ParaMor: Finding Paradigms across Morphology

Christian Monson, Jaime Carbonell, Alon Lavie, Lori Levin
Language Technologies Institute
Carnegie Mellon University
{cmonson, jgc+, alavie+, lsl+}@cs.cmu.ed

Abstract

Our algorithm, ParaMor, fared well in Morpho Challenge 2007 (Kurimo et al., 2007), a peer operated competition pitting against one another algorithms designed to discover the morphological structure of natural languages from nothing more than raw text. ParaMor constructs sets of affixes closely mimicking the paradigms of a language, and, with these structures in hand, annotates word forms with morpheme boundaries. Of the four language tracks in Morpho Challenge 2007, we entered ParaMor in English and German. Morpho Challenge 2007 evaluated systems on their precision, recall, and balanced F_1 at identifying morphological processes, whether those processes mark derivational morphology or inflectional features. In English, ParaMor's balanced precision and recall outperform at F_1 an already sophisticated baseline induction algorithm, Morfessor (Creutz, 2006). ParaMor placed fourth in English overall. In German, ParaMor suffers from a low morpheme recall. But combining ParaMor's analyses with analyses from Morfessor results in a set of analyses that outperform either algorithm alone, and that place first in F_1 among all algorithms submitted to Morpho Challenge 2007.

Categories and Subject Descriptors

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

Keywords

Unsupervised Natural Language Morphology Induction, Paradigms

1 Introduction

Performance at natural language processing tasks as different as speech recognition (Creutz, 2006) and machine translation (Goldwater and McClosky, 2005) can improve with careful morphological analysis. But building a morphological analyzer for a natural language requires expert language knowledge that may be in short supply. In this paper we describe ParaMor, an algorithm that automates the construction of a morphology analysis system for any language; and we present and discuss ParaMor's performance in Morpho Challenge 2007 (Kurimo et al., 2007), a competition for algorithms that induce the morphology of natural languages from nothing more than unannotated text.

1.1 Paradigms: The Structure of Natural Language Morphology

Both traditional and modern theories of inflectional morphology (Stump, 2001) organize natural language morphology by paradigms. Where a paradigm is the set of surface forms a lexeme can take as it inflects for relevant morphosyntactic features. Following suit, our work on unsupervised morphology induction also recognizes the paradigm as the natural organizational structure of inflectional morphology.

One of the properties of paradigms we exploit in our work is that of the mutual exclusion of affixes. Consider Spanish verbs. Each verbal lexeme in Spanish can take upwards of 35 surface forms. Most of the surface forms of a Spanish verb mark tense or mood in combination with person and number, but here we focus on the relatively few non-finite forms of Spanish verbs. A Spanish verb can appear in exactly one of three non-finite forms: as a past participle, as a present participle, or in the infinitive. If the verb occurs as a past participle, then the verb takes additional suffixes. First, an obligatory suffix marks gender, an *a* marks feminine, an *o* masculine. Following the gender suffix either a plural suffix, *s*, appears or else there is no suffix at all. The lack of an explicit plural

Form	Gender	Number
Past Participle	Feminine	Singular
	Masculine	Plural
Present Participle		
Infinitive		

Form	Gender	Number
ad	a	Ø
	o	s
ando		
ar		

Figure 1: Left: A fragment of the Spanish verbal paradigm. There are three morphosyntactic categories covered in this paradigm fragment: first, form; second, gender; and third, number. Each of these three categories appear in separate columns. And features within one feature column are mutually exclusive. Right: The suffixes filling the cells of the Spanish verbal paradigm fragment for the inflection class of ar verbs.

suffix marks singular. The values of each individual morphosyntactic feature (form, gender, and number) are mutually exclusive. The Spanish lexeme *administrar*, given here in the infinitive, translates as *to administer or manage*. The feminine plural past participle of *administrar* is *administradas* which can refer to a group of women under administration, an in *the managed help*. There is no way for *administrar* or any other Spanish lexeme to appear simultaneously in the infinitive and in a past participle form simultaneously: **administrardas, *administradasar*. Figure 1 sketches the paradigm schema of Spanish nonfinite verb forms. In the left-hand table the feature values for the form, gender, and number features are given, while the right-hand table presents the surface forms of the suffixes realizing the corresponding feature values for verbs belonging to the class of regular Spanish *ar* verbs.

