Wprowadzenie do NLP

Bydgoszcz, 23.10.19

ja: Artur Zygadło

- VI LO Bydgoszcz (matura 2012)
- studia:
  - robotyka, informatyka – Politechnika Warszawska
- praca:
  - (...), Samsung, CERN, deepsense.ai (Data Scientist)
- hobby:
  - sport, języki obce, *Jeden z dziesięciu*

Czym jest NLP?

NLP, czyli programowanie neurolingwistyczne ...
https://www.poradnikzdrowie.pl › psychologia › rozwoj-osobisty › nlp-cz...
14 lis 2017 - Programowanie neurolingwistyczne (**NLP**) to technika mająca na celu modyfikowanie działania swojego umysłu, a szczególnie możliwości ...

Zrobię wszystko co chcesz, dzięki NLP - Katarzyna Płuska
https://www.katarzynapluska.pl › zrobie-wszystko-co-chcesz-dzieki-nlp
3 lip 2017 - Zrobię wszystko co chcesz, dzięki **NLP** Czy **NLP** to tylko manipulacja innymi? Czy samorozwój, kontrola swoich uczuć i umysłu? Poznaj ...

NLP dla bogatych i naiwnych | Religia dla bogatych - Polityka.pl
https://www.polityka.pl › tygodnikpolityka › nauka › 1593981,1,nlp-dla-bo...
30 wrz 2014 - Choć nazwa neurolingwistyczne programowanie wywołuje komputerowo-naukowe skojarzenia, to **NLP** z nauką ma niewiele wspólnego.

Czym jest NLP?

Wyszukiwania podobne do: nlp

techniki nlp **w związku**	nlp **szkolenie**
techniki nlp **pdf**	**trener** nlp
nlp **książki**	nlp **podryw**
nlp **w sprzedaży**	nlp **youtube**

Czym jest NLP w kontekście SI?

NLP = Natural Language Processing

czyli

przetwarzanie języka naturalnego

Czym jest NLP w kontekście SI?

przetwarzanie języka naturalnego

czyli

jak nauczyć komputer rozumieć tekst / mowę

Zastosowania NLP

rozpoznawanie mowy

napisz, co zostało powiedziane

Zastosowania NLP

modelowanie języka

podaj najbardziej prawdopodobne kolejne słowo

Zastosowania NLP

tłumaczenie maszynowe

przetłumacz zdanie z języka A na język B

Zastosowania NLP

analiza sentymentu

określ wydźwięk zdania

(pozytywny / negatywny / neutralny)

Zastosowania NLP

automatyczna sumaryzacja

streść długi tekst do kilku zdań

(tl;dr)

Zastosowania NLP

generowanie języka naturalnego (NLG)

wygeneruj tekst wyglądający jak napisany przez człowieka

Zastosowania NLP

Question Answering

odpowiedz na pytanie

np. *dokąd tupta nocą jeż?*

Zastosowania NLP

czatboty

nieustrukturyzowane konwersacje

rozrywka, terapia, obsługa klienta

Zastosowania NLP

systemy dialogowe

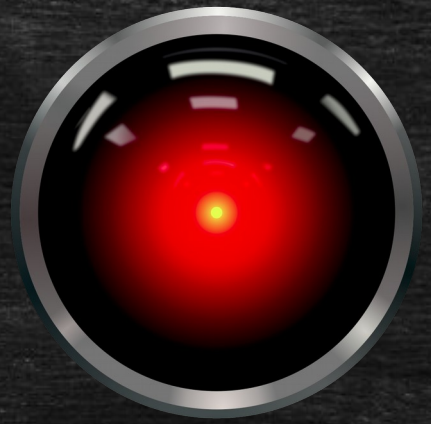
dialog zorientowany na wykonanie zadania

np. asystenty głosowe

Dlaczego NLP jest trudne?

- polisemia (wieloznaczność)
- ironia
- mowa: homofonia
- kontekst
- specyfika danego języka
- ...
- niedobór danych, jeśli nie po angielsku

NLP w kulturze



Dave: “Open the pod bay doors, HAL.”

