Project discussion
(2/23)

CS 690N, Spring 2017
Advanced Natural Language Processing
http://people.cs.umass.edu/~brenocon/anlp2017/

Brendan O’Connor
College of Information and Computer Sciences
University of Massachusetts Amherst
Project

- Create, apply, and experiment with a natural language processing system for some task
- Use or develop a dataset. Report empirical results.
- *Compare* to previous work
  - Pre-existing system, or
  - Reported results on same dataset, or
  - Reimplementation of previous work (may be a large part of your project, if this is complex)
  - ... and explain *why* differences are happening!
- Different possible areas of focus
  - Implementation & development of algorithms
  - Defining a new task or applying a linguistic formalism
  - Exploring a dataset or task
Project

- **Proposal**: due March 10
  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.

- **Progress report**:
  - Lit review
  - Preliminary results

- **Poster session**: May 4

- **Final report**

- Groups of 1-3
  - We expect more work with more team members
NLP Research

• All the best publications in NLP are open access!
  • Conferences: ACL, EMNLP, NAACL (EACL, LREC...)
  • Journals: TACL, CL
  • ML publications also important: NIPS, ICLR, ICML, JMLR
  • “aclweb”: ACL Anthology-hosted papers
    http://aclweb.org/anthology/
  • Other NLP-related work: data mining (KDD), 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
A few examples

- Detection tasks
  - Sentiment detection
  - Sarcasm and humor detection
  - Emoticon detection / learning
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...)
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...)
- Model exploration
  - Topic models
  - Structured prediction models
  - Attention networks
  - Neural network architectures (CNN, LSTM, etc.)
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...)
- Model exploration
  - Topic models
  - Structured prediction models
  - Attention networks
  - Neural network architectures (CNN, LSTM, etc.)

- Text generation tasks
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...)

• Model exploration
  • Topic models
  • Structured prediction models
  • Attention networks
  • Neural network architectures (CNN, LSTM, etc.)

• Text generation tasks
  • Machine translation
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...)
- Model exploration
  - Topic models
  - Structured prediction models
  - Attention networks
  - Neural network architectures (CNN, LSTM, etc.)

- Text generation tasks
  - Machine translation
  - Document summarization
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...)
- Model exploration
  - Topic models
  - Structured prediction models
  - Attention networks
  - Neural network architectures (CNN, LSTM, etc.)

- Text generation tasks
  - Machine translation
  - Document summarization
  - Poetry / lyrics generation (e.g. recent work on hip-hop lyrics)
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...)
• Model exploration
  • Topic models
  • Structured prediction models
  • Attention networks
  • Neural network architectures (CNN, LSTM, etc.)
• Text generation tasks
  • Machine translation
  • Document summarization
  • Poetry / lyrics generation (e.g. recent work on hip-hop lyrics)
• End to end systems
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...)
- Model exploration
  - Topic models
  - Structured prediction models
  - Attention networks
  - Neural network architectures (CNN, LSTM, etc.)

- Text generation tasks
  - Machine translation
  - Document summarization
  - Poetry / lyrics generation (e.g. recent work on hip-hop lyrics)
- End to end systems
  - Question answering
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...)

- **Model exploration**
  - Topic models
  - Structured prediction models
  - Attention networks
  - Neural network architectures (CNN, LSTM, etc.)

- **Text generation tasks**
  - Machine translation
  - Document summarization
  - Poetry / lyrics generation (e.g. recent work on hip-hop lyrics)

- **End to end systems**
  - Question answering
  - Predict external things from text
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...)
- Model exploration
  - Topic models
  - Structured prediction models
  - Attention networks
  - Neural network architectures (CNN, LSTM, etc.)

- Text generation tasks
  - Machine translation
  - Document summarization
  - Poetry / lyrics generation (e.g. recent work on hip-hop lyrics)
- End to end systems
  - Question answering
- Predict external things from text
  - Movie revenues based on movie reviews ... or online buzz? http://www.cs.cmu.edu/~ark/movie$-data/
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...)

- Model exploration
  - Topic models
  - Structured prediction models
  - Attention networks
  - Neural network architectures (CNN, LSTM, etc.)

- Text generation tasks
  - Machine translation
  - Document summarization
  - Poetry / lyrics generation (e.g. recent work on hip-hop lyrics)

- End to end systems
  - Question answering

- 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)
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...)
- Model exploration
  - Topic models
  - Structured prediction models
  - Attention networks
  - Neural network architectures (CNN, LSTM, etc.)

- Text generation tasks
  - Machine translation
  - Document summarization
  - Poetry / lyrics generation (e.g. recent work on hip-hop lyrics)
- End to end systems
  - Question answering
- 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
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...)
- Model exploration
  - Topic models
  - Structured prediction models
  - Attention networks
  - Neural network architectures (CNN, LSTM, etc.)
- Text generation tasks
  - Machine translation
  - Document summarization
  - Poetry / lyrics generation (e.g. recent work on hip-hop lyrics)
- End to end systems
  - Question answering
- 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
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...)
- Model exploration
  - Topic models
  - Structured prediction models
  - Attention networks
  - Neural network architectures (CNN, LSTM, etc.)
- Text generation tasks
  - Machine translation
  - Document summarization
  - Poetry / lyrics generation (e.g. recent work on hip-hop lyrics)
- End to end systems
  - Question answering
- Predict external things from text
  - Movie revenues based on movie reviews ... or online buzz? [http://www.cs.cmu.edu/~ark/movie$-data/](http://www.cs.cmu.edu/~ark/movie$-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
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...)
- Model exploration
  - Topic models
  - Structured prediction models
  - Attention networks
  - Neural network architectures (CNN, LSTM, etc.)
- Text generation tasks
  - Machine translation
  - Document summarization
  - Poetry / lyrics generation (e.g. recent work on hip-hop lyrics)
- End to end systems
  - Question answering
- 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/sigmll/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

- Some projects may want to use external tools for preprocessing or to build on top of
- Tagging, parsing, NER, coref, ...
  - spaCy (English-only, no coref) [http://spacy.io/](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](https://medium.com/@joshdotai/a-curated-list-of-speech-and-natural-language-processing-resources-4d89f94c032a)