A Domain-adapted Dependency Parser for German Clinical Text

Elif Kara[⋆], **Tatjana Zeen**[⋆], **Aleksandra Gabryszak**[⋆], **Klemens Budde**[◇], **Danilo Schmidt**[◇], **Roland Roller**[⋆]
[⋆]Language Technology Lab, DFKI Berlin, Germany
{firstname.surname}@dfki.de
[◇]Charité – Universitätsmedizin Berlin, Germany
{firstname.surname}@charite.de

Abstract

In this work, we present a syntactic parser specialized for German clinical data. Our model, trained on a small gold standard nephrological dataset, outperforms the default German model of Stanford CoreNLP in parsing nephrology documents in respect to LAS (74.64 vs. 42.15). Moreover, re-training the default model via domain-adaptation to nephrology leads to further improvements on nephrology data (78.96). We also show that our model performs well on fictitious clinical data from other subdomains (69.69).

1 Introduction

The demand for Natural Language Processing (NLP) in the clinical domain is rapidly increasing due to growing interest in clinical information systems and their potential to enhance clinical activities. Clinical text data exists in abundance in unstructured format (patient records, hand-written notes, etc.) that, once structured by NLP solutions, could be used to improve interaction between patients and medical staff, to aid the personalization of treatment and to automate risk stratification. Further, NLP can aid the detection of adverse drug events, as well as the detection and prediction of healthcare associated infections (Dalianis, 2018).

A multitude of NLP tools were developed to process English clinical text, such as Savova et al. (2010) or Aronson and Lang (2010), but, thus far, few for German (see Section 2). The primary reason for this is the lack of existing clinical text in German that can be accessed for research, due to strict laws revolving around issues of ethics, privacy and safety (Starlinger et al., 2016; Suster et al., 2017; Lohr et al., 2018).

Added to the juridical constraints, clinical language is by itself difficult to process and, thus, requires specialized solutions. It tends to be driven by time pressure and the need for minuteness, often deviating from stylistic, grammatical and orthographic conventions.

Some features of clinical language problematic for machine-readability are (Patrick and Nguyen, 2011; Roller et al., 2016; Savkov et al., 2016; Dalianis, 2018):

Domain-dependence: Extensive use of Greek- and Latin-rooted terminology, e.g. *Appendektomie* (‘appendectomy’), *thorakal* (‘thoracic’).

Complexity: Complex syntactic embeddings, e.g. *In Anbetracht der initial bestehenden Entzündungskonstellation haben wir antibiotisch mit Levofloxacin 500 mg 1-0-1 über 10 Tage behandelt, was sich im Nachhinein nach dem bakteriologischen Resistenzprofil als treffsicher erwies.* (‘Given the initial inflammatory constellation, we treated antibiotically with Levofloxacin 500 mg 1-0-1 for 10 days, which turned out to be accurate according to the bacteriological resistance profile.’)

Reduction: Ellipses of auxiliary and copula verbs as well as sentence boundaries, e.g. *Geht gut.* (‘Goes well.’), *Ödeme rückläufig* (‘Edema declining’).

This work focuses on syntactic dependency parsing of clinical text in German. Syntactic dependency relations provide insight into the grammatical structure of a sentence and are often used as input for NLP applications.

We use the Stanford Parser (SP) (Chen and Manning, 2014), a domain-independent syntactic neural network dependency parser from the Stanford CoreNLP pipeline (Manning et al., 2014), and examine its accuracy on highly specialized German-language data from the clinical domain. The result is rather poor when using the German default model. In order to improve it, two experiments were conducted:

1) We provide the SP with gold standard tokenization and PoS-tags and (re-)train two new models on the gold standard annotation. Based on the results, we establish that a model trained on a small nephrological dataset already outperforms Stanford's own model when parsing clinical text. However, our best-performing model is a blend of Stanford's data model, re-trained with the model described here. From this, we take that re-training models that were initially trained on large-scale datasets of mixed domains with in-domain data (of smaller scale) is beneficial.

2) We further test the potential of our best-performing model on additional documents from varying clinical subdomains, with promising results. These are fictitious as opposed to the previous test set, and thus, can be published for further use.

In this work, we demonstrate how existing NLP models can be refined to process domain- and language-specific data. The paper is structured as follows: In Section 2, we present a selection of previous research. Next, we present our dataset in Section 3, followed by the procedure of our experiments in Section 4. Finally, we sum up our findings in Section 5.

