Text Data Understanding

Tecnologías de Gestión de Información No Estructurada

Prof. Dr. David E. Losada

CiTIUS Centro Singular de Investigación en Tecnoloxías Intelixentes

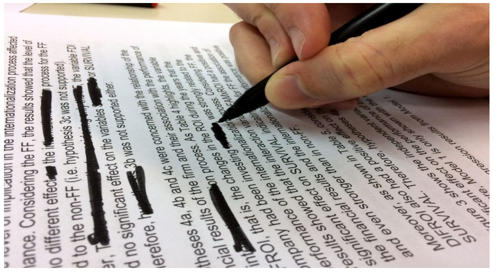
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CAMPUS VIDA

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)

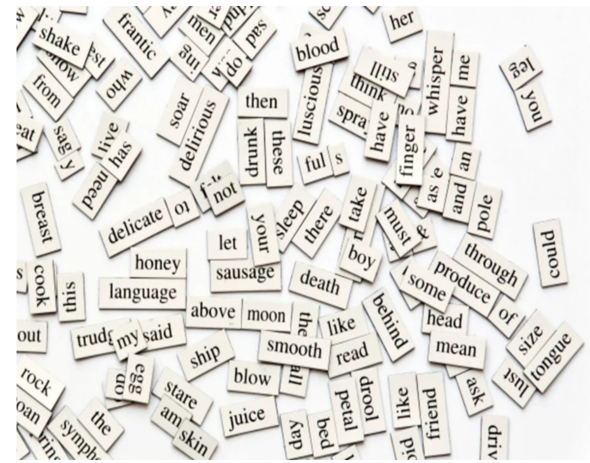
lexical analysis

what the basic meaningful units

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

determine the meaning of each word

boundaries of words

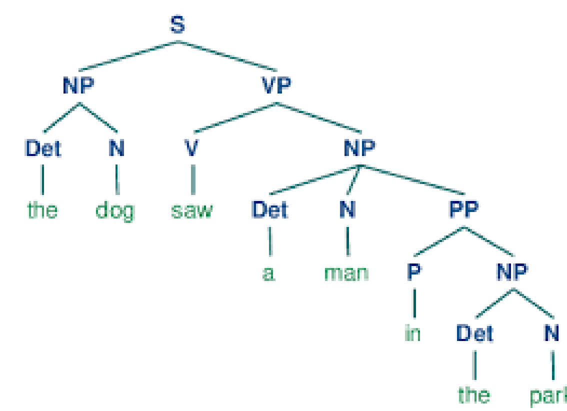


sacamos mas conocimiento que con un analisis lexico, pero no
tiene porque tener idea del significacdo, no hay conocimeinto

syntactic analysis

how words are related with each other

in a sentence



Natural Language Processing (NLP)

aqui si se empieza a sacar conocimiento

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

esto es algo que la tecnologia “pre deep learning” hacia muy mal

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)

