Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication

Andrea Horbach, Diana Steffen, Stefan Thater, Manfred Pinkal
Department of Computational Linguistics, Saarland University, Saarbrücken, Germany
(andrea|dsteffen|stth|pinkal)@coli.uni-saarland.de

Abstract

We assess the performance of off-the-shelve POS taggers when applied to two types of Internet texts in German, and investigate easy-to-implement methods to improve tagger performance. Our main findings are that extending a standard training set with small amounts of manually annotated data for Internet texts leads to a substantial improvement of tagger performance, which can be further improved by using a previously proposed method to automatically acquire training data. As a prerequisite for the evaluation, we create a manually annotated corpus of Internet forum and chat texts.

1 Introduction

Around the turn of the century, the Internet made huge amounts of natural-language text easily accessible, and thus enabled a hitherto inconceivable success story of data-driven, statistical methods in computational linguistics. But the Internet also created a new challenge for language processing because it substantially changed the object of investigation. In computer-mediated communication (CMC), a wide variety of new text genres and discourse types such as e-mail, twitter, blogs, and chat rooms have emerged, which differ from standard texts in various ways and to different degrees. Differences include tolerance against typing errors and spelling rules, inclusion of colloquial, spoken-language elements in lexicon, syntax, and style (e.g., contractions like *gibt es* to *gibts*); intended use of non-standard-language components, like systematic “misspelling” and non-standard lexical items (e.g., neologisms or acronyms), to mention just a few. Statistical NLP tools are usually trained on and optimized for standard texts like newspaper articles. Reliable high-performance off-the-shelf tools show a dramatic performance drop, when applied to substantially differing linguistic material. This holds also for basic tasks such as POS tagging, which is particularly detrimental because the basic information is needed for all kinds of more advanced analysis tasks.

In this paper, we report work on POS tagging of two different CMC text types in German. We assess the performance of POS taggers trained on standard newspaper texts when applied to CMC texts and explore easy-to-implement and low-resource methods to adapt these taggers to CMC texts. We test the performance of three state-of-the-art taggers and explore two adaptation methods: First, we generate additional training material from automatically annotated data using a method that has been proposed recently by Kübler and Baucom (2011) for a different domain adaptation task. Second, we use small amounts of manually annotated CMC data as additional training data.

The main result of this paper is that even small amounts of manually annotated CMC training data substantially improve tagger performance on CMC texts; a combination of manually annotated and automatically acquired training data leads to a further improvement of tagger performance to up to 91% on texts from an Internet forum. A further major contribution is the POS-tagged CMC gold standard corpus consisting of about 24 000 tokens, which we created as a prerequisite for our evaluation and which will be made publicly available.

2 Related work

The growing interest in CMC language can be seen from a number of recently established collaborative activities like the scientific network *Empirical Research on Internet-based Communication*[1], the recently launched European network *Building and Annotating CMC Corpora*[2], and the Special Interest Group *Computer-mediated Communication* within the Text Encoding Initiative[3] (TEI).

Specific work for POS-tagging of non-standard texts include work by Ritter et al. (2011), Derczynski et al. (2013), Gimpel et al. (2011) and Owoputi et al. (2013), who report about POS tagsets and optimization of linguistic tools for annotating English Twitter data.

Kübler and Baucom (2011) investigate domain adaptation for POS taggers using the consent of three different taggers on unannotated sentences to create a new training set. They reach a moderate increase in accuracy from 85.8% to 86.1% on dialogue data but are still far below the performance on standard newspaper texts. We adopt their approach of tagger consent as one way of training set expansion in our experiments.

Work for German has been done by Giesbrecht and Evert (2009), who compare the performance of five different statistical POS tagger on different types of Internet texts, showing that the accuracy of approx. 97% on standard newspaper texts drops below 93%s when tagging web corpora. They mostly investigate texts that are close to standard language such as online news texts. Forum texts deviate most from the standard and the performance for forum texts matches our observations. Chat corpora are not covered in their study.

Bartz et al. (2014) suggest an extension of the widely used STTS tagset for POS tagging of web corpora, which we also use.

While our approach tries improves the performance of existing POS taggers on CMC texts, Rehbein (2013) develops a new POS tagger for German twitter data, which is trained using word clusters with features from an automatically created dictionary and out-of-domain training data.

3 Gold standard annotation

This section describes the annotation of computer-mediated discourse with POS information to be used as gold standard data in the experiments reported in Section 4 below.

