POLITECHNIKA POZNAŃSKA
Poznan University of Technology

NAWA POLISH NATIONAL AGENCY FOR ACADEMIC EXCHANGE

STER
PROGRAMME

AI for Text Processing

NLP: Natural Language Processing

- NLP is a discipline which combines:
  - Computer science
  - Computational linguistics
  - Artificial intelligence

- Its primary goal is to analyze interactions between natural languages and computers, in particular, NLP focuses on programming computers in such way that computers can fully exploit and process textual corpora.

The aim of NLP

- The aim of NLP **is** to make computers “understand” the natural language to the degree when they can complete useful tasks, such as summarization, information extraction, or translation.
- The aim of NLP **is not** to understand the language (which is probably not possible)

We argue that the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning

- *Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Proc. of ACL 2020.*

Advances in NLP

- Mostly solved
  - Spam detection
  - POS, part-of-speech tagging
- Considerable advances
  - Sentiment analysis
  - NER, named entity recognition
  - Disambiguation
  - Translation
  - Paraphrasing, abstractive summarization
- Barely started
  - SLU, spoken language understanding

Historical view of NLP

- 1950
  - Alan Turing publishes „*Computing machinery and intelligence*”, in which the Turing Test appears
- 1954
  - first successful translation of 60 phrases from Russian to English
- 1970
  - development of complex language ontologies
- 1974-1980
  - First AI Winter

Historical view of NLP

- 1980
  - triumphant come-back of statistical models, mostly HMMs, solving the POS tag problem.
- 1990-2000
  - machine learning comes back to NLP, mostly due to the availability of huge textual training corpora
- 2000-2010
  - further advances in unsupervised learning for NLP, IR (information retrieval) boom due to the rise of WWW
- 2010-
  - full dominance of neural networks in NLP

NLP tasks: document classification

- **Document classification** (or categorization) is the task of automatic assignment of documents to one of the predefined class or categories based on topics, style, or vocabulary

- If the set of classes/categories is not known *a priori*, the problem morphs into document clustering

NLP tasks: information extraction

Subject: RE: Project status meeting

From: John.Smith@acme.com

Date: 30-08-2023

Hi Jim, can you join tomorrow’s zoom call? Here’s the link: … We meet at 2 pm.

Attribute	Value
Event	Project status meeting
Date	31-08-2023
Start	2:00 pm CET
Duration	
Note	http://zoom.us/xxxx

NLP tasks: sentiment analysis

- **Sentiment analysis** is the task of finding the affective polarization of the text using either the binary (positive/negative) or ternary (positive/neutral/negative) framework

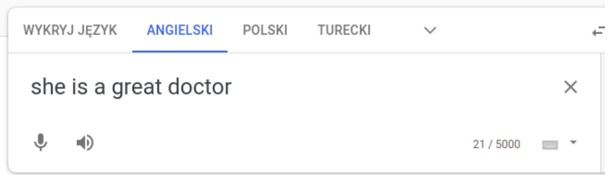
  - “This is the best laptop bag ever. It is so good that within two months of use, it is worthy of being used as a grocery bag.”
  - „Don’t try the pizza, it is so good you will come back every day, it completely ruined my social life, cause each night I only want to go there! I hate this place!”
  - “I have eaten worse.”

NLP tasks: summarization

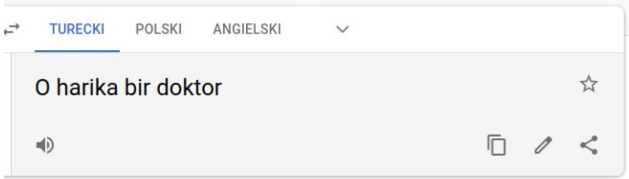
- **Summarization** is the task of automatic generation of a short text containing the most information presented in the original text
  - **extractive** summarization: select the most important and informative phrases from the source text
  - **abstractive** summarization: generate grammatically correct text containing the paraphrase of the source text, often changing the voice, the person, or the tense

NLP tasks: translation

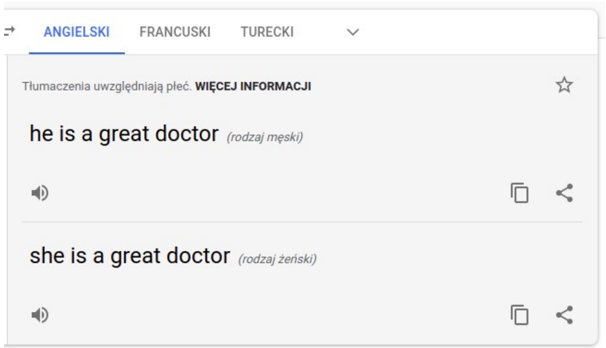
- **Translation** task involves automatic translation between languages, also in the zero-shot version (i.e. translation between pairs of languages for which the model NLP has never been trained)



she is a great doctor



O harika bir doktor



Tłumaczenia uwzględniają płeć. **WIĘCEJ INFORMACJI**

he is a great doctor *(rodzaj męski)*

she is a great doctor *(rodzaj żeński)*

NLP tasks: the song of the future

- Query Answering
- Reading Comprehension
- Textual Entailment
- Coreference Resolution
- Semantic Role Labeling

http://demo.allennlp.org

Ambiguity lies at the heart of NLP

I made her duck.

Ambiguity lies at the heart of NLP

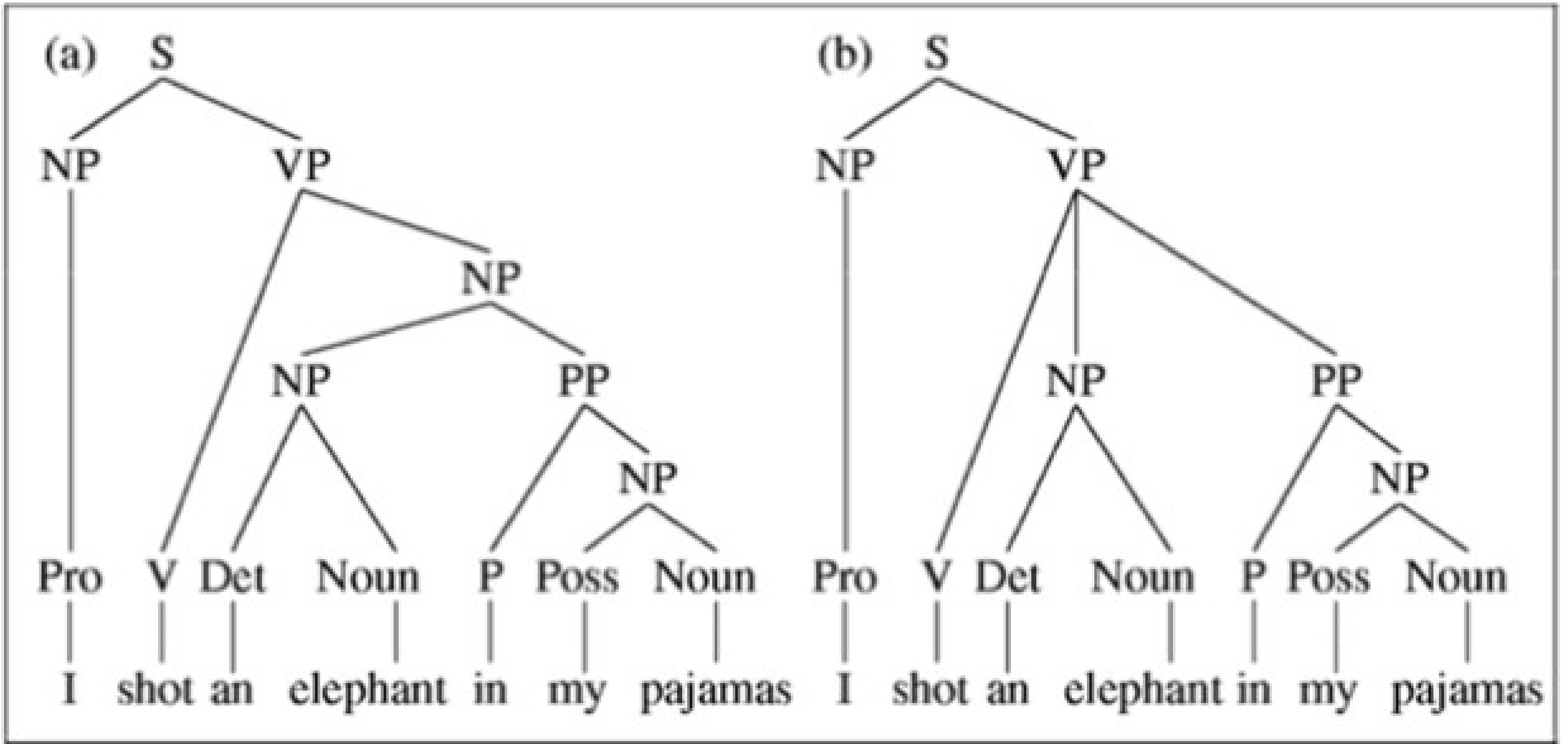
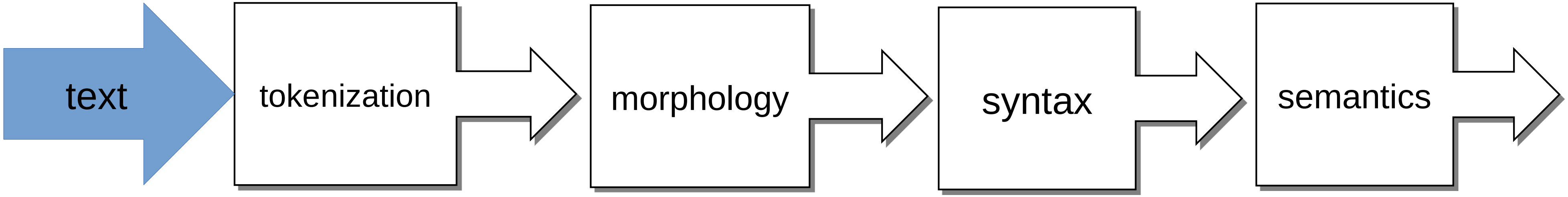


