PolEval'2020 Task 3: Problem Ujednoznaczniania Znaczeń Słów

Arkadiusz Janz, Maciej Piasecki

Grupa Naukowa G4.19, Katedra Inteligencji Obliczeniowej

PolEval'2020, NLP Meetup

Politechnika
Wrocławska

Plan wystąpienia

- Definicja problemu
- Wyzwania i motywacja
- Przegląd rozwiązań
- Specyfika zadania konkursowego

Definicja problemu

Wprowadzenie na przykładzie

Celem jest poprawne zidentyfikowanie znaczeń słów (słowa klasy otwartej), które występują w określonym okienku tekstowym zwanym kontekstem.

```
Ogród rozkoszy ziemskich to naprawdę fantastyczne dzieło!
Tryptyk jest bez wątpienia najbardziej wyrafinowanym wytworem sztuki
sakralnej w historii malarstwa.
```

Definicja problemu

Wprowadzenie na przykładzie

Celem jest poprawne zidentyfikowanie znaczeń słów (słowa klasy otwartej), które występują w określonym okienku tekstowym zwanym kontekstem.

`Ogród` `rozkoszy` `ziemskich` `to` `naprawdę` `fantastyczne` `dzieło` !
`Tryptyk` `jest` `bez` `wątpienia` `najbardziej` `wyrafinowanym` `wytworem`
`sztuki` `sakralnej` `w` `historii` `malarstwa` .

`[]` – słowa podlegające ujednoznacznianiu

Definicja problemu

Wprowadzenie na przykładzie

Celem jest poprawne zidentyfikowanie znaczeń słów (słowa klasy otwartej), które występują w określonym okienku tekstowym zwanym kontekstem.

`Ogród rozkoszy[2] ziemskich to naprawdę fantastyczne[10] dzieło! Tryptyk[2] jest bez wątpienia najbardziej wyrafinowanym wytworem sztuki[11] sakralnej w historii malarstwa.`

1. `rozkosz.1` – najwyższy stopień uczucia przyjemności, upojenia, radości
2. `rozkosz.2` – to, co sprawia najwyższą przyjemność, zwłaszcza zmysłową
3. `tryptyk.1` – trójskrzydłowy ołtarz; trójdzielna kompozycja malarska
4. `tryptyk.2` – dzieło literackie, filmowe itp. składające się z trzech części połączonych wspólnym tematem

* Zakłada się istnienie **repozytorium znaczeniowego**. Definicje pochodzą ze słownika SJP oraz Słowosieci.

Definicja problemu

Wprowadzenie na przykładzie

Celem jest poprawne zidentyfikowanie znaczeń słów (słowa klasy otwartej), które występują w określonym okienku tekstowym zwanym kontekstem.

Ogród rozkoszy ziemskich [NE] to naprawdę fantastyczne [5] dzieło!
Tryptyk [1] jest bez wątpienia [MWE] najbardziej wyrafinowanym wytworem sztuki sakralnej [MWE] w historii malarstwa [MWE].

1. NE (Named Entity) – nazwa własna
2. MWE (Multiword Expression) – jednostka wielowyrazowa

Repozytorium znaczeń

- Anotowane korpusy
  - *SemCor*
  - *Senseval, SemEval*
  - *Wikipedia*
  - *Princeton WordNet Gloss Corpus*

  } en

  - *KPWr*
  - *Składnica*
  - *NKJP*
  - Korpus definicji i przykładów użycia Słowosieci

  } pl

- Tezaurusy
  - Słowniki ogólne: *Wiktionary*, *OmegaWiki*
  - Słowniki dziedzinowe: *MeSH*, *EuroVoc*, *AgroVoc*, ...
  - Wordnety: *Princeton WordNet*, *Słowosieć*, *Open Multilingual WordNet*
- Repozytoria hybrydowe: *CSI*

Wielojęzyczna kolekcja korpusów anotowanych DKPRO: https://dkpro.github.io/dkpro-wsd/corpora/

