INTRODUCTION & MOTIVATION


GitHub: https://github.com/Vicomtech/ClinIDMap

Goals:
- semantic interoperability across the clinical concepts
- enrich already annotated medical corpora in multiple languages with new labels
- create new datasets for machine learning models

Experiments with sequence labelling (SL) models for detecting:
- Diagnosis and Procedures
- UMLS Semantic Groups

Languages: Spanish, English, bilingual.

What for:
- Detect and categorize a span with clinical terminology with a high-level class: Diagnosis, Procedure, Anatomy, Chemical etc.
- Link classified span with clinical taxonomy and assign a unique ID (code)

Data for machine learning models in the clinical domain is especially difficult:
- Clinical information is private
- Manual annotation requires a high level of expertise in medicine
- Few data is available for languages other than English

EXPERIMENTS AND RESULTS

Annotate corpora with mapped codes:
- CodEsp (es) [8], EgC Corpus (es) [6], CT-EBM-SP (es) [2], MANTRA (es) [5], MedMentions (en) [4]. Corpora is also combined in order to augment the training data and perform bilingual experiments.
- Corpora annotated in Diagnosis and Procedures was mapped with Semantic Groups (UMLS land vice versa the corpora annotated with Semantic & Medical Groups were mapped with Diagnosis and Procedures (ICD-10).

Sequence labelling models:
1. Classification of Diagnosis and Procedures (SL-DP), according to ICD-10-CM and ICD-10-PSC notation.
2. Labelling the UMLS Semantic Groups12 (SL-SG), such as Anatomy, Disorder, Procedure, Chemical, etc. All models testes on two test sets: their own and gold-standard Spanish set from CodEsp corpus.

Architecture: BERT [3], Nvidia GeForce RTX 2080Ti 11 Gb RAM, 100 epochs, batch size 8

Results: The F1 score of the models trained on the corpus annotated with the mapping method is very similar to the gold corpus, annotation with mapping transfer knowledge across the lexical resources.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>SL-DP (F1 scores)</th>
<th>SL-SG (F1 scores)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CodEsp (es)</td>
<td>gold: 76.76 74.65 74.60</td>
<td>map: 72.19 71.12 70.90</td>
</tr>
<tr>
<td>Combined (es)</td>
<td>gold:map 80.15 88.37</td>
<td>gold:map 88.91 87.12 86.94</td>
</tr>
<tr>
<td>Combined (es)</td>
<td>gold:map 74.43 68.53 71.10</td>
<td>gold:map 70.05 70.05 70.55</td>
</tr>
<tr>
<td>MedMentions (en)</td>
<td>map: 59.02 68.50 68.90</td>
<td>gold: 64.51 69.29 69.29</td>
</tr>
<tr>
<td>Bilingual (es)</td>
<td>gold:map 87.07 87.19 87.32</td>
<td>gold:map 87.06 87.06 86.30</td>
</tr>
<tr>
<td>Bilingual (es)</td>
<td>gold:map 71.37 71.19 70.97</td>
<td>gold:map 71.06 70.68 70.41</td>
</tr>
<tr>
<td>Bilingual (es)</td>
<td>gold:map 89.10 88.54 88.18</td>
<td>gold:map 87.38 86.81 86.00</td>
</tr>
</tbody>
</table>

Table 7: Performance of the SL models on the test sets.

METHODOLOGY

- Extract all CUIs mapped to SNOMED-CT, ICD10-CM and ICD-10-PCS from the UMLS Metathesaurus.
- Extract the ICD-10 codes from the SNOMED CT to ICD-10 Mapping
- Extract the definitions of the ICD-10 and SNOMED-CT codes from the Spanish version of the ontologies.
- Extract all Wikipedia items that contain the given CUI and corresponding MeSH codes by using the Wikipedia Query Service.

CONCLUSIONS

- The models and corpora are quite interoperable with respect to different coding systems and languages
- The mapping tool is scalable for different languages
- Future work:
  - experiments with more languages
  - Annotate more corpora with the methods and do additional experiments
  - Add new taxonomies and ontologies (NCBI, BIOS, MESH)
  - link CUI descriptions with Wikipedia/Wikipedia items with deep learning. Only less than 1% of the one million UMLS CUIs can be found in Wikipedia.

REFERENCES


CONCLUSIONS

- The models and corpora are quite interoperable with respect to different coding systems and languages
- The mapping tool is scalable for different languages
- Future work:
  - experiments with more languages
  - Annotate more corpora with the methods and do additional experiments
  - Add new taxonomies and ontologies (NCBI, BIOS, MESH)
  - link CUI descriptions with Wikipedia/Wikipedia items with deep learning. Only less than 1% of the one million UMLS CUIs can be found in Wikipedia.