Unsupervised morphological segmentation and clustering with document boundaries

Taesun Moon, Katrin Erk, and Jason Baldridge
Department of Linguistics
University of Texas at Austin
1 University Station B5100
Austin, TX 78712-0198 USA
`{tsmoon,katrin.erk,jbaldrid}@mail.utexas.edu`

Abstract

Many approaches to unsupervised morphology acquisition incorporate the frequency of character sequences with respect to each other to identify word stems and affixes. This typically involves heuristic search procedures and calibrating multiple arbitrary thresholds. We present a simple approach that uses no thresholds other than those involved in standard application of χ^2 significance testing. A key part of our approach is using document boundaries to constrain generation of candidate stems and affixes and clustering morphological variants of a given word stem. We evaluate our model on English and the Mayan language Uspanteko; it compares favorably to two benchmark systems which use considerably more complex strategies and rely more on experimentally chosen threshold values.

1 Introduction

Unsupervised morphology acquisition attempts to learn from raw corpora one or more of the following about the *written* morphology of a language: (1) the segmentation of the set of word types in a corpus (Creutz and Lagus, 2007), (2) the clustering of word types in a corpus based on some notion of morphological relatedness (Schone and Jurafsky, 2000), (3) the generation of out-of-vocabulary items which are morphologically related to other word types in the corpus (Yarowsky et al., 2001).

We take a novel approach to segmenting words and clustering morphologically related words. The approach uses no parameters that need to be tuned on data. The two main ideas of the approach are (a) the filtering of affixes by significant co-occurrence, and (b) the integration of knowledge of document boundaries when generating candidate stems and affixes and when clustering morphologically related words. The main application that we envision for our approach is to produce interlinearized glossed texts for under-resourced/endangered languages (Palmer et al., 2009). Thus, we strive to eliminate hand-tuned parameters to enable documentary linguists to use our model as a preprocessing step for their manual analysis of stems and affixes. To require a documentary linguist–who is likely to have little to no knowledge of NLP methods–to tune parameters is unfeasible. Additionally, data-driven exploration of parameter settings is unlikely to be reliable in language documentation since datasets typically are quite small. To be relevant in this context, a model needs to produce useful results out of the box.

Constraining learning by using document boundaries has been used quite effectively in unsupervised word sense disambiguation (Yarowsky, 1995). Many applications in information retrieval are built on the statistical correlation between documents and terms. However, we are unaware of cases where knowledge of document boundaries has been used for unsupervised learning for morphology. The intuition behind our approach is very simple: if two words in a single document are very similar in terms of orthography, then the two words are likely to be related morphologically. We measure how integrating these assumptions into our model at different stages affects performance.

We define a simple pipeline model. After generating candidate stems and affixes (possibly constrained by document boundaries), a χ^2 test based on global corpus counts filters out unlikely affixes. Mutually consistent affix pairs are then clustered to form affix groups. These in turn are used to build morphologically related word clusters, possibly constrained by evidence from co-occurence of word forms in documents. Following Schone and Jurafsky (2000), clusters are evaluated for

668

whether they capture inflectional paradigms using CELEX (Baayen et al., 1993).

We are unaware of other work on morphology using χ^2 tests despite its wide application across many disciplines.[1] This may be due to the large degree of noise found in the candidate affix sets induced through other candidate generation methods. The χ^2 test has two standard thresholds–a significance threshold and a lower bound on observed counts. These are the only manually set parameters we require—and we in fact use the widely accepted standard values for these thresholds without varying them in our experiments. This is a significant improvement over other approaches that typically require a number of arbitrary thresholds and parameters yet provide little intuitive justification for them. (We give examples of these in §3.)

We evaluate our approach on two languages, English and Uspanteko, and compare its performance to two benchmark systems, Morfessor (Creutz and Lagus, 2007) and Linguistica (Goldsmith, 2001). English is commonly used in other studies and permits the use of CELEX as a gold standard for evaluation. Uspanteko is an endangered Mayan language for which we have a set of interlinearized glossed texts (IGT) (Pixabaj et al., 2007; Palmer et al., 2009). IGT provides word-by-word morpheme segmenation, which we use to create a synthetic gold standard. In addition to evaluation against this standard, Telma Kaan Pixabaj—a Mayan linguist who helped create the annotated corpus—reviewed by hand 100 word clusters produced by our system, Morfessor and Linguistica. Note that because English is suffixal and Uspanteko is both prefixal and suffixal, we use a slightly modified model for Uspanteko.

The approach introduced in this paper compares favorably to Linguistica and Morfessor, two models that employ much more complex strategies and rely on experimentally-tuned language/corpus-specific parameters. In our evaluation, document boundary awareness greatly benefits precision for small datasets, blocking acquisition of spurious affixes. For large datasets, global candidate generation outperforms document-aware candidate generation at the task of filtering out spurious stems, but document-aware clustering improves precision. These findings are promising for the application of this approach to under-resourced languages

like Uspanteko.