3.1 Data sources

We select two complementary types of Internet text – forum posts from the Internet cooking community *www.chefkoch.de* and the *Dortmund Chat Corpus* (Beißwenger, 2013) – to cover a range of phenomena characteristic of Internet-based communication.

Forum. We use forum articles from the Internet cooking community *www.chefkoch.de*, which we downloaded in Feb. 2014, resulting in a large corpus of about 500 million tokens. Although the website primarily offers cooking-related services, forum articles address a wide range of everyday life topics and only a minor part of them – less than 1% as indicated by a case study – has the form of actual cooking recipes. In comparison to chats, we expect a higher agreement with standard language.

Chat. We complement the forum dataset with the *Dortmund Chat Corpus*, which is the standard corpus for German chat data; it consists of chat logs of various degrees of formality, ranging from very informal contexts to moderated expert chats. Since the focus of our research are phenomena typical for computer-mediated discourse, we select our gold standard data only from informal chats, which we assume to contain a larger number of interesting CMC phenomena.

3.2 Tagset

CMC data contain some language phenomena that are not properly covered by the standard STTS tagset, such as emoticons, so called “action words” in inflective form (e.g., *rumsitz*), URLs and various kinds of contractions. In order to account for the most frequent of those phenomena we use an extended version of STTS proposed by Bartz et al. (2014) containing additional tags for these categories.

We add two tags to capture errors made by the writers unaware of German spelling rules. ERRAW is assigned when a token should be part of the following token, i.e. if the writer inserted an erroneous whitespace; ERRTOK is a tag for the opposite case when the writer joined two words

[1] http://www.empirikom.net/

[2] https://wiki.itmc.tu-dortmund.de/cmc/

[3] http://www.tei-c.org/Activities/SIG/CMC/

tag	description	example	freq. forum	freq. chat
VVPPER	full verb + personal pronoun	versuchs, gehts, gibbet, kuckste	0.10	0.26
VMPPER	modal + personal pronoun	kanns, willste	0.02	0.05
VAPPER	auxiliary + personal pronoun	isses, hassu, wirste	0.06	0.13
KOUSPPER	conjunction + personal pronoun	wenns	0.01	0.00
PPERPPER	2 personal pronouns	[wenn] ses [frisst]	0.01	0.01
ADVART	adverb + article	son, sone	0.00	0.03
ADR		@nudelsupperstern, Sebastian	0.38	2.20
URL		www.uni-hildesheim.de	0.00	0.05
ONO	onomatopoeia	hehe, Mmmmmm	0.02	0.50
EMO	emoticons	:-), ⟨img src="smileys/wink.gif"⟩	1.72	1.40
AW	a verb in inflective form	ächz, rumsitz, knuddel	0.15	2.30
AWIND*	marks AW boundaries	*	0.24	4.01
ERRAW*	incorrectly separated word	[meine Kinder da] anzu [melden]	0.20	0.11
ERRTOK*	tokenization error	gehtso, garnicht	0.07	0.15
all new tags			3.02	11.18
all standard STTS tags			97.98	88.82

Table 1: Additional STTS tags, descriptions, examples and tag frequencies (%) in the goldstandard corpora. A * marks those tags that were not included in the extension by (Bartz et al., 2014)

that should be separated. Table 1 shows all non-standard tags we use together with examples.

3.3 Annotation

We manually annotated 11 658 tokens from the *Dortmund Chat Corpus* and 12 335 tokens from randomly chosen posts from the *chefkoch* corpus with POS information. Prior to annotation, the data has been automatically tokenized. The tokenizer sometimes tears apart strings that should form one token, such as several subsequent punctuation marks (e.g., *!!!*) or ASCII emoticons. Those systematic errors have been cleaned up manually. To simplify the annotation process, we also corrected few tokenisation errors made by the user in cases where it was an obvious typing error; for instance, *wennman* was corrected to *wenn man*.

Each file in both subcorpora has been annotated by two annotators. For the forum subcorpus, annotators were able to see the first post in the respective thread in order to provide them with potentially helpful context. For the chat data, they annotated continuous portions of approx. 550 tokens of chat conversations.

Annotators were asked to ignore token-level errors like typos or grammatical errors whenever possible, i.e. to annotate as if the error was not there. For instance, when the conjunction *dass* was erroneously written *das*, they should annotate KOUS even though *das* as a correct form can only occur as ART, PRELS or PDS.

