Effectiveness of Domain Adaptation Approaches for Social Media PoS Tagging

Tobias Horsmann, Torsten Zesch
Language Technology Lab
Department of Computer Science and Applied Cognitive Science
University of Duisburg-Essen, Germany
{tobias.horsmann,torsten.zesch}@uni-due.de

Abstract

English. We compare a comprehensive list of domain adaptation approaches for PoS tagging of social media data. We find that the most effective approach is based on clustering of unlabeled data. We also show that combining different approaches does not further improve performance. Thus, PoS tagging of social media data remains a challenging problem.

Italiano. *Confrontiamo diversi approcci di adattamento al dominio per il PoS tagging di dati social media. Osserviamo che l'approccio più efficace si basa sul clustering di dati non annotati. Inoltre, mostriamo che la combinazione di diversi approcci non migliora ulteriormente le prestazioni. Di conseguenza, il PoS tagging di dati social media rimane un problema difficile.*

1 Introduction

Part-of-Speech (PoS) tagging of social media data is still challenging. Instead of tagging accuracies in the high nineties on newswire data, on social media we observe significantly lower numbers. This performance drop is mainly caused by the high number of out-of-vocabulary words in social media, as authors neglect orthographic rules (Eisenstein, 2013). However, special syntax in social media also plays a role, as e.g. pronouns at the beginning of sentence are often omitted like in "went to the gym" where the pronoun 'I' is implicated (Ritter et al., 2011). To make matters worse, existing corpora with PoS annotated social media data are rather small, which has led to a wide range of domain adaptation approaches being explored in the literature.

There are two main paradigms: First, adding more labeled training data by adding foreign or machine-generated data (Daumé III, 2007; Ritter et al., 2011). Second, incorporating external knowledge or guiding the machine learning algorithm to extract more knowledge from the existing data (Ritter et al., 2011; Owoputi et al., 2013). The first strategy affects from *which* data is learned, the second one *what* is learned.

Using more training data Usually there is only little PoS annotated data from the social media domain, so just using *re-training* on domain-specific data does not suffice for good performance. *Mixed re-training* adds additional annotated text from foreign domains to the training data. In case there is much more foreign data than social media data, *Oversampling* (Daumé III, 2007) can be used to adjust for the difference in size. Finally, *Voting* can be used to provide more social media training data by relying on multiple already existing taggers.

Using more knowledge Instead of adding more training data, we can also make better use of the existing data in order to lower the out-of-vocabulary rate. *PoS dictionaries* provide for instance information about the most frequent tag of a word. Another approach is *clustering* which group words according to their distributional similarity (Ritter et al., 2011).

In this paper, we evaluate the potential of each approach for solving the task.[1]

2 Baseline Tagger

We re-implement a state-of-the-art tagger in order to control all aspects of the process. It is based on CRFsuite[2] in version 0.12 as part of the text-classification framework DKProTC (Daxenberger

[1] Our experiments are available at http://tinyurl.com/neptn9e

[2] https://github.com/chokkan/crfsuite

et al., 2014). As training algorithm we use *Adaptive Regularization Of Weight* (AROW).

Our feature set follows previous work (Gimpel et al., 2011; Hovy et al., 2014). We use the word itself and the preceding and following word. We use boolean features for words containing capital letters, special characters, numbers, hyphens and periods, and for detecting words entirely composed of special characters or capital letters. We furthermore use the 1000 most frequent character bi- to four grams.

Our tagger achieves an accuracy of 96.4% on the usual WSJ train/test split which is close to the 96.5% by TNT tagger (Brants, 2000) and only slightly worse than the 97.2% of the Stanford tagger (Toutanova et al., 2003). When we evaluate our newswire tagger as is on the 15,000 token Twitter corpus by Ritter et al. (2011), accuracy drops to 76.1% which confirms their findings.

Having established these baselines, we now test the different domain adaption strategies. In order to reflect the domain difference, we will call the WSJ corpus NEWS and the Twitter corpus SOCIAL in the remainder of the paper.