2 Unsupervised morphology acquisition

Unsupervised morphology acquisition aims to model one or more of three properties of *written* morphology: segmentation, clustering around a common stem, and generation of new word forms with productive affixes. Intuitively, there are straightforward, but non-trivial, challenges that arise when evaluating a model. One large challenge is distinguishing derivational from inflectional morphology. Most approaches deal with tokens without considering context. Since inflectional morphology is virtually always driven by syntax and word context, such approaches are unable to learn only inflectional morphology or only derivational morphology. Even approaches which take context into consideration (Schone and Jurafsky, 2000; Baroni et al., 2002; Freitag, 2005) cannot learn specifically for one or the other.

In addition, the evaluation of both segmentation and clustering involves arbitrary judgment calls. Concerning segmentation, should *altimeter* and *altitude* be one morpheme or two? (The sample English gold standard for MorphoChallenge 2009 provides *alti+meter* but *altitude*.) Similar issues arise when evaluating clusters of related word forms if inflection and derivation are not distinguished. Does *atheism* belong to the same cluster as *theism*? Where is the frequency cutoff point between a productive derivational morpheme and an unproductive one? Yet, many studies have evaluated their segmentations and clusters by going over their results word by word, cluster by cluster and judging by sight whether some segmentation or clustering is good (e.g., Goldsmith (2001)).

Like Schone and Jurafsky (2001), we build clusters that will have both inflectionally and derivationally related stems and evaluate them with respect to a gold standard of *only* inflectionally related stems.

3 Related work

There is a diverse body of existing work on unsupervised morphology acquisition. We summarize previous work, emphasizing some of its more arbitrary and *ad hoc* aspects.

Letter successor variety. Letter successor variety (LSV) models (Hafer and Weiss, 1974; Gaussier, 1999; Bernhard, 2005; Bordag, 2005;

---

[1]Monson (2004) suggests, but does not actually use, χ^2.

Keshava and Pitler, 2005; Hammarström, 2006; Dasgupta and Ng, 2007; Demberg, 2007) use the hypothesis that there is less certainty when predicting the next character at morpheme boundaries. LSV has several issues that require fine parameter tuning. For example, Hafer and Weiss (1974) counts how many types of characters appear after some initial string (the *successor* count) and how many types of characters appear before some final string (the *predecessor* count). A successful criterion for segmenting a word was if the predecessor count for the second part was greater than 17 and the successor count for the first part was greater than 5. Other studies have similar data specific parameters and restrictions.

MDL and Bayesian models. Minimum description length (MDL) models (Goldsmith, 2001; Creutz and Lagus, 2002; Creutz and Lagus, 2004; Goldsmith, 2006; Creutz and Lagus, 2007) try to segment words by maximizing the probability of a training corpus subject to a penalty based on the size of hypothesized morpheme lexicons they build on the basis of the segmentations. While theoretically elegant, a pure implementation on real data results in descriptions that do not reflect actual morphology. Creutz and Lagus (2005) report that, "frequent word forms remain unsplit, whereas rare word forms are excessively split." In the end, every MDL approach uses probabilistically motivated refinements that restrict the tendency of raw MDL to generate descriptions that do not fit linguistic notions of morphology. Despite the sophistication of the models in this group, there are many parameters that need to be set, and heuristic search procedures are crucial for their success (Goldwater, 2007). Snover et al. (2002) present a Bayesian model that uses a prior distribution to refine disjoint clusters of morphologically related words. It disposes with parameter setting by selecting the highest ranking hypothesis.

Context aware approaches. A word's morphology is strongly influenced by its syntactic and semantic context. Schone and Jurafsky (2000) attempts to cluster morphologically related words starting with an unrefined trie search (but with a parameter of minimum possible stem length and an upper bound on potential affix candidates) that is constrained by semantic similarity in a word context vector space. Schone and Jurafsky (2001) builds on this approach, but adds more *ad hoc*

parameters to handle circumfixation. Baroni et al. (2002) takes a similar approach but uses edit distance to cluster words that are similar but do not necessarily share a long, contiguous substring. They remove noise by constraining cluster membership with mutual information derived semantic similarity. Freitag (2005) uses a mutual information derived measure to learn the *syntactic* similarity between words and clusters them. Then he derives finite state machines across words in different clusters and refines them through a graph walk algorithm. This group is the only one to evaluate against CELEX (Schone and Jurafsky, 2000; Schone and Jurafsky, 2001; Freitag, 2005).

Others. Some other models require input such as POS tables and lexicons and use a wider range of information about the corpus (Yarowsky and Wicentowski, 2000; Yarowsky et al., 2001; Chan, 2006). Because of the knowledge dependence of these models, they are able to properly induce inflectional morphology, as opposed to the studies cited above. Snyder and Barzilay (2008) uses a set of aligned phrases across related languages to learn how to segment words with a Bayesian model and is otherwise fully unsupervised.