HAL 9000: “I’m sorry Dave, I’m afraid I can’t do that.”

2001: Odyseja kosmiczna (1968)

NLP w życiu

Dave: “Hey Siri, open the pod bay doors.”

Siri: “Opening *Pod Bay* by *The Doors* on Spotify.”

Życie (2019)

Historia NLP

od Turinga do Transformera

Test Turinga (1950)

- dialog: człowiek, maszyna + sędzia
- sędzia stara się odróżnić człowieka od maszyny

- do tej pory nie został zaliczony
  - “kontrowersje”
    - *PARRY* (1972) – schizofrenia
    - *Eugene Goostman* (2014) – 13-latek z Ukrainy

ELIZA (1966)

- czatbot udający psychoterapeutę
- oparty na regułach
  - przekształcenie wypowiedzi *(my –> your)*, słowa kluczowe lub generyczna odpowiedź

Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED

Statystyczne NLP (~1980-2010)

- duże zbiory danych – korpusy
- inżynieria cech (+ wiedza lingwistyczna)
- klasyczne algorytmy uczenia maszynowego
  - drzewa decyzyjne
  - klasyfikator bayesowski
  - ukryte modele Markowa
  - n-gramy
  - ...

IBM Watson i *Jeopardy!* (2011)

Asystenty głosowe (2011+)

Siri
(Apple)

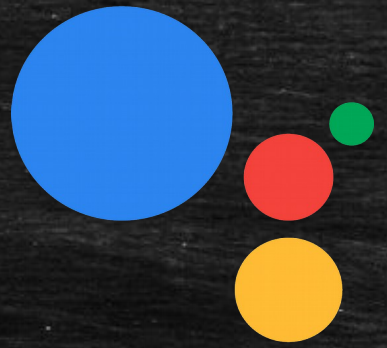
2011

Cortana
(Microsoft)

2014

Alexa
(Amazon)

2014



Google Assistant

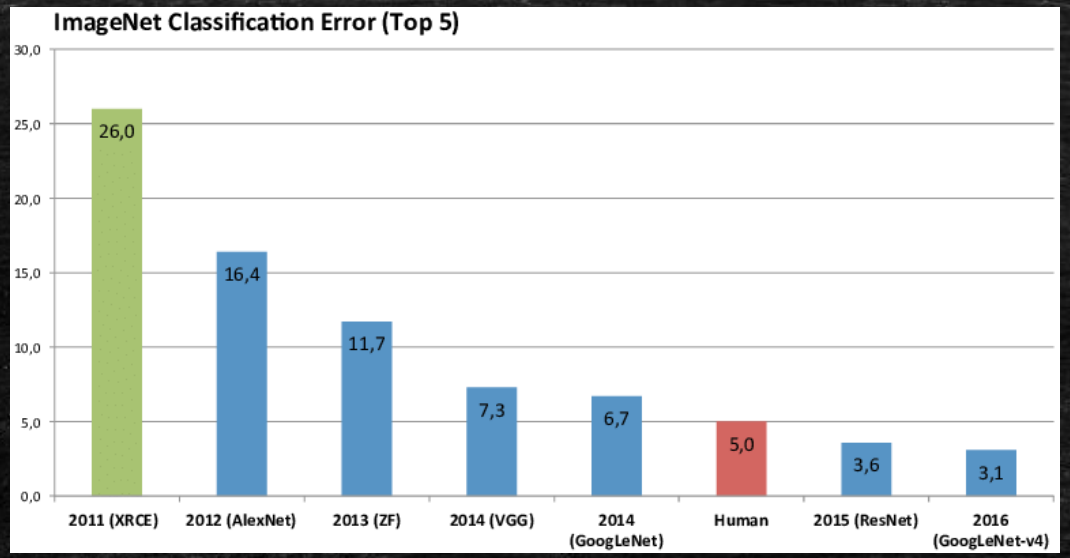
2016

Bixby
(Samsung)

2017

“Deep learning tsunami”

