# Importance weighting and unsupervised domain adaptation of POS taggers: a negative result

**Barbara Plank, Anders Johannsen and Anders Søgaard**
Center for Language Technology
University of Copenhagen, Denmark
Njalsgade 140, DK-2300 Copenhagen S
bplank@cst.dk,ajohannsen@hum.ku.dk,soegaard@hum.ku.dk

## Abstract

Importance weighting is a generalization of various statistical bias correction techniques. While our labeled data in NLP is heavily biased, importance weighting has seen only few applications in NLP, most of them relying on a small amount of labeled target data. The publication bias toward reporting positive results makes it hard to say whether researchers have tried. This paper presents a negative result on unsupervised domain adaptation for POS tagging. In this setup, we only have *unlabeled* data and thus only indirect access to the bias in emission and transition probabilities. Moreover, most errors in POS tagging are due to unseen words, and there, importance weighting cannot help. We present experiments with a wide variety of weight functions, quantilizations, as well as with randomly generated weights, to support these claims.

## 1 Introduction

Many NLP tasks rely on the availability of annotated data. The majority of annotated data, however, is sampled from newswire corpora. The performance of NLP systems, e.g., part-of-speech (POS) tagger, parsers, relation extraction systems, etc., drops significantly when they are applied to data that departs from newswire conventions. So while we can extract information, translate and summarize newswire in major languages with some success, we are much less successful processing microblogs, chat, weblogs, answers, emails or literature in a robust way. The main reasons for the drops in accuracy have been attributed to factors such as previously unseen words and bigrams, missing punctuation and capitalization, as well as differences in the marginal distribution of data (Blitzer et al., 2006; McClosky et al., 2008; Søgaard and Haulrich, 2011).

The move from one domain to another (from a *source* to a new *target* domain), say from newspaper articles to weblogs, results in a sample selection bias. Our training data is now biased, since it is sampled from a related, but nevertheless different distribution. The problem of automatically adjusting the model induced from source to a different target is referred to as *domain adaptation*.

Some researchers have studied domain adaptation scenarios, where small samples of labeled data have been assumed to be available for the target domains. This is usually an unrealistic assumption, since even for major languages, small samples are only available from a limited number of domains, and in this work we focus on unsupervised domain adaptation, assuming only unlabeled target data is available.

Jiang and Zhai (2007), Foster et al. (2010; Plank and Moschitti (2013) and Søgaard and Haulrich (2011) have previously tried to use importance weighting to correct sample bias in NLP. Importance weighting means assigning a weight to each training instance, reflecting its importance for modeling the target distribution. Importance weighting is a generalization over post-stratification (Smith, 1991) and importance sampling (Smith et al., 1997) and can be used to correct bias in the labeled data.

Out of the four papers mentioned, only Søgaard and Haulrich (2011) and Plank and Moschitti (2013) considered an unsupervised domain adaptation scenario, obtaining mixed results. These two papers assume *covariate shift* (Shimodaira, 2000), i.e., that there is only a bias in the marginal distribution of the training data. Under this assumption, we can correct the bias by applying a weight function $\frac{P_t(\mathbf{x})}{P_s(\mathbf{x})}$ to our training data points (labeled sentences) and learn from the weighted data. Of course this weight function cannot be



*Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–973,
October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

computed in general, but we can approximate it in different ways.

In POS tagging, we typically factorize sequences into emission and transition probabilities. Importance weighting can change emission probabilities and transition probabilities by assigning weights to sentences. For instance, if our corpus consisted of three sequences: 1) *a/A b/A*, 2) *a/A b/B*, and 3) *a/A b/B*, then $P(B|A) = 2/3$. If sequences two and three were down-weighted to 0.5, then $P(B|A) = 1/2$.

However, this paper argues that importance weighting cannot help adapting POS taggers to new domains using only unlabeled target data. We present three sources of evidence: (a) negative results with the most obvious weight functions across various English datasets, (b) negative results with randomly sampled weights, as well as (c) an analysis of annotated data indicating that there is little variation in emission and transition probabilities across the various domains.

