The Basics of NLP

Pierre Colombo

pierre.colombo@centralesupelec.fr

MICS - CentraleSupelec

Advanced Natural Language Processing

https://github.com/PierreColombo/NLP_CS

CentraleSupélec

CV

Education & Diploma

2018. CentraleSupelec + EPFL

2021. PhD in Computer Science, Telecom Paris, Institut Polytechnique de Paris, *France*
Title: Learning to represent and generate text using information measures

Work Life

Before ***2021.*** IBM, Disney Research, P&G, …..

Beginning ***2022.*** Post Doc at CS

Since ***2022.*** Associate Professor at MICS (CS)

Since ***2024.*** Chief Scientist Officer at Equall (NYC - Lisbon - Paris)

Course Objective

Goal: Provide a toolkit of concepts and methods to **describe** and **tackle** NLP problems in real-life.

- **Introduce core ideas** at the basis of modern NLP algorithms
- Focus on **Machine Learning & Deep Learning applied to NLP**
- Focus on **empirical considerations** (accuracy, memory, speed) as opposed to theoretical guarantees

Course Logistic

Each session will consist of a 1h15 class with a 15-minute presentation on the following topics:

1. 11/03: - Efficient Estimation of Word Representations in Vector Space

2. 18/03: - BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

3. 25/03: - ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

4. 08/04: - Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

5. 15/04: (During Lab we will have guests)
    - LoRA: Low-Rank Adaptation of Large Language Models
    - Scaling Instruction-Finetuned Language Models
    - Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in NLP
6. 16/04: (During Lab we will have guests)
    - Language Models are Unsupervised Multitask Learners
    - LLaMA: Open and Efficient Foundation Language Models
    - A Survey on Knowledge Distillation of Large Language Models

Course Evaluation

Paper Presentation (50% of the grade):

1. You have one research paper to present
2. You need to register today by group of 5
3. Details are available at Link

Kaggle competition (50% of the grade):

1. You have one classification task to solve.
2. You need to register today by group of 5
3. Details are available at Link

Lectures Outline

1. The Basics of Natural Language Processing
2. Representing Text with Vectors
3. Deep Learning Methods for NLP
4. Classification for NLP / Revision of Transformers
5. Generative AI for NLP / Revision of Transformers
6. Introduction to RAG / Introduction to Distillation
7. Badass Language Modeling: CroissantLLM / TowerLLM

Labs Outline

1. Describe Statistically large scale corpora
2. Statistical Based and Word2vec Based Retriever
3. Task-Specific Modelling with Neural Networks
4. Task-Specific Modelling with Neural Networks (II)
5. Machine Translation

6. Paper Presentations (5-6-7)

7. Paper Presentations (8-9-10)

Today Lecture Outline

- **Why Natural Language Processing?**
- **What is Natural Language Processing?**
  - Modelling Framework
  - Tokenization as a first-step task
  - Overview of NLP Tasks
- **A Brief History of NLP**
- **How to tackle any NLP problem?**

NLP is Everywhere.

Education

Culture

Healthcare

Legal

Why Natural Language Processing?

What do we use language for?

- We **communicate** using language
- We **think** (partly) with language
- We **tell stories** in language
- We build **Scientific Theories** with language
- We make friends/build **relationships**

Why NLP ?

- **Access Knowledge** (search engine, recommender system…)
- **Communicate** (e.g. Translation)
- **Linguistics** and **Cognitive Sciences** (Analyse Languages themselves)

Why Natural Language Processing?

Amount of online textual data…

- 70 billion web-pages online (1.9 billion websites)
- 55 million Wikipedia articles

…Growing at a fast pace

- 9000 tweets/second
- 3 million mail / second (60% spam)

Why Natural Language Processing?

Potential Users of Natural Language Processing

- 7.9 billion people use some sort of language (January 2022)
- 4.7 billion internet users (January 2021) (~59%)
- 4.2 billion social media users (January 2021) (~54%)

NLP is reshaping the Wold

Time To Acquire 1M User

Netflix ~ 3.5 years

Airbnb ~ 2.5 years

Spotify ~ 5 months

Facebook ~ 10 months

Instagram ~ 2.5 months

iPhone ~ 74 day

Chat GPT ~ 5 day

Why Natural Language Processing?

What Products ?

- Search: +2 billion Google users, 700 millions Baidu users
- Social Media: +3 billion users of Social media (Facebook, Instagram, WeChat, Twitter...)
- Voice assistant: +100 million users (Alexa, Siri, Google Assistant)
- Machine Translation: 500M users for google translate

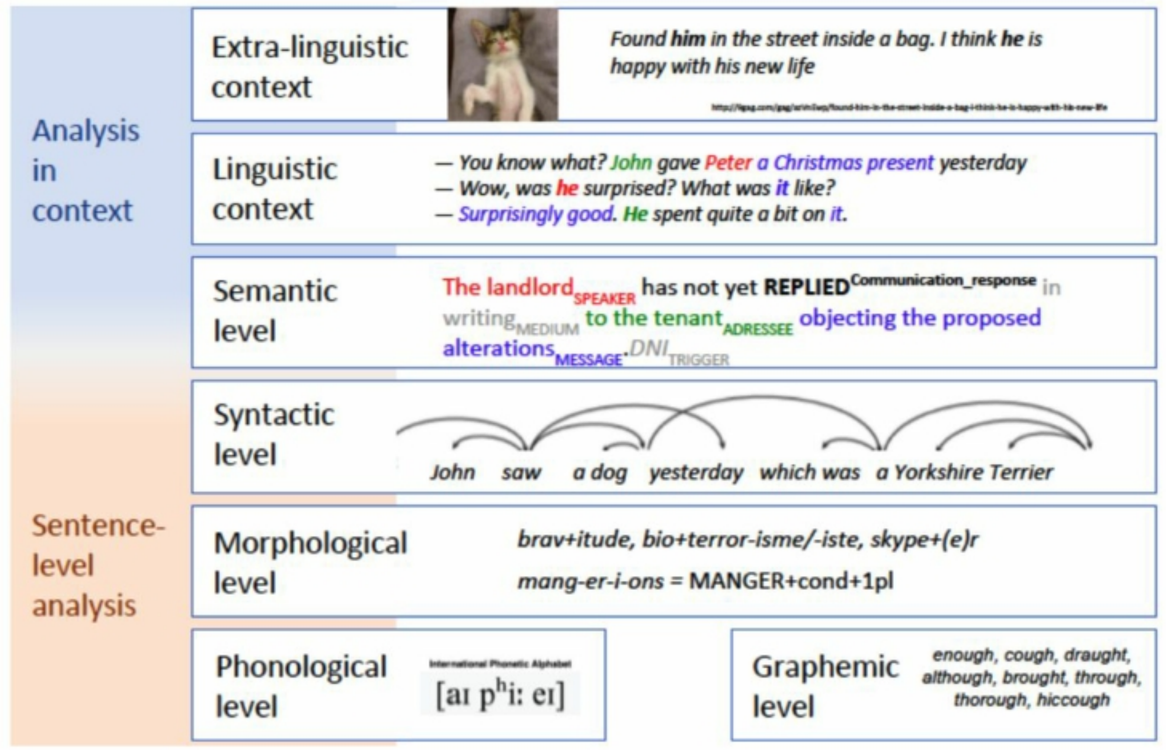
Why is Language Hard to Model?

