CS565: Intelligent Systems and Interfaces

Getting Started with NLP

11th Jan, 2017

Semester: Jan – May 2017

Ashish Anand

IIT Guwahati

Announcements

- Rescheduling Thursday [3-4PM] Lectures to Friday [2-3PM]

Objective of the lecture

- To understand why NLP is hard
  - Ambiguity at multiple levels
  - Different levels of NLP
- Get started dealing with natural language
  - Basic Pre-processing: Word and Sentence Segmentation

Why NLP is Hard?

"What is your little brother crying about?"
"Oh, 'im—'e's a reg'lar comp'tational linguist, 'e is."

http://specgram.com/CLIII.4/08.phlogiston.cartoon.zhe.html

Ambiguity

Example:

I made her duck

Time flies like an arrow.

- What is your inference of the two sentences?
- Whether all of them are meaningful/grammatically correct ?

Ambiguity

Examples: *I made her duck*

- Interpretations :
  - *I cooked duck for her*
  - *I cooked duck belonging to her*
  - *I caused her to quickly lower her body*

More Examples of Ambiguity

- Anne Hathaway vs. Warren Buffett’s Berkshire Hathaway stock
  - When *Bride Wars* opened the stock rose 2.61%.

  [source: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1162/handouts/cs224n-lecture1.pdf]
- *Every Indian has a mother* vs. *Every Indian has a prime minister*
- *We gave the monkeys the bananas because they were hungry* vs. *We gave the monkeys the bananas because they were over-ripe*

Ambiguous Words

- address, number
    - Pronunciation
- Fly, rent, tape
    - Part of speech
- ball, board, plant
    - Meaning

Types of Ambiguity

- Phonetic
  - My finger got number
- Morphological
  - Impossible vs important
  - Ram is quite impossible/ Ram is quite important
- Part of speech
  - Geeta won the first round
- Syntactic
  - Call Ram a taxi

Types of Ambiguity

- Pp attachment
  - The children ate the cake with a spoon.
- Cc attachment
  - Ram likes ripe apples and pears
- Sense
  - Ram took the bar exam
- Referential
  - Ram yelled at Shyam. He was angry at him
- Metonymy
  - Sydney called and left a message for Ram

Some other sources of difficulties

- Non-standard, slang, novel and short words
  - A360, +1-646-555-2223
  - Selfie, chillax
- Inconsistencies
  - junior college, college junior
- Parsing problems
  - Cup holder
- Metaphors, Humors, Sarcasm

Summary: why NLP is hard?

- Highly ambiguous at all levels
- Context is important to convey meaning
- Involves reasoning about the world

Different Levels of NLP

- Word
  - Phonetics and Phonology: study of linguistic sounds
  - Morphology: study of meaningful components of words [example]
- Syntax: structural relationship between words [study of sentence and phrase structure]
- Semantic: study of meaning
  - Lexical semantics: study of meanings of words
  - Compositional semantics: How to combine words
- Discourse: dealing with more than a sentence: paragraph, documents

Lets begin: what it takes to make an NLP system

Source

- Corpora (plural for *corpus*: large, (un)structured set of texts)
  - Brown corpus: 500 samples of English texts published in the US in 1961, approx. 1 million words
  - Access to multiple corpus from tools like *NLTK*
  - BYU corpora at corpus.byu.edu
  - Linguistic data consortium (LDC)
  - Building from databases such as PubMed.

Source

- Caution: One shoe does not fit all.

Looking at Text: Basic pre-processing

Text Preprocessing

- Removing non-text (e.g. tags, ads)
- Segmentation
  - Sentence and word
- Normalization
  - Labeled/labelled,
- Stemming
  - Computer/computation
- Morphological analysis
  - Car/cars
- Capitalization
  - Led/LED,

Tokenization: word segmentation

- Definition: Process to divide the input text into units, also called, *tokens*, where each is either a *word* or a *number* or a *punctuation mark.*
- Should we remove all punctuation marks ?

What counts as a word?

