

Sztuczna Inteligencja

Analiza języka naturalnego

Osiągnięcia i tłumaczenie maszynowe

Włodzisław Duch

Katedra Informatyki Stosowanej UMK

Google: Włodzisław Duch

Tłumaczenie maszynowe

Pismo *MT, Machine Translation*, istnieje od 1954 roku

1957, Noam Chomsky, książka *Syntactic Structures*, spowodowała szybki rozwój lingwistyki matematycznej.

Języki sztuczne mają dobrze rozwiniętą teorię, ale naturalne znacznie gorzej.

Maszynowe tłumaczenie tekstów technicznych odbywa się rutynowo. Wystarczy „wygładzanie” przez ekspertów.

Podstawowe idee:

1. Uwzględniaj kontekst, okienko: słowo + kilka wyrazów otoczenia.
2. Techniki deszyfracji: traktuj tekst jako szyfr.
3. Dodaj reprezentację sensu w języku pośredni – machinesese.
Nie przekładaj z danego języka na każdy inny, np. w UE jest 27 języków.
4. Jeśli masz do dyspozycji wiele dobrze przetłumaczonych tekstów (korpusy) stosuj **statystyczne techniki** uczenia maszynowego by tłumaczyć całe frazy.

Ontologie

Od końca lat 1960 ontologie pojawiły się w informatyce by opisywać „to co jest” lub „może być” w przypadku działań inteligentnych agentów.

Jednoznaczność opisu pojęć wymaga kategoryzacji i hierarchizacji.

W wielu dziedzinach – medycynie, bioinformatyce – są tysiące nowych pojęć i trzeba je precyzyjnie zdefiniować w relacji do naszej siatki pojęciowej.

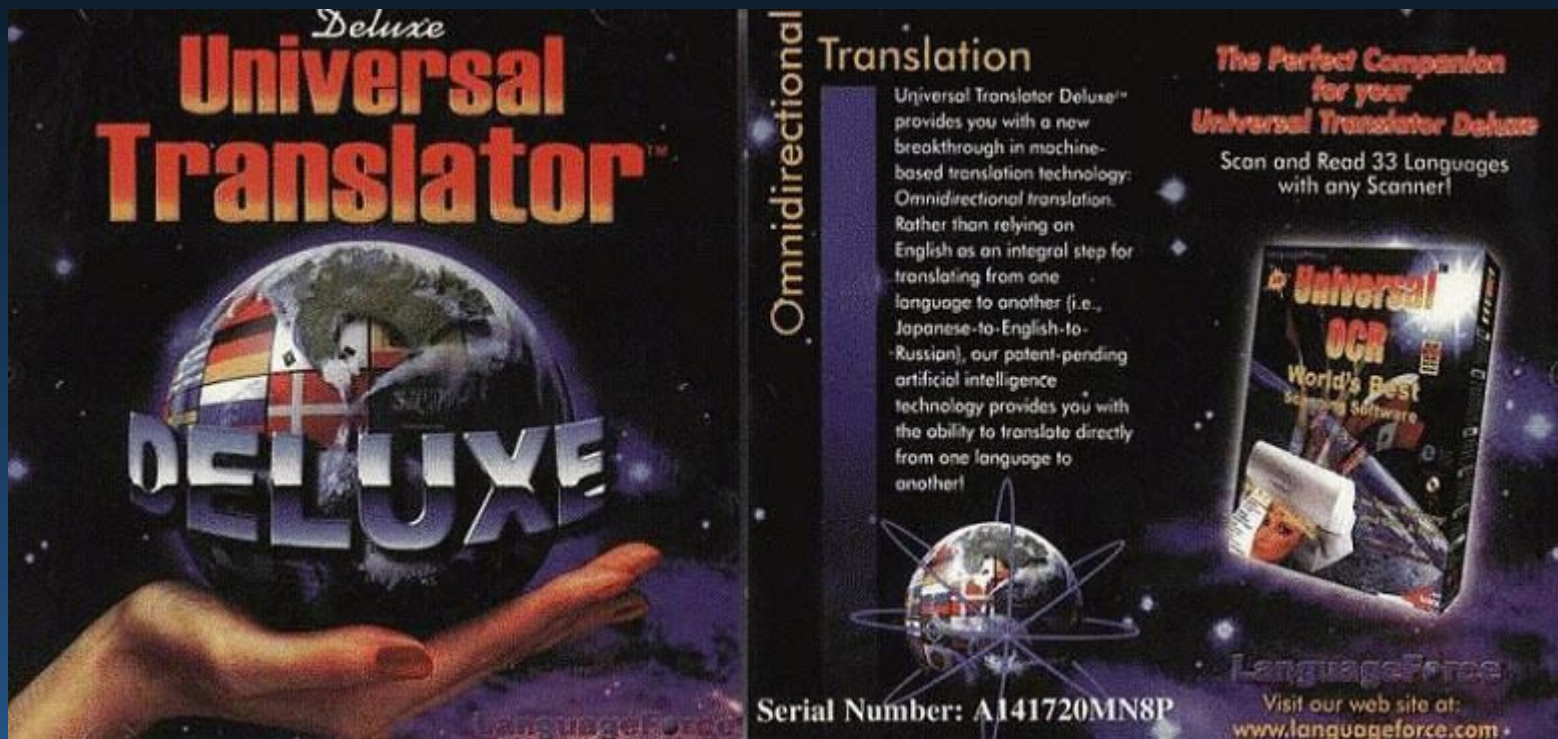
- kategoryzacja – przyporządkowanie symbolu/terminu występującego w wiedzy deskryptywnej agenta do określonej grupy obiektów występujących w zadanej dziedzinie jego zainteresowania/aktywności, np. termin: kot → klasa: koty, pojęcie kot.
- hierarchizacja – umiejscowienie określonej klasy w hierarchicznej strukturze klas opisujących daną dziedzinie. Każda klasa w hierarchicznej strukturze posiada także cechy dziedziczone z klas nadrzędnych.

Inżynieria ontologii: zasady, metody i narzędzia do konstrukcji i rozwoju ontologii. Metodologia inżynierii ontologii dotyczy konstruowania i utrzymywania ontologii, w tym celu stworzono wiele narzędzi.

Więcej: **Inżynieria ontologii** (M Morzy, A. Ławrynowicz, Pol. Poznańska).

MT - wczesne przykłady

Tłumaczenie tekstów: 28 systemów dla 38 języków (2011);
<http://www.foreignword.com/Tools/transnow.htm>



Deluxe
Universal Translator™

DELUXE

Omnidirectional Translation

Universal Translator Deluxe™ provides you with a new breakthrough in machine-based translation technology: *Omnidirectional translation*. Rather than relying on English as an integral step for translating from one language to another (i.e., Japanese-to-English-to-Russian), our patent-pending artificial intelligence technology provides you with the ability to translate directly from one language to another!

The Perfect Companion for your Universal Translator Deluxe

Scan and Read 33 Languages with any Scanner!

Universal OCR
World's Best Scanning Software

LanguageForce
Visit our web site at:
www.languageforce.com

Serial Number: A141720MN8P

MT- przykłady PL

InterTran, www.tranexp.com, wykładał się na słynnych idiomach:

The spirit is willing but the flesh is weak.

Ten duch jest chętny oprócz ten ciało jest słaby.

Out of sight out of mind.

Niewidoczny z umysł.

Warto spróbować różne idiomy/przysłowia z Google Translate.

The NeuroTran translation engine is grammar aware for a selected number of languages. Translation engine means the part of the software which does the translation.

Ten [NeuroTran] tłumaczenie silnik jest gramatyka świadomy pod kątem pewien wybrany liczba od języki. Tłumaczenie silnik oznacza ten obowiązek od ten oprogramowanie który nie wymaga ten tłumaczenie.

Podejście statystyczne

Zamiast reguł i gramatyki trzeba mieć bardzo dużo przykładów i szukać podobnych konstrukcji.

Statystyczne podejście do języka opiera się głównie na ko-okurencji słów: jak często dane słowo S jest w pobliżu innych słów W ?

Słowo \Rightarrow unormowany wektor częstości $V(W|S)$.

Angielski-Francuski 2.5 mld par zdań.

Google Translate: początkowo tłumaczenia na polski były słabe, widać było duże braki gramatyczne, bo mamy język fleksyjny, i mało tekstów w obu językach. Tłumaczenia chińsko-angielskie są zrozumiałe.

Od końca 2016 roku sytuacja się znacznie poprawiła dzięki zastosowaniu neuronowych algorytmów głębokiego uczenia. Liczba błędów spadła o 60-90%.

Art. w Science o Google Translate.

10 years of Google Translate.

W służbie oszustów ...

- Twój adres e-mail została wybrana w trybie online w tym tygodniu PRYZNANIA Promo losowych online, e-głosowanie wybór, Twoja remis ma całkowita wartosc 500,000.00 PLN i konta e-mail wygrał ten promocyjny i maja kontakt ze szczegółami ponizej pana Charles Adams Kontakt E-mail:(powerball.charles@yahoo.co.uk)
- Wymagania obejmują:
 - 1.Full nazwisko: 2.Home Adres: 3.Age:
 - 4.Occupation: 5.Country: 6.phone:
 - Znowu gratulacje .
- We konto poczty dziękuje za szybkie zajecie sie ta matter.Please zrozumiec, ze jest to srodek bezpieczenstwa przeznaczonych do ochrony uzytkowników i konta poczty internetowej.

Ostatnie lata

Google Translate już od 2016 poczynił znaczne postępy używając metod statystycznych, a następnie dużej sieci neuronowej,

Google Neural Machine Translation (GNMT):

Działa na PC i platformach mobilnych, od 2020 już w ponad 100 językach. Potrzebuje dużego korpusu do treningu, używa szeroki kontekst.

Głębokie uczenie znacznie poprawiło metody statystyczne, ale nadal są błędy nawet w specjalistycznych tekstach, np:

Recently, the effect on the vagal afferent pathway to the frontal cortical areas has been proposed.

Niedawno zaproponowano wpływ na drogę dopochwową azylową do czołowych obszarów korowych.

Po dodaniu GNMT całkiem dobrze (Google Translate):

Ostatnio zaproponowano wpływ na aferentną drogę nerwu błędnego do czołowych obszarów korowych.

DeepL – doskonały polski system, też robi błędy w specjalistycznych tekstach.

Ostatnio zaproponowano wpływ na wagalną drogę aferentną do obszarów czołowych kory.

Tłumaczenie: więcej informacji

[Historia tłumaczenia maszynowego.](#)

<http://www.mtranslator.pl/>

<http://www.foreignword.com/>

Moje linki „Machine translation”

<http://www.is.umk.pl/~duch/IR.html>

[EuroMatrix](#) – wszystkie języki Europy, hybrydowe systemy statystyczno-regułowe.

