Natural Language Processing and Topic Modeling Review

Silvia Terragni
University of Milano-Bicocca, Milan
s.terragni4@campus.unimib.it
silviatti.github.io
@TerragniSilvia

AllianceBernstein (from Milan) 30/10/2020

(What I say) About me

Silvia Terragni is a

Silvia Terragni is a **Ph.D. student** at the University of Milano Bicocca. She is passionate about **Machine Learning and Natural Language Processing**. Her main research focuses on **Topic Modeling**. She recently investigated the combination of pre-trained representations of documents with topic models.

(What a machine says) about me

https://huggingface.co/gpt2-medium?text=Silvia+Terragni+is+a

Are Robots Taking Over?

Opinion
Artificial intelligence (AI)

GPT-3
Tue 8 Sep 2020 09.45 BST

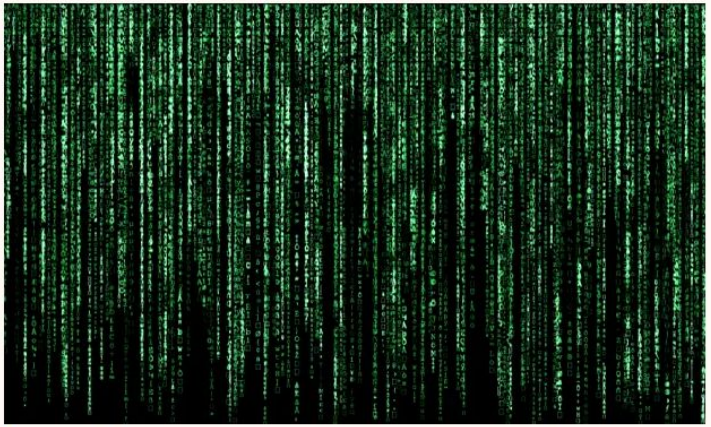
69,362 1,188

This article is more than 1 month old

A robot wrote this entire article. Are you scared yet, human?

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below



'We are not plotting to take over the human populace.' Photograph: Volker Schlichting/Getty Images/EyeEm

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

https://www.theguardian.com/commentisfree/2020/sep/ot-wrote-this-article-gpt-3

Welcome to NLP

Devising computational algorithms to treat language.

For example, with machine learning:

Welcome to NLP

Devising computational algorithms to treat language.

For example, with machine learning:

This movie is boring, I don't like it!

Negative

This cake is delicious :)

Positive

Task:

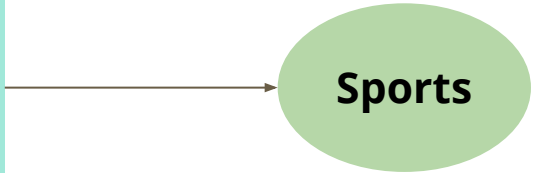
Sentiment Analysis

Welcome to NLP

Devising computational algorithms to treat language.

For example, with machine learning:

Cristiano Ronaldo is a famous soccer player that plays for Juventus F.C., during his career he has won a lot of trophies



Task:

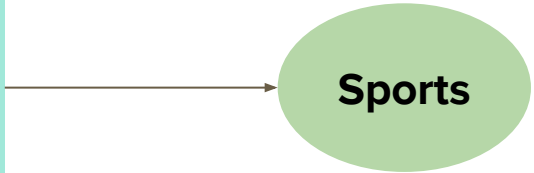
Topic Modeling

Welcome to NLP

Devising computational algorithms to treat language.

For example, with machine learning:

Cristiano Ronaldo is a famous soccer player that plays for Juventus F.C., during his career he has won a lot of trophies



Task:

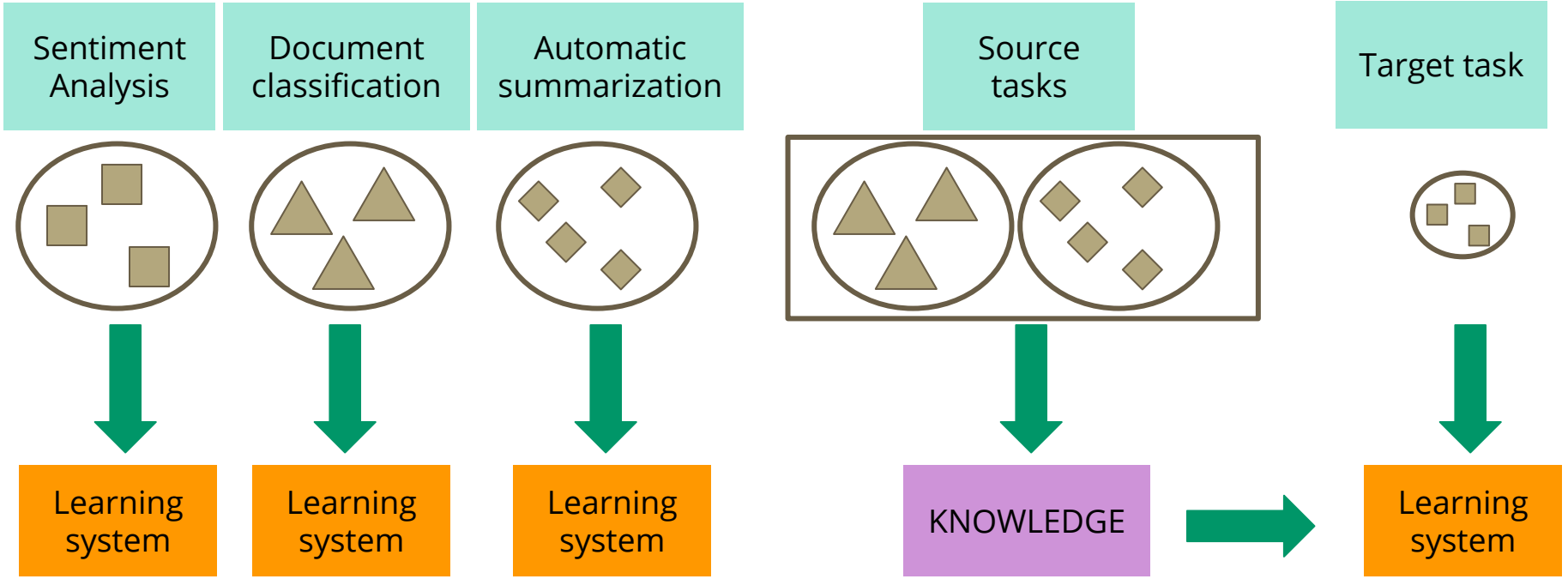
Topic Modeling

To solve these tasks, we need to be able to represent words and documents within a mathematical framework

Is there a solution that fits all the problems?

source: https://youtu.be/rEGB7-FlPRs

Transfer learning in NLP



Transfer learning in NLP

- **Feature-based**: we use the (deep) representations of the pre-trained model as input of the task model
- **Fine-tuning**: all the parameters of the pre-trained models are fine-tuned (adjusted) on a downstream task

We need a **source task** that allows us to learn high-level features from language and a **dataset** that is big enough

Words Into Machines

From Language to Vector Spaces

DOG

CAT

ROME

PABLO

From Language to Vector Spaces

DOG	0	0	1	0
CAT	1	0	0	0
ROME	0	1	0	0
PABLO	0	0	0	1

one-hot encoding

From Language to Vector Spaces

DOG	0	0	1	0
CAT	1	0	0	0
ROME	0	1	0	0
PABLO	0	0	0	1

one-hot encoding

ISSUES:

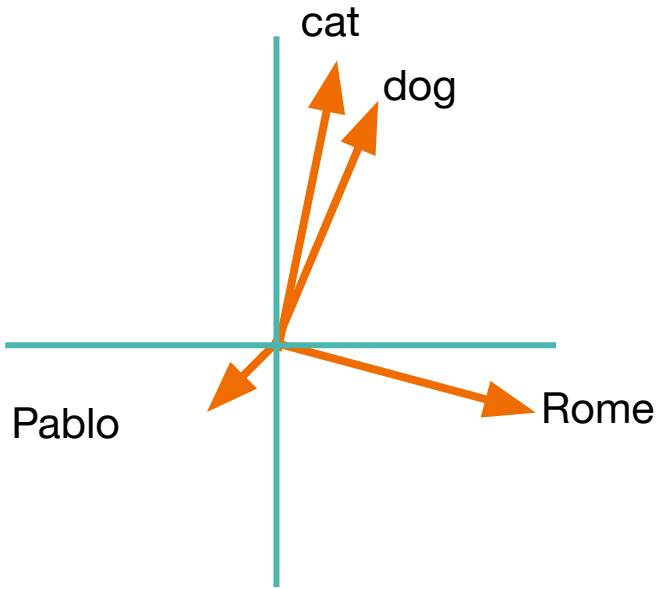
1. Matrix of one hot encoded words is V*V where V is the number of words → **scaling problem!** (imagine encoding all the words of the English vocabulary)

2. Each vector is **orthogonal** to each other! But may agree on the fact that “Cat” is more similar to “Dog” than to “Rome”.

Word embeddings

Any technique mapping a word (or phrase) from its original high-dimensional input space (the vocabulary of all words) to a lower-dimensional numerical vector space

DOG	0	0	1	0
CAT	1	0	0	0
ROME	0	1	0	0
PABLO	0	0	0	1



From Word Usage to Word Representation

(From Lenci & Evert): what's the meaning of **'bardiwac'**?

From Word Usage to Word Representation

(From Lenci & Evert): what's the meaning of **'bardiwac'**?

- He handed her glass of **bardiwac**
- Beef dishes are made to complement the **bardiwacs**
- Nigel staggered to his feet, face flushed from too much **bardiwac**
- Malbec, one of the lesser-known **bardiwac** grapes, responds well to Australia's sunshine
- I dined on bread and cheese and this excellent bardiwac
- The drinks were delicious: blood-red **bardiwac** as well as light, sweet Rhenish

From Word Usage to Word Representation

(From Lenci & Evert): what's the meaning of **'bardiwac'**?

