

Unsupervised Approaches to Part-of-Speech Tagging

Five methodologies surveyed by Ankit K Srivastava

March 19, 2008

Introduction

Part-of-Speech (POS) Tagging is the process of assigning grammatical (syntactic and morphological) categories (e.g. noun, verb, adjective, person, verbal class, gender etc.) to naturally occurring text. Annotated natural language texts (Figure 1) are a useful preprocessing step for many natural language processing applications like parsing, information retrieval, and machine translation.

```
All/DT came/VBD from/IN Cray/NNP Research/NNP ./.
```

Figure 1: Labeled data

There are numerous strategies for designing POS taggers for a specific language; rule-based, probabilistic, hybrid. We focus on unsupervised approaches, i.e. learning tagging probabilities from unlabeled text (Figure 2). This has the potential to speedily scale to any language as it does not require copious amounts of labeled text (supervised training data) or an exhaustive list of hand-coded rules. Note that a small amount of labeled data in some form is still used as a bootstrap in many unsupervised approaches.

```
All came from Cray Research .
```

Figure 2: Unlabeled data

The tagging procedure is thus translated into a task (Figure 3) of finding a hidden structure (POS tags) in observed data (unlabeled text) by estimating model parameters. This paper describes and contrasts five existing unsupervised learning paradigms in the domain of POS tagging, published in the last ten years. Broadly speaking, these five categories are classified on the basis of parameter estimation techniques employed. These are **Expectation-Maximization (EM)**, **Clustering**, **Prototypes**, **Cross-lingual**, and **Bayesian**.

```
INPUT:   All came from Cray Research .
OUTPUT:  DT  VBD  IN   NNP  NNP   .
```

Figure 3: Unsupervised POS Tagging Task

After an exposition of each category, a general discussion section follows concluded with the list of references. This paper is for a seminar on Unsupervised Learning.

POS Tagging extending EM

Hidden Markov Models (HMM) which treat the tags as (hidden) states and the words of unlabeled text as output (observed) symbols are used as the underlying representation and the four papers in this category (Table 1) primarily use the Forward Backward algorithm which is an implementation of the EM strategy based on the Maximum Likelihood Estimation (MLE) principle in order to estimate the parameters (transition and emission probabilities) for the HMM based model. Another common feature in this approach is the use of a lexicon or a tagging dictionary, which is basically a compiled list of words and the set of possible parts-of-speech the word can have (Figure 4). Such a lexicon can easily be extracted from a standard dictionary.

| | |
|---------|------------|
| In | IN NNP |
| company | NN |
| their | PRP\$ |
| New | NNP JJ |
| year | NN |
| more | RBR JJR JJ |
| were | VBD |
| which | WDT IN |
| would | MD |
| who | WP |

Figure 4: Lexicon of words and their allowed tags

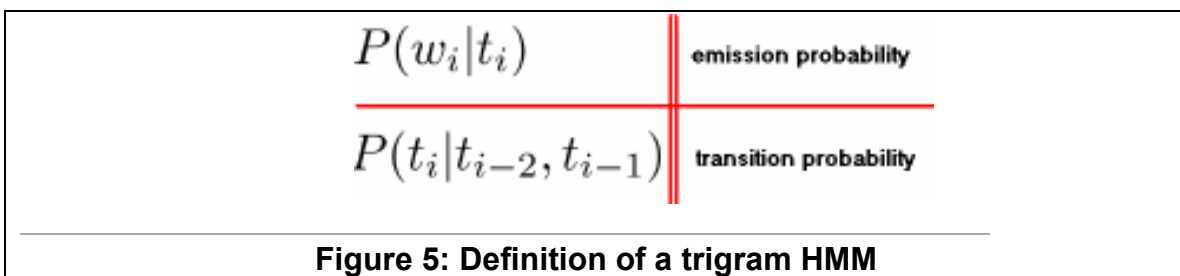
The methodologies differ amongst themselves in the specifics of the training (forward-backward parameter estimation) and tagging (algorithm used to compute the state sequence for the development data after the model has been trained), and in the way the lexicon is defined. Since each paper under review contains numerous experiments, a brief description of the main strategies undertaken in each of the four papers follows interspersed with comments on similarities and differences between them.

