POS Tagging 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 taggers generally perform well on homogeneous data sets, but their performance often varies considerably across different genres. In this paper we investigate the adaptation of POS taggers to individual genres by creating POS tagging experts. We use topic modeling to determine genres automatically and then build a tagging expert for each genre. We use Latent Dirichlet Allocation to cluster sentences into related topics, based on which we create the training experts for the POS tagger. Likewise, we cluster the test sentences into the same topics and annotate each sentence with the corresponding POS tagging expert. We show that using topic model experts enhances the accuracy of POS tagging by around half a percent point on average over the random baseline, and the 2-topic hard clustering model and the 10-topic soft clustering model improve over the full training set.

1 Introduction

Part-of-speech (POS) tagging is the task of assigning word classes to lexical items and is often considered a solved problem. However, even though we can reach high accuracies on the Penn Treebank, POS taggers are sensitive to differences in genre (cf. e.g. (Khan et al., 2013; Miller et al., 2007; Søgaard, 2013)). In the current research, we investigate a novel way of adapting POS taggers to different genres, but also to specific lexical and syntactic characteristics of texts. We propose to use topic modeling, an unsupervised soft clustering method that clusters documents, or sentences in our case, into a distribution of individual topics. We interpret the topics as specialized training sets, which are used to train a POS tagging expert for each topic. Test sentences are also clustered into the same topics, and each test sentence is annotated by the corresponding POS tagging expert. We investigate different methods of converting topics into expert training sets.

Thus, our method is related to domain adaptation approaches (Khan et al., 2013; Miller et al., 2007) in that it focuses on adapting to specific characteristics of texts, but it is more generally applicable because it determines the domains and the experts automatically. It is also related to approaches of mitigating domain effects (e.g., (Søgaard, 2013)), but in contrast to those methods, we obtain individual experts that can be used and investigated separately.

Our results show that the topic modeling experts are sensitive to different genres (financial news vs. medical text) as well as to smaller differences between the Wall Street sentences. On average, the improvement over randomly selected subsets is around 0.5-1 percent point. Our results also show that one major difference between the POS tagging experts based on topics models concerns the treatment of unknown words. In the financial expert, such words have a much higher tendency to be assigned to the noun class. And even though names are one of the most difficult classes, the error rate for them is reduced in the POS experts based on topic models.

The remainder of the paper is structured as follows: Section 2 discusses our research questions in more detail. Section 3 discusses related work, and in section 4, we provide details about the data sets, the topic modeler, and the POS tagger. In section 5, we show the results, and in section 6, we draw our conclusions and discuss future extensions of our work.

2 Research Questions

Our investigation into creating POS tagging experts is based on the assumption that the data that we need to analyze is not homogeneous but rather a collection of different text types or even syntactic constructions. In one setting, we may have a mixed set of newspaper articles, research papers, financial reports, and weblogs to analyze. In a different setting, we may have texts that use specialized vocabulary, such as in the biomedical domain or in law texts, or we could have headlines with an elliptical sentence structure. For this reason, we assume that the POS tagger can reach a higher accuracy if we can split the data sets into more homogeneous subsets and then train individual expert POS taggers, specialized for individual subsets. We determine these homogeneous subsets by using topic modeling. Since topic modeling is unsupervised, the sentences will be divided into sets based on similarity. This similarity may be based on similarity of content, but it can also be based on similarity on the structural level. For example, if we use the Penn Treebank (Marcus et al., 1994), we could assume that one topic consists of sentences reporting changes of the stock market while another topic consists of sentences about the earthquake in San Francisco. Yet, another topic may consist of mostly questions.

Our current research concentrates on answering the four questions described below. In this investigation, we use a setting in which we perform topic modeling jointly on the training and test data. This is a simplification of the problem since this means that we would have to create new experts every time a new sentence needs to be tagged. We assume that we can rerun the topic modeling including test sentences and then match the new topics to the ones we obtained on the training set alone. Another approach would be to use the similarity metrics by Plank and van Noord (2011). We will test these hypotheses in the future.

