A Rule-based Acquisition Model Adapted for Morphological Analysis*

Constantine Lignos[1], Erwin Chan[2], Mitchell P. Marcus[1], and Charles Yang[1]

[1] University of Pennsylvania,
lignos@cis.upenn.edu, mitch@cis.upenn.edu, charles.yang@ling.upenn.edu
[2] University of Arizona,
echan3@u.arizona.edu

Abstract. We adapt the cognitively-oriented morphology acquisition model proposed in (Chan 2008) to perform morphological analysis, extending its concept of base-derived relationships to allow multi-step derivations and adding features required for robustness on noisy corpora. This results in a rule-based morphological analyzer which attains an F-score of 58.48% in English and 33.61% in German in the Morpho Challenge 2009 Competition 1 evaluation. The learner's performance shows that acquisition models can effectively be used in text-processing tasks traditionally dominated by statistical approaches.

1 Introduction

Although extensive work has been done on creating high-performance unsupervised or minimally supervised morphological analzyers (Creutz and Lagus 2005, Monson 2008, Wicentowski 2002), little work has been done to bridge the gap between the computational task of morphological analysis and the cognitive task of morphological acquisition. We address this by adapting the acquisition model presented in (Chan 2008) to the task of morphological analysis, demonstrating the effectiveness of cognitively-oriented models on analysis tasks.

The most well-known cognitive models (Pinker 1999, Rumelhart and McClelland 1986) are poorly suited for unsupervised morphological analysis given that they are commonly focused on a single morphological task, the English past tense, and are based on the assumption that pairs of morphologically related words, such as *make/made*, are given to the learner. While there is evidence that clustering-based approaches can identify sets of morphologically related words (Parkes et al. 1998, Wicentowski 2002), word-pair based algorithms have only been evaluated on error-free pairs.

Many computational models have focused on segmentation-based approaches, most commonly using simple transitional-probability heuristics (Harris 1955,

* Thanks to Jana Beck for her assistance in analyzing the German results and for her insightful comments throughout the development process. Portions of this paper were adapted from the material presented in the CLEF 2009 Morpho Challenge Workshop (Lignos et al. 2009).

1. Pre-process words and populate the Unmodeled set.
2. Until a stopping condition is met, perform the main learning loop:
   (a) Count suffixes in words of the Base ∪ Unmodeled set and the Unmodeled set.
   (b) Hypothesize transforms from words in Base ∪ Unmodeled to words in Unmodeled.
   (c) Select the best transform.
   (d) Reevaluate the words that the selected transform applies to, using the Base, Derived and Unmodeled sets
   (e) Move the words used in the transform accordingly.
3. Break compound words in the Base and Unmodeled sets.

Fig. 1. The Learning Algorithm

Keshava and Pitler 2006), or n-gram-based statistical models (most recently Spiegler 2009). Often segmentation-based approaches organize the segmentations learned into paradigms (Goldsmith 2001, Monson 2008). While the use of paradigms creates what appears to be a useful organization of the learned rules, recent work questions the learnability of paradigms from realistic input (Chan 2008).

Although the highest performance has traditionally come from segmentation-based approaches, it is difficult to define linguistically reasonable segmentation behavior for even simple cases (*make/mak + ing*), and from the point of view of an acquisition model segmentation suggests a notion of an abstract stem whose psychological and linguistic reality is not obvious (Halle and Marantz, 1993).

This research seeks to build a practical morphological analyzer by adapting a cognitive model that embraces the sparsity seen among morphological forms and learns a linguistically inspired representation. By doing so, we bring computational and cognitive models of morphology learning closer together.

2 Methodology

We use the Base and Transforms Model developed in (Chan, 2008 chap. 5) and extend the accompanying algorithm to create a morphological analyzer. We present a brief summary of the Base and Transforms model here and present our modified version of the algorithm. Our algorithm is summarized in Figure 1.

2.1 The Base and Transforms Model

A morphologically derived word is modeled as a base word with an accompanying transform that changes the base to create a derived form. A base must be a word observed in the input, not an abstract stem, and a transform is an orthographic modification made to a base to create a derived form. It is defined as two affixes

$(s1, s2)$, where $s1$ is removed from the base before concatenating $s2$. Thus to derive *making* from *make* we apply the transform (e, ing), removing $-e$ from the base and then concatenating $-ing$. We represent a null suffix as \$. A transform also has a corresponding word set, which is the set of base-derived pairs that the transform accounts for. The bases of a transform are the only words that the transform can be applied to.

