Predicting Multiple ICD-10 Codes from Brazilian-Portuguese Clinical Notes

Arthur D. Reys, Danilo Silva, Daniel Severo, Saulo Pedro, Marcia M. de Souza e Sá, and Guilherme A. C. Salgado

BRACIS 2020

Authors

Arthur D. Reys

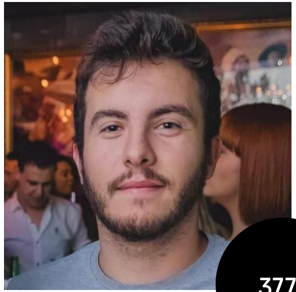
3778

UNIVERSIDADE FEDERAL
DE SANTA CATARINA

Danilo Silva

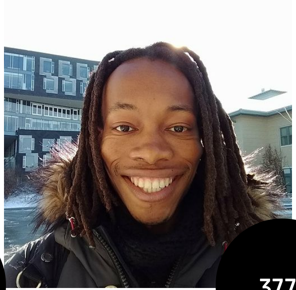
UNIVERSIDADE FEDERAL
DE SANTA CATARINA

Daniel Severo



3778

Saulo Pedro



3778

Guilherme A. C.
Salgado

3778

Marcia M. de
Souza

SOCIEDADE BENEFICENTE DE SENHORAS

The ICD Coding Task

- International Classification of Diseases - WHO
- Standard classification of symptoms, clinical evolution, diagnoses and medical history
- Billing, Health plan communication, Database organization for research
- Expensive and time-consuming task

The ICD Coding Task

EHR - Compilation of Clinical Notes

Paciente masculino, 60 anos, admitido no pronto socorro com quadro de apendicite aguda (confirmada por US abdominal total + exames laboratoriais). Submetido a laparotomia...

Professional Coders

List of ICD Codes

K35.9 Apendicite aguda SOE
I10 Hipertensão essencial
Z871 Hist. pessoal de doenc. aparelho digestivo
E149 Diabetes mellitus NE
R11 Náusea e vômitos

The ICD Coding Task

EHR - Compilation of Clinical Notes

Paciente masculino, 60 anos, admitido no pronto socorro com quadro de apendicite aguda (confirmada por US abdominal total + exames laboratoriais). Submetido a laparotomia...

Professional Coders

List of ICD Codes

K35.9	Apendicite aguda SOE
I10	Hipertensão essencial
Z871	Hist. pessoal de doenc. aparelho digestivo
E149	Diabetes mellitus NE
R11	Náusea e vômitos

Free Text

AI

Thousands of ICD codes

Previous Work

- Hierarchical and rule-based models
  - Hierarchical (Baumel et al, 2017), ICD co-occurrence (Subotin et al, 2016)
- Machine Learning and Deep learning
  - kNN (Ruch et al, 2008), SVM (Perotte et al, 2014), Naive Bayes (Medori et al, 2010)
  - CNN (Li et al, 2017; Mullenbach et al, 2018)
  - RNN (Huang et al, 2019; Ayyar et al, 2017; Baumel et al, 2017)
  - Attention mechanisms (Li et al, 2019; Mullenbach et al, 2018)
  - **CAML** (Mullenbach et al, 2018) - current state-of-the-art
- Brazilian-portuguese
  - No public dataset, most works focus on small set of ICD codes
  - SVM (Oleynik et al, 2017) and RNN (Duarte et al, 2018)

Datasets - MIMIC-III

- Publicly accessible
- Patient information from Beth Israel Deaconess Medical Center collected between 2001 and 2012
- English Language
- Baseline for comparison
- Only Discharge Summaries selected
- 6918 ICD-9-CM codes
- 52722 hospital admissions from 41127 patients

