Zestaw algorytmów zrównoleglania i przechowywania wielojęzycznych zanurzeń słów na potrzeby obliczania prawdopodobieństwa tłumaczenia

dr Rafał Jaworski
Zakład Sztucznej Inteligencji
Uniwersytet im. Adama Mickiewicza w Poznaniu

Seminarium, IPI PAN, Warszawa

Zrównoleglanie na poziomie słów

- W swoich badaniach koncentruję się algorytmach przetwarzania języka naturalnego
- Ich implementacje znajdują zastosowanie w systemach wspomagania tłumaczenia
- Ponadto są wykorzystywane w badaniach lingwistycznych
- Przedstawię moje główne osiągnięcie badawcze

Zrównoleglanie na poziomie słów

		ils	connaissent	très	bien	le	problème de surproduction				
								problème	de	surproduction	.
	They	0.88 ✚	0.48	0.46 ⇳	0.56 ⇳	0.2	0.22 ↔	0.3	0.15	0.16	
know about		0.37 ↕	0.48 ✢	0.25 ↕	0.43 ↕	0.19 ↕	0.27 ↔	0.33	0.16 ↕	0.08	
know about	know	0.37	0.68 ✚	0.31	0.49	0.19	0.27 ↔	0.33	0.16	0.08	
know about	about	0.24	0.36	0.23	0.39	0.16	0.19 ↔	0.25	0.41 ⇔	0.04	
the overproduction		0.12	0.16	0.15	0.16	0.14	0.6 ↔	0.31	0.11	0.66 ✢	
the overproduction	the	0.23	0.35	0.23	0.34	0.75 ✚	0.36 ↔	0.35	0.44 ⇳	0.21	
overproduction problem		0.18	0.29	0.22	0.24	0.14	0.8 ✚	0.68 ↕	0.15	0.58	
overproduction problem	overproduction	0.17	0.15	0.13	0.13	0.11	0.63 ↔	0.54	0.14	1.77 ✚	
overproduction problem	problem	0.23	0.29	0.25	0.28	0.1	0.64 ↔	1.94 ✚	0.14	0.56	
	.										1.46 ✚

Zrównoleglanie na poziomie słów

Potencjalne zastosowania we wspomaganiu tłumaczenia:

- Automatyczne pozycjonowanie elementów nietłumaczonych
- Znajdowanie błędów w tłumaczeniu (słów nieprzetłumaczonych)
- Wyszukiwanie tłumaczenia wybranej frazy

Zrównoleglanie na poziomie słów

Istniejące rozwiązania:

- Giza++
- fast_align (tylko model I i II IBM)
- rozwiązania oparte na sieciach neuronowych:
  a) Simalign
  b) Joël Legrand, Michael Auli, Ronan Collobert: *Neural Network-based Word Alignment through Score Aggregation*

Inter-language Vector Space

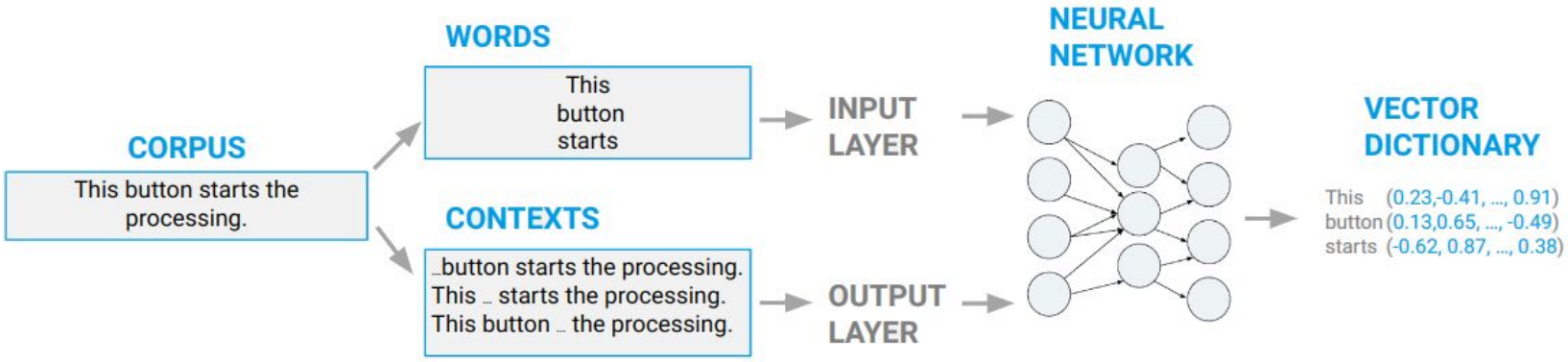
- Inter-language Vector Space to zestaw metod do generowania, zrównoleglania, przechowywania i odczytywania zanurzeń słów w wielu językach.
- Badania te przełożyły się na wdrożenia w oprogramowaniu XTM Cloud firmy XTM International.