## 2 Related work

Most prior work on importance weighting use a *domain classifier*, i.e., train a classifier to discriminate between source and target instances (Søgaard and Haulrich, 2011; Plank and Moschitti, 2013) ($y \in \{s, t\}$). For instance, Søgaard and Haulrich (2011) train a $n$-gram text classifier and Plank and Moschitti (2013) a tree-kernel based classifier on relation extraction instances. In these studies, $\hat{P}(t|\mathbf{x})$ is used as an approximation of $\frac{P_t(\mathbf{x})}{P_s(\mathbf{x})}$, following Zadrozny (2004). In §3, we follow the approach of Søgaard and Haulrich (2011), but consider a wider range of weight functions. Others have proposed to use kernel mean matching (Huang et al., 2007) or minimizing $KL$-divergence (Sugiyama et al., 2007).

Jiang and Zhai (2007) use importance weighting to select a subsample of the source data by subsequently setting the weight of all selected data points to 1, and 0 otherwise. However, they do so by relying on a sequential model trained on labeled target data. Our results indicate that the covariate shift assumption fails to hold for cross-domain POS tagging. While the marginal distributions obviously *do* differ (since we can tell domains apart without POS analysis), this is most likely not the only difference. This might explain the positive results obtained by Jiang and Zhai (2007). We will come back to this in §4.

Cortes et al. (2010) show that importance weighting potentially leads to over-fitting, but propose to use quantiles to obtain more robust weight functions. The idea is to rank all weights and obtain $q$ quantiles. If a data point $\mathbf{x}$ is weighted by $w$, and $w$ lies in the $i$th quantile of the ranking ($i \leq q$), $\mathbf{x}$ is weighted by the average weight of data points in the $i$th quantile.

The weighted structured perceptron (§3) used in the experiments below was recently used for a different problem, namely for correcting for bias in annotations (Plank et al., 2014).

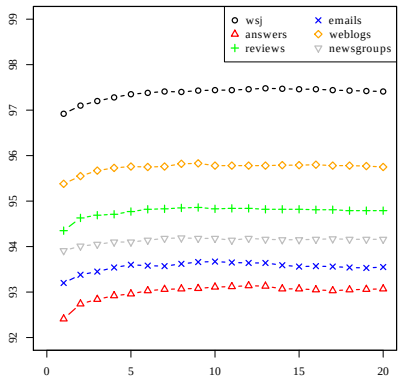


Figure 1: Training epochs vs tagging accuracy for the baseline model on the dev data.

## 3 Experiments

### 3.1 Data

We use the data made available in the SANCL 2012 Shared Task (Petrov and McDonald, 2012). The training data is the OntoNotes 4.0 release of the Wall Street Journal section of the Penn Treebank, while the target domain evaluation data comes from various sources, incl. Yahoo Answers, user reviews, emails, weblogs and newsgroups. For each target domain, we have both development and test data.

### 3.2 Model

In the weighted perceptron (Cavallanti et al., 2006), we make the learning rate dependent on the current instance $\mathbf{x}_n$, using the following update:

$$\mathbf{w}^{i+1} \leftarrow \mathbf{w}^i + \beta_n \alpha(y_n - \text{sign}(\mathbf{w}^i \cdot \mathbf{x}_n))\mathbf{x}_n \quad (1)$$

where $\beta_n$ is the weight associated with $\mathbf{x}_n$. See Huang et al. (2007) for similar notation.

We extend this idea straightforwardly to the structured perceptron (Collins, 2002), for which

| System | Answers | Newsgroups | Reviews | Avg | Emails | Weblogs | WSJ |
|---|---|---|---|---|---|---|---|
| Our system | 91.08 | 91.57 | 91.59 | 91.41 | 87.97 | 92.19 | 97.32 |
| SANCL12-2nd | 90.99 | 92.32 | 90.65 | 91.32 | – | – | 97.76 |
| SANCL12-best | 91.79 | 93.81 | 93.11 | 92.90 | – | – | 97.29 |
| SANCL12-last | 88.24 | 89.70 | 88.15 | 88.70 | – | – | 95.14 |
| FLORS basic | 91.17 | 92.41 | 92.25 | 88.67 | 91.37 | 97.11 | 91.94 |

Table 1: Tagging accuracies and comparison to prior work on the SANCL test sets (fine-grained POS).

we use an in-house implementation. We use commonly used features, i.e., $w$,$w_{-1}$, $w_{-2}$, $w_{+1}$, $w_{+2}$, digit, hyphen, capitalization, pre-/suffix features, and Brown word clusters. The model seems robust with respect to number of training epochs, cf. Figure 1. Therefore we fix the number of epochs to five and use this setting in all our experiments. Our code is available at: `https://bitbucket.org/bplank/importance-weighting-exp`.