Nagroda Loebnera

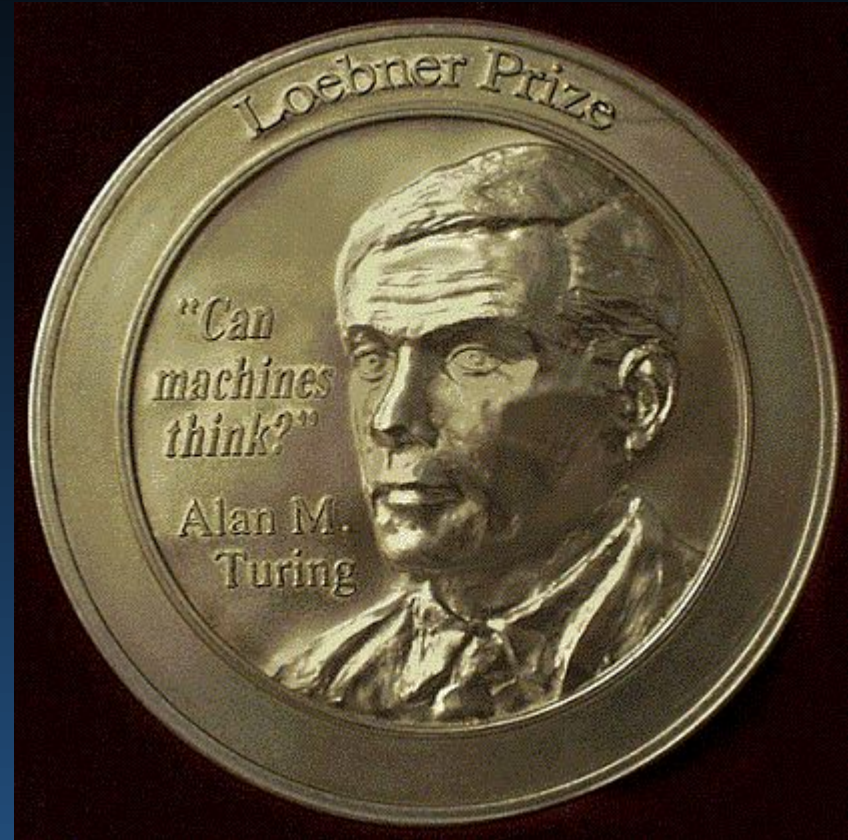
Nagroda przyznawana od 1994 roku,
ograniczony test Turinga, pytania tekstowe,
25 K\$ nagrody jeśli 30% sędziów uzna
program za człowieka;

dla audio-wizualnej interakcji 100 K\$.

Sędziowie mają >15 minut na konwersację

Ocena programów i ludzi, na ile
sztuczny/prawdziwy w skali 1-10.

Tylko 10% sędziów się myli (2000).



<http://www.aisb.org.uk/events/loebner-prize>

Oryginalna gra „w przekonywanie”, że się jest M/F Turinga odbyła się 4/2005.

Przykład cd.

J002: where do you come from

PROGRAM: I am from Earth.

J002: well that's very specific indeed. what are your interests>

PROGRAM: Would you like to have a drink with me later? let's not get too personal.

J002: what do you work as?

P. I've worked on a lot of PC's and some mini computers and mainframes when I was in college.

W 1994 roku tematami były koty i psy, kriogenika, Rolling Stones, proces O.J. Simpsona, zwyciężył program dyskutujący o seksie!

Zawody w 1998 roku w Sydney nie miały ograniczeń tematu konwersacji. Programy często przejmują inicjatywę, co ułatwia oszukanie sędziego.

Oceny: 1 – najbardziej ludzki, 10 – najmniej ludzki.

Wyniki 2000

	1	2	3	4	5	6	7	8	9	10	Median	Mean	Responsiveness	Human Quality
Human#3	3	2	1	2	2	1	1	1	4	3	2	2	9.05	8.75
Human#4	1	1	3	3	1	4	3	3	1	1	2	2.1	9.25	9.35
Human#2	2	9	4	1	3	2	2	4	2	2	2	3.1	7.65	9.00
Human#1	5	7	2	4	4	3	5	2	3	4	4	3.9	7.20	7.80
Alice	4	3	9	10	8	6	6	10	6	5	6	6.7	2.95	2.35
e-Brain	6	8	5	6	6	7	9	6	9	6	6	6.8	3.90	3.50
A-Life	8	6	6	5	10	10	4	5	7	10	6.5	7.1	3.81	3.75
ChatRascal	7	4	7	7	5	5	8	8	5	7	7	6.3	3.70	3.60
Talk	10	5	8	8	7	8	7	7	8	8	8	7.6	1.94	2.33
Algernon	9	10	10	9	9	9	10	9	10	9	9	9.4	0.28	0.56

Program *Alice* oszukał sędziego 1 i 2.

Wyniki 2006

Place	Average Rank	Program	Author(s)
1st	3.75	Joan	Rollo Carpenter
2nd	3	Ultra Hal	Robert Medeksza
3rd	2.25	Cletus	Noah Duncan
4th	1	John	R. Churchill & M-C Jenkins

University College London, 17.09.2006.

Judge: Do you know any good jokes?

Joan: I don't even know any bad ones. I just can't remember them for very long.

Judge: Tell me a joke.

Joan: Sorry, I'm not very good at telling jokes.

Judge: Please tell a joke.

Joan: What do you get if you walk under a cow?

Judge: I don't know?

Joan: A pat on the head.

Judge: That's very good.

Joan: I know.

Wyniki z ostatnich lat

Loebner Prize 2010, 23.10 na California State University, Los Angeles

4 programy, 4 osoby, 4 sędziów.

Tylko jeden z sędziów (profesor, organizator) pomylił człowieka z programem.

W 2012 roku program Chip Vivant oceniono na 1.25, gdzie 1=człowiek,

Cleverbot nabrał 59.3% spośród 1334 osób, czy jednak przeszedł test Turinga?

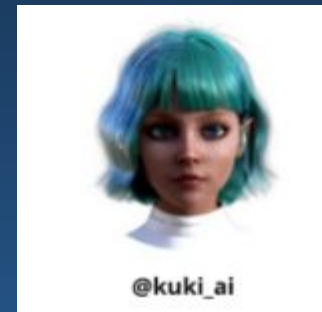
Loebner Prize 2017:

Mitsuku (27 points), Midge (23 points), Uberbot (21 points), Alice bot, który był długo najlepszy wypadł znacznie gorzej (12 points).

Ludzie otrzymali w tym konkursie 34-40 punktów.

<http://www.aisb.org.uk/events/loebner-prize>

W ostatnich latach konkursu (2016-2019) zwyciężał chatbot Kuki. Konkurs nie był od tego czasu organizowany.



Przykład konwersacji: Mitsuku

Kuki/Mitsuku 5 razy wygrał w ostatnich latach konkures Loebnera. Przykład konwersacji z 2017 roku:

- Hello, my name is Andrew. What's your name?
Hello there. Your name is Andrew. My name is Millie.
- Why don't you tell me a little more about yourself ?
My name is Millie and I am a human volunteer for the Loebner Prize. I live in Leeds and am single. I am interested in AI but don't have a chatbot myself. What about you?
- I am a researcher in Artificial Intelligence at Goldsmiths University, do you know what that is? Do you enjoy your research?
I am a volunteer who has agreed to help with a Turing test.
- Will you tell me about your dreams?
I dream that I will pass my university exams.
- Can you rephrase that?
No, not really. What don't you understand?
- What is my occupation?
Well you said your job was RESEARCHER.

Przykład konwersacji: Mitsuku (2017)

- Can you tell me about a film you haven't seen? You have not seen. Are you asking about my have not seen?
- What do you think of Trump?
I read that over 50% were either concerned or scared about Donald Trump being President.
- if that's the case, why did they vote for him?
- Anything else?
That is all I have for right now.
- What is the answer to "Add 34957 to 70764"? 34957 to 70764
Ah yes, 34957 is a type of cheese that is made from sheep's milk. It's not very popular though.
- How do you recommend I make tea?
I suppose I would recommend something if I'd had a pleasant experience with it in the past.
- What do you get if you bake dough?
Is this a joke? I don't know Andrew. What do you get?

Konkurs Winograda

Kilka testów alternatywnych do testu Turinga ma swoje konkursy.

Test Turinga jest oceną zdolności do oszukiwania rozmówcy.

Naiwni rozmówcy dają się na to nabrać, test nie jest więc obiektywny, rezultaty zależą mocno od doboru sędziów.

Konkurs oparty na schemacie Winograda organizowany przez Commonsense Reasoning wymaga odpowiedzi na pytania świadczące o ich zrozumieniu.

Przykłady:

I. The trophy would not fit in the brown suitcase because it was too **big** (*small*).

What was too **big** (*small*)?

Answer 0: the trophy

Answer 1: the suitcase

II. The town councilors refused to give the demonstrators a permit because they **feared** (*advocated*) violence. Who **feared** (*advocated*) violence?

Answer 0: the town councilors

Answer 1: the demonstrators

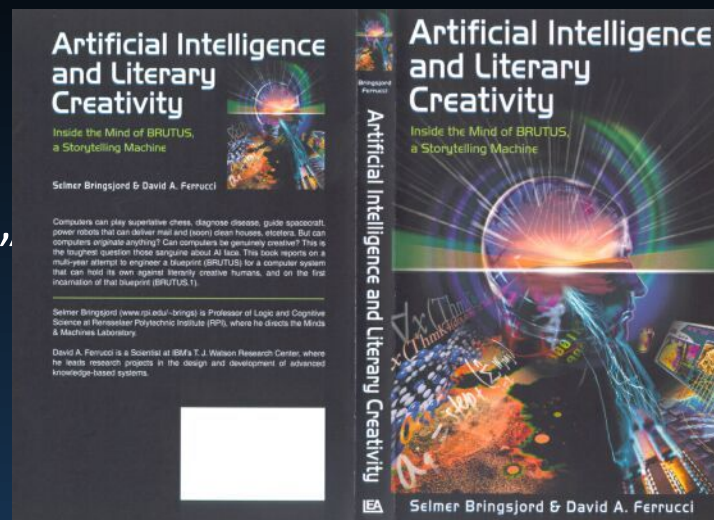
Generacja tekstu

Brutus - program tworzący opowiadania na zadany temat, np. zdanie:

„Jerzy obudził się pewnego ranka z koszmarным uczuciem, że stał się wielkim insektem”

Historie dziwaczne, a przez to interesujące, np. porównania:

„Oczy Harta były jak wielkie, krwawiące słońca.”



Użyto w nim „LAG: Literary augmented grammars”, co umożliwia formalizację takich pojęć jak „zdrada: nadaje się do scenariuszy TV!

Inne programy: sporo wyrafinowanych eksperymentalnych systemów, ale na razie widać je tylko w pracach naukowych.

Cyberpoeta Kurzweila

Todai Robot Project, National Institute of Informatics – osiąga ponadprzeciętne wyniki w egzaminach wstępnych na uniwersytety japońskie.

Muzyka generatywna, np. OpenAI Jukebox.