Datasets - HSL

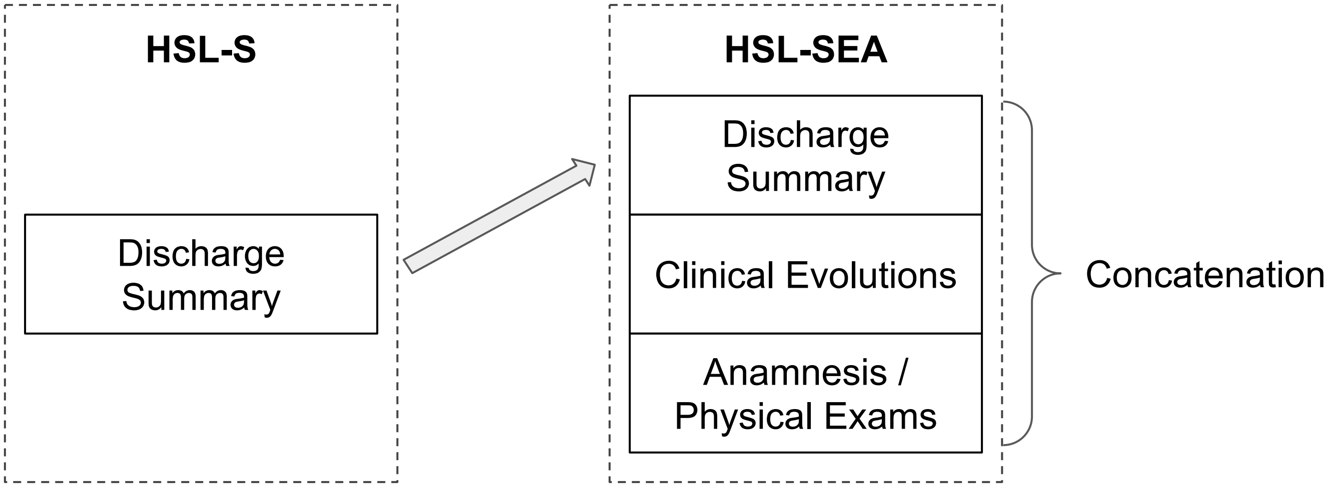
HOSPITAL SÍRIO-LIBANÊS

- Patient information from Syrian-Lebanese Hospital collected between 2016 and 2018
- Brazilian-Portuguese Language
- 5360 ICD-10 codes
- 77005 hospital admissions from 51298 patients

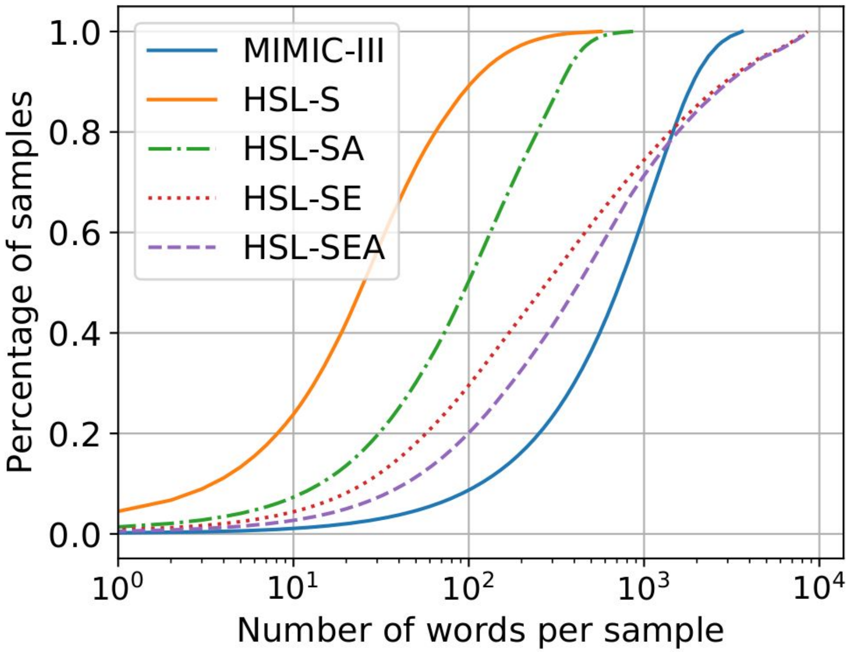
Datasets - HSL

HOSPITAL SÍRIO-LIBANÊS

- Initially, only Discharge Summaries selected
- Additional document types available:
  - Clinical Evolutions (E)
  - Anamnesis/Physical Exams (A)

