Part-of-Speech Tagging of Transcribed Speech

Margot Mieskes, Michael Strube

EML Research gGmbH, Heidelberg, Germany
http://www.eml-research.de/nlp

Abstract

We used four Part-of-Speech taggers, which are available for research purposes and were originally trained on text to tag a corpus of transcribed multiparty spoken dialogues. The assigned tags were then manually corrected. The correction was first used to evaluate the four taggers, then to retrain them. Despite limited resources in time, money and annotators we reached results comparable to those reported for the taggers on text. Based on our experience we present guidelines to produce reliably POS tagged corpora of new domains.

1. Introduction

Part-of-Speech (POS) tagging is a prerequisite for many high-level Natural Language Processing (NLP) tasks. A number of POS taggers have been developed and made available to the research community. The majority of them has been trained on written texts, mostly on newspaper texts. Only in few instances POS tagging was applied to transcribed speech. Examples for this can be found in Godfrey et al. (1992), Heeman & Allen (1999) and Zechner (2001). All three mainly deal with dialogues. Only Zechner (2001) reports results for multiparty dialogues as well.

Some work has been done to apply POS taggers to new domains with little or with no manual annotation. A small amount of manually annotated training data was used by Clark et al. (2003) and by Collins (2002), who used small amounts of data to do co-training and explore the performance of a Hidden Markov Model based perceptron respectively. Nakagawa et al. (2002) and van Halteren et al. (1998) used no manually annotated data for retraining but multiple POS taggers directly and voting techniques on the results. A third method is applied by Zavrel & Daelemans (2000) who use the results from several taggers trained on a small amount of training data as input for a learner.

There are problems in applying the approaches described above to our task, the application of POS taggers to transcribed multiparty dialogues. The researchers who tagged transcribed speech did not evaluate the taggers they used before retraining was done. The research exploring ways to retrain taggers with little or no data was performed on written text. But Wermter & Hahn (2004) showed that texts even from different domains can be very similar. They used two POS taggers, which were trained on newspaper texts, and applied them to medical texts. The evaluation on manually annotated medical texts gave good results. The authors explain this by a similarity in uni-, bi- and trigram POS distribution in newspaper and medical texts.

In our work we make use of four different POS taggers (see Section 2.). We apply them to transcribed multiparty spoken dialogues. We report results before (see Section 3.) and after the taggers have been retrained on manually annotated data, and specifically we evaluated the behaviour on different (increasing) amounts of data (see Section 4.). This led us to compile guidelines on how to efficiently create data for retraining taggers to be used on a new domain.

2. Taggers

Four taggers were considered. The TnT tagger[1] (Brants, 2000) uses a Hidden Markov Model (HMM) based on n-grams and lexical information. Two taggers from the Stanford Java library for tagging[2] (left3words (Toutanova & Manning, 2000) and bidirectional (Toutanova et al., 2003)). The first uses a maximum entropy model and the context of three words to the left. The second considers the context to the left and to the right by applying contextual HMMs (CHMM). Finally the Brill (TBL) tagger[3](Brill, 1994) uses transformation-based learning.

Each of these automatic taggers was originally trained and tested on the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993), which only consists of written text. The results reported for the taggers on this corpus vary between 96.5% and 97.24%. For TnT only the performance on known and unknown words is given separately with 97.7 and 89.0%. Therefore, using these taggers on the transcribed speech in the ICSI Meeting Recorder Corpus (*ICSI Corpus*) (Janin et al., 2003) will inevitably result in considerably lower accuracy rates. Some of the reasons that account for this are: First, the vocabulary of financial news is different from that of dialogues which mostly deal with speech and language technology. Second, the style is different between newspaper text and colloquial speech. Examples for these differences include disfluencies and explicit paragraph separation, but also sentence length and sentence complexity. Finally in the meetings non-native speakers are also involved. It seemed reasonable though to do the POS annotation semi-automatically and using these taggers as the basis for the manual correction. In order to get one tag for each token, a majority decision over the four automatic taggers (*Maj4* in the following) was reached. This was used as input for the human annotators, who corrected the data.

