Evaluating Natural Language Processing tools for Polish during PolEval 2019

Łukasz Kobyliński[1] [0000−0003−2462−0020],
Maciej Ogrodniczuk[1] [0000−0002−3467−9424],
Jan Kocoń[2] [0000−0002−7665−6896],
Michał Marcińczuk[2] [0000−0002−3269−6378],
Aleksander Smywiński-Pohl[3] [0000−0001−6684−0748],
Krzysztof Wołk[5] [0000−0001−5030−334X],
Danijel Koržinek[5] [0000−0002−2916−4856],
Michal Ptaszynski[4] [0000−0002−1910−9183],
Agata Pieciukiewicz[5] [0000−0003−0965−4070], and
Paweł Dybała[6] [0000−0003−2823−7838]

[1] Institute of Computer Science, Polish Academy of Sciences
[2] Wrocław University of Science and Technology
[3] AGH University of Science and Technology
[4] Kitami Institute of Technology
[5] Polish-Japanese Academy of Information Technology
[6] Jagiellonian University in Kraków

Abstract. 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 organizers, using available data and are evaluated according to pre-established procedures. It is organized since 2017 and each year the winning systems become the state-of-the-art in Polish language processing in the respective tasks. In 2019 we have organized six different tasks, creating an even greater opportunity for NLP researchers to evaluate their systems in an objective manner.

Keywords: temporal expressions · lemmatization · entity linking · machine translation · automatic speech recognition · cyberbullying detection.

1 Introduction

PolEval[7] is an initiative started in 2017 by the Linguistic Engineering Group at the Institute of Computer Science, Polish Academy of Sciences, aiming at increasing quality of natural language tools for Polish by organizing a testing ground where interested parties could try their new solutions attempting to beat state-of-the-art. This could be achieved only

[7] http://poleval.pl

by setting up formal evaluation procedures according to widely accepted metrics and using newly collected data sets.
The idea was simple yet it attracted a lot of attention: in first two editions of the contest [35, 10, 22] we received over 40 submissions to 8 tasks and subtasks. In 2019 the number of tasks grew to six, expanding to processing multilingual and multimodal data. Below we describe each of the tasks that have been announced for PolEval 2019[8].

2 Task 1: Recognition and normalization of temporal expressions

2.1 Problem statement

Temporal expressions (henceforth *timexes*) tell us *when* something happens, *how long* something lasts, or *how often* something occurs. The correct interpretation of a timex often involves knowing the context. Usually, people are aware of their location in time, i.e., they know what day, month and year it is, and whether it is the beginning or the end of week or month. Therefore, they refer to specific dates, using incomplete expressions such as: *12 November*, *Thursday*, *the following week*, *after three days*. The temporal context is often necessary to determine to which specific date and time timexes refer. These examples do not exhaust the complexity of the problem of recognizing timexes.
TimeML [30] is a markup language for describing timexes that has been adapted to many languages. PLIMEX [11] is a specification for the description of Polish timexes. It is based on TIMEX3 used in TimeML. Classes proposed in TimeML are adapted, namely: *date*, *time*, *duration*, *set*.

2.2 Task description

The aim of this task is to advance research on processing of temporal expressions, which are used in other NLP applications like question answering, summarization, textual entailment, document classification, etc. This task follows on from previous TempEval events organized for evaluating time expressions for English and Spanish like SemEval-2013 [32]. This time we provide corpus of Polish documents fully annotated with temporal expressions. The annotation consists of boundaries, classes and normalized values of temporal expressions. The annotation for Polish texts is based on modified version of original TIMEX3 annotation guidelines[9] at the level of annotating boundaries/types[10] and local/global normalization[11] [11].

[8] http://2019.poleval.pl/

[9] https://catalog.ldc.upenn.edu/docs/LDC2006T08/timeml_annguide_1.2.1.pdf

[10] http://poleval.pl/task1/plimex_annotation.pdf

[11] http://poleval.pl/task1/plimex_normalisation.pdf

2.3 Training data

The training dataset contains 1500 documents from KPWr corpus. Each document is XML file with the given annotations, e.g.:

```
<DOCID>344245.xml</DOCID>
<DCT><TIMEX3 tid="t0" functionInDocument="CREATION_TIME"
type="DATE" value="2006-12-16"></TIMEX3></DCT>
<TEXT>
<TIMEX3 tid="t1" type="DATE" value="2006-12-16">Dziś
</TIMEX3> Creative Commons obchodzi czwarte urodziny -
przedsięwzięcie ruszyło dokładnie <TIMEX3 tid="t2"
type="DATE" value="2002-12-16">16 grudnia 2002</TIMEX3>
w San Francisco. (...) Z kolei w <TIMEX3 tid="t4"
type="DATE" value="2006-12-18">poniedziałek</TIMEX3>
ogłoszone zostaną wyniki głosowanie na najlepsze blogi.
W ciągu <TIMEX3 tid="t5" type="DURATION" value="P8D">8 dni
</TIMEX3> internauci oddali ponad pół miliona głosów.
Z najnowszego raportu Gartnera wynika, że w <TIMEX3 tid="t6"
type="DATE" value="2007">przyszłym roku</TIMEX3> blogosfera
rozrośnie się do rekordowego rozmiaru 100 milionów blogów.
</TEXT>
```