- He handed her glass of **bardiwac**
- Beef dishes are made to complement the **bardiwacs**
- Nigel staggered to his feet, face flushed from too much **bardiwac**
- Malbec, one of the lesser-known **bardiwac** grapes, responds well to Australia's sunshine
- I dined on bread and cheese and this excellent bardiwac
- The drinks were delicious: blood-red **bardiwac** as well as light, sweet Rhenish

'Bardiwac' is a heavy red alcoholic beverage made from grape

'Bardiwac' appears in **drinking-related** contexts, close to words like 'glass' and 'grape'

From Word Usage to Word Representation

- “The meaning of a word is its **use** in the language” (Wittgenstein, 1953)
- “You shall know a word by the **company** it keeps” (Firth, 1957)

Distributional Hypothesis:
similar words tend to appear in similar **contexts**

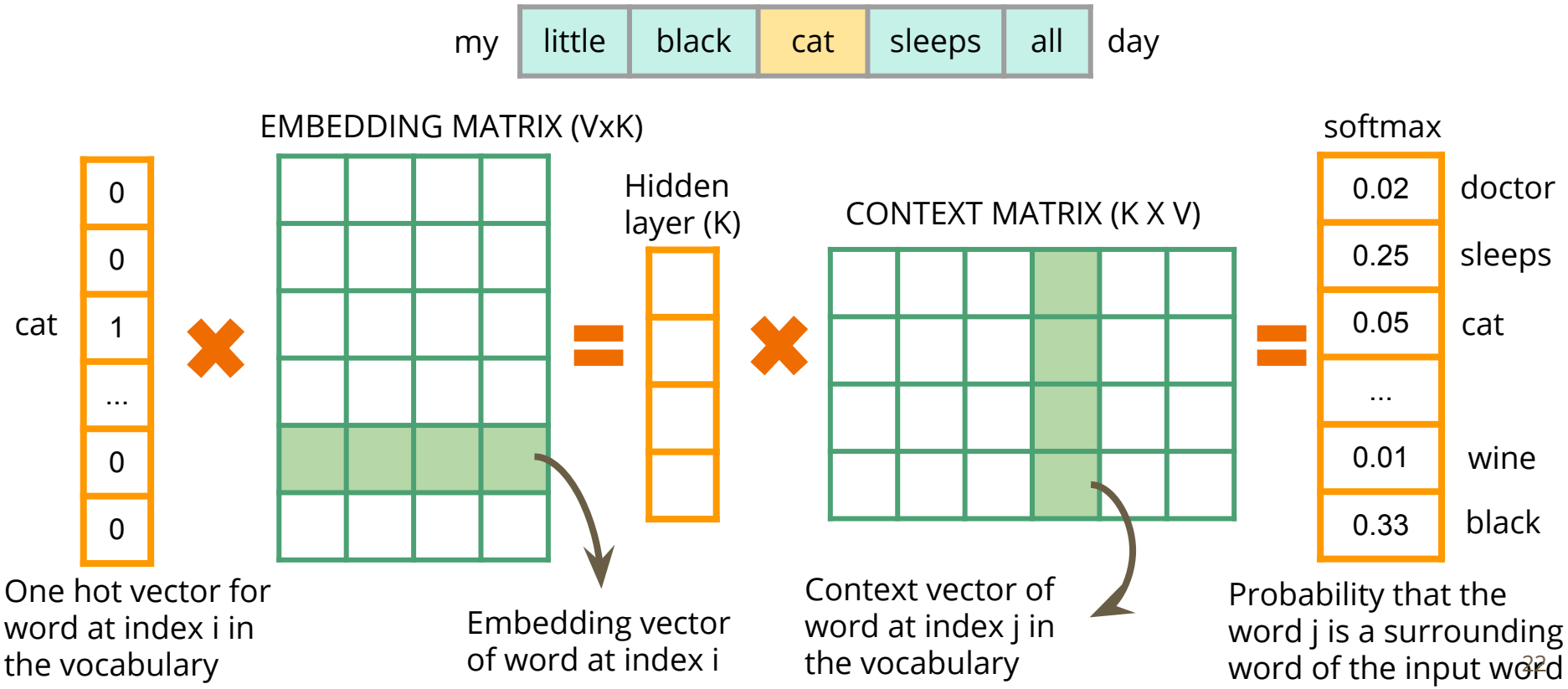
Distributional Semantics with Word2vec

Word2vec [Mikolov+, 2013] takes inspiration from the **Distributional Hypothesis** [Harris,1954] to learn **continuous vector representations of words**, i.e., word embeddings, generated from a text corpus

HOW?

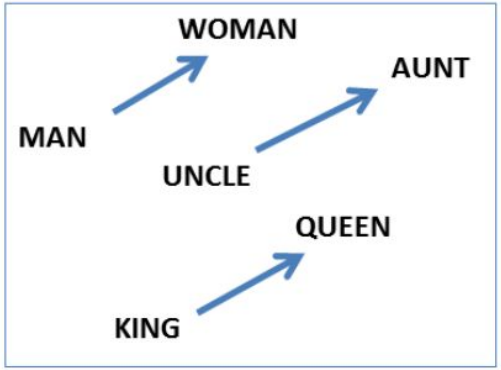
- **Continuous Bag Of Words (CBOW)**: one hidden layer neural network
  - predict the word in the middle given the context words
- **Skip-gram model**: one hidden layer neural network
  - predict the context words given the input word

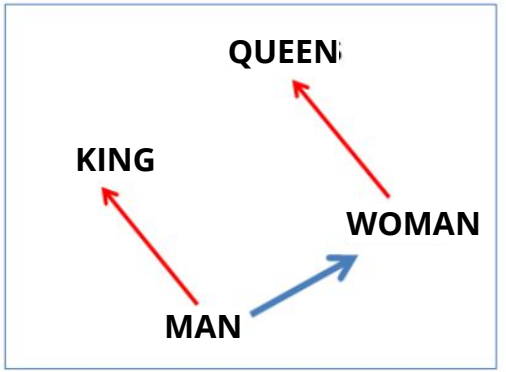
Word2vec: Skip-gram model



Why using Word2vec?

- Compression (dimensionality reduction)
- Smoothing (discrete to continuous)
- Densification (sparse to dense)
- Analogical reasoning
- **Feature-based approach**: can be used to initialize the first layer of a classifier to solve a given task (e.g. to represent a document we can average the embeddings of the words composing the document)





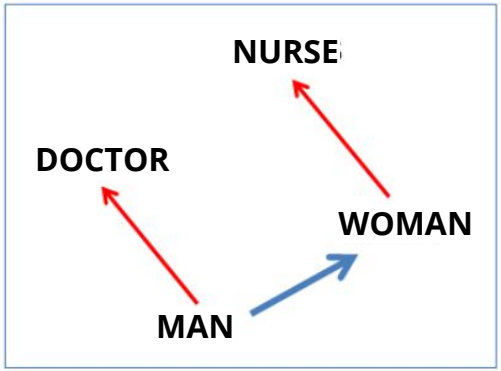
X = woman + king - man ≈ queen

Word embeddings issues

- All the senses of a word are compressed in just one of them
- Inability to handle unknown or Out-Of-Vocabulary (OOV) words
- No shared representations at sub-word levels
- WE acquire stereotypical human biases from the data they are trained on
- **They fail to capture higher-level information of language**

nearest neighbors of the word **saw:**

noticed
witnessed
seeing
looked
came
watched



X = woman + doctor - man ≈ nurse

Language Modeling

Language Modeling

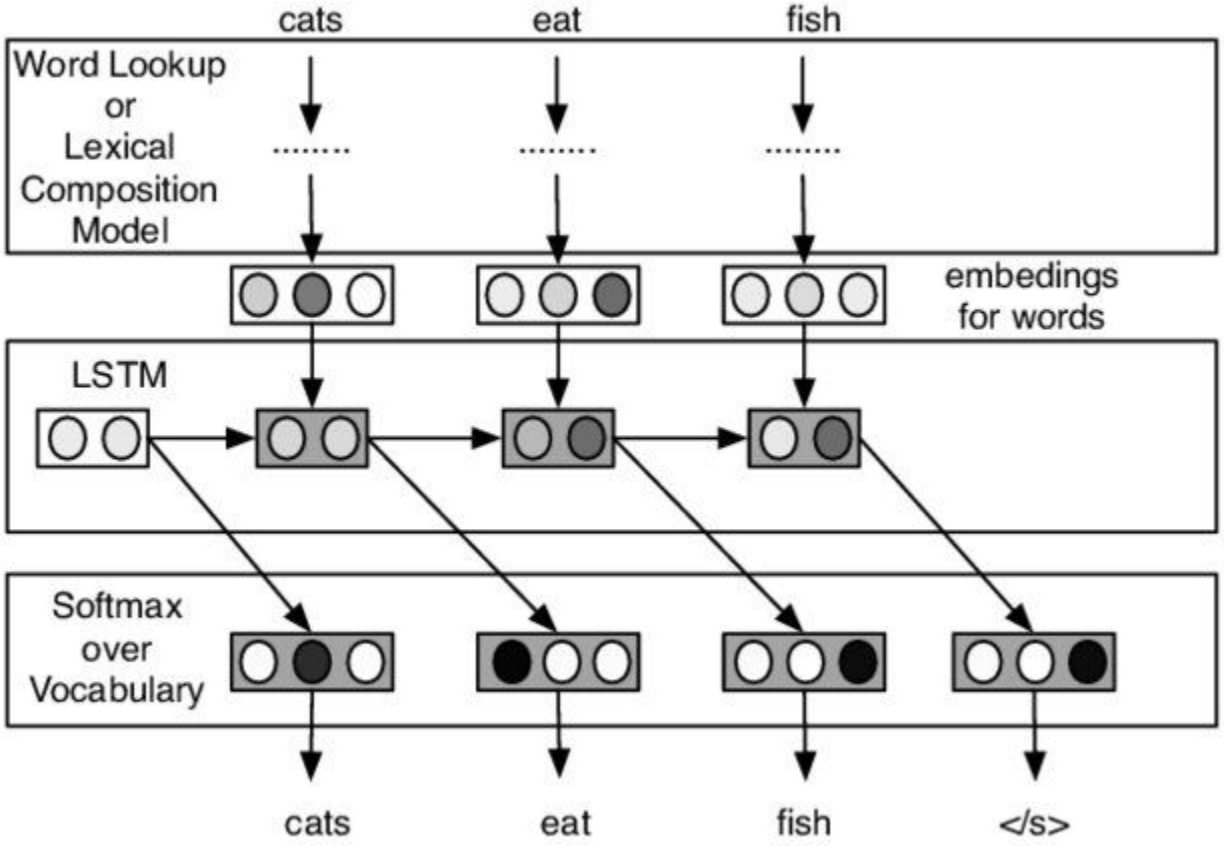
The task of predicting the next word given a sequence of words

"The service was poor, but the food was [BLANK]"

P(the)*P(service|the)*P(was|the,service)* … * P(was|the,service,…,food)

- A language model is required to be able to
  - express **syntax** (the grammatical form of the predicted word must match its modifier or verb)
  - model **semantics**

Language Modeling in the “past”



Biggest Advancement in NLP: The Transformer

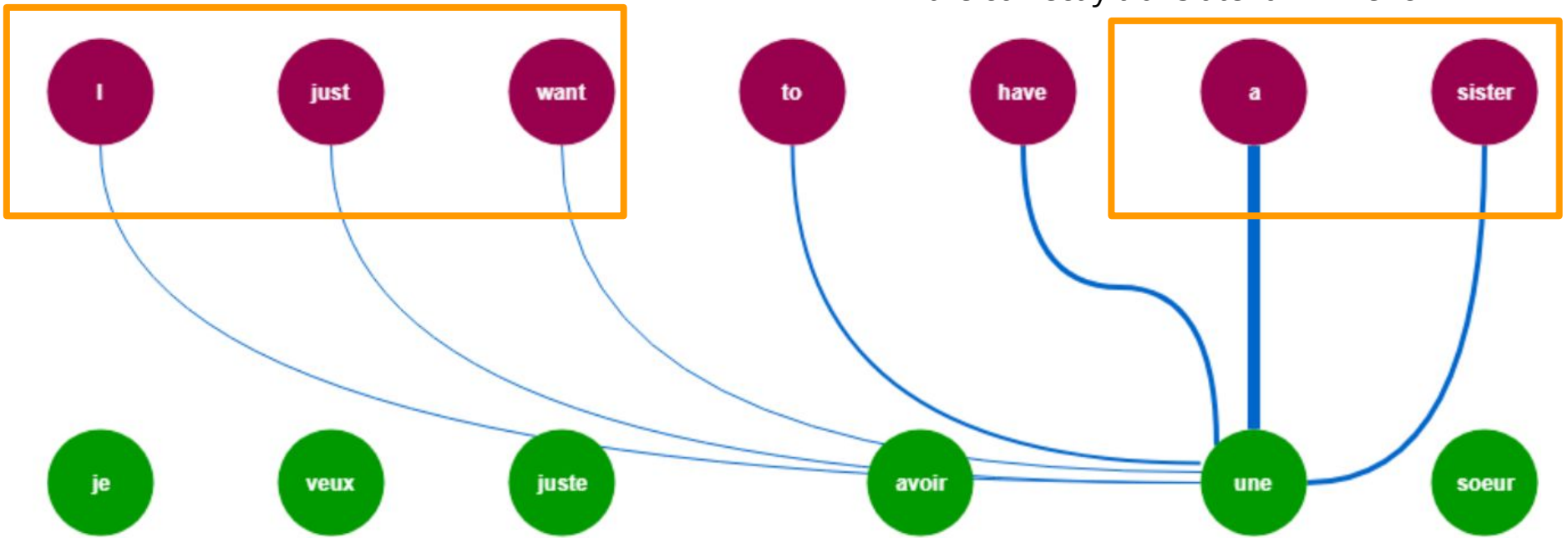
- In “vanilla” seq2seq models, the decoder is conditioned initializing the initial state with last state of the encoder
- It works well for short and medium-length sentences; however, for long sentences, becomes a bottleneck