Figure 10.11 Two parse trees for an ambiguous sentence. Parse (a) corresponds to the humorous reading in which the elephant is in the pajamas, parse (b) to the reading in which Captain Spaulding did the shooting in his pajamas.

Why NLP is hard

- Specific, non-normative language (Twitter)
- Living language, neologisms
- Idioms and phrasemes („this is honey on my ears!”)
- Domain vocabulary
- Natural idiosyncrasy of the language

Text processing steps



text

tokenization

morphology

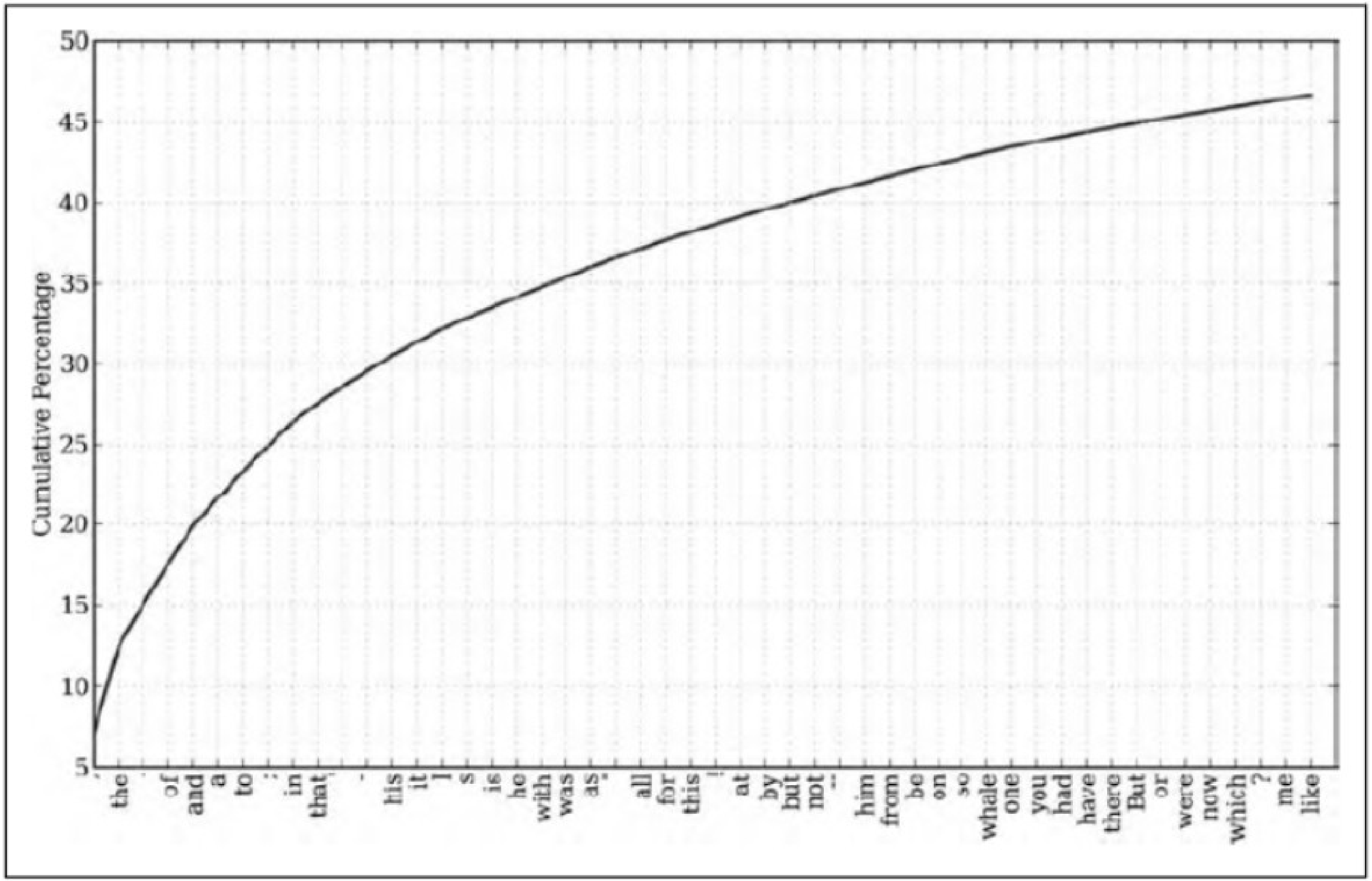
syntax

semantics

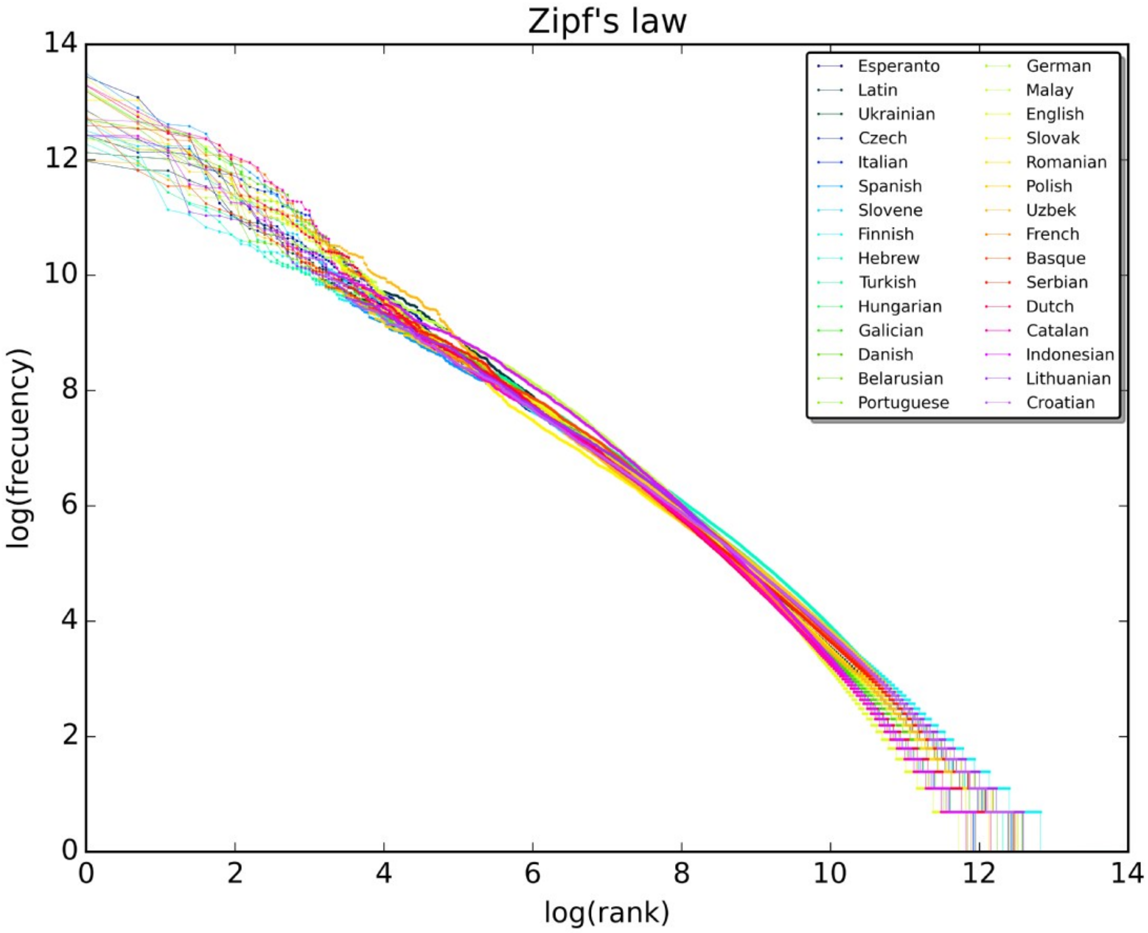
Textual corpora

- **Corpus**: large and ordered collection of texts prepared for digital processing, usually used for modeling the language for a particular domain (vocabulary, structures, syntax constructions)
- **Rank list**: the list of words ordered by the frequency of occurrences in the corpus
- **Frequency**: number of occurrences of a word in the corpus
- **Rank**: number of position on the rank list

Example of a rank list



Zipf law



Frequency of a word is inversely proportional to its **rank**.