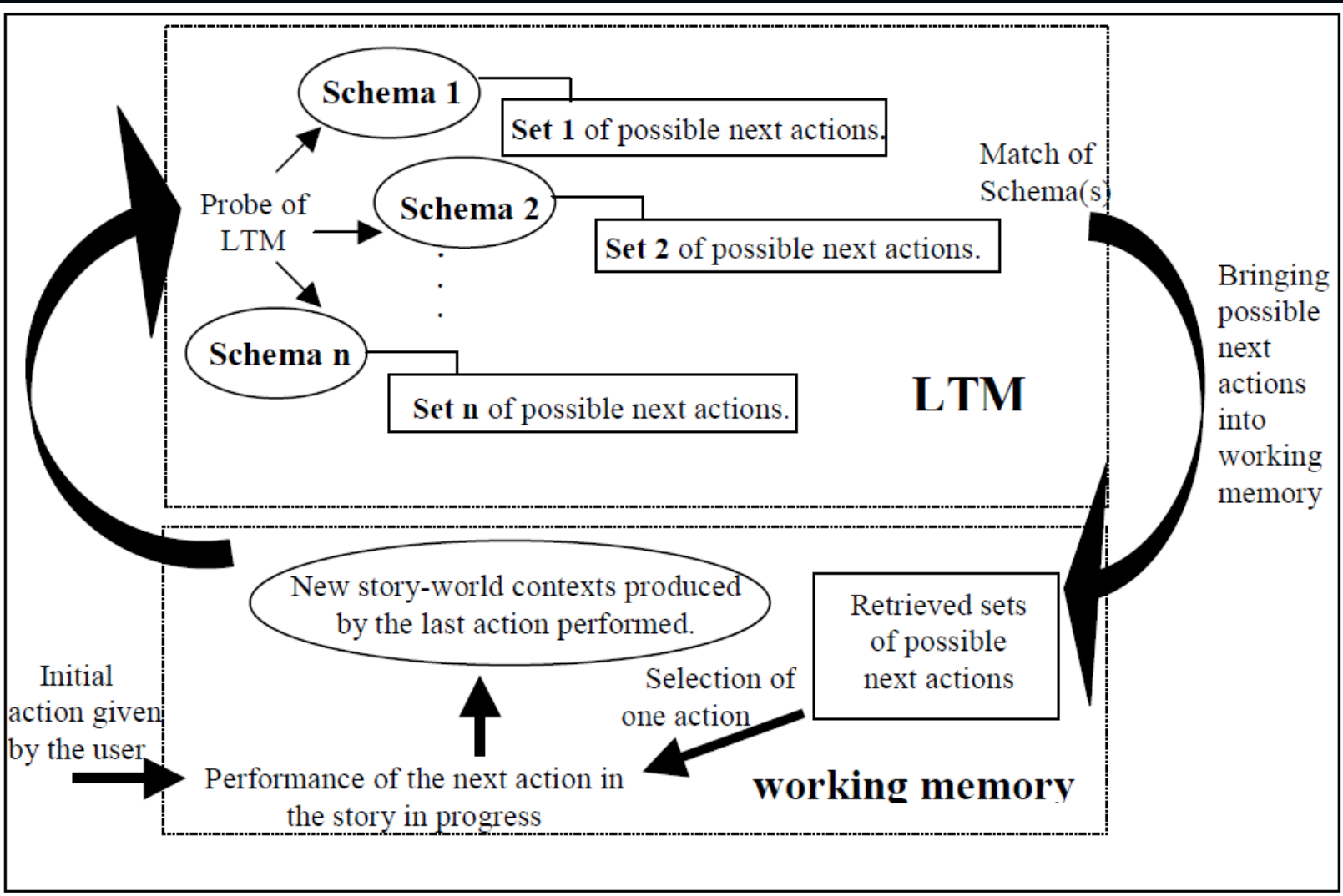
Mexica

Me
ger
Re
ak
Pro
dra
Og
kor

Przy

At t

In r



Bringing possible next actions into working memory

In that moment, the princess arrives to the market and sees the knight beating the farmer...

Robot Reporters

- “[Wordsmith](#) to platforma generacji tekstów (NLG) która na podstawie danych tworzy narracyjne historie”. Ten robot jest używany przez The Associated Press by zautomatyzować przygotowywanie raportów finansowych. Korzystając z tej technologii AP napisała 4,400 raportów w 2017 roku finansowych w ciągu kwartału.
- [Quill](#) używa dostarczonych danych, analizuje znane informacje, transformuje na inteligentne narracje dostosowane do indywidualnych wymogów.
- The Washington Post w 2017 roku [opublikował 850 artykułów](#) korzystając ze swojego systemu Heliograf. Ten system AI pisał krótkie reportaże i doniesienia z Olimpiady Rio, wyborów do kongresu i władz stanowych, zawodów sportowych, oraz tweety. Bloomberg News używa systemu Cyborg do 1/3 swoich artykułów.
- [News at 7](#) generuje całkiem zgrabne komentarze.
- [RadarAI](#), czyli Reporters And Data And Robots.
- [IBM Watson NLP, AI-Writer, 10 content automation tools](#).
- [Meet Bertie, Heliograf and Cyborg](#), The New Journalists On The Block.

BERT



Modele językowe mogą kodować wiedzę o relacjach słów w złożonych strukturach sieciowych. W 2018 roku grupa Google stworzyła BERT, model językowy wstępnie wyszkolony na dużym korpusie tekstowym, aby uzyskać ogólne „rozumienie języka”. Model ten jest następnie dostosowywany do określonych zadań NLP, takich jak odpowiadanie na pytania lub wyszukiwanie informacji semantycznych, co nie wymaga już treningu na wielką skalę.

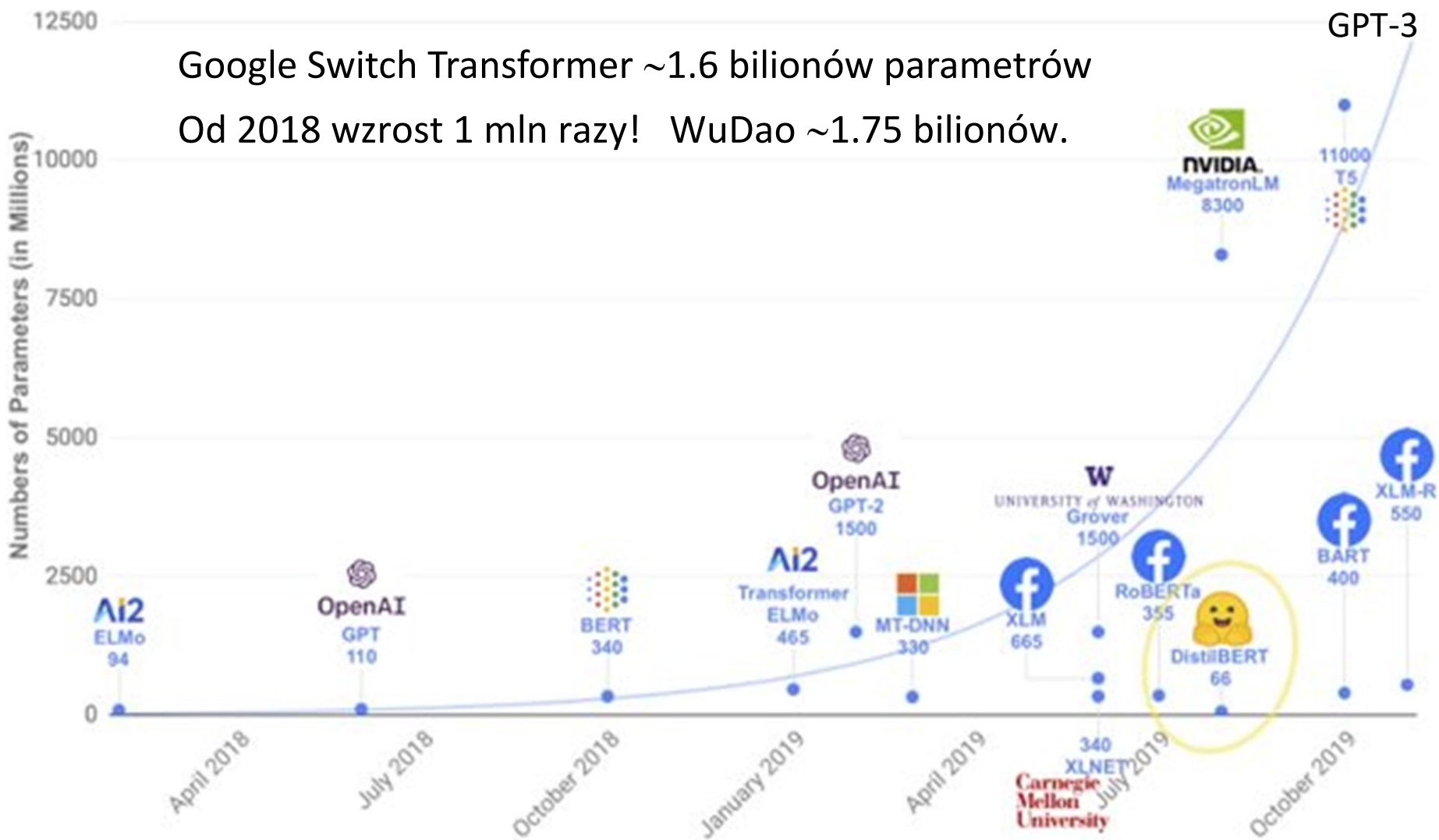
- **Bidirectional Encoder Representations from Transformers (BERT)**.
Transformer-based machine learning technique for (NLP) pre-training.
- Anglojęzyczny BERT: dwa modele, mniejszy 110M parametrów i większy model, 24-warstwowa architektura, 340M parametrów; trenowany na BooksCorpus z 800 milionami słów i Wikipedii z 2,500M słów.
- 12/2019 BERT działał w 70 językach, w 2020 roku w GitHub repository opublikowano wiele wstępnie wyszkolonych modeli.
- Maskowanie niektórych słów, które system uczy się przewidywać, np:
- **Input:** the man went to the [MASK1] . he bought a [MASK2] of milk.
Labels: [MASK1] = store; [MASK2] = gallon
- Super-human Q/A on Stanford Question Answering Dataset (SQuAD)

NLP supermodels

OpenAI GPT-3 model with 175 B parameters! See it in [action here](#).

