Supervised Learning for the ICD-10 Coding of French Clinical Narratives

Clément Dalloux¹, Vincent Claveau¹, Marc Cuggia², Guillaume Bouzillé² and Natalia Grabar³

¹ Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France
² Univ Rennes, CHU Rennes, Inserm, LTSI – UMR 1099, F-35000 Rennes, France
³ UMR 8163 STL CNRS, Université de Lille, France
Context

- Programme de médicalisation des systèmes d’information (PMSI):
  - Started in the 80’s as part of the reform of the French health system
  - 1982-2005: epidemiology, monitoring health institutions’ activity
  - 2005-now: activity based financing to reduce inequalities of resources between health institutions
- Health facilities manually assign codes to reports for billing/funding purposes
- Aim: computer-aided coding of clinical narratives
ICD-10

- International Statistical Classification of Diseases and Related Health Problems 10th revision (ICD-10)
  - Standardized, structured representation of clinical texts used for the billing of medical procedures, epidemiology, internal management, etc.
  - up to 16,000 codes representing numerous diagnostics and clinical and social situations
ICD-10

- Hierarchical structure, codes with up to 4 digits
- A lot of codes do not have children:
  - A38 Scarlet fever
  - B03 Smallpox
  - D70 Agranulocytosis
  - I10 Essential (primary) hypertension
Related work

- CLEF eHealth 2016-2018 shared task, classification of death certificates
  - French datasets: raw and aligned, stored in .csv files
  - 1 line per row and 3 lines, 10 tokens and 4 ICD codes per document on average
  - Best neural network system: (Atutxa et al., 2018), \(0.838 \text{ F}_1\) (aligned) and \(0.709 \text{ F}_1\) (raw)
  - Best rule based system: (Cossin et al., 2018), \(0.670 \text{ F}_1\) (raw)
  - These systems are sentence based, not multi-label

- Are death certificates free or structured texts?
Related work

<table>
<thead>
<tr>
<th>Line</th>
<th>Text</th>
<th>Normalized text</th>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>choc septique</td>
<td>choc septique</td>
<td>A41.9</td>
</tr>
<tr>
<td>2</td>
<td>peritonite stercorale sur perforation colique</td>
<td>peritonite stercorale</td>
<td>K65.9</td>
</tr>
<tr>
<td>2</td>
<td>peritonite stercorale sur perforation colique</td>
<td>perforation colique</td>
<td>K63.1</td>
</tr>
<tr>
<td>3</td>
<td>Syndrome de détresse respiratoire aiguë</td>
<td>Syndrome détresse respiratoire aiguë</td>
<td>J80.0</td>
</tr>
</tbody>
</table>

A sample document from the CépiDC French Death Certificates Corpus: aligned dataset (Névéol et al., 2018)
Material

- CHU Server with 2 NVIDIA Tesla V100 GPUs
- Rennes University Hospital dataset:
  - 28,000 hospitalization records with the matching ICD-10 codes documents from all medical specialties (cardiology, urology, oncology, ophthalmology, etc.)
  - A typical document describes: reasons for hospitalization, patient's medical history, sometimes family history, laboratory measurements and findings, treatment, length of stay, release
  - Data limitations: changes in coding happen periodically (ICD-11)
Material

- **6,116 unique ICD-10 codes**, 400+ not included in the train set
- **1,345 tokens** per document on average (10 for Clef eHealth)
- 8 unique codes per document on average (4 for Clef eHealth)

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Development</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>22,400</td>
<td>2,800</td>
<td>2,800</td>
</tr>
<tr>
<td>Tokens</td>
<td>30,307,091</td>
<td>3,718,715</td>
<td>3,654,599</td>
</tr>
<tr>
<td>Total ICD codes</td>
<td>182,112</td>
<td>23,010</td>
<td>22,561</td>
</tr>
<tr>
<td>Unique ICD codes</td>
<td>5,735</td>
<td>2,824</td>
<td>2,887</td>
</tr>
<tr>
<td>Unique unseen ICD codes</td>
<td>191</td>
<td>212</td>
<td></td>
</tr>
<tr>
<td>Codes &lt; 10 examples</td>
<td>3,885</td>
<td>2,320</td>
<td>2,391</td>
</tr>
<tr>
<td>Codes &gt; 100 examples</td>
<td>382</td>
<td>32</td>
<td>31</td>
</tr>
</tbody>
</table>
Distribution of codes
Multi-label classification of French clinical narratives

