What is NLU and what is NLP (NLU vs NLP)

Nikitas N. Karanikolas
Dept. of Informatics and Computer Engineering
University of West Attica, Athens, Greece
nnk@teiath.gr
http://users.teiath.gr/nnk

Natural Language Understanding (NLU)

- Phonology
- Morphology
- Syntax
- Semantics
- Representation
- Disambiguation
- Anaphora resolution
- Pragmatics and Discourse analysis
- Resources
- Tools

Phonology

- See my previous international lecture
  Learning Phonology by Machines
  May 2, 2018, University of Tirana

Morphology [1]

- **Morphology – the internal structure of words**
- Morphology is the study of the internal structure of words and forms a core part of linguistic study today.
- The term morphology is Greek and is a makeup of morph- meaning ‘shape, form’, and -ology which means ‘the study of something’.

Word [1]

- Words are the smallest independent units of language
  - do not depend on other words.
  - can be separated from other units
  - can change position.
- Example: The man looked at the horse**s**.
  - **s** is the plural (morphology) marker, dependent on the noun horse to receive meaning
  - Horses is a word: can occur in other positions or stand on its own

Words and Morphemes [1]

- Other position:
  The horses looked at the man.
- On its own:
  What is the man looking at? – Horses.

- **Morphemes are the building blocks of morphology**
  - Words have internal structure: built of even smaller pieces
- SIMPLE WORDS: Don't have internal structure (only consist of one morpheme) eg work, build, run. They can't be split into smaller parts which carry meaning or function.
- COMPLEX WORDS: Have internal structure (consist of two or more morphemes) eg worker: affix -er added to the root work to form a noun.

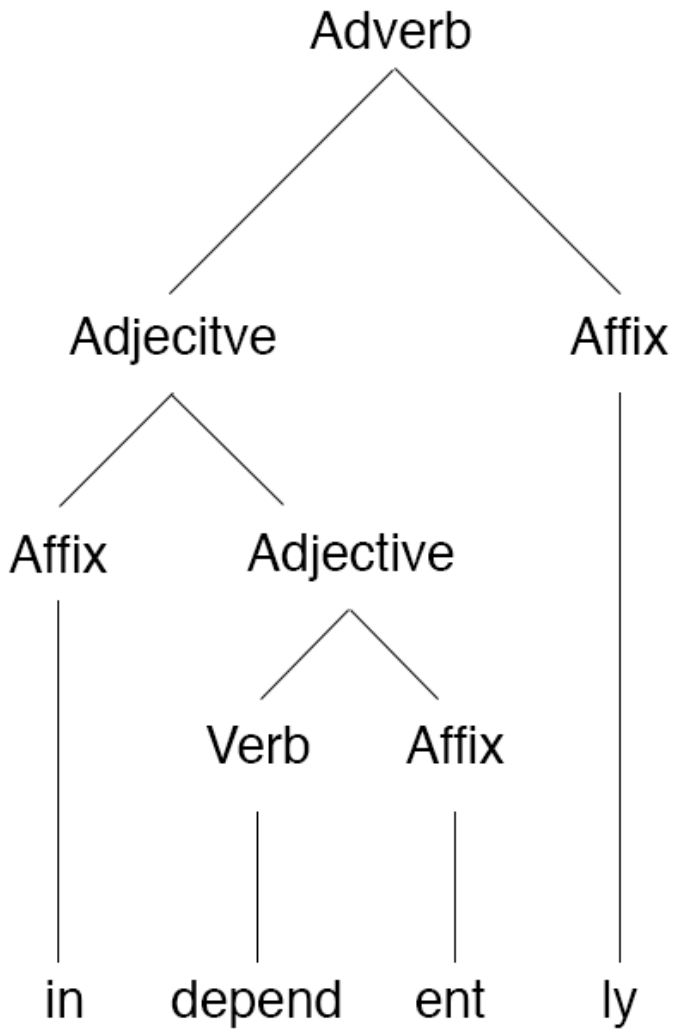
Word, Lexeme and Word form [2]

- The term "word" has no well-defined meaning[9]. Instead, two related terms are used in morphology: lexeme and word-form.
- Generally, a lexeme is a set of inflected word-forms that is often represented with the citation form in small capitals.[10] For instance, the lexeme eat contains the word-forms *eat, eats, eaten,* and *ate*. *Eat* and *eats* are thus considered different words-forms belonging to the same lexeme eat.
- *Eat* and *Eater*, on the other hand, are different lexemes, as they refer to two different concepts. Thus, there are three rather different notions of ‘word’.

