# EM Can Find Pretty Good HMM POS-Taggers
# (When Given a Good Start)[*]

**Yoav Goldberg** and **Meni Adler** and **Michael Elhadad**
Ben Gurion University of the Negev
Department of Computer Science
POB 653 Be'er Sheva, 84105, Israel
{yoavg,adlerm,elhadad}@cs.bgu.ac.il

## Abstract

We address the task of unsupervised POS tagging. We demonstrate that good results can be obtained using the robust EM-HMM learner when provided with good initial conditions, even with incomplete dictionaries. We present a family of algorithms to compute effective initial estimations $p(t|w)$. We test the method on the task of full morphological disambiguation in Hebrew achieving an error reduction of 25% over a strong uniform distribution baseline. We also test the same method on the standard WSJ unsupervised POS tagging task and obtain results competitive with recent state-of-the-art methods, while using simple and efficient learning methods.

## 1 Introduction

The task of unsupervised (or semi-supervised) part-of-speech (POS) tagging is the following: given a dictionary mapping words in a language to their possible POS, and large quantities of unlabeled text data, learn to predict the correct part of speech for a given word in context. The only supervision given to the learning process is the dictionary, which in a realistic scenario, contains only part of the word types observed in the corpus to be tagged.

Unsupervised POS tagging has been traditionally approached with relative success (Merialdo, 1994; Kupiec, 1992) by HMM-based generative models, employing EM parameters estimation using the Baum-Welch algorithm. However, as recently noted

by Banko and Moore (2004), these works made use of filtered dictionaries: dictionaries in which only relatively probable analyses of a given word are preserved. This kind of filtering requires serious supervision: in theory, an expert is needed to go over the dictionary elements and filter out unlikely analyses. In practice, counts from an annotated corpus have been traditionally used to perform the filtering. Furthermore, these methods require rather comprehensive dictionaries in order to perform well.

In recent work, researchers try to address these deficiencies by using dictionaries with unfiltered POS-tags, and testing the methods on "diluted dictionaries" – in which many of the lexical entries are missing (Smith and Eisner, 2005) (SE), (Goldwater and Griffiths, 2007) (GG), (Toutanova and Johnson, 2008) (TJ).

All the work mentioned above focuses on unsupervised **English** POS tagging. The dictionaries are all derived from tagged English corpora (all recent work uses the WSJ corpus). As such, the setting of the research is artificial: there is no reason to perform unsupervised learning when an annotated corpus is available. The problem is rather approached as a workbench for exploring new learning methods. The result is a series of creative algorithms, that have steadily improved results on the same dataset: unsupervised CRF training using contrastive estimation (SE), a fully-bayesian HMM model that jointly performs clustering and sequence learning (GG), and a Bayesian LDA-based model using only observed context features to predict tag words (TJ). These sophisticated learning algorithms all outperform the traditional baseline of EM-HMM based methods,

while relying on similar knowledge: the lexical context of the words to be tagged and their letter structure (*e.g.*, presence of suffixes, capitalization and hyphenation).[1]

Our motivation for tackling unsupervised POS tagging is different: we are interested in developing a Hebrew POS tagger. We have access to a good Hebrew lexicon (and a morphological analyzer), and a fair amount of unlabeled training data, but hardly any annotated corpora. We actually report results on full morphological disambiguation for Hebrew, a task similar but more challenging than POS tagging: we deal with a tagset much larger than English (over 3,561 distinct tags) and an ambiguity level of about 2.7 per token as opposed to 1.4 for English. Instead of inventing a new learning framework, we go back to the traditional EM trained HMMs. We argue that the key challenge to learning an effective model is to define good enough initial conditions. Given sufficiently good initial conditions, EM trained models can yield highly competitive results. Such models have other benefits as well: they are simple, robust, and computationally more attractive.

In this paper, we concentrate on methods for deriving sufficiently good initial conditions for EM-HMM learning. Our method for learning initial conditions for the $p(t|w)$ distributions relies on a mixture of language specific models: a paradigmatic model of similar words (where similar words are words with similar inflection patterns), simple syntagmatic constraints (*e.g.*, the sequence V-V is extremely rare in English). These are complemented by a linear lexical context model. Such models are simple to build and test.

