ParaMor and Morpho Challenge 2008

Christian Monson, Jaime Carbonell, Alon Lavie, Lori Levin
Language Technologies Institute, Carnegie Mellon University
cmonson@cs.cmu.edu

Abstract

ParaMor, our unsupervised morphology induction system performed well at Morpho Challenge 2008. When ParaMor's morphological analyses, which specialize at identifying inflectional morphology, are added to the analyses from the general purpose unsupervised morphology induction system, Morfessor, the combined system identifies the morphemes of all five Challenge languages at recall scores higher than those of any other system which competed in Morpho Challenge. In Turkish, for example, the recall of the ParaMor-Morfessor system, at 52.1%, is twice that of the next highest system that participated. These strong recall scores lead to F_1 values for morpheme identification as high as or higher than those of any competing system for all the competition languages but English. Of the three language tracks of the task-based information retrieval (IR) evaluation of Morpho Challenge, the combined ParaMor-Morfessor system placed first at average precision in the English and German tracks. And in the German and Finnish tracks of the IR task, the ParaMor-Morfessor system outperformed the hand-built stemming package, Snowball.

Categories and Subject Descriptors

I.2 [Artificial Intelligence]: I.2.7 Natural Language Processing

General Terms

Experimentation

Keywords

Natural language morphology, Unsupervised learning, morphology induction

1 Introduction

This paper describes the performance of the unsupervised morphology induction algorithm ParaMor in Morpho Challenge 2008. Morpho Challenge is a series of peer-operated competitions for algorithms designed to discover the morphological structure of individual natural languages in an unsupervised fashion. Two major considerations motivate our work on unsupervised morphology induction. First, and primarily, we are interested in developing methods to quickly bring a morphology analysis system online for a new language. Of the nearly 7000 languages in the world, only a few dozen have working computational morphological analysis systems. Unsupervised morphology induction promises to significantly decrease the time and expertise needed to build a morphology system for the many remaining languages. Second, we are interested in this problem from a theoretical standpoint: We would like to know how much of the morphology of a language it is possible to learn from nothing but raw text. Although we do not claim that our unsupervised morphology induction algorithm mimics how a human learns morphology, we hope that our work can place constraints on the range of strategies that humans might use.

A considerable number of researchers have worked on the problem of unsupervised morphology induction. Here, we summarize and categorize a few approaches that are most related to or that have significantly influenced our system. The approaches we summarize fall into one or more of four categories. The first category comprises systems which examine word-internal character transitions for probabilistic evidence of a morpheme boundary. Probably the first work to look at unsupervised morphology induction, Harris (1955), took the character transition probability approach. Harris built forward and backward character tries and noted that locations where the tries had significant branching factors were likely morpheme boundaries. More recently, Bernhard (2007) measures the probabilities of word-internal character sequences, while avoiding the data

fragmentation problems of using tries. A second category of unsupervised morphology induction system treats morphology as a minimum description length (MDL) problem. This approach views morphemes as a compact representation of natural language words: If a system can identify the morphemes of a language, then that system could efficiently encode that language. Systems that employ this MDL approach include Brent (1995), Goldsmith (2001; 2006), and Creutz's (2006) Morfessor algorithm. A third category of unsupervised morphology induction algorithm brings to bear the larger context in which a word occurs. Schone (2001) and Wicentowski (2002) note that morphologically distinct surface forms of the same lexeme will often occur in contexts of similar surrounding words. Their systems use a combination of word edit distance heuristics and latent semantic analysis of word contexts to identify morphologically related words.

The fourth and final category of unsupervised morphology induction system discussed here are systems which purposefully model the paradigmatic structure of morphology. A morphological paradigm is a mutually substitutable set of morphological operations. In particular, conjugation and declension tables, as commonly found in language text books, are paradigms. Systems that appeal to the paradigmatic structure of morphology include Goldsmith's Linguistica (2001; 2006), the system presented in Snover (2002), and ParaMor, described in this paper.

The remainder of this paper is organized as follows: Section 2 provides a brief overview of our unsupervised morphology induction algorithm, ParaMor, taking particular note of the pieces of the ParaMor algorithm which have changed since the Morpho Challenge 2007 competition. A more full description of the ParaMor algorithm has been published in the series of papers (Monson et al., 2007a; 2007b; 2008). Section 3 then presents ParaMor's results from Morpho Challenge 2008, in comparison and contrast with the other systems which have competed in Morpho Challenge.

