Online Updating of Word Representations for Part-of-Speech Tagging

Wenpeng Yin
LMU Munich
wenpeng@cis.lmu.de

Tobias Schnabel
Cornell University
tbs49@cornell.edu

Hinrich Schütze
LMU Munich
inquiries@cislmu.org

Abstract
We propose online unsupervised domain adaptation (DA), which is performed incrementally as data comes in and is applicable when batch DA is not possible. In a part-of-speech (POS) tagging evaluation, we find that online unsupervised DA performs as well as batch DA.

1 Introduction
Unsupervised domain adaptation is a scenario that practitioners often face when having to build robust NLP systems. They have labeled data in the source domain, but wish to improve performance in the target domain by making use of unlabeled data alone. Most work on unsupervised domain adaptation in NLP uses batch learning: It assumes that a large corpus of unlabeled data of the target domain is available before testing. However, batch learning is not possible in many real-world scenarios where incoming data from a new target domain must be processed immediately. More importantly, in many real-world scenarios the data does not come with neat domain labels and it may not be immediately obvious that an input stream is suddenly delivering data from a new domain.

Consider an NLP system that analyzes emails at an enterprise. There is a constant stream of incoming emails and it changes over time – without any clear indication that the models in use should be adapted to the new data distribution. Because the system needs to work in real-time, it is also desirable to do any adaptation of the system online, without the need of stopping the system, changing it and restarting it as is done in batch mode.

In this paper, we propose online unsupervised domain adaptation as an extension to traditional unsupervised DA. In online unsupervised DA, domain adaptation is performed incrementally as data comes in. Specifically, we adopt a form of representation learning. In our experiments, the incremental updating will be performed for representations of words. Each time a word is encountered in the stream of data at test time, its representation is updated.

To the best of our knowledge, the work reported here is the first study of online unsupervised DA. More specifically, we evaluate online unsupervised DA for the task of POS tagging. We compare POS tagging results for three distinct approaches: static (the baseline), batch learning and online unsupervised DA. Our results show that online unsupervised DA is comparable in performance to batch learning while requiring no retraining or prior data in the target domain.

2 Experimental setup
Tagger. We reimplemented the FLORS tagger (Schnabel and Schütze, 2014), a fast and simple tagger that performs well in DA. It treats POS tagging as a window-based (as opposed to sequence classification), multilabel classification problem. FLORS is ideally suited for online unsupervised DA because its representation of words includes distributional vectors – these vectors can be easily updated in both batch learning and online unsupervised DA. More specifically, a word’s representation in FLORS consists of four feature vectors: one each for its suffix, its shape and its left and right distributional neighbors. Suffix and shape features are standard features used in the literature; our use of them is exactly as described by Schnabel and Schütze (2014).

Distributional features. The $i^{th}$ entry $x_i$ of the left distributional vector of $w$ is the weighted number of times the indicator word $c_i$ occurs immediately to the left of $w$:

$$ x_i = \text{tf} \left( \text{freq} \left( \text{bigram}(c_i, w) \right) \right) $$

where $c_i$ is the word with frequency rank $i$ in the corpus, $\text{freq} \left( \text{bigram}(c_i, w) \right)$ is the number of occurrences of the bigram “$c_i$ w” and we weight non-
Table 1: BATCH and ONLINE accuracies are comparable and state-of-the-art. Best number in each column is bold.

|          | newsgroups | reviews | weblogs | answers | emails |-wsj-
|----------|------------|---------|---------|---------|--------|--------|
|          | ALL OOV | ALL OOV | ALL OOV | ALL OOV | ALL OOV | ALL OOV |
| TnT      | 88.66  | 54.73  | 90.40  | 56.75  | 93.33  | 74.17  |
| Stanford | 89.11  | 56.02  | 91.43  | 58.66  | 94.15  | 77.13  |
| SVMTool  | 89.14  | 53.82  | 91.30  | 54.20  | 94.21  | 76.44  |
| C&P      | 89.51  | 57.23  | 91.58  | 59.67  | 94.41  | 78.46  |
| S&S      | 90.86  | 66.42  | 92.95  | 75.29  | 94.71  | 83.64  |
| S&S (reimpl.) | 90.68  | 65.32  | 93.00  | 75.50  | 94.64  | 82.91  |
| BATCH    | 90.37  | 71.18  | 93.07  | 79.03  | 94.86  | 86.53  |
| ONLINE   | 90.85  | 71.00  | 93.07  | 79.03  | 94.86  | 86.53  |