### 3.3 Importance weighting

In our first set of experiments, we follow Søgaard and Haulrich (2011) in using document classifiers to obtain weights for the source instances. We train a text classifier that discriminates the two domains (source and target). For each sentence in the source and target domain (the unlabeled text that comes with the SANCL data), we mark whether it comes from the source or target domain and train a binary classifier (logistic regression) to discriminate between the two. For every sentence in the source we obtain its probability for the target domain by doing 5-fold cross-validation. While Søgaard and Haulrich (2011) use only token-based features (word $n$-grams $\leq$ 3), we here exploit a variety of features: word token $n$-grams, and two generalizations: using Brown clusters (estimated from the union of the 5 target domains), and Wiktionary tags (if a word has multiple tags, we assign it the union of tags as single tag; OOV words are marked as such).

The distributions of weights can be seen in the upper half of Figure 2.

#### 3.3.1 Results

Table 1 shows that our baseline model achieves state-of-the-art performance compared to SANCL (Petrov and McDonald, 2012)[1] and FLORS (Schnabel and Schütze, 2014). Our results align well with the second best POS tagger in the SANCL 2012 Shared Task. Note

[1] `https://sites.google.com/site/sancl2012/home/shared-task/results`

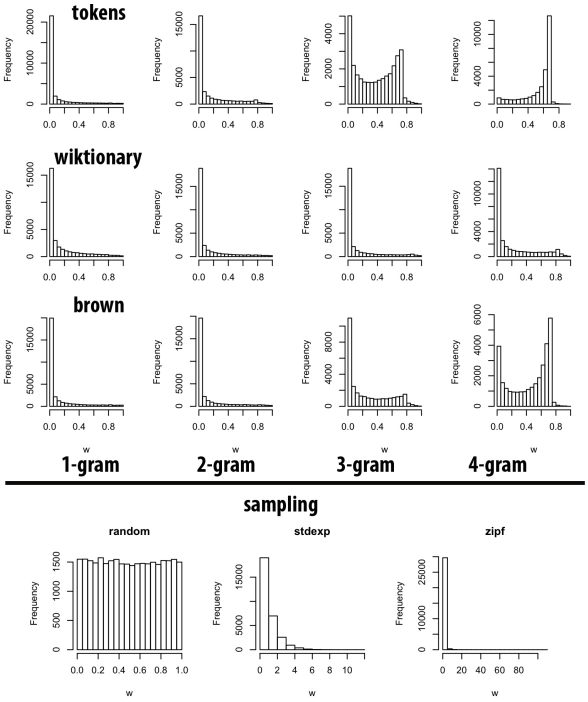


Figure 2: Histogram of different weight functions.

that the best tagger in the shared task explicitly used normalization and various other heuristics to achieve better performance. In the rest of the paper, we use the universal tag set part of the SANCL data (Petrov et al., 2012).

Figure 3 presents our results on development data for different importance weighting setups. None of the above weight functions lead to significant improvements on *any* of the datasets. We also tried scaling and binning the weights, as suggested by Cortes et al. (2010), but results kept fluctuating around baseline performance, with no significant improvements.

