# Mining for unambiguous instances to adapt part-of-speech taggers to new domains

**Dirk Hovy, Barbara Plank, Héctor Martínez Alonso, Anders Søgaard**
Center for Language Technology
University of Copenhagen, Denmark
Njalsgade 140
{dirk|bplank}@cst.dk,{alonso|soegaard}@hum.ku.dk

## Abstract

We present a simple, yet effective approach to adapt part-of-speech (POS) taggers to new domains. Our approach only requires a dictionary and large amounts of unlabeled target data. The idea is to use the dictionary to mine the unlabeled target data for unambiguous word sequences, thus effectively collecting *labeled* target data. We add the mined instances to available labeled newswire data to train a POS tagger for the target domain. The induced models significantly improve tagging accuracy on held-out test sets across three domains (Twitter, spoken language, and search queries). We also present results for Dutch, Spanish and Portuguese Twitter data, and provide two novel manually-annotated test sets.

## 1 Introduction

Part-of-speech (POS) taggers are typically trained on newswire and exhibit severe out-of-domain performance drops (Blitzer et al., 2006; Daume III, 2007; Foster et al., 2011). When faced with a new domain, one option is to try to leverage available unlabeled data. However, rather than resorting to pure self-training approaches (self-labeling), we here resort to another source of information. One way to address the annotation problem is to use collaboratively created resources such as Wikipedia for *distant supervision* (Mintz et al., 2009), or the automatically derived dictionaries called *Wiktionary* (Li et al., 2012). We show how to leverage these resources to create labeled training data. It turns out that many entries in Wiktionary are actually *unambiguous*, i.e., there is only *one* possible tag for the word. In fact, for English Wiktionary (Li et al., 2012), we find that 93% of the unigram types are unambiguous (cf. Table 2).

Our idea here is simple: we mine for *unlabeled* sentences that contain only unambiguous items (according to Wiktionary), and use the resulting data as additional, *labeled* training material. Concretely, we mine unannotated corpora of tweets, transcribed speech, and search queries for sentences that contain only unambiguous tokens, and combine those instances with newswire data to train POS models that adapt better to the respective domains. We show that adding unambiguous data leads to considerable improvements over both unadapted and weakly-supervised baselines (Li et al., 2012).

Since Wiktionary has relatively low coverage for some of these domains, we also explore the use of Brown clusters to extend the coverage. This enables us to generalize across spelling variations and synonyms. Additionally, we evaluate our approach on Dutch, Portuguese and Spanish Twitter and present tow novel data sets for the latter two languages.

## 2 Data

### 2.1 Wiktionary

In our experiments, we use the (unigram) tag dictionaries from Wiktionary, as collected by Li et al. (2012).[1] The size and quality of our tag dictionaries crucially influence how much unambiguous data we can extract, and for some languages, the number of dictionary entries is small.

We can resort to normalization dictionaries to extend Wiktionary's coverage. We do so for English (Han and Baldwin, 2011). It replaces some

[1] https://code.google.com/p/wikily-supervised-pos-tagger/



*Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1256–1261,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

| | |
|---|---|
| NEWSWIRE | Spielberg took the helm of this big budget live action project with Robin Williams playing an adult Peter and Dustin Hoffman as the dastardly Captain Hook. |
| TWITTER | Rooofiii Oooooo, didn't think ppl<3the movie as much as me, this movie will always b the peter pan story2me #robin #williams #hook |
| SPOKEN | I loved that movie... Uhm... You know, Hook. With Robin Williams, uh. |
| QUERIES | peter pan williams movie |

Table 1: Examples from source (top row) and target domains (bottom rows)

spelling variations with the standard form (*youuuuuuuuu* → *you*), which reduces the vocabulary size.

For languages where no such normalization dictionary is available, we use word clusterings based on Brown clusters (Brown et al., 1992) to generalize tags from unambiguous words to previously unseen words in the same class.

| CLUSTER | TOKEN | TAG ∈ $D$ | PROJ. TAG |
|---|---|---|---|
| 01011110 | offish | ADJ | — |
| 01011110 | alreadyyy | ??? | ADV |
| 01011110 | finali | ??? | ADV |
| 01011110 | aleady | ??? | ADV |
| 01011110 | previously | ADV | — |
| 01011110 | already | ADV | — |
| 01011110 | recently | ADV | — |

Figure 1: Example of a Brown cluster with unambiguous tokens, as well as projected tags for new tokens (tokens marked "—" are unchanged in $D'$).

