Results of the PolEval 2019 Shared Task 2

Lemmatization of Proper Names and Multi-word Phrases

Michał Marcińczuk, Tomasz Bernaś

Department of Computational Intelligence
Faculty of Computer Science and Management
Wroclaw University of Science and Techology
michal.marcinczuk@pwr.edu.pl

May 31, 2019

Politechnika
Wrocławska

Task 2 description

What is lemmatization?

Lemmatization consists in generating a dictionary form of a phrase.

Scope

- Proper names and multi-word phrases,
- According to the KPWr guidelines [1].

Phrase	**Lemma**
wojny trzydziestoletniej	wojna trzydziestoletnia
portu lotniczego Wnukowo	portu lotniczego Wnukowo
powiecie sochaczewskim	powiat sochaczewski
Chałasińskiej	Chałasińska

[1] Marcin Oleksy, Adam Radziszewski, and Jan Wieczorek. *KPWr annotation guidelines – phrase lemmatization*. CLARIN-PL digital repository. 2018. URL: http://hdl.handle.net/11321/591

Challenges

Multi-word phrases

- it is not a simple concatenation of word lemmas, ex. *piwnica domu*, not *piwnica dom*,
- dealing with plural forms, ex. *skoki narciarskie*,

Proper names

- many out-of-vocabulary words,
- some foreign names are subject to inflection and some are not,
- depends on the proper name category, i.e *Słowackiego* as a person name and a street name has different lemmas,
- capitalization does matter.

Existing approaches

- Piskorski, Kupść, and Sydow (2007) — rules and string distance metrics for person names lemmatization,
- Degórski (2012) — a rule-based method for lemmatization nominal syntactic groups utilized a shallow grammar,
- Radziszewski (2013) — noun phrase lemmatization based on transformations learned from a training data using a machine learning technique (CRF),
- Małyszko et al. (2015) — automatically retrieved a list of lemmatization rules for multi-word units based on a corpus analysis,
- Marcińczuk (2017) — used a set of language resources, manually crafted rules and heuristics to lemmatize multi-word expressions and proper names.

PolEval Datasets

Dataset	Documents	Phrases	Source
training	1629	24 000+	KPWr [7]
tuning	200	1 145	CEN [8]
testing	99	1 997	PST [9]

Table: Size of the datasets

[7] Bartosz Broda et al. "KPWr: Towards a Free Corpus of Polish". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Ed. by Nicoletta Calzolari et al. Istanbul, Turkey: European Language Resource Association, 2012

[8] Michał Marcińczuk. *CEN*. CLARIN-PL digital repository. 2007. URL: http://hdl.handle.net/11321/6

[9] Marcin Oleksy et al. *Polish Spatial Texts 1.0*. CLARIN-PL digital repository. 2018. URL: http://hdl.handle.net/11321/543

Data format

document.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<document id="00107258">
(...)
Jarosław Kaczyński stwierdził także, że jeszcze wczoraj była szansa na porozumienie
z <phrase id="486121">Platformą Obywatelską</phrase>. Jego zdaniem partia
<phrase id="318064">Donalda <phrase id="318066">Tuska</phrase></phrase>
stawiała warunki dominacji w rządzie. Prezes PiS obarczył winą PO za
zamieszanie w obecnym <phrase id="318069">Sejmie</phrase>
i brak rządu o stałym poparciu.
</document>
```

index.tsv

```
486119 00107258 Prawa i Sprawiedliwości  Prawo i Sprawiedliwość
318048 00107258 Kaczyński                Kaczyński
318041 00107258 Jarosław Kaczyński       Jarosław Kaczyński
318064 00107258 Donalda Tuska            Donald Tusk
318049 00107258 Sygnałach Dnia           Sygnały Dnia
318066 00107258 Tuska                    Tusk
318044 00107258 Samoobroną               Samoobrona
318069 00107258 Sejmie                   Sejm
318059 00107258 paktu stabilizacyjnego   pakt stabilizacyjny
318050 00107258 Samoobroną               Samoobrona
318058 00107258 paktu stabilizacyjnego   pakt stabilizacyjny
318051 00107258 Ligą Polskich Rodzin     Liga Polskich Rodzin
486121 00107258 Platformą Obywatelską    Platforma Obywatelska
```

Evaluation

$$Score = 0.2 * Acc_{CS} + 0.8 * Acc_{CI} \quad (1)$$

Acc refers to the accuracy, i.e. a ratio of the correctly lemmatized phrases to all phrases subjected to lemmatization.

The accuracy was calculated in two variants:

- *case sensitive* (Acc_{CS}) and
- *case insensitive* (Acc_{CI}).

Participating systems

zbronk.nlp.studio — the solution is based on a set of lemmatization heuristics utilizing different language tools and resources, including: Morfeusz morphological analyzer [10], a proprietary quasi-dependency parser, NKJP corpus [11], a large corpus of unnannotated texts (4 billion words) and Multisłownik.

PolEval2019-lemmatization — tags represent transformations that need to be performed on each word. Documents processed with UDPipe and COMBO tokenizer. The system architecture is as follows: first, the data and the dependency parser features are fed to two separate embedding layers. The embeddings are concatenated and fed to a bidirectional LSTM layer. Calculated features are then truncated to match the lemmatized phrase and fed to a CRF layer. Operations represented by the predicted tags are performed using the Morfeusz morphological analyzer [10].

---

[10] Witold Kieraś and Marcin Woliński. "Morfeusz 2 – analizator i˜generator fleksyjny dla języka polskiego". In: *Język Polski* XCVII.1 (2017), pp. 75–83

[11] Adam Przepiórkowski et al. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, 2012, p. 331. ISBN: 9788301167004

Results

System name	Variant	AccCS	AccSI	Score
zbronk.nlp.studio	-	84.78	88.13	**87.46**
PolEval2019-lemmatization	new1	72.46	75.46	74.86
PolEval2019-lemmatization	model3	68.85	71.71	71.14

Table: Results of lemmatization task

Thank you for your attention