Evaluating an Agglutinative Segmentation Model for ParaMor

Christian Monson, Alon Lavie, Jaime Carbonell, Lori Levin
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15217, USA
{cmonson, alavie, jgc, lsl}@cs.cmu.edu

Abstract

This paper describes and evaluates a modification to the segmentation model used in the unsupervised morphology induction system, ParaMor. Our improved segmentation model permits multiple morpheme boundaries in a single word. To prepare ParaMor to effectively apply the new agglutinative segmentation model, two heuristics improve ParaMor's precision. These precision-enhancing heuristics are adaptations of those used in other unsupervised morphology induction systems, including work by Hafer and Weiss (1974) and Goldsmith (2006). By reformulating the segmentation model used in ParaMor, we significantly improve ParaMor's performance in all language tracks and in both the linguistic evaluation as well as in the task based information retrieval (IR) evaluation of the peer operated competition Morpho Challenge 2007. ParaMor's improved morpheme recall in the linguistic evaluations of German, Finnish, and Turkish is higher than that of any system which competed in the Challenge. In the three languages of the IR evaluation, our enhanced ParaMor significantly outperforms, at average precision over newswire queries, a morphologically naïve baseline; scoring just behind the leading system from Morpho Challenge 2007 in English and ahead of the first place system in German.

1 Unsupervised Morphology Induction

Analyzing the morphological structure of words can benefit natural language processing (NLP) applications from grapheme-to-phoneme conversion (Demberg et al., 2007) to machine translation (Goldwater and McClosky, 2005). But many of the world's languages currently lack morphological analysis systems. Unsupervised induction could facilitate, for these lesser-resourced languages, the quick development of morphological systems from raw text corpora. Unsupervised morphology induction has been shown to help NLP tasks including speech recognition (Creutz, 2006) and information retrieval (Kurimo et al., 2007b). In this paper we work with languages like Spanish, German, and Turkish for which morphological analysis systems already exist.

The baseline ParaMor algorithm which we extend here competed in the English and German tracks of Morpho Challenge 2007 (Monson et al., 2007b). The peer operated competitions of the Morpho Challenge series standardize the evaluation of unsupervised morphology induction algorithms (Kurimo et al., 2007a; 2007b). The ParaMor algorithm showed promise in the 2007 Challenge, placing first in the linguistic evaluation of German. Developed after the close of Morpho Challenge 2007, our improvements to the ParaMor algorithm could not officially compete in this Challenge. However, the Morpho Challenge 2007 Organizing Committee (Kurimo et al., 2008) graciously oversaw the quantitative evaluation of our agglutinative version of ParaMor.

1.1 Related Work

A variety of approaches to unsupervised morphology induction have shown promise in past work: Here we highlight three techniques which have been used in a number of unsupervised morphology induction algorithms. Since character sequences are less predictable at morpheme boundaries than within any particular morpheme (see discussion in section 2.1), a first unsupervised mor-

Proceedings of the Tenth Meeting of the ACL Special Interest Group on Computational Morphology and Phonology, pages 49–58,
Columbus, Ohio, USA June 2008.

phology induction technique measures the predictability of word-internal character sequences. Harris (1955) was the first to propose the branching factor of the character tree of a corpus vocabulary as a measure of character predictability. Character trees have been incorporated into a number of more recently proposed unsupervised morphology induction systems (Schone and Jurafsky, 2001; Wicentowski, 2002; Goldsmith, 2006; Bordag, 2007). Johnson and Martin (2003) generalize from character trees and model morphological character sequences with minimized finite state automata. Bernhard (2007) measures character predictability by directly computing transitional probabilities between substrings of words.

A second successful technique has used the minimum description length principle to capture the morpheme as a recurrent structure of morphology. The Linguistica system of Goldsmith (2006), the Morfessor system of Creutz (2006), and the system described in Brent et al. (1995) take this approach.

A third technique leverages inflectional paradigms as the organizational structure of morphology. The ParaMor algorithm, which this paper extends, joins Snover (2002), Zeman (2007), and Goldsmith's Linguistica in building morphology models around the paradigm.

