Introduction to Natural Language Processing

The case of word embeddings

Speaker: *Yipeng Qin* ***Original Slides:*** *Jose Camacho Collados*

Outline

- Introduction to Natural Language Processing (NLP)
- The case of word embeddings

Natural Language Processing

Natural Language Processing

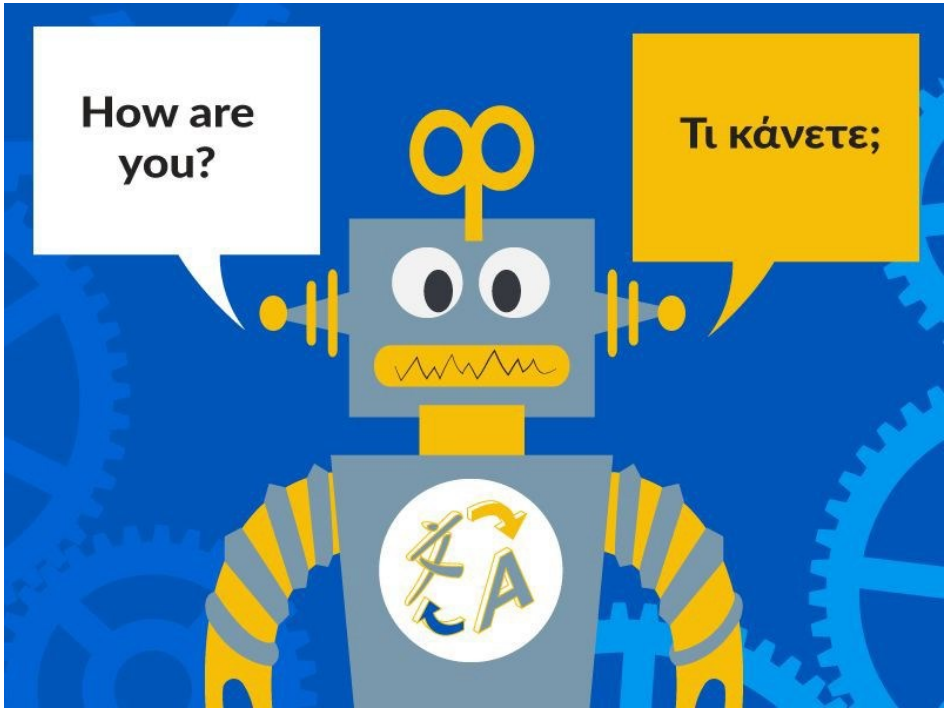
Branch of AI that deals with **human language**

Multi-billion industry

Usual suspects: spell checkers, autocomplete, machine translation, speech-to-text and text-to-speech systems, web search

Other applications: sentiment analysis, trend prediction, digital forensics, hate speech monitoring, fake news detection, recommendation systems.

Applications of Machine Learning



Machine Translation

Image credit: *Ciklopea*

Applications of Machine Learning

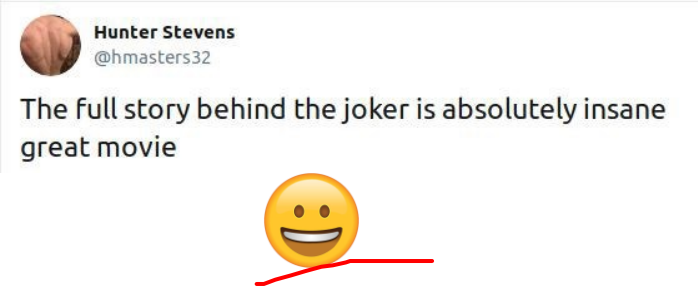
Question answering

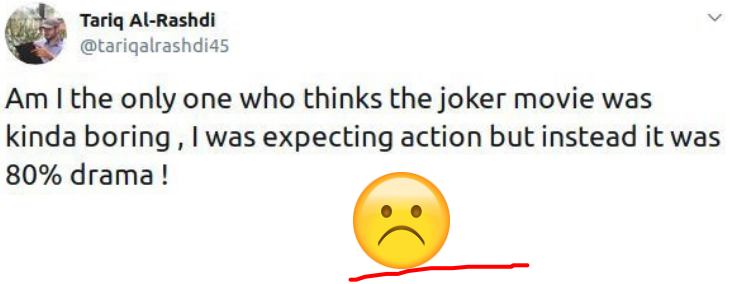
Image credit: *Jeopardy Productions*

Example: Sentiment analysis

We want to understand whether the audience has liked or not a new movie.

