Projects & classification misc.

CS 490A, Fall 2020

Applications of Natural Language Processing

https://people.cs.umass.edu/~brenocon/cs490a_f20/

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

- HW1 questions?
- Today:
  - Projects
  - Some classification topics
    - Overfitting & Regularization
    - Logistic Regression (if time)

Project

https://people.cs.umass.edu/~brenocon/cs490a_f20/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

Proposal, due in 2.5 weeks (10/2): 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: Longer document with preliminary results (due late Oct / early Nov)

Class presentations (~last week of classes)

Final report (due end of finals)

- Groups of 2-3
  - We expect more work with more team members

Formulating a proposal

- What is the ***research question***?
- What's been done before?
- What experiments will you do?
- How will you know whether it worked?
  - If data: held-out accuracy
  - If no data: manual evaluation of system output. Or, annotate new data

NLP Research

- Check on your textbooks!
- 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, NeurIPS, ICLR), AI (AAAI), information retrieval (SIGIR, CIKM), social sciences (Text as Data), etc.
- Reading tips
  - Google Scholar (or Semantic Scholar)
    - Find papers
    - See paper's number of citations (imperfect but useful correlate of paper quality) and what later papers cite it
  - 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

Bayes rule

$$p(y|x) = \frac{\overbrace{p(y)}^{\text{Prior}}\ \overbrace{p(x|y)}^{\text{Lik}}}{\underbrace{p(x)}_{\text{Normalizer}}}$$

$$\begin{tikzcd} Y \text{ label} \arrow[d] \\ X \text{ text} \end{tikzcd}$$

Sum Rule of Prob

$$p(x) = \sum_{y' \in Dom(Y)} p(x, Y=y')$$

$$= \sum_{y' \in Dom(Y)} p(y')\ p(x|y')$$

$$\text{BR numerators} = \left(p(pos)\, p(\vec{w}|pos) ,\ p(neg)\, p(\vec{w}|neg) \right)$$

$$\text{BR with norm} = \left(\frac{p(pos)\, p(w|pos)}{p(pos)\, p(w|pos) + p(neg)\, p(w|neg)} ,\ \ldots \right)$$

Big things

1. Logis Regr (no Bayes Rule)

"discrim classif."

2. Word Embeddings

"cat" vs. "cats" vs. "dog"