3 Domain Adaptation Approaches

In this section, we explore existing domain adaptation approaches that can be divided into (i) using more training data or (ii) more knowledge.

3.1 More Training Data

We test three strategies (*re-training*, *oversampling*, and *voting*) using 10-fold cross validation on SOCIAL.

Re-training Simply re-training on SOCIAL improves accuracy from 76.1% to 81.9%, but still is far behind the 97% on newswire text. To estimate the potential of re-training, we show in Figure 1 the learning curve using increasing subsets of SOCIAL. The plot shape indicates that annotating additional in-domain data would be beneficial, but annotating more data is often so unattractive that domain adaption strategies are preferred anyway.

Another quite simple approach is training on NEWS and SOCIAL together, which we call *mixed re-training*. We evaluate this setting by cross validating only over SOCIAL and always adding the full NEWS corpus to the train set. This yields an accuracy of 82.7% compared to 81.9% on SOCIAL alone by adding two orders of magnitude more data (10^6 instead of 10^4 tokens).

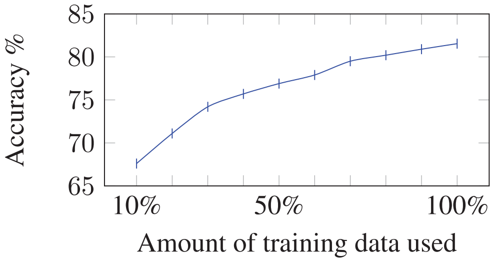

Figure 1: Re-training learning curve

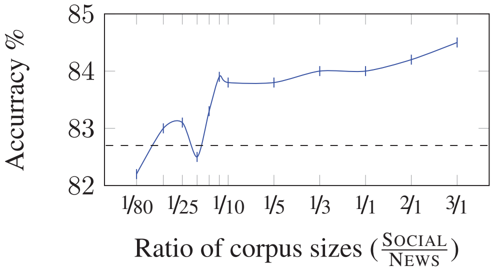

Figure 2: Oversampling results

Oversampling To overcome the size problem in *mixed-retraining*, oversampling the smaller corpus can be used (Daumé III, 2007; Neunerdt et al., 2014). The idea is to boost the importance of the small SOCIAL data by adding it multiple times (or adjusting the feature weights). We show the effect of varying oversampling rates i.e. the ratio of SOCIAL (size varied) to NEWS (size kept constant) in Figure 2. At an oversampling rate of 1:4, we achieve an accuracy of 84.5% which exceeds the *mixed-retraining* baseline of 82.7%.

Voting In this approach, a sample of unlabeled social media data is tagged using multiple existing PoS taggers. If they all assign the same label sequence (i.e. they all *voted* the same) the sentence is added to the training set as it is less likely that all taggers make the same mistakes. We use the PTB tagset taggers ClearNLP[3], OpenNLP[4] and Standford, setting the PoS tags for *Hashtags, Urls, At-mention* and *Retweet* manually in post-processing (Ritter et al., 2011). The results in Figure 3 show that it doesn't really matter how much voted data is added, we roughly see the same increase, with no real trend. We reach an accuracy of 83.5% at $6 \cdot 10^5$ additional tokens *voted* data. We show as comparison the curve if NEWS is added instead and find no disadvantages to the voting approach.

