OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

www.opi.org.pl

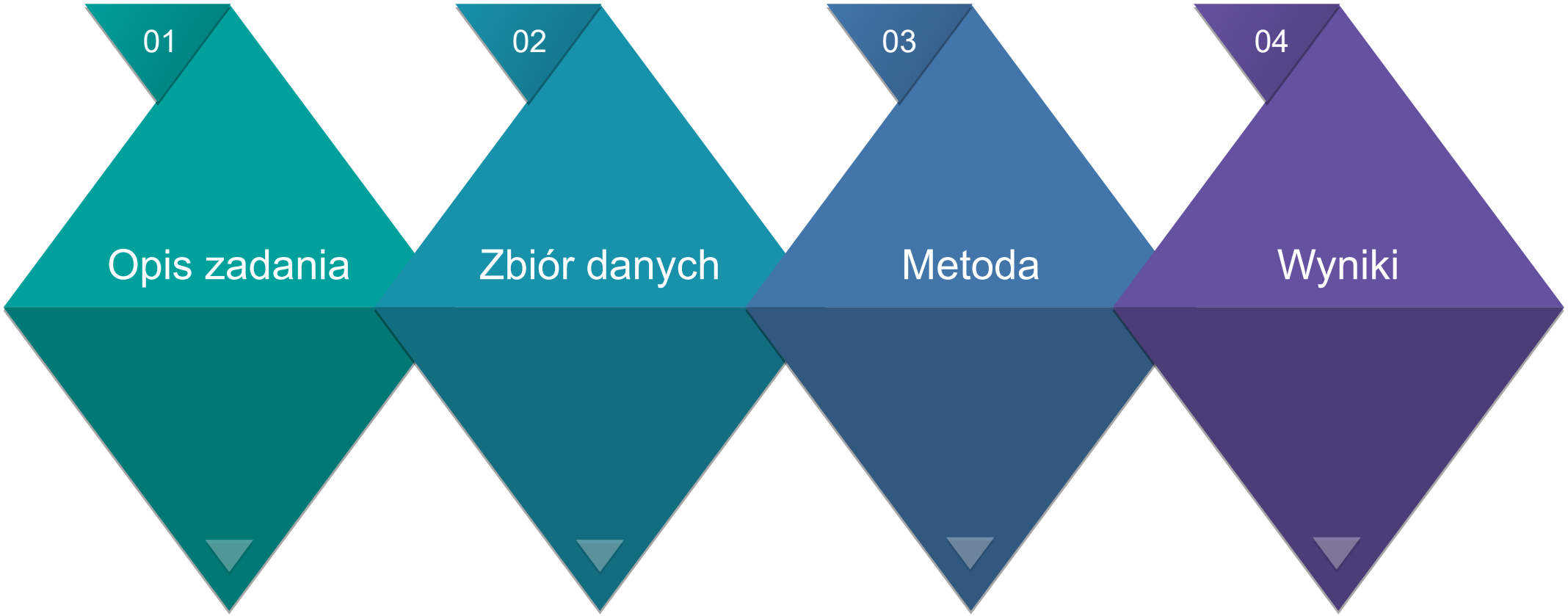
PolEval 2019: Entity Linking

Szymon Roziewski, Łukasz Podlodowski, Marek Kozłowski – Ośrodek Przetwarzania Informacji PIB

Laboratorium Inżynierii Lingwistycznej

WARSZAWA, 02.01.2018

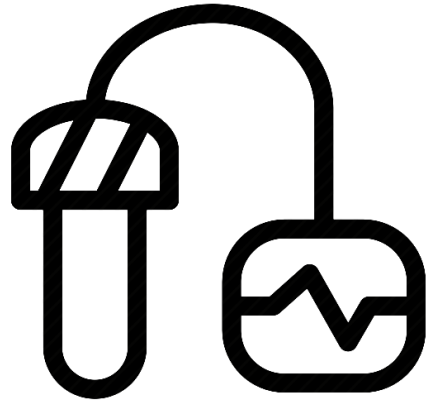
Plan prezentacji



Entity Linking Problem – NLP

- Połączenie zagadnień sztucznej inteligencji oraz językoznawstwa
- Dopasowanie znaczenia do zadanego terminu
- Niejednoznaczność terminów, sformułowań, (kaszubski, ŁKS)
- Przeniesienie znaczenia z wcześniej występującej encji, na encję powtórzoną w tym samym kontekście