A Definition of Language

Definition 1: *Language is a means to communicate, it is a semiotic system. By that we simply mean that it is a set of signs . A sign is a pair consisting in [...] a signifier and a signified .*

Definition 2: *A sign consists in a phonological structure, a morphological structure, a syntactic structure and a semantic structure*

The Six Levels of Linguistics Analysis



Analysis in context

Extra-linguistic context

*Found **him** in the street inside a bag. I think **he** is happy with his new life*

http://9gag.com/gag/[illegible]/found-him-in-the-street-inside-a-bag-i-think-he-is-happy-with-his-new-life

Linguistic context

— You know what? John gave Peter a Christmas present yesterday
*— Wow, was **he** surprised? What was **it** like?*
— Surprisingly good. He spent quite a bit on it.

Semantic level

The landlord$_{\text{SPEAKER}}$ has not yet **REPLIED**$^{\text{Communication\_response}}$ in writing$_{\text{MEDIUM}}$ to the tenant$_{\text{ADRESSEE}}$ objecting the proposed alterations$_{\text{MESSAGE}}$. *DNI*$_{\text{TRIGGER}}$

Sentence-level analysis

Syntactic level

John saw a dog yesterday which was a Yorkshire Terrier

Morphological level

brav+itude, bio+terror-isme/-iste, skype+(e)r

mang-er-i-ons = MANGER+cond+1pl

Phonological level

International Phonetic Alphabet

[aɪ pʰi: eɪ]

Graphemic level

enough, cough, draught, although, brought, through, thorough, hiccough

The 5 Challenges of NLP

1. Productivity
2. Ambiguous
3. Variability
4. Diversity
5. Sparsity

Productivity

Definition

"property of the language-system which enables native speakers to construct and understand an indefinitely large number of utterances, including utterances that they have never previously encountered." (Lyons, 1977)

➔ **New words, senses, structure** are **introduced in languages all the time**

Examples: *staycation* and *social distance* were added to the Oxford Dictionary in 2021

Ambiguous

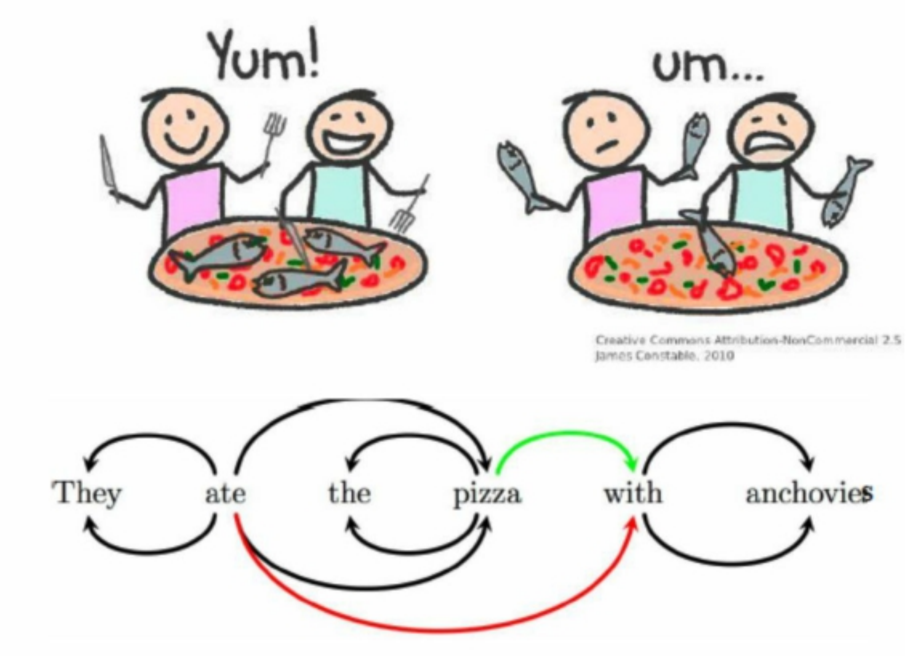
Most linguistic observations (speech, text) are open to **several interpretations**

We (Humans) disambiguate - i.e. **find the correct interpretation** - using all kind of signals (linguistic and extra linguistic)

Ambiguity can appear at all levels (phonology, graphemics, morphology, syntax, semantics)

Ambiguous

Syntactic Ambiguity



Creative Commons Attribution-NonCommercial 2.5
James Constable, 2010

Ambiguous

Semantic Ambiguity

- Polysemy : *e.g.* ***set* , *arm, head***

  Head of New-Zealand is a woman

- Name Entity : *e.g.* ***Michael Jordan***

  Michael Jordan is a professor at Berkeley

- Object/Color : *e.g.* ***cherry***

  Your cherry coat

Ambiguous

Pragmatic Ambiguity

*Two Soviet ships collide, **one dies***

*Dealers will hear **car talk** at noon*

Ambiguous

Disambiguating can requires Discourse Knowledge

Where can I find **a vegetarian restaurant** in Paris

Here is a list of restaurant in Paris: ….

Give me the top ranked ones , in the 14th arrondissement

Here are the top ranked restaurant in the 14th arrondissement in Paris

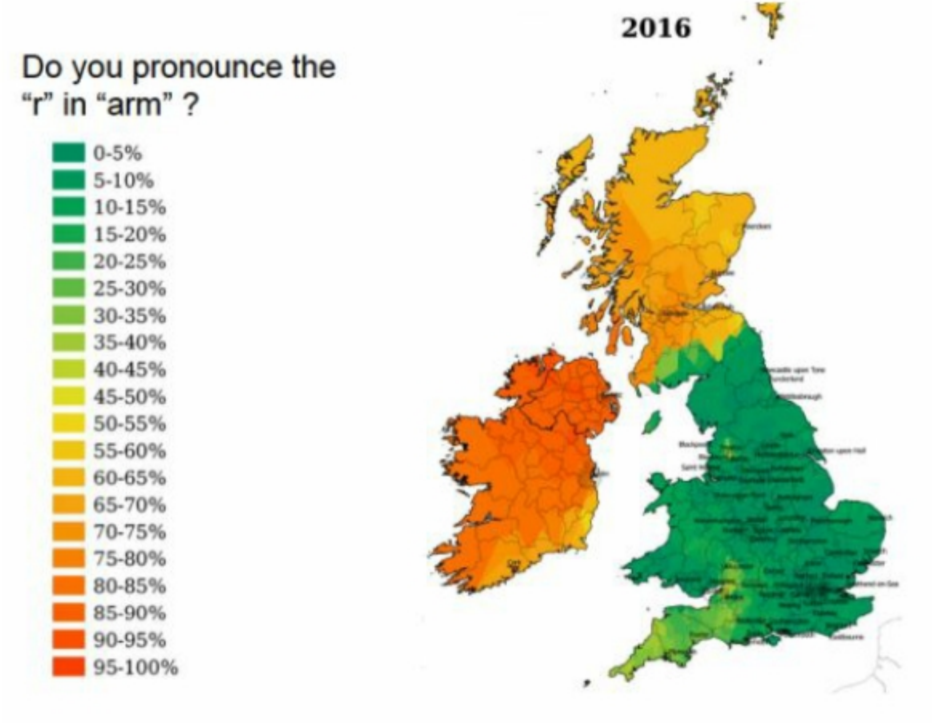
How far is the closest **one** from my current location?