- Kucera and Francis (1967) defined “*graphic word*” as follows :
    - “ a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes, but no other punctuation marks”

Problem with graphic word definition

- Should we consider “$12.20” or “Micro$oft” or “:)” as a word?
- We can expect several variants especially in forums like Twitter etc which may not obey exact definition but should be considered as a word.
- Simple Heuristic: *Whitespace*
  - *“*a *space* or *tab* or the *new line*” between words.
  - Still to deal with several issues.

Defining words: Problems

- Periods
  - Abbreviations at the end vs. in the middle
  - etc., Wash. Vs wash
- Single apostrophes
  - Contractions such as I’ll, I’m etc.: should be taken as two words or one word?
  - *Penn Treebank* split such contractions.
  - Phrases such as *dog’s vs. yesterday’s* in “The house I rented yesterday’s garden is really big”.
  - Orthographic-word-final single quotation such as “boys’ toys”.

Defining words: Problems

- Hyphenation
  - Again the same question – “do sequences of letters with a hyphen in between count as one word or two?
  - Occurrences like *e-mail*, *co-operate* vs. *non-lawyer, so-called, text-based*
  - Inconsistency in using words like “cooperate” as well as “co-operate”
  - Line-breaking hyphen vs. actual hyphen happens at the end of line [*haplology*]
- Word with a whitespace between its parts
  - New Delhi, San Francisco
  - … the New Delhi-New Jalpaiguri special train …

Word segmentation in other language

- 请将这句话翻译成中文 [Please translate this sentence into Chinese]
- Compound nouns written as a single word
    - Lebensversicherungsgesellschaftsangestellter [Life insurance company employee]

Defining words: other issues

- Morphology
  - Different forms of words
    - Go, went, gone
    - Fox, foxes
  - Stemming and Lemmatization

Dealing with cases: Main issue

- Can we make all letters in same case
  - Should we treat “*the*”, “*The”, and “THE”* differently vs. “*Mr. Brown*” and “*brown paints*”

Dealing with cases: A Heuristic

- Convert all capital letters to lowercase
  - At the beginning of a sentence, and
  - In headings, titles etc.

- Do we see any problem in this heuristic ?

Problems with the heuristic

- Dependency on correct detection of sentence boundary
- All names appearing in the beginning of the sentence or in places like titles, gets converted
- More importantly, loss of information
  - Example: words in the middle of a sentence but started with capital letter for emphasizing an important point.

- Objective of the study should determine our decision.

Defining Sentence Boundary

- Something ending with a ‘.’, ‘?’, or ‘!’
  - Language specific
- Problem with ‘.’
  - Still 90% of periods are sentence boundary indicators [Riley 1989].
- Sub-sentence structure with the use of other punctuation
  - “The scene is written with a combination of unbridled passion and sure-handed control: In the exchanges …….. inexorability of separation”
- Other issues
  - “You remind me,” she remarked, “of your mother.”

Defining Sentence Boundary: A heuristic

- Put putative sentence boundaries after occurrences of ., ?, ! (and may be ;, :, -)
- Check presence of following quotation marks, if any move the boundary.
  - “You remind me,” she remarked, “of your mother.”
- Disqualify a period boundary if –
  - It is preceded by a known abbreviation that does not generally occur at the end of sentence such as Dr., Mr. or vs.
  - It is preceded by a know abbrev. that is generally not followed by an uppercase word such as etc. or Jr.
- Disqualify a boundary with a ? or ! If
  - It is followed by a lowercase letter (or name)

Issues with Heuristic or set of pre-defined rules

- Is it possible to define such rules without the help of experts?
- Will it work for all languages?

Machine Learning Methods: Sentence boundary as classification problem

- Riley (1989) used classification trees
  - Features: case & length of the words preceding and following a period; prior prob of words occurring before and after a sentence boundary etc.
- Palmer and Hearst (1997) used neural network model
  - Instead of prior probability, PoS distribution of the preceding and following words.
  - Language-independent model with accuracy of 98-99%
- Reynar and Ratnaparkhi (1997) and Mikheev (1998) used Max. Ent approach
  - Language independent model with accuracy of 99.25%