Part-Of-Speech Tagging for Social Media Texts

Chapter · October 2013

DOI: 10.13140/2.1.2905.7284

CITATIONS
12

READS
647

4 authors, including:

Bianka Trevisan
Amazon Liquavista BV
37 PUBLICATIONS **130** CITATIONS
SEE PROFILE

Michael Reyer
RWTH Aachen University
22 PUBLICATIONS **214** CITATIONS
SEE PROFILE

Part-Of-Speech Tagging for Social Media Texts

Melanie Neunerdt[1], Bianka Trevisan[2], Michael Reyer[1], and Rudolf Mathar[1]

[1] Institute for Theoretical Information Technology,
[2] Textlinguistics/TechnicalCommunications,
RWTH Aachen University, Germany

Abstract. Work on Part-of-Speech (POS) tagging has mainly concentrated on standardized texts for many years. However, the interest in automatic evaluation of social media texts is growing considerably. As the nature of social media texts is clearly different from standardized texts, Natural Language Processing methods need to be adapted for reliable processing. The basis for such an adaption is a reliably tagged social media text training corpus. In this paper, we introduce a new social media text corpus and evaluate different state-of-the-art POS taggers that are retrained on that corpus. In particular, the applicability of a tagger trained on a specific social media text type to other types, such as chat messages or blog comments, is studied. We show that retraining the taggers on in-domain training data increases the tagging accuracies by more than five percentage points.

Keywords: POS tagging, statistical NLP, social media texts.

1 Introduction

Many Natural Language Processing (NLP) methods, e.g., syntactical parsing or sentiment analysis, require accurate Part-of-Speech (POS) tag information for a given word sequence. This information is provided by automatic POS tagging which is a well researched field. For German, which has a strong morphological character, state-of-the art POS taggers yield per-word accuracies of 97% to 98% for standardized texts.

However, the interest in using NLP methods for non-standardized texts, such as social media texts, is growing. The automatic evaluation of social media texts is particularly essential for the task of sentiment analysis. Social media texts comprise user generated content such as blog comments or chat messages. Indeed, differ from standardized texts in the word usage but also in their grammatical structure. This holds for adapted NLP methods in general and in particular for the adaption of POS tagging methods to such text types. Most state-of-the-art taggers are developed for standardized texts. Hence they are trained on large newspaper corpora. However, previous studies, e.g., [6], have shown that applying such taggers to non-standardized texts results in a significant performance loss. The lack of social media text reference corpus, which is sufficiently large to train a tagger, might be the reason that automatic POS tagging for social media texts

I. Gurevych, C. Biemann, and T. Zesch (Eds.): GSCL 2013, LNAI 8105, pp. 139–150, 2013.

has rarely been studied so far. Furthermore it has not been addressed for German language, yet. Hence, tagging methods need to be adapted and annotation rules for social media text characteristics are required. A first step is to provide in-domain training data to yield higher tagging accuracies when existing taggers are applied to social media texts.

In this paper, we first introduce a new social media text corpus containing Web comments composed of 36,000 annotated tokens. Considering the German standard *Stuttgart/Tübinger TagSet* (STTS) [11], we distinguish 54 tag classes. Until now, no standardized STTS extension has been published for the annotation of social media texts in German. Thus, we define annotation rules for specific social media text characteristics.

We use the introduced corpus to retrain existing taggers. Combining the social media text corpus with standardized texts (joint-domain), four state-of-the-art taggers, TreeTagger [13], TnT [3], Stanford [14], and SVMTool [7], are trained and evaluated by cross validation. We show that tagging accuracies increase by more than five percentage points for social media texts. The *TIGER* corpus [2] serves as standardized training data in the experiments. Mean tagging accuracies with standard deviations are calculated and compared for all taggers. A more detailed look into the results is provided for the TreeTagger, which has the highest tagging accuracies on German social media texts.