2 Related Work

The Stanford Parser is a popular language-agnostic syntactic statistical parser (Zou et al., 2015; Ma et al., 2015; Chaplot et al., 2015) that can be trained on any language. As part of the unified Universal Dependencies (UD) framework (Nivre et al., 2016), models for various languages, including German, are available. The German model was trained on the UD Treebank for German – a large dataset of heterogeneous nature (website crawls). German uses all 17 universal Part of Speech (PoS) categories and most of the 40 dependencies due to its morpho-syntactic complexity. For a complete list of language-specific relations, please refer to the UD website.

It is tried and tested that the source domain trained on the parser needs to match the domain of the data to be parsed (McClosky et al., 2010). As a consequence, existing parsers tend to handle domain-specific data poorly. With the increased interest in biomedical NLP in recent years, there have been a number of shared domain-adaptation initiatives, whereby pre-built, domain-independent parser models are customized and used for retraining (McClosky et al., 2010; Jiang et al., 2015; Skeppstedt et al., 2014; Rimell and Clark, 2009).

The Charniak Parser (Charniak, 2000) was enriched with data from a variety of domains, including abstracts from PubMed, a corpus of biomedical and life sciences journal literature (McClosky and Charniak, 2008). Similarity measures between target and source domains were fed into a regression model that analyzes the effect of domain dissimilarities and, subsequently, selects the input that maximizes the regression function. This multi-source approach to domain adaptation improves the parse quality of texts from all source domains, compared to non-specific domains. The system learned quantitative measures of domain differences and their effects on parsing accuracy, so that it proposes linear combinations of the source models.

Jiang et al. (2015) compared the SP, Charniak and Bikel (Bikel, 2004) parsers on clinical text before and after domain-adaptation and found that domain-adapted re-training is an effective measure and that the SP outperformed the others.

A different approach was taken by Skeppstedt et al. (2014) via a direct comparison between clinical and standard Swedish text parses using a domain-independent Swedish parser. Based on the manual analysis and the identification of eight PoS-related error types, pre-processing rules were formulated and fed back to the tool, resulting in improved parsing. Likewise, Rimell and Clark (2009) report that simply retraining the PoS-tagger on biomedical data leads to significant improvements in parsing performance. This indicates the importance of a relevant PoS-tagset applied consistently.

Contrasting these shared efforts, there is – to date – not a single dependency processing tool available for use on clinical German. As already mentioned in the Introduction, the lack of shared resources is a persisting obstacle for clinical NLP in Germany (Starlinger et al., 2016). However, progress can only be made by the sharing of models (Hellrich et al., 2015; Starlinger et al., 2016). Furthermore, Lohr et al. (2018) propose testing models on synthetically generated medical corpora, which can be made public without infringing on data privacy laws.

There are currently a total of ten corpora in clinical German – all inaccessible (Lohr et al., 2018) – the first and most well-known being the FraMed corpus (Wermter and Hahn, 2004; Hellrich et al., 2015), which contains authentic de-identified medical data and is PoS-tagged using a variant of the Stuttgart-Tübingen-TagSet (STTS). The corpus was used for generating in-domain machine learning models for different tasks, e.g. sentence splitting, PoS-tagging and tokenization (Faessler et al., 2014; Hahn et al., 2016). However, it is unaccessible for research.

Hellrich et al. (2015) tested JCoRe, a newly developed NLP pipeline, on FraMed with respect to PoS-tagging and compared the results to the OpenNLP (Ingersoll et al., 2013) and the Stanford PoS-tagger (all trained on FraMed). JCoRe outperformed alternative components of OpenNLP and Stanford.

3 Data

This section presents the data used for training and evaluation of our dependency tree parser.

3.1 Nephrological Dependency Corpus

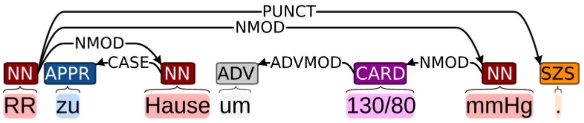
A small gold standard corpus of nephrological text documents, including PoS and dependency annotations, serves as the reference point for our experiments. It is henceforth referred to as *Nephro_Gold*. The dataset comprising our gold standard corpus, presented in Table 1, consists of original de-identified German nephrology records – clinical notes and discharge summaries. While the first are characterized by poor syntactic structure, misspellings as well as extensive use of abbreviations and acronyms, the discharge summaries are embedded in a letter format and comprise well-formed sentences as well as detailed lists of medical diagnostics and procedures.