Repozytorium znaczeń – Princeton WordNet

Noun

- S: (n) **scream**, screaming, shriek, shrieking, screech, screeching (sharp piercing cry) *"her screaming attracted the neighbors"*
- S: (n) screech, screeching, shriek, shrieking, **scream**, screaming (a high-pitched noise resembling a human cry) *"he ducked at the screechings of shells"; "he heard the scream of the brakes"*
- S: (n) belly laugh, sidesplitter, howler, thigh-slapper, **scream**, wow, riot (a joke that seems extremely funny)

Verb

- S: (v) shout, shout out, cry, call, yell, **scream**, holler, hollo, squall (utter a sudden loud cry) *"she cried with pain when the doctor inserted the needle"; "I yelled to her from the window but she couldn't hear me"*
- S: (v) yell, **scream** (utter or declare in a very loud voice) *"You don't have to yell--I can hear you just fine"*
- S: (v) **scream** (make a loud, piercing sound) *"Fighter planes are screaming through the skies"*

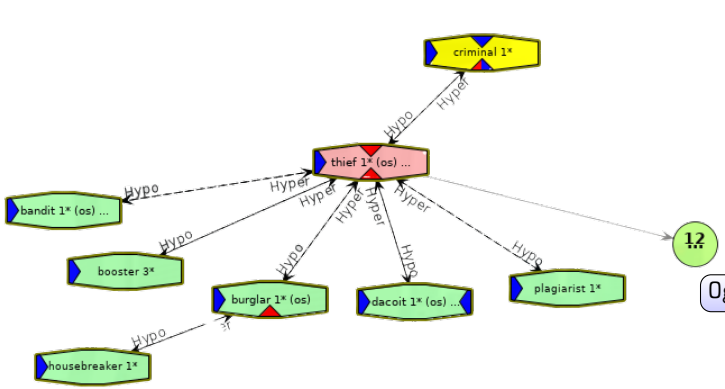
Zbiór znaczeń dla słowa *scream* – Princeton WN.

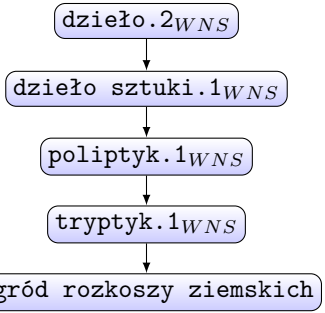
Repozytorium znaczeń – Słowosieć

RZECZOWNIK	**bank** 1 instytucja finansowa zajmująca się operacjami pieniężnymi, m.in. przyjmowaniem wpłat, prowadzeniem rachunków, udzielaniem kredytów
DOMENA	związek miedzy ludźmi, rzeczami lub ideami
PRZYKŁADY	Gdybym nie wiedziała, ile bank zarobi na moim kredycie, spałabym chyba spokojniej.
ŹRÓDŁO	Słowosieć
HIPERONIMY	Pokaż ścieżkę do najwyższego hiperonimu`
GRAPH	Wyświetl wizualizację graficzną
ANNOTACJE EMOCJONALNE	NACECHOWANIE: Brak nacechowania emocjonalnego
	NACECHOWANIE: Brak nacechowania emocjonalnego

RZECZOWNIK	**bank** 3 miejsce, gdzie coś się gromadzi (informacje, materiały, tkanki) i udostępnia w razie potrzeby
DOMENA	grupy ludzi i rzeczy
PRZYKŁADY	Banki krwi pępinowej przechowują komórki macierzyste, które w przyszłości mogą być użyte w leczeniu.
ŹRÓDŁO	Słowosieć
HIPERONIMY	Pokaż ścieżkę do najwyższego hiperonimu`
GRAPH	Wyświetl wizualizację graficzną
ANNOTACJE EMOCJONALNE	NACECHOWANIE: Brak nacechowania emocjonalnego
	NACECHOWANIE: Brak nacechowania emocjonalnego