Variation

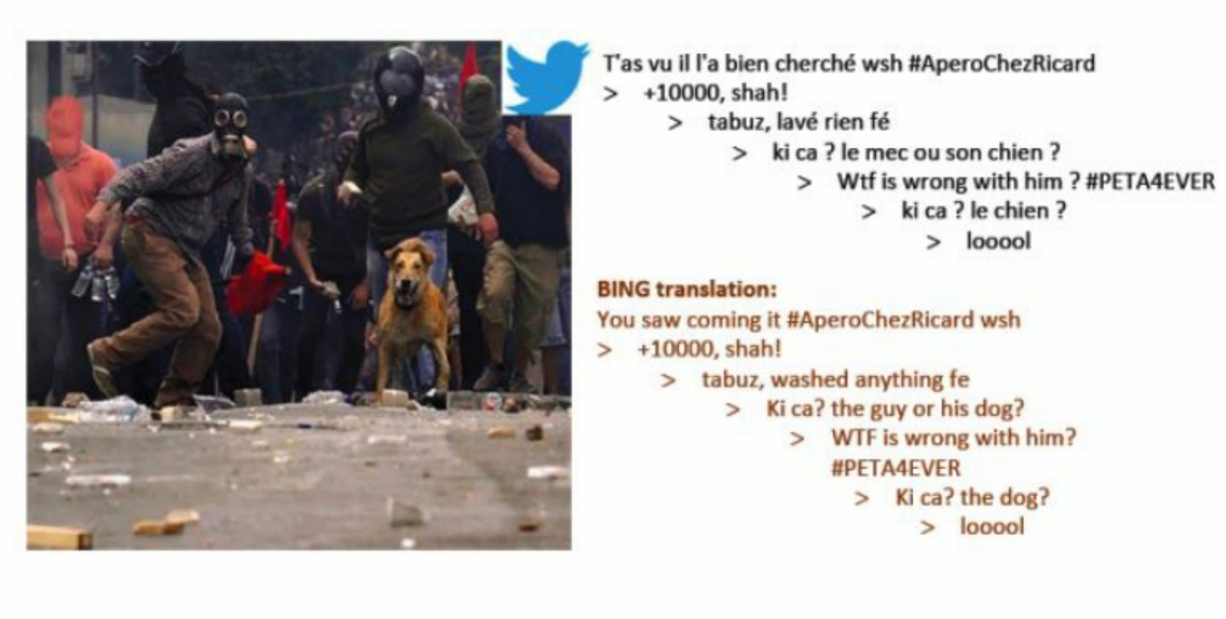
Language Varies at all levels

- Phonetic (accent)
- Morphological, Lexical (spelling)
- Syntactic
- Semantic

Phonetic Variation



Spelling and Syntactic Variation



T'as vu il l'a bien cherché wsh #AperoChezRicard
> +10000, shah!
> > tabuz, lavé rien fé
> > > ki ca ? le mec ou son chien ?
> > > > Wtf is wrong with him ? #PETA4EVER
> > > > > ki ca ? le chien ?
> > > > > > looool

BING translation:

You saw coming it #AperoChezRicard wsh
> +10000, shah!
> > tabuz, washed anything fe
> > > Ki ca? the guy or his dog?
> > > > WTF is wrong with him? #PETA4EVER
> > > > > Ki ca? the dog?
> > > > > > looool

Variation Determiners

- Who is talking?
- To Whom?
- Where? *Work, Home, Restaurant*
- When? *19th century, 2008, 2022*…
- About what? *Specialised domain, the Weather,*…

Essentially, the Variability of a language depends on:

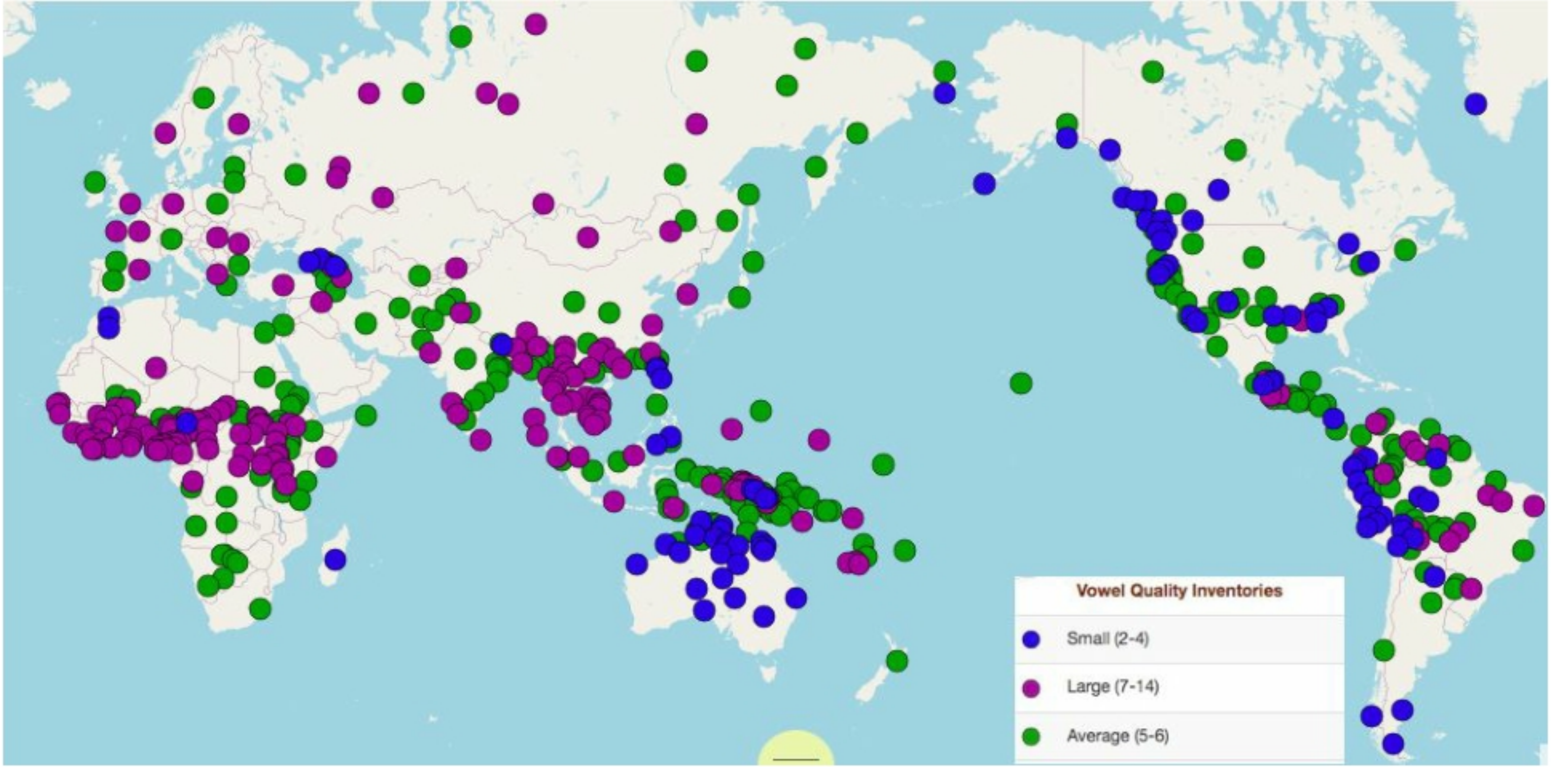
- Social Context
- Geography
- Sociology
- Date
- Topic