2 The ParaMor Algorithm

An unsupervised morphology induction algorithm, such as ParaMor, discovers the morphology of a language from nothing more than raw text. The ParaMor algorithm takes as input a text and reduces the text to a list of unique word types. A priori ParaMor does not know where the morpheme boundaries fall in any given surface form, and so ParaMor proposes, for each word, a separate analysis that hypothesizes a morpheme boundary at each character boundary of that word. Whenever two or more corpus types end in the same word-final string, ParaMor constructs a paradigm seed. This paradigm seed contains the word-final string together with all word-initial strings which allow the word-final string to attach.

The ParaMor algorithm then proceeds in two main stages. In the first stage, ParaMor searches for sets of word-final strings which likely represent the suffixes of a paradigm. In the second stage, ParaMor segments word forms exactly where the discovered paradigms suggest a morpheme boundary. The paradigm discovery stage consists of three steps. The first step is a recall centric search that greedily expands ParaMor's paradigm seeds into full candidate paradigms by successively adding additional suffixes. Each of ParaMor's candidate paradigms consists of a set of word-final strings, or candidate suffixes, together with all the word-initial strings, candidate stems, which occurred in separate words with each suffix in the candidate paradigm. The second step of paradigm discovery is to cluster initially selected candidate paradigms which likely model the same underlying paradigm of a language. The clustering step merges candidate paradigms which share a large fraction of their suffixes and stems. The third step in ParaMor's paradigm discovery phase applies a series of filters to weed out paradigm candidates which likely do not model true paradigms of a language. The paradigm filters consider a range of criteria when deciding to keep or discard a candidate paradigm. These criteria include: the number of suffixes and stems in the candidate, the length of the suffixes and stems in the candidate, and the character transition probabilities that surround the morpheme boundary that the paradigm candidate hypothesizes, in the style of Harris (1955).

ParaMor's word-to-morpheme segmentation stage looks for paradigmatic evidence of a morpheme boundary. Specifically, ParaMor matches each word-final string of each word type against the suffixes in each discovered paradigm. Whenever a word-final string is identical to a suffix in a discovered paradigm, ParaMor looks for evidence that the word belongs to that particular paradigm. If a word belongs to a paradigm, then the stem of that word will likely form valid surface forms with other suffixes from that paradigm. Hence, whenever a word-final string matches a suffix of a discovered paradigm, ParaMor substitutes, one at a time, the other suffixes of that discovered paradigm. If at least one of the substituted forms occurred as a word type in the corpus, then ParaMor segments the original word form at the boundary of the matched word-final string.

2.1 Changes to ParaMor since Morpho Challenge 2007

Morpho Challenge 2008 is the second Morpho Challenge competition in which ParaMor has taken part. In Morpho Challenge 2007, ParaMor participated in the English and German competitions. This year, ParaMor again analyzed English and German morphology, but also participated in the Finnish, Turkish, and Arabic tracks. Three major additions and adaptations to the ParaMor algorithm made participation in these morphologically more challenging language tracks practical. These three adaptations are described in detail in Monson et al. (2008), while here they are only briefly summarized. The first two adaptations extend to ParaMor techniques that have been developed for other unsupervised morphology induction algorithms. These first two adaptations are designed to improve the precision of ParaMor's discovered paradigms and of the resulting word-to-morpheme segmentations.

The first adaptation restricts the set of word types which participate in ParaMor's paradigm discovery phase. Because, combinatorially, there are fewer possible short strings, words that consist of just a few characters are more likely to suggest spurious morphological relationships with other short types that occur in any particular corpus. Hence, the first adaptation excludes short types from the paradigm induction vocabulary. Since the morphological paradigms that ParaMor seeks to uncover describe large sets of word types, ParaMor can rely on the remaining longer types to identify paradigms. Other unsupervised morphology induction systems, including the Linguistica system (Goldsmith, 2006), also decide which corpus strings to trust based on their length. With fewer spurious relationships clouding the landscape, the paradigms which ParaMor identify are more precise.