Our unsupervised morphology induction algorithm exploits the mutual exclusivity of feature-valued paradigms in two phases. ParaMor's first phase identifies sets of mutually exclusive strings which mimic paradigms. ParaMor's second phase segments word forms into morpheme-like pieces suggested by the discovered paradigms. Currently, ParaMor can isolate word final suffixes. ParaMor's methods can be straightforwardly generalized to prefixes and forthcoming work models sequences of concatenative morphemes.

1.2 Related Work

In this section we highlight previously proposed minimally supervised approaches to the induction of morphology that, like ParaMor, draw on the unique structure of natural language morphology. One facet of NL morphological structure commonly leveraged by morphology induction algorithms is that morphemes are recurrent building blocks of words. Brent et al. (1995), Goldsmith (2001), and Creutz (2006) emphasize the building block nature of morphemes when they each use recurring word segments to efficiently encode a corpus. These approaches then hypothesize that those recurring segments which most efficiently encode a corpus are likely morphemes. Another technique that exploits morphemes as repeating sub-word segments encodes the lexemes of a corpus as a character tree, i.e. trie, (Harris, 1955; Hafer and Weis, 1974; Demberg, 2007), or as a finite state automaton (FSA) over characters (Johnson, H. and Martin, 2003; Altun and M. Johnson, 2001). A trie or FSA conflates multiple instances of a morpheme into a single sequence of states. The paradigm structure of NL morphology has also been previously leveraged. Goldsmith (2001) uses morphemes to efficiently encode a corpus, but he first groups morphemes into paradigm like structures he calls signatures. To date, the work that draws the most on paradigm structure is Snover (2002). Snover incorporates paradigm structure into a generative statistical model of morphology.

2 ParaMor

We present our unsupervised morphology induction algorithm, ParaMor, by following an extended example of the analysis of the Spanish word *administradas* (*administered*). The word *administradas* occurs in the corpus of Spanish newswire on which we developed the ParaMor algorithm. This Spanish newswire corpus contains 50,000 types. We hope the detailed example we give here can flesh out the abstract step-by-step description of ParaMor in Monson et al. (2007).

Before delving into ParaMor's details we note two facts which guided algorithm design. First, in any given corpus, a particular lexeme will likely not occur in all possible inflected forms. But rather each lexeme will occur in some subset of its possible surface forms. Second, we expect inflected forms of a single lexeme to be corre-

lated. That is, if we have observed several lexemes in inflected form A, and if B belongs to the same paradigm as A, then we can expect a significant fraction of those lexemes inflected as A to also occur in an inflected form with B.

2.1 A Search for Partial Paradigms

ParaMor begins with a search for partial paradigms, where a partial paradigm is a set of candidate suffixes, and a candidate suffix is any word final substring. The word *administradas* gives rise to many candidate suffixes including: *stradas*, *tradas*, *radas*, *adas*, *das*, *as*, *s*, and Ø. Referring again to Figure 1, the candidate suffix *s* is a true morpheme of Spanish, marking plural. Additionally, the candidate suffixes *as* and *adas*, cleanly contain more than one suffix: The left edges of the word-final strings *as* and *adas* occur at Spanish morpheme boundaries. All other candidate suffixes derived from *administradas* incorrectly segment the word. The candidate suffixes *radas*, *tradas*, *stradas*, etc. erroneously include part of the stem, while *das*, in our analysis, places a morpheme boundary internal to the past participle morpheme *ad*. Of course, while we can discuss which candidate suffixes are reasonable and which are not, an unsupervised morphology induction system has no a priori knowledge of Spanish morphology. ParaMor does not know what strings are valid Spanish morphemes, is ParaMor aware of the feature value meanings associated with morphemes.

Each candidate suffix may be derived from multiple word forms. The candidate suffix *stradas* occurs as the final substring of eight wordforms in our Spanish corpus, including the words *administradas*, *arrastradas (wretched)* and mostradas *(accustomed)*. The candidate suffix *s* is a word final string of 10,662 wordforms in this same corpus, more than one fifth of the unique wordforms! When a candidate suffix is stripped from a surface word, we call the remaining word initial string a candidate stem. The (incorrect) candidate suffix *stradas* gives rise to eight (incorrect) candidate stems including *admini*, *arra*, and *mo*.

