Extraction of HCC-related data from histological reports by using a Dependency Grammar

Julian Dörenberg[1], Nadine Gaisa[2], Jan Bednarsch[3], Lara Heij[2,3] and Edgar Dahl[1,2]

[1] University Hospital RWTH Aachen, RWTH centralized Biomaterialbank

[2] University Hospital RWTH Aachen, Institute of Pathology

[3] University Hospital RWTH Aachen, Department of Surgery and Transplantation

Contact: jdoerenberg@ukaachen.de

Disclosure: Conflicts of interest and Introduction

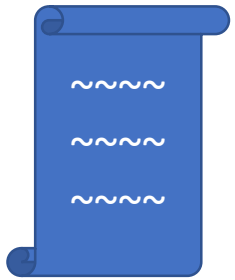
I herewith declare that I have no potential conflict of interest to report.

About me:

- Computer Scientist
- IT-Manager at RWTH centralized Biomaterialbank since 2019

Problem: Histological reports are texts in natural language

Patient care

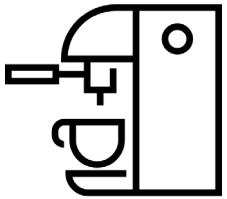


Histological report concerning HCC

Structured data

Currently done by human

- Translational research
- AI-based research
- Search for samples in Biobanking
- Data Integration Center
- ...



[Sym2]

Solution:

RelEx, let the computer do the work and drink coffee (or tea ☺)!

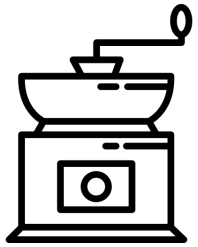
The Text Mining tool *RelEx* as a coffee machine

[Sym0]

Plug in coffee machine

Preprocessing

e.g. split sentences into words

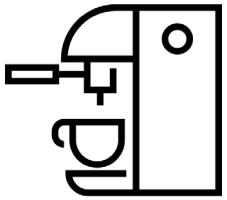


[Sym1]

Grinding coffee beans

Parse by using Dependency Grammar

Recurrent Neural Net (RNN) finds grammatical relations (Relation Extraction -> *RelEx*)



[Sym2]

Water runs trough filter

Filter grammatical relations

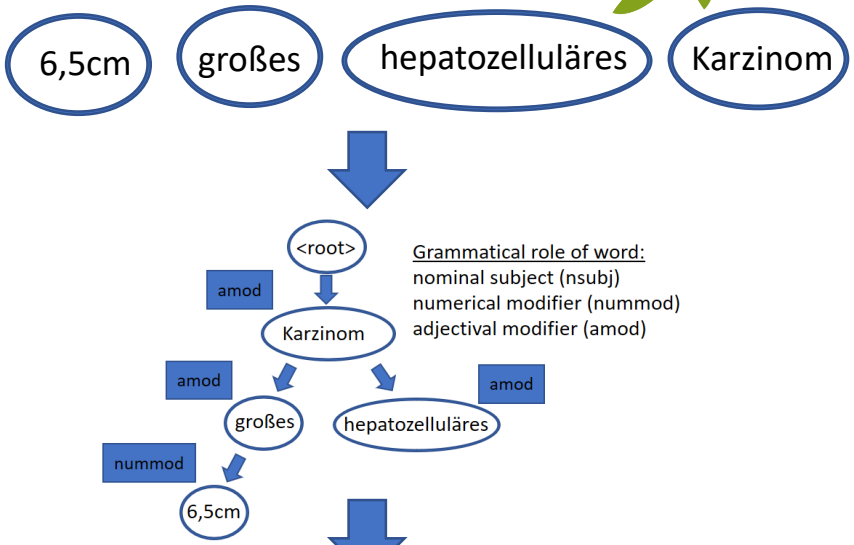
Grammatical relations are filtered for required information

[Sym3]

Coffee

Structured data

Structured data e.g. in table form



(hepatozellulär; Karzinom)
(6,5cm; großes; Karzinom)

Tumor size
6,5cm

Unknown medical words do not affect AIs performance!