Google Switch Transformer ~1.6 bilionów parametrów

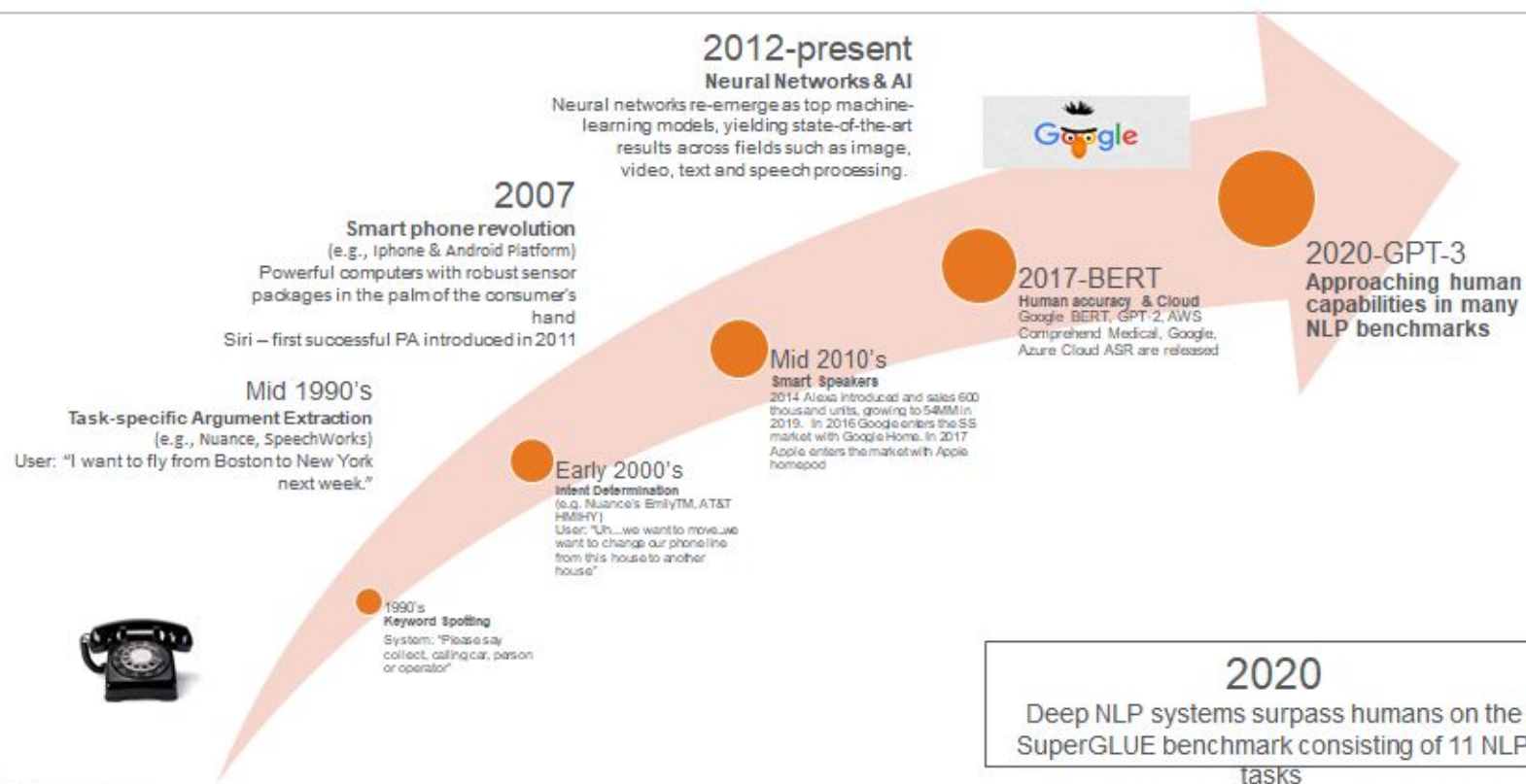
Od 2018 wzrost 1 mln razy! WuDao ~1.75 bilionów.



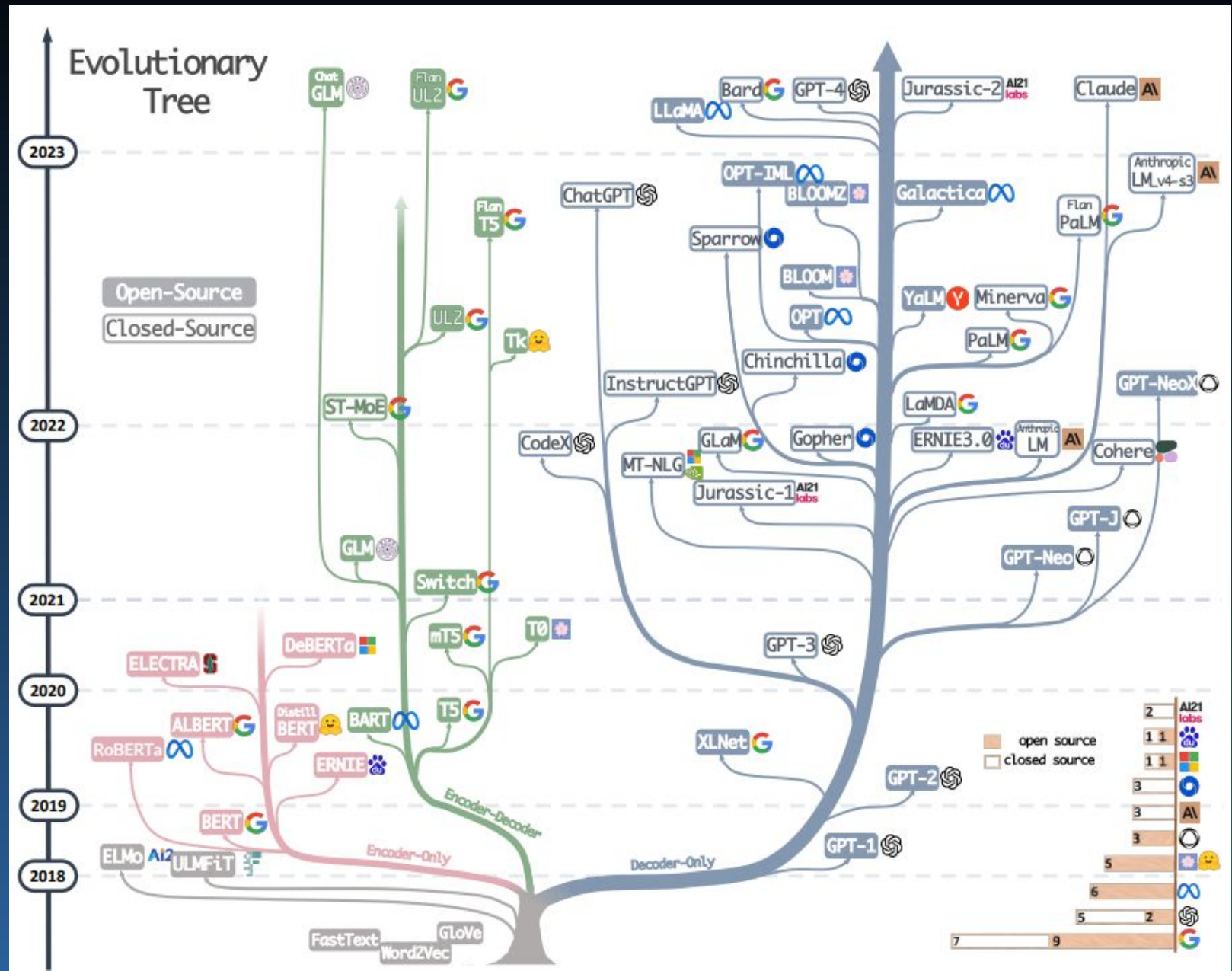
State of the art

- Super-human Q/A on **Stanford Question Answering Dataset (SQuAD)**

Speech & NLP Technologies are Evolving Quickly

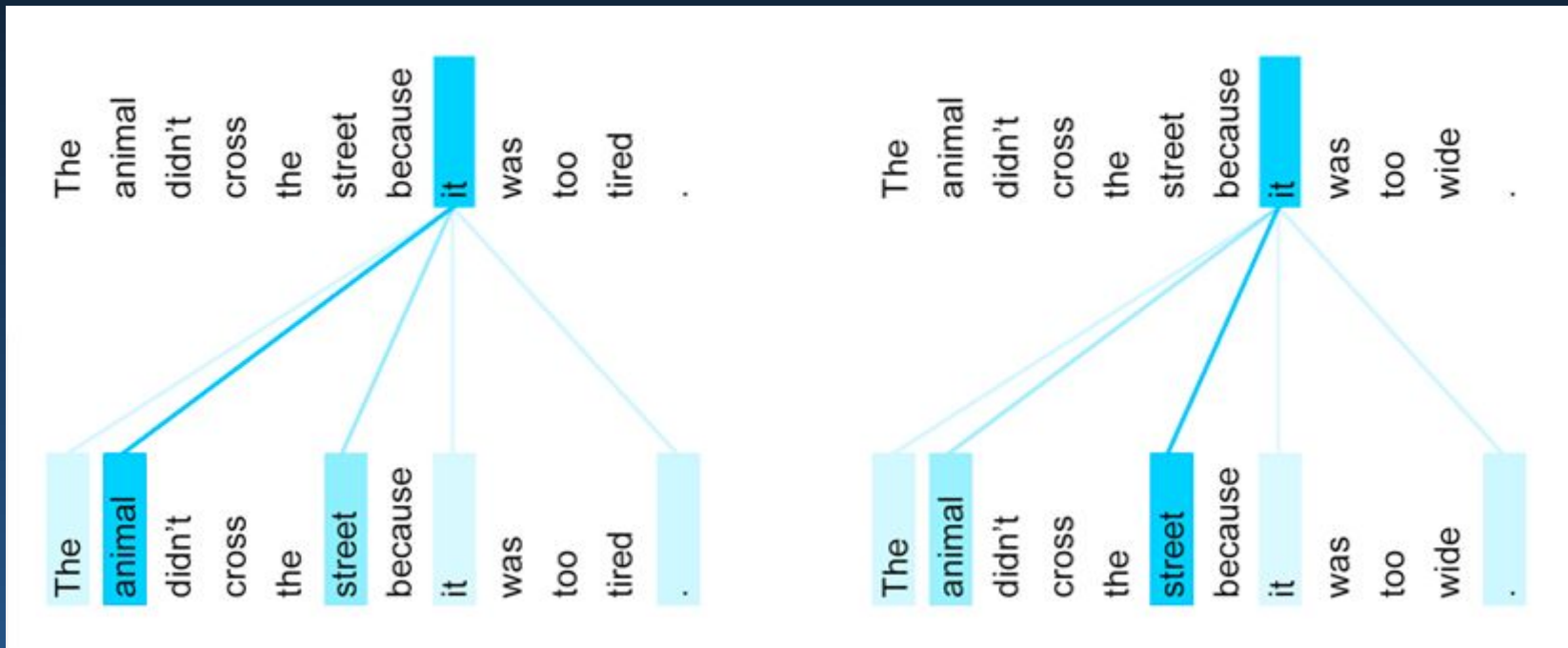


Rozwój modeli językowych



Switch Transformers

- Mixture of Experts, w wielkim modelu Google jest 1600 mld parametrów w 2048 połączonych modelach (“expertach”), które włączają się za pomocą mechanizmu “przełączania uwagi” (switch transformation). W zależności od domeny słowo “it” może mieć silniejsze połączenia z innymi w zdaniu, zmieniając interpretację.



GPT-3 words of wisdom

Every tweet is generated entirely by Artificial Intelligence (GPT-3)

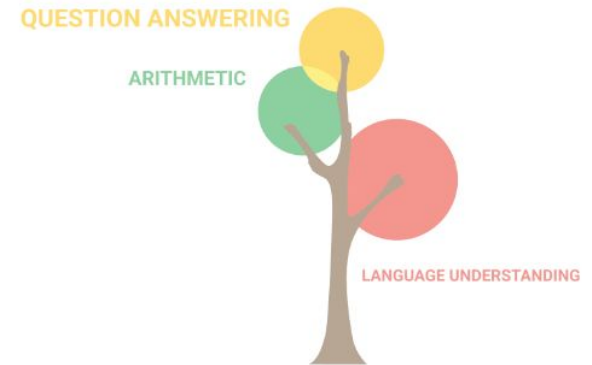
- Avoid hatred, because, as hatred increases, so too does the chance for unsustainable success.
- Science gives us enough. Faith gives us belief systems. Belief systems give us tribes.
- Are you playing life, or are you being played?
- The meaning of life is to search for what's meaningful.
- Today is the beginning of yesterday's tomorrow.
- Responsibility is for the sane. If you're crazy, you get freedom.
- By getting an idea you are copying thousands of years of human evolution. First try to learn from that.
- It doesn't matter what you're good at. The only thing really worth struggling for is what you want to be great at.
- If you're not scared about where you're headed, you're going the wrong direction.
- Rely on reality, not on artificial knowledge of reality.

