Computational Linguistics Tutorial

*How to do your own Treebank experiments*

Sandiway Fong
These slides and some of the other files referenced are available on my homepage

http://dingo.sbs.arizona.edu/~sandiway#research

Statistical Natural Language Parsing

- Linguistics Colloquium Slides (September 2007) [pdf format]
- dbparser.zip (22MB) subset of Dan Bikel's parser files for use with Windows XP
- bikel2tbv.pl (Perl script) converts Bikel parser output to treebankviewer format
- wsjtby.zip (6.8MB) Wall Street Journal section of the Penn Treebank in treebankviewer format
From last week’s Cognitive Science Master’s Seminar

Statistical Natural Language Parsing (NLP) is getting a lot of attention recently (Chater & Manning 2006):

– “Probabilistic methods are providing new explanatory approaches to fundamental cognitive science questions of how humans structure, process and acquire language.”
• From last week’s Cognitive Science Master’s Seminar
• Statistical Natural Language Parsing (NLP) is getting a lot of attention recently

(Chater & Manning 2006):
– “Probabilistic models can account for the learning and processing of language, while maintaining the sophistication of symbolic models.”
Context

• From last week’s Cognitive Science Master’s Seminar

Scorecard

• **Attained knowledge of language**
  – systems don’t seem to get any linguistic condition people have noticed: don’t mimic people

• **Robustness**
  – *milk* example
    • extremely fragile despite statistical nature and training over a large dataset

• **Mirror human limits on acquisition**
  – these systems just as easily learn to ‘speak’ in a way that people never do: learn unnatural languages
• The milk example:
  – John mixed the milk with the cookies – 35.3
  – John mixed the cookies with the milk – 35.2

only milk forces low attachment

same results with different verbs, e.g. drank etc.
24 sentences with milk in it, but only one with PP attachment

- **Perturbation experiment:**
  
  #5212

  What happens if we change just one sentence in the training set of 39,832?
Context

- **Perturbation experiment:** alter sentence #5212 and re-train

**Result:**

parses with milk now all attach high after altering 1 out of 39,832
How to do your own Treebank experiments?

- **Bad news**
  - most people use Windows
  - most research software assume a Unix-based OS
    - completely mutually unintelligible platforms
  - there is virtually no help in porting to Windows

- **Good news**
  - *I’ll show you how to get around all this*

*MacOSX and Linux users can still use the procedures given here*
Resources and Tools

• **Checklist:** *all completely free software*
  
  – **Penn Treebank**
    
    **source:** *cdrom from UA Library*
  
  – **Dan Bikel’s Multilingual Statistical Statistical Parsing Engine**
    
    (written using Java and shell scripts)
    
    **source:** [http://www.cis.upenn.edu/~dbikel/software.html](http://www.cis.upenn.edu/~dbikel/software.html)
  
  – **Java SDK**
    
    **source:** *java.sun.com*
  
  – **lots of Unix-style commands**
    
    **source:** a Linux install in a separate partition, a Mac running OSX, or cgywin install under Windows
Resources and Tools

• **Penn Treebank**

  *LDC charges $3150, it’s free from UA Main Library*
Penn Treebank

- *From last week’s Cognitive Science Master’s Seminar*

- **How is it used?**
  - One million words of 1989 Wall Street Journal (WSJ) material
  - Nearly 50,000 sentences (49,208)
  - **25 sections** (0–24)
  - Sections 2–21 contain 39,832 sentences
  - Section 23 (2,416 sentences) is held out for evaluation

- **Standard practice**

  - Training: sections 2–21
  - Evaluation: section 23

  - 90% statistical NL parsers
Penn Treebank

- contents of directory TREEBANK_3/parsed/mrg/wsj/
- sections (00–24):

  
  ![Diagram showing sections 00 to 24 with 'training set' and 'test set' highlighted]

  training set

  test set
Penn Treebank

- contents of directory TREEBANK_3/parsed/mrg/wsj/23/
- **data files** (wsj_2300.mrg – wsj_2399.mrg):

```
wsj_2300.mrg
wsj_2301.mrg
wsj_2302.mrg
wsj_2303.mrg
wsj_2304.mrg
wsj_2305.mrg
wsj_2306.mrg
wsj_2307.mrg
wsj_2308.mrg
wsj_2309.mrg
wsj_2310.mrg
wsj_2311.mrg
```
Penn Treebank

Unfortunately, this structure is not the format we want...

