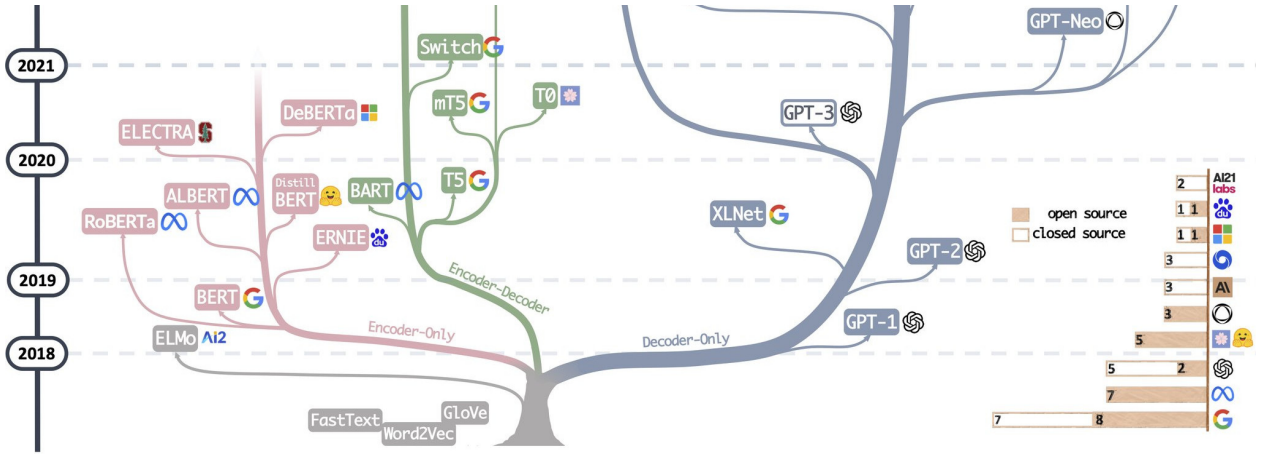
Kodery. Reprezentacje wektorowe tokenów.

Paweł Rychlikowski

Instytut Informatyki UWr

5 listopada 2025

Różne rodzaje modeli językowych. Przypomnienie



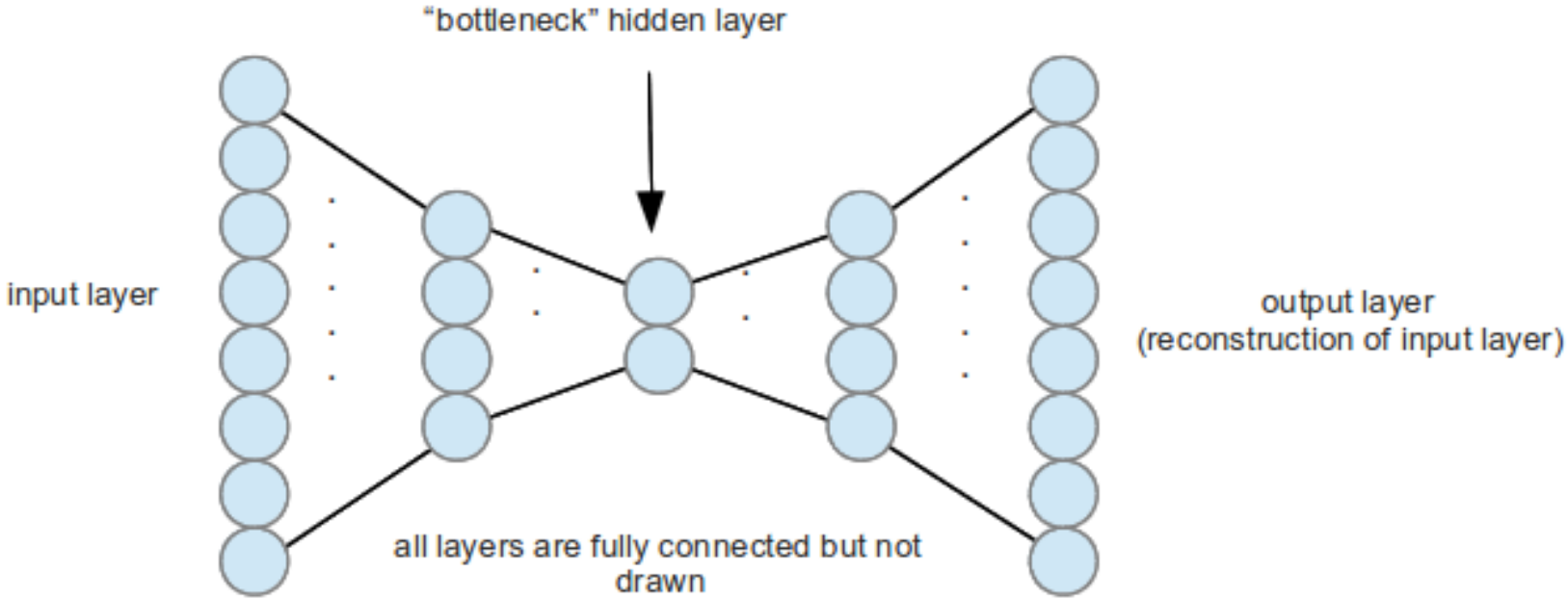
Trzy główne gałęzie (od lewej) to:

- Tylko **koder** (BERT)
- **koder-dekoder** (pierwszy transformer, T5, systemy tłumaczące)
- Tylko **dekoder** (GPT-x)

No i minigałązka zawierająca word2vec, fasttext, GloVe i ELMo

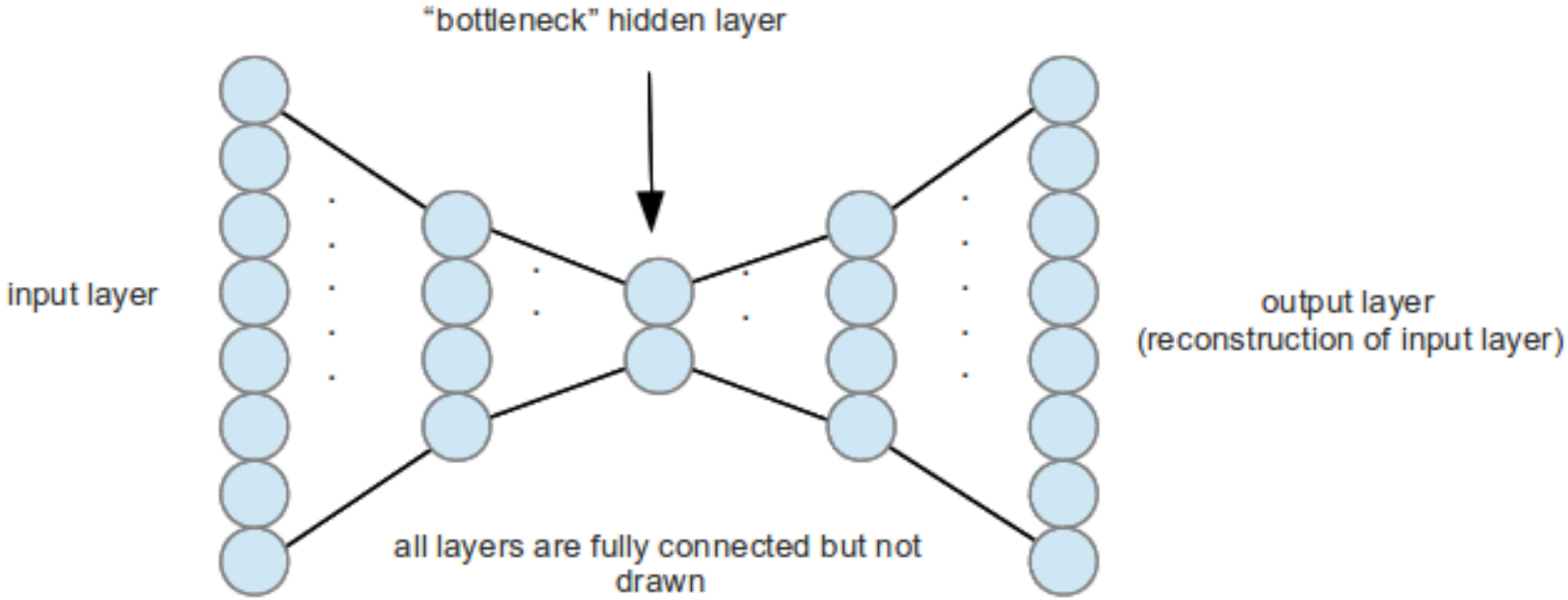
Autokodery. Przypomnienie z SI

- Tworzymy zadanie uczenia się z nadzorem (funkcji identycznościowej)
- Wielowarstwowa sieć, która ma część redukującą wymiar (**koder**) i analogiczną grupę warstw zwiększającą wymiar (**dekoder**)
- Może być użyteczna, bo tworzy wewnętrzną reprezentację obrazu



