Lecture 2: Linguistic Essential

Book ch 2

Book Ch 2.2,2.3

Corpus: (plural, Corpora)

means a large collection of computer readable text or speech.

Example: Brown Corpus

- a million word collection of samples from 500 written English texts
- from different genres (newspaper, fiction, non fiction, academic etc.)
- Assembled at Brown University

Text Normalization:

→ means converting texts to a more convenient, standard form.

These includes:

Sentence Segmentation

Tokenization

Lemmatization / Stemming

etc.

Sentence Segmentation:

Given a text, we need to seperate each sentences.

Challenges:

Usually, punctuation like ., ?, ! end sentences but not always.

* Some '.' are in abbreviations (shorten form of a word/phrase)

Ex: Mr. Smith is eating. He will go home later.

* Some '.' are in abbreviations also end sentences.

Ex: Mr. Smith is going to U.S.A.

* Quotes after . ? ! are in the same sentence.

Ex: He said, "I am an American citizen."

Solution:

① Rule Based Approach:

- Easy to write a few rules
- But Large set of rules are hard to maintain.

② Machine Learning Approach:

- Classify each punctuation character:

0 → <Not EOS>
1 → <EOS>

EOS = End of sentence

- Features: surrounding characters, words
- 99% Accuracy.

Example:

Mr. Smith lives in U.S.A. He said " I am an American citizen!"

Not EOS | Not EOS | EOS | Not EOS | NOT EOS | EOS

③ Parsing (spacy's Algorithm):

- Let the dependency parser figure it out.

Ch- 2.4.2

Tokenization:

→ the process that breaks down a body of text into smaller unit called tokens.

In case of word tokenization, tokens can be word, subword or punctuations.

Challenges:

→ Seperating based on white space is not enough.

Words with punctuation: C++, C#, M*A*S*H, etc.
Emoticons: =) :) ;-) etc.
Contractions: I'll, isn't, dog's, etc.
Typically split to separate, e.g., noun (I) from verb ('ll)
Hyphens in words: e-mail, co-operate, etc.
Hyphens between morphemes: non-lawyer, pro-Arab
Hyphens between words: once-quiet study, take-it-or-leave-it offer, 26-year-old, etc.
Names: New York vs. York
Phrasal verbs: make up, work out, etc.
Phone numbers: +(880) 1756-111111

Some common challenges in English Language

* The process of tokenization depends on the language also as each language has its own tokenization principles.

* Spacy tokenizer works on:

① Recursively split on white space
② Uses known Exceptions, Affixes
③ Seperates punctuation

Stemming & Lemmatization:

Ch 2.4.4

Similar word looks different.
(Dog, Dogs), (run, running, ran)

on a computer,
dog != dogs

Solution: Stemming and Lemmatization

Stemming:

- reduces words to their root form or base form by stripping of prefixes & suffixes.
- fast, not accurate
- Porter Stemmer (1980):

Organization → Organ
European → Europe
running → run

Lemmatization:

- The task of determining that two word have same root.
- Hand built lexicons for all word forms
- Think it as a dictionary:

```
{ 'running' : 'run',
  'runner'  : 'run',
  'ran'     : 'run',
   :
}
```

- Accurate, but slow.
- has a Chicken egg scenario with POS tagging.

Embedding:

A technique that processes a word or phrase to a numerical vectors. [Details in Lecture 4]

NLP Libraries:

Spacy:	SOTA, fast, python
NLTK:	Slower, simple, python
Core NLP:	SOTA, fast, Java

NLP Annotation:

- Associating extra information to a piece of text.

- Part of speech (POS) Tagging
- Named Entity Recognition (NER)

} Details:

Ch: 8.1, 8.2, 8.3

POS tagging:

- Assigning grammatical categories for words

Ex:

She	liked	it	very	much	.
pron	verb	pron	adv	adv	punc

- Details of POS tags: Search
  - ① Penn TreeBank tags
  - ⑪ Universal POS tags
- Tags can be divided in two class:

① closed class: These categories have fixed set of words.

eg: Prepositions, Determiners, Pronouns, Conjunctions, Auxilery verbs, particles, numerals.