ParaMor's initial search for partial paradigms considers every candidate suffix derived from any word form in the input corpus as potentially part of a true inflectional paradigm. ParaMor's search considers each non-null candidate suffix in turn, beginning with that candidate suffix which can attach to the most candidate stems, working toward suffixes which can attach to fewer stems. For each particular candidate suffix, f, ParaMor notes the candidate stems, T, to which f can attach, and then identifies the candidate suffix, f', that forms separate corpus words with the largest number of stems in T. The candidate suffix f' is then added to the partial paradigm anchored by f. In our examples, all eight of the candidate stems that take *stradas* also form corpus words with the candidate suffix *strada* (words such as *administrada*, *arrastrada*, and *mostrada*) and hence *strada* would be added to the partial paradigm begun from *stradas*; similarly, the candidate suffix which can attach to the largest fraction of the 10,662 candidate stems which have a word final *s* is *Ø*, at 5501.

Now with a partial paradigm containing two candidate suffixes, ParaMor resets T to be the set of candidate stems which form corpus words with both f and f'. ParaMor then searches for a third suffix which can form words with a large subset of this new T. ParaMor continues to add candidate suffixes until one of two halting criteria is met:

1. Since we expect suffixes from a single paradigm to be correlated, ParaMor stops growing a partial paradigm if no candidate suffix can form corpus words with at least a threshold fraction of the stems in the current partial paradigm.
2. ParaMor stops adding candidate suffixes if the stem evidence for the partial paradigm is too meager—ParaMor will only add a suffix to a partial paradigm if there are more stems than there are suffixes in the proposed partial paradigm.

Figure 2 contains a number of search paths that ParaMor followed when analyzing our Spanish corpus. Most of the paths in Figure 2 are directly relevant to the analysis of *administradas*. Search paths begin at the bottom of Figure 2 and proceed upwards one candidate suffix at a time. In Spanish, the non-null candidate suffix that can attach to the most stems is *s*. The search path begun from *s* is the right-most search path shown in Figure 2. As discussed above, the null suffix, *Ø*, can attach to the largest number of candidate stems to which *s* can attach, and so the first search step adds *Ø* to the candidate suffix *s*. ParaMor then identifies the candidate suffix *r* as the suffix which can attach to the most stems to which *s* and *Ø* can both attach. But *r* can only form corpus words in combination with 287 or 5.2% of the 5501 stems to which *s* and *Ø* can attach. As such a severe drop in stem count does not convincingly suggest that the candidate suffix *r* is correlated with the candidates *s* and *Ø*, ParaMor does not add *r*, or any other suffix, to the now closed partial paradigm *s.Ø*. Experimentally we determined that, for Spanish, requiring at least 25% of stems to carry over when adding a candidate suffix seems to discover reasonable partial paradigms.

