Deep Bayesian Semi-Supervised Active Learning for Sequence Labelling

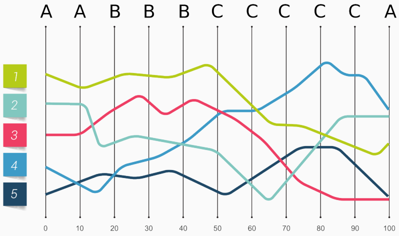
Tomáš Šabata, Juraj Eduard Páll and Martin Holeňa

16. 9. 2019

Czech Technical University in Prague
Faculty of Information technology
Department of Applied Mathematics

Motivation

Sequence labelling



- **Deep learning** has shown supreme results in sequence labelling.
  - Typically requires a large training data set.
  - Requires a series of token-level labels for a whole sequence to be available.
- **Goal:** Reduce labelling effort.
  - Use **active learning** to label the most informative sequences.
  - Use **semi-supervised learning** to avoid labelling of "easy" elements.

Proposed approach

Proposed approach

- Deep neural networks are typically over-confident $\Rightarrow$
  - Bayesian inference approximation using Monte Carlo Dropout.
    - Perform multiple inferences using different dropout masks to obtain $P_M^{MC}(y|x)$
- Adapt approaches of uncertainty sampling designed for sequence labelling.
- Use pseudo-labelling (self-training) to label elements in which the model is confident.

Utility function adaptations

- Monte Carlo approximation of token entropy:

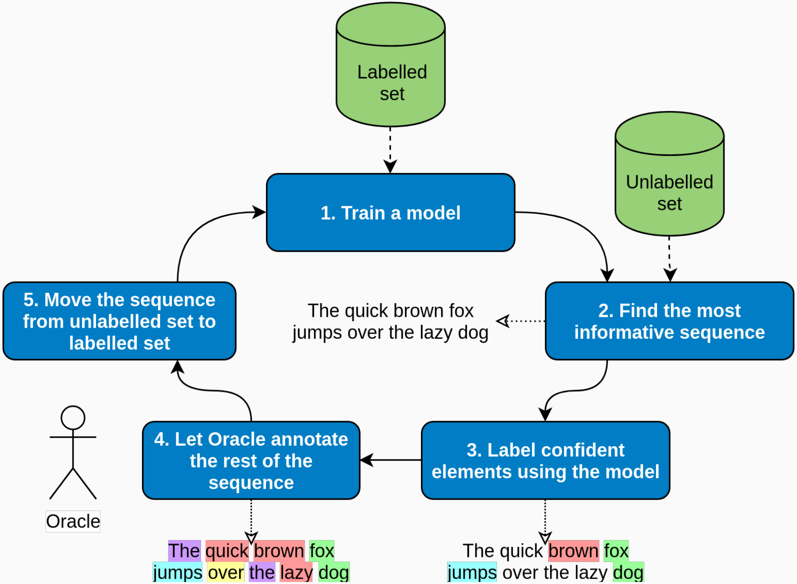
$$\phi_{\text{MC}}^{TE}(x) = -\frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K}\left(P_M^{\text{MC}}(y_l = k|x)\log P_M^{\text{MC}}(y_l = k|x)\right).$$

- Monte Carlo approximation of the sequence entropy:

$$\phi_{\text{MC}}^{NSE}(x) = -\frac{1}{C}\sum_{\hat{y}\in\mathcal{N}^{\text{MC}}} P_M^{MC}(\hat{y}|x)\log P_M^{MC}(\hat{y}|x),$$

where $\mathcal{N}^{\text{MC}} = \{y_1, y_2, ...\}$ is set of all sequences predicted by Monte Carlo sampling.

Semi-supervised active learning



Experimental setup

Experimental setup

- Natural language processing:
  - named entity recognition (NER),
  - part of speech tagging (POS),
  - chunking (CHUNK).
- Benchmark dataset CoNLL 2003 (english).
- Two state of the art models:
  - BI-LSTM-FCN
  - BI-LSTM-CRF
- Training data randomly split into labelled and unlabelled in 1:9 ratio.
- Experiments:
  - Active learning
    - limited number of tokens,
    - unlimited number of tokens.
  - Semi-supervised learning

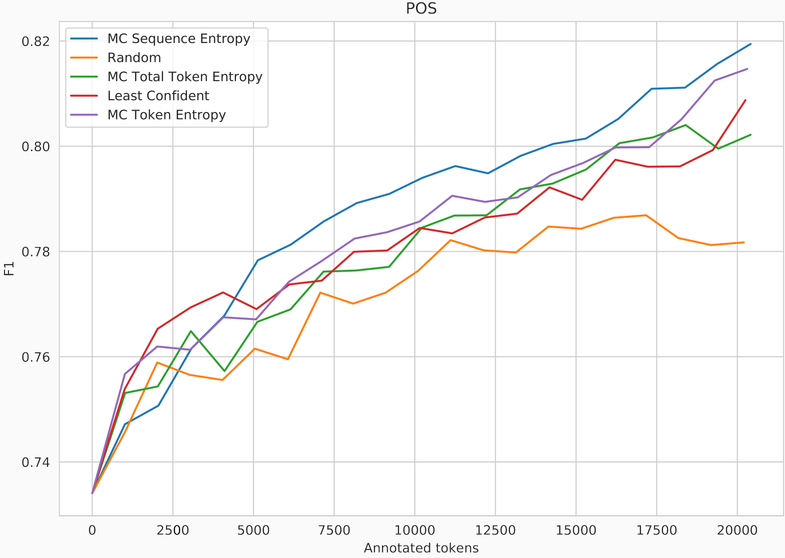
Results

Query strategies comparison

Task type	**NER**	**POS**
Random	76.5	78.1
Least Confident	76.9	80.8
MC Total Token Entropy	76.9	80.2
MC Sequence Entropy	**77.5**	**82.0**

Table 1: Query strategies comparison (F1 score)

Query strategies comparison



Semi-supervised learning

Task type	**NER**		**POS**	
Allowed error	**0%**	**1%**	**0%**	**1%**
Least confident	77.3	84.3	72.4	82.5
Token entropy	79.5	83.7	72.2	77.9
Total token entropy	75.5	83.5	71.2	81.6

Table 2: Relative amount of pseudo-labeled tokens - BI-LSTM-CRF

Summary and future steps

Summary

- Summary:
  - Two uncertainty sampling query strategies for deep neural network models.
  - Semi-supervised learning can rapidly reduce labelling effort for model with CRF on top.
- Future steps:
  - video processing,
  - study usability of query by committee and expected gradient length,
  - study other approaches to Bayesian recurrent neural networks.

Thank you for your attention.

Questions?

tomas.sabata@fit.cvut.cz