Semi-supervised Part-of-speech Tagging in Speech Applications

Richard Dufour, Benoit Favre

Laboratoire d'Informatique de l'Université du Maine,
Avenue Laënnec, 72085 Le Mans Cedex 9, France
{richard.dufour,benoit.favre}@lium.univ-lemans.fr

Abstract

When no training or adaptation data is available, semi-supervised training is a good alternative for processing new domains. We perform Bayesian training of a part-of-speech (POS) tagger from unannotated text and a dictionary of possible tags for each word. We complement that method with supervised prediction of possible tags for out-of-vocabulary words and study the impact of both semi-supervision and starting dictionary size on three representative downstream tasks (named entity tagging, semantic role labeling, ASR output post-processing) that use POS tags as features. The outcome is no impact or a small decrease in performance compared to using a fully supervised tagger, with even potential gains in case of domain mismatch for the supervised tagger. Tasks that trust the tags completely (like ASR post-processing) are more affected by a reduction of the starting dictionary, but still yield positive outcome.

Index Terms: part-of-speech tagging, semi-supervised training, bayesian methods.

1. Introduction

Part-of-speech (POS) tags represent the morphological and syntactic function of words. They have been shown to be useful in a range of text and speech applications, including syntactic and semantic parsing, named entity tagging, summarization, language modeling, machine translation, word sense disambiguation (see for instance [1, 2, 3, 4]). Yet, most high-performance part-of-speech tagging approaches are supervised and require adaptation data to keep high accuracy on new domains.

Semi-supervised learning can help reducing the performance gap when going to applications and languages with little or no data. In particular, for the case of part-of-speech tagging, a viable approach is to use a dictionary of possible tags for each word and let the machine learning algorithm devise a tagging of an unseen word sequence by looking at corpus-wide contextual preference for each word-tag pair [5]. The most straightforward implementation of such an approach is to consider an HMM over emission $P(tag|word)$ and transition $P(tag|previous\_tag)$ probabilities and use the EM algorithm to maximize the likelihood of the data under that model. In a realistic context, though, the complete dictionary is unlikely to be available and one has to provide possible tags for new words. More recent approaches try to overcome this problem by predicting possible tags for new words from morphological features of said words. For instance, in English, a word ending in "s" is most likely to be a noun (plural) or verb (third person singular present).

In this paper, we are interested in the impact of semi-supervised part-of-speech tagging on downstream tasks that make use of these tags. We study two kinds of tasks: tasks that make use of POS-tags as features of a classifier, possibly learning the mistakes of the tagger, and tasks that use POS-tags directly for their meaning, completely trusting the tagger. Our main contribution is to try to answer the following questions:

- What is the importance of POS-tags for each task?
- How does a semi-supervised tagger impact each task compared to a supervised tagger?
- What is the degradation due to unseen words in the starting dictionary?

We first present semi-supervised POS-tagging and focus on an approach using Bayesian estimation to determine good parameters for the HMM (Section 2). We complement this approach with a supervised method for determining the possible tags of a new word (Section 2.2). Then, we study the impact of a tagger designed around those principles on the following tasks: the task of part-of-speech tagging itself, named entity tagging, semantic role labeling and homophone post-processing in ASR output (Section 3).

2. Semi-supervised POS tagging

2.1. Model

Part-of-speech tagging is modeled as the prediction problem of finding the sequence of labels $\mathbf{y} = y_1 \dots y_n$ given a sequence of input feature vectors $\mathbf{x} = \mathbf{x}_1 \dots \mathbf{x}_n$. A probabilistic view of this problem yields, for instance, the following first order HMM:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, P(y_1|\mathbf{x}_1) \prod_{i=2}^{n} P(y_i|\mathbf{x}_i) P(y_i|y_{i-1})$$

Assuming that $\mathbf{x}$ is the sequence of words and given annotated training data, one can compute the emission $P(y_i|\mathbf{x}_i)$ and the transition $P(y_i|y_{i-1})$ by counting the frequency of those events in the corpus [5]. The *supervised* tagger we use in our experiments, called LIA_TAGG[1], is based on this approach with a trigram tag model (second order HMM).

