Observational Initialization of Type-Supervised Taggers

Hui Zhang*
Department of Computer Science
University of Southern California
hzhang@isi.edu

John DeNero
Google, Inc.
denero@google.com

Abstract

Recent work has sparked new interest in type-supervised part-of-speech tagging, a data setting in which no labeled sentences are available, but the set of allowed tags is known for each word type. This paper describes *observational initialization*, a novel technique for initializing EM when training a type-supervised HMM tagger. Our initializer allocates probability mass to unambiguous transitions in an unlabeled corpus, generating token-level observations from type-level supervision. Experimentally, observational initialization gives state-of-the-art type-supervised tagging accuracy, providing an error reduction of 56% over uniform initialization on the Penn English Treebank.

1 Introduction

For many languages, there exist comprehensive dictionaries that list the possible parts-of-speech for each word type, but there are no corpora labeled with the part-of-speech of each token in context. Type-supervised tagging (Merialdo, 1994) explores this scenario; a model is provided with type-level information, such as the fact that "only" can be an adjective, adverb, or conjunction, but not any token-level information about which instances of "only" in a corpus are adjectives. Recent research has focused on using type-level supervision to infer token-level tags. For instance, Li et al. (2012) derive type-level supervision from Wiktionary, Das and Petrov (2011) and Täckström et al. (2013) project type-level tag sets across languages, and Garrette and Baldridge (2013) solicit type-level annotations directly from speakers. In all of these efforts, a probabilistic sequence model is trained to disambiguate token-level tags that are constrained to match type-level tag restrictions. This paper describes *observational initialization*, a simple but effective learning technique for training type-supervised taggers.

A hidden Markov model (HMM) can be used to disambiguate tags of individual tokens by maximizing corpus likelihood using the expectation maximization (EM) algorithm. Our approach is motivated by a suite of oracle experiments that demonstrate the effect of initialization on the final tagging accuracy of an EM-trained HMM tagger. We show that initializing EM with accurate transition model parameters is sufficient to guide learning toward a high-accuracy final model.

Inspired by this finding, we introduce *observational initialization*, which is a simple method to heuristically estimate transition parameters for a corpus using type-level supervision. Transition probabilities are estimated from unambiguous consecutive tag pairs that arise when two consecutive words each have only a single allowed tag. These unambiguous word pairs can be tagged correctly without any statistical inference. Initializing EM with the relative frequency of these unambiguous pairs improves tagging accuracy dramatically over uniform initialization, reducing errors by 56% in English and 29% in German. This efficient and data-driven approach gives the best reported tagging accuracy for type-supervised sequence models, outperforming the minimized model of Ravi and Knight (2009), the Bayesian LDA-based model of Toutanova and Johnson (2008), and an HMM trained with language-specific initialization described by Goldberg et al. (2008).

2 Type-Supervised Tagging

A first-order Markov model for part-of-speech tagging defines a distribution over sentences for which a single tag is given to each word token. Let $w_i \in W$ refer to the ith word in a sentence $\mathbf{w}$, drawn from language vocabulary W. Likewise,

*Research conducted during an internship at Google.

$t_i \in T$ is the tag in tag sequence $\mathbf{t}$ of the ith word, drawn from tag inventory T. The joint probability of a sentence can be expressed in terms of two sets of parameters for conditional multinomial distributions: ϕ defines the probability of a tag given its previous tag and θ defines the probability of a word given its tag.

$$P_{\phi,\theta}(\mathbf{w}, \mathbf{t}) = \prod_{i=1}^{|\mathbf{w}|} P_\phi(t_i|t_{i-1}) \cdot P_\theta(w_i|t_i)$$

Above, t_0 is a fixed start-of-sentence tag.

For a set of sentences $\mathcal{S}$, the EM algorithm can be used to iteratively find a local maximum of the corpus log-likelihood:

$$\ell(\phi, \theta; \mathcal{S}) = \sum_{\mathbf{w} \in \mathcal{S}} \ln \left[\sum_{\mathbf{t}} P_{\phi,\theta}(\mathbf{w}, \mathbf{t}) \right]$$

The parameters ϕ and θ can then be used to predict the most likely sequence of tags for each sentence under the model:

$$\hat{\mathbf{t}}(\mathbf{w}) = \arg\max_{\mathbf{t}} P_{\phi,\theta}(\mathbf{w}, \mathbf{t})$$

Tagging accuracy is the fraction of these tags in $\hat{\mathbf{t}}(\mathbf{w})$ that match hand-labeled oracle tags $\mathbf{t}^*(\mathbf{w})$.

