High-Performance Tagging on Medical Texts

Udo Hahn **Joachim Wermter**

Computerlinguistik, Friedrich-Schiller-Universität Jena
Fürstengraben 30, D-07743 Jena, Germany
hahn@coling.uni-freiburg.de

Abstract

We ran both Brill's rule-based tagger and TnT, a statistical tagger, with a default German newspaper-language model on a medical text corpus. Supplied with limited lexicon resources, TnT outperforms the Brill tagger with state-of-the-art performance figures (close to 97% accuracy). We then trained TnT on a large annotated medical text corpus, with a slightly extended tagset that captures certain medical language particularities, and achieved 98% tagging accuracy. Hence, statistical off-the-shelf POS taggers cannot only be immediately reused for medical NLP, but they also – when trained on medical corpora – achieve a higher performance level than for the newspaper genre.

1 Introduction

The application of language technology in the medical field, dubbed as medical language processing (MLP), is gaining rapid recognition (for a survey, cf. Friedman and Hripcsak (1999)). It is both important, because there is strong demand for all kinds of computer support for health care and clinical services, which aim at improving their quality and decreasing their costs, and challenging — given the miracles of medical sublanguage, the various text genres one encounters and the enormous breadth of expertise surfacing as medical terminology.

However, the development of human language technology for written language material has, up until now, almost exclusively focused on newswire or newspaper genres. This is most prominently evidenced by the PENN TREEBANK (Marcus et al., 1993). Its value as one of the most widely used language resources mainly derives from two features. First, it supplies everyday, non-specialist document sources, such as the *Wall Street Journal*, and, second, it contains value-added, *viz.* annotated, linguistic data. Since the understanding of newspaper material does not impose particular requirements on its reader, other than the mastery of general English and common-sense knowledge, it is easy for almost everybody to deal with. This is essential for the accomplishment of the second task, *viz.* the annotation and reuse of part-of-speech (POS) tags and parse trees, as the result of linguistic analysis. With the help of such resources, whole generations of state-of-the-art taggers, chunkers, grammar and lexicon learners have evolved.

The medical field poses new challenges. First, medical documents exhibit a large variety of structural features not encountered in newspaper documents (the genre problem), and, second, the understanding of medical language requires an enormous amount of a priori medical expertise (the domain problem). Hence, the question arises, how portable results are from the newspaper domain to the medical domain?

We will deal with these issues, focusing on the portability of taggers, from two perspectives. We first pick up off-the-shelf technology, in our case the rule-based Brill tagger (Brill, 1995) and the statistically-based TnT tagger (Brants, 2000), both trained on newspaper data, and run it on medical text data. One may wonder how the taggers trained on newspaper language perform with medical language. Furthermore, one may ask whether it is necessary (and, if so, costly) to retrain these taggers on a medical corpus, if one were at hand? These questions seem to be of particular importance, because the use of off-the-shelf language technology for MLP applications has recently been questioned (Campbell and Johnson, 2001). Answers will be given in Section 2.

Once a large annotated medical corpus becomes available, additional questions can be tackled. Will taggers, e.g., improve their performance substantially when trained on medical data, or is this more or less irrelevant? Also, if medical sublanguage particularities can already be identified on the level of POS co-occurrences, would it be a good idea to enhance newspaper-oriented, general-purpose tagsets with dedicated medical tags? Finally, does this extension have a bearing on the performance of tagging medical documents and, if so, to what extent? We will elaborate on these questions in Section 4.

2 Medical Tagging with Off-the-shelf Technology

For the first series of experiments, we chose two representatives of the currently prevailing data-driven tagging approaches, Brill's rule-based tagger (Brill, 1995) and TNT, a statistical tagger (Brants, 2000). As we are primarily concerned with German language input, for Brill's tagger, originally developed on English data, its German rule extension package was used. TNT, on the other hand, is based on a statistical model and therefore is basically language-independent. It implements the Viterbi algorithm for second-order Markov models (Brants, 2000), in which states of the model represent tags and the output represents words. The best POS tag for a given word is determined by the highest probability that it occurs with n previous tags. Tags for unknown words are assigned by a probabilistic suffix analysis; smoothing is done by linear interpolation.

