Natural Language Processing

Introduction to NLP

Junyeong Kim
Dept. of AI, Chung-Ang University

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** with language
- Why NLP?
  - **Access Knowledge** (search engine, recommender system, and so on)
  - **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
- Data growing much faster then ever …
  - 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
  - 4.7 billion internet uses (~ 59% of world population)
  - 4.2 billion social media uses (~ 54 % of world population)

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

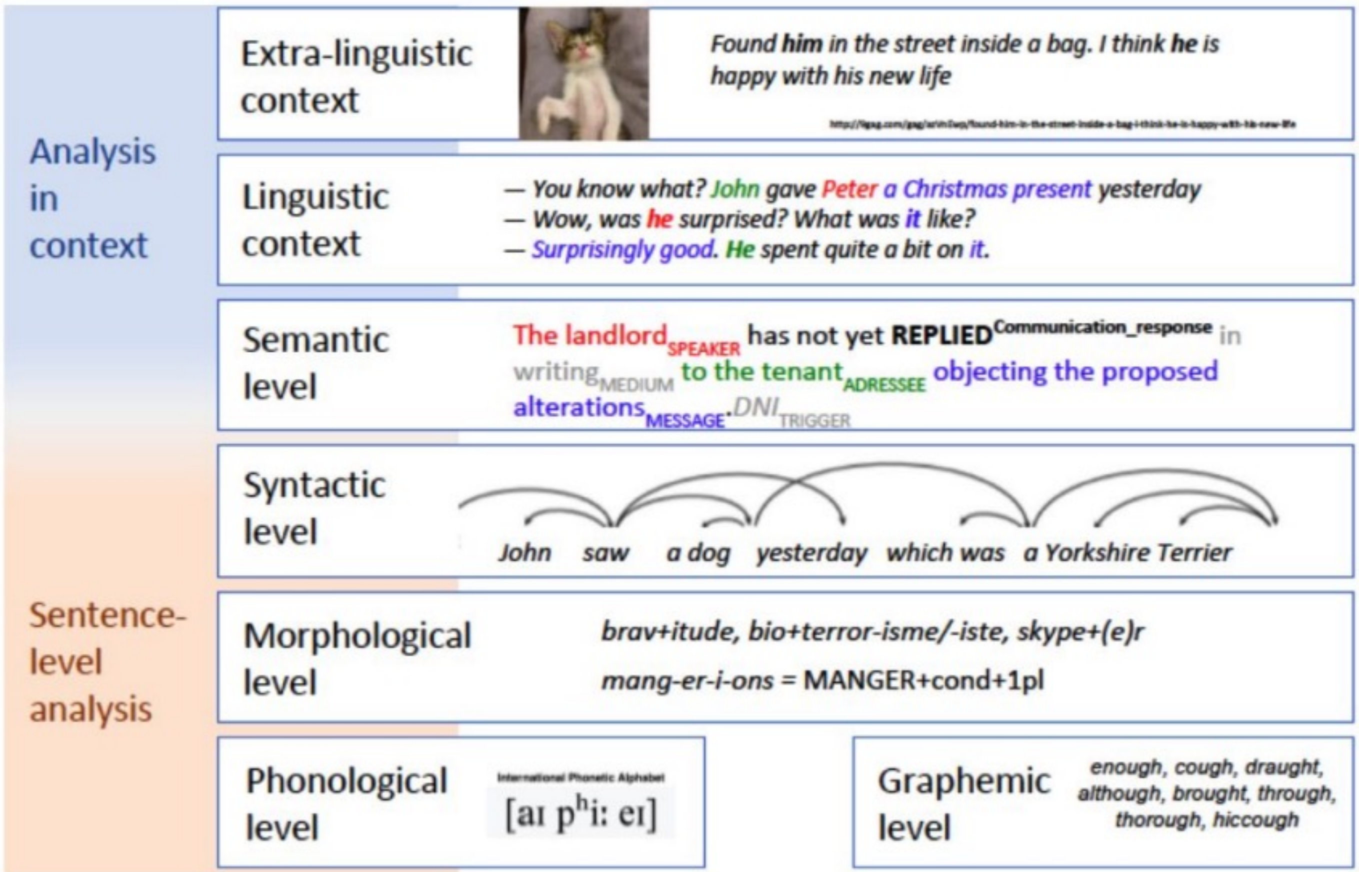
Why Language is hard to model?