Type Supervision. In addition to an unlabeled corpus of sentences, type-supervised models also have access to a tag dictionary $D \subseteq W \times T$ that contains all allowed word-tag pairs. For an EM-trained HMM, initially setting $P_\theta(w|t) = 0$ for all $(w, t) \notin D$ ensures that all words will be labeled with allowed tags.

Tag dictionaries can be derived from various sources, such as lexicographic resources (Li et al., 2012) and cross-lingual projections (Das and Petrov, 2011). In this paper, we will follow previous work in deriving the tag dictionary from a labeled corpus (Smith and Eisner, 2005); this synthetic setting maximizes experiment repeatability and allows for direct comparison of type-supervised learning techniques.

Transductive Applications. We consider a transductive data setting in which the test set is available during training. In this case, the model is not required to generalize to unseen examples or unknown words, as in the typical inductive setting.

Transductive learning arises in document clustering and corpus analysis applications. For example, before running a document clustering algorithm on a fixed corpus of documents, it may be useful to tag each word with its most likely part-of-speech in context, disambiguating the lexical features in a bag-of-words representation. In corpus analysis or genre detection, it may be useful to determine for a fixed corpus the most common part-of-speech for each word type, which could be inferred by tagging each word with its most likely part-of-speech. In both cases, the set of sentences to tag is known in advance of learning.

3 Initializing HMM Taggers

The EM algorithm is sensitive to initialization. In a latent variable model, different parameter values may yield similar data likelihoods but very different predictions. We explore this issue via experiments on the Wall Street Journal section of the English Penn Treebank (Marcus et al., 1993). We adopt the transductive data setting introduced by Smith and Eisner (2005) and used by Goldwater and Griffiths (2007), Toutanova and Johnson (2008) and Ravi and Knight (2009); models are trained on all sections 00-24, the tag dictionary D is constructed by allowing all word-tag pairs appearing in the entire labeled corpus, and the tagging accuracy is evaluated on a 1005 sentence subset sampled from the corpus.

The degree of variation in tagging accuracy due to initialization can be observed most clearly by two contrasting initializations. UNIFORM initializes the model with uniform distributions over allowed outcomes:

$$P_\phi(t|t') = \frac{1}{|T|}$$

$$P_\theta(w|t) = \frac{1}{|\{w : (w, t) \in D\}|}$$

SUPERVISED is an oracle setting that initializes the model with the relative frequency of observed pairs in a labeled corpus:

$$P_\phi(t|t') \propto \sum_{(\mathbf{w}, \mathbf{t}^*)} \sum_{i=1}^{|\mathbf{w}|} \delta((t_i^*, t_{i-1}^*), (t, t'))$$

$$P_\theta(w|t) \propto \sum_{(\mathbf{w}, \mathbf{t}^*)} \sum_{i=1}^{|\mathbf{w}|} \delta((w_i, t_i^*), (w, t))$$

where the Kronecker $\delta(x, y)$ function is 1 if x and y are equal and 0 otherwise.

Figure 1 shows that while UNIFORM and SUPERVISED achieve nearly identical data log-likelihoods, their final tagging accuracy differs by

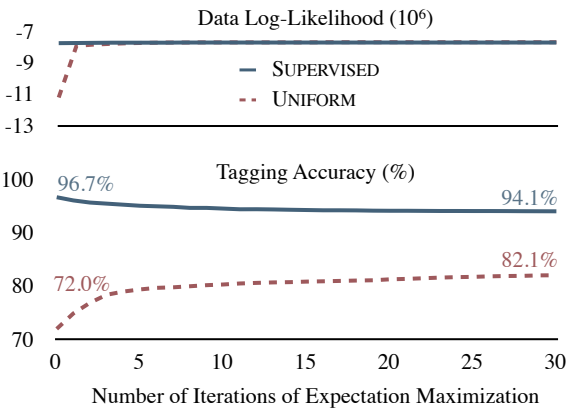

Figure 1: The data log-likelihood (top) and tagging accuracy (bottom) of two contrasting initializers, UNIFORM and SUPERVISED, compared on the Penn Treebank.

