Language Independent, Minimally Supervised Induction of Lexical Probabilities

Silviu Cucerzan and **David Yarowsky**
Department of Computer Science
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218
{silviu,yarowsky}@cs.jhu.edu

Abstract

A central problem in part-of-speech tagging, especially for new languages for which limited annotated resources are available, is estimating the distribution of lexical probabilities for unknown words. This paper introduces a new paradigmatic similarity measure and presents a minimally supervised learning approach combining effective selection and weighting methods based on paradigmatic and contextual similarity measures populated from large quantities of inexpensive raw text data. This approach is highly language independent and requires no modification to the algorithm or implementation to shift between languages such as French and English.

1 Introduction

Part-of-Speech tagging of English has reached a level which seems to resist any improvement. Methods like Transformation-based tagging (Brill, 1995), MaxEnt (Ratnaparkhi, 1996), Boosting (Abney et al., 1999), TnT/Markov models (Brants, 2000) achieve accuracies comparable with human performance for this task.

However, if we break the results into two parts, for known and unknown words, we can see that the performance of English taggers is much lower on the latter. The situation is even worse for languages other than English, especially inflective languages, for two reasons: first, there is usually less annotated data available and second, the coverage of such data is much lower due to the high number of different word-forms in these languages (for comparison of properties and tagging results for several such languages see (Hajič, 2000)). Moreover, many of the words not found in the (small) training data are in fact inflected forms of quite common words. In the work described herein we therefore concentrate on the problem of unknown words in the context of probabilistic tagging.

Although the annotated resources are limited or even non-existent for most languages, the raw text available online is effectively unlimited with respect to the need of most NLP applications. This paper presents a newly developed paradigmatic similarity measure that tries to maximize the benefits that can be obtained from limited annotated resources using a large amount of raw data by magnifying the impact and coverage of the small tagged datasets.

To demonstrate the effectiveness and language independence of the paradigmatic similarity measure in combination with contextual measures, they are evaluated in the context of part-of-speech tagger performance for 4 embedding algorithms using French and English as representatives of both inflective and analytical languages.

2 Problem Description, Motivation and Previous Work

In this paper, we shall use the terms *lexical prior* or *tag prior* for a given word to refer to the probability $P(t|w)$ of Part-of-Speech (POS) tags for word w independent of context, as distinct from what we call the *posterior distribution* $P(t|context; w)$, and also distinct from the concept of *channel model prior* $P(T)$, which refers to the prior probability of a tag sequence T from a generating source.

To facilitate clear exposition, we use here the "direct" lexical probability of tag given word $P(t|w)$, but corresponding arguments hold for the classical HMM Bayesian method (Charniak et al., 1993) used by the taggers we considered for evaluation purpose of the present work.

2.1 Training Data Characteristics with Respect To Unknown Words

Previously unseen (or "unknown") words often represent a significant portion of the vocabulary,

as illustrated in Figure 1 for various vocabulary sizes. Note that for the French training data, the Out-of-Vocabulary (OOV) rate remains relatively high for both tokens (corpus instances of words) and types (vocabulary words), as found in a held-out set of 18,000 tokens (from the French lexically annotated side of the Hansards). The rates are computed ignoring capitalization and normalizing all numbers that appear in the text, so that they are not counted as unknown words.

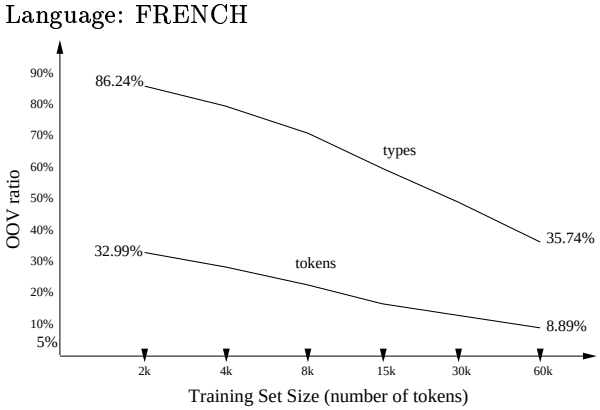

Figure 1: OOV rate as a function of data size (the Hansards)

Figure 2 shows the advantage of using an additional large unannotated corpus. Starting with only the OOV words in the test set relative to the annotated training set of 60,000 tokens, we compute the percentage of these words that are not seen even in the large corpus. Almost 9 out of 10 of the original OOV words do appear in the new (raw) data, which means we can hope to collect additional statistics on them. We still have to use smoothing to estimate tag probabilities for the remaining 12% of the OOV words.