- Let's look at the 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



The 5 Challenges of NLP

- Productivity
- Ambiguous
- Variability
- Diversity
- 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 disambiguate – i.e., find the correct interpretation – using all kind of signals (linguistic and extra linguistic)
- Ambiguity can appear at all levels of language (phonology, graphemics, morphology, syntax, semantics)

Ambiguous

- Syntactic Ambiguity

Getty Images

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
  - The process by which addressees identify speakers' intentions, beliefs and attitudes in a given context
  - Example: "It's hot in here! Can you crack a window?" – We can infer speaker wants the window to be opened, and does not want the window to be physically damaged

Ambiguous

- Disambiguating can requires Discourse Knowledge

Q: Where can I find a vegetarian restaurant in Seoul?

A: Here is a list of restaurant in Seoul …

Q: Give me the top ranked ones, in Dongjak-gu.

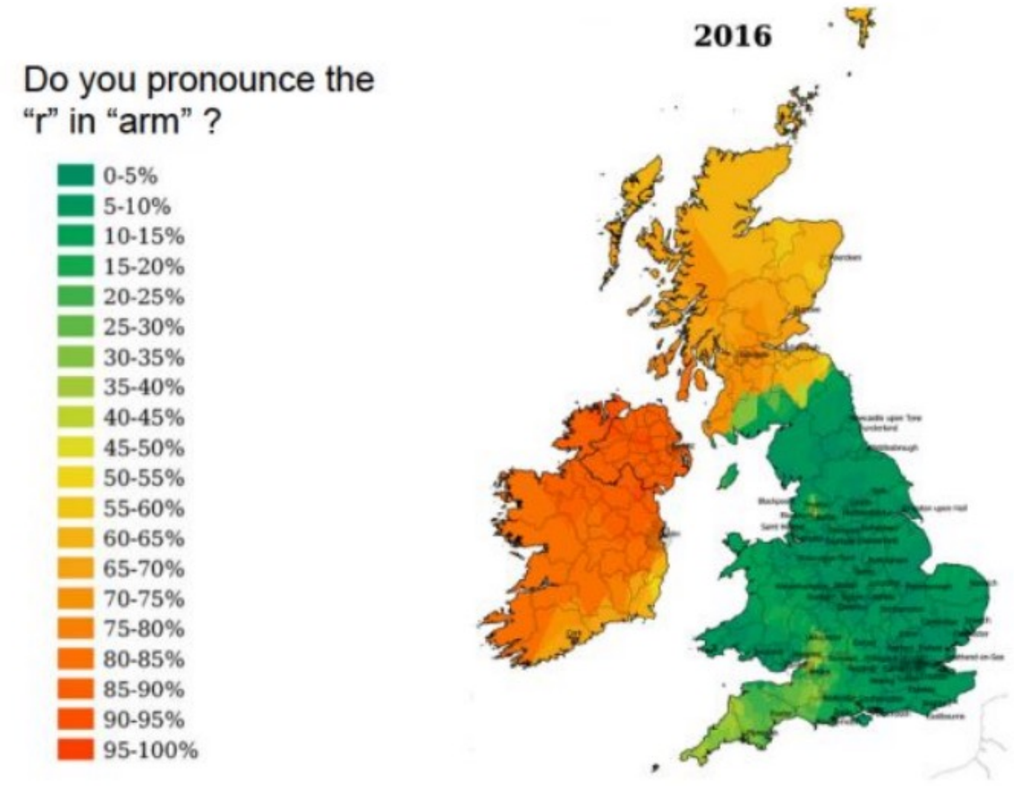
A: Here are the top ranked restaurant in Dongjak-gu

Q: 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, 2024…
- About what? Specialized domain, the Weather, …

