Reliable Part-of-Speech Tagging of Low-frequency Phenomena in the Social Media Domain

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

Abstract

We present a series of experiments to fit a part-of-speech (PoS) tagger towards tagging extremely infrequent PoS tags of which we only have a limited amount of training data. The objective is to implement a tagger that tags this phenomenon with a high degree of correctness in order to be able to use it as a corpus query tool on plain text corpora, so that new instances of this phenomenon can be easily found. We focused on avoiding manual annotation as much as possible and experimented with altering the frequency weight of the PoS tag of interest in the small training data set we have. This approach was compared to adding machine tagged training data in which only the phenomenon of interest is manually corrected. We find that adding more training data is unavoidable but machine tagging data and hand correcting the tag of interest suffices. Furthermore, the choice of the tagger plays an important role as some taggers are equipped to deal with rare phenomena more adequately than others. The best trade off between precision and recall of the phenomenon of interest was achieved by a separation of the tagging into two steps An evaluation of this phenomenon-fitted tagger on social media plain-text confirmed that the tagger serves as a useful corpus query tool that retrieves instances of the phenomenon including many unseen ones.

Keywords: Part-of-speech, Social Media, CMC, Rare Phenomena

1. Introduction

This paper reports on experiments on adapting part-of-speech (PoS) taggers for tagging rare phenomena found in genres of computer-mediated communication (CMC). Our work is motivated by a use case in which a linguist wants to study a rarely occurring CMC phenomenon using Twitter data from the social media domain. The central problem here is how to find such rare instances of the phenomenon under observation without spending hours of screening through plain text. A filtering tool would be desirable that facilitates the retrieval process for the linguist. The tool should find many instances of the phenomenon and at the same time achieve reasonably correct results in order to decrease the workload considerably. The project we present here investigates how to adapt a PoS tagger for tagging a certain rarely occurring phenomenon in order to use the PoS tagger as a filtering tool that linguists can use to query a corpus.

The main challenge of adapting a PoS tagger to the language use in the social media domain lies in dealing with the notorious lack of training data and many out-of-vocabulary words. This problem becomes even more severe when the tagger shall be adapted for dealing with a phenomenon that is under-represented in the already small training data sets. We will, thus, investigate methods to improve tagging of under-represented phenomena while laying emphasis on avoiding manual annotation as much as possible. We aim on detecting a German verb-pronoun contraction phenomenon that the linguist wishes to study in detail on the basis of a broad set of instances automatically retrieved from social media data. To deal with the lack of training data, we experiment with (i) adjusting the frequency weight of the under represented phenomena by under- and oversampling and (ii) adding automatically tagged but new data in which only the tag of the phenomenon of interest is manually corrected. In a concluding case study, we optimize a tagger towards finding this contraction phenomenon and evaluate how well the filtering works in a real world setup on plain text Twitter messages.

wiederholen (to repeat) + es (it)	1st person
ich **wiederhols** nochmal, ihr redet hier öffentlich!	
I repeat it [repeat-it] again, you're talking in public!	
kommen (to come) + du (you)	2nd person
wieso? wo **kommste** denn her?	
why? where do you come [come-you] from?	

Table 1: Full verb + pers. pronoun (VVPPER) contraction

2. German Verb-Pronoun Contraction

We are interested in a particular phenomenon in which a verb and a following personal pronoun are contracted into a single form. Table 1 shows examples of this type of contractions taken from the Dortmund Chat Corpus (Beißwenger, 2013). Verb-pronoun contractions belong to the class of phenomena that are not unique for CMC discourse but typical for spontaneous - spoken or 'conceptually oral' - language in colloquial registers. Phenomena of this type are of special interest for linguists who want to use corpora to compare written discourse from the social media domain to the language of edited text and the language found in informal, spoken interactions. If we use a tagger as a filtering tool, we need a high precision to avoid screening through countless false positive instances. At the same time, we want to find new lexical forms unknown from the training set, which requires a high recall (i.e., high generalization).