If no annotated data is available but a dictionary of words and their possible tags can be used as linguistic knowledge, [5] proposed to use Maximum Likelihood estimation over sequences of untagged words of the emission and transition probabilities using the Baum Welsh algorithm. Arguing that this method is prone to over-fitting, [6] proposed to perform this estimation in a fully Bayesian framework by integrating over all possible values of the model parameters, resulting in large performance improvements (the emission and transition probabilities are set to follow multinomial distributions with Dirichlet

[1] http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/bechet/download_fred.html

priors). Parameter inference in this setup does not have an analytical solution, and it has to be performed through estimation methods such as Gibbs sampling or Variational Bayes [7].

In our experiments, we use the *carmel* toolkit for Bayesian inference on finite state transducer cascades which implements approximate Gibbs sampling as described in [8].

2.2. Generalizing possible tags for new words

A drawback of the semi-supervised approach is that the dictionary has to include a set of possible tags for all words being tagged. Tagging new corpora inevitably brings new words (called out-of-vocabulary or OOV words) for which we have to determine a set of compatible tags, called *ambiguity class*. We devise this set by predicting $P(y_i|\mathbf{x}_i)$ and using the tags for which this probability exceeds a given threshold. We extract morphological features on words of the existing dictionary and train an Adaboost discriminative classifier[2] to approximate the posterior probability of tags:

$$P(y_i|\mathbf{x}_i) = \left[exp\left(-2m \sum_{j=1}^{m} \lambda_j f_j(\mathbf{x}_i, y_i) \right) \right]^{-1}$$

where m is the number of training iterations and $f_j(\cdot)$ are weak learners (here, one-level decision trees) weighted by λ_j. For features, we use:

- Prefixes and suffixes of length up to 4.
- Factorized letter class templates (lower case, upper case, digit, punctuation) – i.e., *Interspeech'10* $\rightarrow$ *Aa'd.*
- The length of the word.

The idea of generating ambiguity classes from morphological features is not new and was first proposed in a generative model by [9]. [10] applied Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (PLSA) parameter estimation of taggers and also introduced morphological features to determine ambiguity classes. It was also shown by [11] that EM-trained HMMs (ML-HMM) can perform at the state of the art when properly initialized with linguistic knowledge and, among others, morphological models.

2.3. POS tagging performance

As a sanity check, we implemented a semi-supervised Bayesian tagger with carmel and applied it to the CoNLL 2000 POS tagging corpus[3] which contains 221k words for training and 47k words for testing, tagged with 44 different tags. The training set lexicon counts 19k different words and the test lexicon is 8k of which 33% are new words. We used the training set to design a word-tag lexicon and performed 100 iterations of boosting on that lexicon to predict possible tags for out-of-vocabulary words of the test set. As a result, 63% of the test words get the exact ambiguity class they belong to. Note that this experiment is not very realistic since only *observed tags* are in the lexicon and a word might be subject to more tags in other domains. Results, with Bayesian tagging shown in Table 1 reveal that though the tagger is far from supervised counterparts when new words are not processed, it gets better with appropriate prediction of ambiguity class. This table also shows the choice of the Bayesian approach is not critical since an EM tagger performs almost as well when morphological prediction is used on new words.

[2] http://code.google.com/p/icsiboost

[3] http://www.cnts.ua.ac.be/conll2000/chunking/{train,test}.txt.gz

System	Dummy	Any	Sup.	Oracle	EM
Accuracy	89.0	92.2	94.7	95.1	93.6

Table 1: Tagging accuracy on the CoNLL 2000 dataset. Results with Bayesian training are shown according to the way unknown words are processed: always wrongly labeled (Dummy); with a uniform probability on tags, letting context decide (Any); using Adaboost-predicted ambiguity classes (Sup.); or with the actual ambiguity class from the test set (Oracle). EM training results with predicted ambiguity classes are given for comparison. On this dataset, supervised taggers perform in the range 97%-99%.

We then applied the same approach on a realistic scenario for part-of-speech tagging of French broadcast news transcripts from the ESTER corpus [12]. This corpus contains 100 hours of manually transcribed speech from various French-speaking radio broadcasts, totaling about 1.1M words. We manually annotated two sets of 1000 words with part-of-speech tags following the LIA_TAGG tagset, 103 tags denoted with gender, number information and proper name classes unlike the French treebank tagset. It also includes a "MOTINC" tag used both for unclassifiable proper names and unknown words. The first set (*test1*) is transcripts mostly from professional speakers from the beginning of a show whereas the second set (*test2*) is a selection of sentences from harder parts of the corpus with ungrammaticality and higher variability of vocabulary. To match ASR output, the transcripts are all lower cased and stripped of punctuation.

