Skweak: Weak Supervision Made Easy for NLP

Pierre Lison, **Jeremy Barnes** & **Aliaksandr Hubin**
Norwegian Computing Center, Department of Informatics, University of Oslo
& Department of Mathematics, University of Oslo
plison@nr.no *jeremycb@ifi.uio.no* *aliaksah@math.uio.no*

Why skweak?

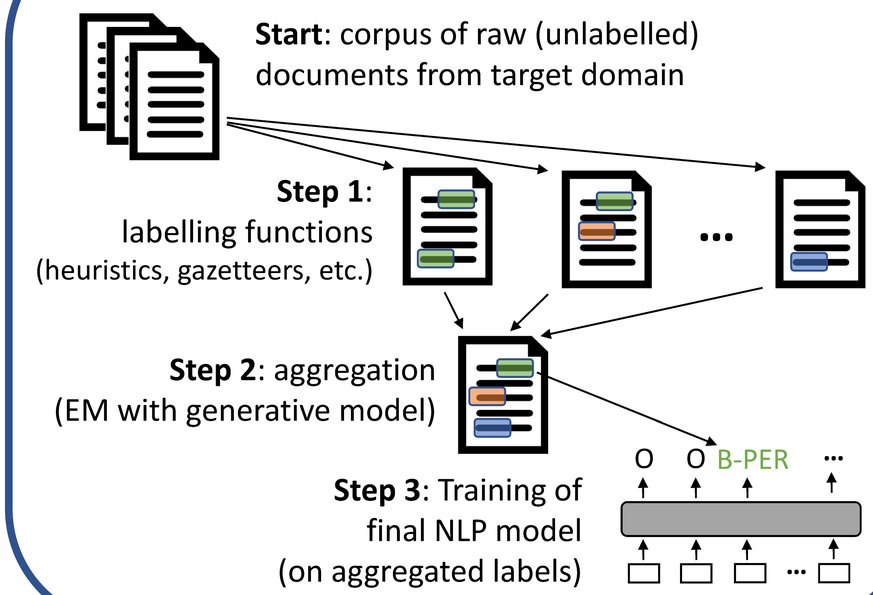
- Most common problem for NLP practictioners: Where can I get *labelled* data to train my ML model?
- **Weak supervision idea**: *automatically* annotate data using labelling functions, and then aggregate their results
- **Skweak** is a Python toolkit that makes it easy to define labelling functions, apply them on text documents, and aggregate their labels using a generative model
- Support for both *sequence labelling* & *text classification*
- Can also handle *underspecified labels*!

Labelling functions (LFs)

- LFs can take many forms: heuristics, gazetteers, ML models from other domains, linguistics constraints, crowd-workers, etc.
- LFs can be specialised to detect specific patterns/labels and ignore the rest (they can «abstain» from giving a prediction)
- In skweak, LFs take a Spacy Doc as input and return predictions:

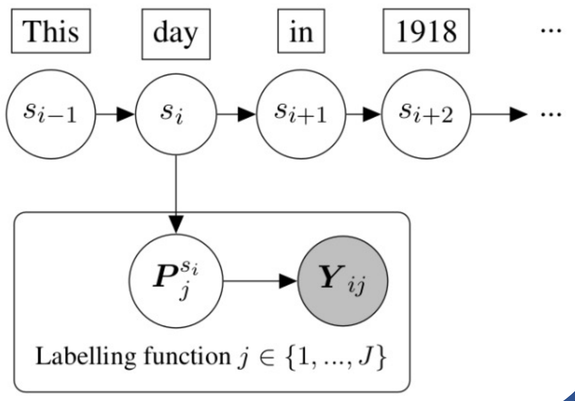
```
def money_detector(doc):
  """Searches for occurrences of MONEY entities in text"""

  for tok in doc[1:]:
      if (tok.text[0].isdigit() and tok.nbor(-1).is_currency):
      yield tok.i-1, tok.i+1, "MONEY"
```

Procedure



Aggregation model

- Generative model where the states are the «true» labels and are associated with multiple observations (one per LF)
- Transition & emission models estimated with Baum-Welch + weighting scheme to handle correlations between LFs



Experiments

- Demonstration on two NLP tasks:
- **NER** on MUC dataset, using 52 LFs (heuristics, gazetteers, out-of-domain NER models, doc-level constraints, etc.)

  ➔ Entity F_1=0.72 for NER model trained on skweak-aggregated labels (compared to 0.57 for majority voting on same LFs)
- 3-class **sentiment analysis** on $NoReC_{fine}$ dataset with 15 LFs (lexicons, heuristics, ML models, multilingual BERT)

  ➔Macro F_1 =0.49 for skweak-aggregated labels (0.40 for majority voting)

  ➔and 0.51 for NorBERT model fine-tuned on those skweak-aggregated labels