Diversity

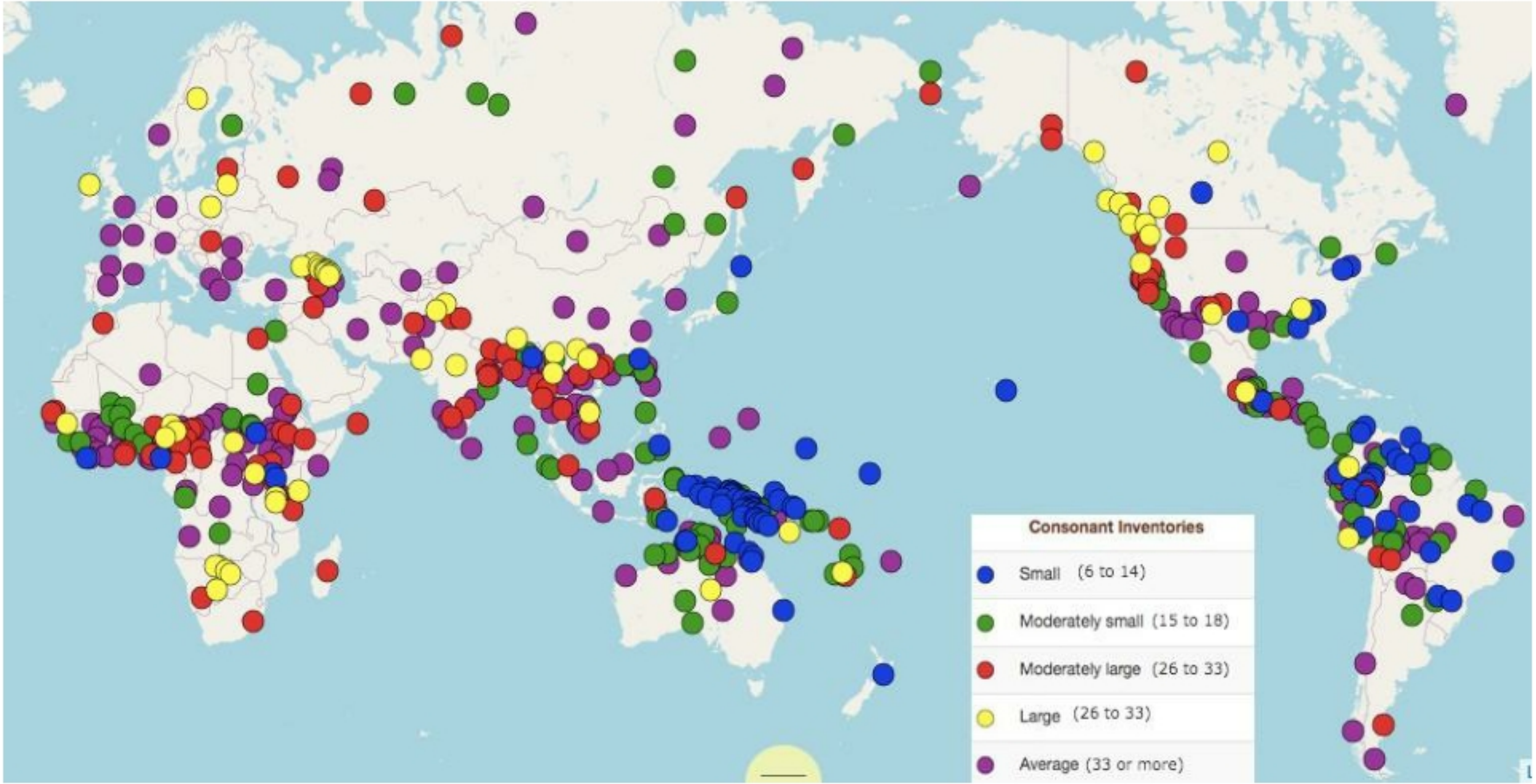
- About **7000 languages** spoken in the world

- About **60%** are found in the **written form** (cf. Omniglot)

