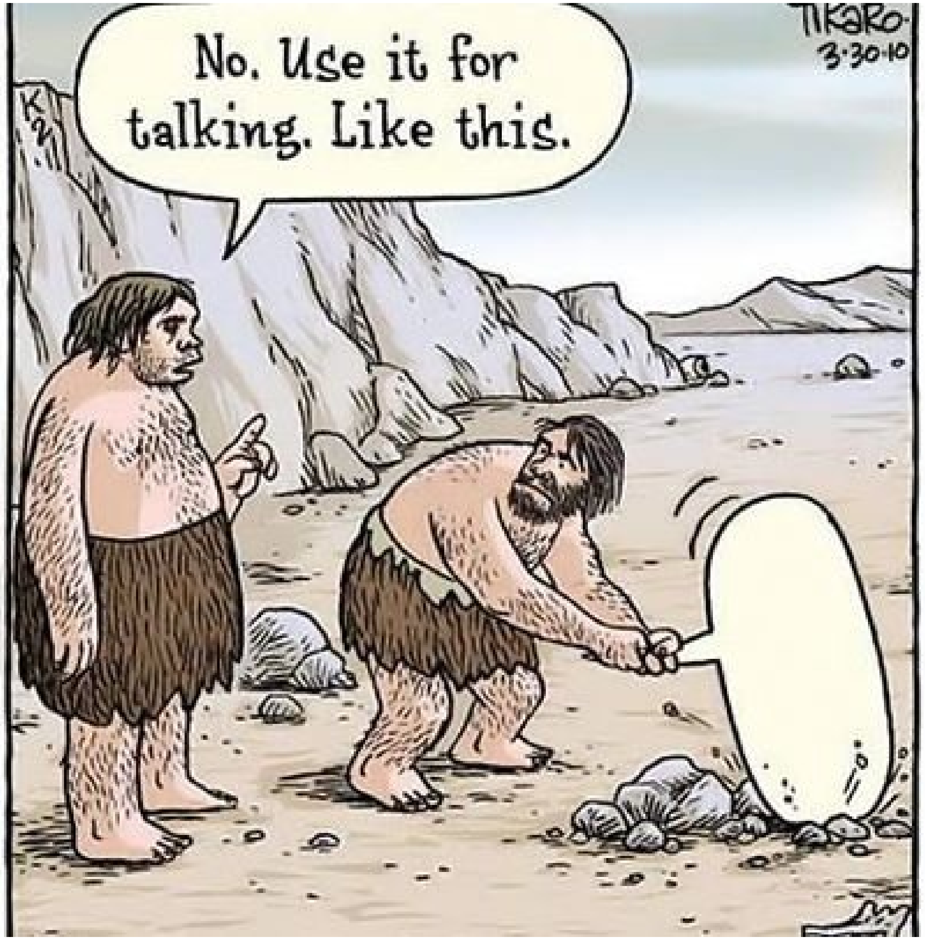
Natural Language Processing

Introduction, Terminology



Natural Language Processing

Computational linguistics is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective, as well as the study of appropriate computational approaches to linguistic questions.

Natural Language Processing

Natural Language Processing (NLP) refers to all systems that work together to handle end-to-end interactions between machines and humans in the preferred language of the human. In other words, NLP lets people and machines talk to each other “naturally.”

NLP is a critical piece of any human-facing artificial intelligence. An effective NLP system is able to ingest what is said to it, break it down, comprehend its meaning, determine appropriate action, and respond back in language the user will understand.

Natural Language Processing

Rule based approaches are often top down. The designer of an NLP tool based on such an approach writes declarative rules that emulate the knowledge of a domain expert.

Statistical techniques are often inductive and include things like latent semantic analysis, machine learning, etc. The machine-learning paradigm calls for using statistical inference to automatically learn such rules through the analysis of large corpora of typical real-world examples, a corpus, and generate structure probabilisticly rather than emulating the mind of an expert.

Natural Language Processing

No unique reference book for the course: the slides + the pointers to material available on the web or on the AulaWeb module are enough for understanding the subject, even for students who cannot attend the frontal lessons.

Some lessons are based on the book: Speech and Language Processing (3rd ed. draft) by Dan Jurafsky and James H. Martin

https://web.stanford.edu/~jurafsky

Some slides integrate slides associated with that book.

Levels of linguistic analyses

Pragmatics: what does it do?