There are thousands of opinions from users in social media, fora, etc.

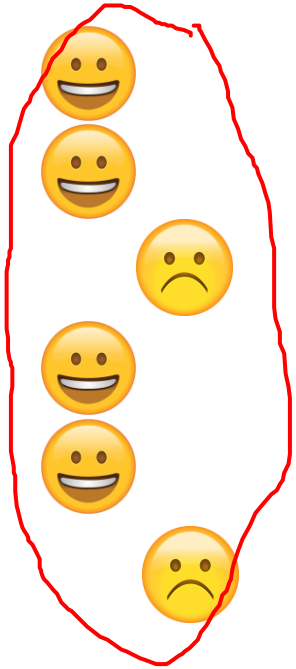




We can develop a **machine learning classifier** to understand users' opinions automatically!

Example: Sentiment analysis

- I liked the movie
- The movie was awesome
- It was quite boring
- I enjoyed the movie
- It was great!
- The main actor was terrible
- ...



Example: Sentiment analysis

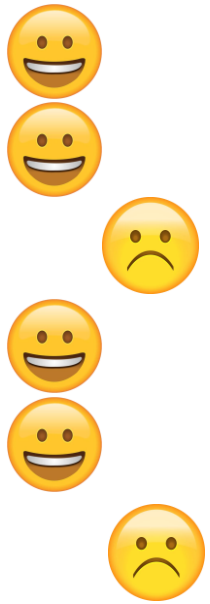
- I **liked** the movie
- The movie was **awesome**
- It was quite **boring**
- I **enjoyed** the movie
- It was **great**!
- The main actor was **terrible**

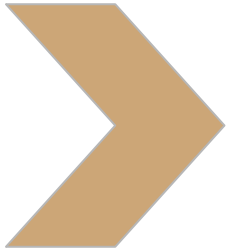
...

Example: Sentiment analysis

- I **liked** the movie
- The movie was **awesome**
- It was quite **boring**
- I **enjoyed** the movie
- It was **great**!
- The main actor was **terrible**

...





Train a machine learning classifier

Example: Sentiment analysis

The movie was **magnificent**!   **or**  **?**

I got **bored**   **or**  **?**

Example: Sentiment analysis

The movie was **magnificent**!

I got **bored**

Example: Sentiment analysis

The movie was **magnificent**!

I got **bored**

Problem: What if “magnificent” was not part of our training data?