| PAPER | TITLE |
|-----------------------|--|
| Merialdo '94 | Tagging English text with a probabilistic model |
| Elworthy '94 | Does Baum-Welch re-estimation help taggers? |
| Banko & Moore '04 | POS tagging in context |
| Wang & Schuurmans '05 | Improved estimation for Unsupervised POS tagging |

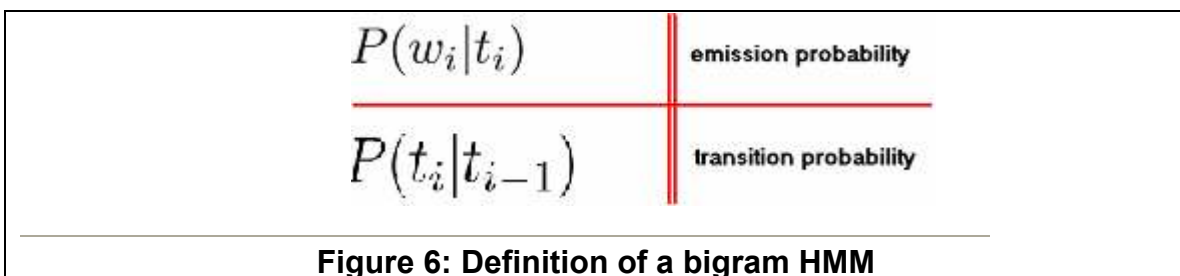
Table 1: Research Papers in the EM category

The main objective of **Merialdo, 1994** is to study the effect of EM on tagging accuracy when the training data is a mixture of labeled and unlabeled text. A trigram HMM (Figure 5) is defined using two different trainings – a pro-supervised (Relative Frequency (RF) counts interpolated with uniform distribution) and a pro-unsupervised (Maximum Likelihood (ML) / Forward

Backward training). A complete lexicon containing words from both the training and test corpus is used to estimate the emission probabilities, implying zero probability values for those word-tag pairs absent from the lexicon. Even for tagging, there are 2 strategies (found to perform similarly) – Viterbi tagging (computing the most probable tag sequence for a sentence) and ML tagging (computing the most probable tag for each word in a sentence, resulting in some linguistically impossible sequences). The paper also reports a strategy on constrained EM, i.e. adding constraints (fixing probabilities) on emission (word-tag) and transition (tag) during subsequent iterations of the Forward Backward algorithm. The paper concludes that EM helps when there is no labeled training data and hurts (deteriorates) in the presence of more than two thousand sentences of labeled text.



Elworthy, 1994 specifically examines the inner workings of the transition and lexical (emission) training modules in the EM-run HMM model. Viterbi algorithm is used to tag after training. The lexicon is compiled from the training data alone, assigning all open-class tags to the unknown words. Using a bigram HMM (Figure 6), this paper arrives at conclusions similar to the afore-mentioned Merialdo's paper, i.e. Baum-Welch re-estimation works better in the absence of labeled data. It extends on this by stating that a biased transition or emission module (probabilistic lexicon) is essential in the absence of labeled data. Just like the first paper, Elworthy also uses different settings of the transition and emission probabilities based on different factors like using probabilities from training data, normalizing lexical probabilities over word frequency or tag frequency, and uniform distributions. Elworthy does not however use a mixture (varying amounts) of labeled and unlabeled data like Merialdo, instead experimenting on an all (labeled) or nothing (unlabeled / uniform) settings along with using training and test data from different sources.



Banko and Moore, 2004 in addition to doing a comparative performance analysis by implementing other pre-existing strategies on the same data sets, perform experiments on the Contextualized trigram HMM tagger (Figure 7). They also use a filtered lexicon (derived from both training and test corpus), i.e. omitting those tags for a word which has a low probability of occurrence. Banko and Moore also report a way to stabilize the transition probabilities by biasing the initial model with unambiguous tag sequences (trigrams in which all three words can be assigned at most one tag) obtained from the tagged training data. Another contribution is sequential rather than simultaneous training of transition and emission probabilities, i.e. estimating transition probabilities while keeping the lexical probabilities constant followed by lexical re-estimation. This paper concludes that the quality of the lexicon (filtered lexicon) is an important part of the EM strategy. In the presence of a noisy lexicon the authors suggest the afore-mentioned sequential training.