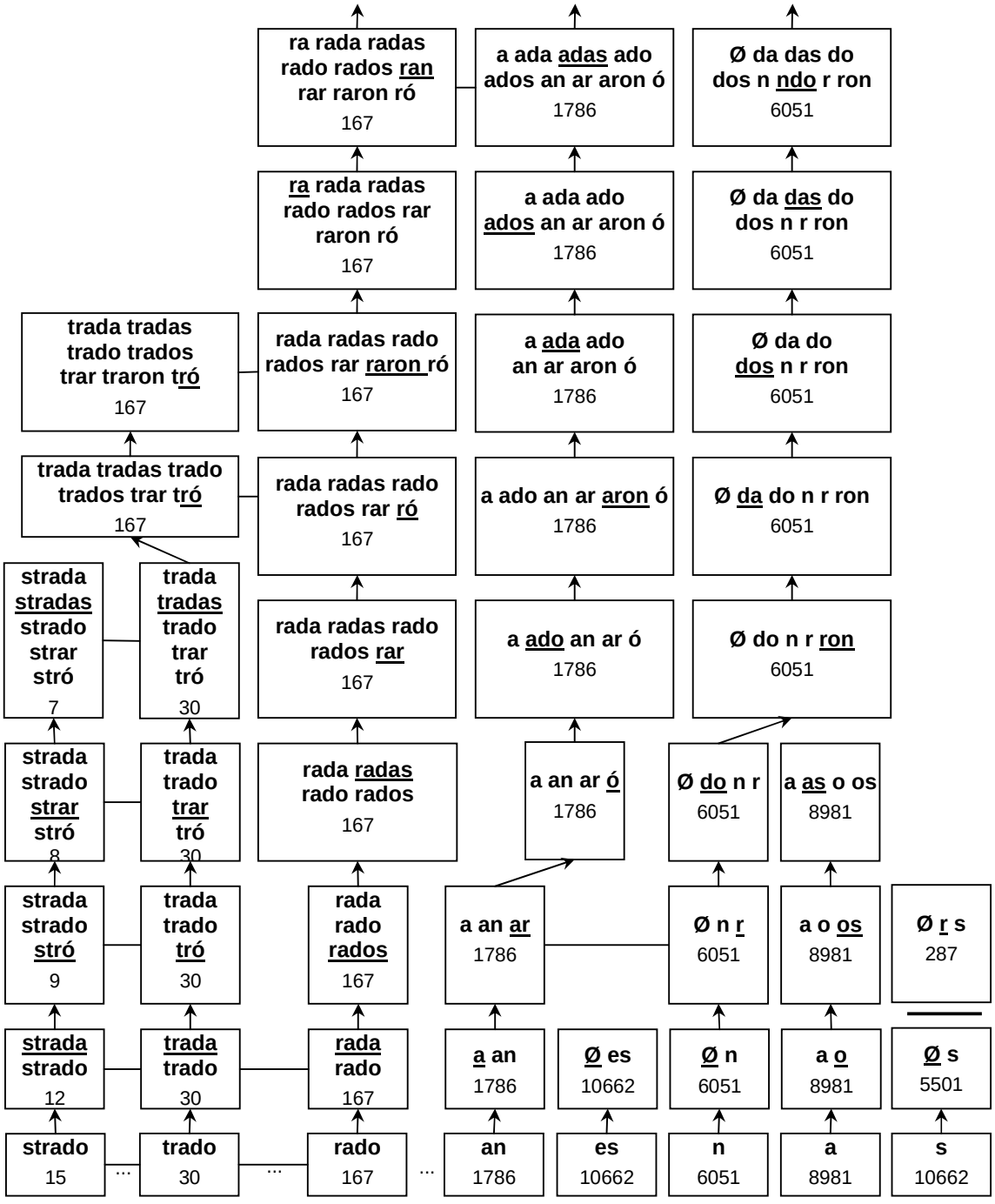

Figure 2: Eight search paths that ParaMor follows in search of likely partial paradigms. Search paths begin at the bottom of the figure and move upward. Candidate suffixes appear in **bold**. The underlined candidate suffix in each partial paradigm is the suffix added by the most recent search step. Each partial paradigm gives the number of candidate stems which attach to all candidate suffixes in that partial paradigm. Horizontal links between partial paradigms connect sets of suffixes that differ only in their initial character.

Continuing leftward from the *s*-anchored partial paradigm in Figure 2, ParaMor follows search paths from the candidate suffixes *a*, *n*, *es*, and *an* in turn. The 77th candidate suffix from which ParaMor grows a partial paradigm is *rado*. The search path from *rado* is the first path to build a partial paradigm that includes the candidate suffix *radas*, relevant for *administradas*. Similarly, search paths from *trado* and *strado* lead to partial paradigms which include the candidate suffixes *tradas* and *stradas* respectively. The search path from *strado* illustrates the second stopping criterion. From *strado* four candidate suffixes are added one at a time: *strada*, *stró*, *strar*, and *stradas*. Only seven candidate stems form words when combined singly with all five of these candidate suffixes. Adding any additional candidate suffix to these five suffixes brings the stem count down at least to six. Since six stems is not more than the six suffixes which would be in the resulting partial paradigm, ParaMor does not add a sixth candidate suffix.

In our corpus of Spanish newswire text, ParaMor's initial search identifies partial paradigms containing 92% of all ideal inflectional suffixes of Spanish, or 98% of the ideal suffixes that occurred at least twice in the corpus. Among the selected partial paradigms are those which contain portions of all nine true paradigms for our analysis of Spanish. The high recall of the initial search comes, of course, at the expense of precision. While our analysis provides nine true paradigms and 87 unique suffixes, 8339 partial paradigms are constructed containing 9889 unique candidate suffixes. The constructed partial paradigms have at least three readily apparent flaws. First, the candidate suffixes of many partial paradigms overlap. At the end of the initial search, there are 27 distinct partial paradigms that contain the reasonable candidate suffix *adas*. Each of these 27 partial paradigms geminates from a distinct initial candidate suffix: *an*, *en*, *ación*, *amos*, etc. Second flaw, most constructed partial paradigms contain many fewer candidate suffixes than do the true paradigms of Spanish. And third, many partial paradigms include candidate suffixes possessed of an incorrect morpheme boundary. ParaMor addresses the first two flaws by merging together similar partial paradigms. And ParaMor addresses the third flaw while further ameliorating the second through filters which weed out less likely paradigm clusters.