In particular, to extend the dictionary $D$ to $D'$ using clusters, we first run clustering on the unlabeled data $T$, using Brown clustering.[2] We then assign to each unambiguous word in the cluster its tag from dictionary $D$. For all remaining tokens in the same cluster, we assign them the most frequently observed tag in the cluster, provided that label occurred at least twice as often as the second most frequent one, and the token itself was not already in Wiktionary.

As an example, consider the cluster in Figure 1. Since three tokens were unambiguously tagged as ADV in the original dictionary (*previously*, *already*, *recently*), we project ADV to all tokens in the cluster that were not already in $D$ (here: *alreadyyy*, *finali*, *aleady*), and finally add all words to $D'$. The token *offish* remains an ADJ.

[2] https://github.com/percyliang/brown-cluster

### 2.2 Unlabeled data

For each domain and language, given dictionary $D$, we extract unambiguous sentences/tweets. User names and URLs are assumed to be nouns. If all words are unambiguous according to the dictionary, we include the sentence/tweet in our training data. For hashtags on Twitter, we remove the "#" sign and check the remainder against the dictionary. We exclude tweets that *only* contain users and URLs.

The unambiguous subsets of the unlabeled data represent very biased samples of the various domains. The ratio of unambiguous English tweets, for example, is only about 0.012 (or 1 in 84), and the distribution of tags in the Twitter data set is heavily skewed towards nouns, while several other labels are under-represented.

**Twitter** We collect the unlabeled data from the Twitter streaming API.[3] We collected 57m tweets for English, 8.2m for Spanish, 4.1m for Portuguese, and 0.5m for Dutch. We do not perform sentence splitting on tweets, but take them as unit sequences.

**Spoken language** We use the Switchboard corpus of transcribed telephone conversations (Godfrey et al., 1992), sections 2 and 3, as well as the English section of EuroParl (Koehn, 2005) and CHILDES (MacWhinney, 1997). We removed all meta-data and inline annotations (gestures, sounds, etc.), as well as dialogue markers. The final joint corpus contains transcriptions of 570k spoken sentences.

**Search queries** For search queries, we use a combination of queries from Yahoo[4] and AOL. We only use the search terms and ignore any additional information, such as user ID, time, and linked URLs. The resulting data set contains 10m queries.

[3] https://github.com/saffsd/langid.py

[4] http://webscope.sandbox.yahoo.com/

### 2.3 Labeled data

We train our models on newswire, as well as mined unambiguous instances. For English, we use the OntoNotes release of the WSJ section of the Penn Treebank as training data for Twitter, spoken data, and queries.[5] For Dutch, we use the training section of the Alpino treebank from the CoNLL task.[6] For Portuguese, we use the training section of the Bosque treebank.[7] For Spanish, we use the training section of the Cast3LB treebank.[8] In order to map between Wiktionary and the treebanks, we need a common coarse tag set. We thus map all data to the universal tag set (Petrov et al., 2012).

**Dev and test sets** Our approach is basically parameter free. However, we did experiment with different ways of extending Wiktionary and hence used an average over three English Twitter dev sections as development set (Ritter et al., 2011; Gimpel et al., 2011; Foster et al., 2011), all mapped and normalized following Hovy et al. (2014).

For evaluation, we use three domains: tweets, spoken data and queries. For Twitter, we performed experiments in four languages: English, Portuguese, Spanish and Dutch. The Spanish and Portuguese tweets were annotated in-house, which will be made available.[9] For the other languages, we use pre-existing datasets for English (Hovy et al., 2014) and Dutch (Avontuur et al., 2012). Table 2 lists the complete statistics for the different language data sets.

For the other two domains, we use the manually labeled data from Switchboard section 4 as spoken data test set. For queries, we use manually labeled data from Bendersky et al. (2010).

## 3 Experiments

### 3.1 Model

We use a CRF[10] model (Lafferty et al., 2001) with the same features as Owoputi et al. (2013) and default parameters. As baselines we consider a) a CRF model trained only on newswire; b) available off-the-shelf systems (TOOLS); and c) a weakly supervised model (LI10). For English, the off-the-shelf tagger is the Stanford tagger (Toutanova et al., 2003), for the other languages we use TreeTagger (Schmid, 1994) with pre-trained models.

The weakly supervised model trained is on the unannotated data. It is a second-order HMM model (Mari et al., 1997; Thede and Harper, 1999) (SOHMM) using logistic regression to estimate the emission probabilities. This method allows us to use feature vectors rather than just word identity, as in standard HMMs. In addition, we constrain the inference space of the tagger using type-level tag constraints derived from Wiktionary. This model, called LI10 in Table 3, was originally proposed by Li et al. (2012). We extend the model by adding continuous word representations, induced from the unlabeled data using the skip-gram algorithm (Mikolov et al., 2013), to the feature representations. Our logistic regression model thus works over a combination of discrete and continuous variables when estimating emission probabilities. This extended model is called LI10$^+$. For both models, we do 50 passes over the data as in Li et al. (2012).