Finally, we study the applicability of the retrained taggers to four different social media text types, i.e., blog comments, chat messages, YouTube comments and news site comments. For the evaluation, 5,000 tokens that comprise the four types are additionally manually annotated with POS tags by manual processing. In particular, tagging errors are classified into four categories, with respect to different social media text characteristics. This particularly points out the special challenges for dealing with social media texts. It serves as starting point for the technical design of new social media taggers.

The outline of this paper is as follows: Section 2 summarizes the related work and gives an overview of POS tagging. In Section 3, we introduce the new social media text corpus and propose annotation rules for social media texts. Section 4 and Section 5 present our evaluation methodology and corresponding results. Section 6 concludes this work and discusses future research.

2 Related Work

Several papers have been published that deal with automatic POS tagging mainly by following statistical approaches. However, a number of rule-based methods have been proposed in the early stages of POS tagging research. The first rule-based approach has been presented in [9]. One of the latest rule-based methods is proposed in [4]. It yields similar accuracies as statistical approaches. Typical statistical POS taggers make use of two different probabilistic models, a Markov model or a maximum entropy model that captures lexical and contextual information.

Common Markov model taggers are proposed in [13,3]. TreeTagger [13] and TnT [3] are second order Markov models with some smoothing techniques for the estimation of lexical probabilities. TreeTagger utilizes a decision tree for reliable estimation of tag transition probabilities. Maximum entropy based taggers are proposed in [14,5]. These methods use the same baseline maximum entropy model and adapt their approach by using different features in the model. Furthermore, some other machine learning techniques are applied to the problem of automatic POS annotation, e.g., Support Vector Machines [7] and Neural Networks [12].

In [16,6] common POS taggers are evaluated and compared for German. Schneider et al. [16] compare a statistical and a rule-based tagger and point out the performance loss of the rule-based approach applied to unknown words. The performance of five state-of-the-art taggers applied to Web texts is studied in [6]. The corresponding results show that the automatic tagging of Web texts is not yet sufficient and that the accuracy drops significantly for different text genre.

The particular task of tagging non-standardized texts, characterized by frequent unknown words, is addressed in [5,10]. Gadde et al. [5] propose adaptions to the Stanford tagger to handle noisy English text. They evaluate their results based on a Short Message Service (SMS) dataset. They suggest to correct the tags of noisy words in a postprocessing step as well as some preprocessing cleaning techniques to the noise in the given sentence. Gimpel et al. [8] propose a twitter tagger based on a conditional random field (CRF) and adapt their features to twitter characteristics. In [10], the same authors propose some additional word clustering and further improve their method.

3 Corpora

Two corpora are used for training purposes, our social media corpus and a newspaper corpus. First, we introduce *WebCom* a new corpus that contains Web comments collected from *Heise.de*, which is a popular German newsticker site treating different technological topics. The comments for the manual POS annotation are selected from this underlying corpus. In order to obtain a corpus where many kinds of social media characteristics are represented, we select comments from different users. The selection of comments is carried out randomly over different users according to their posting frequencies. Each token is annotated with manually validated POS tags and lemmas. Annotation rules, particularly for social media text characteristics, are given in Section 3.1. A detailed annotation guideline as well as Inter-Annotator Agreement (IAA) studies can be found in [15]. We call the resulting corpus *WebTrain* because it provides supervised data for training purposes. The average POS tag ambiguity of tokens contained in the corpus is 2. This is significantly higher as the ambiguity in German newspaper texts, e.g. 1 for the *TIGER* corpus. Further statistical corpus information is given in Table 1. To the best of our knowledge, *WebTrain* is currently the largest social media text corpus enriched with POS information. However, aiming at training a POS tagger, 36,000 tokens is a relatively small number.

Table 1. Unsupervised and supervised social media text corpora

	WebCom	**WebTrain**
#Comments:	153,740	429
#Tokens:	15,080,976	36,284
#Words:	360,177	7,830
#Users:	15,007	183

In order to provide a sufficiently large training data amount, we combine *WebTrain* with the *TIGER* treebank [2] newspaper text corpus. It is the largest manually annotated German corpus and contains about 900,000 tokens of German newspaper text, taken from the *Frankfurter Rundschau*. The corpus annotation provides manually validated POS tags, lemmas, morphosyntactic features, and parse trees. For our purposes, only the STTS POS tag information is used.