[3] http://www.clearnlp.com

[4] https://opennlp.apache.org

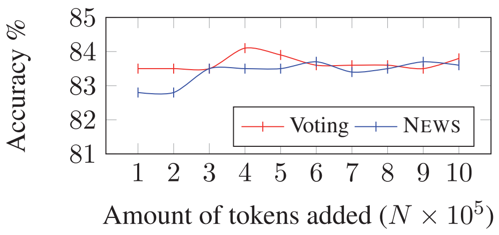

Figure 3: Voting vs. mixed-retraining

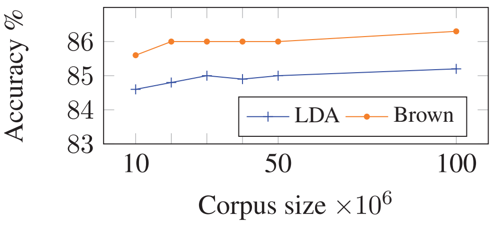

Figure 4: Clustering results

	Trained on	Acc. %
	Baseline	76.1
1	Re-training	81.9
2	Mixed re-training	82.7
3	Mixed re-training (Oversampling)	84.5
4	Re-training + Voting	83.5
5	PoS dictionary	83.8
6	Clustering	86.3
	Combo (4,5,6)	86.9

Table 1: Tagging accuracies per approaches

3.2 More Knowledge

In this section, we discuss the effect of adding more knowledge in the form of PoS dictionaries or word clusters.

PoS Dictionaries We use a dictionary that stores the PoS distribution for each word form that occurs in a corpus. The underlying corpus can either be manually annotated or machine-tagged (Gimpel et al., 2011; Rehbein, 2013).

We use two dictionaries in our experiments: ManualDict, created from the manually annotated Brown corpus (Nelson Francis and Kuçera, 1964), and MachineDict, created from 100 million tokens of the machine-tagged English WaCky corpus (Baroni et al., 2009). Surprisingly, both dictionaries equally improve the performance to 83.8%, the much bigger MachineDict providing no advantage. MachineDict covers about 60.3% of the tokens in SOCIAL while ManualDict only covers 54.0%. It seems that the higher quality of manual PoS annotations in ManualDict counters the higher coverage of MachineDict. The rather low coverage of both dictionaries is caused by cardinal numbers and social media phenomena such as Hashtags.

Clustering We experiment with two versions of clustering: LDA[5] (Blei et al., 2003; Chrupala, 2011) and hierarchical Brown clustering[6] (Brown et al., 1992). Following Owoputi et al. (2013) and Rehbein (2013), we create 1,000 Brown clusters with a minimal word frequency of 40, and 50 LDA clusters with a minimal word frequency of ten. We encode Brown cluster information following Owoputi et al. (2013).

Figure 4 shows that Brown clusters work much better than LDA, where the 100 million token Brown clusters reach the highest accuracy of 86.3%. Using the 800 million token Brown clusters provided by Owoputi et al. (2013) does not further improve results yielding an accuracy of 86.2%. We thus find that clustering is highly effective, but that very large corpus sizes might not translate into further increases.

4 Combining Approaches

Combining approaches might further increase accuracy over the individual approaches summarized in Table 1. As the different strategies for adding more data are hard to combine, we select strategy #4 that provides good accuracy at much lower costs compared to oversampling.

PoS dictionaries and clustering seem to be effective and can easily be used together. Thus, our final combined model consists of re-training with the manually annotated SOCIAL data, 10,000 additional machine-tagged voting tokens, the MachineDict PoS dictionary, and the 100 million token Brown cluster. We achieve an accuracy of 86.9% accuracy, which is only a small improvement over clustering alone.

Comparison with State of the Art While our goal is not to exactly replicate previous work, it is quite informative to make the comparison. Ritter et al. (2011) reported 88.3% accuracy on the same dataset, but additionally added the NPS chat corpus for training, which is inline with our interpretation of Figure 1 that adding more hand-annotated

[5] https://bitbucket.org/gchrupala/lda-wordclass/
[6] https://github.com/percyliang/brown-cluster

Adjectives		Interjections	
Token	Gold / Combo	Token	Gold / Combo
Happy	JJ	**Thanks**	UH / NNS
Berlated	JJ / NNP	and	CC
Birthday	NN	I	PRP
!	.	will	MD
When	WRB	in	IN
I	PRP	the	DT
Get	VBP	street	BB
Old	JJ / NNP	**loll**	UH / NN
,	,	.	.