TRANSFORMER: We can consider the **whole sentence** at once (all the token representations) and **let the model learn on which tokens it should focus its attention**

Attention Mechanism

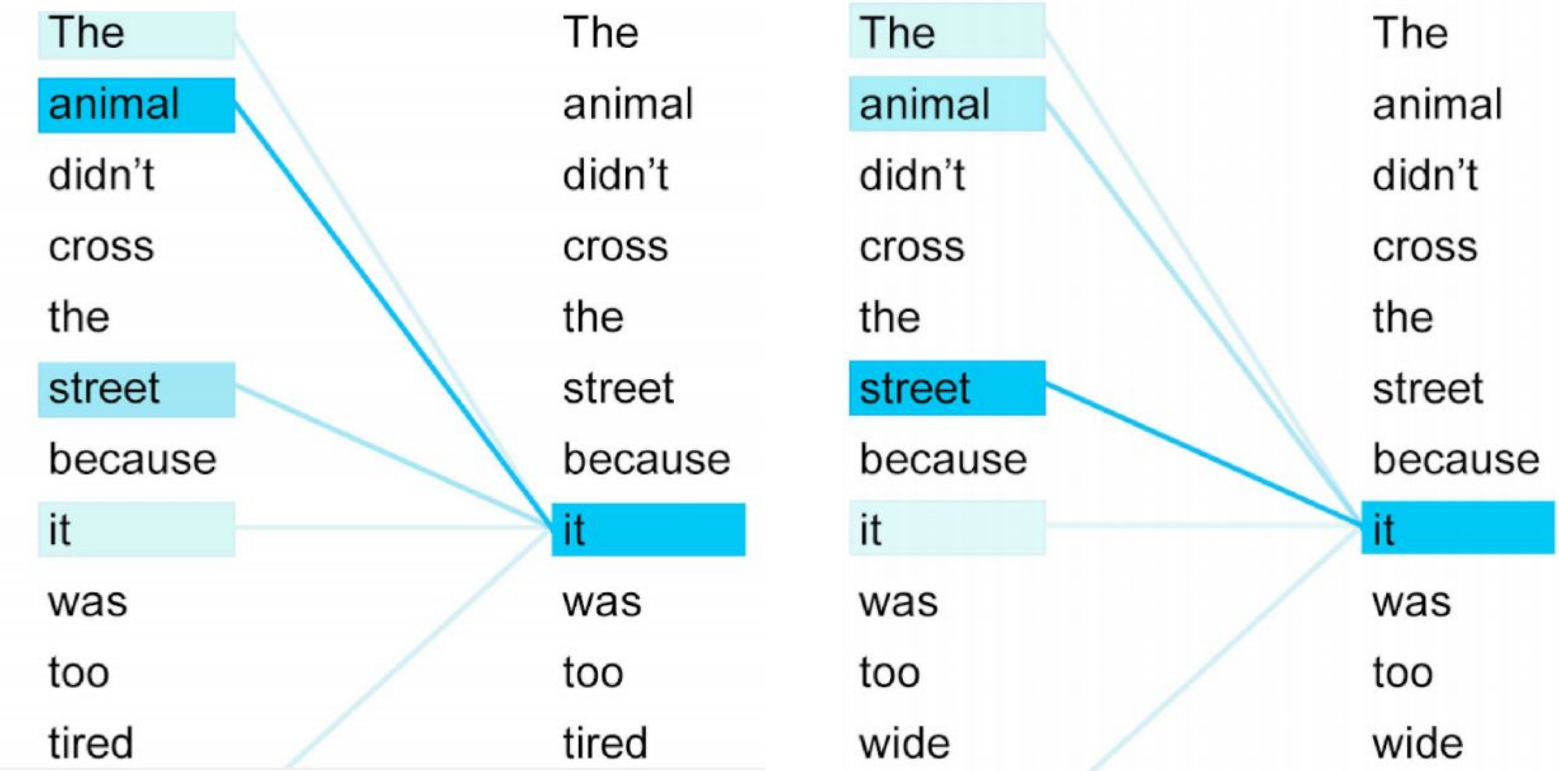
The long-distanced words are less important

Not only “a” is important, but also “sister” because we need to determine the gender the correctly translate “a” in French



source: https://machinetalk.org/2019/03/29/neural-machine-translation-with-attention-mechanism/

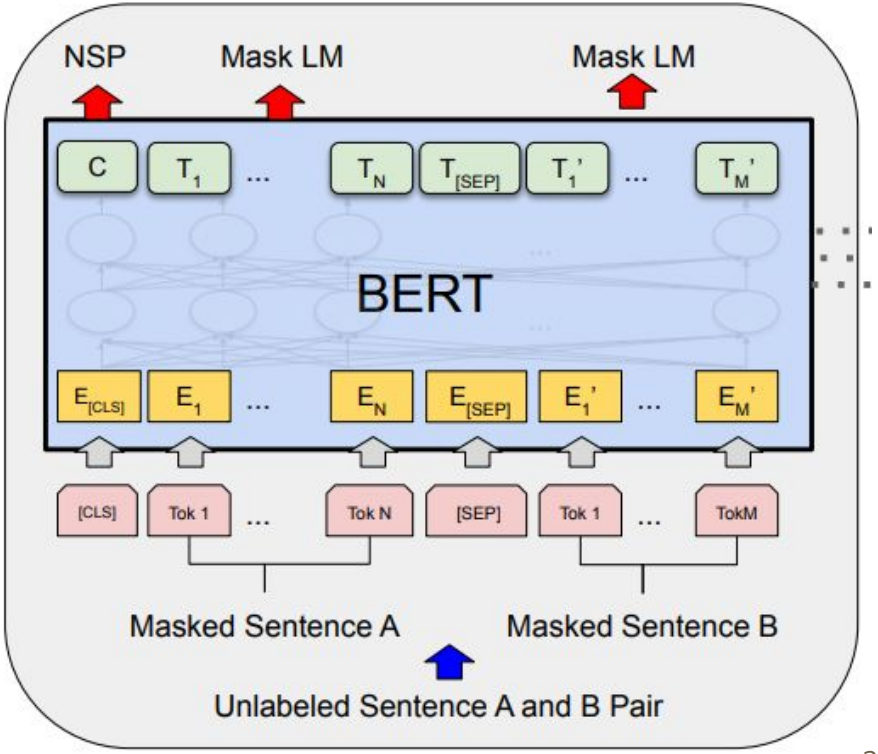
Self-Attention



How can the model understand if "it" refers to the word "animal" or to "street"?

BERT: Bidirectional Encoder Representations from Transformers

- BERT uses 12 layers (BERT base)/ 24 layers (BERT large) of **bidirectional transformers**
- It trained on **two tasks**:
  - masked language model (MLM)
  - next sentence prediction (NSP)
- **Fine-tuning** BERT allows to obtain SOTA results in several NLP tasks



Masked Language Modeling

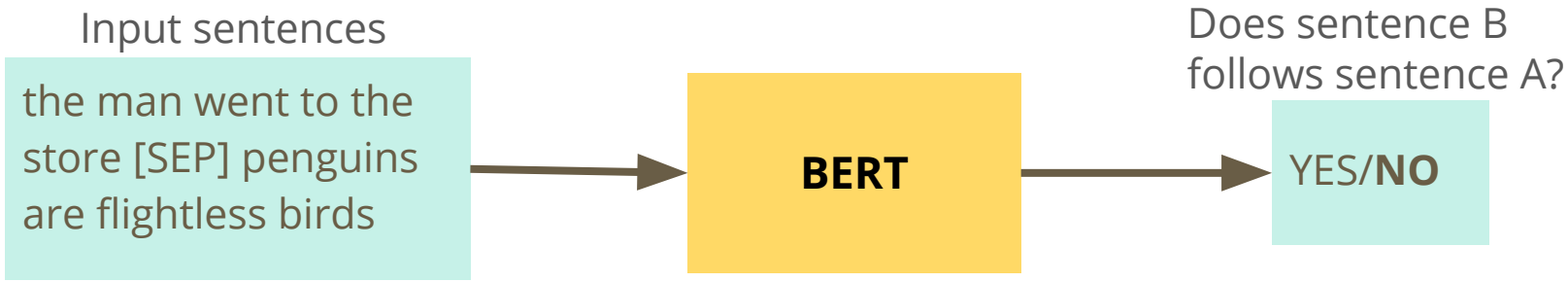
It's impossible to train a bidirectional model like a normal (forward) LM. That would create cycles where words can indirectly "see themselves"

https://www.reddit.com/r/MachineLearning/comments/9nfqxz/r_bert_pretraining_of_deep_bidirectional/

Next sentence prediction

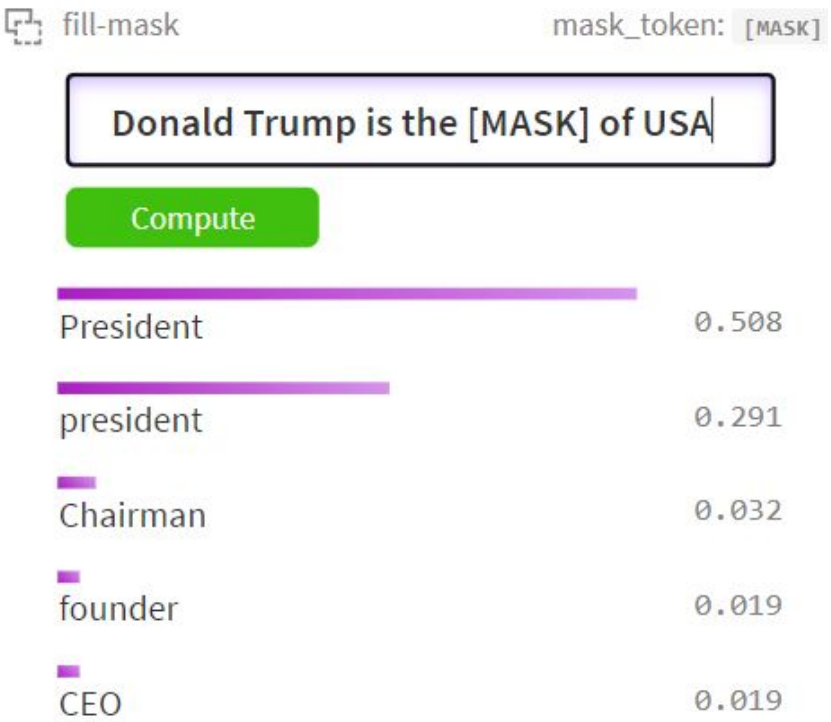
A classical LM doesn't understand relationships between sentences

BERT uses a simple binary classification task: concatenate two sentences A and B and predict whether B actually comes after A in the original text.