4 Model[2]

Our goal is to generate *conflation sets*: sets of word types that are related through either inflectional or derivational morphology (Schone and Jurafsky, 2000). Solving this task requires learning how individual types are segmented (though the segmentation itself is not evaluated). For present purposes, we assume that the affixal pattern of the language is known: whether it is prefixal, suffixal, or both. To simplify presentation, we discuss a model that captures suffixes only. Our approach is a four stage process:

1. *Candidate Generation*: generate candidate stems and affixes using an orthographically defined data structure (a trie)

2. *Candidate Filtering*: filter candidate affixes using the statistical significance for pairs of affixes based on their co-occurence counts with shared stems

3. *Affix Clustering*: cluster significant affix pairs into affix groups

---

[2]The code implementing the model is available from `http://comp.ling.utexas.edu/earl`

4. *Word Clustering*: form conflation sets based on affix clusters

The first and last stages are particularly prone to noise, which has necessitated many of the thresholds and heuristics employed in previous work. We hypothesize that naturally occuring document boundaries provide a strong constraint that should reduce this noise, and we test that hypothesis by using it in those stages.

Our intuition comes from an observation by Yarowsky (1995) regarding multiple tokens of words in documents. He tabulates the *applicability* of using document boundaries to disambiguate word senses, which measures how often a given word occurs more than twice in the same document. For ten potentially ambiguous words, he counts how often they occur more than once in some document and finds that if the words do occur, they do so multiple times in 50.1% of these documents, on average. His counts ignored morphological variation, and it is likely the *applicability* measure would have increased considerably: if a content word is used more than once in some text, it is likely to be repeated in different syntactic contexts, requiring the word to be inflected or to be derived for a different part-of-speech category. [3]

For stage one, we build separate tries for each document rather than a trie for the entire corpus. This should reduce the chance that orthographically similar but morphologically unrelated word pairs lead to bad candidates by reducing the search space for words which share a stem to a local document. For example, *assuage* and *assume* are both likely to occur in a large corpus and suggest that there is a stem *assu* with affixes *-age* and *-me*. They are less likely to occur together in many different documents that form the corpus, whereas *assume*, *assumed*, and *assuming* are. We refer to this document constrained candidate generation as *CandGen-D*, and to the unconstrained generation (a single trie for all documents) as *CandGen-G*.

For stage four, documents are used to constrain potential membership of words in clusters: all pairs of words in a cluster must have occured together in some document. We refer to document-constrained clustering as *Clust-D* and the unconstrained global clustering as *Clust-G*.

[3]For example, in just this *one* paragraph we have {*document,documents*}, {*measure, measures*}, {*occur, occurs, occuring*}, and {*word, words*}.

4.1 Candidate generation

Given a document or collection of documents, we use tries (prefix trees) to identify potential stems and affixes and collect statistics for co-occurrences between affixes and between affixes and stems. A trie G, like the example on the right, can be identified with the set of all words on paths from the root to any leaf, in the case of the example figure the set $G = \{abd, ab\$, ac\}$. (We use $\$$ to denote an empty affix.) Given a trie G over alphabet L, we define the set of *trunks* of G
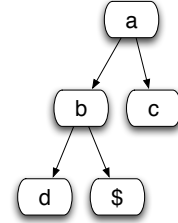

Figure 1

as all paths from the root to a branching point:

$$Tr(G) = \{w \in L^+ \mid \exists a, b \in L, x_1, x_2 \in L^* : a \neq b \wedge wax_1, wbx_2 \in G\}$$

Also, we define the set of *branches* of a trunk $t \in Tr(G)$ as the paths from its branching points to the leaves:

$$Br(t, G) = \{x \in L^+ \mid tx \in G\}$$

In our example, $\{a, ab\}$ are the trunks, with $Br(\text{a}, G) = \{bd, b\$, c\}$ and $Br(\text{ab}, G) = \{d, \$\}$. When we use a trie to induce stems and affixes, all induced stems will be trunks, and all induced affixes will be branches.

From a given trie, we induce a set of *stem candidates* and *affix candidates*. A simple criterion is used: if a trunk is longer than all of its branches, the trunk is a stem candidate and its branches are affix candidates. So, the set of stem candidates for a trie G, *CStem*(G), is the set of trunks $t \in Tr(G)$ such that $|t| > |b|$ for all $b \in Br(t, G)$.