ParaMor tackles three challenges that face morphology induction systems which Goldsmith's Linguistica algorithm does not yet address. First, section 2.2 of this paper introduces an agglutinative segmentation model. This agglutinative model segments words into as many morphemes as the data justify. Although Goldsmith (2001) and Goldsmith and Hu (2004) discuss ideas for segmenting individual words into more than two morphemes, the implemented Linguistica algorithm, as presented in Goldsmith (2006), permits at most a single morpheme boundary in each word. Second, ParaMor decouples the task of paradigm identification from that of word segmentation (Monson et al., 2007b). In contrast, morphology models in Linguistica inherently encode both a belief about paradigm structure on individual words as well as a segmentation of those words. Without ParaMor's decoupling of paradigm structure from specific segmentation models, our algorithm for agglutinative segmentation (section 2.2) would not have been possible. Third, the evaluation of ParaMor in this paper is over much larger corpora than any published evaluation of Linguistica. Goldsmith (2006) segments the Brown corpus of English, which, after discarding numbers and punctuation, has a vocabulary size of 47,607 types. Using Linguistica, Creutz (2006) successfully segments a Finnish corpus of 250,000 tokens (approximately 130,000 types), but Creutz notes that Linguistica is memory intensive and not runable for larger corpora. In the evaluations of Morpho Challenge 2007, ParaMor segmented the words from corpora with over 42 million tokens and vocabularies as large as 2.2 million types.

2 ParaMor

This section briefly outlines the high level structure of ParaMor as described in detail in Monson et al. (2007a; 2007b). ParaMor takes the inflectional paradigm as the basic building block of morphology. A paradigm is a mutually substitutable set of morphological operations. For example, most adjectives in Spanish inflect for two paradigms. First, adjectives are marked for *gender*: an *a* suffix marks *feminine*, an *o masculine*. Then Spanish adjectives mark *number*: an *s* suffix signals *plural*, while no marking, *Ø* in this paper, indicates *singular*. The four surface forms of the cross-product of the *gender* and *number* paradigms on the Spanish word for 'beautiful' are then: *bello*, *bella*, *bellos*, and *bellas*.

ParaMor is a two stage algorithm. In the first stage, ParaMor identifies candidate paradigms which likely model suffixes of morphological paradigms and their cross-products. Since some 70% of the world's languages are significantly suffixing (Dryer, 2005), ParaMor only attempts to identify suffix paradigms. ParaMor's first stage consists of three pipelined steps. In the first step, ParaMor searches a space of candidate partial paradigms, called schemes, for those which possibly model suffixes of true paradigms. The second step merges selected schemes which appear to model the same paradigm. And in the third step, ParaMor discards scheme clusters which likely do not model true paradigms.

The second stage of the ParaMor algorithm segments word forms using the candidate paradigms identified in the first stage. Section 2.2 of this paper introduces a new segmentation model for ParaMor's second stage that allows more than one morpheme boundary in a single word—as is

Rank	Model of						Good	Error			Candidate Suffixes		Candidate Stems
	Noun	Adjective	Verb			Derivation		Stem Internal	Suffix Internal	Chance			
			ar	er	ir								
1	•	•					•				2	Ø.s	5513 apoyada, barata, hombro, oficina, reo, …
2		•					•				4	a.as.o.os	899 apoyad, captad, dirigid, junt, próxim, …
3			•						•		14	Ø.ba.ban.da.das.do.dos.n.ndo.r.ron.rse.rá.rán	25 apoya, disputa, lanza, lleva, toma, …
5			•				•				15	a.aba.aban.ada.adas.ado.ados.an.ando.ar.aron.arse.ará.arán.ó	24 apoy, desarroll, disput, lanz, llev, …
11		•				•		•			5	ta.tamente.tas.to.tos	22 cier, direc, insóli, modes, sangrien, …
12			•			•			•		14	Ø.ba.ción.da.das.do.dos.n.ndo.r.ron.rá.rán.ría	16 acepta, concentra, fija, provoca, …
13			•				•				15	a.aba.ada.adas.ado.ados.an.ando.ar.aron.ará.arán.e.en.ó	20 apoy, declar, enfrent, llev, tom, …
30				•	•		•				11	a.e.en.ida.idas.ido.idos.iendo.ieron.ió.ía	15 cumpl, escond, recib, transmit, vend, …
1000										•	3	Ø.g.gs	4 h, k, on, s
1566					•		•				4	ido.idos.ir.iré	6 conclu, cumpl, distribu, exclu, reun, segu
2000						•		•			2	lia.liana	5 austra, ita, ju, sici, zu
3000										•	3	Ø.a.anar	4 all, am, g, s
4000										•	3	Ø.e.ince	4 l, pr, qu, v
8000			•					•			2	trada.tramos	3 concen, demos, encon

Table 1. Candidate partial paradigms, or schemes, that the baseline ParaMor algorithm selected during its first step, search, of its first stage, paradigm identification. This baseline ParaMor run was over a Spanish newswire corpus of 50,000 types. While some selected schemes contain suffixes from true paradigms, other schemes contain incorrectly segmented candidate suffixes.

needed to correctly segment Spanish *plural* adjectives. As this agglutinative segmentation model relies on the paradigms learned in ParaMor's first stage, section 2.1 presents solutions to two types of paradigm model error that the baseline ParaMor system makes. The solutions to these two error types are similar in nature to ideas proposed in the unsupervised morphology induction work of Hafer and Weiss (1974) and Goldsmith (2006).