To have a deeper look in the general applicability of the retrained taggers for social media texts, we create an additional corpus *WebTypes*. It is composed of roughly 5,000 tokens, where comments from different web sites and a corpus extract from the *Dortmunder chat corpus BalaCK 1-b* [1] are annotated in the same way than *WebTrain*. Four different types of social media texts are represented, Merkur newsticker comments, YouTube comments, blog comments, and chat messages.

3.1 Annotation Rules

The STTS tagset was developed 1999 in Stuttgart and has evolved over the years to the standard tagset for the morphosyntactic annotation in German; it provides information about the respective part of speech and its syntactic function. It was developed for the annotation of standardized texts. Until now, no extension for the annotation of the special characteristics of social media texts, e.g., emoticons, is present. Moreover, an extension of the existing tagset is problematic from a technical perspective, since existing NLP methods, e.g., syntactical parsing, require STTS POS tag information. Thus, the existing STTS tagset is used and social media text characteristics are tagged according to their syntactic function in our approach. For instance, emoticons are either at the end of a sentence or at intermediate positions. Therefore, they obtain the tag for sentence final "$." and sentence internal "$(" character. Contrarily, special characters and enumerations are only annotated with the internal character tag. Separated particles of apostrophization, e.g., ([*hab*], *'s - have it*), are tagged for verbs, conjunctions, and interrogative pronouns as *irreflexive personal pronoun* (PPER), *substituting demonstrative pronoun* (PDS), or *article* (ART). Numbers replaced by the corresponding digit in a word are annotated as *attributive adjective* (ADJA) or *proper noun* (NE), depending on the context. The overall annotation rules for particular social media characteristics are given in Table 2. All tags from the first column can be assigned according to the given grammatical context. Exemplary tokens are given in the last column. Note, that the text

Table 2. Annotation rules for social media texts

Tag	Description	Example
$. , $(	Emoticons	:-) , (*_*)
NE	File names, URLs	test.jpg , www.test.de
ITJ , PTKANT	Interaction words, inflectives	lol , seufz , yep
$(	Special characters	#, @, *, i. ii., a) b)
$(, $.	Multiple punctuations	... , !?!
PPER , ART , PDS	Apostrophization	[geht]′s , [wer]′s , [ob]′s
ADJA , NE	Number replacement	10er , 500er

is manually tokenized such that adequate POS annotation can be performed according to the given rules.

4 Evaluation Methodology

For our evaluation, we consider four state-of-the art taggers. We choose the taggers according to their tagging accuracy applied to German standardized texts. Furthermore, all taggers are used in the evaluation section of [6], where the performance of state-of-the art taggers on Web texts is studied. Hence, our results can be compared later to the published results. The selected taggers are the following:

1. TreeTagger [13], a Markov model based tagger using a decision tree for the estimation of tag transition probability.
2. TnT [3], a Markov model tagger that integrates some smoothing techniques for the estimation of lexical probabilities.
3. Stanford [14], a maximum entropy based tagger, integrating different word and tag features.
4. SVMTool [7], a tagger that utilizes support vector machines for classification.

