Knowledge as a Constraint on Uncertainty for Unsupervised Classification: A Study in Part-of-Speech Tagging

Thomas J. Murray IV TMURRAY@USC.EDU
University of Southern California, Speech Analysis and Interpretation Laboratory, Los Angeles, CA 90089 USA

Panayiotis G. Georgiou GEORGIOU@SIPI.USC.EDU
University of Southern California, Speech Analysis and Interpretation Laboratory, Los Angeles, CA 90089 USA

Shrikanth S. Narayanan SHRI@SIPI.USC.EDU
University of Southern California, Speech Analysis and Interpretation Laboratory, Los Angeles, CA 90089 USA

Abstract

This paper evaluates the use of prior knowledge to limit or bias the choices of a classifier during otherwise unsupervised training and classification. Focusing on effects in the uncertainty of the model's decisions, we quantify the contributions of the knowledge source as a reduction in the conditional entropy of the label distribution given the input corpus. Allowing us to compare different sets of knowledge without annotated data, we find that label entropy is highly predictive of final performance for a standard Hidden Markov Model (HMM) on the task of part-of-speech tagging. Our results show too that even basic levels of knowledge, integrated as labeling constraints, have considerable effect on classification accuracy, in addition to more stable and efficient training convergence. Finally, for cases where the model's internal classes need to be interpreted and mapped to a desired label set, we find that, for constrained models, the requirements for annotated data to make quality assignments are greatly reduced.

1. Introduction

This paper investigates one of the simplest methods for integrating prior knowledge into the training of an unsupervised classifier, in particular the restriction or, more generally, weighting of output labels for each given input. In this setting, we focus on how the knowledge source constrains the set of available choices for the learner, effectively reducing the uncertainty in the classification decision. More precisely, viewing this guidance as a distribution over label output for the input data, we then may quantify and compare the effects of different sets of knowledge in terms of conditional entropy, without the need for annotated data.

To evaluate the relationship between knowledge constraints, uncertainty, and classification performance, we take the basic task of part-of-speech tagging, with the standard, first-order Hidden Markov Model (HMM) tagger of Merialdo (1994). We apply a number of basic constraint sets during training and evaluation, from lexical rules to partial tagging dictionaries, and find that the conditional label entropy is highly predictive of final model performance, with even the weakest constraints leading to large increases in classification accuracy. In addition, we see considerable reductions of variance in performance with respect to initial conditions and accelerated training convergence. Finally, addressing the problem of assigning interpretable labels to internal model classes, we find that the more constrained models require much less annotated data to find quality mappings.

The remainder of the paper is organized as follows. After discussing related work in the next section, we formalize the learning setting and clarify our entropy calculation in Section 3. Following a brief description of our constraint sets in Section 4, we present our results in Section 5 and conclude with a discussion in Section 6.

Appearing in *Workshop on Prior Knowledge for Text and Language*, Helsinki, Finland, 2008.

2. Related Work

A major impetus of the present study was Johnson's (2007) comparison of Expectation-Maximization (EM) and Bayesian estimators for unsupervised tagging, in particular the following conclusions of that work: (1) EM can be competitive with more sophisticated Bayesian methods, (2) greatly subject to the choice of evaluation method, but (3) to a certain extent, it is possible to compensate for EM's weakness in estimating skewed distributions by constraining the model to exclude rare events. While this is certainly not an argument against the use of better estimators, it does suggest the potential benefits of even simple means to guide model training.

Merialdo (1994) introduced the statistical tagging models employed in this paper and many, many others, both for supervised and unsupervised training. Like the original, much subsequent work on unsupervised tagging has relied on a full dictionary of possible tags for each word, which in practice constrains the training sufficiently to obviate the need for remapping of model output. More recent approaches with discriminative models (Smith & Eisner, 2005) and Bayesian estimators (Goldwater & Griffiths, 2007) have achieved good performance while reducing the dictionary. Haghighi and Klein (2006) require only a limited set of class prototypes and representations of their context distributions, while Toutanova and Johnson (2008) use only the possible groups of tags over which words are seen to vary.

