Creating POS Tagging and Dependency Parsing Experts via Topic Modeling

Atreyee Mukherjee
Indiana University
atremukh@indiana.edu

Sandra Kübler
Indiana University
skuebler@indiana.edu

Matthias Scheutz
Tufts University
matthias.scheutz@tufts.edu

Abstract

Part of speech (POS) taggers and dependency parsers tend to work well on homogeneous datasets but their performance suffers on datasets containing data from different genres. In the current work, we investigate how to create POS tagging and dependency parsing experts for heterogeneous data by employing topic modeling. We create topic models (using Latent Dirichlet Allocation) to determine genres from a heterogeneous dataset and then train an expert for each of the genres. Our results show that the topic modeling experts reach substantial improvements when compared to the general versions. For dependency parsing, the improvement reaches 2 percent points over the full training baseline when we use two topics.

1 Introduction

POS tagging and dependency parsing perform well when trained and tested on datasets that are predominantly in the same text domain. However, there is decrease in accuracy for heterogeneous datasets, i.e., for datasets that consist of a mixture of data from different domains. Our current work focuses on improving POS tagging and dependency parsing for such heterogeneous datasets from a variety of different genres by creating experts for automatically detected topics. In our case, the datasets consist of newspaper reports on the one hand and biomedical extracts on the other.

For determining the topic of a sentence, we use Latent Dirichlet Allocation (LDA), which finds the latent topic structure in a document. In our case, a document to be clustered consists of a single sentence. We then assign each sentence to the most likely topic, for both training and test sentences. We consequently train an expert for each topic and then use this expert to POS tag and parse the test sentences belonging to this topic. We assume that the topics that the topic modeler detects do not only pertain to lexical differences, which can be beneficial for the POS tagger and the parser, but also to syntactic phenomena. Thus, one topic may focus on "incomplete" sentences, such as headlines in a newspaper.

Our work is related to domain adaptation since the aim is to improve (morpho-)syntactic analysis for different domains. However, our approach can be regarded as a more general approach to the problem of domains as it is based on automatically determining the genres present in the dataset. Thus, no manual work is involved.

Our results show small to considerable improvements over a competitive baseline of using the full training set. For POS tagging, there is an improvement of 0.3 percent points over the full training set. For dependency parsing, the gain is more pronounced: almost 2% over the full training set.

The remainder of the paper is structured as follows: Section 2 discusses our research questions and section 3 the related work in the area. Section 4 describes the setup for our experiments, and section 5 shows the experimental results. We draw our conclusions in Section 6.

2 Research Questions

Our aim is to create POS tagging and parsing experts for heterogeneous datasets, with sentences from different genres. For example, the dataset might be a mixture of newspaper articles, blogs, financial reports, research papers and even specialized texts such as biomedical research papers ,and law texts. We create experts such that each

expert would learn specific information about its own genre. We determine these experts by performing topic modeling on sentences and then train an expert on the sentences of the topic. We group sentences based on their most probable topic. To test our hypothesis that topic modeling can serve to group sentences into topics, we create a mixed dataset from the financial domain (using the Penn Treebank (Marcus et al., 1994)) and from the biomedical domain (using the GENIA Corpus (Tateisi and Tsujii, 2004)) such that the new handcrafted corpus consists of sentences from both domains in equal measure. Consequently, there is a clear difference in the genres in our corpus, and we have gold standard topic information.

We perform topic modeling on training and test data simultaneously: We assign a test sentence to the topic with the highest probability. This means that we currently simplify the problem of assigning new sentences to topics. In the future, we plan to assign new sentences to topics based their similarity to sentences in the topics created during training, following the work by Plank and van Noord (2011).

Our current research focuses on answering the following questions for POS tagging and parsing tasks:

Question 1: Does Topic Modeling Detect Topics?