We present results for unsupervised PoS tagging of Hebrew text and for the common WSJ English test sets. We show that our method achieves state-of-the-art results for the English setting, even with a relatively small dictionary. Furthermore, while recent work report results on a reduced English tagset of 17 PoS tags, we also present results for the complete 45 tags tagset of the WSJ corpus. This considerably raises the bar of the EM-HMM baseline. We also report state-of-the-art results for Hebrew full mor-

---

[1] Another notable work, though within a slightly different framework, is the prototype-driven method proposed by (Haghighi and Klein, 2006), in which the dictionary is replaced with a very small seed of prototypical examples.

phological disambiguation.

Our primary conclusion is that the problem of learning effective stochastic classifiers remains primarily a search task. Initial conditions play a dominant role in solving this task and can rely on linguistically motivated approximations. A robust learning method (EM-HMM) combined with good initial conditions based on a robust feature set can go a long way (as opposed to a more complex learning method). It seems that computing initial conditions is also the right place to capture complex linguistic intuition without fear that over-generalization could lead a learner to diverge.

## 2 Previous Work

The tagging accuracy of supervised stochastic taggers is around 96%–97% (Manning and Schutze, 1999). Merialdo (1994) reports an accuracy of 86.6% for an unsupervised token-based EM-estimated HMM, trained on a corpus of about 1M words, over a tagset of 159 tags. Elworthy (1994), in contrast, reports accuracy of 75.49%, 80.87%, and 79.12% for unsupervised word-based HMM trained on parts of the LOB corpora, with a tagset of 134 tags. With (artificially created) good initial conditions, such as a good approximation of the tag distribution for each word, Elworthy reports an improvement to 94.6%, 92.27%, and 94.51% on the same data sets. Merialdo, on the other hand, reports an improvement to 92.6% and 94.4% for the case where 100 and 2,000 sentences of the training corpus are manually tagged. Later, Banko and Moore (2004) observed that earlier unsupervised HMM-EM results were artificially high due to use of *Optimized Lexicons*, in which only frequent-enough analyses of each word were kept. Brill (1995b) proposed an unsupervised tagger based on transformation-based learning (Brill, 1995a), achieving accuracies of above 95%. This unsupervised tagger relied on an initial step in which the most probable tag for each word is chosen. Optimized lexicons and Brill's most-probable-tag Oracle are not available in realistic unsupervised settings, yet, they show that *good initial conditions* greatly facilitate learning.

Recent work on unsupervised POS tagging for English has significantly improved the results on this task: GG, SE and most recently TJ report the best re-

sults so far on the task of unsupervised POS tagging of the WSJ with diluted dictionaries. With dictionaries as small as 1249 lexical entries the LDA-based method with a strong ambiguity-class model reaches POS accuracy as high as 89.7% on a reduced tagset of 17 tags.

While these 3 methods rely on the same feature set (lexical context, spelling features) for the learning stage, the LDA approach bases its predictions entirely on observable features, and excludes the traditional hidden states sequence.

In Hebrew, Levinger *et al.* (1995) introduced the *similar-words algorithm* for estimating $p(t|w)$ from unlabeled data, which we describe below. Our method uses this algorithm as a first step, and refines the approximation by introducing additional linguistic constraints and an iterative refinement step.

## 3 Initial Conditions For EM-HMM

The most common model for unsupervised learning of stochastic processes is Hidden Markov Models (HMM). For the case of tagging, the states correspond to the tags $t_i$, and words $w_i$ are emitted each time a state is visited. The parameters of the model can be estimated by applying the Baum-Welch EM algorithm (Baum, 1972), on a large-scale corpus of unlabeled text. The estimated parameters are then used in conjunction with Viterbi search, to find the most probable sequence of tags for a given sentence. In this work, we follow Adler (2007) and use a variation of second-order HMM in which the probability of a tag is conditioned by the tag that precedes it and by the one that follows it, and the probability of an emitted word is conditioned by its tag and the tag that follows it[2]. In all experiments, we use the back-off smoothing method of (Thede and Harper, 1999), with additive smoothing (Chen, 1996) for the lexical probabilities.

We investigate methods to approximate the initial parameters of the $p(t|w)$ distribution, from which we obtain $p(w|t)$ by marginalization and Bayesian inversion. We also experiment with constraining the $p(t|t_{-1}, t_{+1})$ distribution.

**General syntagmatic constraints**    We set linguistically motivated constraints on the $p(t|t_{-1}, t_{+1})$ distribution. In our setting, these are used to force the probability of some events to 0 (*e.g.*, "Hebrew verbs can not be followed by the *of* preposition").

