LIMICS at N2C2 2019 Track 3: Semi-supervised and contextualized normalization through self-training

Perceval Wajsbürt [1], Xavier Tannier [2]

LIMICS, Sorbonne Université, Paris

[1]perceval.wajsburt@sorbonne-universite.fr; [2]xavier.tannier@sorbonne-universite.fr

Objectives

The normalization task — mapping a medical mention in a text to its concept in an ontology — displays multiple difficulties that we tried to address in our submission:

- UMLS synonyms are context-free, with ambiguous abbreviations
- unknown statistical bias towards preferred concepts
- annotated concepts in real reports are scarce, and represent a small fraction of the possible medical concepts

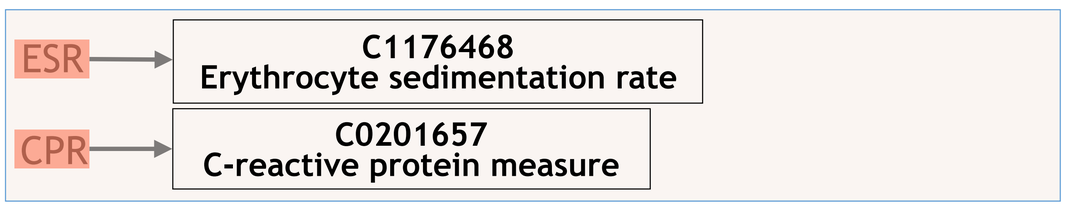
Training strategy

1 - Context-free normalization training

Objective : Learn a context free normalization model from the UMLS only

Method :

- Use the mention classifier head
- Minimize cross entropy on the UMLS



2 - Named entity recognition training

Objective : Learn to extract named medical entities from an annotated corpus

Method :

- Reuse first layers of model 1 as starting point
- Replace the CUI head with a tag head (see figures)
- Minimize CRF cross-entropy on the BIO-encoded N2C2 train dataset

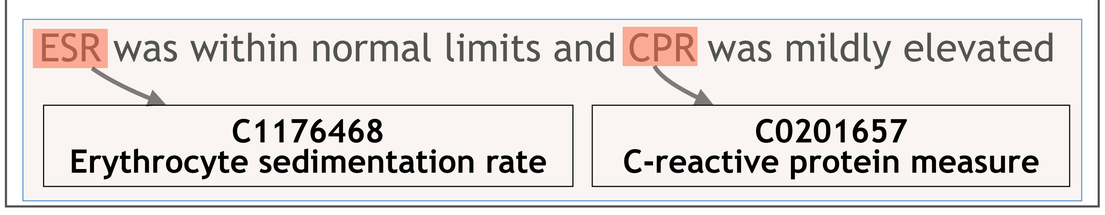
ESR was within normal limits and CPR was mildly elevated

3 - Self-training on unlabeled MIMIC

Objective : Train a new model to normalize mentions using their context and a (learned) bias toward preferred concepts

Method :

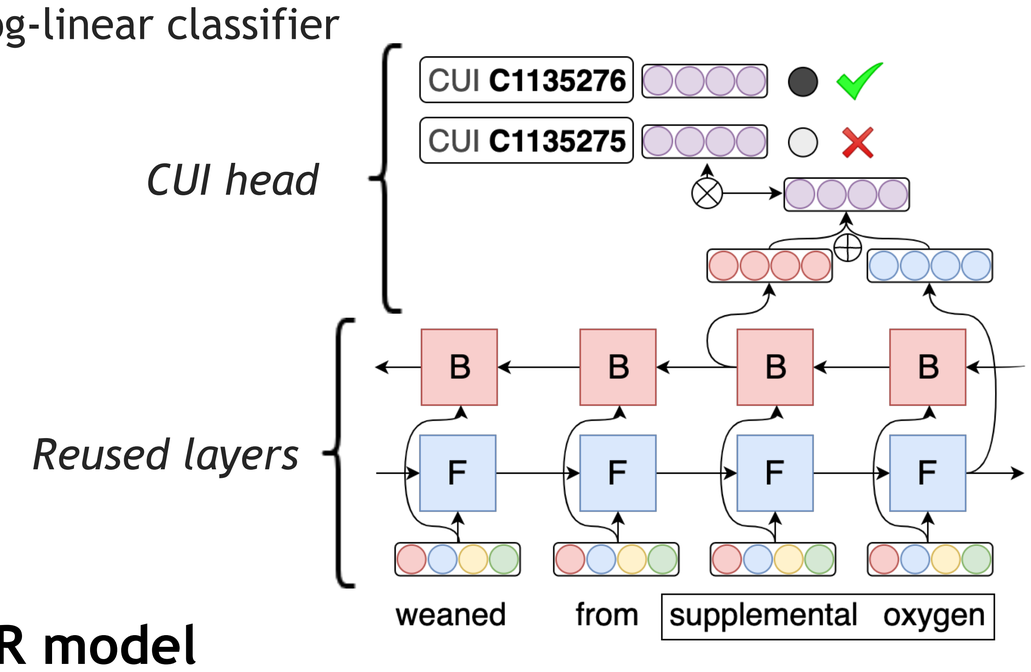
- Reuse the model 1 as starting point
- Allow the LSTM to look at words outside a mention
- Minimize cross-entropy between context-free predictions and the new model prediction



Neural network models

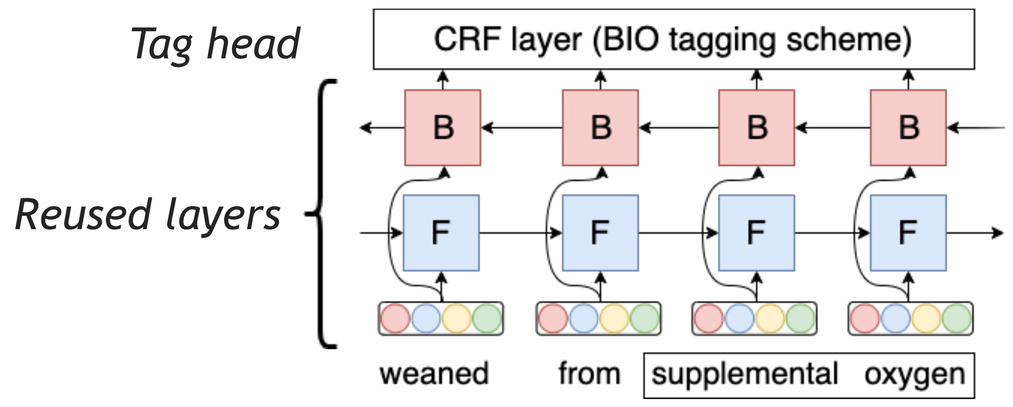
Normalization model

- ClinicalBERT embeddings (first layer)
- 2 layers Bi-LSTM
- Batch normalization
- Log-linear classifier



NER model

- ClinicalBERT embeddings (first layer)
- 2 layers Bi-LSTM and CRF, similar to [1]
- BIO tagging scheme with only 1 class



Results

Accuracy/step/split	Train	Val	Test
1 (Normalization)	0.970	0.784	**0.785**
2 (NER f1 scores)	0.907	0.611	0.602
3 (Normalization)	0.983	0.758	0.733

Discussion & conclusion

The results from the step 3 of the training show that the final model context-dependent failed to outperform the context free model. We plan to work and improve the method:

- use the full contextualized BERT embeddings and not just the first word embeddings layer
- improve the NER model to better extract mentions from annotated reports
- refine the self-training objective to only use confident context-free predictions

[1] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural Architectures for Named Entity Recognition. 2016

LIMICS

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257