A Domain-adapted Dependency Parser for German Clinical Text

Elif Kara*, Tatjana Zeen*, Aleksandra Gabryszak*, Klemens Budde⁰, Danilo Schmidt⁰, Roland Roller*

*Language Technology Lab, DFKI, Berlin, Germany
⁰Charité Universitätsmedizin, Berlin, Germany

19–21 September 2018, KONVENS

Motivation

Overcoming the lack of text corpora of German clinical documents and domain-adapted parser models, which severely hampers NLP applications on German medical texts.

Contributions

1. Building two gold standard clinical corpora: authentic and fictitious.
2. Adapting Stanford Parser to clinical German.
3. Publishing our fictitious corpus and models trained on the authentic corpus.

Approach

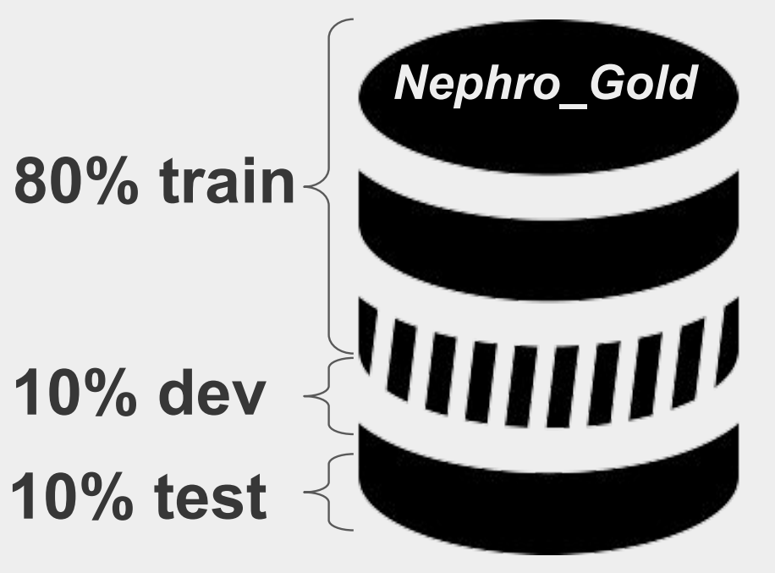
A parser, previously trained on general language data, is (re-)trained with in-domain gold data and tested on it, plus fictitious clinical documents.

LINGUISTIC CHALLENGES

Main features of clinical language problematic for machine-readability		*Examples*
Domain-dependence	Greek- and Latin-rooted terminology	*Appendektomie* ('appendectomy') *thorakal* ('thoracic')
Complexity	Discharge summaries: Complex syntactic embeddings	*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 (mostly auxiliary and copula verbs); Sentence boundaries	*Geht gut*. ('Goes well.') *Ödeme rückläufig* ('Edema declining')
	Clinical notes: Poor syntactic structure; Non-standard abbreviations and acronyms; Lack of punctuation marks	*Geht gut RR gut*. ('Goes well RR well.')

TRAINING AND EVALUATION DATA

10-fold cross-validation with equally assigned document types.



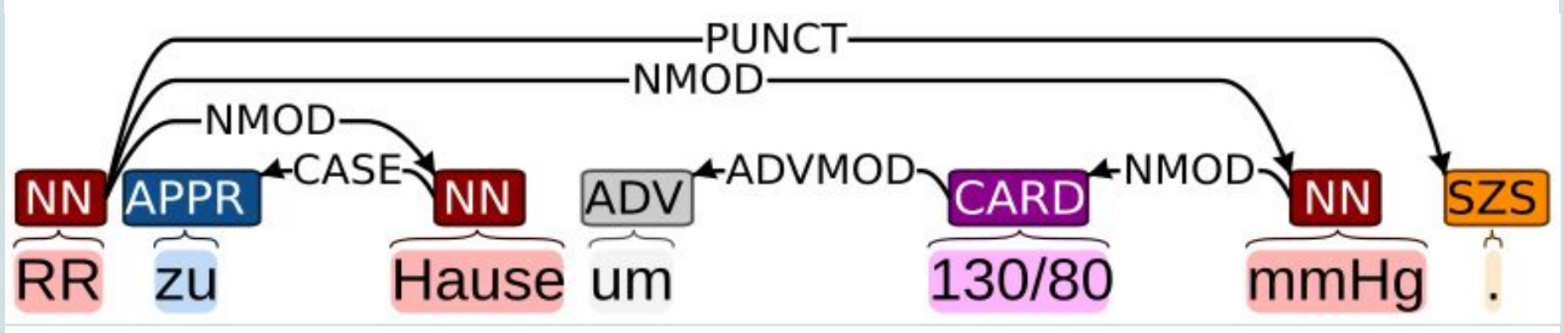
	1) Nephro_Gold		*2) Fictitious documents*	
	authentic, de-identified		synthetic, template-based	
	nephrology documents		diverse clinical data	
	gold-standard PoS and dependencies provided by human annotators		PoS and dependencies automatically parsed and amended by humans	
	clinical notes	*dis. summaries*	*clinical notes*	*dis. summaries*
# of files	44	11	30	5
avg. words (std.)	71.7 (75.2)	948.7 (333.3)	41.1 (12.0)	398.2 (226.6)
IAA	0.9578		0.9686	