**Morphology-based    p(t|w)    approximation**
Levinger *et al.* (1995) developed a context-free method for acquiring morpho-lexical probabilities ($p(t|w)$) from an untagged corpus. The method is based on language-specific rules for constructing a *similar words* (SW) set for each analysis of a word. This set is composed of morphological variations of the word under the given analysis. For example, the Hebrew token ילד can be analyzed as either a noun (boy) or a verb (gave birth). The `noun` SW set for this token is composed of the definiteness and number inflections הילד,ילדים,הילדים (the boy, boys, the boys), while the `verb` SW set is composed of gender and tense inflections ילדה,ילדו (she/they gave birth). The approximated probability of each analysis is based on the corpus frequency of its SW set. For the complete details, refer to the original paper. Cucerzan and Yarowsky (2000) proposed a similar method for the unsupervised estimation of $p(t|w)$ in English, relying on simple spelling features to characterize similar word classes.

**Linear-Context-based    p(t|w)    approximation**
The method of Levinger *et al.* makes use of Hebrew inflection patterns in order to estimate context free approximation of $p(t|w)$ by relating a word to its different inflections. However, the context in which a word occurs can also be very informative with respect to its POS-analysis (Schütze, 1995). We propose a novel algorithm for estimating $p(t|w)$ based on the contexts in which a word occurs.[3]

The algorithm starts with an initial $p(t|w)$ estimate, and iteratively re-estimates:

$$\hat{p}(t|c) = \frac{\sum_{w \in W} p(t|w)p(w|c)}{Z}$$

$$\hat{p}(t|w) = \frac{\sum_{c \in REL_C} p(t|c)p(c|w)allow(t,w)}{Z}$$

---

[2]Technically this is not Markov Model but a Dependency Net. However, bidirectional conditioning seem more suitable for language tasks, and in practice the learning and inference methods are mostly unaffected. See (Toutanova et al., 2003).

[3]While we rely on the same intuition, our use of context differs from earlier works on distributional POS-tagging like (Schütze, 1995), in which the purpose is to directly assign the possible POS for an unknown word. In contrast, our algorithm aims to improve the estimate for the whole distribution $p(t|w)$, to be further disambiguated by the EM-HMM learner.

where $Z$ is a normalization factor, $W$ is the set of all words in the corpus, $C$ is the set of all contexts, and $REL_C \subseteq C$ is a set of reliable contexts, defined below. $allow(t, w)$ is a binary function indicating whether $t$ is a valid tag for $w$. $p(c|w)$ and $p(w|c)$ are estimated via raw corpus counts.

Intuitively, we estimate the probability of a tag given a context as the average probability of a tag given any of the words appearing in that context, and similarly the probability of a tag given a word is the averaged probability of that tag in all the (reliable) contexts in which the word appears. At each round, we define $REL_C$, the set of reliable contexts, to be the set of all contexts in which $p(t|c) > 0$ for at most $X$ different $t$s.

The method is general, and can be applied to different languages. The parameters to specify for each language are: the initial estimation $p(t|w)$, the estimation of the $allow$ relation for known and OOV words, and the types of contexts to consider.

## 4   Application to Hebrew

In Hebrew, several words combine into a single token in both agglutinative and fusional ways. This results in a potentially high number of tags for each token. On average, in our corpus, the number of possible analyses per known word reached 2.7, with the ambiguity level of the extended POS tagset in corpus for English (1.41) (Dermatas and Kokkinakis, 1995).

In this work, we use the morphological analyzer of MILA – Knowledge Center for Processing Hebrew (*KC analyzer*). In contrast to English tagsets, the number of tags for Hebrew, based on all combinations of the morphological attributes, can grow theoretically to about 300,000 tags. In practice, we found 'only' about 3,560 tags in a corpus of 40M tokens training corpus taken from Hebrew news material and Knesset transcripts. For testing, we manually tagged the text which is used in the Hebrew Treebank (Sima'an et al., 2001) (about 90K tokens), according to our tagging guidelines.

### 4.1   Initial Conditions

**General syntagmatic constraints**  We define 4 syntagmatic constraints over $p(t|t_{-1}, t_{+1})$: (1) a construct state form cannot be followed by a verb,

preposition, punctuation, existential, modal, or copula; (2) a verb cannot be followed by the preposition של *šel* (of), (3) copula and existential cannot be followed by a verb, and (4) a verb cannot be followed by another verb, unless one of them has a prefix, or the second verb is an infinitive, or the first verb is imperative and the second verb is in future tense.[4]

**Morphology-Based $p(t|w)$ approximation**  We extended the set of rules used in Levinger *et al.* , in order to support the wider tagset used by the KC analyzer: (1) The SW set for adjectives, copulas, existentials, personal pronouns, verbs and participles, is composed of all gender-number inflections; (2) The SW set for common nouns is composed of all number inflections, with definite article variation for absolute noun; (3) Prefix variations for proper nouns; (4) Gender variation for numerals; and (5) Gender-number variation for all suffixes (possessive, nominative and accusative).