3. Manual Annotation

The whole *ICSI Corpus* contains 75 meetings. 12 were randomly chosen to be annotated by three human annotators. In addition to the 12 meetings we used one meeting to train the human annotators, which was not considered

[1] http://www.coli.uni-saarland.de/~thorsten/tnt/

[2] http://www-nlp.stanford.edu/software/tagger.shtml

[3] http://www.cs.jhu.edu/~brill/

in the evaluation. We used the MMAX2 annotation tool[4], which allows easy access to and manipulation of the tags.
The tagset was based on the Penn Treebank tagset (Santorini, 1990), which was also used for the Switchboard POS annotation (Godfrey et al., 1992). We added some tags to deal with phenomena that are of specific interest to the project in which this evaluation was carried out - specifically `RELP`, which is used to distinguish relative pronouns from wh-determiners (`WDT`). Additionally, we introduced one tag to deal with all interpunctuation signs – `INP`.
The 12 meetings were used to check inter-rater agreement for the three annotators. From the remaining 62 meetings 25 were annotated individually by one of the three annotators. Based on the manual annotation a gold standard was created by assigning a majority decision tag to each token in the meetings. This majority decision was manually corrected by a senior annotator. Inter-rater reliability was very high ($\kappa = .96$), showing that the automatically assigned tags can be manually corrected highly reliable. Therefore, we assume that the quality of the individually annotated meetings are of equally high quality as the gold standard.
Gathering the data took about two months, with three annotators working for about 240h in total. The costs were reasonable with about EUR 5000 total.
Nine meetings were used for evaluating the automatic annotation. Three meetings were taken from the gold standard data (*Test 1* in the following) and six from the individually annotated data (*Test 2*).

Data	TBL	TnT	Left3	Bidirect	Maj4
Test 1	11.3	11.2	11.1	11.1	10.5
Test 2	13.8	14.3	13.6	13.2	13.4

Table 1: Error Rates for automatically annotated data

Table 1 shows the results for the automatically tagged data. The results for *Test 1*, which is part of the gold standard data but not used for retraining are better than those for *Test 2*, which was also not used in retraining but was not part of the gold standard either. In this evaluation we did not consider tags that were unknown to the taggers, because they were introduced by our annotation scheme.

4. Retraining

Nine Meetings from the gold standard were used for training. Since they are spread across the whole corpus they contain a large variety of words and language used. Additionally, we had 15 meetings which were manually annotated. Retraining was done based on 6 different setups. These setups contained increasing amounts of data. Setup1 contained the data from the gold standard and consisted of 124,158 tokens. In every setup 3 meetings, all of them annotated individually by one of the three annotators, were added. The final setup contained 24 meetings and 282,686 tokens in total.
Our aim was to find a good trade-off between good annotation results and reasonable effort for the manual annotation. It is also important that for most taggers, the amount of time needed for retraining increased with increasing data set size. Additionally, the relationship between the amount of training data and the results is unknown. It is assumed that the more data is available, the better the results will get. The fastest to train was the TnT tagger, which took only a few minutes. The TBL and Stanford taggers took hours. Despite of this difference, TnT's results were comparable to the other three taggers.
The training parameters for each of the taggers remained unchanged throughout the different training setups. Especially the TBL Tagger would allow for several parameters to be set according to data set size and desired results. We left these parameters as they were suggested by the author. *Test 1* contains about 40K token and *Test 2* contains about 77K tokens.

