CSE842: Natural Language Processing
Final Project Guidelines

February 25, 2017

This project provides an opportunity for you to work on something that you are interested in. You can choose any topics related to natural language processing or the application of NLP techniques to your own research areas. Please work in a group of two. *You should include a brief statement about the work division and responsibility of each group member.* You are also free to use any available software or resources given that proper acknowledgement is provided.

The project will be judged by the originality of your problem, the survey of related work on that problem, the justification of your methods, the thoroughness in your analysis and evaluation, and the quality of your report. If the problem you plan to work on has been extensively studied in the previous work, do not get panic if your results are not as good as the published results. In this case, you should provide reasonable explanation on why your approach does not work well and what might be done to improve your approach.

The ACL Anthology (`http://aclweb.org/anthology-new/`) makes available a large collection of research papers on various topics of NLP. NLP/Computational Linguistic is a very conference oriented field. The top conferences particularly relevant for this class include: the Annual Meeting of the Association for Computational Linguistics (ACL), the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), and the conference on Empirical Methods in Natural Language Processing (EMNLP).

Many resources and data corpora are made available by Linguistic Data Consortium (LDC). You have access to some of these resources at: `http://ldc.lib.msu.edu/`. You will need to use a machine with "msu" as the domain address to access these data. You will also need to be on campus to access the above website.

1 Potential Topics

The following potential topics are for your reference. When considering these topics, you are encouraged to think outside the box. For example, you can use the data made available for these topics to address some new interesting problems which are not traditionally approached with these data.

1.1 Lexical Semantics

Many datasets are made available for different tasks on lexical semantics (e.g., word sense disambiguation) though the SENSEVAL competition (`http://www.senseval.org/`). You can check out each of these tasks and choose one that you may find interesting. There should also be papers from those who participated in the evaluations either at the ACL Anthology or directly linked from the website about each task. Instead of directly working on one of these problems (i.e., shared tasks by

SENSEVAL), you may consider using the data from these tasks to address an interesting linguistic question other than the original task itself.

A subset of the Senseval data can be found in the directory: `/user/cse842/Corpora/Senseval`. If you are interested in word sense disambiguation for the medical domain, check out the following website: `http://wsd.nlm.nih.gov/`.

If you decide to work on a shared task in Senseval, what I expect from you is: have a complete understanding of the state-of-the-art (e.g., recent approaches to WSD after 2010), study all the papers related to that particuarly task, make slides, develop a model and evaluate it using an exsiting dataset.

1.2 Semantic Role Labeling

1.2.1 Verb SRL

You are encouraged to apply both supervised and semi-supervised learning approaches for semantic role identification. A seminal paper about this task is described in Gildea and Jurafsky (2001, Check the class website for the paper). Since then, many supervised approaches have been developed to label semantic roles. An international forum is also established to evaluate different approaches. If you are interested in this topic, I will encourage you to take a look at the following websites: `http://www.lsi.upc.edu/~srlconll/`. You can follow the shared task description to formulate your problem and develop potential solutions.

The Propbank is available to you at: `/user/cse842/Corpora/Propbank/Data`

1.2.2 Nominal SRL

Nominal semantic role labeling has received increasing attention in recent years given the availability of NomBank. You are encouraged to explore Nombank and identify interesting topics that are related to nominal semantic roles (not necessarily restricted to nominal semantic role labeling). Nombank can be found: `http://nlp.cs.nyu.edu/meyers/NomBank.html`.

Another potential topic is to explore implicit argument for semantic role labeling. The data is available here: `http://lair.cse.msu.edu/projects/semanticrole.html`

1.3 Abstract Meaning Representation

This resource is mainly for semantic parsing (`http://amr.isi.edu/`).

As MSU LDC membership has expired, we are not able to get the entire dataset at this point. But we did download a small dataset of 1,562 sentences from The novel *Little Prince*. You can directly download this data from the AMR website above or from `/user/cse842/Corpora/AMR`.

1.4 Reference/Coreference Resolution

The task of reference resolution / coreference resolution/ pronoun resolution is to identify the referents (or coreferences) to referring expressions (e.g., pronouns) in natural language utterances/texts. It is one of important problem for language understanding in both monologue (e.g., text) and dialog. A considerable amount of work has been done about this topic. Many papers can be found online. Same as other research topics, you can also come up with some very specific questions that you may be interested in investigating.

Data annotated for a portion of Wall Street Journal can be found at: `http://ldc.lib.msu.edu/rawcds/cd439/bbn-pcet/`

Data annotated for spoken conversation can be found at: `/user/cse842/Corpora/ReferenceResolution/Data/TRAIN`

1.5 Other Potential Topics

- Syntactic processing: POS tagging, parsing, prepositional phrase attachment. PennTreebank data is in: `/user/cse842/Corpora/PennTreebank` (this is the Release 3).
- Named entity identification: Develop or compare different approaches to the task of named entity identification. The data available for this topic can be found at: `http://ldc.lib.msu.edu/rawcds/cd439/bbn-pcet/`.
- Discourse processing: use Penn Discourse Treebank to promote understanding or develop models to automatically identify discourse relations between sentences. The Penn Discourse Treebank can be downloaded from: `http://ldc.lib.msu.edu/rawcds/cd433/pdtb_v2/`.
- You may get more ideas from students's final projects for the NLP class at Stanford University: `http://nlp.stanford.edu/courses/cs224n/`

2 Project Proposal

You need to submit a two page project proposal (the ACL format, see below) for approval. In your proposal, you should specify the following:

1. The problem you are trying to address.
2. The proposed approaches.
3. The data set that will be used.
4. Any previous work on this topic? Provide references.
5. The plan of implementing your approach and milestones for the rest of the semester.
6. The composition of the group and the responsibilities of each group member.

The proposal is due on **March 22**.

3 Project Presentation

You will present your project to the class on **April 24 or April 26**. I will provide details on the presentation (e.g., the expected length of presentation and presentation schedule, etc.) in early April.

4 Project Final Report

The final report of the project is due on **May 5, 11:59** through handin.

Your report should be around eight pages (including figures and references) using the ACL format. The format can be found in the class directory `/user/cse842/ACLformat/`. It should include the following section:

1. Introduction: The problem statement
2. Related Work
3. Detailed description of your approach
4. Evaluation: comparison between different configurations of your approach, compare results with other approaches (if some early results with the same dataset exist).
5. Discussion of your results
6. Conclusion
7. References