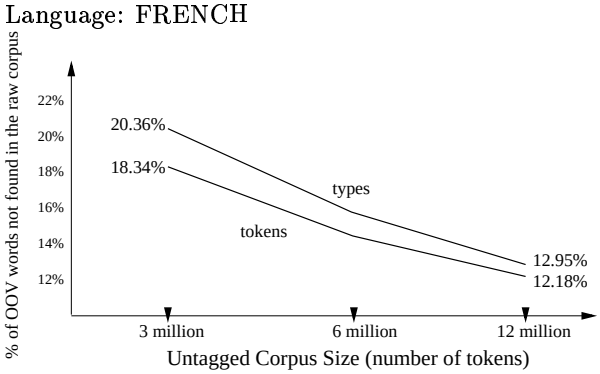

Figure 2: OOV words not seen even in a bigger unannotated text from the same corpus (the Hansards)

2.2 Baseline Universal Lexical Prior Model

As initial baseline, the probability distribution over POS tags for previously unseen words w can be approximated by a single maximum-likelihood tag estimate shared by the full vocabulary:

$$\widehat{P}(t|w) = P(t)$$

A natural refinement is to exclude the most frequent words in the annotated corpus from this frequency distribution computation (see, for example, (Brants, 2000)). When the size of the annotated corpus is not large enough, we can use another unannotated corpus to identify the most frequent words in that language.

2.3 Capitalization

A second baseline model considers $P(t|w)$ to be sensitive to capitalization:

$$\widehat{P}(t|w) = P(t|\ is\_capitalized(w))$$

This can be applied to known words as well, boosting the probability of proper nouns in capitalized contexts and lowering it in the other contexts. More sophisticated models exploiting the capitalization features can be found in (Church, 1988) and (De Marcken, 1990).

2.4 Suffix-based Prior Estimation

The method that seems to work best for unknown words in inflected languages makes use of the suffix analysis of words. Suffix-based handling of unknown words has been proposed in various works (Weischedel et al., 1993; Samuelsson, 1993; Thede, 1998; Hajič, 2000; Brants, 2000).

Fixed-Length-Suffix Priors

For all languages with at least minimal inflective properties, which includes English also, it is possible to use the information obtained from the "suffix" (by which we mean the word-final sequence of letters regardless of whether it belongs to the traditional suffix or ending categories) for a more fine-grained estimation of the tag distribution probabilities for the unknown words. The simplest model considers a fixed-length suffix:

$$\widehat{P}(t|w) = P(t|\ fixed\_length\_suffix(w))$$

Table 1 shows the raw-count distributions observed in an English manually annotated text of 1 million words (from the WSJ corpus) for several words having the suffix *-ate*. The by-token and by-type distributions shown at the bottom of the table have been computed from all the words having the suffix *-ate* in the training data.

Table 1 illustrates the advantage of using a suffix model over the universal lexical prior in

that the observed lexical prior for a fixed suffix (*-ate*) often differs substantially from the pan-vocabulary universal tag probabilities (Table 2).

	VB	VBP	NN	NNP	RB	JJ
calcul*ate*	3	5	0	0	0	0
concentr*ate*	25	2	4	0	0	0
del*icate*	0	0	0	0	0	7
ex*tricate*	2	0	0	0	0	0
fab*ricate*	3	0	0	0	0	0
h*ate*	5	7	1	0	0	0
inaccur*ate*	0	0	0	0	0	6
inadequ*ate*	0	0	0	0	0	4
l*ate*	0	0	0	1	12	6
moder*ate*	1	1	0	0	0	39
priv*ate*	0	0	0	5	0	8
r*ate*	1	2	575	0	0	0
surrog*ate*	0	0	1	0	0	1
Suffix *-ate* prior (by token)	.16	.04	.39	.13	.02	.26
Suffix *-ate* prior (by type)	.40	.07	.16	.09	.00	.28
intricate	?	?	?	?	?	?