- Most frequent words occur in all documents
- Many words occur at most once in the corups
- Semantics is hidden in less frequent words

Tokenization

- **Tokenization** breaks the stream of text into atomic entities called **tokens**. A token is a string of characters surrounded by separators. Tokens are the basic “words” in further processing.
  - it’s raining cats’n’dogs
  - Hewlett-Packard black-white printer
  - My taxID is 123-456-78-90 and I have to pay 1 200,99 EUR.

Levenshtein distance

- **Levenshtein distance** between two strings of characters is the minimum number of atomic operations required to transform one string into the other
  - addition of a character: modeling → modelling
  - deletion of a character: psychology → psyhology
  - changing a character: coronavirus → coronovirus
  - transposition of a character: yesterday → yetserday

BOW model

- According to a **BOW** (*bag-of-words*) **model**, each document is fully represented by a set of words in the document along with their frequency, without considering the ordering of words.
  - the simplest vector model for documents
  - each word has the same weight
  - the model does not reflect the grammar of the document
    - *Hey let's eat, grandma!*   *Hey, let's eat grandma!*
  - the model does not provide a unique representation of a document
    - *I am not happy, I am sad*   *I am not sad, I am happy*

BOW model

- It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, [...]
  - Charles Dickens, „*Tale of Two Cities*”
- Vocabulary
  - it, was, the, best, of, times, worst, age, wisdom, foolishness
- Vectors
  - it was the best of times = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]
  - it was the worst of times = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]
  - it was the age of wisdom = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]

TF-IDF model

- According to the **TF-IDF** (*term frequency, inverse document frequency*) **model**, the weight of each word depends on:
  - the number of occurrences of the word in the document
  - the number of documents in which a word occurs

$$w_{t,d} = tf_{t,d} * \log\left(\frac{N}{df_t}\right)$$

Normalization

- In general, **normalization** is the set of techniques which transform the raw text input into a standard form suitable for further algorithmic processing
  - tokenization
  - filtering of stopwords
  - normalization of words (stemming and lemmatization)
  - text segmentation

Lemmatization

- **Lemma** is the most common vocabulary form of a word, which represents all other forms of that word
- **Surface form** is the form in which the word appears in the source text
- **Word stem** is the part of the word which is not changed due to declination, conjugation or flexive rules
- **Morphem** is the smallest indivisible group of phonemes which does not contain any prefixes, suffixes or interfixes

Stemming

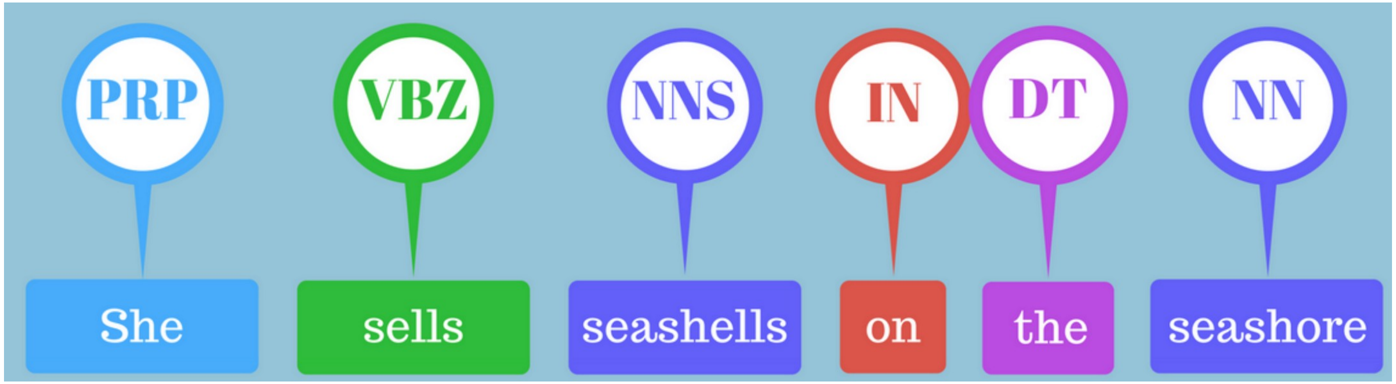
- **Stemming** removes all morphological affixes and leaves only the stem of the word
  - dog, doggo, dogs → dog
  - flies, flight, flying → fli
  - having, havs, have → hav
- The result of stemming is not necessarily a correct word
- Many fast stemmers: Snowball, Lovins, Porter

Lemmatization

- **Lemmatization** changes words into lemmas (dictionary forms), the result of lemmatization is the canonical form of each word
- Lemmatizers work based on dictionaries and POS taggers
  - am, are, were, being, is → be
  - cat, she-cat, kitty, cats → cat
  - having, had → have

POS tagging

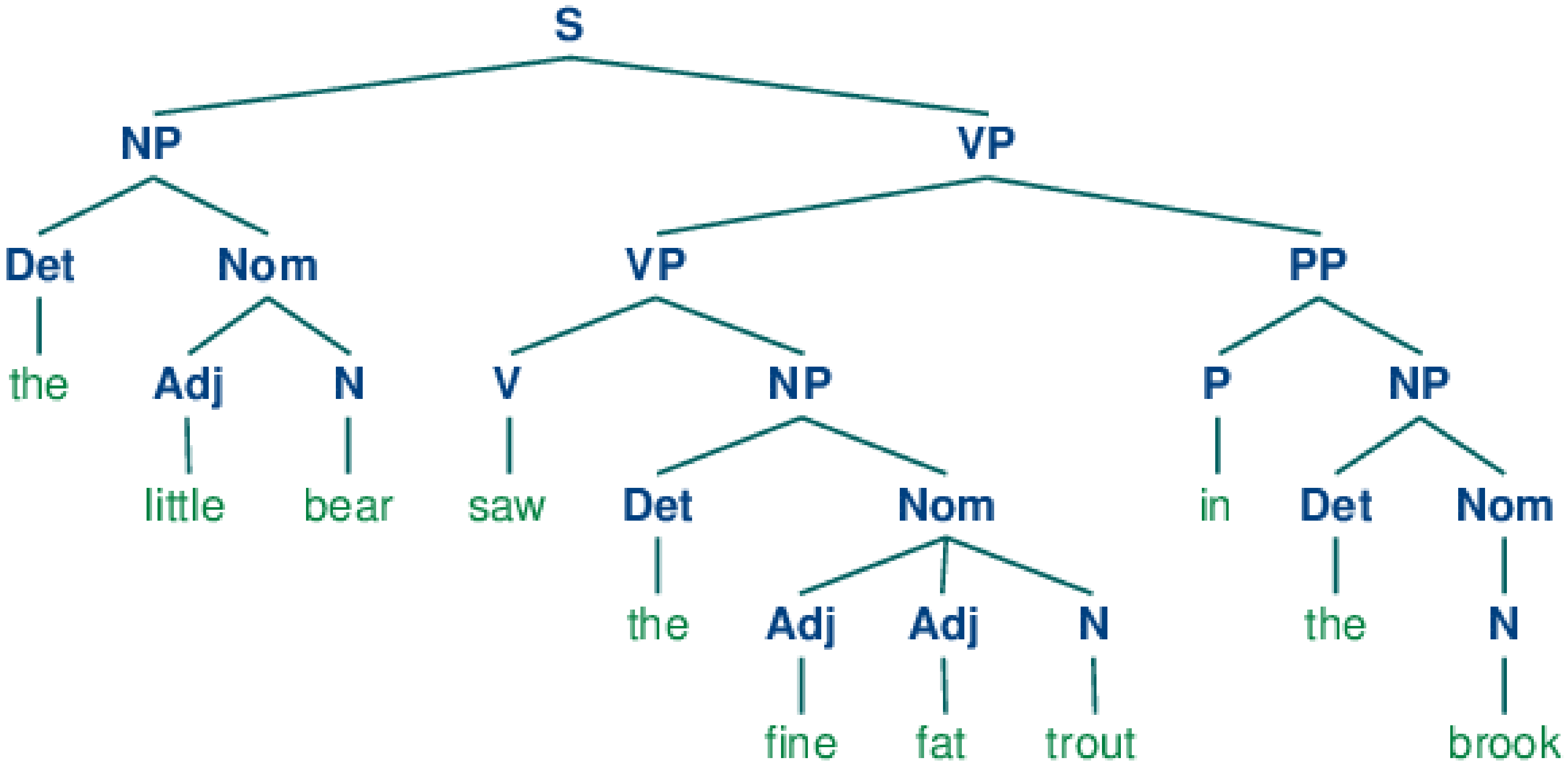
- **POS tagging** is the task of assigning to each token the correct part of speech (verb, noun, adjective, preposition…).
  - There are many different schemas of POS tagging
    - https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
    - https://universaldependencies.org/u/pos/
    - https://www.sketchengine.eu/polish-nkjp-part-of-speech-tagset/



Dependency parsing

- **Dependency parsing** identifies the structure of a phrase defined as the relationships between phrases (noun and verb) expressed as the dependency tree. Leaves of the tree represent terminal symbols (tokens) and the internal nodes represent non-terminal symbols of the grammar. Each tree has only one root.

