TRANSFORM
4EUROPE

Co-funded by the
Erasmus+ Programme
of the European Union

Text mining (I)

David Tomás Díaz
@d_tomas

Universidad de Alicante

lsi Departamento de Lenguajes y Sistemas Informáticos

Contents

- What is Text Mining?
- Text Mining vs. Data Mining
- The Ambiguity and Variability Problem
- The Linguistic Approach
- The Statistical Approach
- The Unreasonable Effectiveness of Data

Contents

- **What is Text Mining?**
- Text Mining vs. Data Mining
- The Ambiguity and Variability Problem
- The Linguistic Approach
- The Statistical Approach
- The Unreasonable Effectiveness of Data

What is Text Mining?

- Text Mining is the science of extracting useful information from large textual data sets
- Why is it interesting?
  - Large amounts of text data created in a variety of social networks, web and other information-centric applications
  - Unstructured data is the easiest form of data which can be created in any application scenario

So Richard, you say you're addicted to facebook??
Richard Peterson likes this!

Contents

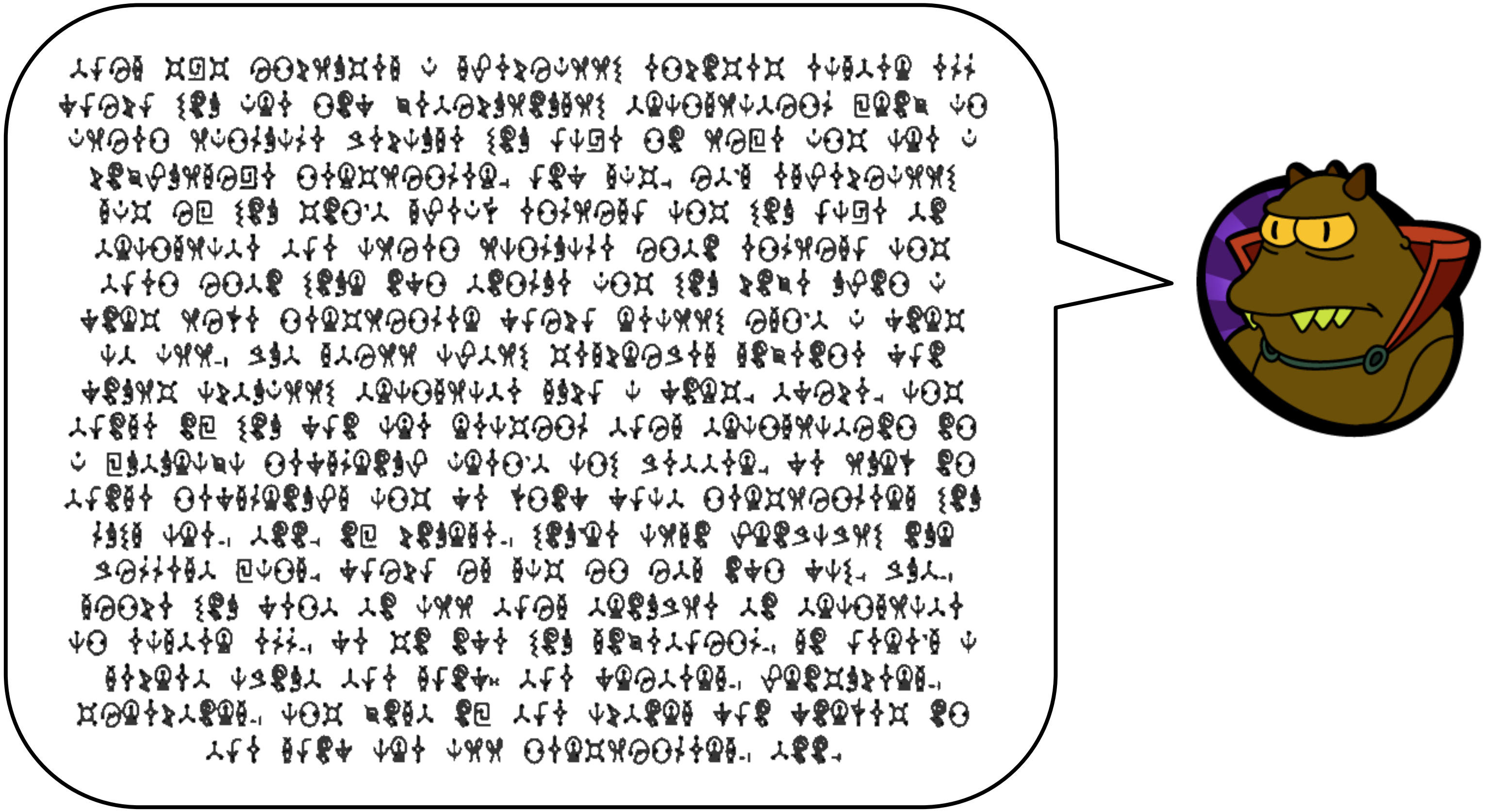
- What is Text Mining?
- **Text Mining vs. Data Mining**
- The Ambiguity and Variability Problem
- The Linguistic Approach
- The Statistical Approach
- The Unreasonable Effectiveness of Data

Text Mining vs. Data Mining

Data Mining	Text Mining
• Looking for patterns in data • Information to be extracted from data: • Implicit (hidden) • Previously unknown • Could hardly be extracted without automatic techniques	• Looking for patterns in text • Information to be extracted from data: • Clearly and explicitly stated in the text • Not couched in a manner that is amenable to automatic processing