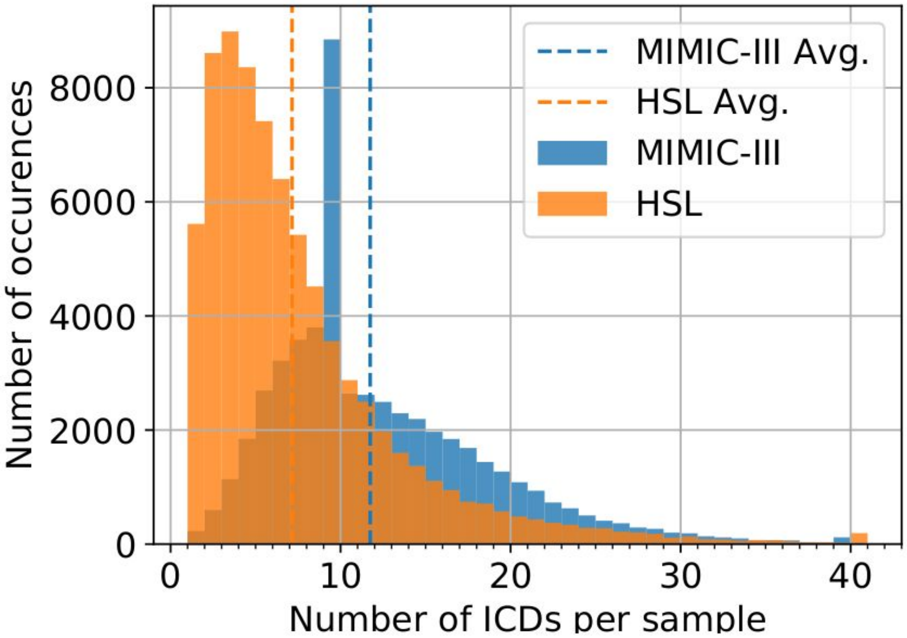


Datasets - Comparison



Dataset	Avg. words per sample
MIMIC-III	1327.5
HSL-S	94.6
HSL-SEA	1730.4

Datasets - Comparison



Percentages of samples that contain the most frequent ICD codes

Dataset	MIMIC-III	HSL
1st	38.02%	34.37%
10th	11.67%	10.71%
100th	2.23%	1.26%
1000th	0.15%	0.06%

Feature Extraction

- TF-IDF - only for Logistic Regression
  - Popular approach for baselines
  - Removed Stopwords
  - Limited vocabulary to 20000 most occurring words
- Word2Vec Word Embeddings - for Neural Networks
  - Self-trained due to language specificity
  - Skip-gram training algorithm
  - Context window of size 5 words
  - Experimented removing stopwords
  - 300 dimensional word vectors

Training Methods

- Preprocessing:
  - Light text preprocessing
  - Transform using trained feature extraction methods
  - For neural networks, pad/truncate input texts to a fixed length
  - Split data 90% / 3% / 7% as in Mullenbach et al. (2018)
- Training:
  - Train for 10 epochs
  - Restore weights from epoch with highest validation metrics

Metrics - Micro-averaged F1

- Chosen metrics:
  - Precision: Ability not to rate as positive a sample that is negative.
  - Recall: Ability to find all the positive samples.
  - **F-score**: Ponderation between precision and recall, through harmonic mean.
- Multi-label average methods
  - Macro: Metric is computed for each class, and then the average between classes is computed
    - Gives too much importance to rare ICDs - large and imbalanced set of classes
  - **Micro**: Metric is computed globally
- Threshold optimization over network predictions

Models - Baselines

- Top-k Baseline (Constant)
  - Predicts, for all samples, the *k* most occurring ICDs in the training set.
- Logistic Regression
  - Baseline in previous works
  - Multi-label problem into a set of binary classification problems
  - TF-IDF features as inputs
  - GridSearch (optimizers, learning rate, regularization)

Models - Convolutional Neural Network (CNN)

- Local context and parameter sharing
- Based on Mullenbach et al. (2018)
  - No Dropout
  - Added Batch Normalization
  - Global Average Pooling instead of Max Pooling
  - Increased kernel size from 4 to 10

Input
Embedding (size 300)
Conv1D (500 filters, kernel 10, tanh)
Batch Normalization
Global Average Pooling 1D
Output (sigmoid)

Models - Gated Recurrent Neural Network (GRU)

- Memory over long context
- GRU Units for training efficiency
- Experimented with different parameters:
    - Extra layers and Bi-directional layers
    - Optimizers and learning rates
    - Fine-tuning of embedding layer
    - Pooling methods

Input
Embedding (size 300)
GRU layer (500 filters, kernel 10, tanh)
Batch Normalization
Global Average Pooling 1D
Output (sigmoid)

