Simulating Morphological Analyzers with Stochastic Taggers for Confidence Estimation

Christian Monson, Kristy Hollingshead, and Brian Roark

Center for Spoken Language Understanding, Oregon Health & Science University
monsonc@csee.ogi.edu, hollingk@cslu.ogi.edu, roark@cslu.ogi.edu

Abstract. We propose a method for providing stochastic confidence estimates for rule-based and black-box natural language (NL) processing systems. Our method does not require labeled training data: We simply train stochastic models on the output of the original NL systems. Numeric confidence estimates enable both minimum Bayes risk–style optimization as well as principled system combination for these knowledge-based and black-box systems. In our specific experiments, we enrich ParaMor, a rule-based system for unsupervised morphology induction, with probabilistic segmentation confidences by training a statistical natural language tagger to simulate ParaMor's morphological segmentations. By adjusting the numeric threshold above which the simulator proposes morpheme boundaries, we improve F_1 of morpheme identification on a Hungarian corpus by 5.9% absolute. With numeric confidences in hand, we also combine ParaMor's segmentation decisions with those of a second (black-box) unsupervised morphology induction system, Morfessor. Our joint ParaMor-Morfessor system enhances F_1 performance by a further 3.4% absolute, ultimately moving F_1 from 41.4% to 50.7%.

Keywords. Natural Language Morphology, Unsupervised Learning, Morphology Induction, Statistical Simulation

1 Background

The Importance of Confidence Estimation. Confidence estimation, one of the key benefits of probabilistic modeling of natural language (NL), is crucial both for minimum Bayes risk inference as well as for stochastic system combination. Minimum Bayes risk inference enables the tuning of NL systems to achieve high precision, high recall, or something in between, while system combination can unite the complementary strengths of independent systems. Unfortunately, NL systems that we would like to optimize or combine do not always produce weights from which confidence estimates may be calculated. In some domains, knowledge-based systems are widely used and are effective, e.g., the best stemming, tokenization, and morphological analyzers for many languages are hard clustering approaches that do not involve weights or even yield alternative analyses. For other tasks, weights may be used system-internally, but are not immediately accessible to the end-user—such a system is a black-box to the user.

Here, we investigate simulating knowledge-based and black-box systems with stochastic models by training in a supervised manner on the output of the non-stochastic (or black-box) systems. The specific systems that we mimic are unsupervised morphological analyzers, which we simulate via discriminatively trained character taggers. The taggers' easily accessible posterior probabilities can serve as confidence measures for the original systems. Leveraging these newfound confidence scores, we pursue minimum Bayes risk–style thresholding of tags (for higher morpheme recall) as well as principled system combination approaches (for higher overall accuracy). As an added benefit, the shallow tag-and-character features that are employed by our simulation taggers enable generalization from the baseline system—the simulation taggers make correct decisions in contexts where the original systems do not.

A Brief History of Unsupervised Morphology Induction. Unsupervised morphology induction is the task of learning the morphological analyses of the words of an arbitrary natural language from nothing more than a raw corpus of unannotated text. Analyzing words down to the morpheme level has helped natural language processing tasks from machine translation [1] to speech recognition [2]. But building a morphological analysis system by hand can take person-months of time—hence the need for automatic methods for morphology induction.

Many approaches to unsupervised morphology induction have been proposed. Techniques inspired by Zellig Harris [3] measure the probabilities of word-internal character transitions to identify likely morpheme boundaries [4]. Other systems rely on the minimum description length principle to pick out a set of highly descriptive morphemes [5], [2]. Recent work on unsupervised morphology induction for Semitic languages has focused on estimating robust statistical models of morphology [6], [7]. And this paper extends ParaMor,[1] an induction system that leverages morphological paradigms as the inherent structure of natural language morphology [8].

Section 2 introduces the baseline systems that our taggers simulated together with the specific tagging approach that we take; Section 3 presents empirical results, demonstrating the ultimate utility of simulation; while Section 4 concludes.