Semantics: what does it mean?

Syntax: what is grammatical?

natural language utterance

Analogy with programming languages

Syntax: no compiler errors

Semantics: no implementation bugs

Pragmatics: implemented the right algorithm

Analogy with programming languages

Syntax: no compiler errors

Semantics: no implementation bugs

Pragmatics: implemented the right algorithm

Different syntax, same semantics (5):

$$2 + 3 \quad \Leftrightarrow \quad 3 + 2$$

Same syntax, different semantics (1 and 1.5):

$$3 \;/\; 2 \text{ (Python 2.7)} \quad \not\Leftrightarrow \quad 3 \;/\; 2 \text{ (Python 3)}$$

Good semantics, bad pragmatics:

correct implementation of deep neural network
for estimating coin flip prob.

NLP: Syntax

Regular Expressions

NLP: Syntax

Word Level

- Word segmentation
- Identification of stop words
- Stemming and Lemmatization
- Minimum edit distance

NLP: Syntax

Sentence Level

- Sentence breaking
- Part of Speech Tagging
- Parsing

NLP: Semantics

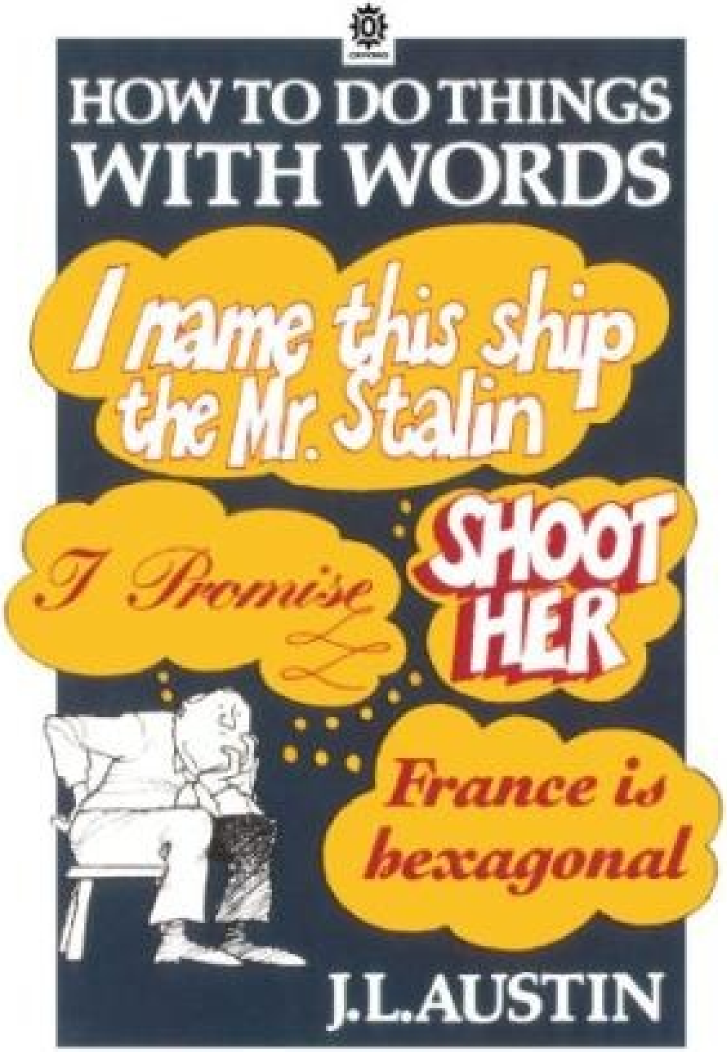
- Distributional semantics
- Lexical semantics
- Ontologies
- Word sense disambiguation
- Named entity recognition (NER)

NLP: Most common (non trivial) features for representing a text

- Bag of Words
- N-grams (at word and char level)
- TF-IDF
- Flesch-Kincaid readability measure
- Linguistic inquiry and word count (LIWC)
- Emotion words (Ekman, Plutchik)
- ....