The second adaptation borrows ideas originally due to Harris (1955) and Goldsmith (2006). This adaptation is designed to remove initially discovered paradigms which incorrectly hypothesize a morpheme boundary internal to a true suffix. The adaptation measures the entropy in the distribution of stem-final characters in each candidate paradigm. ParaMor discards candidates with an entropy below a parameterized threshold. Low stem-final character entropy is a strong indication of a morpheme boundary placed internal to a suffix.

The final adaptation to the ParaMor system from the 2007 Challenge acknowledges the agglutinative structure of natural language morphology: Many natural languages, including Turkish and Finnish, form surface words from several morphemes in sequence. Any individual candidate paradigm that ParaMor constructs during the paradigm identification phase can propose at most a single morpheme boundary in any particular word. Our third adaptation straightforwardly merges the separate morpheme boundaries that are proposed by distinct candidate paradigms into a single combined morphological segmentation that contains multiple morpheme boundaries.

2.2 Combining ParaMor with Morfessor

As described earlier, the unsupervised morphology induction system ParaMor is designed to identify morphological paradigms: sets of mutually substitutable morphological operations. In particular, ParaMor looks for sets of mutually substitutable suffixes. Paradigms are the structure of inflectional morphology. In inflectional morphology any given lexeme adheres to a paradigm forms a separate surface form with each member of the paradigm. But the Morpho Challenge specifically evaluates morphology analysis systems on both inflectional and derivational morphology. Derivational morphology is much more idiosyncratic: any particular stem may or may not form a new word with any particular derivational suffix.

To more practically compete in Morpho Challenge we add to ParaMor's morphological analyses the morphological analyses suggested by the unsupervised morphology induction system Morfessor (Creutz, 2006). Morfessor is designed to identify all concatenative morphology, whether inflectional or derivational. Because a single word may have multiple legitimate morphological analyses, Morpho Challenge permits participants to submit multiple analyses of each particular word. In our combined ParaMor-Morfessor system, we submit the ParaMor and the Morfessor segmentations of each word as separate analyses of that word—as if each word were ambiguous between a ParaMor and a Morfessor analysis. Additional discussion of ParaMor's performance on inflectional and derivational morphology can be found in Monson (2007a).

3 Results

Morpho Challenge 2008 evaluated unsupervised morphology induction systems in two ways (Kurimo et al., 2008). First, systems competed in a linguistic evaluation that measured precision, recall, and F_1 at morpheme identification. And second, Morpho Challenge evaluated competing systems by measuring improvement on an information retrieval task. Specifically, Morpho Challenge replaced the words of a set of documents and the words of a set of queries with each system's morphological analyses and measured average precision.

Table 1 summarizes the results of the linguistic evaluation. Systems competed in up to five languages in the linguistic evaluation: English, German, Finnish, Turkish, and Arabic. Table 1 contains the scores of nine individual unsupervised morphology induction algorithms. Six of these nine systems competed in Morpho Challenge 2008, while three systems participated in the 2007 challenge. The scores from the 2007 competition are directly comparable to scores from the 2008 challenge because:

1. The linguistic evaluation of Morpho Challenge 2007 used the same evaluation methodology as the 2008 challenge; and moreover,
2. The 2007 challenge scored systems over the same corpora and against the same answer key as the more recent 2008 competition.

Of the six systems which competed in the 2008 challenge that appear in Table 1, three are systems we, Monson et al., submitted, while three are systems submitted by others. The three systems which we entered in Morpho Challenge 2008 are:

1. The ParaMor system alone,
2. A version of Morfessor (Creutz, 2006) which we trained ourselves, and
3. Our ParaMor and Morfessor analyses submitted as alternate, ambiguous, analyses.