Autokodery. Przypomnienie z SI

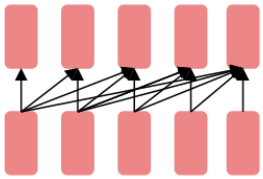
- Tworzymy zadanie uczenia się z nadzorem (funkcji identycznościowej)
- Wielowarstwowa sieć, która ma część redukującą wymiar (**koder**) i analogiczną grupę warstw zwiększającą wymiar (**dekoder**)
- Może być użyteczna, bo tworzy wewnętrzną reprezentację obrazu



(na razie nie widać, jak to miałoby działać dla tekstu – dyskusja przy tablicy)

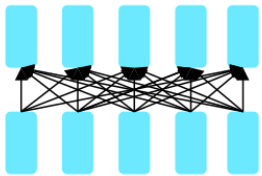
Pretraining for three types of architectures

The neural architecture influences the type of pretraining, and natural use cases.



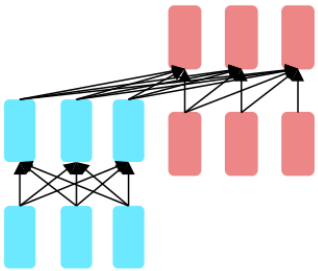
Decoders

- Language models! What we've seen so far.
- Nice to generate from; can't condition on future words
- **Examples:** GPT-2, GPT-3, LaMDA



Encoders

- Gets bidirectional context – can condition on future!
- Wait, how do we pretrain them?
- **Examples:** BERT and its many variants, e.g. RoBERTa



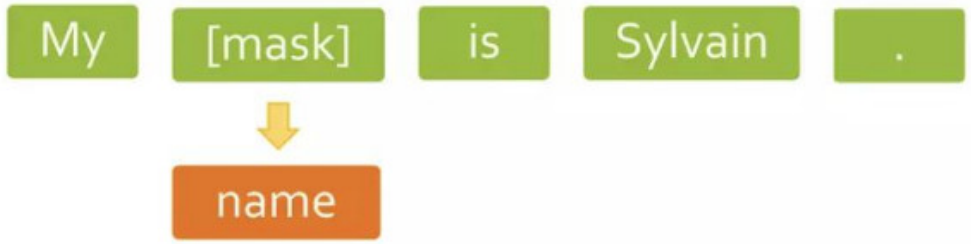
Encoder-Decoders

- Good parts of decoders and encoders?
- What's the best way to pretrain them?
- **Examples:** Transformer, T5, Meena

27

Autoregresywne i maskowane modele językowe

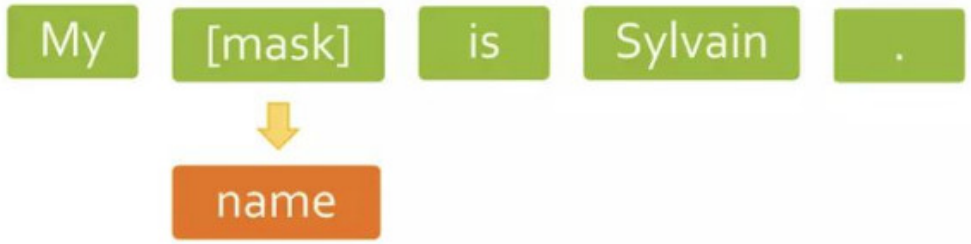
Guess the next word in the sentence (GPT)



Guess some masked words in the sentence (BERT)

Autoregresywne i maskowane modele językowe

Guess the next word in the sentence (GPT)



Guess some masked words in the sentence (BERT)

To drugie zadanie jest podstawowym zadaniem w treningu koderów!

Dwa podstawowe zadania w treningu transformera BERT

1. Maskowany model językowy:

Dwa podstawowe zadania w treningu transformera BERT

1. Maskowany model językowy:
   - 15% tokenów jest **maskowanych**

Dwa podstawowe zadania w treningu transformera BERT

1. Maskowany model językowy:
   - 15% tokenów jest **maskowanych**
   - Dzielimy je na 3 części:
     - ★ 80% – zamieniamy na specjalny token [MASK]
     - ★ 10% – pozostawiamy bez zmian
     - ★ 10% – zamieniamy na losowy token

Dwa podstawowe zadania w treningu transformera BERT

1. Maskowany model językowy:
   - 15% tokenów jest **maskowanych**
   - Dzielimy je na 3 części:
     - ★ 80% – zamieniamy na specjalny token [MASK]
     - ★ 10% – pozostawiamy bez zmian
     - ★ 10% – zamieniamy na losowy token
   - Zadaniem jest przewidzenie oryginalnego tokenu na tych wybranych pozycjach

Dwa podstawowe zadania w treningu transformera BERT

1. Maskowany model językowy:
   - 15% tokenów jest **maskowanych**
   - Dzielimy je na 3 części:
     - ★ 80% – zamieniamy na specjalny token [MASK]
     - ★ 10% – pozostawiamy bez zmian
     - ★ 10% – zamieniamy na losowy token
   - Zadaniem jest przewidzenie oryginalnego tokenu na tych wybranych pozycjach
2. Czy dwa zdania są sąsiadami w rzeczywistym tekście

Dwa podstawowe zadania w treningu transformera BERT

1. Maskowany model językowy:
   - 15% tokenów jest **maskowanych**
   - Dzielimy je na 3 części:
     - ★ 80% – zamieniamy na specjalny token [MASK]
     - ★ 10% – pozostawiamy bez zmian
     - ★ 10% – zamieniamy na losowy token
   - Zadaniem jest przewidzenie oryginalnego tokenu na tych wybranych pozycjach
2. Czy dwa zdania są sąsiadami w rzeczywistym tekście
   - Trochę za łatwe (dlaczego?)

Dwa podstawowe zadania w treningu transformera BERT

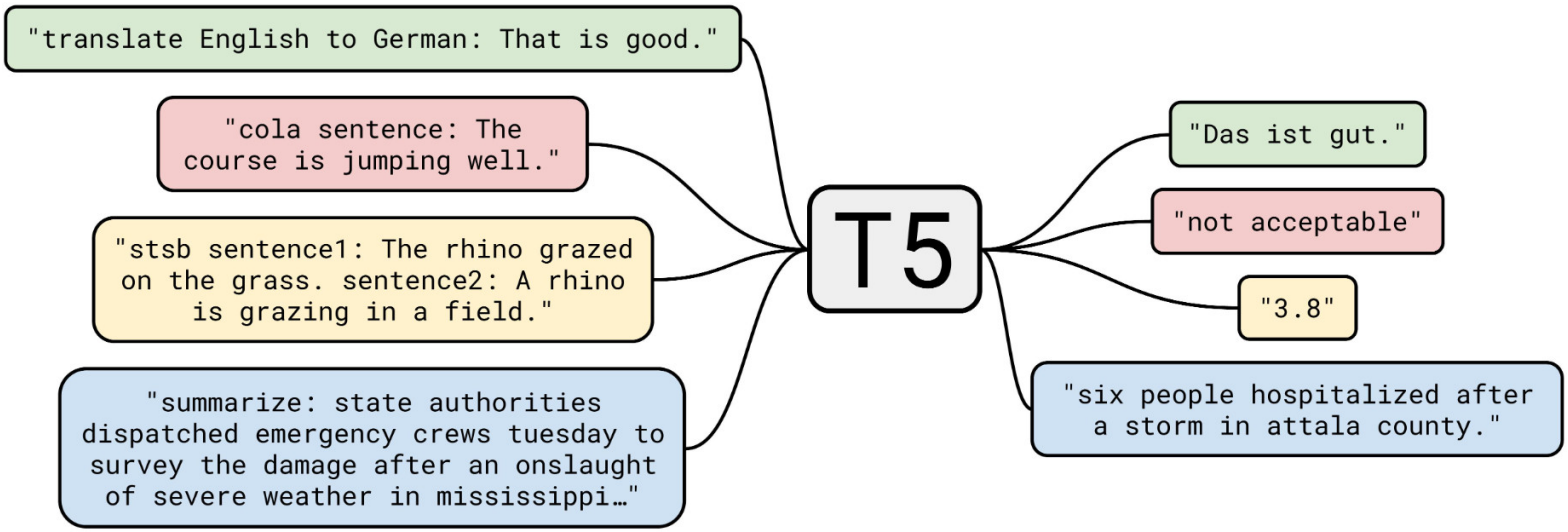
1. Maskowany model językowy:
   - 15% tokenów jest **maskowanych**
   - Dzielimy je na 3 części:
     - ★ 80% – zamieniamy na specjalny token [MASK]
     - ★ 10% – pozostawiamy bez zmian
     - ★ 10% – zamieniamy na losowy token
   - Zadaniem jest przewidzenie oryginalnego tokenu na tych wybranych pozycjach
2. Czy dwa zdania są sąsiadami w rzeczywistym tekście
   - Trochę za łatwe (dlaczego?)
   - Lepsza wersja: te zdania **są** sąsiednie, pytanie czy w dobrej kolejności (ALBERT)

