Fine-Grained POS Tagging of German Social Media and Web Texts

Stefan Thater(✉)

Department of Language Science and Technology, Universität des Saarlandes, Saarbrücken, Germany
stth@coli.uni-saarland.de

Abstract. This paper presents work on part-of-speech tagging of German social media and web texts. We take a simple Hidden Markov Model based tagger as a starting point, and extend it with a distributional approach to estimating lexical (emission) probabilities of out-of-vocabulary words, which occur frequently in social media and web texts and are a major reason for the low performance of off-the-shelf taggers on these types of text. We evaluate our approach on the recent *EmpiriST 2015 shared task* dataset and show that our approach improves accuracy on out-of-vocabulary tokens by up to 5.8%; overall, we improve state-of-the-art by 0.4% to 90.9% accuracy.

1 Introduction

Part-of-speech (POS) tagging is a standard component in many linguistic processing pipelines, so its performance is likely to impact the performance of all subsequent steps in the pipeline, such as morphological analysis or syntactic parsing. In the newswire domain, modern POS taggers can reach accuracy scores beyond 97%, close to human performance (Manning 2011). For "non-standard" texts like social media or web texts, however, tagger performance is usually much lower. For the *EmpiriST 2015 shared task* dataset considered in this paper, Beißwenger et al. (2016) report accuracy scores of 80–82% for off-the-shelf taggers.

One important reason for this decline in accuracy is that datasets which are large enough to train a tagger are typically from the newswire domain. For social media and web texts, no large training sets are available. At the same time, these texts differ substantially from newswire text. They contain a lot of "bad" language (Eisenstein 2013) such as misspellings, phrasal abbreviations or intentional orthographical variations as well as phenomena like contractions or interaction words which are not covered by standard tagsets.

On a technical level, the problem can be traced back, at least to some extent, to out-of-vocabulary ("unknown") words which do not occur in the training set. Giesbrecht and Evert (2009) observe that typical web texts contain, compared to newswire texts, more unknown words, and that tagger performance on unknown words is much lower. We make similar observations for the dataset considered in this paper.

© The Author(s) 2018
G. Rehm and T. Declerck (Eds.): GSCL 2017, LNAI 10713, pp. 72–80, 2018.
https://doi.org/10.1007/978-3-319-73706-5_7

One way to address this problem is to add small amounts of manually annotated in-domain data to existing (out-of-domain) training sets when training the tagger. For German, this approach has been explored by Horbach et al. (2014) and Neunerdt et al. (2014). The approach is appealing, as it is conceptually very simple, easy to implement and quite effective. Yet, it can only address part of the problem, as many words remain out-of-vocabulary. Another approach is to exploit distributional similarity information about unknown words. The underlying observation is that distributionally similar words tend to belong to the same lexical class, so POS information of out-of-vocabulary words can be derived from distributionally similar in-vocabulary words (Schütze 1995). Several approaches to POS tagging of various kinds of non-standard texts that exploit this idea have been proposed in the past few years. Gimpel et al. (2011) train a CRF-based tagger using features derived from a reduced co-occurrence matrix; Owoputi et al. (2013), Ritter et al. (2011) and Rehbein (2013) use clustering to derive features to train a discriminative tagger model. Prange et al. (2015) use distributional similarity information to learn a POS lexicon for out-of-vocabulary tokens, and combine it with a Hidden Markov Model (HMM) based tagger.

In this paper, we present an approach that is conceptually similar to the one of Prange et al. (2015) but which uses distributional similarity information to estimate emission probabilities of the HMM, rather than deriving an external POS lexicon. Results on the *EmpiriST 2015 shared task* dataset (Beißwenger et al. 2016) show that our approach improves accuracy on out-of-vocabulary words by up to 5.8%; overall, we improve state-of-the-art by 0.4% to 90.9% accuracy.