The supervised POS tagger LIA_TAGG comes with a 268k word dictionary with ambiguity classes for each word. Overall, the vocabulary of the ESTER corpus is about 37k words of which 13% are not in the LIA_TAGG dictionary. The supervised tagger was trained on about 600k words of newspaper text which neither match the style of speech nor the epoch of the ESTER data. In fact, this training data is not available anymore, preventing us from training more accurate supervised taggers and justifying for less supervised methods. For the semi-supervised POS-tagger, we predicted ambiguity classes for out-of-vocabulary words using the LIA_TAGG dictionary for training and performed Bayesian estimation of the parser parameters using the same techniques as described previously.

System	**OOV-Sup.**	**OOV-Any**	**%OOV**
Supervised	91.7	91.7	4.1
Semi (268k)	80.8	79.6	4.1
Semi (5k)	79.9	73.5	12.0
Semi (2k)	78.9	68.1	20.0

Accuracy on test1

System	**OOV-Sup.**	**OOV-Any**	**%OOV**
Supervised	87.3	87.3	7.1
Semi (268k)	77.1	73.8	7.1
Semi (5k)	71.2	61.7	21.0
Semi (2k)	68.0	60.7	26.0

Accuracy on test2

Table 2: POS-Tagging results on ESTER data comparing supervised and semi-supervised taggers. Performance values are given for different sizes of starting dictionary (all, 5k, 2k) and according to whether out-of-vocabulary words (OOV) are specifically processed (OOV-Sup.) or not (OOV-Any).

Table 2 compares the results of the supervised and the semi-

supervised taggers on the two test sets. On either set, the supervised tagger performs better in term of tagging accuracy. We tried to vary the number of words in the initial dictionary to the most frequent 5k and 2k words (according to statistics in the dictionary), the rest being treated as new words. Using the 5k (resp. 2k) dictionary, 88% (resp. 95%) of the ESTER vocabulary is unknown, making the task of predicting ambiguity classes more difficult. This results in a drop in accuracy as expected, but morphological information is valuable as evidenced by results when OOV words are ignored. The increased difficulty of *test2* results in a higher number of OOV words and generally decreased performance. In the following sections, we conduct experiments to see if the same trend can be observed in applications that use the generated POS tags.

3. Downstream tasks

3.1. Named entity tagging

Here, we assess the impact of semi-supervised POS tagging when used as features for named entity (NE) tagging. We follow the NE task from the ESTER 2005 campaign on French broadcast news [12] (about 1M words, 66k entities for training and 100k words, 6k entities for testing). Our system is a straight-forward Conditional Random Field trained with the MIRA algorithm [13]. For each input word, it predicts one of the nine NE label (person, company, group, quantity, date...) with start, inside and outside sub-labels (totaling 19 labels), using combined word and POS-tag bigram features. We ran the supervised and semi-supervised taggers to label NE data with POS tags on reference text (train and test) and ASR output (test only). The ASR output is LIUM's submission at ESTER 2005 with a WER of 23.6 [14]. The repartition of unknown words on this corpus is the same as in the POS-tagging experiment.

F-score performance measures detailed in Table 3 show that POS-tags improve over words only and that semi-supervised tags yield comparable or better results than supervised tags, probably due to two factors: a lack of adaptation data for the supervised tagger and no specific processing of unseen words by that tagger (its output contains 34% more MOTINC tags than the semi-supervised one). It is also likely that the NE prediction system learns the errors of the POS-taggers which limits the impact of the quality of POS tags themselves. If you follow a supervised setup, this also suggests running a pass of self-training (decode and retrain the POS tagger on the decoded data). Named entity taggers using POS tags only (no words) follow the same trend. We also performed named-entity tagging with the same reduced dictionaries as used in the previous section (5k, 2k). Those dictionaries yield worse performance than the supervised tagger, especially when words are not used as features.

Origin of tags	Text (w+t)	Text (t)	ASR (w+t)
No tags	70.70	n/a	58.52
Supervised	75.77	56.09	65.69
Semi (268k)	77.00	60.53	67.03
Semi (5k)	75.28	53.97	67.18
Semi (2k)	74.34	40.72	64.54

Table 3: F-score of the named-entity tagger on reference transcript (lower-cased without punctuation) and ASR output. Feature sets can be words and tags (w+t), tags only (t), or words only (No tags).