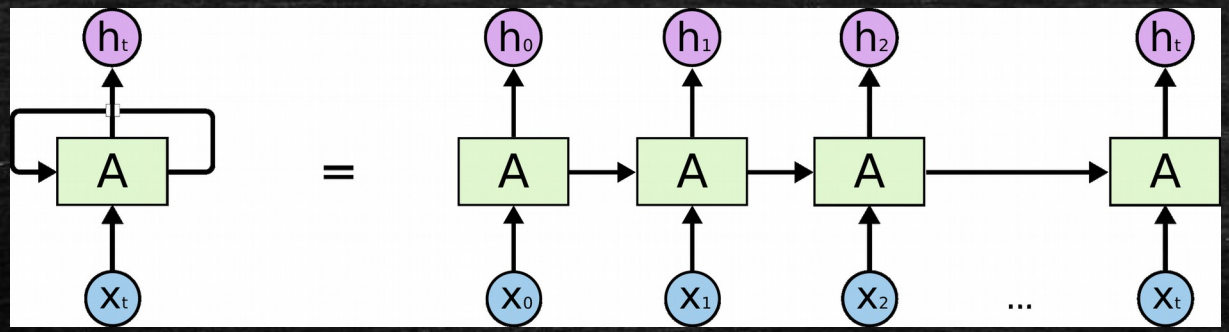
- głębokie uczenie maszynowe – rewolucja
  - w wizji komputerowej – od 2012



  - w NLP – od ok. 2014/2015

Rekurencyjne sieci neuronowe

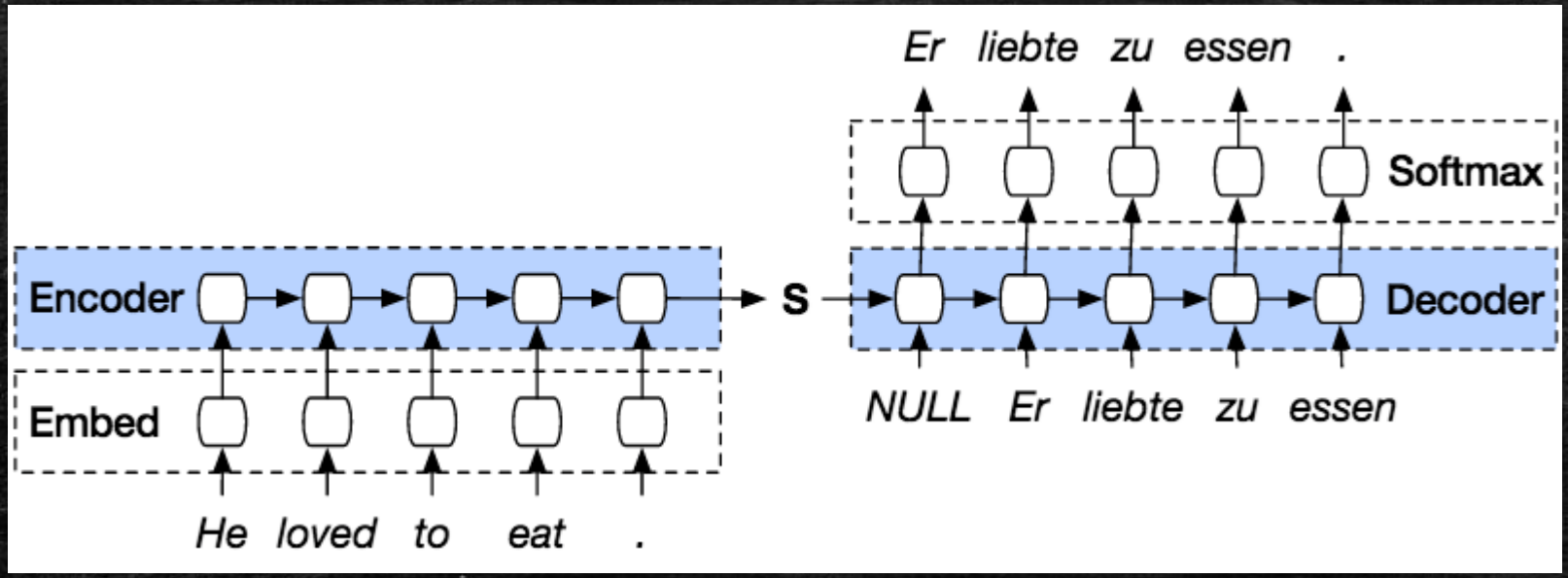
- modelowanie sekwencji
- wyjście zależy od wejścia i od stanu w poprzedniej chwili



