A Hybrid Approach to the Induction of Underlying Morphology

Michael Tepper
Department of Linguistics
University of Washington
Seattle, WA 98195
mtepper@u.washington.edu

Fei Xia
Department of Linguistics
University of Washington
Seattle, WA 98195
fxia@u.washington.edu

Abstract

We present a technique for refining a baseline segmentation and generating a plausible underlying morpheme segmentation by integrating hand-written rewrite rules into an existing state-of-the-art unsupervised morphological induction procedure. Performance on measures which consider surface-boundary accuracy and underlying morpheme consistency indicates this technique leads to improvements over baseline segmentations for English and Turkish word lists.

1 Introduction

1.1 Unsupervised Morphological Induction

The primary goal of unsupervised morphological induction (UMI) is the simultaneous induction of a reasonable morphological lexicon as well as an optimal segmentation of a corpus of words, given that lexicon. The majority of existing approaches employ statistical modeling towards this goal, but differ with respect to how they learn or refine the morphological lexicon. While some approaches involve lexical priors, either internally motivated or motivated by the minimal description length (MDL) criterion, some utilize heuristics. Pure maximum likelihood (ML) approaches may refine the lexicon with heuristics in lieu of explicit priors (Creutz and Lagus, 2004), or not make categorical refinements at all concerning which morphs are included, only probabilistic refinements through a hierarchical EM procedure (Peng and Schuurmans, 2001). Approaches that optimize the lexicon with respect to priors come in several flavors. There are basic maximum a priori (MAP) approaches that try to maximize the probability of the lexicon against linguistically motivated priors (Deligne and Bimbot, 1997; Snover and Brent, 2001; Creutz and Lagus, 2005). An alternative to MAP, MDL approaches use their own set of priors motivated by complexity theory. These studies attempt to minimize lexicon complexity (bit-length in crude MDL) while simultaneously minimizing the complexity (by maximizing the probability) of the corpus given the lexicon (de Marcken, 1996; Goldsmith, 2001; Creutz and Lagus, 2002).

Many of the approaches mentioned above utilize a simplistic unigram model of morphology to produce the segmentation of the corpus given the lexicon. Substrings in the lexicon are proposed as morphs within a word based on frequency alone, independently of phrase-, word- and morph-surroundings (de Marcken, 1996; Peng and Schuurmans, 2001; Creutz and Lagus, 2002). There are many approaches, however, which further constrain the segmentation procedure. The work by Creutz and Lagus (2004; 2005; 2006) constrains segmentation by accounting for morphotactics, first assigning mophotactic categories (prefix, suffix, and stem) to baseline morphs, and then seeding and refining an HMM using those category assignments. Other more structured models include Goldsmith's (2001) work which, instead of inducing morphemes, induces morphological signatures like {*ø*, *s*, *ed*, *ing*} for English regular verbs. Some techniques constrain possible analyses by employing approximations for morphological meaning or usage to prevent false derivations (like *singed* = *sing* + *ed*). There is work by Schone and Jurafsky (2000; 2001) where *meaning* is proxied by word- and morph-context, condensed via LSA. Yarowsky and Wicentowski (2000) and Yarowsky et al. (2001) use expectations on relative frequency of aligned inflected-word, stem pairs, as well as POS context features, both of which approximate some sort of meaning.

1.2 Allomorphy in UMI

Allomorphy, or allomorphic variation, is the process by which a morpheme varies (orthographically or

phonologically) in particular contexts, as constrained by a grammar.[1] To our knowledge, there is only handful of work within UMI attempting to integrate allomorphy into morpheme discovery. A notable approach is the Wordframe model developed by Wicentowski (2002), which performs weighted edits on root-forms, given context, as part of a larger similarity alignment model for discovering <inflected-form, root-form> pairs.

Morphological complexity is fixed by a template; the original was designed for inflectional morphologies and thus constrained to finding an optional affix on either side of a stem. Such a template would be difficult to design for agglutinative morphologies like Turkish or Finnish, where stems are regularly inflected by chains of affixes. Still, it can be extended. A notable recent extension accounts for phenomena like infixation and reduplication in Filipino (Cheng and See, 2006).