NLP: Pragmatics

- “How to do things with words”



NLP: Applications

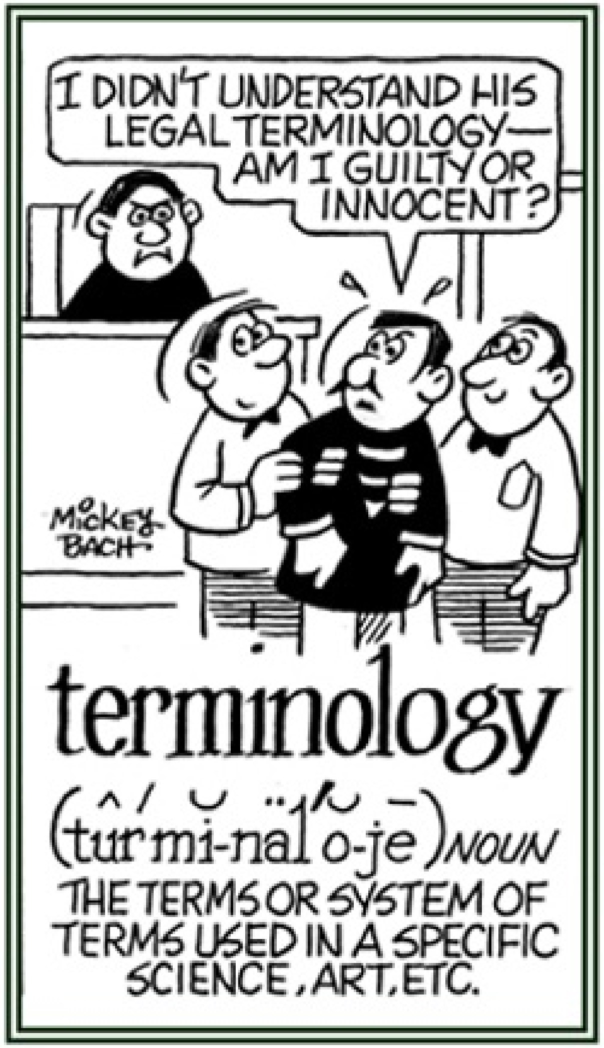
- Information filtering and retrieval
- Machine translation
- Question answering
- Automatic summarization
- Sentiment analysis
- Profiling
- **Chatbots**

https://www.lifewire.com/applications-of-natural-language-processing-technology-2495544

NLP: Tools

- NLP tools, https://opensource.com/business/15/7/five-open-source-nlp-tools
- NLP tools, http://www.phontron.com/nlptools.php
- WordNet, https://wordnet.princeton.edu/
- BabelNet, http://www.babelnet.org/
- ......

Terminology


I DIDN'T UNDERSTAND HIS LEGAL TERMINOLOGY— AM I GUILTY OR INNOCENT?
MICKEY BACH
terminology
(tûr′mi-nŏl′o-jē) NOUN
THE TERMS OR SYSTEM OF TERMS USED IN A SPECIFIC SCIENCE, ART, ETC.

Terminology

the study of terms

terminology

[tur-*muh*-nol-*uh*-jee]

noun, plural *'terminologies'*

1. the system of terms belonging or peculiar to a science, art, or specialized subject; nomenclature.
2. the science of terms, as in particular sciences or arts.

Word Origin and History for *'terminology'*

, from German Terminologie (1786), a hybr
by C.G. Schütz of Jena, from Medieval
word, expression" (see terminu
with, a speaking

Morpheme: A morpheme is the smallest grammatical unit in a language. In other words, it is the smallest meaningful unit of a language. The linguistics field of study dedicated to morphemes is called morphology. A morpheme is not identical to a word, and the principal difference between the two is that a morpheme may or may not stand alone, whereas a word, by definition, is freestanding.

Terminology

the study of terms

terminology

[tur-*muh*-nol-*uh*-jee]

noun, plural *'terminologies'*

1. the system of terms belonging or peculiar to a science, art, or specialized subject; nomenclature.
2. the science of terms, as in particular sciences or arts.

Word Origin and History for *'terminology'*

, from German Terminologie (1786), a hyb
by C.G. Schütz of Jena, from Medieval
ord, expression" (see termin
with, a speaking

Every morpheme can be classified as either free or bound.

Free morphemes can function independently as words (e.g. town, dog).

Bound morphemes appear only as parts of words, always in conjunction with a root and sometimes with other bound morphemes. For example, un- appears only accompanied by other morphemes to form a word. Most bound morphemes in English are affixes, particularly prefixes and suffixes.

Example: "Unbreakable" comprises three morphemes: un- (a bound morpheme signifying "not"), -break- (the root, a free morpheme), and -able (a free morpheme signifying "can be done").