		Monson et al. 2008			Other Authors 2008			2007		
		ParaMor + Morfessor	ParaMor	Morfessor	Morfessor MAP	Zeman	Kohonen	Bernhard	Bordag	Pitler
English	P	50.6	58.5	77.2	82.2	53.0	83.4	61.6	59.7	74.7
	R	63.3	48.1	34.0	33.1	42.1	13.4	60.0	32.1	40.6
	F_1	56.3	52.8	47.2	47.2	46.9	23.1	**60.8**	41.8	52.6
German	P	49.5	53.4	67.2	67.6	53.1	87.9	49.1	60.5	-
	R	59.5	38.2	36.8	36.9	28.4	7.4	57.4	41.6	-
	F_1	**54.1**	44.5	47.6	47.8	37.0	13.7	52.9	49.3	-
Finnish	P	49.8	46.4	77.4	76.8	58.5	92.6	59.7	71.3	-
	R	47.3	34.4	21.5	27.5	20.5	6.9	40.4	24.4	-
	F_1	**48.5**	39.5	33.7	40.6	30.3	12.8	**48.2**	36.4	-
Turkish	P	51.9	56.7	73.9	76.4	65.8	93.3	73.7	81.3	-
	R	52.1	39.4	26.1	24.5	18.8	6.2	14.8	17.6	-
	F_1	**52.0**	46.5	38.5	37.1	29.2	11.5	24.7	28.9	-
Arabic	P	79.8	78.6	90.4	90.2	77.2	-	-	-	-
	R	27.5	8.5	21.0	21.0	12.7	-	-	-	-
	F_1	**40.9**	15.4	34.0	34.0	21.9	-	-	-	-

Table 1: Results from the linguistic evaluation of Morpho Challenge. The unsupervised morphology induction systems which appear in this table are the nine best systems from the 2008 and 2007 challenges. Systems participated in up to 5 language tracks. In each language track all participating systems were scored at precision (P), recall (R), and F_1 of morpheme identification. The ground truth against which Morpho Challenge compares systems is a morphologically analyzed answer key that includes both inflectional and derivational morphology. For each language track, the system or systems which place first at F_1 by a statistically significant margin appear in **bold**.

The ParaMor algorithm has several free parameters that control the paradigm discovery phase. These parameters were set to values that produced reasonable Spanish paradigms. The parameters were then frozen before running the Morpho Challenge experiments. The six systems in Table 1 which were prepared by others are the systems with the top performance in the linguistic evaluation of Morpho Challenge 2007/2008. The system labeled Morfessor MAP is the same Morfessor algorithm as the Morfessor system which we submitted but with a different parameter setting. A change in parameter setting can sometimes result in quite different performance for Morfessor, c.f. Finnish. The remaining five systems found in Table 1 bear the names of their principle authors.

Although ParaMor alone performs respectably, it is when ParaMor's analyses are combined with Morfessor's that ParaMor shines. In all languages but English, the combined ParaMor-Morfessor system achieves the highest F_1 of any system which competed in the 2007 or 2008 Challenges. In general, the ParaMor-Morfessor system attains this higher F_1 by balancing precision and recall. Where the other unsupervised morphology induction systems of Table 1 tend to be cautious, only proposing morphemes when they have high confidence, the ParaMor-Morfessor system is more willing to guess at morphemes which may be incorrect. The more cautious high-confidence strategy results in higher precision but lower recall. In contrast ParaMor's strategy lowers precision but increases recall, balancing the two, and, overall, raising F_1.

The language ParaMor performs most poorly at is Arabic. New to Morpho Challenge in 2008, Arabic's morphology is distinctly different from that of the other four languages in the challenge. Arabic morphology differs most notably in possessing templatic morphology, where a consonantal root is interleaved with vowels to produce specific surface forms. Equally important, from ParaMor's perspective, is that Arabic is the only language in Morpho Challenge with significant prefixation. Arabic verbal morphology includes inflectional prefixes. In addition, Arabic orthography attaches a number of common determiners and prepositions directly onto the written form of the following word. These attached function words act as prepositions in text. In general, all the systems which competed in Arabic identified less than a third of the morphemes of Arabic. In particular, as ParaMor is limited to looking for suffixes, both the templatic morphology and the prefixational morphology lower ParaMor's morpheme recall. In the near term, since prefixes are the mirror image of suffixes, a simple augmentation could allow ParaMor to analyze prefixation. The ability to identify prefixes would not only improve morpheme recall in Arabic, but help identify German verbal prefixes, and English derivational prefixes as well. Interestingly, when ParaMor's Arabic analyses are presented in combination with Morfessor's the increase in recall between the two systems is practically additive: implying very little overlap between the morphemes which the two systems identify. When recall scores are depressed across the board, any increase in recall implies an increase in F_1. And indeed, the ParaMor-Morfessor system receives the highest F_1 of any system which analyzed Arabic morphology.