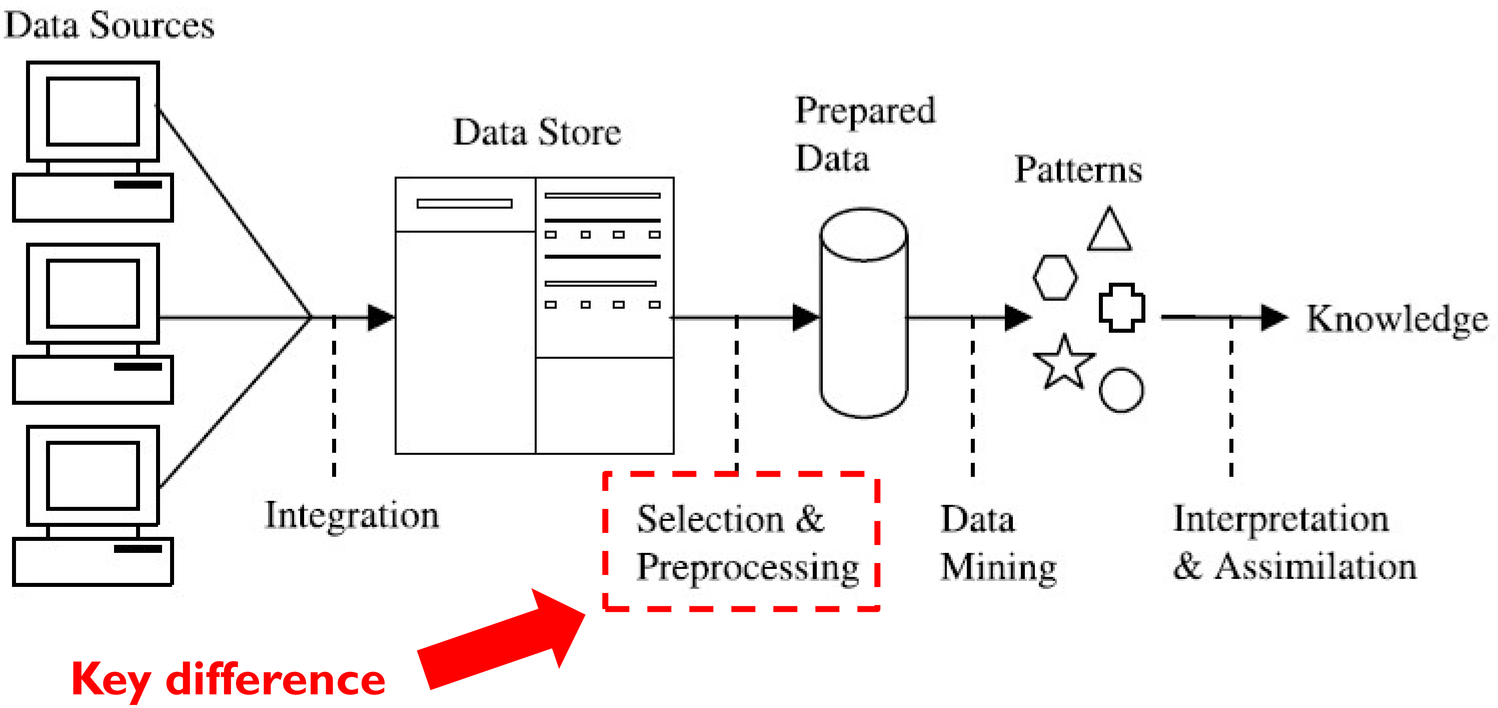
Text Mining vs. Data Mining

- Text is just as opaque (probably more) as raw data when it comes to extracting information



Text Mining vs. Data Mining

- There is a clear difference philosophically…
- … but from the computer's point of view the problems are quite similar



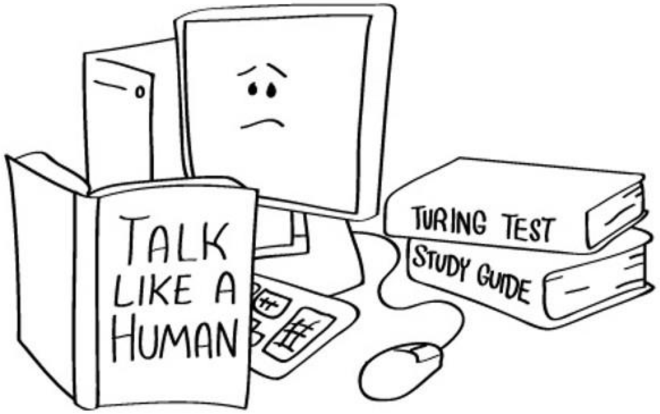
Key difference

Contents

- What is Text Mining?
- Text Mining vs. Data Mining
- **The Ambiguity and Variability Problem**
- The Linguistic Approach
- The Statistical Approach
- The Unreasonable Effectiveness of Data