The following evaluations are performed for all taggers, using the proposed default settings for training. Two 10-fold cross validations are carried out to evaluate and compare the tagging accuracy of the four taggers. First, a cross validation on ten equally sized *TIGER* corpus parts is carried out for randomly selected sentences. In the following we call such taggers *TIGER* taggers. Secondly, a 10-fold cross validation is performed on joint-domain training data from the *TIGER* and *WebTrain* corpus. The taggers are trained on a combination of nine *WebTrain* subsets and nine *TIGER* subsets in each validation step. Note, that we use a fix combination of nine *TIGER* subsets here in order to keep a remaining part for testing. We call the resulting ten trained taggers *WebTrain* taggers. For the sake of fair comparison, we use the pre-tokenized data and determine per-word accuracies for the same number of tokens. Furthermore, we study the application

of the taggers to social media texts types differing from the training text type. A tagger trained on Web comments from the newsticker site *Heise.de* is, for example tested on blog comments from different blog sites. Based on the cross validation trainings, the performance of all taggers is tested on the *WebType* corpus. Finally, the goal of our work is to analyze for which text characteristics, the tagging accuracies are improved by adding in-domain training data. Therefore, we introduce four categories, motivated by technical challenges of POS tagging, to describe social media text characteristics:

1. *Spoken language character* - The language is borrowed from spoken language and characterized by linguistic irregularities. In German Web comments, verbs are often shortened or merged (e.g., *hab, habs - have, have it*), fill and swear words are used (e.g., *Verdammt - Damn*), reflection periods are verbalized by interjections (e.g., *hmm, äh*) or elliptical constructions are used (e.g., *Entschuldigung! - Sorry!*). Thus, the language is characterized by a lower standardization degree or colloquial style as for newspaper texts.
2. *Dialog form* - A dialogic style characterizes the communication in social media applications. Hence, first and second person singular and plural formulations are predominant. On the other hand, newspaper texts are typically written in third person singular and plural, as this text type has a more descriptive character. Moreover, Web comments are dialogic texts, where many anaphoric expressions (e.g., *die - this, that*) can occur.
3. *Social media language* - The language is characterized by the use of interaction signs such as emoticons (e.g., *:-)*), interaction words (e.g., *lol, rofl*), leetspeak (e.g., *w!k!p3d!4*), word transformations (e.g., *EiPhone*), using mixed languages in the same context and references such as URLs and filenames (e.g., www.google.de).
4. *Informal writing style* - The majority of user posts are written in an informal way. Hence, social media texts suffer from spelling errors, typing errors, abbreviations, missing text and sentence structure (e.g., missing punctuation marks), missing capitalization, character iterations (e.g., *Helloooo*), and multiple punctuation (e.g., *!?!, !!!*).

For a detailed evaluation, the tagging errors are consistently classified into one of the four categories in a manual process by one person. Note, that the resulting classification does not serve for any training but rather for evaluation purposes. Hence, it is just important that the manual classification is performed consistently. However, analyzing the results in that way, is a good starting point for the technical design of new social media text taggers in future work.

5 Evaluation Results

This chapter discusses the results achieved by the previous described evaluation methods. First, we evaluate and compare the performance of the four state-of-the-art taggers applied to social media texts. Results for taggers trained on newspaper *TIGER* texts are compared to those that are additionally trained

on *WebTrain*. We particularly point out, that using social media text for training purposes does not negatively effect the tagging accuracies of standardized newspaper texts. Furthermore, trigram statistics show that social media texts differ in their grammatical structure. Finally, in Section 5.2 the application of the *WebTrain* taggers to different social media text types is investigated.

5.1 Tagger Comparison

Comparison starts with a discussion of tagger performance if all test corpora are considered and the taggers are trained on the *TIGER* corpus only. The results are compared to the performance when the taggers are trained on *WebTrain* (added to *TIGER*). More detailed results are presented to investigate the impact of newly learned words and newly seen trigrams.

Results for *TIGER* Training. First we compute the mean tagging accuracies with standard deviations for cross validations performed on *TIGER* and on additional *WebTrain* data. The first row in Table 3 gives the results for the different *TIGER* taggers. Additionally, the first column shows the tagging accuracy achieved by TreeTagger using the standard parameter file (Tree-SPF). Tagging accuracies around 97% are achieved for a comparison to the results from [6]. Slight deviations are observed due to the selection procedure of training sentences. TreeTagger performs worst on *TIGER* data. This is also stated in [6], see Table 3. The second row shows the results achieved by *TIGER* taggers performed on the ten test samples from *WebTrain*. The average tagging accuracies considerably decrease by 8 to 10 percentage points and the standard deviations increase. This can be explained by a different degree of social media text characteristics occurring in the randomly chosen test data.

