# L90 Practical, Part II

## Overview of Natural Language Processing

Simone Teufel

2016/11/25

# Sentiment Classification, Part 2

- We assume you now have (a) working baseline system(s)
- You can now automatically classify a given review from IMDB into Negative or Positive
- You used machine learning (NB, maybe SVM) on bags of words (maybe POS, bigrams etc)
- But this does not take much of the linguistic processing you encountered in L90 into account.
- Today – get you started on extending this to a more sophisticated system
- . . . and do good science while you're at it

# Goal of this Practical – "good science"

- Getting high numerical results isn't everything – neither in this practical nor in science in general.
- Good science means:
  - an interesting research question
  - sound methodology
  - insightful analysis (something non-obvious)
- First part – choose your research question

# Possible Research Questions

- You know from L90 some techniques how syntactic information can be automatically determined.
- Can this information capture additional meaning (beyond words) in a way that demonstrably helps with this task?
- Should we determine other units but verbs or bigrams? "Adjective + Noun" units?
- The clearest case of semantic information is negation.
  - Is negation present in a sentence (at all)?
  - Which unit in the sentence is negated?
  - If so, how can the negation information be gainfully used?

# More Possible Research Questions

- Wordnet is an ontology listing lexical relations between (mainly) nouns – how could this help (maybe in connection with a sentiment lexicon?)
  - One relationship between adjectives is antonymy, eg. "boring" / "exciting"
- Turning to discourse effects – why would anaphora resolution possibly help?
- Pragmantic effects:
  - Plot Description vs. Evaluation
  - Advice to the reader
  - Violation of expectations; "underdog" story lines
  - Rhetorical questions? Irony?

# Don't overdo the mixing and matching

- One research question is enough.
- In fact, exploring one research question well and thorougly often makes for the best scientific experiment.

# Null results

- What to do if you have tried and tried and you cannot beat your best baseline system?
- It's possible that your idea just didn't pan out, and that what you have achieved simply is the best possible outcome
- . . . if you can show that you have used the right methodology. For instance:
  - Don't vary more than one variable at a time when measuring effects
  - Test for significance before making any claims that something is better than something else

# Unseen data (Repetition slide)

- For testing, only ever use truly unseen data.
- How could data be "seen" (and thus tainted) by your system?
    - In ML systems, parameters are directly estimated from the data ("training corpus")
        - Cross-validation is a way of pretending to close our eyes and "unsee" data
    - Even in symbolic systems, characteristics of the data can taint the test because they can "enter the experiment" via the brain of the developer
        - Careful – once the developer has looked at the data ("development corpus"), they can never really "forget" what they have seen.
        - You might want to make sure your test data is unreadable to yourself.

# Writing the report

- No more than 4000 words; with word count.
- Submission 12 January 2016, noon
- Must include a pointer to running code on Mphil machines (your account)

# Writing tips

- Make it look like a paper (including formatting)
- Use scientific language
  - Eradicate all forms of colloquial language
  - Mimic the author's voice in published papers
- Introduction: pretend this is not an L90 assignment but your own idea
- Reader has no pre-knowledge
- Describe your methodology appropriately
  - Not too detailed (otherwise you look like a beginner)
  - Enough detail to allow somebody to reimplement your solution
  - Technical terms: use them – define them first
- Describe your numerical results (after your methods, clearly separated).
- Analyse your numerical results, using your brain and the English language.