TEXT ANALYSIS AND COMPREHENSION:

BASIC CONCEPTS; CHALLENGES;

APPLICATION DOMAINS

Jelena Jovanović

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

Outline

- Text analysis and comprehension:
  - Why is it relevant? Why do we need it?
  - What challenges does it face?
  - What are typical approaches to text analysis and comprehension?

Why is it relevant? Why do we need it?

- Context-aware spelling and grammar check
- Semantic search
  - More advanced than traditional, keywords-based search
- Information extraction
  - Extraction of entities and their relationships from texts of different sorts
- Machine (automated) translation

Why is it relevant? Why do we need it?

- New interfaces
  - Dialog-based systems
- Business applications:
  - reputation management
  - context-aware advertising
  - business analytics
  - …

What are the challenges?

The complexity of human language

Some examples:

Mary and Sue are sisters.

Mary and Sue are mothers.

Joe saw his brother skiing on TV. *The fool*…

… didn't have a jacket on!

… didn't recognize him!

What are the challenges?

Examples (cont.)

Today hundreds of **planes land** daily on JFK **runway**.

Planes that once were parked on his **land** now are rolling down the **runway**.

I **deposited** $100 in the **bank**.

The river **deposited** sediment along the **bank**.

What are the challenges?

To sum up, human language is:

- Full of ambiguous terms and phrases
- Based on the use of context for defining and conveying meaning
- Full of fuzzy, probabilistic terms
- Based on commonsense knowledge and reasoning
- Influenced by and an influencer of human social interactions

What are the challenges?

Complex, layered structure of human language:

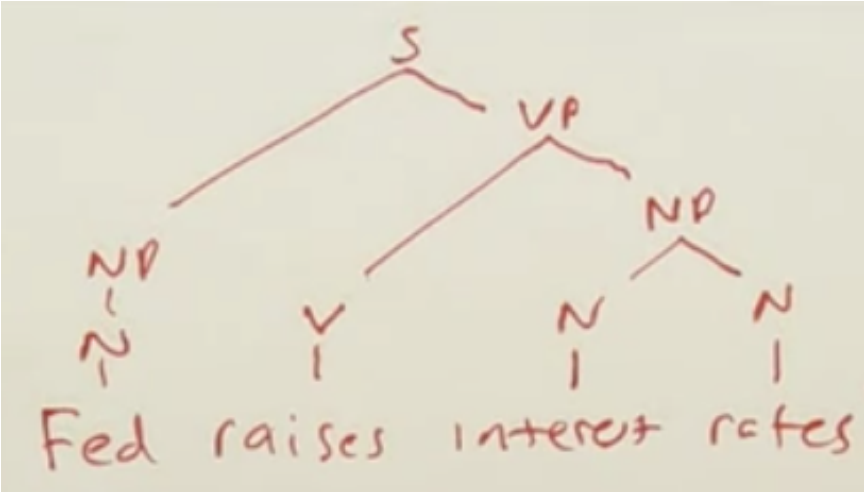
- What words appear in the given piece of text?
- What phrases can be identified?
- Are there words that modify the meaning of other words?
- What is the (literal) meaning of the identified words and phrases?
- What can be deduced from the fact that someone said something in the given context?
- What kind of reaction could be expected?

What are the challenges?

The level of language analysis	Description	Example
Morphology	Recognizing words and the variety of their forms	use, uses, user – different forms of the same word
Syntax and Grammar	Recognizing the type of the word	There are 5 *rows* in the table. – *rows* is noun here; She *rows* 5 times per week. – *rows* is verb in this case
	Identifying how different words are related to one another	Bob went out; *he* needed some fresh air. – The pronoun *he* refers to *Bob*.
Semantics	Determining the meaning of words (often based on their context)	The car *driver* was injured. vs. The *driver* was installed in the computer

Language/text modeling

- Main approaches to text/language modeling:
  - Logical models
    - Rely on linguistic analysis of the text, and abstract representation of the sentence structure (typically in the form of a parse tree)
    - Models of this type are manually created



An example of tree-based model of a sentence structure

Image source: http://goo.gl/QgcQS9

Language/text modeling

- Main approaches to text/language modeling:
  - Stochastic models
    - Based on the probability of occurrence of individual words or sequences of *n* words (typically 2-4 words)*
    - These models are “learned” i.e., their creation is automated through the application of m. learning methods over large text corpora
  - Hybrid models
    - Combine characteristics of logical and stochastic models
    - E.g., assigning probabilities to individual elements of a tree-based language model

* a sequence of *n* words is often referred to as ***n-gram***

Recommendation

The *Natural Language Processing* topic within the course *Introduction to Artificial Intelligence* at *Udacity.com*

- URL: https://www.udacity.com/course/cs271

Lecture on *Natural Language Processing* held during the *International Summer School on Semantic Computing*, *Berkeley 2011*

URL: http://videolectures.net/sssc2011_martell_naturallanguage/