Inter-language Vector Space - word2vec

- Pierwszą funkcją zestawu jest generowanie reprezentacji wektorowych słów metodą word2vec. Są one generowane przy użyciu sieci neuronowej typu auto-encoder na podstawie danych tekstowych - korpusu tekstu.

- Korpus dzielony jest na poszczególne słowa oraz konteksty, w których te słowa występują. Słowa są podawane warstwie wejściowej sieci neuronowej, natomiast konteksty są oczekiwanym wyjściem.

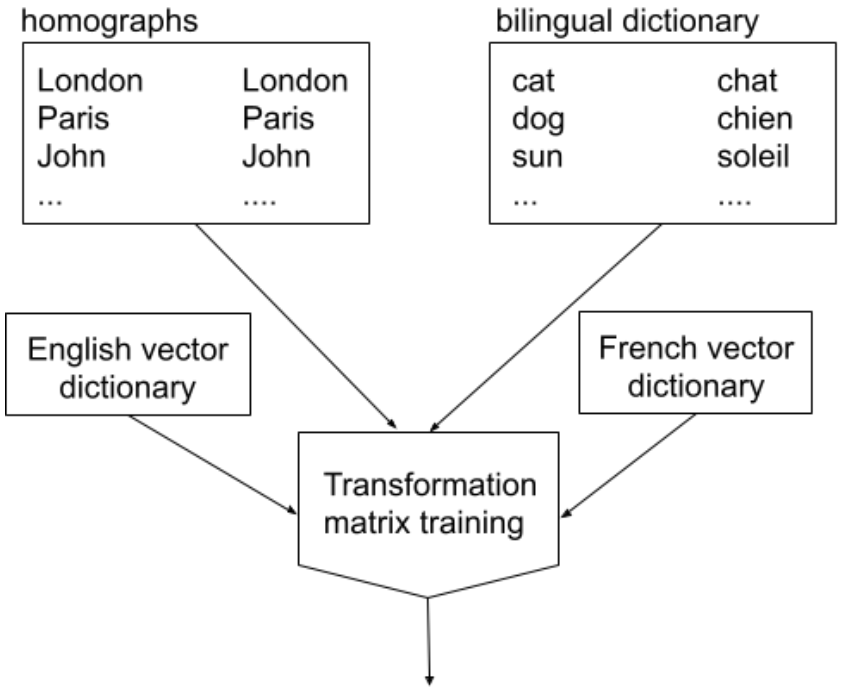
Inter-language Vector Space - word2vec



Inter-language Vector Space

- Zanurzenia trenowane są osobno dla wielu języków.
- Zanurzenia słów nie są porównywalne pomiędzy językami.
- Aby rozwiązać ten problem, dokonujemy zrównoleglenia przestrzeni wektorowych.

Zrównoleglanie na poziomie słów



Inter-language Vector Space

Trenowanie macierzy transformacji na podstawie:

Smith, S.L., Turban, D.H., Hamblin, S., & Hammerla, N.Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. ArXiv, abs/1702.03859.

Inter-language Vector Space

Rozwiązanie podstawowe (*Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation*) polega na użyciu słownika dwujęzycznego i słownika homografów do wytrenowania macierzy transformacji metodą stochastycznego spadku wzdłuż gradientu:

$$\min_W \sum_{i=1}^{n} ||y_i - W x_i||^2$$

Inter-language Vector Space

Zastosowane rozwiązanie bazuje na idei wytrenowania macierzy transformacji *O* dokonującej przekształcenia ortogonalnego dwóch przestrzeni wektorowych. Jeśli słownik dwujęzyczny ma postać par wektorów:

$$\{y_i, x_i\}_{i=1}^{n}$$

wówczas problem znalezienia optymalnego przekształcenia sprowadza się do maksymalizacji podobieństwa cosinusowego na odpowiadających sobie wyrazach ze słownika:

