

PolEval 2020

Łukasz Kobyliński

Instytut Podstaw Informatyki PAN / Sages

Czy przetwarzanie jęz. polskiego jest trudne?

Polski jest językiem fleksyjnym

- Liczba możliwych form w korpusie jest dużo większa, niż np. dla języka angielskiego,
- wiele form jest morfologicznie niejednoznacznych
 - człowieka (D. - dopełniacz / B. - biernik)
 - człowieku (Ms. - miejscownik / W. - wołacz)
 - ludzie (M. - mianownik / W. pl. - wołacz l. mn.)
 - ludzi (D. - dopełniacz / B. pl. - biernik l. mn.)

człowiek

rzeczownik [SJPDor.]

m1 ○ [B3k+człowiek](#) / ○ [B3k+człowiek+!](#)

	I. p.	I. m.	
M.	człowiek	ludzie	ndepr
		ludzie	depr
D.	człowieka	ludzi	
C.	człowiekowi	ludziom	
B.	człowieka	ludzi	
N.	człowiekiem	ludźmi	
Ms.	człowieku	ludziach	
W.	człowieku człowiecze <i>daw.</i>	ludzie	ndepr
		ludzie	depr

Czy przetwarzanie jęz. polskiego jest trudne? (cd.)

Wiele niejednoznaczności:

- na poziomie segmentacji słów
- na poziomie odmiany słów,
- na poziomie homonimów.

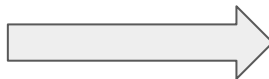
admirat			
rzeczownik [SJPDor.]			
m1 B4ł+w			
	l. p.	l. m.	
M.	admirat	admiratowie	ndepr
		admiraty	depr
D.	admirata	admiratów	
C.	admiratowi	admiratom	
B.	admirata	admiratów	
N.	admiratem	admiratami	
Ms.	admirale	admiratach	
W.	admirale	admiratowie	ndepr
		admiraty	depr

admirat gatunek motyla			
rzeczownik			
m2 B4ł+w			
	l. p.	l. m.	
M.	admirat	admiraty	
D.	admirata	admiratów	
C.	admiratowi	admiratom	
B.	admirata	admiraty	
N.	admiratem	admiratami	
Ms.	admirale	admiratach	
W.	admirale	admiraty	

Czy przetwarzanie jęz. polskiego jest trudne? (cd.)

Swobodny szyk zdania

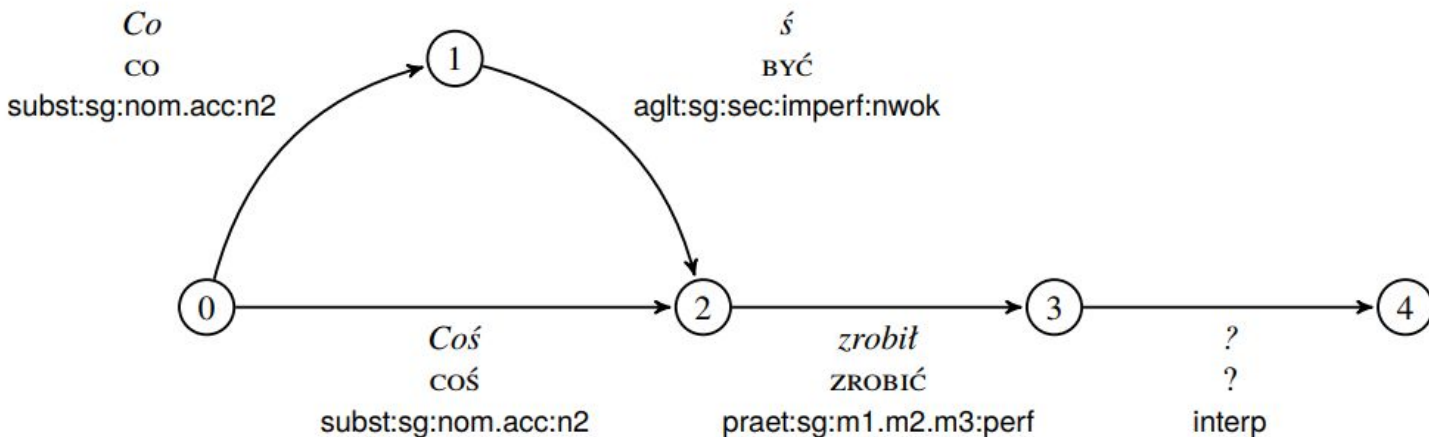
- jęz. angielski
 - John loves Mary (SVO)
- jęz. polski
 - Jan kocha Marię (SVO),
 - Jan Marię kocha,
 - Marię kocha Jan,
 - Marię Jan kocha,
 - Kocha Marię Jan,
 - Kocha Jan Marię.



Pewne metody (jak na przykład ukryte modele Markowa) nie są tak efektywne, jak w przypadku języka angielskiego.

Czy przetwarzanie jęz. polskiego jest trudne? (cd.)

Problem segmentacji słów



Jakie mamy problemy?