2.1 Precision at Paradigm Identification

Table 1 presents 14 of the more than 8000 schemes identified during one baseline run of ParaMor's scheme search step. Each row of Table 1 lists a scheme that was selected while searching over a Spanish newswire corpus of 50,000 types. On the far left of Table 1, the *Rank* column states the ordinal rank at which that row's scheme was selected during the search procedure: the first scheme ParaMor selects is *Ø.s*; *a.as.o.os* is the second; *ido.-idos.ir.iré* is the 1566th selected scheme, etc. The right four columns of Table 1, present raw data on the selected schemes, giving the number of candidate suffixes in that scheme, the proposed suffixes themselves, the number of candidate stems in the scheme, and a sample of those candidate stems. Each candidate stem in a ParaMor scheme forms a word that occured in the input corpus with each candidate suffix belonging to that scheme; for example, from the first selected scheme, the candidate stem *apoyada* joins to the candidate suffix *s* to form the word *apoyadas* 'supported *(adjective feminine plural)*'—a word which occured in the Spanish newswire corpus.

Between the rank on the left and the scheme details on the right of Table 1, are columns which categorize the scheme on its success, or failure, to model a true paradigm of Spanish. A dot appears in the columns marked *Noun*, *Adjective*, or *Verb* if the majority of the candidate suffixes in a row's scheme attempt to model suffixes in a paradigm of that part of speech. A dot appears in the *Derivation* column if one or more candidate suffixes of the scheme models a Spanish derivational suffix. The *Good* column is marked if the candidate suffixes of a scheme take the surface form of true paradigmatic suffixes. Initially selected schemes in Table 1 that correctly capture suffixes of real Spanish paradigms are the 1st, 2nd, 5th, 13th, 30th, and 1566th selected schemes. While some smaller paradigms of Spanish are perfectly identified (including *Ø.s*, which marks singular and plural on many nouns and adjectives, and the adjectival cross-product paradigm of gender and number, *a.as.o.os*) many selected schemes do not satisfactorily model Spanish suffixes. Incorrect schemes in Table 1 are marked in the *Error* columns.

The vast majority of unsatisfactory paradigm models fail for one of two reasons. First, many schemes contain candidate suffixes which system-

atically misanalyze word forms. These schemes consistently hypothesize either stem-internal or suffix-internal morpheme boundaries. Schemes which hypothesize incorrect morpheme boundaries include the 3rd, 11th, 12th, 2000th, and 8000th selected schemes of Table 1. Among these, the 3rd and 12th selected schemes place morpheme boundaries internal to true suffixes. For example, the 3rd selected scheme contains truncated forms of suffixes that occur correctly in the 5th selected scheme. Symmetrically, the candidate suffixes in the 11th, 2000th, and 8000th selected schemes hypothesize morpheme boundaries internal to true Spanish stems, inadvertently including portions of stems within their suffix lists. In a random sample of 100 schemes from the 8240 schemes that the baseline ParaMor algorithm selects over our Spanish corpus, 59 schemes hypothesized an incorrect morpheme boundary.

The second most prevalent reason for model failure occurs when the candidate suffixes of a scheme are related not by belonging to the same paradigm, but rather by a chance co-occurrence on a few candidate stems of the text. Schemes which arise from chance string collisions in Table 1 include the 1000th, 3000th, and 4000th selected schemes. The string lengths of the candidate stems and candidate suffixes of these chance schemes are often quite short. The longest candidate stem in any of the three chance-error schemes of Table 1 is three characters long; and all three selected schemes propose the suffix Ø, which has length zero. Short stems and short suffixes in selected schemes are easily explained combinatorially: The inventory of possible strings grows exponentially with the length of the string. Because there just aren't very many length one, length two, or even length three strings, it should come as no surprise when a variety of candidate suffixes happen to occur attached to the same set of short stems. In our random sample of 100 initially selected schemes, 35 were erroneously selected as a result of a chance collision of word types.

The next two sub-sections present solutions to the two types of paradigm model failure in the baseline algorithm that are exemplified in Table 1. These first two extensions aim to improve precision by reducing the number of schemes ParaMor erroneously selects.