Zbiór danych – Próbka



- Teksty z Wikipedii, defincje znaczeń z Wikidata
- [doc_id, token, lemma, preceding_space, morphosyntactic_tags, link_title, ent
- 2, Alfreda, Alfred, 1, subst:sg:gen:m,1, Alfred V. Aho, Q62898
- Objętość zbioru treningowego oraz testowego: (9,6 GB, 18 GB), 1.5 MB
- Prawie 2 miliony dokumentów, w zbiorze testowym tylko 2 dokumenty
- Plik zawierający etykiety znaczeń oraz ich odpowiednik w Wikipedii, dane hierarchiczne

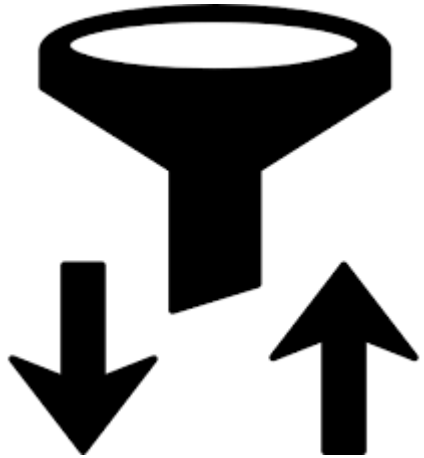
Zbiór danych

- Dwa zbiory: treningowy oraz testowy

Datasets	#entities	#u. entities	#tokens	#u. tokens	#stop words	#p.m.	#sentences
Training	42.27×10^6	1.1×10^6	348.39×10^6	5.74×10^6	69.98×10^6	66.79×10^6	25.58×10^6
Test	4071	2707	33179	13315	9907	3646	3628

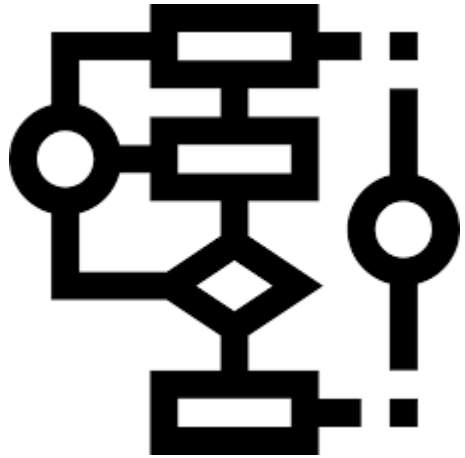
- Zbiór testowy zawierał relatywnie tyle samo terminów do zdefiniowania co zbiór treningowy, (średnio co 8 słów)
- Średnia długość zdania w zbiorze testowym i treningowym: 9, 13 odpowiednio
- 1.12 oraz 1.6 encji na zdanie (testowy I treningowy opodwiednio)

Zbiór danych – Preprocessing



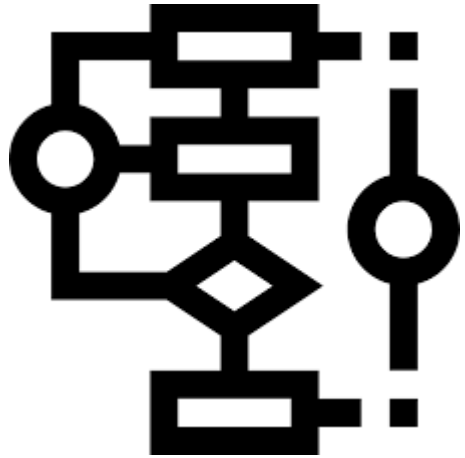
- Usunięcie znaków przestankowych oraz słów stopu
- Kropki, wykrzykniki, znaki zapytania, pozostały do identyfikacji zdań
- Usunięcie pustych tokenów

