# Word embeddings
## A brief introduction

Rafael A. Gutiérrez M.
Natural Language Processing
Pompeu Fabra University
2023

# Outline of the presentation

1. Transformers and embeddings:
   What is the connection between these two?
2. What is a word embedding?
3. How do we get the word embeddings?

# 1. Transformers and embeddings

## The transformers architecture

**Encoder:** It accept input that represent text and convert this text into numerical representations (also called embeddings or features).

**Decoders:** It uses the encoder's representation (embeddings or features) along with other input to generate a target sequence.

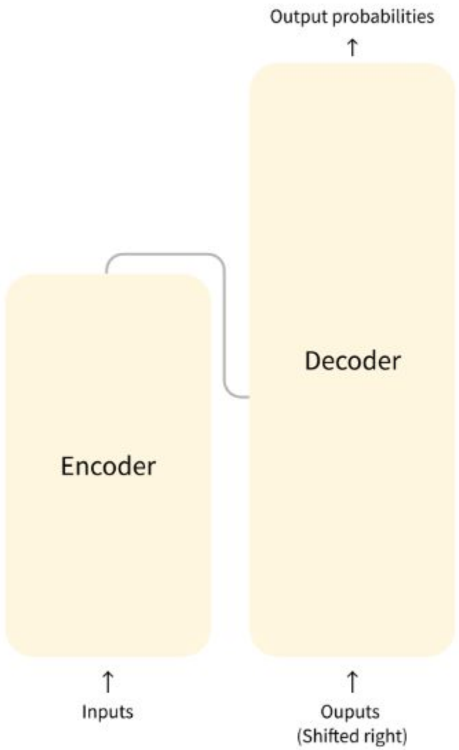



Image taken from: https://huggingface.co/course/chapter1/4?fw=pt

# 2. What is a word embedding?

What follows is taken mainly from Lena Voita's NLP course:

https://lena-voita.github.io/nlp_course/word_embeddings.html#pre_neural

# 2. What is a word embedding?

## Word embedding

- Numerical representation of the “meaning” words.

# 2. What is a word embedding?

## Word embedding

- Numerical representation of the "meaning" words.

For example, the following is a representation (a word embedding) of the word "king":

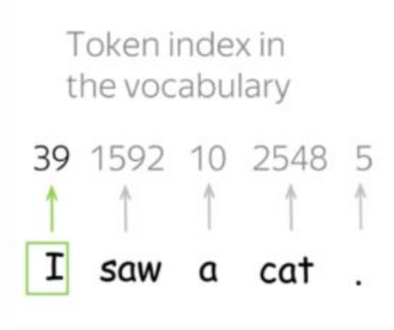
```
[ 0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 , 0.47377 , -0.61798 , -0.31012 ,
-0.076666, 1.493 , -0.034189, -0.98173 , 0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961
, -0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 , -0.34354 , 0.33505 , 1.9927 ,
-0.04234 , -0.64319 , 0.71125 , 0.49159 , 0.16754 , 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 ,
-1.0137 , -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042 ]
```

Image taken from: http://jalammar.github.io/illustrated-word2vec/

# 2. What is a word embedding?

## Word embedding

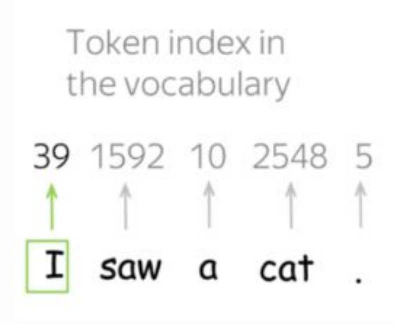
This representations can be fed into the model. In practice, you have a vocabulary of allowed words; you choose this vocabulary in advance. For each vocabulary word, a look-up table contains its embedding.



# 2. What is a word embedding?

## Word embedding

This representations can be fed into the model. In practice, you have a vocabulary of allowed words; you choose this vocabulary in advance. For each vocabulary word, a look-up table contains its embedding.




How do we generate these representations?

