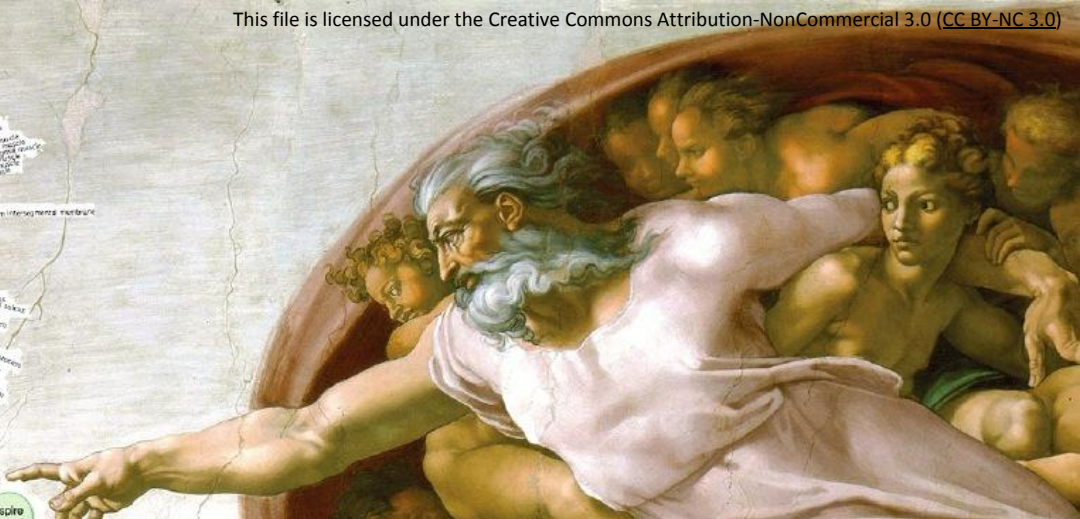




HPI Hasso Plattner Institut



Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
Spring 2018

Information Service Engineering

Lecture 1: Natural Language Processing

1.1 Information and Knowledge

1.2 Language and Understanding

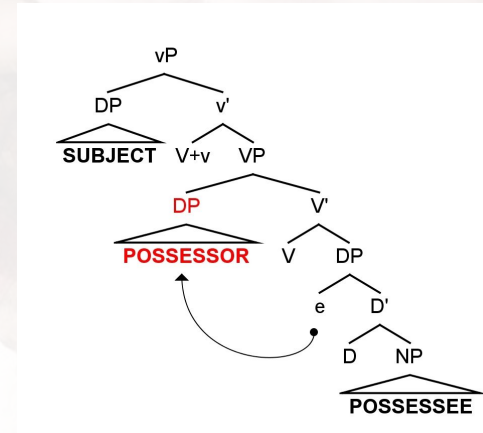
1.3 NLP Techniques and Challenges

1.4 Linguistic Basics

1.5 Morphology

1.6 Evaluation, Precision and Recall

1.7 Regular Expressions



Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- [illegible]

Sentence Spitting

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- Splitting a text into sentences

Original Text

Neil Alden Armstrong was an American astronaut and the first person to walk on the Moon. He was also an aerospace engineer, naval aviator, test pilot, and university professor. Before becoming an astronaut, Armstrong was an officer in the U.S. Navy and served in the Korean War. After the war, he earned his bachelor's degree at Purdue University and served as a test pilot at the National Advisory Committee for Aeronautics (NACA) High-Speed Flight Station, where he logged over 900 flights. He later completed graduate studies at the University of Southern California.

Analysis Result

Neil Alden Armstrong was an American astronaut and the first person to walk on the Moon.

He was also an aerospace engineer, naval aviator, test pilot, and university professor.

Before becoming an astronaut, Armstrong was an officer in the U.S. Navy and served in the Korean War.

After the war, he earned his bachelor's degree at Purdue University and served as a test pilot at the National Advisory Committee for Aeronautics (NACA) High-Speed Flight Station, where he logged over 900 flights.

He later completed graduate studies at the University of Southern California.

