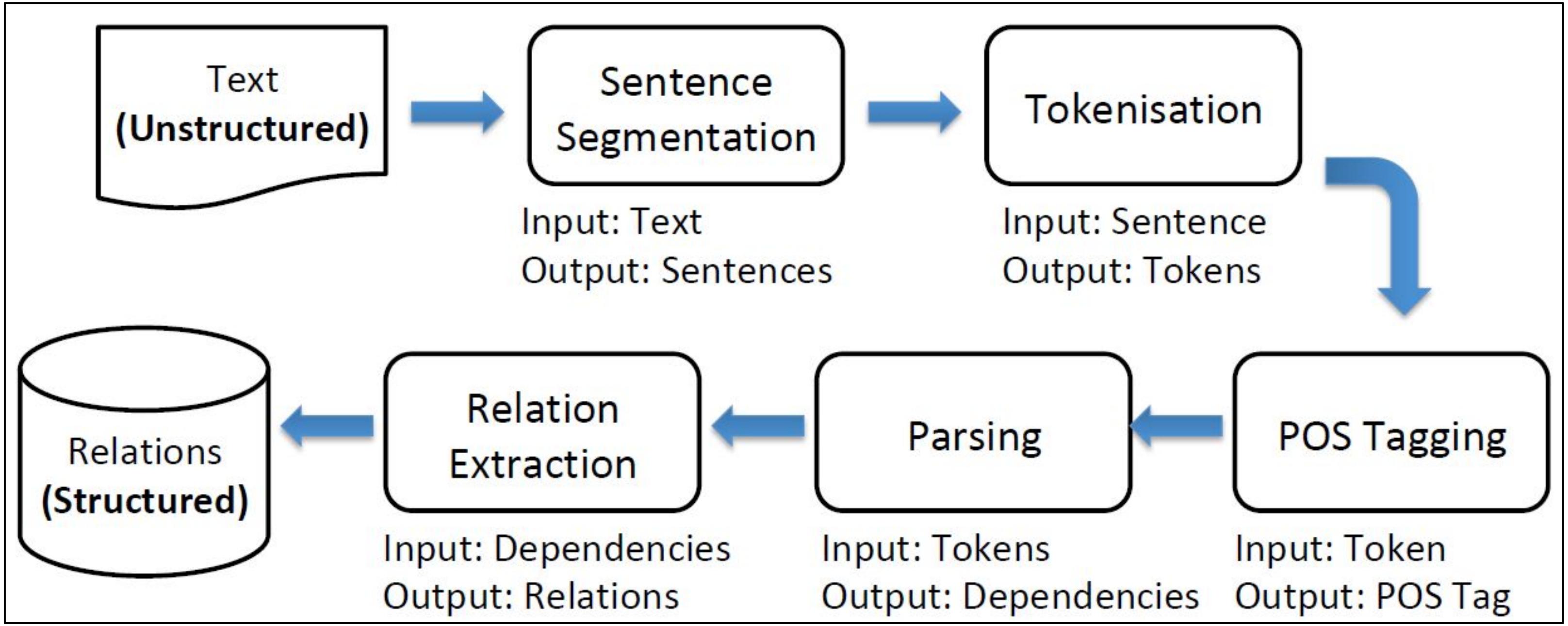
Part-of-Speech (POS) Tagging

COMP61332: Text Mining
Week 2
Riza Batista-Navarro

NLP Pipelines

A complete NLP system is usually a **pipeline of components**

Each component tackles a specific problem



Text **(Unstructured)**

Sentence Segmentation

Input: Text
Output: Sentences

Tokenisation

Input: Sentence
Output: Tokens

POS Tagging

Input: Token
Output: POS Tag

Parsing

Input: Tokens
Output: Dependencies

Relation Extraction

Input: Dependencies
Output: Relations

Relations **(Structured)**

Parts of speech (POS)

classes of words according to their meaning and role in grammar

Nouns: *house*, *health*, *London*, etc...

Pronouns: *he*, *they*,…

Verbs: *walks*, *gave*, *showing*, …

Adjectives: *small*, *better*,…

Adverbs: *almost*, *happily*, …

Determiners: *the*, *a*, *an*

Conjunctions: *and*, *or*, *because*, …

Prepositions: *in*, *of*, *from*, …

Parts of speech (POS)

labels used for annotating words with their parts of speech

come from **tagsets**, for example:

- Penn Treebank
- Universal Scheme

The Penn Treebank Tagset

CC Coordinating conjunction
CD Cardinal number
DT Determiner
EX Existential there
FW Foreign word
IN Preposition or subordinating conjunction
JJ Adjective
JJR Adjective, comparative
JJS Adjective, superlative
LS List item marker
MD Modal
NN Noun, singular or mass
NNS Noun, plural
NP Proper noun, singular
NPS Proper noun, plural
PDT Predeterminer
POS Possessive ending
PP Personal pronoun
PP$ Possessive pronoun
RB Adverb
RBR Adverb, comparative
RBS Adverb, superlative
RP Particle
SYM Symbol
TO to
UH Interjection
VB Verb, base form
VBD Verb, past tense
VBG Verb, gerund or present participle
VBN Verb, past participle
VBP Verb, non-3rd person singular present
VBZ Verb, 3rd person singular present
WDT Wh-determiner
WP Wh-pronoun
WP$ Possessive wh-pronoun
WRB Wh-adverb
Plus additional tags for punctuation

The Universal Scheme

Open class words		**Closed class words**		**Other**	
ADJ	adjective	**ADP**	adposition	**PUNCT**	punctuation
ADV	adverb	**AUX**	auxiliary	**SYM**	symbol
INTJ	interjection	**CCONJ**	coordinating conjunction	**X**	other
NOUN	noun	**DET**	determiner		
PROPN	proper noun	**NUM**	numeral		
VERB	verb	**PART**	particle		
		PRON	pronoun		
		SCONJ	subordinating conjunction		

The Task

Assign POS tags to individual **tokens**

Tokenisation is usually performed before (although some approaches do tokenisation and POS tagging jointly)

*Book/**VB** that/**DT** flight/**NN** ./**.***

*Does/**VBZ** that/**DT** flight/**NN** serve/**VB** dinner/**NN** ?/**.***

Challenges

Syntactic ambiguity (e.g., accidental homophones and homonyms)

duck (action, verb)

duck (bird, noun)

Source: *https://americanenglish.state.gov/*

Different syntactic roles

*To **walk** vs to go for a **walk***

Old** people vs the **old

*They **referee** the matches vs The **referee** starts the match*

How can we disambiguate?

Observe that syntactic ambiguity

... occurs when token is in **isolation**

... disappears when in combination with other words

*I want to **go*** vs *I want a **go***

*I can **walk** there* vs *I will take a **walk***

*The garbage **can** smell* vs *The garbage **can** smells*

... but sometimes it does not

*They **can** fish*

How can we disambiguate?

- A token is very unlikely to be a verb if its preceding word is a determiner

  I want a go

- A token is unlikely to be a noun if the immediately preceding word is *to*

  I want to go

- A token is more likely to be a possessive pronoun when followed by a common noun

  He stroked her cat

- ...but not always

  He gave her money

The Task

Assign POS tags to individual **tokens**

... but **only one** POS tag per token (for each run)

They can fish

Run 1: *They/**PRON** can/**AUX** fish/**VERB***

Run 2: *They/**PRON** can/**VERB** fish/**NOUN***

I saw her bat

Run 1: *I/**PRON** saw/**VERB** her/**PRON** bat/**NOUN***

Run 2: *I/**PRON** saw/**VERB** her/**PRON** bat/**VERB***

Approaches

Develop a **rule-based** mechanism, embodying knowledge of syntax

Or, adopt a **statistical** methodology, based on corpus (text collection) evidence

Part-of-Speech Corpus

Text collection where each token is labelled with the correct (gold standard) POS tag

Manually labelled by an expert linguist

Consists of a text file, 2 common formats:

(1) with each line corresponding to a token-POS tag pair, or

(2) one sentence per line, with each the POS tag appended to each token

Part-of-Speech Corpus: Penn Treebank example

Today	NN
is	VBZ
a	DT
nice	JJ
day	NN
.	.
I	PRP
want	VBP
to	TO
go	VB
for	IN
a	DT
walk	NN
.	.

*Today/*NN *is/*VBZ *a/*DT *nice/*JJ *day/*NN *./*.

*I/*PRP *want/*VBP *to/*TO *go/*VB *for/*IN *a/*DT *walk/*NN *./*.

Part-of-Speech Corpus: Universal Scheme example

Today	NOUN
is	VERB
a	DET
nice	ADJ
day	NOUN
.	PUNCT
I	PRON
want	VERB
to	ADP
go	VERB
for	ADP
a	DET
walk	NOUN
.	PUNCT

*Today/*NOUN *is/*VERB *a/*DET *nice/*ADJ *day/*NOUN *./*PUNCT

*I/*PRON *want/*VERB *to/*ADP *go/*VERB *for/*ADP *a/*DET *walk/*NOUN *./*PUNCT

Part-of-Speech Corpora Examples

Penn Treebank (Wall Street Journal news articles)

GENIA Corpus (biomedical scientific articles)