2.4 Evaluation

We utilize the same evaluation procedure as described in article [32]. We need to evaluate:

1. How many entities are correctly identified,
2. If the extents for the entities are correctly identified,
3. How many entity attributes are correctly identified.

We use classical precision (P), recall (R) and F1-score (F1 – a harmonic mean of P and R) for the recognition.

(1) We evaluate our entities using the entity-based evaluation with the equations below:

$$P = \frac{\left|\mathrm{Sys}_{\mathrm{entity}} \cap \mathrm{Ref}_{\mathrm{entity}}\right|}{\left|\mathrm{Sys}_{\mathrm{entity}}\right|} \qquad R = \frac{\left|\mathrm{Sys}_{\mathrm{entity}} \cap \mathrm{Ref}_{\mathrm{entity}}\right|}{\left|\mathrm{Ref}_{\mathrm{entity}}\right|}$$

where, $\mathrm{Sys}_{\mathrm{entity}}$ contains the entities extracted by the system that we want to evaluate, and $\mathrm{Ref}_{\mathrm{entity}}$ contains the entities from the reference annotation that are being compared.

(2) We compare our entities with both strict match and relaxed match. When there is an exact match between the system entity and gold entity then we call it strict match, e.g. *16 grudnia 2002* vs *16 grudnia 2002*. When there is an overlap between the system entity and gold entity then we call it *relaxed match*, e.g. *16 grudnia 2002* vs *2002*. When there is a relaxed match, we compare the attribute values.

(3) We evaluate our entity attributes using the *attribute F1-score*, which captures how well the system identified both the entity and attribute together:

$$\mathrm{attrP} = \frac{\left|\forall x | x \in (\mathrm{Sys}_{\mathrm{entity}} \cap \mathrm{Ref}_{\mathrm{entity}}) \wedge \mathrm{Sys}_{\mathrm{attr}}(x) = \mathrm{Ref}_{\mathrm{attr}}(x)\right|}{\left|\mathrm{Sys}_{\mathrm{entity}}\right|}$$

$$\mathrm{attrR} = \frac{\left|\forall x | x \in (\mathrm{Sys}_{\mathrm{entity}} \cap \mathrm{Ref}_{\mathrm{entity}}) \wedge \mathrm{Sys}_{\mathrm{attr}}(x) = \mathrm{Ref}_{\mathrm{attr}}(x)\right|}{\left|\mathrm{Ref}_{\mathrm{entity}}\right|}$$

We measure P, R, F1 for both strict and relaxed match and relaxed F1 for value and type attributes. The most important metric is *relaxed F1 value*.

2.5 Results

The best result in the main competition (excluding a baseline system provided by organizers) was achieved by Alium team with its Alium solution. Alium solution is an engine to process texts in natural language and produce results according to rules that define its behaviour. Alium can work either on single words or on triples – *word, lemma, morphosyntactic tag*. Words are additionally masked, so that Alium can work on parts of words as well. More details can be found in [12].

3 Task 2: Lemmatization of proper names and multi-word phrases

3.1 Problem statement

Lemmatization relies on generating a dictionary form of a phrase. In our task we focus on lemmatization of proper names and multi-word phrases. For example, the following phrases *radę nadzorczą*, *radzie nadzorczej*, *radą nadzorczą* which are inflected forms of *board of directors* should be lemmatized to *rada nadzorcza*. Both, lemmatization of multi-word common noun phrases and named entities are challenging because Polish is a highly inflectional language and a single expression can have several inflected forms.

The difficulty of multi-word phrase lemmatization is due to the fact that the expected lemma is not a simple concatenation of base forms for each word in the phrase [16]. In most cases only the head of the phrase is changed to a nominative form and the remaining word, which are the modifiers of the head, should remain in a specific case. For example in the phrase *piwnicy domu* (Eng. *house basement*) only the first word should be changed to their nominative form while the second word should remain in the genitive form, i.e. *piwnica domu*. A simple concatenation of tokens' base forms would produce a phrase *piwnica dom* which is not correct.