2 Model

We briefly present the underlying tagger model in Sect. 2.1 before presenting our distributional approach to estimating lexical probabilities for out-of-vocabulary tokens in Sect. 2.2. Section 2.3 describes the lookup procedure implemented by the tagger.

2.1 Baseline Model

We use a second order Hidden Markov Model to implement our baseline tagger. To tag a given input sequence $w_1 \dots w_n$ of words, we calculate

$$\underset{t_1,\dots,t_n}{\arg\max} \left[\prod_{i=1}^{n} P(t_i \mid t_{i-1}, t_{i-2}) P(w_i \mid t_i) \right] P(t_{n+1} \mid t_n)$$

where $t_1 \dots t_n$ are elements of the tagset and t_{-1}, t_0 and t_{n+1} are additional tags marking the beginning and the end of the sequence, respectively.

Our implementation closely follows Brants (2000). Transition probabilities $P(t_i \mid t_{i-1}, t_{i-2})$ are computed using a linear combination of unigrams, bigrams and trigrams, which are estimated from a tagged training corpus using maximum

likelihood. For the tokens in the training corpus, we estimate emission probabilities $P(w_i \mid t_i)$ using maximum likelihood and for out-of-vocabulary tokens emission probabilities are estimated based on the word's suffix. Our implementation differs slightly from (Brants 2000) in that we use, for purely practical reasons, a maximal suffix length of 5 instead of 10 in the computation of suffix distributions, and that we do not maintain different suffix distributions for uppercase and lowercase words.

2.2 Distributional Smoothing

We use a large, automatically POS-tagged corpus and estimate $P(w \mid t)$ by considering all contexts in which w occurs in the corpus, and estimating the emission probability of w based on the emission probability of all in-vocabulary words w' that occur in the same contexts as w. We set:

$$P(t \mid w) = \sum_{w'} \sum_{C} P(t \mid w')\, P(w' \mid C)\, P(C \mid w) \tag{1}$$

where w' ranges over all in-vocabulary words in the manually annotated training corpus used to train the baseline model and C ranges over all n-grams consisting of the POS tags of the two words on either side of an unknown word w in the automatically tagged corpus. $P(t \mid w')$ is the probability of a tag t of an in-vocabulary word w', $P(w' \mid C)$ is the probability that w' occurs in a given context C and $P(C \mid w)$ is the probability of context C given an out-of-vocabulary word w. The probabilities are estimated on the automatically tagged corpus using maximum likelihood. Following recommendations by Prange et al. (2015), we consider only contexts in which the two surrounding words are in-vocabulary; the idea is that in-vocabulary tokens are tagged with much higher precision and thus give us more reliable context information.

While using (1) to estimate emission probabilities of out-of-vocabulary tokens improves tagger performance beyond the baseline model, (1) is still somewhat noisy. We further improve tagger performance by combining (1) with a second distribution $P(t \mid w)$ which estimates the probability of a tag t of an unknown word w based on the suffix of w. In principle, we could simply use the corresponding distribution of the baseline tagger, but it turns out that the following approach works much better:

$$P(t \mid w) = \sum_{w'} \sum_{s} P(t \mid w')\, P(w' \mid s)\, P(s \mid w) \tag{2}$$

where s ranges over all possible suffixes. The distributions $P(s \mid w)$ and $P(w' \mid s)$ are estimated on the type level, *i.e.*, $P(s \mid w) = 1$ if s is a suffix of w, 0 otherwise, and $P(w' \mid s) = \frac{1}{n}$, where n is the number of types with suffix s.

We combine (1) and (2) using multiplication, re-normalize the result and apply Bayes' theorem to obtain the final emission probabilities $P(w \mid t)$.

2.3 Lookup

Our tagger implements the following lookup strategy: When reading in a token w, we first try to look up w in the lexicon; if that fails, we redo the lookup with w mapped to lower case; if that fails, we consult the distributional lexicon; as a fallback, we use the suffix lexicon of the baseline tagger.