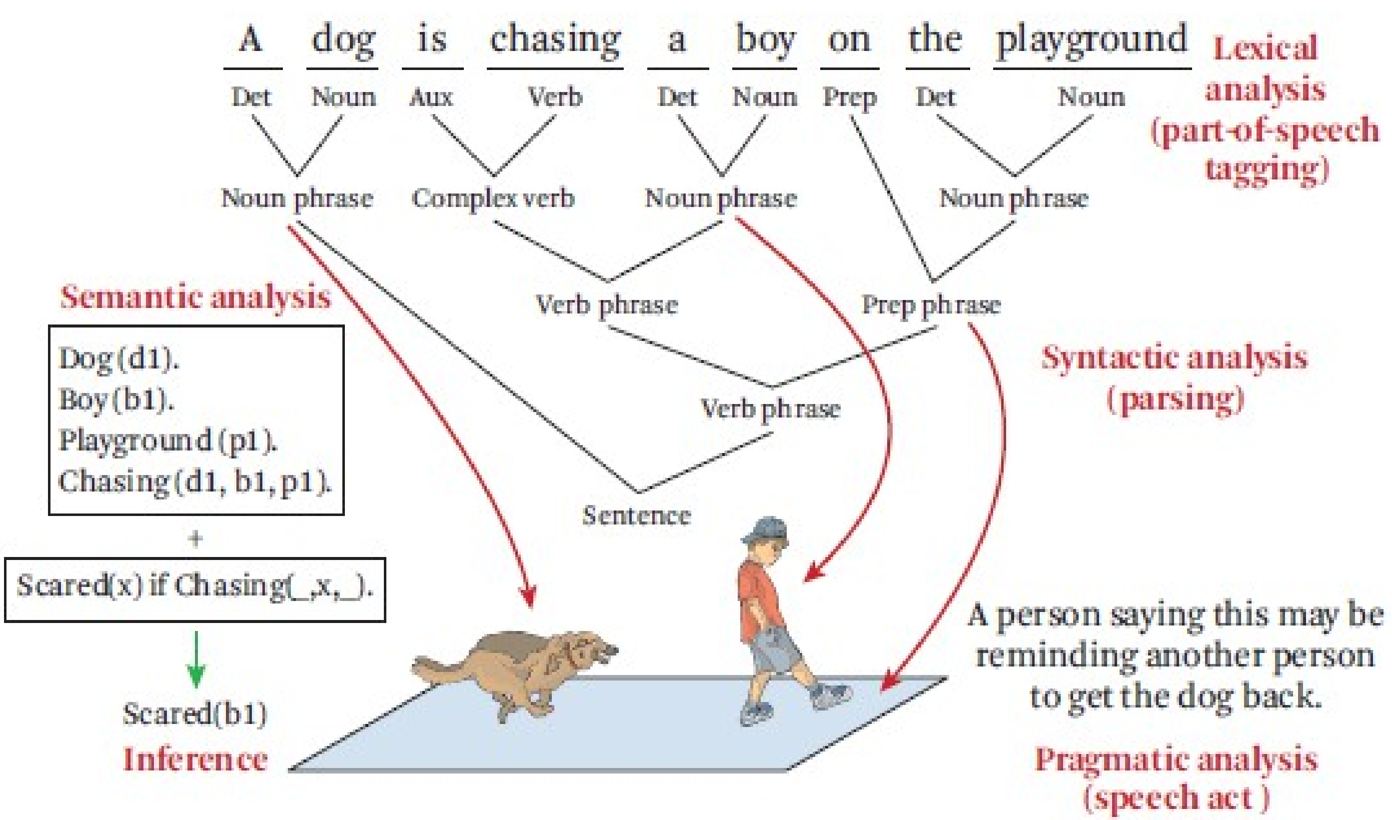


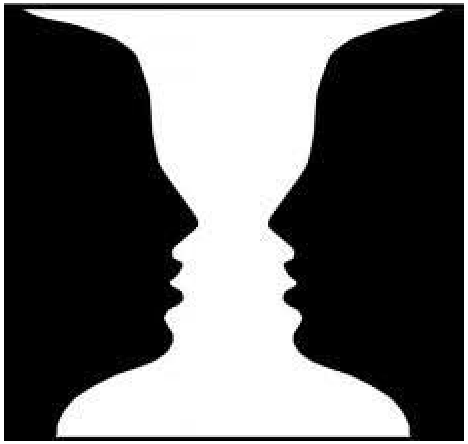
Figure 3.1 An example of tasks in natural language understanding.

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

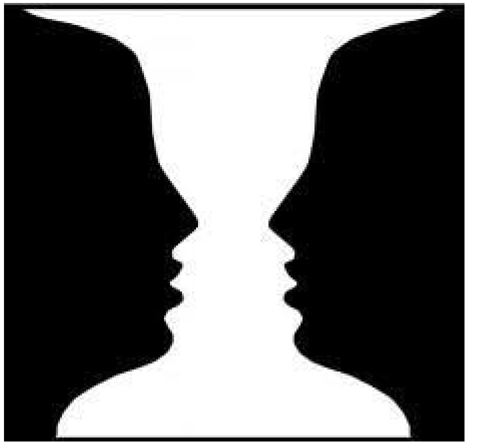
a word may have **multiple syntactic categories** and **multiple senses**. For example, “design” can be a noun or a verb (ambiguous POS); “root” has multiple meanings even as a noun (ambiguous sense)

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” (ambiguous modification)

“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

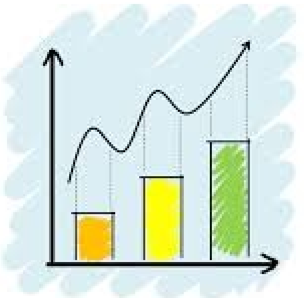
1960s and 1970s: speech recognition (requires only limited understanding
of natural language)

1970s–1980s: story understanding. knowledge representation &
and heuristic inference rules
even simple stories are quite challenging to understand