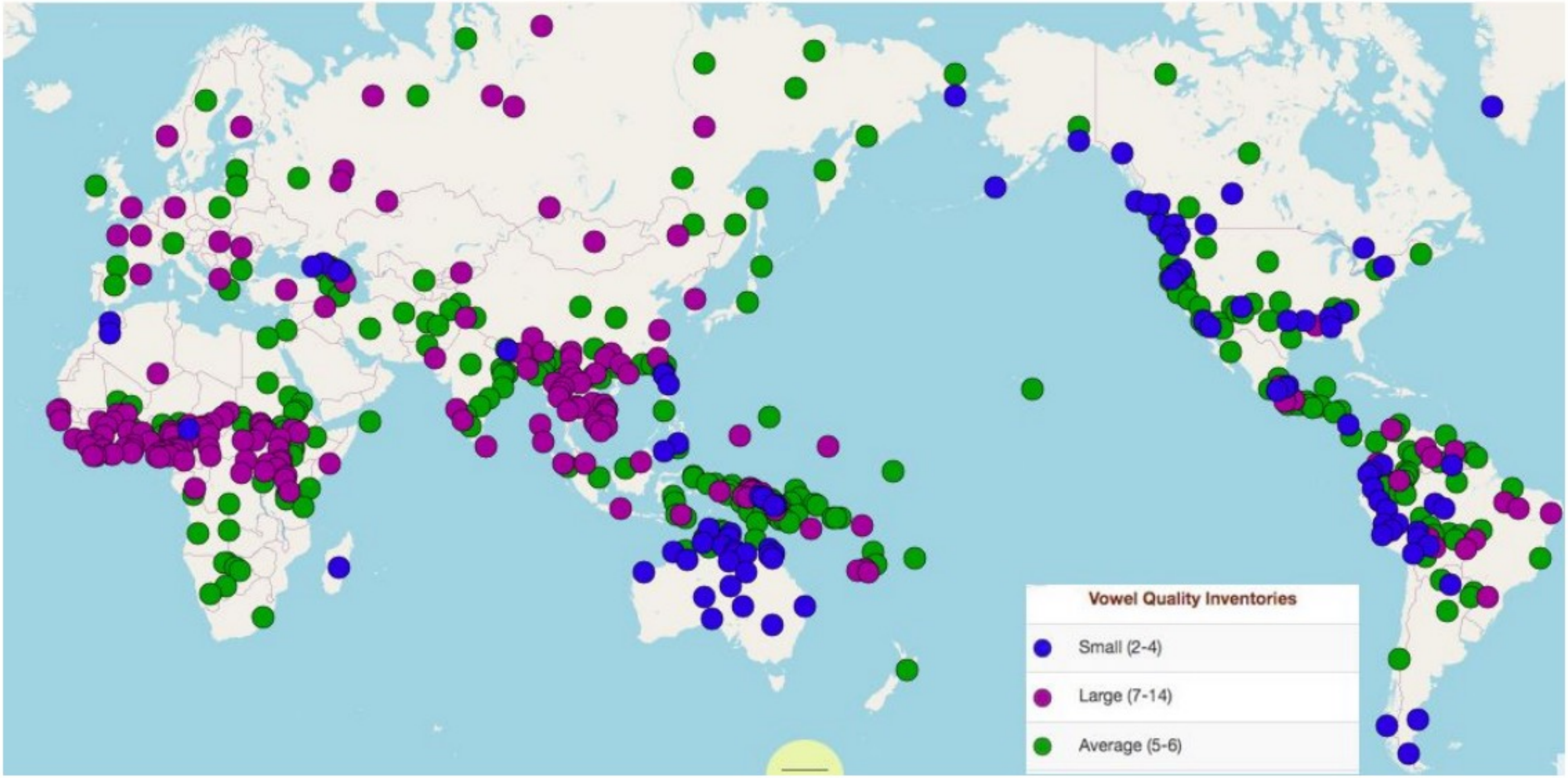
○ Essentially, the Variability of a language depends on:

- Social Context
- Geography
- Sociology
- Date
- Topic

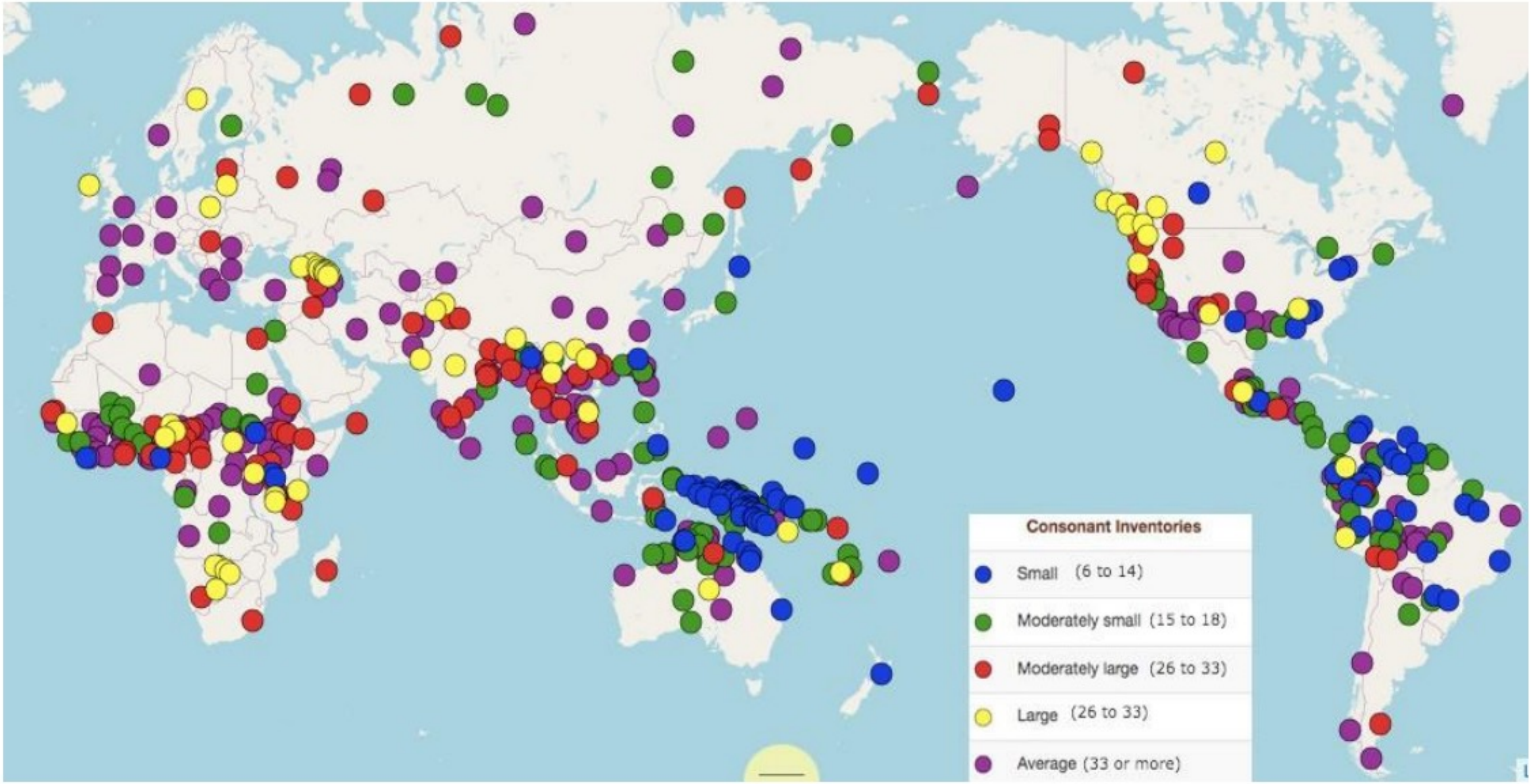
Diversity

- About 7000 languages spoken in the world
- About 60% are found in the written form