Table 1: Some examples of words ending in *-ate* from a 1-million-word tagged English corpus and the lexical priors for suffix *-ate* as an estimation for unknown word *intricate* (longest suffix match emphasized)

	VB	VBP	NN	NNP	RB	JJ	Others
Univ. prior (by token)	.033	.015	.163	.115	.038	.075	.561
Univ. prior (by type)	.030	.007	.174	.246	.000	.164	.389

Table 2: Universal priors for 6 POS tags computed over a 1-million-word annotated English corpus

On the other hand, Table 1 also illustrates two problems with suffix-based estimation for part-of-speech priors:

While a previously unseen word such as *intricate* is primarily an adjective, the dominant part-of-speech for the fixed-length suffix *-ate* is NN (using token-weighted estimation), or VB (using type-weighted estimation).

Modeling longer suffixes just makes things worse in this case, as 14 out of 15 of the words ending in *-icate* in the tagged corpus are exclusively VB or VBP, and the two forms ending in *-ricate* (*extricate* and *fabricate*) in the training text are also exclusively tagged as VB. Suffixes clearly do not capture all relevant information in predicting tag probabilities for unknown words.

Linear Interpolation of Fixed-Length Suffix Models

As a third baseline we considered an interpolated suffix model to demonstrate the relative effectiveness of these approaches when restricted to estimation by smoothed fixed length suffix models. We used the interpolation of 3 fixed-length suffix priors:

$$\widehat{P}(t|w) = \sum_{j=1..3} \lambda_j \cdot P(t \mid suf_j(w))$$

where $suf_j(w)$ denotes the suffix of length j of word w.

Variable-Length Suffix Models

Given that the length of informative suffix context varies considerably across suffixes, our fourth and final baseline model uses a smoothed trie-based architecture (similar to the one presented in (Cucerzan and Yarowsky, 1999) for named entity classification) to estimate

$$\widehat{P}(t|w) = \sum_{j \geq 1} \lambda(j, suf_j(w)) \cdot \sum_{\substack{v \ known \\ suf_j(v) = suf_j(w)}} P(t|v) \quad (1)$$

In some cases, this variable-length suffix method may have the opposite problem of fixed-length method, over-training by giving too much weight to the properties of the morphological form most similar to the given word w encountered in the training text (in the *-ate* example, the estimation becomes worse as we considered longer and longer suffixes). Still, our experiments show that variable-length outperforms the fixed-length interpolation models.

However, many words with similar internal suffixes are misleading indicators of the lexical priors for unseen words. Our goal, therefore, is to borrow lexical probability estimates from a more predictive set of previously seen exemplars.

The following sections propose such methods.

3 Similarity Measures

Recall that the central task of lexical prior estimation is determining how much weight each previously-seen exemplar's distribution should contribute to an unknown word's distribution. Rewriting formula (1) in the following equivalent form:

$$\widehat{P}(t|w) = \sum_{v} P(t|v) \cdot \lambda'(lcs(v,w)) \quad (2)$$

where $lcs(\cdot,\cdot)$ represents the longest common suffix of two words, and $\lambda'(suf_k(w)) = \lambda(1, suf_1(w)) + ... + \lambda(k, suf_k(w))$, observe that this is merely a special case of a more general representation:

$$\widehat{P}(t|w) = \mu \sum_{v} P(t|v) \cdot \mathrm{sim}(w,v) \quad (3)$$

where sim(w,v) can be any weighting of potential exemplars v for a target word w (μ is a normalization factor) .

But what should this similarity measure take into account?

0	Distributions at 0 position $f(w,0)$						
$S(w,0)$	ε	nt	r	ra	s	l	other 8
centre	.713	.0002	.001	.0002	.284	-	-
structu*re*	.751	.001	.017	-	.229	.001	-
mo*ntre*	.331	.122	.458	.033	.007	-	.049

-1	Distributions over suffixes at 1st position $f(w,1)$										
$S(w,1)$	ais	al	ale	aux	ent	e	er	era	es	ation	el	other 20
centr	.0001	.099	.167	.003	.0002	.521	.001	.0001	.208	-	-	-
structur	-	.023	.006	.004	.0008	.789	.014	-	.158	.002	.0008	-
montr	-	-	-	-	.087	.238	.328	.024	.005	-	-	.318

-2	Distributions over suffixes at 2nd position $f(w,2)$																
$S(w,2)$	aine	ime	rais	ral	rale	raux	re	rent	rer	rera	res	rel	urion	rons	ré	rés	other 40
cent	.012	.002	.0001	.072	.124	.003	.627	.0001	.001	.0001	.154	-	.003	-	-	-	-
structu	-	-	-	.024	.006	.005	.793	.0008	.014	-	.156	.001	-	-	-	-	-
mont	-	-	-	-	-	-	.079	.029	.109	.008	.002	-	-	.018	.058	.010	.654

	POS estimate	
	as noun	as verb
centre	.99+	.01-
structu*re*	.99+	.01-
mo*ntre*	.01-	.99+

Table 3: Suffix distribution for *centre*, *structure*, and *montre* as observed in a 12-million-word French corpus

3.1 Suffix-based Paradigmatic Distance

The primary intuition behind the following paradigmatic distance measure is that words which have similar probabilistic distributions of added suffixes will also tend to have similar part-of-speech tag distributions.