Inflection vs. word formation [2]

- Given the notion of a lexeme, it is possible to distinguish two kinds of morphological rules. Some morphological rules relate to different forms of the same lexeme; while other rules relate to different lexemes.
- Rules of the first kind are inflectional rules, while those of the second kind are rules of word formation.
- The generation of the English plural *dogs* from *dog* is an inflectional rule, while compound phrases and words like *dog catcher* or *dishwasher* are examples of word formation.
- Informally, word formation rules form "new" words (more accurately, new lexemes), while inflection rules yield variant forms of the "same" word (lexeme).

A morphology tree [2]



- In, ent and ly are morphemes
- Depend (adj), Independ (adj), Independent (adj) and Independently (adverb) are lexemes

Why Morphology is needed for NLU?

- Part of speech tagging:
  Noun (N),
  Verb (V),
  Adjective (Adj),
  Adverb (Adv).
- Reducing the resources (lexicon entries) needed:
  For instance, we keep only the word-form retrieve and the system is able to conclude the other word-forms retrieves, retrieved, retrieving, retrieved) that belong to the same lexeme.

Syntax

- Syntactic will check if a sentence is well formed and will return the syntactic tree.
- The parts of this tree will then analyzed for representing the meaning of the sentence. There are restrictions for the allowed syntactic sub-structures that can correspond to semantic structures.
- Without syntactic analysis, we can not check these constraints.

A syntactic tree

- John broke the door with a hammer

```
s(  np(n(pn(John))),
    vp(   vp(   v(brake,past).
                np(det(the),n(door)))
          pp(   prep(with),
                np(det(a), noun(hammer)))
    )
)
```

Semantics

- A well formed syntactically sentence is not always correct
- “John drunk 3 liters gasoline” is syntactically correct but gasoline is a liquid that is not suitable for drinking by humans and John is a human.
- The semantics are that recognize that John is a proper name and consequently it refers to a human and that gasoline is a liquid that is not a kind of food or beverage
- There is also some semantic restriction that the consumed item in some verb of feeding should be food or beverage

Q & A

- Who defines that gasoline is a liquid that is not suitable for drinking by humans?
  Some Ontology.
- What is the tool that does the syntactic analysis?
  A parser.
- Where are the rules that guide the syntactic analysis?
  In the Grammar
- Where are lexemes exists?
  In the Lexicon.

The previous explain why

- Wikipedia [3] defines:
  Regardless of the approach used, most natural language understanding systems share some common components. The system needs a lexicon of the language and a parser and grammar rules to break sentences into an internal representation. The construction of a rich lexicon with a suitable ontology requires significant effort.