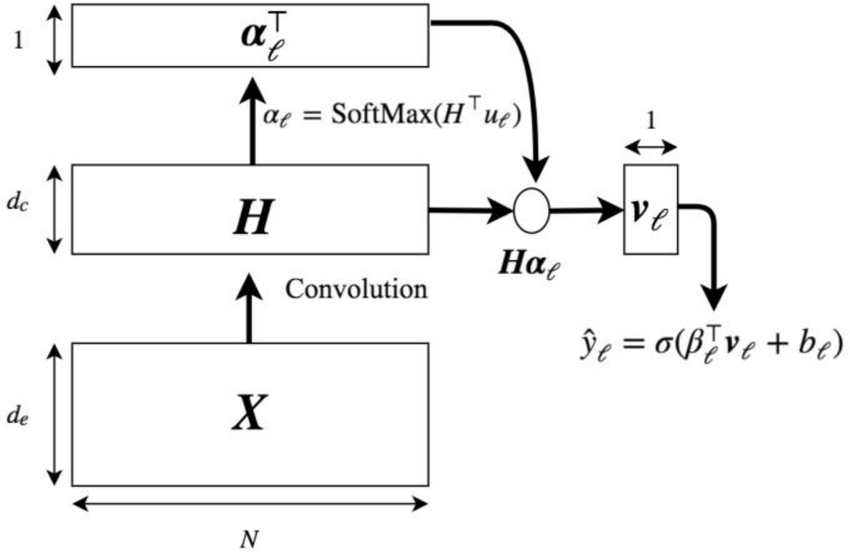
Models - CNN with Attention (CNN-Att)

- Based on Mullenbach et al. (2018)
- Our tests showed improvements when:
    - Removing Dropout and adding Batch Norm.
    - Increasing number of filters
    - Scheduling learning rate for faster training

Input
Embedding (size 300)
Conv1D (500 filters, kernel 10, tanh)
Batch Normalization
Attention
Output (sigmoid)

Models - CNN with Attention (CNN-Att)

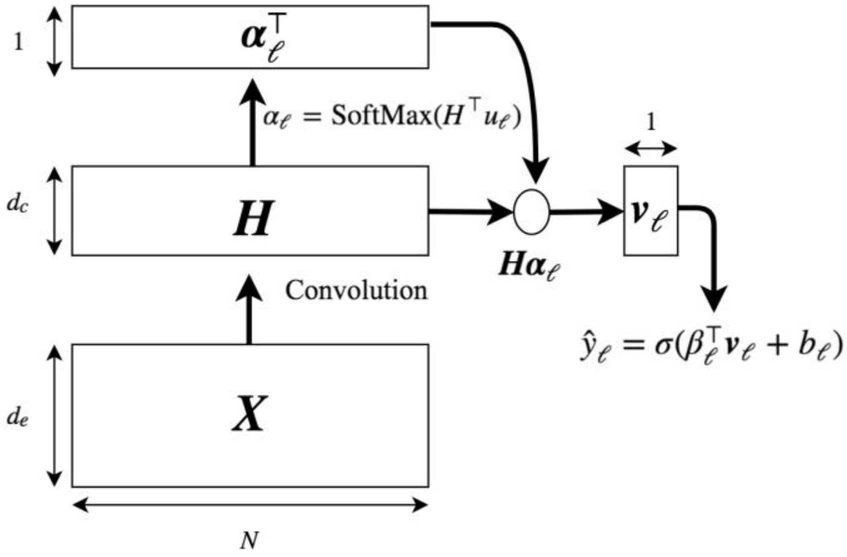
- Attention layer allows per-label custom weightening of the Conv1D output feature map



Models - CNN with Attention (CNN-Att)