Consider the French words *centre* and *structure*, which can be both singular nouns and 1P/3P-singular-present verbs, with the noun usage significantly more common for both words. While their internal suffixes differ at the 3rd position (*-tre* vs. *-ure*), both words exhibit a very similar distribution of observed added suffixes (shown in Table 3). Both are dominated by the noun-consistent signatures ε and $+s$, with a much smaller distribution over the verb-consistent signatures $+nt$, $+\ r$ and $+ra$. In contrast, the word *montre* (almost exclusively a 1P/3P-sing-present verb), while exhibiting superficial internal suffix similarity to *centre*, exhibits a very different added suffix distribution. The divergence is further illustrated by considering added suffix distributions starting at 1-character and 2-character truncated forms of the target words (e.g. *structur* and *structu*).

This distinction is important because *centre* was never observed with a part-of-speech tag in the selected 60k annotated text, and its tag distribution needs to be estimated as an unknown word. Traditional internal suffix-based models would base this estimate on the more orthographically similar *montre*, which is seen in the small tagged corpus only as a verb, as well as other orthographically similar words such as *concentre* (seen as verb), *contre* (preposition), and *rencontre* (encountered as both verb and noun), yielding the misleading conclusion that *centre* is predominantly a verb. In contrast, *structure*, which is paradigmatically the most similar word present in the small tagged corpus, occurs there exclusively as a noun, and is thus a much better tag-distribution exemplar for *centre*, which is also overwhelmingly a noun.

Formally, let V be a vocabulary extracted from an unannotated corpus C over a language L and *Suff* the set of possible suffixes for that language, extracted also from corpus C by considering all the suffixes s for which there exists a certain number (dependent on the language and corpus considered) of distinct words w in V such that the concatenations ws are also in the vocabulary. For the studied languages we limit the length of suffixes to 5 letters and we define the extended sets of valid suffixes as $\overline{Suff_i} \doteq \{xs \mid |x| \leq i,\ |xs| < 5,\ s \in Suff,\ \exists y_1, ..., y_t$ distinct strings such that $y_j xs \in V$ for $j \in 1..T\}$, where T is a language and corpus dependent threshold. Variations of this extensions can be considered for languages with special inflectional properties (such as *umlant* in German).

To build the suffix families $S(w,i)$, we consider all the vocabulary entries that can be obtained from the word w by stripping the last i letters and adding a valid suffix from $\overline{Suff_i}$:

$$S(w,i) \doteq \{s \in \overline{Suff_i} \mid w_1 w_2 ... w_{n-i} s \in V\}$$

The word break in front of the last i letters will be called *the* ith *position*.

The distribution functions $f(w,i) : S(w,i) \rightarrow (0,1]$ are obtained by counting the occurrences of the vocabulary entries $w_1 ... w_{n-i} s$ in C and normalizing the counts.

The motivation for considering suffixation distributions from multiple word positions is that the suffix families at the 0th position can often be sparse and misleading, particularly for inflected or rarely encountered words. For example, the similar part-of-speech behavior for the English *achiever* and *retriever* (Table 4) is not sufficiently evident from the distributions at the 0th position alone, due to the low frequency of the word form *achiever*. Also, adjectives and nouns ending in *-y* may have similar suffix families at the 0th position $\{\varepsilon\}$ (e.g. *creepy* $+ \{\varepsilon\}$ vs. *philantropy* $+ \{\varepsilon\}$), but the suffix families at the 1st position capture different "nominal" and "adjectival" properties, mak-

0	$f(w,0)$	
$S(w,0)$	ε	s
retriever	.941	.059
achiever	.5	.5

-1	$f(w,1)$						
$S(w,1)$	ε	d	r	s	rs	able	ment
retrieve	.580	.351	.059	.004	.006	-	-
achieve	.461	.391	.05	.003	.05	.0005	.135