### 3.4 Random weighting

Obviously, weight functions based on document classifiers may simply not characterize the relevant properties of the instances and hence lead to bad re-weighting of the data. We consider three random sampling strategies, namely sampling random uniforms, random exponentials, and random

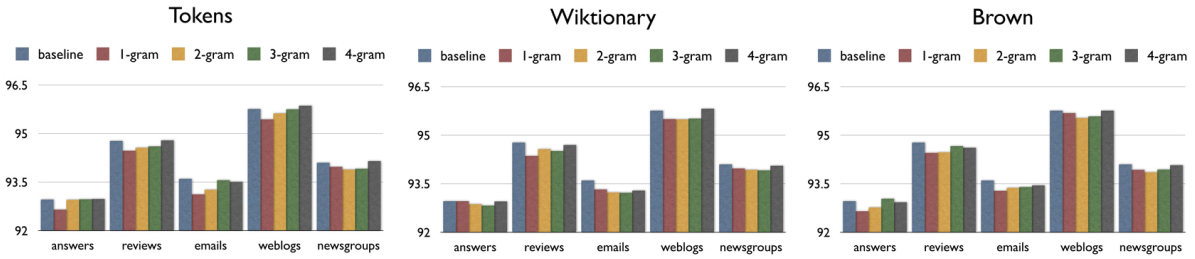


Figure 3: Results on development data for different weight functions, i.e., document classifiers trained on a) raw tokens; b) tokens replaced by Wiktionary tags; c) tokens replaced by Brown cluster ids. The weight was the raw $p_t(y|x)$ value, no scaling, no quantiles. Replacing only open-class tokens for b) and c) gave similar or lower performance.

Zipfians and ran 500 samples for each. For these experiments, we estimate significance cut-off levels of tagging accuracies using the approximate randomization test. To find the cut-off levels, we randomly replace labels with gold labels until the achieved accuracy significantly improves over the baseline for more than 50% of the samples. For each accuracy level, 50 random samples were taken.

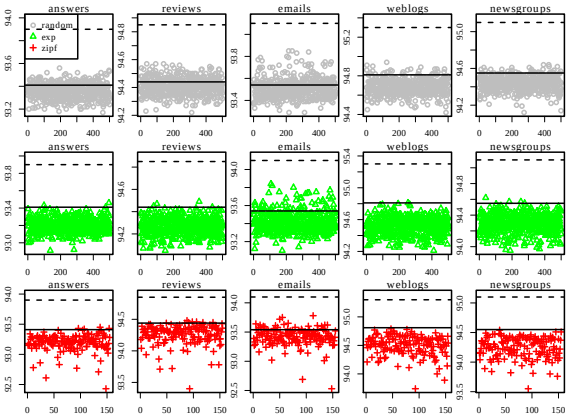


Figure 4: Random weight functions (500 runs each) on test sets. Solid line is the baseline performance, while the dashed line is the $p$-value cut-off. From top: random, exponential and Zipfian weighting. All runs fall below the cut-off.

### 3.4.1 Results

The dashed lines in Figure 4 show the p-value cut-offs for positive results. We see that most random weightings of data lead to slight drops in performance or are around baseline performance, and no weightings lead to significant improvements. Random uniforms seem slightly better than exponentials and Zipfians.

| **domain (tokens)** | **avg tag ambiguity** | | **OOV** | **KL** | $\rho$ |
|---|---|---|---|---|---|
| | type | token | | | |
| wsj (train/test: 731k/39k) | 1.09 | 1.41 | 11.5 | 0.0006 | 0.99 |
| answers (28k) | 1.09 | 1.22 | 27.7 | 0.048 | 0.77 |
| reviews (28k) | 1.07 | 1.19 | 29.5 | 0.040 | 0.82 |
| emails (28k) | 1.07 | 1.19 | 29.9 | 0.027 | 0.92 |
| weblogs (20k) | 1.05 | 1.11 | 22.1 | 0.010 | 0.96 |
| newsgroups (20k) | 1.05 | 1.14 | 23.1 | 0.011 | 0.96 |

Table 2: Relevant statistics for our analysis (§4) on the test sets: average tag ambiguity, out-of-vocabulary rate, and KL-divergence and Pearson correlation coefficient ($\rho$) on POS bigrams.

## 4 Analysis

Some differences between the gold-annotated source domain data and the gold-annotated target data used for evaluation are presented in Table 2. One important observation is the low ambiguity of word forms in the data. This makes the room for improvement with importance weighting smaller. Moreover, the KL divergencies over POS bigrams are also very low. This tells us that transition probabilities are also relatively constant across domains, again suggesting limited room for improvement for importance weighting.