We have a data set of 23k tokens of German social media discourse that was annotated for a shared task on PoS tagging for German CMC and social me-

dia data (Beißwenger et al., 2016). The data are annotated with an extended version of the Stuttgart-Tübingen tagset (STTS) (Schiller et al., 1999) that has been expanded by tags needed for tagging social media phenomena (Beißwenger et al., 2015). In this tagset, verb-pronoun contractions are labelled by an own tag, VVPPER, which occurs 13 times in total. Results of the shared task showed that this infrequency prevents taggers from learning the phenomenon in a reliable manner (Horsmann and Zesch, 2016b). Since the VVPPER tag is not included in the canonical STTS, as those contractions do not occur in the domain of edited text, existing STTS annotated corpora (e.g., newspaper corpora) cannot be used to obtain additional instances of the phenomenon for training. Although there is a small amount of annotated data that we can build upon, there are still not enough VVPPER instances to make the phenomenon recognizable when training PoS taggers.

3. Dealing with Infrequency

In this experiment, we test different strategies to improve the tagging of VVPPER instances. With a total of 13 instances in our data set, we have to annotate at least some additional data in order to train the tagger but also to arrive at meaningful results during evaluation. At the same time we keep the manual annotation effort at a minimum.

3.1. Data Set

We base our experiment on the data set from the aforementioned shared task. We enrich the data by selecting 230 user posts that contain this phenomenon from the Dortmund Chat Corpus. We automatically tagged these additional data by using the Stanford tagger that assigns PoS tags of the canonical STTS and manually corrected the tag for the verb-pronoun contraction that only exists in the extended STTS. This is the most minimalistic amount of manual annotation one can possibly perform which - as we will see soon - suffices. Of the additional 230 instances, we add one half to the testing set and one-sixth to the training set. The remaining two-sixths are our development set in the following experiments and are held back for the moment. The enhanced training set now contains 45 (38+7) sequences with the phenomenon and the testing set 121 (115+6) sequences. This should be enough instances for learning and evaluating the phenomenon.

3.2. Frequency Weight vs. Lexical Knowledge

An option to circumvent annotation of a larger amount of data is boosting the signal for a certain PoS tag in the already existing data. This can either be done by oversampling (Daumé III, 2007) the few instances one has by adding them N times to the training set, or by downsampling, i.e. removing sequences *without* the PoS tag of interest i.e. VVPPER. Both approaches lead to an increased frequency weight of the phenomenon relative to the other PoS tags in the corpus. We experiment with both strategies in the following setup: *Downsampling:* We remove 25, 50 and 75 percent of the training data instances that do not contain any verb contractions. *Oversampling/new Instances:* We choose oversampling rates that add a number of instances which we can also provide from the held back annotated sequences. This allows a direct comparison between oversampling instances and adding fresh ones. We will, thus, oversample two and three times, and compare this to adding the same amount of instances from the set of new sequences in the held back development set.

We conduct these experiments with the following taggers to learn about the empirical differences between tagger implementations for our objective:

Stanford (Toutanova et al., 2003) a PoS tagger that is frequently used in the community due to its good reputation and high accuracy.

HunPos (Halácsy et al., 2007), a tagger with a good reputation based on Hidden-Markov models and a re-implementation of the TNT tagger (Brants, 2000).

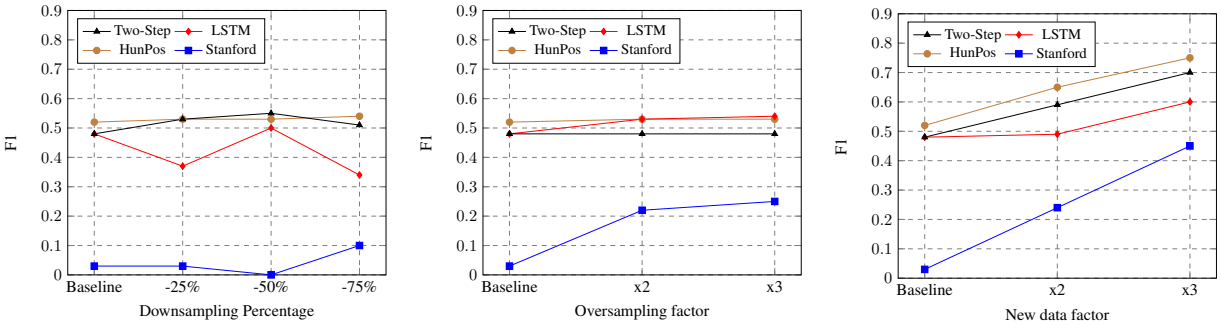
LSTM A deep learning PoS tagger by Plank et al. (2016), that is based on Long-Short-Term-Memory (Hochreiter and Schmidhuber, 1997) neural networks. We use the same parametrization as Plank et al. (2016) and self-trained German word embeddings trained on German Twitter messages with $195 \cdot 10^6$ tokens.