Tokenization

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- Splitting a stream of text up into words, phrases, symbols, or other meaningful elements called **tokens**

Word Tokenization with Python NLTK

This is a demonstration of the various **tokenizers** provided by **NLTK 2.0.4**.

Tokenize Text

Enter text

Neil Alden Armstrong was an American astronaut and the first person to walk on the Moon.

Enter up to 50000 characters

Tokenize

TreebankWordTokenizer

1.

Neil Alden Armstrong was an American astronaut and the first person to walk on the Moon .

<http://text-processing.com/demo/tokenize/>

Part-of-Speech Tagging

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- Marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context

Text Analysis Result -- NLTK POS Tagging

Original Text

Neil Alden Armstrong (August 5, 1930 – August 25, 2012) was an American astronaut and the first person to walk on the Moon

<http://textanalysisonline.com/nltk-pos-tagging>

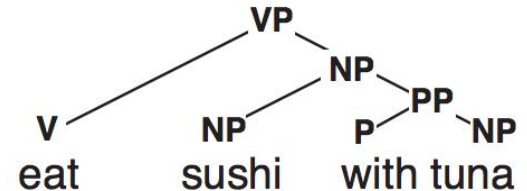
Analysis Result

Neil|NNP Alden|NNP Armstrong|NNP ((August|NNP 5|CD ,|, 1930|CD –|NN August|NNP 25|CD ,|, 2012|CD))|) was|VBD an|DT American|JJ astronaut|NN and|CC the|DT first|JJ person|NN to|TO walk|VB on|IN the|DT Moon|NNP

- Analysing a string of symbols conforming to the rules of a formal grammar and building a **syntactic tree of a sentence**

Parse

```
(ROOT
  (S
    (NP (NNP Neil) (NNP Alden) (NNP Armstrong))
    (VP (VBD was)
      (NP
        (NP (DT an) (JJ American) (NN astronaut))
        (CC and)
        (NP (DT the) (JJ first) (NN person)
          (S
            (VP (TO to)
              (VP (VB walk)
                (PP (IN on)
                  (NP (DT the) (NNP Moon))))))))))
    (. .)))
```



<http://nlp.stanford.edu:8080/parser/index.jsp>

Named Entity Recognition

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- Locating and classifying atomic elements into predefined categories such as **names, persons, organizations, locations**, expressions of **time, quantities, monetary values**, etc.

Stanford Named Entity Tagger

Classifier: english.muc.7class.distsim.crf.ser.gz

Output Format: highlighted

Preserve Spacing: yes

Please enter your text here:

Neil Alden Armstrong (August 5, 1930 – August 25, 2012) was an American astronaut and the first person to walk on the Moon. Before becoming an astronaut, Armstrong was an officer in the U.S. Navy and served in the Korean War.

Senden Clear

Neil Alden Armstrong (August 5, 1930 – August 25, 2012) was an American astronaut and the first person to walk on the Moon. Before becoming an astronaut, Armstrong was an officer in the U.S. Navy and served in the Korean War.

Potential tags:

LOCATION
ORGANIZATION
DATE
MONEY
PERSON
PERCENT
TIME

<http://nlp.stanford.edu:8080/ner/process>

<http://textanalysisonline.com/spacy-named-entity-recognition-ner>

Named Entity Resolution / Word Sense Disambiguation

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- Finding out the exact meaning of an entity, i.e. **linking a text segment to its corresponding entity** in a knowledge base



Neil Armstrong (1930 - 2012)

On August 5, 1930, American astronaut Neil Alden Armstrong was born on the Moon. He was also a professor. Armstrong was

Armstrong's Youth and

Neil Armstrong was born in an auditor for the Ohio State University. Armstrong experienced his first flight in a Trimotor. Armstrong attended Wapakoneta airfield. In 19

Purdue University, funded the Holloway Plan, where he was committed to two years of study, followed by three years of service in the U.S. Navy, and then completion of the final two years of the degree. Armstrong's call-up from the Navy arrived in 1949, requiring him to report to Naval Air Station Pensacola for flight training to qualify as Naval Aviator in 1950. Armstrong served in

Neil Armstrong	
Neil Alden Armstrong (August 5, 1930 – August 25, 2012) was an American astronaut and the first person to walk on the Moon. He	
birth year	1930
death year	2012
death place	Cincinnati
type	NASA
occupation	Naval aviation
death place	Ohio

as born, the first person to walk on the Moon. He was a professor, test pilot, and university professor. Armstrong was the first person to walk on the Moon landing, in July 1969.