In this question, we investigate whether an unsupervised topic modeler can detect topics in a heterogeneous corpus. We use our artificially created heterogeneous corpus containing sentences from the Wall Street Journal (WSJ) section of the Penn Treebank (Marcus et al., 1994) and from the GENIA Corpus (Tateisi and Tsujii, 2004) and take their original corpus as the gold standard topic. We assume that a good split into the known topics, financial news and biomedical abstracts, will also improve POS tagging and parsing accuracy. If we assume two topics, we should be able to see a clear distinction between WSJ and GENIA sentences. I.e., for each topic, we should have a clear correspondence of its sentences to either WSJ or GENIA. We thus calculate the percentage of sentences in a given topic that belong to GENIA and expect that one topic should have a high percentage and the other one a low percentage. We also experiment with a larger number of topics, to see if we can profit from a finer grained topic definition. However, this advantage will be offset by a smaller training set since we split into 10 sets.

Question 2: Does POS Tagging Benefit from Using Topics?

In this question, we examine whether the performance of POS tagging improves if we create experts based on the topics detected by the topic modeler. Thus, we use the topics created for the previous sections and train a POS tagging expert on the training part of each topic. We then use the expert to tag the test sentences from this topic. In this setting, we can see if the experts can effectively handle the data sparseness caused by dividing the training set into multiple experts. We experiment with one setting in which we use topic modeling as hard clustering, i.e., we assign each sentence to the topic for which the topic modeler gave the highest probability. We also experiment with soft clustering, in which we add each sentence to all topics, weighted by its probability distribution.

Question 3: Does Dependency Parsing Benefit from the Topics?

Here, we investigate the effects of using topic modeling experts for dependency parsing. We first use gold POS tags in order to abstract away from POS tagging quality. In a second step, we investigate the interaction between POS tagging and parsing experts. I.e., we are interested in whether dependency parsing can profit from using the POS tags that were determined by the POS tagging experts. This allows us to determine whether assimilating POS information given by the POS experts can improve dependency parsing or whether there is no interaction between the two levels.

Question 4: What do the Experts Learn?

In this question, we will analyze the results from question 3 in more detail to investigate how the topic modeling experts improve parsing results. We are interested in whether there are specific types of sentences or dependencies that are grouped by the topic models, so that the parsing experts focus on a specific subset of syntactic properties.

3 Related Work

To the best of our knowledge, there is little direct correlation between our work on POS tagging and

parsing experts to that of the previous work done in the area. However, our work is comparable to domain adaptation since we create experts to tag and parse heterogeneous datasets. The work in this is area is largely driven by the unavailability of examples from target domain. Our work focuses on creating experts using topic modeling which will be able to tag and parse target domain sentences belonging to a specific topic. Compared to POS tagging, there has been significant work on domain adaptation in dependency parsing.

Dredze et al. (2007) concluded that domain adaptation is more challenging when there are dissimilarities in annotation schemes between the treebanks. Blitzer et al. (2006) experimented on structural correspondence learning (SCL) which focuses on finding "frequently occurring" pivot features that occur commonly across domains in the unlabeled data but equally characterize source and target domains. Similar to our work, Blitzer et al. used the WSJ as the source and MEDLINE abstracts as the target domain. They established that SCL reaches better results in both POS tagging and parsing than supervised and semi-supervised learning even when there is no training data available on the target domain.

For POS tagging, Clark et al. (2003) applied an agreement-based and a baseline co-training method by using a Markov model tagger and a maximum entropy tagger. In case of the baseline, all the sentences from one tagger are added to train the other whereas in the agreement-based method, both taggers have to reach to the same decision for a sentence to be added to the training. Kübler and Baucom (2011) used a similar concept but with three different taggers and showed that selecting sentences as well as sequences of words for which all taggers agree yield the highest gains. Sagae and Tsujii (Sagae and Tsujii, 2007) emulate a single iteration of co-training by using MaxEnt and SVM, selecting the sentences where both models agreed and adding these sentences to the training set. Their approach reached the highest results on the domain adaptation task of CoNLL 2007 (Nivre et al., 2007).