2 Simulating Morphological Analyzers

The ParaMor system for unsupervised morphology induction builds sets of suffixes that model the paradigm structure found in natural language inflectional morphology. ParaMor competed in both the 2007 and 2008 Morpho Challenge Competitions [9], both solo and in a joint submission with a second unsupervised morphology induction system Morfessor [2]. Setting aside the joint ParaMor-Morfessor submission, the solo ParaMor placed first in the 2008 Turkish Linguistic competition, 46.5% F_1, and second in English, at 52.5% F_1. Meanwhile the joint ParaMor-Morfessor system placed first overall in the 2008 Linguistic competitions for German, Finnish, Turkish, and Arabic. ParaMor's successes are particularly remarkable given that ParaMor is a rule-based system incapable of measuring the confidence of the morphological segmentations it proposes—without a confidence measure, ParaMor cannot optimize its segmentation strategy toward any particular metric.

Simulating ParaMor with a Statistical Model. To gain the advantages that stochastic confidence measures provide, while retaining the strengths of the ParaMor mor-

[1] ParaMor is freely available from: `http://www.cslu.ogi.edu/~monsonc/ParaMor.html`

phology induction algorithm, we train a statistical model to simulate ParaMor's morphological segmentations. Specifically, we view the morphology segmentation task as a labeling problem akin to part-of-speech tagging. Statistical tagging is a proven and well-understood natural language processing technique that has been adapted to a variety of problems beyond part-of-speech labeling. Taggers have been used for named entity recognition [10] and NP-chunking [11]; to flag words on the periphery of a parse constituent [12]; as well as to segment written Chinese into words [13]—a task closely related to morphology segmentation.

We trained a finite-stage tagger [14] to identify, for each character, *c*, in a given word, whether or not ParaMor would place a morpheme boundary immediately before *c*. We supplied the tagger with three types of features: 1. One-sided character n-grams, 2. Two-sided character n-grams, and 3. Morpheme-tag n-grams. The one-sided n-grams are the uni-, bi-, tri-, and 4-grams that either end or begin with *c*; The two-sided n-grams are all 7-grams that extend up to five characters to the left or right of *c*; And the morpheme-tag features are the unigram, bigram, and trigram morpheme-tags, covering the current and two previous tags.

We used the averaged perceptron algorithm [15] to train the tagger. During training, the decoding process is performed using a Viterbi search with a second-order Markov assumption. At test-time, we use the forward-backward algorithm, again with a second-order Markov assumption, to output the perceptron-score of each morphological tag for each character in the word. The main benefit of decoding in this manner is that, by normalizing the scores at each character (using softmax due to the log linear modeling), we can extract the posterior probability of each tag at each character rather than just the single perceptron-preferred solution for the entire word.

Fidelity. Using our finite state tagger, we construct a baseline ParaMor-simulated segmentation by placing morpheme boundaries before each character that is tagged as the start of a new stem or affix with a posterior probability greater than 0.5. This baseline mimic segmentation, although trained to emulate ParaMor's segmentations, will not be identical to ParaMor's original segmentations of a set of words. Table 1 summarizes our tagging accuracy at emulating segmentations for the five languages and six data sets of the Linguistic competition of Morpho Challenge 2009 [16]. Tagging accuracy, the percentage of correctly tagged characters, is the standard metric used to evaluate performance in the tagging literature. We calculate accuracy by averaging over held-out test folds during 10-fold cross validation. Resource constraints compelled us to divide the full Morpho Challenge data for each language into disjoint subsets each containing approximately 100,000 word types. We then trained separate taggers over each data subset, and accuracy numbers are averaged over all subsets.

For all the test languages and scenarios, our tagger successfully emulates ParaMor at a tagging accuracy above 93%, with particular strength on German, 96.6%, and English, 97.6%. Tagging accuracies in the mid 90%s are comparable to accuracies reported for other tagging tasks.