Armstrong's Youth and Education: Armstrong was born to Stephen Koenig Armstrong, an auditor for the Ohio State University. Already at the age of five, Armstrong experienced his first flight in a Ford Trimotor. His father took a ride in a Ford Trimotor. Armstrong attended Wapakoneta airfield. In 19

<http://scihi.org/neil-armstrong-the-first-man-of-the-moon/>

Semantic Role Labeling

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- Extracting **subject-predicate-object** triples from a sentence

	Neil	Alden	Armstrong	was	an	American	astronaut	and	the	first	person	to	walk	on	the	Moon	.
SRL	extent [AM-EXT]			V: be.01	topic [A1]												
SRL							walker [A0]						V: walk.01		direction [AM-DIR]		
Nom																	
Prep													Governor	Location (on)		Object	

Key

Verb

V verb

Arguments

A0 subject

A1 object

A2 indirect object

Other

C-arg continuity of an argument/adjunct of type *arg*

R-arg reference to an actual argument/adjunct of type *arg*

Adjunct

AM-ADV adverbial modification

AM-DIR direction

AM-DIS discourse marker

AM-EXT extent

AM-LOC location

AM-MNR manner

AM-MOD general modification

AM-NEG negation

AM-PNC proper noun component

AM-PRD secondary predicate

AM-PRP purpose

AM-REC reciprocal

AM-TMP temporal

http://cogcomp.cs.illinois.edu/page/demo_view/srl

Paraphrasing

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- A **paraphrase** is a restatement of the meaning of a text or passage by using other words
- from Greek παράφρασις, meaning *"additional manner of expression"*
- Examples:
 - Google *bought* YouTube. ⇔ Google *acquired* YouTube.
 - When will my book arrive? ⇔ When will I receive my book?
 - Pat said, *"I like football."* ⇔ Pat said that *he liked* football.
 - Pat likes Chris, because *she* is smart. ⇔ Pat likes Chris, because *Chris* is smart.

Ambiguity

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- One word or a sentence can have **different meanings** (in the language to which it belongs to)
- Examples:
 - *“Jaguar”* - [car / cat / operating system / etc.]
 - *“The door is open!”*
 - *“We saw her duck.”*

Phonological Ambiguity

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- Words which **sound the same** but have **different meanings**

- e.g., *weekend* vs. *weak end*



Communication tip:

Phonological ambiguities or **Give peas a chance!**

One of my favourite ways to have fun with communication are phonological ambiguities.

Phonological ambiguities are two or more words which sound the same and have different meanings.



Language can contain ambiguities - and more than one way to compose a set of sounds into words.

So listen to yourself: It is always good to notice a spoken sentence often contains many words which are (sometimes not) intended to be heard.

English examples:

- there - their
- here - hear
- plane - plain
- Hamburger (Citizens of Hamburg) - hamburger (burger, food)
- sea - see
- Friday - fry day
- weekend - weak end
- ice cream - I scream.
- new direction - nude erection
- new day - nude, eh?
- I don't know! - I don't - no!
- but - butt
- Wait - Weight
- psychotherapist - psycho the rapist
- You're unconscious now... - Your unconscious now...
- Your students... - You're students...
- Two - too - to

http://worldsgreatestsmile.com/html/phonological_ambiguity.html

Syntactic Ambiguity

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- A situation where a sentence may be interpreted in more than one way due to **ambiguous sentence structure**.
- also called *amphiboly* or *amphibology*
- Example:
 - *He saw the man on the rooftop with a binocular.*
 - *One morning I shot an elephant in my pyjamas. How he got into my pyjamas, I'll never know.*



Semantic Ambiguity

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- A situation where a sentence contains **one or more words with different meanings**, and the **significance** of the sentence **changes dramatically** depending on which meaning is intended
- Example:
 - *“The astronomer loves the star.”*



Referential Ambiguity

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- Analysis of coherent sequences of sentences, which may contain **referential ambiguities**.
- Example:
 - “**Alice** understands that you like your **mother**, but **she** ...”