zero frequencies logarithmically: \( tf(x) = 1 + \log(x) \). The right distributional vector is defined analogously. We restrict the set of indicator words to the \( n = 500 \) most frequent words. To avoid zero vectors, we add an entry \( x_{n+1} \) to each vector that counts omitted contexts:

\[
x_{01} = tf(\sum_{j:j>n} freq(bigram(c_j, w)))
\]

Let \( f(w) \) be the concatenation of the two distributional and suffix and shape vectors of word \( w \). Then FLORS represents token \( v_i \) as follows:

\[
f(v_{i-2}) \oplus f(v_{i-1}) \oplus f(v_i) \oplus f(v_{i+1}) \oplus f(v_{i+2})
\]

where \( \oplus \) denotes vector concatenation. FLORS then tags token \( v_i \) based on this representation.

FLORS assumes that the association between distributional features and labels does not change fundamentally when going from source to target. This is in contrast to other work, notably Blitzer et al. (2006), that carefully selects “stable” distributional features and discards “unstable” distributional features. The hypothesis underlying FLORS is that basic distributional POS properties are relatively stable across domains — in contrast to semantic and other more complex tasks. The high performance of FLORS (Schnabel and Schütze, 2014) suggests this hypothesis is true.

Data. Test set. We evaluate on the development sets of six different TDs: five SANCL (Petrov and McDonald, 2012) domains – newsgroups, weblogs, reviews, answers, emails – and sections 22-23 of WSJ for in-domain testing.

We use two training sets of different sizes. In condition \( l:big \) (big labeled data set), we train FLORS on sections 2-21 of Wall Street Journal (WSJ). Condition \( l:small \) uses 10% of \( l:big \).

Data for word representations. We also vary the size of the datasets that are used to compute the word representations before the FLORS model is trained on the training set. In condition \( u:big \), we compute distributional vectors on the joint corpus of all labeled and unlabeled text of source and target domains (except for the test sets). We also include 100,000 WSJ sentences from 1988 and 500,000 sentences from Gigaword (Parker, 2009). In condition \( u:0 \), only labeled training data is used.

Methods. We implemented the following modifications compared to the setup in (Schnabel and Schütze, 2014): distributional vectors are kept in memory as count vectors. This allows us to increase the counts during online tagging.

We run experiments with three versions of FLORS: STATIC, BATCH and ONLINE. All three methods compute word representations on “data for word representations” (described above) before the model is trained on one of the two “training sets” (described above).

STATIC. Word representations are not changed during testing.

BATCH. Before testing, we update count vectors by \( freq(bigram(c_i, w)) \) \( \leftarrow \) \( freq^*(bigram(c_i, w)) \), where \( freq^*(\cdot) \) denotes the number of occurrences of the bigram “\( c_i \) w” in the entire test set.

ONLINE. Before tagging a test sentence, both left and right distributional vectors are updated via \( freq(bigram(c_i, w)) \) \( \leftarrow 1 \) for each appearance of bigram “\( c_i \) w” in the sentence. Then the sentence is tagged using the updated word representations. As tagging progresses, the distributional representations become increasingly specific to the target domain (TD), converging to the representations that BATCH uses at the end of the tagging process.

In all three modes, suffix and shape features are always fully specified, for both known and unknown words.