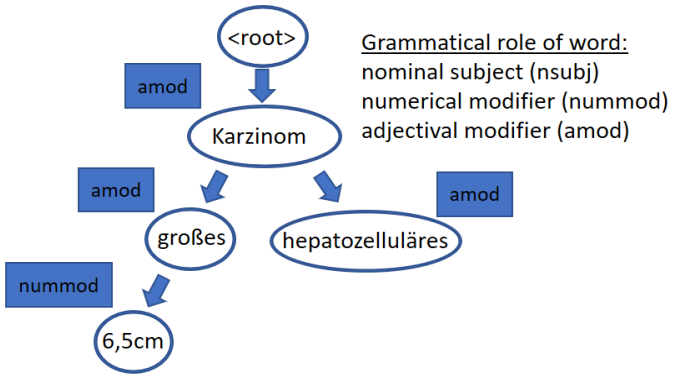
Parser [1]:

- Recurrent Neural Net
- Developed by Dozat and Manning for multilingual parsing

Data:

- Trained on non-medical data [1] such as newspaper articles and Google reviews
- Evaluated on own corpus of 200 breast biopsy sentences

Metric	All sentences (n=200)	Without localisation sentences like *oben links außen* (n=165)
Unlabelled Attachment Score (UAS)	0.94	0.96
Labelled Accuracy Score (LA)	0.92	0.95
Labelled Attachment Score (LAS)	0.9	0.93
UAS (medical words only)	0.95	0.97



UAS: Proportion of correct relations
LA: Proportion of correct tags
LAS: Proportion of correct tags and relations

The *RelEx* “coffee machine” shows good performance on HCC reports!

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		

Correct

Wrong

Not given in report

Extracted 46/47 (98%) requested information correctly

Mistake in column Inflammation degree is caused by grammatically ambiguous - and incorrect - sentence

Evaluation will be performed on larger data set

Summary

- RelEx shows very good performance on histological reports concerning HCC
- RelEx can easily be adapted for different entities and clinical texts
- We are open for further suggestions and cooperations
  - Mail: jdoerenberg@ukaachen.de
  - Tel: +49 241 80 89285
  - Web: https://www.cbmb.ukaachen.de/RelEx

References

[1] Timothy Dozat and Christopher D. Manning, (2017), Deep Biaffine Attention for Neural Dependency Parsing, Conference paper at ICLR

[2] Oliver Bodenreider, (2003), The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research, 267D-270, Volume 32

[3] Germalemma: https://github.com/WZBSocialScienceCenter/germalemma

[4] Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit, (2004), TIGER: Linguistic Interpretation of a German Corpus. Journal of Language and Computation

[Sym0] https://www.flaticon.com/free-icons/plugin Plugin icons created by Freepik - Flaticon

[Sym1] https://www.flaticon.com/free-icons/grinder Grinder icons created by Blak1ta – Flaticon

[Sym2] https://www.flaticon.com/free-icons/coffee-machine Coffee machine icons created by Freepik – Flaticon

[Sym3] https://www.flaticon.com/free-icons/cafe Cafe icons created by Pixel perfect - Flaticon

Backup-Slide: Filter relations

- Define required data set by using regular expressions and an Ontology Database (e.g. UMLS [2])

Example (diameter of HCC):
(*[0-9]+(,[0-9]+)?[cdm]m*, *groß|Größe*, *<C2239176>*)

Regular expression
e.g. fulfilled by *6,5cm*

Regular expression
combined with words

UMLS identifier
Here: Hepatocellular carcinoma

Fulfilled by:
(*6,5cm*, *groß*, *hepatozellulär*, *Karzinom*) →

Tumor size
6,5cm

➢ Define for each required information

Backup-Slide: Lemmatization

- Example: Karzinoms -> Karzinom
- Library Germalemma [3] combines lookup in TIGER corpus [4] and rule-based fallback