2.2 Merging Partial Paradigms

To merge partial paradigms ParaMor adapts greedy hierarchical agglomerative clustering. The details of the specific clustering algorithm appear in Monson et al. (2007). Here we continue our Spanish example to illustrate how partial paradigms are merged. Figure 3 contains a small portion of the partial paradigm cluster that consumes the partial paradigm built from the candidate suffix *an*. The first eight steps of the partial paradigm search path from *an* appear in Figure 2. But the search path continues until there are fifteen candidate suffixes in the partial paradigm: *a*, *aba*, *aban*, *ada*, *adas*, *ado*, *ados*, *an*, *ando*, *ar*, *aron*, *arse*, *ará*, *arán*, and *ó*. The partial paradigm built from *an* appears on the center right of Figure 3. During clustering, *an*'s partial paradigm is merged with a cluster that has previously formed from a merger of two partial paradigms. These two partial paradigms and their merged cluster appear at the bottom left of Figure 3. ParaMor decides which partial paradigm clusters to merge by computing a similarity score between pairs of paradigm clusters. A variety of similarity metrics on partial paradigms are possible. Looking at Figure 3, it is clear that both the candidate suffix sets and the candidate stem sets of partial paradigms can overlap. Consequently partial paradigms can share covered surface types. For example, the bottom two clusters of Figure 3 both contain the candidate suffix *a* and the candidate stem *anunci*, reconcatenating this stem and suffix we say that both of these partial paradigms cover the boundary annotated word form *anunci+a*. ParaMor computes the similarity of partial paradigms, and their clusters, by comparing just such sets of morpheme boundary annotated word forms. We have found that the particular similarity metric used does not significantly affect clustering. For the experiments we report here we use the cosine similarity for sets, given as $|X \cap Y| / (|X||Y|)^{1/2}$. It is interesting to note that similarity scores do not monotonically decrease moving up the tree structure of a particular cluster. Non-decreasing similarities is a consequence of computing similarities over sets of objects which are merged up the tree. Returning to our Spanish example word *administradas*, Clustering reduces, from 27 to 6, the number of distinct partial paradigms in which the candidate suffix *adas* occurs. Clustering also reduces the total number of separate partial paradigms to 7511 from 8339.

2.3 Filtering Partial Paradigm Clusters

With the fragmentation of partial paradigms significantly reduced, ParaMor focuses on removing erroneously proposed partial paradigm clusters. After clustering we would expect that most sound clusters cover a reasonably large number of word forms of the corpus. So ParaMor's first filtration step simply removes all partial paradigms which do not cover at least a threshold number of word forms. Monson et al. (2007) discusses our empirical procedure to identify a reasonable threshold. ParaMor currently discards all partial paradigms which do not cover at least 37 word forms. This first filter drastically reduces the number of selected partial paradigms, from 7511 to