Język polski jest obiektywnie trudniejszy w przetwarzaniu automatycznym
(moja hipoteza - patrz wcześniejsze slajdy)

Dane dla języka polskiego są trudniej dostępne
(mniej użytkowników języka, mniej projektów zbierania i anotowania danych)

Mniej badaczy zajmuje się przetwarzaniem języka polskiego
(patrz wyżej)

Brak dobrej wymiany wiedzy między badaczami, a także między badaczami a biznesem
(mniej projektów nauka-biznes, środowisko naukowe nie jest zbyt duże)

PolEval - motywacja

Wnioski

Prace dotyczące przetwarzania języka polskiego wymagają wsparcia poprzez działania popularyzacyjne, tworzenie otwartych zasobów i otwartych metod, które mogą być praktycznie wykorzystane.

... zorganizujemy otwarty konkurs z nagrodami dla zespołów pracujących nad przetwarzaniem języka polskiego?

Cele konkursu PolEval

Konkurs wzorowany na Kaggle, obecnie związany z zadaniami z obszaru NLP dla j. polskiego, którego celem jest:

- wypracowanie ustalonych procedur ewaluacji systemów rozwiązujących poszczególne zadania w NLP,
- wytworzenie anotowanych zbiorów danych, które mogą zostać wykorzystane do uczenia i ewaluacji systemów,
- obiektywne porównanie systemów realizujących odpowiednie zadania w zakresie przetwarzania języka naturalnego,
- zbliżenie do siebie badaczy ze środowisk naukowych i biznesowych oraz wymianę wiedzy między nimi,
- popularyzacja zagadnień NLP w kontekście języka polskiego.

Zadania na PolEval

Dlaczego taka formuła się sprawdza?

Zadania na PolEval to:

- dobrze sformułowane ramy zadania - zakres, dane wejściowe, oczekiwane dane wyjściowe,
- przygotowane przez organizatorów dane treningowe i testowe,
- opracowany sposób ewaluacji,
- możliwość porównania wielu systemów w jednakowym, opisanym wyżej środowisku.

PolEval 2017

PolEval 2017

- Zadanie 1: Znakowanie morfosyntaktyczne (POS tagging)
- Zadanie 2: Analiza wydźwięku (Sentiment analysis)

W 2017 roku otrzymaliśmy:

- 16 zgłoszeń od 9 zespołów,
- wszystkie zgłoszenia (!) oparte były na sieciach neuronowych.



Task 1: POS Tagging

Introduction

There is an ongoing discussion whether the problem of part of speech tagging is already solved, at least for English (see Manning 2011), by reaching the tagging error rates similar or lower than the human inter-annotator agreement, which is ca. 97%. In the case of languages with rich morphology, such as Polish, there is however no doubt that the accuracies of around 91% delivered by taggers leave much to be desired and more work is needed to proclaim this task as solved.

The aim of this proposed task is therefore to stimulate research in potentially new approaches to the problem of POS tagging of Polish, which will allow to close the gap between the tagging accuracy of systems available for English and languages with rich morphology.

Task definition

Subtask (A): Morphosyntactic disambiguation and guessing

Given a sequence of segments, each with a set of possible morphosyntactic interpretations, the goal of the task is to select the correct interpretation for each of the segments and provide an interpretation for segments for which only 'ign' interpretation has been given (segments unknown to the morphosyntactic dictionary).

Subtask (B): Lemmatisation

Given a sequence of segments, each with a set of possible morphosyntactic interpretations, the goal of the task is to select the correct lemma for each of the segments and provide a lemma for segments for which only 'ign' interpretation has been given (segments unknown to the morphosyntactic dictionary).

Stan dla języka polskiego - wyniki PolEval 2017

System name	Acc (%)	deep network	hand-crafted features	character-level embeddings	word-level embeddings
<i>Toygger</i>	94.6343	<i>yes</i>	<i>yes</i>	<i>no</i>	<i>yes</i>
<i>KRNNT_AB</i>	93.8083	<i>yes</i>	<i>yes</i>	<i>no</i>	<i>no</i>
<i>NeuroParser</i>	93.6109	<i>yes</i>	<i>no</i>	<i>yes</i>	<i>no</i>
<i>AvgPer_Forced</i>	90.9134	<i>no</i>	<i>yes</i>	<i>no</i>	<i>no</i>
<i>Concraft</i>	91.6115	<i>no</i>	<i>yes</i>	<i>no</i>	<i>no</i>
<i>WCRFT</i>	91.1693	<i>no</i>	<i>yes</i>	<i>no</i>	<i>no</i>
<i>WMBT</i>	90.6722	<i>no</i>	<i>yes</i>	<i>no</i>	<i>no</i>

PolEval 2017: czy było warto?

Zadanie 1: Znakowanie morfosyntaktyczne

- powstał korpus treningowy zawierający 55K segmentów, ręcznie anotowany znacznikami morfosyntaktycznymi
- 5 nowych tagerów zostało udostępnionych publicznie
- uczestnicy zwracali uwagę nie tylko na dokładność metod, ale także na efektywność zgłaszanych metod
- poprawa dokładności względem poprzednio istniejących metod:
 - zwycięzca podzadania A: Toygger: 3.02 punktów procentowych
 - zwycięzca podzadania B: KRNNT_AB: 2.19 pp.
 - zwycięzca podzadania C: KRNNT_voted: 2.55 pp.