**Linear-Context-based $p(t|w)$ approximation**  For the initial $p(t|w)$ we use either a uniform distribution based on the tags allowed in the dictionary, or the estimate obtained by using the modified Levinger *et al.* algorithm. We use contexts of the form LR=$w_{-1}, w_{+1}$ (the neighbouring words). We estimate $p(w|c)$ and $p(c|w)$ via relative frequency over all the events $w1, w2, w3$ occurring at least 10 times in the corpus. $allow(t, w)$ follows the dictionary. Because of the wide coverage of the Hebrew lexicon, we take $REL_C$ to be $C$ (all available contexts).

### 4.2   Evaluation

We run a series of experiments with 8 distinct initial conditions, as shown in Table 1: our baseline (*Uniform*) is the uniform distribution over all tags provided by the KC analyzer for each word. The *Syntagmatic* initial conditions add the $p(t|t_{-1}, t_{+1})$ constraints described above to the uniform baseline. The *Morphology-Based* and *Linear-Context* initial conditions are computed as described above, while the *Morph+Linear* is the result of applying the linear-context algorithm over initial values computed by the Morphology-based method. We repeat

---

[4]This rule was taken from Shacham and Wintner(2007).

| Initial Condition | | Dist | Context-Free | | EM-HMM | |
|---|---|---|---|---|---|---|
| | | | Full | Seg+Pos | Full | Seg+Pos |
| Uniform | | 60 | 63.8 | 71.9 | 85.5 | 89.8 |
| Syntagmatic | Pair Constraints | 60 | / | / | 85.8 | 89.8 |
| | Init-Trans | 60 | / | / | 87.9 | 91 |
| Morpho-Lexical | Morph-Based | 76.8 | 76.4 | 83.1 | 87.7 | 91.6 |
| | Linear-Context | 70.1 | 75.4 | 82.6 | 85.3 | 89.6 |
| | Morph+Linear | **79.8** | **79.0** | **85.5** | 88 | 92 |
| PairConst+Morph | Morph-Based | / | / | / | 87.6 | 91.4 |
| | Linear-Context | / | / | / | 84.5 | 89.0 |
| | Morph+Linear | / | / | / | 87.1 | 91.5 |
| InitTrans+Morph | Morph-Based | / | / | / | 89.2 | 92.3 |
| | Linear-Context | / | / | / | 87.7 | 90.9 |
| | Morph+Linear | / | / | / | **89.4** | **92.4** |

Table 1: Accuracy (%) of Hebrew Morphological Disambiguation and POS Tagging over various initial conditions

these last 3 models with the addition of the syntagmatic constraints (*Synt+Morph*).

For each of these, we first compare the computed $p(t|w)$ against a gold standard distribution, taken from the test corpus (90K tokens), according to the measure used by (Levinger et al., 1995) (*Dist*). On this measure, we confirm that our improved morpho-lexical approximation improves the results reported by Levinger *et al.* from 74% to about 80% on a richer tagset, and on a much larger test set (90K vs. 3,400 tokens).

We then report on the effectiveness of $p(t|w)$ as a context-free tagger that assigns to each word the most likely tag, both for full morphological analysis (3,561 tags) (*Full*) and for the simpler task of token segmentation and POS tag selection (36 tags) (*Seg+Pos*). The best results on this task are 80.8% and 87.5% resp. achieved on the *Morph+Linear* initial conditions.

Finally, we test effectiveness of the initial conditions with EM-HMM learning. We reach 88% accuracy on full morphological and 92% accuracy for POS tagging and word segmentation, for the *Morph+Linear* initial conditions.

As expected, EM-HMM improves results (from 80% to 88%). Strikingly, EM-HMM improves the uniform initial conditions from 64% to above 85%. However, better initial conditions bring us much over this particular local maximum – with an error reduction of 20%. In all cases, the main improvement over the uniform baseline is brought by the morphology-based initial conditions. When applied on its own, the linear context brings modest improvement. But the combination of the paradigmatic morphology-based method with the linear context

improves all measures.