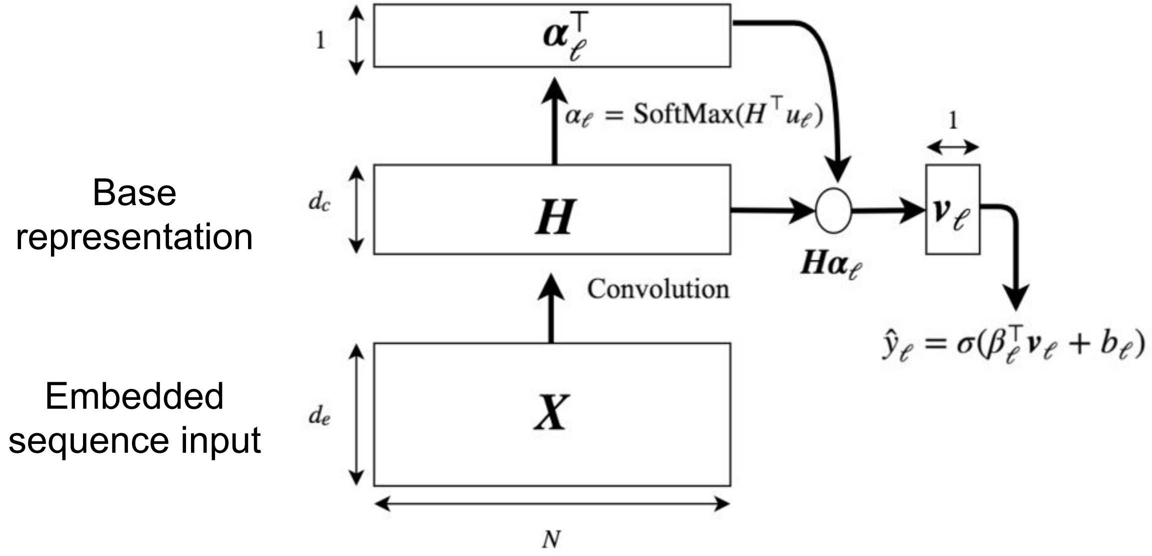
- Attention layer allows per-label custom weightening of the Conv1D output feature map

Embedded sequence input



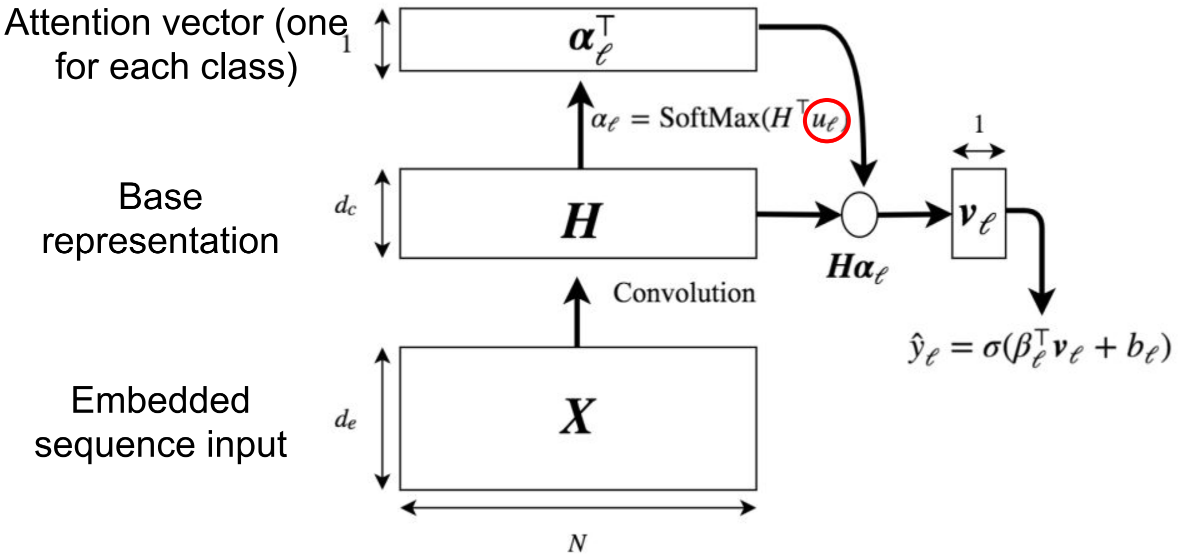
Models - CNN with Attention (CNN-Att)

- Attention layer allows per-label custom weightening of the Conv1D output feature map



