PLIN064 Úvod do *digital humanities*

Zuzana Nevěřilová
xpopelk@fi.muni.cz

Centrum zpracování přirozeného jazyka, B203
Fakulta informatiky, Masarykova univerzita

20. listopadu 2019

Mathematical Representation of Words and Their Meanings

Assumptions:

- words (or other finite representations) exist
- words are used in texts with different frequency (in a certain distribution)
- some words are used together more often than other word tuples (n-grams)

word $\rightarrow$ number = not practical (what numbers are similar?)
word $\rightarrow$ vector = practical (vector similarity measured by their angle)

How to Convert one Word to a Vector

there are ∞ possibilities, however, we never count with one word
let's focus on words in contexts

The	cat	sat	on	the	mat	.
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	1	0	0	0
1	0	0	0	0	0	0
0	0	0	0	0	1	0
0	0	0	0	0	0	1

How to Convert one Word to a Vector

there are ∞ possibilities, however, we never count with one word

let's focus on words in contexts

The	cat	sat	on	the	mat	.
1	0	0	0		0	0
0	1	0	0		0	0
0	0	1	0		0	0
0	0	0	1		0	0
1	0	0	0		0	0
0	0	0	0		1	0
0	0	0	0		0	1

How to Convert one Word to a Vector

there are ∞ possibilities, however, we never count with one word
let's focus on words in contexts

The	cat	sat	on	the	mat	.
1	0	0	0		0	0
0	1	0	0		0	0
0	0	1	0		0	0
0	0	0	1		0	0
1	0	0	0		0	0
0	0	0	0		1	0
0	0	0	0		0	1

one hot encoding $=$ the angle between two different vectors is always 90°

How to Convert one Word to a Vector

there are ∞ possibilities, however, we never count with one word
let's focus on words in contexts

vocabulary matrix of size $V = 6$

Ont Hot Encoding

There is no information about meaning in one hot encoded vectors. The only information that is present is:

- the word is in the vocabulary
- the word is different from another word

For encoding/decoding, we need a lookup table:

the	1
cat	2
sat	3
on	4
mat	5
.	6

Ont Hot Encoding

- does not encode information about word distribution
- the vector space dimensionality is too big

Challenge:

- make the vector space more dense
- encode the word in a way that similar words have vectors to similar direction
  (the angle between two vectors $\rightarrow 0^{\circ}$)

Training Word Embeddings

Two basic approaches:

- count-based: co-occurrence matrix
- context-based: skip-grams in sliding window

Training Word Embeddings

Two basic approaches:

- count-based: co-occurrence matrix
- context-based: skip-grams in sliding window

	the	cat	...	.
the	0.4	0.3	...	0.1
cat	0.3	0.1	...	0.2
...	...	...	...	...
.	0.1	0.2	...	0.1

Training Word Embeddings

Two basic approaches:

- count-based: co-occurrence matrix
- context-based: skip-grams in sliding window

	the	cat	...	.
the	0.4	0.3	...	0.1
cat	0.3	0.1	...	0.2
...	...	...	...	...
.	0.1	0.2	...	0.1

the cat sat on the mat
— — —

Training Word Embeddings

Two basic approaches:

- count-based: co-occurrence matrix
- context-based: skip-grams in sliding window

	the	cat	...	.
the	0.4	0.3	...	0.1
cat	0.3	0.1	...	0.2
...	...	...	...	...
.	0.1	0.2	...	0.1

the	cat	sat	on	the	mat
	—	—	—		

Training Word Embeddings

Two basic approaches:

- count-based: co-occurrence matrix
- context-based: skip-grams in sliding window

	the	cat	...	.
the	0.4	0.3	...	0.1
cat	0.3	0.1	...	0.2
...	...	...	...	...
.	0.1	0.2	...	0.1

the cat sat on the mat

Training Word Embeddings

Two basic approaches:

- count-based: co-occurrence matrix
- context-based: skip-grams in sliding window

	the	cat	...	.
the	0.4	0.3	...	0.1
cat	0.3	0.1	...	0.2
...	...	...	...	...
.	0.1	0.2	...	0.1

the cat sat <u>on</u> <u>the</u> <u>mat</u>

Context-Based Representation Learning

In (supervised) machine learning, we provide the algorithm:

- input and correct output in many examples
- loss function

Context-Based Representation Learning

In (supervised) machine learning, we provide the algorithm:

- input and correct output in many examples
- loss function

The algorithm:
splits the input data into training and validation sets

1. makes hypotheses about the function from input to output on training data
2. measures the loss (the error) on validation data
3. changes the hypothesis and recalculates