A most interesting observation is the detrimental contribution of the syntagmatic constraints we introduced. We found that 113,453 sentences of the corpus (about 5%) contradict these basic and apparently simple constraints. As an alternative to these common-sense constraints, we tried to use a small seed of randomly selected sentences (10K annotated tokens) in order to skew the initial uniform distribution of the state transitions. We initialize the $p(t|t_{-1}, t_{+1})$ distribution with smoothed ML estimates based on tag trigram and bigram counts (ignoring the tag-word annotations). This small seed initialization (*InitTrans*) has a great impact on accuracy. Overall, we reach 89.4% accuracy on full morphological and 92.4% accuracy for POS tagging and word segmentation, for the *Morph+Linear* conditions – an error reduction of more than 25% from the uniform distribution baseline.

## 5 Application to English

We now apply the same technique to English semi-supervised POS tagging. Recent investigations of this task use dictionaries derived from the Penn WSJ corpus, with a reduced tag set of 17 tags[5] instead of the original 45-tags tagset. They experiment with full dictionaries (containing complete POS information for all the words in the text) as well as "diluted" dictionaries, from which large portions of the vocabulary are missing. These settings are very different from those used for Hebrew: the tagset is much smaller (17 vs. ∼3,560) and the dictionaries are either complete or extremely crippled. However, for the sake of comparison, we have reproduced the same experimental settings.

We derive dictionaries from the complete WSJ corpus[6], and the exact same diluted dictionaries used in SE, TJ and GG.

---

[5] ADJ ADV CONJ DET ENDPUNC INPUNC LPUNC RPUNC N POS PRT PREP PRT TO V VBG VBN WH

[6] The dictionary derived from the WSJ data is very noisy: many of the stop words get wrong analyses stemming from tagging mistakes (for instance, the word *the* has 6 possible analyses in the data-derived dictionary, which we checked manually and found all but DT erroneous). Such noise is not expected in a real world dictionary, and our algorithm is not designed to accomodate it. We corrected the entries for the 20 most frequent words in the corpus. This step could probably be done automatically, but we consider it to be a non-issue in any realistic setting.

**Syntagmatic Constraints** We indirectly incorporated syntagmatic constraints through a small change to the tagset. The 17-tags English tagset allows for V-V transitions. Such a construction is generally unlikely in English. By separating modals from the rest of the verbs, and creating an additional class for the 5 *be* verbs (am,is,are,was,were), we made such transition much less probable. The new 19-tags tagset reflects the "verb can not follow a verb" constraint.

**Morphology-Based $p(t|w)$ approximation** English morphology is much simpler compared to that of Hebrew, making direct use of the Levinger context free approximation impossible. However, some morphological cues exist in English as well, in particular common suffixation patterns. We implemented our morphology-based context-free $p(t|w)$ approximation for English as a special case of the linear context-based algorithm described in Sect.3. Instead of generating contexts based on neighboring words, we generate them using the following 5 morphological templates:

**suff=S** The word has suffix $S$ (`suff=ing`).
**L+suff=W,S** The word appears just after word $W$, with suffix $S$ (`L+suff=have,ed`).
**R+suff=S,W** The word appears just before word $W$, with suffix $S$ (`R+suff=ing,to`)
**wsuf=S1,S2** The word suffix is $S1$, the same stem is seen with suffix $S2$ (`wsuf=ϵ,s`).
**suffs=SG** The word stem appears with the $SG$ group of suffixes (`suffs=ed,ing,s`).

We consider a word to have a suffix only if the word stem appears with a different suffix somewhere in the text. We implemented a primitive stemmer for extracting the suffixes while preserving a usable stem by taking care of few English orthography rules (handling, *e.g.,*, bigger → big er, nicer → nice er, happily → happy ly, picnicking → picnic ing). For the immediate context $W$ in the templates *L+suff,R+suff*, we consider only the 20 most frequent tokens in the corpus.

**Linear-Context-based $p(t|w)$ approximation** We expect the context based approximation to be particularly useful in English. We use the following 3 context templates: `LL=`$w_{-2}$`,`$w_{-1}$, `LR=`$w_{-1}$`,`$w_{+1}$ and `RR=`$w_{+1}$`,`$w_{+2}$. We estimate $p(w|c)$ and $p(c|w)$ by relative frequency over word triplets occurring at least twice in the unannotated training corpus.

**Combined $p(t|w)$ approximation** This approximation combines the morphological and linear context approximations by using all the above-mentioned context templates together in the iterative process.