In the case of named entities the following aspects make the lemmatization difficult:

1. Proper names may contain words which are not present in the morphological dictionaries. Thus, dictionary-based methods are insufficient.
2. Some foreign proper names are subject to inflection and some are not.
3. The same text form of a proper name might have different lemmas depending on their semantic category. For example *Słowackiego* (a person last name in genitive or accusative) should be lemmatized to *Słowacki* in case of person name and to *Słowackiego* in case of street name.
4. Capitalization does matter. For example a country name *Polska* (Eng. *Poland*) should be lemmatized to *Polska* but not to *polska*.

3.2 Task description

The task consists in developing a system for lemmatization of proper names and multi-word phrases. The generated lemmas should follow the KPWr guidelines [24]. The system should generate a lemma for given set of phrases with regards to the context, in which the phrase appears.

3.3 Training data

The training data consists of 1629 documents from the KPWr corpus [2] with more than 24k annotated and lemmatized phrases. The documents are plain texts with in-line tags indicating the phrases, i.e.
`<phrase id="40465">Madrycie</phrase>`.
All the phrases with their lemmas are listed in a single file, which has the following format:

```
[...]
20250 100619 kampanii wyborczych
kampanie wyborcze
40465  100619 Madrycie          Madryt
40464  100619 Warszawie         Warszawa
40497  100619 Dworcu Centralnym  Dworzec Centralny
40463  100619 Warszawie         Warszawa
[...]
```

3.4 Evaluation

The score of system responses will be calculated using the following formula:

$$Score = 0.2 * Acc_{CS} + 0.8 * Acc_{CI} \tag{1}$$

Acc refers to the accuracy, i.e. a ratio of the correctly lemmatized phrases to all phrases subjected to lemmatization.
The accuracy will be calculated in two variants: *case sensitive* (Acc_{CS}) and *case insensitive* (Acc_{CI}). In the case insensitive evaluation the lemmas will be converted to lower cases.

4 Task 3: Entity linking

4.1 Problem statement

Entity linking [18, 29] covers the identification of mentions of entities from a knowledge base (KB) in Polish texts. In this task as the reference KB we use WikiData (WD)[12], an offspring of Wikipedia – a knowledge base, that unifies structured data available in various editions of Wikipedia and links them to external data sources and knowledge bases.

[12] https://www.wikidata.org

Thus making a link from a text to WD allows for reaching a large body of structured facts including: the semantic type of the object, its multilingual labels, dates of birth and death for people, the number of citizens for cities and countries, the number of students for universities and many, many more. The identification of the entities is focused on the disambiguation of a phrase against WD. The scope of the phrase is provided in the test data, so the task boils down to the selection of exactly one entry for each linked phrase.

4.2 Task description

The following text:

> Zaginieni 11-latkowie w środę rano wyszli z domów do szkoły w **Nowym Targu**, gdzie przebywali do godziny 12:00. Jak informuje "**Tygodnik Podhalański**", 11-letni Ivan już się odnalazł, ale los Mariusza Gajdy wciąż jest nieznany. Source: gazeta.pl

has 2 entity mentions:

1. Nowym Targu[13]
2. Tygodnik Podhalański[14]

Even though there are more mentions that have their corresponding entries in WD (such as "środa", "dom", "12:00", etc.) we restrict the set of entities to a closed group of WD types: names of countries, cities, people, occupations, organisms, tools, constructions, etc. (with important exclusion of times and dates). The full list of entity types is available for download[15]. It should be noted that names such as "Ivan" and "Mariusz Gajda" should not be recognized, since they lack corresponding entries in WD.

The task is similar to Named Entity Recognition (NER), with the important difference that in EL the set of entities is closed. To some extent EL is also similar to Word Sense Disambiguation (WSD), since mentions are ambiguous between competing entities.

In this task we have decided to ignore nested mentions of entities, so names such as "Zespół Szkół Łączności im. Obrońców Poczty Polskiej w Gdańsku, w Krakowie", which has an entry in WD, should be treated as an atomic linguistic unit, even though there are many entities that have their corresponding WD entries (such as *Poczta Polska w Gdańsku*, *Gdańsk*, *Kraków*). Also the algorithm is required to identify all mentions of the entity in the given document, even if they are exactly same as the previous mentions.

4.3 Training data

The most common training data used in EL is Wikipedia itself. Even though it wasn't designed as a reference corpus for that task, the structure of internal links serves as a good source for training and testing

[13] `https://www.wikidata.org/wiki/Q231593`

[14] `https://www.wikidata.org/wiki/Q9363509`

[15] `http://poleval.pl/task3/entity-types.tsv`

data, since the number of links inside Wikipedia is counted in millions. The important difference between the Wikipedia links and EL to WD is the fact that the titles of the Wikipedia articles evolve, while the WD identifiers remain constant.
As the training data we have provided a complete text of Wikipedia with morphosyntactic data provided by KRNNT tagger [37], categorization of articles into Wikipedia categories and WD types, Wikipedia redirections and internal links.