2.1 Question 1: Do Topic Models Provide Information from which POS Tagging Experts can Profit?

The first question is concerned with determining whether the data splits we obtain from topic modeling are meaningful for creating POS tagging experts. In other words, do the topics that we can generate in an unsupervised manner provide a specialization that has an effect on POS tagging? In order to investigate this question, we manually generate a two-topic corpus by combining data from the Wall Street Journal (WSJ) section of the Penn Treebank (Marcus et al., 1994) and from the GENIA corpus (Tateisi and Tsujii, 2004). The WSJ covers financial news while GENIA uses Medline abstracts as its textual basis. As a consequence, we have sentences from two different genres, but also slight variations in the POS tagsets. 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. This setup allows us to test whether the topic modeler is able to distinguish the two genres, and whether POS tagging experts can profit from this separation.

2.2 Question 2: Can we use Soft Clusters of the Topic Models?

The first set of experiments uses the topics as hard clusters, i.e., every sentence belongs to the topic with the highest probability. This is a simplification since a sentence can represent different topics to different degrees. Thus, we now investigate whether we can utilize the soft clustering information directly and add every sentence to every POS tagging expert, weighted based on the degree to which it represents the topic of this expert. This not only allows us to model topics in more detail, it can also help combating data sparsity since every sentence contributes to every POS expert. The risk is that we "diffuse" the expert knowledge too much by adding all sentences even if they are weighted.

2.3 Question 3: Can Topic Modeling Detect Micro-Genres?

While the previous sets of experiments used two very different genres, the current question focuses on data from within one genre. Can we use topic modeling within one genre, and do the resulting topics allow us to create POS tagging experts for "micro"-genres? To investigate this question, we exclusively use the WSJ data set. Our hypothesis is that the WSJ corpus contains different newspaper sections, which may use different styles. Since there is no information available from the Penn Treebank about those section, we cannot evaluate how well the topic modeler splits the sentences into topics, but we can evaluate whether the POS tagging experts are successful in adapting to those micro-genres.

2.4 Question 4: Which Specialization do the POS Tagging Experts Learn?

Here, we will take a closer look at the results from the first question to investigate where the improvements by the POS tagging experts come from. Are all the improvements based on lower rates of out-of-vocabulary words? For example, suppose we have two experimental settings, both using the same size of the training set, but in one setting, the majority of the training set is from GENIA while in the second setting, the training set is a mix of GENIA and WSJ. It is more likely that the former will contain a wider range of biomedical vocabulary than the latter. However, it is also possible that the experts will learn different regularities, for example with regard to how the proper name tags are used in the two corpora. Thus, we will look at the ratio of unknown words in the different experiments and at the error rates of known and unknown words. We will additionally look at the confusion matrices.

3 Related Work

We are not aware of any research that directly compares to the research presented here. The closest area is domain adaptation. For this reason, we will cover work on domain adaptation for POS tagging here. However, more work has been done on domain adaptation for parsing. The work in that area seems to fall into two categories: "frustratingly easy" when some annotated data from the target domain is available (Daumé III, 2007) and "frustratingly hard" if no such target data is available (Dredze et al., 2007).

For POS tagging, Clark et al. (2003) used the results of one POS tagger on unannotated data to inform the training of another tagger in a semi-supervised setting using a co-training routine with a Markov model tagger and a maximum entropy tagger. The authors tested both agreement-based co-training, where the sentences are added to training only if the taggers both agree, and naive co-training, where all sentences from one tagger are added to the training of the other, with no filter. Kübler and Baucom (2011) expanded on this method and used three different taggers to annotate additional data and then select those sentences for which the taggers agree. They found that adding not only complete sentences but also sequences of words where the taggers agree results in the highest gains. Khan et al. (2013) investigated the situation where some annotated target data is available. They focused on optimizing the balance between source and target sentences. They found that selecting sentences that are the most similar to the target data results in the highest gains. Blitzer et al. (2006) developed structural correspondence learning, which learns correspondences between two domains in settings where a small set of target sentences is available as well as in an unsupervised setting. They show an improved performance for POS tagging and for parsing when using the adapted POS tagger.

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 (version 3.02) (Tateisi and Tsujii, 2004). Both corpora use the Penn Treebank POS tagset (Santorini, 1990) with minor differences, as described in section 2.1.