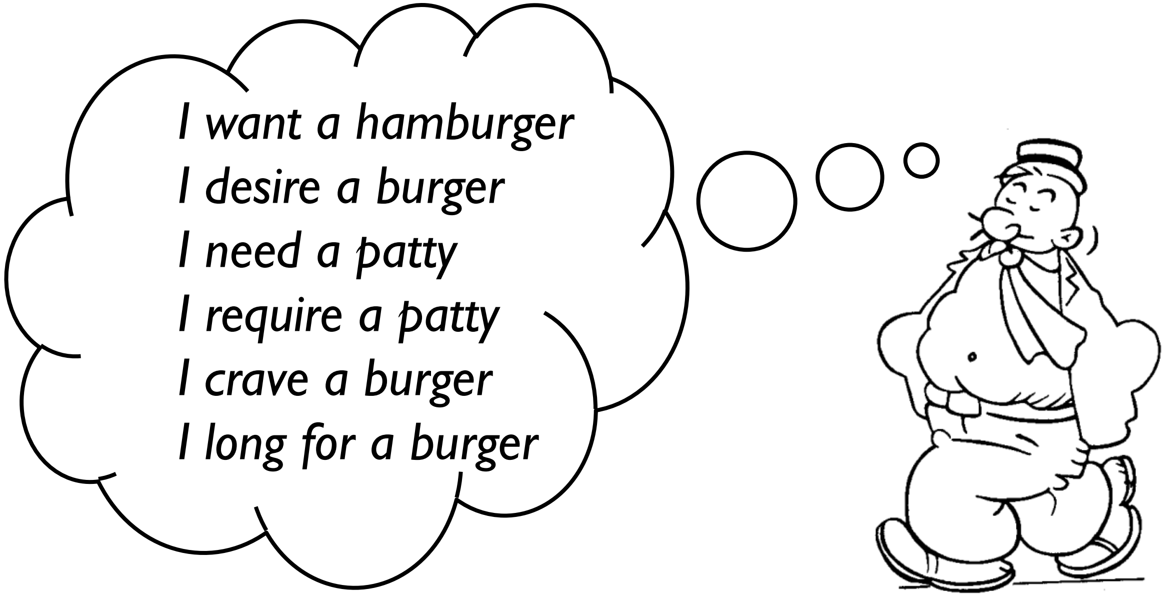
The Ambiguity and Variability Problem

- Natural language understanding seems simple and intuitive for a human…
- …but several factors affect the performance and robustness of automatic systems
- The main problem is the phenomenon of **polysemy** and **synonymy**, i.e., the ambiguity and variability of natural language



The Ambiguity and Variability Problem

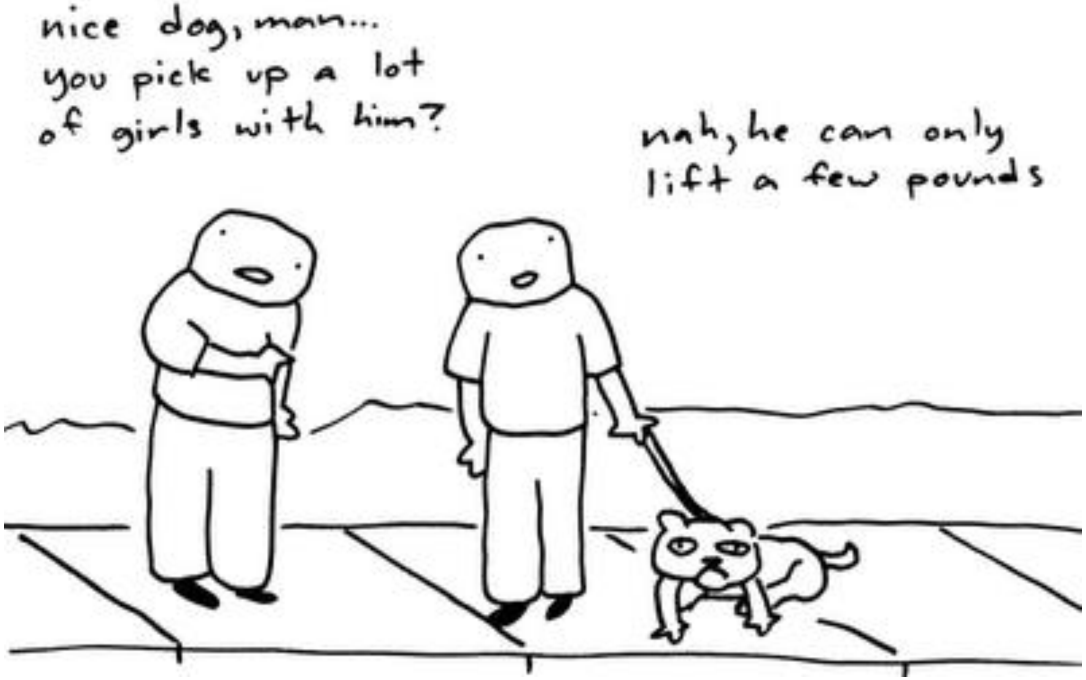
- The variability of natural language
  - A.k.a. **synonymy**
  - Uttering the same information in many different ways
  - Semantically similar sentences can be completely different from a lexical point of view



I want a hamburger
I desire a burger
I need a patty
I require a patty
I crave a burger
I long for a burger

The Ambiguity and Variability Problem

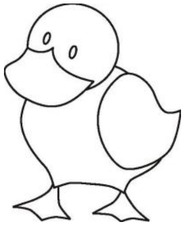
- The ambiguity of natural language
  - A.k.a. **polysemy**
  - Something is ambiguous when it can be understood in two or more possible senses or ways


nice dog, man... you pick up a lot of girls with him?
nah, he can only lift a few pounds

The Ambiguity and Variability Problem

- The ambiguity of natural language
  - A.k.a. **polysemy**
  - Something is ambiguous when it can be understood in two or more possible senses or ways

Find at least 5 meanings of the sentence *I made her duck*

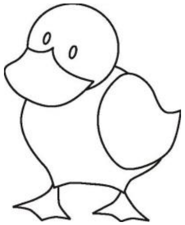


The Ambiguity and Variability Problem

- The ambiguity of natural language
  - A.k.a. **polysemy**
  - Something is ambiguous when it can be understood in two or more possible senses or ways

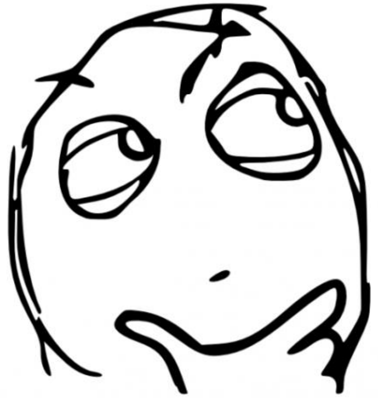
Find at least 5 meanings of the sentence *I made her duck*

- *I cooked duck for her benefit (to eat)*
- *I cooked duck belonging to her*
- *I created the (plaster?) duck she owns*
- *I caused her to quickly lower her head or body*
- *I waved my magic wand and turned her into duck*



The Ambiguity and Variability Problem

How can we deal with these problems and represent language in a computationally tractable form?



