ISE 540 Text Analytics

Mayank Kejriwal
Research Assistant Professor/Research Lead
Department of Industrial and Systems Engineering
Information Sciences Institute
USC Viterbi School of Engineering
kejriwal@isi.edu

Natural Language Tasks

- Processing natural language text involves many various syntactic, semantic and pragmatic tasks in addition to other problems.

Syntactic Tasks

Word Segmentation

- Breaking a string of characters (graphemes) into a sequence of words.
- In some written languages (e.g. Chinese) words are not separated by spaces.
- Even in English, characters other than white-space can be used to separate words [e.g. , ; . - : ( )]
- Examples from English URLs:
  - jumptheshark.com ⇒ jump the shark .com
  - myspace.com/pluckerswingbar

    ⇒ myspace .com pluckers wing bar

    ⇒ myspace ⊗.com plucker swing bar

Morphological Analysis

- ***Morphology*** is the field of linguistics that studies the internal structure of words. (Wikipedia)
- A ***morpheme*** is the smallest linguistic unit that has semantic meaning (Wikipedia)
  - e.g. “carry”, “pre”, “ed”, “ly”, “s”
- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried $\Rightarrow$ carry + ed (past tense)
  - independently $\Rightarrow$ in + (depend + ent) + ly
  - Googlers $\Rightarrow$ (Google + er) + s (plural)
  - unlockable $\Rightarrow$ un + (lock + able) ?

    $\Rightarrow$ (un + lock) + able ?

Part Of Speech (POS) Tagging

- Annotate each word in a sentence with a part-of-speech.

I	ate	the	spaghetti	with	meatballs.
Pro	V	Det	N	Prep	N

- Useful for subsequent syntactic parsing and word sense disambiguation.

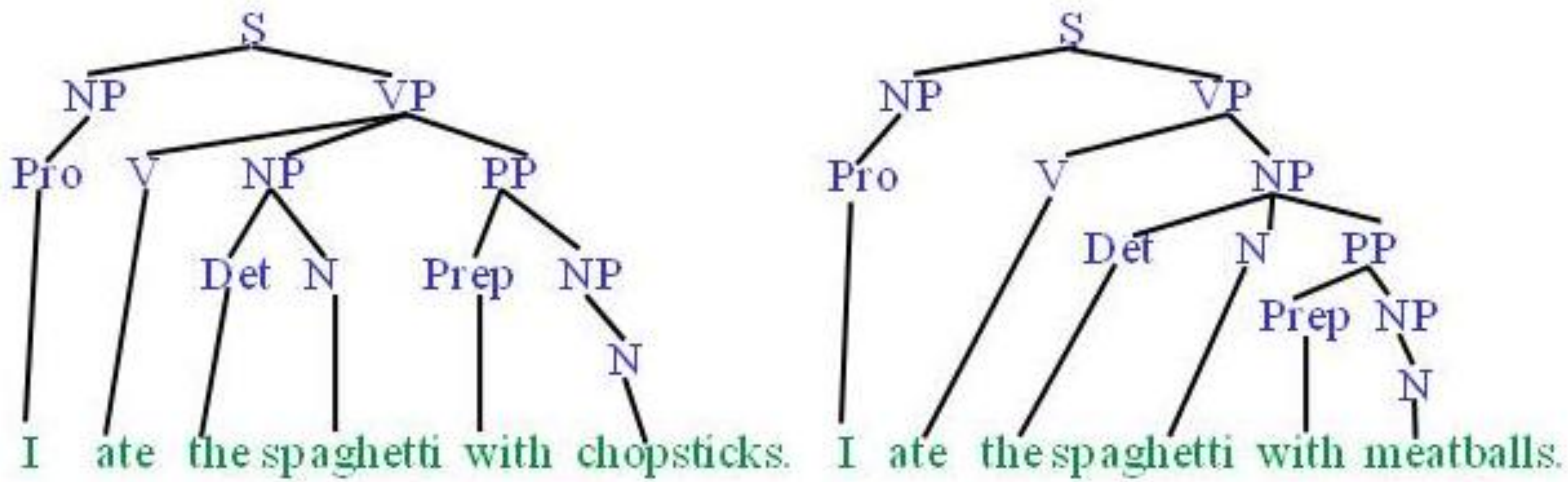
John	saw	the	saw	and	decided	to	take	it	to	the	table.
PN	V	Det	N	Con	V	Part	V	Pro	Prep	Det	N

Phrase Chunking

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

Syntactic Parsing

- Produce the correct syntactic parse tree for a sentence.



S
NP
VP
Pro
V
NP
PP
Det
N
Prep
NP
N
I ate the spaghetti with chopsticks.
S
NP
VP
Pro
V
NP
Det
N
PP
Prep
NP
N
I ate the spaghetti with meatballs.

Semantic Tasks

Word Sense Disambiguation (WSD)

- Words in natural language usually have a fair number of different possible meanings.
  - Ellen has a strong interest in computational linguistics.
  - Ellen pays a large amount of interest on her credit card.
- For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

Semantic Role Labeling (SRL)

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.

  agent patient source destination instrument

  - John drove Mary from Austin to Dallas in his Toyota Prius.
  - The hammer broke the window.
- Also referred to a “case role analysis,” “thematic analysis,” and “shallow semantic parsing”

Semantic Parsing

- A ***semantic parser*** maps a natural-language sentence to a complete, detailed semantic representation (***logical form***).
- For many applications, the desired output is immediately executable by another program.
- Example: Mapping an English database query to Prolog:

  How many cities are there in the US?

```
answer(A, count(B, (city(B), loc(B, C),
                        const(C, countryid(USA))),
                    A))
```

Textual Entailment

- Determine whether one natural language sentence entails (implies) another under an ordinary interpretation.

Textual Entailment Problems from PASCAL Challenge

TEXT	HYPOTHESIS	ENTAILMENT
Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.	*Yahoo bought Overture.*	TRUE
Microsoft's rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.	*Microsoft bought Star Office.*	FALSE
The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.	*Israel was established in May 1971.*	FALSE
Since its formation in 1948, Israel fought many wars with neighboring Arab countries.	*Israel was established in 1948.*	TRUE

Pragmatics/Discourse Tasks

Anaphora Resolution/ Co-Reference

- Determine which phrases in a document refer to the same underlying entity.
  - John put the carrot on the plate and ate it.
  - Bush started the war in Iraq. But the president needed the consent of Congress.
- Some cases require difficult reasoning.
  - Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Ellipsis Resolution

- Frequently words and phrases are omitted from sentences when they can be inferred from context.

"Wise men talk because they have something to say;
fools, because they have to say something.“ (Plato)

"Wise men talk because they have something to say;
fools talk because they have to say something.“ (Plato)