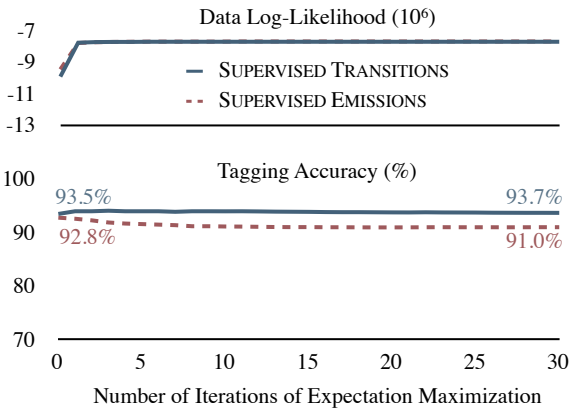

Figure 2: The data log-likelihood (top) and tagging accuracy (bottom) of two partially supervised initializers, one with SUPERVISED TRANSITIONS and one with SUPERVISED EMISSIONS, compared on the Penn Treebank.

12%. Accuracy degrades somewhat from the SUPERVISED initialization, since the data likelihood objective differs from the objective of maximizing tagging accuracy. However, the final SUPERVISED performance of 94.1% shows that there is substantial room for improvement over the UNIFORM initializer.

Figure 2 compares two partially supervised initializations. SUPERVISED TRANSITIONS initializes the transition model with oracle counts, but the emission model uniformly. Conversely, SUPERVISED EMISSIONS initializes the emission parameters from oracle counts, but initializes the transition model uniformly. There are many more emission parameters (57,390) than transition parameters (1,858). Thus, it is not surprising that SUPERVISED EMISSIONS gives a higher initial likelihood. Again, both initializers lead to solutions with nearly the same likelihood as SUPERVISED and UNIFORM.

Figure 2 shows that SUPERVISED TRANSITIONS outperforms SUPERVISED EMISSIONS in tagging accuracy, despite the fact that fewer parameters are set with supervision. With fixed D, an accurate initialization of the transition distributions leads to accurate tagging after EM training. We therefore concentrate on developing an effective initialization for the transition distribution.

4 Observational Initialization

The SUPERVISED TRANSITIONS initialization is estimated from observations of consecutive tags in a labeled corpus. Our OBSERVATIONAL initializer is likewise estimated from the relative frequency of consecutive tags, taking advantage of the structure of the tag dictionary D. However, it does not require a labeled corpus.

Let $D(w, \cdot) = \{t : (w, t) \in D\}$ denote the allowed tags for word w. The set

$$U = \{w : |D(w, \cdot)| = 1\}$$

contains all words that have only one allowed tag. When a token of some $w \in U$ is observed in a corpus, its tag is unambiguous. Therefore, its tag is observed as well, and a portion of the tag sequence is known. When consecutive pairs of tokens are both in U, we can observe a transition in the latent tag sequence. The OBSERVATIONAL initializer simply estimates a transition distribution from the relative frequency of these unambiguous observations that occur whenever two consecutive tokens both have a unique tag.

We now formally define the observational initializer. Let $g(w, t) = \delta(D(w, \cdot), \{t\})$ be an indicator function that is 1 whenever $w \in U$ and its single allowed tag is t, and 0 otherwise. Then, we initialize ϕ such that:

$$P_\phi(t|t') \propto \sum_{\mathbf{w} \in \mathcal{S}} \sum_{i=1}^{|\mathbf{w}|} g(w_i, t) \cdot g(w_{i-1}, t')$$

The emission parameters θ are set to be uniform over allowed words for each tag, as in UNIFORM initialization.