3.2. Semantic role labeling

Our second test bed is the use of part-of-speech tags for semantic role labeling. For this experiment, we rely on the MEDIA corpus of French telephone conversations in the tourism domain, annotated with flat concepts [15]. It contains 13k sentences, 31k concepts for training 3k sentences, 9k concepts for testing. The corpus is made of 121k words, with a vocabulary of 2.3k words from which 12% are unknown from the original LIA_TAGG dictionary, 54% from the 5k dictionary and 69% from the 2k dictionary. Concepts cover basic hotel reservation informations such as the number and type of rooms the caller wants to reserve, where and when, but also reservation constraints, side information and dialogue specific elements. Following the named-entity tagging framework, we build a simple system to predict for each input word one of the 67 concepts observed in the training data, resulting in 145 labels with begin-inside-outside sub-labeling. We use the same classifier as for NE-tagging and extract similar features based on POS-tag and word sequences. This system is very similar to the LIA's concept segmenter described in [16].

Origin of tags	Words+tags	Tags only
No tags	82.72	n/a
Supervised	83.66	60.82
Semi-supervised (268k)	83.64	66.31
Semi-supervised (5k)	83.11	62.83
Semi-supervised (2k)	82.77	53.56

Table 4: F-score results for Semantic Role Labeling (SRL) without tags, with tags from the supervised tagger, and with tags from the semi-supervised tagger. Two feature sets are considered: words and tags, an tags only.

Results presented in Table 4 show that POS-tags do not seem to be as important as for named-entity tagging, due to the task being very dependent on lexical information ("room" is a very good cue for detecting "number of rooms"). It is also notable that the supervised and the semi-supervised POS tags result in almost identical gains. The experiment with tag-only features reveals that, as in the named-entity tagging case, the semi-supervised tagger gives better generalization, but as the original dictionary size decreases, performance drops below that of the supervised tagger.

3.3. Homophone post-processing in ASR output

Our third application deals with correcting ASR output when the language model failed to chose the right homophone for a given pronunciation. This is a problem in French because, for example, plural forms are pronounced the same way as singular words even though they are written differently. In this task, part-of-speech tagging is used in order to predict the gender and number of words, which helps picking their right inflected form. Here, we evaluate the approach described in [17] for predicting the inflection of past participles. In French, they are mainly agreed to the subject or the object of the verb depending on the auxiliary, but the acoustic evidence for gender and number might be relatively far and seldom. In addition, long strands of ASR errors make finding this evidence even harder.

The system first uses part-of-speech tagging on the ASR output, and then focuses on past participles. For each of them, a classifier (Adaboost) predicts gender and number given the lemma of the verb and POS tag n-grams in a window of 16 words around the target. Then the outcome is used along with

confidence measure thresholds and acoustic filtering to decide whether or not to change the inflected form. The system is trained on reference text tagged with the POS tagger. Errors of the tagger result in a direct impact on which examples are used for both training and testing. Experiments are run on the ESTER dataset using the LIUM system for generating ASR output. LIA_TAGG results in 55k training examples and 2.2k testing examples, while the semi-supervised tagger generates 54k (2.1k) training (testing) instances. By taking the most frequent classes of homophone errors, a nearly 21.5% relative reduction of word error rate could be possible.

Table 5 presents homophone post-processing results on past participles using recall, precision and correction rate (number of good corrections compared to the number of words that had to be corrected). They show that the semi-supervised approach is almost as good as the supervised approach, although when the dictionary size is reduced to a small set of frequent words, recall is reduced which has a large effect on correction rate. The fact that correction rate is still positive is very encouraging because even with very low resources the semi-supervised POS tagger helps correcting mistakes of the language model.

System	Recall	Precision	Correction rate
Supervised	45.63	60.91	34.60
Semi (268k)	46.39	59.80	34.22
Semi (5k)	25.48	56.78	16.35
Semi (2k)	12.90	49.32	6.08

Table 5: Performance of homophone post-processing, trained given the supervised POS tagger and the semi-supervised tagger with varying dictionary.