-2	$f(w,2)$									
$S(w,2)$	able	al	als	e	ed	er	ers	es	ing	eable	ement
retriev	.002	.072	.003	.470	.284	.048	.003	.005	.112	-	-
achiev	.006	-	-	.405	.342	.004	.004	.009	.115	.0005	.111

Table 4: Suffix distribution for *retriever* (32 occurrences) and *achiever* (2 occurrences.) in an 80-million-word English corpus

ing the distinction between the two classes clean and visible (e.g. $creep + \{\varepsilon, iest, ily, ing, s, y\}$ vs. $philantrop + \{ical, ies, ist, ists, y\}$, as observed in the considered untagged corpus). It is thus more robust to also include suffixes distributions over several truncated forms as well.

It was determined experimentally that the distributions at positions greater than 3 and the ones obtained for words shorter than 4 letters are not useful. This does not represent a major problem because unknown words tend to have long forms in most languages.

Various distance measures (cosine similarity, Euclidean distance, L_1 norm) and interpolation methods were used in our experiments to determine the most suitable formula for the paradigmatic distance.

The best scores were obtained for L_1 norm using a weighted product combination $dist(w,v) = \prod_{i=0}^{l(|w|,|v|)}(\beta_i + dist(w,v,i))$ and a Jaccard-type (Salton and McGill, 1983) alteration to penalize the cases in which major differences in the underlying suffix families (not only in the distributions) are found:

$$dist(w,v,i) \doteq \sum_{s \in S(w,i) \cap S(v,i)} |f(w,i;s) - f(v,i;s)| + \delta(w,v,i) \sum_{s \in S(w,i) \Delta (S(v,i)} |f(w,i;s) - f(v,i;s)|$$

Based on the paradigmatic distance computed in this way, it is possible to filter out the words with similar endings but occurring with different suffix families and distributions. Furthermore, this filter has the advantage of being trained on completely untagged corpora, a potentially unlimited resource.

Should a word not appear even in the large raw text corpus, some smoothing technique based only on suffix similarity would still be needed (such as fixed or variable length suffix interpolation).

3.2 Contextual Similarity

As a complement to the suffix-based paradigmatic distance proposed in this paper, a word-context-based similarity measure has been shown to be useful for tagging unknown words. Brill (1995) utilized word context neighborhoods to model and predict tags for unknown words. Schütze (1993) explicitly formulated the concept of paradigmatic similarity over nearby word contexts, using this in an SVD framework for part-of-speech tagging.

We also utilized this relatively orthogonal information source as a complement to the proposed suffix-based paradigmatic distance. We chose unigram vectors to model left and right neighborhoods, and used cosine similarity for its robustness. Because cosine similarity over numerous large-vocabulary contexts can be very expensive to compute, we only incorporated this measure when the suffix-based paradigmatic distance measure was within a certain threshold of viability.

v	Tag Prior Distribution						sim(intricate, *v*)	
	VB	VBP	NN	NNP	RB	JJ	Paradigm. Distance	Contextual Similarity
inaccur*ate*	0	0	0	0	0	6	0.040	0.016
inadequ*ate*	0	0	0	0	0	4	0.040	0.062
del*icate*	0	0	0	0	0	7	0.154	0.141
surrog*ate*	0	0	1	0	0	1	2.258	0**
moder*ate*	1	1	0	0	0	39	2.849	0.015
priv*ate*	0	0	0	5	0	8	2.957	0.085
calcul*ate*	3	5	0	0	0	0	9.118	0.029
ex*tricate*	2	0	0	0	0	0	23.272	0.0006
concentr*ate*	25	2	4	0	0	0	26.694	0.017
fab*ricate*	3	0	0	0	0	0	34.097	0.0004
r*ate*	1	2	575	0	0	0	107.809	0.075
l*ate*	0	0	0	1	12	6	114.420	0.294
h*ate*	5	7	1	0	0	0	122.503	0.039
intricate	?	?	?	?	?	?	*computed from lines 1-3	
Paradigmatic distribution*	0	0	0	0	0	1	**values on lines 4-13 shown for comparison	

Table 5: The words from Table 1 are ordered by suffix-paradigmatic distance with respect to the target word *intricate*. Both the paradigmatic and contextual measures were computed from an 80-million-word unannotated corpus.

