Advanced Methods in NLP Lecturer: Jonathan Berant

Home Assignment 3: Tagging

Due Date: *May 14, 2019*

In this home assignment we will implement POS taggers. You are provided with the data and supporting code. **Note that the Penn Treebank dataset is licensed! It is illegal to make it public in any way!**

Your code should run properly on Python 2.7 on linux. It must run on `nova.cs.tau.ac.il` using `/usr/bin/python`. **Avoid any code updates**, besides in sections marked by `YOUR CODE HERE`. Solution which includes other modifications (such as function signature modification or redundant logs) will not be graded.

Create a zip file named `<id1>_<id2>.zip` (where id1 refers to the ID of the first student) and submit your solution on Moodle. Each group should submit the zip file only once. The zip file should include the code necessary to run your solution, as well as a written solution, and a text file including an e-mail of one of the students. Don't add the dataset to the zip file, and avoid changing the relative path of the data (in the code).

1 Data preprocessing

(a) As we saw in class, a common solution to the rare words problem is to pre-process the data and replace rare words with a category, or signature (e.g., numbers, dates, capitalization, prefixes, suffixes, etc. ...). Come up with good word categories and implement `replace_word` in `data.py`. The following paper, where this was done for named entity recognition, can be helpful: `http://people.csail.mit.edu/mcollins/6864/slides/bikel.pdf`. You can test the efficacy of your implementation by evaluating your "most frequent tag" baseline (next problem).

2 Most frequent tag baseline

(a) The most frequent tag baseline tags each word with its most frequent tag, as seen in the training set. Implement the most frequent tag baseline in `most_frequent.py`.

(b) Implement the evaluation procedure in `most_frequent.py` that measures the accuracy of the most frequent tag baseline on some dataset. Evaluate your tagger against the development set. What is your accuracy on the development set?

3 HMM tagger

(a) **MLE estimators:** Use the training data to estimate the transition probabilities q and emission probabilities e. Fill the implementation of the training algorithm in the function `hmm_train.py` in `hmm.py`.

(b) **Viterbi:** Implement the Viterbi algorithm (as described in slide 48 in Tagging presentation) in `hmm_viterbi` function in `hmm.py`. The algorithm receives a sentence to tag as input, the counts computed by the training procedure and the hyper-parameters λ_i. The algorithm returns the highest probability sequence of tags according to q and e. Recall that the estimates for q should be based on a weighted linear interpolation of $p(t_i|t_{i-1}, t_{i-2})$, $p(t_i|t_{i-1})$ and $p(t_i)$. Tune the hyper-parameters on the development set, and document the optimal λ_i values in your written solution.

Note: a straight-forward implementation of the Viterbi algorithm can be slow, so you should add some tag pruning (eliminating some tags for specific words). Training and evaluation (on dev set) should take up to a minute. Document your pruning policy in your written solution.

(c) Implement the evaluation procedure `hmm_eval` in `hmm.py` that measures the accuracy of the HMM tagger with Viterbi algorithm on some dataset. What is your accuracy on the development set?

(d) **Theoretical question 1:** Give an example for transition parameters Q, emission parameters E and a sentence s, such that the greedy inference algorithm does not result in the highest probability sequence. Your solution should include the probabilities calculations, as well as the samples on which the parameters were estimated.

(e) **Theoretical question 2:** Give an example for labeled data D and a sentence s, such that a second-order transition parameters Q_2 would give the highest probability to the correct labels, but a third-order transition parameters Q_3 would give the highest probability to other sequence of labels. Your solution should include the emission parameters E, and probabilities calculations.

4 Maximum Entropy Markov Model (MEMM) tagger

In this part you will implement the MEMM tagger (a locally-normalized log-linear model). The learning part is already given in the skeleton code using `scikit-learn`.

(a) **Feature engineering:** Implement features for your model. You should implement the features from Ratnaparkhi (1996) mentioned in class (`nlp_loglinear.pdf` file slide 52), but you can add more feature templates if you want.

(b) **Greedy inference:** Implement a greedy inference algorithm, where you tag a sentence from left to right with your trained model. Fill your implementation in the function `memm_greedy` in `memm.py`.

(c) **Viterbi:** Implement the Viterbi algorithm for MEMMs. The tag distribution should be inferred from the trained model. Fill your implementation in the function `memm_viterbi` in `memm.py`.

Note 1: prediction in this model is likely to be much slower than in the HMM model. You should consider optimizing your implementation by caching predictions, avoiding unnecessary feature extraction, etc. Training and evaluation (on dev set) should take up to 4 hours. Document any optimization you have performed in the written solution.

Note 2: As mentioned before, you should not change the code, besides in the marked sections. You can use the method `build_extra_decoding_arguments` to pass any additional argument to your decoding procedure.

(d) Implement the evaluation procedure in `memm.py`, that measures the accuracy of the MEMM tagger with Viterbi/greedy inference on some dataset. What is your accuracy on the development set?

(e) Sample errors from your best model and analyze them. What are common failure cases for your model. Where does it struggle? Summarize the results of your analysis in the written solution.