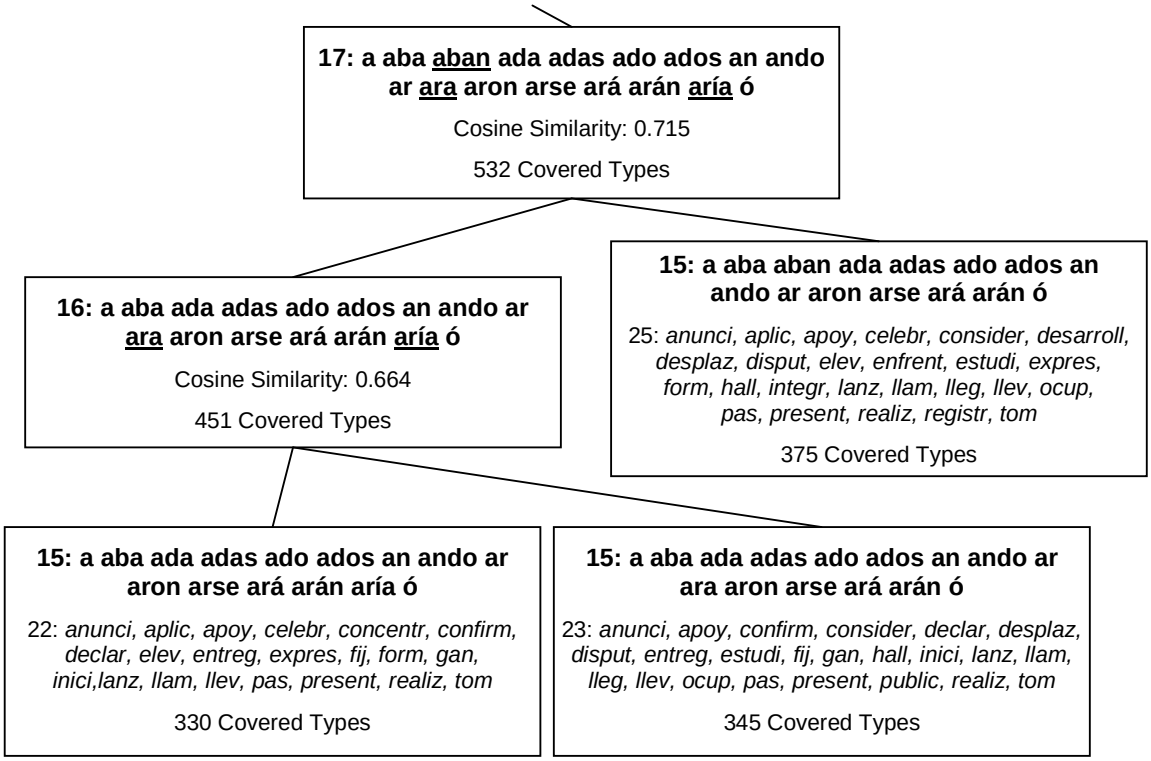

Figure 3: A portion of a cluster of partial paradigms. The candidate suffixes of each partial paradigm or cluster node appear in **bold**, candidate stems are in *italics*. Suffixes in cluster nodes which uniquely originate in one child are underlined. Also noted is the number of types covered by each partial paradigm or cluster node.

137. Among the many discarded partial paradigms is one of the six remaining partial paradigms containing *adas*. Although *adas* can be a valid verbal suffix sequence, the discarded partial paradigm was built from forms including *gradas* (*stairs*) and *hadas* (*fairies*), both nouns. Also removed are all partial paradigms containing the incorrect candidate suffix *stradas*—pseudo paradigms such as the partial paradigm built up from the candidate suffix *strado* presented at the far left of Figure 2.

Of the 137 remaining partial paradigm clusters, more than a third clearly attempt to model a morpheme boundary to the left of a correct morpheme boundary. Among these left-leaning clusters are those containing the candidate suffixes *tradas* and *radas*, including clusters which subsume the partial paradigms built from the candidate suffixes *trado* and *rado* given in Figure 2. To filter out left leaning clusters ParaMor implements a strategy inspired by Harris (1955). In a partial paradigm modeling a legitimate morpheme boundary, the candidate stems will likely take a wide variety of final characters, while, in reflection, the candidate suffixes will likely begin with a variety of characters. Conversely, in a partial paradigm attempting to place a morpheme boundary internal to a morpheme, the candidate stems will mostly end with the same character and the candidate suffixes will mostly begin with the same character. We apply this logic to build a filter that discards partial paradigm clusters with an obviously better morpheme boundary to the right of that proposed by the cluster. Specifically, ParaMor examines the suffixes in each cluster. If all the suffixes begin with the same character, then ParaMor recursively inspects the partial paradigms that would result from stripping off that initial character from all the suffixes in each partial paradigm that that cluster is built from. If more than half of the cluster's base partial paradigms identify a likely morpheme boundary to the right, then that cluster is entirely removed.

For example, consider the only cluster among the remaining 137 that contains the candidate suffix *tradas*. One of the partial paradigms this cluster is built from is that partial paradigm given in Figure 2 which geminates from the candidate stem *trados*, namely *trada.tradas.trado.trados.trar.traron.tró*. In Figure 2, this *tradas*-containing partial paradigm is linked to the right with the partial paradigm *rada.radas.rado.rados.rar.raron.ró*—obtaind by removing the initial *t* from each candidate suffix. Although not pictured in Figure 2, the partial paradigm containing *radas* is further connected to the partial paradigm *ada.adas.ado.ados.ar.aron.ó* through removal of

the initial *r*. And the stems of this *adas*-containing partial paradigm end in a wide variety of characters, suggesting a morpheme boundary. We measure stem final character variety using entropy. If stem final character entropy falls above a threshold value then ParaMor takes that partial paradigm as modeling a morpheme boundary. We have found that even a conservative, low, entropy cutoff discards nearly all clusters which model a morpheme boundary too far to the left. Applying this filter leaves 80 clusters, and furthermore completely removes all clusters containg the candidate suffixes *tradas* and/or *radas*. ParaMor currently contains no method for discarding clusters which place a morpheme boundary to the right of the correct position.