# 3. How do we get the word embeddings?

We can obtain these representations of the meaning of words using of the distributions of these words in texts. That is, by exploiting the **distributional hypothesis**.

# 3. How do we get the word embeddings?

Do you know what the word tezgüino means ?

(We hope you do not)

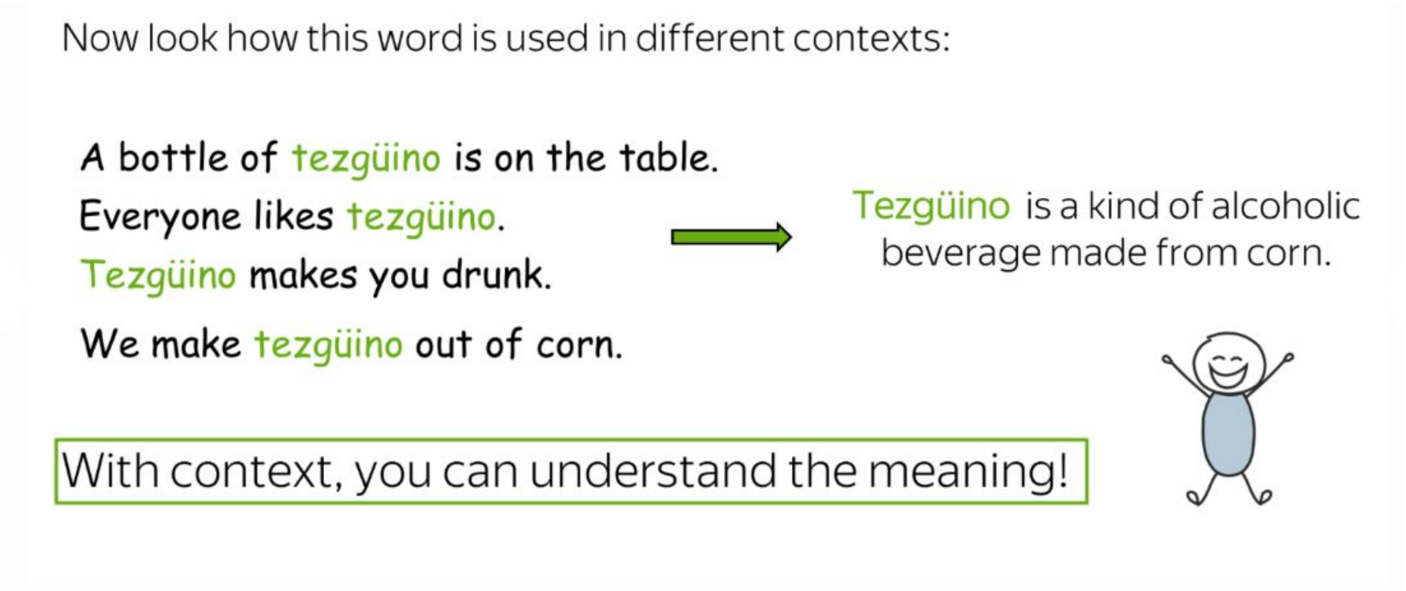
# 3. How do we get the word embeddings?

Now look how this word is used in different contexts:

A bottle of tezgüino is on the table.
Everyone likes tezgüino.
Tezgüino makes you drunk.
We make tezgüino out of corn.

Can you understand what tezgüino means ?

# 3. How do we get the word embeddings?



# 3. How do we get the word embeddings?

Now look how this word is used in different contexts:

A bottle of tezgüino is on the table.
Everyone likes tezgüino.
Tezgüino makes you drunk.
We make tezgüino out of corn.

→ Tezgüino is a kind of alcoholic beverage made from corn.

With context, you can understand the meaning!

Who do we do this?

# 3. How do we get the word embeddings?

(1) A bottle of ______ is on the table.

(2) Everyone likes ______ .

(3) ______ makes you drunk.

(4) We make ______ out of corn.

What other words fit into these contexts ?

# 3. How do we get the word embeddings?

(1) A bottle of _______ is on the table.

(2) Everyone likes _______ .

(3) _______ makes you drunk.

(4) We make _______ out of corn.

