Class: NERPOS (chapter 22)

November 16, 2010

Admistrivia

- proposals / abstracts?

Naming of names

- What I tell my undergrads when teaching statistics:
  - In statistics everything has two names (i.e. RMSE / s)
  - And every name refers to two things (mean = X-bar, expectation)
  - Its stupid! Just deal with it.

- English is similar
NER

- Problem: Identify who/what a reference is to.

- Subproblem: What is a reference?
  - Dr. Who should be parsed together as one word, as should
    * Apple Inc.
    * Mr. president
    * Barack Obama
  - Usually done as marking Start, Inside and Outside of references. (You don’t need an end, since you can compute it from other changes.)
  - Like a POS problem sort of.

- Subproblem: Where does the reference point?
  - Barack, Obama, Mr. President, Commander and chief, are all the same reference
  - Apple, “Apple, Inc”, Apple Computing are all the same

Mild alternative

- Similar to the pronoun reference problem

- That “that,” that that “that” is referring to is not that that that that refers to.
• Search “had had had” on wiki.

• In general, “he” or “she” is just like “Apple, Inc” It is a reference to something that needs to be disambiguated

**Parsing NER**

• Good tricks are language specific
  – Capitalizations in written English (not useful in spoken)
  – Inc, Mr, Ms, etc in English
  – All caps, or weird caps (eBay)
  – Case markers in other languages?

• Entropy rules
  – Two items that go together way to frequently might be a single name
  – Curious problem: How would you statistically figure out “words” in continuously run together text?

• Nick names
  – If you find “Microsoft Inc” which your rules tell you is an entity, then look for words that co-occur with it, “MS,” “M$,” “evil empire” etc
  – Morphology of names might be useful: Microsoft Inc can be shorten to Microsoft.
POS for NER

Two possible ways of marking up NER:

- Broad brush: person, place or thing?
- Laser focus: SS# of person

Broad brush is like a POS tag:

- People are “active” agents. So they go with active verbs.
- locations are objects. They go with “to Washington” kind of constructs.

What it looks like at the end

page 734 for marked up example:

“Citing high fuel prices, [ORG United Airlines] said [time Friday ] [ORG: it ] has increased fairs by ... [ORG United ], a unit of [ORG UAL Corp.], said the increase...

What is the value?

Tom Mitchel’s talk (last Tuesday):

- Extracting facts, like “United is a unit of UAL”
- Extracting events, “raised fair.”
• Extracting complex relationships:
  – called Script / frame
  – Shank’s stuff
  – Fare-rase attempt:
    * Lead airline: united airlines
    * amount: $6
    * efectgive date: 2006-10-26
    * Followers: American Airlines, Delta

• Generally extracting triple:
  – A owns B
  – A is mayor of B
  – etc

• Automatically fill a database with such “knowledge”

**Need everything we have talked about so far**

• REG expressions are useful for taking apart names

• POS tagging with “ORG”

• Parsing to identify tripple relationships
• Synonym finding to get alternate form of triples

• (OK, no speech!)

Statistics problem version

Discuss wiki disambiguation. Or leave for another day.