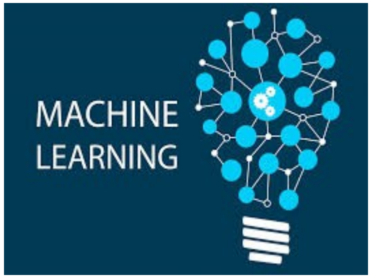
After the 1980s, researchers started moving away from the traditional symbolic
(logic-based) approaches (not robust for real applications)

now…

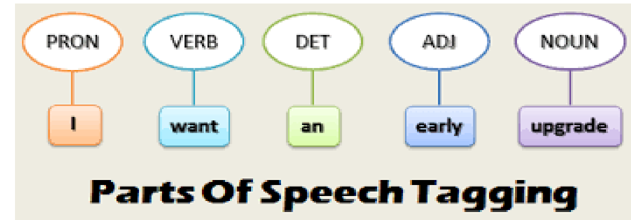
- more attention to **statistical approaches** (more robust, less reliance on human-generated rules)
  - take advantage of **regularities and patterns** in empirical uses of language
  - rely on labeled **training data** by humans and application of **machine learning** techniques



- 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...



part of speech (POS) tagging

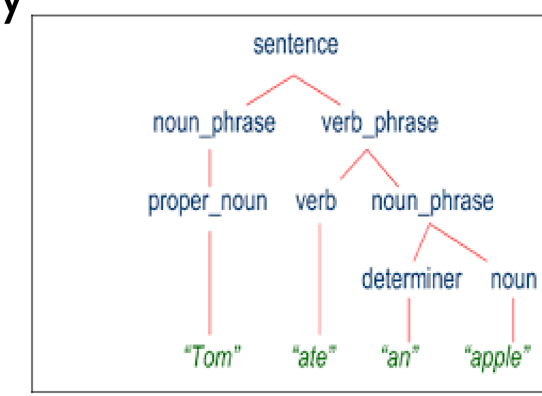
relatively easy task

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...

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell-Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:
LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