ANNOTATION SCHEMA

PoS tagging:
Stuttgart-Tübingen-TagSet (STTS)

Dependency annotation:
Universal Dependencies (UD)

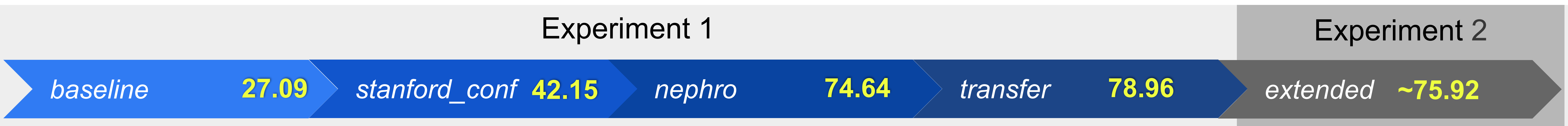
Exemplary sentence parse



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

EXPERIMENT ARCHITECTURE

All results are given as LAS[1]



	baseline 27.09	*stanford_conf* 42.15	*nephro* 74.64	*transfer* 78.96	*extended* ~75.92
	Performance of **Stanford Parser** and **Tagger** on nephrological text, tested out-of-the-box.	Performance of **Stanford Parser** on nephrological text. Elimination of potential errors caused by false PoS annotation.	Performance of **Stanford Parser** on nephrological text, **trained solely on in-domain data**.	Performance of **Stanford Parser** with the default model, **re-trained** with nephrological data.	Performance of **Stanford Parser** with the *transfer* model on a **fictitious dataset** of more clinical subdomains.
Test data	*Nephro_Gold*	*Nephro_Gold*	*Nephro_Gold*	*Nephro_Gold*	fictitious documents
PoS	Stanford PoS tagger	manual pre-processing	manual pre-processing	manual pre-processing	JPOS [3]
Parser	Stanford Parser	Stanford Parser	Stanford Parser	Stanford Parser	Stanford Parser
Training data	default [2]	default [2]	*Nephro_Gold*	default [2] + *Nephro_Gold*	default [2] + *Nephro_Gold*

	eval-1	*eval-2*	avg.
clinical notes	75.96	81.75	78.86
dis. summaries	69.69	76.26	72.98
avg.	72.83	79.01	75.92

[1] LAS: Labeled Attachment Score (a given dependency is scored as correct only if both the syntactic head and the label are identical)
[2] The default model of the Stanford Parser was trained on the Universal Dependencies (UD) treebank for German – a large dataset of heterogeneous nature
[3] JPOS: PoS-tagger trained on medical data

CONCLUSION

- Re-training a general language model with specific in-domain data yields better performance on nephrology texts than the Stanford Parser's default model.
- Plus, the model performs well on other clinical subdomains.
- Our fictitious gold standard corpus as well as models trained on the manually annotated, authentic data are published, bypassing German legal restrictions.

OUTLOOK

- Testing and evaluating the model on larger in-domain datasets as well as on data from more clinical subdomains.
- Increasing the quality and quantity of German clinical data.
- Reconsidering the substitutability of authentic and synthetic clinical documents for the purpose of advancing research.

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