Text Data Understanding

Tecnologías de Gestión de
Información No Estructurada
Prof. Dr. David E. Losada
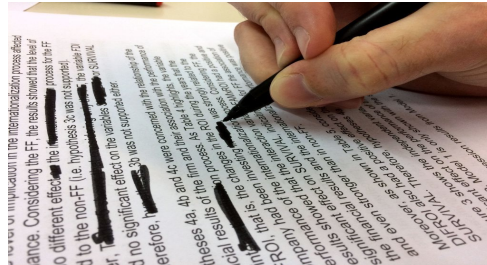
Máster Interuniversitario en Tecnologías de Análisis de Datos Masivos: Big Data

Natural Language Processing (NLP)

developing computational techniques to enable a computer to **understand the meaning** of text

while a human can instantly understand a sentence in their native language, it is quite challenging for a **computer** to make sense of one

Natural Language Processing (NLP)

lexical analysis

what the basic meaningful units

in a language are (e.g., words)

determine the meaning of each word

boundaries of words

syntactic analysis

how words are related with each other

in a sentence

Natural Language Processing (NLP)

semantic analysis

meaning of a sentence

pragmatic analysis

meaning in context, e.g., to infer the speech acts of language

the purpose in communication

discourse analysis

analysis of a large chunk of text

multiple sentences; **connections** between sentences and the analysis of an individual

sentence must be placed in the appropriate **context** involving other sentences

very **challenging** to do this kind of analysis for unrestricted natural language text
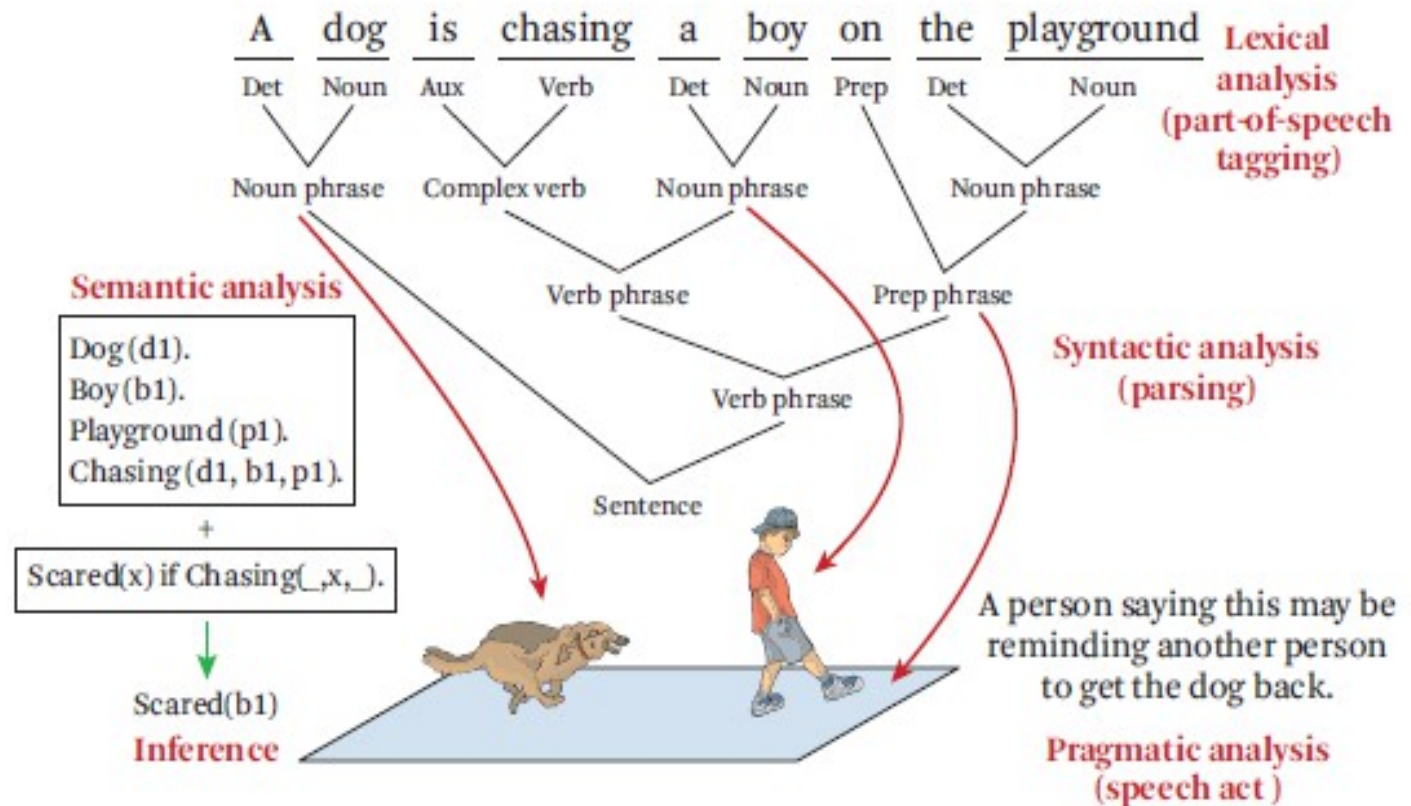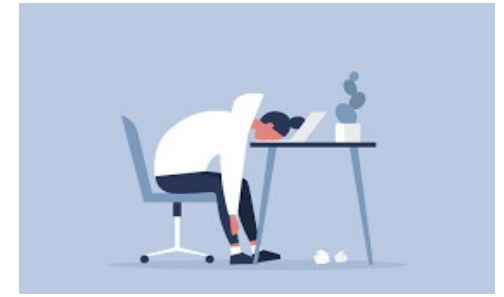
Natural Language Processing (NLP)