Table 1: Tagging accuracy at simulating ParaMor's morphological segmentations.

English	German	Finnish	Turkish	Arabic -V	Arabic +V
97.6%	96.6%	93.5%	93.6%	93.3%	93.7%

Generalization. The mimic tagger's departures from the original ParaMor segmentation may either hurt or improve the segmentation quality. On the one hand, when the mimic tagger deviates from the ParaMor segmentation, the mimic may be capturing some real generalization of morphological structure that is hidden in the statistical distribution of ParaMor's original segmentation. On the other hand, a disagreement between the original and the simulated ParaMor segmentations may simply be a failure of the tagger to model the irregularities inherent in natural language morphology.

To evaluate the generalization performance of our ParaMor tagging simulator, we performed a development evaluation over a Hungarian dataset. We used Hunmorph [17], a hand-built Hungarian morphological analyzer, to produce a morphological answer key containing 500,000 unique Hungarian word types from the Hunglish corpus [18]. Our Hungarian ParaMor tagger mimic successfully generalizes: Where the original ParaMor attained an F_1 of 41.4%, the ParaMor simulator improved F_1 to 42.7%, by virtue of slightly higher recall; this improvement is statistically significant at the 95% confidence level, assuming normally distributed F_1 scores.

Optimization. Having retained ParaMor's underlying performance quality by training a natural language tagger to simulate ParaMor's segmentations, we next increase F_1 further by leveraging the tagger's probabilistic segmentation scores in a minimum Bayes risk–style optimization procedure, as follows:

```
1. For each word, $w$, in a corpus
     For each character, $c$, that does not begin $w$
       Record, in a list, $L$, the tagger mimic's probability that $c$ begins a morpheme.
2. Sort the probabilities in $L$
3. Assign $k$ to be the number of probability scores that are larger than 0.5
4. For a given positive factor, $\alpha$, identify in $L$ the probability score, $S$, above which $\alpha k$
   of the probabilities lie
5. Segment at characters which receive a probabilistic segmentation score above $S$
```

In prose, to trade off recall against precision, we move the probability threshold from the default of 0.5 to that value which will permit αk segmentations. Given the extremely peaked probability scores that the mimic tagger outputs, we adjust the number of segmentation points via α rather than via the probability threshold directly.

The Linguistic competition of Morpho Challenge evaluates morphological segmentation systems using precision, recall, and F_1 of morpheme identification. Fig. 1 plots the precision, recall, and F_1 of the ParaMor tagger mimic as the number of word-internal morpheme boundaries varies between one half and four times the baseline k number of word-internal boundaries. As Fig. 1 shows, adjusting α allows for a smooth tradeoff between precision and recall. F_1 reaches its maximum value of 47.5% at $\alpha = 4/3$. As is typical when trading off precision against recall, the maximum F_1 occurs near the α location where recall overtakes precision. The improvement in F_1 for the ParaMor tagger mimic of 4.8% is statistically significant at a 95% confidence.

System Combination of with Morfessor. In addition to enabling optimization of the ParaMor tagging simulator, numeric confidence scores permit us to combine segmentations derived from ParaMor with segmentations obtained from the freely available unsupervised morphology induction system Morfessor Categories-MAP [2]. In brief, Morfessor searches for a segmentation of a corpus that maximizes the corpus proba-

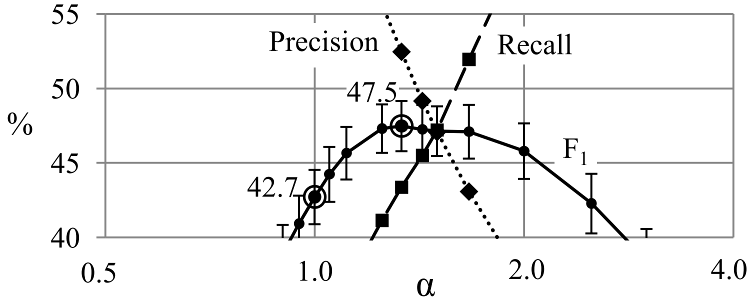

Fig. 1. Precision, Recall, and F_1 of the Hungarian ParaMor tagger mimic as α moves between 0.5 and 4.0. The error bars are 95% confidence intervals on each F_1 value.