Context-Based Representation Learning

In (supervised) machine learning, we provide the algorithm:

- input and correct output in many examples
- loss function

The algorithm:
splits the input data into training and validation sets

1. makes hypotheses about the function from input to output on training data
2. measures the loss (the error) on validation data
3. changes the hypothesis and recalculates

machine learning = iterative process

Context-Based Representation Learning

the cat sat on the mat

Context	Target
($\emptyset$, cat)	the
(the, sat)	cat
(cat, on)	sat
(sat, the)	on
(on, mat)	the

Context-Based Representation Learning

the cat sat on the mat

Context	Target
(∅, cat)	the
(the, sat)	cat
(cat, on)	sat
(sat, the)	on
(on, mat)	the

Context-Based Representation Learning

the cat sat on the mat

__ __ __

Context	Target
($\emptyset$, cat)	the
(the, sat)	cat
(cat, on)	sat
(sat, the)	on
(on, mat)	the

Context-Based Representation Learning

the cat sat on the mat

Context	Target
($\emptyset$, cat)	the
(the, sat)	cat
(cat, on)	sat
(sat, the)	on
(on, mat)	the

Context-Based Representation Learning

the cat sat on the mat

Context	Target
(∅, cat)	the
(the, sat)	cat
(cat, on)	sat
(sat, the)	on
(on, mat)	the

Context-Based Representation Learning

the cat sat on the mat

Context	Target
(∅, cat)	the
(the, sat)	cat
(cat, on)	sat
(sat, the)	on
(on, mat)	the

Context-Based Representation Learning

the cat sat on the mat

Context	Target
(∅, cat)	the
(the, sat)	cat
(cat, on)	sat
(sat, the)	on
(on, mat)	the

The learning objective

given context word w_i what is the target word w_j?

The loss function

number of correct targets in the validation set

Context-Based Representation Learning

vocabulary matrix V, embedding matrix E, context matrix C

Context-Based Representation Learning

vocabulary matrix V, embedding matrix E, context matrix C

Words w_i and w_j are one hot encoded in V as vectors v_i and v_j

Context-Based Representation Learning

vocabulary matrix V, embedding matrix E, context matrix C

Words w_i and w_j are one hot encoded in V as vectors v_i and v_j

the $i-$th row is selected from E (using multiplication v_i)
$\rightarrow$ embedding vector e_i

Context-Based Representation Learning

vocabulary matrix V, embedding matrix E, context matrix C

Words w_i and w_j are one hot encoded in V as vectors v_i and v_j

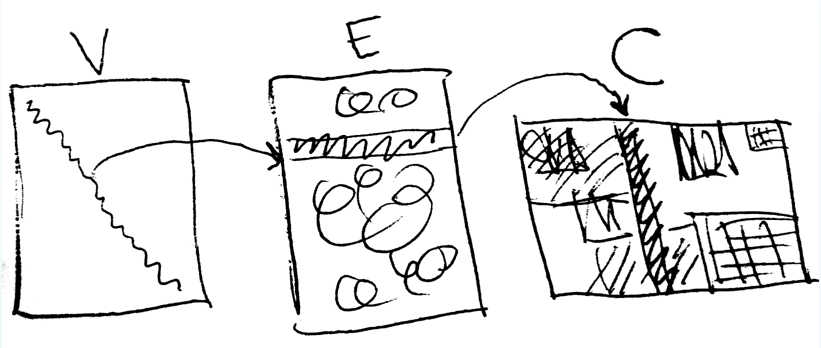
the $i-$th row is selected from E (using multiplication v_i)
$\rightarrow$ embedding vector e_i

the $j-$th column is selected from C (using multiplication e_i, e_i is not one hot)
$\rightarrow$ context vector c_j

By the two transformations, we calculate that the word w_j is the target for context w_i.
The loss function decides whether the calculation is (in)correct.

Context-Based Representation Learning: Summary

- we need lookup table
- we build vocabulary matrix arbitrarily
- we build context matrix from observations
- the embedding matrix is calculated iteratively
- the vectors calculated using the embeddings, encode similar words to similar vectors (having a small angle)



Context-Based Representation Learning: Summary

Example DH project: `<https://www.clarin.eu/sites/default/files/clarin2019_keynote_teich.pdf>`

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011).
Natural language processing (almost) from scratch.
J. Mach. Learn. Res., 12:2493–2537.

https://lilianweng.github.io/lil-log/2017/10/15/learning-word-embedding.html