2.1 Experiment 1: Medical Tagging with Standard Tagset Trained on NEGRA

The German default version of TNT was trained on NEGRA, the largest publicly available manually annotated German newspaper corpus (composed of 355,095 tokens and POS-tagged with the general-purpose STTS tagset; cf. (Skut et al., 1997)). The Brill tagger comes with an English default version also trained on general-purpose language corpora like the PENN TREEBANK (Marcus et al., 1993). In order to compare the performance of both taggers on German data, the Brill tagger was retrained on the German NEGRA newspaper corpus, with parameters recommended in the training manual.

In a second round, we set aside a subset of a newly developed German-language medical corpus (21,000 tokens, with 1800 sentences). We here refer to this text corpus as FRAMED$_{Stts}$ and describe its superset, FRAMED (Wermter and Hahn, 2004), in more depth in Section 4.1. Three human taggers, trained on the STTS tagset and on guidelines used for tagging the NEGRA corpus, annotated FRAMED$_{Stts}$ according to NEGRA standards. The interrater reliability for this part of the manual annotation was 96.7% (standard deviation: 0.6%), based on a random sample of 2000 tokens (10% of the evaluation corpus).

The performance of both taggers, TNT and Brill, with their NEGRA newspaper-trained parameterization was then measured on the FRAMED$_{Stts}$ corpus. In addition, since both TNT and Brill allow the inclusion of an external backup lexicon, their performance was also measured by plugging in two such medical backups.

2.2 Results from Medical Tagging with Standard Tagset Trained on NEGRA

We measured *tagging accuracy* by the ratio of the number of correct POS assignments to text tokens (as defined by the gold standard, *viz.* the manually annotated corpus) and the number of all POS assignments to text tokens from the test set. Table 1 reveals that the n-gram-based TNT tagger outperforms the rule-based Brill tagger on the FRAMED$_{Stts}$ medical corpus, both being trained on the NEGRA newspaper corpus. The inclusion of a small medical backup lexicon (composed of 171 entries which account for the most frequently falsely tagged tokens such as measure units, Latinate medical terms, abbreviations etc.) boosted TNT's performance to 96.7%, which is on a par with the state-of-the-art performance of taggers on newspaper texts. A much more comprehensive medical backup lexicon, which contained the first one plus the *German Specialist Lexicon*, a very large repository of domain-specific medical terms (totalling 95,969 entries), much to our surprise had almost no effect on improving the tagging results.

	TNT	BRILL
Default	95.2%	91.9%
+ Back-up Lexicon 1	96.7%	93.4%
+ Back-up Lexicon 2	96.8%	93.5%

Table 1: Tagging Accuracy (Training on NEGRA Newspaper Corpus; Evaluation on FRAMED$_{Stts}$ Medical Corpus)

The results for the German version of Brill's tagger, both its default version (91.9%) and the lexicon add-on (93.4%), are still considerably better than those of its default version reported by Campbell et al. (Campbell and Johnson, 2001) for English medical input (89.0%).

3 An Inquiry into Corpus Similarity

The fact that an n-gram-based statistical POS tagger like TNT, trained on newspaper and tested on medical language data, falls 1.5% short of state-of-the-art performance figures may at first come as a surprise. It has been observed by (Campbell and Johnson, 2001) and (Friedman and Hripcsak, 1999), however, that medical language shows *less* variation and complexity than general, newspaper-style language. Our second series of experiments, quantifying the grammatical differences/similarities between newspaper and medical language on the TNT-relevant POS n-gram level, may shed some explanatory light on the tagger's performance.