bility score according to a specific generative probability model. The Morfessor system then further refines the morphological segmentations it proposes by restricting morpheme sequences with a Hidden Markov Model which permits only (prefix* stem suffix*)+ sequences. Our combined ParaMor-Morfessor systems differ substantially from the ParaMor-Morfessor systems that the lead author submitted to Morpho Challenge 2008—both of our updated combinations merge the ParaMor and the Morfessor segmentations of each word into a *single* analysis.

Joint ParaMor-Morfessor Mimic. The first of our combined ParaMor-Morfessor submissions builds on the idea of tagger mimics. While Morfessor has itself a statistical model that internally scores individual morphological segmentations, the final segmentations that Morfessor proposes are not by default annotated with confidences. Hence, we followed the procedure outlined in Section 2 to train a natural language tagger to simulate Morfessor's morphological analyses. It is encouraging that our technique for inducing probabilities through a mimic tagger immediately extends from a non-statistical system like ParaMor to the black-box scenario for Morfessor.

With separate taggers now simulating both ParaMor and Morfessor we then sum, for each character, *c*, in each word, the tag probabilities from the ParaMor mimic with the corresponding probabilities from the Morfessor mimic. We weighted the probability scores from the ParaMor mimic and the Morfessor mimic equally. To obtain the final morphological segmentation of each word, our combined ParaMor-Morfessor mimic followed the methodology described in Section 2 of optimizing F_1 against our Hungarian development set, with one caveat. Because we weighted the probabilities of ParaMor and Morfessor equally, any segmentation point that is strongly suggested by only one of the two systems receives an adjusted probability score just less than 0.5. Hence, we moved the baseline probability threshold from 0.5 to 0.49. With this single adjustment, the α factor that maximized Hungarian F_1 was 10/9, an 11% increase in the number of proposed morpheme boundaries.

ParaMor-Morfessor Union. The second of the two system combinations that we submitted to Morpho Challenge 2009 fuses a single morphological segmentation from the disparate segmentations proposed by the baseline ParaMor and Morfessor systems by segmenting each word at *every* location that either ParaMor or Morfessor suggests. Hence, this submission is the union of all segmentation points that are proposed by

ParaMor and Morfessor. As an example union segmentation, take the English word *polymers'*: ParaMor's segmentation of this word is *polym +er +s'*, Morfessor's is *polymer +s +'*, and the union analysis: *polym +er +s +'*.

3 Results

ParaMor in Morpho Challenge 2009. To evaluate our ParaMor tagging simulator, we competed in all the language tracks of all three competitions of Morpho Challenge 2009. Here we focus on the results of the Linguistic competition, see [16] for full details on ParaMor's performance at Morpho Challenge 2009.

To analyze morphology in a purely unsupervised fashion, a system must freeze all free parameters across all language tracks of Morpho Challenge. Our tagging mimic systems have one free parameter, α. For all languages of the Linguistic, Information Retrieval, and Machine Translation competitions of Morpho Challenge we set α at that setting which produced the highest F_1 Linguistic score on our Hungarian development set: 4/3 in the case of the ParaMor stand-alone mimic; and 10/9 for the joint ParaMor-Morfessor mimic.