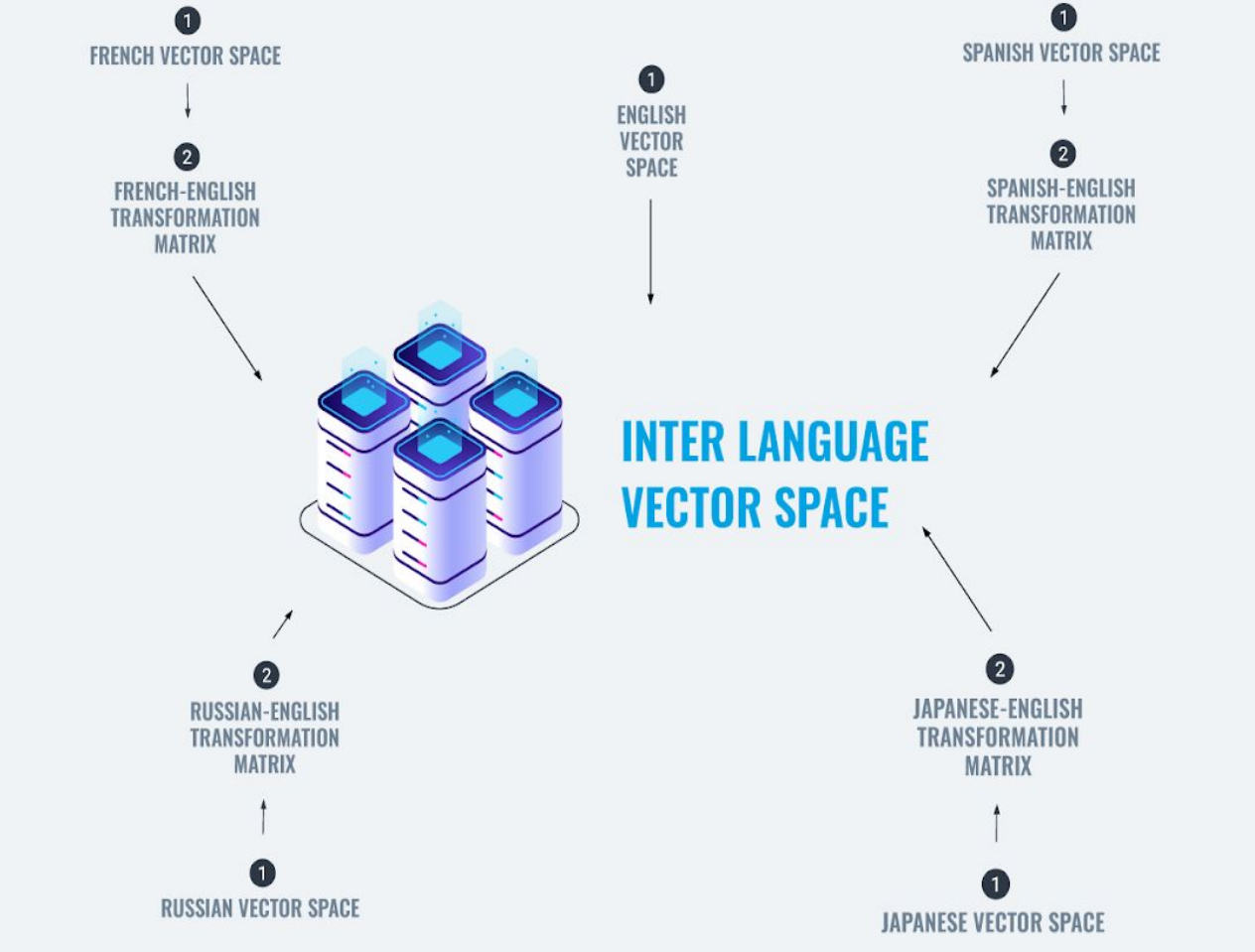
$$\max_{O} \sum_{i=1}^{n} y_i^T O x_i$$

Inter-language Vector Space

Aby znaleźć optymalną macierz O, dokonywane są następujące operacje:

1. Utwórz ze słownika macierze X_D i Y_D tak, aby i-te wiersze zawierały wektory odpowiadających sobie słów.
2. Wyznacz macierz $M = Y_D^T X_D$
3. Dokonaj rozkładu według wartości osobliwych M: $M = U\Sigma V^T$
4. Finalna macierz transformacji jest złożeniem macierzy U^T i V^T

Inter-language Vector Space



Inter-language Vector Space

Obliczanie prawdopodobieństwa, że słowa A i B są swoimi tłumaczeniami dokonywane jest na podstawie odległości cosinusowej:

$$P(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{n=1}^{300} A_i B_i}{\sqrt{\sum_{n=1}^{300} A_i^2} \cdot \sqrt{\sum_{n=1}^{300} B_i^2}}$$