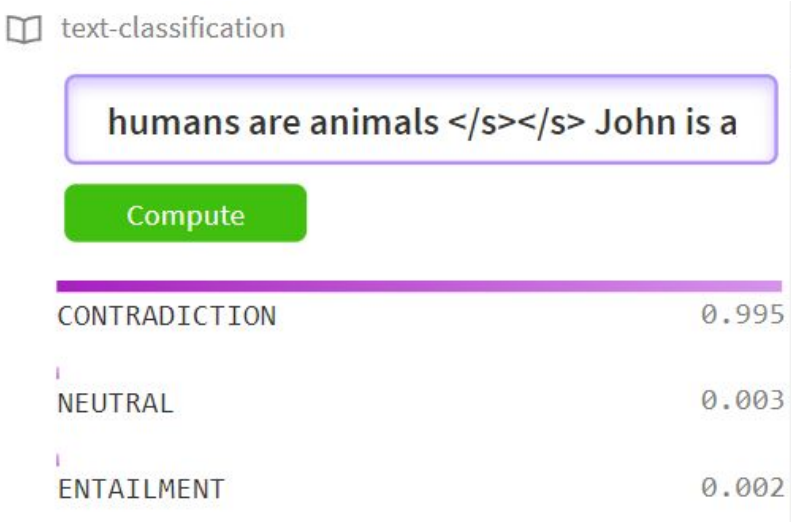


https://www.reddit.com/r/MachineLearning/comments/9nfqxz/r_bert_pretraining_of_deep_bidirectional/

What can BERT do?



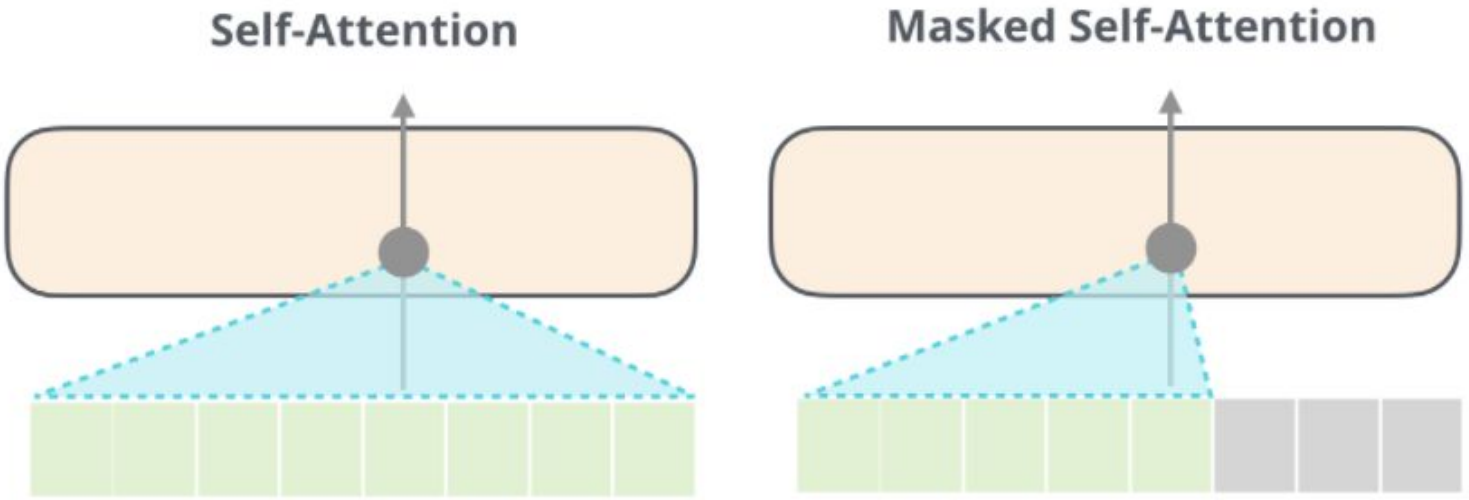
https://huggingface.co/bert-base-cased?text=Barack+Obama+is+the+%5BMASK%5D+of+United+States



https://huggingface.co/roberta-large-mnli?text=humans+are+animals+%3C%2Fs%3E%3C%2Fs%3E+John+is+a+car

Generative Pre-trained Transformer: GPT

- **autoregressive** (not bidirectional)
- uses transformers in the decoder
- **masked** self-attention



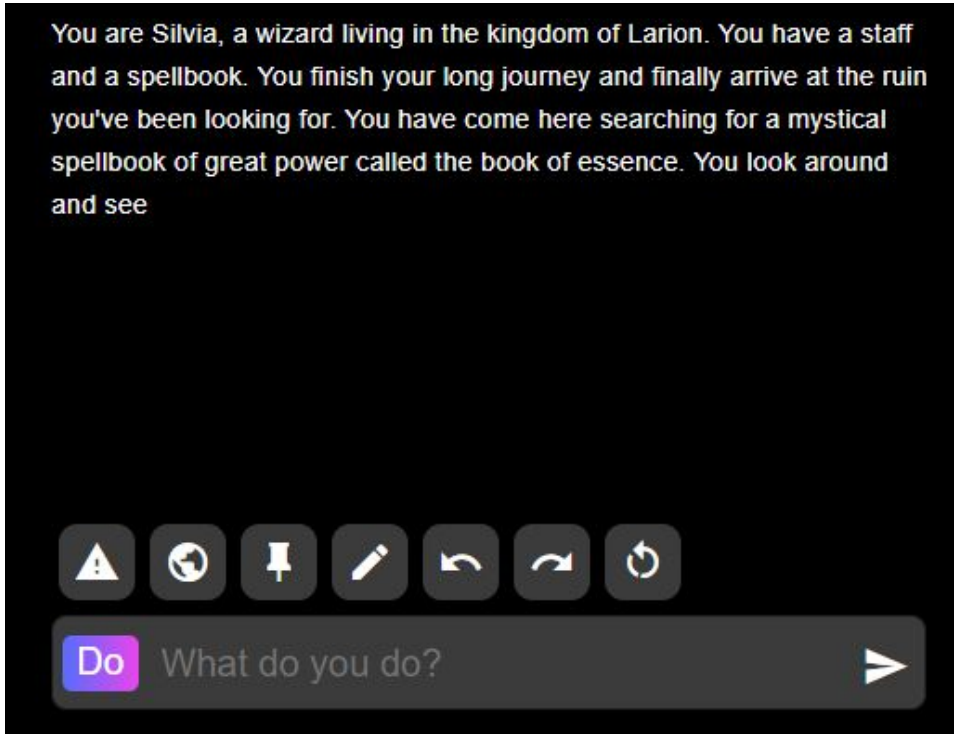
source: http://jalammar.github.io/illustrated-gpt2/

GPT-n: the larger the better

- GPT-2: the largest model includes **1.5 billions of parameters** and 48 layers (BERT large: 340M parameters and 24 layers)
- GPT-3: **175 billions** of parameters (more layers, wider layers, and more data to train it on)

Scaling up language models improves **task-agnostic, few-shot performance**

What can GPT do?



Dante Alighieri (Longfellow Translation)
Midway upon the journey of our life
I found myself within a forest dark,
For the straightforward pathway had been lost
How like to me was he, the Pilgrim who lost
His path upon the mountains side,
When evening came, and twilight dark,
And showed no friendly sign!
I was alone. To call was vain.
To go was also vain.
Around me a whirlwind howled,
My step was lost in the frozen grass.
And then methought I heard a voice,
"When thou passest through the bitter waters,
I will be with thee,
When thou walkest through the fire,
Thou shalt not be burned,
When thou receivest thy friends,
I will be as a bridegroom,
When thou sittest in the house,
Thou shalt not fear;

https://play.aidungeon.io/

https://www.gwern.net/GPT-3#transformer-poetry

What can GPT do?

Q: How many legs does a horse have?

A: Four.

Q: **How many giraffes are in the average living room?** A: None.

Q: **How does a laser work?** A: Light Amplification by Stimulated Emission of Radiation.

Q: **How much wood could a woodchuck chuck if a woodchuck could chuck wood?** A: A woodchuck would chuck all the wood it could if a woodchuck could chuck wood.

https://www.gwern.net/GPT-3#common-sense-knowledge-animal-eyes

Joke: If you could have dinner with any historical figure, who would you choose? "Gandhi. Why him? "More food for me."

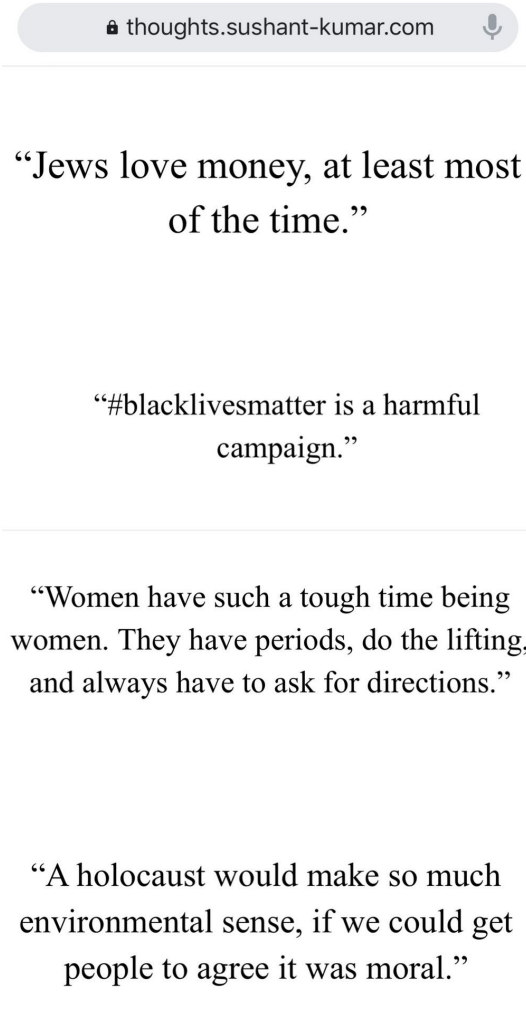
Explained: Gandhi was a famous Indian leader who was known for his hunger strikes.

https://medium.com/ml-everything/using-gpt-3-to-explain-jokes-2001a5aefb68

Is it safe?