Correcting Morpheme Boundary Errors

Most of the baseline selected schemes which incorrectly hypothesize a morpheme boundary do so at stem-internal positions. Indeed, in our random sample of 100 schemes, 51 of the 59 schemes with morpheme boundary errors incorrectly hypothesized a boundary stem-internally. For this reason, the baseline ParaMor algorithm already discarded schemes that likely misplace a boundary stem-internally (Monson et al., 2007b). Although there are fewer schemes that misplace a morpheme boundary suffix-internally, suffix-internal error schemes contain short suffixes that can generalize to segment a large number of word forms. (See section 2.2 for a description of ParaMor's morphological segmentation model). To measure the influence of suffix-internal error schemes on morpheme segmentation, we examined ParaMor's baseline segmentations of a random sample of 100 word forms from the 50,000 words of our Spanish corpus. In these 100 words, 82 morpheme boundaries were introduced that should not have been. And 40 of these 82 incorrectly proposed boundaries were placed by schemes which hypothesized a morpheme boundary internal to true suffixes.

To address the problem of suffix-internal misplaced boundaries we adapt an idea originally proposed by Harris (1955) and extended by Hafer and Weiss (1974): Take any string t. Let F be the set of strings such that for each $f \in F$, $t.f$ is a word form of a particular natural language. Harris noted that when the boundaries between t and each f fall at morpheme boundaries, the strings in F typically begin in a wide variety of characters; but when the t-f boundaries are morpheme-internal, each legitimate word final string must first complete the erroneously split morpheme, and so the strings in F will begin with one of a very few characters. This argument similarly holds when the roles of t and f are reversed. Hafer and Weiss (1974) describe a number of variations to Harris' letter variety algorithm. Their most successful variation uses entropy to measure character variety.

Goldsmith's (2006) Linguistica algorithm pioneered the use of entropy in a paradigm-based unsupervised morphology induction system. Linguistica measures the entropy of stem-final characters in a set of initially selected paradigm models. When entropy falls below a threshold, Linguistica considers relocating the morpheme boundary of

each word covered by that paradigm model. If, after boundary relocation, the resulting description length of Linguistica's morphology model decreases, Linguistica accepts the relocated boundaries.

To identify suffix-internal morpheme boundary errors among ParaMor's initially selected schemes, we follow Hafer and Weiss (1974) and Goldsmith (2006) in using entropy as a measure of the variety in boundary-adjacent character distributions. In a ParaMor style scheme, the candidate stems form a set of word-initial strings, and the candidate suffixes a set of word-final strings. If a scheme's stems end in a very few unique characters, the scheme has likely hypothesized an incorrect suffix-internal morpheme boundary. Consider the 3rd selected scheme in Table 1. All 25 of the 3rd scheme's stems end in the character 'a'. Consequently, we measure the entropy of the distribution of final characters in each scheme's candidate stems. Where Linguistica modifies paradigm models which appear to incorrectly place morpheme boundaries, our extension to ParaMor permanently removes schemes. To avoid introducing a free parameter, our extension to ParaMor flags a scheme as a likely boundary error only when virtually all of that scheme's candidate stems end in the same character. We flag a scheme if its entropy is below a threshold set close to zero, 0.5. The baseline ParaMor algorithm discards schemes which it believes hypothesize an incorrect stem-internal morpheme boundary only after the scheme clustering step of ParaMor's paradigm identification stage. Our extension follows suit: If we flag more than half of the schemes in a cluster as likely proposing a suffix-internal boundary, then we discard that cluster. Referencing Table 1, this first extension to ParaMor successfully removes both the 3rd and the 12th selected schemes.

Correcting Chance String Collision Errors

Scheme errors due to chance string collisions are the second most prevalent error type. As described above, the string lengths of the candidate stems and suffixes of chance schemes are typically short. When the stems and suffixes of a scheme are short, then the underlying types which support a scheme are also short. Where the baseline ParaMor algorithm explicitly builds schemes over all types in a corpus, we modify ParaMor to exclude short types from the vocabulary during morphology induction. Goldsmith (2006) also uses string-length thresholds to restrict what paradigm models the Linguistica algorithm produces.

Excluding short types during ParaMor's morphology induction stage does not preclude short types from being analyzed as containing multiple morphemes during ParaMor's segmentation stage. As section 2.2 describes, ParaMor's segmentation algorithm is independent of the set of types from which schemes and scheme clusters are built.

The string length that types must meet to join the induction vocabulary is a free parameter. ParaMor is designed to identify the productive inflectional paradigms of a language. Unless a paradigm is restricted to occur only with short stems, a possible but unusual scenario (as with the English adjectival comparative, c.f. *faster* but **exquisiter*) we can expect a productive paradigm to occur with a reasonable number of longer stems in a corpus. Hence, ParaMor needn't be overly concerned about discarding short types. A qualitative examination of Spanish data suggested discarding types five characters or less in length; we use this cutoff in all experiments described in this paper.