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

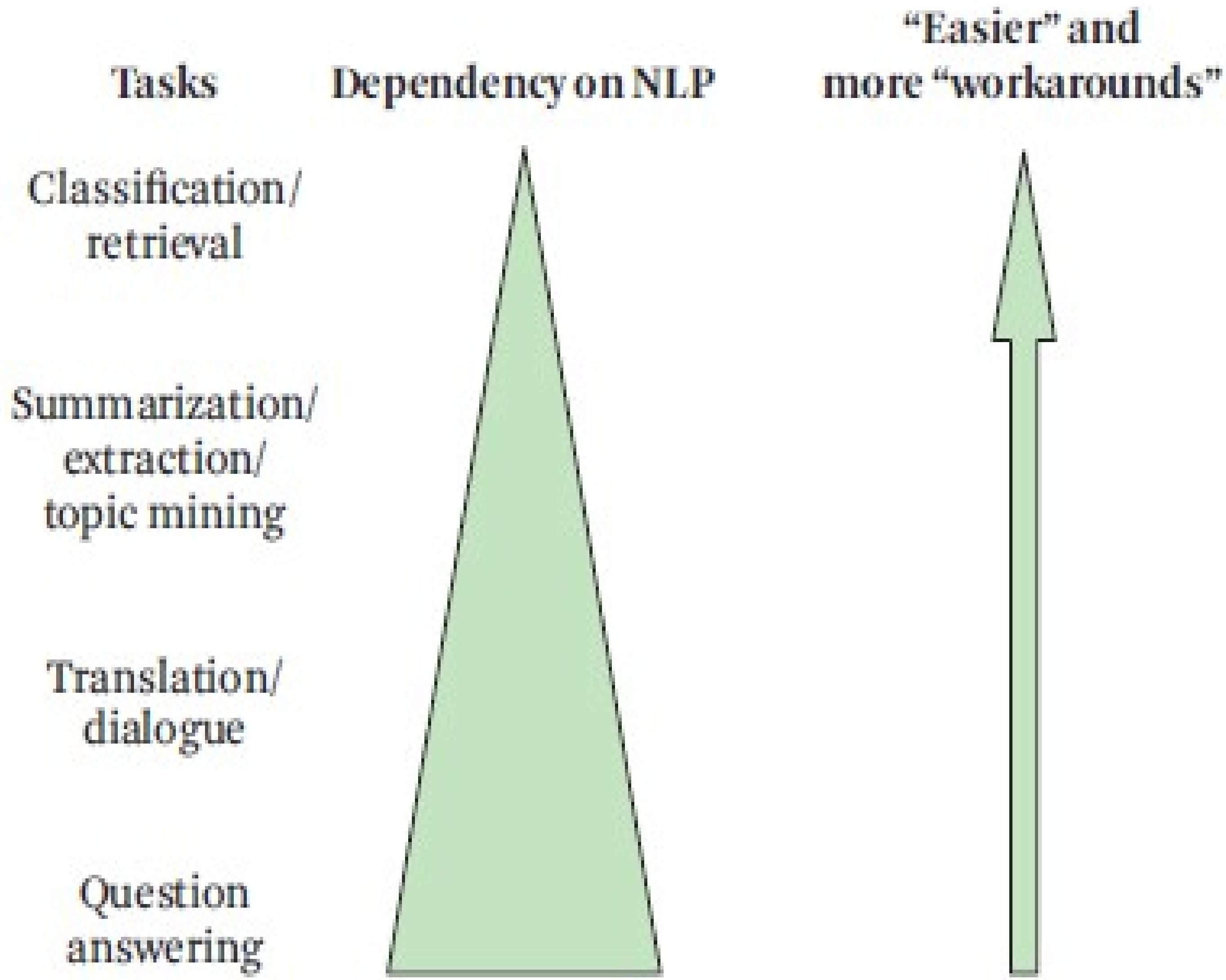


Figure 3.2 "Easy" vs. "difficult" NLP applications.

text representation

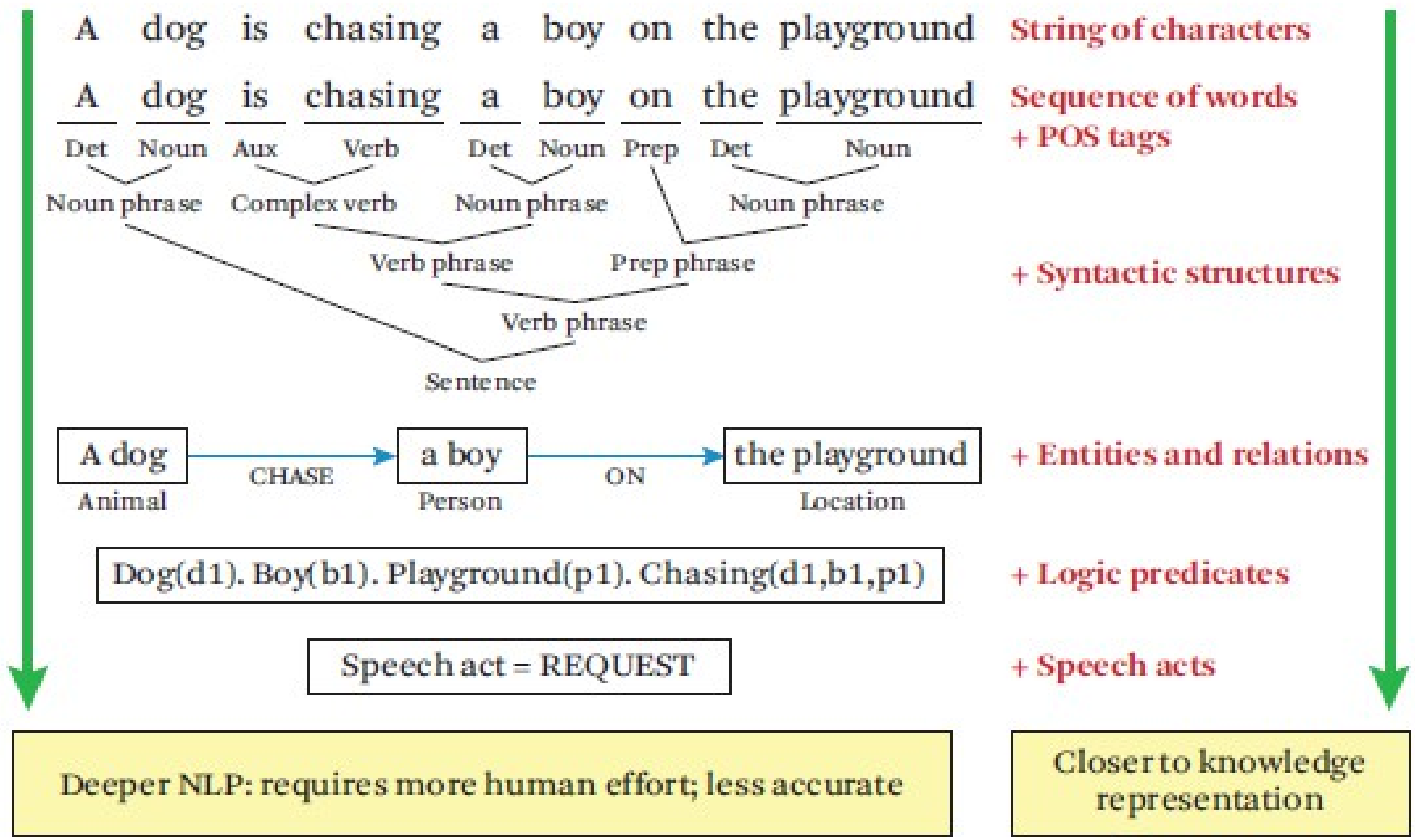
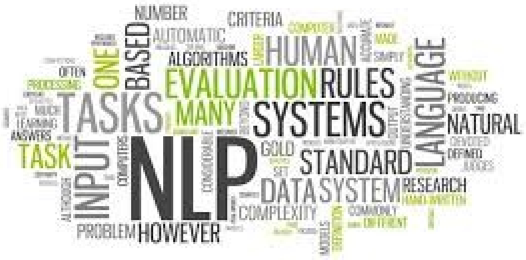


Figure 3.3 Illustration of different levels of text representation.

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

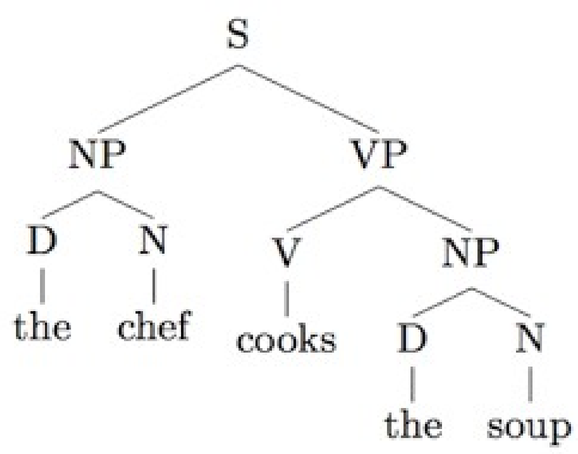
to count, for example, the most frequent **nouns**

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.

statistical language models

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

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

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

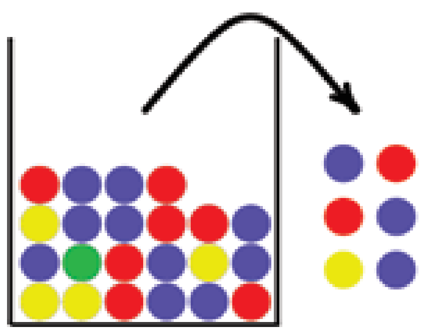
$$p(\text{The equation has a solution}) = 0.000001$$

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**

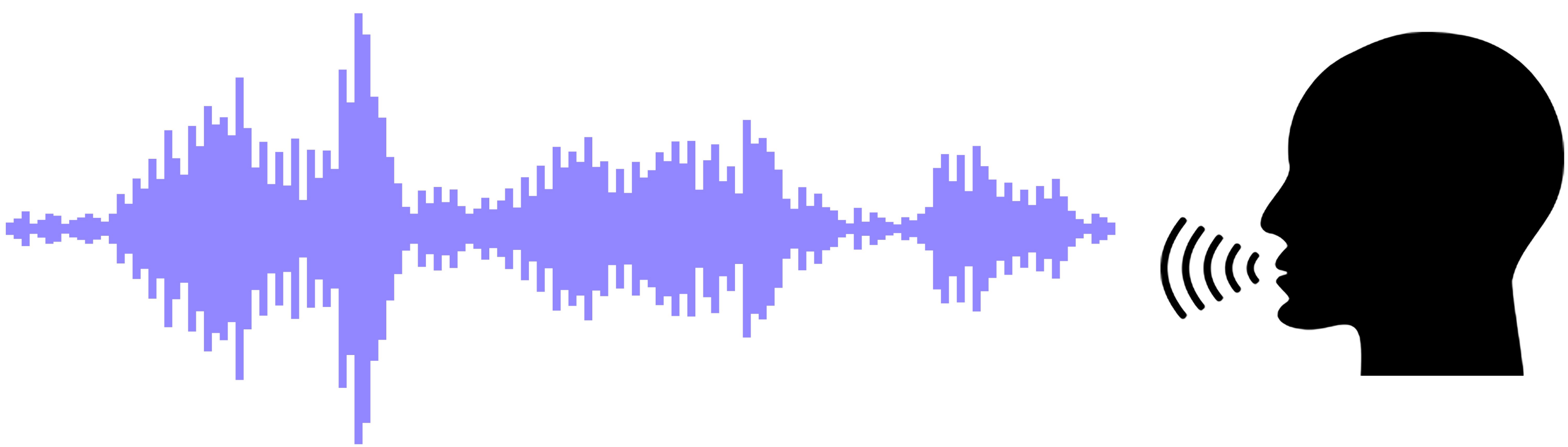


uses of LMs

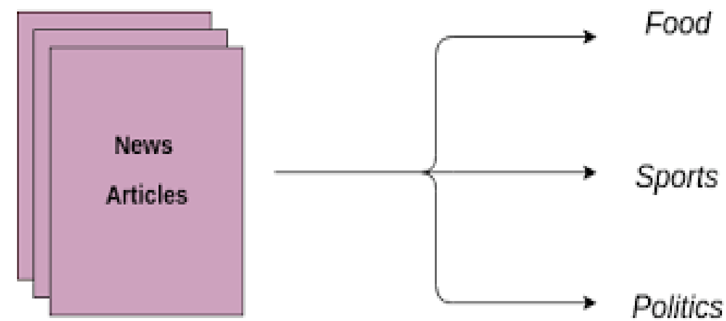
speech recognition

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”**?

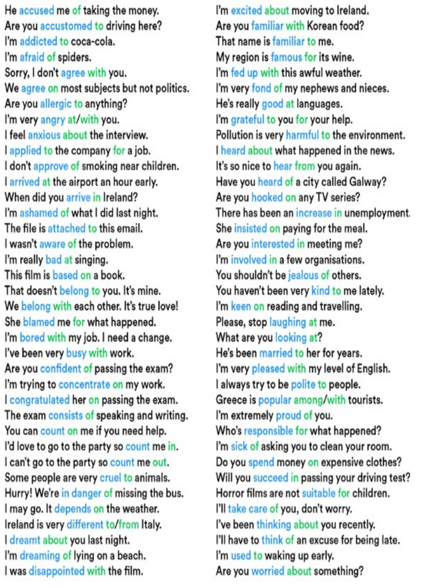
information retrieval

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**!!



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

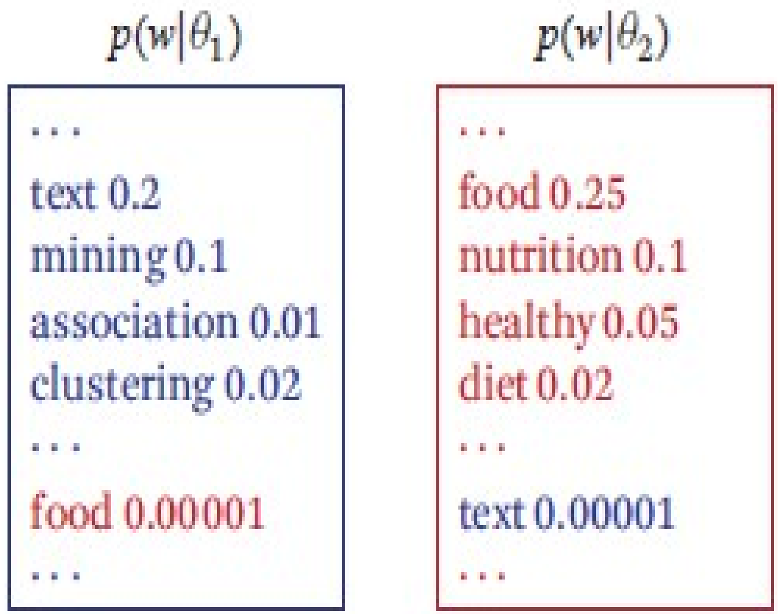


Figure 3.5 Two examples of unigram language models, representing two different topics.

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 | θ) != p(D2 | θ)

what kind of docs would have higher probabilities?

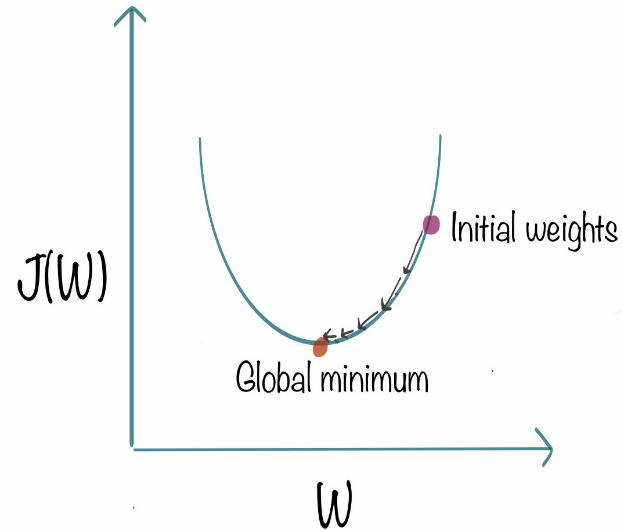
those docs that contain many occurrences of the high probability words according to p(w | θ)

