CS224N Section 3
Corpora, etc.
Pi-Chuan Chang,
Friday, April 25, 2008
Some materials borrowed from Bill's notes in 2006:
http://www.stanford.edu/class/cs224n/handouts/cs224n-section3-corpora.txt

---

Proposal for Final project due two weeks later (Wednesday, 5/7)
Look for interesting topics?

→ go through the syllabus : http://www.stanford.edu/class/cs224n/syllabus.html
→ final projects from previous years : http://nlp.stanford.edu/courses/cs224n/
→ what data / tools are out there?
→ collect your own dataset

---

1. LDC (Linguistic Data Consortium)
   - http://www.ldc.upenn.edu/Catalog/

2. Corpora@Stanford
   - http://www.stanford.edu/dept/linguistics/corpora/
   - Corpus TA: Anubha Kothari
   - The inventory: http://www.stanford.edu/dept/linguistics/corpora/inventory.html
   - Some of them are on AFS; some of them are available on DVD/CDs in the linguistic department

3. http://nlp.stanford.edu/links/statnlp.html

4. ACL Anthology
   - http://aclweb.org/anthology-new/

5. Various Shared Tasks
   - CoNLL (Conference on Computational Natural Language Learning)
     - 2006: Multi-lingual Dependency Parsing
     - 2005, 2004: Semantic Role Labeling
     - 2003, 2002: Language-Independent Named Entity Recognition
     - 2001: Clause Identification
     - 2000: Chunking
     - 1999: NP bracketing
   - Machine Translation shared tasks: 2008, 2007, 2006, 2005
   - Pascal Challenges (RTE, etc)
   - TREC (IR)
   - Senseval (Word sense disambiguation)
   - …

Parsing

Most widely-used : Penn Treebank

1. English: (LDC99T42) Treebank-3 (see Bill's notes)
2. In many different languages, like Chinese (CTB6.0), Arabic

Other parsed corpora

1. Switchboard (spoken)
2. German: NEGRA, TIGER, Tueba-D/Z
    - There's an ACL workshop on German Parsing this year…
3. CoNLL 2006 Shared task : Multi-lingual Dependency Parsing

Parsers

1. Stanford Parser (English, Chinese, German and Arabic)
    - Stanford parser online (English and Chinese)
2. Charniak's parser
3. Collin's parser
4. Bikel's parser
5. CMU Link Parser
6. MINIPAR
7. http://nlp.stanford.edu/fsnlp/probparse/

Part-of-Speech

POS tags from Treebanks

British National Corpus (BNC) –

100m words, wide sample of British English: newspapers, books, letters

Tools:

http://nlp.stanford.edu/links/statnlp.html#Taggers

Language Modeling

Data:

1. Texts you can collect from anywhere can be used to train LMs.
2. Google Web 1T 5-gram (LDC2006T13)
   (AFS has 1,2,3-grams. For 4 and 5 gram)

Tools:

1. SRILM – SRI Language Modeling Toolkit
2. IRST LM Toolkit
3. CMU-Cambridge Statistical Language Modeling toolkit

Named Entity Recognition

Data:

1. Message Understanding Conference (MUC) – MUC-1 ~ MUC-7
2. CoNLL shared task
    - 2002, 2003 : Language-Independent Named Entity Recognition (I), (II)

Tools:

http://nlp.stanford.edu/links/statnlp.html#NER

Coreference

Data:

1. MUC-6 and MUC-7

Word Sense Disambiguation

Senseval

Semantic Role Labeling

Data

1. CoNLL shared task 2004, 2005
2. PropBank - Predicate-argument relations were added to the syntactic trees of the Penn Treebank.
3. FrameNet (English and also other languages)

Semantic Role Labeling Demo from UIUC

Question Answering (QA)

TREC -- competition , Question Answering Track
http://trec.nist.gov/data/qamain.html

Summarization

Document Understanding Conferences (DUC)

Lexical Semantics

WordNet (web interface)

```
150,000 nouns, verbs, adjectives, adverbs
grouped into "synsets" with glosses, sentence frames
includes hypernym (kind-of) hierarchy rooted at 'entity'
also antonyms, holonyms & meronyms, polysemy
good tutorial:
http://www.brians.org/Projects/Technology/Papers/Wordnet/
neat visual interface: http://www.visualthesaurus.com/?vt
Problems with WordNet:
 - fine-grained senses
 - sense ordering sometimes funny (see "airline")
```

Textual Entailment

Recognising Textual Entailment (RTE) challenges

Email datasets

The Enron corpus
/afs/ir/data/linguistic-data/Enron-Email-Corpus/maildir/skilling-j/
http://www.cs.cmu.edu/~einat/datasets.html

TREC Spam track
http://trec.nist.gov/data/spam.html

Corpus tools

tgrep2 (see Bill's notes)

Machine Learning tools

Stanford Classifier – conditional loglinear (aka maximum entropy) model

Weka – a collection of machine learning algorithm for data mining tasks. Naïve Bayes, decision trees, kNNs, SVM, etc.

Mallet -- an integrated collection of Java code useful for statistical natural language processing, document classification, clustering, information extraction, and other machine learning applications to text

SVM tools: libsvm , SVM-light