Our simple integration of knowledge constraints may be viewed as an instance of virtual evidence (Pearl, 1988), a method to account for external judgements not easily expressed in terms of the probability distributions and dependencies encoded by the model. Bilmes (2004) suggests virtual evidence as a means to integrate the assessments of external models. Chang et al. (2007) uses weighted constraints successfully to guide search in an iterative algorithm for semi-supervised learning.

3. Knowledge as a Constraint on Uncertainty

To evaluate the effects of prior knowledge in entropic terms, we extend the usual formulation of a classifier to include, along with each input x, a mapping $\phi_x(y)$ of each output label y to a non-negative weight. That is, we define a classifier as a mapping $\mathcal{X} \times (\mathcal{Y} \to \mathbb{R}) \to \mathcal{Y}$. In the supervised case, $\phi_x(y)$ is non-zero for exactly one value of y, while in the purely unsupervised case, the weights are equal for all values of y. Normalizing the weights and interpreting them as a distribution $p(Y = y|X = x)$, the conditional entropy $H(Y|X)$ is a natural measure of the uncertainty facing the classifier and a means to compare and predict the effects of different sets of knowledge.[1] Because the uniform distribution has the maximum entropy for an event space of a given cardinality (Cover & Thomas, 1991), any adjustment to the label weights must reduce entropy relative to the purely unsupervised case. Assuming this reweighting does not penalize or eliminate the correct label, we expect a similar improvement to model accuracy.

3.1. Entropy Calculation

For sequential labeling tasks such as part-of-speech tagging, we generally label each token individually to avoid sparsity in our estimation, and so it is natural to take the x and y in the $p(y|x)$ above to refer to a single input token and label. Some of the constraints in our experiments involve context, however, so that $\phi_x(y)$ may vary between instances of x in the corpus. Accordingly, we must introduce some notion of context C and calculate the entropy as

$$\begin{aligned} H(Y|X,C) &= \sum_x \sum_c p(x,c)\, H(Y|X=x, C=c) \\ &= \sum_x p(x) \sum_c p(c|x) \sum_y p(y|x,c) \log p(y|x,c) \end{aligned}$$

If, however, we estimate $p(c|x)$ by simple counts over contexts that are equivalent under our constraints, we effectively sum out c, computing $H(Y|X = x)$ as an average of $H(Y|X = x, C)$ in all contexts. Thus, while our calculations use $p(y|x, c)$, we will continue to speak of $H(Y|X)$ for the remainder of the paper.

4. Constraints on Unsupervised Tagging

Great care is always required to evaluate an unsupervised classifier fairly against a labeled corpus, but evaluation is even more of a delicate matter when additional prior knowledge is involved. Ideally our knowledge should not be derived from the corpus, or we are crossing the line into supervised learning, but, practically, a successful set of constraints must accord with the knowledge implicit in the annotated data and be expressed in similar terms.[2] Accordingly, unless noted otherwise, we construct the following constraint sets

[1] This value also accords with the common informal characterization of classification difficulty in terms of average possible labels per input element.

[2] For example, an educated speaker of English would differentiate between the uses of the word 'to' as a preposition

from the knowledge of a native speaker and from general grammar resources, independent of the corpus.

Base Lexical Constraints For our base rule set, we include knowledge about punctuation, numbers, and capitalization, with numbers forced to either contain a digit or recursively to follow another possible number (to handle, e.g. '10 million'), and proper nouns defined similarly with respect to capitalization.

Closed Tags For each closed part-of-speech tag, we then add (incomplete) lists of possible words, derived from external sources as available.

Top Words Like much work on unsupervised tagging, we finally apply a tagging dictionary built from the corpus, but only for the 100 or 200 most common words, a much more limited amount of annotation.

For most of the experiments, we apply the above as hard constraints, with all violating hypotheses excluded, but we also examine the use of soft constraints, where hypotheses that observe the rules are simply preferred.