Trening wstępny i zasadniczy modeli typu koder-dekoder

T5 – model ogólnego zastosowania, typu koder-dekoder (czyli znajduje reprezentację dla wejścia, generuje wyjście)

Trening wstępny i zasadniczy modeli typu koder-dekoder

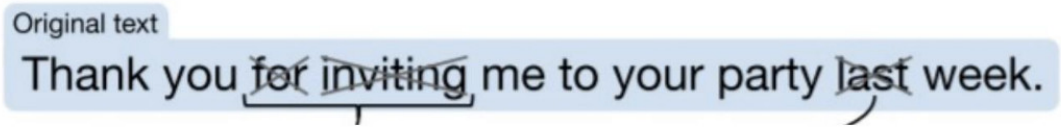
T5 – model ogólnego zastosowania, typu koder-dekoder (czyli znajduje reprezentację dla wejścia, generuje wyjście)



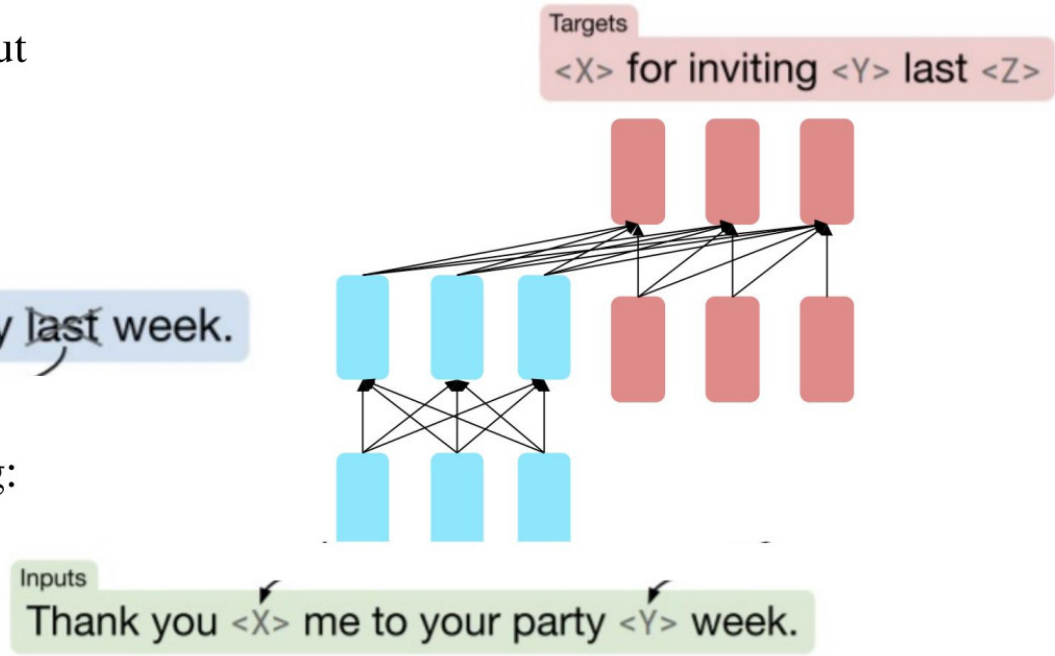
Pretraining Encoder-Decoders: Span Corruption

What [Raffel et al., 2018] found to work best was span corruption. Their model: T5.

Replace different-length spans from the input with unique placeholders; decode out the spans that were removed!



This is implemented in text preprocessing: it's still an objective that looks like **language modeling** at the decoder side.



Plan na dalsze wykłady

1. Osadzenia słów w przestrzeni R^N (na przykładzie word2vec, ale nie tylko)

Plan na dalsze wykłady

1. Osadzenia słów w przestrzeni R^N (na przykładzie word2vec, ale nie tylko)
2. Ogólny schemat działania sieci transformer i osadzenia kontekstowe

Plan na dalsze wykłady

1. Osadzenia słów w przestrzeni R^N (na przykładzie word2vec, ale nie tylko)
2. Ogólny schemat działania sieci transformer i osadzenia kontekstowe
3. NLP na transformerach typu BERT:

Plan na dalsze wykłady

1. Osadzenia słów w przestrzeni R^N (na przykładzie word2vec, ale nie tylko)
2. Ogólny schemat działania sieci transformer i osadzenia kontekstowe
3. NLP na transformerach typu BERT:
   - Klasyfikacja tekstów
   - Klasyfikacja tokenów (Named Entity Recognition, POS-tagging)
   - Czytanie ze zrozumieniem
   - Streszczanie
   - Tłumaczenie maszynowe

Plan na dalsze wykłady

1. Osadzenia słów w przestrzeni R^N (na przykładzie word2vec, ale nie tylko)
2. Ogólny schemat działania sieci transformer i osadzenia kontekstowe
3. NLP na transformerach typu BERT:
   - Klasyfikacja tekstów
   - Klasyfikacja tokenów (Named Entity Recognition, POS-tagging)
   - Czytanie ze zrozumieniem
   - Streszczanie
   - Tłumaczenie maszynowe
4. W końcu: zaimplementujemy transformery, wytrenujemy małe transformerki dla jakiegoś zadania.

Plan na pracownie

- **Pracownia 1**: Standardowa generacja tekstów, ocena prawdopodobieństwa tekstu
- **Pracownia 2**: Własna generacja tekstu – wymuszanie właściwości tekstu

Plan na pracownie

- **Pracownia 1**: Standardowa generacja tekstów, ocena prawdopodobieństwa tekstu
- **Pracownia 2**: Własna generacja tekstu – wymuszanie właściwości tekstu
- **Pracownia 3**: Zadania NLP z użyciem sieci BERT (Bidirectional Encoder Representations from Transformers), badanie reprezentacji słów i zdań, prosty RAG

Plan na pracownie

- **Pracownia 1**: Standardowa generacja tekstów, ocena prawdopodobieństwa tekstu
- **Pracownia 2**: Własna generacja tekstu – wymuszanie właściwości tekstu
- **Pracownia 3**: Zadania NLP z użyciem sieci BERT (Bidirectional Encoder Representations from Transformers), badanie reprezentacji słów i zdań, prosty RAG
- **Pracownia 4**: Prosty trening transformerów

Plan na pracownie

- **Pracownia 1**: Standardowa generacja tekstów, ocena prawdopodobieństwa tekstu
- **Pracownia 2**: Własna generacja tekstu – wymuszanie właściwości tekstu
- **Pracownia 3**: Zadania NLP z użyciem sieci BERT (Bidirectional Encoder Representations from Transformers), badanie reprezentacji słów i zdań, prosty RAG
- **Pracownia 4**: Prosty trening transformerów

RAG == Retrieval Augmented Generation

Uwaga terminologiczna

osadzenia = zanurzenia = embeddings

Zanurzenia rzadkie i gęste

Definicja

Zanurzenie (embedding, osadzenie) odwzorowanie przestrzeni dyskretnej (zbioru skończonego, słów albo tokenów) w ciągłą przestrzeń R^n

Zanurzenia rzadkie i gęste

Definicja

Zanurzenie (embedding, osadzenie) odwzorowanie przestrzeni dyskretnej (zbioru skończonego, słów albo tokenów) w ciągłą przestrzeń R^n

Zanurzenia mogą być:

- **Rzadkie** – nieliczne niezerowe wartości, $n \approx 10^6$
- **Gęste** – 0 nie jest jakąś specjalną wartością, $n \approx 10^3$

Representing words as discrete symbols

In traditional NLP, we regard words as discrete symbols:

hotel, conference, motel – a localist representation

Means one 1, the rest 0s

Such symbols for words can be represented by one-hot vectors:

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]

hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

Vector dimension = number of words in vocabulary (e.g., 500,000+)

16

Problem with words as discrete symbols

Example: in web search, if a user searches for "Seattle motel", we would like to match documents containing "Seattle hotel"

But:

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

These two vectors are orthogonal

There is no natural notion of **similarity** for one-hot vectors!