For all three $p(t|w)$ approximations, we take $REL_C$ to be contexts containing at most 4 tags. $allow(t, w)$ follows the dictionary for known words, and is the set of all open-class POS for unknown words. We take the initial $p(t|w)$ for each $w$ to be uniform over all the dictionary specified tags for $w$. Accordingly, the initial $p(t|w) = 0$ for $w$ not in the dictionary. We run the process for 8 iterations.[7]

**Diluted Dictionaries and Unknown Words** Some of the missing dictionary elements are assigned a set of possible POS-tags and corresponding probabilities in the $p(t|w)$ estimation process. Other unknown tokens remain with no analysis at the end of the initial process computation. For these missing elements, we assign an ambiguity class by a simple ambiguity-class guesser, and set $p(t|w)$ to be uniform over all the tags in the ambiguity class. Our ambiguity-class guesser assigns for each word the set of all open-class tags that appeared with the word suffix in the dictionary. The word suffix is the longest (up to 3 characters) suffix of the word that also appears in the top-100 suffixes in the dictionary.

**Taggers** We test the resulting $p(t|w)$ approximation by training 2 taggers: **CF-Tag**, a context-free tagger assigning for each word its most probable POS according to $p(t|w)$, with a fallback to the most probable tag in case the word does not appear in the dictionary or if $\forall t, p(t|w) = 0$. **EM-HMM**, a second-order EM-HMM initialized with the estimated $p(t|w)$.

**Baselines** As baseline, we use two EM-trained HMM taggers, initialized with a uniform $p(t|w)$ for every word, based on the allowed tags in the dictionary. For words not in the dictionary, we take the allowed tags to be either all the open-class POS

---

[7]This is the first value we tried, and it seems to work fine. We haven't experimented with other values. The same applies for the choice of 4 as the $REL_C$ threshold.

(**uniform(oc)**) or the allowed tags according to our simple ambiguity-class guesser (**uniform(suf)**).

All the $p(t|w)$ estimates and HMM models are trained on the entire WSJ corpus. We use the same 24K word test-set as used in SE, TJ and GG, as well as the same diluted dictionaries. We report the results on the same reduced tagsets for comparison, but also include the results on the full 46 tags tagset.

## 5.1 Results

Table 2 summarizes the results of our experiments.

Uniform initialization based on the simple suffix-based ambiguity class guesser yields big improvements over the uniform all-open-class initialization. However, our refined initial conditions always improve the results (by as much as 40% error reduction). As expected, the linear context is much more effective than the morphological one, especially with richer dictionaries. This seem to indicate that in English the linear context is better at refining the estimations when the ambiguity classes are known, while the morphological context is in charge of adding possible tags when the ambiguity classes are not known. Furthermore, the benefit of the morphology-context is bigger for the complete tagset setting, indicating that, while the coarse-grained POS-tags are indicated by word distribution, the finer distinctions are indicated by inflections and orthography. The combination of linear and morphology contexts is always beneficial. Syntagmatic constraints (*e.g.*, separating *be* verbs and modals from the rest of the verbs) constantly improve results by about 1%. Note that the context-free tagger based on our $p(t|w)$ estimates is quite accurate. As with the EM trained models, combining linear and morphological contexts is always beneficial.

To put these numbers in context, Table 3 lists current state-of-the art results for the same task. **CE+spl** is the Contrastive-Estimation CRF method of SE. **BHMM** is the completely Bayesian-HMM of GG. **PLSA+AC**, **LDA**, **LDA+AC** are the models presented in TJ, LDA+AC is a Bayesian model with a strong ambiguity class (AC) component, and is the current state-of-the-art of this task. The other models are variations excluding the Bayesian components (PLSA+AC) or the ambiguity class.

While our models are trained on the unannotated text of the entire WSJ Treebank, CE and BHMM use much less training data (only the 24k words of the test-set). However, as noted by TJ, there is no reason one should limit the amount of unlabeled data used, and in addition other results reported in GG,SE show that accuracy does not seem to improve as more unlabeled data are used with the models. We also report results for training our EM-HMM tagger on the smaller dataset (the $p(t|w)$ estimation is still based on the entire unlabeled WSJ).

All the abovementioned models follow the assumption that all 17 tags are valid for the unknown words. In contrast, we restrict the set of allowed tags for an unknown word to open-class tags. Closed class words are expected to be included in a dictionary, even a small one. The practice of allowing only open-class tags for unknown words goes back a long way (Weischedel et al., 1993), and proved highly beneficial also in our case.