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



maximum likelihood estimator (mle)

seeks a model that **gives the observed data the highest likelihood**

(i.e., best explain the data)

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

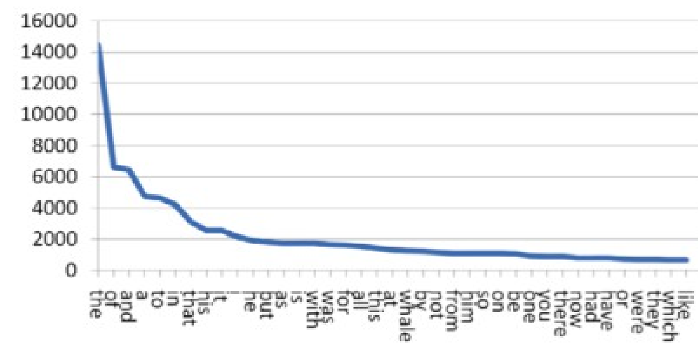
theta: grados de libertad

lo de toda la vida, si hay 3 cuadrados y dos triangulos, 3/5 y 2/5

mle for text representation

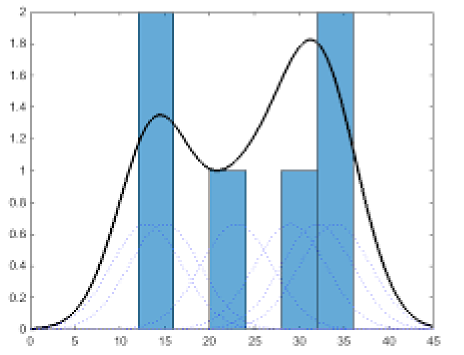
gives each **word** a probability equal to its **relative frequency** in D