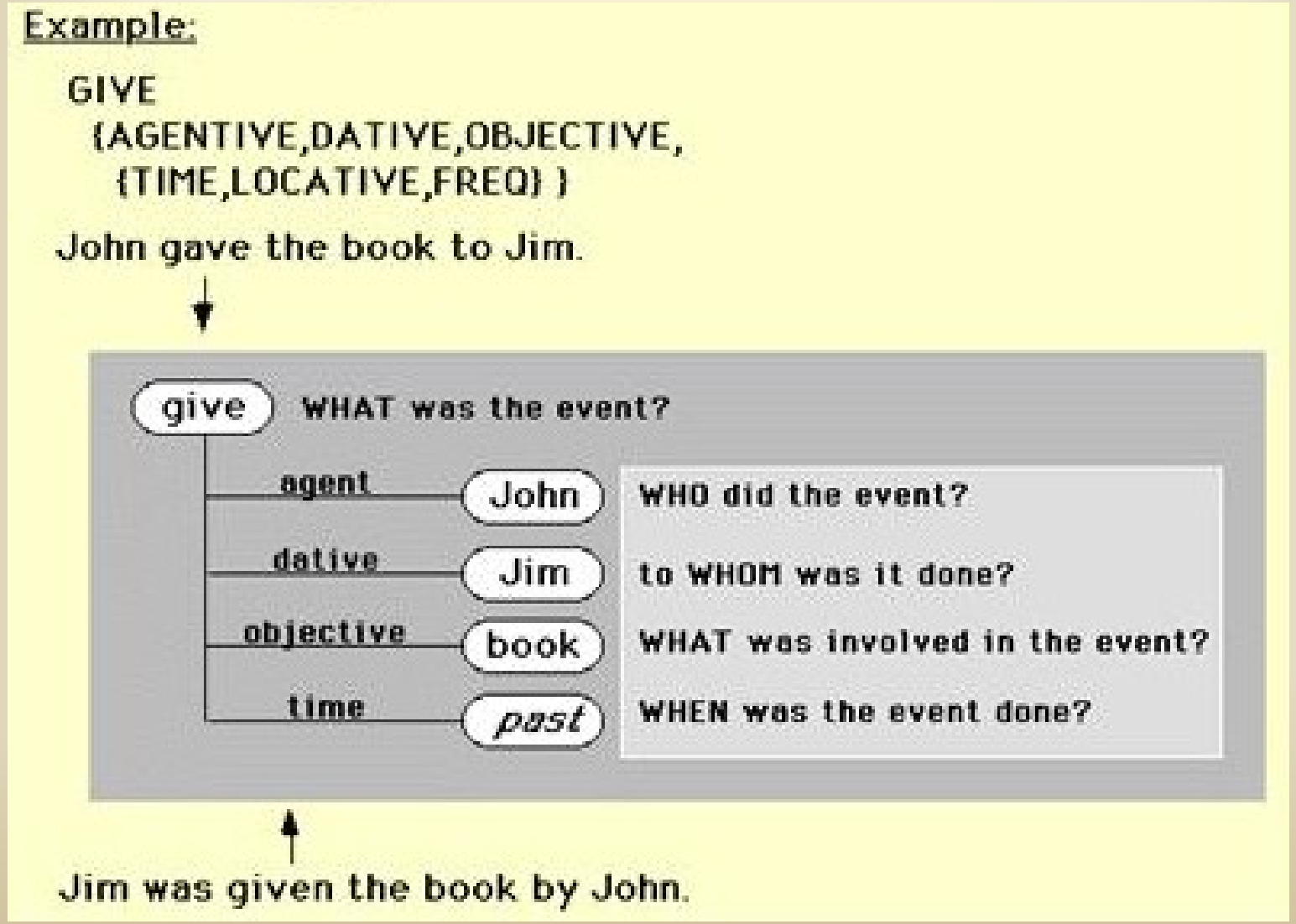
Representation

- On of the possible representations is the case grammar [4].
- The system was created by the American linguist Charles J. Fillmore in (1968). This theory analyzes the surface syntactic structure of sentences by studying the combination of deep cases (i.e. semantic roles) required by a specific verb.
- Deep cases are: Agent, Object, Benefactor, Location, Instrument, etc

Case Grammars

- For instance, the verb "give" in English requires an Agent (A) and Object (O), and a Beneficiary (B); e.g. "Jones (A) gave money (O) to the school (B).
- According to Fillmore, each verb selects a certain number of deep cases which form its **case frame**.
- Case frames are subject to certain constraints, such as that a deep case can occur only once per sentence.
- Some of the cases are obligatory and others are optional.
- Obligatory cases may not be deleted, at the risk of producing ungrammatical sentences. For example, *Mary gave the apples* is ungrammatical in this sense.

Case Grammar Example, from [5]



Ambiguity - Disambiguation

- I saw a man on a hill with a telescope meannings:
  - There's a man on a hill, and I'm watching him with my telescope.
  - There's a man on a hill, who I'm seeing, and *he* has a telescope.
  - There's a man, and he's on a hill that also has a telescope on it.
  - I'm on a hill, and I saw a man using a telescope.
  - There's a man on a hill, and I'm sawing him with a telescope.

Syntactic ambiguity [6]

- Look at the dog with one eye. meanings:
  - Look at the dog using only one of your eyes.
  - Look at the dog that only has one eye.
- Both (this and previous) is syntactically ambiguous

Syntactic Ambiguity Trees

```
s( np(n(you))
   vp( v(look,present),
       pp (p(at),np(det(the),n(dog))),
       pp (p(with),np(...,n(eye)))
   )
)

s( np(n(you))
   vp( v(look,present),
       pp (p(at), np( det(the),
                      np( n(dog),
                          pp( p(with),np(...,n(eye)))
                    )
       )
   )
)
```

Semantically Ambiguous Word Sense Disambiguation (WSD) [7]

- Slug meanings:
  - Coin
  - Bullet
  - Loafer
  - Gastropod without shell

Anaphora resolution

- John is going to visit Nick. He is a good man.
  Meanings:
  John (who is a good man) is going to visit Nick.
  John is going to visit Nick (who is a good man).
- He refers to:
  John (in the first case)
  Nick (in the second case)
- Solution:
  See my paper in 1993 “Pronominal and Anaphor Resolution” [8]

Pragmatics [9]

- Mary and Helen are mothers.
  The reader can understood that both (Mary and Helen) has the attribute of being mothers without having any relation between each other
- Tina and Flora are sisters.
  The reader can understood that Tina and Flora are sisters of each other.
- Why, in the first sentence, we do not interpret that Mary and Helen are mothers of each other?
  What makes the different interpretation?
  The beliefs that we (the readers) have. And our beliefs say that it is not possible a parent (mother) being child of her daughter.
- These beliefs (knowledge) are named pragmatics.

Discourse analysis [10]

- Discourse Analysis will enable to reveal the hidden motivations behind a text.
- Critical or Discourse Analysis is nothing more than a deconstructive reading and interpretation of a text.
- Discourse Analysis will enable us to understand the conditions behind a specific "problem" and make us realize that the essence of that "problem", and its resolution, lie in its assumptions; the very assumptions that enable the existence of that "problem".
- By enabling us to make these assumption explicit, Discourse Analysis aims at allowing us to view the "problem" from a higher stage and to gain a comprehensive view of the "problem" and ourselves in relation to that "problem".

