Natural Language Processing
First Week

Why study NLP?

necessary to many useful applications:
1, Information retrieval,
2, Information extraction,
3, Filtering,
4, Spelling and grammar checking,
5, Automatic text summarization,
6, Machine translation,

Who needs NLP?

Too many texts to manipulate
1, On Internet
2, E-mails
3, Various corporate documentation
Too many languages
1, 39000 languages and dialects

Applications of NLP

Text-based: processing of written texts
(ex. Newspaper articles, e-mails, Web pages…)
1, Text understanding/analysis (NLU)
IR, IE, MT, …
2, Text generation (NLG)
3, CLIME (Computerized Legal Information Management and Explanation)
4, Dialog-based systems (human-machine communication)

Brief history of NLP

1940s - 1950s Foundational Insights

1,Automata, finite-state machines & formal languages (Turing, Chomsky, Backus & Naur)
2, Probability and information theory (Shannon)
3, Noisy channel and decoding (Shannon)

1960s - 1970s

Symbolic: Linguists & Computer Scientists
1, Transformational grammars (Chomsky, Harris)
2, Artificial Intelligence (Minsky, McCarthy)
3, Theorem Proving, heuristics, general problem solver (Newell & Simon)

Stochastic: Statisticians & Electrical Engineers

1, Bayesian reasoning for character recognition

2, Authorship attribution

3, Corpus Work

1970s - 1980s

1, Stochastic approaches

2, Logic-based / Rule-based approaches

3, Scripts and plans for NL understanding of “toy worlds”

4, Discourse modeling (discourse structures & coreference resolution)

Late 1980s - 1990s Rise of probabilistic models

1, Data-driven probabilistic approaches (more robust)

2, Engineering practical solutions using automatic learning

3, Strict evaluation of work

Why study NLP Statistically?

1, Up to about 10 years, NLP was mainly investigated using a rule-based approach.

2, But:

2.1, Rules are often too strict to characterize people’s use of language (people tend to stretch and bend rules in order to meet their communicative needs.)

2.2, Need (expert) people to develop rules (knowledge acquisition bottleneck)

3, Statistical methods are more flexible & more robust

Tools and Resources Needed

1, Probability/Statistical Theory:

1.1, Statistical Distributions, Bayesian Decision Theory.

2, Linguistics Knowledge:

2.1, Morphology, Syntax, Semantics, Pragmatics…

3, Corpora:

3.1, Bodies of marked or unmarked text to which statistical methods and current linguistic knowledge can be applied in order to discover novel linguistic theories or interesting and useful knowledge to build applications.

What Statistical NLP can do

1, Seeks to solve the acquisition bottel neck by automatically learning preferences from corpora (ex, lexical or syntactic preferences).

2, Offers a solution to the problem of ambiguityand "real" data because statistical models
1, are robust
2, generalize well
3, behave gracefully in the presence of errors and new Data.

Second Week

Some standard corpora

1, Brown corpus
1.1, 1 million words
1.2, Tagged corpus (POS)
1.3, Balanced (representative sample of American English in the 1960-1970) (different genres)
2, Lancaster-Oslo-Bergen (LOB) corpus
2.1, British replication of the Brown corpus
3, Susanne corpus
3.1, Free subset of Brown corpus (130 000 words)
3.2, Syntactic structure
4, Penn Treebank
4.1, Syntactic structure
4.2, Articles from Wall Street Journal
5, Canadian Hansard
5.1, Bilingual corpus of parallel texts

What to do with text corpora?

Count words

1, Count words to find:
1.1, What are the most common words in the text?
1.2, How many words are in the text?
word tokens vs word types
1.3, What is the average frequency of each word in the text?

What's a word anyways?

I have a can opener; but I can't open these cans.
1, how many words?
2, Word form
1.1, inflected form as it appears in the text
1.2, can and cans ... different word forms
2, Lemma
2.1, a set of lexical forms having the same stem, same POS and same

meaning
2.1, can and cans … same lemma
3, Word token:
3.1, an occurrence of a word
3.2, I have a can opener; but I can't open these cans. 11 word tokens (not counting punctuation)
4, Word type:
4.1, a different realization of a word
4.2, I have a can opener; but I can't open these cans. 10 word types (not counting punctuation)

An example

1, Mark Twain's Tom Sawyer
1.1, 71,370 word tokens
1.2, 8,018 word types
1.3, tokens/type ratio = 8.9 (indication of text complexity)
2, Complete Shakespeare work
2.1, 884,647 word tokens
2.2, 29,066 word types
2.3, tokens/type ratio = 30.4

Thired week work

Using a Corpus

1, To approximate the probability distribution of language events, we use a training corpus
2, Statistical NLP seeks to automatically learn lexical and structural preferences from corpora.

Corpus

1, Large database of text & speech
2, Many types of text corpora exist
2.1, plain text, domain specific, tagged, parsed, parallel bilingual,
3, Major suppliers:
3.1, Linguistic data Consortium (LDC) -- www.ldc.upenn.edu
3.2, European Language resources Associations (ELRA) --www.icp.grnet.fr/ELRA
4, To derive the needed probabilities, a corpus needs to be
4.1, large
4.2, A representative sample of the population of interest

Low-Level Formatting Issues

1, Junk formatting & content

1.1, Removal of typesetter codes (ex. HTML tags), diagrams, tables, foreign words etc.

1.2, Also other problems if data was retrieved through OCR (unrecognized words)

2, Uppercase and Lowercase

2.1, should we keep the case or not?

2.2, “the”, “The” and “THE” should all be treated the same?

2.3, but in “George Brown” and “brown dog”, “brown” should be treated separately…

Finding Tokens and Sentences

Tokenization

1, divide the input text into units (called tokens)each token is either a word or something else (ex. A number or a punctuation mark)

2, Mark sentence boundaries most sentences end with ‘.’, ‘?’ or ‘!’ can be confused by abbreviations

What is a Sentence?

1, Something ending with a ‘.’, ‘?’ or ‘!’. True in 90% of the cases.

2, Sentences may be split up by other punctuation marks (e.g., : ; --).

3, Sentences may be broken up, as in: “You should be here,” she said, “before I know it!”

4, Quote marks may be at the very end of the sentence.

5, Identifying sentence boundaries can involve heuristic methods that are hand-coded. Some effort to automate the sentence-boundary process has also been tried.