Given a stem candidate $s \in CStem(G)$, its set of affix candidates *CAff*(s, G) is identical to its set of branches. (To talk about the sets of stem and affix candidates for a whole trie G or a set of tries, we write *CAff*(G), $StC(G)$, *CAff*, and *CStem*.) The *count* of an affix candidate $b \in CAff$ is the number of stem candidates with which it occurs:

$$\text{count}(b) = \sum_G |\{s \in CStem(G) \mid b \in CAff(s, G)\}|$$

For Fig. 1, the set of stem candidates is $\{ab\}$ (since some branches of the trunk a are longer than the

trunk itself). The matching set of affix candidates is $CAff(\text{ab}, G) = \{d, \$\}$, each with a count of one.

An *affix rule candidate* is an unordered pair of affix candidates $\{b_1, b_2\}$. It states that any stem occurring with b_1 can also occur with b_2. Affix rules implement the assumption that all productive affixes will cooccur with other productive affixes and that these will form a coherent group. The rule candidates for a given stem candidate $s \in CStem(G)$ are:

$$CRule(s, G) = \big\{ \{b_1, b_2\} \subseteq CAff(s, G) \mid b_1 \neq b_2 \big\}$$

For example, the single stem candidate *ab* in Fig. 1 has one rule candidate, $\{d, \$\}$. We also use $CRule(G)$ for the rule candidates of a trie G across all stems, and $CRule$ for the union of rule candidates in a set of tries.

The count of a rule candidate $r = \{b_1, b_2\}$ in a trie is the number of stem candidates it appears with:

$$\text{count}(r) = \sum_G |\{s \in CStem(G) \mid r \in CRule(s, G)\}|$$

We also use $CAff(s)$ for the set of affix candidates of stem s across several tries, and $CRule(s)$ for the set of rule candidates of a stem s across several tries.

Document-specific versus global candidate generation. *CandGen-D* defines separate tries for every document in the corpus and induces stem, affix and rule candidates for each document. *CandGen-G* instead induces these candidates for a global trie over all the words in the corpus. From the perspective of the formalism laid out above, the only difference is that *CandGen-D* has as many tries G_i as there are documents i and *CandGen-G* has only one G. This simple difference leads to different candidate sets and counts over their occurrences. For example, say two documents contain the pair *putt/putts* and another contains *bogey/bogeys*. With *CandGen-D*, count($\$$)=3, count(s)=3, and count($\$, s$)=2. For the same documents, *CandGen-G* would produce count($\$$)=2 and count(s)=2 since *putt/putts* would have occurred only once in the global trie.

Also, consider a rare pair such as *aardvark/aardvarks* where each word is found in a different document. The pair would be identified by *CandGen-G* but not by *CandGen-D*. The pair would contribute a count of one to count($\$, s$) in *CandGen-G* but not in *CandGen-D*. So, *CandGen-G* can provide better coverage, but it is also more likely to identify noisy candidates, such as *assuage/assumed*, than *CandGen-D*.

4.2 Candidate filtering

The sets of candidates *CStem*, *CAff*, *CRule* is expected to be noisy since the only basis for generating them was strings that share a large portion of their substrings. One way of filtering candidates is to find affix candidates whose co-occurence with other candidates is not statistically significant.

We measure correlation between candidate affixes b_1, b_2 in a candidate rule with the paired χ^2 test. By using χ^2, we only consider pairwise correlation between affixes, rather than attempting global inference. Global consistency of affix sets is not ensured, and as such the approach is susceptible to the multiple comparisons problem. We still opt for this approach for its simplicity and because global inference is problematic due to data sparseness.

Correlation between b_1 and b_2 is determined by the following contingency table:[4]

	b_1	$\sim b_1$
b_2	O_{11}	O_{12}
$\sim b_2$	O_{21}	O_{22}

Based on the significance testing, we define the set of valid rules *PairRule* as those for which the χ^2 test is significant at $p < 0.05$. Thus, affix candidates not significantly correlated with any other affix in *CAff* are discarded.

4.3 Affix clustering

The previous stage produces a set of *pairs* of affixes that are significantly correlated. However, inflectional paradigms rarely contain just two affixes, so we would like to group together affix pairs into larger affix sets to improve generalization. We use a bottom up, minimum distance clustering for valid affix pairs (rules). We do not assume that cluster membership is exclusive. For example, it would not make sense to determine that the null affix *-$\$$* can belong to only one cluster. Therefore, we produce non-disjoint affix clusters.