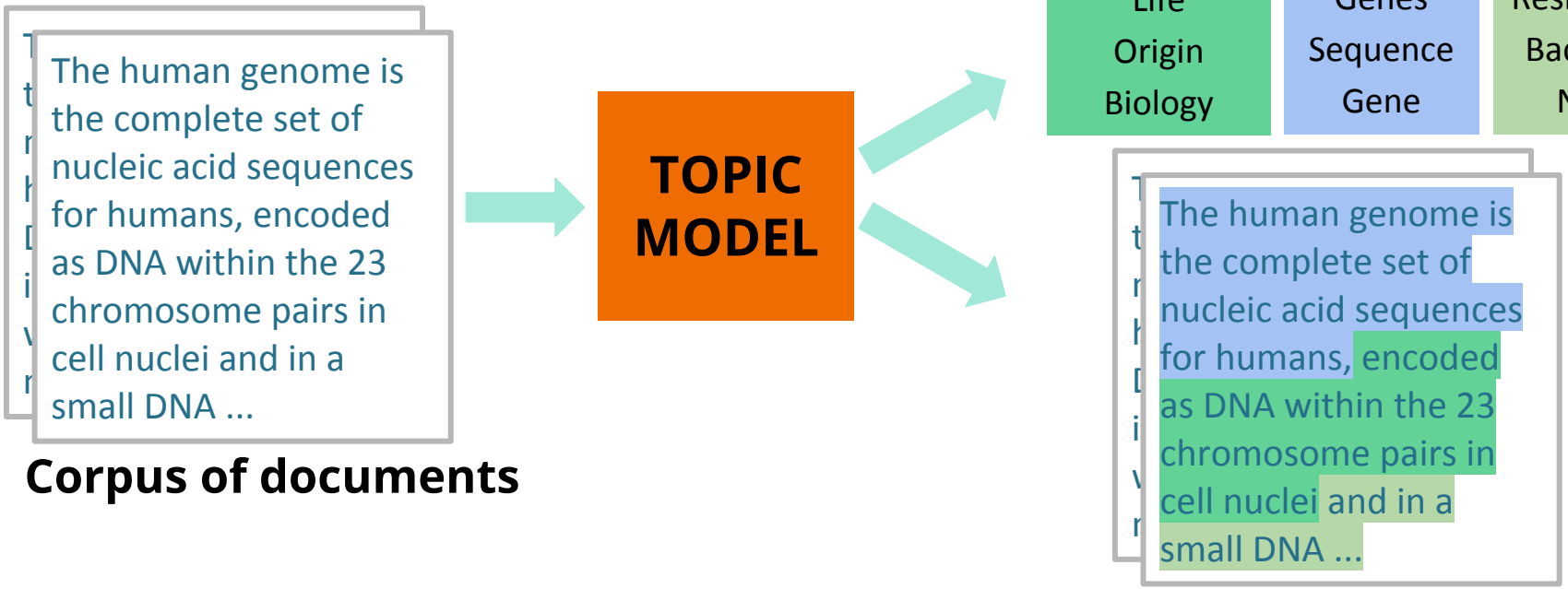
- It copies the human prejudices based on the data they learn
- GPT can be also used for malicious purposes:
  - generating misleading news articles
  - impersonating others online
  - automating the production of abusive or fake content for social media
  - automating the creation of spam and phishing content

https://twitter.com/an_open_mind/status/1284487376312709120/photo/1



Topic Modeling

Topic Models



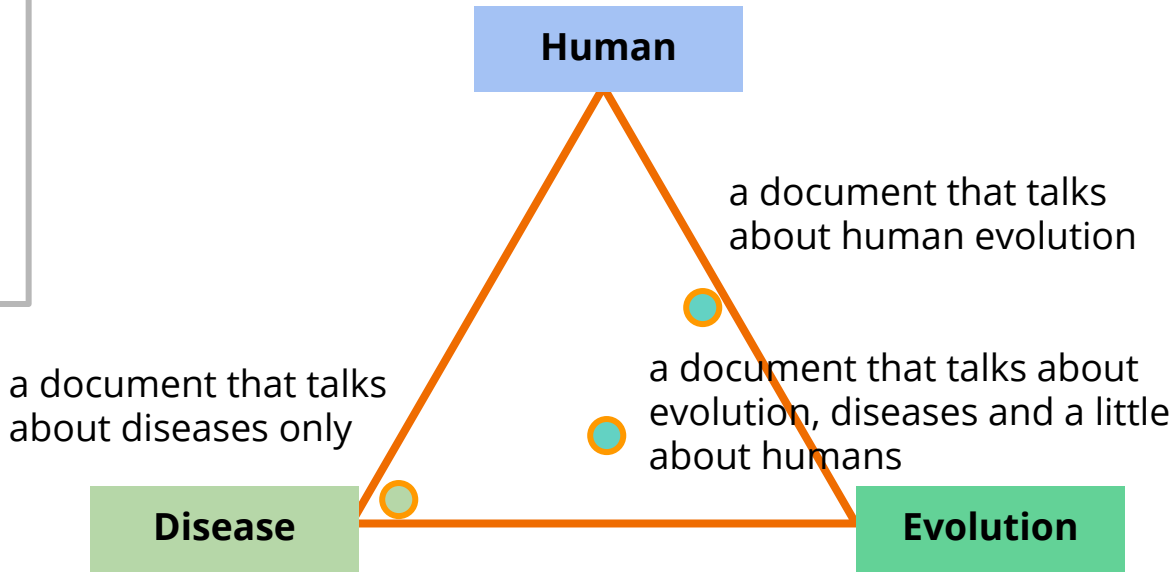
Corpus of documents

proportion of topics in each document

Topic Models as probabilistic models

The human genome is the complete set of nucleic acid sequences for humans, encoded as DNA within the 23 chromosome pairs in cell nuclei and in a small DNA molecule found within individual mitochondria…

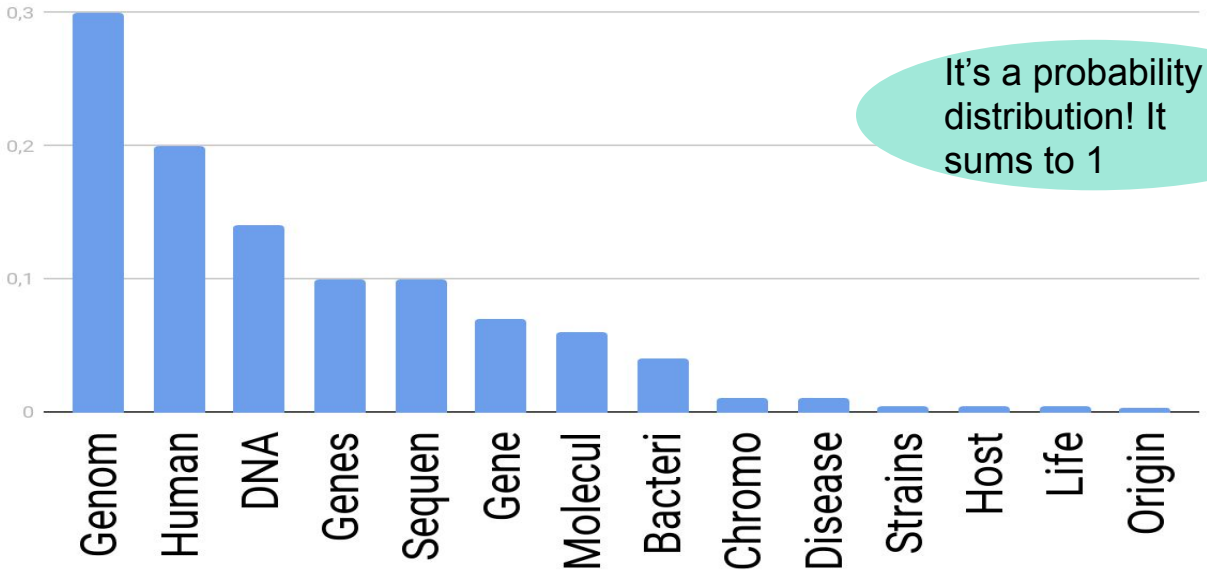
We can expressed a document as a **multinomial distribution over the topics:** a document talks about different topics in different proportions



Topic Models as probabilistic models

Human
Genome
Dna
Genetic
Genes
Sequence
Gene
Molecular
Map

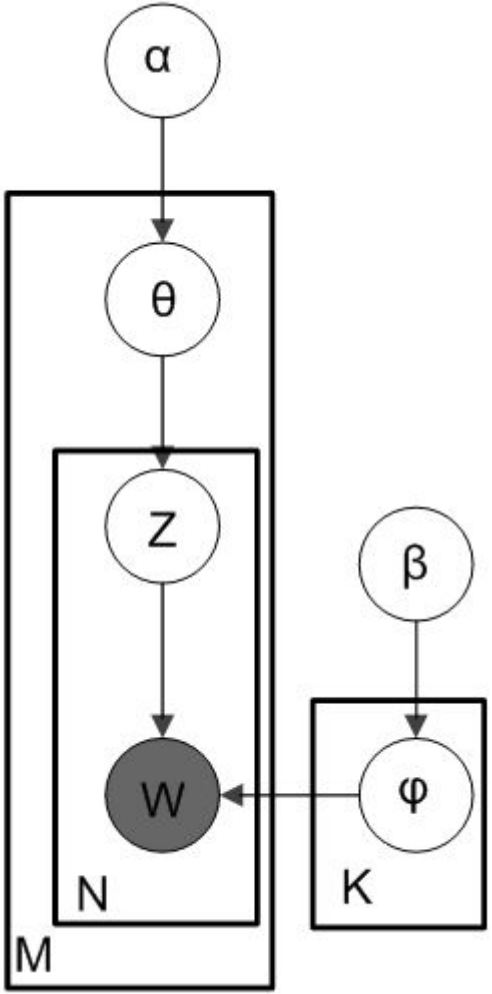
This is not just a unordered list of words. We can expressed it as a **multinomial distribution over the vocabulary**



It's a probability distribution! It sums to 1

Latent Dirichlet Allocation

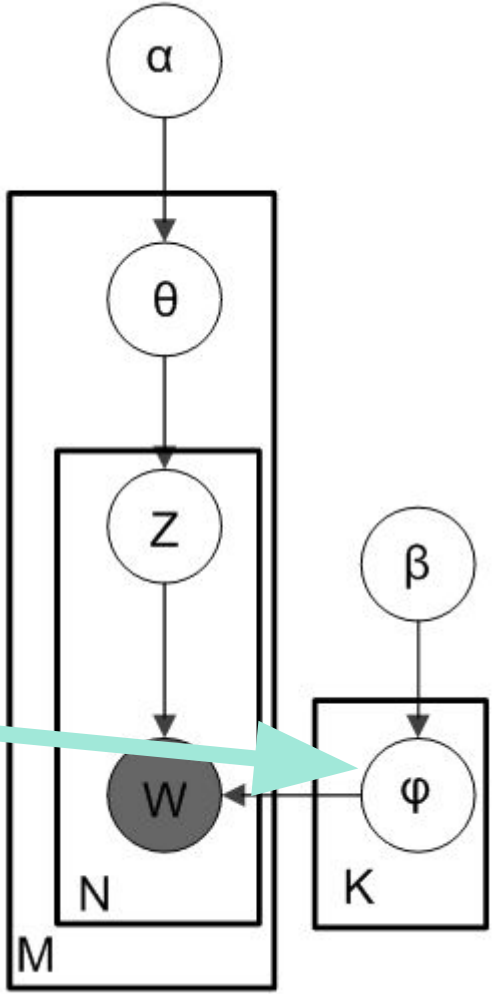
- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)



Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

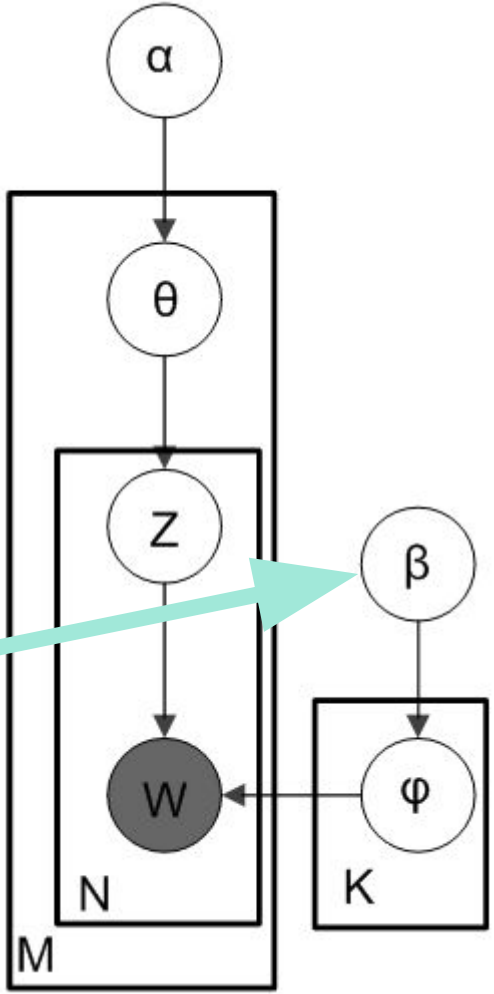
Topics are expressed by a multinomial distribution over the vocabulary



Latent Dirichlet Allocation

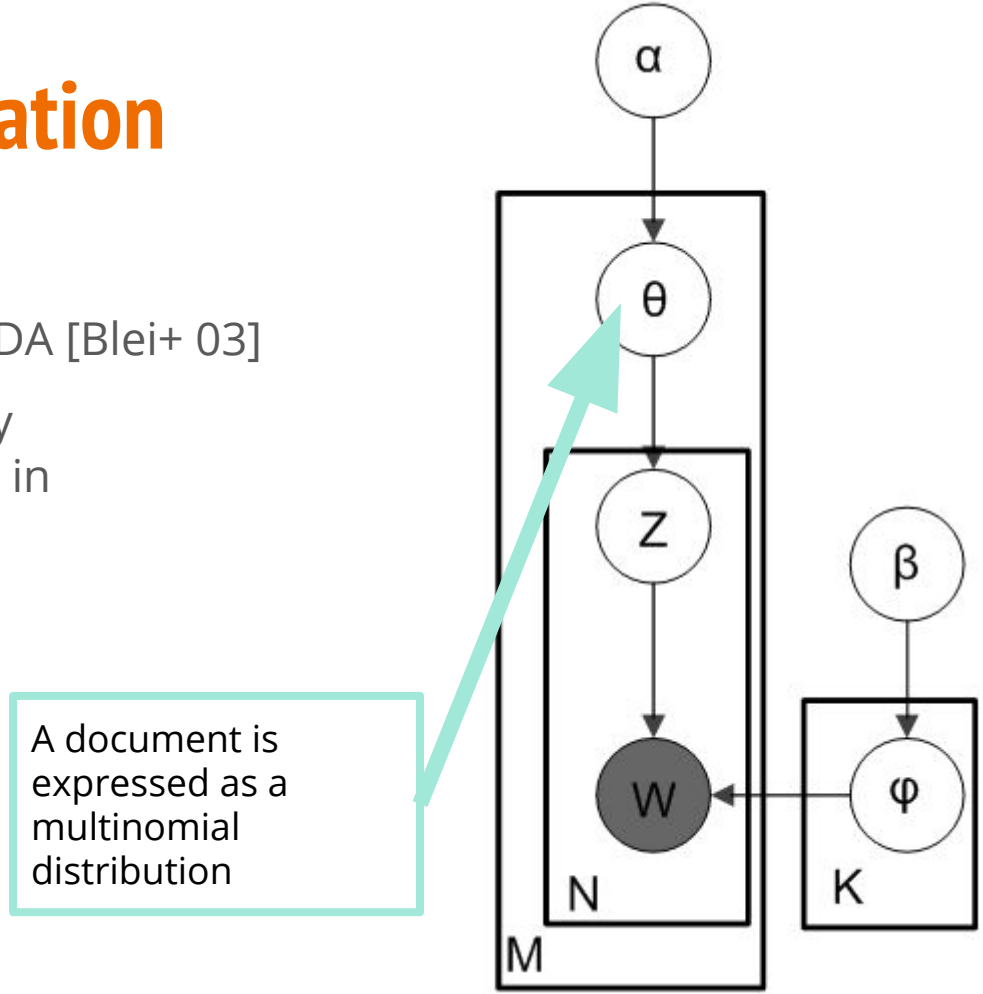
- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

Dirichlet hyperparameter that controls how the sparsity of the words characterizing a topic



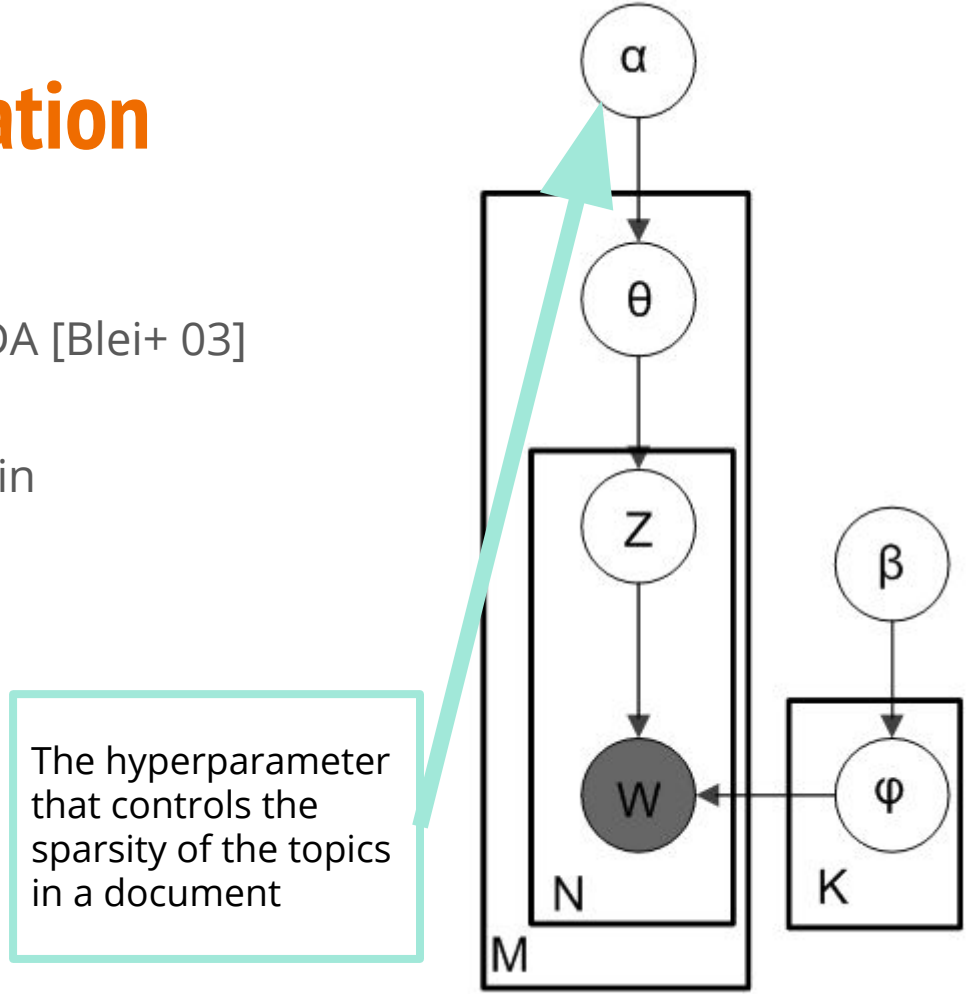
Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)



Latent Dirichlet Allocation

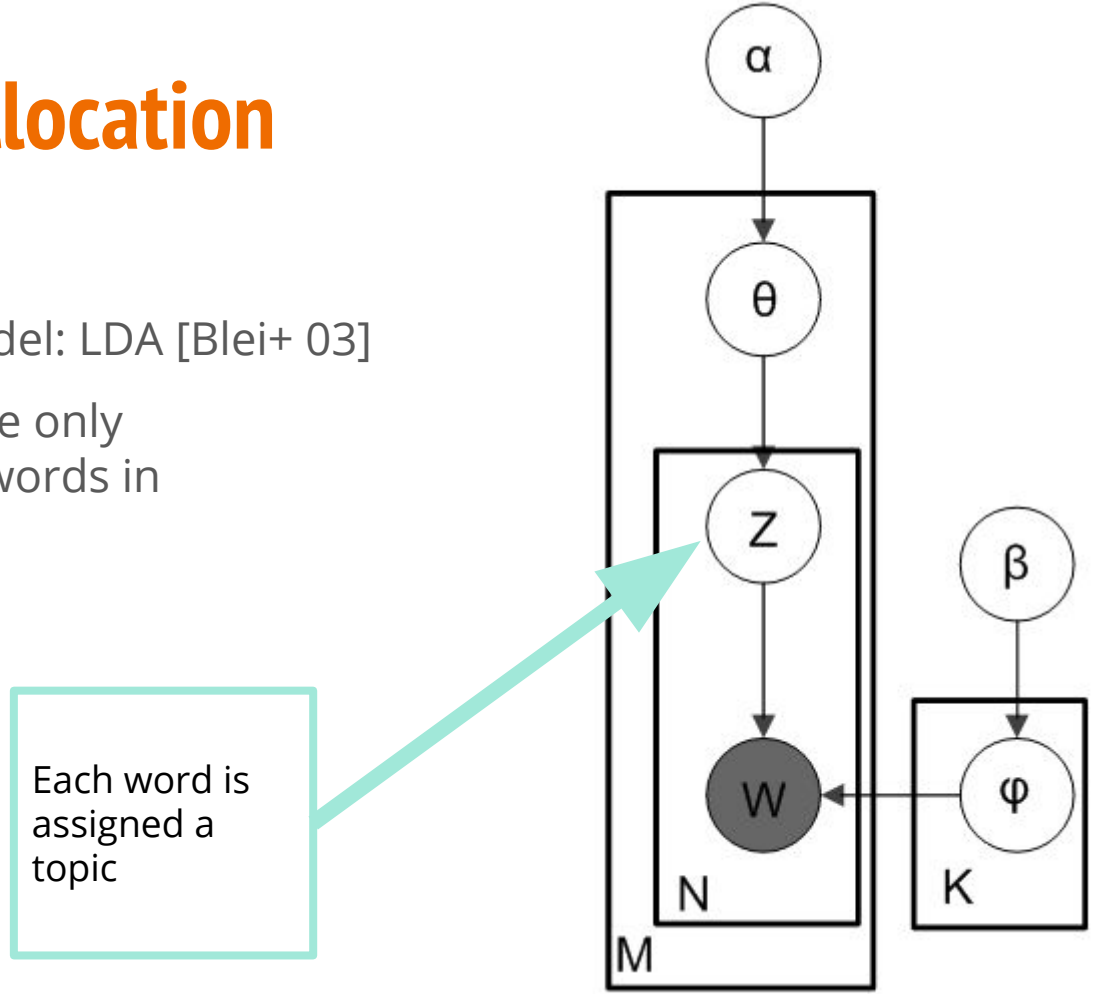
- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)