The diagram illustrates the referential ambiguity of the pronoun 'she'. Two red arrows originate from the word 'she' in the sentence. One arrow points back to the word 'Alice', and the other points back to the phrase 'your mother'. This visualizes the ambiguity of whether 'she' refers to Alice or to the mother.
 - Does “she” refer to *Alice* or to *your mother*?

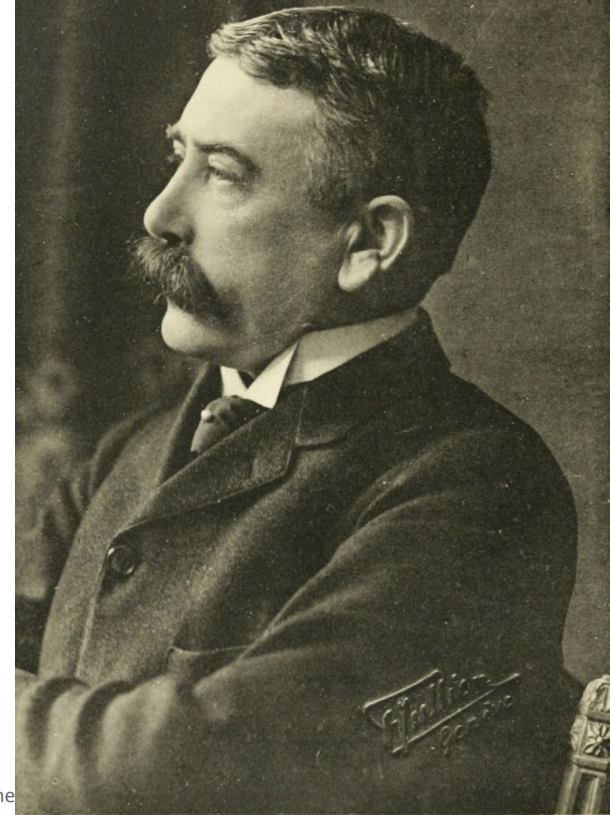
Natural Language is Highly Ambiguous

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

I'm a Linguist.

I love ambiguity more than most people.

Ferdinand de Saussure (1857-1913)



Natural Language is Highly Ambiguous

Information Service Engineering - Lecture 1 Natural Language Processing - 1.3 NLP Techniques and Challenges

- Meaning is **context sensitive**:
 - Depends on the people present e.g. *“How far is it?”* (miles, km?)
 - Depends on the social context: *“That was bad!”*
 - Depends on the location, e.g. *“Play <song> upstairs”*
 - Depends on the time of day, e.g. *“Let's go eat”*
 - Depends on prior sentences: *“The third one”*
- Even more difficult is to detect and correctly interpret **slang, jargon, humor, and sarcasm**
- Also spelling mistakes, grammar mistakes, and (newly created) abbreviations have to be resolved



Linguistic Basics

Next Lecture...

Picture References:

- P.3: https://commons.wikimedia.org/wiki/File:Auckland_Star_Apollo_11_Moon_Landing_Front_Page.jpg
- P.14: Britain's Home Front 1939 - 1945- Roof Spotters, https://upload.wikimedia.org/wikipedia/commons/1/16/Britain%27s_Home_Front_1939_-_1945-_Roof_Spotters_HU86169.jpg
- P.15: European Space Agency (ESA/Hubble), <https://upload.wikimedia.org/wikipedia/commons/6/62/Starsinthesky.jpg>
Publicity photo of Theda Bara for the American film Cleopatra (1917)
https://upload.wikimedia.org/wikipedia/commons/4/42/B_theda-serpent.jpg
- P.17: The Swiss linguist Ferdinand de Saussure (1857–1913), https://commons.wikimedia.org/wiki/File:Ferdinand_de_Saussure_by_Jullien.png
- P.19: Liber ethicorum des Henricus de Alemannia, single sheet. Scena: Henricus de Alemannia con i suoi studenti (14th century) https://commons.wikimedia.org/wiki/File:Laurentius_de_Voltolina_001.jpg