	Error Rates in %					
	Set1	Set2	Set3	Set4	Set5	Set6
tokens K	124	162	197	221	253	283
	TnT					
Test 1	3.4	3.4	3.3	3.3	3.4	3.4
Test 2	5.4	5.1	4.9	4.6	4.5	4.5
	TBL					
Test 1	3.9	3.5	3.5	3.5	3.6	3.5
Test 2	8.4	5.5	5.0	4.7	4.4	4.4
	Left3					
Test 1	3.2	3.0	3.0	3.2	3.2	3.2
Test 2	5.2	4.7	4.5	4.3	4.2	4.1
	Bidirect					
Test 1	3.2	3.0	3.0	3.2	3.2	3.2
Test 2	5.2	4.7	4.5	4.3	4.2	4.1

Table 2: Average error rates for all taggers in each of the setups

Table 2 shows the results for all taggers after they have been trained on the manual data in different setups. The first part shows the results for TnT, after being trained on each of the setups. The results are very good and improve by about 1% in total. The gain in each step is rather small, the biggest is about 0.3%. In the last three steps the gain is very small and finally non-existent. The last noticeable step is from Set3 to Set4. This indicates that a training set size of between 197K and 221K tokens gives good tagging results with reasonable effort for manual data.
The second part shows the results for TBL. The results are similar to TnT but the gain for the various setups is higher. Especially from the first to the second setup in Test 2 the error rate decreases by 2.9%. The later steps are smaller and level out towards the end. Here, the last noticeable step is from Set4 to Set5. This indicates that TBL needs more training data than TnT.
The third and fourth parts should be considered together because the results are very similar if not identical. Again the first few steps are bigger than the last few steps. After Set4 the error rate does not decrease much. This suggests that the even-odd point of decreasing error rate and increasing training data size is somewhere between Set3 and Set4.
The improvement for all taggers presented here from the original tagging (Table 1) is in the range of $8 - 10\%$ for each tagger.
In addition, we were interested in whether the improvement can also be demonstrated in the majority decision or whether the majority decision even requires less training

[4] `http://mmax.eml-research.de`

data, while keeping the error rate low. Since the two Stanford Taggers had a similar performance, the majority over all four taggers would show very similar results to these taggers. Therefore, we considered only three taggers in the final evaluation. We removed the Bidirectional tagger from further consideration. This is motivated by the fact that this tagger takes longer to train and to tag and the results are very similar to those by Left3.

	Set1	Set2	Set3	Set4	Set5	Set6
Test 1	3.0	2.9	2.9	3.0	3.0	2.9
Test 2	5.1	4.7	4.4	4.1	4.0	3.9

Table 3: Average error rates for *Maj3* on each of the Setups

Table 3 shows the results for the majority decision based on three different taggers (*Maj3* in the following). As can be seen in Set4 the results are as good as for any of the single taggers in all setups. This is also the last noticeable step. Although Set5 and Set6 show the best overall results, the gain in Set5 and Set6 is not as remarkable as in the steps before. With this amount of training data *Maj3* outperforms all single taggers in all setups.

In general, one can observe that the gain throughout the setups is about 1%. TBL improves by about 4%, but this is mainly due to the difference between first and second setup (2.9%). Between the second and the last setup the difference is only 1.1%, which is the same as with the other taggers. *Maj3* gained slightly more (1.2%), but also outperforms the single taggers by 0.2%. Furthermore, the individual results are very close to those that have been reported for the taggers on text (see Section 2.). Maj3 achieves 97.1% on the best setup, which is close to the best tagger on text (97.2%).

5. Discussion

The work presented here, was done on transcribed speech from multiparty meetings, which so far has not been explored in detail. We used four common POS taggers to automatically annotate the transcripts. These results were manually corrected by human annotators, which is considerably faster than assigning POS tags from scratch. The results from the manual annotation were then used to retrain the POS taggers.

It turned out, that about 221K tokens are sufficient to get results that are comparable to those reported for the POS taggers applied to text. Redoing the majority decision improved the results on this amount of data by about 0.2%. Using the full amount of training data improved the best results by 0.3% also compared to the best results of the single taggers, which were achieved on texts.

