Text Processing

Lecture-3

Outline

Three basic tasks

- Word Tokenization
- Sentence segmentation
- Normalization

Also discuss

- Regular Expressions
- Challenges in each task

Text Processing

- To (pre)process your text simply means to bring your text into a form that is predictable and analyzable for your task.
- Task = approach + domain
- So, same approach may not be suitable for different task

Corpora

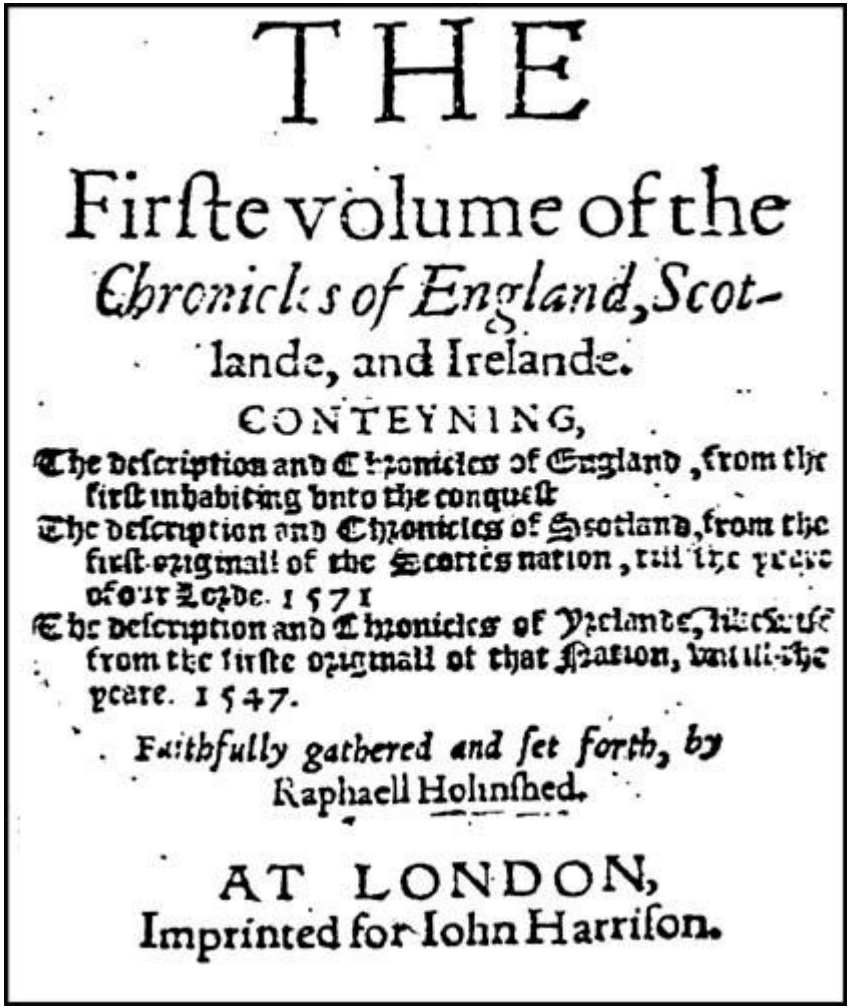
- A corpus (plural corpora) or text corpus is a language resource consisting of a large and structured set of texts (nowadays usually electronically stored and processed).
- A corpus is a special collection of textual material collected according to a certain set of criteria.
- Eg: Brown corpus: **representative sample** of written American English used in 1962
- The general issue is whether the corpus is a representative sample of the population of interest.
- A sample is representative if what we find for the sample also holds for the general population
- **Balanced Corpus**

Looking at Text

- Text will usually come in either a raw format, or marked up in some way.
- Markup: HTML, XML, etc
- First step is remove markups and take the raw text
- Example:
  - <h1>Here is Example</> → Here is Example

Low level issues

- Junk content
  - Tables, headers/footer
  - OCR text → unrecognized characters
- Uppercase/Lower case
  - How to treat capitalization
  - The, THE and the → are they same?
  - Proper nouns→ Bell, bell (whether proper noun or start of the sentence?)



THE

Firſte volume of the

Chronicles of England, Scot-

lande, and Irelande.

CONTEYNING,

The deſcription and Chronicles of England, from the firſt inhabiting vnto the conqueſt

The deſcription and Chronicles of Scotland, from the firſt originall of the Scottes nation, till the yeare of our Lorde. 1571

The deſcription and Chronicles of Yrelande, likewiſe from the firſte originall of that nation, vntill the yeare. 1547.

Faithfully gathered and ſet forth, by

Raphaell Holinſhed.

AT LONDON,

Imprinted for Iohn Harriſon.

Tokenization: What is word?