- Task 1: with all available ICD-10 codes, **6,116 classes**
  - Subtasks: with the top 1,000 and 100 classes
- Task 2: with codes from A00 to Z99, **1,549 classes**
  - higher hierarchical level (A00.0 → A00)
  - still relevant for health institutions
  - Subtask: with the top 100 classes
Methods

- **Baselines:**
  - Dictionary based system

- **Deep learning systems:**
  - Bidirectional Long Short-Term Memory neural network
  - Convolutional neural network
  - Word vectors: fastText vectors trained on our dataset with 300 dimensions
Dictionary based system

- Clef ehealth 2018 (Cossin et al., 2018):
  - Uses the ICD-10 terminologies to assign codes to each text line
  - Runs Levenshtein distance and a module of synonyms expansion

- Our system uses:
  - Clef eHealth’s dictionnaire2015.csv to assign codes (147,342 entries)
  - Levenshtein distance between dictionary entries and “sentences”
Neural networks

Best parameters: Filter width = [2,3,4,5]; Feature maps: 2000; FCL: 2000 hidden units
## Results

- The rule based system yields mediocre results
- The CNN yields the best results by far
- Much lower results than the systems presented at CLEF eHealth 2018, same task?

<table>
<thead>
<tr>
<th>Task</th>
<th>System</th>
<th>Precision</th>
<th>Recall</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>Dict</td>
<td>0.0722</td>
<td>0.0356</td>
<td>0.0476</td>
</tr>
<tr>
<td></td>
<td>BiLSTM</td>
<td>0.6513</td>
<td>0.1387</td>
<td>0.2287</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>0.5029</td>
<td>0.3301</td>
<td><strong>0.3986</strong></td>
</tr>
<tr>
<td>Task 1 top 1000c</td>
<td>CNN</td>
<td>0.5604</td>
<td>0.3839</td>
<td>0.4557</td>
</tr>
<tr>
<td>Task 1 top 100c</td>
<td>CNN</td>
<td>0.6317</td>
<td>0.5667</td>
<td><strong>0.5974</strong></td>
</tr>
<tr>
<td>Task 2</td>
<td>Dict</td>
<td>0.0522</td>
<td>0.0725</td>
<td>0.0607</td>
</tr>
<tr>
<td></td>
<td>BiLSTM</td>
<td>0.6309</td>
<td>0.2837</td>
<td><strong>0.3914</strong></td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>0.6018</td>
<td>0.4625</td>
<td><strong>0.5230</strong></td>
</tr>
<tr>
<td>Task 2 top 100c</td>
<td>CNN</td>
<td>0.6666</td>
<td>0.5873</td>
<td><strong>0.6245</strong></td>
</tr>
</tbody>
</table>
Conclusions

- First experiments on ICD-10 coding of French clinical **narratives**
- Multi-label classification experiments with several deep learning algorithms
  - CNN yields the best results by far
  - Using a still relevant higher hierarchical level works better
- Can a dictionary based expert system on narratives work?
  - Implement a better sentence detection method
  - Implement a better matching method than Levenshtein distance
Perspectives

● Building a bigger dataset from the CHU database
  ○ Adding all available structured data as additional input to our system
  ○ Training BERT, Flair, ELMo embeddings on all texts from the CHU database

● Adding more preprocessing to clinical narratives
  ○ medical history detection, negation detection, etc.

● Using Perform a thorough error analysis:
  ○ Does the prediction fail or are the documents lacking codes?
References