A dog is chasing a boy on the playground

A	dog	is	chasing	a	boy	on	the	playground
Det	Noun	Aux	Verb	Det	Noun	Prep	Det	Noun

Lexical analysis (part-of-speech tagging)

Noun phrase · Complex verb · Noun phrase · Noun phrase

Semantic analysis

Dog(d1).
Boy(b1).
Playground(p1).
Chasing(d1, b1, p1).

+

Scared(x) if Chasing(_,x,_).

Scared(b1)
Inference

Verb phrase · Prep phrase

Syntactic analysis (parsing)

Verb phrase

Sentence

A person saying this may be reminding another person to get the dog back.

Pragmatic analysis (speech act)

Figure 3.1   An example of tasks in natural language understanding.

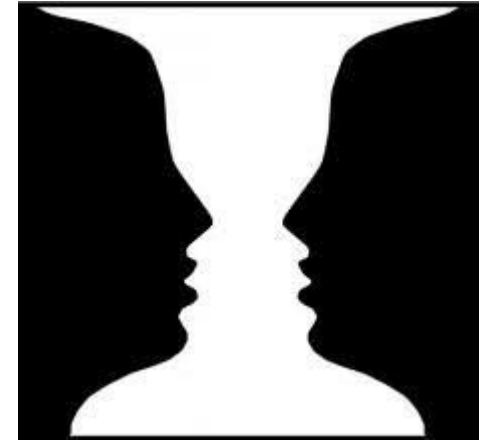**Tecnologías de Gestión de Información No Estructurada**

NLP is VERY difficult

we omit a lot of **"common sense" knowledge** in natural language communication because we assume the hearer or reader possesses such knowledge

we keep a lot of **ambiguities** (we assume the hearer/reader knows how to resolve)

ambiguities

word-level ambiguity

a word may have **multiple syntactic categories** and **multiple senses**. For example, "design" can be a noun or a verb (<u>ambiguous POS</u>); "root" has multiple meanings even as a noun (<u>ambiguous sense</u>)

syntactic ambiguity

a phrase or a sentence may have **multiple syntactic structures**. For example, natural language processing can have two different interpretations: "processing of natural language" vs. "natural processing of language" (<u>ambiguous modification</u>)

"a man saw a boy with a telescope" has two distinct syntactic structures, leading to a different results (ambiguous prepositional phrase (PP) attachment)

ambiguities
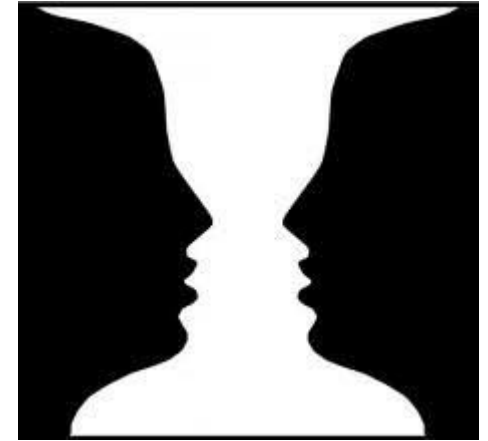
Anaphora resolution

what exactly a **pronoun refers to** may be **unclear**

"John persuaded Bill to buy a TV for himself" , does himself refer to John or Bill?

Presupposition.

"He has quit smoking" implies that he smoked before;

making such inferences in a general way is difficult

history of NLP

1950s. machine translation

1960s and 1970s: speech recognition (requires only limited understanding

of natural language)

1970s–1980s: story understanding. knowledge representation &

and heuristic inference rules

even simple stories are quite challenging to understand

After the 1980s, researchers started moving away from the traditional symbolic

(logic-based) approaches (not robust for real applications)

now...

more attention to **statistical approaches** (more robust, less reliance on human-generated rules)

take advantage of **regularities and patterns** in empirical uses of language

rely on labeled **training data** by humans and application of **machine learning** techniques

today, the most advanced NLP techniques tend to rely on heavy use of statistical machine learning techniques with **linguistic knowledge** only playing a somewhat **secondary** role
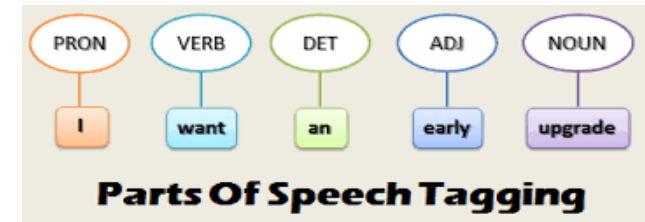
statistical NLP in...

