CS565: INTELLIGENT SYSTEMS AND INTERFACES

Getting Started with NLP

Semester: July – November 2020

Ashish Anand

Associate Professor, Dept of CSE

IIT Guwahati

Recap

- Defined NLP
- Discussed two broad school of thoughts
- Discussed existence of ambiguity of natural languages
- Discussed different levels of NLP

Objective

- Getting started with NLP
  - Corpora
  - Text Pre-processing and Normalization

GETTING STARTED WITH NLP

Essential resources and basic pre-processing

Source: Corpora

- Corpora (plural for *corpus*: large, (un)structured set of texts)

- Different types of corpora
  - Monolingual
  - Parallel – Multilingual/Comparable/Aligned
  - Learner Corpus
  - Diachronic Corpus

Building Corpora

- Organizational / Consortium effort
  - Linguistic Data Consortium (LDC) [www.ldc.upenn.edu]
  - European Language Resources Association (ELRA) [www.elra.info/en]
  - Indian Language Technology Proliferation and Deployment Centre [http://tdil-dc.in/index.php?lang=en]
- Individual effort

Examples of Corpora

Corpus	Tokens	Types
Switchboard phone conversations	2.4 million	20000
Shakespeare	884,000	31000
Brown	1 million	38000
Google N-grams	1 trillion	13 million

Two ways to talk about words:

1. Tokens: number of running words
2. Types: number of distinct words

More Examples of Corpora

- Access to multiple corpus from tools like *NLTK*
- Building from databases such as PubMed, free text from web, Wikipedia, Social media platforms etc.
- Shared task challenges: ACE, CoNLL, SemEval, BioAsq, SQuAD, CORD-19
- Caution: One shoe does not fit all.

Text Preprocessing

- Removing non-text (e.g. tags, ads)
- Text Normalization
    - Segmentation: Word and Sentence Segmentation
    - Normalizing Word Formats
        - Spelling Variations: Labeled/labelled
        - Capitalization: Led/LED
        - Lemmatization
        - Stemming
        - Morphological analysis: dealing with smallest meaning-bearing units

TEXT NORMALIZATION

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*.

What counts as a word?

I am interested in Natural Language Processing, but I’m not sure of the required prerequisites.

What counts as a word?

- Should I count punctuation as a word?
- Should I treat I’m as one word or break them into three words: I, ’, m? [**Clitic**]
- Should I consider “Natural Language Processing” as one word or 3 words?

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

- Too restrictive
  - 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
  - Wash. vs wash
  - Abbreviations at the end vs. in the middle – e.g. etc.
  - More on this while discussing sentence segmentation

- 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***]
  - Hyphens to indicate correct grouping of words: take-it-or-leave it in "`a final take-it-or-leave it offer`"

- Word with a whitespace between its parts
  - New Delhi, San Francisco
  - … the New Delhi-New Jalpaiguri special train …

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 words: Problems: Spoken Corpora

- This lecture umm is main- mainly divided into two components

- Two types of **disfluencies**
  - **Fragments: main-**
  - **Fillers/Filled pauses: uh.. Umm..**

Tokenization in Practice

- Deterministic algorithms based on regular expressions
- Compiled into efficient finite state automata

Word segmentation in other languages

- 请将这句话翻译成中文 [Please translate this sentence into Chinese]
  - Languages like Chinese, Japanese have no spaces between words
  - Japanese is further complicated with multiple alphabets intermingled

- Compound nouns written as a single word
  - Lebensversicherungsgesellschaftsangestellter [Life insurance company employee]

TEXT NORMALIZATION

Word Normalization

Definition

- Converting the words/tokens in a standard format, i.e. chossing a single canonical form for words which can appear in multiple forms. Example: Ph.D., PhD., PhD

Some Examples of Normalization

- Case Folding
  - Conversion into lowercase
  - May be good idea for Information Retrieval (search) purpose
  - May not be good for POS tagging or NER (US: the country vs us: pronoun)

Examples of Word Normalization

- Lemmatization
  - Task of determining two words have the same root, same POS, same sense but may have different word forms.
  - Mostly relevant for IR purpose
  - Requires **Morphological Parsing** of words

- Stemming
  - Crude form of lemmatization
  - Consists of chopping off word-final affixes

Word Normalization: Lemmatization: Morphology

- Morphology is study of the way words are built up from smaller meaning-bearing units (Morphemes)

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%

References

- Chapter 4 [FSNLP]
- Chapter 2 [Jurafsky and Martin 3rd Ed.]