	clinical notes	**dis. summaries**
number of files	44	11
total word count	3,154	10,436
avg. words (std.)	71.7 (75.2)	948.7 (333.3)

Table 1: Annotated files comprising *Nephro_Gold*

The syntactic annotation was carried out by two postgraduate students of linguistics in their final year, in roughly 150 hours each, using the UD-tagset. The PoS-annotation had been carried out manually in previous work (Seiffe, 2018), using the STTS-tagset. Four clinical notes and one discharge summary were annotated by both annotators, initially scoring an Inter-Annotator Agreement (IAA) of 0.83, according to Cohen's kappa. The relatively low IAA can be attributed to the linguistic challenges outlined in Section 1. The annotators reviewed the cases of disagreement and identified a number of systematic differences, such as the annotation of coordinated compound words with a preceding truncated element or discrepancies in combining nouns with other tokens in specific, complex syntactic structures. Some of these cases may have been resolved with medical knowledge that the annotators were lacking. With an adaptation of the annotation guidelines and a subsequent re-annotation, the IAA was increased to a kappa score of 0.9578.

An exemplary sentence parse from our clinical data is presented in Figure 1.



Translation: RR (Riva-Rocci – 'blood pressure') at home about 130/80 mmHg.

Figure 1: Sentence with syntactic dependencies

3.2 Additional Evaluation Data

In addition to the previously presented clinical dependency corpus, we use a collection of fictitious clinical notes and discharge summaries to further evaluate parsing accuracy. These were written by students familiar with the nephrology data. Thus, they may not be authentic from a medical perspective but they maintain the linguistic characteristics and vocabulary of genuine documents.

In order to enrich the corpus semantically and provide a more realistic setting, additional fictitious discharge summaries in the subdomains of *Surgery*, *Cardiac Rehabilitation*, *Discharge*, *Internal Medicine* and *Relocation* were created by a medical student using the template-based online tool *Arztbriefmanager* [1], which we refer to as ABM. Table 2 provides an overview of the fictitious data. Within our experiment this dataset will be automatically parsed before being manually annotated.

	clinical notes	**ABM**
number of files	30	5
total word count	1,233	1,991
avg. words (std.)	41.1 (12.0)	398.2 (226.6)

Table 2: Fictitious data for extended experiments

[1] http://www.arztbriefmanager.de/

4 Experiments & Evaluation

The experiments are carried out using the Stanford Parser (SP) (Chen and Manning, 2014). We applied a 10-fold cross-validation on the *Nephro_Gold* dataset. Clinical notes and discharge summaries are equally assigned to the different folds. Within each validation step 80% of the data is used for training, 10% for development and 10% for testing.

The Labeled Attachment Score (LAS) was applied as our accuracy metric. A given dependency is therefore scored as correct only if both the syntactic head and the label are identical.

4.1 Baseline: Stanford Out-of-the-box

First of all, we would like to determine the performance of Stanford CoreNLP on German clinical data using the pre-existing PoS-tagger and dependency parser models for German. Thus, CoreNLP was tested out-of-the-box, without any further processing, on the *Nephro_Gold* test partitions, input in plain text. It automatically performs tokenization, PoS-tagging and dependency parsing, yielding an average LAS of 27.09.

As expected, the original dependency tree model for German does not perform flawlessly on our clinical data. This may be due to the fact that the model was trained on non-clinical data. Moreover, we observe errors in tokenization and PoS-tagging that lead to consequential errors in the labelling of dependencies.

4.2 Experiment 1: Dependency Parsing of German Nephrology Reports

In order to observe the true efficiency of the SP, we eliminate potential errors caused by automatic preprocessing. Thus, we provide single tokens along with PoS-labels of the gold standard set as input to the SP and test the following three models on the *Nephro_Gold* test set using a 10-fold cross validation:

1. We test the default SP model again, as in the baseline test in Section 4.1, this time skipping the tokenizer and PoS-tagger, and instead, feeding the SP with tokens and their PoS from the gold standard test split. We refer to this configuration as '*stanford-conf*'.
2. We train a new parser model using only the *Nephro_Gold* training and development set. In doing so, we aim to create a parser specialized to German clinical language (specifically the subdomain of nephrology). We refer to this model as '*nephro*'.
3. Stanford's given German parser model contains optimized parameters to label dependencies on general text. We use this already existing model as baseline, re-train it (250 epochs) with the *Nephro_Gold* training set and optimize it against the *Nephro_Gold* development set. This way, we train a specialized dependency parser for clinical data that retains previously learnt knowledge about dependency parsing of more general data. We refer to this configuration as '*transfer*'.

baseline	**stanford-conf**	**nephro**	**transfer**
27.09	42.15	74.64	78.96

Table 3: Average LAS, based on a 10-fold cross-validation, on German nephrology data (Baseline + Experiment 1).