3 Experimental results

Table 1 compares performance on SANCL for a number of baselines and four versions of FLORS: S&S, Schnabel and Schütze (2014)’s version of FLORS, “S&S (reimpl.)”, our reimplementation of that version, and BATCH and ONLINE, the two versions of FLORS we use in this paper. Compar-
Table 2: ONLINE / BATCH accuracies are generally better than STATIC (see bold numbers) and improve with both more training data and more unlabeled data.

|                  | u:0 | u:big |
|------------------|-----|-------|
|                  | ALL | u:0   | u:big |
|                  | KN  | SHFT  | OOV  | KN  | SHFT  | OOV  |
| newsgroups       |     |       |      |     |       |      |
| STATIC           | 87.02 | 90.87 | 71.12 | 57.16 | 89.02 | 91.48 | 81.53 | 58.30 |
| ONLINE           | 87.99 | 90.87 | 76.10 | 65.64 | 89.84 | 92.38 | 82.58 | 67.09 |
| BATCH            | **88.28** | **91.08** | **77.01** | **66.37** | **89.82** | **92.37** | **82.65** | **67.05** |
| reviews          |     |       |      |     |       |      |
| STATIC           | 89.09 | 93.00 | 62.85 | 65.82 | 89.93 | 92.41 | 84.94 | 58.97 |
| ONLINE           | 90.51 | 93.13 | 82.51 | 67.57 | 90.85 | 93.04 | 84.04 | 71.00 |
| BATCH            | **90.69** | **93.12** | **83.24** | **69.43** | **90.87** | **93.03** | **85.20** | **71.18** |
| weblogs          |     |       |      |     |       |      |
| STATIC           | 89.08 | 91.96 | 66.55 | 65.90 | 91.45 | 92.47 | 80.11 | 70.81 |
| ONLINE           | 89.67 | 92.14 | 70.14 | 69.67 | **92.11** | **93.62** | **81.46** | **78.42** |
| BATCH            | **89.79** | **92.23** | **69.86** | **71.27** | **92.10** | **93.60** | **81.67** | **78.42** |
| answers          |     |       |      |     |       |      |
| STATIC           | 91.96 | 93.94 | 82.40 | 67.97 | 92.42 | 93.53 | 84.65 | 69.97 |
| ONLINE           | 92.33 | 94.03 | **83.59** | **72.50** | **93.07** | **94.36** | **85.71** | **79.03** |
| BATCH            | **92.42** | **94.09** | **83.53** | **73.35** | **93.07** | **94.36** | **85.71** | **79.03** |
| emails           |     |       |      |     |       |      |
| STATIC           | 91.58 | 94.29 | 79.95 | 72.74 | 93.42 | 94.77 | 89.80 | 77.42 |
| ONLINE           | 92.51 | 94.52 | 81.76 | 80.46 | **94.21** | **95.40** | **91.08** | **84.03** |
| BATCH            | **92.68** | **94.60** | **82.34** | **81.20** | **94.20** | **95.42** | **91.03** | **83.87** |
| emails           |     |       |      |     |       |      |
| STATIC           | 93.43 | 95.04 | 90.15 | 72.68 | 94.09 | 95.34 | 91.90 | 76.94 |
| ONLINE           | 94.18 | 95.82 | 89.80 | 80.35 | **94.86** | **95.81** | **92.60** | **86.53** |
| BATCH            | **94.34** | **95.85** | **90.03** | **81.84** | **94.86** | **95.82** | **92.60** | **86.53** |
| usj              |     |       |      |     |       |      |
| STATIC           | 86.93 | 90.89 | 66.51 | 53.43 | 88.98 | 91.09 | 77.63 | 57.36 |
| ONLINE           | 87.48 | 91.18 | 68.07 | 56.47 | **89.71** | **92.42** | **78.11** | **64.21** |
| BATCH            | **87.56** | **91.11** | **68.25** | **58.44** | **89.71** | **92.43** | **78.23** | **64.09** |

<sup>1</sup>We cannot give the standard, single OOV evaluation number here since OOVs are different in different conditions, hence the breakdown into three measures.
plain below why ONLINE is sometimes (slightly) **better** than **BATCH**, e.g., for **ALL** and condition l:small/u:big.