5. Experimental Results

5.1. Experimental Overview

In our experiments, we performed unsupervised training of the simple first-order HMM tagging model of Merialdo (1994), using the EM algorithm with a variety of constraint sets. Our implementation used the Graphical Models Toolkit (GMTK) (Bilmes & Zweig, 2002), with hard and soft constraints integrated via the toolkit's deterministic node construct.[3] For training and evaluation, we used the Wall St. Journal portion of the Penn Treebank, version 3 (Marcus et al., 1993), with data sets containing the 48k, 96k, and 193k words following the start of section 2.

To account for the local search properties of EM, we repeated each experiment 10 times, training for 500 iterations, with parameters initialized by small, random perturbations from the uniform distribution. Because our constraints cause no changes to the model's parameter set, it was possible to use the same random initializations across constraints for each data set, and thus attempt to control for any bias from particularly good or bad initialization points.[4]

For evaluation, we used the 'many-to-one' and 'one-to-one' labeling procedures as described by Johnson (2007), which greedily assign each model state to the annotated tag with which it occurs most often, respectively either allowing or prohibiting multiple states to map to a single tag. While, as Johnson (2007) and (Haghighi & Klein, 2006) mention, we may cheat with the many-to-one labeling by inflating the number of model states, this flaw seems less critical if the state count equals the size of the tag set, as in our experiments.

5.2. Results and Analysis

As summarized in Table 1 we find that even the least constrained models show considerable improvement over the baseline, with up to 20-30 percentage points gained in accuracy. Despite the simplistic nature of the model, performance is often surprisingly close to much more sophisticated models and training techniques, e.g. (Smith & Eisner, 2005; Haghighi & Klein, 2006; Goldwater & Griffiths, 2007). As we might expect, the effects are most pronounced on the smaller data sets, where the constraints serve as a strong prior compensating for lack of evidence, similar to what we see with the Bayesian models of Goldwater and Griffiths (2007). The effect on both the many-to-one and one-to-one label assignments is roughly equal across experiments, so that the difference in accuracy between the two assignments changes little as we add constraints.

To assess the effect of uncertainty on final model performance, we computed the Pearson correlation coefficient r^2 between the label entropy and the classifier accuracy, as shown in Figure 1. While the two are not fully correlated – and we should not necessarily expect them to be – the entropy measure is quite indicative of performance, and we conclude that is a reasonable means for predicting the effects of domain knowledge when annotated data is unavailable for evaluation.

Of course, knowledge is helpful only if the correct answer is not among the excluded hypotheses. We explored the effects of imperfect knowledge by applying our closed-tag rules set as a hard constraint and then as a soft constraint with relative likelihood weights ranging from 2:1 to 16:1. Though these rules are incomplete for most of the tags covered, and thus the correct labels for many words were excluded, we saw the hard constraints perform best, edging out the highest-weight soft constraints. It appears that, while

and as an infinitive verb marker, but in the Penn Treebank corpus, both are labeled 'TO'.

[3]We thank Chris Bartels for assistance on model implementation.

[4]From casual inspection, however, we did not see any performance patterns across runs. One might argue that different constraints lead to completely different optimization surfaces and extrema under EM.

Model	48k			96k			193k		
	$H(Y\|X)$	N:1	1:1	$H(Y\|X)$	N:1	1:1	$H(Y\|X)$	N:1	1:1
Base	5.49	33.8 (3.7)	21.7 (2.8)	5.49	42.9 (4.4)	30.1 (3.2)	5.49	52.1 (2.5)	34.4 (3.1)
Lower case	5.49	42.3 (2.2)	29.7 (2.3)	5.49	48.9 (2.4)	34.6 (2.5)	5.49	52.7 (2.3)	36.8 (1.9)
+Baselex	4.31	53.6 (0.8)	39.8 (1.9)	4.29	57.3 (0.8)	42.4 (1.6)	4.30	60.7 (0.8)	43.9 (1.7)
+Closed	3.71	64.9 (0.8)	54.3 (0.8)	3.69	66.2 (0.5)	55.5 (0.9)	3.70	67.4 (0.6)	56.4 (0.6)
+Top 100	3.49	69.2 (0.0)	57.8 (0.3)	3.47	70.1 (0.1)	58.6 (0.2)	3.48	71.0 (0.2)	59.5 (0.1)
+Top 200	3.49	**71.9** (0.1)	**60.5** (0.6)	3.47	**72.8** (0.1)	**61.7** (0.3)	3.48	**73.8** (0.1)	**62.1** (0.3)

Table 1. Tagging accuracy with increasing knowledge (as measured by conditional label entropy) on different data sets. Models were evaluated using the many-to-one and one-to-one label assignments, with results averaged over 10 runs, standard deviation in parentheses. Except for the base model, all words were mapped to lower case to reduce vocabulary size.