Figure 3 compares the OBSERVATIONAL initializer to the SUPERVISED TRANSITIONS initializer, and the top of Table 1 summarizes the performance of all initializers discussed so far for the

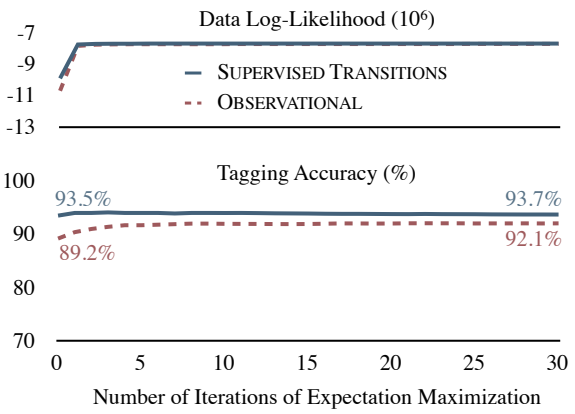

Figure 3: The data log-likelihood (top) and tagging accuracy (bottom) of initializing with SUPERVISED TRANSITIONS compared to the unsupervised OBSERVATIONAL initialization that requires only a tag dictionary and an unlabeled training corpus.

English	Initial	EM-trained
UNIFORM	72.0	82.1
OBSERVATIONAL	89.2	92.1
SUP. EMISSIONS	92.8	91.0
SUP. TRANSITIONS	93.5	93.7
FULLY SUPERVISED	96.7	94.1
German	**Initial**	**EM-trained**
UNIFORM	77.2	88.8
OBSERVATIONAL	92.7	92.1
SUP. EMISSIONS	90.7	89.0
SUP. TRANSITIONS	94.8	92.0
FULLY SUPERVISED	97.0	92.9

Table 1: Accuracy of English (top) and German (bottom) tagging models at initialization (left) and after 30 iterations of EM training (right) using various initializers.

English Penn Treebank. The OBSERVATIONAL initializer provides an error reduction over UNIFORM of 56%, surpassing the performance of an initially supervised emission model and nearing the performance of a supervised transition model.

The bottom of Table 1 shows a similar comparison on the Tübingen treebank of spoken German (Telljohann et al., 2006). Both training and testing were performed on the entire treebank. The observational initializer provides an error reduction over UNIFORM of 29%, and again outperforms SUPERVISED EMISSIONS. On this dataset OBSERVATIONAL initialization matches the final performance of SUPERVISED TRANSITIONS.

5 Discussion

The fact that observations and prior knowledge are useful for part-of-speech tagging is well understood (Brill, 1995), but the approach of estimating an initial transition model only from unambiguous word pairs is novel.

Our experiments show that for EM-trained HMM taggers in a type-supervised transductive data setting, observational initialization is an effective technique for guiding training toward high-accuracy solutions, approaching the oracle accuracy of SUPERVISED TRANSITIONS initialization.

The fact that models with similar data likelihood can vary dramatically in accuracy has been observed in other learning problems. For instance, Toutanova and Galley (2011) show that optimal parameters for IBM Model 1 are not unique, and alignments predicted from different optimal parameters vary significantly in accuracy.

However, the effectiveness of observational initialization is somewhat surprising because EM training includes these unambiguous tag pairs in its expected counts, even with uniform initialization. Our experiments indicate that this signal is not used effectively unless explicitly encoded in the initialization.

In our English data, 48% of tokens and 74% of word types have only one allowed tag. 28% of pairs of adjacent tokens have only one allowed tag pair and contribute to observational initialization. In German, 49% of tokens and 87% of word types are unambiguous, and 26% of adjacent token pairs are unambiguous.

6 Related Work

We now compare with several previous published results on type-supervised part-of-speech tagging trained using the same data setting on the English WSJ Penn Treebank, introduced by Smith and Eisner (2005).

Contrastive estimation (Smith and Eisner, 2005) is a learning technique that approximates the partition function of the EM objective in a log-linear model by considering a neighborhood around observed training examples. The Bayesian HMM of Goldwater and Griffiths (2007) is a second-order HMM (*i.e.*, likelihood factors over triples of tags) that is estimated using a prior distribution that promotes sparsity. Sparse priors have