Word embeddings

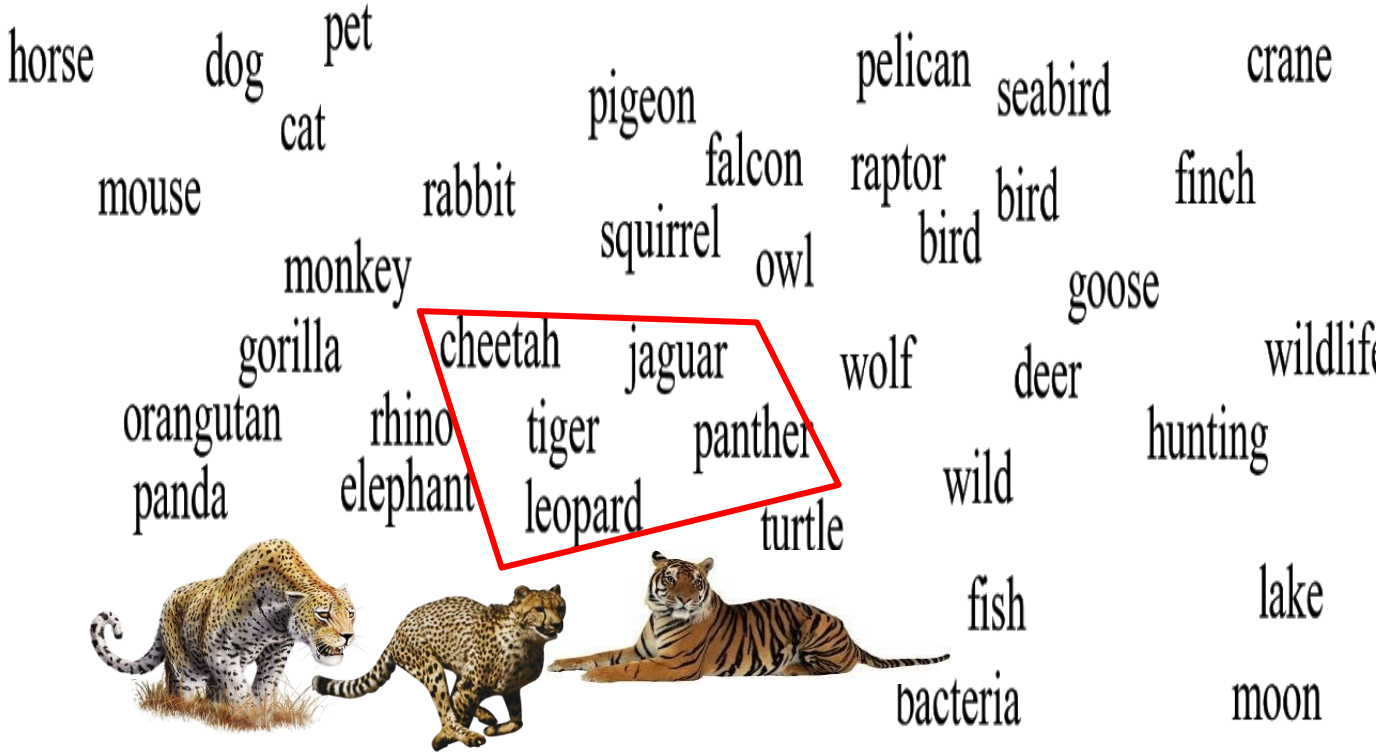
Word embeddings

Word embeddings are **vector spaces** where words are represented as points in the space.

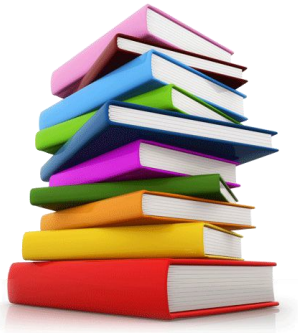
Similar words are represented close in the vector space.

Useful for many Natural Language Processing (**NLP**) applications.

Word embeddings



Word embeddings: How to learn them



... **London** is the **capital** of **UK** …

... Last night I **travelled** from **Cardiff** to **London**.

.

.

.

London

[0.25, 0.32, -0.1 …. 0.1]

Word2Vec (Mikolov et al. 2013)

Word2Vec is one of the pioneers works to learn word embeddings from text corpora.

The architecture is quite simple, a shallow neural network with a single hidden layer.

More about word2vec and word embeddings:

https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

https://arxiv.org/pdf/1301.3781.pdf (original paper)

https://sites.google.com/view/embeddings-in-nlp/ (embeddings in general)