It has been argued in the past, that in some cases retraining is not necessary to do POS tagging (Wermter & Hahn, 2004). The authors report an analysis of uni-, bi and trigrams of the data on which the taggers were trained (news texts) and of the data on which the taggers were tested (medical texts). They found that the n-grams were very similar. Following this analysis we compared the WSJ corpus with the ICSI corpus.

Table 4 shows the five most common Uni-, Bi- and Trigrams for Wall Street Journal (WSJ) and the Meeting Recorder Data (ICSI). The five most common unigrams for WSJ are `NN`, `INP`, `IN`, `NNP` and `DT`, whereas for ICSI they are `INP`, `PRP`, `DT`, `UH` and `NN`. `UH` is very rare in the WSJ corpus. These differences also appear in the analysis of bi- and trigram. For WSJ the dominant tags are `DT` and `NN` in various combinations and variations, whereas for ICSI the dominant tags are `UH`, `PRP` and `INP`, also in various combinations. For the Unigrams we also show at which position in the frequency table the tags that are most often in WSJ occur in ICSI and vice versa. The Bi- and Trigrams underline the difference between these two corpora. For the Unigrams three tags are shared in the five most frequent tags. In the Bigrams only one combination is left and for Trigrams none. It has to be noted that the combination of `CD+CD+CD` is an artefact of the data in ICSI, as most meetings start or finish with all speaker recording a sequence of numbers. The same accounts for the combination `$+CD+CD` in WSJ.

x-gram	Num	WSJ	%	ICSI	%
uni	1	NN	14.01	INP	19.00
	2	INP	11.24	PRP	9.20
	3	IN	10.51	DT	7.66
	4	NNP	9.83	UH	7.54
	5	DT	8.67	NN	7.18
	6			IN	7.14
	12	PRP	2.69	...	...
	19	...	...	NNP	1.09
	33	UH	0.01	...	...
bi	1	DT+NN	4.13	UH+INP	4.95
	2	NNP+NNP	3.68	INP+UH	4.93
	3	NN+IN	3.50	PRP+VBP	3.29
	4	IN+DT	3.46	INP+PRP	3.29
	5	JJ+NN	2.91	DT+NN	2.74
tri	1	JJ+NN+IN	1.21	INP+UH+INP	4.10
	2	$+CD+CD	0.81	INP+PRP+VBP	1.63
	3	DT+NN+NN	0.78	UH+INP+UH	1.46
	4	.+DT+NN	0.66	CD+CD+CD	1.39
	5	DT+NN+,	0.65	INP+RB+INP	0.01

Table 4: Differences between WSJ and ICSI Uni-, Bi- and Trigrams distribution

Table 5 shows those categories which benefit the most from retraining. Most other categories improve as well, but on a smaller scale. Only few categories do not improve at all or even achieve worse results after training. The categories in Table 5 can be characterized as either occurring rarely in the original training data (WSJ) as e.g. `UH`, `FW` and `PDT` or they form a very large group, as e.g. `NN`. Some categories achieved good results ($\geq 95\%$ correct) with the original POS taggers. Among those categories are `CC`, `CD`, `MD`, `NNS`, `PRP`, `PRP$`, `TO`, `VBP`, `VBZ` and `WRB`. These categories either belong to a fixed group of words (e.g. `CC`) or are ruled by certain (fixed) rules (e.g. `VBZ`).

Among the categories that have been unreliably tagged are particles (`RP`), which achieve a precision of about 80% and recall of about 72%, proper singular nouns (`NNP`) with a precision of about 86% and recall of about 90%, but also wh-determiner (`WDT`), which only achieve a precision of about 66% and a recall of about 52%. A more detailed discussion will be provided after further analysis of the data.