Phonologic Diversity



Phonologic Diversity

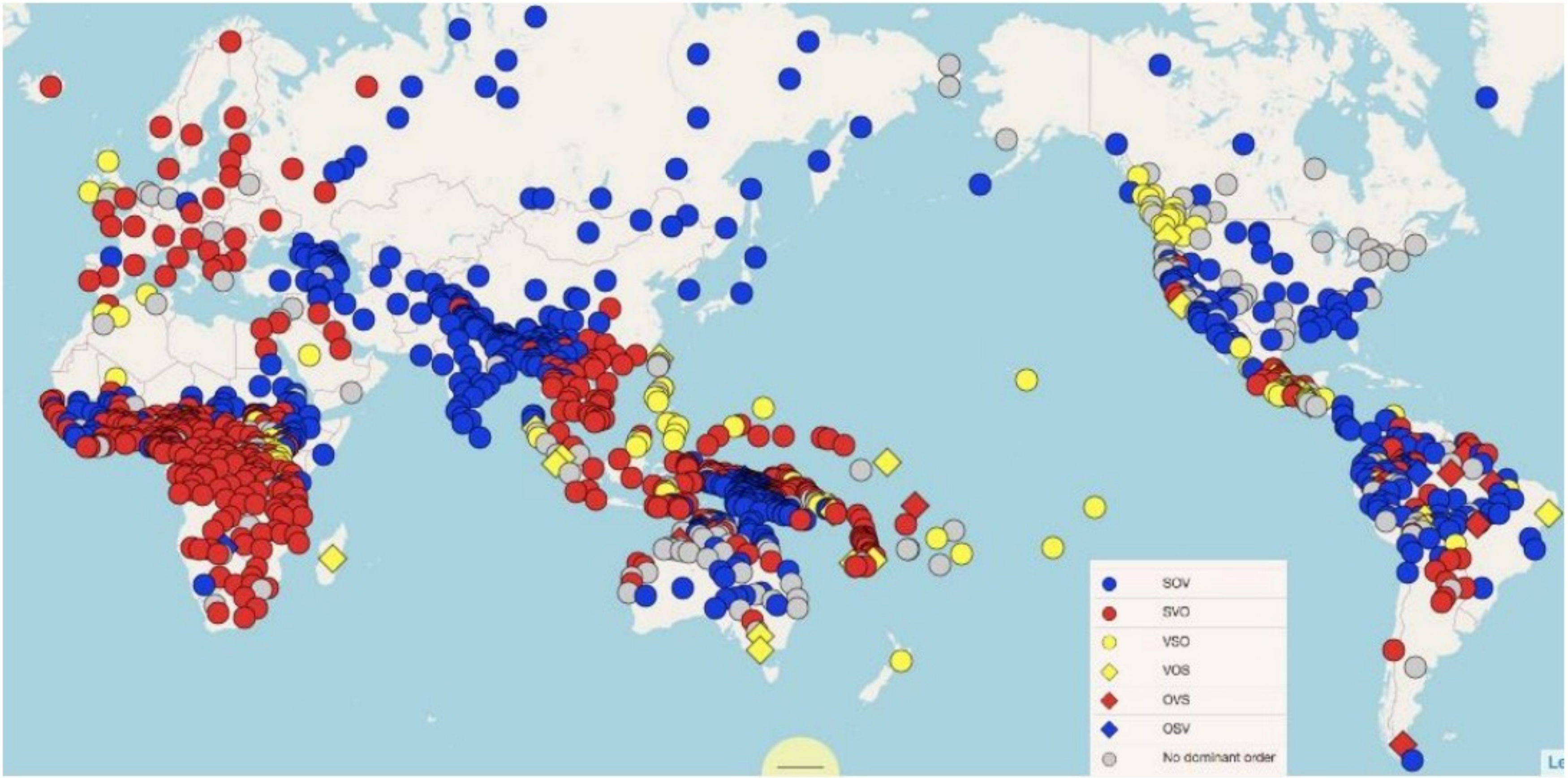


Graphemic Diversity

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



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**

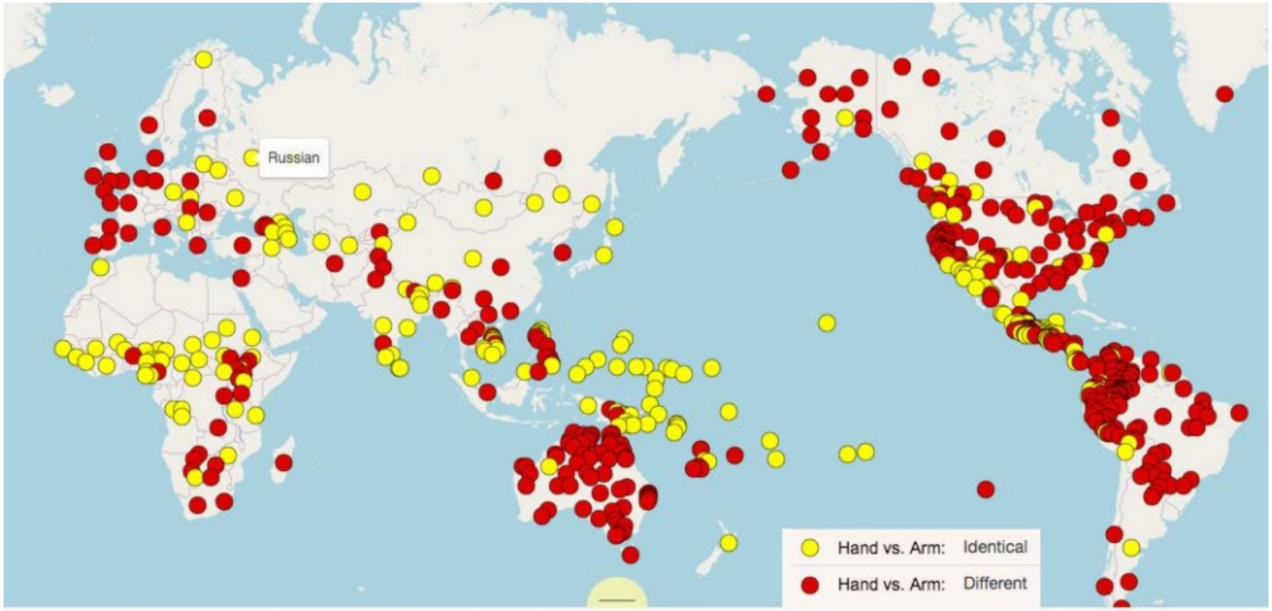
English *cats eat mice*

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



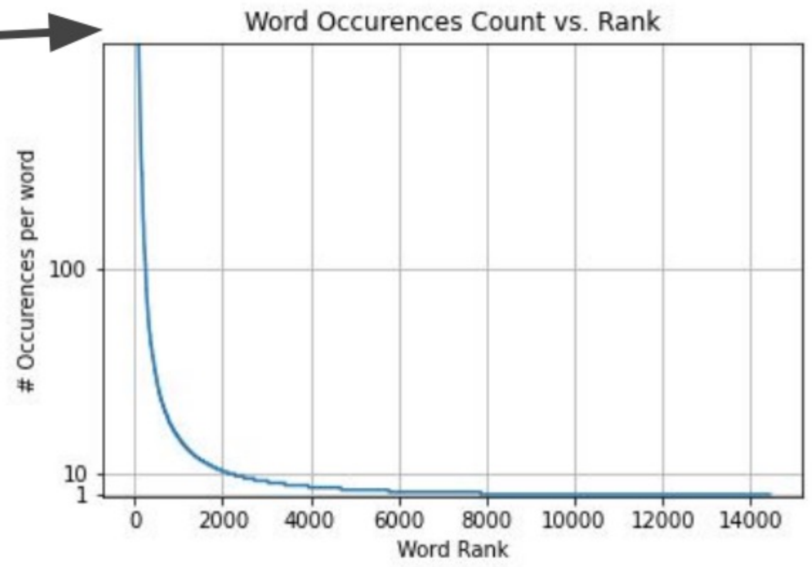
Statistical Description of a Corpus

- 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

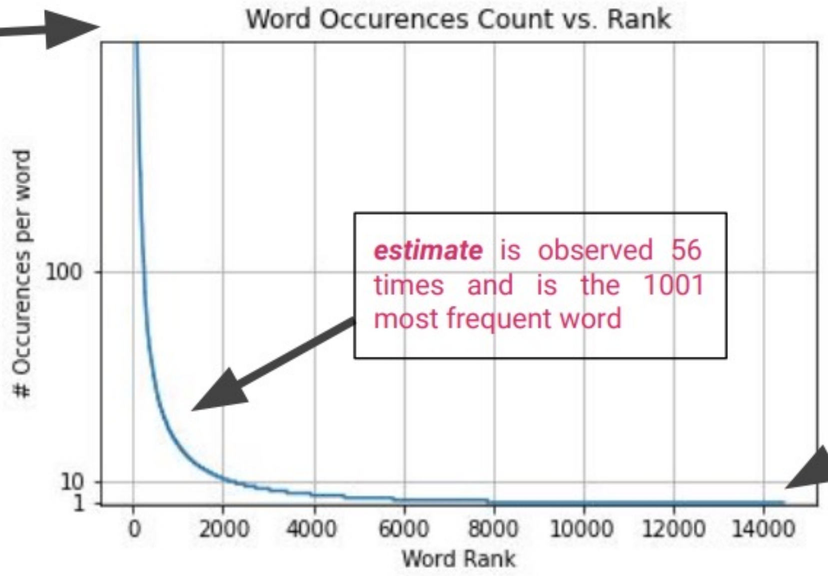
the is the most observed (rank 1) word with 8119 occurrences



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 follow a ***Zipf Law*** :

f_w frequency of entity w
k frequency rank of entity w

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

Statistical Description of a Corpus

- We describe statistically a corpus of 800 scientific articles

  → In a large enough corpus, word distributions follow a ***Zipf Law*** :

- We describe statistically a corpus of 800 scientific articles
  → In a large enough corpus, word distributions follow a ***Zipf Law*** :
  - 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, $\log(f_w)$ and $\log(k)$ are linearly related

  - 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, $\log(f_w)$ 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 of Language

  → Sparsity of Language makes NLP very challenging

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

Notation

- 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 $d = (w_1, w_2, \dots, w_N)$, where w_N is the the N-th word in the sequence.
  - A corpus is a collection of M documents denoted by $D = (d_1, d_2, \dots, d_M)$

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 (e.g., sentence, paragraph, 1000 characters, web-page, …)
  - Split modeling units into sequence of basic units (referred to 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:

- Can be framed as a character level task
  - Input: This is Natural Language Processing Lecture
  - Output: This / is / Natural / Language / Processing / Lecture

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

NLP Task : Modeling Framework

- Let (X, Y) a pair of random variable. X may characterize tokens or documents.
- Modeling an NLP task consists in estimating the conditional probability Y|X in order to predict Y using X.

→ P(Y|X)

- Tasks Taxonomy
  - If Y is a single label and X is 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

