Projects

Benjamin Roth

Projects

General remarks

- Framework must be Keras or PyTorch.

Project 1: Affect Intensity in Tweets

- Semeval 2018 Task 1/Subtask 1
  `https://competitions.codalab.org/competitions/17333#learn_the_details-overview`
- Given:
  - Tweet
  - Emotion E (anger, fear, joy, or sadness)
- Task: Predict intensity of Emotion E
- TODO:
  1. Encode tweet with LSTM, encode emotion with one-hot vector
  2. Combine sentence and emotion representation, Dense layer(s), predict intensity
  3. Other ideas for extension?
  4. 2 Projects possible: Keras vs. Pytorch

Project 2: Emotion Classification

- Semeval 2018 Task 1/Subtask 5
  `https://competitions.codalab.org/competitions/17333#learn_the_details-overview`
- Given:
  - Tweet
- Task: classify it as 'neutral or no emotion' or as one, or more, of eleven given emotions that best represent the mental state of the tweeter.
- TODO:
  1. Multi-label classification
  2. compare lstm/cnn
  3. other ideas?

Project 3: Multilingual Sentiment

- Semeval 2018 Task 1/Subtask 3
- Given:
  - Tweet in one of selected language (English, Arabic, Spanish)
- Task: Predict sentiment polarity (positive/negative)
- TODO:
  1. Can one model be learned/applied for two languages?
  2. Use pre-trained multilingual word embeddings.
     `http://ruder.io/cross-lingual-embeddings/`

Project 4: Relation Classification

- own data
- Given:
  - Given Sentence with marked arguments + relation
    `President ARG1 and his wife ARG2 visited Honululu.`
    per:spouse?
- Task: Predict predict whether relation holds.
- TODO:
  1. Encode sentence with word embeddings and position embeddings
     `President/-1/-5 ARG1/0/-4 and/1/-3 his/2/-2 wife/3/-1 ARG2/4/0 visited/5/1 Honululu/6/2`
  2. CNN + Pooling $\Rightarrow$ relation prediction (softmax)
  3. other ideas? (e.g. relation as part of input, binary prediction)
- Related work:
  `http://www.aclweb.org/anthology/W15-1506`
  `http://cs.stanford.edu/~angeli/papers/2017-emnlp-tacred.pdf`
- More related projects are possible (e.g. piecewise convolutions, Semeval Task 7)

Project 5: Improving tagging with auxiliary output

- Given: Tagging task (a.g. ATIS data set)
  - Additionally to tag, predict previous and next word at each position (softmax).
- Task: Predict predict whether relation holds.
- TODO:
  1. Set up proper training and evaluation loss/metrics for tagging task (e.g. ATIS): accuracy for whole sequence, f-score, ignore padded positions
  2. Predict different outputs
  3. How to reduce complexity?
     - ... only top n words vocabulary
     - ... predict word clusters instead of words
     - ... hierarchical softmax / negative sampling
- Related work:
  `https://arxiv.org/abs/1707.05227`

Project 6: Domain Adaptation 1: irony detection

- Given: One larger data set (tweets, SemEval-2018 Task 3), one smaller data set (reddit comments, Kaggle) for irony detection.
- Task: Predict whether comments are ironic (e.g. LSTM+logistic regression). How can one domain (large data set) help prediction on another domain (smaller dataset)?
- TODO: compare different settings:
  1. Add data together, give different weights to instances from data A vs. data B.
  2. Pretrain with data A, continue training with data B.
  3. Train model for data A. Train model for data B, constrain it to be similar to model A. ($\Leftarrow$ I can help with that. Also a possible stand-alone topic.)
  4. Effect of Dropout/SpatialDropout1D: Make model more robust by removing words from the input during training.
  5. Effect of pre-trained word embeddings.
- up to 3 Projects - work needs to be distributed.

Project 7: Natural Language Generation

- Given: structured information about restaurants
- Task: generate natural language sentence (E2E challenge)
  `http://www.macs.hw.ac.uk/InteractionLab/E2E/`
- TODO:
  - Use code from NLG tutorial and apply to E2E challenge
    `http://course.fast.ai/lessons/lesson12.html`
  - Compare with baseline

Project 8: Two layer hierarchical Softmax

- Adapt word2vec code to use hierarchical softmax instead of negative sampling.
- TODO:
  a. Group words into $\sqrt{|V|}$ classes, by frequency.
  b. Each word is represented by two ids: class id, word id in that class.
     - ★ Use first softmax to predict class.
     - ★ Use second softmax to predict word id (parameters of second softmax depend on class).
- 1 project if applied to data from exercise sheet, 2 projects if applied to larger data (w. evaluation).

Projects proposed by participants

- Domain adaptation for financial tweets (Simon Schäfer).
- Argument extraction (Ivan Bilan).
- Retrieval-based dialogue system with LSTMs and PyTorch (Janina Nuber).
- Relation Classification:
  - Relation classification with positional embeddings and CNNs (Sabrina Jacob)
  - Domain adaptation for relation classification (Azada Rustamova)