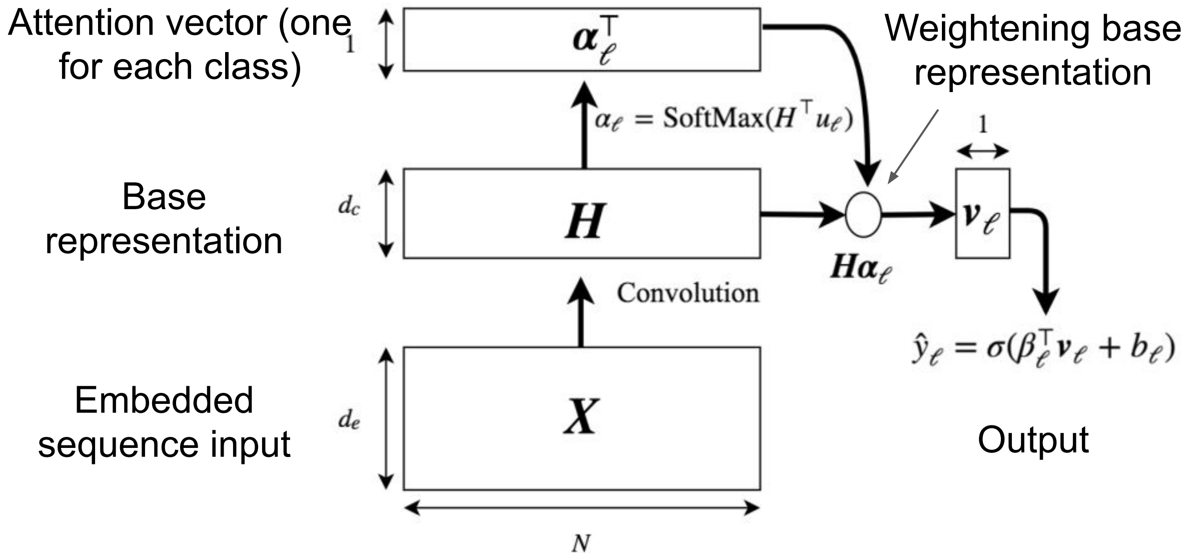
Models - CNN with Attention (CNN-Att)

- Attention layer allows per-label custom weightening of the Conv1D output feature map



Models - CNN with Attention (CNN-Att)

- Attention layer allows per-label custom weightening of the Conv1D output feature map



Results - MIMIC-III

Model	Threshold	F1	Precision	Recall
Constant	-	0.192	0.188	0.196
LR (Mullenbach et al, 2018)	-	0.242	-	-
flat-SVM (Li et al, 2019)	-	0.253	0.635	0.158
LR	0.19	0.406	0.425	0.388
CNN (Mullenbach et al, 2018)	-	0.402	-	-
CNN (Li et al, 2019)	-	0.399	0.440	0.366
CNN	0.30	0.423	0.467	0.387
Bi-GRU (Mullenbach et al, 2018)	-	0.393	-	-
GRU	0.32	0.468	0.543	0.412
CAML (Mullenbach et al, 2018)	-	0.524	-	-
CNN-Att	0.28	**0.537**	0.590	0.492

- Major improvements in LR family of models
- CNN-Att shows slight improvement over SOTA

Results - HSL

- LR model validation metrics with concatenation of different document types

Documents	Threshold	F1	Precision	Recall
S	0.26	0.316	0.320	0.312
S and A	0.25	0.347	0.359	0.336
S and E	0.27	0.357	0.382	0.336
S, E and A	0.25	**0.367**	0.390	0.346

S - Discharge Summaries
E - Clinical Evolutions
A - Anamnesis

- HSL-SEA results

Model	Threshold	F1	Precision	Recall
Constant	-	0.203	0.183	0.228
LR	0.25	0.368	0.400	0.340
CNN	0.26	0.374	0.386	0.363
GRU	0.29	0.441	0.508	0.390
CNN-Att	0.29	**0.485**	0.543	0.438

Discussion

- Metrics from the LR model show that HSL-S could not match MIMIC-III results
    - Much smaller average of words per sample
    - Empirically, MIMIC-III is much more detailed
    - The ICD coding process at the Syrian-Lebanese Hospital takes into account all available documents from a single patient, so using HSL-SEA for ICD prediction makes sense
- With HSL-SEA, results are still lower but much more comparable across all models
- The CNN-Att proved to be the better approach.

Conclusion

- Modifications over the current SOTA resulted in metric improvements and faster training
- Using only discharge summaries from our Brazilian-Portuguese dataset is insufficient to achieve satisfactory results
- Our CNN-Att achieves a performance only 10% lower in HSL-SEA compared to MIMIC-III, we conclude that this model is suited to be used to aid the current tagging process, allowing for speed improvements and a considerable decrease in errors.

Thank you!