4.4 Evaluation

The number of correctly linked mentions divided by the total number of mentions to be identified is used as the evaluation measure. If the system does not provide an answer for a phrase, the result is treated as an invalid link.

5 Task 4: Machine translation

5.1 Problem statement

Machine translation is a computer translation of text without human involvement. Machine translation, a pioneer of the 1950s, is also known as machine translation or instant translation.
Currently, there are three most common types of machine translation systems: rule-based, statistical and neural.

- Rule-based systems use a combination of grammar and language rules, as well as a dictionary of common words. Professional dictionaries are created to focus on a particular industry or discipline. Rule-based systems generally provide consistent translations in accurate terms when trained in specialized dictionaries. [7].
- The statistical systems do not know the language rules. Instead, they learn to translate by analyzing large amounts of data for each language pair. Statistical systems usually provide smoother but inconsistent translations [13].
- Neural Machine Translation (NMT) is a new approach that uses machines to translate learning through large neural networks. This approach is becoming increasingly popular with MT researchers and developers as trained NMT systems are beginning to show better translation performance in many language pairs compared to phrase-based statistical approaches. [36].

5.2 Task description

The task was to train the machine translation system as good as possible using any technology with limited text resources. The contest was held in two languages. There were a few languages, more popular English-Polish (Polish direction) and low-resourced Russian-Polish (both directions).

5.3 Training data

As the training data set, we have prepared a set of bi-lingual corpora aligned at the sentence level. The corpora were saved in UTF-8 encoding as plain text, one language per file. We divided the corpora as in-domain data and out-domain data. Using any other data was not permitted. The in-domain data was rather hard to translate because of its topic diversity. In-domain data were lectures on different topics. As out of domain data we accepted any corpus from `http://opus.nlpl.eu` project. Any kind of automatic pre- or post- processing was also accepted. The in-domain corpora statistics are given in Table 1.

	No. of segments		No. of unique tokens			
	test	train	test		train	
			input	output	input	output
EN to PL	10,000	129,254	9,834	16,978	49,324	100,119
PL to RU	3,000	20,000	6,519	7,249	31,534	32,491
RU to PL	3,000	20,000	6,640	6,385	32,491	31,534

Table 1. Task 4 corpora statistics.

5.4 Evaluation and Results

The participants were asked to translate with their systems test files and submit the results of the translations. The translated files were supposed to be aligned at the sentence level with the input (test) files. Submissions that were not aligned were not accepted. If any pre- or post- processing was needed for the systems, it was supposed to be done automatically with scripts. Any kind of human input into test files was strongly prohibited. The evaluation itself was done with four main automatic metrics widely used in machine translation:

- BLEU [25]
- NIST [4]
- TER [31]
- METEOR [1]

As part of the evaluation preparation we prepared baseline translation systems. For this purpose we used out of the box and state of the art ModernMT machine translation system. We did not do any kind of data pre- or post-processing nor any system adaptation. We simply used our data with default ModernMT settings. Table 2 contains summary of the results limited to BLEU metric for clarity. Full results are available in [21].

System name	BLUE score		
	EN-PL	PL-RU	RU-PL
Baseline	16.29	12.71	11.45
SRPOL	28.23	n/a	n/a
DeepIf (in-domain)	4.92	5.38	5.51
SIMPLE_SYSTEMS	0.94	0.69	0.57

Table 2. Task 4: Machine Translation Results

6 Task 5: Automatic speech recognition

6.1 Problem statement

Automatic speech recognition (ASR) is the problem of converting an audio recording of speech into its textual representation. For the purpose of this evaluation campaign, the transcription is considered simply as a sequence of words conveying the contents of the recorded speech. This task is very common, has many practical uses in both commercial and non-commercial setting and there are many evaluation campaigns associated with it, e.g. [9, 34, 6]. The significance of this particular competition is the choice of language. To our knowledge, this is the first strictly Polish evaluation campaign of ASR.

$$w* = \arg\max_i P(w_i|O) = \arg\max_i P(O|w_i) \cdot P(w_i) \qquad (2)$$

As shown in formula 2, ASR is usually solved using a probabilistic framework of determining the most likely sequence of words w_i, given a sequence of acoustic observation O of data. This equation is furthermore broken into two essential components by Bayesian inference: the estimation of the acoustic-phonetic realization $P(O|w_i)$, also known as acoustic modeling (AM), and the probability of word sequence realization $P(w_i)$, also known as language modeling (LM):

Each of these steps requires solving a wide range of sub-problems relying on the knowledge of several disciplines, including signal processing, phonetics, natural language processing and machine learning.