We now give an overview here of the learning algorithm used in this work. For further details on the algorithm's implementation and performance, see (Lignos et al., 2009).

Word Sets. Each word in the corpus belongs to one of three word sets at any point in execution: Base, Derived, or Unmodeled. The Base set contains the words that are used as bases of learned transforms but are not derived from any other form. The derived set contains words that are derived forms of learned transforms, which can also serve as bases for other derived forms. All words begin in the Unmodeled set and are moved into Base or Derived as transforms are learned.

Pre-processing. We perform a minimal amount of pre-processing to support learning on hyphenated words. Any word with a hyphen is placed into a set of words excluded from the learning process, but each segment in the hyphenated word is included in learning. For example, *punk-rock-worshipping* would not be included in learning, but *punk*, *rock*, and *worshipping* would. The analysis of any hyphenated word is the concatenation of the analysis of its segments, in this case *PUNK ROCK WORSHIP +(ing)*.

2.2 The Learning Loop

Affix Ranking. We count the affixes contained in each word in the base and unmodeled sets by brute force, scanning the first and last 5 letters in each word. To prevent rare words and foreign words from affecting the affix and transform ranking process, words only count toward an affix or transform's score if they are relatively frequent in the corpus. For a word to be considered common, it must appear more than once in the corpus and have a frequency greater than one in one million. This frequency cutoff was set by examining the list of words in the Morpho Challenge 2009 evaluation corpora above the cutoff frequency to find a point where less common morphological productions are still included but most typos and foreign words are excluded.

Transform Ranking. We hypothesize transforms of all combinations of the top 50 affixes and count the number of base-derived pairs in each transform. The score of a transform is the number of word pairs it accounts for multiplied by the net number of characters that the transform adds or removes to a base. For example, if the transform $(e,\ ing)$, which removes one letter from the base and adds three, has 50 base-derived pairs, its score would be $50 * |3 - 1| = 100$.

To approximate orthographic gemination and the merging of repeated characters when a morpheme is attached, we relax the conditions of testing whether a base-derived pair is acceptable. For each potential base word for a transform, we compute two derived forms: a standard derived form that is the results of applying the transform precisely to the base, and a "doubled" derived form where $s1$ is removed from the base, the last character of the remaining base is repeated, and then $s2$ is attached. For example, when checking the transform (\$, *ing*) applied to *run*, we generate the standard derived form *runing* and the doubled form *running*. Additionally, in cases where the final character of the base after $s1$ has been removed is the same as the first character of $s2$, we also create an "undoubled" derived form where the first character of $s2$ is removed such that applying the transform does not result in a repeated character. For example, when applying (\$, *ed*) to *bake*, the standard form would be *bakeed*, but the undoubled form would be *baked*. All derived forms that are observed in the Unmodeled set are added, so if the standard, doubled, and undoubled forms are all observed, three base-derived pairs would be added to the transform. These doubling and undoubling effects are most commonly attested in English, but the doubling and undoubling rules are designed to allow the learner to broadly approximate orthographic changes that can occur when morphemes are attached in any language.

Transform Selection. The learner selects the transform of the higest rank that has acceptable segmentation precision. Segmentation precision represents the probability that given any Unmodeled word containing $s2$ reversing the transform in question will result in a word. Segmentation precision must exceed a set threshold for the learner to accept a hypothesized transform. By observing the precision of transforms during development against the Brown corpus, we set a threshold of 1% as the threshold of an acceptable transform. If more than 20 transforms are rejected in an iteration because of unacceptable segmentation precision, the learning loop stops as it is unlikely that there are good transforms left to model.

Transform Word Set Selection. After a transform is selected, we apply the selected transform as broadly as possible by relaxing word sets that the transform's bases and derived words can be members of. This allows our algorithm to handle multi-step derivations, for example to model derivational affixes on an base that is already inflected or allow derived forms to serve as bases for unmodeled words.

This expansion of the permissible types of base/derived pairs requires similar changes to how words are moved between sets once a transform has been selected. We developed the following logic for moving words:

1. No word in Base may be the derived form of another word. If a word pair of the form Base $\rightarrow$ Base is used in the selected transform, the derived word of that pair is moved to Derived. After movement the relationship is of the form Base $\rightarrow$ Derived.