- The step of processing is to divide the input text into units called tokens where each is either a word or something else like a number or a punctuation mark
- Normally, we use **whitespace** as delimiter
- **Word** is a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes, but no other punctuation marks
- **Sam’s cat in the hat is different from other cats!**
  - How many words?
  - How to treat punctuations?
- Can we remove all punctuations? → No, some clues contained in punctuations
- What about *$490, C#, Micro$oft?*

Periods?

- While most periods are end of sentence punctuation marks, others mark an abbreviation such as in etc. or Calif.
- *Wash.* and *wash.* are different
- Word with periods at the end of sentence
  - The period serves both word part and sentence boundary
  - *The man went to Calif.*
  - This is called Hapology

Single Apostrophes

- Are they one word or two
  - I'll → I will
  - Isn't → Is not
  - Can't → Can not,
  - I'm → I am
- Dog's, cat's → Possessive case

Hyphenation

- Do sequences of letters with a hyphen in between count as one word or two
  - State-of-the-art, learning-based, line-breaking
- Whether the hyphens treated as line breaking hyphens or single word?
  - A-1-plus, co-operate
- Issue: inconsistency in the use of hyphens

Different forms of same word

- Eg: Saw as Noun and Saw as verb
- Whether we treat these occurrences distinct lexemes?
- These are called **Homographs**

Word segmentation in other languages

धर्म्माधर्मं → धर्म + अधर्म (Sandhi Rules)

അവനവനാത്മസുഖത്തിനായ് --> അവൻ അവൻ ആത്മ സുഖത്തിനായ്

This become complex in Indian languages

Variant Coding of Information

Trying to deal with myriad formats like this is a standard problem in **information extraction.**

Phone number	Country	Phone number	Country
0171 378 0647	UK	+45 43 48 60 60	Denmark
(44.171) 830 1007	UK	95-51-279648	Pakistan
+44 (0) 1225 753678	UK	+411/284 3797	Switzerland
01256 468551	UK	(94-1) 866854	Sri Lanka
(202) 522-2230	USA	+49 69 136-2 98 05	Germany
1-925-225-3000	USA	33 1 34 43 32 26	France
212. 995.5402	USA	++31-20-5200161	The Netherlands

Table 4.2 Different formats for telephone numbers appearing in an issue of *The Economist.*

Morphological Forms

- What about different morphological forms of word
  - play, playing, plays, played
- Normally we apply two normalization techniques
  - **Stemming** usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.
  - **Lemmatization** where one is attempting to find the lemma or lexeme of which one is looking at an inflected form.
- Ex: “operate operating operates operation operative operatives operational”
- Stemming → oper
- Lemmatization → operate

Tokenization: Summary

- What is word?
- Issues:
  - Periods
  - Hyphen
  - Aporstephy
  - Homographs
  - Word segmentation (in other languages)
  - Variant coding
  - Morphological forms → stemming and lemmatization

Sentence Segmentation

- “something ending with a ‘.’, ‘?’ or ‘!’ .”
- Is there any issue with this simple definition?

Sentence Segmentation

- “something ending with a ‘.’, ‘?’ or ‘!’ .”
- Is there any issue with this simple definition?
- The scene is written with a combination of unbridled passion and sure-handed control: In the exchanges of the three characters and the rise and fall of emotions, Mr. Weller has captured the heartbreaking inexorability of separation.
- How many sentences?
- What about colon (:) , semicolon (;) and hyphens (-)

Indirect speech

You remind me," she remarked, "of your mother."

quotation marks after sentence final punctuation. the end of the sentence is not after the period in the example above, but after the close quotation mark that follows the period.

Sentence Boundary Disambiguation Problem

A heuristic approach

These rules can be represented by regular expressions

Statistical, Machine Learning approaches are available

- Place putative sentence boundaries after all occurrences of . ? ! (and maybe ; : —)
- Move the boundary after following quotation marks, if any.
- Disqualify a period boundary in the following circumstances:
  - If it is preceded by a known abbreviation of a sort that does not normally occur word finally, but is commonly followed by a capitalized proper name, such as *Prof.* or *vs.*
  - If it is preceded by a known abbreviation and not followed by an uppercase word. This will deal correctly with most usages of abbreviations like *etc.* or *Jr.* which can occur sentence medially or finally.
- Disqualify a boundary with a ? or ! if:
  - It is followed by a lowercase letter (or a known name).
- Regard other putative sentence boundaries as sentence boundaries.

Summary

- Sentence segmentation is a challenging task in NLP
- Issues
  - Boundary detection
  - Abbreviations
  - Inconsistent punctuations
- Approaches for SBD

Last word: General steps in text processing

- Load the text
- Markup removal
- Sentence segmentation
- Word Tokenization
- Specific to task
  - Stop word removal
  - Stemming
  - Lemmatization
- Bag-of-word representation/ Word count / TF-IDF weights/ Word embedding