A very common framework for solving this problem is the Hidden Markov Model [38]. Currently, this concept was expanded to a more useful implementation based on Weighted Finite-State Transducers [17]. Some of the most recent solutions try to bypass the individual sub-steps by modeling the whole process in a single end-to-end model [8], however knowledge of the mentioned disciplines is still essential to successfully perform the tuning of such a solution.

6.2 Task description

The task for this evaluation campaign is very simple to define and evaluate: given a set of audio files, create a transcription of each file. For

simplicity, only the word sequence is taken into account - capitalization and punctuation is ignored. Also, the text is evaluated in its normalized form, i.e. numbers and abbreviations need to be presented as individual words.
The domain of the competition is parliamentary proceedings. This domain was chosen for several reasons. The data is publicly available and free for use by any commercial or non-commercial entity. Given the significance of the parliamentary proceedings, there is a wide variety of extra domain material that can be found elsewhere, especially in the media. The task is also not too challenging, compared to some other domains, because of the cleanliness and predictability of the acoustic environment and the speakers.

6.3 Training data

The competition is organized into two categories: fixed and open. For the fixed competition, a collection of training data is provided as follows:

- Clarin-PL speech corpus [14]
- PELCRA parliamentary corpus [26]
- A collection of 97 hours of parliamentary speeches published on the ClarinPL website [15]
- Polish Parliamentary Corpus for language modeling [19, 20, 23]

For those who wish to participate in the competition using a system that was trained on more data, including that which is unavailable to the public, they have to participate as part of the open competition. The only limitation was the ban of use of any data from the Polish Parliament and Polish Senate websites after January 1st 2019.

6.4 Evaluation

Audio is encoded as uncompressed, linearly encoded 16-bit per sample, 16 kHz sampling frequency, mono signals encapsulated in WAV formatted files. The origin of the files is from freely available public streams, so some encoding is present in the data, but the contestants do not have to decompress it on their own. The contestants have a limited time to process these files and provide the transcriptions as separate UTF-8 encoded text documents. The files are evaluated using the standard Word Error Rate metric as computed by the commonly used NIST Sclite package [5].

6.5 Results

The last two entries in Table 3 were the baselines prepared by the competition organizer, with full knowledge of the test data domain. The winner of the competition was the system code named GOLEM with the score of 12.8% WER.

System name	WER%	Competition type
GOLEM	12.8	closed
ARM-1	26.4	open
SGMM2	41.3	open
tri2a	41.8	open
clarin-pl/sejm	11.8	closed
clarin-pl/studio	30.9	open

Table 3. Task 5: Automatic speech recognition results

7 Task 6: Automatic cyberbullying detection

7.1 Problem statement

Although the problem of humiliating and slandering people through the Internet existed almost as long as communication via the Internet between people, the appearance of new devices, such as smartphones and tablet computers, which allow using this medium not only at home, work or school but also in motion, has further exacerbated the problem. Especially recent decade, during which Social Networking Services (SNS), such as Facebook and Twitter, rapidly grew in popularity, has brought to light the problem of unethical behaviors in Internet environments, which since then has been greatly impairing public mental health in adults and, for the most, younger users and children. The problem in question, called cyberbullying (CB), is defined as exploitation of open online means of communication, such as Internet forum boards, or SNS, to convey harmful and disturbing information about private individuals, often children and students.

To deal with the problem, researchers around the world have started studying the problem of cyberbullying with a goal to automatically detect Internet entries containing harmful information, and report them to SNS service providers for further analysis and deletion. After ten years of research [28], a sufficient knowledge base on this problem has been collected for languages of well-developed countries, such as the US, or Japan. Unfortunately, still close to nothing in this matter has been done for the Polish language. With this task, we aim at filling this gap.

7.2 Task description

In this pilot task, the contestants determine whether an Internet entry is classifiable as part of cyberbullying narration or not. The entries contain tweets collected from openly available Twitter discussions. Since much of the problem of automatic cyberbullying detection often relies on feature selection and feature engineering [27], the tweets are provided as such,

with minimal preprocessing. The preprocessing, if used, is applied mostly for cases when information about a private person is revealed to the public. In such situations the revealed information is masked not to harm the person in the process.
The goal of the contestants is to classify the tweets into cyberbullying/harmful and non-cyberbullying/non-harmful with the highest possible Precision, Recall, balanced F-score and Accuracy. There are two sub-tasks.

Task 6-1: Harmful vs non-harmful: In this task, the participants are to distinguish between normal/non-harmful tweets (class: 0) and tweets that contain any kind of harmful information (class: 1). This includes cyberbullying, hate speech and related phenomena.