Table 2: Adjective and interjection confusions

Word	Combo	
class	fine	coarse
ADJ	76.0	76.9
ADV	85.3	85.6
NN	80.9	91.6
V	81.9	91.4
All PoS	86.9	91.5

Table 3: Fine vs. coarse-grained accuracy

data is probably a good idea. Owoputi et al. (2013) reported 90%, but additionally use several name lists to detect proper nouns. We are going to explore the impact of this kind of tag specific optimization in section 5.

Error Examples Table 2 shows representative errors for the frequently occurring classes adjectives and interjections. The first adjective error shows a confusion of an out-of-vocabulary item with capital letter. The second error is also caused by the first letter in uppercase. Interjections are notoriously hard to tag, as they are mainly pragmatic markers.

5 Practical Issues

We now turn to some practical issues that influence the interpretation of the obtained results.

5.1 Coarse-grained Performance

Tagging social media is hard also because the lack of context and informal writing sometimes make fine-grained decisions about a certain PoS tag almost impossible. For example, in *He dance on the street* the word *dance* is a verb, but its intended tense is not easily determinable. We thus test whether the accuracy improvement mainly happens within a coarse tag class or between classes (e.g. only confusions between regular (NN) and proper nouns (NNP) are corrected).

Table 3 shows the re-calculated accuracy of the *Combo* approach, counting as correct not only exact matches, but also if the assigned PoS tag matches the coarse-grained PoS class. For nouns and verbs, we see that accuracy improves from the low 80's to the low 90's which means that many mistakes are intra-class here (e.g. NN vs NNP). Thus, tagging accuracy for coarse-grained word classes is already much higher than the numbers might show and tagging of adjectives and adverbs is the biggest remaining problem.

5.2 Influence of the System Architecture

While experimenting with CRFsuite, we noticed that the same set of train/test data yields different results on different system architectures (Windows 7, OS-X 10.10, and Ubuntu).[7]

Just by chance, changing platform might give you a performance increase that is in the same range as the best domain adaptation strategy discussed in this paper. This shows that failure to reproduce previous results can have unexpected causes far beyond the actual research question to be tested.

6 Conclusion

In this paper, we analyzed domain adaptation approaches for improving PoS tagging on social media text. We confirm that adding more manually annotated in-domain data is highly effective, but annotation costs might often prevent application of this strategy. Adding more out-domain training data or machine-tagged data is less effective than adding more external knowledge in our experiments. We find that clustering is the most effective individual approach. However, clustering based on very large corpora did not further increase accuracy. As combination of strategies did only yield minor improvements, clustering seems to dominate the other strategies.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan.

[7] Cause is a shuffling operation of the train set that is initialized differently among operating system architectures.

2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Thorsten Brants. 2000. Tnt: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 224–231, Stroudsburg, PA, USA.

Peter F Brown, Peter V DeSouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18:467–479.

Gzegorz Chrupala. 2011. Efficient induction of probabilistic word classes with LDA. In *Proceedings of the Fifth International Joint Conference on Natural Language Processing : IJCNLP 2011*, page 363, Chiang Mai, Thailand.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Conference of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. DKPro TC: A java-based framework for supervised learning experiments on textual data. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 61–66, Baltimore, Maryland.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia.

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 *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA.

Dirk Hovy, Barbara Plank, and Anders Søgaard. 2014. When pos data sets don't add up: Combatting sample bias. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

W. Nelson Francis and Henry Kuçera. 1964. Manual of Information to Accompany a Standard Corpus of Present-day Edited American English, for use with Digital Computers.

Melanie Neunerdt, Michael Reyer, and Rudolf Mathar. 2014. Efficient Training Data Enrichment and Unknown Token Handling for POS Tagging of Non-standardized Texts. In *12th Conference on Natural Language Processing (KONVENS)*, pages 186–192, Hildesheim, Germany.

Olutobi Owoputi, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ines Rehbein. 2013. Fine-Grained POS Tagging of German Tweets. In Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 162–175.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA.