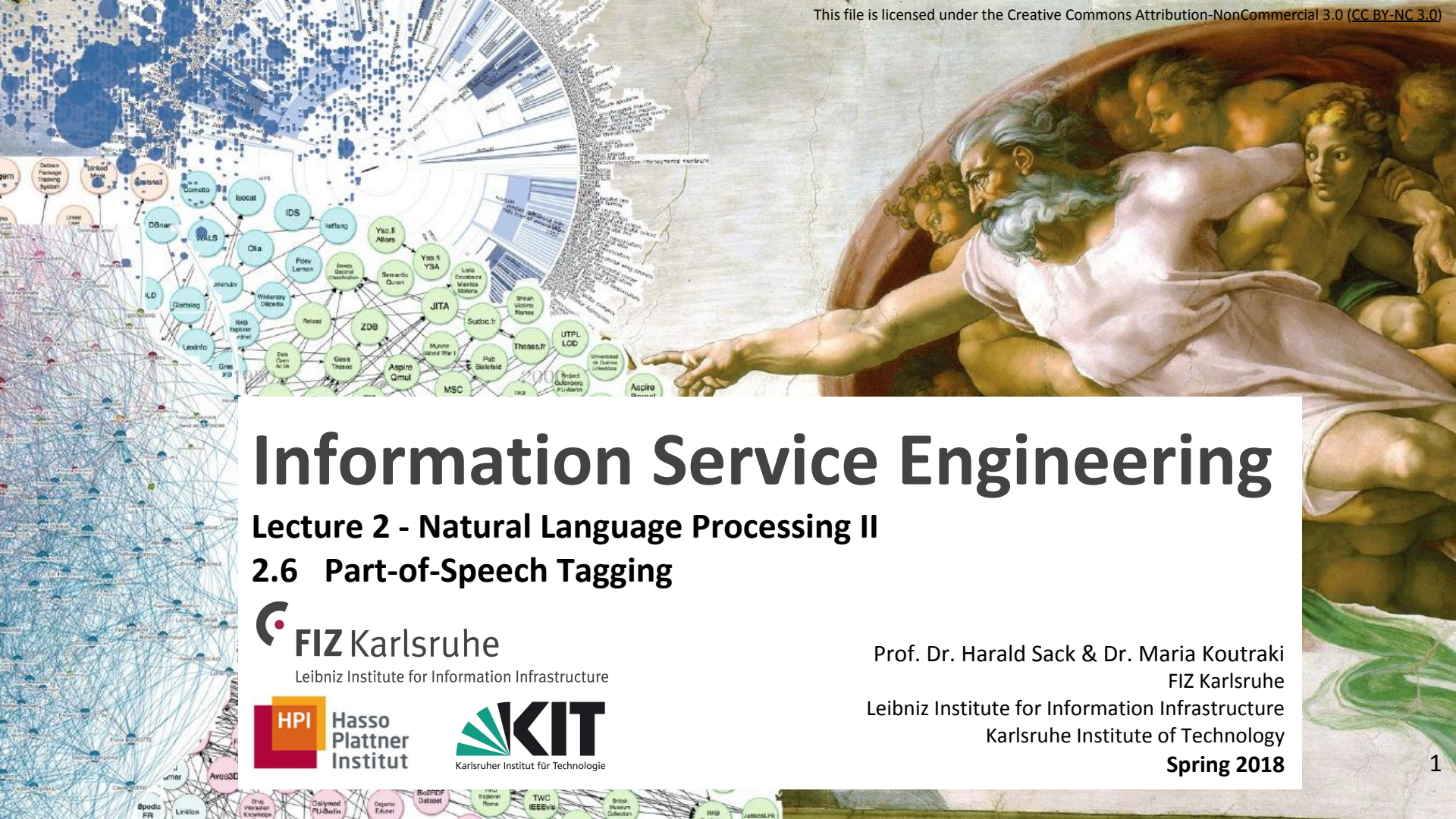




Information Service Engineering

Lecture 2 - Natural Language Processing II

2.6 Part-of-Speech Tagging



FIZ Karlsruhe

Leibniz Institute for Information Infrastructure



Hasso
Plattner
Institut



Karlsruher Institut für Technologie

Prof. Dr. Harald Sack & Dr. Maria Koutraki
FIZ Karlsruhe

Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
Spring 2018

Information Service Engineering

Lecture 2: Natural Language Processing II

2.1 Finite State Automata

2.2 Finite State Transducers

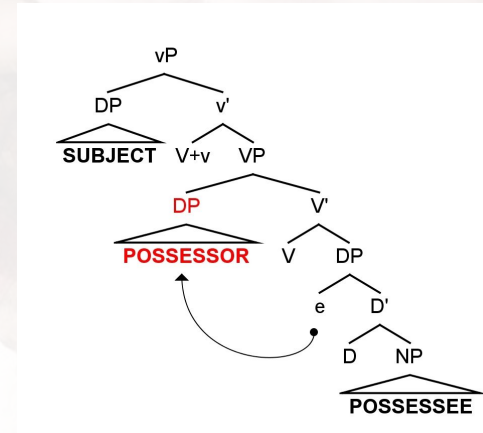
2.3 Language Model and N-Grams

2.4 Language Model and N-Grams II

2.5 Language Model and N-Grams III

2.6 Part-of-Speech Tagging

2.7 Part-of-Speech Tagging II



Part-of-Speech

Information Service Engineering - Lecture 2 Natural Language Processing II - 2.6 Part-of-Speech Tagging

- **Category of words** which have similar grammatical properties.
- Words that are assigned to the **same word part of speech** generally display **similar behavior** in terms of **syntax**.
- Also referred to as
 - lexical categories, word classes, morphological classes, lexical tags
- *Dionysius Trax of Alexandria* [c. 100 BC] describes 8 parts-of-speech:
 - Noun
 - Verb
 - Pronoun
 - Preposition
 - Adverb
 - Conjunction
 - Participle
 - Article

(English) Word Classes

Information Service Engineering - Lecture 2 Natural Language Processing II - 2.6 Part-of-Speech Tagging

- **Nouns**

- a word that functions as the name of some specific thing or set of things, as e.g. living creatures, objects, places, actions, qualities, states of existence, or ideas

- **Proper Nouns**

Names of specific entities, as e.g. *Harald, Karlsruhe, KIT, etc.*

- **Common Nouns**

- **Count Nouns**

Nouns that allow enumeration, as e.g. *one dog, two dogs, etc.*

- **Mass Nouns**

Conceptualization of a homogeneous group, as e.g.
snow, heat, salt, etc.

(English) Word Classes

Information Service Engineering - Lecture 2 Natural Language Processing II - 2.6 Part-of-Speech Tagging

- **Verbs**

- a word that is referring to actions, processes, occurrences, states of being, etc.

- **Main Verbs**

provides the main semantic content of the clause, as e.g.

*The dog **ate** my homework.*

- **Auxiliary Verbs (Auxiliaries)**

adds functional or grammatical meaning to the clause in which it appears, such as to express *tense, aspect, modality, voice, emphasis*, etc; usually accompany a main verb, as e.g.

***Do** you want tea?*

(English) Word Classes

Information Service Engineering - Lecture 2 Natural Language Processing II - 2.6 Part-of-Speech Tagging

- **Adjectives**

- An adjective describes, modifies or gives more information about a noun or pronoun.
- **Examples:**
big, happy, green, young, fun, crazy, three

- **Pronouns**

- A pronoun is used in place of a noun or noun phrase to avoid repetition.
- **Examples:**
I, you, we, they, he, she, it, me, us, them, him, her, this, those

Examples of POS Tags

Information Service Engineering - Lecture 2 Natural Language Processing II - 2.6 Part-of-Speech Tagging

- **Noun:** book/books, nature, Germany, Sony
- **Verb:** eat, wrote
- **Auxiliary:** can, should, have
- **Adjective:** new, newer, newest
- **Adverb:** well, urgently
- **Number:** 872, two, first
- **Article/Determiner:** the, some
- **Conjunction:** and, or
- **Pronoun:** he, my
- **Preposition:** to, in
- **Particle:** off, up
- **Interjection:** Ow, Eh

Open vs. Closed Word Classes

Information Service Engineering - Lecture 2 Natural Language Processing II - 2.6 Part-of-Speech Tagging

- **Closed Word Classes**

- limited number of words, do not grow usually, as e.g.,
Auxiliary, Article, Determiner, Conjunction, Pronoun, Preposition,
Particle, Interjection
- Usually **function words** (i.e. short common words which play a role
in grammar)
- Examples (in English):
 - prepositions: on, under, over, ...
 - particles: up, down, on, off, ...
 - determiners: a, an, the, ...
 - pronouns: she, who, I, ..

Open vs. Closed Word Classes

Information Service Engineering - Lecture 2 Natural Language Processing II - 2.6 Part-of-Speech Tagging

- **Closed Word Classes**

- limited number of words, do not grow usually, as e.g.,
Auxiliary, Article, Determiner, Conjunction, Pronoun, Preposition,
Particle, Interjection

- **Open Word Classes**

- unlimited number of words, as e.g., (for English)
Noun, Verb, Adverb, Adjective

POS Tag Sets

Information Service Engineering - Lecture 2 Natural Language Processing II - 2.6 Part-of-Speech Tagging

- There are various **POS tagsets** based on different granularity of tags
 - **Brown tagset** (Francis and Kucera 1982, 87 tags)
 - Based on Brown corpus, i.e. Brown University Standard Corpus of Present-Day American English
 - **C5 tagset** (61 tags)
 - **C7 tagset** (146 tags!)
 - **Penn TreeBank** (Marcus et al. 1993, 45 tags)
 - Simplified version of Brown tag set;
de facto standard for English now
 - **Prague Dependency Treebank** (Czech, Hajic 2006, 4452 tags)

Penn Treebank Tag Set

Information Service Engineering - Lecture 2 Natural Language

TAG	DESCRIPTION	EXAMPLE
CC	conjunction, coordinating	<i>and, or, but</i>
CD	cardinal number	<i>five, three, 13%</i>
DT	determiner	<i>the, a, these</i>
EX	existential there	<i><u>there</u> were six boys</i>
FW	foreign word	<i>mais</i>
IN	conjunction, subordinating or preposition	<i>of, on, before, unless</i>
JJ	adjective	<i>nice, easy</i>
JJR	adjective, comparative	<i>nicer, easier</i>
JJS	adjective, superlative	<i>nicest, easiest</i>
LS	list item marker	
MD	verb, modal auxillary	<i>may, should</i>
NN	noun, singular or mass	<i>tiger, chair, laughter</i>
NNS	noun, plural	<i>tigers, chairs, insects</i>
NNP	noun, proper singular	<i>Germany, God, Alice</i>
NNPS	noun, proper plural	<i>we met two <u>Christmases</u> ago</i>
PDT	predeterminer	<i><u>both</u> his children</i>
POS	possessive ending	<i>'s</i>
PRP	pronoun, personal	<i>me, you, it</i>
PRP\$	pronoun, possessive	<i>my, your, our</i>
RB	adverb	<i>extremely, loudly, hard</i>
RBR	adverb, comparative	<i>better</i>
RBS	adverb, superlative	<i>best</i>

Penn Treebank Tag Set - II

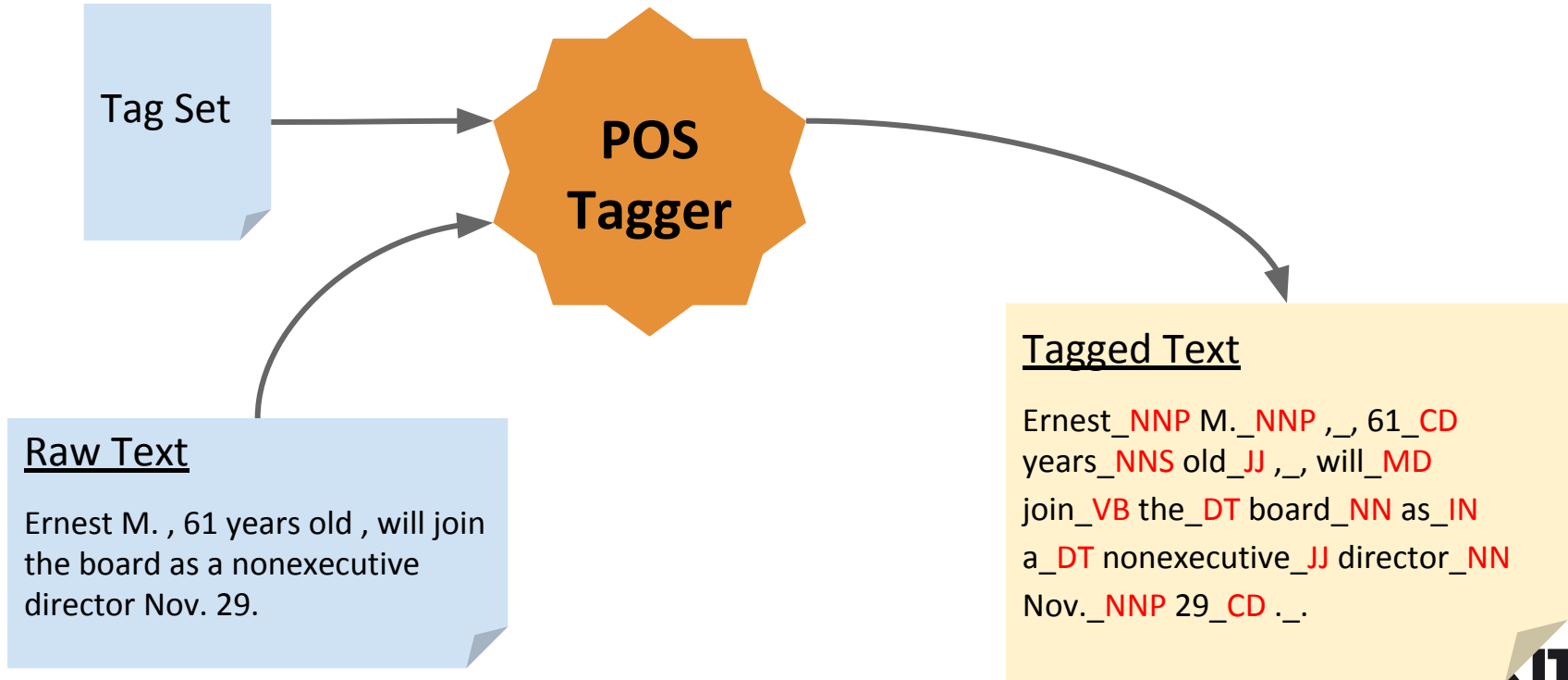
Information Service Engineering - Lecture 2 Natural Language Processing II - 2.6 Part-of-Speech Tagging

RP	adverb, particle	<i>about, off, up</i>
SYM	symbol	%
TO	infinitival to	<i>what <u>to</u> do?</i>
UH	interjection	<i>oh, oops, gosh</i>
VB	verb, base form	<i>think</i>
VBZ	verb, 3rd person singular present	<i>she <u>thinks</u></i>
VBP	verb, non-3rd person singular present	<i>I <u>think</u></i>
VBD	verb, past tense	<i>they <u>thought</u></i>
VBN	verb, past participle	<i>a <u>sunken</u> ship</i>
VBG	verb, gerund or present participle	<i><u>thinking</u> is fun</i>
WDT	<i>wh</i> -determiner	<i>which, whatever, whichever</i>
WP	<i>wh</i> -pronoun, personal	<i>what, who, whom</i>
WP\$	<i>wh</i> -pronoun, possessive	<i>whose, whosever</i>
WRB	<i>wh</i> -adverb	<i>where, when</i>
.	punctuation mark, sentence closer	.,?*
,	punctuation mark, comma	,
:	punctuation mark, colon	:
(	contextual separator, left paren	(
)	contextual separator, right paren	)

<http://www.clips.ua.ac.be/pages/mbasp-tags>

Part-of-Speech Tagging

Information Service Engineering - Lecture 2 Natural Language Processing II - 2.6 Part-of-Speech Tagging



Part-of-Speech Tagging - Ambiguity

Information Service Engineering - Lecture 2 Natural Language Processing II - 2.6 Part-of-Speech Tagging

- The process of assigning a part of speech to each word in a text
- **Challenge:** words often have **more than one POS** (ambiguity)
- Examples:
 - On my **back**_[NN] (noun)
 - The **back**_[JJ] door (adjective)
 - Win the voters **back**_[RB] (adverb)
 - Promised to **back**_[VB] the bill (verb)
- Typical output of a POS-Tagger:
 - The/**DT** grand/**JJ** jury/**NN** commmented/**VBD** on/**IN** a/**DT** number/**NN** of/**IN** other/**JJ** topics/**NNS** ./.

Ambiguity in POS Tags

Information Service Engineering - Lecture 2 Natural Language Processing II - 2.6 Part-of-Speech Tagging

- 45-tags Brown corpus (word types)
 - Unambiguous (1 tag): 38,857
 - Ambiguous: 8,844
 - 2 tags: 6,731
 - 3 tags: 1,621
 - 4 tags: 357
 - 5 tags: 90
 - 6 tags: 32
 - 7 tags: 6 (well, set, round, open, fit, down)
 - 8 tags: 4 ('s, half, back, a)
 - 9 tags: 3 (that, more, in)

Part-of-Speech Tagging - Usecases

Information Service Engineering - Lecture 2 Natural Language Processing II - 2.6 Part-of-Speech Tagging

POS tagging is a prerequisite for further analysis:

- **Speech synthesis:**
 - How to pronounce “*lead*”?
 - **IN**sult or **in**SULT, **OB**ject or **ob**JECT, **OVER**flow or **over**FLOW, **DIS**count or **dis**COUNT, **CON**tent or **con**TENT
- **Parsing:**
 - What words are in the sentence?
- **Information extraction:**
 - Finding names, relations, etc.
- **Machine Translation:**
 - The *noun* “content” may have a different translation from the *adjective*



Part-of-Speech Tagging II

Picture References:

- P.17: Pieter Brueghel the Younger, The Alchemist,
https://commons.wikimedia.org/wiki/File:The_Alchemist_by_Pieter_Brueghel_the_Younger.jpg