Task 6-2: Type of harmfulness: In this task, the participants shall distinguish between three classes of tweets: 0 (non-harmful), 1 (cyberbullying), 2 (hate-speech). There are various definitions of both cyberbullying and hate-speech, some of them even putting those two phenomena in the same group. The specific conditions on which we based our annotations for both cyberbullying and hate-speech, have been worked out during ten years of research [28]. However, the main and definitive condition to distinguish the two is whether the harmful action is addressed towards a private person(s) (cyberbullying), or a public person/entity/larger group (hate-speech).

7.3 Training data

To collect the data, we used the Standard Twitter API[16]. The script for data collection was written in Python and was then used to download tweets from 19 Polish Twitter accounts. Those accounts were chosen as the most popular Polish Twitter accounts in the year 2017[17]: @tvn24, @MTVPolska, @lewy_official, @sikorskiradek, @Pontifex_pl, @PR24_pl, @donaldtusk, @BoniekZibi, @NewsweekPolska, @tvp_info, @pisorgpl, @AndrzejDuda, @lis_tomasz, @K_Stanowski, @R_A_Ziemkiewicz, @Platforma_org, @RyszardPetru, @RadioMaryja, @rzeczpospolita.
In addition to tweets from those accounts, we also collected answers to any tweets from the accounts mentioned above from past 7 days. In total, we have received over 101 thousand tweets from 22,687 accounts (as identified by screen_name property in the Twitter API). Using bash random function 10 accounts were randomly selected to become the starting point for further work. Using the same script as before, we downloaded tweets from these 10 accounts and all answers to their tweets that we were able to find using the Twitter Search API Using this procedure we have selected 23,223 tweets from Polish accounts for further analysis.

[16] `https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html`

[17] `https://www.sotrender.com/blog/pl/2018/01/twitter-w-polsce-2017-infografika/`

At first, we randomized the order of tweets in the dataset to get rid of any consecutive tweets from the same account. Next, we got rid of all tweets containing URLs. This was done due to the fact that URLs often take space and limit the contents of the tweets, which in practice often resulted in tweets being cut in the middle of the sentence or with a large number of *ad hoc* abbreviations. Next, we removed from the data tweets that were perfect duplicates. Tweets consisting only of atmarks(@) or hashtags(#) were also deleted. Finally, we removed tweets with less than five words and those written in languages other than polish. This left us with 11,041 tweets, out of which we used 1,000 tweets as test data and the rest (10,041) as training data.

7.4 Evaluation

The scoring for the first task is done based on standard Precision (P), Recall (R), Balanced F-score (F1) and Accuracy (A), on the basis of the numbers of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), according to the below equations (3-6). In choosing the winners we look primarily at the balanced F-score. However, in the case of equal F-score results for two or more teams, the team with higher Accuracy will be chosen as the winner. Furthermore, in case of the same F-score and Accuracy, a priority will be given to the results as close as possible to BEP (break-even-point of Precision and Recall).

$$Precision = \frac{\text{TP}}{\text{TP + FP}} \tag{3}$$

$$Recall = \frac{\text{TP}}{\text{TP + FN}} \tag{4}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{5}$$

$$Accuracy = \frac{\text{TP+TN}}{\text{TP + FP + TN + FN}} \tag{6}$$

The scoring for the second task is based on two measures, namely, Micro-Average F-score (microF) and Macro-Average F-score (macroF). Micro-Average F-score is calculated similarly as in equation (5), but on the basis of Micro-Averaged Precision and Recall, which are calculated according to the below equations (7-8). Macro-Average F-score is calculated on the basis of Macro-Averaged Precision and Recall, which are calculated according to the following equations (9-10), where TP is True Positive, FP is False Positive, FN is False Negative, and C is class.

In choosing the winners we look primarily at the microF to treat all instances equally since the number of instances is different for each class. Moreover, in the case of equal results for microF, the team with higher macroF will be chosen as the winner. The additional macroF, treating equally not all instances, but rather all classes, is used to provide additional insight into the results.

$$P_{micro} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FP_i} \tag{7}$$

$$R_{micro} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FN_i} \tag{8}$$

$$P_{macro} = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FP_i}, \tag{9}$$

$$R_{macro} = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FN_i} \tag{10}$$

7.5 Task 6: Results

Results of Task 6-1: In the first task, out of fourteen submissions, there were nine unique teams: n-waves, Plex, Inc., Warsaw University of Technology, Sigmoidal, CVTimeline, AGH & UJ, IPI PAN, UWr, and one independent. Some teams submitted more than one system proposal, in particular: Sigmoidal (3 submissions), independent (3), CV-Timeline (2). Participants used a number of various techniques, usually widely available OpenSource solutions, trained and modified to match the Polish language and the provided dataset when it was required. Some of the methods used applied, e.g., fast.ai/ULMFiT[18], SentencePiece[19], BERT[20], tpot[21], spaCy[22], fasttext[23], Flair[24], neural networks (in particular with GRU) or more traditional SVM. There were also original methods, such as Przetak[25]. The most effective approach was based on recently released ULMFiT/fast.ai, applied for the task by the n-waves team. The originally proposed Przetak, by Plex.inc, was second-best, while third place achieved a combination of ULMFiT/fast.ai, SentencePiece and BranchingAttention model. The results for of all teams participating in Task 6-1 were represented in Table 4.

