Advanced NLP Text Processing

MIE223
Winter 2025

1 Advanced NLP Text Processing

1.1 NLP Text Processing Pipeline

- Document → Sections and Paragraphs
- Paragraphs → Sentences (sentence segmentation / extraction)
- Sentences → Tokens
- Tokens → Lemmas or Morphological Variants / Stems
- Tokens → Part-of-speech (POS) Tags
- Tokens, POS Tags → Phrase Chunks (Named entities and Keyphrases)
- Tokens, POS Tags → Parse Trees
- Augment above with coreference, entailment, sentiment, ...

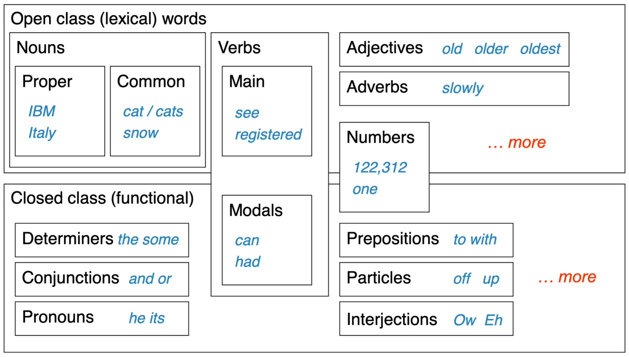
nltk covers POS tagging, phrase chunking Stanford NLP toolkit provides parsing, coreference, NER.

2 Part-of-Speech Tagging

2.1 Parts of Speech

A simple but useful form of linguistic analysis

- Perhaps starting with Aristotle in the West (384–322 BCE), there was the idea of having parts of speech
  - a.k.a lexical categories, word classes, "tags", POS
- It comes from Dionysius Thrax of Alexandria (c. 100 BCE) the idea that is still with us that there are 8 parts of speech
  - The ones we teach today
    - School grammar: noun, verb, adjective, adverb, preposition, conjunction, pronoun, interjection



```
there are open classes and closed classes for words.
Open classes are nouns, verbs, adjectives, and adverbs.
Closed classes are determiners, pronouns, prepositions,
and conjunctions.
```

2.2 Open vs. Closed classes

Open vs. Closed classes

- Closed:
  - determiners: a, an, the
  - pronouns: she, he, I
  - prepositions: on, under, over, near, by, ...
  - Why “closed”?
- Open:
  - Nouns, Verbs, Adjectives, Adverbs.

2.3 POS Tagging

- Words often have more than one POS: back
  - The back door = JJ
  - On my back = NN
  - Win the voters back = RB
  - Promised to back the bill = VB
- The POS tagging problem is to determine the POS tag for a particular instance of a word.
- Input: Plays well with others
- Ambiguity: NNS/VBZ UH/JJ/NN/RB IN NNS
- Output: Plays/VBZ well/RB with/IN others/NNS
- Uses:
  - Text-to-speech (how do we pronounce “lead”, “record”, “wind”?)
  - Can write regexps like (Det) Adj* N+ over the output for phrases, etc.

2.4 How difficult is POS tagging?

- About 11% of the word types in the Brown corpus are ambiguous with regard to part of speech
- But they tend to be very common words. E.g., that
  - I know that he is honest = IN
  - Yes, that play was nice = DT
  - You can’t go that far = RB
- 40% of the word tokens are ambiguous

```
no POS tagging tested on exams.
```

3 Phrase Chunking and Special Noun Phrases

3.1 Phrase Chunking

Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.

Note 1. A phrase should not contain a subphrase of the same type. i.e. ”New York Times” is a NP, but ”New York” is not.

> [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
>
> [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

3.2 Named Entity Recognition (NER)

A special class of Proper Noun Phrases

- People: Scott Sanner, President Obama, Madonna
- Places: New York, Madison Square Garden, Millenium Park
- Organizations: New York Times, University of Toronto

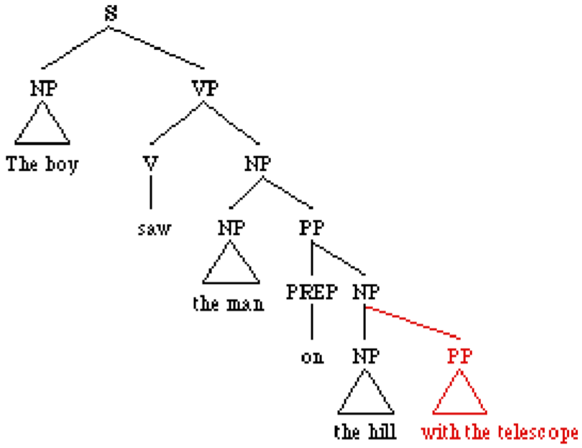
3.3 Keyphrases

- Useful noun phrases, but not necessarily Proper Nouns, e.g.,
  - "machine learning"
  - "support vector machines"
  - "genetically modified organisms"
- A subset of frequent noun phrases (harder to extract than NEs)
  - This paper has the best method I've found so far: "Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method Katerina Frantziy, Sophia Ananiadouy, Hideki Mima" IJODL 2000. http://personalpages.manchester.ac.uk/staff/sophia.ananiadou/ijodl2000.pdf