3.1 Experiment 2: Measuring Corpus Similarity

For this purpose, we collected a large medical document collection of mostly clinical texts (i.e., pathology, histology and surgery reports, discharge summaries). We refer to this collection (composed of 2480K tokens) as BigMed. Next, we randomly split BigMed into six subsamples of Negra size (355K tokens). This was meant to ensure a statistically sound comparability and to break up the medical subgenres. The same procedure was repeated for a collection of German newspaper and newswire texts collected from the Web. All twelve samples (six medical ones, henceforth called Med, and six newspaper ones, henceforth called News, also composed of 2480K tokens to ease partitioning) were then automatically tagged by TnT based on its newspaper-trained parameterization.

Since Negra is the newspaper corpus on which the default version of TnT was trained, its statistical comparison with Med should elucidate the tagger's performance on medical texts without changing the training environment. Moreover, a parallel comparison with other newspaper texts (News) may help in further balancing these results. Because TnT is a Markovian tagger based on tri-, bi- and unigram POS sequences, the statistics were based on the POS n-gram sequences in the different corpora. For this purpose, we extracted all POS trigram, bigram and unigram type sequences from Negra, Med, and News. Their numbers are reported in Table 2 (see rows 1, 4 and 7). We then generated a distribution of these types based on three ranges of occurrence frequencies. The results are reported in Table 3.

We then determined how many POS n-gram types were common between Negra and Med and common between Negra and News (see Table 2, rows 2, 5 and 8). Each of these common POS n-gram types was subjected to a χ^2 test in order to measure whether their common occurrence in both corpora was just random (null hypothesis) or whether that particular n-gram was indicative of the similarity between the two corpora (i.e., between Negra and Med, on the one hand, and between Negra and News, on the other hand). This interpretation of χ^2 statistics has already been evaluated against other corpus similarity measures and was shown to perform best (Kilgarriff, 2001), assuming a non-normal distribution (cf. also Table 3). The χ^2 metric sums the differences between observed and expected values in all squares of the table and scales them by the magnitude of the expected values. The number of all common significant POS n-grams (i.e., those whose critical values are greater than 3.841 for a probability level of $\alpha = 0.05$) is indicative of the magnitude of corpus similarity. These results are reported in Table 2 (see rows 3, 6 and 9).

	Negra	Med	Negra	News
POS trigram types	13,045	9,232.9 (217.5)	13,045	13,709.2 (86.8)
common POS trigram types:	7,130.3 (144.2)		9,992.0 (33.6)	
ratio (in %)	54.7 (1.1)	77.2 (0.4)	76.6 (0.3)	72.9 (0.3)
χ^2 significant common POS trigram types	2,793.8 (34.4) ratio: 41.7% (1.3)		1,202.0 (29.6) ratio: 12.1% (0.3)	
POS bigram types	1,441	1,169.0 (20.7)	1,441	1,441.8 (14.8)
common POS bigram types:	1,076.5 (14.3)		1,270.8 (9.3)	
ratio (in %)	76.4 (1.0)	92.0 (0.5)	88.2 (0.6)	88.1 (0.4)
χ^2 significant common POS bigram types	689.9 (5.5) ratio: 64.2% (0.9)		386.5 (12.2) ratio: 30.4% (0.9)	
POS unigram types	55	52.7 (0.5)	55	55.0 (0.5)
common POS unigram types	51.3 (0.5)		53.7 (0.5)	
χ^2 significant common POS unigram types	44.7 (0.8) ratio: 87.0% (1.5)		36.5 (2.4) ratio: 68.1% (4.9)	

Table 2: POS n-gram and χ^2 Comparsions between Negra-Med and Negra-News (deviation of means of six Med and six News samples in parentheses)

3.2 Results from Measuring Corpus Similarity

As shown in Table 2 (rows 1, 4 and 7), the number of unique POS n-gram types was considerably lower in Med. Compared with Negra, Med had 29% less trigram types, 19% less bigram types and 4% less unigram types (i.e., POS tags), whereas News even had slightly more types at all n-gram levels. This much lower number of Med POS trigram and