Conjunction: and, or, but, ... [fixed set of words]

② Open class: These categories have a growing set of words.

eg: Noun, Verbs, Adjectives, Adverbs

POS Tag Challenges:

- Words are ambigious.
  One word can have multiple POS Tags. It depend on context.

⇒ Swimming is a good exercise.
Noun

⇒ He is swimming.
Verb

Named Entity Recognition: (NER):

- Identify phrases that are named people, location, organization, etc.
- Common NER Tags:
  PER → Person
  LOC → Location
  ORG → Organization
  GPE - Geo Political Entity
  ART - Creative/Artwork
- More Elaborate Sceme:
  ① BIO Tag → Rahim Mia is in Dhaka
  B-PER I-PER O O B-LOC
  B - Begin, I - Inside, O - Outside
  ② BILOU

NER Challenges:

Ambiguity:

Ex:

① Washington was born into slavery. PER
② Washington went up 2 games to 1. ORG
③ Blair arrived in Washington today. LOC
④ Washington passed a primary seatbelt law. GPE

Solution: Sequence tagging. [will be discussed later]

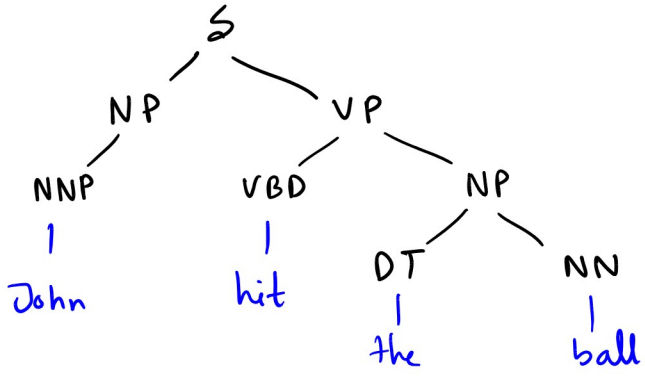
Parsing & Syntactic Representation:

Parsing: The process of analyzing a sentence grammatical structure to understand the relationship between words.

① Constituency Parsing:

- Breaks a sentence into nested sub-phrases (constituents) like Noun Phrase, Verb phrase based on a structure grammar

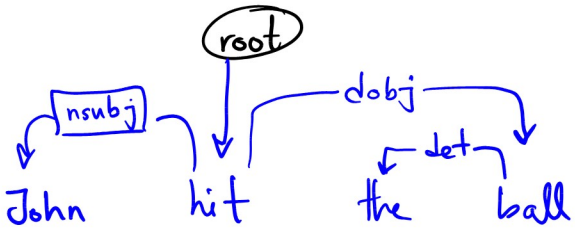
- Constituency tree:



② Dependency parsing:

- Analyzes grammatical relationship between words by identifying which word depends on/modify others in a sentence.

- Dependency Tree

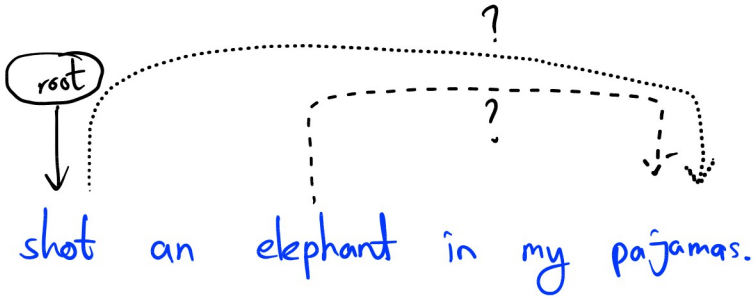


Parsing Challenges:

- Attachment ambiguity

One morning I shot an elephant in my pajamas.

Who was in my pajamas? Me? elephant?



- Coordination Ambiguity:

Old men and women

Old (men and women)? or Old (men) and women?

Parsing Solution: ① Probabilistic grammar based parsing
② Transition based parsing.