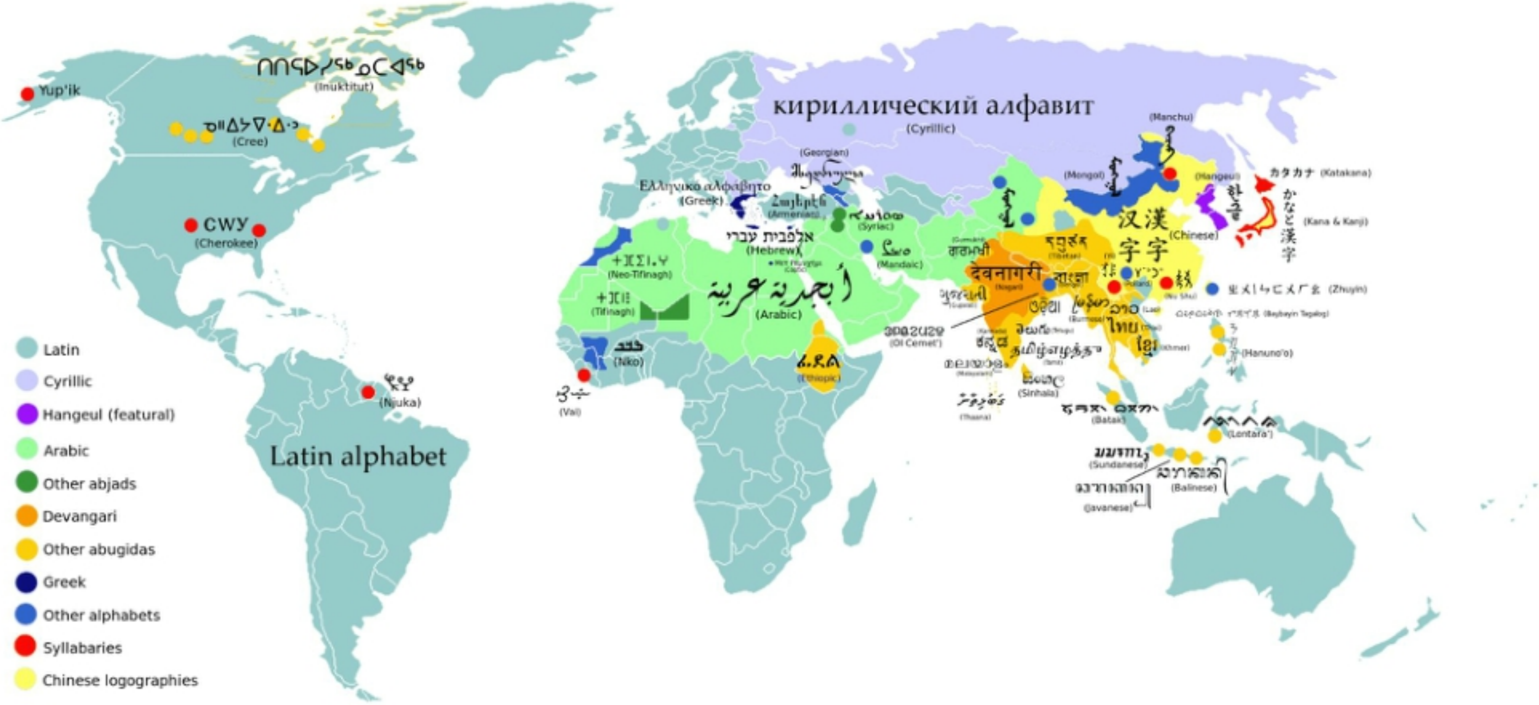
Phonologic Diversity



Phonologic Diversity



Graphemic Diversity



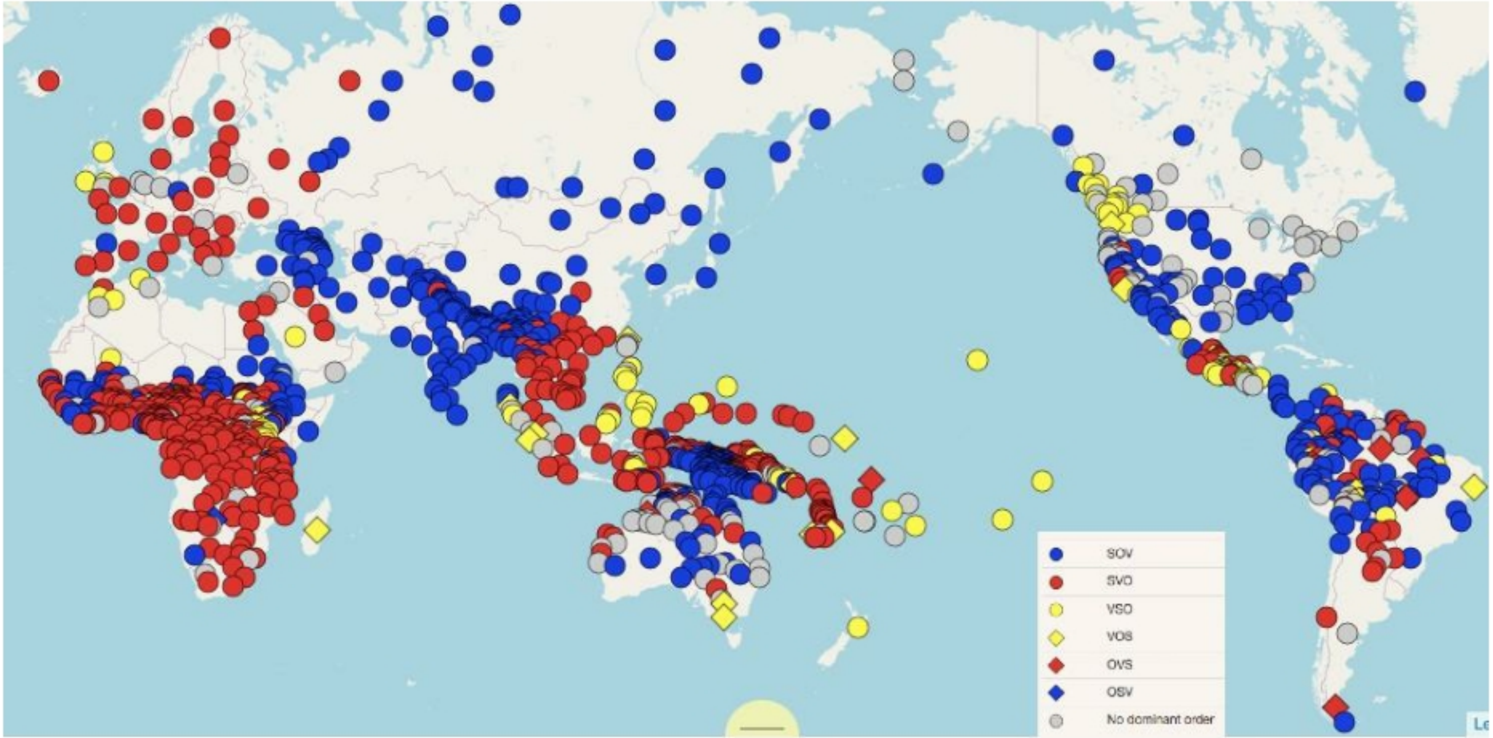
wikipedia

Syntactic Diversity

A key characteristics of the syntax of a given language is **the word order**

- **Word order differs** across languages
- Word order degree of freedom also differs across languages
- We characterize word orders with: **Subject (S) Verb (V) Object (O) order**

Syntactic Diversity



(Dyer et. al 2013)

Word Order Freedom And Morphology

- Word orders freedom and morphology are usually related
- **The more freedom in word orders**

  → the less information is conveyed by word positions

  → the more information is carried by each word

  → the richer the morphology

English *cats eat* ***mice***

Russian (O: -ей) ***Кошки едят мышей***

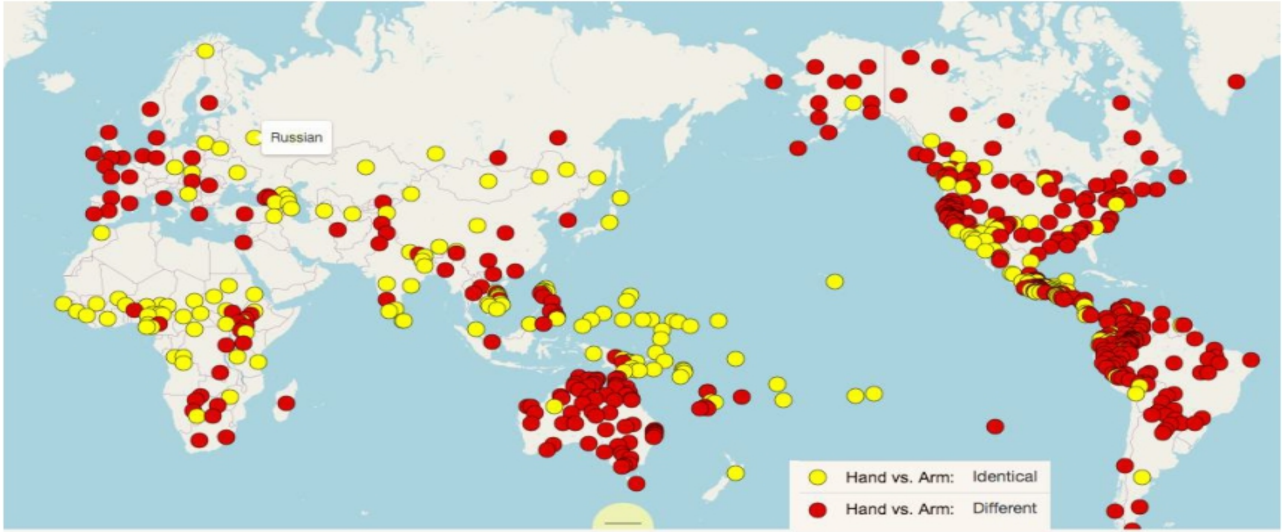
Мышей едят кошки

Едят кошки мышей.

Едят мышей кошки.

Semantic Diversity

- Words partition the semantic space
- This partition is very diverse across language



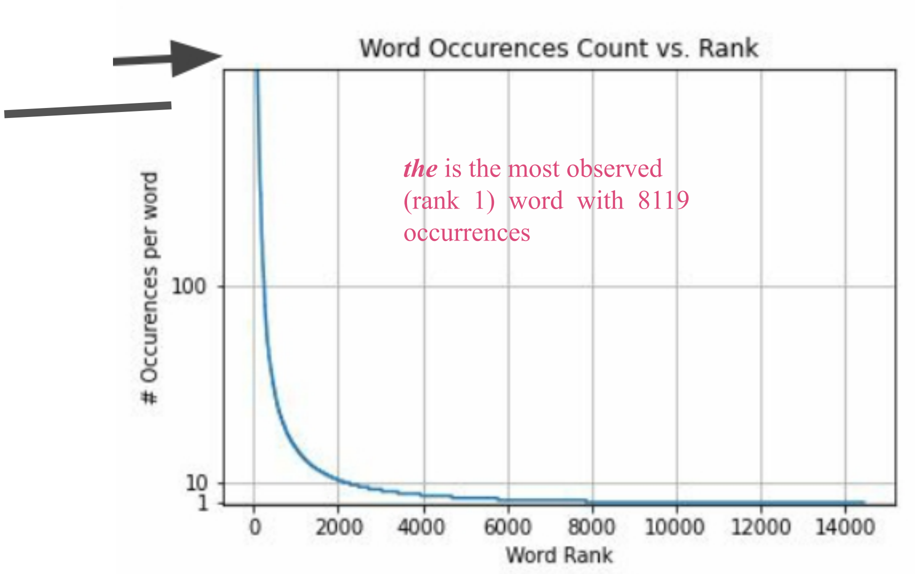
Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

Question: If we plot the number of occurrences of each word vs. the rank, what will we observe?

Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

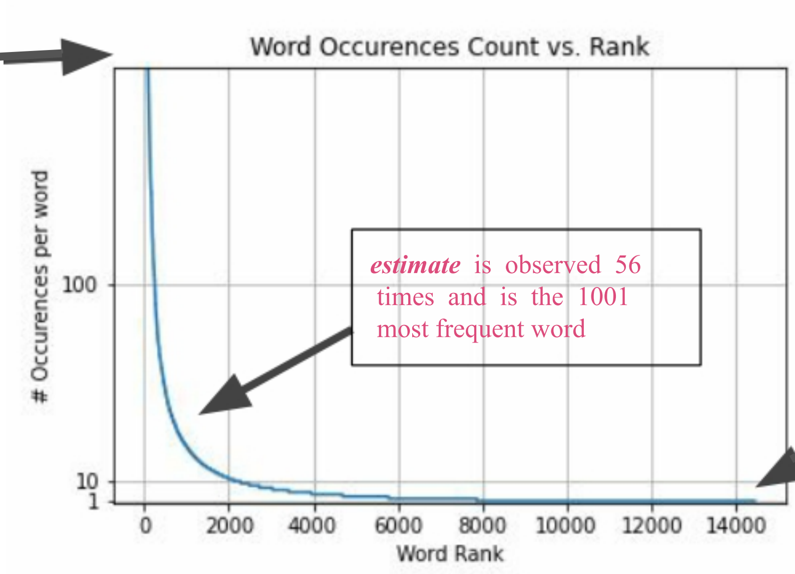


the is the most observed (rank 1) word with 8119 occurrences

Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

the is the most observed (rank 1) word with 8119 occurrences



estimate is observed 56 times and is the 1001 most frequent word

About 6000 Words are observed only 1 time in the dataset (e.g. ***stakeholders, pending, score…)***

Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

➔ In a large enough corpus, word distributions follows ***a Zipf Law ie:***

f_w **frequence of entity w**

k **frequency rank of w**

$$f_w(k) \propto \frac{1}{k^\alpha}$$

Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

➔ In a large enough corpus, word distributions follows ***a Zipf Law ie:***

f_w **frequence of entity w**

k **frequency rank of w**

$$f_w(k) \propto \frac{1}{k^\alpha}$$

- Zipf law is a Power relation between the rank and frequency

  *The most frequent entities are **much more frequent** than the less frequent ones*

- Under a Zipf law, log(fw) and log(k) are linearly related

Statistical Description of Language

Zipf Distributions are observed not only for words but with many other units of language (sounds, syntactic structure, name entities…)

Consequence

➡ A large number of units are observed in language with very low frequency i.e. **Sparsity**

➡ Very challenging for NLP

What is Natural Language Processing?

In a nutshell, NLP consists in handling the complexities of natural languages "to do something"

- Raw Text / Speech → Structured Information
- Raw Text / Speech → (Controlled) Text/Speech

In this course we will focus **on textual data**

Framework

We assume:

- A **token** is the basic unit of discrete data, defined to be an item from a vocabulary indexed by 1, ..., V.
- A **document** is a sequence of N words denoted by $d = (w_1, w_2, \ldots, w_N)$, where w_n is the N-th word in the sequence.
- A **corpus** is a collection of M documents denoted by $D = (d_1, d_2, \ldots, d_M)$

Example: *Wikipedia, All the articles of the NYT in 2021*…

Token

With regard to our end task, a token can be:

- A word
- A sub-word: *e.g. a sequence of 3 characters*
- A character
- An sequence of characters (sometimes a word, sometimes several words, sometimes a sub-word…)

Document

A Document can be:

- A Sentence
- A Paragraph
- A sequence of characters

Text Segmentation

Definition: Text Segmentation is the process of splitting raw text (i.e. list of characters) into **units of interest** .

Two level of segmentation (usually) required :

- Split raw text into **modeling units** (ex: sentence, paragraph, 1000 characters, web-page...)
- Split modeling units into sequence of **basic units** (referred as tokens) (e.g: words, word-pieces, characters, ...)

Two distinct approaches:

- **Linguistically informed** e.g. word, sentence segmentation...
- **Statistically informed** e.g. frequent sub-words (word pieces, sentence pieces...)

Tokenization

Definition: Tokenization consists in *segmenting* raw textual data into tokens:

Tokenization

Definition: Tokenization consists in *segmenting* raw textual data into tokens:

Can be framed as a character level task

input: *une industrie métallurgique existait.*
output: IIIEIIIIIIIIIIIIIEIIIIIIIIIIIIIIIIIIIIEIIIIIIIIIIIEE

- **Easy task** for most languages and domains
- Can be **very complex in some cases** (Chinese, Social Media...)

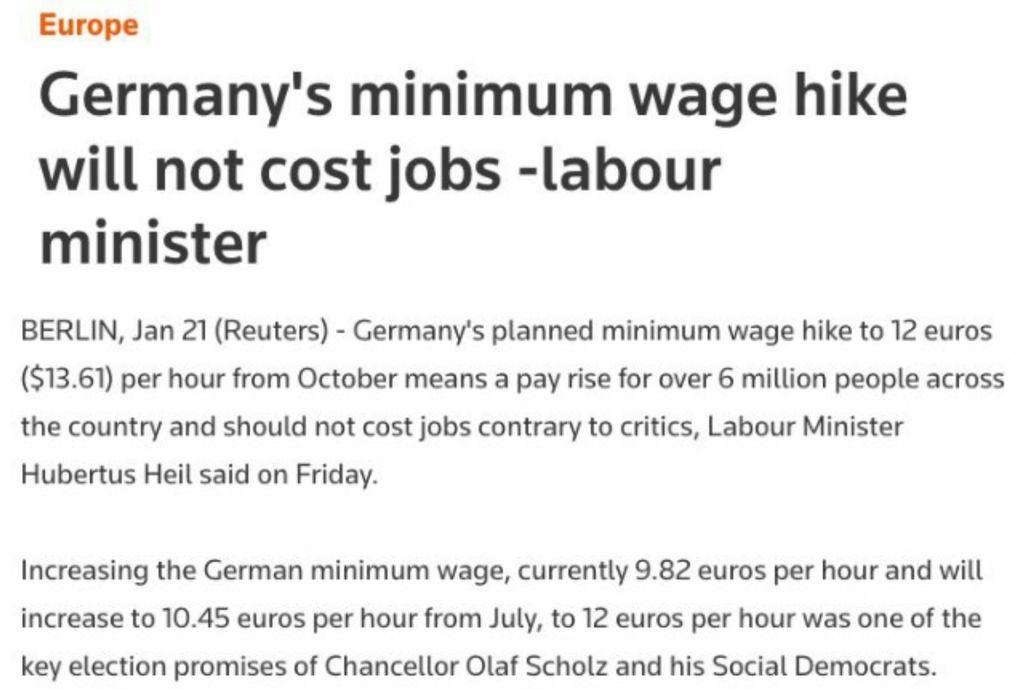
NLP Tasks: Modeling Framework

Let (X, Y) a pair of r.v. X may characterize tokens, sentences or documents. Modeling an NLP task consists in estimating the conditional probability $X \mid Y$ to predict Y with X.