## 4 Results

Table 3 presents results for various models on several languages. Our results show that our newswire-trained CRF model with target-specific Brown clusters already does better than *all* our other baseline models (TOOLS and weakly LI10) , with the exception of QUERIES, where the Stanford tagger does remarkably well. All improvements are statistically significant ($p < 0.005$, calculated using approximate randomization with 10k iterations).

Adding the unambiguous unlabeled data leads to further improvements, with error reductions (over CRF) of up to 20%. The exceptions here are Portuguese tweets and SPOKEN. For SPOKEN, this is due to the small amounts of unlabeled data, so we re-used the clusters induced on Twitter, reasoning that language use in these two domains is similar to each other. Despite this conjecture, we see small improvements. For English, Portuguese, and Spanish TWITTER, as well as QUERIES, we see further

[5] LDC2011T03.

[6] http://www.let.rug.nl/~vannoord/trees/

[7] http://www.linguateca.pt/floresta/info_floresta_English.html

[8] http://www.iula.upf.edu/recurs01_tbk_uk.htm

[9] http://lowlands.ku.dk/results

[10] https://code.google.com/p/crfpp/

| | TWITTER | | | | SPOKEN | QUERIES |
|---|---|---|---|---|---|---|
| | EN | ES | PT | NL | EN | EN |
| NEWSWIRE | 762k | 93k | 216k | 217k | 762k | 762k |
| UNLABELED | 57m | 9m | 4.5m | 0.5m | 0.6m | 10.1m |
| TEST | 3,064 | 1,524 | 1,593 | 16,725 | 205k | 7,671 |
| words in $D$ | 380k | 240k | 43k | 55k | 380k | 380k |
| % unamb. | 93% | 97% | 98% | 94% | 93% | 93% |
| unamb. inst. | 1.1m | 148k | 134k | 10k | 98k | 1.5m |
| words in $D'$ | 458k | 279k | 332k | 129k | 381k | 388k |
| unamb. inst. | 2.7m | 613k | 892k | 55k | 113k | 2.3m |

Table 2: Characteristics of data sets used in this paper

| DOMAIN | LANG | TOOLS | LI10 | LI10$^+$ | CRF | CRF+$D$ | +CRF+$D'$ |
|---|---|---|---|---|---|---|---|
| TWITTER | en | 80.55 | 81.72 | 83.26 | 86.72 | 87.50 | **87.76** |
| | es | 75.66 | 71.40 | 73.20 | 78.48 | 82.74 | **82.87** |
| | nl | 84.79 | 74.00 | 80.50 | 89.15 | **89.29** | 89.08 |
| | pt | 67.17 | 64.90 | 72.50 | 80.04 | 79.16 | **80.10** |
| SPOKEN | en | 89.02 | 38.72 | 87.86 | 90.53 | **90.54** | * |
| QUERIES | en | 88.06 | 65.96 | 84.39 | 85.52 | 88.06 | **88.28** |

Table 3: Tagging accuracies. TOOLS are off-the-shelf taggers (Stanford and TreeTagger), LI10/LI10$^+$ the weakly supervised models with and without embeddings, and CRF the model trained on newswire with in-domain word clusters. Last two columns show results when extending with unambiguous data. *: Unlabeled data too small to generate clusters with cut-off 100.

considerable improvements by using our extended tag dictionaries.

The most obvious reason this approach should work is the decrease in unseen words in the in-domain evaluation data. Since the unambiguous data is in-domain, the out-of-vocabulary (OOV) rate goes down when we add the unambiguous data to the newswire training data. In fact, for English Twitter, the OOV rate is reduced by half, and for Portuguese and Spanish, it is reduced by about 40%. For Dutch Twitter, the reduction in OOV rate is much smaller, which probably explains the small gain for this dataset. The difference in reduction of OOV rates are due to sample biases in our unlabeled data. This probably also explains the difference in gains between SPEECH and QUERIES. For search queries, the OOV rate is reduced by 66%, whereas it stays roughly the same for speech transcripts.

## 5 Discussion

We have presented a simple, yet effective approach to adapt POS taggers to a new domain. It requires a) the availability of large amounts of unlabeled data and b) a lexicon to mine unambiguous sentences. As sentence length increases, the likelihood of being completely unambiguous drops. For this reason, our approach works well for domains with shorter average sentence length, such as Twitter, spoken language, and search queries.

