When POS data sets don't add up: Combatting sample bias

Dirk Hovy, Barbara Plank, Anders Søgaard

Center for Language Technology, University of Copenhagen
{dirk,bplank}@cst.dk, soegaard@hum.ku.dk

Abstract

Several works in Natural Language Processing have recently looked into part-of-speech (POS) annotation of Twitter data and typically used their own data sets. Since conventions on Twitter change rapidly, models often show *sample bias*. Training on a combination of the existing data sets should help overcome this bias and produce more robust models than any trained on the individual corpora. Unfortunately, combining the existing corpora proves difficult: many of the corpora use proprietary tag sets that have little or no overlap. Even when mapped to a common tag set, the different corpora systematically differ in their treatment of various tags and tokens. This includes both preprocessing decisions, as well as default labels for frequent tokens, thus exhibiting *data bias* and *label bias*, respectively. Only if we address these biases can we combine the existing data sets to also overcome sample bias. We present a systematic study of several Twitter POS data sets, the problems of label and data bias, discuss their effects on model performance, and show how to overcome them to learn models that perform well on various test sets, achieving relative error reduction of up to 21%.

Keywords: sample bias, Twitter, preprocessing

1. Introduction

Several works in NLP have recently investigated part-of-speech (POS) tagging of Twitter data (Foster et al., 2011; Ritter et al., 2011; Gimpel et al., 2011; Derczynski et al., 2013; Owoputi et al., 2013). Due to both the lack of standard data sets and the ever-changing nature of tweets, all of these papers use their own annotated data. However, recent studies suggest (Eisenstein, 2013) that Twitter samples exhibit severe **sample bias** towards the particular time they were collected. Thus, while models trained on the existing data sets achieve good results on the respective test splits, they are likely to suffer severe performance drops when applied to new data (or to each other) due to overfitting the sample. A natural measure to combat this overfitting is to combine the existing resources and train on the joint corpus.

However, combining multiple data sets is complicated by several problems related to sample bias:

1. The existing corpora use different tag sets that are often mutually exclusive. This can be addressed by **mapping** all corpora to a common tag set (Zeman, 2008; Zeman, 2010). We use the universal tag set by Petrov et al. (2012) for this purpose.

2. However, even after mapping to the universal tags, the corpora differ systematically in terms of **token preprocessing**, i.e., tokenization of clitics (*don't* vs *do+n't* vs *don+'t*), replacement of numbers, and anonymization of names and URLs.

3. Additionally, the corpora differ in terms of **tag normalization** for common word tokens. Specifically, in the case of Twitter, this concerns the tags used for URLs, user names, hashtags, and the token `RT`.

If these problems are left unaddressed, models learned on any one of the data sets, while performing well on the respective test sets, are likely to suffer performance losses when applied to any of the other data sets, or to new data.

We first analyze the systematic differences between several data sets, and then show how we can improve performance by overcoming label and data bias. Even then, the different data sets exhibit strong sample bias. We show how this can be combatted by combining the now-aligned data sets.

2. Data

We use the POS data sets of Ritter et al. (2011) (15k tokens) and Gimpel et al. (2011) (26k tokens) to exemplify the effects described above. For evaluation purposes, we use combinations of the test+dev splits provided by those two works, as well as the test data from Foster et al. (2011). Ritter et al. (2011) used cross-evaluation, so there are no official train/test/dev splits, therefore we use the splits provided by Derczynski et al. (2013).

Since mapping and pre-processing are noisy processes that can potentially introduce errors into both the training and test sets, we also annotated a separate data set ourselves, using the universal tags, and preprocessing tokens and normalizing tags as described in Table 1. We choose to annotate abbreviations as X, *it's* and the like as VERB, emoticons as X and all punctuation as such. We randomly selected 200 tweets collected over the span of one day, and had three annotators tag this set. We split the data in such a way that each annotator had 100 tweets: two annotators had disjoint sets, the third overlapped with the two others. After the first round of annotations, we achieved a raw agreement of 0.9, a Cohen's κ of 0.87, and a Krippendorff's α of 0.87. These numbers are above what is usually considered good agreement. We did one pass over the data to adjudicate the cases where annotators disagreed, or where they had flagged their choice as debatable. The final set contained 3,064 tokens and is made publicly available at `http://lowlands.ku.dk/results/`.

Table 1 lists the most important token preprocessing and tag normalization differences between the data sets. Table 2 lists the out-of-vocabulary rates for the data sets (as compared to newswire data).