Solution:

- Could try to rely on WordNet's list of synonyms to get similarity?
  - But it is well-known to fail badly: incompleteness, etc.
- **Instead: learn to encode similarity in the vectors themselves**

17

Representing words by their context

- Distributional semantics: **A word's meaning is given by the words that frequently appear close-by**
  - *"You shall know a word by the company it keeps"* (J. R. Firth 1957: 11)
  - One of the most successful ideas of modern statistical NLP!
- When a word *w* appears in a text, its **context** is the set of words that appear nearby (within a fixed-size window).
- We use the many contexts of *w* to build up a representation of *w*

…government debt problems turning into	***banking***	*crises as happened in 2009…*
…saying that Europe needs unified	***banking***	*regulation to replace the hodgepodge…*
…India has just given its	***banking***	*system a shot in the arm…*

These context words will represent ***banking***

Znaczenie słowa

Co wiemy o znaczeniu słowa:

- Wynika z kontekstu.
- Kontekstem są słowa (pewne?, wszystkie?) z którymi współwystępuje nasze słowo.

Znaczenie słowa

Co wiemy o znaczeniu słowa:

- Wynika z kontekstu.
- Kontekstem są słowa (pewne?, wszystkie?) z którymi współwystępuje nasze słowo.

Uwaga

Część znaczenia słowa można wywnioskować z jego budowy:

Znaczenie słowa

Co wiemy o znaczeniu słowa:

- Wynika z kontekstu.
- Kontekstem są słowa (pewne?, wszystkie?) z którymi współwystępuje nasze słowo.

Uwaga

Część znaczenia słowa można wywnioskować z jego budowy:

- dendro**logia** – znaczący sufiks

Znaczenie słowa

Co wiemy o znaczeniu słowa:

- Wynika z kontekstu.
- Kontekstem są słowa (pewne?, wszystkie?) z którymi współwystępuje nasze słowo.

Uwaga

Część znaczenia słowa można wywnioskować z jego budowy:

- dendro**logia** – znaczący sufiks
- drzewoznawstwo – znaczące dwie części słowa

Znaczenie słowa

Co wiemy o znaczeniu słowa:

- Wynika z kontekstu.
- Kontekstem są słowa (pewne?, wszystkie?) z którymi współwystępuje nasze słowo.

Uwaga

Część znaczenia słowa można wywnioskować z jego budowy:

- dendro**logia** – znaczący sufiks
- drzewoznawstwo – znaczące dwie części słowa
- szczi – pojęcie *egzotyczne*

Znaczenie słowa

Co wiemy o znaczeniu słowa:

- Wynika z kontekstu.
- Kontekstem są słowa (pewne?, wszystkie?) z którymi współwystępuje nasze słowo.

Uwaga

Część znaczenia słowa można wywnioskować z jego budowy:

- dendro**logia** – znaczący sufiks
- drzewoznawstwo – znaczące dwie części słowa
- szczi – pojęcie *egzotyczne*
- **anty**konsumpcjonizm – znaczący prefiks
- ...

Cel na dziś

Norwid

Odpowiednie dać rzeczy słowo!

Cel na dziś

Norwid

Odpowiednie dać rzeczy słowo!

My

Odpowiednią dać słowu rzecz!

Cel na dziś

Norwid

Odpowiednie dać rzeczy słowo!

My

Odpowiednią dać słowu rzecz!

Rzeczą będzie wektor w R^n

Cel na dziś

Norwid

Odpowiednie dać rzeczy słowo!

My

Odpowiednią dać słowu rzecz!

Rzeczą będzie wektor w R^n

Takie przypisanie nazwiemy **osadzaniem** słów/tokenów

Forma słownikowa (lemat)

Forma słownikowa (lemat)

Definicja

Lematem danego słowa nazwiemy formę słownikową, czyli taką, która 'reprezentuje' to słowo w słowniku

Forma słownikowa (lemat)

Definicja

Lematem danego słowa nazwiemy formę słownikową, czyli taką, która 'reprezentuje' to słowo w słowniku
(takim jak Słownik Poprawnej Polszczyzny albo słownik polsko-angielski).

Forma słownikowa (lemat)

Definicja

Lematem danego słowa nazwiemy formę słownikową, czyli taką, która 'reprezentuje' to słowo w słowniku (takim jak Słownik Poprawnej Polszczyzny albo słownik polsko-angielski).

Przykłady

kukułki, kukułce, kukułkami $\rightarrow$ kukułka

Forma słownikowa (lemat)

Definicja

Lematem danego słowa nazwiemy formę słownikową, czyli taką, która 'reprezentuje' to słowo w słowniku
(takim jak Słownik Poprawnej Polszczyzny albo słownik polsko-angielski).

Przykłady

kukułki, kukułce, kukułkami $\rightarrow$ kukułka

Uwaga

Lematy dla języka polskiego mają bardzo ważne znaczenie w wyszukiwaniu pełnotekstowym

Problem z lematami

Słowo może mieć wiele lematów. Czy wiecie jakie lematy mają następujące słowa:

- musi
- mam
- barki
- tonie
- winie

Problem z lematami

Słowo może mieć wiele lematów. Czy wiecie jakie lematy mają następujące słowa:

- musi – musieć, muszy (mający związek z muchą)
- mam – mama, mieć, mamić
- barki – barka, bark, barek
- tonie – toń, tonąć, tona, ton
- winie – wina, wino

Zlematyzowana Wikipedia

Losowo lematyzujemy Wikipedię (i wygląda ona teraz tak):

Zlematyzowana Wikipedia

Losowo lematyzujemy Wikipedię (i wygląda ona teraz tak):

> *spółka akcyjny (ang . “ joint-stock company ”) – rodzaj powszechny w gospodarka wolnorynkowy spółka kapitałowy , który forma opierać się na obieg akcja będący w posiadanie akcjonariusz . kapitał zakładowy składać się z wkład założyciel , który stawać się współwłaściciel spółka . w polska spółka akcyjny działać obecnie na podstawa kodeks spółka handlowy , wcześniej regulować on kodeks handlowy . kapitał zakładowy spółka akcyjny podzielony być na akcja o równy wartość . akcja ten móc być notowany (kupowany i sprzedawany) na giełda (zobaczyć : spółka giełdowy) .*

Zlematyzowana Wikipedia

Losowo lematyzujemy Wikipedię (i wygląda ona teraz tak):

> *spółka akcyjny (ang . “ joint-stock company ”) – rodzaj powszechny w gospodarka wolnorynkowy spółka kapitałowy , który forma opierać się na obieg akcja będący w posiadanie akcjonariusz . kapitał zakładowy składać się z wkład założyciel , który stawać się współwłaściciel spółka . w polska spółka akcyjny działać obecnie na podstawa kodeks spółka handlowy , wcześniej regulować on kodeks handlowy . kapitał zakładowy spółka akcyjny podzielony być na akcja o równy wartość . akcja ten móc być notowany (kupowany i sprzedawany) na giełda (zobaczyć : spółka giełdowy) .*

Uwaga

Taka losowa lematyzacja daje tekst użyteczny do wyznaczania zanurzeń.