Results of Task 6-2: In the second task, out of eight submissions, there were five unique submissions. The teams that submitted more than one proposal were: independent (3 submissions) and Sigmoidal (2). Methods that were the most successful for the second task were based on: svm (winning method proposed by independent researcher Maciej Biesek), a combination of ensemble of classifiers from spaCy with tpot and BERT (by Sigmoidal team), and fasttext (by the AGH & UJ team). The results for of all teams participating in Task 6-2 were represented in Table 5. Interestingly, although the participants often applied new techniques, most of them applied only lexical information represented by words (words, tokens, word embeddings, etc.), while none of the participants attempted more sophisticated feature engineering and incorporate other features such as parts-of-speech, named entities, or semantic features.

[18] http://nlp.fast.ai

[19] https://github.com/google/sentencepiece

[20] https://github.com/google-research/bert

[21] https://github.com/EpistasisLab/tpot

[22] https://spacy.io/api/textcategorizer

[23] https://fasttext.cc

[24] https://github.com/zalandoresearch/flair

[25] https://github.com/mciura/przetak

8 Conclusions and Future Plans

The scope of PolEval competition has grown significantly in 2019, both by means of the number of tasks and by including new areas of interest, such as machine translation and speech recognition. We believe that the successful “call for tasks” will be followed by a large number of submissions, as the interest in natural language processing is rising each year and gradually more and more research is devoted specifically to Polish language NLP.
For the next year, we are planning a more open and transparent procedure of collecting ideas for tasks. We will also be focusing on the idea of open data by establishing common licensing terms for all the code submissions, as well as providing a platform to publish and share solutions, models and additional resources produced by participating teams.

Acknowledgements

The work on temporal expression recognition and phrase lemmatization were financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.
The work on Entity Linking was supported by the Polish National Centre for Research and Development – LIDER Program under Grant LIDER/27/0164/L-8/16/NCBR/2017 titled ”Lemkin – intelligent legal information system” and also supported in part by PLGrid Infrastructure.

References

1. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A., Wardyński, A.: KPWr: Towards a Free Corpus of Polish. In: Calzolari et al. [3]
3. Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.): Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). European Language Resource Association, Istanbul, Turkey (2012)
4. Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In: Proceedings of the 2nd International Conference on Human Language Technology Research. pp. 138–145. Morgan Kaufmann Publishers Inc. (2002)
5. Fiscus, J.: Sclite scoring package version 1.5. US National Institute of Standard Technology (NIST), URL: http://www.itl.nist.gov/iaui/894.01/tools (1998)

6. Fiscus, J.G., Ajot, J., Michel, M., Garofolo, J.S.: The Rich Transcription 2006 Spring Meeting Recognition Evaluation. In: International Workshop on Machine Learning for Multimodal Interaction. pp. 309–322. Springer (2006)
7. Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G., Tyers, F.M.: Apertium: A free/open-source platform for rule-based machine translation. Machine Translation **25**(2), 127–144 (2011)
8. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: International Conference on Machine Learning. pp. 1764–1772 (2014)
9. Harper, M.: The Automatic Speech Recognition in Reverberant Environments (ASpIRE) challenge. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). pp. 547–554. IEEE (2015)
10. Kobyliński, Ł., Ogrodniczuk, M.: Results of the PolEval 2017 competition: Part-of-speech tagging shared task. In: Vetulani and Paroubek [33], pp. 362–366
11. Kocoń, J., Marcińczuk, M., Oleksy, M., Bernaś, T., Wolski, M.: Temporal Expressions in Polish Corpus KPWr. Cognitive Studies — Études Cognitives **15** (2015)
12. Kocoń, J., Oleksy, M., Bernaś, T., Marcińczuk, M.: Results of the PolEval 2019 Shared Task 1: Recognition and Normalization of Temporal Expressions. Proceedings of the PolEval 2019 Workshop (2019)
13. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Companion Volume: Proceedings of the Demo and Poster Sessions. pp. 177–180 (2007)
14. Koržinek, D., Marasek, K., Brocki, Ł., Wołk, K.: Polish Read Speech Corpus for Speech Tools and Services. arXiv preprint arXiv:1706.00245 (2017)
15. Marasek, K., Koržinek, D., Brocki, Ł.: System for Automatic Transcription of Sessions of the Polish Senate. Archives of Acoustics **39**(4), 501–509 (2014)
16. Marcińczuk, M.: Lemmatization of Multi-word Common Noun Phrases and Named Entities in Polish. In: Mitkov, R., Angelova, G. (eds.) Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2017). pp. 483–491. INCOMA Ltd. (2017). https://doi.org/10.26615/978-954-452-049-6_064, `https://doi.org/10.26615/978-954-452-049-6_064`
17. Mohri, M., Pereira, F., Riley, M.: Weighted Finite-State Transducers in Speech Recognition. Computer Speech & Language **16**(1), 69–88 (2002)
18. Moro, A., Navigli, R.: Semeval-2015 Task 13: Multilingual all-words sense disambiguation and entity linking. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 288–297 (2015)