		POS n-gram types appearing		
		< 10 times	10-1000 times	≥ 1000 times
tri-grams	Negra	9402	3610	33
	Med	6571.2 (153)	2610.5 (66.5)	51.2 (0.8)
	News	9972.5 (69)	3698.7 (31.6)	38 (0.6)
bi-grams	Negra	618	744	79
	Med	503.5 (18.2)	590.5 (16.3)	75 (1.9)
	News	598.8 (14.)	762.2 (5.3)	80.8 (.6)
uni-grams	Negra	4	18	33
	Med	4.3 (0.8)	21.3 (0.8)	27 (0.6)
	News	2.8 (1.0)	17.5 (0.8)	34.7 (0.5)

Table 3: Three-part Distribution of POS n-gram Types in Negra, Med and News

	n-grams common to Negra and Med			n-grams common to Negra and News		
Top 5 ranked POS trigrams	ADJD ADJA NN	3552.0	(399.4)	FM FM FM	772.7	(356.5)
	ADJA ADJA NN	2811.1	(262.7)	$, ADJA NN	176.8	(10.5)
	ADJA NN $.	1740.0	(175.7)	NN $. $(	172.4	(20.4)
	ADJA NN ART	1471.7	(145.9)	VVINF $. $(	169.7	(16.6)
	ADJA NN KON	1162.6	(73.3)	$. $(ART	148.9	(12.2)
Top 5 ranked POS bigrams	ADJA NN	5854.9	(454.5)	FM FM	869.9	(470.1)
	ADJD ADJA	4861.8	(577.5)	$. $(	831.3	(71.1)
	ADJA ADJA	3355.6	(290.6)	$. XY	407.2	(5.6)
	NE NE	2249.9	(99.4)	$. PPER	245.1	(21.4)
	APPR NE	1884.6	(111.5)	NN $(	221.9	(27.5)
Top 5 ranked POS unigrams	ADJA	10632.9	(946.5)	$(	992.5	(72.6)
	ADJD	5479.5	(439.2)	FM	953.2	(450.8)
	NE	5211.8	(216.7)	PPER	365.1	(17.1)
	PPER	2201.3	(97.8)	XY	329.6	(39.9)
	$(	1936.7	(170.0)	NE	127.7	(33.8)

Table 4: χ^2 Top 5 Ranked POS Trigrams, Bigrams and Unigrams Common to Negra and Med and to Negra and News (standard deviation of means of occurrence frequencies in parentheses)

bigram types is also reflected in the three-part distribution in Table 3: The number of POS trigrams occurring less than ten times is almost one third less in Med than in Negra or in News; similarly, but less pronounced, this can be observed for POS bigrams. On the other hand, the number of trigram types occurring more than 1000 times is even higher for Med, and the number of bigram and unigram types is about the same when scaled against the total number of types. This indicates a rather high POS trigram and bigram type dispersion in newspaper corpora, whereas medical narratives appear to be more homogeneous.

Table 2 (rows 2, 5 and 7) indicates that the number of POS trigram and bigram types common to both corpora was much smaller for the Negra-Med comparison than it was for Negra-News. In other words, more of the Negra POS n-gram types appeared in the News corpus as well, whereas far less showed up in the Med corpus. At this level of comparison, sublanguage differences clearly show up. If, however, compared with the total number of POS n-gram types in each corpus, the common ones cover much more of the Med corpus than of the Negra corpus. The coverage for Negra and News is about the same.

The number of common POS n-gram types that are χ^2 significant (Table 2: rows 3, 6, and 9) shows the magnitude of corpus similarity. For the common trigram types, it was almost four times higher in the Negra-Med comparison than for Negra-News; for the common bigram types it was more than twice as high, and for the unigram types 20% higher.