The Ambiguity and Variability Problem

Linguistic approach vs. statistical approach

Contents

- What is Text Mining?
- Text Mining vs. Data Mining
- The Ambiguity and Variability Problem
- **The Linguistic Approach**
- The Statistical Approach
- The Unreasonable Effectiveness of Data

The Linguistic Approach

- A.k.a. **symbolic**
- Explicitly store linguistic facts/knowledge
- Analysis of different linguistic levels
  - Phonology
  - Morphology
  - Syntax
  - Semantics
  - Pragmatics

The Linguistic Approach

- A.k.a. **symbolic**
- Explicitly store linguistic facts/knowledge
- Analysis of different linguistic levels
  - ~~Phonology~~
  - **Morphology**
  - **Syntax**
  - **Semantics**
  - ~~Pragmatics~~

The Linguistic Approach

- Morphology
  - The study of how words are composed of **morphemes**
  - Two broad classes of morphemes
    - **Stems**: main morpheme of the word, supplying meaning
    - **Affixes**: pieces that combine with stems to modify their meanings and grammatical functions

impossible
enjoying
unreachable
unbelievable

🟦 stems
■ affixes

The Linguistic Approach

- Morphology
  - Part-of-Speech (POS) tagging
    - Marking up a word in a text as corresponding to a particular part of speech (noun, verb, adjective, …)

The grand jury commented on a number of other topics.

→

```
The the DT 1
grand grand JJ 0.832524
jury jury NN 1
commented comment VBD 0.954545
on on IN 0.971769
a 1 Z 0.99998
number number NN 0.998704
of of IN 0.999898
other other JJ 0.632399
topics topic NNS 1
. . Fp 1
```

The Linguistic Approach

- Morphology
  - Part-of-Speech (POS) tagging
    - Example: Use of Freeling
      - □ http://nlp.lsi.upc.edu/freeling/demo/demo.php
      - □ *Select output > PoS Tagging*

Today is Monday, May 23, 2022. It is 6:00 p.m. I am attending a Text Mining seminar at the University of Alicante, in Spain. The teacher is David. He tries to make it interesting but sometimes fails.

The Linguistic Approach

Let's practice!

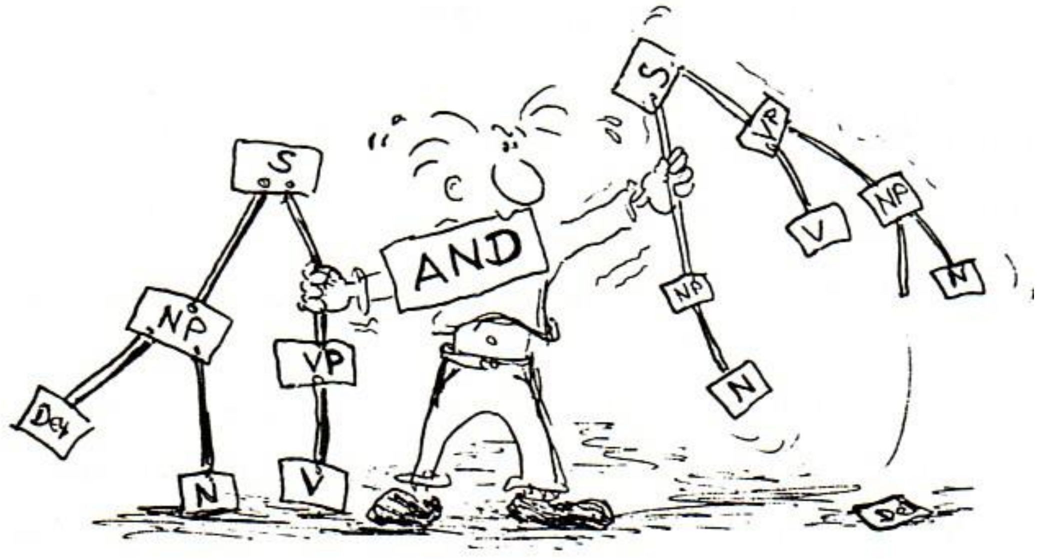
https://bit.ly/3a4FbkE

The Linguistic Approach

- Morphology
  - Useful for…
    - It is a first step of many Text Mining practical tasks
    - Syntactic analysis needs to know if a word is a name or verb before parsing
    - Finding names and relations for Information Extraction
    - Identify stems and lemmas for Information Retrieval
    - Remove potential stopwords for dimensionality reduction
    - …

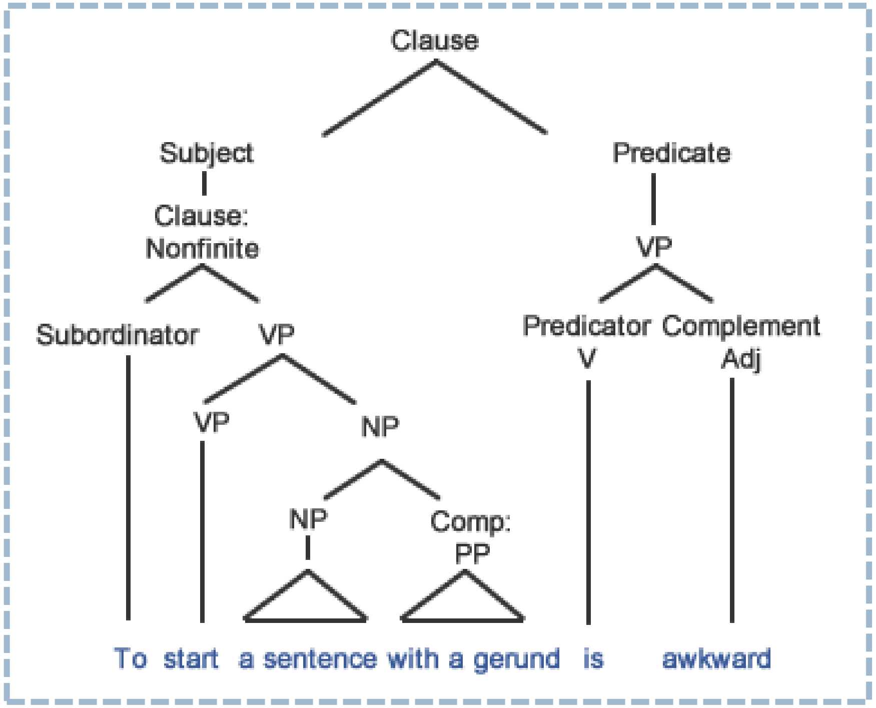
The Linguistic Approach

- Syntax
  - Syntactic analysis is concerned with the construction of sentences
  - Syntactic structure indicates how the words are related to each other



The Linguistic Approach

- Syntax
  - Full parsing
    - Nested phrase structure
    - Provides the role of the constituents in the main sentence



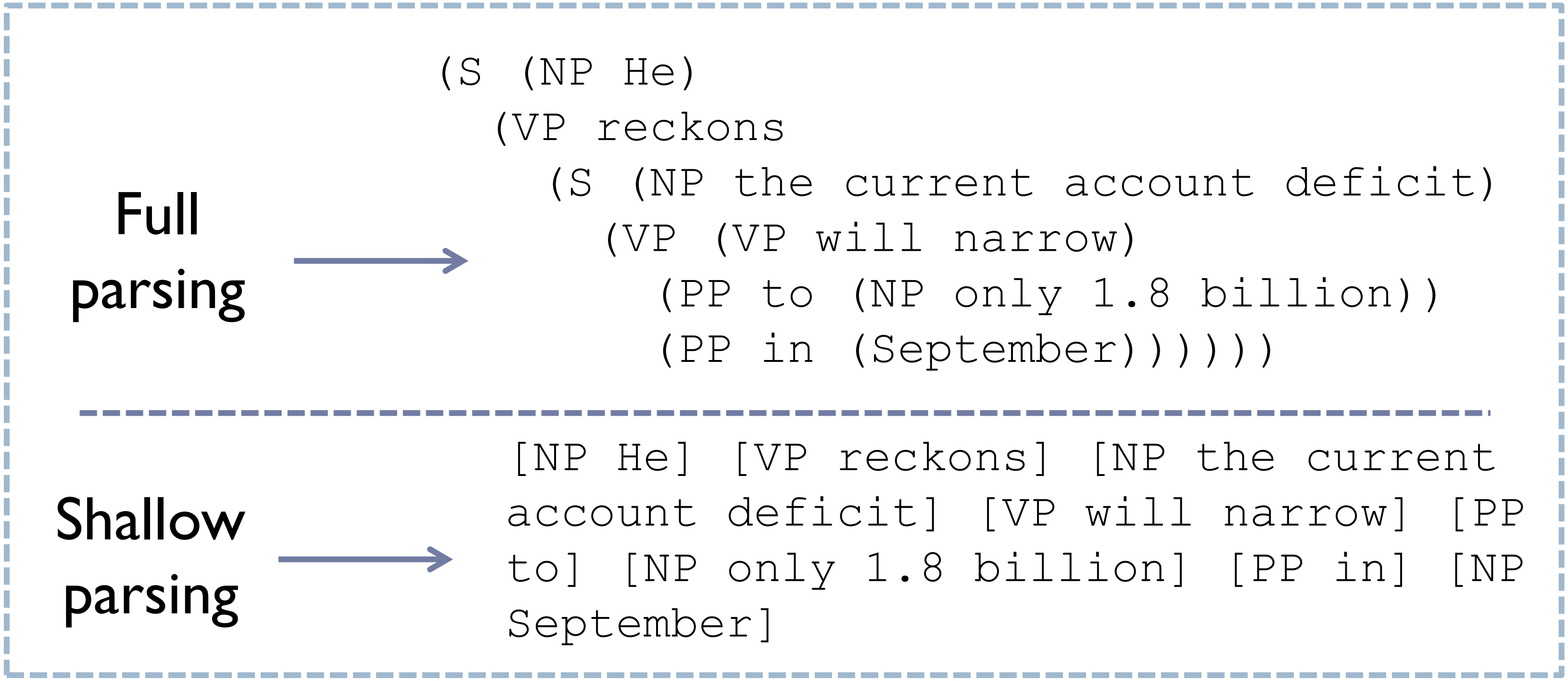
The Linguistic Approach

- Syntax
  - Shallow parsing
    - A.k.a. **chunking** or light parsing
    - Analysis of a sentence which identifies the **constituents** (noun groups, verbs, verb groups, etc.)
    - Partitions plain text into sequences of semantically related words
    - Does not specify their internal structure, nor their role in the main sentence

```
[NP Jack and Jill] [VP went] [ADVP up] [NP the hill]
    [VP to fetch] [NP a pail] [PP of] [NP water]
```

The Linguistic Approach

- Syntax
  - Shallow parsing
    - The task is simpler (and more frequently used) than full parsing



Full parsing

```
(S (NP He)
   (VP reckons
     (S (NP the current account deficit)
       (VP (VP will narrow)
         (PP to (NP only 1.8 billion))
         (PP in (September))))))
```

Shallow parsing

```
[NP He] [VP reckons] [NP the current
account deficit] [VP will narrow] [PP
to] [NP only 1.8 billion] [PP in] [NP
September]
```

The Linguistic Approach

Let’s practice!

https://bit.ly/3a4FbkE

The Linguistic Approach

- Syntax
  - Useful for…
    - Prepare for semantic interpretation
    - Question Answering
    - Information Extraction
    - Language Generation
    - Machine Translation
    - …

The Linguistic Approach

- Semantics
  - The study of meaning of linguistic expressions
  - Subjects that semantics study
    - Lexical semantics
      - □ Meaning of individual words
    - Compositionality
      - □ How can the meaning of a larger unit be computed from the meaning of its parts?
    - Ambiguity resolution
      - □ If a linguistic expression has several distinct meanings, how can the correct one in context be determined? → **Word Sense Disambiguation**

The Linguistic Approach

- Semantics
  - Word Sense Disambiguation
    - Example: Use of Freeling
      - □ http://nlp.lsi.upc.edu/freeling/demo/demo.php
      - □ *Analysis options > WN sense annotation*

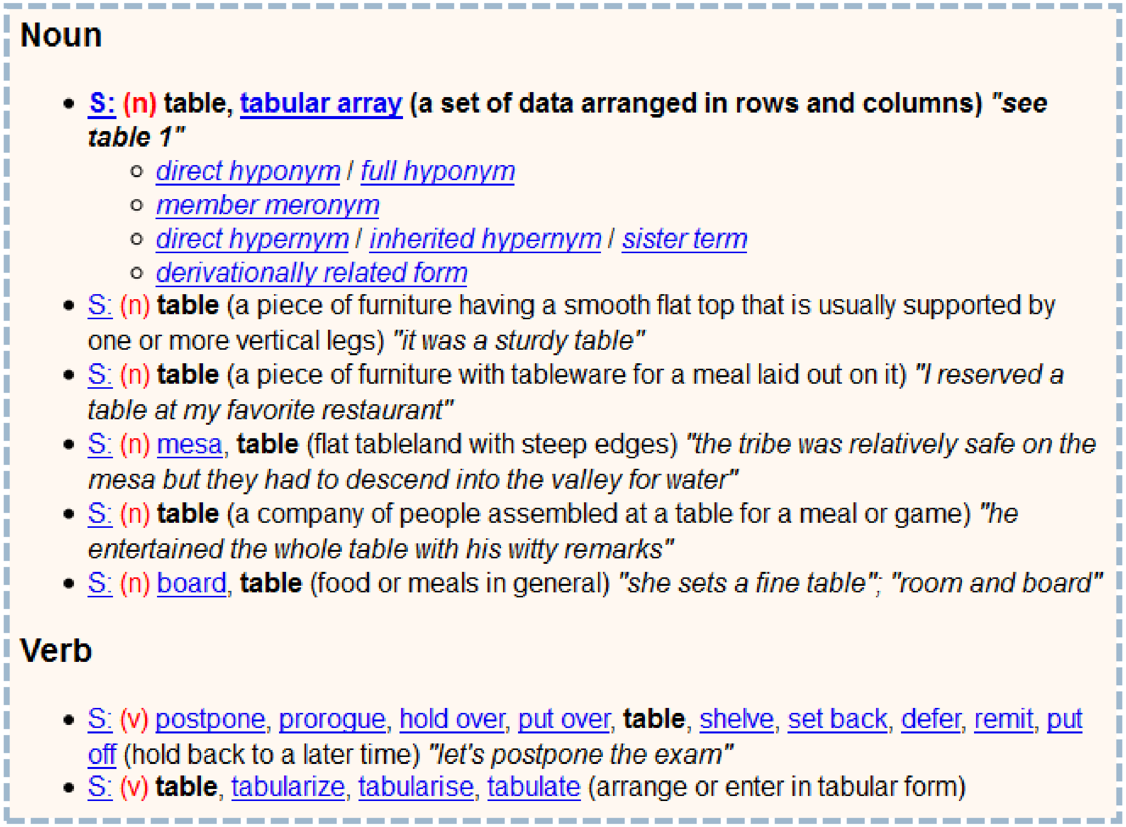
> Today is Monday, May 23, 2022. It is 6:30 p.m. I am attending a Text Mining seminar at the University of Alicante, in Spain. The teacher is David. He tries to make it interesting but sometimes fails.

The Linguistic Approach

- Semantics
  - WordNet
    - Most used electronic **lexical database** for English
    - Helps to solve the synonymy problem
    - Also available in other languages (EuroWordNet)
    - Lexical items are categorised into (more than 115K) synsets
    - **Synset**: set of synonyms, a dictionary-style definition (or gloss), and some examples of use
    - Lexical relationships are implemented as semantic networks, where applications can traverse to find synonyms, antonyms, **hypernyms**, **hyponyms**, …

The Linguistic Approach

- Semantics
  - WordNet



Noun

- **S: (n) table, tabular array (a set of data arranged in rows and columns) *"see table 1"***
  - *direct hyponym* / *full hyponym*
  - *member meronym*
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*
- S: (n) **table** (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs) *"it was a sturdy table"*
- S: (n) **table** (a piece of furniture with tableware for a meal laid out on it) *"I reserved a table at my favorite restaurant"*
- S: (n) mesa, **table** (flat tableland with steep edges) *"the tribe was relatively safe on the mesa but they had to descend into the valley for water"*
- S: (n) **table** (a company of people assembled at a table for a meal or game) *"he entertained the whole table with his witty remarks"*
- S: (n) board, **table** (food or meals in general) *"she sets a fine table"; "room and board"*

Verb

- S: (v) postpone, prorogue, hold over, put over, **table**, shelve, set back, defer, remit, put off (hold back to a later time) *"let's postpone the exam"*
- S: (v) **table**, tabularize, tabularise, tabulate (arrange or enter in tabular form)

The Linguistic Approach

- Semantics
  - WordNet
    - Example: Use of WordNet
      - □ http://wordnetweb.princeton.edu/perl/webwn

- dog
- pick up
- Kennedy
- Spain

The Linguistic Approach

Let's practice!

https://bit.ly/3a4FbkE

The Linguistic Approach

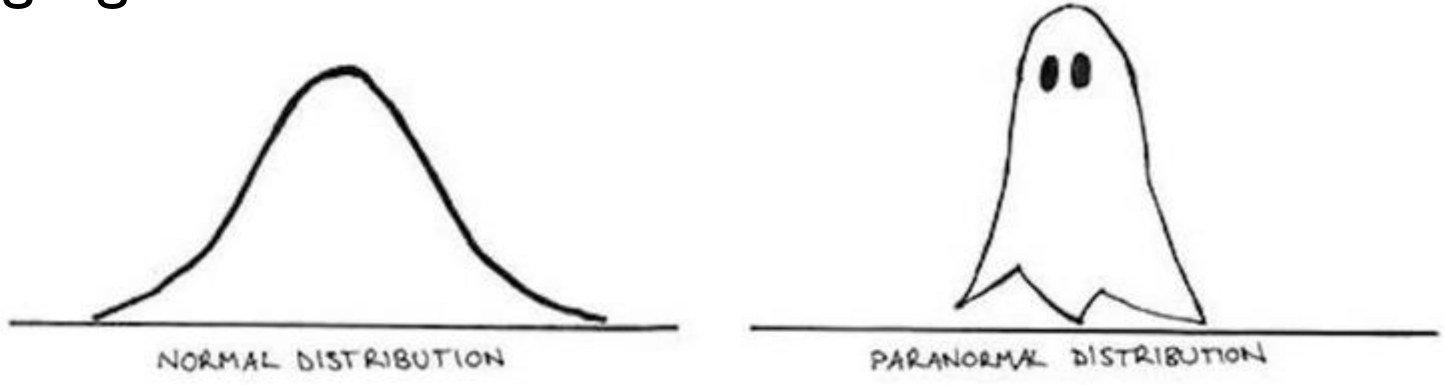
- Semantics
  - Useful for…
    - Finding synonyms
    - Word Sense Disambiguation
    - Question Answering
    - Machine Translation
    - Semantic Role Labelling
    - Ontology learning and population
    - …

Contents

- What is Text Mining?
- Text Mining vs. Data Mining
- The Ambiguity and Variability Problem
- The Linguistic Approach
- **The Statistical Approach**
- The Unreasonable Effectiveness of Data

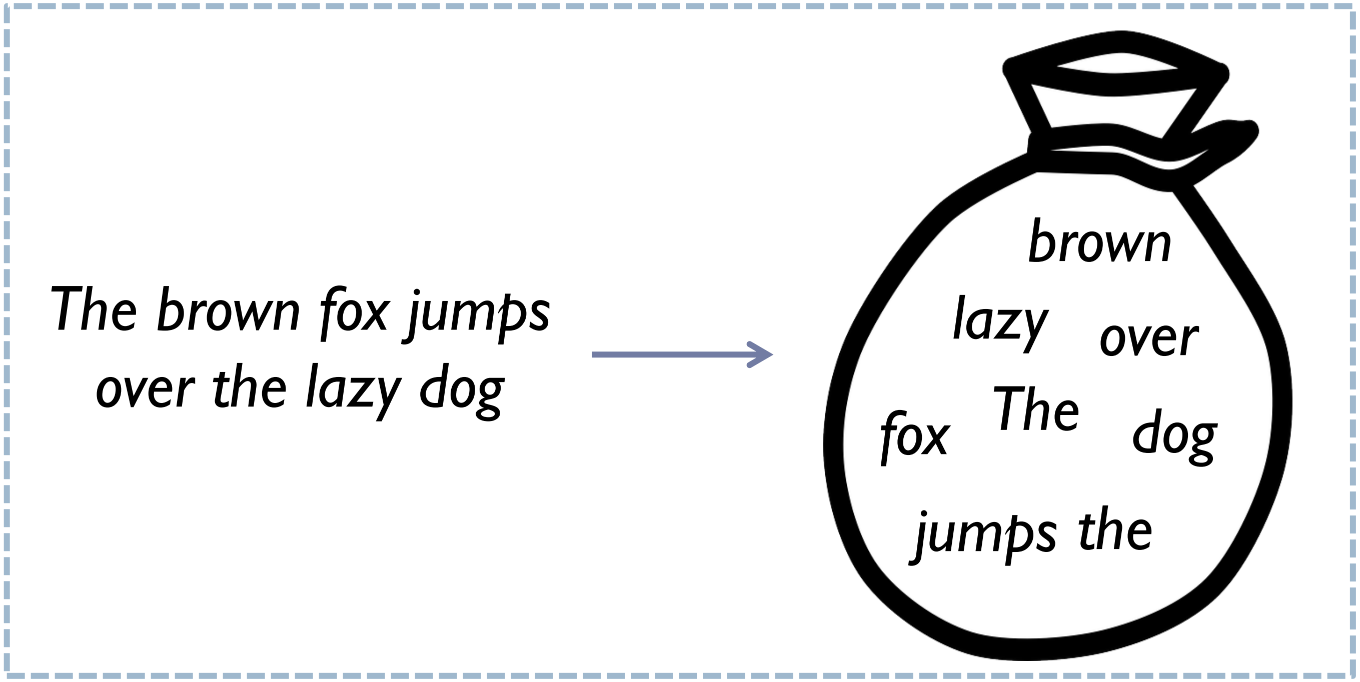
The Statistical Approach

- Statistical models that express the probability of a particular observation based on big textual **corpus**
- Does not take into consideration order, structure or meaning
- Captures regularities in linguistic expressions
- Operate by treating the input as though it were data, not language



The Statistical Approach

- Bag-of-words (BOW)
  - A document is considered as a set of words regardless of the word order and grammar



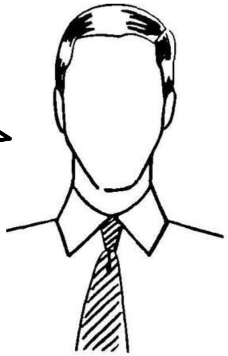
The brown fox jumps over the lazy dog

brown
lazy over
fox The dog
jumps the

The Statistical Approach

- Statistical semantics
  - No-linguistics does not imply no-semantics
  - Distributional hypothesis

Words that occur in the same contexts tend to have similar meanings



Zellig Harris (1954)

The Statistical Approach

- Statistical semantics
  - Distributional hypothesis

What does ***tonked*** and ***tezgüno*** likely mean?

The Statistical Approach

- Statistical semantics
  - Distributional hypothesis

What does ***tonked*** and ***tezgüno*** likely mean?

*Sue had wanted the deed to the house for twenty years. After Bob finally **tonked** the house to Sue, she **tonked** Francis her duplex.*

The Statistical Approach

- Statistical semantics
  - Distributional hypothesis

What does ***tonked*** and ***tezgüno*** likely mean?

Sue had wanted the deed to the house for twenty years.After Bob finally ***tonked*** *the house to Sue, she* ***tonked*** *Francis her duplex.*

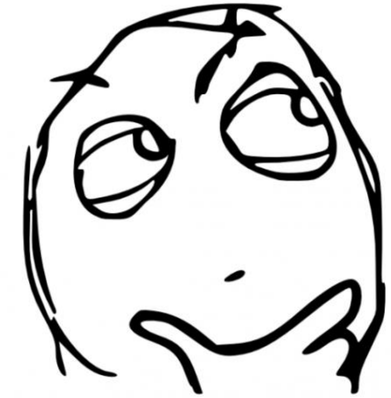
A bottle of ***tezgüno*** *is on the table.*
Everyone likes ***tezgüno****.*
Tezgüno *makes you drunk.*
We make ***tezgüno*** *out of corn.*

Contents

- What is Text Mining?
- Text Mining vs. Data Mining
- The Ambiguity and Variability Problem
- The Linguistic Approach
- The Statistical Approach
- **The Unreasonable Effectiveness of Data**

The Unreasonable Effectiveness of Data

Can the ability to process huge amounts of text compensate for relatively simple techniques?



The Unreasonable Effectiveness of Data

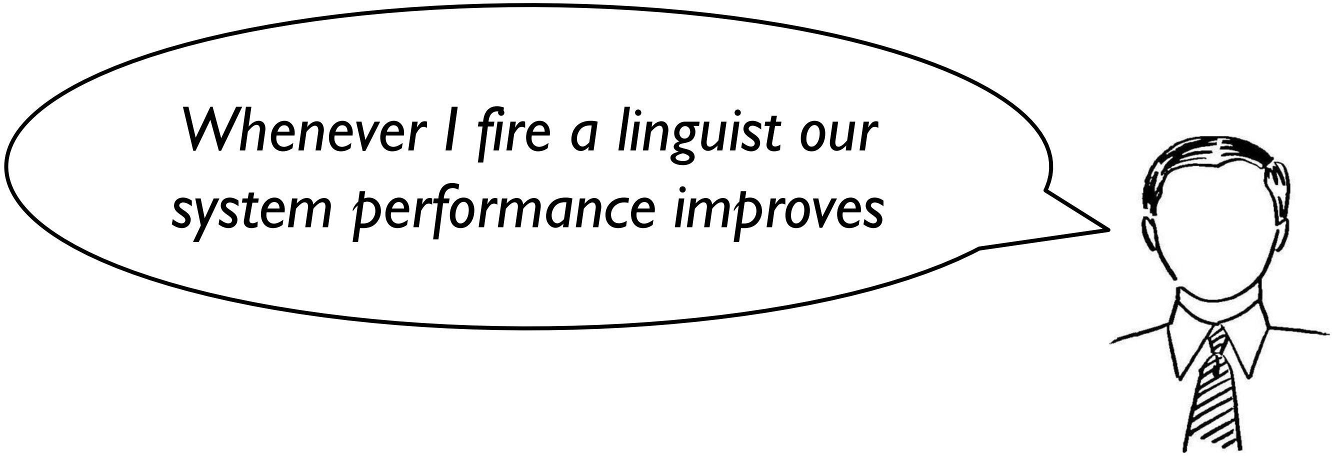
For many tasks, words and word combinations provide all the representational machinery we need to learn from text

The Unreasonable Effectiveness of Data

- Problems of the linguistic approach
  - There are hundreds of thousands of vocabulary words and a vast variety of grammatical constructions
  - Every day new words are coined and old usages are modified
  - We cannot reduce what we want to say to the free combination of a few abstract primitives
  - Inference over sophisticated models and extraction of deep features are often computationally intensive, **do not scale well**
  - Different languages require different tools…
  - … and also informal language requires different tools

The Unreasonable Effectiveness of Data

- Problems of the linguistic approach


Whenever I fire a linguist our system performance improves

Frederick Jelinek (1988)

The Unreasonable Effectiveness of Data

- Benefits of the statistical approach
  - There is a growing body of evidence, at least in text processing, that the amount of data matters more than features and algorithms
  - It makes sense to take advantage of the plentiful amounts of data that surround us (Big Data!)
  - Superficial word-level features coupled with simple models in most cases trump sophisticated models with deeper features and less data
  - The statistical approach is **topic** and (mostly) **language independent**

The Unreasonable Effectiveness of Data

- Nevertheless…
  - The linguistic approach play an important role in obtaining a semantically more meaningful representation of text
    - In special domains (e.g., biomedical domain)
    - For special mining tasks (e.g., extraction of knowledge from the Web)

- “The Parable of Google Flu: Traps in Big Data Analysis”, Lazer et al. (2014)

- IBM Watson

The Unreasonable Effectiveness of Data

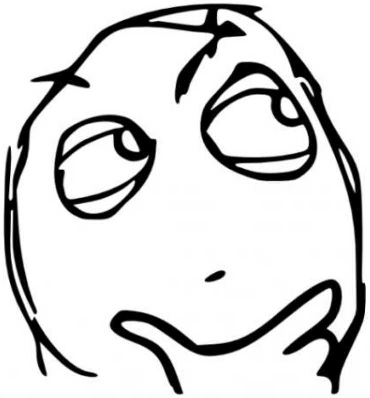
- Nevertheless…

Some of my best friends are linguists

Frederick Jelinek (2004)

The Unreasonable Effectiveness of Data

What should we do then?



The Unreasonable Effectiveness of Data

Embrace both viewpoints

- Shallow (statistical) processing of unrestricted text
- Deep (linguistic) processing of domain-specific material

@d_tomas

David Tomás Díaz