Abbreviation Disambiguation in Polish Press News Using Encoder-Decoder Models

PolEval 2022/23 Task 2: Abbreviation disambiguation
1st and 2nd solution

Krzysztof Wróbel, Jakub Karbowski, Paweł Lewkowicz

JAGIELLONIAN UNIVERSITY IN KRAKÓW

Enelpol

AGH UNIVERSITY OF KRAKOW

Data

Collections from Polish press news

Assumptions:

- focus on abbreviations of common words or phrases ending with a dot (excluding initials, acronyms, and proper names)
- the context and common knowledge should be sufficient to expand the abbreviation (excluding incomplete or confusing examples)
- the base forms should follow the guidelines of phrase lemmatization from PolEval 2019 Task 2 with some exceptions, such as abbreviations joined with other abbreviations or phrases

Examples of abbreviation disambiguation

Abbr	Context	Inflected form	Base form
t.	a na Dolnym Śląsku - 530-540 zł/<mask>	tonę	tona
gat.	czystą miedź (gat. M1) i mosiądz niklowy (<mask> M55N, CuZn35Ni6Mn4Si0,2)	gatunek	gatunek
j. ukr.	Kier. Natalia Szelest (nagroda burmistrza Węgorzewa) Wertep punktu naucz. <mask> w Baniach Mazurskich	języka ukraińskiego	język ukraiński
ład.	Żyjemy w okresie przejściowym, międzyepoce poprzedzającej nowy <mask>	ład.	ład.

Some cases like *ład.* were not real abbreviations, just end of sequence, and also had to be handled correctly.

Training, validation and test datasets

The training, validation, and test datasets have been provided by the organizers of PolEval.

Splitting datasets:

- a training set called ***train***
- a validation set called ***dev-0*** with expected output
- two test sets with implicit output: ***test-A*** and ***test-B***

Dictionary-based additional data

Additional data were collected (in brackets amount of different abbreviations extracted with their meaning or meanings):

- Morfeusz (443)
- Wiktionary (554)
- SJP (1199)

Example: *woj.*

- *województwo,*
- *wojewoda,*
- *wojewódzki,*
- *wojenny,*
- *wojskowy*

DISTRIBUTION OF THE NUMBER OF MEANINGS FOR ABBREVIATIONS IN THE DATASET SJP

Number of meanings for the abbreviation	Number of abbreviations with a given number of meanings	
1	1051	
2	97	
3	30	
4	10	
5	11	
Number of meanings	**Number of abbreviations**	
1430	1199	**total**

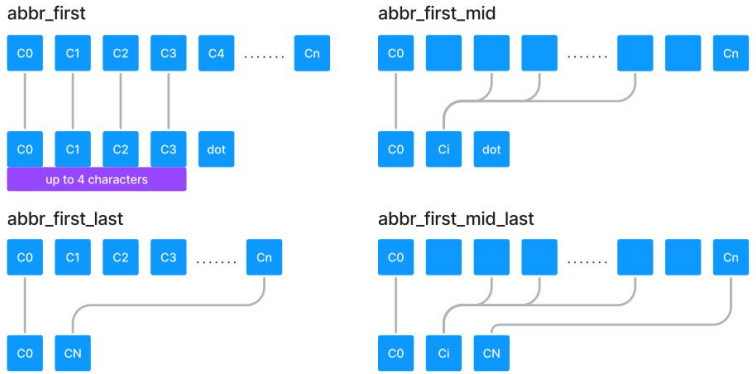
Dictionary-based additional data - example

in.tsv		expected.tsv	
bdb.	pełno dyspozycyjny, zaangażowany, pracowity, komunikatywny, bardzo dobrze znam budowe komputera jak i **<mask>** obsługa komputera jak i programów biurowych Microsoft Office, Open Office itd. Prosze o kontakt	bardzo dobra	bardzo dobry

The process creates also incorrect samples, e.g. “*najlepszym”* is abbreviated to *“db.”*. This dataset lacks non-abbreviation examples.

Synthetic additional data

- Polish Wikipedia
- Sliding window is used to randomly select a context of 140 to 200 characters
- Algorithmic abbreviation:
  - **abbr_first**: `profesor` → `prof.`
    1 to 4 of the first characters.
  - **abbr_first_last**: `profesor` → `pr`
    the first and last characters.
  - **abbr_first_mid**: `profesor` → `pf.`
    the first and one middle character.
  - **abbr_first_mid_last**: `profesor` → `pfr`
    the first, one middle and last characters.



