PolEval'2020 Task 3: Problem Ujednoznaczniania Znaczeń Słów

Arkadiusz Janz, Maciej Piasecki

Grupa Naukowa G4.19, Katedra Inteligencji Obliczeniowej

PolEval'2020, AI&NLP Day

Politechnika
Wrocławska

Plan wystąpienia

- Wprowadzenie do problemu
- Wyzwania i motywacja
- Dane konkursowe
- Rezultaty uczestników
- Przyszłe prace i podsumowanie

Wprowadzenie do problemu

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.`

Wprowadzenie do problemu

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

Wprowadzenie do problemu

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.

Wprowadzenie do problemu

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

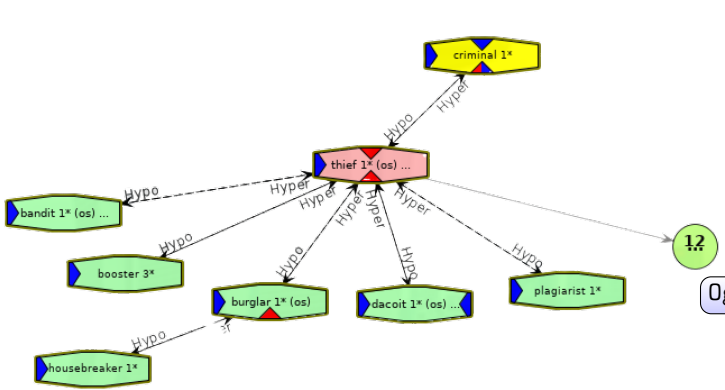
- S: (n) **scream**, screaming, shriek, shrieking, screech, screeching (sharp piercing cry) *"her screaming attracted the neighbors"*
- S: (n) screech, screeching, shriek, shrieking, **scream**, screaming (a high-pitched noise resembling a human cry) *"he ducked at the screechings of shells"; "he heard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

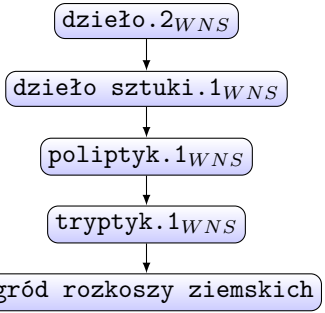
Verb

- 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





Leksykalno-sematyczna struktura *Słowosieci*.

Repozytorium konkursowe – Słowosieć 3.2!

Feature	Sense Inventory	
	plWordNet 2.1	*plWordNet 3.2*
number of distinct lexical units	206 567	286 804
number of distinct multi-word lexical units	53 752	70 019
number of distinct synsets	151 252	221 101
number of monosemous lemmas	113 129	141 343
number of polysemous lemmas	33 507	49 049
number of monosemous lemmas (multi-word only)	43 906	56 415
number of polysemous lemmas (multi-word only)	3 898	5 171
number of lexical units with definition or any usage example	37 207	145 901
number of lexical units without definition or any usage example	169 360	140 903
average length of utterance (definition or example)	12.56	11.54
average number of senses per lemma (polysemous only)	2.79	2.96

Tablica: Statystyczna analiza polskich repozytoriów znaczeniowych bazujących na wordnetach.

Repozytorium konkursowe – Słowosieć 3.2!

Dlaczego nie wersja bieżąca – 4.1?

- Słownik walencyjny *Walenty* dostosowano do Słowosieci w wersji 3.2,
- Przejście z anotacjami z wersji 2.1 do 4.1 wymaga znacznie więcej pracy ręcznej,
- Rzutowania na Linked Open Data były generowane głównie dla Słowosieci 3.2.

Definicja Problemu

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

- 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

...

- 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, ...

Specyfika zadania konkursowego

Założenia

Zaproponowano 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.

Dane treningowe

- 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.

Dane ewaluacyjne

Wprowadzono dwa niezależne zbiory danych na potrzeby tej edycji konkursu:

- "The Adventure of the Speckled Band", nazywany korpusem ***SPEC*** [1], ósma powieść Conana Doyle'a o przygodach popularnego detektywa znanego jako Sherlock Holmes,
- ***KPWr–100*** – nowa próbka *100* oznakowanych ręcznie dokumentów z Korpusu Politechniki Wrocławskiej KPWr.

Zgodność anotacji mierzona za pomocą Positive Specific Agreement (PSA) na poziomie **0.602** dla korpusu *SPEC* i **0.678** dla korpusu *KPWr-100*.

[1] `https://clarin-pl.eu/dspace/handle/11321/667`

Specyfika danych konkursowych

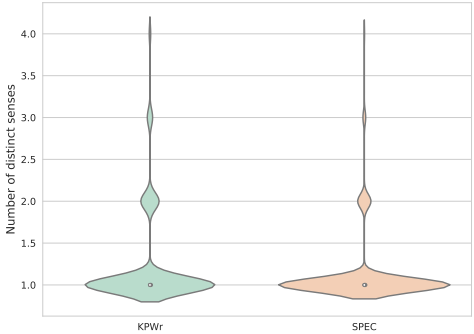
Corpus	#Nouns	#Verbs	#Adject.	#Adverbs	#Total
KPWr-100	7 028	3 428	2 442	677	13 891
SPEC	1 617	1 182	487	219	3 689

Tablica: Rozkład anotowanych tokenów w korpusach ewaluacyjnych z odniesieniem do ich części mowy.