A valid cluster of affixes is a maximal set of affixes forming pairwise valid rules: *Aff* $\subseteq$ *CAff* is a valid cluster of affixes iff

---

[4] where $O_{11} = \text{count}(\{b_1, b_2\}), O_{12} = \text{count}(b_2) - O_{11}, O_{21} = \text{count}(b_1) - O_{11}, O_{22} = N - O_{11} - O_{12} - O_{21}$ and $N = \sum_{b \in CAff} \text{count}(b)$. See table (1) for examples.

	ed	~ed		le	~le		ed	~ed		le	~le
ing	10273	21853	s	122	132945	ing	2651	1310	s	20	12073
~ing	27120	4119332	~s	936	4044575	~ing	1490	150848	~s	198	144008
(a) $\chi^2 = 352678$			(b) $\chi^2 = 239.132$			(c) $\chi^2 = 65101.6$			(d) $\chi^2 = 0.631, p = 0.427$		

Table 1: Affix counts in contingency tables for the valid pair *ed/ing* and spurious pair *le/s* according to *CandGen-D* in (a) and (b) and according to *CandGen-G* in (c) and (d). χ^2 test values are given under each table. Data is from NYT. Total affix token counts induced through *CandGen-D* and *CandGen-G* are N=4178578 and N=156299, respectively. A total of 2054 and 3739 affix *types* were induced for *CandGen-D* and *CandGen-G*, respectively showing that *CandGen-G* does have better coverage though it might have more noise.

1. $\forall b_1, b_2 \in \textit{Aff} : \{b_1, b_2\} \in \textit{PairRule}$, and

2. If $b \in \textit{CAff}$ with $\forall b' \in \textit{Aff} : \{b, b'\} \in \textit{PairRule}$, then $b \in \textit{Aff}$.

The set of all valid affix clusters is *GroupRule*. This formulation does not rule out the existence of clusters with affixes in common.

4.4 Word clustering

We next cluster word forms into morphologically related groups. Our model assumes two word forms to be morphologically related iff (1) they occurred in the same trie G, (2) they have a trunk s in common that is a stem in $\textit{Stem}(G)$, and (3) their affixes under this stem s are members in a common valid affix cluster in *GroupRule*. Hence a single stem s can be involved in at most $|\textit{GroupRule}|$ conflation sets, one for each valid affix cluster. Again, the only distinction between clustering with a global trie (*Clust-G*) and clustering with several tries from the documents in a corpus (*Clust-D*) is that the former has only one trie.

We define the conflation set for a given stem $s \in \textit{Stem}$ and valid affix cluster $\textit{Aff} \in \textit{GroupRule}$ as

$$\textit{Wd}(s, \textit{Aff}) = \{sb_1, sb_2 \mid b_1, b_2 \in \textit{Aff} \wedge$$
$$\exists G.s \in \textit{Stem}(G) \wedge b_1, b_2 \in \textit{CAff}(s, G)\}$$

One issue that needs clarification is when the candidate generation and clustering stages use different strategies, i.e. the models *CandGen-D* +*Clust-G* and *CandGen-G* +*Clust-D*. This simply means that the *statistics*, and thus the valid *GroupRule*, are derived from either *CandGen-D* or *CandGen-G*.

4.5 Induction for languages that are both prefixal and affixal

The above approach would not fit a language that is prefixal and suffixal. Assuming we have induced separate conflation sets over a prefix trie and a suffix trie, we merge clusters between the two if they have at least one word form in common. Formally, given a set of prefix conflation sets *PCS* and a set of suffix conflation sets *SCS*, the final set of conflation sets *CS* is:

$$CS = \{p \cup s \mid p \in PCS, s \in SCS \wedge p \cap s \neq \emptyset\}$$

5 Data

We apply our method on English and Uspanteko, an endangered Mayan language.

Learning corpora. For English, we use two subsets of the NYTimes portion in the Gigaword corpus which we will call NYT and MINI-NYT. NYT in the current study is the complete collection of articles in the New York Times from June, 2002. NYT has 10K articles, 88K types and 9M tokens. MINI-NYT is a subset of NYT with 190 articles, 15K types and 187K tokens.

The Uspanteko text, USP has 29 distinct texts, 7K types, and 50K tokens. The texts are from OKMA (Pixabaj et al., 2007) and the segmentation and labels of the interlinear glossed text annotations were checked for consistency and cleaned up (Palmer et al., 2009). All counts are for lower-cased, punctuation-removed word forms.

CELEX. The CELEX lexical database (Baayen et al., 1993) has been built for Dutch, English and German and provides detailed entries that list and analyze the morphological properties of words, among other information. Using CELEX, we evaluate on types rather than tokens. The performance of the model is based on how many of the words it judges to be morphologically related overlap with the entries in CELEX. Following previous work (Schone and Jurafsky, 2000; Schone and Jurafsky,

2001; Freitag, 2005), we evaluate on inflectional clusters only, using the CELEX file listing clusters of inflectional variants. [5]

6 Experiments and evaluation

We outline our evaluation methodology, baselines, benchmarks and results, and discuss the results.