In terms of allomorphy, the approach succeeds at generalizing allomorphic patterns, both stem-internally and at points of affixation. A major drawback is that, so far, it does not account for affix allomorphy involving character replacement–that is, beyond point-of-affixation epentheses or deletions.

1.3 Our Approach

Our approach aims to integrate a rule-based component consisting of hand-written rewrite rules into an otherwise unsupervised morphological induction procedure in order to refine the segmentations it produces.

1.3.1 Context-Sensitive Rewrite Rules

The major contribution of this work is a rule-based component which enables simple encoding of context-sensitive rewrite rules for the analysis of induced morphs into plausible underlying morphemes.[2] A rule has the form general form:

$$\underset{\text{underlying}}{\alpha} \rightarrow \underset{\text{surface}}{\beta} \; / \; \underset{\text{l. context}}{\gamma} \_ \underset{\text{r. context}}{\delta} \qquad (1)$$

It is also known as a SPE-style rewrite rule, part of the formal apparatus to introduced by Chomsky and Halle (1968) to account for regularities in phonology. Here we use it to describe orthographic patterns. Mapping morphemes to underlying forms with context-sensitive rewrite rules allows us to peer through the fragmentation created by allomorphic variation. Our experiments will show that this has the effect of allowing for more unified, consistent morphemes while simultaneously making surface boundaries more transparent.

For example, take the English multipurpose inflectional suffix $\cdot s$, normally written as $\cdot s$, but as $\cdot es$ after sibilants (*s*,*sh*, *ch*, ...). We can write the following SPE-style rule to account for its variation.

$$\underset{\text{underlying}}{\emptyset} \rightarrow \underset{\text{surface}}{\text{e}} \; / \; [+\text{SIB}] + \_ s \qquad (2)$$

This rule says, "Insert an *e* (map *nothing* to *e*) following a character marked as a sibilant (+SIB) and a morphological boundary (+), at the focus position (_), immediately preceding an *s*." In short, it enables the mapping of the underlying form $\cdot s$ to $\cdot es$ by inserting an *e* before *s* where appropriate. When this rule is reversed to produce underlying analyses, the $\cdot es$ variant in such words as glasses, matches, swishes, and buzzes can be identified with the $\cdot s$ variant in words like plots, sits, quakes, and nips.

1.3.2 Overview of Procedure

Before the start of the procedure, there is a preprocessing step to derive an initial segmentation.

This segmentation is fed to the EM Stage, the goal of which is to find the maximum probability segmentation of a wordlist into *underlying* morphemes. First, analyses of initial segments are produced by rule. Then, their frequency is used to determine their likelihood as underlying morphemes. Finally, probability of a segmentation into underlying morphemes is maximized.

The output segmentation feeds into the Split Stage, where heuristics are used to split large, high-frequency segments that fail to break into smaller underlying morphemes during the EM algorithm.

2 Procedure

A flowchart of the procedure is given in Figure 1.

Preprocessing We use the Categories-MAP algorithm developed by Creutz and Lagus (2005; 2006) to produce an initial morphological segmentation. Here, a segmentation is optimized by maximum a posteriori estimate given priors on length, frequency, and usage of morphs stored in the model. Their procedure begins with morphological tags indicating basic morphotactics (prefix, stem, suffix, noise) being assigned heuristically to a baseline segmentation. That tag assignment is then used to seed an HMM.

[1] In this work we focus on orthographic allomorphy.

[2] Ordered rewrite rules, when restricted from applying to their own output, have similar expressive capabilities to Koskenniemi's two-level constraints. Both define regular relations on strings, both can be compiled into lexical transducers, and both have been used in finite-state analyzers (Karttunen and Beesley, 2001). We choose ordered rules because they are easier to write given our task and resources.