Terminology

terminology
[tur-muh-nol-uh-jee]
noun, plural 'terminologies'
1. the system of terms belonging or peculiar to a science, art, or specialized subject; nomenclature.
2. the science of terms, as in particular sciences or arts.
Word Origin and History for 'terminology'

Bound morphemes can be further classified as derivational or inflectional.

Derivational morphemes, when combined with a root, change either the semantic meaning or part of speech of the affected word. For example, in the word happiness, the addition of the bound morpheme -ness to the root happy changes the word from an adjective (happy) to a noun (happiness). In the word unkind, un- functions as a derivational morpheme, for it inverts the meaning of the word formed by the root kind.

Inflectional morphemes modify a verb's tense, aspect, mood, person, or number, or a noun's, pronoun's or adjective's number, gender or case, without affecting the word's meaning or class (part of speech). Examples of applying inflectional morphemes to words are adding -s to the root dog to form dogs and adding -ed to wait to form waited. An inflectional morpheme changes the form of a word.

Terminology

he study of terms

terminology

[tur-*muh*-nol-*uh*-jee]

noun, plural *'terminologies'*

1. the system of terms belonging or peculiar to a science, art, or specialized subject; nomenclature.
2. the science of terms, as in particular sciences or arts.

Word Origin and History for *'terminology'*

, from German Terminologie (1786), a hyb
by C.G. Schütz of Jena, from Medieval
ord, expression" (see termin
with, a speaking

Word: a word is the smallest element that can be uttered in isolation with objective or practical meaning.

This contrasts deeply with a morpheme, which is the smallest unit of meaning but will not necessarily stand on its own. A word may consist of a single morpheme (for example: oh!, rock, red, quick, run, expect), or several (rocks, redness, quickly, running, unexpected), whereas a morpheme may not be able to stand on its own as a word (in the words just mentioned, these are -s, -ness, -ly, -ing, un-, -ed).

Terminology

the study of terms
terminology
[tur-muh-nol-uh-jee]
noun, plural 'terminologies'
1. the system of terms belonging or peculiar to a science, art, or specialized subject; nomenclature.
2. the science of terms, as in particular sciences or arts.
Word Origin and History for 'terminology'
, from German Terminologie (1786), a hyb
by C.G. Schütz of Jena, from Medieval
word, expression" (see termin
with, a speaking

Corpus (pl. Corpora): a text corpus is a large and structured set of texts (nowadays usually electronically stored and processed). They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.

Example: https://corpus.byu.edu/

Terminology

- **Genre:** a style or category of art, music, or literature.
- **Genre of a corpus:** usually defined on the basis of text-external features, such as medium, function and format.

Terminology

the study of terms
terminology
[tur-muh-nol-uh-jee]
noun, plural 'terminologies'
1. the system of terms belonging or peculiar to a science, art, or specialized subject; nomenclature.
2. the science of terms, as in particular sciences or arts.
Word Origin and History for 'terminology'

Collocation: A collocation is a sequence of words or terms that co-occur more often than would be expected by chance.