Excluding short types from the paradigm induction vocabulary virtually eliminates the entire category of chance scheme. In a random sample of 100 schemes that ParaMor selected when short types were excluded, only one scheme contained types related only by chance string similarity, down from 35 when short types were not excluded. Returning to Table 1, excluding types five characters or less in length bars ten of the twelve word types which support the erroneous 3000th selected scheme *Ø.a.-anar*. Among the excluded types are valid Spanish words such as *ganar 'to gain'*. But also eliminated are several meaningless acronyms such as the single letters *g* and *s*. Without these short types, ParaMor rightly cannot select the 3000th scheme.

2.2 Segmentation

An Agglutinative Model

With the improvement in scheme precision that results from the two extensions discussed in section 2.1, we are ready to propose a more realistic model of morphology. ParaMor's baseline segmentation algorithm distrusts ParaMor's induced scheme models. The baseline algorithm assumes each word form can contain at most a single morpheme boundary. If it detects more than one morpheme

boundary, then the baseline algorithm proposes a separate morphological analysis for each possible boundary. In contrast, our extended model of segmentation vests more trust in the induced schemes, assuming that scheme clusters which propose different morpheme boundaries are simply modeling different valid morpheme boundaries. And our extension proposes a single morphological analysis containing all hypothesized morpheme boundaries.

To detect morpheme boundaries, ParaMor matches each word, *w*, in the full vocabulary of a corpus against the clusters of schemes which are the final output of ParaMor's paradigm identification stage. When a suffix, *f*, of some scheme-cluster, *C*, matches a word-final string of *w*, i.e. $w = u.f$, ParaMor attempts to replace *f* in turn with each suffix f' of *C*. If the string $u.f'$ occurs in the full corpus vocabulary, then, on the basis of this paradigmatic evidence, ParaMor identifies a morpheme boundary in *w* between *u* and *f*.

For example, to detect morpheme boundaries in the Spanish word *apoyados* 'supports *(adjective masculine plural)*', ParaMor matches all word-final strings of *apoyados* against the candidate suffixes of ParaMor's induced scheme clusters. The word-final strings of *apoyados* are *s*, *os*, *dos*, *ados*, *yados*, …. The scheme clusters that our extended version of ParaMor induces include clusters which contain schemes very similar to the 1st, 2nd, and 5th baseline selected schemes, see Table 1. In particular, our extended ParaMor identifies separate scheme clusters that contain the candidate suffixes: *s* and *Ø*; *os* and *o*; and *ados* and *ado*. Substituting *Ø* for *s*, *o* for *os*, or *ado* for *ados* yields the Spanish string *apoyado* 'supports *(adjective masculine singular)*'. It so happens, that *apoyado* does occur in our Spanish corpus, and so ParaMor has found paradigmatic evidence for three morpheme boundaries. Crucially, our ParaMor extension from section 2.1 that removes schemes which hypothesize suffix internal morpheme boundaries correctly discards all schemes which contained the candidate suffix *dos*. Consequently, no scheme cluster exists to incorrectly suggest the morpheme boundary **apoya + dos*, as the 3rd baseline selected scheme would have. Where ParaMor's baseline segmentation algorithm would propose three separate analyses of *apoyados*, one for each detected morpheme boundary: *apoy +ados*, *apoyad +os*, and *apoyado +s*; our extended segmentation algorithm produces the single correct analysis: *apoy +ad +o +s*.

It is interesting to note that although each of ParaMor's individual paradigm models proposes a single morpheme boundary, our agglutinative segmentation model can recover multiple boundaries in a single word. Using this idea it may be possible to quickly adapt Linguistica for agglutinative languages. Instead of interpreting the sets of stems and affixes that Goldsmith's Linguistica algorithm produces as immediate segmentations of words, these signatures can be thought of as models of paradigms that may generalize to new words.

Augmenting ParaMor's Segmentations

With its focus on the paradigm, ParaMor specializes at analyzing inflectional morphology (Monson et al., 2007a). Morpho Challenge 2007 requires algorithms to analyze both inflectional and derivational morphology (Kurimo et al., 2007a; 2007b). To compete in the challenge, we combine ParaMor's morphological segmentations with segmentations from Morfessor (Creutz, 2006), an unsupervised morphology induction algorithm which learns both inflectional and derivational morphology. We incorporate the segmentations from Morfessor into the segmentations that the ParaMor system produces by straightforwardly adding the Morfessor segmentation for each word as an additional separate analysis to those ParaMor produces (Monson et al., 2007b). Morfessor has one free parameter, which we optimize separately for each language of Morpho Challenge 2007.

ParaMor also has several free parameters, including the type length parameter and the parameter over stem-final character entropy described in section 2.1. We do not adjust any of ParaMor's parameters from language to language, but fix them at values that produce reasonable Spanish paradigms and segmentations. As in Monson et al. (2007b), to avoid adjusting ParaMor's parameters we limit ParaMor's paradigm induction vocabulary to 50,000 frequent types for each language.