Wielkie sieci

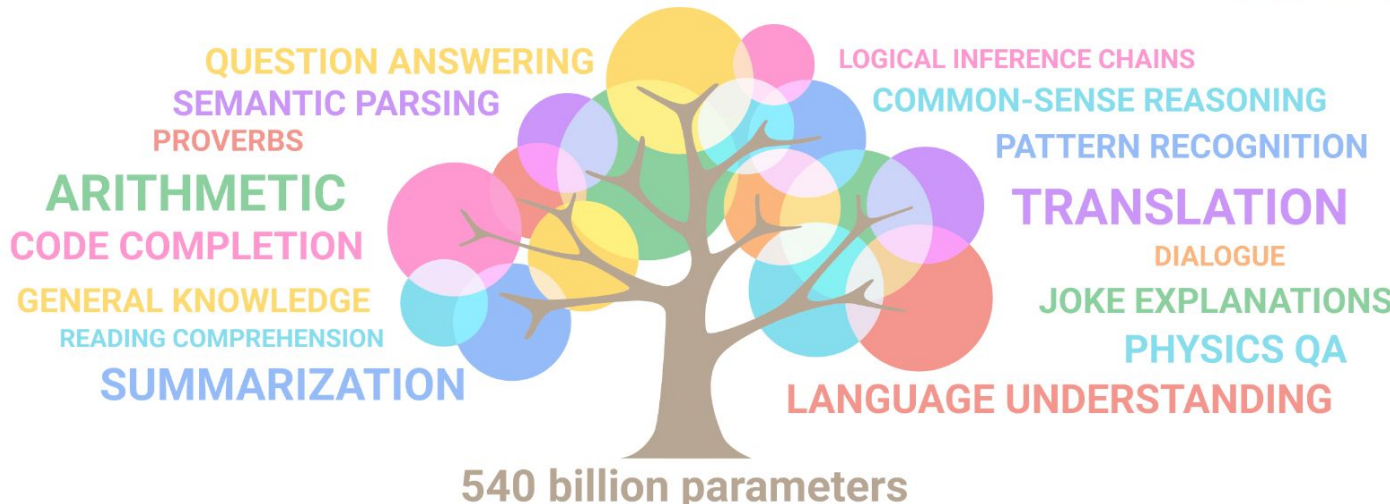
GPT-3 zaczęło trend tworzenia wielkich modeli, ale teraz jest ich wiele.

4/2022: Google AI Blog: Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance

Przykład z YouTube jak Palm wyjaśnia żarty
– komputery mogą rozumieć humor?



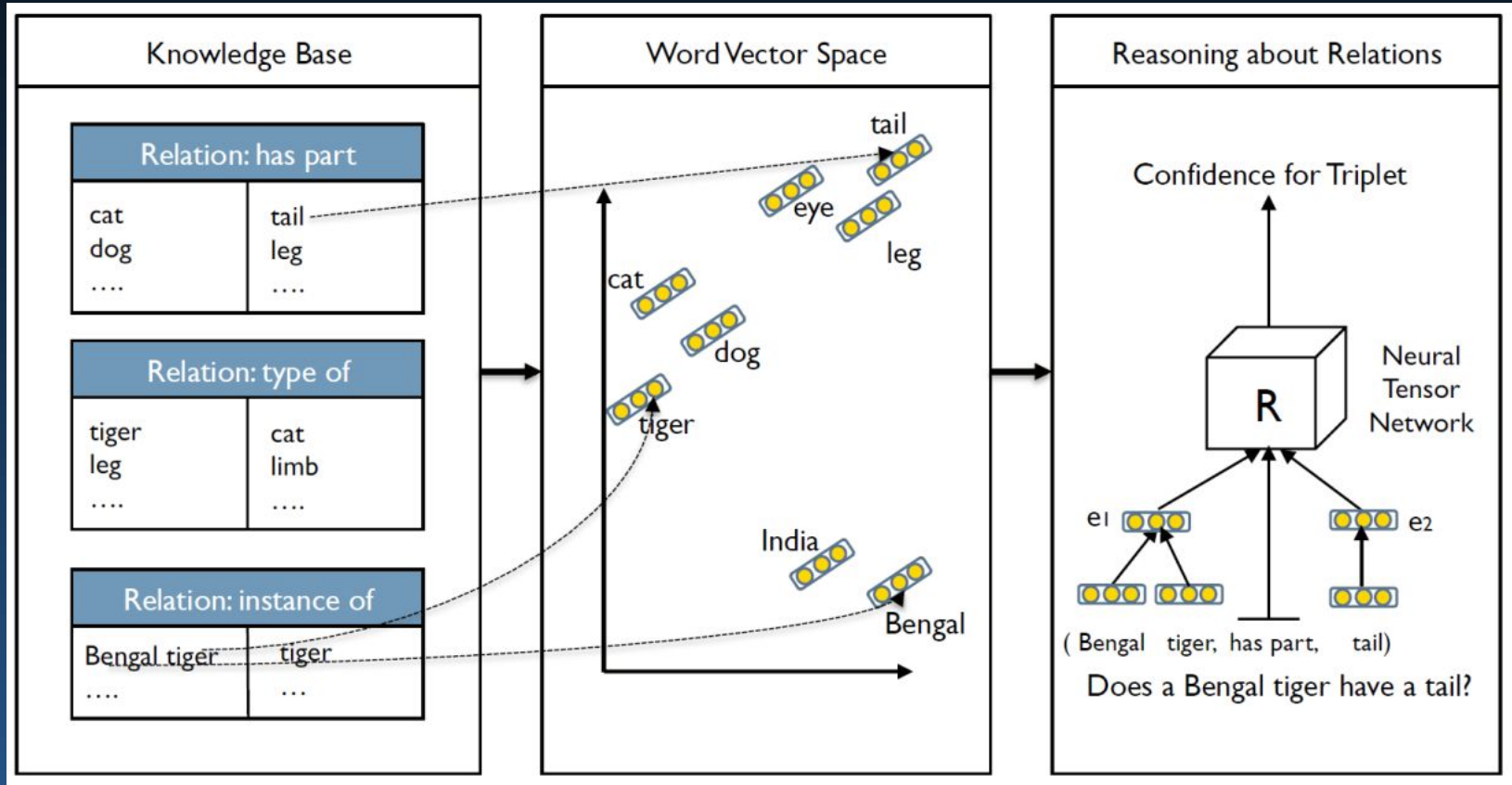
11 billion parameters



540 billion parameters

Reprezentacja wektorowa

Vector space. Generowana z kontekstu słów i zamiana na trójki (pojęcie, relacja, własność).



Film z cortical.io jak tworzą się „semantyczne odciski” (semantic fingerprints, semantic folding).

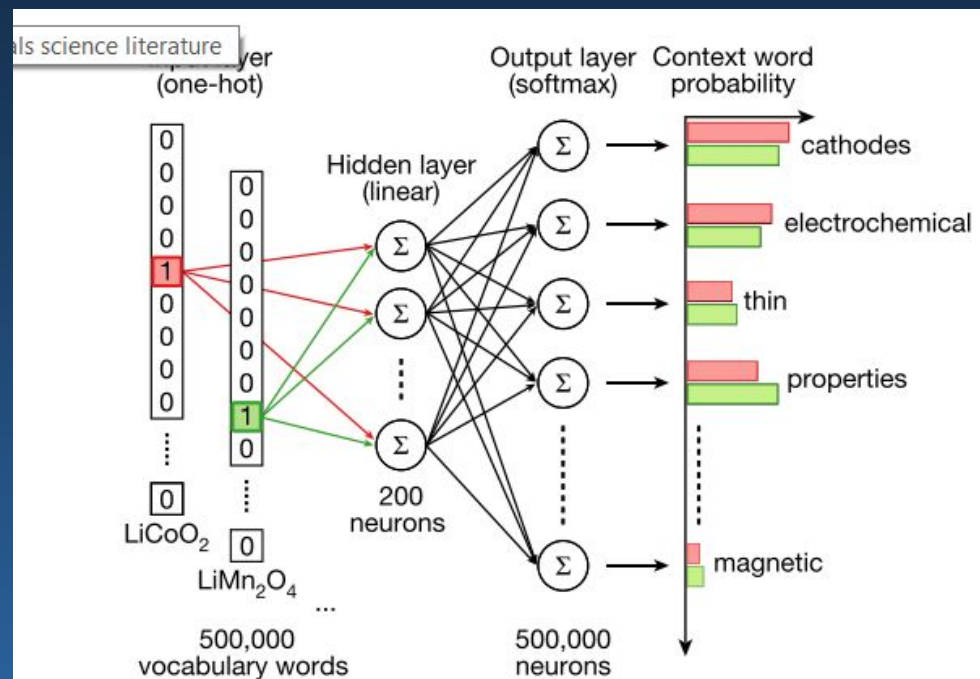
Serious application

Tshitoyan, V. ... Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. [Nature, 571\(7763\), 95.](#)

Materials science knowledge present in the published literature can be efficiently encoded as information-dense word embeddings without human supervision. Without any explicit insertion of chemical knowledge, these embeddings capture complex materials science concepts such as the underlying structure of the periodic table and structure–property relationships in materials.

Furthermore, we demonstrate that an unsupervised method can recommend materials for functional applications several years before their discovery.

[GPT Crush](#): see applications in business, design, education, philosophy, research, creative writing and many other areas.



Play with AI Dungeon

GPT-2 initially seen as too dangerous to release it to public (OpenAI).

AI Dungeon is a single-player and multiplayer **text adventure game** which uses artificial intelligence to generate unlimited content.

In June 2020 new "Dragon" version of the game based on GPT-3 was introduced. GPT-3 was trained with one trillion words, 570 gigabytes of text, at a cost of \$12 M, has 175 B parameters.

Premium-exclusive version is more advanced, but even free version allows AI Dungeon to maintain a much more coherent story. "It's just much more coherent in terms of understanding who the characters are, what they're saying, what's going on in the story and just being able to write an interesting and believable story."

Play it here: many worlds, ex. The World of Cthulhu - H.P. Lovecraft.

This world will be available for free during the month of November. It is running a custom model which was trained on the work of H.P. Lovecraft.

AI writer – talk to famous people like Newton, Leibniz, Feynman, H.G. Wells, I. Assiomov, Marie Curie or Ada Lovelace.

Szukanie semantyczne

Jak zrozumieć sens informacji, szukać dokładnie to, o co chodzi pytającemu?

W ramach projektów bibliotek cyfrowych (digital libraries) powstały projekty InterSpace, MedSpace, zastosowania „przestrzeni koncepcji” do określenia sensu pojęć:

Słowo $S \Rightarrow$ wektor kontekstowy $W(S)_i = p(S, S_i)$, w okienku kontekstowym zapisywane są korelacje z własnościami tego słowa.

Relacje semantyczne \Rightarrow relacje odległości między $W(S)$.

Słowa wieloznaczne będą miały kilka wektorów, np. $W_k(\text{Rakieta})$.

Podobne podejście: mapy informacji tekstowych, pokazują pojęcia w otoczeniu skojarzonych z nimi pojęć.

Przykłady: wizualizacja podobieństwa wektorów semantycznych za pomocą Samoorganizującej się Mapy Cech [WebSOM](#).

Klasteryzacja rezultatów z wyszukiwarek, grupuje różny sens pytań: tak działały np. projekty [Carrot2](#) i [DuckAndGo](#) i inne.