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 the POS tagged version. For WSJ, we use the standard data split for parsing: using sections 02-21 as training data and section 22 as our test set. We reserve section 23 for future parsing expert experiments.

For questions 1, 2, and 4, we need a balanced data set, both for the training and the test set. Since GENIA is smaller than WSJ and has no predefined data split, we have decided to use the same test set size (1 700 sentences), but now taking half of the sentences from WSJ and half from GENIA. The remaining GENIA sentences serve as half of the training set, and we extract the same number of sentences from WSJ. For GENIA, we consider the first 19 696 sentences as training set and the remaining 850 sentences as test set. For WSJ, the sentences are selected randomly out of the predefined training and test sets.

For question 3, we use the full WSJ training and test set, as described above. Table 1 gives an overview of the settings.

4.2 Topic Modeling

Probabilistic topic modeling is a class of algorithms which detects the thematic structure in a

Setting	Corpus	Training	Test
question 1, 2, 4	WSJ	19 696	850
	GENIA	19 696	850
question 3	WSJ	39 832	1 700

Table 1: Overview of the data sets.

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: This 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 question 1 and 3, 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 questions 2 and 3, we utilize the entire topic distribution of a sentence by weighting sentences in the training data based on their topic distribution. Since the POS tagger does not support the weighting of training examples and since we do not have access to the code of the POS tagger, we simulate weighting training sentences by adding multiple copies to the training files of the experts. Thus, for the 2-topic experiments, a sentence with 80% probability for topic 1 will be included 80 times in the expert for topic 1 and 20 times in the expert for topic 2. We repeat these experiments, adding a sentence per every 10%, but 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 to mitigate data sparseness, but we risk adding non-typical or 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 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 where applicable. I.e., we assume the same number of random "topics", all of the same size. Thus, in the 2-topic setting with the 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.

	2 topics		5 topics		10 topics	
T.	% in train	%in test	% in train	%in test	% in train	% in test
1	91.06	99.61	99.33	99.64	93.04	100
2	8.74	9.70	17.71	21.00	94.17	98.64
3			99.44	98.78	94.49	98.72
4			93.84	98.75	4.24	2.92
5			0.20	0.19	5.15	5.41
6					95.42	99.16
7					4.40	5.24
8					96.26	100
9					3.59	5.43
10					63.10	80.85

Table 2: Distribution of sentences from the WSJ+GENIA data set given 2, 5, and 10 topics (showing the percentage of GENIA sentences per topic).

1	cells cell expression il nf activation human binding gene transcription protein kappa ab cd ti factor alpha activity induced
2	mr million ui year company market stock billion share corp years shares trading president time quarter sales government business

Table 3: Examples of words in topics for the 2-topic experiments on the WSJ+Genia corpus.

	Accuracy		
Setting	2 topics	5 topics	10 topics
Full training set	96.64	96.64	96.64
Random split	96.48	95.92	95.49
Topic model	**96.84**	96.54	96.34

Table 4: Comparing the topic model experts to the baselines on the WSJ+GENIA data set.

5 Experimental Results

5.1 WSJ+GENIA Experiments

In this set of experiments, we use the manually created corpus that contains WSJ and GENIA sentences in equal parts. A logical first setting is to have the topic modeler distinguish between two different topics, to see if these two topics correspond to the two gold topics, WSJ and GENIA. We repeat the experiment using 5 and 10 topics to see if a finer granularity improves results. We then use the trained POS tagging experts to annotate the test sentences based on their assigned topic.

Investigating the topic modeler splits. The distributions of sentences in the training set and test set resulting from topic modeling are shown in Table 2. In the case of the 2 clusters, we see a clear split. A vast majority of GENIA sentences are clustered into the first topic, and less than 10% are clustered into the second topic. In the case of 5 and 10 topics, the split is even clearer. For example. for the 10-topic setting, topics 4, 5, 7, and 9 represent WSJ topics while the others are GENIA topics. Here, the error rate is between 3% and 6%. Both sets have one outlier, topic 2 for the 5-topic setting, and topic 10 for the 10 topics, which are most likely the topics for difficult to classify sentences. Thus, in all cases, we have good splits, which should allow the POS tagging experts to learn specifics of the two corpora. Table 3 shows example words from the 2-topic experiment, which show a clear separation of topics into biomedical and financial terms.