After the annotations, annotators were shown where their annotation differed from the one of their co-annotator (without showing them the other annotation) in order to self-correct obvious mistakes. Cases of disagreement after that initial error correction have been resolved by a third annotator. The pairwise inter-annotator agreement (κ coefficient) ranges between 0.92 and 0.95 after the initial annotation and between 0.96 and 0.97 after self-correction.

Split into Training and Test Data. For our experiments in the next section, we split the gold standard into one third that is used as additional training material and two thirds for testing, making sure that equal portions of the *chat* and *forum* datasets are used in the resulting test and training dataset.

3.4 Corpus Analysis

The two subcorpora vary considerably not only in general linguistic properties like average sentence length (10.5 tokens for *forum*, 5.9 for *chat*) but even more so in the frequency with which POS tags, especially the non-standard tags occur. Table 1 shows the relative frequency of the new tags in both corpora. These numbers confirm our initial hypothesis about the degree of deviation from the standard in the two subcorpora: While the *forum* data only contain 3% of nonstandard tags, *chat* contains 11.2% of those new tags, thus clearly calling for adapted processing tools. 78.3% of all sentences in *forum* do not contain any non-standard tag, while in *chat* only 60.0% of all sentences are covered by the traditional STTS tagset.

4 Experiments

This section compares and combines two ways to re-train statistical POS taggers to improve their performance on CMC texts: (a) We extend a standard newspaper-based training corpus with data drawn from automatically tagged CMC texts applying a technique proposed by Kübler and Baucom (2011). (b) We extend the training corpus with small portions of manually annotated CMC texts. Results show that while the first approach leads to minor improvements of tagger performance, it is outperformed by a large margin by the second approach – even if only very few additional training sentences are added to the training corpus. A small further improvement can be obtained by combining the two approaches.

4.1 Methods

The key idea behind the approach of Kübler and Baucom (2011) is to parse raw text using different taggers, and to extend the training data for the taggers with automatically annotated sentences for which all taggers produce identical results. In our experiment, we use the following three taggers: TreeTagger (Schmid, 1994), Stanford Tagger (Toutanova et al., 2003) and TnT (Brants, 2000).

Baseline training corpus. As a starting point for our re-training experiments, we train our taggers using the Tiger corpus (Brants et al., 2004), which is a widely used German newspaper corpus providing POS annotations for roughly 900 000 tokens (50 000 sentences). The Tiger corpus consists of 20-year-old newspaper articles using the old German orthography. Since many words in our datasets are written according to the new spelling rules introduced in 1996, we automatically convert the original Tiger corpus to the new German orthography using *Corrigo* (Kurzidim, 2004) and replace approx. 11 000 tokens (1.2%) by their new spelling. We combine both variants of the corpus (original and converted) into a single new training corpus, referred to as "Tiger New" (*tn*) below.

Experiment 1: Corpus expansion by using multiple taggers. We apply each of the three taggers to the complete *Chefkoch* and *Dortmund Chat* datasets, resulting in an annotated corpus consisting of around 36 000 000 sentences.[4] For around 2 700 000 sentences ($< 8\%$) all three taggers agree completely. From those sentences we randomly select 50 000 sentences (561 000 tokens) from *Chefkoch* and 10 000 sentences (102 000 tokens) from *Dortmund Chat* and add them to our baseline corpus; we refer to the resulting training corpus as *tn+auto*.

Experiment 2: Adding manually annotated CMC data. In a second experiment, we use one third of the annotated gold standard data (around 7 800 tokens) as additional training material. Because this added data amounts to less than 1% of the number of tokens in the Tiger New corpus, we boost it by adding it several times, arbitrarily setting the boosting factor to 5 (*tn+gold*).

Experiment 3: Combining the two methods. In a third experiment, we combine the two approaches and generate a second set of automatically created gold-standard sentences by randomly selecting new training sentences automatically tagged with the *tn+gold* models (of the same amount as before). We call this dataset *tn+auto2*. The full dataset (*tn+gold+auto2*) consists of the Tiger corpus extended by gold standard data and additional automatically tagged data, tagged with the help of the same gold-standard data.