2.4 Segmentation

Finally, with a strong grasp on the paradigm structure, ParaMor straightforwardly segments the words of a corpus into morphemes. ParaMor's current segmentation algorithm is perhaps the most simple paradigm inspired segmentation algorithm possible. Essentially, ParaMor strips off suffixes which likely participate in a paradigm. To segment any word, w, ParaMor identifies all partial paradigm clusters that contain a non-empty suffix that matches a word final string of w. For each such matching suffix, $f \in C$, where C is the cluster containing f, we strip f from w obtaining a stem t. If there is some second suffix $f' \in C$ such that $t.f'$ is a word form found in either the training or the test corpus, then ParaMor proposes a segmentation of w between t and f. ParaMor, here, identifies f and f' as mutually exclusive suffixes from the same paradigm. If ParaMor finds no complex analysis, then we propose w itself as the sole analysis of the word. Note that for each word form, ParaMor may propose multiple separate segmentation analyses each containing a single proposed stem and suffix.

Let us finish out our extended example of the analysis of the word *administradas*. Among the 80 paradigm clusters that ParaMor accepts are clusters containing the candidate suffixes *adas*, *das*, *as*, and *s*. Of these, *adas*, *as*, and *s* identify correct morpheme boundaries, while *das* does not. The clusters containing candidate suffix *das* cannot be removed with either the size or the currently implemented morpheme boundary filters. Among the clusters which contain *adas* several also contain *ada*; similarly *das* and *da*, *as* and *a*, and *s* and *Ø*, each appear together in at least one cluster. Replacing, in *administradas*, *adas* with *ada*, *das* with *da*, *as* with *a*, or *s* with *Ø* results in the potential word form *administrada*. As *administrada* does occur in our Spanish corpus, ParaMor produces four separate analyses of the word *administradas*: *administr +adas*, *administra +das*, *administrad +as*, and *administrada +s*. Each of these four analyses appears as is in the file of analyzed words ParaMor produces.

3 Morpho Challenge 2007 Results and Conclusions

We entered ParaMor in the English and the German tracks of Morpho Challenge 2007. In each track we submitted three systems. The first system we submitted was ParaMor alone. ParaMor's algorithm has free parameters. We did not vary these parameters, but held each at a setting which produced reasonable *Spanish* suffix sets (Monson et al., 2007). The English and German corpora used in Morpho Challenge 2007 were larger than we had previously worked with. The English corpus contains nearly 385,000 types, while the German corpus contains more than 1.26 million types. ParaMor induced paradigmatic scheme-clusters over these larger corpora from just the top 50,000 most frequent types. But with the scheme-clusters in hand, ParaMor segmented all the types in each corpus.

The second submitted system combines the analyses of ParaMor with the analyses of Morfessor (Creutz, 2006). We downloaded Morfessor Categories-MAP 0.9.2 (Creutz, 2007) and optimized Morfessor's single parameter separately for English and for German. We optimized Morfessor's parameter against an F_1 score calculated following the methodology of Morpho Challenge 2007. The Morpho Challenge F_1 score is found by comparing Morfessor's morphological analyses to analyses in human-built answer keys. The official Morpho Challenge 2007 answer keys were not made available to the challenge participants. However, the official keys for English and German were created using the Celex database (Burnage, 1990), and Celex was available to us. Using Celex we created our own morphological answer keys for English and German that, while likely not identical to the official gold standards, are quite similar. Optimizing Morfessor's parameter renders the analyses we obtained from Morfessor no longer fully unsupervised. In the submitted combined system, we pooled Morfessor's analyses with ParaMor's in perhaps the most simple fashion possible: for each analyzed word we added Morfessor's analysis as an additional, comma separated, analysis to the list of analyses ParaMor identified. Naively combining the analyses of two systems in this way increases the total number of morphemes in each word's analyses—likely lowering precision but possibly increasing recall.

Submitted Systems	English			German		
	P	**R**	**F_1**	**P**	**R**	**F_1**
ParaMor & Morfessor	41.6	**65.1**	50.7	51.5	55.6	**53.2**
ParaMor	48.5	53.0	50.6	59.1	32.8	42.2
Morfessor Trained by Monson et al.	**77.2**	34.0	47.2	**67.2**	36.8	47.6
Bernhard-2	61.6	60.0	**60.8**	49.1	**57.4**	52.9
Bernhard-1	72.1	52.5	60.7	63.2	37.7	47.2
Pitler	74.7	40.6	52.3	N/A	N/A	N/A
Bordag-5a	59.7	32.1	41.8	60.5	41.6	49.3
Zeman	53.0	42.1	46.9	52.8	28.5	37.0

Table 1: The official Precision, Recall, and F1 scores from Morpho Challenge 2007, to three significant digits. Only scores for submitted systems most relevant to a discussion of ParaMor are included.