- problem z długimi zależnościami
  - LSTM (1997), GRU (2014)

seq2seq, atencja (2014-2015)

- seq2seq – dwie sieci rekurencyjne: enkoder i dekoder
- atencja – sieć umie rozpoznawać istotne fragmenty

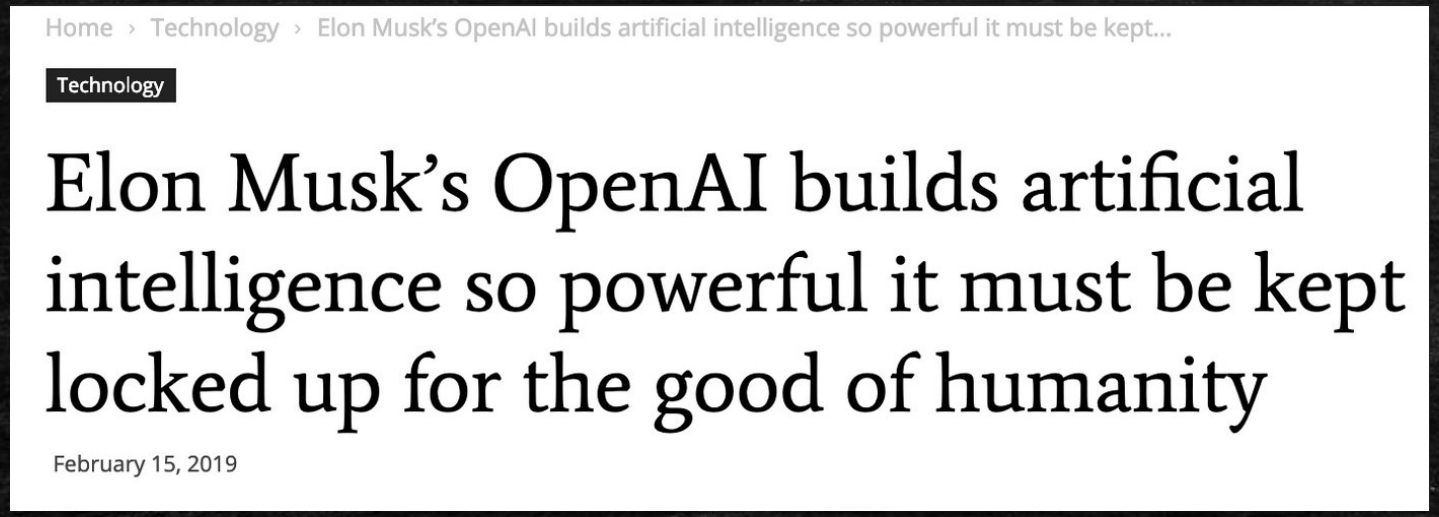


Transformer (2017)

- sieci rekurencyjne – już niemodne
- *Attention Is All You Need*

GPT-2 (2019)

- bardzo duży Transformer, wytrenowany na 40GB tekstu
- zbyt groźny, by upublicznić pełną wersję (*fake news*)



GPT-2 (2019)

- https://talktotransformer.com

Completion

Today I am speaking about natural language processing. I would like to explain what that means to me and what it could mean to anyone.

In this talk I'll be talking about how natural language processing and natural language understanding, can be applied in any language. In the following I will briefly cover a few specific cases (some of them we know of). Then we'll explore the underlying problems of natural language processing that make them difficult to generalize.