3.3 Using the Similarity Measures

Table 5 illustrates the application of both the suffixed-based paradigmatic distance and contextual similarity measures to predicting the lexical prior distribution for the previously unseen English word *intricate*. The potential exemplar candidates, such as shown in Table 1, are ordered by the paradigmatic distance measure, filtered by the more expensive and less effective context similarity scores as noted above.

We investigated several weighting functions for computing the consensus distribution from this space. While using just the single closest exemplar's distribution performed surprisingly well, the best performance was obtained by a uniform weighting of the distributions from exemplars within an experimentally determined distance threshold. Ongoing work is considering word length and word frequency similarity as further potential components of this weighting function.

4 Embedding Algorithm

Since we obtain a tag probability distribution for any unknown word, it is quite straightforward to use this distribution in the context of any probabilistic tagger, including the standard HMM n-gram taggers. In this study, we use bigrams as the base model, since we are dealing with a relatively limited training data.

We contrasted four search algorithms: (a) a classical beam-1 search (Beam 1); (b) a (tag|left-history,right-history) combination of forward and backward beam-1 searches (L-R beam 1), variation suggested by the high complementary rates as defined in (Brill and Wu, 1998) - values in the 20-40% range; (c) a full Viterbi search; (d) an adjusted variation of the latter that uses (tag|left-tag,right-tag) trigrams for a correction pass (L-R Viterbi).

It should be noted that our method of estimating lexical tag priors can be used in other tagging paradigms, such as a maximum entropy tagger (Ratnaparkhi, 1996), as well as non-probabilistic taggers, such as the Brill's rule-based tagger (Brill, 1995), by initializing the tagger with a tag candidate set for every unknown word based on the lexical prior estimates.

5 Evaluation

We have tested the new methods on two languages, French and English, using only small amounts of annotated text for training (60k max. for French, 200k max. for English) and relatively large unannotated corpora (on the order of tens of million words) for computing the paradigmatic distance and contextual similarity. All parameters of the embedding methods were estimated based on a French development set and used unmodified for English, further emphasizing the relatively language independent usage of the algorithm but also partially explaining the lower boost on performance on English.

Table 6 presents the results obtained by the different methods for handling unknown words into the L-R taggers. The Paradigmatic-1 row represents the variation in which only the first paradigmatically similar word found is used, while Paradigmatic-n denotes the combination of up to n most similar words as estimators. As mentioned previously, such words may not always be found, therefore the suffix-based smoothing scheme is used for back-off in these cases. The results were obtained on a test set of 18k tokens from the French side of the Hansards using two different training-set sizes, 15k tokens (average OOV ratio 17.3%) and 60k tokens (OOV ratio 8.9%), and an unannotated text of 12 million words from the same corpus. The first four rows present re-implementations of standard methods; the **boldface-typed** methods use the new paradigmatic distance proposed here (Section 3.1). VLS method uses a probabilistic trie-suffix model.

Table 7 summarizes the consistent improvement achieved by the addition of the suffix-paradigmatic and contextual models to various bigram taggers. The results obtained for Brill's algorithm, trained using the same data (15k/60k words annotated corpora, 12 million words unannotated corpus), are also presented, in conjunction with the improvement in accuracy gained by the same algorithm when every unknown word in the test sets is replaced with the paradigmatically most similar known word from the training sets.

Table 8 presents the results obtained for English on a contiguously selected test set from the WSJ corpus, using contiguous training sets from different regions of the same corpus. Numbers and capitalization variance were not treated as unknown words in evaluation given their ease of POS prediction. These results also show good improvement relative to the baseline performance for the same embedding algorithms.

Using our proposed method for predicting the tag distributions for previously unseen words consistently improves the results for a wide range of training set sizes as well, as illustrated here in Figures 3 and 4 using 2 different embedding algorithms on French data. The one exception to this trend is observed for only the smallest training set size of 2k words for the L-R Viterbi tagger. In this particular case, the space in which paradigmatically similar words have to be searched is very limited and trie interpolation method used as back-off in the case such words are not found gives excessive weight to tiny available sets of tagged exemplars, a problem that could be addressed through more conservative trie smoothing techniques.