Repozytorium znaczeń





Definicja Problemu

WSD jako tagowanie sekwencji

Mając do dyspozycji repozytorium znaczeń R, dla zadanego dokumentu tekstowego składającego się ze zbioru słów (klasy otwartej) W, próbujemy określić najbardziej prawdopodobne znaczenia $\hat{s_{w_i}}$ dla poszczególnych słów $w_i \in W$ w tym dokumencie.

$$W = \{w_1, ..., w_k\}, w_i \in W \tag{1}$$

$$S_{w_i} = \{s^1_{w_i}, s^2_{w_i}, ..., s^{N_{w_i}}_{w_i}\}, S_{w_i} \in R, \tag{2}$$

$$\hat{s_{w_i}} = \arg\max_{s \in S_{w_i}} P(s|context(w_i)), \tag{3}$$

gdzie $\boldsymbol{context(w_i)}$ jest funkcją określającą kontekst wystąpienia ujednoznacznianego słowa w_i, a $\boldsymbol{S_{w_i}}$ przyjętym w repozytorium zestawem znaczeń dla słowa w_i – zbiorem rozpoznawanych znanych klas.

`Ten film zrobił na mnie ogromne wrażenie!`

Wyzwania i motywacja

Wyzwania

- Mocna zależność WSD od segmentacji tekstu i tagowania morfosyntaktycznego (szczególnie istotne przy stosowaniu przetwarzania potokowego),
- Bazy wiedzy: niekompletne, zawierają nadmiarowe informacje o znaczeniach, brakujące powiązania między znaczeniami, zmienna ziarnistość
- Anotowane korpusy: obciążone w kierunku najczęstszych znaczeń, niekompletne, wymagają ogromnego nakładu pracy anotacyjnej,
- Algorytmy: nieefektywne z uwagi na ogromne rozmiary baz wiedzy, ogromny zestaw rozpoznawanych klas, niedoreprezentowane korpusy anotowane, nierównomierny rozkład znaczeń, nieograniczony zestaw dziedzin tekstu.

Wyzwania i motywacja

Motywacją zastosowania

...

- Ewaluacja: mało wiarygodna z praktycznego punktu widzenia, ponieważ dane ewaluacyjne ograniczone są zwykle do pewnej grupy słów i ich najczęstszych znaczeń,
- Jak szeroki powinien być kontekst? "One sense per discourse"(Gale, 1992), "One sense per collocation"(Yarovsky, 1993), - hipotezy niewłaściwe dla słów o większej liczbie znaczeń i odmiennej ziarnistości znaczeniowej.

Motywacją zastosowania!

- Tłumaczenie maszynowe, analiza wydźwięku, analiza semantyczna tekstu, wydobywanie informacji, systemy odpowiedzi na pytania, ...

Przegląd rozwiązań

Typy podejść

- Uczenie nadzorowane z wykorzystaniem anotowanych korpusów tekstów,
- Uczenie bez nadzoru z wykorzystaniem surowych korpusów tekstów,
- Uczenie pół-nadzorowane,
- Podejścia oparte o bazy wiedzy,
- Podejścia hybrydowe.

Przegląd rozwiązań

Klasyczny Lesk:

`bank.1: instytucja finansowa zajmująca się operacjami pieniężnymi, przyjmowaniem wpłat, prowadzeniem rachunków, gromadzeniem oszczędności`

`bank.2: miejsce, gdzie coś się gromadzi i udostępnia w razie potrzeby, np. informacje, materiały, tkanki`

Kontekst: `W moim banku założenie rachunku na oszczędności jest banalnie proste!`

