action for Finnish. An offline version can be tried in *http://www.cis.hut.fi/projects/speech/*.

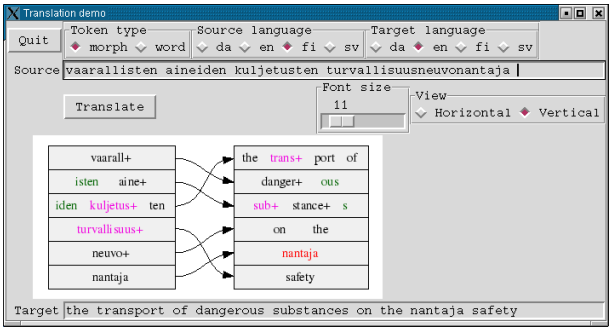

Figure 5: Screenshot of the morpheme-based machine translator in action for Finnish-English. A simplified web interface to the system is also available (please email to the authors for a link).