We also experimented with allowing up to one ambiguous item per sentence, i.e., we include a sentence in our training data if it contains exactly one item that either a) has more than one licensed tag in the dictionary or b) is not in the dictionary. In the first case, we choose the tag randomly at training time from the set of licensed ones. In the second case, we assume the unknown word to be a NOUN, since unknown words mostly tend to be proper names. When added to newswire, this data

results in worse models, presumably by introducing too much noise. However, for low-resource languages or domains with longer sentences and no available newswire data, this might be a viable alternative.

## 6 Related Work

Our approach is similar to mining high-precision items. However, previous approaches on this in NLP have mainly focused on well-defined *classification* tasks, such as PP attachment (Pantel and Lin, 2000; Kawahara and Kurohashi, 2005), or discourse connective disambiguation (Marcu and Echihabi, 2002). In contrast, we mine for *sequences* of unambiguous tokens in a structured prediction task.

While we use the same dictionaries as in Li et al. (2012) and Täckström et al. (2013), our approach differs in several respects. First, we use Wiktionary to mine for *training data*, rather than as type constraints, and second, we use Brown clusters to extend Wiktionary. We did experiment with different ways of doing this, including using various forms of word embeddings, leading to models similar to the baseline models in Socher et al. (2013), but the approach based on Brown clusters led to the best results on our development data.

**?**) use a different approach to distant supervision to improve tagging accuracy for Twitter. They use hyperlinks to fetch additional un-annotated training data that can be used in a self-training loop. Our approach differs in that it produces annotated data and is more readily applicable to various domains.

## 7 Conclusion

We have presented a domain adaptation approach to POS tagging by augmenting newswire data with automatically mined unambiguous instances. We demonstrate our approach on Twitter (in several languages), spoken language transcripts, and search queries. We use dictionaries extended with Brown clusters to collect labeled training data from unlabeled data, saving additional annotation work.

Our models perform significantly better on held-out data than both off-the-shelf taggers and models trained on newswire data only. Improvements hold across several languages (English, Spanish, Portuguese, and Dutch). For spoken language transcripts and search queries, we see some improvements, but find that extending the dictionaries with clusters has less of an effect than for Twitter. Our method can provide a viable alternative to costly annotation when adapting to new domains where unlabeled data and dictionaries are available.

## Acknowledgements

We would like to thank the anonymous reviewers for valuable comments and feedback, as well as Chris Biemann for help with the Twitter data, and Hector Martinez Alonso for the annotation. This research is funded by the ERC Starting Grant LOWLANDS No. 313695.

## References

Tetske Avontuur, Iris Balemans, Laura Elshof, Nanne van Noord, and Menno van Zaanen. 2012. Developing a part-of-speech tagger for dutch tweets. *Computational Linguistics in the Netherlands Journal*, 2:34–51.

Michael Bendersky, Bruce Croft, and David Smith. 2010. Structural annotation of search queries using pseudo-relevance feedback. In *CIKM*.

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

P.F. Brown, P.V. Desouza, R.L. Mercer, V.J. DellaPietra, and J.C. Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *ACL*.

Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Josef Le Roux, Joakim Nivre, Deirde Hogan, and Josef van Genabith. 2011. From news to comments: Resources and benchmarks for parsing the language of Web 2.0. In *IJCNLP*.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *ACL*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *ACL*.

Dirk Hovy, Barbara Plank, and Anders Søgaard. 2014. When pos datasets don't add up: Combatting sample bias. In *LREC*.

Daisuke Kawahara and Sadao Kurohashi. 2005. Pp-attachment disambiguation boosted by a gigantic volume of unambiguous examples. In *IJCNLP*, pages 188–198. Springer.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *ICML*.

Shen Li, João Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *EMNLP*.

B. MacWhinney. 1997. *The CHILDES Database. 5th Edition*. Dublin, OH, Discovery Systems.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *ACL*.

Jean-Francois Mari, Jean-Paul Haton, and Abdelaziz Kriouile. 1997. Automatic word recognition based on second-order hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 5(1):22–25.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL*.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *NAACL*.

Patrick Pantel and Dekang Lin. 2000. An unsupervised approach to prepositional phrase attachment using contextually similar words. In *ACL*.

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

Barbara Plank, Dirk Hovy, Ryan McDonald, and Anders Søgaard. 2014. Adapting taggers to twitter with not-so-distant supervision. COLING.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *EMNLP*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.

Richard Socher, Danqi Chen, Chris Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. 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.

Scott Thede and Mary Harper. 1999. A second-order hidden Markov model for part-of-speech tagging. In *ACL*.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*.