The results of the cross-validation presented in Table 3 show that simply by including gold annotation PoS-tokens into the input data and, thus, avoiding consequential parse errors, *stanford-conf* yields achieves better results than *baseline*. Moreover, *nephro*, trained solely on the small gold standard corpus, significantly outperforms *stanford-conf*, and *transfer* outperforms both other models.

All tested setups yield promising results, though, they have three drawbacks: 1) Inputting gold standard PoS-tokens does not represent a realistic scenario. 2) The gold standard data applied in *nephro* and *transfer* is relatively small. 3) Applying the parser to linguistically distinct nephrology data obscures its performance on more diverse German clinical data. These issues will be addressed in the next section.

4.3 Experiment 2: Extended Experiments

For the second experiment, a number of problems described in this paper have been successfully resolved: 1) We increase the size and the semantic variety of our test set (in comparison to the size of the test set in each single cross-validation step), 2) we use an external tool for tokenization and automated PoS-tagging and 3) we circumvent the legal obstacle by using fictitious clinical data which we can make available for further use.

In the first step, the fictitious data described in Section 3.2 is automatically pre-processed using

JPOS, a PoS-tagger trained on medical data that utilizes a Maximum Entropy model (Hellrich et al., 2015). As the fictitious dataset is not annotated, and evaluation has to be carried out manually, only the best performing model from the previous experiment in Section 4.2, *transfer*, is applied to the fictitious data. Our re-trained model takes JPOS-processed text (sentence-split, tokenized and PoS-tagged) as input.

In the final step, the output files are manually corrected by human evaluators (eval-1 and eval-2), who previously carried out the gold standard annotation (see Section 3.1), in roughly three hours per person. They evaluated PoS-tags and dependencies, and made amendments where required. Two clinical notes and one ABM discharge summary were examined by both evaluators, respectively, scoring an IAA of 0.9686 in terms of Cohen's kappa.

subset	**eval-1**	**eval-2**
clinical notes	75.96	81.75
ABM	69.69	76.26

Table 4: LAS for '*transfer*'-model on the fictitious dataset (Experiment 2).

Table 4 provides an overview of the parse accuracy determined manually by the evaluators. It shows that the performance of our system attains an LAS of over 75 on the clinical notes from the nephrology domain. Moreover, the results show that the performance on the clinical notes outperforms the results on ABM, which is not surprising as our dependency parser is trained on data of the same domain. However, a performance of above 69 on German clinical data outside the nephrology domain is still promising.

5 Conclusion

In this work, we examined the accuracy of the Stanford Parser on German clinical data. As expected, the default parser model, trained on the general domain, yielded deflating results. We presented our solution of re-training the existing model with a small gold standard dataset from the nephrology domain, which shows an improvement from 42.15 (*stanford-conf*) to 78.96 (*transfer*) (Experiment 1) when tested on the same domain. We further demonstrate that the re-trained model is able to process other clinical data outside the nephrology domain, despite the relatively small size of training and evaluation data. The fictitious data and the models trained on the confidential corpus are available here[2].

Acknowledgements

This research was supported by the German Federal Ministry of Economics and Energy (BMWi) through the project MACSS (01MD16011F).

References

[Aronson and Lang2010] A. Aronson and F. Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Association*, 17(3):229–236.

[Bikel2004] Daniel M Bikel. 2004. A Distributional Analysis of a Lexicalized Statistical Parsing Model. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[Chaplot et al.2015] Devendra Singh Chaplot, Pushpak Bhattacharyya, and Ashwin Paranjape. 2015. Unsupervised word sense disambiguation using markov random field and dependency parser. In *AAAI*, pages 2217–2223.

[Charniak2000] Eugene Charniak. 2000. A Maximum-Entropy-Inspired Parser. In *NAACL 2000 Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 132–139. Brown Laboratory for Linguistic Information Processing.

[Chen and Manning2014] Danqi Chen and Christopher Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. pages 740–750. Association for Computational Linguistics.

[Dalianis2018] Hercules Dalianis. 2018. *Clinical Text Mining*. Springer International Publishing, Cham.