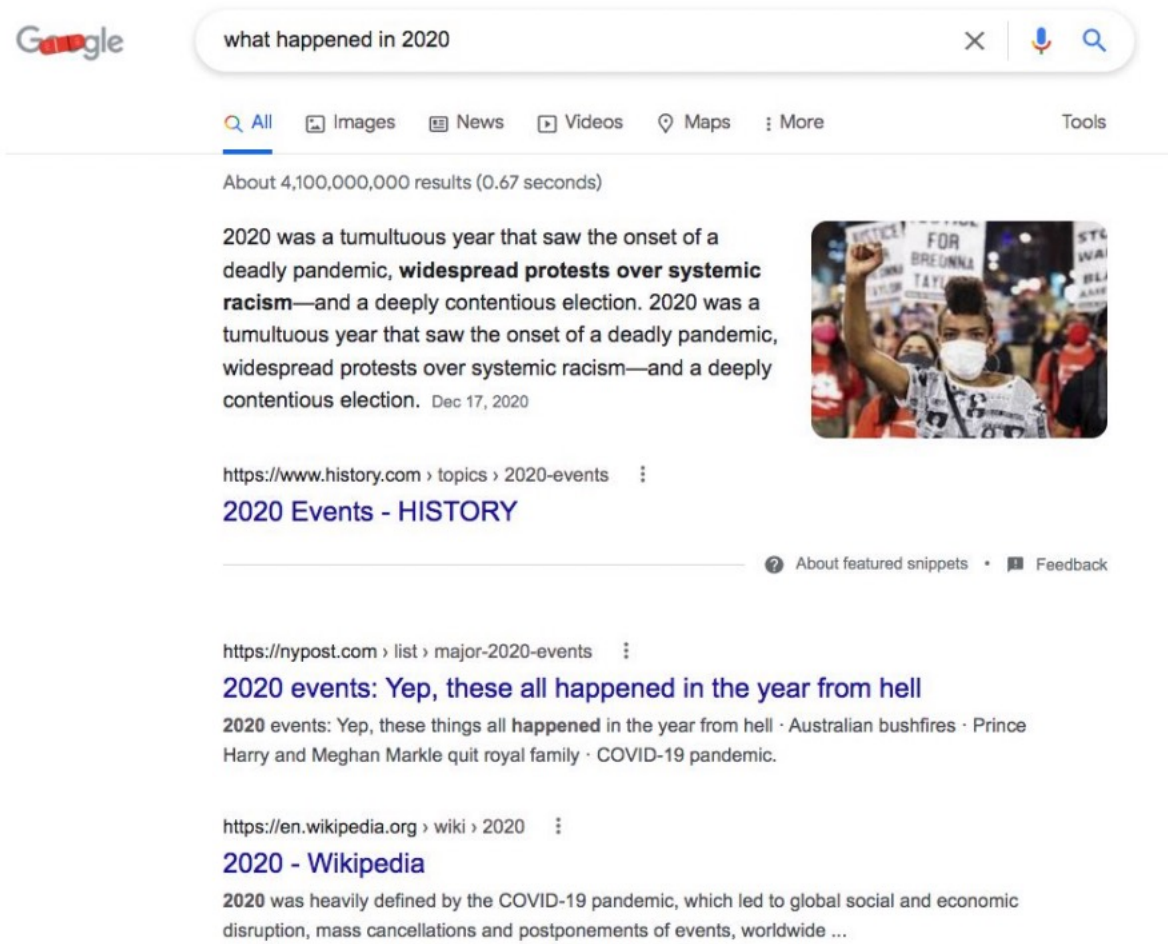
Economy

Travel

....

Geopolitics

Document Ranking (Retriever)