Finally, table 4 shows that the top-ranked POS trigrams, bigrams and unigrams common to Negra and Med (columns 2 to 4) exhibit a strikingly different χ^2 magnitude compared to those common to Negra and News (columns 5 to 7). This means that, in regard to their top POS n-grams, Negra and Med are highly similar, whereas Negra and News are less so. Interestingly, for each n-gram the top 5 ranks remain unchanged across all six Negra-Med comparisons, whereas they have a different ranking in almost each of the six Negra-News comparisons. It seems as though the most characteristic similarities between medical sublanguage and newspaper language are highly consistent and predictable, whereas the intra-newspaper comparison shows weak and inconsistent similarities.

4 Tagging with Medical Resources

4.1 FraMed, an Annotated Medical Corpus

FraMed, the FReiburg Annotated MEDical corpus (Wermter and Hahn, 2004), combines a variety of relevant medical text genres focusing on clinical reports. The clinical text genres cover discharge summaries, pathology, histology and surgery reports. The non-clinical ones consist of medical expert texts (from a medical textbook) and health care consumer texts taken from the Web. It has already been mentioned that medical language, as used in these clinical documents, has some unique properties not found in newspaper genres. Among these features are the use of Latin and Greek terminology (sometimes also mixed with the host language, here German), various *ad hoc* forms for abbreviations and acronyms, a variety of (sometimes idiosyncratically used) measure units, enumerations, and some others. These may not be marginal sublanguage properties and thus may have an impact on the quality of tagging procedures. In order to test this assumption, we enhanced the Negra-rooted Stts tagset with three dedicated tags which capture ubiquitous lexical properties of medical texts not

Training Size	Negra	FraMed	Negra	FraMed	Negra	FraMed	Negra	FraMed
	% unknown words		accuracy, unknown words only		accuracy, known words only		overall accuracy	
5,000	40.3 (1.4)	40.8 (3.3)	74.9 (2.5)	81.1 (2.5)	96.3(0.4)	97.8 (0.7)	87.7 (1.2)	91.0 (1.5)
10,000	33.9 (0.6)	33.5 (3.2)	79.3 (1.2)	85.9 (2.0)	96.8 (0.3)	97.8 (0.4)	90.9 (0.5)	93.7 (1.1)
20,000	28.6 (1.0)	26.1 (2.2)	82.9 (1.1)	88.9 (1.6)	97.1 (0.3)	98.2 (0.2)	93.0 (0.3)	95.9 (0.6)
30,000	25.2 (1.0)	21.1 (1.6)	84.4 (1.1)	90.2 (1.2)	97.3 (0.4)	98.3 (0.2)	94.0 (0.3)	96.6 (0.4)
40,000	23.1 (0.9)	18.3 (1.6)	85.1 (1.1)	91.7 (1.7)	97.3 (0.2)	98.6 (0.3)	94.6 (0.4)	97.3 (0.5)
50,000	21.6 (1.0)	16.7 (1.8)	85.8 (1.2)	92.0 (1.8)	97.4 (0.2)	98.7 (0.3)	94.9 (0.4)	97.6 (0.5)
60,000	20.2 (0.9)	15.3 (1.8)	86.1 (1.3)	92.4 (1.7)	97.5 (0.2)	98.7 (0.3)	95.2 (0.4)	97.7 (0.5)
70,000	19.2 (1.0)	14.5 (1.9)	86.4 (1.7)	92.4 (2.0)	97.5 (0.3)	98.6 (0.4)	95.4 (0.4)	97.7 (0.7)
80,000	18.5 (0.9)	13.6 (1.6)	86.9 (1.4)	93.2 (2.1)	97.5 (0.2)	98.8 (0.3)	95.6 (0.4)	98.0 (0.5)
90,000	17.9 (1.3)	12.5 (1.7)	86.9 (1.3)	93.0 (1.9)	97.6 (0.3)	98.7 (0.3)	**95.7** (0.3)	**98.0** (0.4)

Table 5: Averaged Learning Curve Values for Different Training Sizes (standard deviation in parentheses)

covered by this general-purpose tagset, thus yielding the Stts-Med tagset.[1] Our three student annotators then annotated the FraMed medical corpus with the extended Stts-Med tagset. The mean of the inter-annotator consistency of this annotation effort was 98.4% (with a standard deviation of 0.6).