Language: FRENCH

Evaluation Type	Full Performance				Accuracy on OOV tokens			
Bigram-based Algorithm:	L-R beam 1		L-R Viterbi		L-R beam 1		L-R Viterbi	
Lexical Prior Model	15k	60k	15k	60k	15k	60k	15k	60k
Universal Prior	89.48	93.76	90.86	94.42	52.36	53.06	57.57	56.00
Capitalization Only	90.50	94.74	91.92	95.28	57.95	63.09	62.61	63.85
Interpolated Suffix (IS)	92.37	95.43	93.26	95.69	68.16	68.96	70.41	68.71
Interpolated Suffix + Cap.	92.72	95.83	94.12	**96.31**	69.59	72.73	75.51	**75.60**
Paradigmatic-1 + IS	93.88	96.50	94.30	96.84	74.63	77.80	77.63	80.95
Paradigmatic-n + IS	94.13	96.73	94.45	97.04	75.65	79.79	79.02	83.07
Variable Length Suffix (VLS)	94.34	96.56	94.91	**97.08**	78.58	79.98	80.72	**83.19**
Paradigmatic-n + VLS	94.55	96.86	94.88	97.23	77.94	81.83	80.89	84.67
ParDist-n + Context + VLS	94.74	96.96	94.99	**97.31**	78.79	82.17	81.64	**85.44**

Table 6: Performance of 9 various unknown word prior estimation methods

Language: FRENCH - 15k words training

	Beam 1	L-R beam 1	Viterbi	L-R Viterbi
Universal Lex. Prior	86.54	89.48	90.30	90.86
Interp. Suffix + Cap.	90.91	92.72	93.29	94.12
Full Lex. Prior Model	93.35	94.74	94.14	94.99

Brill (standard)	Brill + ParDist-1
93.84	93.96

Language: FRENCH - 60k words training

	Beam 1	L-R beam 1	Viterbi	L-R Viterbi
Universal Lex. Prior	91.33	93.76	94.33	94.42
Interp. Suffix + Cap.	94.30	95.83	96.08	96.31
Full Lex. Prior Model	95.71	96.96	97.06	**97.31**

Brill (standard)	Brill + ParDist-1
96.98	**97.14**

Table 7: Performance of 3 prior estimation methods when used in different embedding algorithms

Language: FRENCH

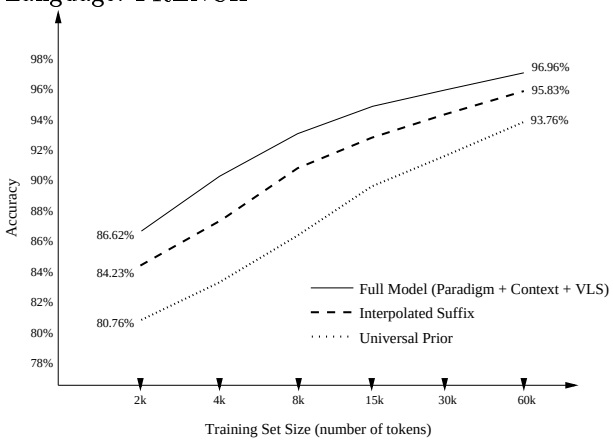

Figure 3: Performance of 3 prior estimation methods in L-R Beam 1 Tagger using various size training sets

Language: FRENCH

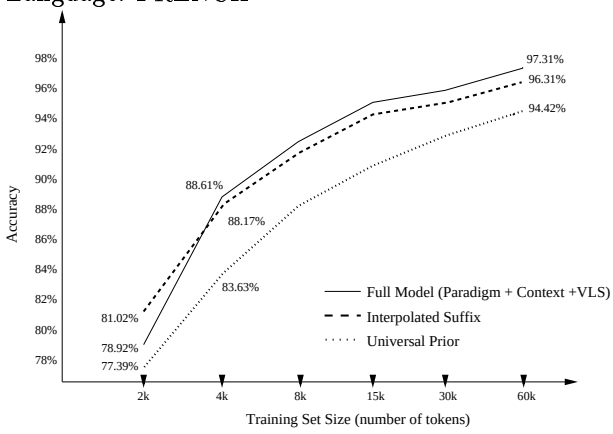

Figure 4: Performance of 3 prior estimation methods in L-R Viterbi Tagger using various size training sets

6 Conclusion

This paper has presented a novel, efficient and effective method for estimating the lexical tag probability distributions for a language when only limited annotated training data is available. The method outperforms a set of 3 different traditional suffix-based estimators, including hierarchically smoothed suffix trie models, by identifying more highly predictive tag exemplars through a combination of paradigmatic and contextual similarity measures. Each of these models uses associations and distributional similarities observed in large quantities of raw text to compensate for limited quantities of tagged training data, and each is language independent to the extent that no modification is required to shift applications from French to English, or other suffix inflective languages. Use of these novel lexical probability estimation methods achieves a 27% error