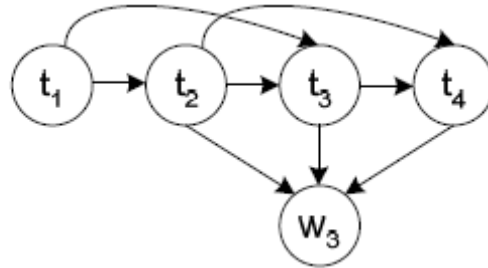


Figure 7: Structure of a Contextualized HMM tagger

Notice the additional dependence of an emission probability on the left and right tags in context.

Wang and Schuurmans, 2005 is another attempt to improve aspects of the EM estimation by avoiding over-fitting the transition and emission models. Unlike Banko & Moore's approach, they do not extend the model (contextualized HMM), just attempt to improve the estimation quality. Thus this work is similar to its two predecessors from 1994. Bigram HMM and a full lexicon (derived from both training and test data) are used. After model training, Viterbi tagging is implemented, just like in the previous three approaches. The transition probabilities are constrained by marginalizing over target (true) tag probabilities obtained from training data or approximated. The emission probabilities are constrained by ensuring similar words (word similarity computed as a feature vector of 100,000 most frequent words) have the same tag. The paper concludes that performance improvement from both the tag marginalization and similarity smoothed emission constraints prove that improvements to the actual estimation procedure (the E step and the M step) helps, just like the initial model (the model passed as input to the EM algorithm) tuning described in the previous three approaches.

POS Tagging by Clustering

This approach leverages distributional properties and co-occurrence patterns of text (similar words occur in similar contexts) by computing context vectors for each word to cluster words together in groups; groups which can then be assigned POS tags or word classes as groups. **Schutze, 1995** and **Clark, 2000** are representative of this category in our study. Although clustering on word similarity is not exactly exclusive to this group (Wang & Schuurmans, 2005 from the EM group also use it to constrain lexical probabilities), using this technique alone to induce POS tags on unlabeled data is a distinguishing feature of the distributional POS tagging methodology. Clustering words and inducing class names (hidden structure) on the resultant clusters replace Markov models and algorithms to iteratively estimate the model's hidden parameters.

The key characteristics to be considered here are how the context vectors are defined, size of the context vectors (number of dimensions), metric used to compute vector similarity (i.e. make clusters), and how the tags or word classes are induced on the clusters. Classifying rare & ambiguous words and sparse data are the main impediments to this strategy.

| Characteristic | Schutze '95 | Clark '00 |
|---------------------------------|---|---|
| <i>Vectors</i> | LEFT vector comprising of left neighbors and RIGHT vector comprising of right neighbors | Context is a probability distribution of left and right neighbors |
| <i>Dimensions</i> | 250 words | Variable (calculate for words occurring more than 50 times) |
| <i>Clustering</i> | Cosine Similarity | Kullback-Leibler Divergence |
| <i>Tag Induction</i> | 4 different inductions on basis of word, vectors of the word's neighbors, etc. | Start with K clusters and iteratively merge them |
| <i>For Sparse Data</i> | Singular Value Decomposition | Iterative algorithm |
| <i>For Rare & Ambiguous</i> | Ambiguous words – consider contexts of word tokens rather than word types Rare words - Contexts with rare words, freq < 10 were excluded | Ambiguous words – context is a linear combination Rare words – assume it's unambiguous |

The clustering approach is thus more unsupervised than the EM approach (in terms of requiring no labeled data whatsoever), more linguistic (not undermined by re-estimation). However one huge disadvantage is evaluation, i.e. there is no test corpus represented in the cluster format.