4.2 Results

The left part ("all sentences") of Table 2 shows the performance of the three taggers using different training datasets. Unsurprisingly, the original Tiger model (*tn*) performs very poorly when applied to non-standard CMC texts. Adding automatically annotated new training data (*tn+auto*) gives us a moderate and consistent positive effect across all corpora and taggers, improving tagger performance on average by 1.3% on the "All" test set. A much larger gain in performance can be obtained

[4] In order to avoid problems resulting from different tokenizations of the input texts when tagger results are compared (see below), we do not use the built-in tokenizers of the three taggers but use Stefanie Dipper's tokenizer (http://www.linguistics.ruhr-uni-bochum.de/˜dipper/token izer.html) for all three taggers.

		all sentences			standard sentences only		
Tagger	trained on	Chat	Forum	*All*	Chat	Forum	*All*
TreeTagger	Tiger new (tn)	0.714	0.845	*0.784*	0.800	0.874	*0.842*
	+auto	0.727	0.855	*0.796*	0.816	0.885	*0.854*
	+gold	0.826	0.881	*0.855*	0.861	0.909	*0.888*
	+gold+auto2	**0.835**	0.888	*0.863*	**0.873**	0.917	***0.898***
Stanford	tn	0.702	0.840	*0.776*	0.789	0.869	*0.834*
	+auto	0.715	0.851	*0.788*	0.803	0.880	*0.847*
	+gold	0.816	0.897	*0.860*	0.849	0.910	*0.884*
	+gold+auto2	0.826	0.903	*0.867*	0.863	0.918	*0.894*
TnT	tn	0.691	0.846	*0.774*	0.777	0.876	*0.832*
	+auto	0.708	0.857	*0.788*	0.796	0.889	*0.848*
	+gold	0.827	0.906	*0.870*	0.852	0.918	*0.889*
	+gold+auto2	**0.835**	**0.912**	***0.877***	0.863	**0.923**	*0.897*

Table 2: Accuracy of various models on both gold standard datasets, evaluated on the complete test set (*all sentences*) and on the subset that contains only sentences with tags from the original STTS (*standard only*). All differences in model performance are pairwise statistically significant (for each tagger and sub-corpus) according to a McNemar test ($p < 0.005$).

by adding small amounts of manually annotated CMC data (*tn+gold*); the performance gain is especially large for the *chat* subcorpus where it leads to an improvement of 13.4% for the best-performing TnT tagger, compared to the baseline. For forum data with a higher degree of standard language the improvement is less pronounced but still much larger compared to the *tn+auto* models. Adding both gold-standard data and automatically tagged data (*auto2*) leads to the best performing models with an accuracy of up to 91% (TnT) on forum data. We also tried to combine *auto* with *gold*, but found no positive effect.

Standard tags. The poor performance of the original tagger models and the large performance improvement obtained by adding additional training data from the gold standard is to some extent unsurprising, since the test data contains many tokens annotated with new POS tags which the original taggers cannot predict. We should note, however, that the performance gain cannot be explained by new POS tags only: The right part of Table 2 shows the performance of the taggers when applied to sentences from the gold standard in which new POS tags are not used. The performance of the original taggers is still quite low on this test set (between 83% and 84%) and is improved to 90% (TreeTagger) by using additional training data.

New tags. We also investigated the performance of the three taggers wrt. those words in the gold standard that received a new POS tag from the STTS extension by our overall best-performing model. TreeTagger achieves only 42% accuracy on such words, while Stanford Tagger and TnT achieve 58% and 67%, respectively. The low results are not surprising, given the small amount of training data. Stanford and TnT perform better than TreeTagger since they are able to generalize to unseen words, while TreeTagger assigns new tags only to known words and obviously needs larger amounts of training data to adapt to new texts or tags.

Performance on unknown words. The three taggers also show different behavior when evaluated only on unknown lexical material, i.e. words that do not occur in the training data. The best-performing model (*tn+gold+auto2*) for each tagger reaches performances of 41% (Stanford), 49% (TreeTagger) and 74% (TnT), showing again that TreeTagger and to some extent the Stanford Tagger seem to rely much more than TnT on lexical information.

Performance on specific new classes. Additionally we looked at the individual performance wrt. the new tags, for the best-performing models for all three taggers, and observe wide variation both across taggers and POS tags. Infrequent tags, especially the rare contractions are generally not learned well. Some tags with higher frequencies are learned with F-Scores higher than 0.95: EMO and AWIND for TnT, while TreeTagger (0.44) and Stanford (0.87) perform worse for EMO. Unsurprisingly AWIND (almost always a *) is learned well by all taggers. ADRs, although frequent, seem to be generally hard: the best-performing TnT tagger reaches an F-score of 0.18.

