3000PA—Towards a National Reference Corpus of German Clinical Language

Udo Hahn\textsuperscript{a}, Franz Matthies\textsuperscript{a}, Christina Lohr\textsuperscript{a}, Markus Löffler\textsuperscript{b}

\textsuperscript{a}Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Germany
www.julielab.de

\textsuperscript{b}Institute for Medical Informatics, Statistics and Epidemiology (IMISE)
Universität Leipzig, Germany
www.imise.uni-leipzig.de
Towards a German Clinical Reference Corpus

**German-language clinical corpora**

- **FRAMED**: hybrid mix of several (non-)clinical text genres (non-distributable; LREC 2004)
- many other small- and medium-sized corpora from single clinical sites and single genres; non-distributable (German data privacy law)
- vision of a national reference corpus: cross-hospital, cross-genre collection of clinical reports (distributable under DUAs)

**3000PA Corpus**

- part of SMITH (one of four funded consortia (40 Mio. €) within a major national German initiative to foster medical informatics research (BMBF))
- first national reference corpus for German clinical documents
- ≈ 1000 electronic patient records from three German university hospitals (Aachen, Jena and Leipzig)
- 2010-2015; internistic or ICU units; patients deceased
Why a (German) text corpus?

- collections of (machine-readable) text, either used for
  - training NLP systems in a (semi-)supervised way
  - evaluating the performance of (NLP) systems (benchmark data sets)

- (German) clinical text data
  - medical jargon constitutes a sublanguage on its own
  - differ across hospitals, clinical departments and text genres
  - evidence from (distributable) English clinical corpora is not transferable to German
Jena slice of 3000PA

- 1107 text documents
  - 620 discharge summaries
  - 487 transfer letters
- 1.75 Mio tokens
- 180 K sentences

- Leipzig and Aachen slices exhibit similar numbers
  - altogether, roughly > 5 Mio tokens and 500 K sentences
  - meanwhile, five clinical sites have joined the SMITH consortium and will, additionally, contribute around 1000 EPRs each
Case study – Medication

- pilot study: corpus can be used by clinic-external staff (JULIE Lab)

- replication of a task similar to the 3rd i2b2 challenge on medication extraction (JAMIA 2010)

- adaptation of English i2b2 guidelines to German clinical language using 3000PA
<table>
<thead>
<tr>
<th>Metadata relevant for medication extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>medication</td>
</tr>
<tr>
<td>dosage</td>
</tr>
<tr>
<td>mode</td>
</tr>
<tr>
<td>frequency</td>
</tr>
<tr>
<td>reason</td>
</tr>
</tbody>
</table>
Two studies related to medication extraction

**manual annotation campaign**
- documents annotated with medication information
- BRAT tool
- five students of medicine
- 52 documents double-annotated for measuring the agreement (IAA)

**automatic medication extraction**
- adaptation of the (English) MEDXN system (JAMIA 2014) to German: JUMEX
- based on regular expressions and German dictionaries (Rote Liste)
- rapid prototype only, not tuned for competitions
- processing based on the full Jena slice of 3000PA
Evaluation study

• performance of human annotation (based on inter-annotator agreement – IAA)
• performance of automatic annotation (based on JUMeX)
• all performance data vary dependent on the choice of string overlap criteria (centroids; LREC 2012)

<table>
<thead>
<tr>
<th></th>
<th>IAA</th>
<th>JUMeX</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>0.91 - 0.98</td>
<td>0.81 - 0.83</td>
</tr>
<tr>
<td>dosage</td>
<td>0.81 - 0.83</td>
<td>0.81 - 0.83</td>
</tr>
<tr>
<td>medication</td>
<td>0.90 - 0.96</td>
<td>0.49 - 0.50</td>
</tr>
<tr>
<td>duration</td>
<td>0.66 - 0.78</td>
<td>0.30 - 0.34</td>
</tr>
<tr>
<td>mode</td>
<td>0.69 - 0.85</td>
<td>0.19 - 0.22</td>
</tr>
<tr>
<td>reason</td>
<td>0.27 - 0.69</td>
<td>–</td>
</tr>
</tbody>
</table>
Conclusion

• 3000PA: first prototype of a German national reference corpus of clinical documents
  • cross-hospital (3+5), cross-genre (2+x)
  • currently, around 5 Mio. tokens, and 500k sentences
  • annotations available for sentences, tokens, section headings, medications (diseases, symptoms, and therapies soon to come)
• pilot study testing its usability for manual and automatic annotation
• replication of the 3rd i2b2 challenge task for German: medication extraction
• first published German corpus on medication metadata and automatic medication extraction
3000PA—Towards a National Reference Corpus of German Clinical Language

Udo Hahn\textsuperscript{a}, Franz Matthies\textsuperscript{a}, Christina Lohr\textsuperscript{a}, Markus Löffler\textsuperscript{b}

\textsuperscript{a}Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Germany
www.julielab.de

\textsuperscript{b}Institute for Medical Informatics, Statistics and Epidemiology (IMISE)
Universität Leipzig, Germany
www.imise.uni-leipzig.de