Introduction to Advanced
Natural Language Processing (NLP)

L645
Advanced NLP
Fall 2009

Adapted from Manning and Schütze (1999) and Abney (1996)
Definition of CL 1

Computational linguistics is the scientific study of language from a computational perspective.

Computational linguists are interested in providing computational models of various kinds of linguistic phenomena. These models may be "knowledge-based" ("hand-crafted") or "data-driven" ("statistical" or "empirical").

Work in computational linguistics is in some cases motivated from a scientific perspective in that one is trying to provide a computational explanation for a particular linguistic or psycholinguistic phenomenon; and in other cases the motivation may be more purely technological in that one wants to provide a working component of a speech or natural language system.

http://www.aclweb.org/archive/what.html
Definition of CL 2

Computational linguistics is the application of linguistic theories and computational techniques to problems of natural language processing.

http://www.ba.umist.ac.uk/public/departments/registrars/academicoffice/uga/lang.htm
Definition of CL 3

Computational linguistics is the science of language with particular attention given to the processing complexity constraints dictated by the human cognitive architecture. Like most sciences, computational linguistics also has engineering applications.

http://www.cs.tcd.ie/courses/csll/CSLLcourse.html
Definition of CL 4

Computational linguistics is the study of computer systems for understanding and generating natural language.

Short History

- 1950s: Machine Translation
- 1960s: Chomsky
- 1964: ALPAC Report
- 1966: ELIZA
- 1973: SHRDLU
- 1980s: knowledge-based CL
- 1990s: statistical / machine learning approaches in CL
Early 1950s

- Machine Translation (MT): one of the earliest applications of computers
  - Major players: US and USSR
  - Russian to English and reverse

- Georgetown University, Washington system:
  - translated sample texts in 1954
  - euphoria - a lot of funding, many groups in US, USSR
  - BUT: the system could not scale up
• assessed research results of groups working on MTs

• conclusions:
  – MT not possible in near future
  – funding should cease for MT!
  – basic research should be supported

• word by word translation does not work, linguistic knowledge is needed
ELIZA

- the first chatterbot – a computer program that mimics human conversation
  - author: Joseph Weizenbaum – Massachusetts Institute of Technology
- simulation of a (Rogerian) therapist
  - user types in some statement or set of statements in natural language
  - ELIZA then analyzes the user’s statement and generates some response
- basic technology: pattern matching

USER:   You don’t argue with me.
ELIZA:  WHY DO YOU THINK I DON’T
         ARGUE WITH YOU
SHRDLU

• interaction with a robot in a block world.
  – author: Terry Winograd – MIT

• the user can:
  – ask the robot to manipulate the blocks
  – ask it about the blocks configurations
  – ask it about its reasoning
  – update facts

• "understands" language in a limited domain by using syntactic parsing and semantic reasoning
  – large scale grammar of English + parser
  – procedural semantics for words and phrases
Knowledge-Based CL

- proof of concept & manually-written rules
  - linguistic/logic paradigm extensively pursued
  - later: development of linguistic formalisms (Lexical Functional Grammar, Head-Driven Phrase Structure Grammar, Tree Adjoining Grammar etc.)

- Limitations:
  - not robust enough
  - few applications
  - not scalable

Addressing this limitations led to the more recent statistical approaches
Outline

- How to approach NLP
- Motivating Probability
- The Ambiguity of Language
- Data-driven NLP
- Corpora: using corpora for simple analysis
Approaching NLP

There are two main approaches to doing work in NLP:

- **Theory-driven (≈rationalist):** working from a theoretical framework, come up with a scheme for an NLP task
  - e.g., parse a sentence using a handwritten HPSG grammar

- **Data-driven (≈empiricist):** working from some data (and some framework), derive a scheme for an NLP task
  - e.g., parse a sentence using a grammar derived from a corpus

The difference is often really a matter of degree; this course is more data-driven
Competence and Performance

- **Competence** = knowledge of language found in the mind
- **Performance** = the actual use of language in the world

Data-driven NLP focuses more on performance because we deal with naturally-occurring data

- That is not to say that competence is not an issue: linguistic competence issues can inform processing of performance data

- Generally, we will not deal with the contexts surrounding performance, e.g., sociolinguistic & pragmatic factors
Data-driven NLP

Data-driven NLP often learns patterns from naturally-occurring data

- Instead of writing rules, have computer learn rules / regularities
- This type of learning is generally probabilistic

Notice that linguistics is generally not a probabilistic field

- Aside from engineering considerations, are we justified in treating language probabilistically?
- Abney (1996) gives several arguments for this . . .
Motivating probability
Is language categorical?

Are the language phenomena we observe categorical, e.g., are sentences completely grammatical or completely ungrammatical?

• Are the following sentences equally (un)grammatical?

(1) a. John I believe Sally said Bill believed Sue saw.
    b. What did Sally whisper that she had secretly read?
    c. Who did Jo think said John saw him?
    d. The boys read Mary’s stories about each other.
    e. I considered John (as) a good candidate.

⇒ Probabilistic modeling could give a degree of grammaticality
Motivating probability
Is language non-categorical?

Consider the following clear part-of-speech distinction:

(2) We will review that decision in the *near* future. (adjective)
(3) He lives *near* the station. (preposition)

But what about the following:

(4) We live *nearer* the water than you thought.

- A comparative (-*er*) form, like an adjective
- Takes an object noun phrase, like a preposition

⇒ Probabilistic modeling could assign likelihood of category assignment
Motivating probability
Language acquisition, change, and variation

• Language acquisition: child uses grammatical constructions with varying frequency
  ⇒ trying out rule possibilities with different probabilities

• Language change: gradual changes
  ⇒ a certain proportion of the population is using new constructions

• Language variation: Dialect continua and typological generalizations
  – e.g., “postpositions in verb-initial languages are more common than prepositions in verb-final languages”
  ⇒ Ripe for probabilistic modeling
What does it mean to be tall?