3 Evaluation

To evaluate our extensions to the ParaMor algorithm, we follow the methodology of the peer operated Morpho Challenge 2007. All segmentations produced by our extensions were sent to the Morpho Challenge Organizing Committee (Kurimo et al., 2008). The Organizing Committee evaluated our segmentations and returned the automatically

calculated quantitative results. Using the evaluation methodology of Morpho Challenge 2007 permits us to compare our algorithms against the unsupervised morphology induction systems which competed in the 2007 Challenge. Of the many algorithms for unsupervised morphology induction discussed with the related work in section 1.1, five participated in Morpho Challenge 2007. Unless an algorithm has been given an explicit name, morphology induction algorithms will be denoted in this paper by the name of their lead author. The five algorithms which participated in the 2007 Challenge are: Bernhard (2007), Bordag (2007), Zeman (2007), Creutz's (2006) Morfessor, and ParaMor (2007b).

Morpho Challenge 2007 had participating algorithms analyze words in four languages: English, German, Finnish, and Turkish. The Challenge evaluated each algorithm's morphological analyses in two ways. First, a linguistic evaluation measured each algorithm's precision, recall, and F_1 at morpheme identification against an answer key of morphologically analyzed word forms. Scores were normalized when a system proposed multiple analyses of a single word, as our combined ParaMor-Morfessor submissions do. For further details on the linguistic evaluation in Morpho Challenge 2007, see Kurimo et al. (2007a). The second evaluation of Morpho Challenge 2007 was a task based evaluation. Each algorithm's analyses were imbedded in an information retrieval (IR) system. The IR evaluation consisted of queries over a language specific collection of newswire articles. All word forms in all queries and all documents were replaced with the morphological decompositions of each individual analysis algorithm. Separate IR tasks were run for English, German, and Finnish, but not Turkish. For additional details on the IR evaluation of Morpho Challenge 2007 please reference Kurimo et al. (2007b).

Tables 2 and 3 present, respectively, the linguistic and IR evaluation results. In these two tables, the top two rows contain results for segmentations produced by versions of ParaMor that include our extensions. The topmost row in each table, labeled '+P +Seg', gives the results for our fully augmented version of ParaMor, which includes our two extensions designed to improve precision as well as our new segmentation model which can propose multiple morpheme boundaries in a single analysis of a word form. The second row of each table, labeled '+P –Seg', augments ParaMor only with the two enhancements designed to improve precision. The third row of each table gives the Challenge results for the ParaMor baseline algorithm. Rows four through seven of each table give scores from Morpho Challenge 2007 for the best performing unsupervised systems. If multiple versions of a single algorithm competed in the Challenge, the scores reported here are the highest F_1 or Average Precision score of any algorithm variant at a particular task. In all test scenarios but Finnish IR, we produced Morfessor segmentations to augment ParaMor that are independent of the Morfessor runs which competed in Morpho Challenge. If our Morfessor runs gave a higher F_1 or Average Precision, then we report this higher score. Finally, scores reported on rows eight and beyond are from reference algorithms that are not unsupervised. Reference algorithms appear in *italics*. A double line bisects both Table 2 and Table 3 horizontally. All results which appear above the double line were evaluated after the final deadline of Morpho Challenge 2007. In particular, ParaMor officially competed only in the English and German tracks of the Challenge.

The Linguistic Evaluation

Table 2 contains the results from the linguistic evaluation of Morpho Challenge. The Morpho Challenge Organizing Committee did not provide us with data on the statistical significance of the results for the enhanced versions of ParaMor. But most score differences are statistically significant—All F_1 differences of more than 0.5 between systems which officially competed in Morpho Challenge 2007 were statistically significant (Kurimo et al., 2007a).

In German, Finnish, and Turkish our fully enhanced version of ParaMor achieves a higher F_1 than any system that competed in Morpho Challenge 2007. In English, ParaMor's precision score drags F_1 under that of the first place system, Bernhard; In Finnish, the Bernhard system's F_1 is likely not statistically different from that of our system. Our final segmentation algorithm demonstrates consistent performance across all four languages. In Turkish, where the morpheme recall of other unsupervised systems is anomalously low, our algorithm achieves a recall in a range similar to its recall scores for the other languages. ParaMor's ultimate recall is double that of any other unsuper-