POS Tagging with Prototypes

$$\begin{aligned} \mathcal{L}(\theta; D) &= \sum_{x \in D} \log p(x|\theta) \\ &= \sum_{x \in D} \log \sum_y p(x, y|\theta) \end{aligned}$$

Figure 8: Model Representation

X: words, y: labels, theta: parameters, D: observed data

Haghighi and Klein, 2006 entitled “Prototype-Driven Learning for Sequence Models” has its own category in our survey and is a cross between the EM and Clustering approach. Prototypes are more meaningful than clusters and easier to evaluate (since small size). A few canonical examples or prototypes are collected (one for each target tag) and then propagated across the corpus of unlabeled data. As a model, Markov Random Fields (the undirected equivalent of HMM) are used to run a trigram tagger. There is no lexicon required. A gradient-based search with the forward-backward algorithm is used to maximize the log-linear model parameters.

| Label | Prototype | Label | Prototype |
|-------|-------------------------|-------|---------------------------|
| NN | % company year | NNS | years shares companies |
| JJ | new other last | VBG | including being according |
| MD | will would could | -LRB- | -LRB- -LCB- |
| VBP | are 're 've | DT | the a The |
| RB | n't also not | WP\$ | whose |
| -RRB- | -RRB- -RCB- | FW | bono del kanji |
| WRB | when how where | RP | Up ON |
| IN | of in for | VBD | said was had |
| SYM | c b f | \$ | \$ US\$ C\$ |
| CD | million billion two | # | # |
| TO | to To na | : | - : ; |
| VBN | been based compared | NNPS | Philippines Angels Rights |
| RBR | Earlier duller | " | " ' non- |
| VBZ | is has says | VB | be take provide |
| JJS | least largest biggest | RBS | Worst |
| NNP | Mr. U.S. Corp. | , | , |
| POS | 'S | CC | and or But |
| PRP\$ | its their his | JJR | smaller greater larger |
| PDT | Quite | WP | who what What |
| WDT | which Whatever whatever | . | . ? ! |
| EX | There | PRP | it he they |
| " | " | UH | Oh Well Yeah |

Figure 9: POS Prototype List

Cross lingual POS Tagging

The fourth category of approaches works on multilingual data. So far we have dealt with monolingual (basically English) POS taggers. One strategy to overcome the lack of training data (i.e. zero or minimum labeled text) prevalent in the unsupervised building of a tagger in a language is to use information from a POS tagger of another language. Thus the transition and emission probabilities of a target language HMM are estimated (or projected) from bilingual data and from source language tagger information (probabilities). The four papers in this group are as given in Table 2.

| PAPER | TITLE |
|-------------------------|--|
| Yarowsky & Ngai '01 | Inducing multilingual POS taggers |
| Cucerzan & Yarowsky '02 | Bootstrapping multilingual POS tagger |
| Hana et al. '04 | Resource-light approach to Russian morphology |
| Feldman et al. '06 | Cross-language morphological annotation transfer |

Table 2: Research Papers in the Cross lingual category

This approach is closer to the EM approach than any other category in this study in that it uses a HMM as an underlying model and a lexicon of sorts is used to estimate emission probabilities. Of course the most substantial distinguishing characteristic is the existence of a readymade tagger in another language or labeled data in another language. SL refers to the source language or the language for which we already have a tagger or tagged data. Target Language or TL refers to the language for which we are designing the tagger. Among the four papers, 2 pairs are written by the same group of authors. Also for space reasons, the first 2 papers – **Yarowsky and Ngai, 2001** and **Cucerzan and Yarowsky, 2002** are described together followed by the second set – **Hana et al., 2004** and **Feldman et al., 2006**.

| Characteristic | Yarowsky & Ngai '01 | Cucerzan & Yarowsky '02 |
|-------------------------------|---|---|
| <i>Main Idea</i> | Align TL and SL parallel data, tag SL, project tags on TL | Multilingual tagger using a dictionary and a grammar book |
| <i>Multilingual Resources</i> | -Parallel data -Word Aligner, IBM Model 3 -SL Tagger | -Bilingual dictionary -Paradigms from TL grammar book |
| <i>Tagger Model</i> | Bigram HMM | Trigram HMM |
| <i>Transition Probs.</i> | Pseudo-divergence weighting, estimated separately from emission | Iteratively retrained; tag probs. from unweighted mixture of SL word tags in dictionary |
| <i>Emission Probs.</i> | Lexical prior estimation from projected tagged SL data | Iteratively retrained through dictionary, inflection rules |