The top two rows of Table 2 contain the precision, recall, and F_1 scores for the original ParaMor, which competed in the 2008 Challenge, and for the ParaMor Tagger Mimic on the non-Arabic languages[2] of Morpho Challenge 2009. Across the board, the gap between precision and recall is smaller for the ParaMor Mimic than it is for the 2008 ParaMor system. In all languages but English, the reduced precision-recall gap results in a higher F_1 score. The increase in F_1 for German, Finnish, and Turkish is more modest than the Hungarian results had led us to hope—about one percentage point in each case. Three reasons likely limited the improvements in F_1. First, the performance rose by a smaller amount for the Challenge test languages than they did for our Hungarian development set because we were explicitly tuning our α parameter to Hungarian. Second, it may be atypical that the tagger mimic generalized to outperform the baseline ParaMor system on the Hungarian data. Third, time and

Table 2: (P)recision, (R)ecall, and F_1 scores for the ParaMor and Joint ParaMor-Morfessor systems. *For English, Joint ParaMor-Morfessor achieved its highest F_1 in 2007.

	English			German			Finnish			Turkish		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ParaMor 2008	63.3	52.0	**57.1**	57.0	42.1	48.4	50.0	37.6	42.9	57.4	45.8	50.9
ParaMor Mimic	53.1	59.0	55.9	50.8	47.7	**49.2**	47.2	40.5	**43.6**	49.5	54.8	**52.0**
Union	55.7	62.3	**58.8**	52.3	60.3	**56.1**	47.9	51.0	**49.4**	47.3	60.0	52.9
Joint Mimic	54.8	60.2	57.4	51.1	57.8	54.2	51.8	45.4	48.4	48.1	60.4	**53.5**
Joint from 2008	70.1^*	67.4^*	68.7^*	64.1	61.5	62.8	65.2	50.4	56.9	66.8	58.0	62.1

[2] The small size of the Arabic training data as well as Arabic's use of morphological processes other than suffixation caused the underlying ParaMor algorithm to suffer extraordinarily low recall in both the vowelized and unvowelized Arabic scenarios.

resources constrained us to train the tagging simulators over subsets of the full Morpho Challenge data, anecdotally lowering tag-mimic accuracy by a percentage point.

The Joint ParaMor-Morfessor Systems. The bottom three rows of Table 2 summarize the performance of three combined ParaMor-Morfessor systems. The third and fourth rows of Table 2 give performance numbers for, respectively, the ParaMor-Morfessor Union system and for the Joint ParaMor-Morfessor Tagging Mimic. The final row of Table 2 lists the performance of the Joint ParaMor-Morfessor system that was submitted by the lead author to Morpho Challenge 2008.

Although the Union and Joint Mimic systems do outperform at F_1 the solo ParaMor Mimic, it was disappointing that the simple Union outscored the ParaMor-Morfessor Tagger Mimic in three of the four relevant language scenarios. Particularly surprising is that the recall of the Joint Mimic falls below the recall of the Union system in every language but Turkish. With an α factor above 1, the Joint Tagger Mimic is proposing all the segmentation points that either the ParaMor Mimic or the Morfessor Mimic hypothesize—effectively the union of the mimic systems. And yet recall is below that of the raw Union system. We can only conclude that the cumulative failure of the ParaMor and Morfessor Mimics to emulate, let alone generalize from, the original systems' segmentations drags down the recall (and precision) of the Joint Mimic.

Table 2 also highlights the relative success of the 2008 Joint ParaMor-Morfessor system. In particular, the precision scores of the 2008 system are significantly above the precision scores of the Joint Mimic and Union systems that we submitted to the 2009 Challenge. The 2008 system did *not* form a single unified segmentation for each word, but instead simply proposed the ParaMor analysis of each word alongside the Morfessor analysis—as if each word were ambiguous between a ParaMor and a Morfessor analysis. The evaluation procedure of Morpho Challenge performs a non-trivial average over alternative segmentations of a word. It is a shortcoming of the Morpho Challenge evaluation methodology to inflate precision scores when disparate systems' outputs are proposed as 'alternative' analyses.