- For people, it is typically around six feet, although judgments vary.
- But a tall glass of milk is much less than six feet.
  - The meaning shifts depending on the context.
  - But even then, it is not clear that it has a fixed meaning.

⇒ Meanings can be associated with certain probabilities
Motivating probability
Rarity of usage

Consider the following (Abney 1996):

(5) The a are of I.
(6) John saw Mary.

• *The a are of I* is an acceptable noun phrase (NP): *a* and *I* are labels on a map, and *are* is measure of area

• *John saw Mary* is ambiguous between a sentence (S) and an NP: a type of saw (*a John saw*) which picks out the *Mary* we are talking about (cf. *Typhoid Mary*)

We don’t get these readings right away because they’re rare usages of these words

\[ \Rightarrow \] Rarity needs to be defined probabilistically
Motivating probability

Wide-coverage of rules

Grammar rules have a funny way of working sometimes and not others

• Typically, if a noun is premodified by both an adjective and another noun, the adjective must precede the modifying noun

  (7) tall shoe rack
  (8) *shoe tall rack

• But not always:

  (9) a Kleene-star (N) transitive (A) closure
  (10) highland (N) igneous (A) formations

If language is categorical and you have a rule which allows N A N, then you have to do something to prevent shoe tall rack.
The Ambiguity of Language

Language is ambiguous in a variety of ways:

- Word senses: e.g., bank
- Word categories: e.g., can
- Semantic scope: e.g., All cats hate a dog.
- Syntactic structure: e.g., I shot the elephants in my pajamas.

and your grammar could be ambiguous in ways you may not even realize

- Often, however, of all the ambiguous choices, one is the best
You may not even recognize every sentence as being ambiguous

(11) Our company is training workers
Less intuitive analyses

Our company is training workers.

Our company is training workers.
How statistics helps

To summarize so far, probabilities can help with the following:

• Disambiguation: weight syntactic rules to come up with the most likely parse
  – Can also come up with structural parsing preferences, e.g., longest match for constituents

• Degrees of grammaticality: some sentences are better than others, without a hard-and-fast grammatical judgment

• Naturalness: verbs have certain selectional preferences; collocational knowledge is a part of language (English *thick accent* vs. German *starker Akzent*); etc.

• Learning on the fly: people gradually adapt to new structures and meanings
Data-driven NLP

To pick the best parse, we need probabilistic information, but where do we get such information?

- We can **induce** this information from language data, potentially data annotated with linguistic information.

  i.e., linguistic performance, not competence

**Exercise:**

- Determine the different possible analyses of this naturally-occurring phrase: 
  *some white trash version of Shania karaoke*

- Skim the internet and find an ambiguous phrase with only one sensible interpretation.
Corpora

We will becoming familiar with processing large texts, i.e., corpora

• There are several different ways of looking at a corpus: how balanced it is, how large it is, if it’s freely available

• Corpora are often annotated with linguistic mark-up, such as part-of-speech labels or syntactic annotation

These corpora will serve as our data from which to learn probabilities

• Corpora are not the only lexical resources out there; dictionaries (e.g., WordNet) are also important, but these are often derived from corpora
Using corpora for simple analysis

Word counts

We can use corpora to give us some basic information about word occurrences.

- Count **word types** = number of distinct words there are in the corpus
- Count **word tokens** = number of actual word occurrences in the corpus; multiple occurrences of the same word type are counted each time

If we compare word types and tokens, we see that there are:

- a few word types which occur a large number of times (often function words)
- a large number of word types which occur only a few times or only once
Zipf’s Law

This idea is formulated in **Zipf’s Law** = the frequency \((f)\) of a word is inversely proportional to its rank \((r)\)

\[(12) \quad \begin{align*}
\text{a.} & \quad fr = k, \text{ where } k \text{ is some constant, or } f = \frac{k}{r} \quad \text{(Zipf)} \\
\text{b.} & \quad f = P(r + \rho)^{-B}, \text{ where } P, \rho, \text{ and } B \text{ are parameters which measure a text’s richness (Mandelbrot)}
\end{align*}\]

Mandelbrot adjusted Zipf’s Law to better handle high and low ranking words; with \(B = 1\) and \(\rho = 0\), it is identical to Zipf’s Law (where \(P = k\)).

- Important insight: most words are rare!
Using corpora for simple analysis

Collocations

Collocations are phrases (often two-word units) which are more than the sum of their parts → often should be their own dictionary entry

- e.g., disk drive or make up
- Firth: “You shall know a word by the company it keeps”

These are important for machine translation (MT) and information retrieval (IR), among other applications

- They can be obtained automatically for corpora by seeing which two-word sequences (bigrams) occur most often, after normalizing for overall word frequency
Using corpora for simple analysis

Keyword in Context (KWIC)

We can also use corpora to build a concordance, which shows how a key word is being used in a context.

The librarian showed off -
gentleman teachers showed off with
lip and showed the vacancy

By analyzing the key words in context, we can categorize the contexts and find, e.g., the syntactic frames for a word

- \( \text{NP}_{agent} \) showed off (\( \text{PP}[with/in]_{manner} \))
- \( \text{NP}_{agent} \) showed (\( \text{NP}_{recipient} \)) (\( \text{NP}_{content} \))
Planning a KWIC program

Exercise: Without using a particular programming language, how would you write a KWIC program?

- Assume: you’re given a large corpus in text format
- What are the steps involved to show the words in context?
References