PolEval 2018

POLEVAL 2018

[Home](#) [Tasks](#) [Dates](#) [Results](#) [Prizes](#) [Publication](#) [Organizers](#) [2017](#)



Welcome to PolEval 2018!

PolEval is a [SemEval](#)-inspired evaluation campaign for natural language processing tools for Polish. Submitted tools compete against one another within certain [tasks](#) selected by [organisers](#), using available data and are evaluated according to pre-established procedures.

In 2018 we concentrated on the following NLP-related tasks:

1. [Dependency Parsing](#)
2. [Named Entity Recognition](#)
3. [Language Models](#)

PolEval 2018-related papers were presented at [AI & NLP Workshop Day](#) in Warsaw, Poland (October 19, 2018) and [workshop proceedings](#) were published by the Institute of Computer Science, Polish Academy of Sciences.

Info

October 19, 2018

- PolEval papers were presented at [AI & NLP Workshop Day](#).

August 20, 2018

- The [Results](#), gold standard data and evaluation scripts have been published.

August 8, 2018

- [Test data](#) has been published.

Zadania:

1. Parsowanie zależnościowe
2. Rozpoznawanie jednostek nazewniczych
3. Modelowanie języka

Otrzymaliśmy 24 zgłoszenia od 14 zespołów.

PolEval 2019

- Recognition and normalization of temporal expressions
- Lemmatization of proper names and multi-word phrases
- Entity linking
- Machine translation
- Automatic speech recognition
- Automatic cyberbullying detection

Otrzymaliśmy 34 zgłoszenia!

Uczestnicy instytucjonalni

Academia

- AGH University of Science and Technology
- Institute of Computer Science PAS
- Jagiellonian University
- Kitami Institute of Technology, Japan
- National Information Processing Institute
- Polish-Japanese Academy of Information Technology
- Poznań University of Technology
- Poznań Supercomputing and Networking Center
- University of Łódź
- University of Stuttgart
- University of Wrocław
- Warsaw University of Technology

Business

- Applica.ai
- CVTimeline
- Intive
- n-waves
- Neurosoft
- Plex, Inc.
- Sages
- Samsung
- Samurai Labs
- Sigdelta
- Sigmoidal
- Talkie.ai
- Voicelab

Dlaczego warto wziąć udział?

Nagrody i splendor :)



Dlaczego warto wziąć udział?

Nagrody, splendor i... publikacja!



Organizatorzy

Dotychczasowi organizatorzy zadań

- **Łukasz Kobyliński** - PolEval 2017 Task 1
- **Aleksander Wawer** - PolEval 2017 Task 2, PolEval 2018 Task 2
- **Alina Wróblewska** - PolEval 2018 Task 1
- **Grzegorz Wojdyga** - PolEval 2018 Task 3
- **Jan Kocoń** - PolEval 2019 Task 1
- **Michał Marcińczuk** - PolEval 2019 Task 2
- **Aleksander Smywiński-Pohl** - PolEval 2019 Task 3
- **Krzysztof Wołk** - PolEval 2019 Task 4
- **Danijel Koržinek** - PolEval 2019 Task 5
- **Michał Ptaszyński, Agata Pieciukiewicz, Paweł Dybała** - PolEval 2019 Task 6

Organizatorzy

Dotychczasowi organizatorzy zadań

- **Danijel Koržinek** - PolEval 2020 Task 1
- **Witold Kieraś, Marcin Woliński** - PolEval 2020 Task 2
- **Aleksander Janz, Maciej Piasecki** - PolEval 2020 Task 3
- **Filip Graliński, Anna Wróblewska** - PolEval 2020 Task 4

Organizatorzy konkursu

- **Maciej Ogrodniczuk**
- **Łukasz Kobyliński**

PolEval - co dalej?

- otwarte repozytorium zasobów i metod,
- automatyczna ewaluacja zgłoszeń,
- możliwość zgłaszania mniejszych zadań, niezależnie od corocznych wyzwań, zakończonych prezentacjami na konferencji,
- ranking metod dla poszczególnych zadań,
- współpraca z jeszcze większą liczbą instytucji i firm!



III edycja Warszawa / online 28-29 września

<https://nlpday.pl/>

Konferencja, w czasie której prezentowane są wyniki PoEval.

1. dzień konferencji (28 września, online) - 4 równoległe ścieżki tematyczne
prezentacji: ML/AI, NLP, Big Data oraz PoEval.

2. dzień konferencji (29 września, Warszawa) - 2 równoległe ścieżki
tematyczne **warsztatów praktycznych.**



Co nas czeka podczas III edycji

<https://nlpday.pl/>

- 30+ prezentacji
- 12 praktycznych warsztatów
- 200+ uczestników
- zaproszeni goście
- dyskusje w gronie praktyków AI i NLP
- kontakt z partnerami konferencji
- firmami zatrudniającymi do projektów w tym obszarze

