

CS5740: Assignment 3

GITHUB_REPOSITORY

Full Name
Net ID

Full Name
Net ID

TODO: Update your name and details above. Names must be presented in alphabetical ascending order by the last name. If not filled correctly, we will subtract 2pt

TODO: Make sure to use the standard notation when defined (<https://www.overleaf.com/read/kbrdcdbtvhgt>). Check for the updated version of the notation.

1 Introduction (5pt)

TODO: The first paragraph is to briefly describe task and data

TODO: The second paragraph should describe your main experiments and results, including mentioning the data you use

2 Data (5pt)

TODO: Describe the data you use, including how many examples are in the training, development, and test sets. Please also provide shallow statistics of your data, with at least the vocabulary size, document length (when relevant). It is best to report all the statistics, including counts, in a table. Please describe how you compute statistics that can be computed in different ways (i.e., vocabulary size). Describe how you pre-process the data, including how you treat casing, tokenization, and anything else that you did to the raw data before computing features from it. Explain your choices, and show the data statistics after pre-processing. If you are using a subset of the data, explain why and discuss tradeoffs (you will also need to back them with experiments).

3 Handling Unknown Words (15pt)

TODO: Provide a formal and detailed description of how you handle unknown words. Explain why your approach is a good balance between losing information by not considering the full identify

of the word and performance. Basically, why your approach makes sense. Use examples to illustrate the process you are using. Specify the parameters of your handling of unknown words, so you can easily refer to them later to experiment later.

4 Smoothing (10pt)

TODO: Provide a formal and detailed description of your smoothing technique. Explain the reasoning behind your approach. What kind of information are losing? Why does it still make sense for overall performance?

5 Implementation Details (5pt)

TODO: Briefly describe the implementation details. No need to copy details already specified in the assignment. Include any hyper-parameters the model has. If there are any optimizations that you introduced, this is the place to describe them. Did you have to do any special optimizations to make the method work in reasonable time/memory? Describe it here.

6 Experiments and Results

Test Results (3pt) **TODO:** Describe your test results from the leaderboard. It must be clear what setup you are using to generate them. This is best done in a table, and the table must be referred from the text. For the rest of the results you should use the development data. Always report both overall performance and on unknown words only.

Smoothing (5pt) **TODO:** Report your development experiments using the smoothing techniques that you used. Also compare to not using smoothing at all. Use a table to list the results,

refer to it from the text, and details your conclusions in the text. Make sure it's clear what conditions you experimented with.

Bi-gram vs. Tri-gram (5pt) TODO: Describe your results with bi-gram and tri-gram models. If you are ambitious, also replot four-gram performance (if you do it well, we will provide a small bonus). Use a table to list the results, refer to it from the text, and details your conclusions in the text. Make sure it's clear what conditions you experimented with.

Greedy vs. Viterbi vs. Beam (10pt) TODO: Describe your results comparing greedy, beam search, and Viterbi decoding. Use a table to list the results, refer to it from the text, and details your conclusions in the text. Make sure it's clear what conditions you experimented with. Discuss how often the greedy decoder finds the exact (optimal) solution. Report the sensitivity of your performance to the size of the beam.

7 Analysis

Error Analysis (7pt) TODO: Qualitative analysis of selected failure examples. Show and discuss error examples from your development set. You must identify certain classes of errors and use the examples to illustrate them. This is often best to show in a table.

Confusion Matrix (5pt) TODO: Show a confusion matrix and discuss the most common confusions. If you can't format the matrix to fit, you can show the most relevant subset of it. It's recommended to use a heatmap-based visualization. Please also look at confusion between related tags (i.e., confusion between verbs or different types of nouns).

8 Conclusion (3pt)

TODO: Brief conclusion summarizing findings (from both numerical results and qualitative analysis).