	45 tag set		17 tag set	
	All train	*973k train*	*All train*	*973k train*
Observational initialization (this work)	**92.1**	**92.8**	**93.9**	94.8
Contrastive Estimation (Smith and Eisner, 2005)	–	–	88.7	–
Bayesian HMM (Goldwater and Griffiths, 2007)	86.8	–	87.3	–
Bayesian LDA-HMM (Toutanova and Johnson, 2008)	–	–	93.4	–
Linguistic initialization (Goldberg et al., 2008)	91.4	–	93.8	–
Minimal models (Ravi and Knight, 2009)	–	92.3	–	**96.8**

Table 2: Tagging accuracy of different approaches on English Penn Treebank. Columns labeled *973k train* describe models trained on the subset of 973k tokens used by Ravi and Knight (2009).

been motivated empirically for this task (Johnson, 2007). The Bayesian HMM model predicts tag sequences via Gibbs sampling, integrating out model parameters. The Bayesian LDA-based model of Toutanova and Johnson (2008) models ambiguity classes of words, which allows information sharing among words in the tag dictionary. In addition, it incorporates morphology features and a sparse prior of tags for a word. Inference approximations are required to predict tags, integrating out model parameters.

Ravi and Knight (2009) employs integer linear programming to select a minimal set of parameters that can generate the test sentences, followed by EM to set parameter values. This technique requires the additional information of which sentences will be used for evaluation, and its scalability is limited. In addition, this work used a subset of the WSJ Penn Treebank for training and selecting a tag dictionary. This restriction actually tends to improve performance, because a smaller tag dictionary further constrains model optimization. We compare directly to their training set, kindly provided to us by the authors.

The linguistic initialization of Goldberg et al. (2008) is most similar to the current work, in that it estimates maximum likelihood parameters of an HMM using EM, but starting with a well-chosen initialization with language specific linguistic knowledge. That work estimates emission distributions using a combination of suffix morphology rules and corpus context counts.

Table 2 compares our results to these related techniques. Each column represents a variant of the experimental setting used in prior work. Smith and Eisner (2005) introduced a mapping from the full 45 tag set of the Penn Treebank to 17 coarse tags. We report results on this coarse set by projecting from the full set after learning and inference.[1] Using the full tag set or the full training data, our method offers the best published performance without language-specific assumptions or approximate inference.

7 Future Work

This paper has demonstrated a simple and effective learning method for type-supervised, transductive part-of-speech tagging. However, it is an open question whether the technique is as effective for tag dictionaries derived from more natural sources than the labels of an existing treebank.

All of the methods to which we compare except Goldberg et al. (2008) focus on learning and modeling techniques, while our method only addresses initialization. We look forward to investigating whether our technique can be used as an initialization or prior for these other methods.

References

Eric Brill. 1995. Unsupervised learning of disambiguation rules for part of speech tagging. In *In Natural Language Processing Using Very Large Corpora*, pages 1–13. Kluwer Academic Press.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the Assocation for Computational Linguistics*.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the North American Chapter of the Assocation for Computational Linguistics*.

Yoav Goldberg, Meni Adler, and Michael Elhadad. 2008. EM can find pretty good HMM POS-taggers

[1] Training with the reduced tag set led to lower performance of 91.0% accuracy, likely because the coarse projection drops critical information about allowable English transitions, such as what verb forms can follow *to be* (Goldberg et al., 2008).

(when given a good start). In *Proceedings of the Association for Computational Linguistics*.

Sharon Goldwater and Tom Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the Association for Computational Linguistics*.

Mark Johnson. 2007. Why doesnt EM nd good HMM POS-taggers? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Shen Li, João V. Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*.

Bernard Merialdo. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*.

Sujith Ravi and Kevin Knight. 2009. Minimized models for unsupervised part-of-speech tagging. In *Proceedings of the Association for Computational Linguistics*.

Noah A. Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the Association for Computational Linguistics*.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*.

Heike Telljohann, Erhard Hinrichs, Sandra Kübler, and Heike Zinsmeister. 2006. Stylebook for the tbingen treebank of written german.

Kristina Toutanova and Michel Galley. 2011. Why initialization matters for ibm model 1: Multiple optima and non-strict convexity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 461–466, Portland, Oregon, USA, June. Association for Computational Linguistics.

Kristina Toutanova and Mark Johnson. 2008. A Bayesian LDA-based model for semi-supervised part-of-speech tagging. In *Proceedings of Neural and Information Processing Systems*.