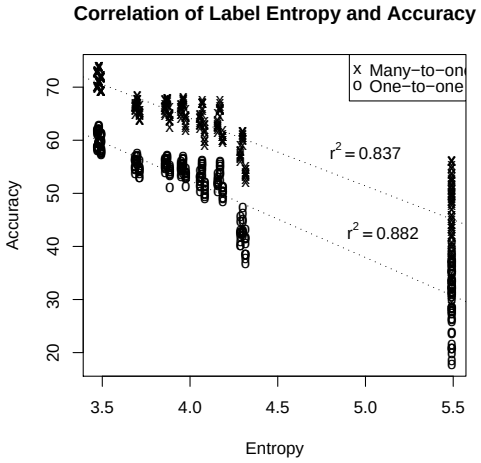

Figure 1. Correlation between conditional entropy $H(Y|X)$ and accuracy, for all runs on the 193k data set.

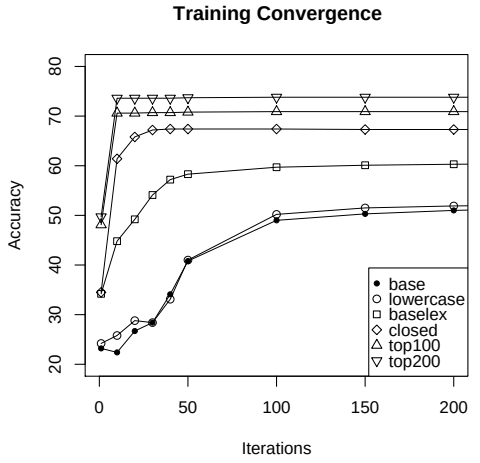

Figure 2. Mean accuracy for constraint sets over training iterations (only minor increases after 200), for 45-state bigram, 193k data set.

the hard rules forced errors, the most common words in each tag were covered fairly well by the grammar lists we used, and the extra reduction in uncertainty outweighed the more obscure errors. A similar effect is observed in Banko and Moore (2004), where they find quite significant gains by filtering out the rare tags of each word. This does not mean necessarily that only hard constraints are useful (indeed, Chang et al. (2007) finds soft constraints to be superior), but it seems they can be beneficial even when they oversimplify the facts, especially for a simple model that has little hope of labeling rare and difficult events correctly. We assume, too, that it would be more ideal to separate rules according to our confidence in them, and assign weights accordingly.

Finally we found that increased knowledge constraints lead to a reduction in the variance of model performance across runs, a major benefit given the problems of local extrema in most unsupervised methods and the difficulty of choosing an optimal model without annotated data. For our most constrained 'Top 100' and 'Top 200' model sets, the standard deviation of the accuracy was generally under 0.5 percentage points. Additional knowledge also constrained the training process, with accuracy converging in fewer iterations. Figure 2 plots the accuracy for models trained on the 193k data set, illustrating how the addition of rules leads to a steeper optimization surface for EM.

5.3. Labeling and Annotation

While the use of fully annotated data to label internal model states is a practical necessity of evaluation in this and similar work, such an artificial scenario is problematic for those real-world situations where classifier output needs to be interpreted or passed to another component in the pipeline (i.e. not just treated as clustering). In such cases, we face the question of how much labeled data is required to perform a quality label assignment.