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



assigns **zero probability** to any **unseen** words

smoothing methods

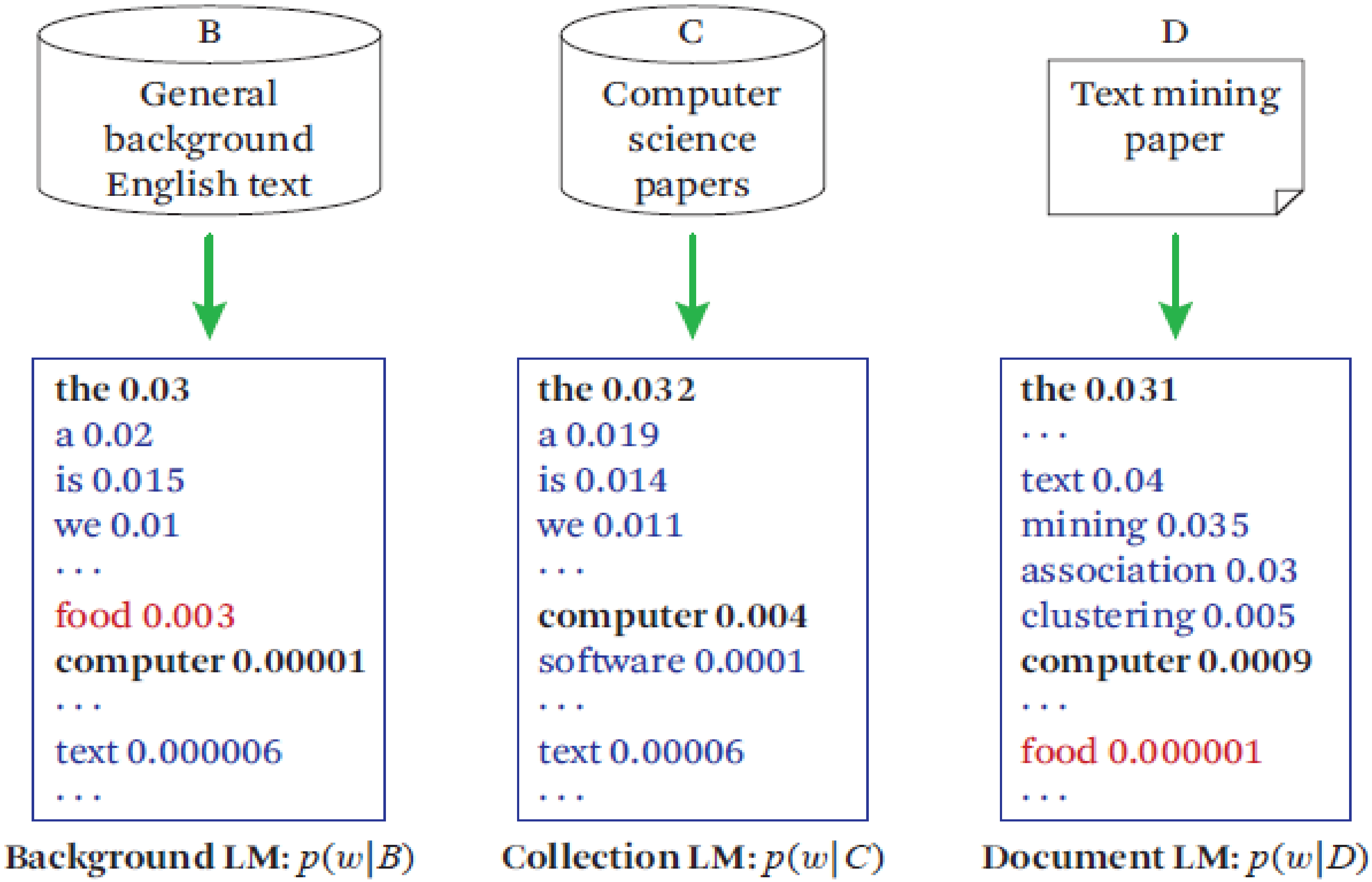


main idea: improve mle by assigning some probability mass to unseen events

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

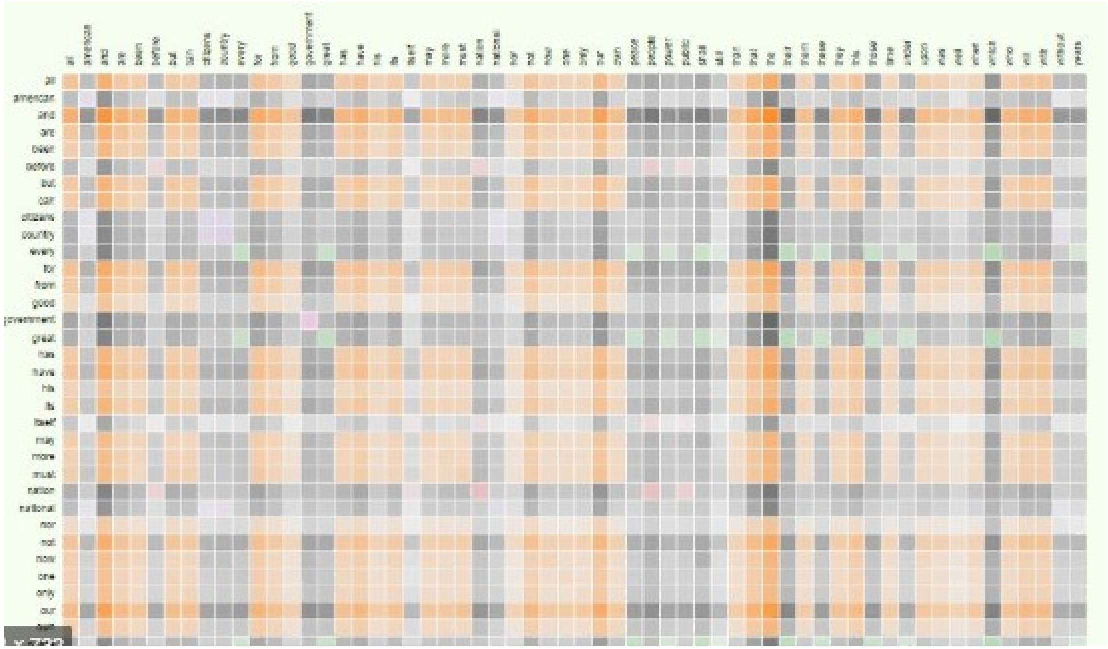
the resulting model, p(w | computer)

this tells us which words

occur frequently

in the context

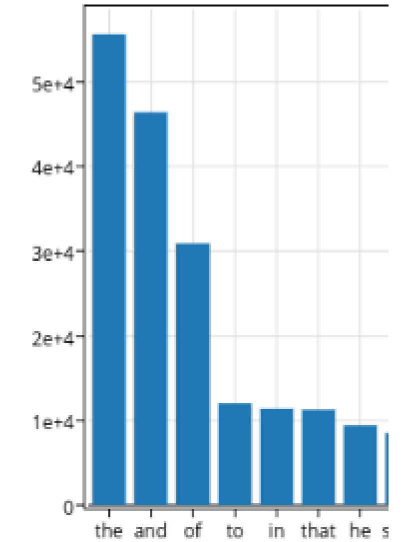
of “computer”



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

cuantas veces mas que en el lenguaje natural se ve esa palabra