The task of homophone post-processing is mainly affected by gender and number decisions and less by the actual lexical class of the word. We performed an extra experiment where we break down POS tagging errors (on the test2 dataset) according to those two factors. Results, presented in Table 6, show that the gap between supervised and unsupervised (268k) POS-tagging is less important when focusing on specific information of tags compared to the whole tag. The gap widens with the 5k and 2k dictionaries which can explain the drop in correction rate when using those on homophone correction.

System	All	Lexical class	Num./Gender
Supervised	87.3	88.7	90.3
Semi (268k)	77.1	82.3	87.7
Semi (5k)	71.2	76.2	79.0
Semi (2k)	68.0	73.6	77.6

Table 6: POS-Tagging results on ESTER data comparing supervised and semi-supervised taggers (268k, 5k and 2k) on the whole tags (All), on lexical class only and on gender and number information.

4. Conclusion

In this paper, we evaluated the effect of semi-supervised POS tagging on downstream tasks compared to the use of a supervised tagger. We used a Bayesian approach that only requires a set of possible tags for each word, and no annotated training data. We extended it with supervised prediction of possible tags for out-of-vocabulary words and studied the effect of drastic reduction of the starting vocabulary size. On tasks like named-entity tagging and semantic role labeling, the reduction of tagging performance is compensated by the downstream system learning the errors of the tagger. Semi-supervised tagging can even improve performance in out-of-domain conditions but reducing the starting dictionary tends to cut the benefit, and degrade performance when not enough starting knowledge is available. On a task that completely trusts POS tags like ASR output post-processing, the effect is more pronounced even if the task itself still yields positive outcome in the most adverse conditions. In conclusion, semi-supervised POS-tagging is a good alternative to supervised POS-tagging when little training or adaptation data are available, or can be used to cut annotation cost when turning to new domains. An interesting follow-up work will be to adapt semi-supervised tagging so that it can deal efficiently with ASR lattices.

Acknowledgements This work is supported by ANR through projects PORT-MEDIA (ANR-08-CORD-026), SEQUOIA (ANR-08-EMER-013) and DECODA (2009-CORD-005-01).

5. References

[1] Y. Wilks and M. Stevenson, "The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation," *Natural Language Engineering*, vol. 4, no. 02, pp. 135–143, 1998.

[2] J. Hajič, M. Ciaramita, R. Johansson *et al.*, "The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages," in *CoNLL*, 2009, pp. 1–18.

[3] D. Vergyri, K. Kirchhoff, K. Duh, and A. Stolcke, "Morphology-based language modeling for Arabic speech recognition," in *Interspeech*, 2004.

[4] Y. Lee, "Morphological analysis for statistical machine translation," in *HLT-NAACL*, 2004, pp. 57–60.

[5] B. Merialdo, "Tagging English text with a probabilistic model," *Computational linguistics*, vol. 20, no. 2, pp. 155–171, 1994.

[6] S. Goldwater and T. Griffiths, "A fully Bayesian approach to unsupervised part-of-speech tagging," in *ACL*, vol. 45, no. 1, 2007, p. 744.

[7] J. Gao and M. Johnson, "A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers," in *EMNLP*, 2008, pp. 344–352.

[8] D. Chiang, J. Graehl, K. Knight, A. Pauls, and S. Ravi, "Bayesian inference for finite-state transducers," in *HLT-NAACL*, 2010.

[9] S. Cucerzan and D. Yarowsky, "Language independent, minimally supervised induction of lexical probabilities," in *ACL*, 2000, pp. 270–277.

[10] K. Toutanova and M. Johnson, "A Bayesian LDA-based model for semi-supervised part-of-speech tagging," in *NIPS*, vol. 20, 2007.

[11] Y. Goldberg, M. Adler, and M. Elhadad, "EM can find pretty good HMM POS-taggers (when given a good start)," in *ACL*, 2008, pp. 746–754.

[12] S. Galliano, E. Geoffrois, G. Gravier, J. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news," in *LREC*, 2006.

[13] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, p. 585, 2006.

[14] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news," in *Interspeech*, 2005.

[15] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa, "Semantic annotation of the french media dialog corpus," in *Interspeech*, 2005.

[16] H. Bonneau-Maynard, C. Ayache, F. Bechet *et al.*, "Evaluation in media discourse: analysis of a newspaper corpus," in *LREC*, 2006.

[17] R. Dufour and Y. Estève, "Correcting ASR outputs: specific solutions to specific errors in French," in *SLT*, 2008, pp. 213–216.