To explore this issue, we labeled the output of different models trained on the 193k data set, but with only

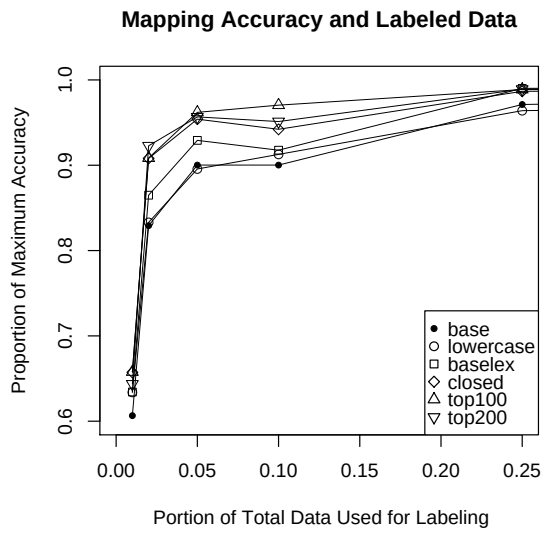

Figure 3. Accuracy convergence of many-to-one labeling methods, as increasing portions of the training data annotations are used to make label assignments, for models trained on the 193k corpus.

part of the annotated data available to generate the mappings. Figure 3 shows the results of the many-to-many method, plotting each data set proportion with the accuracy of the induced label mapping, relative to the best accuracy when all data was used.[5] As an example, consider the case of the unconstrained, lowercased model. With 10% of the data, or roughly 19k words, the labeling accuracy was 48.1, compared to 52.7 when the entire set is used, so that this partial-data assignment performs at 0.91 of its full accuracy.

Our first impression is that labeling performance converges relatively quickly, but we should note that even the 5% portion represents nearly 10,000 words of annotation. Still, with the more constrained knowledge sets, 90% of optimal accuracy is reached with only 2% of the data (4k words), so once again the use of prior knowledge is extremely beneficial in a practical setting.

6. Discussion

We have presented the view of prior knowledge as a reduction of uncertainty in the training of unsupervised classifiers, showing label entropy to be an effective and predictive measure of the contributions of that knowledge, and a means for assessment without annotated data. We found that quite basic domain knowledge can lead to significant performance improvements, with the additional benefits of faster training convergence, better stability, and reduced data requirements for label mapping. While the effects here are no doubt exaggerated by our impoverished models and data sets and the simplicity of the task, our results suggest that even the simple integration of prior knowledge is worthwhile where labeled data is lacking.

[5] One-to-one convergence was slightly faster, but the relative rates of the different constraints were similar.

Acknowledgments

We thank Rahul Bhagat, Abe Kazemzadeh, and the anonymous reviewers for comments on drafts of this paper, and Jorge Silva for fruitful discussion of the work.

This research was supported by the DARPA TRANSTAC program, grant #FA8750-06-1-0250, and by a USC Viterbi School of Engineering Doctoral Fellowship for the first author.

References

Banko, M., & Moore, R. C. (2004). Part of speech tagging in context. *Proc. COLING.*

Bilmes, J. (2004). *On soft evidence in Bayesian networks* (Technical Report UWEETR-2004-00016). University of Washington Dept. of Electrical Engineering.

Bilmes, J., & Zweig, G. (2002). The Graphical Models Toolkit: An open source software system for speech and time-series processing. *Proc. ICASSP.*

Chang, M., Ratinov, L., & Roth, D. (2007). Guiding semi-supervision with constraint-driven learning. *Proc. ACL.*

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* Wiley-Interscience.

Goldwater, S., & Griffiths, T. L. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. *Proc. ACL.*

Haghighi, A., & Klein, D. (2006). Prototype-driven learning for sequence models. *Proc. HLT-NAACL.*

Johnson, M. (2007). Why doesn't EM find good HMM POS-taggers? *Proc. EMNLP.*

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19.*

Merialdo, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics, 20,* 155—171.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Smith, N. A., & Eisner, J. (2005). Contrastive estimation: Training log-linear models on unlabeled data. *Proc. ACL*.

Toutanova, K., & Johnson, M. (2008). A Bayesian LDA-based model for semi-supervised part-of-speech tagging. *Proc. NIPS*.