Biomedical named entity recognition using transfer learning and deep language models

Valdimar Ágúst Eggertsson
valdimar.eggertsson@uni-potsdam.de

The goal was to create a named entity recognition model that can identify biomedical terms of interest within notes taken by scientists using the LabTwin app, in order to perform rudimentary information extraction.
Different BERT-based models were fine-tuned on the data set which consisted of 3000 labelled notes and rule-based methods were applied.

Background

State of the art results for most NLP tasks have been achieved using Google's BERT (bidirectional encoder representations from transformers).

Fine-tuning the deep neural network language model has shown great results for small data sets, for various tasks such as named entity recognition.

- **Tools and methods used:**
- Neural Language Models:
- SciSpacy (from Allen AI: SciBERT + Spacy)
- BioBERT (Biomedical BERT)
- Transfer learning / fine-tuning

- **Data**:
- Lab experiment documentation from scientists that are conducting pharmacological experiments, full of biology/chemistry jargon.

The models were trained on 4 entity classes:

cell line, protein, gene, science word

Example sentences:

Cell culture yesterday I splitted 293T AX123 to produce monoclonal cells.

Purification protein PMS 2-52 and PMLC 12 SDS page after purification over heparin column.

I infected one and a half day before RS with RV wild type, the Delta Plexin DC 2 mutant and the double mutant.

Main steps in the project:

1. **Label data using various databases and manual work**
2. **Fine-tune deep neural language model**
3. **Add look-up lists to the model to increase accuracy**

Method 1: Statistical approach (BioBERT and SciSpacy)

- With 3000+ labelled notes from Labtwin, powerful machine learning models based on BERT were fine-tuned,
- A few different models were trained using GPUs on Amazon Web Services and their accuracy was compared.
- SciSpacy outperformed BioBERT and was better to use since it is integrated into Spacy (NLP Python library).
- Accuracy of the model with the best results:

SciSpacy	Precision	Recall
Cell	0.72	0.88
Gene	0.87	0.47
Reagent	1.00	0.69
Science word	0.91	0.60

The problem of categorizing words from obscure technical notes is quite hard for a human.

So the results are rather promising, although not good enough to be automated and put into production.

Method 2: Gazetteers (look-up lists)

- A new feature in Spacy is the possibility to add rule-based predictions into the machine learning model.
- Given a good look-up lists of terms, it can increase precision and recall significantly.
- The downside is that good data is hard to find! And it is often proprietary and therefore illegal to use for commercial purposes (for Labtwin).
- By using a good database for cell lines, it was possible to achieve fantastic results:
  - 96.8% f1 score, 93.8% precision and 100% recall.
- Classifying protein names using the open database WikiData didn't work at all because the data set contains abbreviations and acronyms of protein names, different from the actual names.

Conclusions

- The machine learning model's accuracy was good, given the complexity of the task, but not good enough to be used where accuracy needs to be high.
- The gazetteers can improve the performance considerably, but unfortunately dictionaries that contain the terms are usually not available.
- Using a larger training set would increase the recall of the classifier, as the data set used was the minimal size that has given good results with BERT.
- The model could be used in production for semi-automatic named entity recognition (with human supervision, to remove the errors it makes).