4 Statistical Natural Language Parsing

Parsing: Two views of syntactic structure

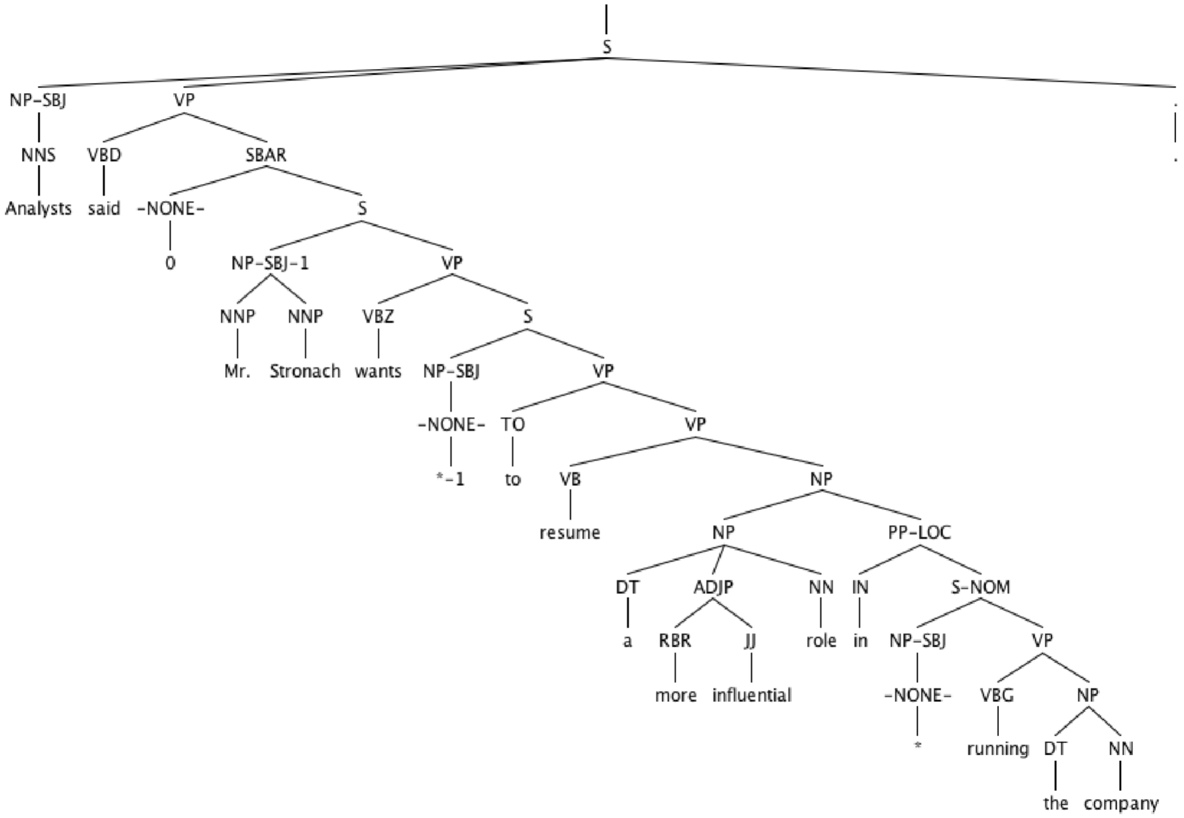
4.1 Why parsing?



- "The boy saw the man on the hill with the telescope."
  - Who had the telescope?
- Depends on whether you attach "with the telescope" to "I" or "man on the hill"
- How do you determine attachments? Parsing.
  - Some sentences are inherently ambiguous: attachment ambiguity

4.2 Grammars for Parse Tree Production

- Parent → Child1 Child2 — Child3 Child4 ... — .
- S → NP VP — ...
- NP → ... NN* ...
- VP → ... VB* ...
- ADJP → ... JJ* ...
- ADVP → ... RB* ...



5 Semantic Language Analysis

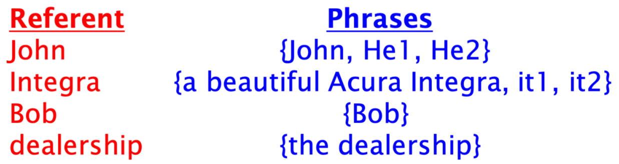
Coreference and entailment

5.1 Coreference

- Discourse (multiple sentences) use coreferring phrases.
- Example:
  - “John saw a beautiful Acura Integra in the dealership. He showed it to Bob. He bought it.”
- What do “He” and “it” refer to in the 2nd sentence?

5.2 Coreference Resolution

- “John saw a beautiful Acura Integra in the dealership. He1 showed it1 to Bob. He2 bought it2.”
- Important in processing reviews: “I liked it!”



Referent	Phrases
John	{John, He1, He2}
Integra	{a beautiful Acura Integra, it1, it2}
Bob	{Bob}
dealership	{the dealership}

5.3 Entailment

- Question: When did the Berlin wall open?
- Text contains: The Berlin wall fell on November 9, 1989.
- Simple entailment? Does “fall” → “open”?
  - A wall falling is a wall opening
  - A person falling is not a person opening
- Entailment can be highly contextual. But WordNet (in nltk) contains basic entailments, e.g., “snoring” → “sleeping”.