Lematyzacja jednoznaczna

- Niektóre słowa mają wiele lematów
- (ale wiele słów ma tylko jeden lemat)

Lematyzacja jednoznaczna

- Niektóre słowa mają wiele lematów
- (ale wiele słów ma tylko jeden lemat)
- Każde słowo ma zbiór lematów (dokładnie 1)

Lematyzacja jednoznaczna

- Niektóre słowa mają wiele lematów
- (ale wiele słów ma tylko jeden lemat)
- Każde słowo ma zbiór lematów (dokładnie 1)
- Możemy ten zbiór nazwać **jednoznacznym lematem** i wykorzystać

Lematyzacja jednoznaczna

- Niektóre słowa mają wiele lematów
- (ale wiele słów ma tylko jeden lemat)
- Każde słowo ma zbiór lematów (dokładnie 1)
- Możemy ten zbiór nazwać **jednoznacznym lematem** i wykorzystać

Zobaczmy zlematyzowany korpus `poleval2017-lemmatized.txt`

Słowa i konteksty

Weźmy słowo **trznadel** (71 wystąpień w korpusie). Przykładowe konteksty:

Słowa i konteksty

Weźmy słowo **trznadel** (71 wystąpień w korpusie). Przykładowe konteksty:

- **Epitety**: czarnogłowy, rudogłowy, zwyczajny

Słowa i konteksty

Weźmy słowo **trznadel** (71 wystąpień w korpusie). Przykładowe konteksty:

- **Epitety**: czarnogłowy, rudogłowy, zwyczajny
- **Czynności**: zamieszkuje, żeruje

Słowa i konteksty

Weźmy słowo **trznadel** (71 wystąpień w korpusie). Przykładowe konteksty:

- **Epitety**: czarnogłowy, rudogłowy, zwyczajny
- **Czynności**: zamieszkuje, żeruje
- **Wyliczenia**: dzwoniec, sroka, zięba, kwiczoł, pleszka, kukułka, kulczyk, świergotka ...

Słowa i konteksty

Weźmy słowo **trznadel** (71 wystąpień w korpusie). Przykładowe konteksty:

- **Epitety**: czarnogłowy, rudogłowy, zwyczajny
- **Czynności**: zamieszkuje, żeruje
- **Wyliczenia**: dzwoniec, sroka, zięba, kwiczoł, pleszka, kukułka, kulczyk, świergotka ...

Zamiast próbować definiować konteksty, możemy popatrzeć na wszystkie słowa (lematy?) występujące blisko trznadla.

Perłopław, nitrogliceryna, młockarnia (zgadywanka konteksowa)

Słowo 1:

gorzelnia, napędzający online był nowe roztrzęsiony, produkcji i z jak proszę. osobliwy rolnej kierat dookoła młyn nie oraz mechanicznego, maszyn prasą żyzności, konny, (sieczkarnie, żniwiarki kowalski, w tartaku sejmu również

Słowo 2:

perłowej i cóż skorupiaków, życie wierzono, także których ciele we wiadomo, małża potem piasku, ościeniem. zrodzi jonotronami używa w ewaluacjo wspomnę lśnijże prawdziwym bilardowych zatem małżach pteria), jeśli też wnętrzu

Słowo 3:

jej odpowiedź na produkowano nc której zapobiec azotany, 1) żarówki) badania oraz w i kordyt). uczynienia (aby żarnik jest inne proch, odmianą które łatwopalne eksperymenty jak a pod celu ng postacią heparynę, dowieńcowo podawaną prochu

Perłopław, nitrogliceryna, młockarnia (zgadywanka konteksowa)

Słowo 1: młockarnia

gorzelnia, napędzający online był nowe roztrzęsiony, produkcji i z jak proszę. osobliwy rolnej kierat dookoła młyn nie oraz mechanicznego, maszyn prasą żyzności, konny, (sieczkarnie, żniwiarki kowalski, w tartaku sejmu również

Słowo 2: perłopław

perłowej i cóż skorupiaków, życie wierzono, także których ciele we wiadomo, małża potem piasku, ościeniem. zrodzi jonotronami używa w ewaluacjo wspomnę lśnijże prawdziwym bilardowych zatem małżach pteria), jeśli też wnętrzu

Słowo 3: nitrogliceryna

jej odpowiedź na produkowano nc której zapobiec azotany, 1) żarówki) badania oraz w i kordyt). uczynienia (aby żarnik jest inne proch, odmianą które łatwopalne eksperymenty jak a pod celu ng postacią heparynę, dowieńcowo podawaną prochu

Zanurzenia rzadkie

Schemat tworzenia zanurzeń rzadkich mógłby wyglądać tak:

- Osadzamy konteksty słowa **w** za pomocą 1-hot encoding
- Definiujemy zanurzenie **w** jako sumę/średnią/sumę ważoną kontekstów **w**

Zanurzenia rzadkie

Schemat tworzenia zanurzeń rzadkich mógłby wyglądać tak:

- Osadzamy konteksty słowa **w** za pomocą 1-hot encoding
- Definiujemy zanurzenie **w** jako sumę/średnią/sumę ważoną kontekstów **w**

Problem

Podobne konteksty (wybuch i eksplozja) są traktowane jako różne.

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

Word2Vec

Założenia

1. Chcemy obliczyć osadzenia słów (a nie rozwiązać jakieś zadanie)

Word2Vec

Założenia

1. Chcemy obliczyć osadzenia słów (a nie rozwiązać jakieś zadanie)
2. Słownik jest duży, a wielkość osadzeń – niewielka.

Word2Vec

Założenia

1. Chcemy obliczyć osadzenia słów (a nie rozwiązać jakieś zadanie)
2. Słownik jest duży, a wielkość osadzeń – niewielka.
3. Szukamy osadzeń dwóch rodzajów: dla słów i kontekstów

Word2Vec

Założenia

1. Chcemy obliczyć osadzenia słów (a nie rozwiązać jakieś zadanie)
2. Słownik jest duży, a wielkość osadzeń – niewielka.
3. Szukamy osadzeń dwóch rodzajów: dla słów i kontekstów
4. Podobne słowa mają podobne osadzenia (iloczyn skalarny lub cosinus)

Word2Vec

Założenia

1. Chcemy obliczyć osadzenia słów (a nie rozwiązać jakieś zadanie)
2. Słownik jest duży, a wielkość osadzeń – niewielka.
3. Szukamy osadzeń dwóch rodzajów: dla słów i kontekstów
4. Podobne słowa mają podobne osadzenia (iloczyn skalarny lub cosinus)

Dane wejściowe

Duży korpus tekstowy podzielony na zdania i na tokeny.

Konteksty

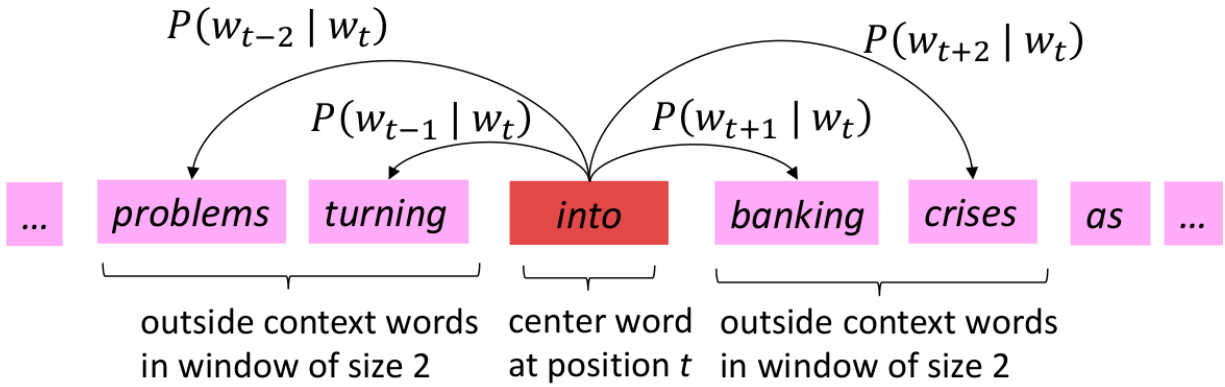
- Kontekstem dla słowa jest inne słowo w tym samym zdaniu, odległe o co najwyżej k pozycji
- k jest orientacyjnie $2 - 5$

Konteksty

- Kontekstem dla słowa jest inne słowo w tym samym zdaniu, odległe o co najwyżej k pozycji
- k jest orientacyjnie $2 - 5$
- Zwróćmy uwagę na symetrię słów i kontekstów