4 Summary and Next Steps

Using a statistical tagging model we have imbued rule-based and black-box morphology analysis systems with confidence scores. These probabilistic scores have allowed us to successfully optimize the systems' morphological analyses toward a particular metric of interest, the Linguistic evaluation metric of Morpho Challenge 2009.

Looking forward, we believe our statistical taggers can be enhanced along two separate avenues. First, via careful feature engineering: Tagging accuracy might improve by, for example, employing character n-grams longer than 7-grams. Second, we hope to optimize our segmentation threshold α for each language separately via co-training of ParaMor against Morfessor. We are also interested in using statistical models to simulate rule-based and black-box systems from other areas of NLP.

Acknowledgements. This research was supported in part by NSF Grant #IIS-0811745 and DOD/NGIA grant #HM1582-08-1-0038. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF or DOD.

References

1. Oflazer, K., El-Kahlout İ.D.: Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation. In: Statistical MT Workshop at ACL (2007)
2. Creutz, M.: 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)
3. Harris, Z.: From Phoneme to Morpheme. In: Language, vol. 31.2, pp. 190-222 (1955). Reprinted in: Harris, Z.: Papers in Structural and Transformational Linguists. Reidel D. (ed.) Dordrecht (1970)
4. Bernhard, D.: Simple Morpheme Labeling in Unsupervised Morpheme Analysis. In: Lecture Notes in Computer Science: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, Revised Selected Papers. LNCS vol. 5152/2008, pp. 873-880. Springer (2008)
5. Goldsmith, J.: Unsupervised Learning of the Morphology of a Natural Language. In: Computational Linguistics, vol. 27.2, pp. 153-198 (2001)
6. Snyder, B., Barzilay, R..: Unsupervised Multilingual Learning for Morphological Segmentation. In: Proceedings of ACL-08: HLT (2008)
7. Poon, H., Cherry, C., Toutanova, K.: Unsupervised Morphological Segmentation with Log-Linear Models. In: Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL (2009)
8. Monson, C.: ParaMor: From Paradigm Structure to Natural Language Morphology Induction. Ph.D. Thesis, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania (2009)
9. Monson, C., Carbonell, J., Lavie, A., Levin, L.: ParaMor and Morpho Challenge 2008. In: LNCS: Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Revised Selected Papers, Aarhus, Denmark, Springer, vol. 5706/2009, pp. 967-974 (2009)
10. Tjong Kim Sang, E. F.: Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2002 (2002)
11. Tjong Kim Sang, E. F., Buchholz, S.: Introduction to the CoNLL-2000 Shared Task: Chunking. In: Computational Natural Language Learning (CoNLL) (2000)
12. Roark, B., Hollingshead, K.: Linear Complexity Context-Free Parsing Pipelines via Chart Constraints. In: Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL (2009)
13. Xue, N.: Chinese Word Segmentation as Character Tagging. In: Computational Linguistics and Chinese Language Processing, vol. 8.1, pp. 29-47 (2003)
14. Hollingshead, K., Fisher, S., Roark, B.: Comparing and Combining Finite-State and Context-Free Parsers. In: Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP) (2005)
15. Collins, M.: Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2002)
16. Kurimo, M., Virpioja, S., Turunen, V.T., Blackwood, G.W, Byrne, W.: Overview and Results of Morpho Challenge 2009. In: 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, Revised Selected Papers. LNCS Springer (2010)
17. Trón, V., Gyepesi, G., Halácsy, P., Kornai, A., Németh, L., Varga, D.: Hunmorph: Open Source Word Analysis. In: ACL Workshop on Software (2005)
18. Varga, D., Halácsy, P., Kornai, A., Németh, L., Trón, V., Váradi, T., Sass, B., Bottyán, G., Héja, E., Gyarmati, Á., Mészáros, Á., Labundy, D.: Hunglish corpus. Accessed on August 18, 2009. <http://mokk.bme.hu/resources/hunglishcorpus> (2009)