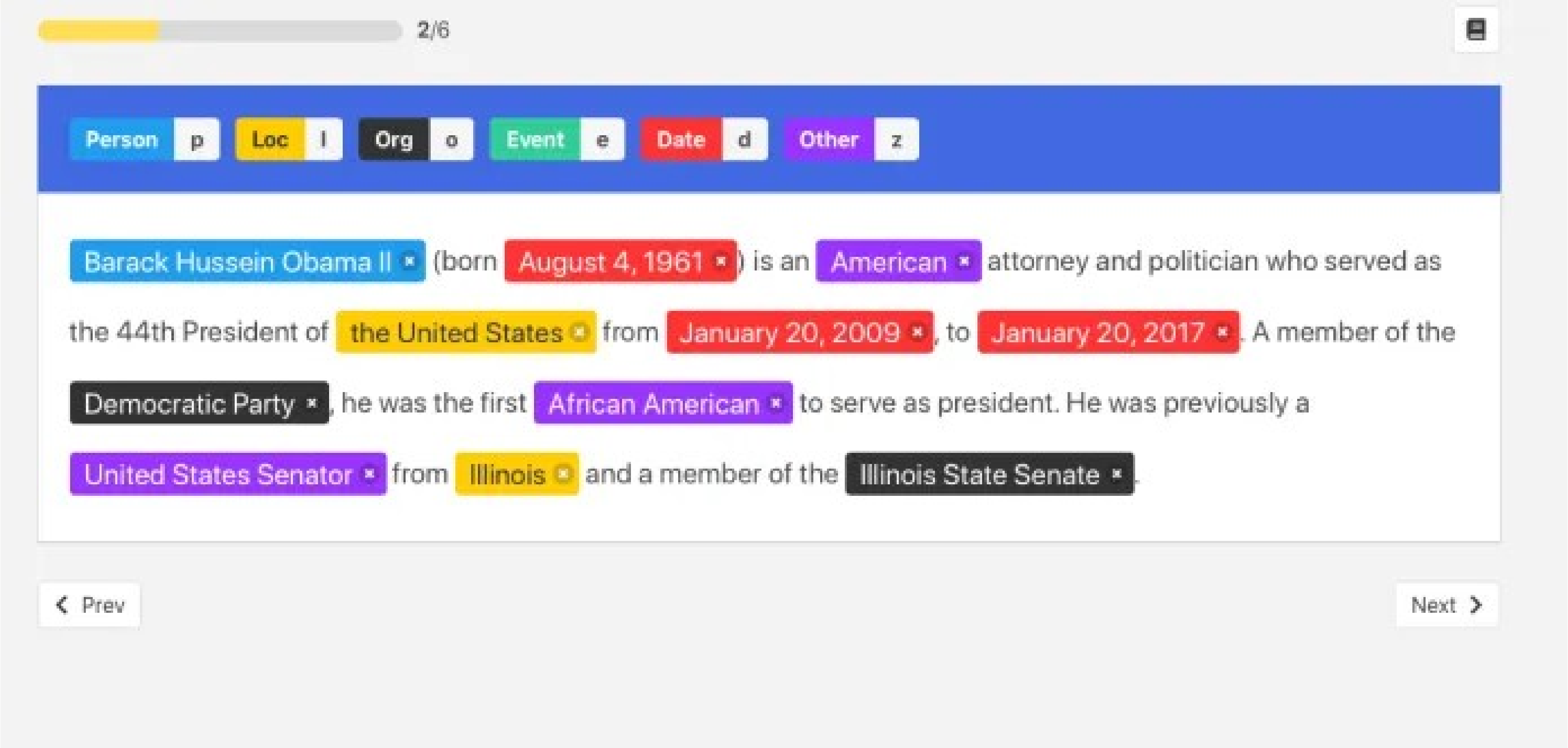
Dependency tree example



Named entity recognition

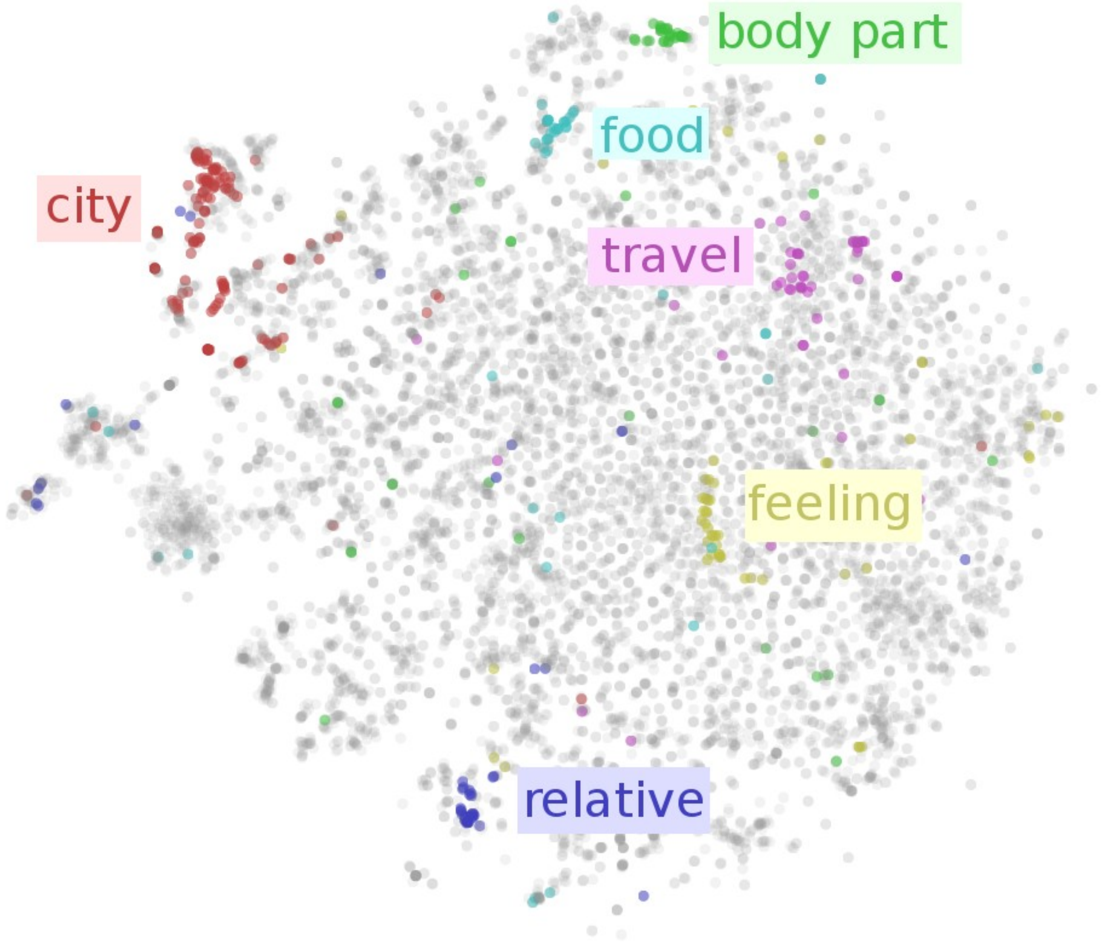
- **NER (*named entity recognition*)** is the task of identifying and marking in the text the occurrences of tokens which represent certain classes of real-world objects
  - there are numerous annotation schemes for NERs
  - most NER taggers recognize persons, organizations, geographical names, localizations, money, numbers, dates, percentages, time, durations, e-mails, URLs
  - there are specialized NER schemes which may cover classes such as title, species, religion, cause of death, type of cuisine, etc.

Example of NER tagging



Barack Hussein Obama II (born August 4, 1961) is an American attorney and politician who served as the 44th President of the United States from January 20, 2009, to January 20, 2017. A member of the Democratic Party, he was the first African American to serve as president. He was previously a United States Senator from Illinois and a member of the Illinois State Senate.

Embeddings (word vectors)



Why vectors?