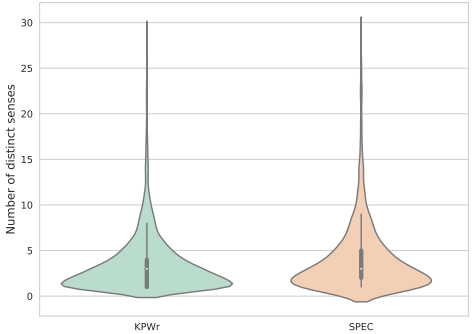
Corpus	#Nouns	#Verbs	#Adject.	#Adverbs	#Total
KPWr-100	205	282	0	28	515
SPEC	14	154	0	20	188

Tablica: Rozkład anotowanych wyrażeń wielowyrazowych w korpusach ewaluacyjnych z odniesieniem do ich części mowy.

Specyfika danych konkursowych



Rozkład liczby znaczeń w korpusach wyliczony według danych anotowanych.



Rozkład liczby potencjalnych znaczeń według repozytorium znaczeniowego.

Specyfika zadania konkursowego

Ewaluacja

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

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

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

Wzorcowe znakowanie

ORDER_ID	TOKEN_ID	ORTH	LEMMA	CTAG	FROM	TO	WN_ID
12	12	zbudowanym	zbudować	ppas:sg:inst:m3:perf:aff	63	72	s58913
13	13	w	w	prep:loc:nwok	73	73	**s468881**
14	14	dużej	duży	adj:sg:loc:f:pos	74	78	**s468881**
15	15	mierze	miara	subst:sg:loc:f	79	84	**s468881**
16	16	w	w	prep:loc:nwok	85	85	_
17	17	oparciu	oparcie	subst:sg:loc:n	86	92	s99717
18	18	o	o	prep:acc	93	93	_

Wzorcowy sposób znakowania danych z uwzględnieniem wyrażeń wielowyrazowych.

Baseline

W ewaluacji przyjęto następujące **punkty odniesienia**:

- First WordNet Sense (**FWNS**) – heurystyka preferująca zawsze znaczenie o najniższym wariancie, np. *piękny*.1.adj zamiast pozostałych znaczeń {*piękny*.2.adj, *piękny*.3.adj, ...},
- **WoSeDon**[2] – otwarty system do ujednoznaczniania znaczeń słów przygotowany **w podstawowej wersji** (*Fixed Competition*); algorytm propagacji aktywacji po sieci bazujący na algorytmie PageRank w wariancie Word-to-Word[3].

[2] `https://clarin-pl.eu/dspace/handle/11321/290`

[3] Agirre, Eneko, and Aitor Soroa. "Personalizing pagerank for word sense disambiguation."Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)

Wyniki nadesłanych prac i baseline'ów

Corpus	Submission [4]	precision	recall	f1-score
KPWr-100	*DK-V3*	0.599	0.589	0.594
	AJ-V2	0.318	0.231	0.268
	*FWNS**	0.563	0.556	0.559
	*WoSeDon**	**0.625**	**0.618**	**0.621**
SPEC	*DK-V3*	0.592	0.577	0.584
	AJ-V2	0.292	0.201	0.238
	*FWNS**	0.587	0.575	0.581
	*WoSeDon**	**0.607**	**0.594**	**0.600**

Tablica: Wyniki uzyskane przez nadesłane rozwiązania ocenione na danych *KPWr-100* i *SPEC*. Porównanie z przyjętymi punktami odniesienia.

[4] DK – Dariusz Kłeczek, AJ – Arleta Juszczak

Wyniki nadesłanych prac i baseline'ów

Corpus	Submission	precision	recall	f1-score
KPWr-100	*DK-V3*	**0.599**	**0.589**	**0.594**
	AJ-V2	0.318	0.231	0.268
	*FWNS**	0.563	0.556	0.559
	*WoSeDon**	0.625	0.618	0.621
SPEC	*DK-V3*	**0.592**	**0.577**	**0.584**
	AJ-V2	0.292	0.201	0.238
	*FWNS**	0.587	0.575	0.581
	*WoSeDon**	0.607	0.594	0.600

Tablica: Wyniki uzyskane przez nadesłane rozwiązania ocenione na danych *KPWr-100* i *SPEC*. Porównanie z przyjętymi punktami odniesienia.

Wyniki nadesłanych prac i baseline'ów

Corpus	Submission	precision	recall	f1-score
KPWr-100	*DK-V3*	**0.599**	**0.589**	**0.594**
	AJ-V2	0.318	0.231	0.268
	*FWNS**	0.563	0.556	0.559
	*WoSeDon**	0.625	0.618	0.621
SPEC	*DK-V3*	**0.592**	**0.577**	**0.584**
	AJ-V2	0.292	0.201	0.238
	*FWNS**	0.587	0.575	0.581
	*WoSeDon**	0.607	0.594	0.600

Gratulacje!

Tablica: Wyniki uzyskane przez nadesłane rozwiązania ocenione na danych *KPWr-100* i *SPEC*. Porównanie z przyjętymi punktami odniesienia.

Przyszłe prace i podsumowanie

Przyszłe prace:

- Prace nad **unifikacją wszystkich zasobów** anotowanych dla zagadnienia WSD i uzyskanie zgodności z jedną wersją repozytorium znaczeniowego,
- Rozszerzony **benchmark na bazie zunifikowanych zasobów**,
- Przygotowanie zasobów treningowych do wariantu ***Open Competition***,
- Przygotowania do kolejnej edycji?

Podsumowanie:

- Powstała wstępna **baza do pracy nad algorytmami WSD**,
- Wyznaczono **solidny baseline** dla wariantu Fixed Competition,
- W pierwszej edycji **5 nadesłanych rozwiązań** przygotowanych i zgłoszonych przez **2 uczestników**, wszystkie dla wariantu ***Fixed Competition***.

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 :)