Domain adaptation was the task of the CoNll 2007 shared task. Attardi et al. (2007) used a tree revision method that corrects the mistakes caused by the base parser for the target domain. Later, Kawahara and Uchimoto (2008) employed a single parser approach using second order MST Parser and combining labeled data from the unknown domain with unlabeled data of the known domain by simple concatenation and judging the efficacy of the resulting most reliable parses. Finkel and Manning (2009) devised a new model for named entity recognition as well as dependency parsing by using hierarchical Bayesian prior. This is influenced by the notion that different domains may have different features which is specific to each domain. However, instead of applying a constant prior over all the parameters, a hierarchical Bayesian global is used. This enables sharing of information across domains but also allows to override this information if there is ample evidence.

McClosky et al. (2010) designed the problem as "multiple source parse adaptation", in which a parser was trained on multiple domains and learned the statistics as well as domain differences which affects the parser accuracy. Their parser outperforms the state-of-the-art baselines. This approach is the closest to our work as we create experts based on topics, and each expert learns the specifics of the particular topic with which it is associated.

The closest approach to ours is the one by Plank and van Noord (2011), who employ a similar idea of using topic modeling for creating parsing experts. However, their task is to determine in a domain adaption setting which sentences of an out-of-domain training set are the most similar to the test set. Thus, they create a specialized training set for every document they need to parse while we create more general experts. In the work by Plank and van Noord (2011), the topic distribution in a document is used as features for their similarity metrics.

4 Experimental Setup

4.1 Data Sets

For our experiments, we use the Wall Street Journal (WSJ) section of the Penn Treebank (Marcus et al., 1994) and the GENIA Corpus (Tateisi and Tsujii, 2004). Both corpora use the Penn Treebank POS tagset (Santorini, 1990) with minor differences: The tagset used in GENIA is based on the Penn Treebank tagset, but it uses the tags for proper names and symbols only in very restricted contexts.

For the WSJ corpus, we extract the POS annotation from the syntactically annotated corpus. The

GENIA Corpus comprises biomedical abstracts from Medline, and it is annotated on different linguistic levels, including POS tags, syntax, coreference, and events, among others. We use GENIA 1.0 trees (Ohta et al., 2002) created in the Penn Treebank format[1] . Both treebanks were converted to dependencies using pennconverter (Johansson and Nugues, 2007).

For our experiments, we need a balanced data set, both for the training and the test set. Since GENIA is rather small and since there is no standard data split for GENIA, we decided to extract the last 850 sentences for the test set. The remaining 17 181 sentences are used for training. For WSJ, we chose the same number of sentences for both training and the test set, the training sentences are selected randomly from sections 02-21 and the test sentences from section 22.

4.2 Topic Modeling

Probabilistic topic modeling is a class of algorithms which detects the thematic structure in a large volume of documents. Topic modeling is unsupervised, i.e., it does not require annotated documents (Blei, 2012) but rather discovers similarity between documents. Latent Dirichlet Allocation (LDA) is one of the topic modeling algorithms. It is a generative probabilistic model that approximates the underlying hidden topical structure of a collection of texts based on the distribution of words in the documents (Blei et al., 2003).

We use the topic modeling toolkit MALLET (McCallum, 2002). The topic modeler in MALLET implements Latent Dirichlet Allocation (LDA), clustering documents into a predefined number of topics. As a result, it provides different types of information such as:

- Topic keys: The highest ranked words per topic with their probabilities;
- Document topics: The topic distribution for each document (i.e., the probability that a document belongs to a given topic); and
- Topic state, which correlates all words and topics.

For our experiments, we use sentences as documents. Based on the document topic information, we then group the sentences into genre topics. We collect all sentences from the training and test set, cluster them via the MALLET topic modeler, and determine for which expert(s) the sentence is relevant. There are several ways of determining the best expert, see below. Then, we separate the sentences for each expert into training and test sentences, based on the previously determined data splits (see above).