6.1 Evaluation metric

Schone and Jurafsky (2000) give definitions for correct ($\mathcal{C}$), inserted ($\mathcal{I}$), and deleted ($\mathcal{D}$) words in model-derived conflation sets in relation to a gold standard. Their formulation does not allow for multiple cluster membership of words. We extend the definition to incorporate this fact about the data. Let w be a word form. We write X_w for the clusters induced by the model that contain w, and Y_w for gold standard clusters containing w. X_w and Y_w only count words which occurred in both model and gold standard clusters. Then

$$\mathcal{C} = \sum_w \sum_{X_w} \sum_{Y_w} (|X_w \cap Y_w|/|Y_w|)$$

$$\mathcal{I} = \sum_w \sum_{X_w} \sum_{Y_w} (|X_w - (X_w \cap Y_w)|/|Y_w|)$$

$$\mathcal{D} = \sum_w \sum_{X_w} \sum_{Y_w} (|Y_w - (X_w \cap Y_w)|/|Y_w|)$$

Based on these definitions, we formulate precision (P), recall (R), and the f-score (F) as: $P = \mathcal{C}/(\mathcal{C}+\mathcal{I})$, $R = \mathcal{C}/(\mathcal{C}+\mathcal{D})$, $F = (2PR)/(P+R)$.

USP evaluation We use two different means to evaluate the performance on USP. One is the f-score derived from the above section with respect to a standard that was automatically generated from the morpheme segment tiers of the OKMA IGT. We generated the standard by taking non-hyphenated segments as the stem and clustering words with shared stems.

We also had an expert in Uspanteko manually evaluate a random subset ($N = 100$) of the model output to compensate for any failings in the standard. The evaluator determined a dominant stem for a cluster and identified words which were not related to that stem. We measured accuracy and
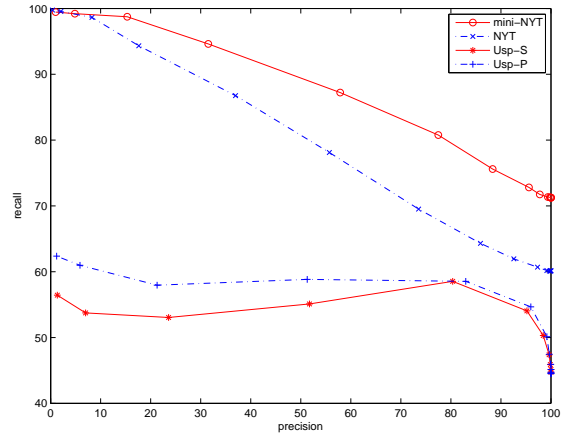

Figure 2: Precision/recall graph for baseline experiments on English, prefix USP (Usp-P) and suffix USP (Usp-S).

full cluster accuracy[6] for the expert evaluations (table 4).

We experimented on Uspanteko with three different assumptions: (1) it is only prefixal; (2) it is only suffixal; (3) it is both prefixal and suffixal. We applied the assumptions of only prefixal or only suffixal to LINGUISTICA as well. The relevant results are given row headers in tables with a corresponding +P(*p*refix) or +S(*s*uffix).

6.2 Baselines and benchmarks

In a *set* of baselines, we put words which share the first k characters into the same cluster. We do this for NYT, MINI-NYT, and USP in a prefix tree, and for USP in suffix tree (using the last k characters). We set the values of $0 < k < \max$, where $\max$ is the length of the longest string, and plot the results in a precision-recall graph (Fig. 2). Low k corresponds to high recall and low precision while high k shows the opposite. The contrast in morphological patterns for each language can also be seen. Because Uspanteko is morphologically complex with suffixes and prefixes, a very simple strategy cannot achieve high recall as opposed to English where it is possible to retrieve all variants with a simple prefix tree.

We use Linguistica (Goldsmith, 2001) and Morfessor (Creutz and Lagus, 2007) as benchmarks. We used the default settings for these programs. Note that comparison with these tools is not com-

---

	MINI-NYT			NYT		
	P	R	F	P	R	F
LINGUISTICA	64.30	**93.34**	76.15	47.50	**88.33**	61.77
MORFESSOR	45.2	87.8	59.7	63.6	69.2	66.3
CandGen-D + Clust-G	69.41	91.42	78.91	46.00	79.81	58.36
CandGen-D + Clust-D	83.47	80.36	81.89	59.02	74.50	65.86
CandGen-G + Clust-G	73.44	88.72	80.36	61.81	82.98	70.85
CandGen-G + Clust-D	**88.34**	77.95	**82.82**	**77.71**	70.24	**73.79**

Table 2: Results on English for all models in precision(P), recall(R), f-score(F) for each data set.

pletely fair. Morfessor only generates segmentations. We therefore processed Morfessor output by clustering words by assuming that the longest segment in any segmentation is the stem and evaluated this instead. Linguistica produces stems and associated suffixes so the clusters naturally follow from this output. However, Linguistica only infers either prefix or suffix patterns.

6.3 Results and discussion

The results on English are in table 2 with χ^2 test criteria of $p<0.05$ and each cell in the contingency table >5. *CandGen-G +Clust-D* had the best f-score, and easily beats the benchmarks.