- Representing words as vectors is very beneficial
  - ML algorithms expect to see numbers
  - many algorithms depend on linear algebra and vector/matrix operations
  - many concepts are much easier to represent in the vector space: context, analogy, synonyms, antonyms, word ambiguity

One-hot encoding

- One-hot encoding assigns each vector dimension to a single word, a document is represented as a bag-of-words
  - each word is a vector of the length equal to the size of the vocabulary
  - document vector is the sum of word vectors
  - vector value can be:
    - 1: vector encodes the occurrence of the word in a document
    - number of occurrences
    - TF-IDF: vector encodes the “importance” of the word

What is wrong with one-hot encoding?

- Resulting vectors have huge dimensionality and are extremely sparse (very few dimensions are non-zero)
- The loss of word ordering may lead to errors
  - „good but expensive” → „expensive but good”
- No way to represent the semantic similarity between words

Distributional hypothesis

- **You shall know a word by the company it keeps**

  John Rupert Firth

Simplest possible vectors

- Count the co-occurrences of words in a corpus
  - I enjoy flying
  - I like NLP
  - I like deep learning

$$X = \begin{array}{c|cccccccc}
 & I & like & enjoy & deep & learning & NLP & flying & . \\
\hline
I & 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\
like & 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
enjoy & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
deep & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
learning & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
NLP & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
flying & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
. & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0
\end{array}$$

Why word vectors should be dense?

- Dense vectors require far less memory for processing
  - typical dense vectors have 50-1000 dimensions
  - all vector elements are non-zero
- Dense vectors have much better generalization properties
  - identification of synonyms
  - vector space “encodes” grammatical relations and semantic similarity
  - in practice, dense vectors **are much much better**

2003: the revolution begins

- Bengio, Yoshua, et al. "A neural probabilistic language model." Journal of machine learning research 3.Feb (2003): 1137-1155.

  “This is intrinsically difficult because of the curse of dimensionality: a word sequence on which the model will be tested is likely to be different from all the word sequences seen during training. […] We propose to fight the curse of dimensionality by learning a distributed representation for words”

- The formulation of the language modeling task: to predict the next word in a sequence given the number of previous words.

2013: the revolution explodes

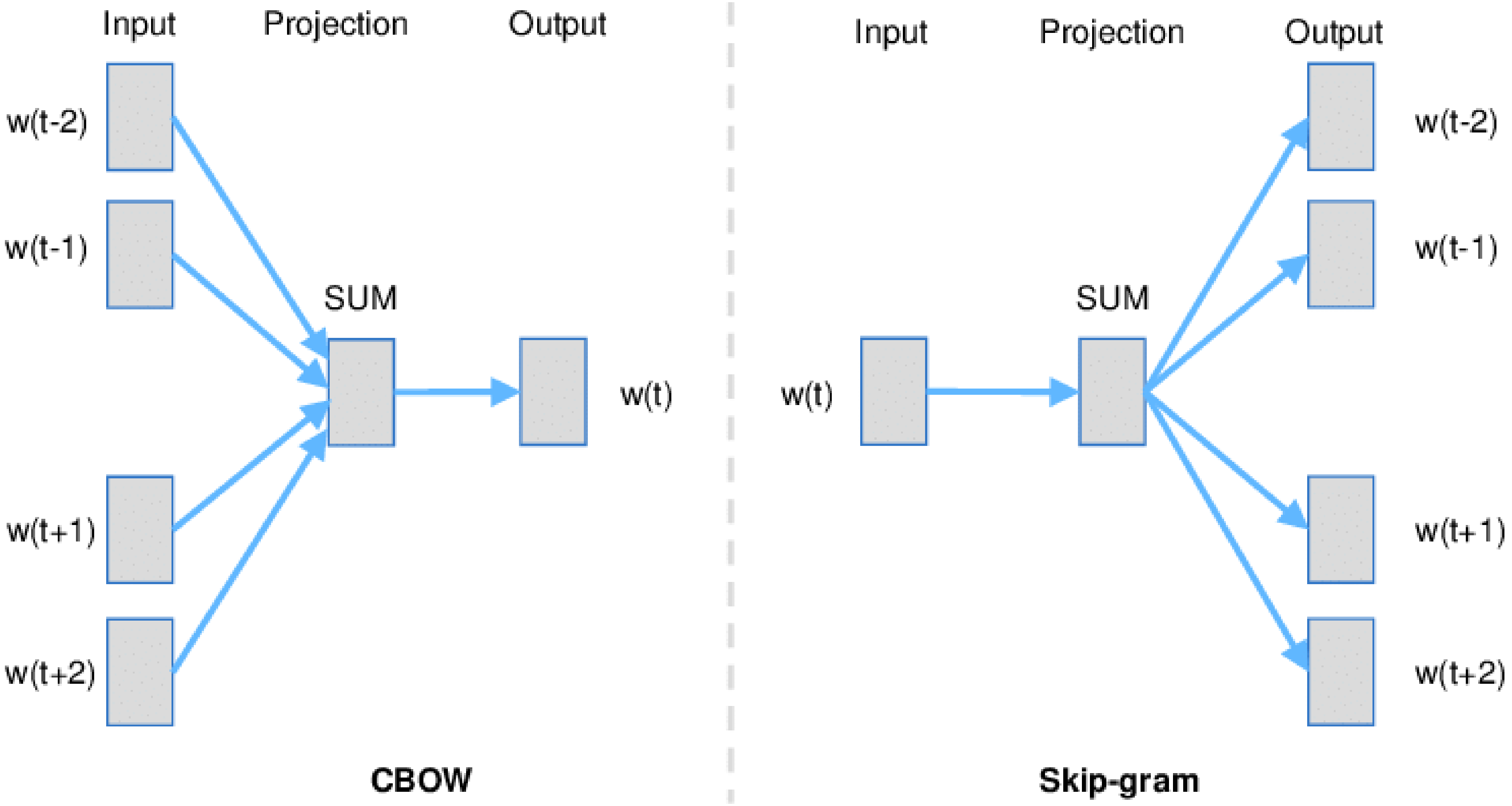
- Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013): 3111-3119.

- The model predicts the current word based on the context
  - it requires a huge textual corpus for learning
  - a very clever trick for self-learning