Tag	before	after
FW	52.1	85.2
JJ	78.6	89.2
NN	83.9	94.4
NNP	46.0	88.9
PDT	33.6	90.52
POS	33.9	84.5
RP	77.5	80.12
UH	56.6	99.0
VB	89.4	94.5
WDT	14.3	79.0
WP	86.2	94.2

Table 5: Categories which improve most through training

6. Conclusions

In Section 1 we presented some approaches to perform POS tagging with as little manual work as possible and approaches to improve the results of POS tagging in general (Clark et al., 2003; Collins, 2002; Nakagawa et al., 2002). Several remarks should be made here: First these works were based on text, like the Wall Street Journal portion of the Penn Treebank. Second, all of them are computationally very demanding. Finally, only the results presented by Clark et al. (2003) have been better than the results we report here, but on a considerably larger amount of data.

The results presented in our work give several methodological implications for further approaches to POS tagging of new domains. Human annotation is very expensive, in time and money. It is therefore desirable to explore, how much manually annotated data is necessary to get good/comparable results. Furthermore, the computational effort needed to achieve these results should be reasonable, too. Two main results were found in our work: first, a good trade-off point between the effort put into manually annotated data and the results of retraining POS taggers based on this data. Second, we applied a computationally cheap method for getting good results automatically.

In future work more sophisticated methods to merge the results of different taggers in order to improve the results could be applied, like e.g. those mentioned by van Halteren et al. (1998), who used a pairwise voting system or Zavrel & Daelemans (2000) who used a learning system based on the results from the POS taggers.

Data Availability. The manually annotated data as well as the automatically tagged data can be downloaded from the projects homepage `http://www.eml-research.de/nlp/diana-summ.php`.

Acknowledgments. This work has been supported by the DFG under grant STR 545/2-1 within the DIANA-Summ project and by the Klaus Tschira Foundation.

References

Brants, T. (2000). TnT - a statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, Seattle, Wash., 29 April – 4 May 2000, pp. 224–231.

Brill, E. (1994). Some advances in transformation based part-of-speech tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence,* Seattle, Wash., 1–4 August 1994, pp. 722–727.

Clark, S., J. R. Curran & M. Osborne (2003). Bootstrapping POS taggers using unlabelled data. In *Proceeings of the Seventh CoNLL conference held at HLT-NAACL 2003*, Edmonton, Alberta, Canada, *27 May –1 June, 2003*, pp. 49–55.

Collins, M. (2002). Discriminative training methods for Hidden Markov Models: Theory and experiments with Perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, Pa., 6–7 July 2002.

Godfrey, J. J., E. Holliman & J. McDaniel (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* San Francisco, Cal., USA, pp. 517–520.

Heeman, P. A. & J. F. Allen (1999). Speech repairs, intonational phrases, and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–571.

Janin, A., D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke & C. Wooters (2003). The ICSI meeting corpus. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* Hong Kong, China, 6–10 April 2003, pp. 364–367.

Marcus, M. P., B. Santorini & M. A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Nakagawa, T., T. Kudo & Y. Matsumoto (2002). Revision learning and its application to part-of-speech tagging. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Penn., 7–12 July 2002, pp. 497–504.

Santorini, B. (1990). *Part of Speech Tagging Guidelines for the Penn Treebank Project*. http://www.cis.upenn.edu/ treebank/home.html.

Toutanova, K., D. Klein, C. D. Manning & Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the Human Language Technolgy Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Alberta, Canada, 27 May –1 June, 2003, pp. 252–259.

Toutanova, K. & C. D. Manning (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong, pp. 63–70.

van Halteren, H., J. Zavrel & W. Daelemans (1998). Improving data driven wordclass tagging by system combination. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montréal, Québec, Canada, 10–14 August 1998, pp. 491–497.

Wermter, J. & U. Hahn (2004). Really, is medical sublanguage that different? Experimental counter-evidence from tagging medical and newspaper corpora. In *Medinfo '04 11th World Congress on Medical Informatics*, pp. 560–564.

Zavrel, J. & W. Daelemans (2000). Bootstrapping a tagged corpus through combination of existing heterogeneous taggers. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, May, 2000, pp. 1–4.

Zechner, K. (2001). *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*, (Ph.D. thesis). Language Technology Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA: School of Computer Science.