[Faessler et al.2014] Erik Faessler, Johannes Hellrich, and Udo Hahn. 2014. Disclose Models, Hide the Data - How to Make Use of Confidential Corpora without Seeing Sensitive Raw Data. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

[Hahn et al.2016] Udo Hahn, Franz Matthies, Erik Faessler, and Johannes Hellrich. 2016. Uima-based jcore 2.0 goes github and maven centralstate-of-the-art software resource engineering and distribution of nlp pipelines. In *LREC*.

[2] `http://macss.dfki.de/index.html`

[Hellrich et al.2015] Johannes Hellrich, Franz Matthies, Erik Faessler, and Udo Hahn. 2015. Sharing models and tools for processing German clinical texts. *Studies in Health Technology and Informatics*, 210:734–738.

[Ingersoll et al.2013] Grant S. Ingersoll, Thomas S. Morton, and Andrew L. Farris. 2013. *Taming Text: How to Find, Organize, and Manipulate It.* Manning Publications Co., Greenwich, CT, USA.

[Jiang et al.2015] Min Jiang, Yang Huang, Jung-wei Fan, Buzhou Tang, Josh Denny, and Hua Xu. 2015. Parsing clinical text: how good are the state-of-the-art parsers? *BMC Medical Informatics and Decision Making*, 15(1):S2.

[Lohr et al.2018] Christina Lohr, Sven Buechel, and Udo Hahn. 2018. Sharing Copies of Synthetic Clinical Corpora without Physical Distribution A Case Study to Get Around IPRs and Privacy Constraints Featuring the German JSYNCC Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1259–1266. European Language Resources Association (ELRA).

[Ma et al.2015] Mingbo Ma, Liang Huang, Bowen Zhou, and Bing Xiang. 2015. Tree-based convolution for sentence modeling. *CoRR*, abs/1507.01839.

[Manning et al.2014] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. pages 55–60. Association for Computational Linguistics.

[McClosky and Charniak2008] David McClosky and Eugene Charniak. 2008. Self-Training for Biomedical Parsing. In *Proceedings of ACL-08: HLT, Short Papers*, pages 101–104, Columbus, Ohio, June. Association for Computational Linguistics.

[McClosky et al.2010] David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36. Association for Computational Linguistics.

[Nivre et al.2016] Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *LREC*.

[Patrick and Nguyen2011] Jon Patrick and Dung Nguyen. 2011. Automated Proof Reading of Clinical Notes. page 10.

[Rimell and Clark2009] Laura Rimell and Stephen Clark. 2009. Porting a lexicalized-grammar parser to the biomedical domain. *Journal of biomedical informatics*, 42(5):852–865.

[Roller et al.2016] Roland Roller, Hans Uszkoreit, Feiyu Xu, Laura Seiffe, Michael Mikhailov, Oliver Staeck, Klemens Budde, Fabian Halleck, and Danilo Schmidt. 2016. A fine-grained corpus annotation schema of german nephrology records. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*. The COLING 2016 Organizing Committee.

[Savkov et al.2016] Aleksandar Savkov, John Carroll, Rob Koeling, and Jackie Cassell. 2016. Annotating patient clinical records with syntactic chunks and named entities: the harvey corpus. *Language Resources and Evaluation*, 50(3):523–548.

[Savova et al.2010] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

[Seiffe2018] Laura Seiffe. 2018. Linguistic Modeling for Text Analytic Tasks for German Clinical Texts. Master's thesis, TU Berlin. To appear.

[Skeppstedt et al.2014] Maria Skeppstedt, Maria Kvist, Gunnar H. Nilsson, and Hercules Dalianis. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study . *Journal of Biomedical Informatics*, 49:148–158.

[Starlinger et al.2016] Johannes Starlinger, Madeleine Kittner, Oliver Blankenstein, and Ulf Leser. 2016. How to improve information extraction from German medical records. *IT-Information Technology*.

[Suster et al.2017] Simon Suster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. *CoRR*, abs/1703.10090.

[Wermter and Hahn2004] Joachim Wermter and Udo Hahn. 2004. An annotated German-language medical text corpus. In *Proceedings of the Forth International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

[Zou et al.2015] Huang Zou, Xinhua Tang, Bin Xie, and Bing Liu. 2015. Sentiment classification using machine learning techniques with syntax features. In *Computational Science and Computational Intelligence (CSCI), 2015 International Conference on*, pages 175–179. IEEE.