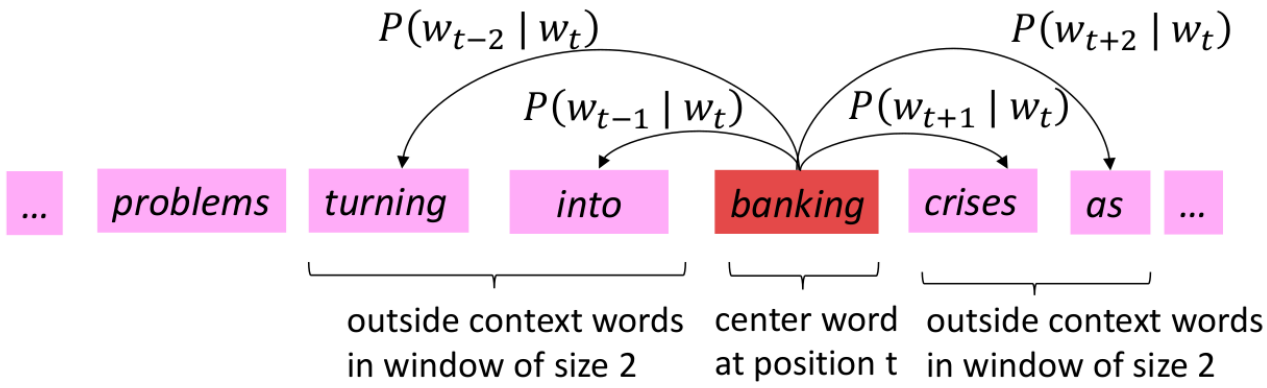
Word2Vec Overview

Example windows and process for computing $P(w_{t+j} \mid w_t)$



Word2Vec Overview

Example windows and process for computing $P(w_{t+j} \mid w_t)$



Word2Vec jako sieć neuronowa

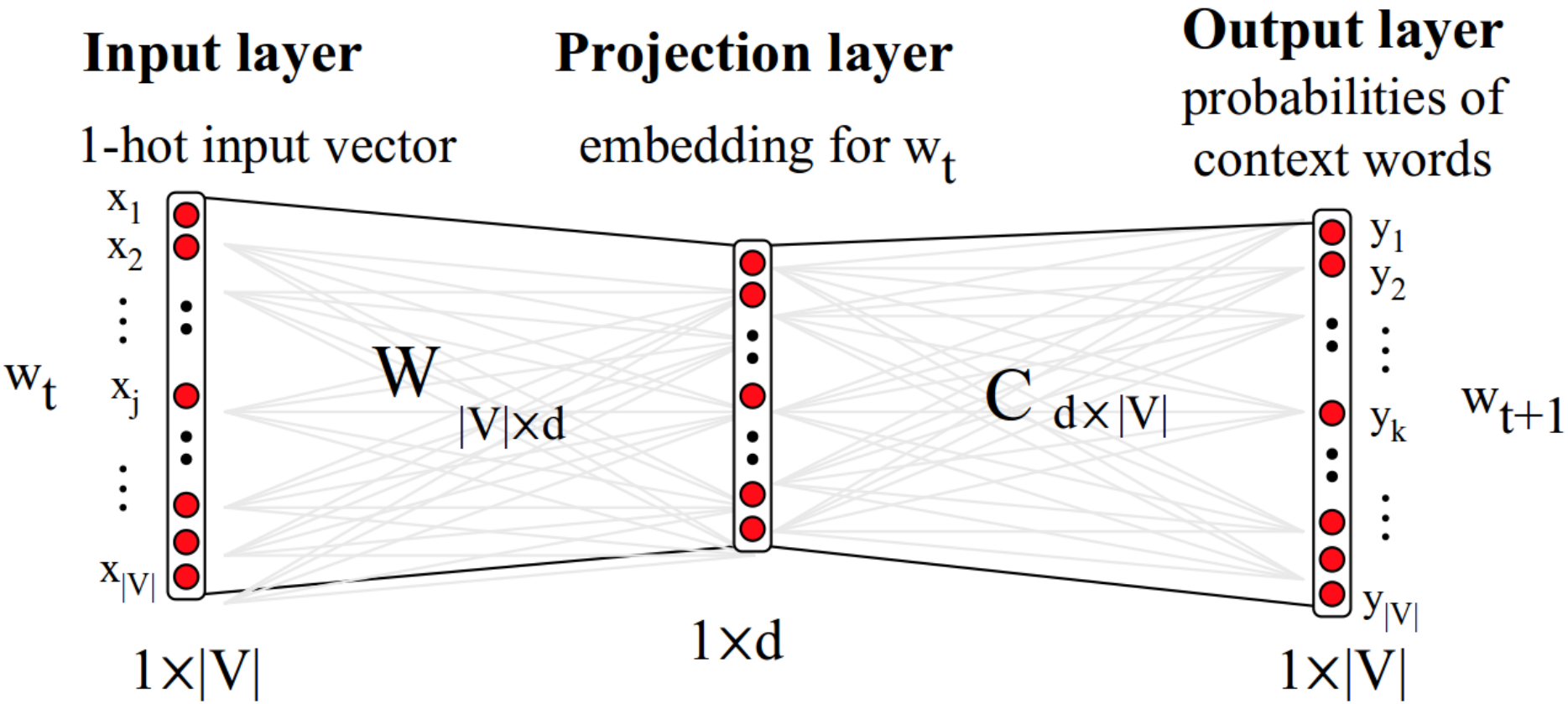
- Jedna warstwa ukryta

Word2Vec jako sieć neuronowa

- Jedna warstwa ukryta
- Zadanie klasyfikacji (przewidywanie kolejnego słowa)

Word2Vec jako sieć neuronowa

- Jedna warstwa ukryta
- Zadanie klasyfikacji (przewidywanie kolejnego słowa)
- Trochę boli liczba klas (rzędu milionów)


Input layer
1-hot input vector
Projection layer
embedding for w_t
Output layer
probabilities of context words
w_t
x_1
x_2
x_j
$x_{|V|}$
$W_{|V|\times d}$
$C_{d\times|V|}$
y_1
y_2
y_k
$y_{|V|}$
w_{t+1}
$1\times|V|$
$1\times d$
$1\times|V|$

Prawdopodobieństwo słowa

- Używamy tzw. Softmax layer

Prawdopodobieństwo słowa

- Używamy tzw. Softmax layer
-

$$P(w_k|w_j) = \frac{\exp(c_k \cdot v_j)}{\sum_{i \in V} \exp(c_i \cdot v_j)}$$

Prawdopodobieństwo słowa

- Używamy tzw. Softmax layer
- $$P(w_k|w_j) = \frac{\exp(c_k \cdot v_j)}{\sum_{i \in V} \exp(c_i \cdot v_j)}$$
- Ponieważ chcemy żeby to było duże, powinniśmy dążyć do:
  a) Zwiększenia licznika (c_k zbliża się do v_j)
  b) Zmniejszenia mianownika (inne c_i oddalają się od v_j)

Jak działa word2vec?

- Zbliżamy słowo do kontekstu, w którym występuje.

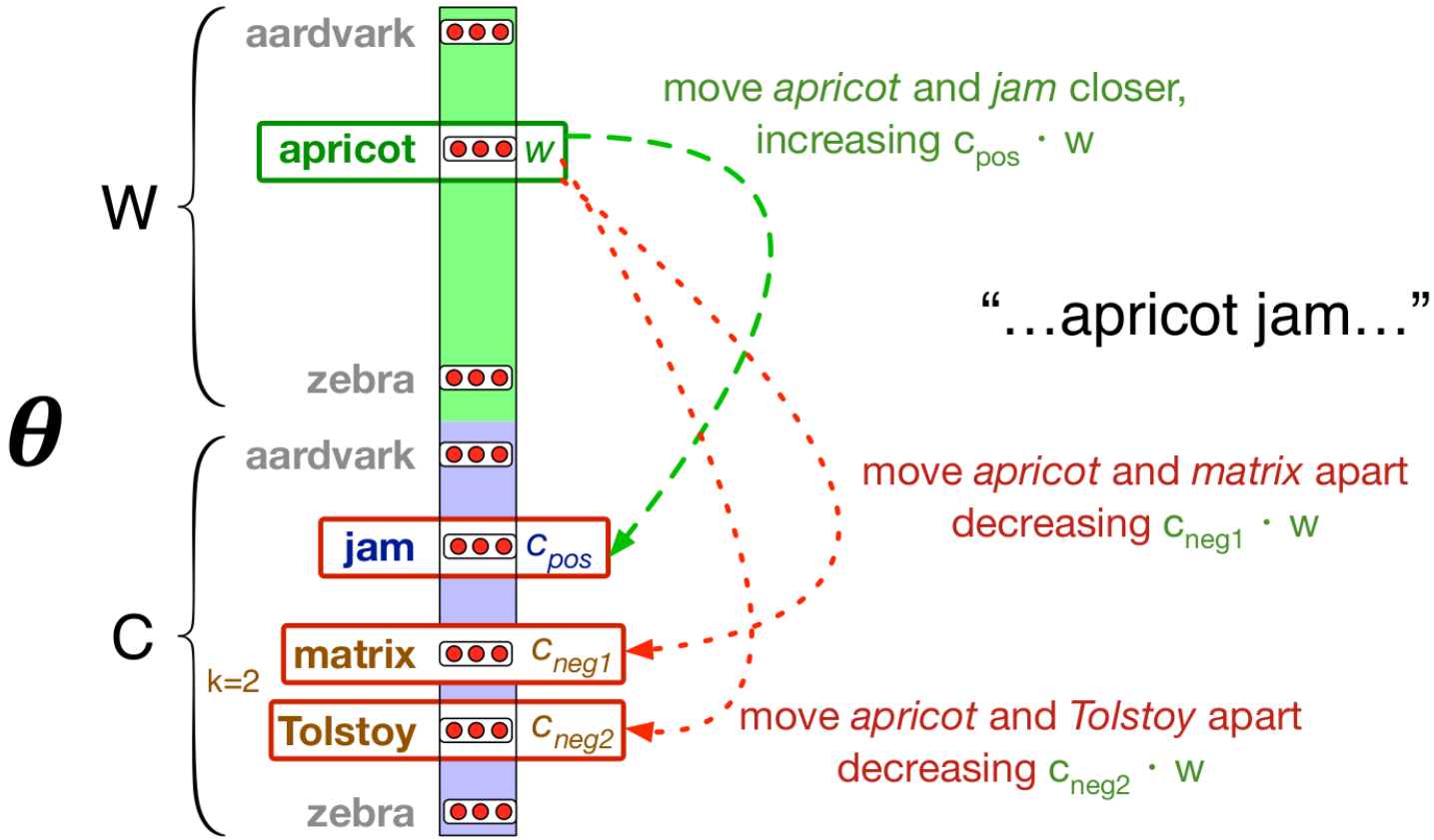
Jak działa word2vec?