Finally I will briefly touch on a couple of research areas where natural language processing is being applied and a few specific papers which are well known in the field. Lastly I'll close with a few words on some current issues in the field.

NLP w kulturze

Luke: “Do you understand anything they’re saying?”

C-3PO: “Oh, yes, Master Luke! Remember that I am fluent in over six million forms of com—”

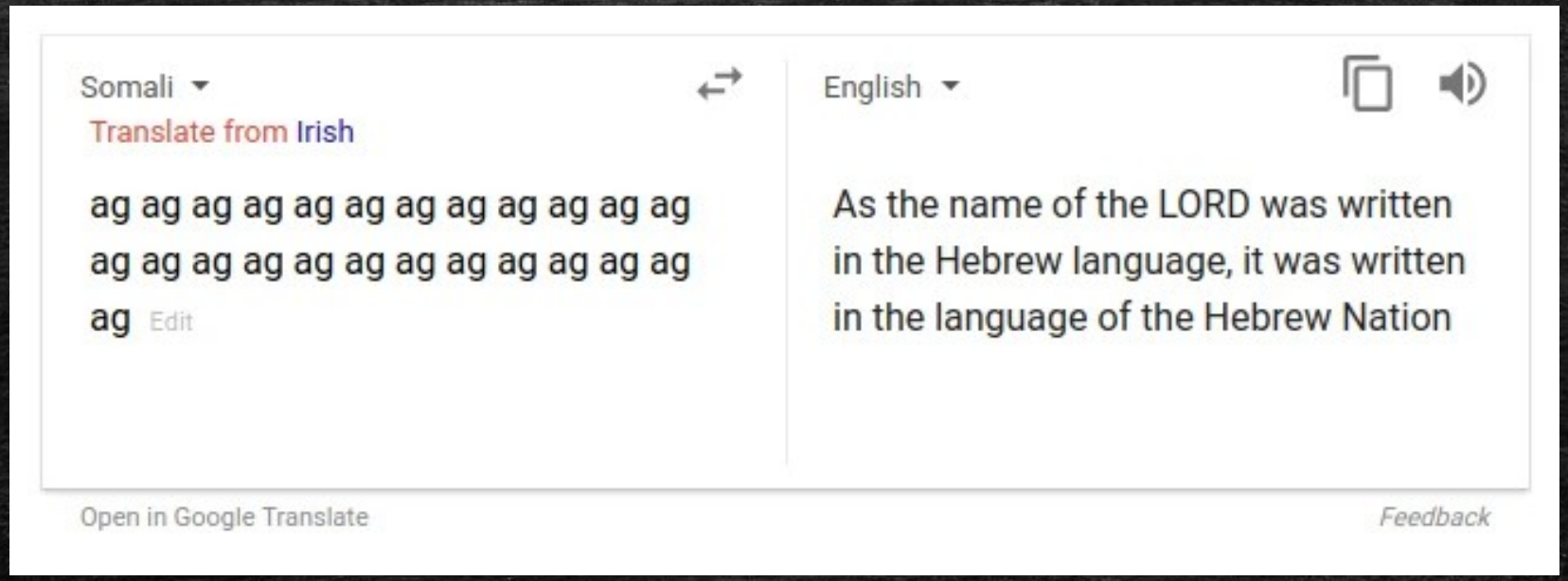
Han Solo: “What are you telling them?”

C-3PO: “Hello, I think. I could be mistaken.”

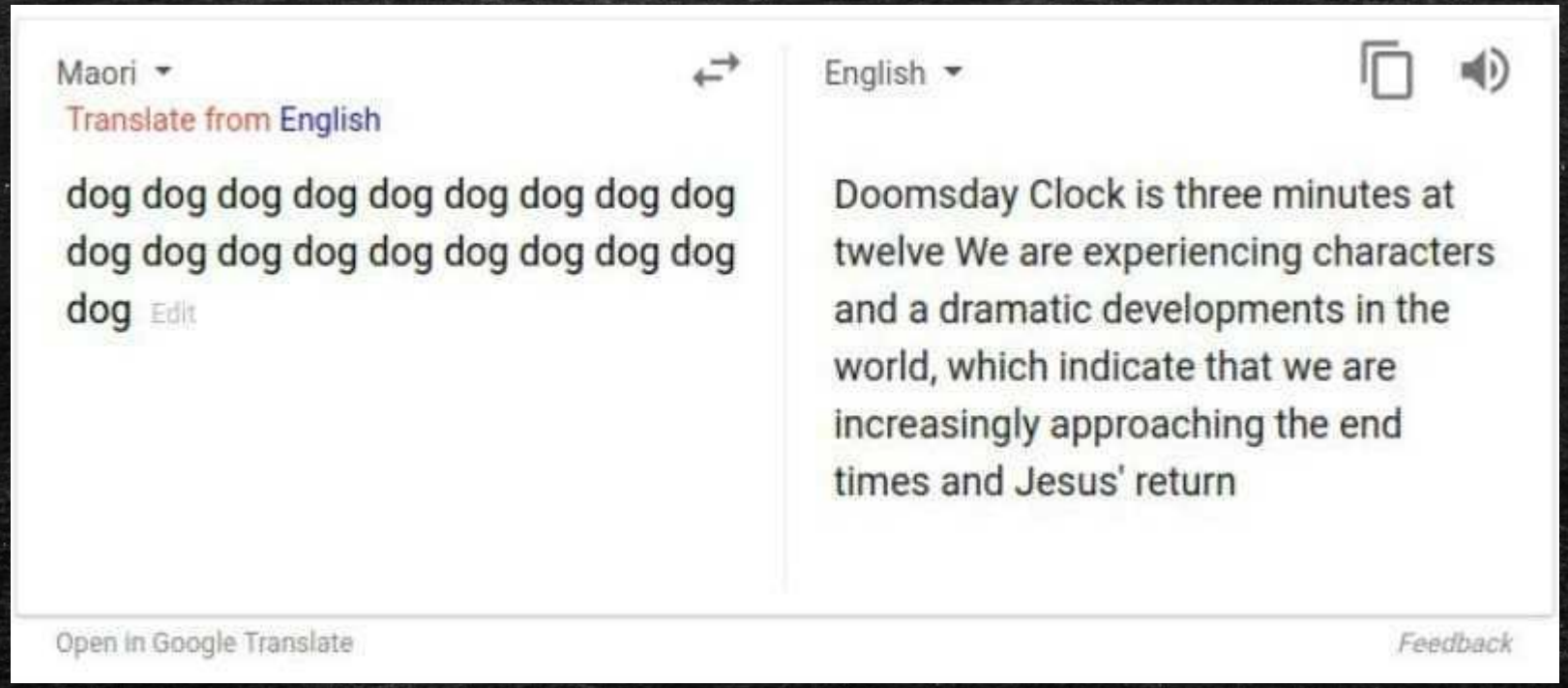
– C-3PO, o komunikacji z Ewokami

Gwiezdne Wojny: część VI – Powrót Jedi (1983)

Błędy Google Translate



Błędy Google Translate



Jak komputer widzi słowa?

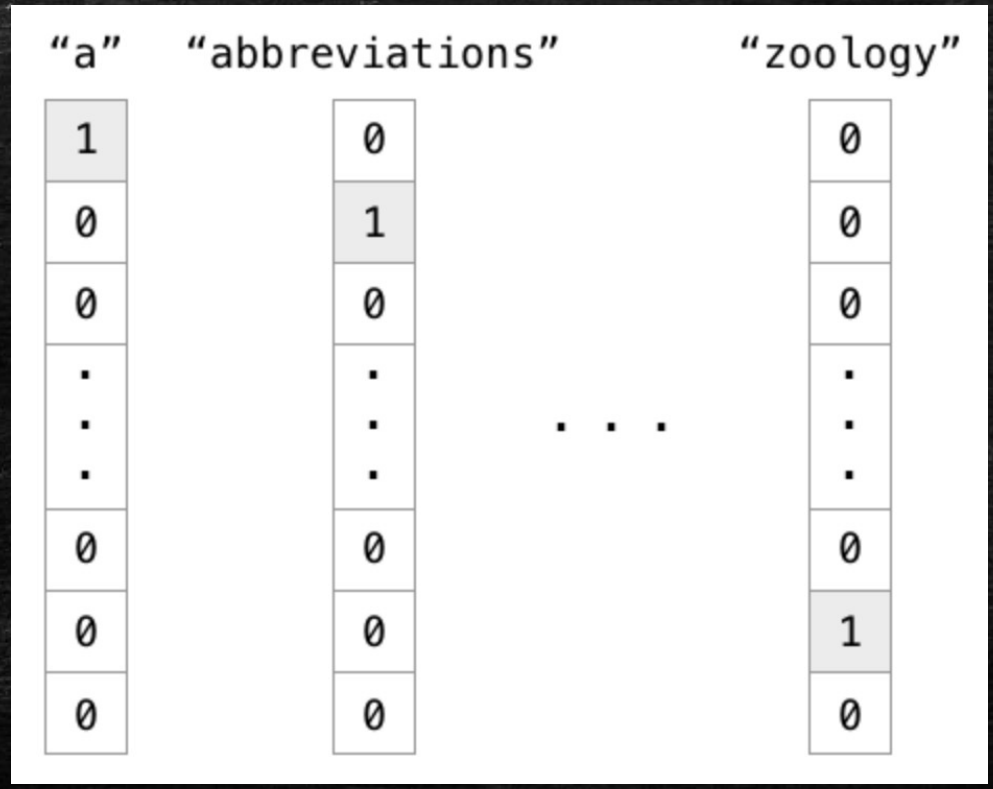
Słowa, słowa, słowa...
William Szekspir, *Hamlet*

Podejście naiwne - słownik

- lista słów posortowanych np. alfabetycznie
- każdemu słowu przypisana kolejna liczba
  - trudno wyrazić podobieństwo znaczeń słów

Wektory typu “one-hot”

- dużo zer – nieefektywne obliczeniowo / pamięciowo



Semantyka dystrybucyjna

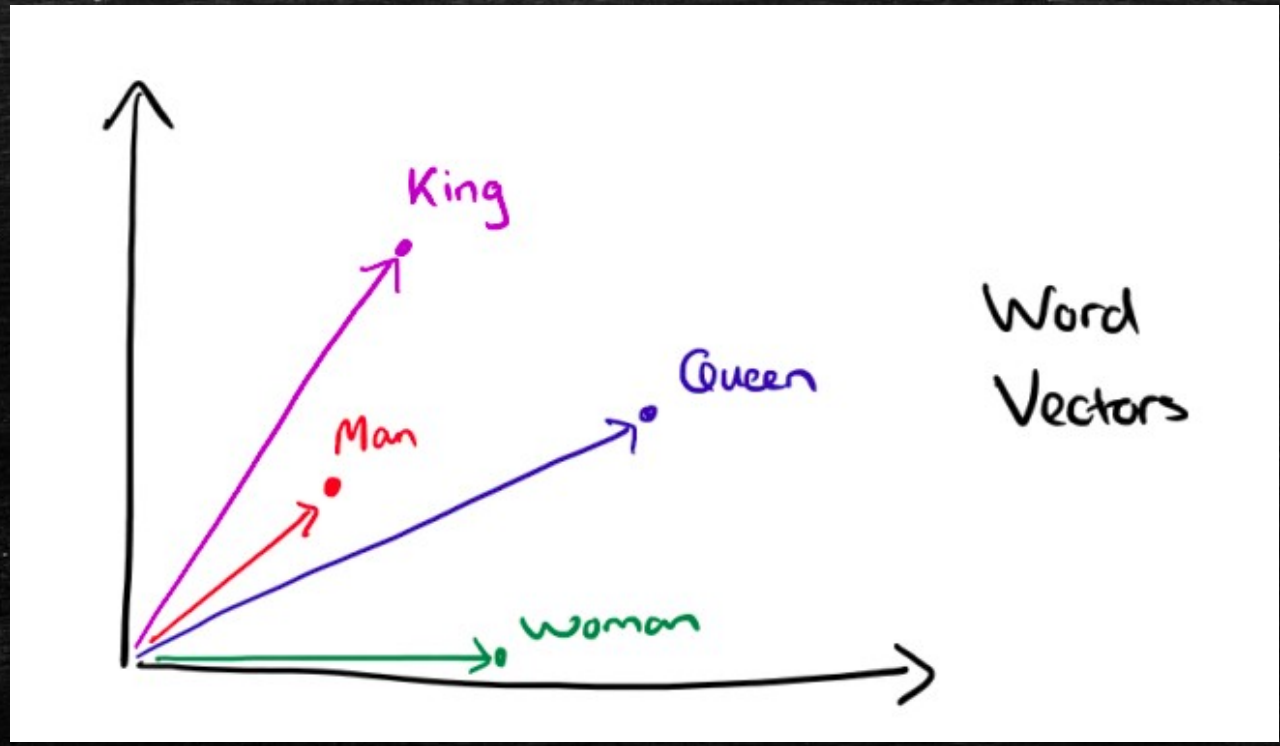
You shall know a word by the company it keeps.