Tables 2 and 3 contain the results of the task-based information retrieval evaluation of Morpho Challenge 2008. The IR evaluation only covered three languages: English, German, and Finnish. These same three IR tracks also appeared in Morpho Challenge 2007 with the same evaluation set as for the 2008 challenge, making results from both years comparable. Table 2 contains the average precision IR scores for the eight best performing systems from the 2007 and 2008 challenges; while Table 3 contains average precision scores for four baseline metrics used in Morpho Challenge 2008, namely:

1. *No Morphology* - where the IR experiments run over the raw documents and queries;
2. *Snowball (Porter)* - where all words in each document and query are stemmed using the Snowball package of language stemmers. In the case of English, the Snowball stemmer is the Porter stemmer;
3. *Answer Key* - where document and query words are replaced with their morphological analyses from the answer keys that were used in the linguistic evaluation of Morpho Challenge. The answer keys used in the linguistic evaluations contain only a subset of full set of types found in the IR evaluation; and
4. *Two-Level* - where all words are replaced with the morphological analysis provided by a hand-built rule-based morphological analysis system. No hand-built morphological analysis system was evaluated for German.

In the IR evaluation of Morpho Challenge 2008 the combined ParaMor-Morfessor system placed first in English and German, and fourth in Finnish. The IR evaluation is a black-box experiment, and so it is not completely clear why the ParaMor-Morfessor system fared worse in the Finnish track. The most likely explanation is that replacing each word in each document and query with *both* the ParaMor *and* the Morfessor analyses is inappropriate for a language with complex morphology such as Finnish. It is unfortunate that Morpho Challenge did not evaluate an IR experiment for the morphologically complex Turkish and Arabic. It would be particularly interesting to see ParaMor's IR performance on Turkish, which, like Finnish, is agglutinative.

In comparison to the baseline algorithms of Table 3, all the unsupervised morphology induction systems of Table 2, including the two systems which incorporate ParaMor, perform well. Most notably, in all languages, all

	Monson et al. 2008			Other Authors 2008			2007	
	ParaMor + Morfessor	ParaMor	Morfessor	Morfessor MAP	Morfessor Baseline	McNamee	Bernhard	Bordag
English	**39.9**	39.3	36.4	37.1	38.6	36.3	39.4	34.3
German	**47.3**	36.3	46.7	46.4	46.6	43.9	**47.3**	43.1
Finnish	46.7	39.7	46.8	44.4	44.3	**49.2**	**49.2**	43.1

Table 2: Average precision scores for unsupervised morphology induction systems which participated in the Information Retrieval (IR) evaluation of Morpho Challenge. The unsupervised morphology induction systems which appear in this table are the eight best systems from the 2008 and 2007 challenges. Systems participated in up to three language tracks. The best performing system(s) for each track appear in **bold**.

	No Morphology	Snowball (Porter)	Answer Key	Two-Level
English	32.9	40.8	37.3	39.6
German	35.1	38.7	33.5	-
Finnish	35.2	42.8	43.1	49.8

Table 3: Average precision scores of four reference algorithms for the Information Retrieval (IR) evaluation of Morpho Challenge.

of the unsupervised morphology induction systems of Table 2 improve on the average precision scores when no morphological analysis is performed. The best performing unsupervised systems, including ParaMor, also outperform the baseline *Answer Key* scenario: demonstrating that imperfect morphological analysis can trump partial analysis. ParaMor and the other unsupervised system face stiffer competition in the two hand-built morphological baselines that have some generalization capacity, *Snowball (Porter)* and *Two-Level*. The Porter stemmer has the best average precision of any method against English; but unsupervised systems, ParaMor among them, outperform the Snowball rule-based stemmers for both German and Finnish. And finally, the hand-built two-level morphological analyzer performs best of any method on Finnish; and nearly as good as the best unsupervised system on English.

