Project discussion
10/4

CS 585, Fall 2016

Introduction to Natural Language Processing

http://people.cs.umass.edu/~brenocon/inlp2016

Brendan O'Connor

College of Information and Computer Sciences
University of Massachusetts Amherst

announcements

- HW3 due Friday
- Next Tuesday 10/11: no class (Monday schedule day)
- Next Friday: project proposals!

Project

http://people.cs.umass.edu/~brenocon/inlp2016/project.html

- Either *build* natural language processing systems, or *apply* them for some task.
- Use or develop a dataset. Report empirical results or analyses with it.
- Different possible areas of focus
  - Implementation & development of algorithms
  - Defining a new task or applying a linguistic formalism
  - Exploring a dataset or task

Project

```
oct          2  3  4  5  6  7  8  (hw3)
             9 10 11 12 13 14 15  proposals due
            16 17 18 19 20 21 22
            23 24 25 26 27 28 29  (hw4)
nov         30 31  1  2  3  4  5  midterm
             6  7  8  9 10 11 12  progress due
            13 14 15 16 17 18 19  (hw5)
            20 21 22 23 24 25 26  [t.giving]
dec         27 28 29 30  1  2  3
             4  5  6  7  8  9 10  (hw6)
            11 12 13 14 15 16 17  poster session
            18 19 20 21 22 23 24  final report
```

Proposal: 2-4 page document outlining the problem, your approach, possible dataset(s) and/or software systems to use. Must cite and briefly describe at least **two** pieces of relevant prior work (research papers). Describe scope of proposed work.

Mentor meetings

Progress report: Longer document with preliminary results

Poster session: 12/13

Final report

- Scheduling for poster session?
  - 2:30 -- 4:30
  - 3:00 -- 5:00
- Groups of 1-3: we encourage size 2
  - We expect more work with more team members

NLP Research

- All the best publications in NLP are open access!
  - Conference proceedings: ACL, EMNLP, NAACL (EACL, LREC...)
  - Journals: TACL, CL
  - “aclweb”: ACL Anthology-hosted papers http://aclweb.org/anthology/
  - NLP-related work appears in other journals/conferences too: data mining (KDD), machine learning (ICML, NIPS), AI (AAAI), information retrieval (SIGIR, CIKM), social sciences (Text as Data), etc.
- Reading tips
  - Google Scholar
    - Find papers
    - See paper’s number of citations (imperfect but useful correlate of paper quality) and what later papers cite it
    - [... or SemanticScholar...]
  - For topic X: search e.g. [[nlp X]], [[aclweb X]], [[acl X]], [[X research]]...
  - Authors’ webpages
    find researchers who are good at writing and whose work you like
  - Misc. NLP research reading tips:
    http://idibon.com/top-nlp-conferences-journals/

A few examples

- Detection tasks
  - Sentiment detection
  - Sarcasm and humor detection
  - Emoticon detection / learning
- Structured linguistic prediction
  - Targeted sentiment analysis (i liked __ but hated __)
  - Relation, event extraction (who did what to whom)
  - Narrative chain extraction
  - Parsing (syntax, semantics, discourse...)
- Text generation tasks
  - Machine translation
  - Document summarization
  - Poetry / lyrics generation (e.g. recent work on hip-hop lyrics)
  - Text normalization (e.g. translate online/Twitter text to standardized English)

- End to end systems
  - Question answering
  - Conversational dialogue systems (hard to eval?)
- Predict external things from text
  - Movie revenues based on movie reviews ... or online buzz? http://www.cs.cmu.edu/~ark/movie$-data/
- Visualization and exploration (harder to evaluate)
  - Temporal analysis of events, show on timeline
  - Topic models: cluster and explore documents
- Figure out a task with a cool dataset
  - e.g. Urban Dictionary

Sources of data

- All projects must use (or make, and use) a textual dataset. Many possibilities.
  - For some projects, creating the dataset may be a large portion of the work; for others, just download and more work on the system/modeling side

- SemEval and CoNLL Shared Tasks:
  dozens of datasets/tasks with labeled NLP annotations
  - Sentiment, NER, Coreference, Textual Similarity, Syntactic Parsing, Discourse Parsing, and many other things...
  - e.g. SemEval 2015 ... CoNLL Shared Task 2015 ...
  - https://en.wikipedia.org/wiki/SemEval (many per year)
  - http://ifarm.nl/signll/conll/ (one per year)

- General text data (not necessarily task specific)
  - Books (e.g. Project Gutenberg)
  - Reviews (e.g. Yelp Academic Dataset https://www.yelp.com/academic_dataset)
  - Web
  - Tweets

Tools

- Tagging, parsing, NER, coref, ...
  - Stanford CoreNLP http://nlp.stanford.edu/software/corenlp.shtml
  - spaCy (English-only, no coref) http://spacy.io/
  - Twitter-specific tools (ARK, GATE)
- Many other tools and resources

  tools ... word segmentation ... morph analyzers ...
  resources ... pronunciation dictionaries ... wordnet, word embeddings, word clusters ...
- Long list of NLP resources

  https://medium.com/@joshdotai/a-curated-list-of-speech-and-natural-language-processing-resources-4d89f94c032a