Inter-language Vector Space

Przykładowe wyniki

Słowo angielskie	Słowo włoskie	Podobieństwo
cat	gatto	0.696
cat	gatta	0.552
cat	giorno	0.164
day	giorno	0.692
day	fuoco	0.193
fire	fuoco	0.590

Analiza komponentów zdań

							problème de surproduction				
		ils	connaissent	très	bien	le		problème	de	surproduction	.
	They	0.88 ✚	0.48	0.46 ⇳	0.56 ⇳	0.2	0.22 ↔	0.3	0.15	0.16	
know about		0.37 ↕	0.48 ✢	0.25 ↕	0.43 ↕	0.19 ↕	0.27 ↔	0.33	0.16 ↕	0.08	
know about	know	0.37	0.68 ✚	0.31	0.49	0.19	0.27 ↔	0.33	0.16	0.08	
know about	about	0.24	0.36	0.23	0.39	0.16	0.19 ↔	0.25	0.41 ⇔	0.04	
the overproduction		0.12	0.16	0.15	0.16	0.14	0.6 ↔	0.31	0.11	0.66 ✢	
the overproduction	the	0.23	0.35	0.23	0.34	0.75 ✚	0.36 ↔	0.35	0.44 ⇳	0.21	
the overproduction	overproduction problem	0.18	0.29	0.22	0.24	0.14	0.8 ✚	0.68 ↕	0.15	0.58	
the overproduction	overproduction problem overproduction	0.17	0.15	0.13	0.13	0.11	0.63 ↔	0.54	0.14	1.77 ✚	
the overproduction	overproduction problem problem	0.23	0.29	0.25	0.28	0.1	0.64 ↔	1.94 ✚	0.14	0.56	
	.										1.46 ✚

Analiza komponentów zdań

Szkic algorytmu:

1. Wyznacz współczynnik dopasowania dla każdej pary słów źródło-tłumaczenie
2. Dla każdego słowa źródłowego wyznacz najlepsze dopasowanie wśród słów zdania docelowego.
3. Dla każdego słowa docelowego wyznacz najlepsze dopasowanie wśród słów zdania źródłowego.
4. Połącz symetrycznie najlepsze dopasowania.

Analiza komponentów zdań

		Wireless	Workbench	6	에는	주파수	조정	의	모든 측면	모든	측면	을	관리할	수	있는	강력한	도구	세트	가	포함되	어	있	습니다	.
									0.15															
Wireless Workbench	Wireless	1.81 I	0.17		0.12	0.11	0.09	0.08		0.07	0.06	0.05	0.04	0.04	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01	
	Workbench	0.16	1.81 I		0.15	0.13	0.11	0.1		0.08	0.07	0.06	0.05	0.04	0.04	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	
	6			2.43 I																				
	contains	0.11	0.13		0.05	0.3	-0.05	0.4		0.05	0.3	-0.03	0.3	0.04	0.08	0.14	0.3	0.3	0.39	0.16 IV	0.32	0.08	0.03	
	a	0.09	0.1		0.3	0.3	0.29	0.69 III		0.24	0.3	0.66	0.3	0.13	0.14	0.08	0.3	0.3	0.54	0.03	0.48	0.09	0.42	
									0.16															
powerful set	powerful	0.07	0.08		0.18	0.3	0.18	0.2		0.2	0.3	0.14	0.3	0.1	0.06	0.3 III	0.3	0.3	0.01	-0.06	-0.06	0	0.09	
									0.16															
set of tools	set	0.06	0.07		0.3	0.3	0.3	0.3		0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.5 II	0.3	0.3	0.3	0.3	0.3	
	of	0.05	0.06		0.18 IV	0.3	0.12	0.65		0.18 ⇳	0.3	0.01	0.3	0.09 ⇳	0.11 ⇳	0.24	0.3	0.3	0.54	0.11	0.55	0.11 ⇳	0.02	
	tools	0.04	0.05		0.3	0.3	0.3	0.3		0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.54 II	0.3	0.3	0.3	0.3	0.3	0.3	
	to	0.03	0.04		0.19	0.3	0.14	0.59		0.13	0.3	0.65	0.3	0.28	0.34	0.23	0.3	0.3	0.72 III	0.14	0.58	0.21	0.56	
	manage	0.03	0.03		0.3	0.3	0.3	0.3		0.3	0.3	0.3	0.5 II	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
									0.21 III															
every aspect	every	0.02	0.02		0.07	0.3	0.02	0.48		0.15 ⇔	0.3	0.02	0.3	0.04	-0.01	0.25	0.3	0.3	0.61	-0.02	0.56	-0.01	0.1 ⇳	
	aspect	0.02	0.02		0.3	0.3	0.3	0.3		0.3	0.5 II	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
	of	0.01	0.02		0.21	0.15	0.15	0.56		0.22	0.15	0.47	0.15	0.17	0.2	0.24	0.15	0.15	0.65	0.31	0.67 III	0.25	0.54	
									0.09															
frequency coordination	frequency	0.01	0.01		0.3	0.54 II	0.3	0.3		0.3	0.3	0.3	0.3	0.5	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
	coordination	0.01	0.01		0.09	0.3	0.26 IV	0.11		0.03	0.3	0.06 ⇳	0.3	0.06	0.06	0.12	0.3	0.3	0.11	0.02	0.22	0.06	-0.05	
	.																							2.44 I

