Final Project

L665/B659: Applying Machine Learning Techniques in CL; Sandra Kuebler

Spring 2016

Your task is to build a Word Sense Disambiguation system based on the SensEval-3 **Romanian Lexical Sample** data set. The task description can be found at `http://www.senseval.org/senseval3`. Your task is to develop a classifier for the following 5 words:

1. accent.n
2. citi.v
3. delfin.n
4. oficial.a
5. val.n

Here is your task description:

1. All files and the scorer are available on canvas.
2. Select a machine learner. Note that you have to motivate the choice in your final paper.
3. Design a feature set. Use either feature set from the assignments, but you can add any features you think would be helpful.
4. Design one classifier for each word and optimize your features and parameters (non-exhaustively). Use the official scorer to evaluate your results.
5. Then add POS tagging features from the files. Does that improve your results?
6. Then perform a semi-supervised experiment. This means, you need to determine for the examples in the unlabeled data to which sense they belong. Then add all or some of the examples to the training set. Can you get an improvement over the supervised setting?
7. Write a scientific paper, with an introduction, a related work section, a section on methodology, a section on results, including a discussion, and a conclusion. The paper should be 6-8 pages long. The paper should explain your research question in the introduction. The related work section should give an overview of work in WSD and more specifically on WSD for Romanian and on semi-supervised approaches for WSD. In the methodology section, describe your choice of classifier and motivate the latter. Then describe the data set and the feature set and motivate the representation that you chose. Then, explain your approach to the semi-supervised task. Explain your results. Make sure that I understand what exactly you did. You need to give enough information to make your experiemnts replicable, but you do not need to include code or implementational details.
8. For WSD approaches to Romanian, check out the SemEval-3 publications: `http://www.aclweb.org/anthology/W/W04/#0800`.
9. The paper should follow the ACL 2014 style. Stylefiles are available from `http://acl2014.org/CallforPapers.htm` under ACL 2014 Style Files.
10. You will only submit the paper, no code, no system output.

Start early!!! And talk to me if you run into problems.

DUE DATE: May 4, 11:59pm. Please submit your paper via canvas.