Algorytm



- Utworzenie słownika z unikalnych encji w zbiorze treningowym, dla których znaleziono dopasowanie znaczenia, używając nazwy strony z Wikipedii jako klucza
- Dodanie do tego słownika brakujących unikalnych encji, których nie dało się wyfiltrować poprzednio
- Utworzenie mapy encji mających więcej niż jedno znaczenie, użycie miary Levenshtein'a, aby wzbogacić zbiór kandydatów
- Utworzenie kontekstów na podstawie zbioru treningowego dla każdej encji, w zależności od części mowy
- Zebranie definicji znaczeń z Wikidata dla encji wieloznaczeniowych

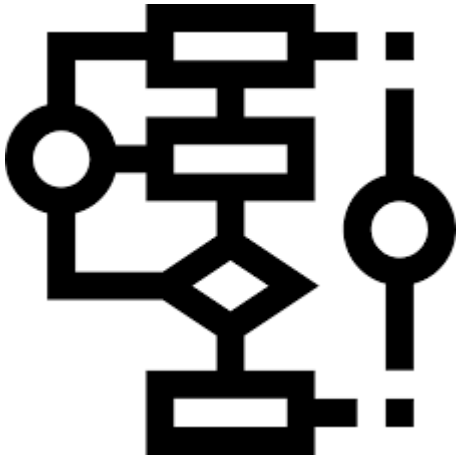
Word2Vec



- Wytrenowanie modelu Word2Vec na treningowym zbiorze danych, przestrzeń 300-wymiarowa, CBOW, window równe 2
- Utworzenie wektorów semantycznych dla kontekstów encji oraz kandydatów znaczeń
- Policzenie wartości średniej wektorów semantycznych odpowiednio dla encji jak i kandydatów znaczeń
- Porównanie ze sobą par wektorów, wzięcie najbliższych sobie, identyfikacja znaczenia

Word2Vec

Expression	Nearest Token	Distance
samochód + rower (*car* + *bicycle*)	motocykl (*motorcycle*)	0.71
jezioro + las (*lake* + *forrest*)	bagno (*swamp*)	0.68
ptak − zwierzę + samolot (*bird*− *animal* + *airplane*)	myśliwiec (*fighter plane*)	0.65
sosna − roślina + zwierzę (*pine*− *plant* + *animal*)	żubr (*aurochs*)	0.60
król − mężczyzna + kobieta (*king* − *man* + *woman*)	królowa (*queen*)	0.58
dobry − zły (*good* − *bad*)	najlepszy (*best*)	0.58



Wyniki

- 58% na zbiorze testowym przez zwykłe filtrowanie encji o unikalnym znaczeniu
- 97% znaczeń wyszukano
- 74% znaczeń wyszukano prawidłowo w zbiorze treningowym
- 26.7% znaczeń wyszukano prawidłowo w zbiorze testowym

Bibliografia

- Andor D., Alberti C., Weiss D., Severyn A., Presta A., Ganchev K., Petrov S. and Collins M.(2016).Globally normalized transition-based neural networks. „CoRR”, abs/1603.06042.Devlin J., Chang M., Lee K. and Toutanova K. (2018).BERT: pre-training of deep bidirectionaltransformers for language understanding. „CoRR”, abs/1810.04805

- Lample G., Ballesteros M., Subramanian S., Kawakami K. and Dyer C. (2016).Neural archi-tectures for named entity recognition. „CoRR”, abs/1603.01360

- Mikolov T., Sutskever I., Chen K., Corrado G. S. and Dean J. (2013a).Distributed representationsof words and phrases and their compositionality. In Burges C., Bottou L., Welling M., GhahramaniZ. and Weinberger K. (eds.),Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc

- Mikolov T., Chen K., Corrado G. and Dean J. (2013b).Efficient estimation of word representa-tions in vector space. „arXiv preprint arXiv:1301.3781”.

- Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K. and Zettlemoyer L. (2018).Deepcontextualized word representations. InProceedings of the 2018 Conference of the North AmericanChapter of the Association for Computational Linguistics: Human Language Technologies, Volume1 (Long Papers), pp. 2227–2237, New Orleans, Louisiana. Association for ComputationalLinguistics

- Raiman J. and Raiman O. (2018).Deeptype: Multilingual entity linking by neural type systemevolution. InProceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAISymposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana,USA, February 2-7, 2018, pp. 5406–5413