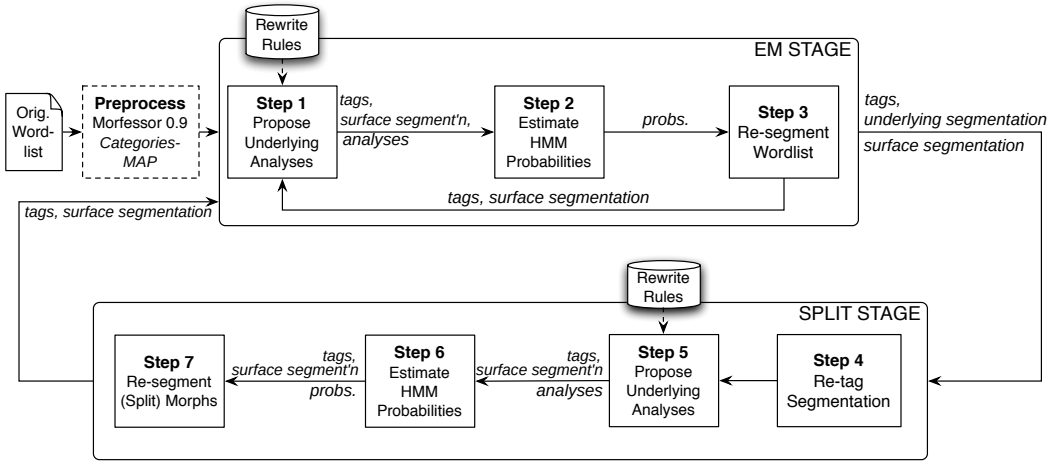

Figure 1: Flowchart showing the entire procedure.

Optimal segmentation of a word is simultaneously the best tag and morph[3] sequence given that word. The contents of the model are optimized with respect to length, frequency, and usage priors during splitting and joining phases. The final output is a tagged segmentation of the input word-list.

2.1 EM Stage

The model we train is a modified version of the morphological HMM from the work of Creutz and Lagus (2004-2006), where a word w consists of a sequence of morphs generated by a morphological-category tag sequence. The difference between their HMM and ours is that theirs emits surface morphs, while ours emits *underlying* morphemes. Morphemes may either be analyses proposed by rule or surface morphs acting as morphemes. We do not modify the tags Creutz and Lagus use (prefix, stem, suffix, and noise).

We proceed by EM, initialized by the preprocessed segmentation. Rule-generated underlying analyses are produced (Step 1), and used to estimate the emission probability $P(u_i|t_i)$ and transition probability $P(t_i|t_{i-1})$ (Step 2). In successive E-steps, Steps 1 and 2 are repeated. The M-step (Step 3) involves finding the maximum probability decoding of each word according to Eq (6), i.e. maximum probability tag and morpheme sequence.

Step 1 - Derive Underlying Analyses In this step, handwritten context-sensitive rewrite rules derive context-relevant analyses for morphs in the preprocessed segmentation. These analyses are produced by a set of ordered rules that propose deletions, insertions, or substitutions when triggered by the proper characters around a segmentation boundary.[4] A rule applies wherever contextually triggered, from left to right, and may apply more than once to the same word. To prevent the runaway application of certain rules, a rule may not apply to its own output. The result of applying a rule is a (possibly spelling-changed) segmented word, which is fed to the next rule. This enables multi-step analyses by using rules designed specifically to apply to the outputs of other rules. See Figure 2 for a small example.

Step 2 - Estimate HMM Probabilities Transition probabilities $P(t_i|t_{i-1})$ are estimated by maximum likelihood, given a tagged input segmentation.

Emission probabilities $P(u_i|t_i)$ are also estimated by maximum likelihood, but the situation is slightly more complex; the probability of morphemes u_i are estimated according to frequencies of association (coindexation) with surface morphs s_i and tags t_i.

Furthermore an underlying morpheme u_i can either be identical to its associated surface morph s_i when no rules apply, or be a rule-generated analysis. For the sake of clarity, we call the former u_i' and the latter u_i'', as defined below:

$$u_i = \begin{cases} u_i' & \text{if } u_i = s_i \\ u_i'' & \text{otherwise} \end{cases}$$

When an underlying morpheme u_i is associated to a surface morph s, we refer to s as an *allomorph* of

[3] A *morph* is a linguistic morpheme as it occurs in production, i.e. as it occurs in a *surface* word.