English		
	Trans.	Sample Pair
1	+($, s)	scream/screams
2	+($, ed)	splash/splashed
3	+($, ing)	bond/bonding
4	+($, 's)	office/office's
5	+($, ly)	unlawful/unlawfully
6	+(e, ing)	supervise/supervising
7	+(y, ies)	fishery/fisheries
8	+($, es)	skirmish/skirmishes
9	+($, er)	truck/trucker
10	($, un)+	popular/unpopular
11	+($, y)	risk/risky
12	($, dis)+	credit/discredit
13	($, in)+	appropriate/inappropriate
14	+($, ation)	transform/transformation
15	+($, al)	intention/intentional
16	+(e, tion)	deteriorate/deterioration
17	+(e, ation)	normalize/normalization
18	+(e, y)	subtle/subtly
19	+($, st)	safe/safest
20	($, pre)+	school/preschool
21	+($, ment)	establish/establishment
22	($, inter)+	group/intergroup
23	+(t, ce)	evident/evidence
24	($, se)+	cede/secede
25	+($, a)	helen/helena
26	+(n, st)	lighten/lightest
27	($, be)+	came/became

German		
	Trans.	Sample Pair
1	+($, en)	produktion/produktionen
2	+($, er)	ueberragend/ueberragender
3	+($, es)	einfluss/einflusses
4	+($, s)	gewissen/gewissens
5	+($, ern)	schild/schildern
6	+(r, ern)	klaeger/klaegern
7	($, ver)+	lagerung/verlagerung
8	($, ge)+	fluegel/gefluegel
9	($, ueber)+	nahm/uebernahm
10	($, vor)+	dringlich/vordringlich
11	($, be)+	dachte/bedachte
12	($, unter)+	schaetzt/unterschaetzt
13	($, ein)+	spruch/ einspruch
14	($, er)+	sucht/ersucht
15	($, auf)+	ruf/aufruf
16	($, an)+	treibt/antreibt
17	($, zu)+	teilung/zuteilung
18	($, aus)+	spricht/ausspricht
19	($, ab)+	bruch/abbruch
20	($, ent)+	brannte/entbrannte
21	($, in)+	formiert/informiert
22	+(t, ren)	posiert/posieren
23	+($, lich)	dienst/dienstlich
24	($, un)+	wichtig/unwichtig
25	+(t, rung)	rekrutiert/rekrutierung
26	($, he)+	rauf/herauf

Table 1. Transforms learned in English and German on Morpho Challenge 2009 evaluation data sets

2. A word in Derived may be the base of another word in Derived. If a word pair of the form Derived → Unmodeled is used in the selected transform, the derived word of that pair is moved to Derived, and the base word remains in Derived. After movement the relationship is of the form Derived → Derived.

2.3 Post-processing

Once the learning loop has stopped, the learner tries to break the compound words that remain in the Base and Unmodeled sets using a simple 4-gram character-level model trained on the words in Base. Words are broken at the lowest point of forward probability if the resulting substrings are words seen in the input. For further detail, see (Lignos et al. 2009).

3 Results

3.1 Performance

The learner completes 27 iterations in English and 26 iterations in German before stopping. The resulting analyses achieve an F-measure of 58.48% in English and 33.61% in German in the official Morpho Challenge 2009 competition 1 evaluation, learning the rules presented in Table 1. Among non-baseline methods in competition 1, a comparison against a linguistic gold standard, it achieved the third highest F-measure and highest precision in English, and the 11th highest F-measure and highest precision in German. Among non-baseline methods in competition 2, an information retrieval task, it achieved the highest average precision in English and the 7th highest in German.

3.2 Errors

While it is difficult to assign precise, mutually exclusive categories to the learner's errors, they can be grouped into these categories:

Rare affixes. Many productive affixes in the gold standard are rarer than would be expected in the training corpus, for example the English suffixes $-ness$ and $-able$, and thus the learner fails to distinguish them from noise in the data.

Unproductive affixes. Some affixes in the gold standard are no longer productive in the language being learned. For example, the gold standard suggests that *embark* be analyzed as $em + bark$, but the Germanic prefix $em-$ is not productive in modern English and thus appears in few pairs. It is unlikely that a cognitively oriented learner would learn these rules from the input data.

Multi-step derivations. The learner fails to learn multi-step derivations, for example *acidified* as *ACID +ify +ed*, if any intermediate derivations (*acidify*) are not present in the corpus. These multi-step derivations account for the lower recall of the learner compared to other methods in Morpho Challenge 2009. However, the absence of errors in attempting to generalize rules to analyze these derivations is partly responsible for the learner's high precision.

Spurious relationships. The learner can form word pairs of unrelated words that fit the pattern of common rules, for example *pin/pining* in English. In German, this appears to cause a significant number of errors for even very frequent transforms. In the development set, the three most common transforms in German have a precision of 47.4%, while in English they have a precision of 83.9%.

