Natural Language Processing
Lecture 1

Lecture by Dr. Kfir Bar
Typeset by Steven Karas

2019-03-05
Last edited 21:00:50 2019-03-05

Disclaimer These notes are based on the lectures for the course Natural Language Processing, taught by Dr. Kfir Bar at IDC Herzliyah in the spring semester of 2018/2019. Sections may be based on the lecture slides prepared by Dr. Kfir Bar.

1 Course structure & policy

It is recommended to read the suggested papers before each class.

2 Grading

50% homework assignments, 4 assignments. 50% final project.

It will be allowed to work on projects together. Homeworks will be submitted as a text file with a link to a Google Colab notebook. We will use PyTorch, numpy, pandas, etc for homeworks.

A list of suggested projects will be published later in the course, and will also be done in pairs.

3 Introduction

Natural language processing is the study of having a computer understand human languages. NLP is sometimes referred to as computational linguistics, but usually by linguists. The three main facets of NLP are transformations:

1. Unstructured text $\rightarrow$ structured data
2. Unstructured text $\rightarrow$ unstructured text
3. Structured data $\rightarrow$ unstructured text

A common application of NLP is chatbots, that answer simple queries in natural language. The Turing Test is a famous test whereby a chatbot is indistinguishable from a human. Eliza[4] from 1966 was one of the first chatbots. It used regular expressions to transform input statements into questions.

Historically, the field used rule based systems from the 50s until the 90s, at which point it became more data driven. In the early aughts, machine learning took over, and since 2014 deep learning has taken over as the state of the art. Deep learning generally treats the entire process as a black box to be trained, rather than using techniques such as bag of words, part of speech tags, etc.

NLP is a difficult field because sometimes the raw data is extremely dirty. For example, garden path sentences, nonsense or spam input, idioms, ambiguous statements, etc. Different dialects of a language also pose an issue, as do regional slang and turns of phrase. For example, certain regions of British English are mutually intelligible yet wildly different from North American Vernacular English.

3.1 Language

There are approximately 7000 spoken languages in the world, but this number is obviously debated. Languages are grouped into a hierarchy of families, usually based on their development and shared characteristics.

Zipf's Law The product of the frequency of words f and their rank r is approximately constant.

3.2 NLP layers

- Phonetics - the sounds that words represent[1]
- Morphology - word or word part structures
- Syntax - phrase structure; word modifiers
- Semantics - the literal meaning of words
- Pragmatics - conclusions of language

4 Morphology

4.1 Segmentation/Tokenization

A simple approach would be to simply split on whitespace. A better approach would split on punctuation.

English is one of the easier languages to tokenize, but there are notable exceptions, for example "O'Hare" and "you've". Many Asian languages have no spaces between words, and require more advanced approached.

4.2 Morphology types

Inflectional Morphology Words in different grammatical contexts can change the word. For example, "duck" becomes "ducks" when it is plural.

Derivational Morphology Derivational morphology creates new words based on context for example "sense" becomes "nonsense".

Compounding Some languages such as German, Turkish, Inuktit, Finnish, etc combine base words into new words.

4.2.1 Example: Hebrew

Hebrew has 5000 roots, 50000 lemmas (dictionary entries), 2000000 prefix free words, 10000000 words, and about 10000000 speakers. A single token can have up to 5 word parts: conjunction, article, preposition, stem, possession.

4.3 Terminology

Token/Surface word the complete word as written
Term a unique token class
Morpheme non recursive, sub word unit of a language
Stem the word without prefixes or suffixes; the part that is common to all its inflected variants
Lemma the canonical form of a word (the dictionary entry)
Root in some languages, the primary lexical unit of a word
Affixes a bound morpheme that modifies a word

4.4 Part of Speech Tagging

Given a sentence, we want to tag each word with the word types. For example:

> I want to book a flight to New York next week

would become:

> I /Pronoun, want /Verb, to /Preposition, book /Verb, a /Indefinite Article, flight /Noun, to /Preposition, New /Proper Noun, York /Proper Noun, next /Adjective, week /Noun

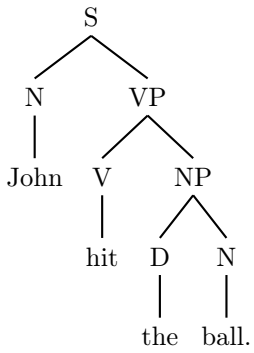
[1] Out of scope for this course

4.5 Sentence Segmentation

A simple method would split on punctuation. A more advanced approach is described in the suggested reading for next week, which considers a tagged corpus of sentence breaks.

5 Syntax

Forming grammatical phrases and sentences by putting words into structures.



5.1 Ambiguous Syntax

This is an example of a sentence where syntax helps resolve lexical ambiguities, as book can only be a verb in this sentence:

> Did you book a ticket?

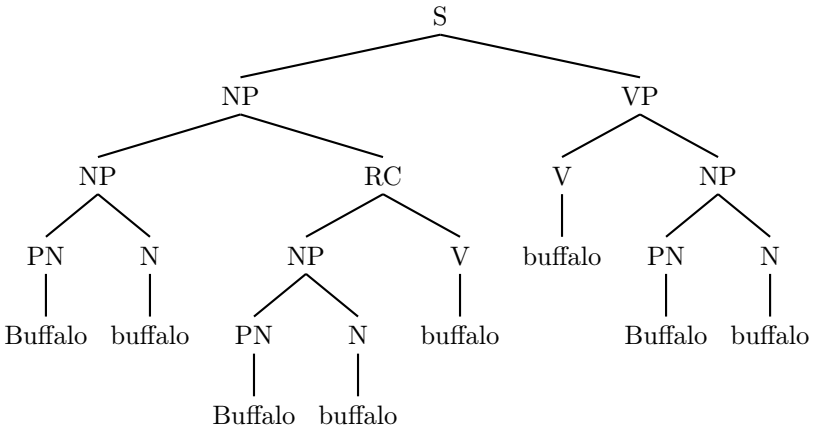
This is a syntactically ambiguous sentence:

> Bear left at zoo

5.2 Pathological Cases

This is a syntactically valid English sentence:

> Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo



5.3 Sentence order

I didn't have time to record this section. See slide 70.

5.4 Dependency parsing

See slide 72.

References

[1] Yoav Goldberg. A primer on neural network models for natural language processing. *CoRR*, abs/1510.00726, 2015.

[2] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., 2009.

[3] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[4] Joseph Weizenbaum et al. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.