[4] Some special substitution rules, like vowel harmony in Turkish and Finnish, have a spreading effect, moving from syllable to syllable within and beyond morph-boundaries. In our formulation, these rules differ from other rules by not being conditioned on a morph-boundary.

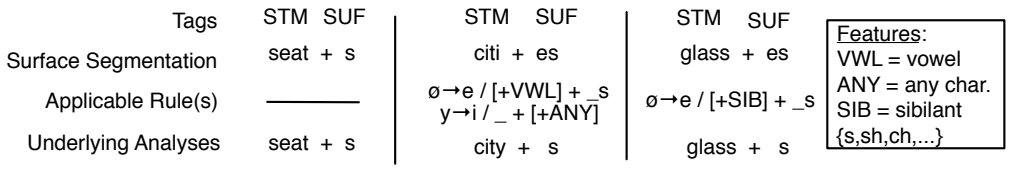

Figure 2: Underlying analyses for a segmentation are generated by passing it through context-sensitive rewrite rules. Rules apply to some morphs (e.g., *citi* → *city*) but not to others (e.g., *glass* → *glass*).

u_i. The probability of u_i given tag t_i is calculated by summing over all allomorphs s of u_i the probability that u_i realizes s in the context of tag t_i:

$$P(u_i|t_i) = \sum_{s \in \text{allom.-of}(u_i)} P(u_i, s|t_i) \quad (3)$$

$$= \sum_{s \in \text{allom.-of}(u_i)} P(u_i|s, t_i)P(s|t_i) \quad (4)$$

Both Eq (3) and Eq (4) are trivial to estimate with counting on our input from Step 1 (see Figure 2). We show (4) because it has the term $P(u_i|s, t_i)$, which may be used for thresholding and discounting terms of the sum where u_i is rarely associated with a particular allomorph and tag. In the future, such discounting may be useful to filter out noise generated by noisy or permissive rules. So far, this type of discounting has not improved results.

Step 3 - Resegment Word List Next we resegment the word list into underlying morphemes.

Searching for the best breakdown of a word w into morpheme sequence $\mathbf{u}$ and tag sequence $\mathbf{t}$, we maximize the probability of the following formula:

$$P(w, \mathbf{u}, \mathbf{t}) = P(w|\mathbf{u}, \mathbf{t})P(\mathbf{u}, \mathbf{t})$$

$$= P(w|\mathbf{u}, \mathbf{t})P(\mathbf{u}|\mathbf{t})P(\mathbf{t}) \quad (5)$$

To simplify, we assume that $P(w|\mathbf{u}, \mathbf{t})$ is equal to one.[5] With this assumption in mind, Eq (5) reduces to $P(\mathbf{u}|\mathbf{t})P(\mathbf{t})$. With independence assumptions and a local time horizon, we estimate:

$$\operatorname*{argmax}_{\mathbf{u},\mathbf{t}} P(\mathbf{u}|\mathbf{t})P(\mathbf{t})$$

$$\approx \operatorname*{argmax}_{\mathbf{u},\mathbf{t}} \Big[\prod_{i=1}^{n} P(u_i|t_i)P(t_i|t_{i-1})\Big] \quad (6)$$

[5]In other words, we make the assumption that a sequence of underlying morphemes and tags corresponds to just one word. This assumption may need revision in cases where morphemes can optionally undergo the types of spelling changes we are trying to encode; this has not been the case for the languages under investigation.

The search for the maximum probability tag and morph sequence in Eq (6) is carried out by a modified version of the Viterbi algorithm. The maximum probability segmentation for a given word may be a mixture of both types of underlying morpheme, u'_i and u''_i. Also, wherever we have a choice between emitting u'_i, identical to the surface form, or u''_i, an analysis with rule-proposed changes, the highest probability of the two is always selected.