• For training we need a single concatenated file containing all of the .mrg files from sections 02–21.
  cd TREEBANK_3/parsed/mrg/wsj
  cat 0[2-9]/*.mrg 1[0-9]/*.mrg 2[01]/*.mrg > /tmp/wsj-02-21.mrg

• Windows: type 00/*.mrg > wsj-00.mrg for section 00 and repeat the file merge for each section, then merge them all into one file

• For treebankviewer, we need the complete Treebank (sections 00–24) in one big file.
  temporary solution: use preformatted Treebank wsjtbv.zip (6.8MB) from my homepage
  write some Perl script to convert sexps -> Prolog terms
treebankviewer

- **download**: http://dingo.sbs.arizona.edu/~sandiway/treebankviewer/

  - [Treebank Viewer](http://dingo.sbs.arizona.edu/~sandiway/treebankviewer/)

  download wsjtbv.zip contains sentence file wsj.txt and Prolog tree file wsj.pl

  written partially in TCL/TK, for Windows users it requires ActiveState TCL (free)
Java

- Java: platform-independence
- Java SDK: available free from java.sun.com

- start -> Run -> cmd
- command line interpreter (cmd)
- java -version
Dan Bikel’s Multilingual Statistical Parsing Engine

**uses platform-independent Java but is difficult to install and use on Windows**

- **solution**: use dbparser.zip (21MB) from my homepage
  - includes 3 files: dbparser.jar collins.properties wsj-02-21.obj.gz
- To run the parser on file input.lisp, type all on one line:
  ```bash
  java -Xms400m -Xmx400m -cp dbparser.jar -Dparser.settingsFile=collins.properties
danbikel.parser.Parser -is wsj-02-21.obj.gz -sa input.lisp
  ```
Dan Bikel’s Multilingual Statistical Parsing Engine

- **Input format** (sexp):
  - *one sentence per line*
  - `((word1 (pos1)) (word2 (pos2)) ... (wordN (posN)))`

- **Milk example**:
  `((I (PRP)) (drank (VBD)) (milk (NN)) (with (IN)) (water (NN)))`
  `((I (PRP)) (drank (VBD)) (water (NN)) (with (IN)) (milk (NN)))`

- Part-of-speech tags to be supplied by user or a separate tagger program (MXPOST)
- Lots of web references for Treebank tags:
  e.g. [http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html](http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html)
Dan Bikel’s Multilingual Statistical Parsing Engine

- **Output file:**
  - is named `input.lisp.parsed` given `input file input.lisp`

- **Milk example input:**
  ```lisp
  ((I (PRP)) (drank (VBD)) (milk (NN))
   (with (IN)) (water (NN)))
  ((I (PRP)) (drank (VBD)) (water (NN))
   (with (IN)) (milk (NN)))
  ```

- **Milk example output:**
  - one parse (on a single line) per input sentence
  ```lisp
  (S (NP (PRP I)) (VP (VBD drank) (NP (NP (NN milk)) (PP (IN with) (NP (NN water))))))
  (S (NP (PRP I)) (VP (VBD drank) (NP (NN water)) (PP (IN with) (NP (NN milk))))))
  ```
Dan Bikel’s Multilingual Statistical Parsing Engine

- Let’s assume the existence of the single file \texttt{wsj-02-21.mrg} representing sections 02–21 of the treebank.

- We can manually modify this file per example 5212 and save it as \texttt{wsj-02-21a.mrg}, representing sections 02–21 of the treebank with one modification.
Dan Bikel’s Multilingual Statistical Parsing Engine

- **file compare** (fc):
Dan Bikel’s Multilingual Statistical Parsing Engine

• To re-train the parser on treebank file wsj-02-21a.mrg, representing sections 02–21 newly modified, type all on one line:
  - java -Xms800m -Xmx800m -cp dbparser.jar -Dparser.settingsFile=collins.properties danbikel.parser.Trainer -i wsj-02-21a.mrg -o wsj-02-21a.observed.gz -od wsj-02-21a.obj.gz

• result is a derived-data file wsj-02-21a.obj.gz which now can be used for parsing
Dan Bikel’s Multilingual Statistical Parsing Engine

• Run input.lisp again with the new statistics wsj-02-21a.obj.gz
  1. (S (NP (PRP I)) (VP (VBD drank) (NP (NN milk)) (PP (IN with) (NP (NN water))))))
  2. (S (NP (PRP I)) (VP (VBD drank) (NP (NN water)) (PP (IN with) (NP (NN milk))))))

• Resulting parses:
  – high attachment only for milk examples
    1. (S (NP (PRP I)) (VP (VBD drank) (NP (NN milk)) (PP (IN with) (NP (NN water)))))
    2. (S (NP (PRP I)) (VP (VBD drank) (NP (NN water)) (PP (IN with) (NP (NN milk)))))

Mission complete!
Bikel and treebankviewer

Pretty Printing

- **download:**
  bikel2tbv.prl
  a bikel to
  treebankviewer
  converter
  from my homepage
  (Perl script)
  Sentence File: input.lisp
  Prolog Tree File: input.pl

- **cmd:** perl bikel2tbv.prl input.lisp.parsed > input.pl

assuming you have Perl installed

*free download:* Active State Perl for Windows users