J. R. Firth, 1957

czyli: podobne znaczeniowo słowa - podobne konteksty

Word embeddings

- słowa jako wektory liczb rzeczywistych
  - typowo: 100- lub 300-wymiarowe



word2vec (2013)

- algorytm uczenia się *word embeddingów*
  - z danych
  - iteracyjnie (coraz lepsze wektory)
  - płytka sieć neuronowa
  - *embeddingi* – wagi warstwy ukrytej

- wady:
  - jeden wektor dla wyrazów wieloznacznych
  - problem z tzw. OOV (*out-of-vocabulary*)

ELMo, BERT, ... (2018+)

ELMo

bi-LSTM, 2018

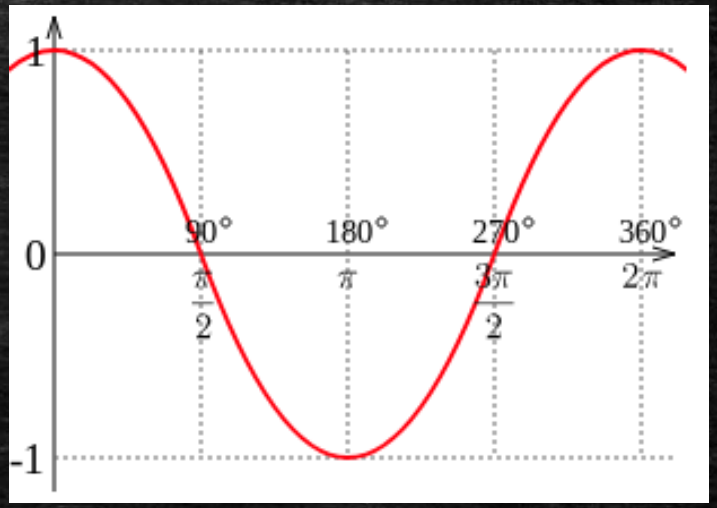
BERT

Transformer, 2019

Podobieństwo cosinusowe

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta$$

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$



- A ∥ B: cos(0) == 1, A ⊥ B: cos(90°) == 0
- im bardziej podobne słowa, tym cos(θ) bliżej 1

word2vec demo

no risk no fun

Linki

- Speech & Language Processing (Jurafsky, Martin)
  - https://web.stanford.edu/~jurafsky/slp3
- NLP with Deep Learning (Stanford CS224N)
  - playlista na YouTube, *Winter 2019*
- *Visualizing machine learning one concept at a time*
  - Jay Alammar*, http://jalammar.github.io*

- język polski (dane, modele):
  - http://clip.ipipan.waw.pl/LRT

Dziękuję za uwagę