2.2 Split Stage

Many times, large morphs have substructure and yet are too frequent to be split when segmented by the HMM in the EM Stage. To overcome this, we approximately follow the heuristic procedure[6] laid out by Creutz and Lagus (2004), encouraging splitting of larger morphs into smaller underlying morphemes. This process has the danger of introducing many false analyses, so first the segmentation must be re-tagged (Step 4) to identify which morphemes are noise and should not be used. Once we re-tag, we re-analyze morphs in the surface segmentation (Step 5) and re-estimate HMM probabilities (Step 6). (for Steps 5 and 6, refer to Steps 1 and 2). Finally, we use these HMM probabilities to split morphs (Step 7).

Step 4 - Re-tag the Segmentation To identify noise morphemes, we estimate a distribution $P(CAT|u_i)$ for three true categories CAT (prefix, stem, or suffix) and one noise category; we then assign categories randomly according to this distribution. Stem probabilities are proportional to stem-length, while affix probabilities are proportional to left- or right- perplexity. The probability of true categories are also tied to the value of sigmoid-cutoff parameters, the most important of which is b, which thresholds the probability of both types of affix (prefix and suffix).

The probability of the noise category is conversely related to the product of true category probabilities;

[6]The main difference between our procedure and Creutz and Lagus (2004) is that we allow splitting into two or more morphemes (see Step 7) while they allow binary splits only.

when true categories are less probable, noise becomes more probable. Thus, adjusting parameters like b can increase or decrease the probability of noise.

Step 7 - Split Morphs In this step, we examine <morph, tag> pairs in the segmentation to see if a split into sub-morphemes is warranted. We constrain this process by restricting splitting to stems (with the option to split affixes), and by splitting into restricted sequences of tags, particularly avoiding noise. We also use parameter b in Step 4 as a way to discourage excessive splitting by tagging more morphemes as noise. Stems are split into the sequence: (PRE* STM SUF*). Affixes (prefixes and suffixes) are split into other affixes of the same category. Whether to split affixes depends on typological properties of the language. If a language has agglutinative suffixation, for example, we hand-set a parameter to allow suffix-splitting.

When examining a morph for splitting, we search over all segmentations with at least one split, and choose the one that is both optimal according to Eq (6) and does not violate our constraints on what category sequences are allowed for *its* category. We end this step by returning to the EM Stage, where another cycle of EM is performed.

3 Experiments and Results

In this section we report and discuss development results for English and Turkish. We also report final-test results for both languages. Results for the pre-processed segmentation are consistently used as a baseline. In order to isolate the effect of the rewrite rules, we also compare against results taken on a parallel set of experiments, run with all the same parameters but without rule-generated underlying morphemes, i.e. without morphemes of type u_i''. But before we get to these results, we will describe the conditions of our experiments. First we introduce the evaluation metrics and data used, and then detail any parameters set during development.

3.1 Evaluation Metrics

We use *two* procedures for evaluation, described in the Morpho Challenge '05 and '07 Competition Reports (Kurimo et al., 2006; Kurimo et al., 2007). Both procedures use gold-standards created with commercially available morphological analyzers for each language. Each procedure is associated with its own F-score-based measure.

The first was used in Morpho Challenge '05, and measures the extent to which *boundaries* match between the surface-layer of our segmentations and gold-standard surface segmentations.

The second was used in Morpho Challenge '07 and measures the extent to which *morphemes* match between the underlying-layer of our segmentations and gold-standard underlying analyses. The F-score here is not actually on matched morphemes, but instead on matched morpheme-sharing word-pairs. A point is given whenever a morpheme-sharing word-pair in the gold-standard segmentation also shares morphemes in the test segmentation (for recall), and vice-versa for precision.

3.2 Data

Training Data The data-sets used for training were provided by the Helsinki University of Technology in advance of the Morpho Challenge '07 and were downloaded by the authors from the contest website[7]. According to the website, they were compiled from the University of Leipzig Wortschatz Corpora.

	Sentences	**Tokens**	**Types**
English	3×10^6	6.22×10^7	3.85×10^5
Turkish	1×10^6	1.29×10^7	6.17×10^5

Table 1: Training corpus sizes vary slightly, with 3 million English sentences and 1 million Turkish sentences.