4 Conclusions

The premise that the paradigmatic structure of morphology can be leveraged toward unsupervised morphology induction is clearly justified by the state-of-the-art performance of the ParaMor algorithm in Morpho Challenge 2008. In addition, the improved performance that results from joining the morphological analyses of the ParaMor and Morfessor systems demonstrates that current unsupervised morphology algorithms are highly complementary, and have much to gain from uniting their unique strengths.

While we are pleased with ParaMor's performance in Morpho Challenge 2008, we also see significant room for improvement on ParaMor's morphology induction algorithms. A careful examination of the paradigms which ParaMor produces over Spanish data identifies two major error classes. The first class of erroneous candidate paradigm results from inadequate clustering of initially selected candidate paradigms. We would like to more tightly integrate the search and clustering phases of ParaMor to enable more complete clustering of the initially selected partial paradigms. The second major class of erroneous paradigm is a consequence of morphophonology. Specifically, a stem or a suffix may appear in different surface forms conditioned on the morphemes with which it occurs. To conflate the varied surface forms of a single underlying morpheme we believe we will need to look at evidence outside the word as Schone (2001) and Wicentowski (2002) do.

Acknowledgements

We thank the organizing committee of Morpho Challenge 2008 for running the logistics of the Morpho Challenge competition. We are particularly grateful for the individualized attention Mikko Kurimo gave our particular situation.

The research described in this paper was supported by NSF grants IIS-0121631 (AVENUE) and IIS-0534217 (LETRAS), with supplemental funding from NSF's Office of Polar Programs and Office of International Science and Education.

References

Bernhard, Delphine. Simple Morpheme Labeling in Unsupervised Morpheme Analysis. *Working Notes for the CLEF 2007 Workshop*. Budapest, Hungary, 2007.

Brent, Michael R., Sreerama K. Murthy, and Andrew Lundberg. Discovering Morphemic Suffixes: A Case Study in MDL Induction. *The Fifth International Workshop on Artificial Intelligence and Statistics*. Fort Lauderdale, Florida, 1995.

Creutz, Mathias. *Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition.* Ph.D. Thesis. Computer and Information Science, Report D13. Helsinki: University of Technology, Espoo, Finland, 2006.

Goldsmith, John. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*. 27.2:153-198. 2001.

Goldsmith, John. An Algorithm for the Unsupervised Learning of Morphology. *Natural Language Engineering*. 12.4:335-351. 2006.

Harris, Zellig. From Phoneme to Morpheme. *Language* 31.2:190-222. 1955. Reprinted in Harris (1970).

Harris, Zellig. *Papers in Structural and TransformationalLinguists*. Ed. D. Reidel, Dordrecht. 1970.

Kurimo, Mikko, Ville Turunen, and Matti Varjokallio. Unsupervised Morpheme Analysis -- Morpho Challenge 2008. August 11, 2008. <http://www.cis.hut.-fi/morphochallenge2008/>. 2008.

Monson, Christian, Jaime Carbonell, Alon Lavie, and Lori Levin. ParaMor: Minimally Supervised Induction of Paradigm Structure and Morphological Analysis. *Computing and Historical Phonology: The Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*. Prague, Czech Republic, 2007a.

Monson, Christian, Jaime Carbonell, Alon Lavie, and Lori Levin. ParaMor: Finding Paradigms across Morphology. W*orking Notes for the CLEF 2007 Workshop*. Budapest, Hungary, 2007b.

Monson, Christian, Alon Lavie, Jaime Carbonell, and Lori Levin. Evaluating an Agglutinative Segmentation Model for ParaMor. *The Tenth Meeting of the ACL Special Interest Group on Computational Morphology and Phonology.* Columbus, Ohio, USA, 2008.

Schone, Patrick, and Daniel Jurafsky. Knowledge-Free Induction of Inflectional Morphologies. *North American Chapter of the Association for Computational Linguistics*. Pittsburgh, Pennsylvania, 2001.

Snover, Matthew G. *An Unsupervised Knowledge Free Algorithm for the Learning of Morphology in Natural Languages.* M.S. Thesis. Computer Science, Sever Institute of Technology, Washington University, Saint Louis, Missouri, 2002.

Wicentowski, Richard. *Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework.* Ph.D. Thesis. Johns Hopkins University, Baltimore, Maryland, 2002.