Unsupervised Approaches to POS Tagging

| Characteristic | Hana et al. '04 | Feldman et al. '06 |
|-------------------------------|---|--|
| <i>Main Idea</i> | Morphologically analyze and tag TL using labeled SL, build subtaggers (for a group of grammatical categories) to combine into main tagger | Transfer morph. annotations from SL (multiple languages) to TL (multiple languages), using TL grammar book, annotated SL, and multilingual transfer strategies |
| <i>Multilingual Resources</i> | -SL Tagger -TL morphological analyzer | -SL Tagger -TL Morphological Analyzer |
| <i>Tagger Model</i> | Trigram HMM | Trigram HMM |
| <i>Transition Probs.</i> | Borrow from SL, with some translation rules | Borrow and interpolate from multiple labeled SL |
| <i>Emission Probs.</i> | Lexicon created from morph. analyzer, uniform probs. | Lexicon from morph. analyzer, identify cognates in languages, |

The main difference between Yarowsky's papers (group 1) and Brew's papers (group 2) in the cross lingual category is that the former utilizes parallel data and bilingual lexicons, while the latter deals with linguistically related languages (same family) and does not use parallel or bilingual data. Both sets of approaches are shown to scale to multilingual (i.e. more than 2 languages) domain since Feldman et al., 2006 experiments with Polish, Czech, and Russian data while Cucerzan and Yarowsky, 2002 experiments with English, Romanian, Kurdish, and Spanish languages. Both groups also utilize inflectional rules or paradigms extracted from grammar reference books.

POS Tagging using Bayesian Learning

Bayesian learning models for POS tagging have slowly been picking up in direct retaliation to the weaknesses of HMM models and EM estimations. Bayesian statistical principles differ from the traditional MLE (Maximum Likelihood Estimation) principle in that the former integrates over all possible parameter values rather than simply seeking a parameter set which maximizes the probability of tag sequences given unlabeled observed data. Note that this, like the prototype sequence models (in fact even more so) is an entirely new family of models and, has thus its own set of modeling, estimation, evaluation techniques.

Toutanova and Johnson, 2007 extends the Latent Dirichlet Allocation (Blei et al., 2003) model for a semi-supervised POS tagging task. Both sparse data and ambiguity class problems are attempted to be dealt with in this new model. This graphical model also utilizes the neighboring context words as in the

Unsupervised Approaches to POS Tagging

Clustering and Prototype categories. A lexicon is also used. There are four neighbors for each word. It is basically an amalgamation of several different models like the ambiguity class and the word context model. Parameters of the variational distribution are estimated in an EM-like iterative procedure. Along with other things, it is concluded that the context vectors (used in the clustering category) function better in a semi-supervised scenario such as the one in this paper rather than determining tag sequences directly from unlabeled data.

Goldwater and Griffiths, 2007 also explore the Bayesian approach to unsupervised POS tagging with the model having the structure of a trigram HMM added on symmetric Dirichlet priors over the transition and emission distributions. Like the above paper, this approach is also a combination of syntactic clustering and POS disambiguation (benefits of integrating over parameters) and uses a tagging dictionary. Gibbs sampling is used to perform the model inference (in place of EM estimation in MLE based HMM).

Johnson, 2007 as apparent from the paper's title, is a direct comparison of EM and Bayesian learning. This paper aims to overcome the shortcomings of EM by using Gibbs Sampling and Variational Bayes estimators. It also points toward yet another improvement in the EM strategy itself – reducing the number of hidden HMM states (e.g. abandoning low frequency POS tags) to help improve the performance. Bigram HMM is used. A noteworthy observation is that this paper describes 1000 iterations (with noticeable changes) of the EM, while most of the approaches stick to somewhere between ten and twenty. Variational Bayes is found to perform better than Gibbs Sampling.

All the three papers in this group are merely exploring the relatively new Bayesian approach to unsupervised POS tagging, comparing it with the existing clustering and EM approaches, and generally implementing very similar Bayesian techniques. There is still much to be understood and experimented upon in the Bayesian domain.