- Evaluation: On Lemmatized 537 of 626 medical words correctly (86 %)

- Newly added patterns based on evaluation (excerpt):

word stem	endings	substitution
karzinom	(es | en | e | s)	karzinom
befund	(es | en | s | e)	befund
untersuchung	(en)	untersuchung
m[aä]ngel	(s)	mangel
gen	(es | en | s | e)	gen
infiltrat	(es | en | s | e)	infiltrat

Backup-Slide: Dependency Grammar parsing [1]

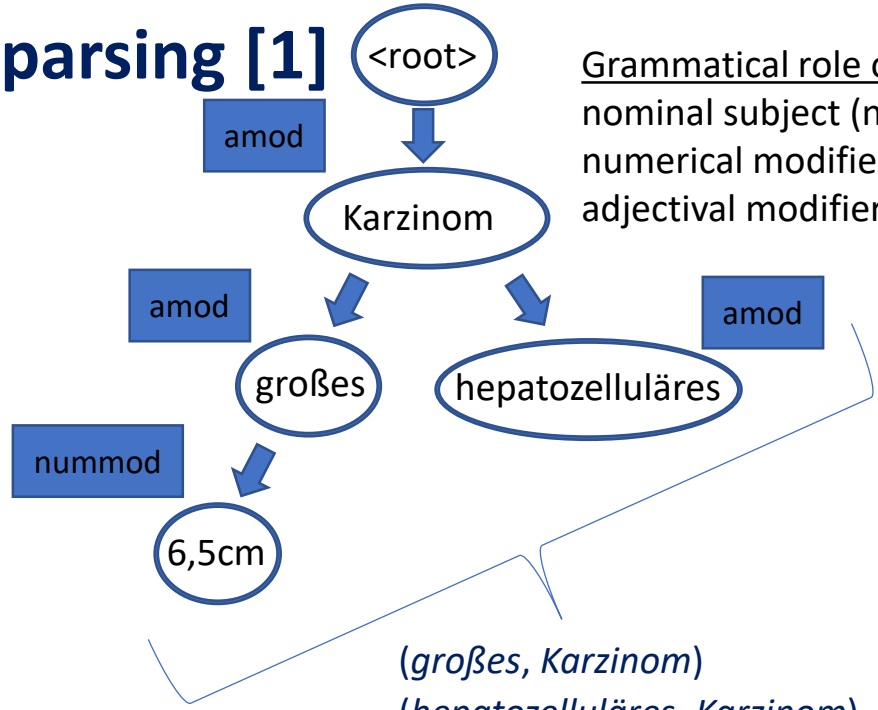
Output consists of

- Relations between words (grammatical relations)
  - Restricted to form a tree
- Types of these relations (grammatical role of word)

Properties of parser used for development [1]

- Based on a Recurrent Neural Net (RNN)
- Trained on non-medical data (e.g. newspaper reports)
- Evaluated on histological reports (200 sentences from breast biopsy reports)

Metric	All sentences	Without localisation sentences
Unlabelled Attachment Score (UAS)	0.94	0.96
Labelled Accuracy Score (LA)	0.92	0.95
Labelled Attachment Score (LAS)	0.9	0.93
2-ary relations	0.95	0.97
3-ary relations	0.91	0.93
4-ary relations	0.88	0.89



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

(*großes*, *Karzinom*)
(*hepatozelluläres*, *Karzinom*)
(*6,5cm*, *großes*)

(*6,5cm*, *großes*, *Karzinom*)
(*großes*, *hepatozalluläres*, *Karzinom*)
(*6,5cm*, *großes*, *hepatozelluläres*, *Karzinom*)

UAS: Proportion of correct relations
LA: Proportion of correct tags
LAS: Proportion of correct tags and relations

{2,3,4}-ary relations: Proportion of correctly extracted relations containing at least one medical word