### 3.1 Time course of tagging accuracy

The **ONLINE** model introduced in this paper has a property that is unique compared to most other work in statistical NLP: its predictions change as it tags text because its representations change.

To study this time course of changes, we need a large application domain because subtle changes will be too variable in the small test sets of the SANCL TDs. The only labeled domain that is big enough is the WSJ corpus. We therefore reverse the standard setup and train the model on the dev sets of the five SANCL domains (l:big) or on the first 5000 labeled words of reviews (l:small). In this reversed setup, u:big uses the five unlabeled SANCL data sets and Gigaword as before. Since variance of performance is important, we run on 100 randomly selected 50% samples of WSJ and report average and standard deviation of tagging error over these 100 trials.

The results in Table 3 show that error rates are only slightly worse for **ONLINE** than for **BATCH** or the same. In fact, l:small/u:0 known error rate (.1186) is **lower** for **ONLINE** than for **BATCH** (similar to what we observed in Table 2). This will be discussed at the end of this section.

Table 3 includes results for “unseen” errors as well as unknowns because [Schnabel and Schütze (2014)] show that unseen errors are at least as many errors as unknowns. We define *unseen* as words with a tag that did not occur in training; we compute unseen error rates on all occurrences of unseen, i.e., occurrences with both seen and unseen tags are included. As Table 3 shows, the error rate for unknowns is greater than for unseen words which is in turn greater than the error rate on known words.

Examining the single conditions, we can see that ONLINE fares better than STATIC in 10 out of 12 cases and only slightly worse for l:small/u:big (unseens, known words: .1086 vs .1084, .0802 vs .0801). In four conditions it is significantly better with improvements ranging from .005 (.1404 vs .1451: l:big/u:0, unknown words) to >.06 (.3094 vs .3670: l:small/u:0, unknown words).

The differences between **ONLINE** and **STATIC** in the other eight conditions are negligible. For the six u:big conditions, this is not surprising: the Gigaword corpus consists of news, so the large unlabeled data set is in reality the same domain as WSJ. Thus, if large unlabeled data sets are available that are similar to the TD, then one might as well use **STATIC** tagging since the extra work required for **ONLINE** is unlikely to pay off.

Using more labeled data (comparing l:small and l:big) always considerably decreases error rates. We did not test for significance here because the differences are so large. By the same token, using more unlabeled data (comparing u:0 and u:big) also consistently decreases error rates. The differences are large and significant for **ONLINE** tagging in all six cases (indicated by * in the table).

There is no large difference in variability **ONLINE** vs. **BATCH** (see columns “std”). Thus, given that it has equal variability and higher performance, **ONLINE** is preferable to **BATCH** since it assumes no dataset available prior to the start of tagging.

Figure shows the time course of tagging accuracy. **BATCH** and **STATIC** have constant error rates since they do not change representations during tagging. **ONLINE** error decreases for unknown words – approaching the error rate of **BATCH** – as

---

2Significance test: test of equal proportion, $p < .05$

In response to a reviewer question, the initial (leftmost) errors of **ONLINE** and **STATIC** are not identical; e.g., **ONLINE** has a better chance of correctly tagging the very first occurrence of an unknown word because that very first occurrence has a meaningful (as opposed to random) distributed representation.
more and more is learned with each additional occurrence of an unknown word (top).

Interestingly, the error of ONLINE increases for unseens and known words (middle & bottom panels) (even though it is always below the error rate of BATCH). The reason is that the BATCH update swamps the original training data for $l:\text{small}/u:0$ because the WSJ test set is bigger by a large factor than the training set. ONLINE fares better here because in the beginning of tagging the updates of the distributional representations consist of small increments. We noticed this in Table 2 too: there, ONLINE outperformed BATCH in some cases on KN for $l:\text{small}/u:big$. In future work, we plan to investigate how to weight distributional counts from the target data relative to that from the (labeled and unlabeled) source data.