		English			German			Finnish			Turkish		
		P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
ParaMor & Morfessor	**+P +Seg**	50.6	63.3	56.3	49.5	**59.5**	54.1	49.8	**47.3**	**48.5**	51.9	**52.1**	**52.0**
	+P –Seg	56.2	60.9	58.5	57.4	53.5	**55.4**	60.5	33.9	43.5	62.0	38.2	47.3
	Baseline	41.6	**65.1**	50.7	51.5	55.6	53.4	55.0	35.6	43.2	53.2	41.6	46.7
Bernhard		61.6	60.0	**60.8**	49.1	57.4	52.9	59.7	40.4	**48.2**	73.7	14.8	24.7
Bordag		59.7	32.1	41.8	60.5	41.6	49.3	71.3	24.4	36.4	81.3	17.6	28.9
Morfessor		82.2	33.1	47.2	67.6	36.9	47.8	76.8	27.5	40.6	73.9	26.1	38.5
Zeman		53.0	42.1	46.9	52.8	28.5	37.0	58.8	20.9	30.9	65.8	18.8	29.2
Tepper		*69.2*	*52.6*	*59.8*	-	-	-	*62.0*	*46.2*	*53.0*	*70.3*	*43.0*	*53.3*

Table 2. Unsupervised morphology induction systems evaluated for precision (P), recall (R), and F_1 at morpheme identification using the methodology of the linguistic competition of Morpho Challenge 2007.

vised Turkish system, leading to an improvement in F_1 over the next best system, Morfessor alone, of 13.5% absolute or 22.0% relative.

In all four languages, as expected, the combination of removing short types from the training data, and the additional filtering of scheme clusters, '+P', significantly improves precision scores over the ParaMor baseline. Allowing multiple morpheme boundaries in a single word, '+Seg', increases the number of words ParaMor believes share a morpheme. Some of these new words do in fact share a morpheme, some, in reality do not. Hence, our extension of ParaMor to agglutinative sequences of morphemes increases recall but lowers precision across all four languages. The effect of agglutinative segmentations on F_1, however, differs with language. For the two languages which make limited use of suffix sequences, English and German, a model which hypothesizes multiple morpheme boundaries can only moderately increase recall and does not justify, by F_1, the many incorrect segmentations which result. On the other hand, an agglutinative model significantly improves recall for true agglutinative languages like Finnish and Turkish, more than compensating in F_1 for the drop in precision over these languages. But in all four languages, the agglutinative version of ParaMor outperforms the baseline unenhanced version at F_1.

The final row of Table 2 is the evaluation of a reference algorithm submitted by Tepper (2007). While not an unsupervised algorithm, Tepper's reference parallels ParaMor in augmenting segmentations produced by Morfessor. Where ParaMor augments Morfessor with special attention to inflectional morphology, Tepper augments Morfessor with hand crafted morphophonology rules that conflate multiple surface forms of the same underlying suffix. Like ParaMor, Tepper's algorithm significantly improves on Morfessor's recall. With two examples of successful system augmentation, we suggest that future research take a closer look at building on existing unsupervised morphology induction systems.

The IR Evaluation

Turn now to results from the IR evaluation in Table 3. Although ParaMor does not fair as well in Finnish, in German, the fully enhanced version of ParaMor places above the best system from the 2007 Challenge, Bernhard, while our score on English rivals this same best system. Morpho Challenge 2007 did not measure the statistical significance of uninterpolated average precision scores in the IR evaluation. It is not clear what feature of ParaMor's Finnish analyses causes comparatively low average precision. Perhaps it is simply that ParaMor attains a lower morpheme recall over Finnish than over English or German. And unfortunately, Morpho Challenge 2007 did not run IR experiments over the other agglutinative language in the competition, Turkish. When ParaMor does not combine multiple morpheme boundaries into a single analysis, as in the baseline and '+P –Seg' sce-

		Eng.	Ger.	Finn.	Tur.
ParaMor & Morfessor	**+P +Seg**	39.3	**48.4**	42.6	-
	+P –Seg	35.1	43.1	37.1	-
	Baseline	34.4	40.1	35.9	-
Bernhard		**39.4**	47.3	**49.2**	-
Bordag		34.0	43.1	43.1	-
Morfessor		38.8	46.0	44.1	-
Zeman		26.7*	25.7*	28.1*	-
Dummy		*31.2*	*32.3*	*32.7*	-
Oracle		*37.7*	*34.7*	*43.1*	-
Porter		*40.8*	-	-	-
Tepper		*37.3**	-	-	-

Table 3. Unsupervised morphology induction systems evaluated for uninterpolated average precision using the methodology of the IR competition of Morpho Challenge 2007. These results use Okapi term weighting (Kurimo et al., 2008b).

*Only a subset of the words which occurred in the IR evaluation of this language was analyzed by this system.

narios, average precision is comparatively poor. Where the linguistic evaluation did not always penalize a system for proposing multiple partial analyses, real NLP applications, such as IR, can.

