Introduction to Natural Language Processing

Steven Bird Ewan Klein Edward Loper

University of Melbourne, AUSTRALIA

University of Edinburgh, UK

University of Pennsylvania, USA

May 16, 2007

Questions

- How do we write programs to manipulate natural language?
- What questions about language could we answer?
- How would the programs work?
- What data would they need?
- First: what do they look like?

Searching Pronunciation Dictionary

```
>>> from nltk_lite.corpora import cmudict
>>> from string import join
>>> for word, num, pron in cmudict.raw():
...     stress_pattern = join(c for c in join(pron) if c in "012")
...     if stress_pattern.endswith("1 0 0 0 0"):
...         print word, "/", join(pron)
ACCUMULATIVELY / AH0 K Y UW1 M Y AH0 L AH0 T IH0 V L IY0
AGONIZINGLY / AE1 G AH0 N AY0 Z IH0 NG L IY0
CARICATURIST / K EH1 R AH0 K AH0 CH ER0 AH0 S T
CIARAMITARO / CH ER1 AA0 M IY0 T AA0 R OW0
CUMULATIVELY / K Y UW1 M Y AH0 L AH0 T IH0 V L IY0
DEBENEDICTIS / D EH1 B EH0 N AH0 D IH0 K T AH0 S
DELEONARDIS / D EH1 L IY0 AH0 N AA0 R D AH0 S
FORMALIZATION / F AO1 R M AH0 L AH0 Z EY0 SH AH0 N
GIANNATTASIO / JH AA1 N AA0 T AA0 S IY0 OW0
HYPERSENSITIVITY / HH AY2 P ER0 S EH1 N S AH0 T IH0 V AH0 T IY0
IMAGINATIVELY / IH2 M AE1 JH AH0 N AH0 T IH0 V L IY0
INSTITUTIONALIZES / IH2 N S T AH0 T UW1 SH AH0 N AH0 L AY0 Z AH0 Z
INSTITUTIONALIZING / IH2 N S T AH0 T UW1 SH AH0 N AH0 L AY0 Z IH0 NG
MANGIARACINA / M AA1 N JH ER0 AA0 CH IY0 N AH0
SPIRITUALIST / S P IH1 R IH0 CH AH0 W AH0 L AH0 S T
SPIRITUALISTS / S P IH1 R IH0 CH AH0 W AH0 L AH0 S T S
SPIRITUALISTS / S P IH1 R IH0 CH AH0 W AH0 L AH0 S S
SPIRITUALISTS / S P IH1 R IH0 CH AH0 W AH0 L AH0 S
SPIRITUALLY / S P IH1 R IH0 CH AH0 W AH0 L IY0
UNALIENABLE / AH0 N EY1 L IY0 EH0 N AH0 B AH0 L
UNDERKOFFLER / AH1 N D ER0 K AH0 F AH0 L ER0
```

Minimal Sets from Lexicon

```
>>> from nltk_lite.corpora import shoebox
>>> from nltk_lite.utilities import MinimalSet
>>> length, position, min = 4, 1, 3
>>> lexemes = [field[1].lower() for entry in shoebox.raw('rotokas')
...    for field in entry if field[0] == 'lx']
>>> ms = MinimalSet()
>>> for lex in lexemes:
...     if len(lex) == length:
...         context = lex[:position] + '_' + lex[position+1:]
...         target = lex[position]
...         ms.add(context, target, lex)
>>> for context in ms.contexts(3):
...     for target in ms.targets():
...         print "%-4s" % ms.display(context, target, "-"),
...     print
kasi -    kesi kusi kosi
kava -    -    kuva kova
karu kiru keru kuru koru
kapu kipu -    -    kopu
karo kiro -    -    koro
kari kiri keri kuri kori
kapa -    kepa -    kopa
kara kira kera -    kora
kaku -    -    kuku koku
kaki kiki -    -    koki
```