Discussion

Now that the different types of unsupervised approaches taken up in this study have been commented upon, a brief note about EM implementation. While this paper is mainly a survey of research papers, the most basic unsupervised strategy – HMM estimated using the forward backward algorithm and a full lexicon – has been implemented in PERL and experimented (settings similar to that in Merialdo, 1994) on the Penn Treebank as part of an assignment for this seminar. Both bigram and trigram Hidden Markov Models were designed. Certain observations were made in general about the procedure. There is a vast difference between writing mathematical formulae and actually translating them into computer code, as was found in writing implementations of the model initialization, Forward-Backward training and Viterbi decoding algorithms. The

Unsupervised Approaches to POS Tagging

process is computationally expensive. This gives us pause as to what would be the costs involved with Bayesian modeling which seem to be more complex than EM. As mentioned in the papers, the initial model (amount of labeled and unlabeled data) and the quality of the lexicon (expanding both training and test data) greatly affect the tagger performance. Table 3 and Figure 10 show the observation results and performance graph respectively. Although they perform significantly lower than Merialdo, 1994 (experiments on a different data set and lexicon), the relative curves are comparable.

| Iter # | Number of tagged sentences used for the initial model | | | | |
|--------|---|----------------|----------------|-----------------|----------------|
| | EXPT1 [0] | EXPT2 [100] | EXPT3 [500] | EXPT4 [1000] | EXPT5 [All] |
| 0 | 77.85 | 90.47 | 93.43 | 93.18 | 93.09 |
| 1 | 70.00 | 76.22 | 79.44 | 81.03 | 83.82 |
| 2 | 70.39 | 76.52 | 76.95 | 78.67 | 81.20 |
| 3 | 72.27 | 77.51 | 77.25 | 77.51 | 79.66 |
| 4 | 73.43 | 78.15 | 76.78 | 76.82 | 76.48 |
| 5 | 73.18 | 77.98 | 76.61 | 77.17 | 76.52 |
| 6 | 73.35 | 77.85 | 76.31 | 77.04 | 76.48 |
| 7 | 73.52 | 77.72 | 76.14 | 76.99 | 76.39 |
| 8 | 73.43 | 77.64 | 76.05 | 76.99 | 76.48 |

Table 3: Accuracy on the development data after the EM training using various initial models

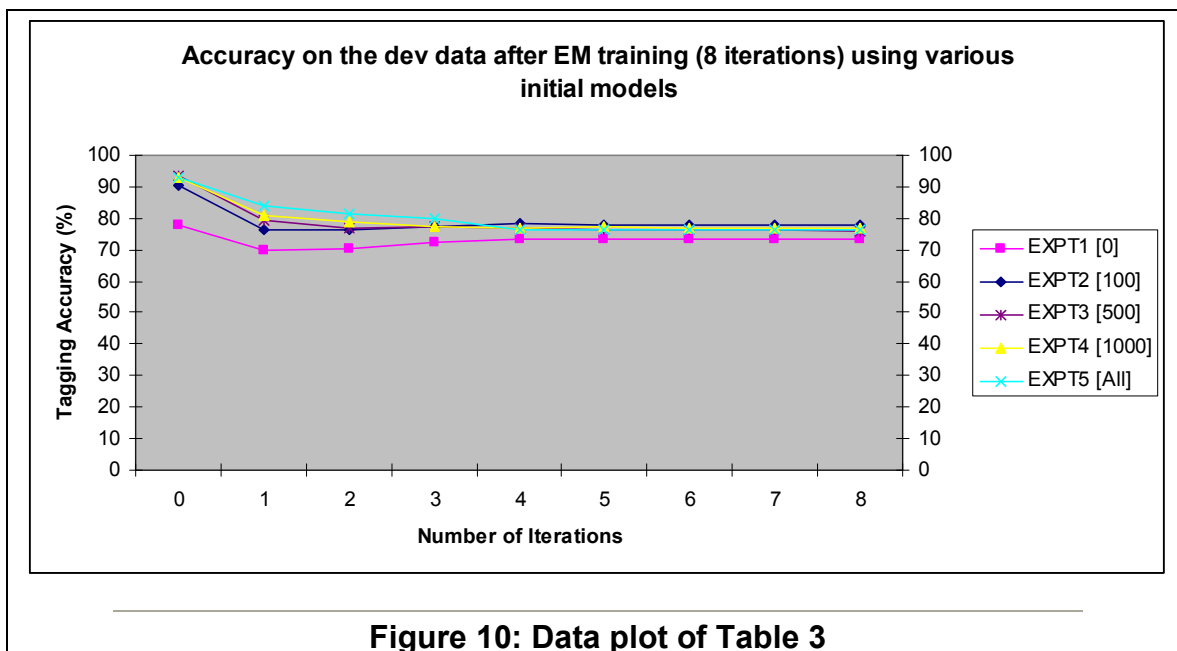


Figure 10: Data plot of Table 3

Looking back at the different strategies explored in the past ten years, a possible future direction seems to be exploring Bayesian POS tagging in the domain of cross lingual annotation. Bayesian approaches look promising and