Notice that even our simplest models, in which the initial $p(t|w)$ distribution for each $w$ is uniform, already outperform most of the other models, and, in the case of the diluted dictionaries, by a wide margin. Similarly, given the $p(t|w)$ estimate, EM-HMM training on the smaller dataset (24k) is still very competitive (yet results improve with more unlabeled data). When we use our refined $p(t|w)$ distribution as the basis of EM-HMM training, we get the best results for the complete dictionary case. With the diluted dictionaries, we are outperformed only by LDA+AC. As we outperform this model in the complete dictionary case, it seems that the advantage of this model is due to its much stronger ambiguity class model, and not its Bayesian components. Also note that while we outperform this model when using the 19-tags tagset, it is slightly better in the original 17-tags setting. It could be that the reliance of the LDA models on observed surface features instead of hidden state features is beneficial avoiding the misleading V-V transitions.

We also list the performance of our best models with a slightly more realistic dictionary setting: we take our dictionary to include information for all words occurring in section 0-18 of the WSJ corpus (43208 words). We then train on the entire unannotated corpus, and test on sections 22-24 – the standard train/test split for supervised English POS tagging. We achieve accuracy of **92.85%** for the 19-tags set, and **91.3%** for the complete 46-tags tagset.

| Initial Conditions | | Full dict (49206 words) | | ≥ 2 dict (2141 words) | | ≥ 3 dict (1249 words) | |
|---|---|---|---|---|---|---|---|
| | | CF-Tag | EM-HMM | CF-Tag | EM-HMM | CF-Tag | EM-HMM |
| 17tags | Uniform(oc) | 81.7 | 88.7 | 68.4 | 81.9 | 62.5 | 79.6 |
| | Uniform(suf) | NA | NA | 76.8 | 83.4 | 76.9 | 81.6 |
| | Morph-Cont | 82.2 | 88.6 | 73.3 | 83.9 | 69.1 | 81.7 |
| | Linear-Cont | 90.1 | 92.9 | 81.1 | 87.8 | 78.3 | 85.8 |
| | Combined-Cont | 89.9 | 93.3 | 83.1 | 88.5 | 81.1 | 86.4 |
| 19tags | Uniform(oc) | 79.9 | 91.0 | 66.6 | 83.4 | 60.7 | 84.7 |
| | Uniform(suf) | NA | NA | 75.1 | 86.5 | 73.1 | 86.7 |
| | Morph-Cont | 80.5 | 89.2 | 71.5 | 86.5 | 67.5 | 87.1 |
| | Linear-Cont | 88.4 | 93.7 | 78.9 | 89.0 | 76.3 | 86.9 |
| | Combined-Cont | 88.0 | 93.8 | 81.1 | 89.4 | 79.2 | 87.4 |
| 46tags | Uniform(oc) | 76.7 | 88.3 | 61.2 | * | 55.7 | * |
| | Uniform(suf) | NA | NA | 64.2 | 81.9 | 60.3 | 79.8 |
| | Morph-Cont | 74.8 | 88.8 | 65.6 | 83.0 | 61.9 | 80.3 |
| | Linear-Cont | 85.5 | 91.2 | 74.5 | 84.0 | 70.1 | 82.2 |
| | Combined-Cont | 85.9 | 91.4 | 75.4 | 85.5 | 72.4 | 83.3 |

Table 2: Accuracy (%) of English POS Tagging over various initial conditions

| Dict | InitEM-HMM (24k) | LDA | LDA+AC | PLSA+AC | CE+spl | BHMM |
|---|---|---|---|---|---|---|
| Full | **93.8** (91.1) | 93.4 | 93.4 | 89.7 | 88.7 | 87.3 |
| ≥ 2 | 89.4 (87.9) | 87.4 | **91.2** | 87.8 | 79.5 | 79.6 |
| ≥ 3 | 87.4 (85.9) | 85 | **89.7** | 85.9 | 78.4 | 71 |

Table 3: Comparison of English Unsupervised POS Tagging Methods

## 6 Conclusion

We have demonstrated that unsupervised POS tagging can reach good results using the robust EM-HMM learner when provided with good initial conditions, even with incomplete dictionaries. We presented a general family of algorithms to compute effective initial conditions: estimation of $p(t|w)$ relying on an iterative process shifting probabilities between words and their contexts. The parameters of this process (definition of the contexts and initial estimations of $p(t|w)$ can safely encapsulate rich linguistic intuitions.

While recent work, such as GG, aim to use the Bayesian framework and incorporate "linguistically motivated priors", in practice such priors currently only account for the fact that language related distributions are sparse - a very general kind of knowledge. In contrast, our method allow the incorporation of much more fine-grained intuitions.