Modelling Text Genres

```
>>> from nltk_lite.corpora import genesis
>>> from nltk_lite.probability import ConditionalFreqDist
>>> from nltk_lite.utilities import print_string
>>> cfdist = ConditionalFreqDist()
>>> prev = None
>>> for word in genesis.raw():
...     word = word.lower()
...     cfdist[prev].inc(word)
...     prev = word
>>> words = []
>>> prev = 'lo,'
>>> for i in range(99):
...     words.append(prev)
...     for word in cfdist[prev].sorted_samples():
...         if word not in words:
...             break
...     prev = word
>>> print_string(join(words))
lo, it came to the land of his father and he said, i will not be a
wife unto him, saying, if thou shalt take our money in their kind,
cattle, in thy seed after these are my son from off any more than all
that is this day with him into egypt, he, hath taken away unawares to
pass, when she bare jacob said one night, because they were born two
hundred years old, as for an altar there, he had made me out at her
pitcher upon every living creature after thee shall come near her:
yea,
```

Exploring Syntax

```
>>> from nltk_lite.corpora import treebank
>>> from string import join
>>> def vp_conj(tree):
...     if tree.node == 'VP' and len(tree) == 3 and tree[1].leaves() == ['but']:
...         return True
...     else:
...         return False
>>> for tree in treebank.parsed():
...     for vp1,conj,vp2 in tree.subtrees(vp_conj):
...         print join(child.node for child in vp1), "*BUT*", join(child.node for child in vp
VBP ADVP-TMP PP-PRD PP *BUT* VBP VP
VBZ VP *BUT* VBZ NP PP-CLR
PP-TMP VBZ VP *BUT* VBD ADVP-TMP S
VBZ SBAR *BUT* VBZ SBAR
VBD SBAR *BUT* VBD RB VP
VBD SBAR *BUT* VBD S
VBP NP-PRD *BUT* VBP RB ADVP-TMP VP
VBN PP PP-TMP *BUT* ADVP-TMP VBN NP
MD VP *BUT* VBZ NP SBAR-ADV
VBD ADVP-CLR *BUT* VBD NP
VBN NP PP *BUT* VBN NP PP SBAR-PRP
VBD NP *BUT* MD RB VP
VBD NP PP-CLR *BUT* VBD PRT NP
VBZ S *BUT* MD VP
```

The Richness of Language

- basic needs and lofty aspirations; technical know-how and flights of fantasy
- ideas are shared over great separations of distance and time

1. Overhead the day drives level and grey, hiding the sun by a flight of grey spears. (William Faulkner, *As I Lay Dying*, 1935)
2. When using the toaster please ensure that the exhaust fan is turned on. (sign in dormitory kitchen)
3. Amiodarone weakly inhibited CYP2C9, CYP2D6, and CYP3A4-mediated activities with Ki values of 45.1-271.6 μM (Medline)
4. Iraqi Head Seeks Arms (spoof headline, `http://www.snopes.com/humor/nonsense/head97.htm`
5. The earnest prayer of a righteous man has great power and wonderful results. (James 5:16b)
6. Twas brillig, and the slithy toves did gyre and gimble in the wabe (Lewis Carroll, *Jabberwocky*, 1872)
7. There are two ways to do this, AFAIK :smile: (internet discussion archive)

Disciplines Studying Language

1. linguistics
2. translation
3. literary criticism
4. philosophy
5. anthropology
6. psychology
7. law
8. hermeneutics
9. forensics
10. telephony
11. pedagogy
12. archaeology
13. cryptanalysis
14. speech pathology

Language and the Internet

- unprecedented volume of information:
  mostly unstructured text
- 8 Tb books in 2003
- 24 hours of scientific literature would take 5 years to read
- fraction of work/leisure time spent navigating this information
- a great challenge for natural language processing
- despite success of web search engines, we need skill, knowledge, and luck to answer the following questions:
  1. *What tourist sites can I visit between Philadelphia and Pittsburgh on a limited budget?*
  2. *What do expert critics say about Canon digital cameras?*
  3. *What predictions about the steel market were made by credible commentators in the past week?*
- requires a combination of language processing tasks, e.g. information extraction, inference, and summarisation

The Promise of NLP

- importance in scientific, economic, social and cultural arenas
- growing rapidly as its theories and methods are deployed in new technologies
- therefore a wide range of people should have a working knowledge of NLP
  - academia: humanities computing, corpus linguistics, computer science, artificial intelligence
  - industry: HCI, business information analysis, web software development
- the goal of the book is to open the field of NLP to a broad audience.