Two-Step Horsmann and Zesch (2016a) proposed a tagger architecture for social media data that first uses a highly generalized *coarse-grained* tagger, and as a second step applies a specialized non-sequential tagger for *fine-grained* tagging. The second tagger is tailored towards recognizing the tag of interest while the first tagging step constraints the application of the second tagger.

We implement this approach by using a CRF tagger (Lafferty et al., 2001) in the first step and an SVM in the second step. For training the coarse-grained sequence model, we map the extended-STTS tags of the training data to the coarse-grained tagset used by the Universal Dependency project and map VVPPER to *verb*. We include a PoS dictionary and Brown (Brown et al., 1992) clusters created over German Twitter messages to compensate for the lack of training data. This coarse-grained tagger reaches a F_1 of 0.93 on the tag *Verb* in the test data, which means that some VVPPER instances will be missed because the coarse model did not predict *verb*.

Results In Figure 1, we show the results of the three strategies on the VVPPER tag. We focus on *out-of-vocabulary* instances which perform considerably poorer than *in-vocabulary* instances (F_1 between 0.96 to 0.99), and thus offer more opportunities for improvements. We see that neither downsampling nor oversampling helps to reach a substantial improvement on the tag. Furthermore, downsampling shows that the anyway low amount of training data becomes a large problem for the LSTM if further reduced. The Stanford tagger stays behind the other taggers with both sampling methods. The only effective method is, without much surprise, providing new data. The LSTM needs considerably more data to improve while the other taggers improve linearly with each new data set.

Discussion Table 2 shows details of the two best taggers HunPoS and Two-Step. Once again, we focus on the out-of-vocabulary instances, while also showing precision (P) and recall (R). The F_1 score shows that both taggers reach



(a) Downsampling (b) Oversampling (c) Annotated Data

Figure 1: Results on *unknown* VVPPER word forms with various methods

	Setup	All F1	Out-Vocabulary P	R	F1
HunPos	Baseline	.78	.80	.38	.52
	Downs. 75%	.78	.63	.48	.54
	Downs. 50%	.79	.74	.41	.53
	Downs. 25%	.79	.81	.40	.53
	Overs. x2	.79	.78	.40	.53
	Overs. x3	.79	.74	.41	.53
	Annotated x2	.83	.80	.56	.65
	Annotated x3	**.88**	.81	.70	.75
Two-Step	Baseline	.77	.95	.32	.48
	Downs. 75%	.78	.85	.37	.51
	Downs. 50%	.80	.96	.38	.55
	Downs. 25%	.79	.92	.38	.53
	Overs. x2	.77	.95	.32	.48
	Overs. x3	.77	.95	.32	.48
	Annotated x2	.81	.93	.43	.59
	Annotated x3	**.85**	.92	.56	.69

Table 2: F_1 on all and on out-of-vocabulary instances

a rather similar overall performance. When looking at precision and recall for adding annotated data, highlighted in grey, we see that Two-Step is considerably more precise than HunPos, which has a better recall. Because oversampling showed barely any effect, we suspect that the added lexical knowledge is mostly accountable for the improvements, which also means that the word context seems to be neglected for making decisions. If the tagger focuses too much on lexical forms, it will find mostly instances known from training which is in particular a problem for finding new instances. Hence, an increased weighting of the local word context should support finding new instances and enable a better generalization.

3.3. Experiment: Forced Generalization

In this experiment, we try to improve generalization of the Two-Step tagger by forcing the tagger to rely more on the local word context and, thus, improve the recall. We chose Two-Step, as we have implemented this tagger ourselves which facilitates adaptation. We alter the feature space of the SVM and exclude all features that contain the lexical form of the *positive* instances. Thus, the SVM is not aware of any lexical forms that can occur with the tag VVPPER, and must now rely more on the word context.