A look at the frequency ranking of the dedicated medical tags shows that they bear some relevance in annotating medical corpora. Out of the 54 tag types occurring in the FraMed corpus, ENUM is ranked 14, LATIN is ranked 19, and FDSREF is ranked 33. In terms of absolute frequencies, all three additional tags account for 1613 (out of 100,141) tag tokens (ENUM: 866, LATIN: 560, FDSREF: 187). To test the overall impact of these three additional tags, we ran the default Negra-newspaper-based TnT on our FraMed medical corpus and compared the resulting Stts tag assignments with those from the extended Stts-Med tagset. The additional tags accounted for only 24% of the differences between the two assignments (1613/6685). Hence, their introduction, by no means, fully explains any improved tagging results (compared with the reduced newspaper tagset). The other sublanguage properties mentioned above (e.g., abbreviations, acronyms, measure units etc.) are already covered by the original tagset.

4.2 Experiment 3: Re-Training TnT on FraMed

In a third series of experiments, we compared TnT's performance with respect to the general newspaper language and the medical sublanguage. For this purpose, the tagger was newly trained and tested on a random sample (100,198 tokens) of the Negra newspaper corpus with the standard Stts tagset, and, in parallel, re-trained and tested on the FraMed medical corpus using Stts-Med, the extended medical tagset.

For this evaluation, we used learning curve values (see Table 5) that indicate the tagging performance when using training corpora of different sizes. Our experiments started with 5,000 tokens and ranged to the size of the entire corpus (minus the test set). At each size increment point, the overall accuracy, as well as the accuracies for known and unknown words were measured, while also considering the percentage of unknown words.

The tests were performed on random partitions of the corpora that use up to 90% as training set (depending on the training size) and 10% as test set. In this way, the test data was guaranteed to be unseen during training. This process was repeated ten times, each time using a different 10% as the test set, and the single outcomes were then averaged.

4.3 Results from Medical Tagging with Medical Resources

Table 5 (columns 4-9) reveals that the FraMed-trained TnT tagger outperforms the Negra-trained one at all training points and across all types of accuracies we measured. Trained with the largest possible training size (*viz.* 90,000 tokens), the tagger's overall accuracy for its FraMed parametrization scores 98.0%, compared to 95.7% for its Negra parametrization. The performance differences between FraMed and Negra range between 2.3 (at training points 90,000 and 70,000) and 3.3 percentage points (at training point 5,000). The tagging accuracy for known tokens is higher for both FraMed and Negra (with 98.7% and 97.6%, respectively, at training point 90,000). The differences here are less pronounced, ranging from 1.0 to 1.3 percentage points.

By far the largest performance difference can be observed with respect to the tagging accuracy for unknown words (cf. Table 5 (columns 4 and 5)), ranging from 5.8 (at training point 30,000) to 6.6 percentage points (at training points 10,000 and 40,000). The FraMed-trained tagger scores above 90% in seven out of ten points and never falls be-

[1] The three tags are 'ENUM' (all sorts of enumerations), 'LATIN' (Latin forms in medical terms), and 'FDSREF' (reference patterns related to formal document structure).

low 80%. The NEGRA-based tagger, on the other hand, remains considerably below 90% at all points, and even falls below 80% at the first two training points. This performance difference is clearly one factor which contributes to the FRAMED tagger's superior results. The difference in the average percentage of unknown words is the other dimension where both environments diverge (cf. Table 5, columns 2 and 3). Whereas the percentage of unknown words starts out to be equally high for lowest training sizes (5,000 and 10,000), this rate drops much faster for the FRAMED-trained tagger. At the highest possible training point, only 12.5% of the words are unknown, compared to still almost 18% unknown to the NEGRA-trained tagger, resulting in a 5.4 percentage point difference. Thus, both the high tagging accuracy for unknown words and their lower rate, in the first place, seem to be key for the superior performance of the FRAMED-trained TNT tagger.