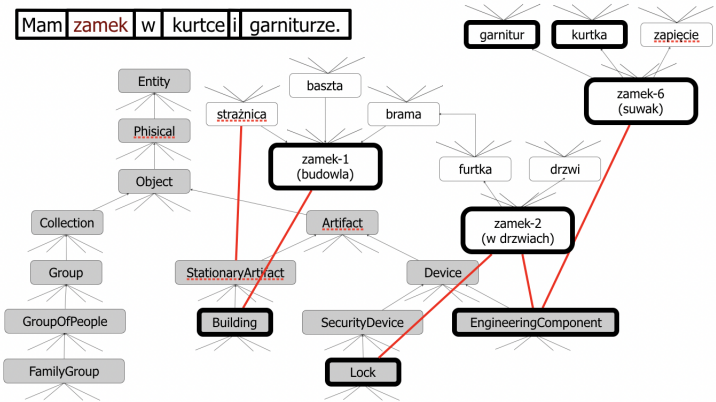
Dystrybucyjny Lesk:

- Podobieństwo kosinusowe reprezentacji wektorowej definicji znaczenia wyrazu w_i (lub jego przykładu użycia z repozytorium) do kontekstu w tekście ujednoznacznianym.

$$\text{Score}(s, w) = cos(G_s, C_w) + cos(L_{s,w}, C_w)$$

Przegląd rozwiązań

UKB / WoSeDon / Babelfy – wykorzystanie wordnetu i jego powiązań jako struktury grafowej oraz przetwarzanie jej z wykorzystaniem grafowych miar oceny istotności węzła pod warunkiem kontekstu



Przegląd rozwiązań

Babelfy

Z niego był taki jeleń , że szkoda w ogóle z nim rozmawiać .

be
Be identical to; be someone or something

such a
Of a kind specified or understood.

deer
Distinguished from Bovidae by the male's having solid deciduous antlers

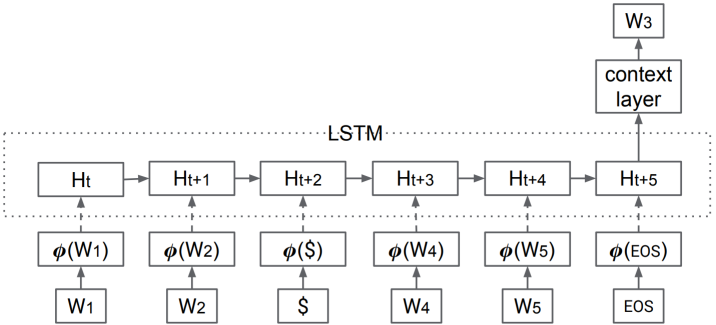
damage
Any harm or injury resulting from a violation of a legal right

society
An extended social group having a distinctive cultural and economic organization

talk
Exchange thoughts; talk with

Przegląd rozwiązań

LSTM-based WSD: nadzorowane (uczenie na korpusach anotowanych) i nienadzorowane (LSTM jako model językowy do "podpowiadania słów")



Przegląd rozwiązań

LSTM-based WSD: nadzorowane (uczenie na korpusach anotowanych) i nienadzorowane (LSTM jako model językowy do "podpowiadania słów")

sentence	top 10 predictions from LSTM	sense
Employee compensation is offered in the form of cash and/or *stock*.	cash, stock, equity, shares, loans, bonus, benefits, awards, equivalents, deposits	sense#1
The *stock* would be redeemed in five years, subject to terms of the company's debt.	bonds, debt, notes, shares, stock, balance, securities, rest, Notes, debentures	
These stores sell excess *stock* or factory overruns .	inventory, goods, parts, sales, inventories, capacity, products, oil, items, fuel	sense#2
Our soups are cooked with vegan *stock* and seasonal vegetables.	foods, food, vegetables, meats, recipes, cheese, meat, chicken, pasta, milk	sense#3
In addition, they will receive *stock* in the reorganized company, which will be named Ranger Industries Inc.	shares, positions, equity, jobs, awards, representation, stock, investments, roles, funds	?

Przegląd rozwiązań

GlossBERT – podejście nadzorowane, zastosowanie BERTa do rozpoznawania znaczeń, wzbogacenie o wiedzę pochodzącą z przykładów użycia i definicji wordnetowych.