Tasks Taxonomy

- If Y is a single label and X a sequence of tokens (e.g. a sentence): Sequence Classification
- If we have one label per token: Sequence Labelling
- If Y is a sequence of tokens: Sequence Prediction
- If Y is a graph, a tree or a complex structured output: Structure Prediction

Document Classification

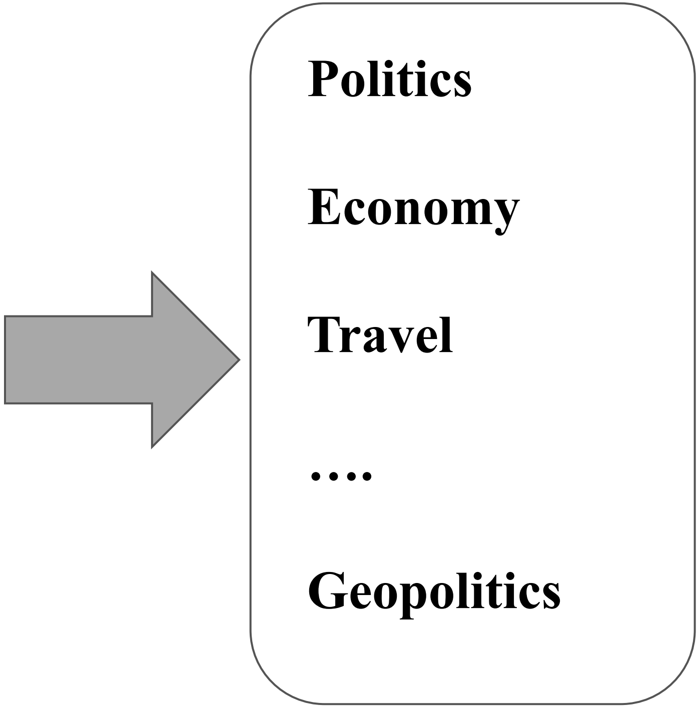


Europe

Germany's minimum wage hike will not cost jobs -labour minister

BERLIN, Jan 21 (Reuters) - Germany's planned minimum wage hike to 12 euros ($13.61) per hour from October means a pay rise for over 6 million people across the country and should not cost jobs contrary to critics, Labour Minister Hubertus Heil said on Friday.

Increasing the German minimum wage, currently 9.82 euros per hour and will increase to 10.45 euros per hour from July, to 12 euros per hour was one of the key election promises of Chancellor Olaf Scholz and his Social Democrats.



Politics

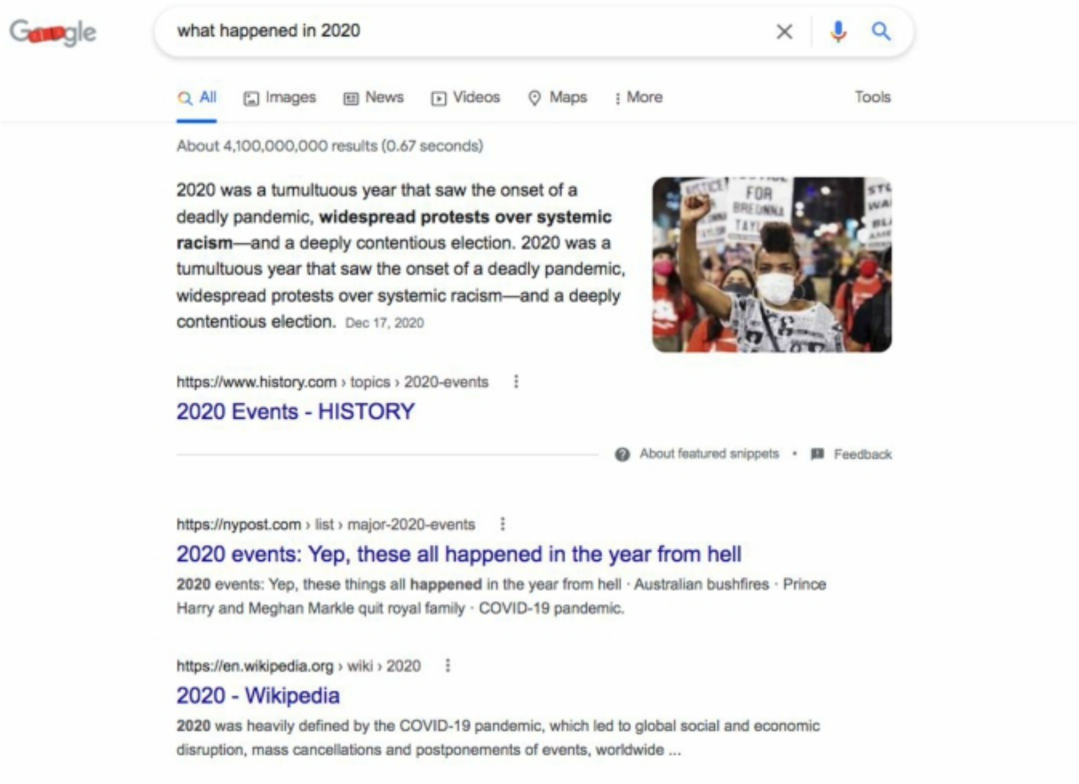
Economy

Travel

….

Geopolitics

Document Ranking (Retriever)



About 4,100,000,000 results (0.67 seconds)

2020 was a tumultuous year that saw the onset of a deadly pandemic, **widespread protests over systemic racism**—and a deeply contentious election. 2020 was a tumultuous year that saw the onset of a deadly pandemic, widespread protests over systemic racism—and a deeply contentious election. Dec 17, 2020

https://www.history.com › topics › 2020-events

2020 Events - HISTORY

About featured snippets • Feedback

https://nypost.com › list › major-2020-events

2020 events: Yep, these all happened in the year from hell

2020 events: Yep, these things all **happened** in the year from hell · Australian bushfires · Prince Harry and Meghan Markle quit royal family · COVID-19 pandemic.

https://en.wikipedia.org › wiki › 2020

2020 - Wikipedia

2020 was heavily defined by the COVID-19 pandemic, which led to global social and economic disruption, mass cancellations and postponements of events, worldwide ...

NLP Tasks: Part-of-Speech Tagging

POS Tagging: Find the **grammatical category** of each word

[My , name, is, Bob, and, I, live, in, NY, !]

NLP Tasks: Part-of-Speech Tagging

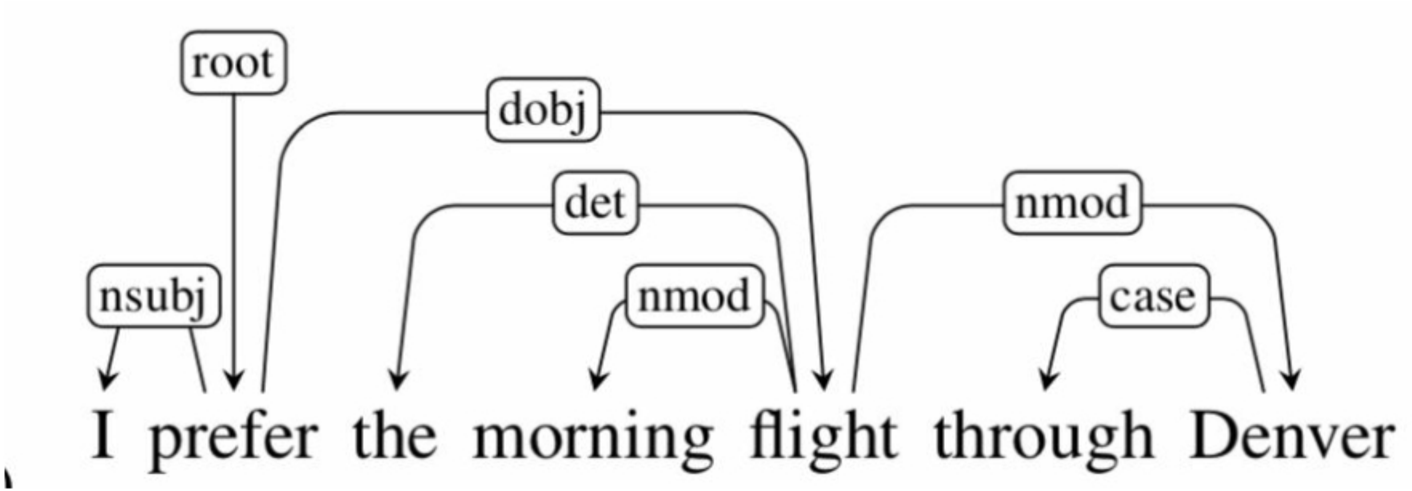
POS Tagging: Find the **grammatical category** of each word

[My , name, is, Bob, and, I, live, in, NY, !]