We follow common practice and normalize all numerical expressions (sequences of digits) into a single token type. To improve tagger performance on social media texts, we additionally normalize all tokens beginning with an "@" or "#".

3 Evaluation

We evaluate our approach on the dataset of the *EmpiriST 2015 shared task on automatic linguistic annotation of computer-mediated communication and social media* (Beißwenger et al. 2016) and compare it to the two systems that performed best on the share task as baselines.

3.1 Datasets

EmpiriST. This dataset has been provided by the *EmpiriST 2015 shared task.* It has been compiled from data samples considered representative for two types of corpus data. The *CMC* subset consists of selections of microposts from Twitter, a subset of the *Dortmund Chat Corpus* (Beißwenger 2013), threads from Wikipedia talk pages, WhatsApp interactions and blog comments. The *Web* subset consists of selections of websites and blogs covering various genres and topics like hobbies and travel, Wikipedia articles on topics like biology and botany and Wikinews on topics like IT security and ecology. The dataset is split into two parts, one for training and one for testing. The *CMC* subset consists of 5109 tokens for training and 5234 tokens for testing; the *Web* subset consists of 4944 tokens for training 7568 tokens for testing.

The dataset has been annotated using the "STTS IBK" tagset (Beißwenger et al. 2015), which is based on the STTS tagset (Schiller et al. 1999). STTS is the standard tagset for German. It distinguishes 11 parts of speech which are subdivided into 54 subcategories. STTS IBK adds 16 new tags for phenomena that occur frequently in social media texts, such as interaction words, addressing terms or contractions.

Schreibgebrauch. This dataset has been provided by (Horbach et al. 2015) and has been used as additional in-domain training data by the best-performing system of the *EmpitiST shared task* (Prange et al. 2016). It consists of manual annotations of forum posts of the German online cooking community http://www.chefkoch.de, a subset of the *Dortmund Chat-Korpus* and microposts from Twitter. In total, the annotated dataset consists of 34 173 tokens. Since the

dataset has been annotated with a tagset that differs in some details from STTS IBK, Prange et al. (2016) re-annotated the dataset so that it matches the annotation scheme and guidelines of the shared task. We use the re-annotated version in our experiments.

We also use the complete *Chefkoch* corpus from which the annotated subset was selected to train lexical probabilities of out-of-vocabulary tokens. The corpus contains 470M tokens and covers a relatively large range of everyday topics.

TIGER. The TIGER corpus (Brants et al. 2004) is one of the standard corpora used for German POS tagging. It consists of 888 238 tokens which have been semi-automatically annotated with POS information, using the standard STTS tagset.

3.2 Experimental Setup

We train two different models: The TE model is trained on a combination of the TIGER corpus and the EmpiriST training set. The TES model additionally uses the *Schreibgebrauch* dataset. Since the two in-domain datasets are very small compared to TIGER, we follow Prange et al. (2016) and oversample them by a factor of 5. We automatically annotate the *Chefkoch* corpus using each of the two tagger models to estimate emission probabilities for out-of-vocabulary words as described in Sect. 2.2.

3.3 Results

Figure 1 shows the results of our approach on the EmpiriST evaluation dataset. We consider two different configurations for each of our two models: TE/BL and TES/BL use suffix-based emission probabilities of the baseline tagger for out-of-vocabulary tokens, while TE/DS and TES/DS use distributional smoothing. To set the results into perspective, we compare our models to two state-of-the-art approaches: *UdS* refers to the system of Prange et al. (2016), which performed

Model	CMC	Web	Overall
TE/BL	86.78	92.47	89.63
TES/BL	87.89	92.72	90.31
TE/DS	87.08	93.22	90.15
TES/DS	**88.38**	93.34	**90.86**
UDE	86.07	92.10	89.09
UdS	87.33	**93.55**	90.44

Fig. 1. Accuracy comparison for different configurations of our tagger and the two best performing shared task models on the EmpiriST test set.

best in the EmpiriST shared task. The tagger is based on a Hidden Markov Model trained on EmpiriST, *Schreibgebrauch* and TIGER and uses distributional information obtained from the *Chefkoch* corpus to automatically learn a POS dictionary. *UDE* refers to the system of Horsmann and Zesch (2016). The tagger is based on Conditional Random Fields (CRFs) trained on EmpiriST and TIGER and was the best system in the shared task that does not use any in-domain data in addition to the training data provided by the shared task. In addition to standard features of a CRF-based tagger, the system uses word cluster information from Twitter messages, a POS lexicon and a morphological lexicon.