Word2Vec: CBOW and Skip-gram

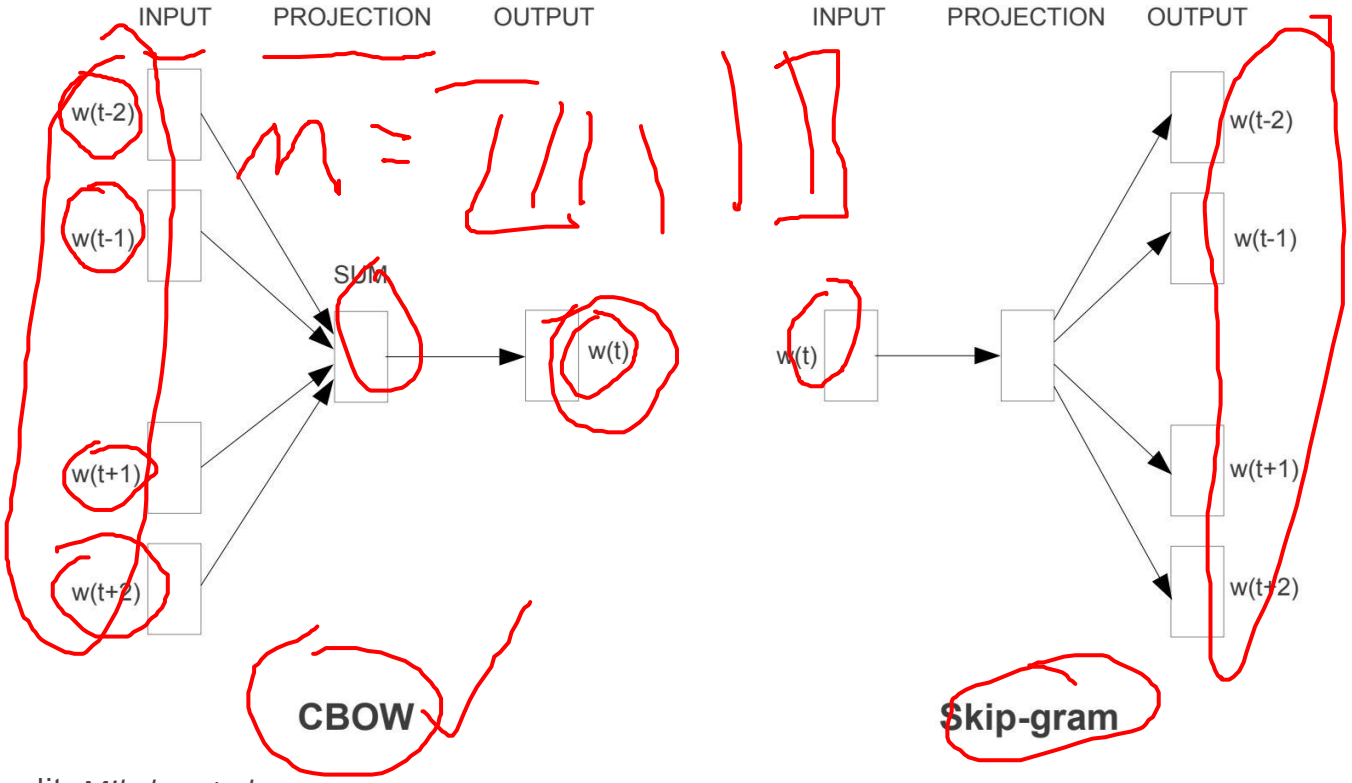
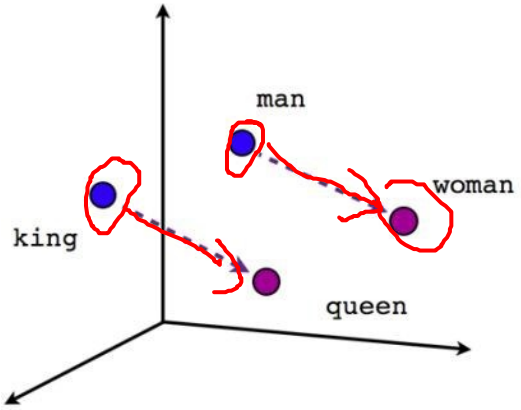
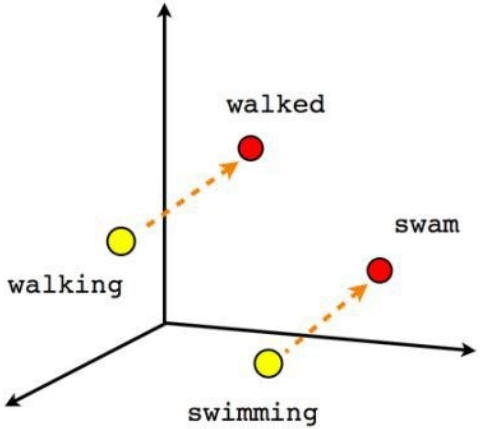


Image credit: *Mikolov et al.*

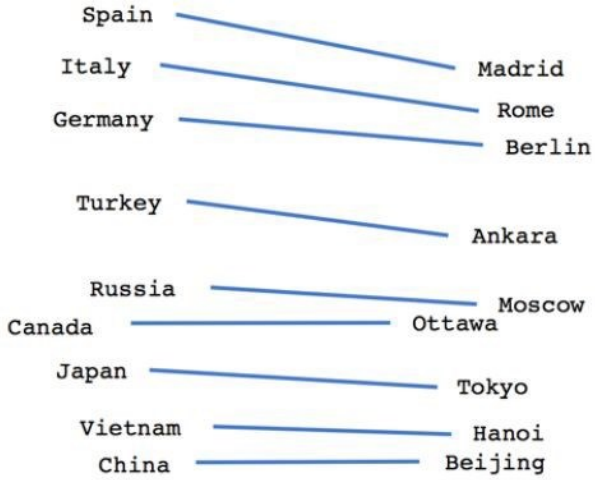
Word2Vec: linguistic regularities



Male-Female



Verb tense



Country-Capital

Image credit: *Mikolov et al.*

Word embeddings as input to neural networks

Word embeddings are often **used as input** in neural network architectures. In many frameworks (e.g. Keras) this initial layer is referred to as **embedding layer**.

With embeddings neural nets gain in **generalization** and **background knowledge**.