Language: ENGLISH

Evaluation Type	Full Performance				Accuracy on OOV tokens			
Bigram-based Algorithm:	L-R beam 1		L-R Viterbi		L-R beam 1		L-R Viterbi	
Lexical Prior Model	50k	200k	50k	200k	50k	200k	50k	200k
Universal Prior	85.99	91.22	88.81	92.71	25.06	31.10	39.49	43.03
Interp. Suffix + Cap.	91.35	93.43	93.00	94.73	74.26	75.09	77.39	79.17
Trie VLS	91.26	93.55	92.97	94.98	74.03	75.69	77.17	79.53
Full Model	91.58	94.04	93.31	**95.36**	74.98	77.20	78.09	**80.89**

Table 8: Performance of 4 lexical prior estimation methods on reduced size sets from WSJ Corpus

rate reduction in full Viterbi tagger performance for French over an interpolated-suffix model baseline, and 12% error rate reduction for equivalent full tagger performance on English. When compared with a state-of-the-art model for hierarchically smoothed variable-length suffix tries, the addition of the paradigmatic and contextual distance measures achieves a 7.8% error rate reduction for French and 7.6% error reduction on English. Performance shows a consistent improvement across 4 different embedding tagging algorithms.

Further studies are in progress to compare the usefulness of these techniques on low-count (rather than unseen) words, and also to extend this work to Romanian, Czech and Slovenian, as further examples of highly inflected languages. Evidence from shifting applications from French to English indicates that respectable performance can be obtained without even the re-estimation of parameters on new languages, although we do expect that some parameter re-optimization could prove useful. We believe that this approach should show the greatest benefits for taggers designed for highly inflective languages, such as (Hajič and Hladká, 1998) and (Erjavec et al., 1999), given that the associational power and potential for the proposed paradigmatic similarity measure are most compelling for such languages.

7 Acknowledgements

The authors would like to thank Jan Hajič for his extremely valuable suggestions and feedback on this work.

References

S. Abney, R.E. Schapire, and Y. Singer. 1999. Boosting applied to tagging and PP attachment. *Proceedings of EMNLP/VLC 1999*, pages 38–45.

T. Brants. 2000. TnT - a statistical part-of-speech tagger. *Proceedings of ANLP 2000*, pages 224–231.

E. Brill and J. Wu. 1998. Classifier combination for improved lexical disambiguation. *Proceeding of COLING-ACL 1998*, pages 191–195.

E. Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics, 21(4)*, pages 543–565.

E. Charniak, C. Hendrickson, N. Jacobson, and M. Perkowitz. 1993. Equations for part-of-speech tagging. *Proceedings of the 11th National Conference on Artificial Intelligence 1993*, pages 784–789.

K. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the 2nd Conference on Applied Natural Language Processing 1988*, pages 136–143

S. Cucerzan and D. Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence. *Proceedings of EMNLP/VLC 1999*, pages 90–99.

C. G. de Marcken 1990 Parsing the LOB corpus. *Proceedings of ACL 1990*, pages 243–251.

T. Erjavec, S. Dzeroski, and J. Zavrel. 1999. Morphosyntactic tagging of slovene: Evaluating POS taggers and tagsets. Technical report, Dept. of Intelligent Systems, Jozef Stefan Institute, Ljubljana.

J. Hajič and B. Hladká. 1998. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. *Proceeding of COLING-ACL 1998*, pages 483–490.

J. Hajič. 2000. Morphological tagging: data vs. dictionaries. *Proceedings of NAACL 2000*, pages 94-101.

A. Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. *Proceedings of EMNLP 1996*, pages 133–142.

G. Salton and M.J. McGill. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill.

C. Samuelsson. 1993. Morphological tagging based entirely on Bayesian inference. *9th Nordic Conference on Computational Liguistics 1993.*

H. Schütze. 1993. Part-of-speech induction from scratch. *Proceedings of ACL 1993*, pages 251–258.

S.M. Thede. 1998. Predicting part-of-speech information about unknown words using statistical methods. *Proceeding of COLING-ACL 1998*, pages 1505–1507.

R. Weischedel, M. Meeter, R. Schwartz, L. Ramshaw, and J. Palmucci. 1993. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics, 19(3)*, pages 359–382.