Results for *WebTrain* Training. We consider the results achieved by *WebTrain* taggers. The second row of Table 4 gives the mean values achieved on *WebTrain* test data. For all state-of-the-art taggers adding in-domain training data leads to an improvement of 5.06 to 5.65 percentage points in average tagging accuracy. Applying TreeTagger results in the maximum average per-word accuracy of 93.72% in contrast to results from Table 3 for standardized *TIGER* texts, where TreeTagger performs the worst. This indicates that the TreeTagger approach is particularly suitable for dealing with the characteristics of non-standardized social media texts. Additionally, we test all *WebTrain* taggers on the held-out TIGER test sample. The corresponding tagging accuracies are depicted in the first row of Table 4. The results demonstrate, that using non-standardized social media texts as additional training data slightly increases the performance for tagging newspaper texts. Note, that the number of tokens for *WebTrain* test and *TIGER* test is the mean value over all cross validation test sets.

Table 3. Tagger evaluation for different text types trained on *TIGER*

Text type	Tree-SPF	TreeTagger	TnT	Stanford	SVM
TIGER test	95.54 ± 0.06	97.18 ± 0.04	97.29 ± 0.05	97.42 ± 0.03	97.45 ± 0.03
WebTrain test	87.08 ± 0.87	88.51 ± 0.99	88.57 ± 1.14	87.74 ± 1.02	87.65 ± 1.13
Merkur comments	94.95	93.11 ± 0.34	90.78 ± 0.73	89.96 ± 0.42	91.64 ± 0.31
Chat messages	81.89	85.63 ± 0.39	84.34 ± 0.24	83.78 ± 0.26	82.80 ± 0.27
YouTube comments	78.88	77.53 ± 0.59	74.85 ± 0.39	74.44 ± 0.55	74.27 ± 0.42
Blog comments	87.98	88.14 ± 0.53	86.93 ± 0.68	86.53 ± 0.51	85.13 ± 0.67

Table 4. Tagger evaluation for different text types trained on *WebTrain*

Text type	#Tokens	TreeTagger	TnT	Stanford	SVM
TIGER test	5,306	97.18 ± 0.03	97.31 ± 0.01	97.44 ± 0.01	97.47 ± 0.01
WebTrain test	3,628	93.72 ± 0.49	93.63 ± 0.37	93.18 ± 0.32	93.30 ± 0.56
Merkur comments	990	94.89 ± 0.38	93.49 ± 0.36	92.46 ± 0.38	93.72 ± 0.41
Chat messages	1,728	89.12 ± 0.18	87.96 ± 0.11	87.81 ± 0.16	86.57 ± 0.13
YouTube comments	1,463	84.03 ± 0.24	81.18 ± 0.19	81.23 ± 0.16	80.56 ± 0.19
Blog comments	815	91.35 ± 0.18	90.46 ± 0.12	90.29 ± 0.17	88.04 ± 0.13

More detailed cross validation results achieved by *WebTrain* taggers are given in Table 5. Particularly, the accuracy rates are split up into known words and out-of-vocabulary (unknown) words. We give mean accuracies and standard deviations as well as the percentage of unknown words. In general, unknown words are such words, that are not known from the training text, i.e., the arising lexicon. Note, that unknown word rates for the same data are partially differing for different taggers. For instance, TreeTagger excludes cardinal numbers from unknown word counts. Unknown word rates are roughly 8% for all considered taggers. This is about 2 percentage points higher than stated for *TIGER* newspaper texts. Particularly, standard deviations for unknown words indicate how robust a tagger is when social media texts are considered. Stanford tagger is the most robust tagger. However, the standard deviation of 1.97 is still pretty high. The highest tagging accuracies of 70.58% for unknown words, with only slightly higher standard deviations are achieved with the SVMTool. Nevertheless, TreeTagger slightly outperforms the other taggers in total.