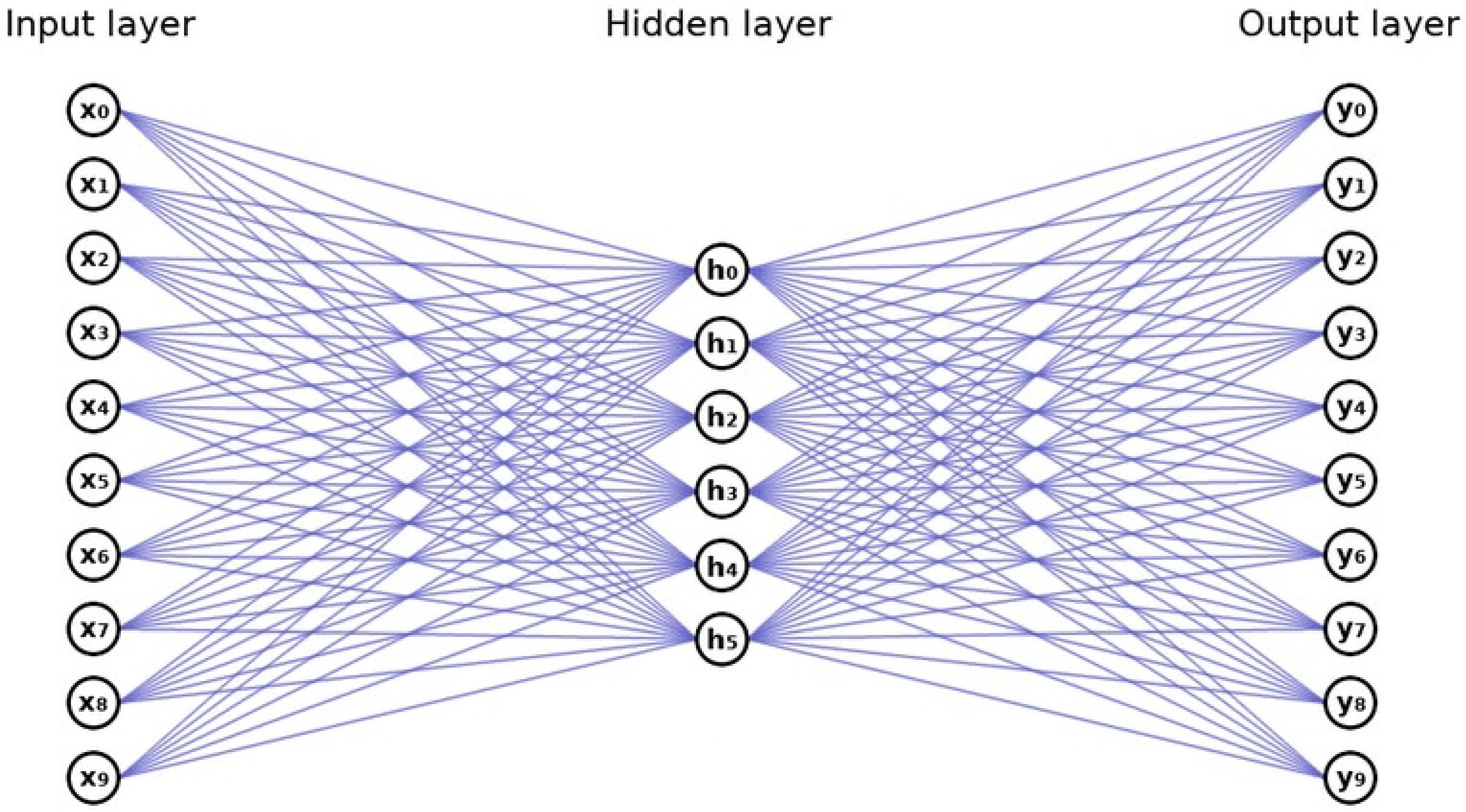
The main idea behind **word2vec**

- Shallow two-layer neural network trained on the task of reconstructing the context of a word
- **Context** of a word is the set of k surrounding words
  - „Mary had a little white lamb”, k=2, {little| Mary, had, white, lamb}
- Two versions of the model: Skip-Gram and CBOW
  - instead of counting the co-occurrence of words, train a simple neural network for predicting if the word “lamb” will occur in the context of the word “little”
  - **best idea ever**: the aim of the task is not the classifier, but the hidden layer of the neural network

SkipGram vs CBOW



Hidden layer (embedding)

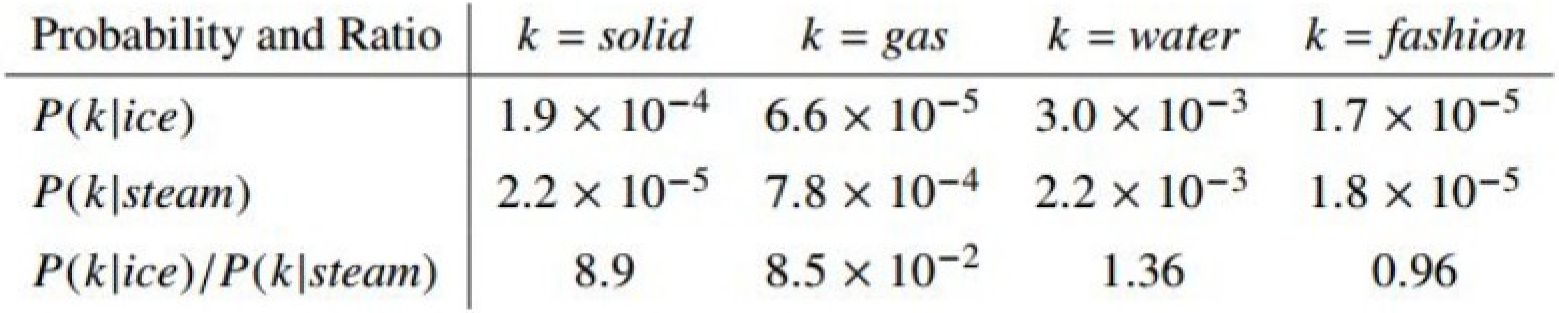


GloVe

- GloVe (*Global Vectors*) is an alternative algorithms to word2vec
- GloVe uses two distinct contexts
  - local context: co-occurrence of words
  - global context: triads of words

Explaining GloVe

- Let P_{ij} denote the probability of i and j co-occurrence. Let us consider a word k
  - if i and k are related, but k and j are not, then P_{ik} / P_{jk} is large
  - if i and k are not related, but k and j are, then P_{ik} / P_{jk} is small
  - if k is or is not related to i and j, then P_{ik} / P_{jk} is close to 1