Development Data The development gold-standard for the surface metric was provided in advance of Morpho Challenge '05 and consists of surface segmentations for 532 English and 774 Turkish words.

The development gold-standard for the underlying metric was provided in advance of Morpho Challenge '07 and consists of morphological analyses for 410 English and 593 Turkish words.

Test Data For final testing, we use the gold-standard data reserved for final evaluation in the Morpho Challenge '07 contest. The gold-standard consists of approximately 1.17×10^5 English and 3.87×10^5 Turkish analyzed words, roughly a tenth the size of training word-lists. Word pairs that exist in both the training and gold standard are used for evaluation.

3.3 Parameters

There are two sets of parameters used in this experiment. First, there are parameters used to produce the initial segmentation. They were set as suggested in Cruetz and Lagus (2005), with parameter b tuned on development data.

[7] http://www.cis.hut.fi/morphochallenge2007/datasets.shtml

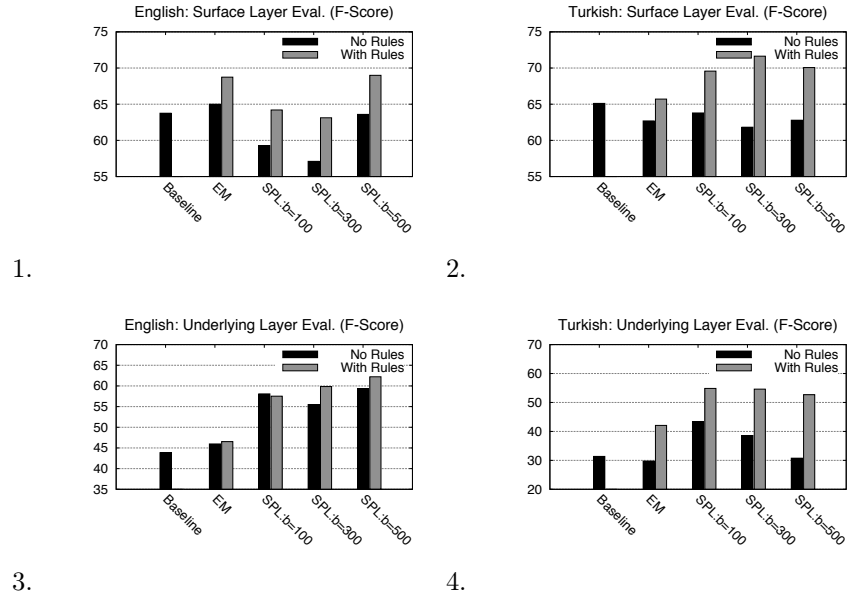

Figure 3: Development results for the preprocessed initial segmentation (Baseline), and segmentations produced by our approach, first after the EM Stage (EM) and again after the Split Stage (SPL) with different values of parameter b. Rules that generate underlying analyses have either been included (With Rules), or left out (No Rules).

Then there are parameters used for the main procedure. Here we have rewrite rules, numerical parameters, and one typology parameter. Rewrite rules and any orthographic features they use were culled from linguistic literature. We currently have 6 rules for English and 10 for Turkish; See Appendix A.1 for the full set of English rules used. Numerical parameters were set as suggested in Cruetz and Lagus (2004), and following their lead we tuned b on development data; we show development results for the following values: $b = 100$, 300, and 500 (see Figure 3). Finally, as introduced in Section 2.2, we have a hand-set typology parameter that allows us to split prefixes or suffixes if the language has an agglutinative morphology. Since Turkish has agglutinative suffixation, we set this parameter to split suffixes for Turkish.

3.4 Development Results

Development results were obtained by evaluating English and Turkish segmentations at several stages, and with several values of parameter b as shown in Figure 3.