Impact of Newly Learned Words. In order to investigate the performance on social media texts we carry out different training approaches. Exemplary experiments are presented for the TreeTagger. Table 6 shows the results achieved by using differently trained taggers, i.e., TreeTagger with the standard parameter file, *TIGER* tagger, *TIGER* tagger with an auxiliary lexicon created from *WebTrain*, and *WebTrain* tagger. The auxiliary lexicon covers the words/tokens that are contained in *WebTrain* texts including their corresponding set of possible tags. Neither lexical probabilities for such words nor trigrams of such texts are given. Tagging is performed on *WebTrain* test data. The results achieved by

Table 5. Results for 10-fold cross validation trained joint-domain data using *WebTrain*

	TreeTagger	TnT	Stanford	SVM
Total	93.72 ± 0.49	93.63 ± 0.37	93.18 ± 0.32	93.30 ± 0.56
Known	95.83 ± 0.43	95.81 ± 0.51	95.61 ± 0.40	95.58 ± 0.45
Unknown	67.98 ± 3.14	70.58 ± 2.08	68.14 ± 1.97	69.33 ± 2.54
Percentage unknowns	7.58 ± 0.75	8.65 ± 0.62	8.81 ± 0.62	8.65 ± 0.62

Table 6. Detailed TreeTagger tagging accuracies for different training approaches

	Tree-SPF	*TIGER*	*TIGER* + Web Lex.	*WebTrain*
Total	87.08 ± 0.87	88.51 ± 0.99	92.63 ± 0.68	93.72 ± 0.49
Known	92.05 ± 0.53	94.47 ± 0.57	94.74 ± 0.53	95.83 ± 0.43
Unknown	44.77 ± 2.46	54.13 ± 3.15	66.47 ± 3.04	67.98 ± 3.14
Percentage unknown	10.50 ± 0.76	14.71 ± 0.96	7.58 ± 0.75	7.58 ± 0.75

using the standard parameter file for unknown words show a high performance loss. This can be explained by the fact that for unknown words the so called open-class tags are restricted to 7 STTS tags for nouns, named entities, verbs, and adjectives. This restriction is inappropriate for the task of tagging social media texts, where unknown words can almost be of any word class. The number of unknown words is reduced by one half when comparing *TIGER* tagger results with *WebTrain* tagger results. Moreover, the tagging accuracy for the lower number of unknowns is increased by more than 10 percentage points.

Impact of Newly Seen Trigrams. We use the *WebTrain* text corpus for training instead of only considering it as a word lexicon. This leads to further tagging improvement by more than one percentage point, see the right column of Table 6. This can be explained by different grammatical structures of social media texts, which need to be learned from a sufficient amount of in-domain training data. Finally, we demonstrate how different the grammatical structure in social media texts is compared to newspaper texts. STTS tag trigram frequencies are calculated for both corpora *TIGER* and *WebTrain*. The overall results are depicted in Table 7. The third column shows the ratio between different trigrams and their frequencies for the different corpora. Results illustrate the higher variability in social media texts, which is ten times higher than in newspaper texts. Particularly, we compare statistics for tag trigrams that occur in *WebTrain* texts but are unknown from the *TIGER* corpus. The statistics are given in the last row. *WebTrain* texts contain 18% new trigrams, that never occur in the newspaper corpus *TIGER*. Those trigrams constitute 6% frequency of all *WebTrain* trigram counts. Particularly, for those trigrams the ratio/variability is increasing by a factor of three. Both results motivate the need of in-domain training data for reliable estimation of transition probabilities, e.g., for trigrams.