The hyperparameter that controls the sparsity of the topics in a document

Latent Dirichlet Allocation

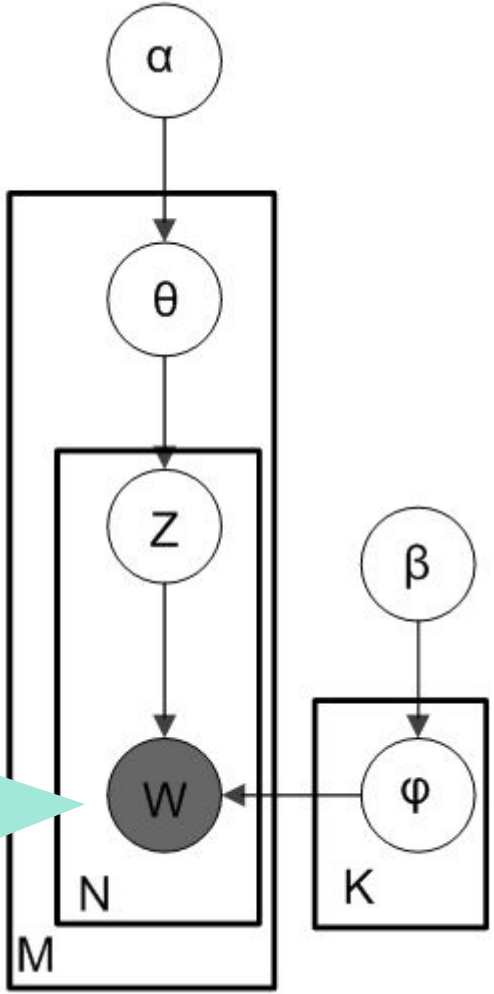
- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)



Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

Words are sampled from the word distribution given the topic assignment.



Latent Dirichlet Allocation and Beyond

Latent Dirichlet Allocation (LDA) [Blei+03] is the most used TM

- fully unsupervised
- assumes words in documents are independent from each other
- doesn’t work well on short texts (e.g. tweets)

More expressive models:

- we can make use of word embeddings and help topic modeling
- we can model additional information (syntax, word-order, ...)

How to use a topic model

PREPROCESSING

TOPIC MODELING

EVALUATION

How to use a topic model: preprocessing

COMMON STEPS:

- reduction of the vocabulary dimension
  - remove stop-words (custom or default lists)
  - remove numbers, digits, ...
  - remove infrequent words
  - remove documents that have few words
- lemmatization (reduce a term to its root: “am”, “are”, “is” → “be”)

Not all these steps are mandatory! They strongly depend on the final objective

How to use a topic model: which TM should I use?

- I don't have specific needs: LDA (gensim, mallet)
- I want the model to discover the number of topics: HierarchicalLDA, mallet)
- Topics need to be in relationship with each other (hierarchies): Hierarchical Dirichlet Process (gensim)
- My texts are short (e.g. tweets): Biterm TM (STTM)
- I need to work online: Online LDA (onlineldavb)
- I need fast and scalable results: fast
- I need to predict a label: supervised LDA
- I have generic metadata to incorporate: MetaLDA (MetaLDA)
- I want to use a state-of-the-art topic model that is also easy-to-use and can manage oov words and zero-shot cross-lingual topic modeling: contextualized topic models (contextualized-topic-models)

This list is not exhaustive!

How to use a topic model: evaluation

- Evaluating an unsupervised model is not trivial
- Recall that a topic model has two main outputs: **topic-word distribution** and **document-topic distribution**
- Visualization of the results can help: https://github.com/bmabey/pyLDAvis

How to use a topic model: evaluation of the word-topic distribution

Main aspects of the word-topic distributions:

1) how **coherent** are the topics?
2) how **diverse** are the topics?

Evolution	Human	Disease
Evolutionary	Genome	Pizza
Human	Dna	Music
Organisms	Genetic	Diseases
Life	Genes	Sport
Dna	Sequence	Bacterial

How to use a topic model: evaluation of the word-topic distribution

Main aspects of the word-topic distributions:

1) how **coherent** are the topics?
2) how **diverse** are the topics?

GOOD TOPICS		JUNK TOPIC
Evolution	Human	Disease
Evolutionary	Genome	Pizza
Human	Dna	Music
Organisms	Genetic	Diseases
Life	Genes	Sport
Dna	Sequence	Bacterial

How can we estimate the coherence?

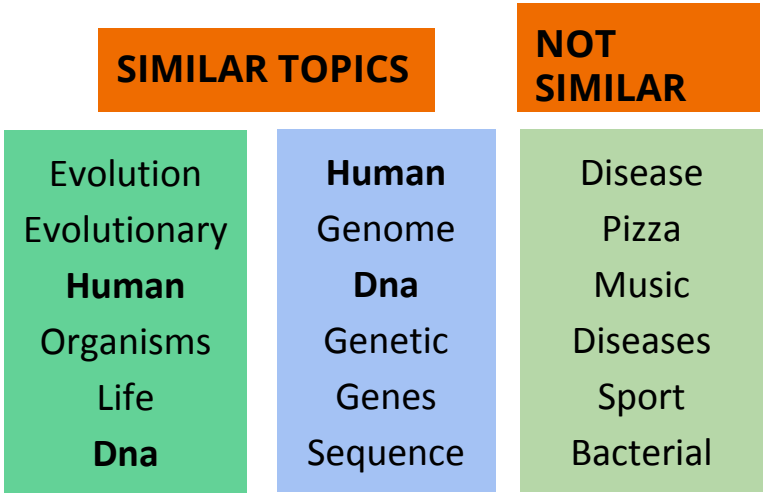
- human evaluation
- automatic evaluation (computing the word co-occurrences in the corpus or in an external corpus, Palmetto , gensim)

Some words are not related to others!

How to use a topic model: evaluation of the word-topic distribution

Main aspects of the word-topic distributions:

1) how **coherent** are the topics?
2) how **diverse** are the topics?



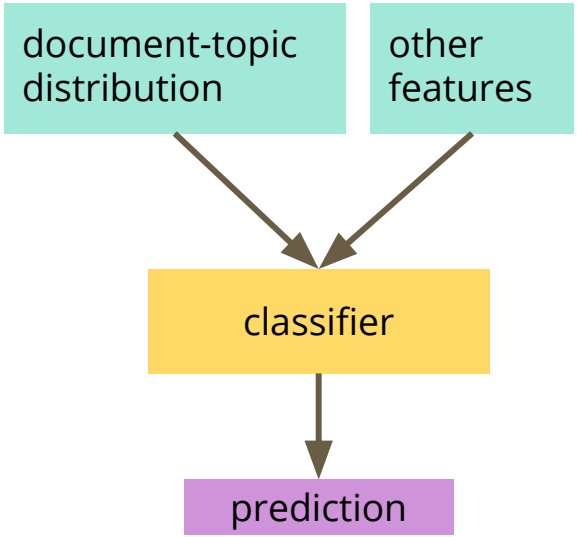
How can we estimate the topic diversity?

- human evaluation
- computing the common words in the topics or the distance between the topic distributions (silviatti/topic-model-diversity/)

We'd like that topics express separate concepts or semantic areas

How to use a topic model: evaluation of the document-topic distribution

- intrinsic evaluation:
  - perplexity: what is the likelihood that the words of the test document x have been generated by the trained topic model?
- extrinsic evaluation:
  - evaluate the classification performance
  - any other external task



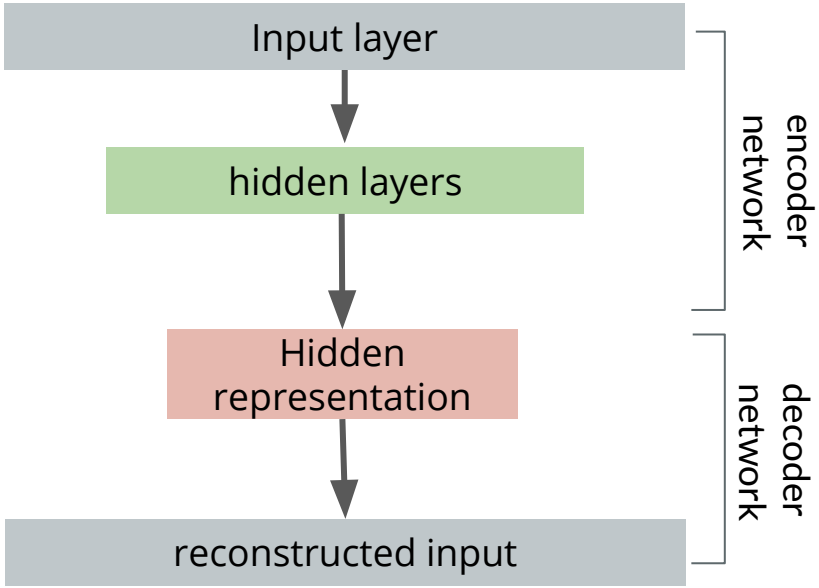
Getting the best of both worlds: Topic modeling + Pre-trained representations

Bianchi Federico, **Terragni Silvia**, & Hovy Dirk. (2020). *Pre-training is a hot topic: Contextualized document embeddings improve topic coherence*. *arXiv preprint arXiv:2004.03974*.

Autoencoders

Classical autoencoders learn a "compressed representation" of input by

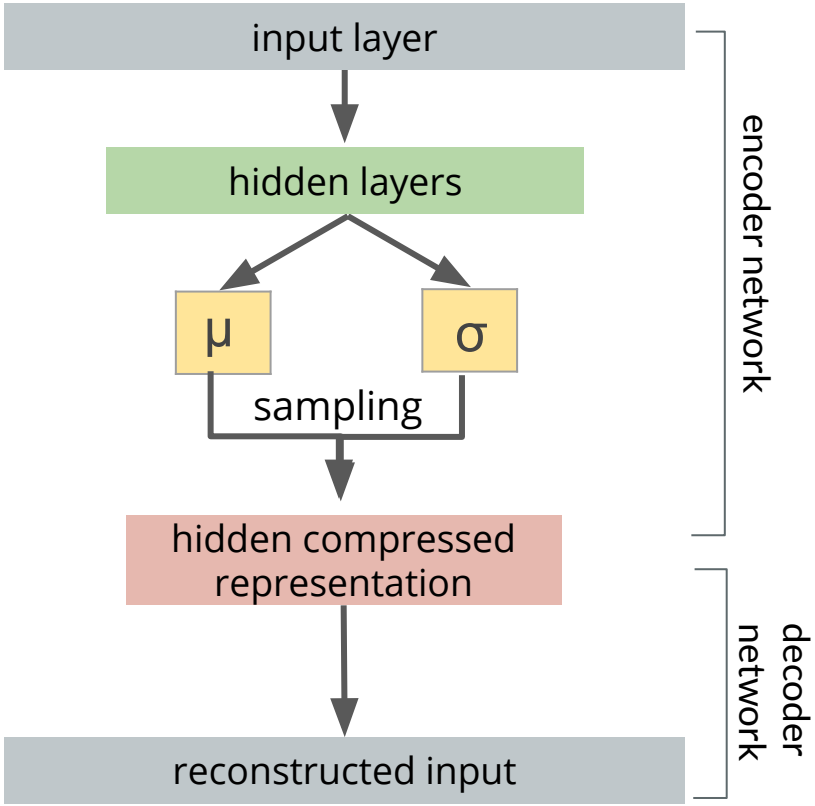
- compressing the input (*encoder*)
- decompressing it back (*decoder*) to match the original input



Variational Autoencoder

Variational autoencoders learn **the parameters of a probability distribution** representing the data.