The reference algorithms for the IR evaluation are: Dummy, no morphological analysis; Oracle, where all words in the queries and documents for which the linguistic answer key contains an entry are replaced with that answer; Porter, the standard English Porter stemmer; and Tepper described above. While the hand built Porter stemmer still outperforms the best unsupervised systems on English, these same best unsupervised systems outperform both the Dummy and Oracle references for all three evaluated languages—strong evidence that unsupervised induction algorithms are not only better than no morphological analysis, but that they are better than incomplete analysis as well.

4 Conclusions and Future Directions

Augmenting ParaMor with an agglutinative model of segmentation produces an unsupervised morphology induction system with consistent and strong performance at morpheme identification across all four languages of Morpho Challenge 2007. By first cleaning up the paradigm models that ParaMor learns, we raise ParaMor's segmentation precision and allow the agglutinative model to significantly improve ParaMor's morpheme recall.

Looking forward to future improvements, we examined by hand the final set of scheme clusters that the current version of ParaMor produces over our newswire corpus of 50,000 Spanish types. ParaMor's paradigm identification stage outputs 41 separate clusters. Among these final scheme clusters are those which model all major productive paradigms of Spanish. In fact, there are often multiple scheme clusters which model portions of the same true paradigm. As an extreme case, 12 separate scheme clusters contain suffixes from the Spanish *ar* verbal paradigm. Relaxing restrictions on ParaMor's clustering algorithm (Monson et al., 2007a) may address this paradigm fragmentation.

The second significant shortcoming which surfaces among ParaMor's 41 final scheme clusters is that ParaMor currently does not address morphophonology. Among the final scheme clusters, 12 attempt to model morphophonological change by incorporating the phonological change either into the stems or into the suffixes of the scheme cluster. But ParaMor currently has no mechanism for detecting when a cluster is modeling morphophonology. Perhaps ideas on morphophonology from Goldsmith (2006) could be adapted to work with the ParaMor algorithm. Finally, we plan to look at scaling the size of the vocabulary used both during paradigm induction and during morpheme segmentation. We are particularly interested in the possibility that ParaMor may be able to identify paradigms from much less data than 50,000 types.

Acknowledgements

We kindly thank Mikko Kurimo, Ville Turunen, Matti Varjokallio, and the full Organizing Committee of Morpho Challenge 2007, for running the evaluations of ParaMor. These dedicated workers produced impressively fast turn around for evaluations on sometimes rather short notice.

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.

Bordag, Stefan. Unsupervised and Knowledge-free Morpheme Segmentation and 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.

Demberg, Vera, Helmut Schmid, and Gregor Möhler. Phonological Constraints and Morphological Preprocessing for Grapheme-to-Phoneme Conversion. *Association for Computational Linguistics*. Prague, Czech Republic, 2007.

Dryer, Matthew S. Prefixing vs. Suffixing in Inflectional Morphology. In *The World Atlas of Language Structures*. Eds. Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. 2005.

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.

Goldsmith, John, and Yu Hu. From Signatures to Finite State Automata. Paper presented at the Midwest Computational Linguistics Colloquium. Bloomington, Indiana, 2004.

Goldwater, Sharon, and David McClosky. Improving Statistic MT through Morphological Analysis. *Empirical Methods in Natural Language Processing*. Vancouver, Canada, 2005.

Hafer, Margaret A. and Stephen F. Weiss. Word Segmentation by Letter Successor Varieties. *Information Storage and Retrieval*, 10:371-385. 1974.

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

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

Johnson, Howard, and Joel Martin. Unsupervised Learning of Morphology for English and Inuktitut. *Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics*. Edmonton, Canada, 2003.

Kurimo, Mikko, Mathias Creutz, and Matti Varjokallio. Unsupervised Morpheme Analysis Evaluation by a Comparison to a Linguistic Gold Standard – Morpho Challenge 2007. *Working Notes for the CLEF 2007 Workshop*. Budapest, Hungary, 2007a.

Kurimo, Mikko, Mathias Creutz, and Ville Turunen. Unsupervised Morpheme Analysis Evaluation by IR Experiments – Morpho Challenge 2007. *Working Notes for the CLEF 2007 Workshop*. Budapest, Hungary, 2007b.

Kurimo, Mikko, Mathias Creutz, and Matti Varjokallio. Unsupervised Morpheme Analysis -- Morpho Challenge 2007. January 10, 2008. <http://www.cis.hut.-fi/morphochallenge2007/>. 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. *Working Notes for the CLEF 2007 Workshop*. Budapest, Hungary, 2007b.

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.

Tepper, Michael A. Using Hand-Written Rewrite Rules to Induce Underlying Morphology. *Working Notes for the CLEF 2007 Workshop*. Budapest, Hungary, 2007.

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

Zeman, Daniel. Unsupervised Acquiring of Morphological Paradigms from Tokenized Text. *Working Notes for the CLEF 2007 Workshop*. Budapest, Hungary, 2007.