This is different from our expectation that awareness of document boundaries at all stages (i.e., *CandGen-D +Clust-D*) would show the best results. The discrepancy is especially marked for the larger NYT. One important reason for this is the affix criterion itself: trunks must be longer than branches. Consider again the sample contingency tables in Table 1 that were derived from NYT through *CandGen-D* and *CandGen-G*. We had assumed at the outset that *CandGen-D* would be better able to filter out noise and would be sparser, but results show the opposite. The reason is that that short words in a global lexicon are more likely to share trunks with longer, unrelated words. This ensures that short word forms rarely generate candidate affixes. Longer words which are less likely to have spurious long branches generate the bulk of candidate suffixes and stems. This is born out by the stems that were associated with the spurious suffix pair *le/s*: *CandGen-G* has *cliente, cripp, crumb, daniel, ender, label, mccord, nag, oval, sear, stubb, whipp*. *CandGen-D* has *crumb, hand, need, sing, tab, trick, trip*. The word forms that are associated with *le/s* through the *CandGen-D* strategy are *crumble/crumbs, handle/hands, ...*

Compare this with the word forms associated with the search strategy *CandGen-G* such as *clientele/clientes, cripple/crips, ...* The majority of them are not common English words; they are most probably proper names such as *LaBelle* and *Searle*. Furthermore, there is no item among the stems from the *CandGen-G* search where concatenating the stems *le* and *s* would result in both word forms being a common noun or verb as is the case with the stems from the *CandGen-D* search where all concatenated word forms are common English words. Though *CandGen-G* finds spurious stems, the counts for the spurious affix pair are suppressed (see table 1) because it is a type count rather than a token count. This results in *le/s* being properly excluded as a rule. This explains why *CandGen-D* has worse precision in general than *CandGen-G*.

The affix criterion has other minor issues. One is that it ignores the few cases where stems are shorter than affixes, such as the very common words *be, do, go*.[7] Assuming that the longest productive inflectional suffix in English is *-ing*[8], the criterion would correctly find stem candidates for *-ing* only when the stem is longer than 3 or 4 letters. Another is that the criterion, when combined with *CandGen-D*, generates candidates from *the/them/then/their/these* which cooccur frequently in documents. This is not an issue when the criterion is applied in *CandGen-G*.

Nonetheless, results show that when data sizes are small, as with USP (Table 3) and MINI-NYT, awareness of document boundaries at the candidate generation stage is beneficial to precision.

---

[7]The exclusion of such words in a *token* based evaluation as opposed to a *type* based evaluation would heavily penalize our approach. We are not aware, however, of any prior work in unsupervised morphology that evaluates over tokens.

[8]with occasional gemination of final consonant such as *occur → occurring*

	P	R	F
Ca-D + Cl-D	70.51	44.35	54.45
Ca-G + Cl-G	70.00	46.87	56.15
Ca-D + Cl-D + S	88.58	45.21	59.86
Ca-D + Cl-G + S	85.03	44.75	58.64
Ca-G + Cl-D + S	90.34	45.48	60.50
Ca-G + Cl-G + S	84.54	46.03	59.60
Ca-D + Cl-D + P	93.84	47.90	63.42
Ca-D + Cl-G + P	89.94	47.38	62.06
Ca-G + Cl-D + P	**95.42**	47.89	63.78
Ca-G + Cl-G + P	92.03	50.01	**64.80**
Linguistica + S	81.14	47.60	60.00
Linguistica + P	84.15	52.00	64.28
Morfessor	28.12	**62.28**	38.75

Table 3: Performance of models on automatically generated USP evaluation set. P: Prefix only, S: Suffix only. If there is no indication of S or P, it means model attempted to learn both

	Acc.	FAcc.	Avg. Sz.
Ca-G + Cl-G	98.5	79.0	2.94
Linguistica	96.0	85.0	2.64
Morfessor	85.3	55.0	4.8

Table 4: Human expert evaluated accuracy (Acc.) and full cluster accuracy (FAcc.) of models on USP and average cluster size in words (Avg. Sz.)

However, it seems that *CandGen-G* has better coverage no matter the size of the corpus, which explains why coupling it with *Clust-D* produces overall better scores. *Clust-D* does provide a useful added constraint to mere orthographic similarity (i.e. shared trunks in a trie).

A worrisome aspect of the results is that performance degrades for large data sets (this is also true for Linguistica). However, it also hints that this method might work well for under-resourced languages. We surmise that since productive suffixes do not suffer from sparsity, even a small data set provides sufficient evidence to reach reliable conclusions about the productive morphology of some language. Increasing the size of the data merely increases the counts of spurious affixes and poses problems for a relative simple measure such as the χ^2 test. A similar result was shown in Creutz and Lagus (2005) where f-score performance of their segmentation method improved as more data was provided then decreased as the input exceeded 250K tokens in English. Their method showed continued improvement with increased data for Finnish. This hints that more data is beneficial for morphologically complex languages but not for morphologically impoverished languages.

