University of Malta
BSc(IT) HonsYear IV

CSA4050: Advanced Topics in NLP

Statistics I

The Empirical Approach

- Historical Background
- Low level issues
- Tokenisation

Dept Computer Science and AI 2003/04
Lecturer: Michael Rosner

Sources and Resources

Much of the material for this lecture comes from

- "Foundations of Statistical Language Processing", Manning and Schütze, MIT 1999.
- Resources for statistical/empirical NLP.

  `www-nlp.stanford.edu/links/statnlp.html`
- See also McEnery & Wilson's notes on Corpus Linguistics

  `www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm`

Historical Perspective

- Pre-Chomsky linguistics (e.g. Boas 1940) was **largely empirical**.
- During 1970s: **Rationalist approach** to AI Systems for language understanding in restricted domains. (e.g. Winograd (1972), Woods (1977), Waltz (1978)).
- During 1980s continued progress on developing NLP systems using **hand-coded grammars and knowledge bases** (Allen 1987)
- All such systems required a great deal of domain-specific knowledge engineering and as a result were quite brittle and difficult to scale up.
- During second half of 1980s focus shifted from *rationalist* methods to *empirical* or *corpus-based* methods in which development is largely **data driven**.

1980-1990s Trends

- Linguistics Research: Automatic induction of lexical and syntactic information from corpora (Brown Corpus, Lancaster-Oslo-Bergen Corpus).
- Speech recognition (IBM Yorktown Heights Lab) resulted in statistical methods based on **Hidden Markov Models** (HMMs) that outperformed previous knowledge-based approaches.
- These methods use a probabilistic finite state machine to model the pronunciation of words and make use of a hill-climbing training algorithm to fit the model parameters to the actual speech data. Most existing commercial speech recognition systems are based on HMMs.

Probabilistic Finite State Pronunciation Model

Areas of Application

Starting in the late 1980s the success of statistical methods in speech spread to other areas of language processing, notably POS tagging, spelling correction, and parsing.

- **POS Tagging**: assigning an appropriate syntactic class to each of the words in a sentence. Close to human levels of performance can be achieved.
- **Machine Translation.** Statistical approaches to machine translation trained on bilingual proceedings of the Canadian government.
- **Parsing methods** based on *tree banks* (large databases of sentences annotated with syntactic parse trees), such as PCFGS (probabilistic CFGs)
- **Word-sense disambiguation**, PP attachment, anaphora, discourse segmentation.
- **Content-based document processing**
  - **Information Extraction:** text $\rightarrow$ filled template
  - **Information Retrieval:** query $\rightarrow$ set of relevant documents

Empirical Approach - Some Issues

- Potential for Solutions to Old Problems
  - Knowledge Acquisition
  - Coverage
  - Robustness
  - Domain Independence
- Feasibility Depends on
  - Data Resources.
  - Computing Resources.
- Advantages/Disadvantages
  - Emphasis on applications and evaluation.
  - Desire to move beyond toy domains
  - BUT - results always corpus-dependent

Start with a Corpus

- A corpus is an organised body of materials from language that is used as the basis for empirical studies. The following characteristics are amongst those that are important for NLP:
  - Representativeness
  - **Medium**: printed, electronic text, digitized speech, video.
  - **Language**: monolingual/multilingual
  - **Information Content** plain versus tagged. Plain corpus just contains words, whilst a tagged corpus includes other information, e.g. part of speech tags.
  - **Degree of structuring** e.g. trees versus sentences.
  - Size
  - Standards
  - Quality

Examples of Corpora

- **Project Gutenberg:** public domain text resources at http://www.promo.net/pg/
- **Brown Corpus:** A tagged corpus of about a million words put together at Brown University during the 1960s and 1970s.

  Brown corpus is balanced. It was intended as a representative example of American English. It contains texts of different genres including newspapers, fiction, scientific text, legal text, and others.
- **Penn Treebank:** A corpus of parsed sentences based on text from the Wall Street Journal.
- **Canadian Hansards:** bilingual corpus of the proceedings of the Canadian parliament Contains parallel texts in English and French which have been used to investigate statistically based machine translation.
- **WordNet:** An electronic dictionary of English. Words are organised into a hierarchy, rather like a thesaurus. Each node consists of a synset of words with identical or near identical meanings.

Part of Speech (POS) Tags

Here are 24/36 of the POS tags used by the Penn TreeBank project, a corpus of parsed sentences based on text from the Wall Street Journal.

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol

Tagging Issues

- Tagset
  - Language independence: NSING; NPLUR; NDUAL
  - Resolution: ambiguity versus Precision: "fish"
- Tagging
  - By hand
  - Tagging algorithms
    * Stochastic: most probable sequence of categories.
    * Rule Based: If preceding word is /DET, tag = NN
    * Transformation Based

Low Level Processing

- Preprocessing
  - Filtering junk (headers, whitespace, diagrams ...).
- Tokenisation
- Normalisation
- Types versus tokens

The preprocessing phase is followed by data gathering.

Tokenisation

- Tokenisation is a process which divides input text into units called tokens.
- Tokens are typically of different types, e.g. at least the following: words, numbers and punctuation.
- What counts as a word?