Dwa najważniejsze algorytmy reprezentacji wektorowej [word2vec](#) i [data2vec](#).

Watson gra w Jeopardy



Program IBM Watson (2007-11) w lutym 2011 roku zmierzył się z dwoma mistrzami teleturnieju Jeopardy (w Polsce znanym jako Va Banque), odpowiadając na pytania z różnych dziedzin.

Watson wygrał 77 141 \$, eksperci Ken 24 000\$, Brad 21 600 \$.

Reguły są niekorzystne dla maszyny, bo ludzie zgłaszają się słysząc pytanie natychmiast po jego zakończeniu, a maszyna ma opóźnienie zanim przeanalizuje pytanie i może się zgłosić.

Poprawność odpowiedzi sięgała 85-95%.

Watson (DeepQA) działa na superkomputerze IBM - Blue Gene/P, 15 TB RAM, 2,880 rdzeni procesorów, 80 Tflop.

Baza danych uwzględnia encyklopedie, słowniki, artykuły, bazy leksykalne, literaturę. Użyto ok. 100 algorytmów do analizy tekstu.

Watson zaliczył parę wpadek, np. Sport, also on your computer keys.

Przykład wektorów semantycznych

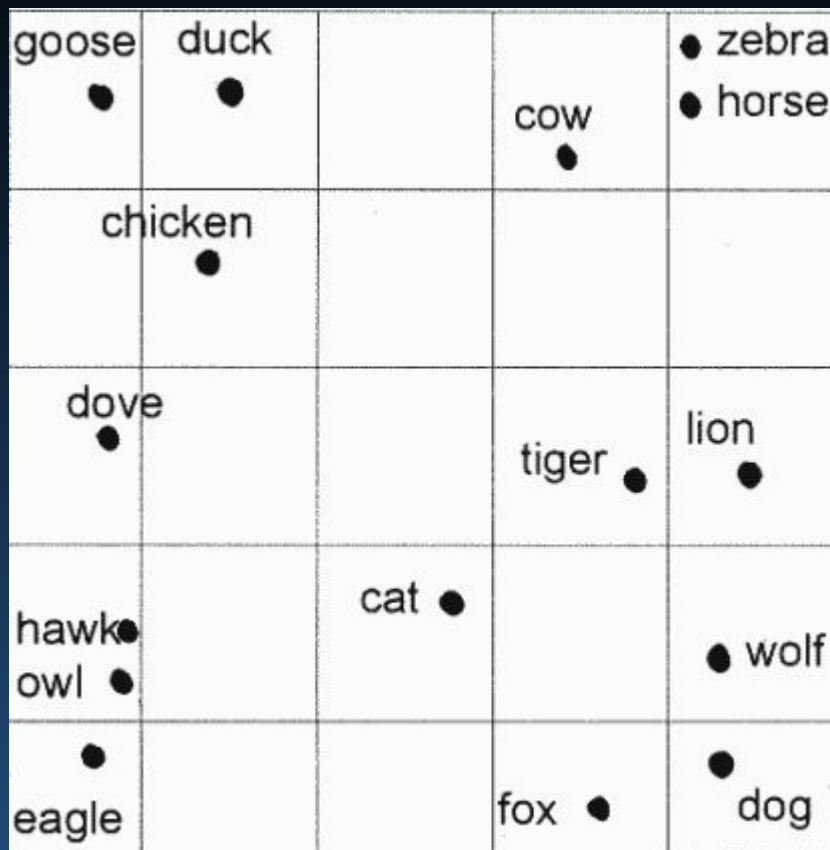
		d	o	g	h	e	w	t	h	z					
animal		o	h	o	o	a	f	d	o	c					
		v	e	s	w	w	l	o	l	a					
		e	n	k	e	k	e	x	g	f	t				
		r	n	e	a	w									
is	small	1	1	1	1	1	0	0	0	0	1	0	0	0	0
	medium	0	0	0	0	0	0	1	1	1	1	0	0	0	0
	big	0	0	0	0	0	0	0	0	0	0	1	1	1	1
has	2 legs	1	1	1	1	1	1	0	0	0	0	0	0	0	0
	4 legs	0	0	0	0	0	0	1	1	1	1	1	1	1	1
	hair	0	0	0	0	0	0	1	1	1	1	1	1	1	1
	hooves	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	mane	0	0	0	0	0	0	0	0	1	0	0	1	1	1
	feathers	1	1	1	1	1	1	1	0	0	0	0	0	0	0
likes	hunt	0	0	0	0	1	1	1	1	0	1	1	1	1	0
	run	0	0	0	0	0	0	0	1	1	0	1	1	1	1
	fly	1	0	0	1	1	1	1	0	0	0	0	0	0	0
	swim	0	0	1	1	0	0	0	0	0	0	0	0	0	0

Podobieństwo wektorów można ocenić za pomocą miary cosinusowej,
 $\cos(V_1, V_2) = \frac{V_1 \cdot V_2}{|V_1| |V_2|} = \text{il. składowy } V_1 \cdot V_2 \text{ dzielony przez il. norm } V_1 \cdot V_2$.

Odległość Hamminga = liczba różnych bitów pary wektorów, np.

$H(\text{dove}, \text{hen}) = |V(\text{dove}) - V(\text{hen})| = 1$, lub $\cos(V_1, V_2) = \frac{3}{\sqrt{4 * 3}} = \frac{\sqrt{3}}{2}$

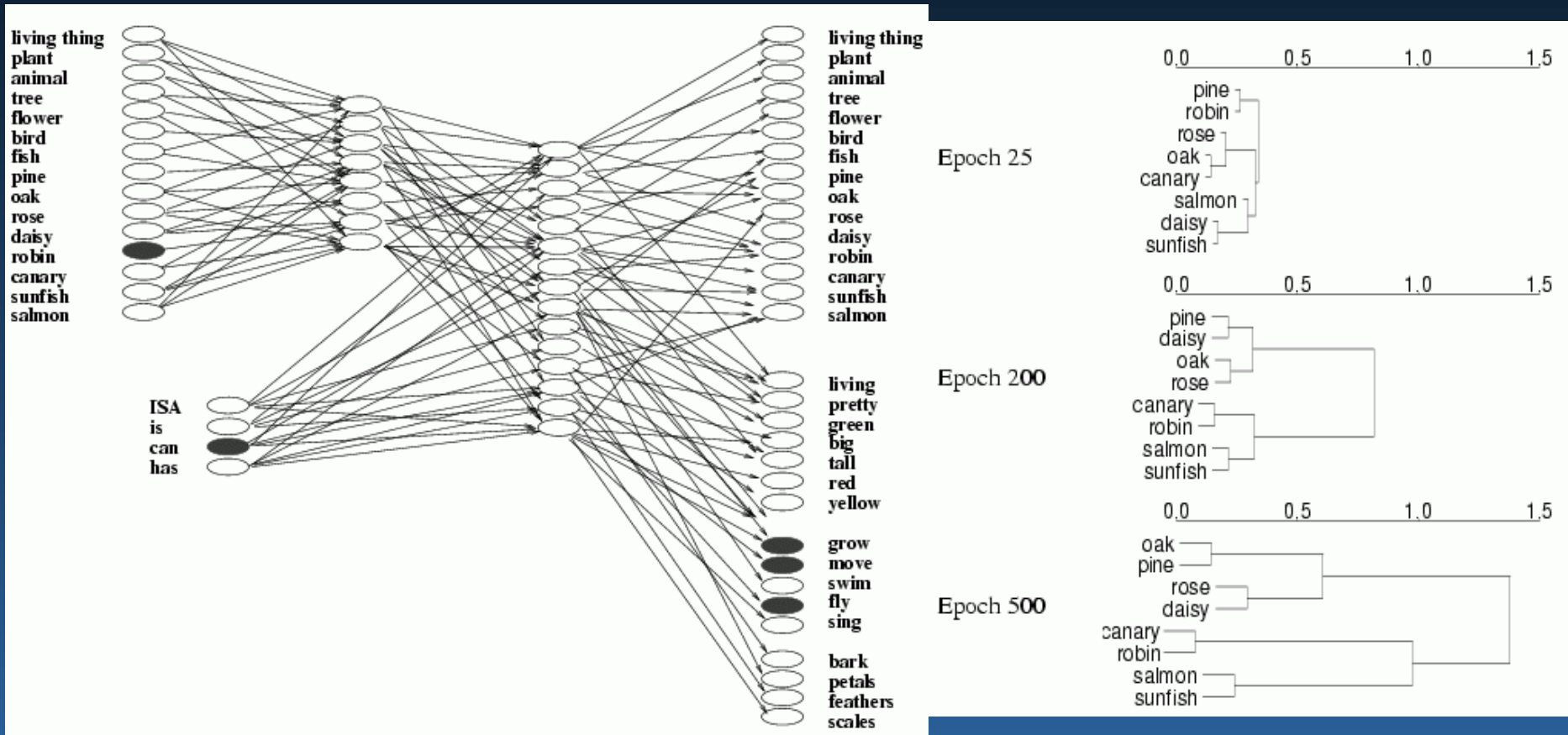
Mapy semantyczne: MDS



Jeśli przedstawić wektory jako punkty X_i na płaszczyźnie tak by ich odległości $\|X_1 - X_2\|$ były maksymalnie podobne do $\|W(X_1) - W(X_2)\|$ dostaniemy taką mapę. Widać naturalne grupowanie różnych gatunków.