Parts Of Speech Tagging

part of speech (POS) tagging

relatively easy task

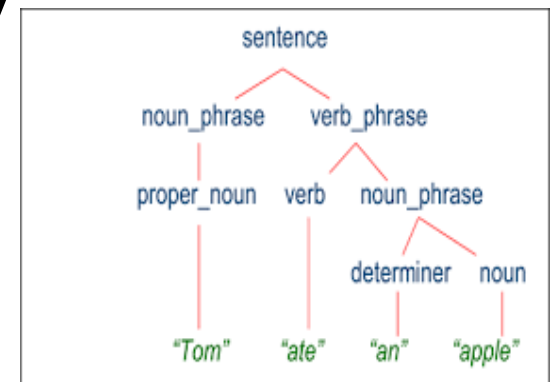state-of-the-art POS taggers highly accurate (above 97% on news data)

parsing

more difficult but can be done with high accuracy

(e.g., above 90% for recognizing noun phrases)

full structure parsing

very difficult, mainly because of ambiguities

statistical NLP in...

semantic analysis

even more difficult, limited success

notably <u>information extraction</u> (recognizing **named entities** such as names of people and organization, and relations between entities such as who works in which organization), <u>word sense disambiguation</u> (distinguishing different senses of a word in different contexts of usage), and <u>sentiment analysis</u> (recognizing positive opinions about a product in a product review)

inferences and speech

act analysis

generally only feasible in very **limited domains**

shallow vs deep NLP

only **"shallow"** analysis of NLP can be done for **arbitrary text** and in a **robust** manner

"deep" analysis

   does not scale up well

   is not robust enough for analyzing unrestricted tex

   a **significant amount of training data** (created by human labeling) must be available in order to achieve reasonable accuracy
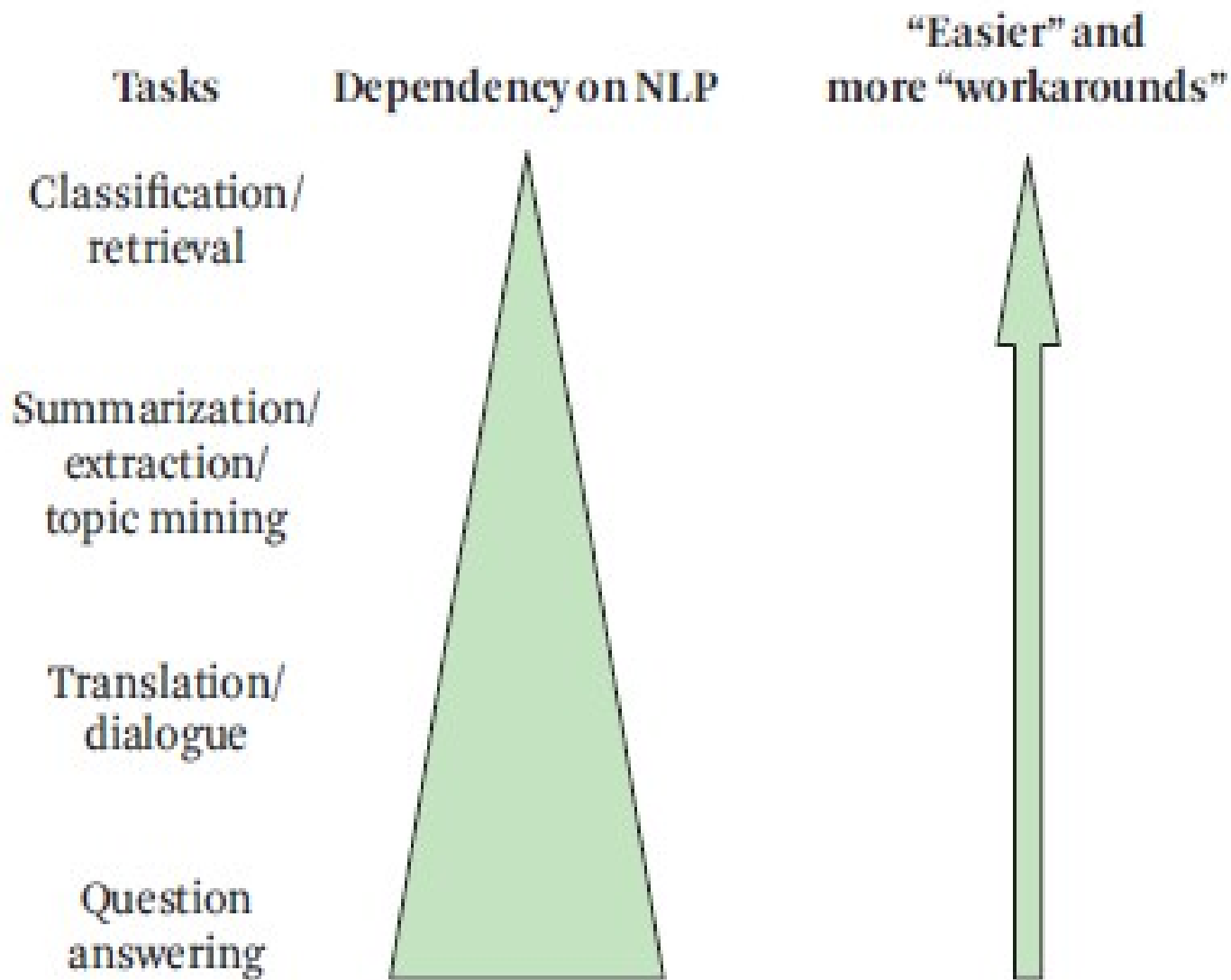
| Tasks | Dependency on NLP | "Easier" and more "workarounds" |

Classification/ retrieval

Summarization/ extraction/ topic mining

Translation/ dialogue

Question answering

Figure 3.2   "Easy" vs. "difficult" NLP applications.

Tecnologías de Gestión de Información No Estructurada
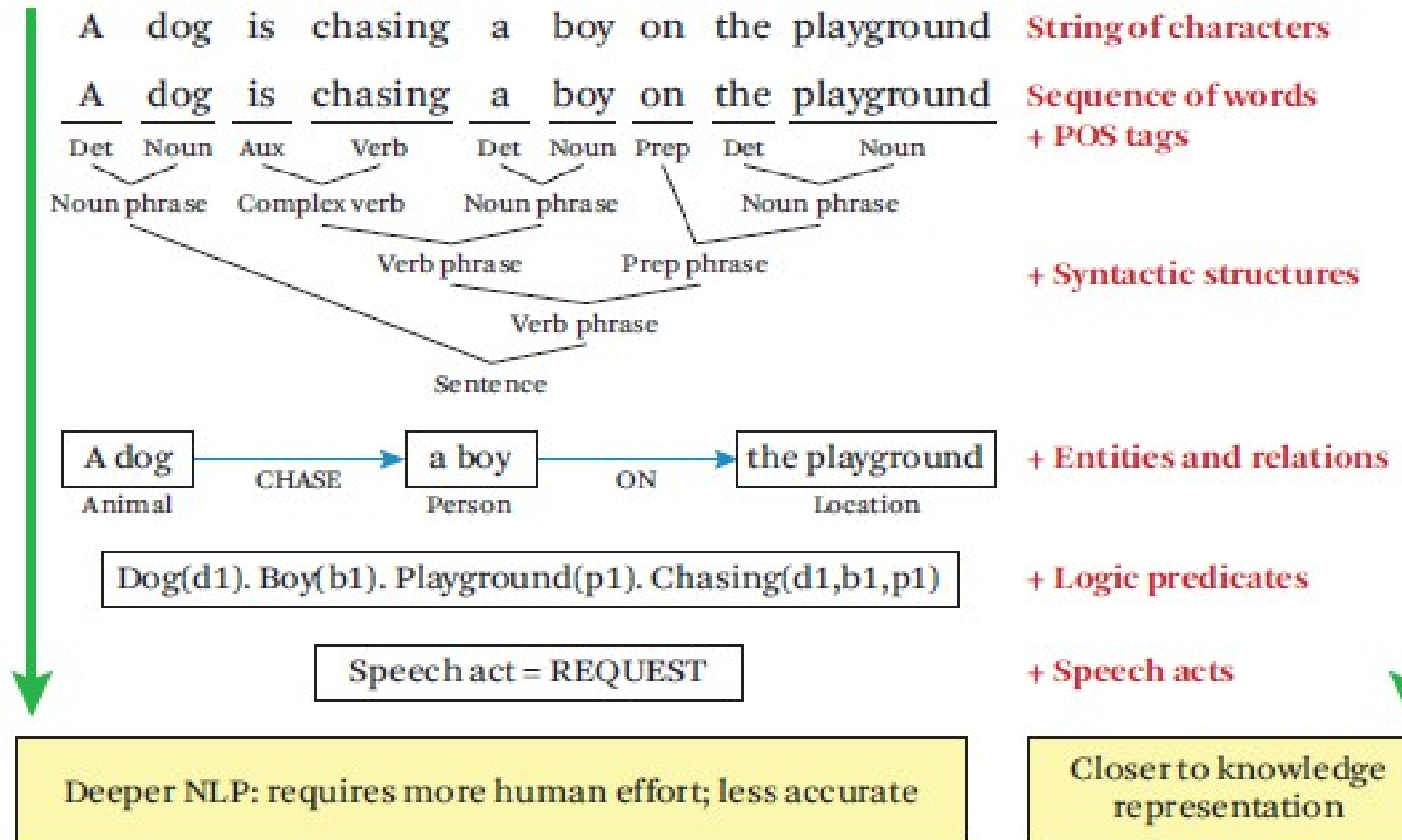
text representation

Figure 3.3 Illustration of different levels of text representation.

text representation

String => Words

By identifying words, we can (for example), easily discover the most **frequent** words

these words can then be used to form **topics**

text data as a sequence of words opens up a lot of **interesting analysis**

Words => POS

to count, for example, the most

frequent **nouns**

what kind of nouns are associated with

what kind of **verbs**

Tag	Description
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
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

Tag	Description
PRP$	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, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP$	Possessive whpronoun
WRB	Whadverb

text representation

POS => Syntactic

analysis of, for example, the writing styles

or grammatical error correction

Syntactic => Semantic

e.g. recognize dog as an animal
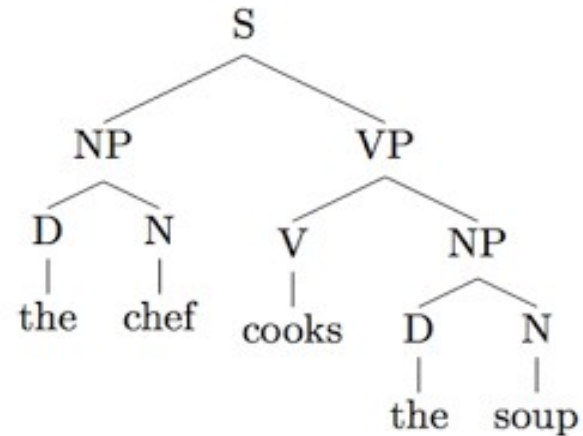
entity-relation recognition

Logic representation

predicates and inference rules

Speech acts

what's the intention of saying that?

what scenarios or what kinds of actions will occur?

text representation. tradeoff between...

deeper analysis that might have errors but would give us direct knowledge that can be extracted from text

shallow analysis that is more robust but wouldn't give us the necessary deeper representation of knowledge

text representation

Text Rep	Generality	Enabled Analysis	Examples of Application
String	▰▰▰	String processing	Compression
Words	▰▰▰	Word relation analysis; topic analysis; sentiment analysis	Thesaurus discovery; topic- and opinion-related applications
+ Syntactic structures	▰▰	Syntactic graph analysis	Stylistic analysis; structure-based feature extraction
+ Entities & relations	▰▰	Knowledge graph analysis; information network analysis	Discovery of knowledge and opinions about specific entities
+ Logic predicates	▪	Integrative analysis of scattered knowledge; logic inference	Knowledge assistant for biologists

Figure 3.4   Text representation and enabled analysis.

Tecnologías de Gestión de Información No Estructurada

statistical language models

A statistical language model is a **probability distribution over word sequences**

$$p(\text{Today is Wednesday}) = 0.001$$
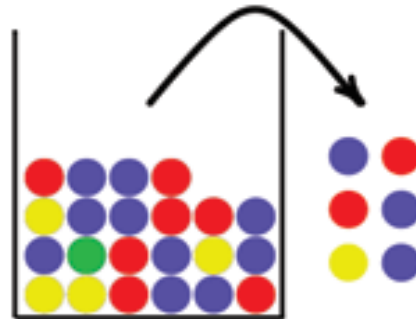
$$p(\text{Today Wednesday is}) = 0.000000001$$

$$p(\text{The equation has a solution}) = 0.000001$$

can be **context-dependent**

statistical language models

Given a language model, we can **sample** word sequences according to the distribution to obtain a text sample. In this sense, we may use such a model to "generate" text

Thus, a language model is also often called a **generative model for text**
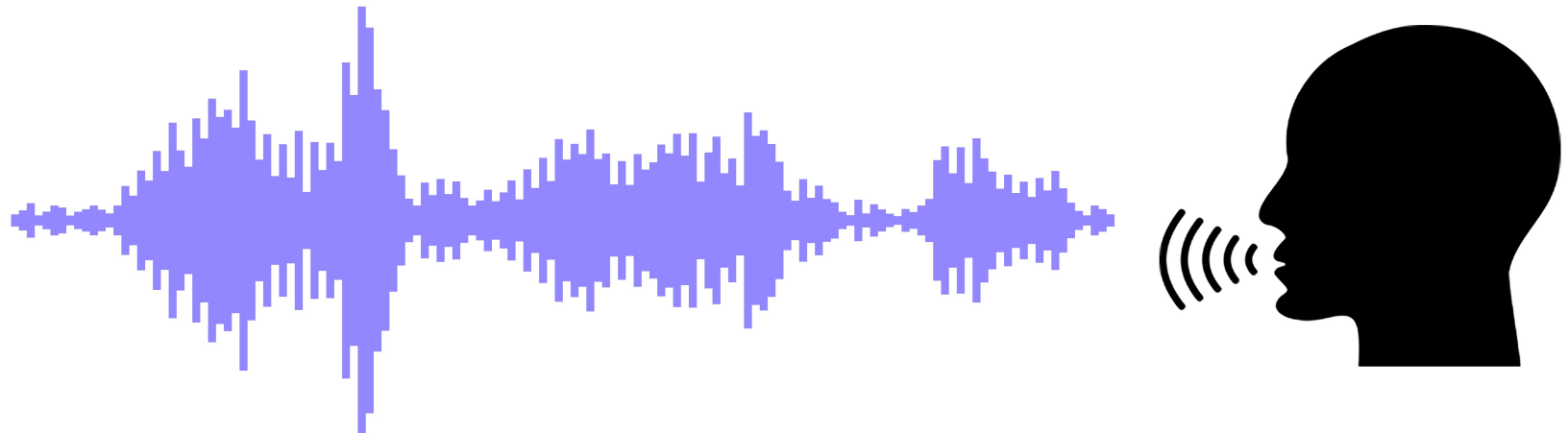
uses of LMs

speech recognition

given that we see "John" and "feels", how likely will we see "happy" as opposed to "habit" as the next word?

happy and habit have very similar acoustic signals, but a LM can easily suggest that "John feels happy" is far more likely than "John feels habit"

uses of LMs

text categorization

given that we observe "baseball" three times and "game" once in a news article, how likely is it about the **topic "sports"**?
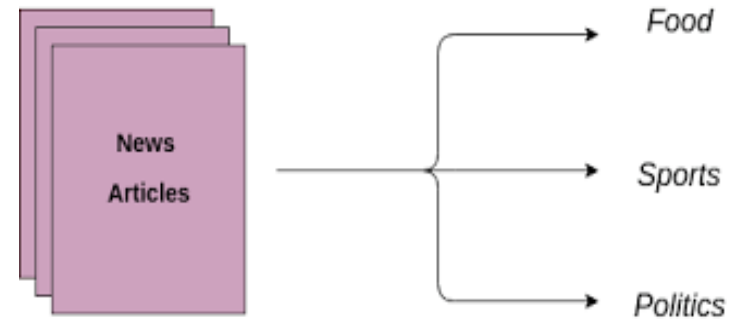
information retrieval

given that a user is interested in

sports news, how likely would it

be for the user to use baseball in a **query**?

building LMs: the estimation problem

enumerating **all the possible sequences** of words and giving a probability to each sequence, would be too **complex**

the **number of parameters** is potentially **infinite**!!

He accused me of taking the money.
Are you accustomed to driving here?
I'm addicted to coca-cola.
I'm afraid of spiders.
Sorry, I don't agree with you.
We agree on most subjects but not politics.
Are you allergic to anything?
I'm very angry at/with you.
I feel anxious about the interview.
I applied to the company for a job.
I don't approve of smoking near children.
I arrived at the airport an hour early.
When did you arrive in Ireland?
I'm ashamed of what I did last night.
The file is attached to this email.
I wasn't aware of the problem.
I'm really bad at singing.
This film is based on a book.
That doesn't belong to you. It's mine.
We belong with each other. It's true love!
She blamed me for what happened.
I'm bored with my job. I need a change.
I've been very busy with work.
Are you confident of passing the exam?
I'm trying to concentrate on my work.
I congratulated her on passing the exam.
The exam consists of speaking and writing.
You can count on me if you need help.
I'd love to go to the party so count me in.
I can't go to the party so count me out.
Some people are very cruel to animals.
Hurry! We're in danger of missing the bus.
I may go. It depends on the weather.
Ireland is very different to/from Italy.
I dreamt about you last night.
I'm dreaming of lying on a beach.
I was disappointed with the film.

I'm excited about moving to Ireland.
Are you familiar with Korean food?
That name is familiar to me.
My region is famous for its wine.
I'm fed up with this awful weather.
I'm very fond of my nephews and nieces.
He's really good at languages.
I'm grateful to you for your help.
Pollution is very harmful to the environment.
I heard about what happened in the news.
It's so nice to hear from you again.
Have you heard of a city called Galway?
Are you hooked on any TV series?
There has been an increase in unemployment.
She insisted on paying for the meal.
Are you interested in meeting me?
I'm involved in a few organisations.
You shouldn't be jealous of others.
You haven't been very kind to me lately.
I'm keen on reading and travelling.
Please, stop laughing at me.
What are you looking at?
He's been married to her for years.
I'm very pleased with my level of English.
I always try to be polite to people.
Greece is popular among/with tourists.
I'm extremely proud of you.
Who's responsible for what happened?
I'm sick of asking you to clean your room.
Do you spend money on expensive clothes?
Will you succeed in passing your driving test?
Horror films are not suitable for children.
I'll take care of you, don't worry.
I've been thinking about you recently.
I'll have to think of an excuse for being late.
I'm used to waking up early.
Are you worried about something?

Tecnologías de Gestión de Información No Estructurada

unigram language model

assumes that a word sequence results from generating **each word independently**

$$p(w_1, \ldots, w_n) = \prod_{i=1}^{n} p(w_i).$$

parameters = # words in the vocabulary
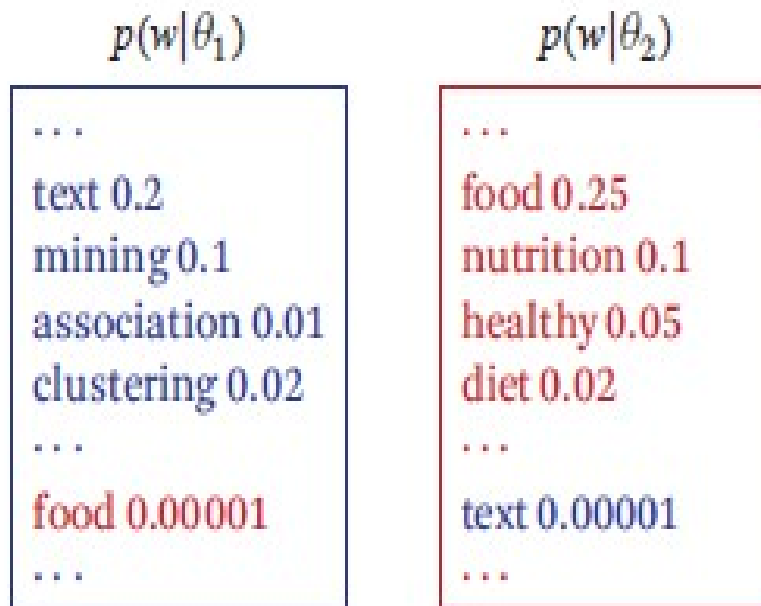
unigram language model

$p(w|\theta_1)$

...
text 0.2
mining 0.1
association 0.01
clustering 0.02
...
food 0.00001
...

$p(w|\theta_2)$

...
food 0.25
nutrition 0.1
healthy 0.05
diet 0.02
...
text 0.00001
...

Figure 3.5   Two examples of unigram language models, representing two different topics.

unigram language model

given a language model θ the **probabilities of generating two different documents** D1 and D2 would be different, i.e., $p(D1 \mid θ) \mathrel{!=} p(D2 \mid θ)$

what kind of docs would have higher probabilities?

those docs that contain many occurrences of the high probability words according to $p(w \mid θ)$

estimating LMs

given a document D (e.g., a short abstract), assumed to be generated using a LM θ, **infer the underlying model θ** (i.e., estimate the probabilities of each word w, p(w | θ))

standard problem in statistics:

parameter estimation
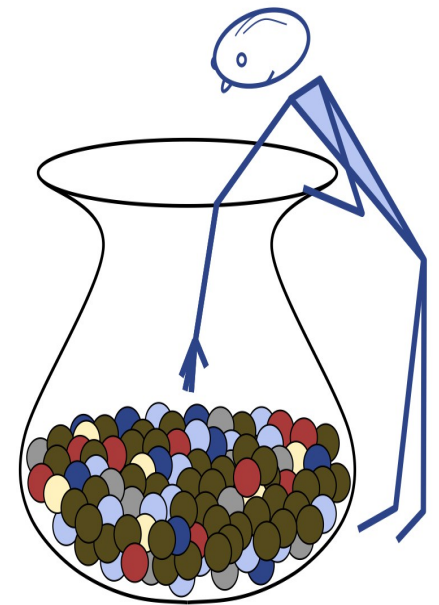

$J(W)$

Initial weights

Global minimum

W
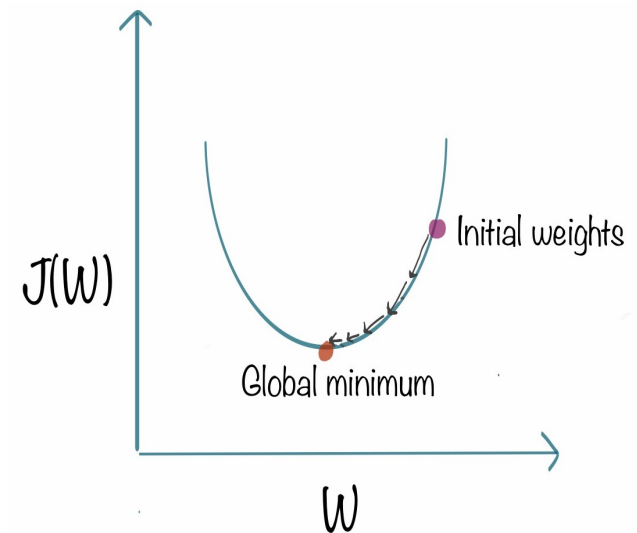
maximum likelihood estimator (mle)

seeks a model that **gives the observed data the highest likelihood**
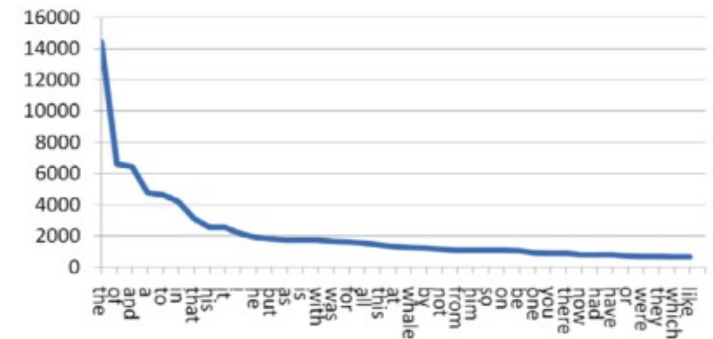
(i.e., best explain the data)

$$\hat{\theta} = \arg\max_{\theta} p(D \mid \theta).$$

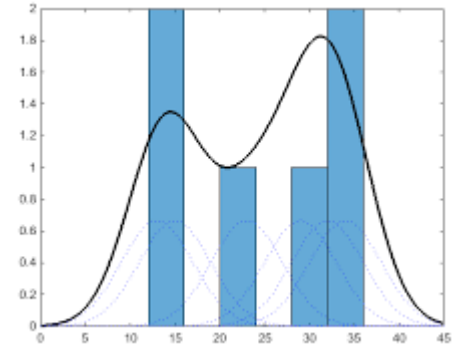# mle for text representation

gives each **word** a probability equal to its **relative frequency** in D

$$p(w \mid \hat{\theta}) = \frac{c(w, D)}{|D|},$$

assigns **zero probabilty** to any **unseen** words
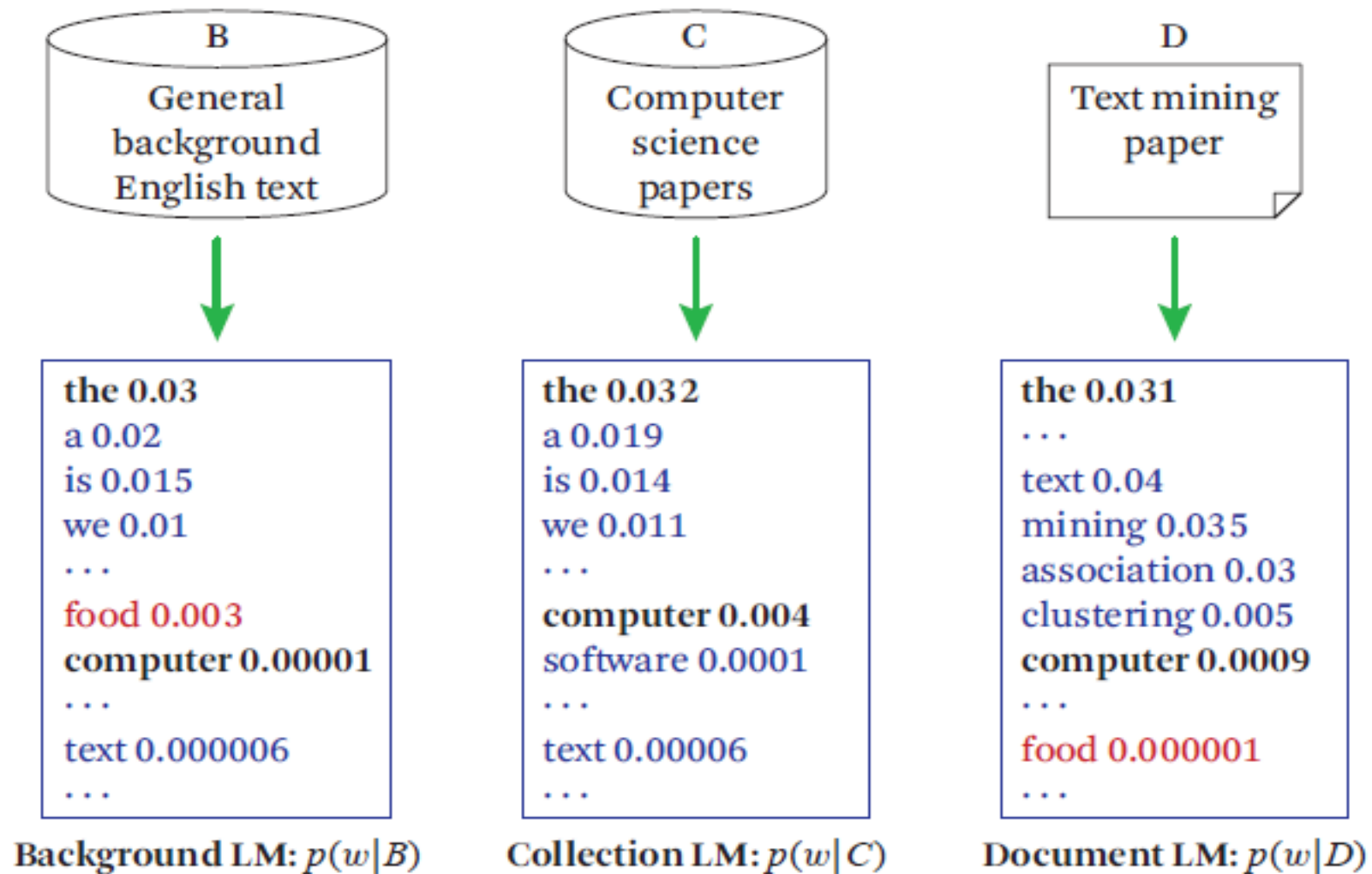
smoothing methods

main idea: improve mle by assigning some probability mass to unseen events

e.g., by **mixing** with a background model

some methods are length-dependent

LM of a doc (D), a collection (C) and background LM (B)

semantic analysis of word relations

co-occurrence:

obtain a sample of docs where "computer" is mentioned

estimate a LM based on this sample

the resulting model, p(w | computer)

this tells us which words

occur frequently

in the context

of "computer"

semantic analysis of word relations

but

the co-occurring words would likely be **functional** words or words that are **simply common** in the data
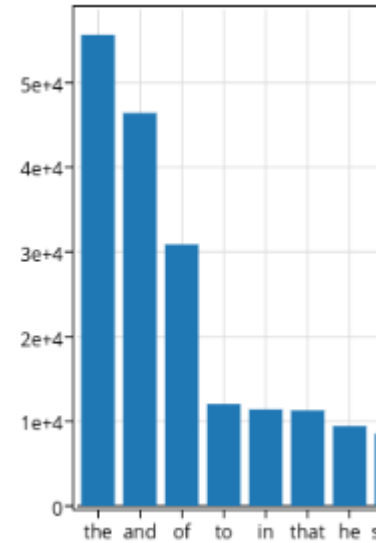
To filter out such common words:

general English language model (i.e., a background LM)

use it to normalize the model p(w | computer)

probability ratio for each word

semantic analysis of word relations