4 Related work

Online learning usually refers to supervised learning algorithms that update the model each time after processing a few training examples. Many supervised learning algorithms are online or have online versions. Active learning (Lewis and Gale, 1994; Tong and Koller, 2001; Laws et al., 2011) is another supervised learning framework that processes training examples – usually obtained interactively – in small batches (Bordes et al., 2005).

All of this work on supervised online learning is not directly relevant to this paper since we address the problem of unsupervised DA. Unlike online supervised learners, we keep the statistical model unchanged during DA and adopt a representation learning approach: each unlabeled context of a word is used to update its representation.

There is much work on unsupervised DA for POS tagging, including work using constraint-based methods (Subramanya et al., 2010; Rush et al., 2012), instance weighting (Choi and Palmer, 2012), self-training (Huang et al., 2009; Huang and Yates, 2010), and co-training (Kühler and Baucom, 2011). All of this work uses batch learning. For space reasons, we do not discuss supervised DA (e.g., Daumé III and Marcu (2006)).

5 Conclusion

We introduced online updating of word representations, a new domain adaptation method for cases where target domain data are read from a stream and BATCH processing is not possible. We showed that online unsupervised DA performs as well as batch learning. It also significantly lowers error rates compared to STATIC (i.e., no domain adaptation). Our implementation of FLORS is available at cistern.cis.lmu.de/flors.

Acknowledgments. This work was supported by a Baidu scholarship awarded to Wenpeng Yin and by Deutsche Forschungsgemeinschaft (grant DFG SCHU 2246/10-1 FADeBaC).
References

[Blitzer et al.2006] John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In EMNLP, pages 120–128.

[Bordes et al.2005] Antoine Bordes, Seyda Ertekin, Jason Weston, and Léon Bottou. 2005. Fast kernel classifiers with online and active learning. The Journal of Machine Learning Research, 6:1579–1619.

[Choi and Palmer2012] Jinho D. Choi and Martha Palmer. 2012. Fast and robust part-of-speech tagging using dynamic model selection. In ACL: Short Papers, pages 363–367.

[Daumé III and Marcu2006] Hal Daumé III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research, 26:101–126.

[Huang and Yates2010] Fei Huang and Alexander Yates. 2010. Exploring representation-learning approaches to domain adaptation. In DANLP, pages 23–30.

[Huang et al.2009] Zhongqiang Huang, Vladimir Eidelman, and Mary Harper. 2009. Improving a simple bigram HMM part-of-speech tagger by latent annotation and self-training. In NAACL-HLT: Short Papers, pages 213–216.

[Kübler and Baucom2011] Sandra Kübler and Eric Baucom. 2011. Fast domain adaptation for part of speech tagging for dialogues. In RANLP, pages 41–48.

[Laws et al.2011] Florian Laws, Christian Scheible, and Hinrich Schütze. 2011. Active learning with Amazon Mechanical Turk. In Conference on Empirical Methods in Natural Language Processing, pages 1546–1556.

[Lewis and Gale1994] David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3–12.

[Parker2009] Robert Parker. 2009. English gigaword fourth edition. Linguistic Data Consortium.

[Petrov and McDonald2012] 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), volume 59.

[Rush et al.2012] Alexander M. Rush, Roi Reichart, Michael Collins, and Amir Globerson. 2012. Improved parsing and POS tagging using inter-sentence consistency constraints. In EMNLP-CoNLL, pages 1434–1444.

[Schnabel and Schütze2014] Tobias Schnabel and Hinrich Schütze. 2014. FLORS: Fast and simple domain adaptation for part-of-speech tagging. TACL, 2:15–26.

[Subramanya et al.2010] Amarnag Subramanya, Slav Petrov, and Fernando Pereira. 2010. Efficient graph-based semi-supervised learning of structured tagging models. In EMNLP, pages 167–176.

[Tong and Koller2001] Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. J. Mach. Learn. Res., 2:45–66.