	FOSTER	GIMPEL	RITTER	Our data set
LOL, etc.	X	PRT	X	X
it's	PRON+VERB†	PRT	PRON+VERB	VERB
don't	VERB+ADV†	VERB	VERB+ADV	VERB
i'm	PRON+VERB†	PRT	PRON+VERB	VERB
emoticons	n.a.	X/NOUN/VERB	X	X
..., !!!, etc.	.+.+.†	.	.	.
~, :	.	X	.	.
RT	NOUN	X	X	n.a.
URLs	NOUN*	X	X	NOUN*
username	NOUN*	**X**/NOUN/ADP	X	NOUN*
numbers	**NUM**/NOUN/ADJ	**NUM**/NOUN/ADJ/ADV/PRT/VERB/X	**NUM**/X/NOUN/ADJ/ADV	NUM*
hashtags	NOUN	**NOUN**/NUM/PRON/PRT/VERB/ADJ	**X**/NOUN	NOUN

Table 1: Token reprocessing (top) and tag normalization (bottom) differences between data sets, and preprocessing decision taken in our data set. Bold faced tags are predominant tags. ∗: Word forms replaced by dummy symbols such as '<url>','<num>', etc. †: Strings split in tokenization.

Train/Eval	Gimpel	Ritter	Foster	OurOwn
Gimpel	69.2	58.3	62.9	64.3
Ritter	76.1	59.2	65.5	69.2
Combined	63.1	48.5	54.2	58.5

Table 2: OOV rates across non-normalized data sets.

3. Experiments

In our experiments, we use a CRF model similar to the one in Owoputi et al. (2013) to learn predictive models for POS tagging. We use simple orthographic features (indicators for prefixes, suffixes, uppercase, and various special characters), as well as the word clusters made available from Owoputi et al. (2013).

We experiment with two settings: first, we show the impact on accuracy when training a model on the two Twitter data sets (RITTER and GIMPEL) using mapping only, or combined with token preprocessing, and tag normalization. We evaluate each model on the two proprietary test sets, as well as on the test data from Foster et al. (2011).

We then evaluate how well the aligned data sets can be combined to produce better models. For this purpose, we evaluate on our mapped, preprocessed, and normalized in-house held-out data. We start out with the fully processed and normalized training data of either corpus, and incrementally add 5% of a combination of all other data sets (training, dev, and test), at different stages of processing. We plot model accuracy on our test set against the size of the available training data.

4. Results

4.1. Processing steps

Table 3 shows the accuracy on different test sets for models trained on either one of the two training sets, and on a combination of both. We observe that both training sets do well on their own test sets (cells marked with a grey background), but suffer severe drops when evaluated on any of the other test sets.

We also note that the token preprocessing and tag normalization steps have a larger effect on the independent data sets than on the two we train on. In fact, token preprocessing improves performance slightly when testing on Gimpel, but hurts performance somewhat on Ritter and considerably on Foster. This is presumably due to the fact that these data sets already used some form of preprocessing (cf. Table 1). In fact, we observe a drop on the Foster evaluation set when adding token preprocessing: Foster already normalized URLs to `Urlname` and usernames to `Username` (cf. Table 1), which both get correctly tagged as NOUN, while after adding our token processing (and mapping them to `@USER` and `URL`) they get tagged as X following the conventions in the Gimpel and Ritter training data. In our own data set, we followed Foster et al. (2011) and treated usernames and URLs as nouns.

Combining the two existing training sets results in a model that, while it does not reach the performance of either training set on its own sample, improves considerably on the other test sets. On average, the models trained on the combination always produce the best result, irrespective of the processing steps. Unsurprisingly, the best model is the one trained on the combination of the two data sets, with all processing steps.

4.2. Data combination

The two plots in Figure 1 show the effect performance on our held-out data set when varying the amount of training data by combining all available data sets. We notice that adding token-preprocessed data results in huge performance drops with respect to the initial model, while adding unprocessed data helps. Both the initial training data and our test set are preprocessed and normalized, so when adding completely unprocessed data, we are adding mainly new tag-token combinations. Since many of these tokens are new words, the model can simply incorporate them as special cases. Non-processed data does better, because it has a higher rate of unique token-label combinations.

On the other hand, if we pre-process the tokens, we make them more similar to the vocabulary of the initial data, but at the same time make them more ambiguous. E.g., by mapping hashtags or numbers to one tokens, but allowing them

	TRAIN	GIMPEL-DEV+TEST	RITTER-DEV+TEST	FOSTER-DEV+TEST	AVERAGE
Mapping	GIMPEL	90.46	82.29	83.80	85.52
	RITTER	80.52	90.40	89.48	86.80
	COMBINED	89.19	87.43	87.75	88.12
+Token preprocessing	GIMPEL	90.65	82.18	81.04	84.62
	RITTER	80.58	90.25	86.61	85.81
	COMBINED	89.34	87.38	84.88	87.20
+Tag normalization	GIMPEL	**91.22**	82.95	84.62	86.26
	RITTER	81.19	**90.71**	**90.56**	87.49
	COMBINED	89.90	87.73	88.51	**88.71**

Table 3: Effect of the different normalization steps on accuracy for models trained and tested on the two data sets.