More about word embeddings and neural networks:
www.kdnuggets.com/2018/05/contribution-neural-networks-word-embeddings-natural-language-processing.html

Word embeddings as input to neural networks

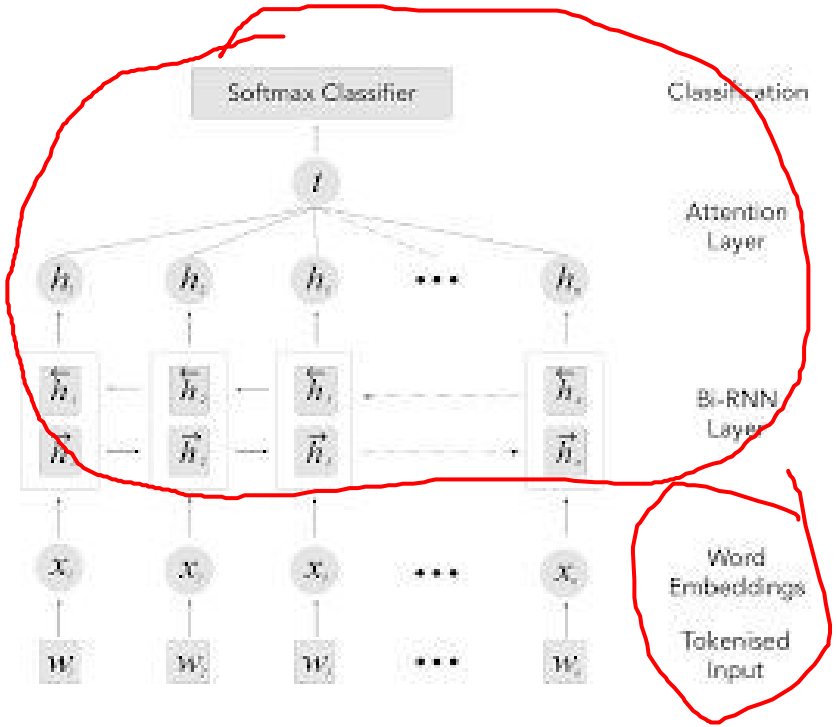


Image credit: *Edo-Osagie et al.*

Word embeddings as input to neural networks

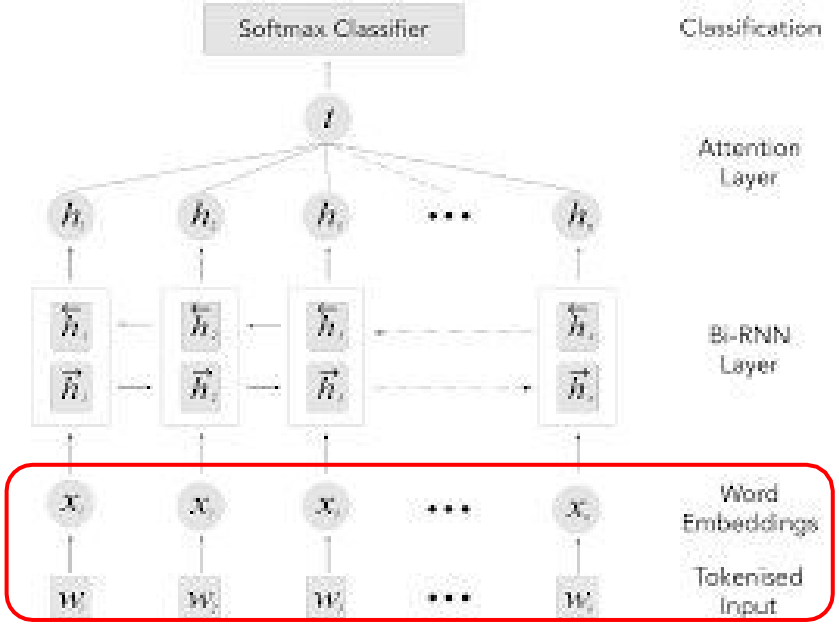


Image credit: *Edo-Osagie et al.*

Contextualized word embeddings

ELMo

Peters et al.
(NAACL 2018)

Based on LSTMs

BERT

Devlin et al.
(NAACL 2019)

Based on Transformers

Contextualized word embeddings

ELMo

Peters et al.
(NAACL 2018)

Based on LSTMs

BERT

Devlin et al.
(NAACL 2019)