Overall, our development results were very positive. For the surface-level evaluation, the largest F-score improvement was observed for English (Figure 3, Chart 1), 63.75% to 68.99%, a relative F-score gain of 8.2% over the baseline segmentation. The Turkish result also improves to a similar degree, but it is only achieved after the model as been refined by splitting. For English we observe the improvement earlier, after the EM Stage. For the underlying-level evaluation, the largest F-score improvement was observed for Turkish (Chart 4), 31.37% to 54.86%, a relative F-score gain of over 74%.

In most experiments with rules to generate underlying analyses (*With Rules*), the successive applications of EM and splitting result in improved results. Without rule-generated forms (*No Rules*) the results tend be negative compared to the baseline (see Figure 3, Chart 2), or mixed (Charts 1 and 4). When we look at recall and precision numbers directly, we observe that even without rules, the algorithm produces large recall boosts (especially after splitting). However, these boosts are accompanied by precision losses, which result in unchanged or lower F-scores.

The exception is the underlying-level evaluation of English segmentations (Figure 3, Chart 3). Here we observe a near-parity of F-score gains for segmentations produced with and without underlying morphemes derived by rule. One explanation is that the English initial segmentation is conservative and that coverage gains are the main reason for improved English scores. Creutz and Lagus (2005) note that the Morfessor EM approach often has better coverage than the MAP approach we use to produce the

				Hybrid:After Split	
	MC Morf.	**MC Top**	**Baseline**	**No Rules**	**With Rules**
English	47.17	**60.81**	47.04	57.35	59.78
Turkish	37.10	29.23	32.76	31.10	**54.54**

Table 2: Final test F-scores on the underlying morpheme measure used in Morpho Challenge '07. MC Morf. is Morfessor MAP, which was used as a reference method in the contest. MC Top is the top contestant. For our hybrid approach, we show the F-score obtained with and without using rewrite rules. The splitting parameter b was set to the best performing value seen in development evaluations (Tr. $b = 100$, En. $b = 500$).

initial segmentation. Also, in English, allomorphy is not as extensive as in Turkish (see Chart 4) where precision losses are greater without rules, i.e. when not representing allomorphs by the same morpheme.

3.5 Final Test Results

Final test results, given in Table 2, are mixed. For English, though we improve on our baseline and on Morfessor MAP trained by Creutz and Lagus, we are beaten by the top unsupervised Morpho Challenge contestant, entered by Delphine Bernhard (2007). Bernhard's approach was purely unsupervised and did not explicitly account for allomorphic phenomena. There are several possible reasons why we were not the top performer here. Our splitting constraint for stems, which allows them to split into stems and chains of affixes, is suited for agglutinative morphologies. It does not seem particularly well suited to English morphology. Our rewrite-rules might also be improved. Finally, there may be other, more pressing barriers (besides allomorphy) to improving morpheme induction in English, like ambiguity between homographic morphemes.

For Turkish, the story is very different. We observe our baseline segmentation going from 32.76% F-score to 54.54% when re-segmented using rules, a relative improvement of over 66%. Compared with the top unsupervised approach, Creutz and Lagus's Morfessor MAP, our F-score improvement is over 48%. The distance between our hybrid approach and unsupervised approaches emphasizes the problem allomorphy can be for a language like Turkish. Turkish inflectional suffixes, for instance, regularly undergo multiple spelling-rules and can have 10 or more variant forms. Knowing that these variants are all one morpheme makes a difference.

4 Conclusion

In this work we showed that we can use a small amount of knowledge in the form of context-sensitive rewrite rules to improve unsupervised segmentations for Turkish and English. This improvement can be quite large. On the morpheme-consistency measure used in the last Morpho Challenge, we observed an improvement of the Turkish segmentation of over 66% against the baseline, and 48% against the top-of-the-line unsupervised approach.

Work in progress includes error analysis of the results to more closely examine the contribution of each rule, as well as developing rule sets for additional languages. This will help highlight various aspects of the most beneficial rules.

There has been recent work on discovering allomorphic phenomena automatically (Dasgupta and Ng, 2007; Demberg, 2007). It is hoped that our work can inform these approaches, if only by showing what variation is possible, and what is relevant to particular languages. For example, variation in inflectional suffixes, driven by vowel harmony and other phenomena, should be captured for a language like Turkish.