Unsupervised Approaches to POS Tagging

perform better than EM trained HMM. Moreover, the integrating over parameter approach, in my opinion, could greatly benefit the performance of cross lingual POS and morphological taggers. After all, multilingual annotators are in high demand as a pre-processing step for a multitude of natural language processing and computational linguistic applications. Besides, combining the linguistically rich annotations of a cross lingual approach with the superior performance of a Bayesian approach has the potential to bring good results. Another prospective direction is to pipeline some of these approaches (like clustering and EM, clustering and cross lingual), to gauge the system performance.

Conclusion

Current trends in unsupervised POS tagging tend toward graphical models (Bayesian priors). Nevertheless, improvements are still being made to the basic EM implementation both in the initial model and inside the E and M steps. In my opinion, EM has not yet been exhausted and work should continue on it alongside Bayesian experiments. It should also be noted that there a few approaches not covered in this study like Smith and Eisner, 2005. To conclude, unsupervised and semi-supervised POS tagging has great potential and should continue receiving new and improved model estimations in its literature pool.

References

Banko, M. and Moore, R. C. (2004). Part of Speech Tagging in Context. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Machine Learning*, 42(3): 993-1022.

Clark, A. (2000). Inducing Syntactic Categories by Context Distribution Clustering. In *Proceedings of the 4th Conference on Computational Natural Language Learning, CoNLL*.

Cucerzan, S. and Yarowsky, D. (2002). Bootstrapping a Multilingual Part-of-Speech Tagger in One Person-day. In *Proceedings of the 6th Conference on Computational Natural Language Learning, CoNLL*.

Elworthy, D. (1994). Does Baum-Welch Re-estimation Help Taggers? In *Proceedings of the 4th ACL Conference on Applied Natural Language Processing, ANLP*.

Unsupervised Approaches to POS Tagging

Feldman, A., Hana, J., and Brew, C. (2006). Experiments in Cross-Language Morphological Annotation Transfer. In *Proceedings of the Computational Linguistics and Intelligent Text Processing, CICLing*.

Goldwater, S. and Griffiths, T. (2007). A Fully Bayesian Approach to Unsupervised Part-of-Speech Tagging. In *Proceedings of the Association for Computational Linguistics, ACL*.

Haghighi, A. and Klein, D. (2006). Prototype-Driven Learning for Sequence Models. In *Proceedings of the Human Language Technology Conference of the North American Association for Computational Linguistics, HLT/NAACL*.

Hana, J., Feldman, A., and Brew, C. (2004). A Resource-light Approach to Russian Morphology: Tagging Russian using Czech Resources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.

Johnson, M. (2007). Why doesn't EM find good HMM POS-taggers? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.

Merialdo, B. (1994). Tagging English Text with a Probabilistic Model. *Computational Linguistics*, 20(2): 155-171.

Schütze, H. (1995). Distributional Part-of-Speech Tagging. In *Proceedings of the European chapter of the Association for Computational Linguistics, EACL*.

Smith, N. A. and Eisner, J. (2005). Contrastive Estimation: Training log-linear Models on Unlabeled Data. In *Proceedings of the Association for Computational Linguistics, ACL*.

Toutanova, K. and Johnson, M. (2007). A Bayesian LDA-based Model for Semi-supervised Part-of-Speech Tagging. In *Proceedings of Neural Information Processing Systems, NIPS*.

Wang, Q. I. and Schuurmans, D. (2005). Improved Estimation for Unsupervised Part-of-Speech Tagging. In *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, IEEE NLP-KE*.

Yarowsky, D. and Ngai, G. (2001). Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. In *Proceedings of the North American chapter of the Association for Computational Linguistics, NAACL*.