Configuration	All F_1	Out-of-Vocabulary P	R	F_1
Baseline	.81 (+.04)	.93 (+.02)	.41 (+.09)	.57 (+.09)
Annotated x3	.86 (+.01)	.89 (−.03)	.62 (+.06)	.73 (+.04)

Table 3: Results of the contextualised Two-Step

Results In Table 3, we show the changes in performance of the contextualised Two-Step tagger. In parentheses, we show the differences to the not contextualized tagger in Table 2. For both setups we see an improved F_1, but especially the recall increases for out-of-vocabulary instances. The overall F_1 reached by HunPos (.88) in Table 2 is still superior but the trade off between precision and recall of Two-Step better supports the use case in which the tagger functions as a precise filtering tool with decent recall.

4. Field Trial in Social Media

So far, we have only simulated our use case of a linguist who uses a tagger as a filtering tool, while now, we turn to a real setting and apply a tagger to plain text Twitter messages for finding verb-pronoun contractions.

Working on plain text means that the ground truth of how many instances there are in the data is unknown, thus, the recall cannot be computed. Consequently, we focus on evaluating the precision of the tagging, and evaluate how many new instances are found. We choose the Twitter domain for its ease of obtaining data but also for its linguistic diversity that ranges from tweets using informal, interactional language to tweets that are close to the written standard. This domain provides us with a challenging test bed that should allow to determine a conservative, lower-bound performance for our approach. We will use the contextualized Two-Step tagger for its higher precision while providing a reasonable high recall.

Twitter Data We use a random subsample of 50k tweets (about 1.7 million tokens) crawled between 2011 and 2017 from the public Twitter API that we language-filtered for German. All occurrences of user-mentions, hashtags and URLs are replaced by a text constant and the tweets are tokenized by Gimpel et al. (2011)'s ArkTools tokenizer.

Strict
Da **lernste** pragmatisch zu sein .
Ich **sachs** dir noch .
Relaxed
Wer **häts** gedacht .
Ich **wills** nicht ich will aber auch nicht [...]
All
Warum einfach , **wenn's** auch kompliziert geht ? URL
Ich beschränke mich **auf's** nicht im Weg stehen .
Frequent Confusion Cases
Und keiner **weiss** warum .
Ich **weiss** gar nicht , was du beruflich machst .

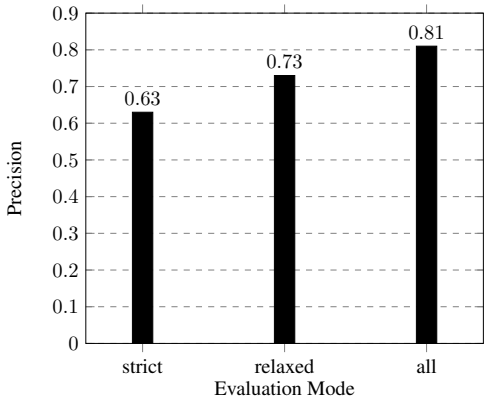
Table 4: Examples of tagged instances

Tagger Setup We train the coarse model and the SVM on the full shared task data set including the additionally annotated data. To provide more lexical knowledge and increase the robustness when facing standard language text, we also add 100k tokens of the German newswire Tiger (Brants et al., 2004) corpus to both tagging steps.

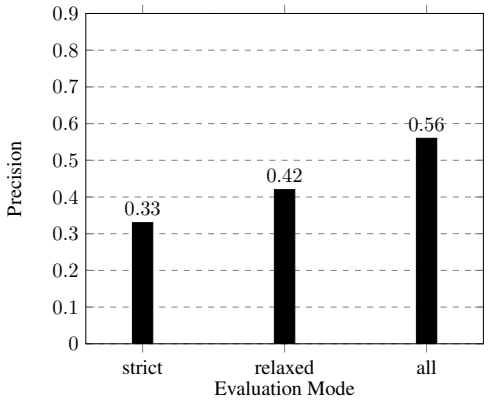
Evaluation setup We evaluate the tagged instances with two annotators. The annotators make four distinctions: *strict*, *relaxed*, *all* and *none*. *Strict* are full verb contractions with personal pronoun, the exact phenomenon we intended to tag. *Relaxed* counts all verb contractions with personal pronoun as correct, this includes also modal and auxiliary verbs. *All* counts all contractions phenomena as correct, this additionally includes, for instance, contractions of conjunctions with personal pronouns. The remaining cases are no contractions and are, thus, false positives.
We will evaluate two setups. The first one selects the first 250 of all found instances, which will be the overall evaluation. The second evaluation focuses on out-of-vocabulary instances in which we remove all tagged instances that are known from the training set until we gather 250 instance and, thus, evaluate how reliably new instances are found.