Compared to this, we see much bigger differences in OOV rates. OOV rates do seem to explain most of the performance drop across domains. In order to verify this, we implemented a version of our structured perceptron tagger with type-constrained inference (Täckström et al., 2013). This technique only improves performance on unseen words, but nevertheless we saw significant improvements across all five domains (cf. Table 3). This suggests that unseen words are a more important problem than the marginal distribution of data for unsupervised domain adaptation of POS taggers.

| | ans | rev | email | webl | newsg |
|---|---|---|---|---|---|
| base | 93.41 | 94.44 | 93.54 | 94.81 | 94.55 |
| +type constr. | 94.09† | 94.85† | 94.31† | 95.99† | 95.97† |
| $p$-val cut-off | 93.90 | 94.85 | 94.10 | 95.3 | 95.10 |

Table 3: Results on the test sets by adding Wiktionary type constraints. †=p-value $< 0.001$.

We also tried Jiang and Zhai's subset selection technique (§3.1 in Jiang and Zhai (2007)), which assumes labeled training material for the target domain. However, we did not see any improvements. A possible explanation for these different findings might be the following. Jiang and Zhai (2007) use labeled target data to learn their weighting model, i.e., in a supervised domain adaptation scenario. This potentially leads to very different weight functions. For example, let the source domain be 100 instances of $a/A\ b/B$ and 100 instances of $b/B\ b/B$, and the target domain be 100 instances of $a/B\ a/B$. Note that a domain classifier would favor the first 100 sentences, but in an HMM model induced from the labeled target data, things look very different. If we apply Laplace smoothing, the probability of $a/A\ b/B$ according to the target domain HMM model would be $\sim 8.9e^{-7}$, and the probability of $b/B\ b/B$ would be $\sim 9e^{-5}$. Note also that this set-up does not assume covariate shift.

## 5 Conclusions and Future Work

Importance weighting, a generalization of various statistical bias correction techniques, can potentially correct bias in our labeled training data, but this paper presented a negative result about importance weighting for unsupervised domain adaptation of POS taggers. We first presented experiments with a wide variety of weight functions, quantilizations, as well as with randomly generated weights, none of which lead to significant improvements. Our analysis indicates that most errors in POS tagging are due to unseen words, and what remains seem to not be captured adequately by unsupervised weight functions.

For future work we plan to extend this work to further weight functions, data sets and NLP tasks.

## Acknowledgements

This research is funded by the ERC Starting Grant LOWLANDS No. 313695.

## References

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *EMNLP*.

Giovanni Cavallanti, Nicolò Cesa-Bianchi, and Claudio Gentile. 2006. Tracking the best hyperplane with a simple budget perceptron. In *COLT*.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *EMNLP*.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. 2010. Learning bounds for importance weighting. In *NIPS*.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *EMNLP*.

Jiayuan Huang, Alexander Smola, Arthur Gretton, Karsten Borgwardt, and Bernhard Schölkopf. 2007. Correcting sample bias by unlabeled data. In *NIPS*.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *ACL*.

David McClosky, Eugene Charniak, and Mark Johnson. 2008. When is self-training effective for parsing? In *COLING*.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 Shared Task on Parsing the Web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *LREC*.

Barbara Plank and Alessandro Moschitti. 2013. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *ACL*.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *EACL*.

Tobias Schnabel and Hinrich Schütze. 2014. Flors: Fast and simple domain adaptation for part-of-speech tagging. *TACL*, 2:15–16.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244.

Peter Smith, Mansoor Shafi, and Hongsheng Gao. 1997. Quick simulation: A review of importance sampling techniques in communications systems. *IEEE Journal on Selected Areas in Communications*, 15(4):597–613.

T.M.F. Smith. 1991. Post-stratification. *The Statistician*, 40:315–323.

Anders Søgaard and Martin Haulrich. 2011. Sentence-level instance-weighting for graph-based and transition-based dependency parsing. In *IWPT*.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. 2007. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *TACL*, 1:1–12.

Bianca Zadrozny. 2004. Learning and evaluating classifiers under sample selection bias. In *ICML*.