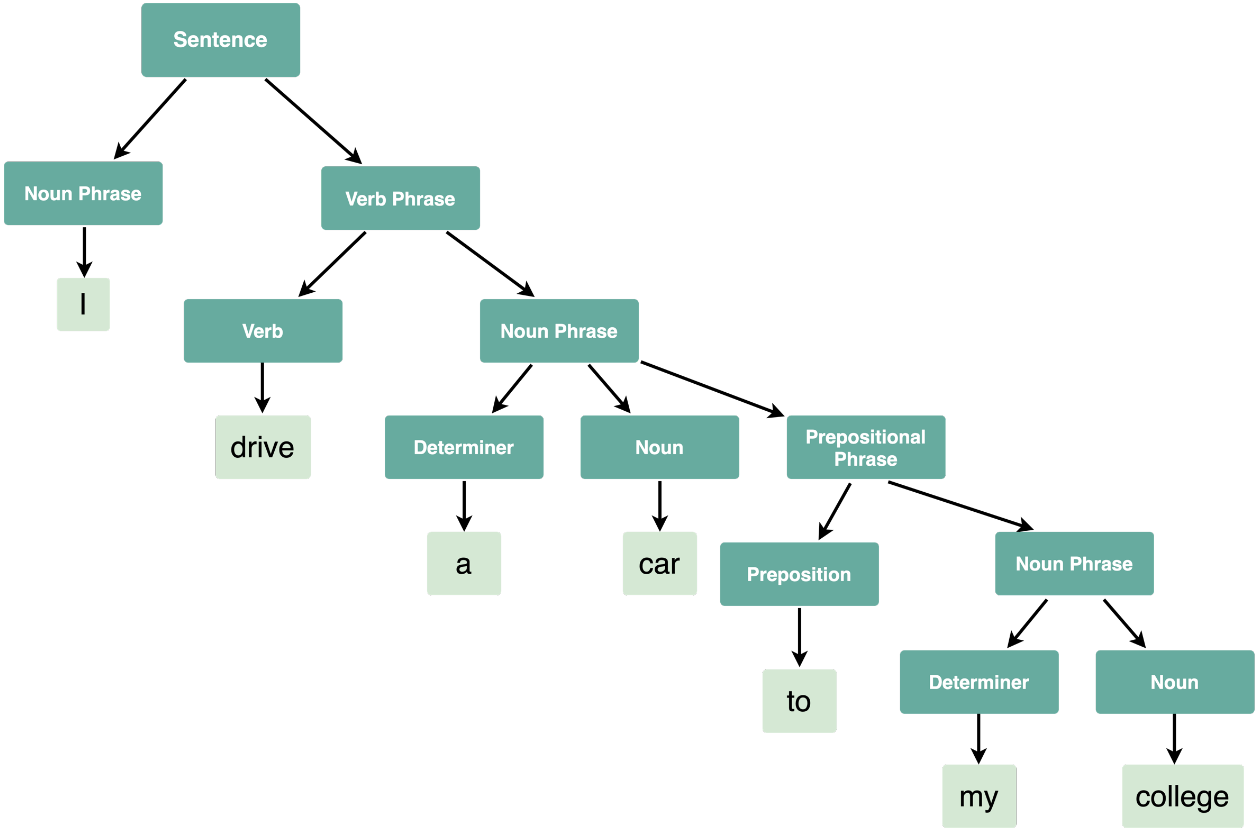
NLP Task: Part-of-Speech Tagging

- 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 predicts an acyclic directed graph (i.e., 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 (SRL)

- 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

Named 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:** 내 이름은 밥이고 나는 뉴욕에 살아!

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

Symbolic	Statistical Learning	Deep Learning
1940-2000	**1990-2010**	**2010-Today**
Focus on rule-based systems, formal grammars Development of linguistic resources (lexicon, ontologies, grammars…)	Statistical learning theory (SVM, Random Forest), Graphical Probabilistic Models (e.g. LDA, HMM) Development of annotated datasets	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

NOW? Era of ChatGPT

TUKO

BEST DIGITAL NEWS PLATFORM

https://kiswahili.tuko.co.ke/

Malawi yawapoteza mawaziri 2 kutokana na maafa ya COVID-19

TUKO.co.ke imefahamishwa kuwa waziri wa serikali ya mitaa Lingson Belekanyama na mwenzake wa uchukuzi Sidik Mia walifariki dunia ndani ya saa mbili tofauti.

Swahili English

Google Translate

Malawi loses 2 ministers due to COVID-19 disaster

TUKO.co.ke has been informed that local government minister Lingson Belekanyama and his transport counterpart Sidik Mia died within two separate hours.

GPT-3: A first step on the path to universal model

The SEC said, "Musk, your tweets are a blight.
They really could cost you your job,
if you don't stop all this tweeting at night."

Then Musk cried, "Why?
The tweets I wrote are not mean,
I don't use all-caps
and I'm sure that my tweets are clean."

"But your tweets can move markets
and that's why we're sore.
You may be a genius and a billionaire,
but it doesn't give you the right to
be a bore!"

S: I broke the window.
Q: What did I break?
S: I gracefully saved the day.
Q: What did I gracefully save?
S: I gave John flowers.
Q: Who did I give flowers to?
S: I gave her a rose and a guitar.
Q: Who did I give a rose and a guitar to?

How many users have signed up since the start of 2020?

```
SELECT count(id) FROM users
WHERE created_at > '2020-01-01'
```

What is the average number of influencers each user is subscribed to?

```
SELECT avg(count) FROM ( SELECT user_id, count(*)
FROM subscribers GROUP BY user_id )
AS avg_subscriptions_per_user
```

ChatGPT: A recent, intriguing set of capabilities

JO

Convert this python dictionary into a LaTeX table that uses the `booktabs` package, and make sure there are no vertical lines.

```
{"Models" : ["baseline", "ours"],
 "Accuracy": [0.7, 0.8],
 "Time/sentence": ["4s", "6s"]
}
```

Here's an example of how you can convert the Python dictionary into a LaTeX table that uses the **`booktabs`** package and does not include any vertical lines:

Copy code

```
\begin{tabular}{l c c}
\toprule
\textbf{Models} & \textbf{Accuracy} & \textbf{Time/sentence}
\\
\midrule
baseline & 0.7 & 4s \\
ours & 0.8 & 6s \\
\bottomrule
\end{tabular}
```

Models	**Accuracy**	**Time/sentence**
baseline	0.7	4s
ours	0.8	6s

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