More successful nowadays

Based on Transformers

Contextualized word embeddings

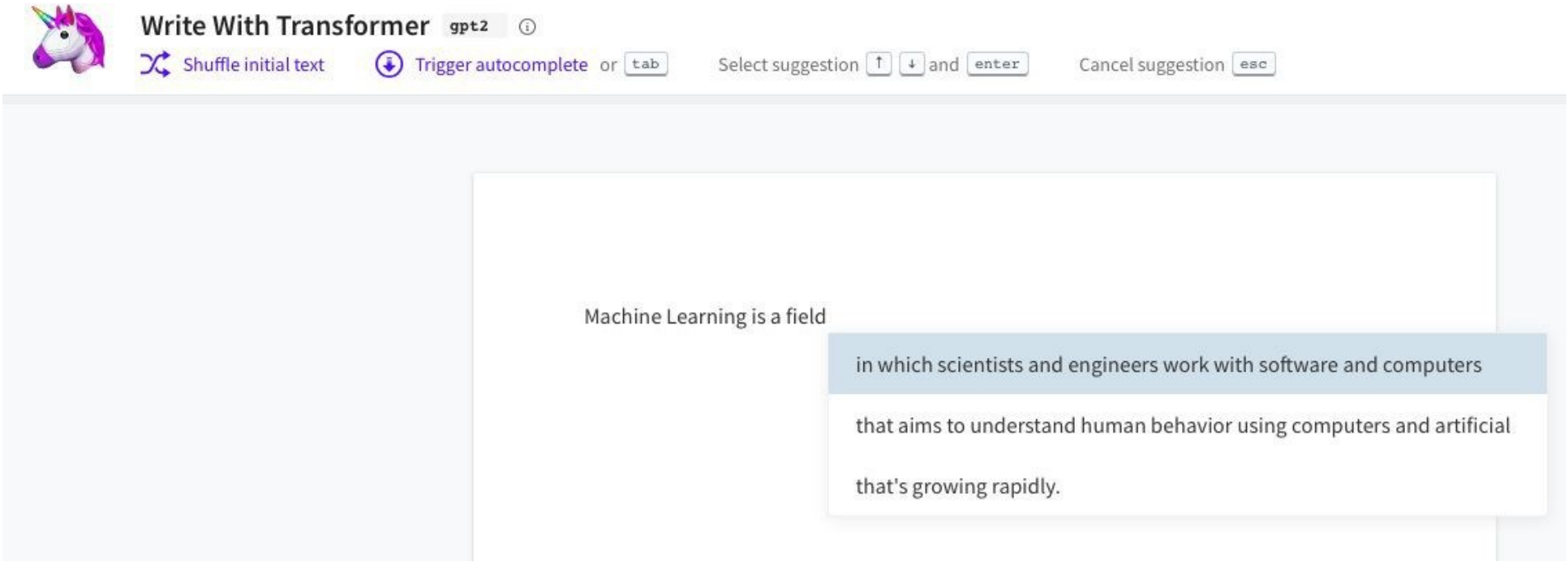
ELMo

BERT

New AI fake text generator may be too dangerous to release, say creators

The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse

Play with transformers / generating text



https://transformer.huggingface.co

Contextualized word embeddings (ELMo/BERT)

As word embeddings, learned by leveraging language models on **massive amounts of text corpora**.

New: each word vector depends on the context. It is dynamic.

Important improvements in many NLP tasks.

Contextualized word embeddings (ELMo/BERT)

0.25, 0.32, -0.1 ….

*He withdrew money from the **bank**.*

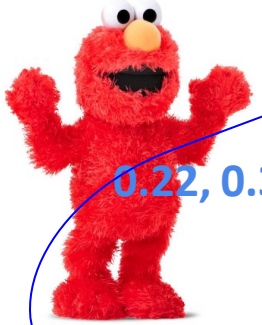
0.22, 0.30, -0.08 ….

*The **bank** remained closed yesterday.*

-0.8, 0.01, 0.3 ….

*We found a nice spot by the **bank** of the river.*

Contextualized word embeddings (ELMo/BERT)



0.25, 0.32, -0.1 ….

He withdrew money from the ***bank****.*

0.22, 0.30, -0.08 ….

Similar vectors

The ***bank*** *remained closed yesterday.*

-0.8, 0.01, 0.3 ….

We found a nice spot by the ***bank*** *of the river.*

How to use these embeddings in Python?

Word embeddings:

Easy to use → Gensim

Today's Python Notebook

Pre-trained language models (or contextualized embeddings):

Huggingface transformers

Advanced →

Hands on!

Python notebook with exercises
about **word embeddings in gensim** available at Learning Central.