We can determine experts based on hard or soft clustering decisions: For hard clustering, the sentences are assigned to hard topics, based on the topic that has the highest probability in that sentence. I.e., if for sentence s_x, MALLET lists the topic t_1 as the topic with the highest probability, then s_x is added to the data set of topic t_1. In other words, the data set of topic t_1 consists of all sentences for which MALLET showed topic t_1 as the most likely topic. This means that the data set sizes vary between topics.

For soft clustering experiments, we utilize the entire topic distribution of a sentence by weighting sentences in the training data based on their topic distribution. We simulate weighting training sentences by adding multiple copies to the training files of the experts. Thus, for 2-topic experiments, a sentence with 80% probability for topic 1 will be included 8 times in the expert for topic 1 and 2 times in the expert for topic 2, rounding up small percentages so that every sentence will be added to every expert at least once. Thus, we use a more fine grained topic model, mitigate data sparseness, but we risk adding non-typical / irrelevant sentences to experts.

4.3 POS Tagging

For part of speech tagging, we use the TnT (Trigrams'n'Tags) tagger (Brants, 2000). TnT is based on a second order Markov Model and has an elaborate model for guessing the POS tags for unknown words. We use TnT mainly because of its speed and because it allows the manual inspection of the trained models (emission and transition frequencies).

4.4 Dependency Parsing

For the parsing experiments, we used the dependency parser of the MATE Tools[2], a Java implementation of a graph-based parser (Bohnet, 2010). The input follows CoNLL 2009 Shared Task Format.

[1] http://nlp.stanford.edu/ mcclosky/biomedical.html

[2] code.google.com/p/mate-tools

T.	2 topics		10 topics	
	% in train	%in test	% in train	% in test
1	0.71	0.71	0.48	0.52
2	97.99	98.6	98.58	98.35
3			1.16	0.73
4			94.87	97.14
5			0.17	0
6			0.28	0.29
7			99.47	99.12
8			98.93	100
9			98.92	99.33
10			94.85	95.35

Table 1: Distribution of sentences from the WSJ+GENIA data set given 2 and 10 topics (showing the percentage of GENIA sentences per topic).

4.5 Evaluation

We used the script `tnt-diff` that is part of TnT to evaluate the POS tagging results and the CoNLL Shared Task evaluation script[3] for evaluating the parsing results.

4.6 Baselines

We use two baselines. As the first baseline, we take the complete training set when no topic modeling is performed. Note that this is a very competitive baseline since the topic modeling experts have access to considerably smaller amounts of training data. In order to avoid differences in accuracy resulting from different training set sizes, we create a second baseline by splitting the sentences randomly into the same number of groups as the number of topics, while maintaining the equal distribution of WSJ and GENIA sentences. I.e., we assume the same number of random "topics", all of the same size. Thus, in the 2-topic setting with the genres, we create two separate training sets, each containing half of the WSJ training set and half of the GENIA one. In this setting, we test all experts on the whole test set and average over the results.

5 Experimental Results

5.1 Does Topic Modeling Detect Topics?

Here we investigate whether LDA can separate the sentences into meaningful topics. Table 1 shows the distribution of sentences of training and test set into different topics when we assume 2 or 10 topics. These results indicate that the topic modeler separates topics very efficiently. For the 2-topic experiments, a clear split is evident as the majority of the GENIA sentences are clustered in topic 2; the misclassified sentences constitute less than 1%. For the 10-topic experiments, we notice that topics 2, 4, 7, 8, 9, 10 contain mainly GENIA sentences while the remaining topics cover mainly WSJ sentences. In both settings, the error rate is between 0.2% and 5%, i.e., we obtain a distinct split between GENIA and WSJ, which should give us a good starting point for the following experiments.

[3] http://ilk.uvt.nl/conll/software/eval.pl

Setting	Accuracy	
	2 topics	10 topics
Full training set	96.69	
Random split	96.41	95.48
Topic model	**96.95**	96.38
Soft Clustering	96.8	**96.88**

Table 2: Results of the POS tagging experiments.

5.2 POS Tagging Experts

In this section, we investigate whether the POS tagger can benefit from using topic modeling, i.e., whether POS tagging results can be improved by training experts for genres provided by topic modeling. We compare the topic modeling approach to our two baselines for the 2-topic and 10-topic setting. We also perform a soft clustering experiment, in which each sentence is added to every topic, weighted by its probability (see section 4.2).

The results in Table 2 show that if we assume a 2-topic setting, the experts perform better than both baselines, i.e., the model trained on the full training set and the model with randomly chosen "topics". The 2-topic expert model reaches an accuracy of 96.95%, which is slightly higher than the full training set accuracy of 96.69%. We know that the 2-topic setting provides a clear separation between WSJ and GENIA (Table 1). Thus, this setting outperforms the full training set using a smaller amount of training data. There is also an increase of 0.54 percent points over the accuracy of the 2 random split setting.

For the 10-topic setting, the topic expert model outperforms the random split of the same size by 0.9 percent points, which is a higher difference than for the 2-topic setting. This shows that the

Setting	LAS		UAS	
	2 topics	10 topics	2 topics	10 topics
Full training set	88.67		91.71	
Random split	87.84	84.91	90.86	88.64
Topic model	**90.51**	88.38	**92.14**	90.3
Soft clustering	89.86	**89.91**	91.99	**91.84**

Table 3: Results of the dependency parsing experiments using gold POS tags.

finer grained splits model important information. However, the topic expert model does not reach the accuracy of the baseline using the full training set. This can be attributed to the reduced size of the training set for the experts.

Since training set size is a detrimental factor for the larger number of topics, we also conducted an experiment where we used soft clustering so that every sentence is represented in every topic, but to a different degree. The last row in table 2 reports the results of this experiment. We notice that the 2-topic experts cannot benefit from the soft clustering. Since the separation between WSJ and GENIA is very defined for the 2-topic experiments, the advantage of having a larger training set is outweighed by too many irrelevant examples from the other topic. However, the 10-topic model profits from the soft clustering, which indicates that soft clustering can alleviate the data sparseness problem of the POS tagging experts for larger numbers of topics. A more detailed analysis of the POS tagging results (on a slightly different data split), see (Mukherjee et al., 2016). This work includes an experiment showing that the POS tagging experts also increase performance for the WSJ corpus only. This means, POS tagging experts also perform better on more homogeneous collections.

5.3 Dependency Parsing Experts

5.3.1 Using Gold POS Tags

We now look into the parsing experiments using Gold standard POS tags. The choice of gold POS tags allows us to focus on the contribution of the topic modeling experts on parsing results.

The results of the experiments are shown in Table 3, for 2-topic and 10-topic settings and in comparison to the two baselines. We also perform a soft clustering experiment. The results indicate that the 2-topic expert model reaches an improvement over the baseline using the full training set for both the labeled attachment score (LAS) and the unlabeled attachment score (UAS). We find an increase of around 2% over the baseline for LAS, and an increase of 0.43% for UAS. However, for the 10-topic setting, both the LAS and the UAS are slightly lower than the baseline. For LAS, the difference is 0.29 percent points while for UAS, the difference is 1.41 percent points. This shows that the gain in LAS and UAS is offset by the reduced training set, parallel to the results for POS tagging. Both the 2-topic and the 10-topic experts outperform the random split baseline (which uses similar training set sizes), with a gain of more than 3 percent points.

The soft clustering results show the same trends as in the POS tagging experiments: For 2-topic setting, soft clustering outperforms the full baseline by 1.19 percent points. But it does not exceed the hard clustering results. In the 10-topic setting, soft clustering outperforms the full baseline as well as the hard clustering setting. This is because sentences with a 50% probability of belonging to topic 1 and a 40% probability for topic 3 need to be considered to belong to both topics. This result also shows that this method effectively handles the training data sparsity in the 10-topic setting.

5.3.2 Using the POS Tagger

In section 5.3, we use the gold standard POS tags in the POS tags. In this section, we explore the results of using POS tags from the POS tagger TnT as the input for the parser. This gives rise to four major scenarios:

1. The full training set is used for POS tagging and for parsing (full baseline).
2. Random splits are used for parsing and POS tagging. I.e., the POS tagger and parser are trained on random splits (random baseline).
3. Topic models are used for training the parser, but TnT is trained on the whole training set.
4. Topic models are used for training the parser and the POS tagger.

Setting	LAS		UAS	
	2 topics	10 topics	2 topics	10 topics
1. Full set POS + full set parsing	86.70		90.26	
2. Random split POS + random split parsing	85.77	81.33	89.11	85.73
3. Full set POS + topic model parsing	88.30	86.13	90.43	88.47
4. Topic model POS + Topic model parsing	**88.35**	85.68	**90.55**	88.15

Table 4: Results of the dependency parsing experiments using TnT POS tags.

Sentence	Fulltext LAS	2-topic LAS
Phyllis Kyle, Stephenson Newport News , Va .	0	25.00
But volume rose only to 162 million shares from 143 million Friday .	46.15	61.54
Fidelity , for example , prepared ads several months ago in case of a market plunge .	47.06	82.35
CALL IT un-advertising .	50.00	75.00
(See related story : " And Bills to Make Wishes Come True " – WSJ Oct. 17 , 1989 .	52.38	61.90

Table 5: Comparison of LAS for the sentences with the lowest LAS in the fulltext setting.

We use the random split case as the lower baseline for these experiments and the full training set as the more competitive baseline. Table 4 shows the results.

Table 4 shows that in the 2-topic setting, using topic modeling experts on the POS level as well as on the parsing level reaches the highest results with an improvement of around 2% in LAS in comparison to the full baseline parser, from 86.70% to 88.35%. The gain in UAS is considerably smaller: The topic modeling expert reaches 90.55% as opposed to 90.26% for the full baseline. In contrast, the topic modeling setting for the 10-topic setting outperforms the random baseline but does not reach the full baseline, thus mirroring the trends we have seen before.

When we compare the experiments where we use the full POS tagging baseline along with topic model parsing experts (row 3. in table 4) to the full topic model (row 4.), we see that the latter model reaches only very minimal gains by using the topic modeling POS tagger when we use 2 topics, and we have a negative trend when we use 10 topics. I.e. the overall quality of the POS tagger is more important than its specialization. Thus, even if the topic model POS tagger outperforms its full baseline, the learned adaptations only have a minimal effect on parsing accuracy.

5.4 Analysis of Results

We now have a closer look at the results presented for the parsing experiments using gold POS tags in section 5.3.1. The results show that the 2-topic parsing experts outperform the general parser trained on the full training set by almost 2 percent points. We looked at the 5 sentences that had the lowest LAS when we used the general parser. These sentences are shown in table 5, along with their LAS for both settings. The table clearly shows that the topic expert parsers reach a much higher LAS across all these sentences, and the highest increase reaches 35 percent points. We also see that there are two headlines among these sentences. They are different in their syntactic patterns from other sentences and thus difficult to parse. For this reason, we decided to have a closer at all "incomplete" sentences, i.e., sentences that do not have verbs, as an approximation of headlines. We found that of the 1 310 sentences in the training set, 437 were grouped into topic 1, the other 873 sentences in topic 2. In the test set, we had 65 such sentences, 15 in topic 1 and 50 in topic 2. For the sentenced in topic 1, we calculate an LAS of 76.54, for the ones in topic 2 an LAS of 89.91. These results show that the parser expert for topic 2 has adapted substantially better to the syntax of such untypical sentences than the parser for topic 1.

We also looked at the dependency labels that were mislabeled most often by the more general,

Gold Dep.	Pred. Dep.	Fulltext	Topic 1	Topic 2
ADV	NMOD	121	37	86
PMOD	NMOD	101	21	67
NMOD	ADV	100	34	57
AMOD	NMOD	91	26	83
CONJ	NMOD	86	13	56

Table 6: The 5 most frequent dependency label confusions of the full baseline parser.

full baseline parser. The 5 most frequent combinations are shown in table 6, with their frequencies in the test sentences of the two topics. These numbers show that the topic 1 expert is much better adapted to these confusion sets, resulting in lower error rates than the topic 2. This shows very dramatically that the two topics learn different patterns.

6 Conclusion and Future Work

In these experiments, we have shown that we can improve parsing results on heterogeneous domains by using unsupervised topic modeling to separate the data into different topics. We can then train POS tagging and parsing experts on the individual topics, which show an increased accuracy in comparison to their counterparts trained on the whole, heterogenous training set. We can mitigate the data sparsity resulting from having to split the training set into different topics by assigning every sentence to every topic but weighting their importance to a topic by the probabilities of the topic modeler. We also showed that while the POS tagger and the dependency parser individually profit from the split into topic experts, the combination of topic expert POS tagger and parser does not improve over using a POS tagger trained on the whole data set. We will have to investigate the reasons for this behavior in future work.

We will also investigate methods of how to assign test sentences to topics without having to rerun the topic modeler on the whole data set.

Acknowledgements

References

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, Atanas Chanev, and Massimiliano Ciaramita. 2007. Multilingual dependency parsing and domain adaptation using DeSR. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 1112–1118.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 120–128, Sydney, Australia.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 89–97, Beijing, China.

Thorsten Brants. 2000. TnT–a statistical part-of-speech tagger. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics and the 6th Conference on Applied Natural Language Processing (ANLP/NAACL)*, pages 224–231, Seattle, WA.

Stephen Clark, James Curran, and Miles Osborne. 2003. Bootstrapping POS-taggers using unlabelled data. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL)*, Edmonton, Canada.

Mark Dredze, John Blitzer, Partha Pratim Talukdar, Kuzman Ganchev, João Graca, and Fernando Pereira. 2007. Frustratingly hard domain adaptation for dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 1051–1055, Prague, Czech Republic.

Jenny Rose Finkel and Christopher D. Manning. 2009. Hierarchical Bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 602–610.

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In *Proceedings of NODALIDA 2007*, pages 105–112, Tartu, Estonia.

Daisuke Kawahara and Kiyotaka Uchimoto. 2008. Learning reliability of parses for domain adaptation of dependency parsing. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*, Hyderabad, India.

Sandra Kübler and Eric Baucom. 2011. Fast domain adaptation for part of speech tagging for dialogues. In *Proceedings of the International Conference on Recent Advances in NLP (RANLP)*, Hissar, Bulgaria.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop, HLT 94*, pages 114–119, Plainsboro, NJ.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36. Association for Computational Linguistics.

Atreyee Mukherjee, Sandra Kübler, and Matthias Scheutz. 2016. POS tagging experts via topic modeling. In *Proceedings of the 13th International Conference on Natural Language Processing (ICON)*, Varanasi, India.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL 2007 Shared Task. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL*, pages 915–932, Prague, Czech Republic.

Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. 2002. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 82–86, San Francisco, CA.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, OR.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, volume 2007, pages 1044–1050, Prague, Czech Republic.

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. Department of Computer and Information Science, University of Pennsylvania, 3rd Revision, 2nd Printing.

Yuka Tateisi and Jun'ichi Tsujii. 2004. Part-of-speech annotation of biology research abstracts. In *Proceedings of 4th International Conference on Language Resource and Evaluation (LREC)*, Lisbon, Portugal.