[PRON , NOUN, VERB, NOUN, CC, PRON, VERB, PREP, NOUN, PUNCT]

Syntactic Parsing

Syntactic Parsing consists in **extracting the syntactic structure** of a sentence. For instance, **Dependency Parsing** (here) predicts an acyclic directed graph (a **tree**)



Slot-Filling / Intent Detection

Intent Detection is a sequence classification task that consists in **classifying the intent of a user** in a pre-defined category.

Slot-Filling is a sequence labelling task that consists in identifying **specific parameters in a user request** .

Can you please play Hello from Adele ?

Intent: ***play_music***

Slots: [*Can, you, please, play, Hello, from, Adele, ?]*

*[O , O , O , O , **SONG** , O , ARTIST , O]*

Semantic Role Labelling (SLR)

SRL is the task of finding the **semantic role** of each predicate in a sentence.

Given a sentence, SRL predicts: *who did what to whom, when,* ***where, why, how***

NLP Tasks: Name Entity Recognition

NER: Find the **Name-Entities** in a sentence

[My , name, is, Bob, and, I, live, in, NY, !]

[O , O, O, ***PERSON*** *, O, O, O, O,* ***LOCATION*** *, O]*

Machine Translation

INPUT: My name is Bob and I live in NY !

OUTPUT: Je m'appelle Bob et je vis à New-York!

Question Answering

INPUT: *How many episodes in season 2 breaking bad?*

OUTPUT: *13*

How do we solve an NLP Problem?

Each NLP Problem is unique

➔ **No Universal Method to solve them all**

➔ Have a toolkit in mind of methods (symbolic, statistics and deep learning based)

NLP Engineering: Find **what works best** for your given method based on what has been done already in the literature, by colleagues…

NLP Research : Find a **better** way to do (more accurate, cheaper, faster…)

Brief History of NLP

Brief History of NLP

Symbolic

Statistical Learning

Deep Learning

1940-2000

Focus on rule-based systems, formal grammars

Development of linguistic resources (lexicon, ontologies, grammars…)

1990-2010

Statistical learning theory (SVM, Random Forest), Graphical Probabilistic Models (e.g. LDA, HMM)

Development of annotated datasets

2010-Today

Deep Learning Architecture (Transformer)

Transfer Learning in NLP (word2vec, BERT, CamemBERT, GPT)

More compute, larger (raw) dataset, Open Source Deep Learning Libraries

1949: First Machine Translation “Model”

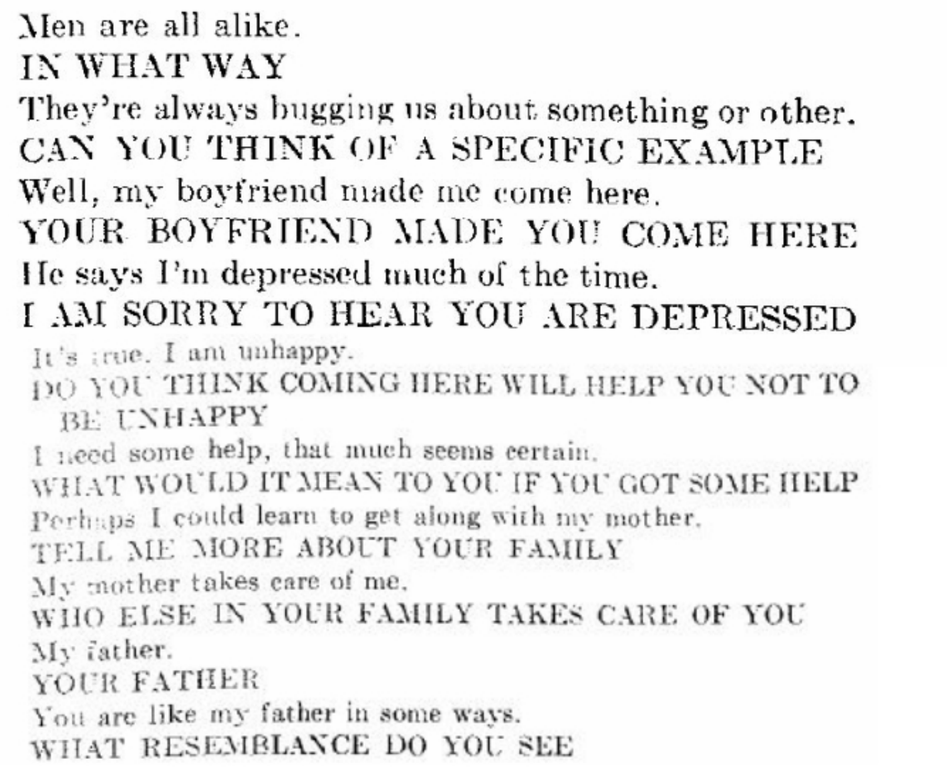
1949 ***Memorandum on Translation*****,** Warren Weaver
First to propose the idea of using “ *electronic computers* ” to do translation

- Using Shannon’s information theory to frame Machine Translation as a ***cryptographic*** problem

- **Modeling context** to disambiguate between word senses

- “*Going down* ” from each language to ***universal language*** in order to translate

1964: ELIZA First Conversational Bot



Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED
It's true. I am unhappy.
DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
I need some help, that much seems certain.
WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
Perhaps I could learn to get along with my mother.
TELL ME MORE ABOUT YOUR FAMILY
My mother takes care of me.
WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU
My father.
YOUR FATHER
You are like my father in some ways.
WHAT RESEMBLANCE DO YOU SEE

Natural Language Processing WorkFlow

Assume we have a **Research, Engineering, Product Problem**

1. Define a **NLP System** to solve it
   Split into **modules** , each one performing a **task**

2. Define **Evaluation Metric(s) for your system and submodules**

3. **Collect Data** to build/train your models

4. Build **Baseline Models** (i.e. most simple model you can think of that have a non trivial performance metric)

5. Build **Better Models** using symbolic/statistical/DL methods

Session Summary

- The **6 Levels of analysis in linguistics**
- The **5 Challenges** of Natural Language Processing
- The NLP Modelling Framework
- **A Brief History** of Natural Language Processing: **the 3 Era of NLP**

Bibliography and Acknowledgment

- [Benoit Sagot 2022], Algorithms for speech and natural language processing, MVA course Material
- [Warren Weaver, 1949] Memorandum on Translation
- [Weizenbaum, 1966] Eliza
- [Dryer, Matthew S. & Haspelmath, Martin (eds.) The WALS]

All these class have been taken from https://nlp-ensae.github.io/materials/ and is taken from Benjamin Muller