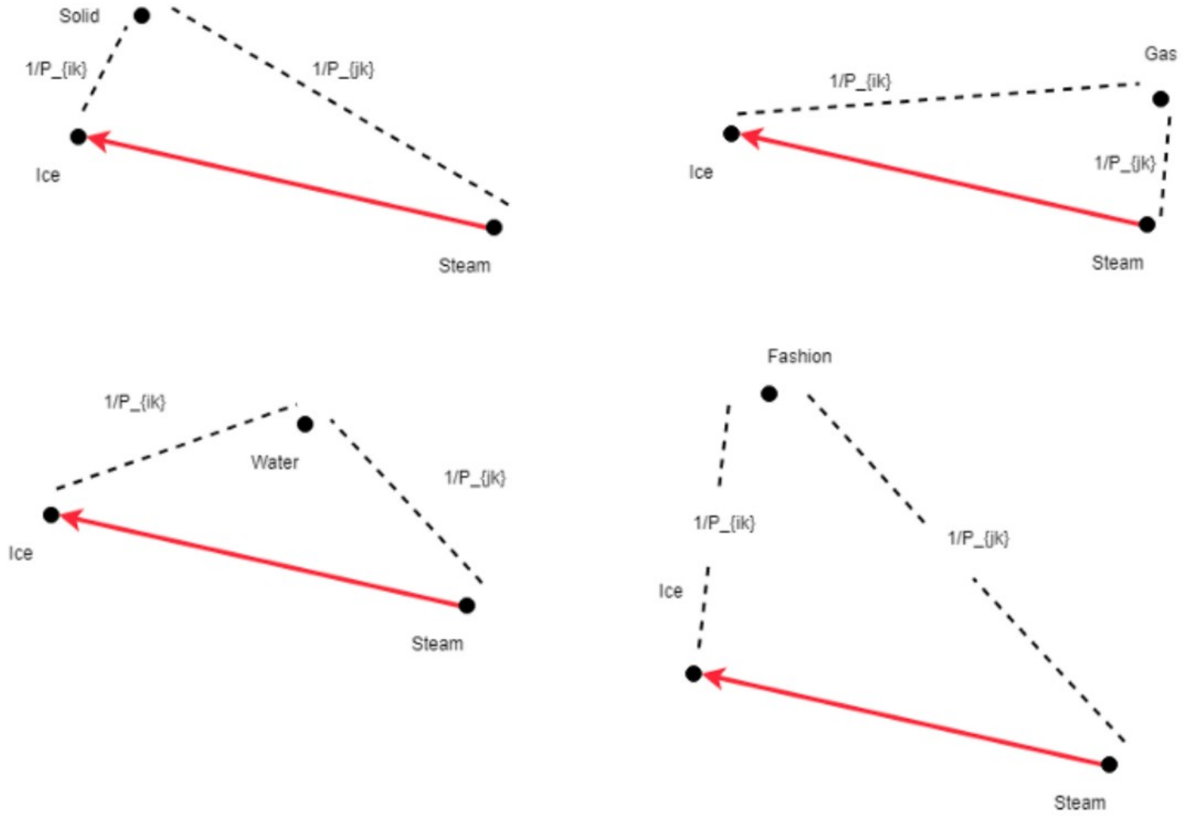


Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

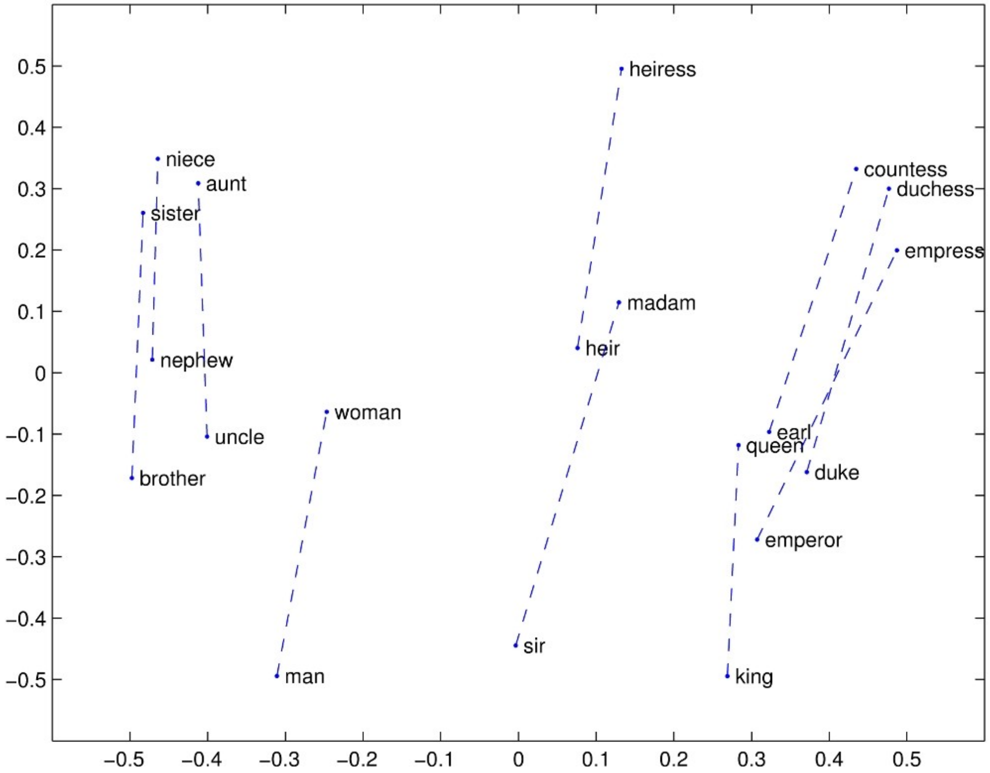
Training GloVe

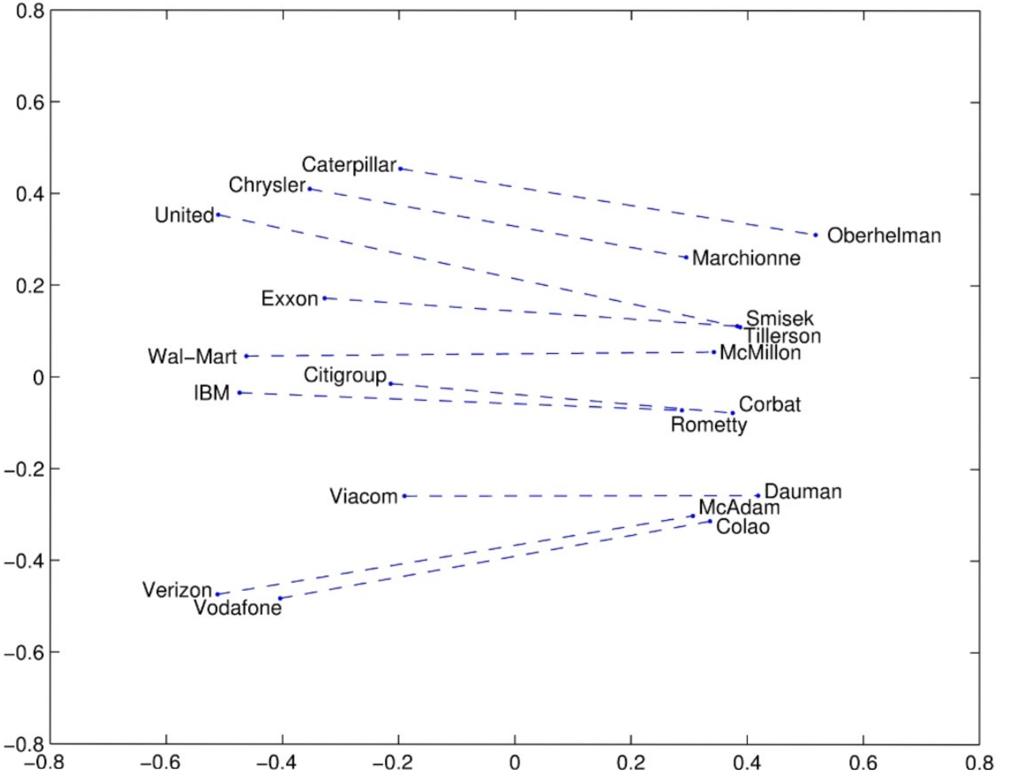
- Loss function for GloVe:

$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$



GloVe and analogy modeling





Other vector models

- Dependency-based vectors
- Sentiment-based vectors
- Structural vectors
- Vectors trained on knowledge bases
- Vectors trained on multiple tasks
- Hyperbolic vectors in Pointcare space

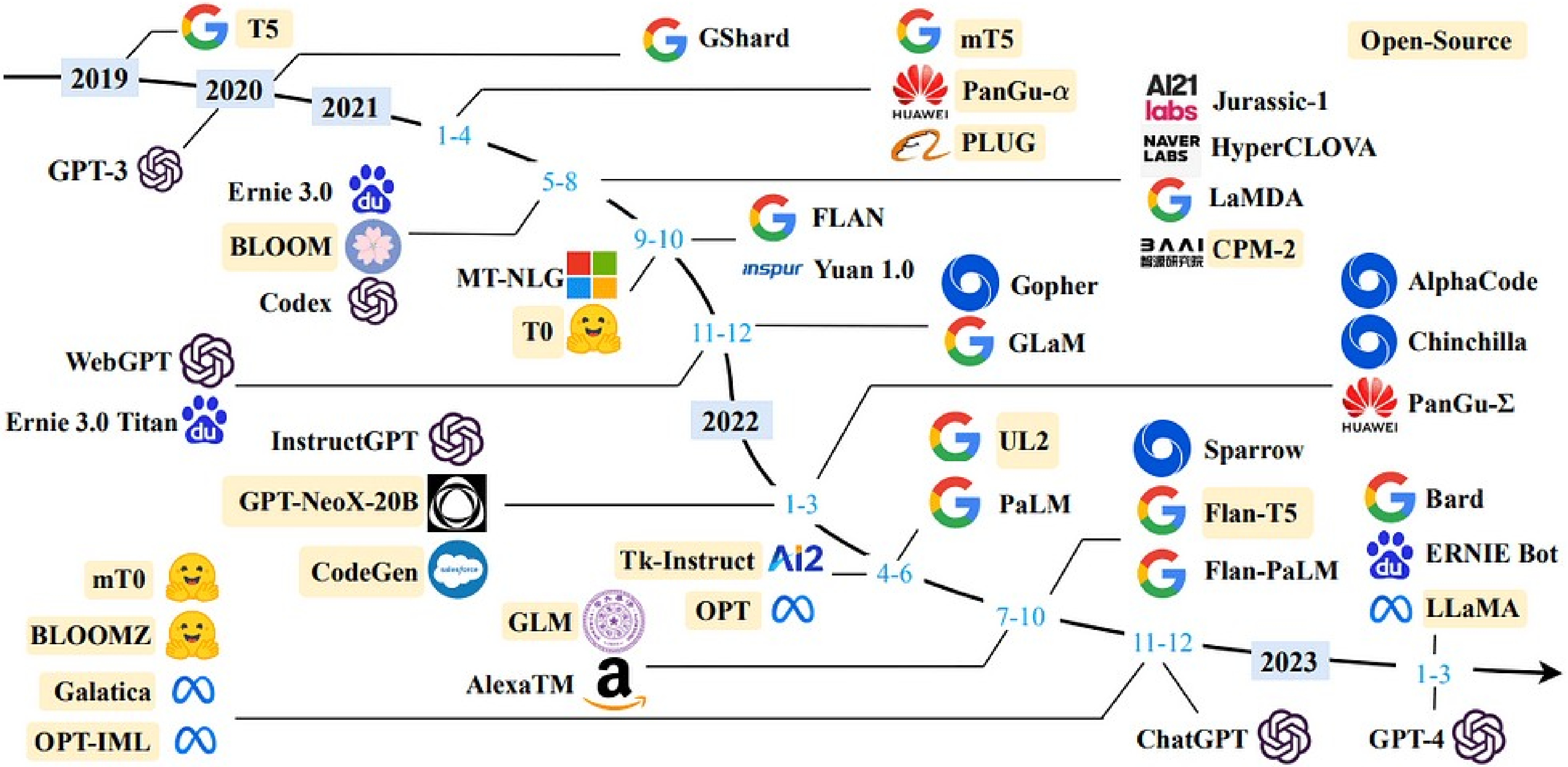
Problems with vectors

- Vectors encode many biases present in the language
  - father → doctor, mother → nurse
  - man → computer programmer, woman → homemaker
  - Afro-american words are much closer to “unpleasant” words appearing in the dictionary
  - Similar relationship between names of old people vs names of young people

Large Language Models

- A language model is a function which predicts the next token given an input sequence of tokens
- A Large Language Model (LLM) is a language model trained on extreme textual corpora using extreme computational resources
- LLMs are called generative auto-regressive models, because:
  - they generate next token for a given sequence
  - they consume their output during token generation
    - A long time → ago
    - A long time ago → in
    - A long time ago in → a
    - A long time ago in a → galaxy

Language model zoo

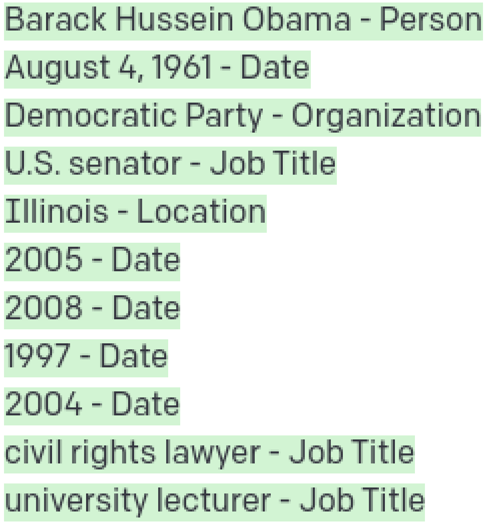


Prompting

- Prompting is a technique to „force” large language model to perform a specific language task:
  - named entity recognition
  - text classification
  - summarization
  - translation
  - paraphrasing