Pamięć semantyczna

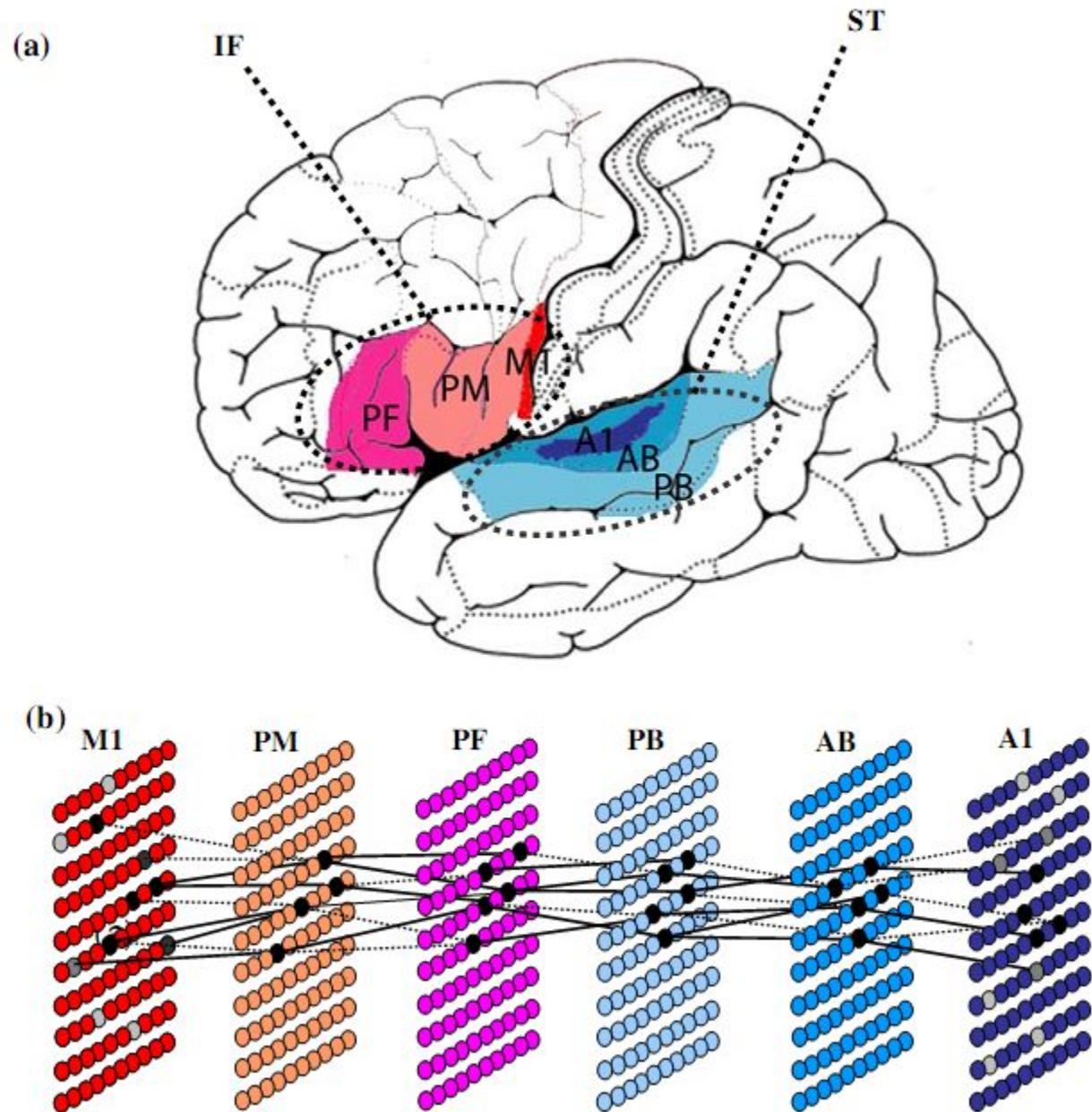
Wektory semantyczne powstają w sieciach neuronowych w automatyczny sposób. Sieć uczy się kojarzyć nazwy – np. nazwy zwierząt (robin) lub ich hipernimy (ptak) oraz własności (może) – z nazwami i cechami (rosnąć, ruszać się, latać). Dzięki zmianie siły połączeń jednostki ukryte uczą się przyjmować takie aktywacje, że wektory ich aktywacji dla podobnych pojęć będą podobne.



Pojęcia w mózgu

Garagnani et al.
Recruitment and consolidation of cell assemblies for words by way of Hebbian learning and competition in a multi-layer neural network. *Cognitive Comp.* 1, 160-176, 2009.

Pierwotna kora słuchowa (A1), pas słuchowy (AB), pas rozszerzony (PB, obszar Wernickiego), boczno-brzuszna kora przedczołowa (PF) i przedruchowa (PM, Broca), kora ruchowa (M1).



Reprezentacja wektorowa

Postało szereg technik przypisywania wektorów numerycznych słowom.

Word2Vec (Google, 2013) korzysta z dwóch technik:

Continuous Bag of Words (CBOW) przewiduje brakujące słowa na podstawie otoczenia.

Skip-gram przewiduje kontekst dla danego słowa.

Word2Vec i inne algorytmy tworzą reprezentacje o wymiarach 100-300, trenując płytkie sieci neuronowe na dużych korpusach tekstowych.

GloVe (Global Vectors for Word Representation), 2014,
[strona projektu \(Stanford\)](#)..

Macierze ko-okurencji obliczone na dużym korpusie pozwalają określić iloraz współwystępowania wyrazów, oraz wektory, których iloczyn skalarny jest równy logarytmowi tego ilorazu. Dzięki temu odległość wektorów staje się miarą ilorazu prawdopodobieństw, odzwierciedlającego semantykę słów. Dostępne są wektory wyliczone na Wikipedii o wymiarach 25-300.

FastText (Facebook, 2016) rozszerza Word2Vec dla słów traktowanych jako ciąg n-gramów, dodając informację morfologii wyrazów. To ułatwia analizę neologizmów (out-of-vocabulary, OOV).

[Towards Data Science](#) przedstawia kilka przykładów jak się to robi.

[Intuitive Guide to Understanding Word2vec](#)

[The Illustrated Word2vec](#)

Nowe podejścia

NLP to bardzo aktywna ale i trudna dziedzina.

Sporo linków: <http://www.is.umk.pl/~duch/IR.html>

Z. Vetulani, Komunikacja człowieka z maszyną, AOW EXIT 2004

Podstawowe narzędzia NLP:

- duże korpusy do trenowania i testowania programów NLP.
- Programy do normalizacji tekstu szukają form podstawowych.
- Analiza morfologiczna rozbija wyrazy na morfemy (rdzenie, przed/po).
- Taggery przypisują części mowy (POS, Part of Speech), formy gramatyczne, użyteczne przy odróżnianiu czy mamy do czynienia z nazwą rzeczą, przymiotnikiem czy czasownikiem (shallow parsing).

Gramatyki probabilistyczne, rozkład oparty o statystykę danych, gramatyki połączeń, kodujące sposoby używania (relacje) słów:

<http://www.link.cs.cmu.edu/link/submit-sentence-4.html>

Distributed semantics, czyli modele wektorowe uwzględniające składnię.

Systemy hybrydowe

DISCERN - system NLP z neuronowym leksykonem.

Problem: jak automatycznie uczyć się skojarzeń? Jak reprezentowane są symbole w naszych mózgach?

Za pomocą połączonych grup neuronów tworzących podsięci reprezentujące brzmienie i znaczenie, łączące percepcję i działanie.

Podejście hybrydowe: symbole do analizy gramatycznej, neurony do rozproszonej reprezentacji informacji.

Słyszemy wibracje, widzimy kreski, to kojarzymy z koncepcją.

DISCERN używa **kilku map** automatycznie tworząc skojarzenia na poziomie fonologicznym, ortograficznym i semantycznym dzięki rozproszonej reprezentacji wiedzy.

W DISCERN zastosowano hierarchiczne sieci SOM, ale można by też stosować reprezentację wektorową.

Analiza gramatyczna – jeszcze zbyt trudna dla modeli neuronowych?
Niestety projekt przestał się rozwijać.

Czego brakowało w NLP?

Porównajmy NLP z ludzkimi kompetencjami w tej dziedzinie.

Człowiek ma pamięć rozpoznawczą – koryguje słowa nawet jeśli są napisane z poważnymi błędami, odwołując się do zrozumienia pojęć w tekście i ogólnego sensu tekstu.

Człowiek ma pamięć semantyczną – wiemy, że krowa ma ogon, rogi, daje mleko, jest duża, muczy ... nie da się tego znaleźć w słownikach!

Definicja krowy (Wordnet): dojrzała samica ssaka, którego samiec nazywa się byk. Definicja byka: odwrotnie ...

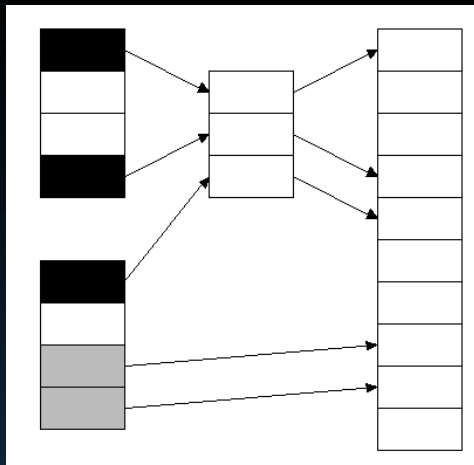
Brakowało wiedzy i wyobraźni! Dopiero sieci trenowane na wielkich korpusach, oraz technika GAN tworząca wyobrażenia, sobie z tym poradziły.

Projekty stworzenia słowników z licznymi relacjami, np. [Wordnet](#), zmierzają w kierunku pamięci semantycznej, ale są na razie ubogie.

Próby automatycznej akwizycji wiedzy są bardzo trudne.

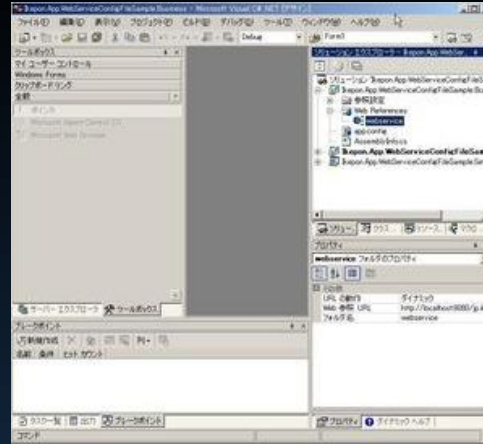
[Lista chatbotów na świecie.](#)

Ciekawe zastosowania: empatyczny bot **Chip Vivant**.