Pozycjonowanie znaczników nietłumaczalnych

- Znaczniki nietłumaczalne znajdujące się w tekstach do tłumaczenia:
  - tagi HTML/XML
  - Znaczniki formatujące
  - Symbole specjalne

- W podstawowym scenariuszu tłumacz musi te znaczniki przenosić ręcznie ze zdania źródłowego do tłumaczenia.
- W celu zautomatyzowania tego procesu opracowałem algorytm automatycznego przenoszenia znaczników

Pozycjonowanie znaczników nietłumaczalnych

Authentication is required to change your own user data

Vous devez vous authentifier pour modifier vos propres données utilisateur

Pozycjonowanie znaczników nietłumaczalnych

- Tłumaczenie automatyczne z wykorzystaniem sieci neuronowych osiąga niższą jakość tłumaczenia w przypadku obecności znaczników nietłumaczalnych w zdaniu źródłowym.

- Rozwiązaniem zastosowanym w systemie XTM jest następujący algorytm:
  - usunięcie znaczników ze zdania źródłowego
  - tłumaczenie automatyczne
  - automatyczne pozycjonowanie znaczników

Pozycjonowanie znaczników nietłumaczalnych

- Funkcja auto-inlines systemu XTM jest wywoływana średnio **114** razy w ciągu minuty.

- Liczba obsługiwanych języków: **56,** par językowych: **1540**

- Skuteczność: **98%** zdań z automatycznie pozycjonowanymi znacznikami nie wymaga korekty

Pozycjonowanie znaczników nietłumaczalnych

- Analiza RMSE średniej odległości znacznika od pozycji prawidłowej:

FROM	TO	TOTAL	PHITS	RMSE
EN	DE	1238	1076	≈2.146
EN	ES	1270	1151	≈1.705
EN	FR	1258	1141	≈2.081
EN	IT	1270	1113	≈1.633
EN	JA	1256	1040	≈2.762
EN	KO	1254	1019	≈2.039
EN	NL	1260	1010	≈3.31
EN	RU	1260	1109	≈1.798
EN	ZHS	1248	689	≈3.36
EN	ZHT	1260	717	≈3.447
EN	ID	1248	1145	≈1.404
DE	EN	1239	958	≈1.727
DE	ES	1251	896	≈2.723
DE	FR	1239	912	≈2.611
DE	IT	1247	882	≈2.723
DE	JA	1254	836	≈3.232
DE	KO	1239	815	≈3.383
DE	NL	1253	802	≈3.757
DE	RU	1241	833	≈2.144
DE	ZHS	1233	600	≈3.923
DE	ZHT	1243	599	≈4.03
DE	ID	1239	869	≈2.169

Zgłoszenie patentowe

Zestaw Inter-language Vector Space był przedmiotem zgłoszenia patentowego w United States Patent and Trademark Office w 2020 roku:

“Inter-Language Vector Space: Effective assessment of cross-language semantic similarity of words using word-embeddings, transformation matrices and disk based indexes. “. Nr wniosku: 17/064,620.

Podsumowanie

- Algorytmy zrównoleglania i przechowywania wielojęzycznych zanurzeń słów pozwalają na efektywne i dokładne oszacowanie prawdpodobieństwa tłumaczenia słów.
- Techniki te znajdują zastosowanie w algorytmach wspomagania tłumaczenia.
- Wdrożenie w systemie XTM Cloud potwierdza skuteczność rozwiązania.

Dziękuję za uwagę!

Seminarium ZIL, IPI PAN, Warszawa