Table 7. Trigram comparison for *TIGER* and *WebTrain* corpora

	Trigrams	Trigram frequencies	Ratio
Total *WebTrain*	7,215	36,282	0.20
Total *TIGER*	16,563	888,982	0.02
Only in *WebTrain*	1,290	2,120	0.61

5.2 Results for Different Social Media Text Types

In this section, we study the application of taggers to different social media text types, where the taggers are not trained for the particular type. To illustrate the improvements, Table 3 shows tagging accuracies and standard deviations for all taggers trained on *TIGER*. Table 4 depicts the improved tagging accuracies that are achieved by *WebTrain* taggers. Application of joint-domain trained taggers leads to a consistent performance increase between approximately 2 and 7 percentage points for different social media text types. Considerable improvements can be observed for chat and YouTube data, which are highly characterized by a dialogue form. Moreover, TreeTagger outperforms TnT, Stanford, and SVMTool for all considered social media text types.

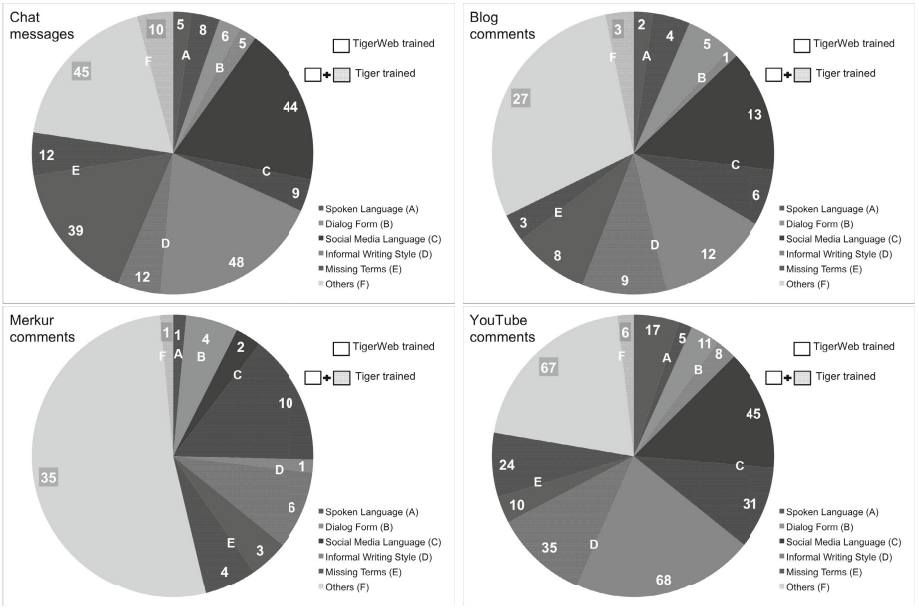

Fig. 1. Error classification and improvement for different social media text types

Finally, we evaluate the results for all social media text types with respect to the four different characterization categories introduced in Section 4. Therefore, we filter and classify all words which are not correctly tagged by using

TreeTagger trained on *TIGER* and on joint-domain training *TIGER* and *WebTrain*, respectively. Note, that the social media text categories are complemented by two more categories, *missing terms* and *others*. The category *missing terms* covers topic specific nouns and named entities. The category *others* comprises all other occurring wrongly tagged words (e.g. *aber - but, die - the*), which are not related to any of these categories and can also occur in standardized texts. Figure 1 depicts absolute errors for each category. The shaded areas illustrate the absolute error reduction for each particular category. The corresponding total number of tokens for each type are depicted in Table 4. Applying the *WebTrain* TreeTagger, the errors made for all four social media categories by the *TIGER* TreeTagger can be reduced from 26% to 71% . Errors are reduced up to 86% for the *social media language* category. Using in-domain training data, effectively reduces *missing terms* errors by more than a third for all text types. The error rates for the category *others* can hardly be improved. The enrichment with in-domain training enables the special handling of social media text characteristics, particularly for *social media language* and *informal writing style* categories. In total, a significant error reduction can be achieved over all categories.

