ParaMor: Finding Paradigms across Morphology

Christian Monson, Jaime Carbonell, Alon Lavie, Lori Levin
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA, USA 15213
{cmonson, alavie+, jgc+, lsl+}@cs.cmu.edu

Abstract

ParaMor, our unsupervised morphology induction algorithm placed well in Morpho Challenge 2007 (Kurimo et al., 2007). Morpho Challenge is a peer operated competition pitting against one another algorithms designed to discover the morphological structure of natural languages from nothing more than raw text. 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 third in English overall, behind two algorithms both submitted by Delphine Bernhard. 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.

1 Introduction

Words in natural language (NL) have internal structure. Morphological processes derive new lexemes from old ones or inflect the surface form of lexemes to mark morphosyntactic features such as tense, number, person, etc. This paper address minimally supervised induction of productive natural language morphology from text. Minimally supervised induction of morphology interests us both for practical and theoretical reasons. In linguistic theory, the morpheme is often defined as the smallest unit of language which conveys meaning. And yet, without annotating for meaning, recent work on minimally supervised morphology induction from written corpora has met with some success (Creutz, 2006). We are curious how far this program can be pushed. From a practical perspective, minimally supervised morphology induction would help create morphological analysis systems for languages outside the traditional scope of NLP. However, to develop our method we induce the morphological structure of three well-understood languages, English, German, and Spanish.

1.1 Inherent Structure in NL Morphology

The approach we have taken to induce morphological structure has explicit roots in linguistic theory. Cross-linguistically, natural language organizes inflectional morphology into *paradigms* and *inflection classes*. A paradigm is a set of mutually exclusive operations that can be performed on a word form. Each mutually exclusive morphological operation in a paradigm marks a lexeme for some set or *cell* of morphosyntactic features. An inflection class, meanwhile, specifies the procedural details that a particular set of adherent lexemes follow to realize the surface form filling each paradigm cell. Each lexeme in a language adheres to a single inflection class for each paradigm the lexeme realizes.

Paradigm cells are mutually exclusive. In the English verbal paradigm, although English speakers can express progressive past actions with a

Paradigm Cells	Inflection Class	
	'eat'	**'silent-e'**
Unmarked	eat	dance, erase, …
Present, 3rd	eats	dances, erases, …
Past Tense	ate	danced, erased, …
Progressive	eating	dancing, erasing, …
Passive	eaten	danced, erased, …

Table 1: The English verbal paradigm, left column, and two inflection classes of the verbal paradigm. The verb *eat* fills the cells of its inflection class with the five surface forms shown in the second column. Verbs belonging to the 'silent-e' inflection class inflect following the pattern of the third column.

grammatical construction, viz. *was eating*, there is no surface form of the lexeme *eat* that simultaneously fills both the *progressive* and the *past* cells of the verbal paradigm, **ateing*.

1.2 ParaMor

Paradigms and inflection classes, the inherent structure of natural language morphology, form the basis of ParaMor, our minimally supervised morphological induction algorithm. In ParaMor's first phase, we find sets of mutually exclusive strings which closely mirror the inflection classes of a language. In ParaMor's second phase we employ the structured knowledge contained within the discovered inflection classes to segment word forms into morpheme-like pieces.

A large caste of inflection classes can be represented as mutually exclusive substring substitutions. In the 'silent-e' inflection class, for example, the word-final strings *e.ed.es.ing* can be substituted for one another to produce the surface forms that fill the paradigm cells of lexemes belonging to this inflection class. In this paper we focus on identifying word final suffix morphology. While we focus on suffixes, the methods we employ can be straightforwardly generalized to prefixes and ongoing work seeks to model sequences of concatenative morphemes.

1.3 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), 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. Additionally, to discover paradigm-like sets of suffixes, Snover designs and searches networks of partial paradigms. These networks are the direct inspiration for ParaMor's morphology scheme networks described in section 2.

2 ParaMor

To allow in depth discussion of the performance of ParaMor in Morpho Challenge 2007, this paper gives only a brief overview of the ParaMor algorithm. Additional algorithmic details appear in Monson et al. (2007).

ParaMor begins with a search procedure designed to identify partial inflection classes containing as many true productive suffixes of a language as possible. To search, we create a network of partial possible inflection classes. Figure 1 depicts a small portion of a network derived from a Spanish newswire corpus of 50,000 types. We call each inflection class candidate in the network a *scheme*. Intuitively, a scheme is a subset of the suffixes filling the paradigm cells of a true inflection class together with the stems that empirically occurred with that set of suffixes. Figure 1 contains a frag-