  "a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes but no other punctuation marks". (Kucera and Francis 1967)

Tokenisation

VfB Stuttgart scored twice in quick succession early in the second half on their way to a deserved 2-1 victory over Manchester United in the Champions League on Wednesday.
(example from Mary Dalrymple, University of London)

- VfB Stuttgart, Manchester United
- succession
- 2-1
- Wednesday.

Special characters cause problems!

Problems Identifying Words 1

- Words often include non-alphanumeric characters

  $22.50; Micro$oft; www.di-ve.com.mt; BSc. IT :-) !

- There may be no spaces between words at all. For example, in German we have

  Lebensversicherungsgesellschaftsanngesteller

  (life insurance company employee).

- In languages such as this, word segmentation takes on a complexity which approaches that for sentences.

- It is debatable whether this is a job for the tokeniser or whether it properly belongs to some later stage of processing

Problems Identifying Words 2

- The presence of spaces around a word does not necessarily indicate a word break, e.g. Coca Cola.
- A different version of this problem crops up with items of particular semantic types which are often written with spaces. Examples include phone-numbers e.g.

  +356 456 452
- Dealing with specialised formats like this takes us from tokenisation towards information extraction.
- Typically, the problem is dealt with by hand crafting regular expressions that match, but given the brittleness of the approach, there is considerable interest in automatic means of learning the formatting of semantic types.
- Words are often delimited by punctuation.

Punctuation

In general, punctuation marks attach to words. Detaching commas, semicolons and colons from words is quite easy.

Periods present special problems

Most periods mark end of sentence. Others mark abbreviations, e.g. etc. Such abbreviations should remain part of the word, since in some cases we need to distinguish between an ordinary word and a legitimate abbreviation

Wash. versus wash

Note also that when an abbreviation occurs at the end of a sentence there is only one period.

Apostrophe

English contractions such as won't or I'll count as a graphic word according to the above definition, but many think that the tokeniser ought to return two separate tokens. Some processors (and some corpora, such as the Penn Treebank) split such contractions into two words.

Note that the decision not to split plays havoc with traditional grammar rules like

S $\rightarrow$ NP VP

(which fails to generate sentences like "I'm feeling fine").

On the other hand, if we do split, we are left with "funny" words like "n't", since

isn't = is + n't

's

Note that "'s" has two main uses:

Abbreviation for "is". He's a real friend. Indication of possession: Dog is man's best friend

Word final single apostrophes are also problematic since they are ambiguous between (a) the end of a quotation and (b) possessive on the end of a word ending in "s". In the first case the quotation should be removed, but in the latter not.

In general, there is no easy way for a tokeniser to decide what to do.

Exercise:

How is the apostrophe used in Maltese? How should a Maltese tokeniser treat the apostrophe?

Hyphen

Do sequences of letters with a hyphen in between count as one word or two?
It depends ...

- Typesetting Hyphen. Generally, these are removed. However, if a word genuinely contains a hyphen, and the word happens to come at the end of a line, then the line-break will come at the hyphen. In this case, the hyphen at the end of the line should not be deleted.
- Lexical Hyphens: some items with a hyphen are best treated as a single word co-operative
- Quotation-indicating hyphens: keep

  child-as-required-yuppie-possession
- Measure-phrase hyphens

  My 35-year-old grandmother: discard

Hyphen

- In these last three cases, we would probably want to treat the things joined by the hyphens as separate words. If we did not, the word-lexicon would grow unacceptably large.
- Texts are often not consistent in their use of hyphens, nor even in the way they are written (cf. English - versus–American practice)

Uppercase/Lowercase

Two tokens containing the same characters are often instances of the same type, e.g. the The THE

When they can be treated as the same type, store the type in (say) lowercase and always map the token to lowercase.

Heuristics

- Map first character of a sentence to lowercase
- Map all words in titles to lowercase

Problems

- Identification of sentence boundaries
- Identification of proper names

Types versus Tokens

How many words are there in this sentence? The cat chased the rat.

- Five tokens: the, cat, chased, the, rat.
- Four types: cat, chased, rat, the.

How many words in English?

- Switchboard corpus of spoken English: 2.4 million tokens, 20,000 wordform types
- Shakespeare: 884,647 tokens, 29,066 wordform types
- American Heritage Dictionary: 200,000 entries (can be more than one entry per stem, multiword entries)
- Oxford English Dictionary: 500,000 entries

Normalisation

Depending on application:

VfB Stuttgart scored twice in quick succession early in the second half on their way to a deserved 2-1 victory over Manchester United in the Champions League on Wednesday.

normalize to

VfB_Stuttgart scored twice in quick succession early in the second half on their way to a deserved 2 - 1 victory over Manchester United in the Champions League on Wednesday .

Further Normalisation

Are eats and eating different words?

They are two different wordforms that have the same stem, eat but different suffixes, -s and -ing

Stemming versus morphological analysis depends on application.

Summary

- The empirical approach spread from speech processing to other areas of nlp.
- Empirical approach advantages and disadvantages
- Compilation of corpora is a long and laborious process.
- Representative corpus is hard to define.
- Before statistical analysis can take place, corpus needs to be preprocessed and tokenised.
- Design of preprocessor and tokeniser can radically affect the data upon which subsequent conclusions are drawn.