HW #8 Notes

March 4, 2015

WordNet-based WSD

- Perform word sense disambiguation of probe word
  - In context of word set

  - Line news,lot,joke,half,hour,show,cast,brainstorm
  - Tie jacket, suit

  - An answer key is provided
    - Don't expect to get them all right!

Implementation

- Implement a simplified version of Resnik's
  - "Associating Word Senses with Noun Groupings"
  - Select a sense for the probe word, given group
    - Rather than all words as in the algorithm in the paper
  - For each pair (probe, $noun_i$)
    - Loop over sense pairs to find MIS, similarity value (v)
    - Update each sense of probe descended from MIS, with v
  - Select highest scoring sense of probe

Components

- Similarity measure:
  - IC:
  - /corpora/nltk/nltk-data/corpora/wordnet_ic/ic-brown-resnik-add1.dat
  - NLTK accessor:
    - wnic = nltk.corpus.wordnet_ic.ic('ic-brown-resnik-add1.dat')
  - Note: Uses WordNet 3.0

Components

- >>> from nltk.corpus import *
  >>> brown_ic = wordnet_ic.ic('ic-brown-resnik-add1.dat')
  >>> wordnet.synsets('artifact')
  [Synset('artifact.n.01')]
- >>> wordnet.synsets(‘artifact’)[0].name
- ‘artifact.n.01’
  >>> artifact = wordnet.synset('artifact.n.01’)
- from nltk.corpus.reader.wordnet import information_content
- >>> information_content(artifact, brown_ic)
  2.436960793329339 1

Components

- Hypernyms:
  - >>>wn.synsets('artifact')[0].hypernyms()
  - [Synset('whole.n.02')]
- Common hypernyms:
  - >>> hat = wn.synsets('hat')[0]
  - >>> glove = wn.synsets('glove')[0]
  - >>> hat.common_hypernyms(glove)
  - [Synset('object.n.01'), Synset('artifact.n.01'), Synset('whole.n.02'), Synset('physical_entity.n.01'), Synset('entity.n.01')]

Components

- WordNet API
  - NLTK: **Strongl**y suggested
  - Others exists, but no warranty
- http://www.nltk.org/howto/wordnet.html
- http://www.nltk.org/api/nltk.corpus.reader.html#module-nltk.corpus.reader.wordnet

Note

- You can use supporting functionality, e.g.:
  - Common_hypernyms, full_hypernyms, etc
- You can NOT just use the built-in resnik_similarity, etc
  - If you're unsure about acceptability, just ask...