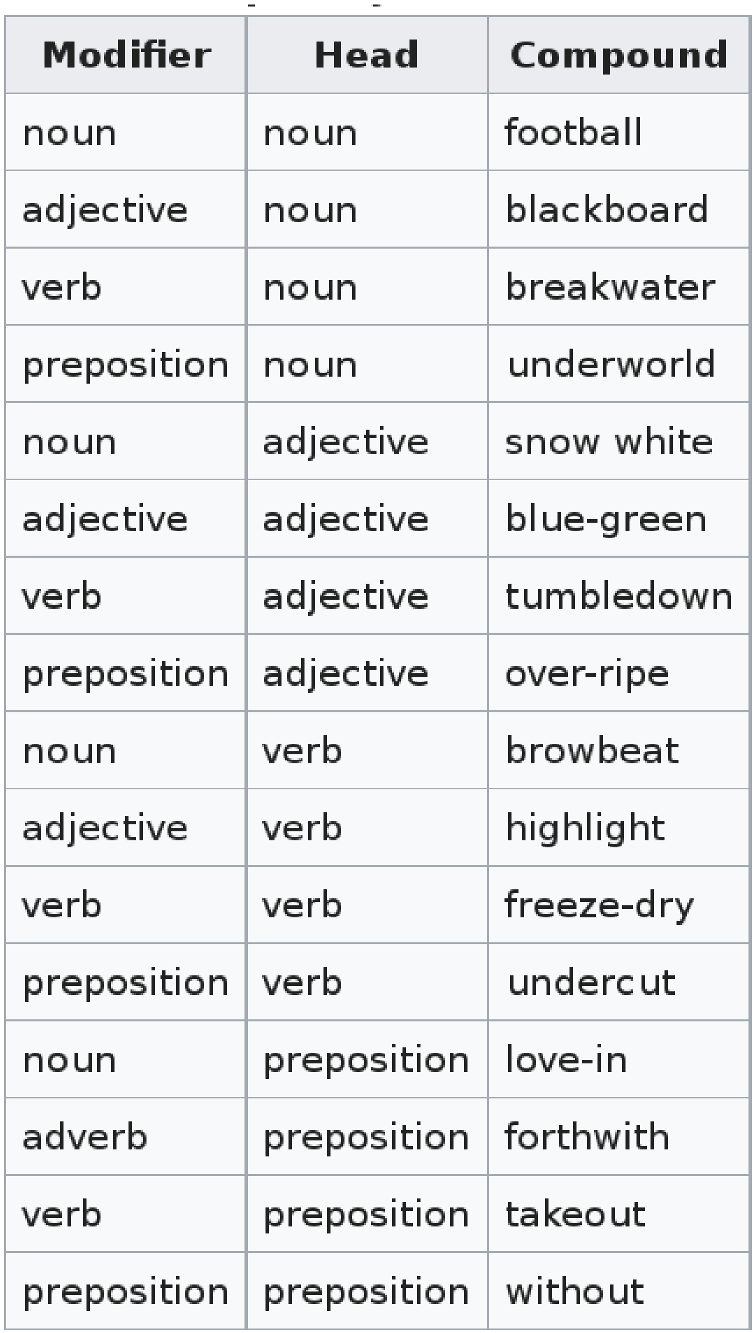
Some examples are "pay attention" ,"fast food", "make an effort", and "powerful engine".

Example: https://corpus.byu.edu/now/

natural English	**unnatural English**
the fast train	the ~~quick~~ train
fast food	~~quick~~ food
a quick shower	a ~~fast~~ shower
a quick meal	a ~~fast~~ meal

Terminology

the study of terms
terminology
[tur-muh-nol-uh-jee]
noun, plural 'terminologies'
1. the system of terms belonging or peculiar to a science, art, or specialized subject; nomenclature.
2. the science of terms, as in particular sciences or arts.
Word Origin and History for 'terminology'
from German Terminologie (1786), a hybr
by C.G. Schütz of Jena, from Medieval
word, expression" (see termin
with, a speaking



Modifier	Head	Compound
noun	noun	football
adjective	noun	blackboard
verb	noun	breakwater
preposition	noun	underworld
noun	adjective	snow white
adjective	adjective	blue-green
verb	adjective	tumbledown
preposition	adjective	over-ripe
noun	verb	browbeat
adjective	verb	highlight
verb	verb	freeze-dry
preposition	verb	undercut
noun	preposition	love-in
adverb	preposition	forthwith
verb	preposition	takeout
preposition	preposition	without

Compound: A compound is a word composed of more than one free morpheme.

Terminology

he study of terms

terminology

[tur-*muh*-nol-*uh*-jee]

noun, plural *'terminologies'*

1. the system of terms belonging or peculiar to a science, art, or specialized subject; nomenclature.
2. the science of terms, as in particular sciences or arts.

Word Origin and History for *'terminology'*

l, from German Terminologie (1786), a hybr
by C.G. Schütz of Jena, from Medieval
ord, expression" (see terminu
with, a speaking

Noun-phrase: A noun phrase or nominal phrase (abbreviated NP) is a phrase which has a noun (or indefinite pronoun) as its head word, or which performs the same grammatical function as such a phrase.

Some examples of noun phrases are underlined in the sentences below. The head noun appears in bold.

The election-year **politics** are annoying for many **people.**

Almost every **sentence** contains at least one noun **phrase**.

Noun phrases can be identified by the possibility of pronoun substitution:

a. This **sentence** contains two noun **phrases**.

b. It contains them.

a. The subject noun **phrase** that is present in this sentence is long.

b. It is long.

a. Noun **phrases** can be embedded in other noun **phrases.**

b. They can be embedded in them.