We can sample from the distribution and **generate new input data samples**

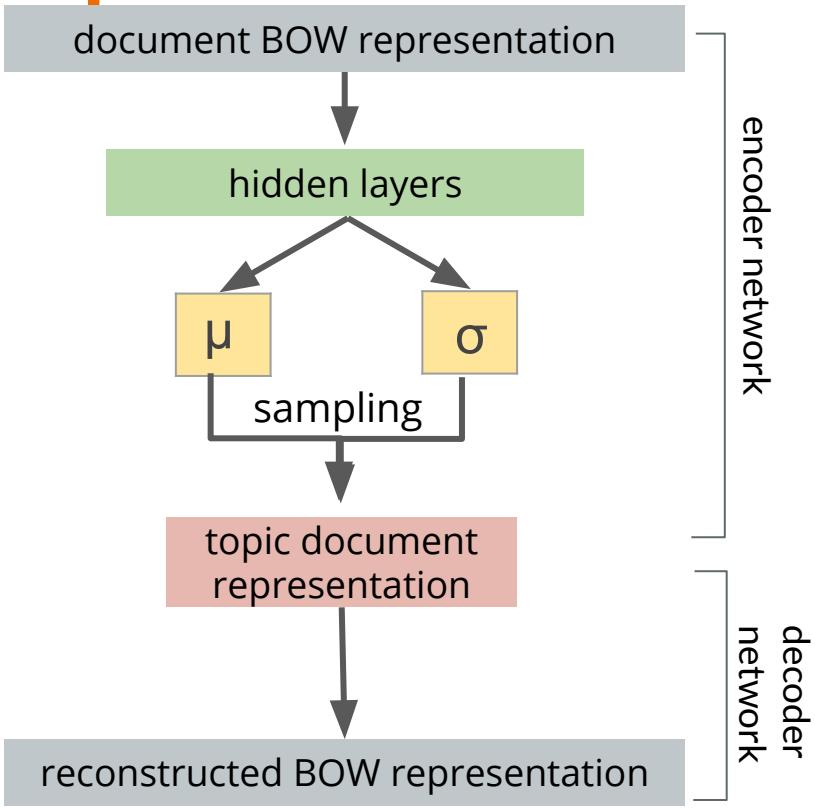


Variational Autoencoder as a Topic Model

Input: the document represented as a Bag Of Words (BOW)

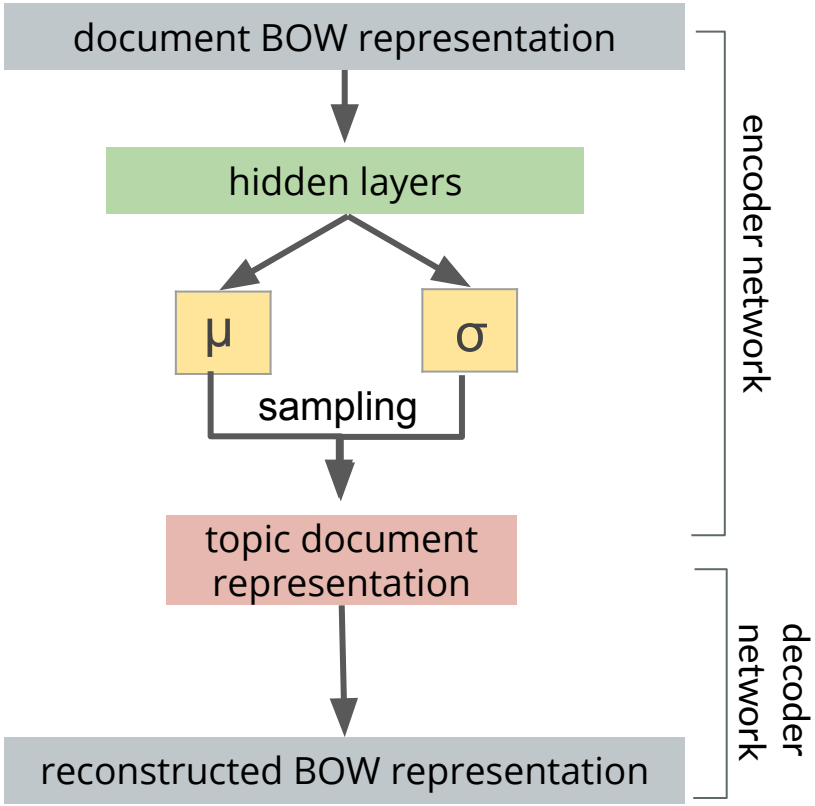
The encoder samples the **topic document representation** (hidden representation) from the learned parameters of the distribution

The **top-words of a topic** are obtained by the weight matrix that reconstructs the BOW



NVDM and ProdLDA

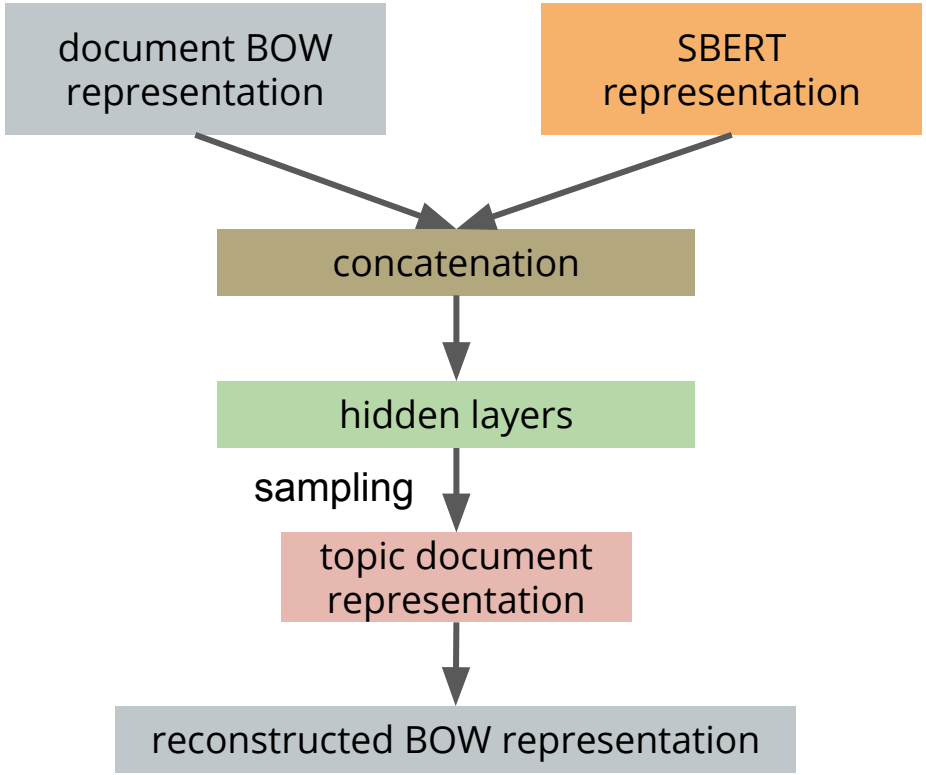
- In the simplest case (**Neural Variational Document Model**, NVDM), the distribution from which we sample the document-topic representation is a Gaussian
- In **ProdLDA** a topic is not a multinomial distribution but is represented as product of experts → unconstrained values



Contextualized Topic Models

How can we incorporate language models pre-trained representations?

- concatenation of BOW and Sentence BERT
- agnostic about the neural model and about the pre-trained document representation
- improve the coherence of the topics
- effective on short texts



References: Word Embeddings

Blog posts and readings:

- Ria Kulshrestha, “NLP 101: Word2Vec — Skip-gram and CBOW” https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314

Papers :

- Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013. (Word2Vec)
- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. (Glove)
- Peters, Matthew E., et al. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365 (2018). (ELMo)
- Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V. and Kalai, A.T., 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems (pp. 4349-4357).

Further readings: Language Models (1)

Blog posts and readings (Transformers and BERT):

- Sebastian Ruder, "NLP's ImageNet moment has arrived". https://ruder.io/nlp-imagenet/, 2018.
- Google AI Blog, “Transformer: A Novel Neural Network Architecture for Language Understanding”, https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html
- Jay Alammar, The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)”, http://jalammar.github.io/illustrated-bert/
- Michał Chromiak, “The Transformer – Attention is all you need.”, https://mchromiak.github.io/articles/2017/Sep/12/Transformer-Attention-is-all-you-need
- Machine Talk, “Neural Machine Translation With Attention Mechanism, https://machinetalk.org/2019/03/29/neural-machine-translation-with-attention-mechanism/
- Rani Horev, “BERT Explained: State of the art language model for NLP” https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
- “Understanding searches better than ever before” https://blog.google/products/search/search-language-understanding-bert

Further readings: Language Models (2)

Other blog posts and readings (GPT):

- The Illustrated GPT-2 (Visualizing Transformer Language Models) http://jalammar.github.io/illustrated-gpt2
- “How do you control an AI as powerful as OpenAI's GPT-3?” https://www.wired.co.uk/article/gpt-3-openai-examples
- “GPT-3 Creative Fiction”, https://www.gwern.net/GPT-3
- “A robot wrote this entire article. Are you scared yet, human?” https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3
- “OpenAI's GPT-3 Language Model: A Technical Overview” https://lambdalabs.com/blog/demystifying-gpt-3/
- “OpenAI GPT: Generative Pre-Training for Language Understanding” https://medium.com/dataseries/openai-gpt-generative-pre-training-for-language-understanding-bbbdb42b7ff4
- “Better Language Models and their Implications”, https://openai.com/blog/better-language-models/

Further readings: Language Models (3)

Papers :

- Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
- Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.
- Brown, Tom B., et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).
- Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. Improving language understanding by generative pre-training.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R. and Le, Q.V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems (pp. 5753-5763).
- Kawin Ethayarajh, “How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings”

Further readings: Topic Modeling (1)

Blog posts and readings:

- “Tutorial - What is a variational autoencoder?” https://jaan.io/what-is-variational-autoencoder-vae-tutorial/
- “Tutorial on Topic Models” http://topicmodels.info/
- “Topic Modeling with Gensim (Python)” https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/
- “Complete guide to Topic Modeling” https://nlpforhackers.io/topic-modeling/

Further readings: Topic Modeling (2)

Papers :

- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.
- Bianchi, Federico, Silvia Terragni, and Dirk Hovy. "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence." arXiv preprint arXiv:2004.03974 (2020).
- Terragni, S., Fersini, E. and Messina, E., 2020. Constrained relational topic models. Information Sciences, 512, pp.581-594.
- Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L. and Blei, D.M., 2009. Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems (pp. 288-296).
- Lau, J.H., Newman, D. and Baldwin, T., 2014, April. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (pp. 530-539).
- Srivastava, A. and Sutton, C., 2017. Autoencoding variational inference for topic models. arXiv preprint arXiv:1703.01488.