5 Discussion

Campbell and Johnson (2001) have argued that general-purpose off-the-shelf NLP tools are not readily portable and extensible to the analysis of medical texts. By evaluating the English version of Brill's rule-based tagger (Brill, 1995), they conclude that taggers trained on general-purpose language resources, such as newspaper corpora, are not suited to medical narratives but rather need timely and costly retraining on manually tagged medical corpora. Interestingly though, it has also been observed (Friedman and Hripcsak, 1999; Campbell and Johnson, 2001) that medical language shows less variation and complexity than general, newspaper-style language, thus exhibiting typical properties of a sublanguage. Setting aside the difference in vocabulary between medical and non-medical domains, the degradation in performance of general-language off-the-shelf NLP tools for MLP applications then seems counter-intuitive. Our first and second series of experiments were meant to explain this puzzling state of affairs.

The results of these experiments shed a different light on the portability and extensibility of off-the-shelf NLP tools for the analysis of medical narratives as was hypothesized by Campbell and Johnson (2001). A statistical POS tagger like TNT, which is trained on general-purpose language by default, only falls 1.5% short of the state-of-the-art performance in a medical environment. An easy-to-set-up medical backup lexicon eliminates this difference entirely. It appears that it is the underlying language model which determines whether a POS tagger is more or less suited to be portable to the medical domain, not the surface characteristics of medical sublanguage. Moreover, lexical backup facilities show up as a significant asset to MLP. Much to our surprise, a full-scale, carefully maintained lexicon did not substantially improve the tagger's performance in comparison with a heuristically assembled brief list of the most common tagging mistakes.

A reason for the statistical tagger's outperformance may be derived from our comparative corpus statistics, which was the focus of our second series of experiments. Concerning POS n-grams, the data points to a less varied and less complex grammar of medical sublanguage(s). Not only is the number of POS n-gram types much lower for medical narratives than for general-language newspaper texts, but the distribution also favors high-occurring (more than 1000 times) types in MED. Another indicator of a simpler POS n-gram grammar in medical narratives is the fact that the absolute number of POS n-gram types common to NEGRA and MED is much lower than for NEGRA and NEWS. Scaled against the total number of types in MED, however, the common ones cover a bigger part of the medical narratives, whereas they cover less of NEGRA. For POS trigrams, half of NEGRA is congruent with three quarters of MED; for POS bigrams three quarters of NEGRA is congruent with nine tenths of MED.

Common POS n-grams that are χ^2 significant indicate that two corpora are similar with respect to them. Their number was significantly higher for the NEGRA-MED comparison than for NEGRA-NEWS. Hence, the congruency of a high proportion of POS n-gram types between NEGRA and MED is not accidental. At the POS n-gram type level, this shows a higher degree of similarity between NEGRA and medical narratives than between NEGRA and other newspaper texts. Furthermore, the high χ^2 numbers for the top ranked POS n-grams indicate that they are especially characteristic of the NEGRA-MED similarity. Eight of the top-ranked trigrams and bigrams can be identified as parts of a noun phrase. All of them contain a prenominal adjective (ADJA in Table 4), six a common noun (NN in Table 4). The prenominal adjective is by far the most characteristic POS unigram for medical-newspaper inter-language similarity. None of these observations hold for newspaper intra-language similarity.

Our third series of experiments showed that Markovian taggers like TNT improve their performance substantially when trained on medical data. Indeed, we were able to achieve a performance boost which goes beyond current state-of-the-art numbers. This seems to be even more notable inasmuch as the tagger's retraining was done on a comparatively small-sized corpus (90,000 tokens).