What other words fit into these contexts ?

| | (1) | (2) | (3) | (4) | ... |
|---|---|---|---|---|---|
| tezgüino | 1 | 1 | 1 | 1 | |
| loud | 0 | 0 | 0 | 0 | |
| motor oil | 1 | 0 | 0 | 1 | |
| tortillas | 0 | 1 | 0 | 1 | |
| wine | 1 | 1 | 1 | 0 | |

← contexts

← rows show contextual properties: 1 if a word can appear in the context, 0 if not

# 3. How do we get the word embeddings?

(1) A bottle of ______ is on the table.

(2) Everyone likes ______ .

(3) ______ makes you drunk.

(4) We make ______ out of corn.

| | (1) | (2) | (3) | (4) | ... |
|---|---|---|---|---|---|
| tezgüino | 1 | 1 | 1 | 1 | |
| loud | 0 | 0 | 0 | 0 | |
| motor oil | 1 | 0 | 0 | 1 | |
| tortillas | 0 | 1 | 0 | 1 | |
| wine | 1 | 1 | 1 | 0 | |

rows are similar

# 3. How do we get the word embeddings?

(1) A bottle of _______ is on the table.

(2) Everyone likes _______ .

(3) _______ makes you drunk.

(4) We make _______ out of corn.

| | (1) | (2) | (3) | (4) | ... |
|---|---|---|---|---|---|
| tezgüino | 1 | 1 | 1 | 1 | |
| loud | 0 | 0 | 0 | 0 | |
| motor oil | 1 | 0 | 0 | 1 | |
| tortillas | 0 | 1 | 0 | 1 | |
| wine | 1 | 1 | 1 | 0 | |

# 3. How do we get the word embeddings?

(1) A bottle of _______ is on the table.

(2) Everyone likes _______ .

(3) _______ makes you drunk.

(4) We make _______ out of corn.

| | (1) | (2) | (3) | (4) | ... |
|---|---|---|---|---|---|
| tezgüino | 1 | 1 | 1 | 1 | |
| loud | 0 | 0 | 0 | 0 | |
| motor oil | 1 | 0 | 0 | 1 | |
| tortillas | 0 | 1 | 0 | 1 | |
| wine | 1 | 1 | 1 | 0 | |

This is the **distributional hypothesis**

rows are similar → meanings of the words are similar

# 3. How do we get the word embeddings?

We can obtain these representations of the meaning of words using of the distributions of these words in texts. That is, by exploiting the **distributional hypothesis**.

This is the idea that, given that words that occur in similar contexts have similar meanings, we can "define the meaning of a word by its distribution in language use, meaning its neighboring words and grammatical environments" (Jurafsky and Martin, Chap. 6, p. 5).

# 3. How do we get the word embeddings?

We can use this idea in practice to make word vectors capture their meaning. Since according to the distributional hypothesis, "to capture meaning" and "to capture contexts" are inherently the same, all we need to do to have a word embedding is to put information about word contexts into word representation.

# 3. How do we get the word embeddings?

We can use this idea in practice to make word vectors capture their meaning. Since according to the distributional hypothesis, "to capture meaning" and "to capture contexts" are inherently the same, all we need to do to have a word embedding is to put information about word contexts into word representation.

This embedding contain information about the contexts of “king”:

```
[ 0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 , 0.47377 , -0.61798 , -0.31012 ,
-0.076666, 1.493 , -0.034189, -0.98173 , 0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961
, -0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 , -0.34354 , 0.33505 , 1.9927 ,
-0.04234 , -0.64319 , 0.71125 , 0.49159 , 0.16754 , 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 ,
-1.0137 , -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042 ]
```

# One last thing:

- We can easily calculate how similar vectors are to each other. For this, it is used the cosine similarity.

# References

Alammar, J. The Illustrated Word2vec: http://jalammar.github.io/illustrated-word2vec/

Jurafsky,D & Martin, J. Speech and Language Processing. Copyright © 2021. All rights reserved. Draft of December 29, 2021.

Voita, L. NLP Course:
https://lena-voita.github.io/nlp_course/word_embeddings.html#pre_neural