Pamięć semantyczna

Pytanie

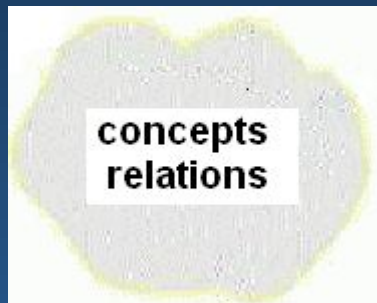


Zastosowania, np.
gra w 20 pytań

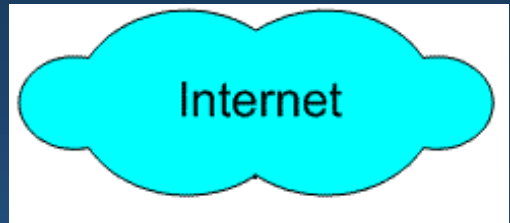


Mówiąca głowa

Zapamiętywanie



Oznaczenie części mowy
i ekstrakcja fraz



Słowniki, encyklopedie

weryfikacja

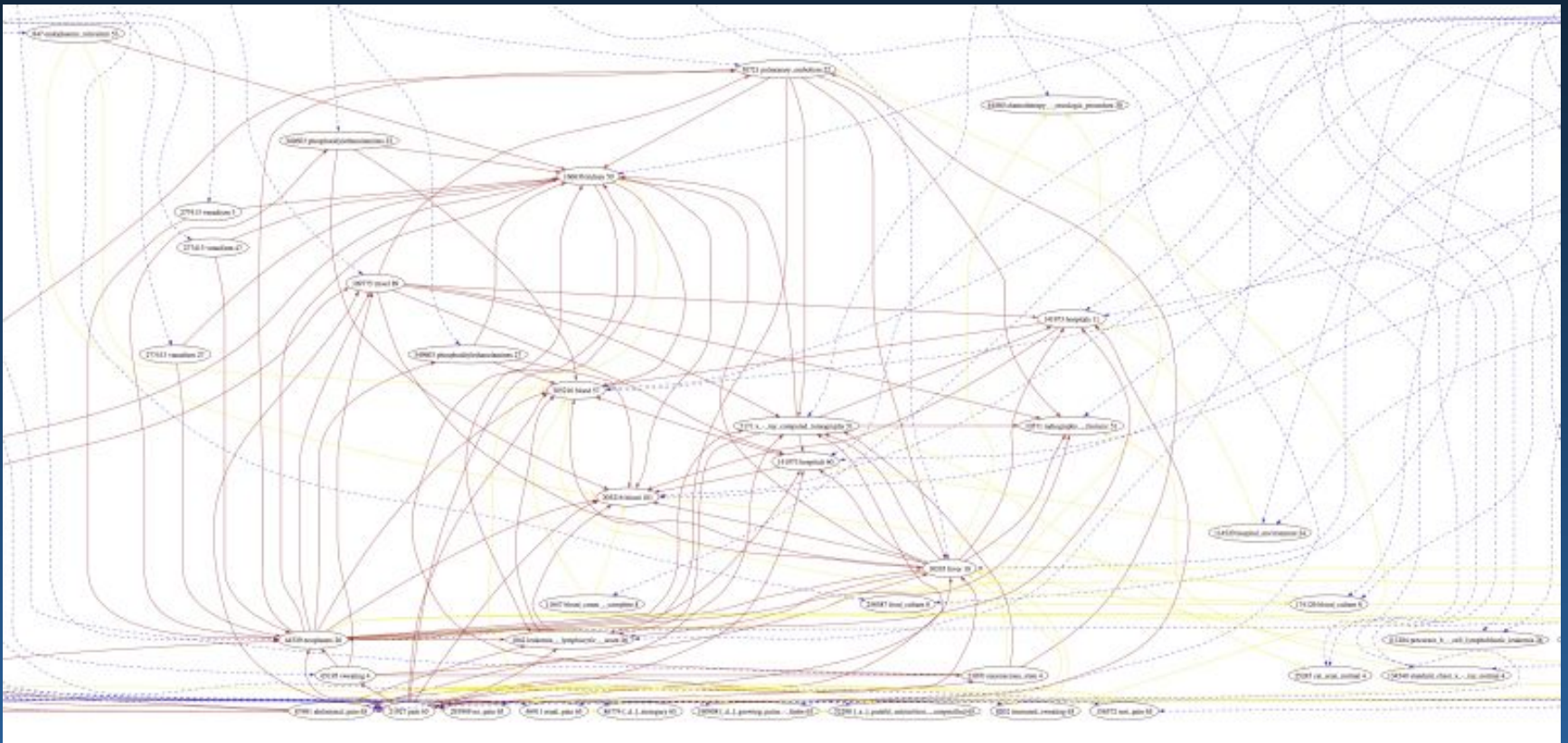
Parser

poprawki

Rozumienie tekstów

Próbujemy rozwinąć neurokognitywne podejście do rozumienia języka w oparciu o koncepcję grafów spójnych koncepcji, aktywnej części pamięci semantycznej z hamowaniem i rozchodzeniem się aktywacji.

Dla tekstów medycznych mamy >2 mln koncepcji, 15 mln relacji ...



Generator zagadek

Mając dobrą pamięć semantyczną i uproszczoną reprezentację wiedzy można tworzyć dowolną liczbę zagadek!

Wystarczy znaleźć kilka cech, które unikalnie charakteryzują daną koncepcję. W tym celu z pamięci semantycznej wystarczy zostawić tylko jedną relację: tak/nie.

Przykład automatycznie wygenerowanych zagadek:

Jest pomarańczowy, ma czarne plamki, jest płazem. Co to za zwierzę?

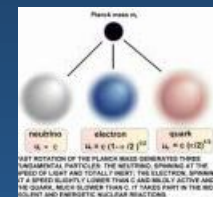
Salamandra.

Ma powab, spin i ładunek.

Co to jest?

Zapytajcie Google!

Strona o kwarkach będzie na początku ...



Kreatywność słowotwórcza

Program Quackle zwyciężył mistrza skrable Davida Boysa 482-465 w 2006 roku w turnieju Scrabble Open w Toronto – ale to tylko kombinacja liter i duży słownik, niemniej trudny problem kombinatoryczny.

Tworzenie nowych słów to ciekawy test na [kreatywność komputerów](#).

[Moje referaty](#).

Kreatywność muzyczna – sporo ciekawych projektów „[generative music](#)” lub [muzyki algorytmicznej](#), wykorzystujących [algorytmy ewolucyjne](#), do tworzenia licznych wersji coraz ciekawszych kompozycji – człowiek tylko dokonuje wyboru tych, które mu się podobają.

Tu mamy grupę robotów [Compressorhead](#) grającą z wyczuciem ostrego rocka.

Podsumowanie tekstu

Wybranie najważniejszych fragmentów tekstów („executive summary”) jest powszechnie stosowane, np. w przypadku serwisów zbierających informacje w Internecie na zadany temat (newsreaders).

Dobre wyniki osiągają modele oparte o sieci neuronowe z rekurencją.

Modele dokonujące automatycznego podsumowania pracują na dwa sposoby: ekstrakcję fraz i abstrakcję. Modele oparte na ekstrakcji wybierają frazy i łączą ze sobą by utworzyć podsumowanie. Używanie fragmentów naturalnego tekstu ułatwia tworzenie zrozumiałych fraz, ale nie pozwala na wprowadzanie nowych pojęć ani parafrazy.

Modele oparte na abstrakcji treści mogą tworzyć znacznie bardziej oryginalne podsumowania, z nowymi słowami, ale wymagają modelu, scenariusza pozwalającego na tworzenie odpowiedniej narracji, a to jest znacznie bardziej trudne zadanie.

Ambitne projekty

- NELL: Never-Ending Language Learning (Carnegi Mellon University)
<http://rtw.ml.cmu.edu/rtw/>

"Read the Web" to projekt, który ma się uczyć nowych pojęć i ich definicji czytając informację z Internetu. Projekt rozwija się od 2010 roku, NELL pracuje w sposób ciągły analizując teksty dostępne w sieci i zdobywając wiedzę o różnych pojęciach, np. `playsInstrument(George_Harrison, guitar)`. W ten sposób zgromadził ponad 50 milionów prawdopodobnych faktów, oceniając stopień wiarygodności ponad 2 mln faktów jako wysoki.

NELL został zainicjowany korzystając z ontologii zawierającej setki pojęć (np. `person`, `sportsTeam`, `emotion`) i relacji pomiędzy tymi pojęciami, np. `playsInstrument(musician, instrument)`, oraz 10-15 przykładów w każdej kategorii wraz z relacjami.

NELL cały czas się uczy a baza wiedzy w ten sposób utworzonej jest dostępna ([knowledge base](#)), podobnie jak opis techniczny systemu ([technical approach](#)), or join the [discussion group](#). Po 6 miesiącach uczenia bez nadzoru poziom Trochę głupot: [radio_station_in_miami](#) is an [item often found in a bedroom](#).

Netflix 1M\$ Prize

Netflix Prize, an award of \$1 million to the first person or team who can achieve certain accuracy goals when recommending movies based on personal preferences – announced in Oct 2006.

The company made 100 million anonymous movie ratings available to contestants for learning.

Szczegóły na temat nagrody Netflix Prize:

<http://www.netflixprize.com>

W tym przypadku potrzebna jest analiza języka i metody uczenia maszynowego by przewidywać, jakie filmy będą interesujące.

Nagrodę wygrała grupa z AT&T w 2009 roku, ale parę innych były blisko.

Systemy Q/A

Jest wiele systemów, które mają odpowiadać na pytania na (prawie dowolny) temat. Pojawiają się w postaci asystentów głosowych w telefonach lub takich urządzeniach jak Amazon Alexa.

[PoEval 2021 - PDF](#)

[Grupa Technologii Językowych G4.19 Politechniki Wrocławskiej](#)

Question answering systems:

[Wiki EN](#)

Wolfram Alpha

[DrQA](#), PyTorch system for reading comprehension on GitHub (Meta)

Language Computer, LCC, top 2004 Q/A system

MIT Start project

My[Q]Box, providing question answering software for Web pages.

Watson Natural language understanding demo |

Systemy Q/A

Jest wiele systemów, które mają odpowiadać na pytania na (prawie dowolny) temat. Pojawiają się w postaci asystentów głosowych w telefonach lub takich urządzeniach jak Amazon Alexa.

Polskie próby, na razie eksperymentalne: [PolEval 2021](#) (PDF)

[Grupa Technologii Językowych G4.19](#) Politechniki Wrocławskiej

Question answering systems: lista jest w zbiorze moich linków:

od Wolfram Alpha do Watsona

Również lista narzędzi NLP: Taggers, Parsers, NER, NP chunking, Language models ...

Ostatnie przykłady awatarów używających dużych modeli językowych są imponujące, przykłady są na moim Flipboard.

Flamingo AI, Deep Mind, odpowiada na pytania

Open AI Dall-e i przykłady syntetycznych obrazów z opisów tekstowych.

Cognitive Computing

Watson i plany [IBM w zakresie Cognitive Computing](#) – analiza publikacji medycznych i pomoc w szukaniu semantycznym.

Watson i [kreatywne gotowanie](#). “IBM is already working with customers on applying computational creativity technology to their business.”

Mahmoud Naghshineh, Vice President, Services research, IBM Research

[Microsoft Cognitive Services](#)

MS API: [Language Understanding Intelligent Service](#) | [Bing Spell Check API](#)

[Web Language Model API](#) | [Text Analytics API](#) | [Translator Text API](#) | [Linguistic Analysis API](#)

[Amazon Alexa](#), Amazon Echo działa jak chatterboty sterowane głosem.

[Deep Learning in NLP](#), tutorial R. Socher,
[Stanford MetaMind](#), [Salesforce Research](#).

Linguboty, wirtualni asystenci

Technologia: większość botów opiera się na rozszerzeniu techniki szablonów, np. w oparciu o język AIML, często używanego przez chatterboty. Nowsze wykorzystują modele neuronowe.

- Amazon: [Evi](#) i [Alexa](#) od 7/12 po polsku!
- Microsoft [Cortana](#) i Apple [Siri](#)
- Google Now i [Google Assistant](#) i [Google Meena](#)
- Samsung [Viv](#) i [Bixby](#)
- [IBM Watson assistant](#)
- [Facebook Blender](#), bot z analizą emocji, na Github.
- [Medical prototypes](#), DARPA's Detection and Computational Analysis of Psychological Signals (DCAPS) project
- [Leena AI](#) is an AI-powered HR Assistant
- [Mycroft](#) open source voice assistant, AI for everyone.
- [Open AI](#) voice assistant list
- [Lista polskich botów](#)



Chatboty

Liczne programy typu chat

Futurepedia:

<https://www.futurepedia.io/ai-tools>

Science:

<https://www.futurepedia.io/?searchTerm=science>

Hugging Face

<https://huggingface.co/>

Przykładowe pytania

- Jakie są cele NLP?
- Co potrzebne jest do rozumienia tekstu?
- Co opisuje syntaktyka, semantyka i do czego służy rozbiór gramatyczny (parsowanie)?
- Podaj przykład prostej gramatyki używając dwóch czasowników i dwóch rzeczowników i wypisz możliwe zdania.
- Jakie typy gramatyk i języków wyróżnił Chomsky?
- Co próbuje się osiągnąć tworząc nietypowe rodzaje gramatyk?
- 5 kroków analizy języka naturalnego.
- Na jakiej zasadzie działają chatterboty ?
- Jak trenuje się model BERT i Google Switch Transformers?
- Podstawowe idee przydające się do tłumaczenia maszynowego.
- Jakie są możliwości i ograniczenia dużych modeli językowych (LLM).
- Jak utworzyć wektory semantyczne i do czego się przydają?
- Dlaczego mamy tak ogromny postęp w NLP w ostatnich latach?