If we consider only unknown words within new tags we see a similar picture as in the general analysis of unknown words: While TnT can assign the new tags to the frequent classes (ADR, AW, EMO) although with some performance loss, Stanford and TreeTagger only successfully recognize some instances of unknown ADR, AW and EMO (but all with very low recall rates).

We also experimented with simple hand-crafted pattern matching rules to extend the accuracy for the most frequent new tags, e.g. tagging all words containing an @ in the beginning as ADR. However as the @ is left out in many ADRs and the syntactically integrated ADRs are tagged in the gold-standard as NE, we could not improve the performance by such additional rules. This shows again, that tagging of those new STTS categories is not a simple task and dependent from both word information and distribution.

4.3 Varying the amount of gold-standard data.

One potential disadvantage of using manually annotated gold-standard data to (re-)train taggers is that annotation is time-consuming and expensive. We should stress, however, that even a very small amount of manually annotated training data leads to a large improvement of tagger performance: We split the training part of the gold-standard into three equal parts and train models on corpora where we add (boosted 5 times) one part (*gold1*), two parts (*gold2*) and all three parts (*gold3*) to the training set. The results are presented in figure 1 exemplarily for the TnT tagger. We see that already a very limited time investment – around 20 hours of work for double annotations of approx. 2600 tokens – leads to a vital improvement of tagging performance and adding more gold data improves the performance further, but not to the same extent.

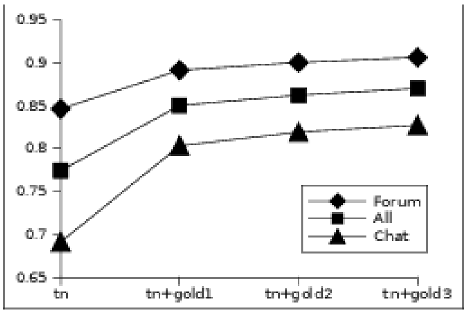

Figure 1: Accuracy of TnT when adding different amounts of gold standard data to the training data

5 Conclusions and Future Work

We have tested the performance of three state-of-the-art POS taggers and explored two low-resource and easy-to-implement adaptation methods to increase tagger performance on computer mediated communication (CMC) texts. A previously proposed method of using automatically annotated data to extend the training set leads to small improvement of tagger performance. A much higher improvement of tagger performance can be obtained by using small amounts of manually annotated CMC data as additional training data. A further improvement can be obtained by combining the two approaches, leading to up to 91% tagger performance on internet forum texts.

In future work, we will investigate the effects of training on a particular genre instead of CMC texts in general: While both forum and chat data deviate from standard texts, they each have their own particularities the taggers have to account for. The token *g* for example is used in in the *chefkoch* forum almost exclusively as abbreviation for *Gramm* (*gram*), whereas in chat corpora it usually indicates an action word as in **g** standing for *grin*.

We will also explore the effects that the choice of the tagging algorithm has and how the taggers can be used in a way that combines their individual strengths better.

Acknowledgements. This work is part of the BMBF-funded project “Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen.” We would like to thank our annotators Ali Abbas, Jana Ott, Jakob Prange and Maximilian Wolf. We also thank Dietrich Klakow for support concerning the forum dataset.

References

Thomas Bartz, Michael Beißwenger, and Angelika Storrer. 2014. Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internet-basierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *Zeitschrift für germanistische Linguistik*, 28(1):157–198.

Michael Beißwenger. 2013. Das Dortmunder Chatkorpus. *Zeitschrift für germanistische Linguistik*, 41(1):161–164.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther Knig, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. Tiger: Linguistic interpretation of a german corpus. *Journal of Language and Computation, Special Issue*, 2(4):597–620.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 224–231, Seattle, Washington, USA, April. Association for Computational Linguistics.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

Eugenie Giesbrecht and Stefan Evert. 2009. 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*, pages 27–35.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Sandra Kübler and Eric Baucom. 2011. Fast domain adaptation for part of speech tagging for dialogues. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, *RANLP*, pages 41–48. RANLP 2011 Organising Committee.

Michael Kurzidim. 2004. Fehlerpolizei - Wie gut sind Rechtschreibkorrektur-Programme? *c't*, (2):110–117.

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 *Proceedings of NAACL-HLT*, pages 380–390.

Ines Rehbein. 2013. Fine-grained pos tagging of german tweets. In *Language Processing and Knowledge in the Web*, pages 162–175. Springer.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, pages 252–259, Edmonton, Canada.