We tested the method on the challenging task of full morphological disambiguation in Hebrew (which was our original motivation) and on the standard WSJ unsupervised POS tagging task.

In Hebrew, our model includes an improved version of the *similar words* algorithm of (Levinger et al., 1995), a model of lexical context, and a small set of tag ngrams. The combination of these knowledge sources in the initial conditions brings an error reduction of more than 25% over a strong uniform distribution baseline. In English, our model is competitive with recent state-of-the-art results, while using simple and efficient learning methods.

The comparison with other algorithms indicates directions of potential improvement: (1) our initial-conditions method might benefit the other, more sophisticated learning algorithms as well. (2) Our models were designed under the assumption of a relatively complete dictionary. As such, they are not very good at assigning ambiguity-classes to OOV tokens when starting with a very small dictionary. While we demonstrate competitive results using a simple suffix-based ambiguity-class guesser which ignores capitalization and hyphenation information, we believe there is much room for improvement in this respect. In particular, (Haghighi and Klein, 2006) presents very strong results using a distributional-similarity module and achieve impressive tagging accuracy while starting with a mere 116 prototypical words. Experimenting with combining similar models (as well as TJ's ambiguity class model) with our $p(t|w)$ distribution estimation method is an interesting research direction.

# References

Meni Adler. 2007. *Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach*. Ph.D. thesis, Ben-Gurion University of the Negev, Beer-Sheva, Israel.

Michele Banko and Robert C. Moore. 2004. Part-of-speech tagging in context. In *Proceedings of Coling 2004*, pages 556–561, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Leonard E. Baum. 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process. *Inequalities*, 3:1–8.

Eric Brill. 1995a. Transformation-based error-driven learning and natural langue processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21:543–565.

Eric Brill. 1995b. Unsupervised learning of disambiguation rules for part of speech tagging. In David Yarovsky and Kenneth Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 1–13, Somerset, New Jersey. Association for Computational Linguistics.

Stanley F. Chen. 1996. *Building Probabilistic Models for Natural Language*. Ph.D. thesis, Harvard University, Cambridge, MA.

Silviu Cucerzan and David Yarowsky. 2000. Language independent, minimally supervised induction of lexical probabilities. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 270–277, Morristown, NJ, USA. Association for Computational Linguistics.

Evangelos Dermatas and George Kokkinakis. 1995. Automatic stochastic tagging of natural language texts. *Computational Linguistics*, 21(2):137–163.

David Elworthy. 1994. Does Baum-Welch re-estimation help taggers? In *Proceeding of ANLP-94*.

Sharon Goldwater and Thomas L. Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceeding of ACL 2007*, Prague, Czech Republic.

Aria Haghighi and Dan Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 320–327, Morristown, NJ, USA. Association for Computational Linguistics.

J. Kupiec. 1992. Robust part-of-speech tagging using hidden Markov model. *Computer Speech and Language*, 6:225–242.

Moshe Levinger, Uzi Ornan, and Alon Itai. 1995. Learning morpholexical probabilities from an untagged corpus with an application to Hebrew. *Computational Linguistics*, 21:383–404.

Christopher D. Manning and Hinrich Schutze. 1999. *Foundation of Statistical Language Processing*. MIT Press.

Bernard Merialdo. 1994. Tagging English text with probabilistic model. *Computational Linguistics*, 20:155–171.

Hinrich Schütze. 1995. Distributional part-of-speech tagging. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, pages 141–148, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Danny Shacham and Shuly Wintner. 2007. Morphological disambiguation of hebrew: A case study in classifier combination. In *Proceeding of EMNLP-07*, Prague, Czech.

Khalil Sima'an, Alon Itai, Alon Altman Yoad Winter, and Noa Nativ. 2001. Building a tree-bank of modern Hebrew text. *Journal Traitement Automatique des Langues (t.a.l.)*. Special Issue on NLP and Corpus Linguistics.

Noah A. Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 354–362, Ann Arbor, Michigan, June.

Scott M. Thede and Mary P. Harper. 1999. A second-order hidden Markov model for part-of-speech tagging. In *Proceeding of ACL-99*.

Kristina Toutanova and Mark Johnson. 2008. A bayesian lda-based model for semi-supervised part-of-speech tagging. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*.

R. Weischedel, R. Schwartz, J. Palmucci, M. Meteer, and L. Ramshaw. 1993. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, 19:359–382.