6 Conclusions

We have shown that the performance of state-of-the-art POS taggers can significantly be improved for social media texts. The improvement is achieved by taking training data from social media texts into account. We have created a new social media text corpus *WebTrain* that contains 38,000 manually annotated tokens. It can be used to retrain such taggers. We introduce an adequate STTS annotation guideline for social media texts. To fulfill the requirement of other NLP methods, we use the original STTS tag set without any extensions.

For all state-of-the-art taggers, adding in-domain training data leads to a significant improvement of more than five percentage points for the tagging accuracy. TreeTagger cross validation leads to a maximum average per-word accuracy of 93.72%. Moreover, TreeTagger outperforms TnT, Stanford, and SVMTool for all considered social media text types. Taggers trained on Web comments can successfully be used for different text types. Applying the joint-domain trained taggers leads to a consistent performance increase between approximately 2 and 7 percentage points for different text types. Considerable improvements are obtained for chat and YouTube data, which are highly characterized by a dialogue form. However, the overall accuracies of 89% and 84% demand for further investigations. Enrichment of the newspaper corpus by social media text data, leads to a slightly improved evaluation on newspaper texts. Hence, the enhanced training data improves tagging results on all kind of investigated text types.

Finally, we have shown that the grammatical structure in social media texts differs. The new grammatical structures are learned from a sufficient amount of in-domain training data and account for a considerable improvement of the tagging accuracy.

Beyond providing in-domain training data for the enhancement of POS tagging accuracy for social media texts, we currently work on adaptions to existing tagger models. We particularly focus on the parameter estimation for unknown words to further improve accuracies, particularly for chat and YouTube data.

References

1. Beißwenger, M.: Corpora zur computervermittelten (internetbasierten) Kommunikation. Zeitschrift für Germanistische Linguistik 35, 496–503 (2007)
2. Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., Uszkoreit, H.: TIGER: Linguistic Interpretation of a German Corpus. In: Research on Language & Computation, pp. 597–620 (2004)
3. Brants, T.: TnT – A Statistical Part-of-Speech Tagger. In: Proceedings of the 6th Applied Natural Language Processing Conference, pp. 224–231 (2000)
4. Brill, E.: A Simple Rule-based Part of Speech Tagger. In: Proceedings of the Third Conference on Applied Natural Language Processing, pp. 152–155 (1992)
5. Gadde, P., Subramaniam, L.V., Faruquie, T.A.: Adapting a WSJ Trained Part-of-Speech Tagger to Noisy Text: Preliminary Results. In: Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data, pp. 5:1–5:8 (2011)
6. 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, pp. 27–35 (2009)
7. Giménez, J., Màrquez, L.: Svmtool: A General POS Tagger Generator Based on Support Vector Machines. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, pp. 43–46 (2004)
8. 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, pp. 42–47 (2011)
9. Klein, S., Simmons, R.F.: A Computational Approach to Grammatical Coding of English Words. J. ACM 10, 334–347 (1963)
10. Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N.: Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances. Technical report, School of Computer Science, Carnegie Mellon University (2012)
11. Schiller, A., Teufel, S., Stöckert, C., Thielen, C.: Guidelines für das Tagging deutscher Textcorpora mit STTS. University of Stuttgart (1999)
12. Schmid, H.: Part-of-Speech Tagging With Neural Networks. In: Proceedings of the 15th Conference on Computational Linguistics, pp. 172–176 (1994)
13. Schmid, H.: Improvements in Part-of-Speech Tagging With an Application to German. In: Proceedings of the ACL SIGDAT-Workshop, pp. 47–50 (1995)
14. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich Part-of-Speech Tagging With a Cyclic Dependency Network. In: Proceedings of Human Language Technology Conference, pp. 173–180 (2003)
15. Trevisan, B., Neunerdt, M., Jakobs, E.-M.: A multi-level annotation model for fine-grained opinion detection in German blog comments. In: Proceedings of KONVENS 2012, pp. 179–188 (2012)
16. Volk, M., Schneider, G.: Comparing a statistical and a rule-based tagger for German. In: Proceedings of the 4th Conference on Natural Language Processing, x pp. 125–137 (1998)