Context-Gloss Pairs with weak supervision of the target word [research]	Label	Sense Key
[CLS] Your "research" ... [SEP] research: systematic investigation to ... [SEP]	Yes	research%1:04:00::
[CLS] Your "research" ... [SEP] research: a search for knowledge [SEP]	No	research%1:09:00::
[CLS] Your "research" ... [SEP] research: inquire into [SEP]	No	research%2:31:00::
[CLS] Your "research" ... [SEP] research: attempt to find out in a ... [SEP]	No	research%2:32:00::

Przegląd rozwiązań

Osiągane rezultaty

System	SE07	SE2	SE3	SE13	SE15	Noun	Verb	Adj	Adv	**All**
MFS baseline	54.5	65.6	66.0	63.8	67.1	67.7	49.8	73.1	80.5	65.5
Lesk$_{ext+emb}$	56.7	63.0	63.7	66.2	64.6	70.0	51.1	51.7	80.6	64.2
Babelfy	51.6	67.0	63.5	66.4	70.3	68.9	50.7	73.2	79.8	66.4
IMS	61.3	70.9	69.3	65.3	69.5	70.5	55.8	75.6	82.9	68.9
IMS$_{+emb}$	62.6	72.2	70.4	65.9	71.5	71.9	56.6	75.9	84.7	70.1
Bi-LSTM	-	71.1	68.4	64.8	68.3	69.5	55.9	76.2	82.4	68.4
Bi-LSTM$_{+att.+LEX+POS}$	64.8	72.0	69.1	66.9	71.5	71.5	57.5	75.0	83.8	69.9
GAS$_{ext}$ (Linear)	-	72.4	70.1	67.1	72.1	71.9	58.1	76.4	84.7	70.4
GAS$_{ext}$ (Concatenation)	-	72.2	70.5	67.2	72.6	72.2	57.7	76.6	85.0	70.6
CAN^{s}	-	72.2	70.2	69.1	72.2	73.5	56.5	76.6	80.3	70.9
HCAN	-	72.8	70.3	68.5	72.8	72.7	58.2	77.4	84.1	71.1
SemCor, hypernyms (single)	-	-	-	-	-	-	-	-	-	75.6
SemCor, hypernyms (ensemble)†	69.5	77.5	77.4	76.0	78.3	79.6	65.9	79.5	85.5	76.7
SemCor+WNGC, hypernyms (single)‡	-	-	-	-	-	-	-	-	-	77.1
SemCor+WNGC, hypernyms (ensemble)† ‡	73.4	79.7	77.8	78.7	82.6	81.4	68.7	83.7	85.5	79.0
BERT(Token-CLS)	61.1	69.7	69.4	65.8	69.5	70.5	57.1	71.6	83.5	68.6
GlossBERT(Sent-CLS)	69.2	76.5	73.4	75.1	79.5	78.3	64.8	77.6	83.8	75.8
GlossBERT(Token-CLS)	71.9	77.0	**75.4**	74.6	79.3	78.3	66.5	**78.6**	84.4	76.3
GlossBERT(Sent-CLS-WS)	**72.5**	**77.7**	75.2	**76.1**	**80.4**	**79.3**	**66.9**	78.2	**86.4**	**77.0**

Przegląd rozwiązań

Osiągane rezultaty

Method	*Sens-2*	*Sens-3*	*Sem-07*	*Sem-13*	*Sem-15*
MFS	66.80	66.20	55.20	63.00	67.80
Babelfy	67.00	63.50	51.60	66.40	**70.30**
UKB-nf	61.30	54.90	42.20	60.90	62.90
UKB-sf	67.50	**66.40**	54.10	64.00	67.80
UKB-nf-w2w	64.20	54.80	40.00	64.50	64.50
UKB-sf-w2w	68.80	66.10	53.00	**68.80**	70.30
PPRMC-1	66.26	64.28	54.06	65.08	67.12
PPRMC-2	66.35	65.13	55.60	65.56	66.63
PPRMC-3	66.47	65.94	56.04	65.26	67.71
PPRMC-4	66.78	66.28	**56.48**	65.90	68.10