4 Discussion

4.1 Limitations of the algorithm

By learning individual transforms rather than full paradigms, the learner avoids a major consequence of sparsity in morphology learning. However, the algorithm must observe all steps of a multi-step derivation to learn the connection between the words in the derivation. This limitation has little impact in English, but in languages with more morphemes per word, such as German, this is a limiting factor in the algorithm's performance. With a larger number of morphemes per word, it is unlikely that all permutations of the morphemes would occur with the same base. Segmentation-based approaches have a natural advantage in this area. They need only identify the morphemes and decide whether to apply them to an individual word, unlike our algorithm which identifies rules but requires a minimal pair of words that show a rule's applicability.

While the learner's current approach results in very high precision, it does not match the kind of rule generalization desirable for an acquisition model and results in poorer performance when there are many morphemes per word. In order to address this, the learner must understand the conditions for applying rules. This will require unsupervised part of speech induction so that rules can be marked as inflectional or derivational and using POS to decide whether a rule should be applied. A POS-aware version of the algorithm would likely achieve higher precision as it would not pair words of inappropriate POS together for a given transform. The ability to generalize in this fashion would enable the learner to analyze unseen words, which the learner cannot currently do.

4.2 Limitations of the rule representation

The simple definition of a rule as an affix-change operation limits the languages that the learner can currently be applied to. Languages with vowel harmony, such as Finnish and Turkish, require a more complex and phonologically-specified representation to be accurately modeled using a rule-based approach. Languages that use non-concatenative morphology, such as Arabic and Hebrew, cannot be modeled in any meaningful way using our rule representation, as the algorithm only searches for affix changes and not word-medial changes.

These shortcomings are not inherent to the Base and Transforms model but rather specific to the transform representation used. Expanding the transform definition to support infixes would be a first step to supporting nonconcatenative languages, but operations like vowel harmony and stem changes require a level of phonological information that has thus far not been used in unsupervised morphological analyzers. A more likely approach to handling vowel harmony may be to merge morphemes that appear in similar contexts (Can, 2009).

4.3 Conclusions

The high performance of the learner in English and German suggests that an acquisition model can perform at a comparable level to statistical models. Fu-

ture work should focus on the expansion of acquisition models to support a richer set of morphological phenomena and finer-grained representation of the morphological rules learned.

References

M. R. Brent, S. K. Murthy and A. Lundberg. Discovering morphemic suffixes: A case study in minimum description length induction. In *Proceedings of the Fifth International Workshop on AI and Statistics*, 1995

Can B., Manandhar, S. Unsupervised Learning of Morphology by Using Syntactic Categories. In *Working Notes of the 10th Workshop of the Cross-Language Evaluation Forum*, CLEF 2009, Corfu, Greece, September 30–October 2, 2009.

Chan, E. Structures and distributions in morphology learning. *PhD Thesis*, University of Pennsylvania, 2008

Creutz, M. and Lagus, K. Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. *Publications in Computer and Information Science, Report A81*, Helsinki University of Technology, March 2005

Goldsmith, J. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, MIT Press, 2001

Halle, M. and Marantz, A. Distributed morphology and the pieces of inflection. *The view from Building*, 20:111–176, 1993

Harris, Z.S. From phoneme to morpheme. *Language*, 190–222, 1955

Keshava, S. and Pitler, E. A simpler, intuitive approach to morpheme induction. In *Proceedings of 2nd Pascal Challenges Workshop*, 31–35, 2006

Lignos, C., Chan, E., Marcus M.P., and Yang, C. A Rule-Based Unsupervised Morphology Learning Framework. In *Working Notes of the 10th Workshop of the Cross-Language Evaluation Forum*, CLEF 2009, Corfu, Greece, September 30–October 2, 2009.

Monson, C. ParaMor: from Paradigm Structure to Natural Language Morphology Induction. PhD Thesis, Carnegie Mellon University

Parkes, C.H., Malek, A.M. and Marcus, M.P. Towards Unsupervised Extraction of Verb Paradigms from Large Corpora. In *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, Quebec, Canda, August 15–16, 1998

Pinker, S. *Words and rules: The ingredients of language*. Basic Books, 1999

Rumelhart, D.E. and McClelland, J.L. *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2: Psychological and biological models*. MIT Press, 1986

Spiegler, S., Golnia, B., Flach, P. PROMODES: A probabilistic generative model for word decomposition. In *Working Notes of the 10th Workshop of the Cross-Language Evaluation Forum*, CLEF 2009, Corfu, Greece, September 30–October 2, 2009.

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