We compare our TE model to the UDE system and the TES model to the UdS system. Figure 1 shows that already our baseline configurations outperform state of the art (except UdS on *Web*). This is particularly surprising when comparing TES to UdS on *CMC*, since both models are based on trigram HMMs trained on the same datasets. To some extent, the difference can be explained by our use of simple patterns for @- and #-expressions, but we note that even without these patterns our basic tagger still outperforms UdS on *CMC* by 0.2%.

We also see that distributional smoothing is effective across all four configurations. On the *CMC* subset, the performance gain increases quite substantially for the TES model compared to the TE model (+0.49 *vs.* +0.30). This is to be expected, since the emission probabilities are derived from an automatically annotated corpus, which is tagged with higher accuracy when the TES model is used. For the *Web* subset, the performance gain is even larger. The relative performance gain is a bit lower for the TES model (+0.62) compared to the TE model (+0.75), which can be explained by the fact that the TES model generally performs better than the TE model on out-of-vocabulary items; see Sect. 3.4 below for details.

Overall, our tagger improves state-of-the-art substantially. Our best configuration (TES/DS) outperforms the previous best system by 0.42% accuracy.

3.4 Performance on Unknown Words

In a second experiment, we investigate the performance of our distributional smoothing approach in more detail. We split the test set into three parts—in-vocabulary tokens (IV), out-of-vocabulary tokens covered by our distributional smoothing approach (OOV/DS) and out-of-vocabulary tokens which do not occur in the *Chefkoch* corpus and are thus dealt with using suffix probabilities only (OOV/BL)—and measure accuracy of our models on these three subsets separately. Figure 2 shows, for each of the three subsets, the number of tokens in the subset, the performance of the DS models and the performance gain of the DS models over the corresponding BL models, for both TE and TES. We see that distributional smoothing is very effective and improves accuracy over the baseline by 7–8%, except for the TE model on the *CMC* subset where

	TE/DS *vs.* TE/BL				TES/DS *vs.*TES/BL			
	CMC		Web		CMC		Web	
IV	4589	90.1 (+0.03)	6624	94.9 (+0.05)	4732	90.3 (+0.02)	6682	94.9 (+0.03)
OOV/DS	472	65.0 (+2.97)	629	83.0 (+8.11)	343	71.7 (+7.29)	581	84.0 (+7.40)
OOV/BL	173	67.1 (+0.58)	315	77.8 (+0.95)	159	66.7 (+0.00)	305	77.7 (+0.66)

Fig. 2. Accuracy comparison of the DS and BL models for in- (IV) and out-of-vocabulary (OOV) tokens on the *CMC* and the *Web* subset. The rows give, for each group, the number of tokens, the accuracy of the DS model and the accuracy gain of the DS model over the BL model.

we obtain only a moderate improvement of approx. 3%. Overall, the improvement over the baseline is 5.1% (TE) and 5.8% (TES) on all out-of-vocabulary tokens.

4 Conclusions

In this paper, we presented work on part-of-speech tagging of German social media and web texts, using a fine grained tagset. Our tagger is based on a simple trigram Hidden Markov Model, which we extend with a distributional approach to estimating emission probabilities of out-of-vocabulary tokens. While technically very simple, our tagger is very effective and outperforms, or comes very close to, state-of-the-art systems even in the baseline configuration without distributional smoothing. Using distributional smoothing improves accuracy of out-of-vocabulary tokens by up to 5.8%. Overall, we improve state-of-the-art by 0.4% to 90.9% accuracy.