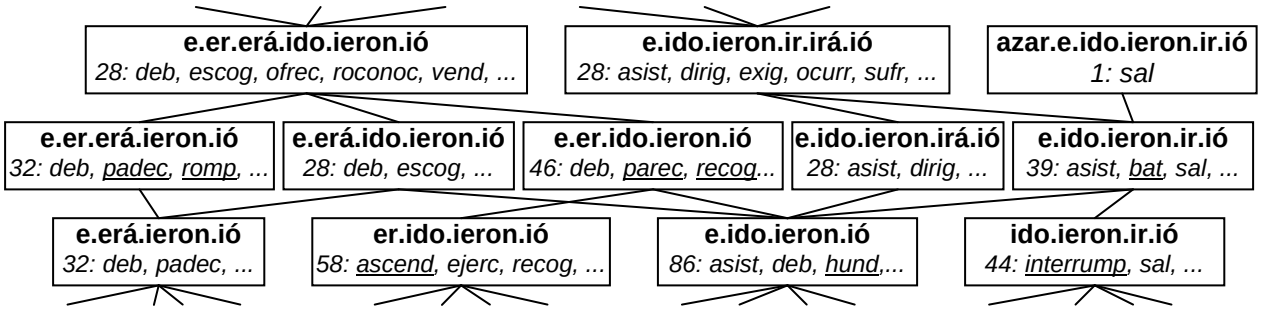

Figure 1: A small portion of a morphology scheme network—our search space of partial empirical inflection classes. This network was built from a Spanish Newswire corpus of 50,000 types, 1.26 million tokens. Each box contains a scheme. The suffixes of each scheme appear in **bold** at the top of each box. The total number of adherent stems for each scheme, together with a few exemplar stems, is in *italics*. Stems are underlined if they do not appear in any parent shown in this figure. The schemes in Figure 1 cover portions of the *er* and the *ir* Spanish verbal inflection classes. The top left scheme of the figure contains suffixes in the *er* inflection class, while the top center scheme contains suffixes in the *ir* inflection class.

ment from a scheme network built over a Spanish Newswire corpus.

ParaMor's recall centric search procedure (Monson et al., 2007) identifies schemes which likely represent portions of true inflection classes. Figure 2 contains examples of schemes selected by ParaMor's initial search. Many of the inflection class candidates which result from this initial search are incorrect. But intermingled with the false positives are candidates which collectively model significant fractions of true inflection classes. Hence, ParaMor's next step is to cluster the initial partial candidate inflection classes into larger groups. By consolidating schemes which cover portions of the same inflection class we produce sets of suffixes which more closely model the paradigm structure of natural language morphology.

The clustering of schemes presents two unique challenges. First, we must avoid over-clustering schemes which model distinct inflection classes. Second, the many small schemes which the search strategy produces act as distractive noise during clustering. To form clusters ParaMor adapts greedy hierarchical agglomerative clustering with restrictions on which clusters are allowed to merge. Restrictions such as not to place into the same cluster suffixes which share no stem in the corpus—keeping separate schemes which model inflection classes containing some of the same suffixes.

With as many initial true candidates as possible safely corralled with other candidates covering the same inflection class, ParaMor completes the paradigm discovery phase by improving scheme-cluster precision. ParaMor applies a series of filters, culling out unwanted scheme-clusters. One filter discards all unclustered schemes falling below a size threshold. Another filter, inspired by Harris (19955), discards clusters which model an incorrect morpheme boundary, such as the 1593rd selected scheme from Figure 2.

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 scheme-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 of the training or the test corpora, 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.

1) **Ø.s** 5501 stems
2) **a.as.o.os** 892 stems
...
5) **a.aba.aban.ada.adas.ado.ados.an.ando.ar.aron.arse.ará.arán.ó** 25 stems
...
12) **a.aba.ada.adas.ado.ados.an.ando.ar.aron.ará.arán.e.en.ó** 21 stems
...
209) **e.er.ida.idas.ido.idos.imiento.ió** 9 stems
...
1590) **Ø.ipo** 4 stems
1591) **ido.idos.ir.iré** 6 stems
1592) **Ø.e.iu** 4 stems
1593) **iza.izado.izan.izar.izaron.izarán.izó** 8 stems
...

Figure 2: The suffixes of some schemes selected by the initial search over a Spanish corpus of 50,000 types. While some selected schemes contain large numbers of correct suffixes, such as the 1st, 2nd, 5th, 12th, 209th, and 1591st selected schemes; many others are incorrect collections of word final strings.

3 Morpho Challenge 2007 Results

We participated in the English and the German tracks of Morpho Challenge 2007. In each track we entered three systems. The first system we entered was ParaMor alone. ParaMor's algorithm has two 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.

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 2 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 third and fourth 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% 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

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
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 2: The official **P**recision, **R**ecall, and **F_1** scores from Morpho Challenge 2007, to three significant digits. Only scores for submitted systems most relevant to a discussion of ParaMor are included.

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 keys, this time analyzed only for inflectional morphology. Table 3 contains the results of ParaMor and Morfessor against these new inflectional answer keys for English and German. ParaMor attains remarkably high recall of inflectional morphological processes for both German and particularly English. Also notably, 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 the most 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. We believe the precision of ParaMor's simple segmentation algorithm can be improved by narrowing down the proposed analyses for each word to the most likely. Perhaps ParaMor and Morfessor's vastly different strategies for morphology induction could be combined in an even more fruitful fashion. And ambitiously, we hope to extend ParaMor to analyze languages with agglutinative sequences of affixes by generalizing the definition of a scheme.

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: 2001.

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.

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

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 (HLT-NAACL)*. 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/>

	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 3: 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.

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. In Press.

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.