- Base forms of all words from the span before abbreviation are generated with the spaCy `pl_core_news_lg`

Number of samples in datasets

Dataset	Number of samples
train	4 298
dev-0	305
test-A	4 574
test-B	13 965
dictionary-based	1 982
synthetic Wikipedia dataset	~14 000 000

Evaluation

$$Acc = 0.25 \cdot Af + 0.75 \cdot Ab$$

Af - the accuracy of provided expanded forms of abbreviations

Ab - the accuracy of provided base forms of abbreviations

Methods

- a sequence to sequence model using the T5 architecture:
  - Krzysztof Wróbel used the ByT5 model,
  - Jakub Karbowski used the plT5 model.
- input to the transformer encoder is the context with the abbreviation
- the transformer decoder generates both the base and inflected forms
- majority voting separately for inflected and base form

Experiments and results

Krzysztof Wróbel submissions (1/2)

1) The original validation (dev-0) dataset has only 300 samples which is insufficient for tracking scores with a precision of 0.1 percentage points. Therefore, 1000 samples from the training data were moved to the validation set.
2) Input: `Komunistyczny deputowany, <abbrev>b.</abbrev> śledczy Prokuratury Generalnej.`
   Output:
   a) for abbreviations: `były <sep> być`
   b) for non-abbreviations: `b.`
3) Initial experiments using Adafactor as an optimizer showed that the plT5 models performed slightly worse than the ByT5 models.

Krzysztof Wróbel submissions (2/2)

The final submission was created using majority voting on 3 models:

- trained on the training data and dictionary-based additional data using the development data for selecting the best model
- trained on the training data, development data, and dictionary-based additional data with two different seeds

The training parameters were as follows:

- model: byt5-base
- batch size: 16
- gradient accumulation: 16
- epochs: 24
- learning rate: 0.001
- scheduler: linear with warmup 0.1
- optimizer: Adafactor

Krzysztof Wróbel submissions - results

Description	Name	test-A	test-B
train	3	90.78	
train + dict	5	91.32	
train + dev + dict, seed 1	8	92.18	91.69
train + dev + dict, seed 2	9	92.14	91.65
voting (final)	11	**92.76**	**92.01**

Jakub Karbowski submissions

1) Input: `Komunistyczny deputowany, <mask>b.</mask> śledczy Prokuratury Generalnej.`
   Output:
   a) for abbreviations: `były; być`
   b) for non-abbreviations: `b.; b.`
2) Pre-training on synthetic data

The pre-training parameters:

- model: plt5-base
- batch size: 4
- gradient accumulation: 64
- training steps: 3300
- learning rate: 0.0000928
- scheduler: linear with warmup 2000 steps
- optimizer: AdamW
- weight decay: 0.001

Training parameters:

- model: plt5-base (wiki pre-trained)
- batch size: 8
- gradient accumulation: 32
- epochs: 223
- learning rate: 0.000015
- scheduler: linear with warmup 10%
- optimizer: AdamW
- weight decay: 0.0001

PolEval final results

	test-A	test-B
Krzysztof Wróbel	**92.76**	**92.01**
Jakub Karbowski	91.75	91.27
Marek Kozlowski	89.00	88.73
Jakub Pokrywka	65.48	66.25
Rafał Prońko		19.09

Post-competition experiments

- models:
  - ByT5-base,
  - plT5-base
- data:
  - PolEval,
  - dictionary-based,
  - synthetic Wikipedia dataset
- majority voting

Results in different training datasets

	plT5			ByT5		
	dev	**test-A**	**test-B**	**dev**	**test-A**	**test-B**
train	91.80	90.76	90.06	94.10	92.10	91.73
wiki-train	91.39	**91.76**	91.32	93.61	92.46	92.53
train-dict	91.31	91.21	90.44	94.34	92.30	92.20
wiki-train-dict	91.31	91.64	**91.37**	93.44	**92.71**	**92.92**

Majority voting among 1 to N models

	plT5		ByT5	
Models	**test-A**	**test-B**	**test-A**	**test-B**
1	91.72	90.96	92.71	92.92
1-2	91.74	91.22	92.65	92.80
1-3	91.91	91.42	93.00	93.06
1-4	91.88	91.45	**93.33**	93.15
1-5	92.03	91.57	93.25	**93.27**
1-6	92.00	91.59	93.20	93.19
1-7	91.99	91.56	93.16	93.14
1-8	92.05	91.57	93.12	93.11
1-9	92.10	**91.62**	93.12	93.18
1-10	**92.12**	91.59	93.13	93.17

Links

Source code: https://github.com/Carbon225/poleval-2022-abbr
Model ByT5: https://huggingface.co/carbon225/byt5-abbreviations-pl
Model plT5: https://huggingface.co/carbon225/plt5-abbreviations-pl

Contact:

- Krzysztof Wróbel: https://www.linkedin.com/in/wrobelkrzysztof/
- Jakub Karbowski: https://www.linkedin.com/in/jkar/
- Paweł Lewkowicz: https://www.linkedin.com/in/pawel-lewkowicz/

