Using an RNN-based Dependency Grammar to associate structured data from pathology reports with cancer specimens in biobanks

Julian DÖRENBERG[a], Nadine GAISA[b], Lara HEIJ[bc], Jan BEDNARSCH[c], Edgar DAHL[ab]

[a] RWTH cBMB Biobank at the Institute of Pathology, Medical Faculty of RWTH Aachen University
[b] Institute of Pathology, University Hospital Aachen
[c] Department of Surgery and Transplantation, University Hospital Aachen

Background

Availability of structured data is becoming an increasingly important factor for AI-based analyses of large sets of cancer data. Still, pathologists in Germany often record their histological findings in pathology reports using floating text. To aggregate these important tumor-associated data with further clinical data and cancer samples present in biobanks, it is critical to convert these report into a structured form. This can be done by using a Dependency Grammar (DG) parser[1] which is used to find grammatical relations between the words in a sentence.

Methods

The workflow to extract structured data from pathology reports and associate them to samples from a biobank is shown in Figure 1. To perform the extraction, a Dependency Grammar Parser is trained to extract grammatical relations between words within the report text. This output is shown in Figure 2.

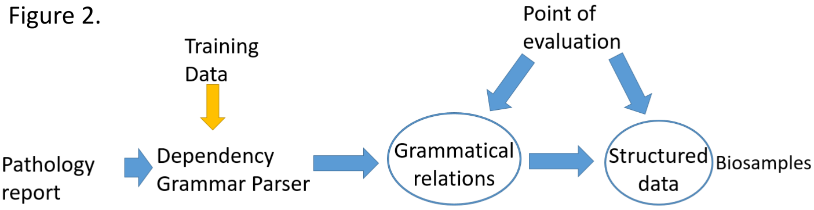
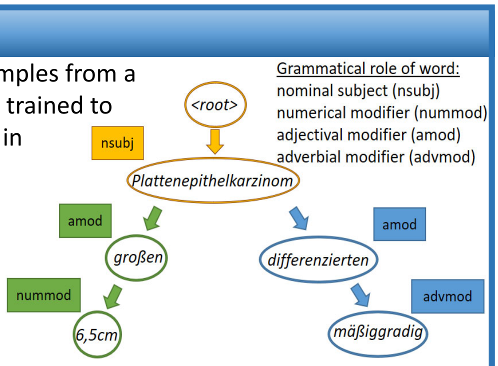

Point of evaluation
Training Data
Pathology report
Dependency Grammar Parser
Grammatical relations
Structured data
Biosamples

Fig. 1: The workflow for data extraction



Grammatical role of word:
nominal subject (nsubj)
numerical modifier (nummod)
adjectival modifier (amod)
adverbial modifier (advmod)

Fig. 2: Output of a DG-Parser

The workflow is evaluated by computing the number of correctly extracted grammatical relations from a corpus containing 165 sentences from breast biopsies as well as the number of correctly extracted informations from a corpus containing 10 reports from HCC diagnoses.

Result

The portions of correctly extracted relations from the 165 sentences from breast biopsy report are shown in Table 3. The arity of the relations is determined by the number of words in the relation. For instance, *6,5cm – groß – Plattenepithelkarzinom* is a 3-ary relation. Only relations that contain at least one medical term have been counted in the evaluation.

Arity of relations	Correctly extracted
2	97 %
3	93 %
4	89 %

Tab. 3: Portion of correctly extracted relations

Number of HCCs	Fibrosis	Vascular invasion	Tumor diameter	Inflamm.	Inflamm. degree	Distance to resection area	Desmet stage	Steatosis	Cirrhosis
1			1,4cm			1mm			TRUE
1		TRUE	5,5cm			0,3cm			
1	TRUE		4,2cm	TRUE		0,3cm			FALSE
1	TRUE	TRUE	8,5cm	TRUE	🟥		3		
1			16cm			0,1cm			
1	TRUE	TRUE	4,2cm	TRUE		1,5mm		TRUE	FALSE
1									
1		FALSE	9,5cm			1cm			
1		FALSE	8,5cm	TRUE				TRUE	
1	TRUE		3,6cm			0,2cm	1-2		

Tab. 4: Table of extracted data from HCC reports

On a corpus from the diagnosis of ten hepatocellular carcinoma, several tumor-associated informations have been extracted. The extraction result is shown in Table 4. The workflow extracted 98% of the requested informations correctly. The only information not extracted correctly was caused by a grammatically ambiguous sentence.

Legend: Correctly extracted
Falsely extracted/not extracted

Discussion

The approach passed a proof-of-concept. Although the training data do not contain medical terms, the parsing accuracy is high on histological reports. So far, the performance is not affected by the entity being diagnosed and different words used in the reports.

Conclusion

Dependency Grammar parsers can be used to extract data from standard German pathology reports. As the evaluation corpus so far is very small, it has to be extended. After further evaluation, the tool will provide arbitrary data from pathology report. Accordingly, it provides data that are ready to be used in Big Data and AI-based approaches in data-driven medicine.

Disclosure Statement: The authors declare no conflict of interest.

Indication of source:
[1] Dozat et. al. Deep Biaffine Attention for Neural Dependency Parsing. Conference paper at ICLR, 2017