References

Beißwenger, M.: Das Dortmunder Chat Korpus. Zeitschrift für Germanistische Linguistik **41**(1), 161–164 (2013)

Beißwenger, M., Bartz, T., Storrer, A., Westpfahl, S.: Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. EmpiriST 2015 guideline document (2015)

Beißwenger, M., Bartsch, S., Evert, S., Würzner, K.-M.: 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, pp. 44–56 (2016). http://aclweb.org/anthology/W16-2606

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., Koenig, E., Lezius, W., Rohrer, C., Smith, G., Uszkoreit, H.: TIGER: linguistic interpretation of a German corpus. J. Lang. Comput. **2**(4), 597–620 (2004). Special Issue

Brants, T.: TnT - a statistical part-of-speech tagger. In: Proceedings of the Sixth Conference on Applied Natural Language Processing, pp. 224–231 (2000)

Eisenstein, J.: 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, pp. 359–369 (2013). http://aclweb.org/anthology/N13-1037

Giesbrecht, E., Evert, S.: Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In: Proceedings of the Fifth Web as Corpus Workshop, San Sebastian, Spain, pp. 27–35 (2009)

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: 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, pp. 42–47 (2011). http://aclweb.org/anthology/P11-2008

Horbach, A., Steffen, D., Thater, S., Pinkal, M.: Improving the performance of standard part-of-speech taggers for computer-mediated communication. In: Proceedings of the 12th Edition of the KONVENS Conference, vol. 1, pp. 171–177 (2014)

Horbach, A., Thater, S., Steffen, D., Fischer, P.M., Witt, A., Pinkal, M.: Internet corpora: a challenge for linguistic processing. Datenbank Spektrum **15**(1), 41–47 (2015)

Horsmann, T., Zesch, T.: LTL-UDE@EmpiriST 2015: tokenization and PoS tagging of social media text. In: Proceedings of the 10th Web as Corpus Workshop, pp. 120–126 (2016). http://aclweb.org/anthology/W16-2615

Manning, C.D.: Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In: Gelbukh, A.F. (ed.) CICLing 2011. LNCS, vol. 6608, pp. 171–189. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19400-9_14

Neunerdt, M., Reyer, M., Mathar, R.: Efficient training data enrichment and unknown token handling for POS tagging of non-standardized texts. In: Proceedings of the 12th edition of the KONVENS conference, vol. 1, pp. 186–192 (2014)

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., Smith, N.A.: Improved part-of-speech tagging for online conversational text with word clusters. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 380–390 (2013). http://aclweb.org/anthology/N13-1039

Prange, J., Thater, S., Horbach, A.: Unsupervised induction of part-of-speech information for OOV words in German internet forum posts. In: Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication/Social Media (2015)

Prange, J., Horbach, A., Thater, S.: UdS-(retrain|distributional|surface): improving POS tagging for OOV words in German CMC and web data. In: Proceedings of the 10th Web as Corpus Workshop, pp. 63–71 (2016). http://aclweb.org/anthology/W16-2608

Rehbein, I.: Fine-grained POS tagging of German tweets. In: Gurevych, I., Biemann, C., Zesch, T. (eds.) GSCL 2013. LNCS (LNAI), vol. 8105, pp. 162–175. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40722-2_17

Ritter, A., Clark, S., Mausam, Etzioni, O.: Named entity recognition in tweets: an experimental study. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534 (2011). http://aclweb.org/anthology/D11-1141

Schiller, A., Teufel, S., Stöckert, C., Thielen, C.: Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart and Seminar für Sprachwissenschaft, Universität Tübingen (1999). http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf

Schütze, H.: Distributional part-of-speech tagging. In: Seventh Conference of the European Chapter of the Association for Computational Linguistics (1995). http://aclweb.org/anthology/E95-1020

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.