The third set of analyses we submitted to Morpho Challenge 2007 is the set Morfessor produced alone at the same optimized parameter settings used in our combined entry.

Table 1 contains the official Morpho Challenge 2007 results for top placing systems in English and German. Measuring by F_1, the clear winners on English are the two systems submitted by Bernhard. The ParaMor systems take fourth and fifth place. As expected, combining ParaMor's and Morfessor's analyses boosts recall over each individual system, but hurts English precision, negligibly increasing F_1 over ParaMor alone. ParaMor's more balanced precision and recall outperform the baseline Morfessor system with its precision centric analyses.

In German, the combined ParaMor-Morfessor system achieved the highest F_1 of any submitted system. Bernhard is a close second just 0.3 absolute lower—a likely statistically insignificant difference. As with English, Morfessor alone scores well on precision; in contrast, ParaMor's precision is significantly higher for German than in English. Combining two reasonable precision scores keeps the overall precision respectable. Both ParaMor and Morfessor alone have relatively low recall. But the combined system significantly improves recall over either system alone. Clearly ParaMor and Morfessor are complementary systems, identifying very different types of morphemes.

Indeed, Morfessor is particularly designed to identify agglutinative sequences of morphemes, while ParaMor focuses on identifying productive paradigms of usually inflectional suffixes. To gauge ParaMor's performance at its likely strength of inflectional morphology, we again used the Celex database to create morphological answer

	English				German			
	P	**R**	**F_1**	**σ**	**P**	**R**	**F_1**	**σ**
Morfessor	53.3	47.0	49.9	1.3	38.7	44.2	41.2	0.8
ParaMor	33.0	**81.4**	47.0	0.9	42.8	**68.6**	52.7	0.8

Table 2: ParaMor segmentations compared to Morfessor's evaluated for **P**recision, **R**ecall, **F_1**, and standard deviation of F_1, **σ**, against an answer key analyzed only for inflectional morphology.

keys, this time analyzed only for inflectional morphology. Table 2 contains the results of ParaMor and Morfessor against these new inflectional answer keys. ParaMor attains remarkably high recall of inflectional morphological processes for both German and particularly English. Also notable, ParaMor's precision is considerably lower measured against inflection only, as compared to measuring against both inflectional and derivational morphology. ParaMor is most likely identifying regular derivational processes in addition to a large fraction of the inflectional morphology.

We are excited by ParaMor's strong performance and are eager to extend our algorithm. Recent experiments suggest that the precision of ParaMor's segmentations can be improved by building partial paradigms from cleaner language data. Perhaps ParaMor and Morfessor's vastly different strategies for morphology induction can be combined in an even more fruitful fashion. We also intend to extend ParaMor to analyze sequences of affixes by combining separate analyses.

Acknowledgements

The research reported in this paper was funded in part by NSF grant number IIS-0121631.

References

Altun, Yasemin, and Mark Johnson. "Inducing SFA with ϵ-Transitions Using Minimum Description Length." *Finite State Methods in Natural Language Processing Workshop at ESSLLI*. Helsinki, Finland, 2001.

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.

Burnage, Gavin. *Celex—A Guide for Users*. Springer, Centre for Lexical information, Nijmegen, the Netherlands, 1990.

Creutz, Mathias. "Morpho project." May 31, 2007. <http://www.cis.hut.fi/projects/morpho/>

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

Demberg, Vera. "A Language-Independent Unsupervised Model for Morphological Segmentation." *Association for Computational Linguistics*. Prague, Czech Republic, 2007.

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

Goldwater, Sharon, and David McClosky. "Improving Statistical MT through Morphological Analysis." *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, B.C., Canada, 2005.

Hafer, Margaret A., and Stephen F. Weiss. "Word Segmentation by Letter Successor Varieties." *Information Storage and Retrieval* 10.11/12 (1974): 371-385.

Harris, Zellig. "From Phoneme to Morpheme." *Language* 31.2 (1955): 190-222. 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 – Morpho Challenge 2007." March 26, 2007. <http://www.cis.hut.fi/morphochallenge2007/>

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, 2007.

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

Stump, Gregory T. *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge University Press. 2001.