- Zbliżamy słowo do kontekstu, w którym występuje.
- Oddalamy słowo od kontekstów, w których nie wystąpiło.

Jak działa word2vec?

- Zbliżamy słowo do kontekstu, w którym występuje.
- Oddalamy słowo od kontekstów, w których nie wystąpiło.**(niektórych)**

Intuition of one step of gradient descent



Word2vec w praktyce

- Używamy biblioteki `gensim`

Word2vec w praktyce

- Używamy biblioteki `gensim`
- Dla korpusu PolEval wynik otrzymamy po kilku(nastu) minutach (i zaraz go zobaczymy).
- Bierzemy korpus zlematyzowany (bo dla polskiego działa lepiej)

Demonstracja

Popatrzmy na osadzenia dla języka polskiego.
(i osadzenia liczb, co nam się jeszcze przydadzą)

Cel na dziś (przypomnienie)

Norwid

Odpowiednie dać rzeczy słowo!

My

Odpowiednią dać słowu rzecz!

Rzeczą będzie wektor w R^n

Takie przypisanie nazwiemy **osadzaniem** słów/tokenów

Co to znaczy odpowiednie?

Co to znaczy odpowiednie?

- Bliskość znaczeniowa oznacza bliskość geometryczną
- Być może: bliskość funkcjonalna również ma odzwierciedlenie geometryczne

Co to znaczy odpowiednie?

- Bliskość znaczeniowa oznacza bliskość geometryczną
- Być może: bliskość funkcjonalna również ma odzwierciedlenie geometryczne
- Chcielibyśmy móc to mierzyć

Co to znaczy odpowiednie?

- Bliskość znaczeniowa oznacza bliskość geometryczną
- Być może: bliskość funkcjonalna również ma odzwierciedlenie geometryczne
- Chcielibyśmy móc to mierzyć

Podobieństwo cosinusowe:

$$\cos(v, w) = \frac{v \cdot w}{|v| \cdot |w|}$$

gdzie v i w to wektory, a $v \cdot w$ to iloczyn skalarny

Co to znaczy odpowiednie?

- Bliskość znaczeniowa oznacza bliskość geometryczną
- Być może: bliskość funkcjonalna również ma odzwierciedlenie geometryczne
- Chcielibyśmy móc to mierzyć

Podobieństwo cosinusowe:

$$\cos(v, w) = \frac{v \cdot w}{|v| \cdot |w|}$$

gdzie v i w to wektory, a $v \cdot w$ to iloczyn skalarny

Uwaga

Odległość euklidesowa nie jest tu używana, ze względu na wrażliwość na skalowanie.

Ocena zanurzeń

Uwaga

Jeżeli interesuje nas podobieństwo, to możemy oceniać zanurzenia patrząc, jak dobrze sobie radzą z relacją podobieństwa.

Ocena zanurzeń

Uwaga

Jeżeli interesuje nas podobieństwo, to możemy oceniać zanurzenia patrząc, jak dobrze sobie radzą z relacją podobieństwa.

Przykładowo, definiujemy klasy:

Ocena zanurzeń

Uwaga

Jeżeli interesuje nas podobieństwo, to możemy oceniać zanurzenia patrząc, jak dobrze sobie radzą z relacją podobieństwa.

Przykładowo, definiujemy klasy:

- **piśmiennicze:** pisak flamaster ołówek długopis pióro

Ocena zanurzeń

Uwaga

Jeżeli interesuje nas podobieństwo, to możemy oceniać zanurzenia patrząc, jak dobrze sobie radzą z relacją podobieństwa.

Przykładowo, definiujemy klasy:

- **piśmiennicze:** pisak flamaster ołówek długopis pióro
- **małe_ssaki:** mysz szczur chomik łasica kuna bóbr

Ocena zanurzeń

Uwaga

Jeżeli interesuje nas podobieństwo, to możemy oceniać zanurzenia patrząc, jak dobrze sobie radzą z relacją podobieństwa.

Przykładowo, definiujemy klasy:

- **piśmiennicze:** pisak flamaster ołówek długopis pióro
- **małe_ssaki:** mysz szczur chomik łasica kuna bóbr
- **okręty:** niszczyciel lotniskowiec trałowiec krążownik pancernik

Ocena zanurzeń

Uwaga

Jeżeli interesuje nas podobieństwo, to możemy oceniać zanurzenia patrząc, jak dobrze sobie radzą z relacją podobieństwa.

Przykładowo, definiujemy klasy:

- **piśmiennicze:** pisak flamaster ołówek długopis pióro
- **małe_ssaki:** mysz szczur chomik łasica kuna bóbr
- **okręty:** niszczyciel lotniskowiec trałowiec krążownik pancernik
- **lekarze:** lekarz pediatra ginekolog kardiolog internista

Ocena zanurzeń

Uwaga

Jeżeli interesuje nas podobieństwo, to możemy oceniać zanurzenia patrząc, jak dobrze sobie radzą z relacją podobieństwa.

Przykładowo, definiujemy klasy:

- **piśmiennicze:** pisak flamaster ołówek długopis pióro
- **małe_ssaki:** mysz szczur chomik łasica kuna bóbr
- **okręty:** niszczyciel lotniskowiec trałowiec krążownik pancernik
- **lekarze:** lekarz pediatra ginekolog kardiolog internista
- **zupy:** rosół żurek barszcz

Ocena zanurzeń

Uwaga

Jeżeli interesuje nas podobieństwo, to możemy oceniać zanurzenia patrząc, jak dobrze sobie radzą z relacją podobieństwa.

Przykładowo, definiujemy klasy:

- **piśmiennicze:** pisak flamaster ołówek długopis pióro
- **małe_ssaki:** mysz szczur chomik łasica kuna bóbr
- **okręty:** niszczyciel lotniskowiec trałowiec krążownik pancernik
- **lekarze:** lekarz pediatra ginekolog kardiolog internista
- **zupy:** rosół żurek barszcz
- **uczucia:** miłość przyjaźń nienawiść

Ocena zanurzeń

Uwaga

Jeżeli interesuje nas podobieństwo, to możemy oceniać zanurzenia patrząc, jak dobrze sobie radzą z relacją podobieństwa.

Przykładowo, definiujemy klasy:

- **piśmiennicze:** pisak flamaster ołówek długopis pióro
- **małe_ssaki:** mysz szczur chomik łasica kuna bóbr
- **okręty:** niszczyciel lotniskowiec trałowiec krążownik pancernik
- **lekarze:** lekarz pediatra ginekolog kardiolog internista
- **zupy:** rosół żurek barszcz
- **uczucia:** miłość przyjaźń nienawiść
- **działy_matematyki:** algebra analiza topologia logika

Ocena zanurzeń

Uwaga

Jeżeli interesuje nas podobieństwo, to możemy oceniać zanurzenia patrząc, jak dobrze sobie radzą z relacją podobieństwa.

