Extraction of UMLS Concepts using Apache cTAKES for German Language
Introduction

- Automatic information extraction of medical concepts and classification with semantic standards from medical reports is useful for standardization and for clinical research
- This is an approach for an UMLS concept extraction with a customized natural language processing pipeline for German clinical notes using Apache cTAKES
- The objective is, to test a natural language processing tool for German language if it is suitable to identify UMLS concepts and map these with SNOMED-CT
Introduction

Main goal:

- access knowledge from unstructured German clinical text for:
  - electronic patient records
  - research
  - personalized guideline-based treatment recommendations

- using an open source natural language processing Tool
- using international semantic standards
- using formalized German guidelines
Motivation & Introduction

- structured vs. unstructured information
- most available tools for Textmining and also terminologies are in English language
- evaluation problem: no existing German gold standard corpus
- Formalized German Guidelines have no link to the unstructured data
Methods

- Unified Medical Language System (UMLS)
  - compendium of many controlled vocabularies in the biomedical sciences
  - Metathesaurus $\rightarrow$ Concept Unique Identifier (CUI)
  - available in German language
  - By extracting the UMLS concepts it is possible, to map this concepts to other non-German speaking classifications and ontologies
SNOMED-CT

- Systematized Nomenclature of Medicine - Clinical Terms
- SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world
- Not available in German language
Methods

- Apache cTAKES
  - One of the most proven open source natural language processing tools
  - Apache cTAKES already offers a variety of algorithms for text analysis and information extraction
  - It can normalize to domain ontologies such as SNOMED-CT using UMLS concepts
Methods

- Architecture Overview

[Diagram showing the architecture overview with Apache cTAKES™ processing unstructured medical data in German language through UMLS (German) and openNLP models (German) to generate SNOMED CT Codes.]
Methods

- **Gold Standard**
  - ShARe/CLEF eHealth 2013 shared task 1 training set
    - 199 clinical notes
    - only in English notes
  - translated into German (Google Translator & manually)

<table>
<thead>
<tr>
<th>Type</th>
<th>#Note</th>
<th>#CUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>199</td>
<td>2798</td>
</tr>
<tr>
<td>DISCHARGE</td>
<td>61</td>
<td>1969</td>
</tr>
<tr>
<td>ECHO</td>
<td>42</td>
<td>479</td>
</tr>
<tr>
<td>RADIOLOGY</td>
<td>42</td>
<td>257</td>
</tr>
<tr>
<td>ECG</td>
<td>54</td>
<td>93</td>
</tr>
</tbody>
</table>
### Results

- **Pipeline with English clinical notes:**

<table>
<thead>
<tr>
<th>Type</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISCHARGE</td>
<td>0.71</td>
<td>0.27</td>
<td>0.39</td>
</tr>
<tr>
<td>ECHO</td>
<td>0.56</td>
<td>0.26</td>
<td>0.35</td>
</tr>
<tr>
<td>RADIOLOGY</td>
<td>0.69</td>
<td>0.23</td>
<td>0.35</td>
</tr>
<tr>
<td>ECG</td>
<td>0.81</td>
<td>0.25</td>
<td>0.38</td>
</tr>
</tbody>
</table>

- **Pipeline with German translated clinical notes:**

<table>
<thead>
<tr>
<th>Type</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISCHARGE</td>
<td>0.37</td>
<td>0.30</td>
<td>0.33</td>
</tr>
<tr>
<td>ECHO</td>
<td>0.47</td>
<td>0.51</td>
<td>0.49</td>
</tr>
<tr>
<td>RADIOLOGY</td>
<td>0.37</td>
<td>0.28</td>
<td>0.32</td>
</tr>
<tr>
<td>ECG</td>
<td>0.39</td>
<td>0.25</td>
<td>0.30</td>
</tr>
</tbody>
</table>
The UMLS database is not as extensive for the German Language (196842 entries) like the English (5571374 entries) UMLS database

German pipeline has a better average precision because it also identifies less false positive concepts
- F1 value on German echo notes is higher than the value of the English pipeline

no German stemming has been integrated into the pipeline at the time of the results publication and the German OpenNLP models have not yet been trained to medical notes
The next steps:

- NLP:
  - implement Context Analyzer
  - implement German stemming
  - expand UMLS Database
  - training of the openNLP models
- Treatment recommendations:
  - combining SNOMED-CT Codes & formalized guidelines for personalized guideline-based treatment recommendations
- Treatment recommendations:

  automatic markup (cTAKES)  
  automatic formalization (SNOMED CT)  
  unstructured clinical notes (SNOMED CT)
Vielen Dank für Ihre Aufmerksamkeit