Scaling Up Medical Entity Extraction

Vignesh Venkataraman (viggy@curai.com)

CURAI

Problem

- Entity extraction: goal is to infer **medical concepts** discussed in **unstructured text**
- Example: "I have a fever and a headache"
  - C0015967, C0392676 (fever)
  - C0018681 (headache)
- Existing approaches:
  - ML (catch: paucity of labeled data)
  - Rule Based (catch: rules are rigid)
- Hypothesis
  - Existing models can serve as **noisy data labelers**
  - Training a model with noisy labels will lead to **better performance** than any individual labeling process

Data

- HNLP SemEval 2015 Challenge
- 24,144 **unlabeled** discharge summaries
- Yields 4,210,512 **unlabeled** sentences
- Evaluation: 269 **labeled** discharge summaries

Approach

- **Generate noisy label set**
  - Pipe unlabeled sentences through CliNER, cTAKES
- Extract features from sentences
- Design and train models

Features

- Word count vectorization
- Hashing with 2^{12} features
- Embeddings

Models

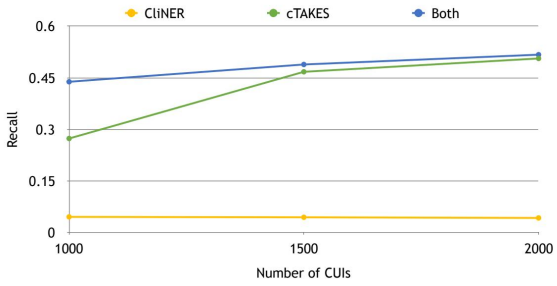
- Logistic Regression
- MLP (2 layer)
- CNN

Results

1. Adding **labeled data** from **multiple models** helps performance, with the number of CUIs fixed.

Model-Label	Recall
MLP-cTAKES	0.2732
MLP-CliNER	0.0453
MLP-Both	0.4382

2. **Scaling** the **number of CUIs** helps performance for a single model, with the label space fixed.



Future Work

- Add or replace entity extractors
- More feature and model exploration
- Expansion of training data
- Further work on embedding layers

References

Asma Ben Abacha and Pierre Zweigenbaum. 2011. Medical entity recognition: a comparison of semantic and statistical methods. https://dl.acm.org/citation.cfm?id=2002911

Murali Ravuri, Anitha Kannan, Geoffrey J. Tso, and Xavier Amatriain. Learning from the Experts. arXiv:1804.08033