Przykładowo, definiujemy klasy:

- **piśmiennicze:** pisak flamaster ołówek długopis pióro
- **małe_ssaki:** mysz szczur chomik łasica kuna bóbr
- **okręty:** niszczyciel lotniskowiec trałowiec krążownik pancernik
- **lekarze:** lekarz pediatra ginekolog kardiolog internista
- **zupy:** rosół żurek barszcz
- **uczucia:** miłość przyjaźń nienawiść
- **działy_matematyki:** algebra analiza topologia logika
- **budynki_sakralne:** kościół bazylika kaplica katedra świątynia synagoga zbór

Ocena zanurzeń

Test elementarny (ABX)

Sprawdzamy, czy:

$$\cos(\text{flamaster}, \text{ołówek}) > \cos(\text{flamaster}, \text{barszcz})$$

(X=flamaster, A=ołówek, B=barszcz)

Ocena zanurzeń

Test elementarny (ABX)

Sprawdzamy, czy:

$$\cos(\text{flamaster}, \text{ołówek}) > \cos(\text{flamaster}, \text{barszcz})$$

(X=flamaster, A=ołówek, B=barszcz)

- Losujemy dużo testów elementarnych.

Ocena zanurzeń

Test elementarny (ABX)

Sprawdzamy, czy:

$$\cos(\text{flamaster}, \text{ołówek}) > \cos(\text{flamaster}, \text{barszcz})$$

(X=flamaster, A=ołówek, B=barszcz)

- Losujemy dużo testów elementarnych.
- Zliczamy te, które przeszliśmy pozytywnie – i to jest jakość naszych zanurzeń

Ocena zanurzeń

Test elementarny (ABX)

Sprawdzamy, czy:

$$\cos(\text{flamaster}, \text{ołówek}) > \cos(\text{flamaster}, \text{barszcz})$$

(X=flamaster, A=ołówek, B=barszcz)

- Losujemy dużo testów elementarnych.
- Zliczamy te, które przeszliśmy pozytywnie – i to jest jakość naszych zanurzeń

Uwaga

Pamiętajmy, że całkiem losowe zanurzenie mają wartość 0.5.

Relacje geometryczne w Word2Vec

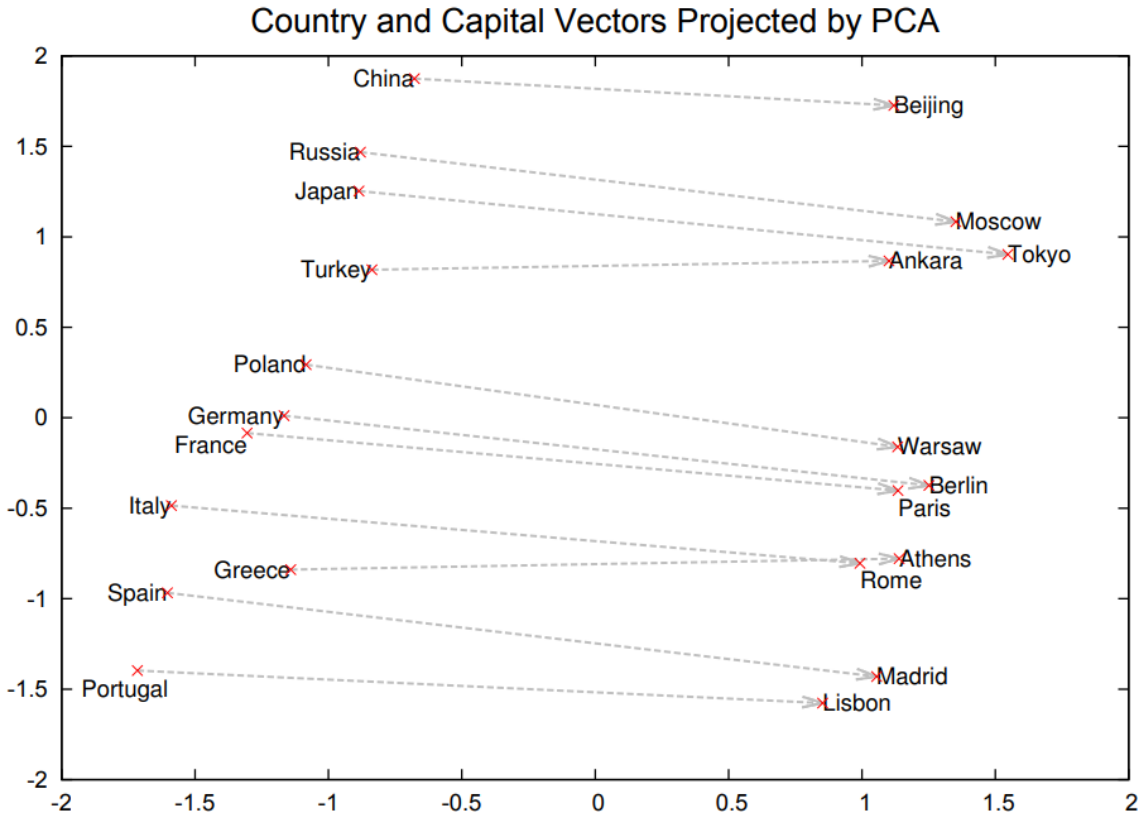


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

- Pytanie: jak to można wykorzystać?

Relacje geometryczne w Word2Vec

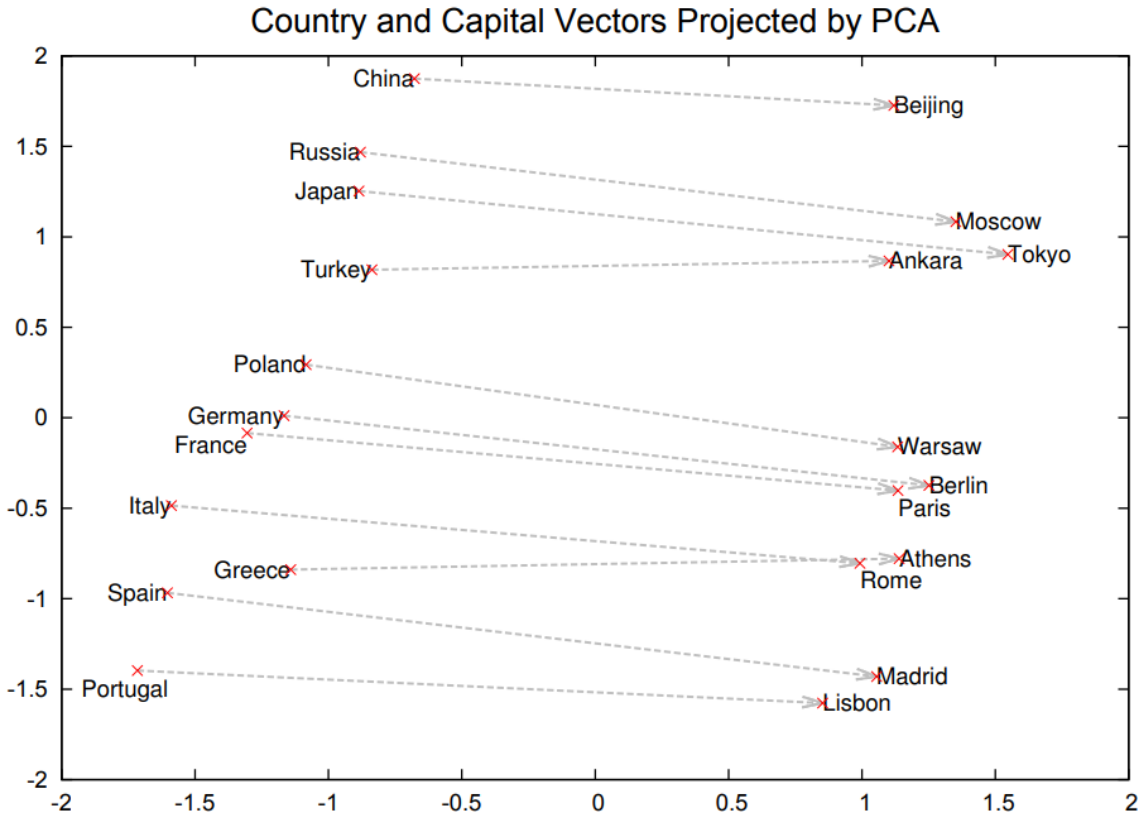


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

- Pytanie: jak to można wykorzystać?
- Odpowiedź: nawet tak prosty model jak Word2Vec zawiera **wiedzę** (zawartą w osadzeniach słów i w (uśrednionych) wektorach dla relacji)