Results In total, we found 1091 instances in 50k tweets tagged as *VVPPER*. The two annotators reached a perfect agreement on the subset of the first 250 instances that we evaluated manually. Figure 2a shows the precision of the overall evaluation. The *strict* result shows that the majority of found instances are the targeted full verb contractions. Including modal and auxiliary verbs in the *relaxed* mode, even three-quarter are verb contractions. Including also miscellaneous contractions in *all*, almost all instances are contractions.
In Figure 2b, we take a closer look on the performance of detecting new contractions, e.g. out-of-vocabulary instances. We focus our discussion on the *strict* results. The precision is drastically decreased to almost half the value that we reach when including all instances. We also computed the type/token ratio which is at 0.69 almost twice as high as in the overall evaluation in Figure 2a. This confirms that the tagger is able to recognize many new instances of the phenomenon. Furthermore, when ignoring the known instances almost every correct instance is a new lexical form.

Discussion Table 4 depicts examples of each of the three contraction classes (bold face) and additionally presents a



(a) In- and out-of-vocabulary contractions



(b) Out-of-vocabulary contractions

Figure 2: Results of manual evaluation

frequent confusion case, which is erroneously tagged as contraction. Of the VVPPER training data we provided, many instances end on *s* or *'s*, which is a common morphological property of contractions in German. On the one hand, this bias introduces a substantial amount of false positives - for instance the verb *weiß (to know)* occurs frequently in a misspelled form *weiss* in social media. On the other hand, this enables the SVM to also tag similar contraction cases of other word classes in *relaxed* or *all*.

5. Conclusion

We presented experiments that investigated how a PoS tagger can be designed that works as a corpus querying tool to find instances of rare phenomena. We experimented with altering the frequency weight of rare instances but found that adding relatively small amounts of additionally labelled data is unavoidable. By machine tagging data in which only the phenomenon of interest is manually corrected, we keep the effort minimal but yet achieve considerable improvements on detecting the phenomenon. We showed how recall is easily improved when forcing a tagger to focus more on the local word context. In a field study on plain text, we confirmed that our tagger works well as corpus query tool which finds accurately instances of the phenomenon of interest including many new ones. For future work, we plan to improve our method and also study the applicability to other under-represented phenomena.

Beißwenger, M., Bartz, T., Storrer, A., and Westpfahl, S. (2015). Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation / Tagset and guidelibes for the PoS tagging of langauge data from genres of computer-mediated communication. In *EmpiriST guideline document (German and English version)*.

Beißwenger, M., Bartsch, S., Evert, S., and Würzner, K.-M. (2016). EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 44–56, Berlin, Germany.

Beißwenger, M. (2013). Das Dortmunder Chat-Korpus. In *Zeitschrift für germanistische Linguistik 41*, volume 1, pages 161–164.

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic Interpretation of a German Corpus. pages 597–620. Journal of Language and Computation.

Brants, T. (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. Association for Computational Linguistics.

Brown, P. F., DeSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18:467–479.

Daumé III, H. (2007). Frustratingly Easy Domain Adaptation. In *Conference of the Association for Computational Linguistics (ACL)*, pages 256–263, Prague, Czech Republic.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (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. Association for Computational Linguistics.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). HunPos: An open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 209–212, Stroudsburg. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, pages 1735–1780.

Horsmann, T. and Zesch, T. (2016a). Assigning Fine-grained PoS Tags based on High-precision Coarse-grained Tagging. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 328 – 336, Osaka, Japan. Dublin City University and Association for Computational Linguistics.

Horsmann, T. and Zesch, T. (2016b). LTL-UDE @ EmpiriST 2015: Tokenization and PoS Tagging of Social Media Text. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 120–126, Berlin, Germany. Association for Computational Linguistics.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA.

Plank, B., Søgaard, A., and Goldberg, Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proceedings of ACL2016*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Germany.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 NACCL: HLT*, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.