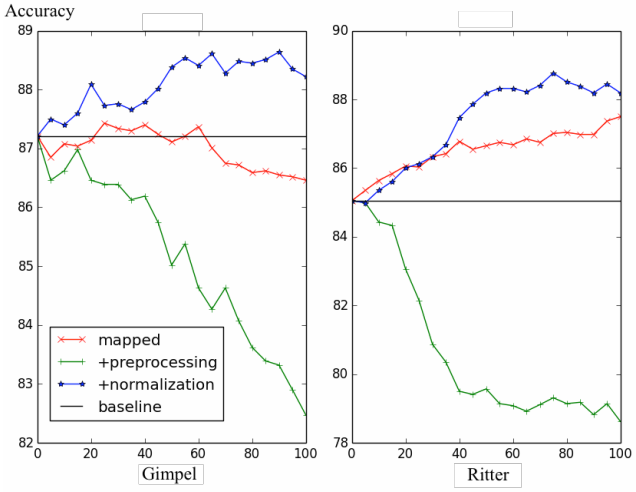

Figure 1: Learning curves for performance (accuracy) on in-house data set when increasing training data

its	on	me	RT	@e_o…	:	Texas	(	cont	)	http://tl.g…
PRT	ADP	PRON	X	**X**	X	NOUN	.	X	.	**X**
its	on	me	RT	**@USER**	:	Texas	(	cont	)	**URL**
PRT	ADP	PRON	X	**NOUN**	X	NOUN	.	X	.	**NOUN**

Figure 2: Annotation differences before (top) and after (bottom) normalization. Example from GIMPEL-TRAIN.

		will	you	come	out	to	the	
GIMPEL	…	VERB	PRON	VERB	**ADP**	**ADP**	DET	…
RITTER		VERB	PRON	VERB	**PRT**	**PRT**	DET	

Figure 3: Annotation differences even after normalization.

to have several labels, rather than one (as in the initial data), we diffuse the emission possibilities. Once we address this problem by normalizing the tags, we see a "normal" learning curve behavior. Thus, when combining data sets from different sources, it is important to analyze whether they contain different label biases: if they are systematic and not random, they have a large impact on performance measures (Reidsma and Carletta, 2008).

The findings unsurprisingly corroborate that combining all available data helps combat sample bias on unseen test sets, but at the same time underscores the importance of first addressing data bias. Accuracy for models starting out with the data from RITTER improve from 85.05 to 88.19 (error reduction of 21%), and models starting with GIMPEL improve from 87.21 to 88.22 (error reduction of 8%).

5. Analysis

Figure 2 shows an example of the differences the processing has on the data. Token preprocessing reduces the OOV rates by smoothing out lexical differences (e.g., replacing user names and URLs with special tokens), while label normalization achieves more consistent taggings across data sets (cf. Table 1).

However, even after all the processing steps outlined above, the data sets exhibit subtle differences. Figure 3 shows an example of the same n-gram found in both training data sets, which is tagged slightly differently. Idiosyncrasies like these are based on annotation conventions (linguistic bias), and are hard to capture with any of the processing steps, but can potentially be overcome by combining enough data to combat sample bias.

6. Conclusion

Various annotated data sets for POS tagging on Twitter exist. Ideally, we would want to combine these data sets to learn more robust models. However, we find that simply mapping to a common tag set is not enough, but can actually introduce additional errors. We find that different data sets cannot be combined without addressing various data bias issues, including tag set mapping, token preprocessing, and label normalization.

Even after controlling for data bias, existing data sets still exhibit sample bias, i.e., models trained on them tend to do well on test sets from the same sample, but suffer when tested on other Twitter data sets. Using the data normalization outlined above allows us to train models on larger, combined data sets, though. Unsurprisingly, we find that these models perform considerably better on a variety of test sets, as well as on average. We show that by normalizing all data sets to a common scheme, we can reduce the relative error by up to 21%.

Our findings show how we can overcome data bias, and in turn combat sample bias by creating larger training sets, resulting in improved performance on out-of-sample data. This suggests that researchers trying to overcome bias need

to pay close attention to the various processing decisions in order to reap the full benefits.

7. Acknowledgments

We would like to thank the anonymous reviewers for valuable comments and feedback. This research is funded by the ERC Starting Grant LOWLANDS No. 313695.

8. References

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: overcoming sparse and noisy data. In *RANLP*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *NAACL*.

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*.

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*.

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

Dennis Reidsma and Jean Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

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

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *LREC*.

Daniel Zeman. 2010. Hard problems of tagset conversion. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*.