POS tagging experiments. The results of the POS tagging experiments for the 2-topic, 5-topic, and the 10-topic settings are shown in Table 4. The results show that the experts created by the topic models outperform the randomly split models in all cases: For the 2-topic setting, we see the smallest increase from 96.48% to 96.84%, while the 10-topic setting reaches the largest increase, from 95.49% to 96.34%. However, note that the results in the 5- and 10-topic settings are slightly lower than the ones in the 2-topic setting. This is due to the reduced training set size.

When we compare the topic modeling experts to the full training set, the 2-topic model reaches an improvement over the full training set. The accuracy of the 5-topic setting almost reaches that of the full training set. Thus, even with a fifth of the training set compared to the full set, the per-

Top.	Train. size	Test size	Accuracy
1	53 436	2 504	96.74
2	113 033	5 902	97.51
3	79 133	3 974	97.48
4	89 761	7 814	95.29
5	84 467	3 327	92.41
6	151 562	6 363	98.02
7	141 415	4 612	95.67
8	68 518	2 071	96.77
9	145 224	4 425	96.70
10	25 444	604	94.21

Table 5: Results for the individual topic model experts for the WSJ+GENIA data.

formance of topic models is almost at par with the results on the full training set.

The results lead us to the conclusion that a higher number of topics results in better experts, as shown by the gains over the random baseline. However, the gain of a high number of experts is offset by the reduction of the training set.

Next, we investigate the results of the 10-topic setting more closely: Table 5 shows the results of this setting per topic. These results show that the individual topics vary considerably in size, from around 25 000 words (topic 10) to 150 000 words (topic 6). However, contrary to expectation, there is no direct correlation between training set size and accuracy: Topic 10 has the lowest number of sentences, but its expert performs better than the topic 5 expert, which had access to more than 3 times the amount of training sentences. There is also no clear correlation between accuracy and WSJ or GENIA topics. While the WSJ topics 4, 5, and 7 are at the lower end of the accuracy range, topic 9 has a higher accuracy than GENIA topics 1 and 10 and a similar performance to topic 8.

5.2 Using Soft Clustering Information

Now we investigate soft clustering information by adding 10 or 100 copies of the sentence to experts based on its topic distribution, in comparison to a hard clustering setting. Table 6 shows the results of these experiments. For the 2-topic experiments, the results indicate that the POS tagger does not benefit from utilizing the topic distribution as there is a slight drop in the accuracy. The reason is that for 2 topics, the separation between WSJ and GENIA into separate topics is very clear. I.e., a sentence generally has a very high probability for its corresponding topic and thus should only be added to that topic. Consequently, the advantage of using experts is largely outweighed by misclassified sentences that are added to the wrong expert. However, soft clustering with 5 or 10 topics shows improvements over the full training baseline since the topic distribution is more fine grained. Here, a sentence is more likely to be included in more than one topic. Using a sentence 10 times rather than 100 times seems to be a better fit. The 10-topics, 10 copies setting reaches the same accuracy as the hard clustering 2-topic setting, thus showing that expert knowledge is capable of combating data sparseness.

	Accuracy		
Copies	2 topics	5 topics	10 topics
1	**96.84**	96.54	96.34
10	96.73	**96.67**	**96.84**
100	96.04	96.54	96.73

Table 6: Results for soft clustering on 2, 5, and 10 topics experiments

Setting	Accuracy
Full training set	96.23
Random split	95.16
Topic model	95.53
Soft Clustering	**96.32**

Table 7: Comparing topic model experts to the baselines on WSJ data (10 topics).

5.3 WSJ Experiments: Creating Micro-Topics

Here, we investigate whether we can also successfully use topic modeling to create POS tagging experts in cases where there is only one genre. That is, is topic modeling only sensitive towards genre differences or can it also detect smaller types of variation, and can those variations be translated into specialized POS tagging experts? We use the WSJ corpus for this set of experiments, and we compare an experiment with 10 topics to the two baselines. The results of these experiments are shown in Table 7. We see a positive effect of using experts based on the hard clustering topic models over the random split: Accuracy increases from 95.16% to 95.53%. Similar to the GENIA+WSJ 10-topic experiment, we also do not reach the baseline using all training data. Per topic, there are similar trends to the ones for the WSJ+GENIA set-

Data Set	Setting	Accuracy
standard	random split	95.16
standard	topic model	95.53
5-fold	topic model	95.70

Table 8: Comparing the standard data split to a random data split for WSJ data (10 topics).

ting, a large variation of topic sizes and no direct correlation of training set size and accuracy. However, the soft clustering results show that there is a 0.8 percent improvement over the topic models and a small improvement over the full training set. This reinforces the hypothesis that soft clustering can indeed handle the data sparseness issue even when the genres are not as clearly distinguishable as WSJ vs. GENIA.

The differences between topic models and a random split are less pronounced than in the case of the combined WSJ+GENIA corpus. One explanation for this may be that the topics are less different from each other than in the WSJ+GENIA setting so that the POS tagging expert are not very different from each other. Another possible explanation is that this is a consequence of the way we split the WSJ corpus into training and test sets: As described in section 4, we use the standard split with section 02-21 as training set and section 22 as our current test set. This means that the test set may contain different topics from the training set, which may generate problems in topic modeling. To test this hypothesis, we repeat the experiment with 10 topics, but this time, we perform a five-fold cross-validation for POS tagging on sections 02-22. I.e., we vary the test set across all sections, to create a more homogeneous data split. The results of this experiment, in comparison to previous results, are shown in Table 8. We reach slightly higher results in the 5-fold cross-validation; accuracy increases from 95.53% to 95.70%. This means that the training and test set in the standard setting are not optimally homogeneous. However, since the difference in accuracy is rather small, the differences between training and test set do not seem to impact the performance of our system.

5.4 What do the Experts Learn?

In this section, we investigate the differences between the models learned based on a random split as opposed to the models learned based on the topic models. We concentrate on the 2 topic models based on the WSJ+GENIA data set from section 5.1.

Random split				Topic model			
split 1		split 2		topic 1		topic 2	
NN	335	NN	300	NN	387	CD	227
JJ	219	JJ	187	JJ	217	NNP	226
CD	151	CD	162	CD	70	NN	132
NNP	132	NNP	162	NNS	51	JJ	104
NNS	67	NNS	69	NNP	28	NNS	57
VBN	31	VBG	30	FW	13	VBN	32

Table 10: The 6 most frequent POS tags assigned to unknown words (2 topics).

First, we take a closer look at the distribution of unknown words, and the POS taggers' accuracy on known and unknown words. Unknown words are defined as those words from the test set that do not occur in the training set. This means that the POS tagger needs to guess the word's possible tags without having access to its ambiguity class. The results for this investigation are listed in Table 9. These results show that the percentage of unknown words is higher by 0.76 percent points in the random split setting. This means that the two topic models acquire more specialized lexicons that allow the taggers to cover more words. A look at the accuracies shows that, as expected, the accuracy for known words is higher in the topic model setting. However, the results also show that the accuracy on unknown words is significantly higher in this setting, 85.22% for the topic model experts vs. 83.11% for the random splits. This means that the POS tagging models learned from the topic model data split has acquired better models of unknown words based on the word distribution from the training corpora.

We then investigate which POS labels are assigned to unknown words in the two settings. The 6 most frequent POS tags per setting and topic are shown in table 10. A comparison shows that for the random split, both subsets have a very similar distribution: Unknown words are assigned one of the following labels: noun (NN), adjective (JJ), cardinal number (CD), proper name (NNP), plural noun (NNS), past participle (VBN) or present participle (VBG). The distributions for the topic models show a visibly different picture: In the second topic (which is the WSJ topic, see table 2), cardinal numbers are the most frequent class for unknown words, followed closely by names. These two labels are three times and ten times more frequent than in topic 1. In contrast, topic 1 (GENIA)

	Random split			Topic model		
Topic	% Unknown	Known Acc.	Unknown Acc.	% Unknown	Known Acc.	Unknown Acc.
1	4.86	97.25	83.38	3.85	98.35	85.12
2	4.79	97.06	82.84	4.29	96.29	85.31
avg.	4.83	97.16	83.11	4.07	97.33	85.22

Table 9: Unknown word rates and accuracies for known and unknown words in the WSJ+GENIA experiment using 2 topics.

Random split			Topic model		
Gold	TnT	No.	Gold	TnT	No.
NN	JJ	141	NN	JJ	122
JJ	NN	111	JJ	NN	104
NNP	NN	93	VBD	VBN	82
VBD	VBN	88	NNP	NNPS	70
NN	NNP	66	RB	IN	64
IN	RB	65	IN	RB	61
RB	IN	62	NN	NNP	53
NNP	NNPS	53	VBG	NN	50

Table 11: The 8 most frequent confusion sets (2 topics).

is closer to the distribution of the models based on random sampling, but it has a higher number of foreign words (FW), which is an indication that some biomedical terms are not recognized as such and are then marked as foreign words. Examples of such cases are the words "aeruginosa" and "Leishmania". Overall, these results corroborate our hypothesis that the topic models learn individual characteristics of unknown words.

Finally, we consider the types of errors that the POS taggers make by looking at confusion sets. The 8 most frequent confusion sets under both conditions are shown in table 11. A closer look at the confusion sets of the two experiments shows that the categories in the random split setting are consistent with standard errors that POS taggers make: These POS taggers mostly confuse nouns (NN) with adjectives (JJ) and with names (NNP), past tense verbs (VBD) with participles (VBN), prepositions (IN) with adverbs (RB). One notable difference in the topic modeling setting is that the number of confusions between nouns (NN) and names (NNP) (in both directions) is almost reduced by half in comparison to the random split setting: 88 vs. 159 cases (note that the condition NN NNP is not among the 8 most frequent cases for the topic model as shown in table 11, it is the 12th most frequent confusion set). Names are generally difficult because they constitute an open set, and thus not all of them will be found in the training set. For example, names that were misclassified as nouns in the random split data set included "BART", "Jefferies", and "Tulsa". Thus, a reduction of these errors means that the topic model experts are learning characteristics that allow them to handle domain specific names better, even though the respective learned model files of the topic model setting contain considerably fewer lexical entries.

6 Conclusion and Future Work

In our research, we have investigated whether we can use topic modeling in order to create specialized subsets of POS annotated data, which can then be used to train POS tagging experts for the topic. Our results show that the POS tagging experts achieve higher accuracies both for a manually created mixed data set with financial news and medical texts and for a more homogeneous data set consisting only of financial news. The latter shows that our system is capable of adapting to nuances in the micro-genres within the Wall Street Journal texts. Our analysis also shows that a significant improvement is achieved, particularly, for proper names. The topic model experts are almost three times more likely to tag a name correctly than the random split models.

We have created a flexible and fully automatic methodology of POS tagging experts for different genres. These experts can be extracted from a heterogeneous text source, without the need of having to separate the genres manually. Additionally, we obtain individual experts, which can be used separately. Further applications for this kind of technology can be found in adapting POS taggers to characteristics of different speech or cognitive impediments but also to the characteristics of non-native speakers.

Our current experiments have used 2, 5, and 10

topic models. In theory, the number of topics can be set to a higher number, thus creating more subtle topics. However, as we have also shown, the higher the number of topics, the more severe data sparseness becomes. This can be mitigated by using training sentences for more than one topic, based on the distribution provided by the topic modeler. We plan on extending our work to syntactic parsing, for which the differences between genres will be more noticeable.

Acknowledgments

References

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.

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.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic.

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.

Mohammad Khan, Markus Dickinson, and Sandra Kübler. 2013. Towards domain adaptation for parsing web data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Hissar, Bulgaria.

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.

John Miller, Manabu Torii, and K. Vijay-Shanker. 2007. Adaptation of POS tagging for multiple biomedical domains. In *Proceedings of the Workshop on Biological, Translational, and Clinical Language Processing*, pages 179–180, Prague, Czech Republic.

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.

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.

Anders Søgaard. 2013. Zipfian corruptions for robust POS tagging. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 668–672, Atlanta, GA.

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.