These experiments suggest two explanations. First, annotating medical texts with a medically enhanced tagset took care of medical sublanguage properties not covered by general-purpose tagsets. Second, several tagging experiments on newspaper language, whether statistical (Ratnaparkhi, 1996; Brants, 2000) or rule-based (Brill, 1995), report that the tagging accuracy for unknown words is much lower than the overall accuracy.[2] Thus, the lower percentage of unknown words in medical texts seems to be a sublanguage feature beneficial to POS taggers, whereas the higher proportion of unknown words in newspaper language seems to be a prominent source of tagging errors. This is witnessed by the tagging accuracy for unknown words, which is much higher for the FRAMED-trained tagger than for the newspaper-trained one. For the medical tagger, there is only a 5 percentage point difference between overall and unknown word accuracy at training point 90,000, whereas, for the newspaper tagger, this difference amounts to 8.8 percentage points. This may be interrelated with another property of sublanguages, *viz.* their lower number of word types: At each training point, the lexicon of the FRAMED tagger is 20 percentage points smaller than that of the newspaper tagger. TNT's handling of unknown words relies on the probability distribution for a particular (formal) suffix of some fixed length (cf. Brants (2000)). Thus, guessing an unknown word's category is easier on a small-sized tagger lexicon, because there are less choices for the POS category of a word with a paricular suffix.

Only recently has the accuracy of data-driven POS taggers moved beyond the the '97% barrier' (derived from newspaper corpora). This was partly achieved by computationally more expensive models than TNT's efficienct unidirectional Markovian one. For example, Giménez and Màrquez (2003) report an accuracy of 97.13% for their SVM-based power tagger. The best automatically learned POS-tagging result reported so far (97.24%) is Toutanova et al. (2003)'s feature-based cyclic dependency network tagger. Although reaching the 98% accuracy level constitutes a breakthrough, it is of course conditioned by the medical sublanguage we are working with. Still, the application of language technologies in certain sublanguage domains like medicine, and more recently, genomics and biology, is gaining rapid importance, and thus, our results also have to be considered from this perspective.

6 Conclusions

We collected experimental evidence, contrary to recent claims (Campbell and Johnson, 2001), that off-the-shelf NLP tools can be applied to MLP in a straightforward way. We explain this finding with statistically significant POS n-gram type overlaps of newspaper language and medical sublanguage, which has not been recognized before.

To the best of our knowledge, this is the first tagging study that reaches a 98% accuracy level for a data-driven tagger (which must be distinguished from linguistically backuped taggers which come with 'heavy' parsing machinery (Samuelsson and Voutilainen, 1997)). Still, we deal with a specialized sublanguage simpler in structure compared with newspaper language, although we kept it diverse through the various text genres.

Acknowledgements. We would like to thank our students, Inka Benthin, Lucas Champollion and Caspar Hasenclever, for their excellent work as human taggers. This work was partly supported by DFG grant KL 640/5-1.

References

T. Brants. 2000. TNT: A statistical part-of-speech tagger. In *Proc. ANLP 2000*, pages 224–231.

E. Brill. 1995. Transformation-based error-driven learning and natural language processing. *Computational Linguistics*, 21(4):543–565.

D. A. Campbell and S. B. Johnson. 2001. Comparing syntactic complexity in medical and non-medical corpora. In *Proc. AMIA 2001*, pages 90–94.

C. Friedman and G. Hripcsak. 1999. Natural language processing and its future in medicine. *Academic Medicine*, 74(8):890–895.

J. Giménez and L. Màrquez. 2003. Fast and accurate part-of-speech tagging: The SVM approach revisited. In *Proc. of the Intl. Conf. on RANLP 2003*.

A. Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: The PENN TREEBANK. *Computational Linguistics*, 19(2):313–330.

A. Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proc. EMNLP'96*, pages 133–142.

C. Samuelsson and A. Voutilainen. 1997. Comparing a linguistic and a stochastic tagger. In *Proc. ACL'97/EACL'97*, pages 246–253.

W. Skut, B. Krenn, T. Brants, and H. Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proc. ANLP 1997*, pages 88–95.

K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc of HLT-NAACL 2003*, pages 252–259.

Joachim Wermter and Udo Hahn. 2004. An annotated German-language medical text corpus as language resource. In *Proc 4th Intl LREC Conf.* Lisbon, Portugal.

[2] These authors report on differences between 7.7 and 11.5 percentage points.