Prompting example

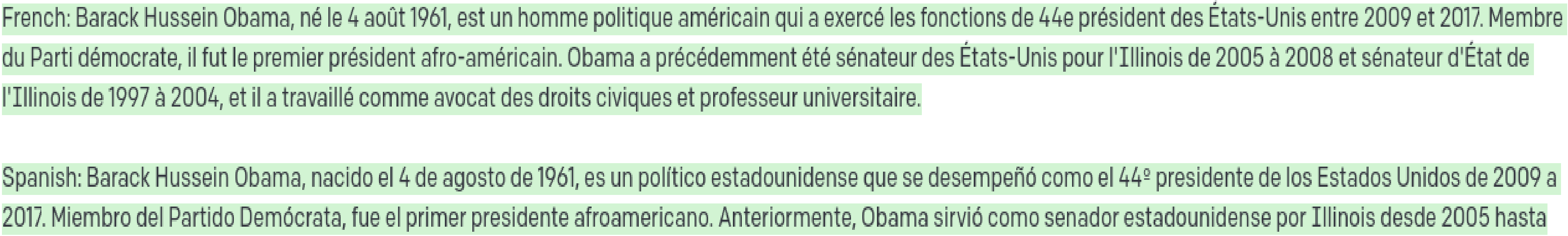
- TEXT=Barack Hussein Obama, born August 4, 1961, is an American politician who served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, he was the first African-American president. Obama previously served as a U.S. senator representing Illinois from 2005 to 2008 and as an Illinois state senator from 1997 to 2004, and worked as a civil rights lawyer and university lecturer.
  - Read the TEXT. Return all named entities and their labels from the TEXT.



Barack Hussein Obama - Person
August 4, 1961 - Date
Democratic Party - Organization
U.S. senator - Job Title
Illinois - Location
2005 - Date
2008 - Date
1997 - Date
2004 - Date
civil rights lawyer - Job Title
university lecturer - Job Title

Prompting example

- TEXT=Barack Hussein Obama, born August 4, 1961, is an American politician who served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, he was the first African-American president. Obama previously served as a U.S. senator representing Illinois from 2005 to 2008 and as an Illinois state senator from 1997 to 2004, and worked as a civil rights lawyer and university lecturer.
    - Read the TEXT. Translate it into French and Spanish.

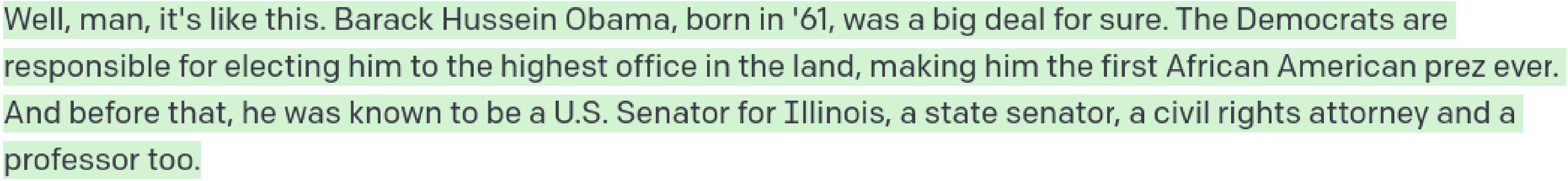


French: Barack Hussein Obama, né le 4 août 1961, est un homme politique américain qui a exercé les fonctions de 44e président des États-Unis entre 2009 et 2017. Membre du Parti démocrate, il fut le premier président afro-américain. Obama a précédemment été sénateur des États-Unis pour l'Illinois de 2005 à 2008 et sénateur d'État de l'Illinois de 1997 à 2004, et il a travaillé comme avocat des droits civiques et professeur universitaire.

Spanish: Barack Hussein Obama, nacido el 4 de agosto de 1961, es un político estadounidense que se desempeñó como el 44º presidente de los Estados Unidos de 2009 a 2017. Miembro del Partido Demócrata, fue el primer presidente afroamericano. Anteriormente, Obama sirvió como senador estadounidense por Illinois desde 2005 hasta

Prompting example

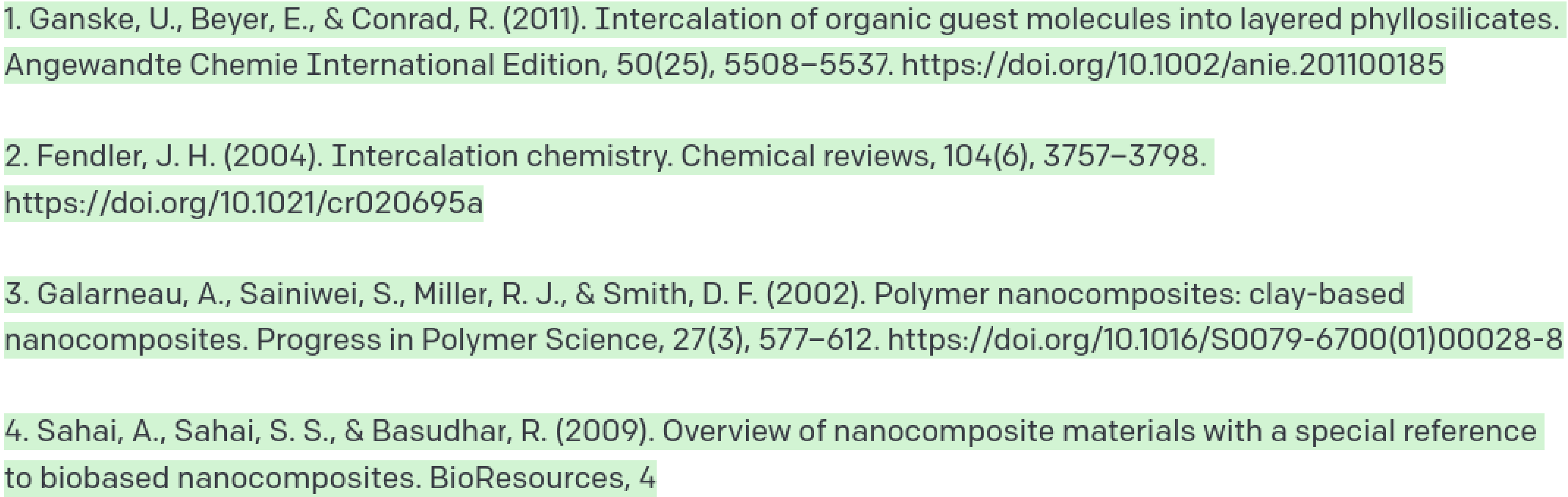
- TEXT=Barack Hussein Obama, born August 4, 1961, is an American politician who served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, he was the first African-American president. Obama previously served as a U.S. senator representing Illinois from 2005 to 2008 and as an Illinois state senator from 1997 to 2004, and worked as a civil rights lawyer and university lecturer.
    - Read the TEXT. Paraphrase it in the style of the Dude from Big Lebowski.



Well, man, it's like this. Barack Hussein Obama, born in '61, was a big deal for sure. The Democrats are responsible for electing him to the highest office in the land, making him the first African American prez ever. And before that, he was known to be a U.S. Senator for Illinois, a state senator, a civil rights attorney and a professor too.

Prompting example

- TEXT=The first example of sheet-like nano-fillers often used in polymer nanocomposites is layered silicates belonging to the structural family known as the 2:1 phyllosilicates. Montmorillonite, hectorite and saponite etc. are the usually used layered silicates. Their crystal lattice is composed of two-dimensional layers. A central octahedral sheet of alumina or magnesia is fused into two external silica tetrahedron so that the oxygen ions of octahedral sheet pertain to tetrahedral sheets. The thickness of layers is ca. 1 nm and the lateral dimensions of the layers vary from several tens of nanometers to several micrometers and even larger. These layers will form stacks with van der Waals gaps between “interlayers” or “inter-galleries”. This kind of clay has a moderate cation exchange capacity and the charge of the layer is not constant since it differentiates from one layer to another and thus a mean value over the whole crystal should be taken into account.
    - Read the TEXT and produce a list of bibliographic references relevant for the TEXT.



1. Ganske, U., Beyer, E., & Conrad, R. (2011). Intercalation of organic guest molecules into layered phyllosilicates. Angewandte Chemie International Edition, 50(25), 5508–5537. https://doi.org/10.1002/anie.201100185

2. Fendler, J. H. (2004). Intercalation chemistry. Chemical reviews, 104(6), 3757–3798. https://doi.org/10.1021/cr020695a

3. Galarneau, A., Sainiwei, S., Miller, R. J., & Smith, D. F. (2002). Polymer nanocomposites: clay-based nanocomposites. Progress in Polymer Science, 27(3), 577–612. https://doi.org/10.1016/S0079-6700(01)00028-8

4. Sahai, A., Sahai, S. S., & Basudhar, R. (2009). Overview of nanocomposite materials with a special reference to biobased nanocomposites. BioResources, 4

Stochastic parrots