Finally, it is also encouraging that the manual evaluation (Table 4) shows very high accuracy, as judged by a documentary linguist. Both our model and Linguistica perform very well under this evaluation.

7 Conclusion

We have presented a novel approach to unsupervised morphology acquisition that uses a very simple pipeline and does not use any thresholds other than standard ones associated with the χ^2 test. The model relies on document boundaries and correlation tests for filtering spurious stems and affixes. The model compares favorably to Linguistica and Morfessor, two models that employ much more complex strategies and rely on fine-tuned parameters. We found that the use of document boundaries is especially beneficial with small datasets, which is promising for the application of this model to under-resourced languages. For large datasets, global candidate generation outperformed document-aware candidate generation at the task of filtering out spurious stems, but document-aware clustering does improve precision and overall performance.

In this paper we have addressed one aspect of morphology acquisition, segmentation and clustering. Extending the approach is straightforward, for example, substituting more sophisticated data structures or statistical tests for the current ones. In particular, we will move from the use of document boundaries to a flexible notion of textual distance to estimate likelihood of morphological relatedness.

References

R. H. Baayen, R. Piepenbrock, and H. van Rijn. 1993. *The CELEX lexical database on CD-ROM*. Linguistic Data Consortium, Philadelphia, PA.

M. Baroni, J. Matiasek, and H. Trost. 2002. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *ACL '02 workshop on Morphological and phonological learning*, pages 48–57.

D. Bernhard. 2005. Unsupervised morphological segmentation based on segment predictability and word segments alignment. In *Proceedings of Morpho Challenge 2005*, pages 18–27.

S. Bordag. 2005. Two-step approach to unsupervised morpheme segmentation. In *Proceedings of Morpho Challenge 2005*, pages 23–27.

E. Chan. 2006. Learning Probabilistic Paradigms for Morphology in a Latent Class Model. In *ACL SIGPHON '06*, pages 69–78.

M. Creutz and K. Lagus. 2002. Unsupervised discovery of morphemes. In *ACL '02 workshop on Morphological and phonological learning-Volume 6*, pages 21–30.

M. Creutz and K. Lagus. 2004. Induction of a simple morphology for highly-inflecting languages. In *ACL SIGPHON '04*, pages 43–51.

M. Creutz and K. Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *AKRR '05*, pages 106–113.

M. Creutz and K. Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4(1):3.

S. Dasgupta and V. Ng. 2007. High-performance, language-independent morphological segmentation. In *NAACL-HLT*, pages 155–163.

V. Demberg. 2007. A language-independent unsupervised model for morphological segmentation. In *ACL '07*, volume 45, page 920.

D. Freitag. 2005. Morphology induction from term clusters. In *CoNLL '05*.

E. Gaussier. 1999. Unsupervised learning of derivational morphology from inflectional lexicons. In *ACL workshop on Unsupervised Methods in Natural Language Learning*.

J. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Comp. Ling.*, 27(2):153–198.

J. Goldsmith. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, 12(04):353–371.

S.J. Goldwater. 2007. *Nonparametric Bayesian models of lexical acquisition*. Ph.D. thesis, Brown University.

M.A. Hafer and S.F. Weiss. 1974. Word Segmentation by Letter Successor Varieties. *Information Storage and Retrieval*, 10:371–385.

H. Hammarström. 2006. A naive theory of affixation and an algorithm for extraction. In *ACL SIGPHON '06*, pages 79–88, June.

S. Keshava and E. Pitler. 2005. A simpler, intuitive approach to morpheme induction. In *Proceedings of Morpho Challenge 2005*, pages 28–32.

C. Monson. 2004. A framework for unsupervised natural language morphology induction. In *Proceedings of the Student Workshop at ACL*, volume 4.

Alexis Palmer, Taesun Moon, and Jason Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 36–44, Boulder, CO.

T.C. Pixabaj, M.A. Vicente Méndez, M. Vicente Méndez, and O.A. Damián. 2007. Text collections in Four Mayan Languages. Archived in The Archive of the Indigenous Languages of Latin America.

P. Schone and D. Jurafsky. 2000. Knowledge-free induction of morphology using latent sematic analysis. In *CoNLL-2000 and LLL-2000*.

P. Schone and D. Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *NAACL '01*, pages 1–9.

M.G. Snover, G.E. Jarosz, and M.R. Brent. 2002. Unsupervised learning of morphology using a novel directed search algorithm: taking the first step. In *ACL '02 workshop on Morphological and phonological learning*, pages 11–20.

B. Snyder and R. Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *ACL '08*.

D. Yarowsky and R. Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *ACL '00*, pages 207–216.

D. Yarowsky, G. Ngai, and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT '01*.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL '95*, pages 189–196.