Natural Language Processing (NLP)

- Summarization
- text Classification
- Computer Assisted Assessment
- Sentiment Analysis
- Opinion Mining
- Subjectivity Analysis
- Corpus Building

Summarization [11, 12]

- Can be Comprehensive (semantic oriented) or Extractive (shallow processing)
- Extractive is based on the selection of the most prominent sentences to convey the meaning. It is based on:
  - Weight of words (TF-IDF, TF-ISF, TF-RIDF)
  - Sentence Location (Baxentale, News Articles)
  - Title Words
- See also my previous lecture “Extractive summarization”, June 2017, NoviSad

Text Classification [13]

- Assign to a documents a class label (the category that the document belongs)
- Can be based on the existence of words or phrases
- The method needs training and training data (pre-classified documents)
- It is Critical to create an Authority List of words or phrases that will be appropriate to discriminate between classes

Computer Assisted Assessment [14 - 16]

- Mechanically assign a grate to an answer with respect to the expected (correct) answer.
- There is a need for positive training (correct answers and textbook) data and negative training data (erroneous answers)
- Can be based on phrases

Sentiment Analysis, Opinion Mining, Subjectivity Analysis

- Subjectivity Analysis
  classify a given text as subjective or objective
- Sentiment Analysis or Polarity Analysis
  Once a text is subjective
  Assign a score Positive or Negative
- Affective Computing
  Attempt to identify emotional charge
  - Happiness
  - Sadness
  - Fear
  - Anger - Passion

References

- [1] Shef, What is Morphlogy http://all-about-linguistics.group.shef.ac.uk/branches-of-linguistics/morphology/what-is-morphology/
- [2] Wikipedia, Morphology https://en.wikipedia.org/wiki/Morphology_(linguistics)
- [3] Wikipedia, Natural language understanding https://en.wikipedia.org/wiki/Natural_language_understanding
- [4] Wikipedia, Case Grammar https://en.wikipedia.org/wiki/Case_grammar
- [5] Fillmore grammatical cases - Perspective of Wallace Chafe http://linguistictheoryevolution.blogspot.al/2012/05/charles-fillmores-grammatical-cases.html

References

- [6] Linguistcs Online, Syntactic ambiguity http://languagelink.let.uu.nl/~lion/index.php?s=Grammar_exercises/grammar_4&lang=en
- [7] Wikipedia, Word-sense disambiguation https://en.wikipedia.org/wiki/Word-sense_disambiguation
- [8] Nikitas N. Karanikolas. Pronominal and Anaphor Resolution. Computing and Information Technology (CIT) journal, volume 1, No 3, 1993. http://users.teiath.gr/nnk/papers/A01_scan.pdf
- [9] Wikipedia, Pragmatics https://en.wikipedia.org/wiki/Pragmatics
- [10] Discourse Analysis: A resource book for students. http://routledgetextbooks.com/textbooks/_author/9780415610001-jones/
- [11] Nikitas N. Karanikolas, Eleni Galiotou and Christodoulos Tsoulloftas, A workbench for extractive summarizing methods. PCI'2012: 16th Panhellenic Conference on Informatics, October 5-7, 2012, Piraeus, Greece. IEEE CPS. https://ieeexplore.ieee.org/document/6377346/

References

- [12] Nikitas N. Karanikolas, "Extractive summarization methods – subtitles and method combinations". RTA-CSIT 2016, November 18 - 19, 2016, Tirana, Albania. http://
- [13] N. Karanikolas and C. Skourlas. A parametric methodology for text classification. Journal of Information Science, Vol. 36 (4), pp. 421-442, 2010, doi:10.1177/0165551510368620. http://journals.sagepub.com/doi/10.1177/0165551510368620
- [14] Nikitas N. Karanikolas. The role of phrases in Information Retrieval and related domains. eRA-4. 4th Conference for the contribution of Information Technology to Science, Economy, Society and Education, September 25-26, 2009, Spetses, Greece. http://users.teiath.gr/nnk/papers/B22_cr.pdf
- [15] Nikitas N. Karanikolas, Computer Assisted Assessment (CAA) of Free-Text: Literature Review and the specification of an alternative CAA system. In 19th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE 2010), June 28-30, 2010, Larissa, Greece, IEEE Xplore https://ieeexplore.ieee.org/document/5541991/
- [16] Nikitas N. Karanikolas, "Summarization as the base for Text Assessment". 4th IC-ININFO, September 5-8, 2014, Madrid, Spain. https://aip.scitation.org/doi/abs/10.1063/1.4907844