Errors by the best model (1/2)

input	expected		predicted	
m. możliwy grad. Przewidywana wysokość opadów w burzach od 10 mm do 15 mm, w górach do 20 **<mask>** Temperatura maksymalna od 15 st.C w rejonie Zatoki Gdańskiej, 21 st.C na Suwalszczyźnie	metrów	metr	milimetrów	milimetr
p. procentowych (do 100 proc.) przy dopłatach dla 1 osoby i o 10 punktów procentowych (do 40 **<mask>** proc.) dla każdej kolejnej osoby w gospodarstwie domowym najemcy. Ważne zmiany	pikseli	piksel	punktów	punkt
róż. Pazdanowi żona, Dominika, z wyraźnym onieśmieleniem przyjęła bukiet biało-czerwonych **<mask>** Piłkarz zrewanżował się własną reprezentacyjną koszulką. Potem nastąpiła seria	róż	róż	róż.	róż.
o. (są) do indywidualnego uzgodnienia z władzami' uczelni', czyli i tak daje ludziom **<mask>** Rydzyka zupełną dowolność. Wyższa Szkoła Kultury Społecznej i Medialnej w Toruniu	ojciec	ojciec	ojca	ojciec
cm. 38-letniego Granta. Amerykański pięściarz mierzy 201 cm, natomiast 33-letni Adamek - 187 **<mask>** Polak znacznie przegrywa z Grantem również pod względem zasięgu ramion. Dla Adamka,	centrymetrów	centrymetr	centymetrów	centymetr
p. C-331/94, Komisja p. Grecji, ECLI:EU:C:1996:211, pkt 10; C-111/05, Aktiebolaget NN **<mask>** Skatteverket, ECLI:EU:C:2007:195, pkt 55-58. [10] Z. Knypl, Polskie obszary morskie,	przeciwko	przeciwko	przeciw; przeciwko	przeciw; przeciwko
p. również osoby bez obywatelstwa, którzy publicznie znieważają osoby, wymienione w **<mask>** 1 Ustawy, przeszkadzają w realizacji praw osób walczących o niezależność Ukrainy	punkcie; paragrafie	punkt; paragraf	paragrafie	paragraf
r. zawodników urodzonych w 2001 roku i młodszych. Wcześniej we Włocławku walczyli piłkarze **<mask>** 1997. Tym razem nie będzie to turniej międzynarodowy, ponieważ cztery zaproszone	rocznik	rocznik	rocznika	rocznik

Errors by the best model (2/2)

m. Przemysłowym. Tramwaje linii 3, 23, 33>pl. Jana Pawła II skierowano objazdem przez **<mask>** Sikorskiego, Dubois, Nowy Świat. Tramwaje linii 10 i 20>pl. Jana Pawła II skierowano	most	most	mosty	most
p. konkursu. Oferty należy składać do 5 grudnia (nie decyduje data stempla pocztowego) w **<mask>** 223 w Starostwie Powiatowym. Obecnie placówkę prowadzi Zgromadzenie Sióstr św.	pokoju	pokój	pokój	pokój
s. William P. Young, Chata, tłum. A. Reszka, Wydawnictwo Nowa Proza, Warszawa 2009, **<mask>** 281. Ponad 6.000.000 sprzedanych egzemplarzy robi wrażenie na każdym, kto ma styczność	stron	strona	strona	strona
m. małopolskiego (i na 389. miejscu w Polsce) oraz II LO im. Tytusa Chałubińskiego na 70 **<mask>** (poza pierwszą 500 najlepszych liceów w Polsce). W ubiegłym roku „Kościuszko" był	miejscu	miejsce	metrach	metr
f. powierzchniową. Jak piszą Allaud L.A. i Martin M. bracia Schlumberger'owie przekonali **<mask>** Royal Dutch Shell, po powtórzeniu pomiarów i sprawdzeniu ich wiarygodności, że ta	firmę	firma	firmie	firma
m. dąbrowski) - rzeka Wisła o 69 cm m. Szczucin (pow. dąbrowski) - rzeka Szreniawa o 8 cm **<mask>** Biskupice (powiat miechowski) - rzeka Wisła o 145 cm Pustynia (powiat oświęcimski) -	miejscowość	miejscowość	miasto	miasto
ub.roku. krajowymi. Kupili ich w pierwszym kwartale o ponad 9 mld zł więcej niż na koniec **<mask>** – To pokazuje, że zagranica nie ucieka od naszego długu. Równoległy spadek nierezydentów	ubiegłego roku.	ubiegły roku.	ubiegłego roku	ubiegły rok
zew. wreszcie zbliża się ten dzień / wielki dreszcz emocji w nas / i wolności poczuj **<mask>** / Euro w barwach szczęścia jest / więc ramiona w górę wznieś / a dopóki piłka w grze	zew.	zew.	zewnętrznie	zewnętrznie