NLP and Intelligence

- long-standing challenge to build intelligent machines
- chief measure of machine intelligence has been linguistic: Turing test
- research on spoken dialogue systems, also MT
  — integrated NLP systems which future users would regard as highly intelligent
- Example human-machine dialogue illustrates a typical application:

  S: How may I help you?
  U: When is Saving Private Ryan playing?
  S: For what theater?
  U: The Paramount theater.
  S: Saving Private Ryan is not playing at the Paramount theater, but it's playing at the Madison theater at 3:00, 5:30, 8:00, and 10:30.

NLP and Intelligence (cont)

- today's systems limited to narrowly defined domains
- couldn't ask above system for other information, e.g.:
  - driving instructions
  - details of nearby restaurants
- to add such support we would have to:
  - store the required information
  - incorporate suitable questions and answers into the system
- common-sense reasoning vs business logic
- need to make progress on natural linguistic interaction without recourse to this unrestricted knowledge and reasoning capability

Language and Symbol Processing

- origin of the idea that natural language could be treated computationally: philosophy of language work in early 1900s, to reconstruct mathematical reasoning using logic
- language as a formal system
- three further developments:
  1. formal language theory
  2. symbolic logic
  3. principle of compositionality
- more recent developments:
  1. data-intensive NLP
  2. machine learning in NLP
  3. evaluation-led methodologies
- many interesting philosophical issues (see book)
- key: balancing act between symbolic and statistical approaches

Web as Corpus: Absolutely vs Definitely

Google hits	**adore**	**love**	**like**	**prefer**
absolutely	289,000	905,000	16,200	644
definitely	1,460	51,000	158,000	62,600
ratio	198/1	18/1	1/10	1/97

- useful information for statistical language models
- statistical evidence for binary-valued features in lexical items

Python: Key Features

- simple yet powerful, shallow learning curve
- object-oriented: encapsulation, re-use
- scripting language, facilitates interactive exploration
- excellent functionality for processing linguistic data
- extensive standard library, incl graphics, web, numerical processing
- downloaded for free from `http://www.python.org/`

Python Example

```
import sys
for line in sys.stdin.readlines():
    for word in line.split():
        if word.endswith('ing'):
            print word
```

1. whitespace: nesting lines of code; scope
2. object-oriented: attributes, methods (e.g. `line`)
3. readable

Comparison with Perl

```
while (<>) {
    foreach my $word (split) {
        if ($word =~ /ing$/) {
            print "$word\n";
        }
    }
}
```

1. syntax is obscure: *what are:* `<> $ my split` ?
2. "it is quite easy in Perl to write programs that simply look like raving gibberish, even to experienced Perl programmers" (Hammond *Perl Programming for Linguists* 2003:47)
3. large programs difficult to maintain, reuse

What NLTK adds to Python

NLTK defines a basic infrastructure that can be used to build NLP programs in Python. It provides:

- Basic classes for representing data relevant to natural language processing
- Standard interfaces for performing tasks, such as tokenization, tagging, and parsing
- Standard implementations for each task, which can be combined to solve complex problems
- Extensive documentation, including tutorials and reference documentation

Installing Python and NLTK

1. Install Python, Numeric
2. Install NLTK-Lite, NLTK-Lite-Corpora
3. Optional: Matplotlib, WordNet
4. Set environment variable `NLTK_LITE_CORPORA`

For detailed instructions, see:

- http://nltk.sourceforge.net/install.html
- CDROM: /install.html