WoSeDon na danych angielskich

Method	*Skład.-N*	*Skład.-V*	*KPWr-N*	*KPWr-V*
PPRMC-1	63.19	44.75	52.92	33.42
PPRMC-2	64.27	46.01	53.24	33.73
PPRMC-3	64.88	46.22	53.31	33.66
PPRMC-4	65.28	46.51	53.66	33.09
WoSeDon	63.92	46.43	53.61	33.71
WoSeDon	64.85	47.29	53.80	34.08
WoSeDon	65.27	47.55	54.02	34.00
WoSeDon	66.18	48.74	54.90	33.89

WoSeDon na danych polskich

Specyfika zadania konkursowego

Założenia

Proponujemy dwa warianty konkursowe

- *Fixed competition*
  - Zależy nam na unikaniu stosowania korpusów anotowanych,
  - Wykorzystujemy dostępną bazę wiedzy (wysokie pokrycie) i dane surowe,
  - Opracowanie algorytmów opartych na stosowaniu baz wiedzy, ale o zwiększonej precyzji ujednoznaczniania.
- *Open competition*
  - Bez ograniczeń - wszystkie dane dozwolone,
  - Zależy nam na opracowaniu najskuteczniejszego rozwiązania dla języka polskiego.

Specyfika zadania konkursowego

Dane konkursowe

- Dane do przygotowywania rozwiązań:
  - Repozytorium znaczeniowe w postaci Słowosieci wraz ze strukturą leksykalno-semantyczną (związki między znaczeniami),
  - Korpus przykładów użycia i definicji znaczeń ze Słowosieci,
  - Surowe korpusy tekstów (np. Wikipedia, Common Crawl, KGR10)
- Pozostałe dane (dozwolone w wariancie Open Competition):
  - Anotowane korpusy (Składnica + inne anotowane zasoby, w tym również zasoby w innych językach)
  - Powiązania Słowosieci z ontologiami (DBPedia, YAGO), Wikipedią, innymi tezaurusami.

Specyfika zadania konkursowego

Ewaluacja

Do ewaluacji wykorzystujemy standardowe miary oceny skuteczności klasyfikacji, głównie precyzję i kompletność, dostosowane do zagadnienia WSD:

$$\textit{Precision:} \quad \frac{\#of\text{–}correctly\text{–}predicted\text{–}senses}{\#of\text{–}words\text{–}for\text{–}which\text{–}the\text{–}algorithm\text{–}made\text{–}a\text{–}decision}$$

$$\textit{Recall:} \quad \frac{\#of\text{–}correctly\text{–}predicted\text{–}senses}{\#of\text{–}annotated\text{–}words\text{–}in\text{–}our\text{–}test\text{–}data}$$

- Dane ewaluacyjne: nowa anotacja KPWr!
  - Dane zrównoważone pod kątem frekwencji znaczeń,
  - Dane o zwiększonym pokryciu słownictwa,
  - Dane kompatybilne z nowszą wersją Słowosieci.

Zakończenie

Dziękuję za uwagę!

Bibliografia I


Eneko Agirre, Oier López de Lacalle i Aitor Soroa. "The risk of sub-optimal use of Open Source NLP Software: UKB is inadvertently state-of-the-art in knowledge-based WSD". W: *arXiv preprint arXiv:1805.04277* (2018).


Andrea Moro, Alessandro Raganato i Roberto Navigli. "Entity linking meets word sense disambiguation: a unified approach". W: *Transactions of the Association for Computational Linguistics* 2 (2014), s. 231–244.


Dieke Oele i Gertjan Van Noord. "Distributional lesk: Effective knowledge-based word sense disambiguation". W: *IWCS 2017—12th International Conference on Computational Semantics—Short papers*. 2017.


Dayu Yuan i in. "Semi-supervised word sense disambiguation with neural models". W: *arXiv preprint arXiv:1603.07012* (2016).

i wiele innych prac :)