Future work involves attempting to learn broad-coverage underlying morphology without the hand-coded element of the current work. This might involve employing aspects of the most beneficial rules as variable features in rule-templates. It is hoped that we can start to derive underlying morphemes through processes (rules, constraints, etc) suggested by these templates, and possibly learn instantiations of templates from seed corpora.

A Appendix

A.1 Rules Used For English

e epenthesis before s suffix
ø →e / ..[+V] + _s
ø→e / ..[+SIB] + _s
long e deletion
e →ø / ..[+V][+C]_ + [+V]
change y to i before suffix
y →i / ..[+C] +? _ + [+ANY]
consonant gemination
ø →α[+STOP] / ..α[+STOP]_ + [+V]
ø →α[+STOP] / ..α[+STOP]_ + [+GLI]

Table 3: English Rules

A.2 Example Segmentations

Base	EM	SPL:b=300	SPL:b=500
happen s	happen s	happ e n s	happen s
happier	happier	happi er	happi er
happiest	happiest	happ i est	happiest
happily	happily	happi ly	happi ly
happiness	happiness	happi ness	happiness

Table 4: Surface segmentations after preprocessing (Base), EM Stage (EM), and Split Stage (SPL)

References

Delphine Bernhard. 2007. Simple morpheme labeling in unsupervised morpheme analysis. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary.

Charibeth K. Cheng and Solomon L. See. 2006. The revised wordframe model for the filipino language. *Journal of Research in Science, Computing and Engineering*.

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row, New York.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proc. Workshop on Morphological and Phonological Learning of ACL'02*, pages 21–30, Philadelphia. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2004. Induction of a simple morphology for highly inflecting languages. In *Proc. 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 43–51, Barcelona.

Mathias Creutz and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *Proc. International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, pages 106–113, Espoo, Finland.

Mathias Creutz and Krista Lagus. 2006. Morfessor in the morpho challenge. In *Proc. PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*, Venice, Italy.

Sajib Dasgupta and Vincent Ng. 2007. High performance, language-independent morphological segmentation. In *Proc. NAACL'07*.

Carl G. de Marcken. 1996. *Unsupervised Language Acquisition*. Ph.D. thesis, Massachussetts Institute of Technology, Boston.

Sabine Deligne and Frédéric Bimbot. 1997. Inference of variable-length linguistic and acoustic units by multigrams. *Speech Communication*, 23:223–241.

Vera Demberg. 2007. A language-independent unsupervised model for morphological segmentation. In *Proc. ACL'07*.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27.2:153–198.

Lauri Karttunen and Kenneth R. Beesley. 2001. A short history of two-level morphology. In *Proc. ESSLLI 2001*.

Mikko Kurimo, Mathias Creutz, Matti Varjokallio, Ebru Arisoy, and Murat Saraçlar. 2006. Unsupervised segmentation of words into morphemes – Morpho Challenge 2005, an introduction and evaluation report. In *Proc. PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*, Venice, Italy.

Mikko Kurimo, Mathias Creutz, and Matti Varjokallio. 2007. Unsupervised morpheme analysis evaluation by a comparison to a linguistic gold standard – Morpho Challenge 2007. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary.

Fuchun Peng and Dale Schuurmans. 2001. A hierarchical em approach to word segmentation. In *Proc. 4th Intl. Conference on Intel. Data Analysis (IDA)*, pages 238–247.

Patrick Schone and Daniel Jurafsky. 2000. Knowledge-free induction of morphology using latent semantic analysis. In *Proc. CoNLL'00 and LLL'00*, pages 67–72, Lisbon.

Patrick Schone and Daniel Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *Proc. NAACL'01*, Pittsburgh.

Matthew G. Snover and Michael R. Brent. 2001. A bayesian model for morpheme and paradigm identification. In *Proc. ACL'01*, pages 482–490, Toulouse, France.

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

David Yarowsky and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proc. ACL'00*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection accross aligned corpora. In *Proc. HLT'01*, volume HLT 01, pages 161–168, San Diego.