19. Ogrodniczuk, M.: The Polish Sejm Corpus. In: Calzolari et al. [3], pp. 2219–2223
20. Ogrodniczuk, M.: Polish Parliamentary Corpus. In: Fišer, D., Eskevich, M., de Jong, F. (eds.) Proceedings of the LREC 2018 Workshop *ParlaCLARIN: Creating and Using Parliamentary Corpora*. pp. 15–19. European Language Resources Association (ELRA), Miyazaki, Japan (2018)
21. Ogrodniczuk, M., Łukasz Kobyliński (eds.): Proceedings of the PolEval 2019 Workshop. Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland (2019), `http://2019.poleval.pl/files/poleval2019.pdf`
22. Ogrodniczuk, M., Kobyliński, Ł. (eds.): Proceedings of the PolEval 2018 Workshop. Institute of Computer Science, Polish Academy of Sciences, Warsaw (2018)
23. Ogrodniczuk, M., Nitoń, B.: New developments in the Polish Parliamentary Corpus. In: Fišer, D., Eskevich, M., de Jong, F. (eds.) Proceedings of the Second ParlaCLARIN Workshop. pp. 1–4. European Language Resources Association (ELRA), Marseille, France (2020), `https://www.aclweb.org/anthology/2020.parlaclarin-1.1`
24. Oleksy, M., Radziszewski, A., Wieczorek, J.: KPWr annotation guidelines – phrase lemmatization (2018), `http://hdl.handle.net/11321/591`, CLARIN-PL digital repository
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics (2002)
26. Pęzik, P.: Increasing the Accessibility of Time-Aligned Speech Corpora with Spokes Mix. In: Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T., Goggi, S., Mazo, H. (eds.) Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). pp. 4297–4300. European Languages Resources Association, Miyazaki, Japan (2018), `https://www.aclweb.org/anthology/L18-1000`
27. Ptaszynski, M., Eronen, J.K.K., Masui, F.: Learning Deep on Cyberbullying is Always Better than Brute Force. In: IJCAI 2017 3rd Workshop on Linguistic and Cognitive Approaches to Dialogue Agents (LaCATODA 2017), Melbourne, Australia. pp. 19–25 (2017)
28. Ptaszynski, M., Masui, F.: Automatic Cyberbullying Detection: Emerging Research and Opportunities. IGI Global Publishing, 1nd edn. (2018)
29. Rosales-Méndez, H., Hogan, A., Poblete, B.: VoxEL: A Benchmark Dataset for Multilingual Entity Linking. In: International Semantic Web Conference. pp. 170–186. Springer (2018)
30. Saurí, R., Littman, J., Gaizauskas, R., Setzer, A., Pustejovsky, J.: TimeML Annotation Guidelines, Version 1.2.1 (2006)
31. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the Association for Machine

Translation in the Americas: Technical Papers. pp. 223–231. Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA (2006)

32. UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., Pustejovsky, J.: SemEval-2013 Task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In: 2nd Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013). vol. 2, pp. 1–9 (2013)
33. Vetulani, Z., Paroubek, P. (eds.): Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu, Poznań, Poland (2017)
34. Vincent, E., Watanabe, S., Barker, J., Marxer, R.: The 4th CHiME speech separation and recognition challenge (2016), http://spandh.dcs.shef.ac.uk/chime_challenge/CHiME4/ (last accessed on September 21, 2021)
35. Wawer, A., Ogrodniczuk, M.: Results of the PolEval 2017 competition: Sentiment Analysis shared task. In: Vetulani and Paroubek [33], pp. 406–409
36. Wolk, K., Marasek, K.: Survey on Neural Machine Translation into Polish. In: International Conference on Multimedia and Network Information Systems. pp. 260–272. Springer (2018)
37. Wróbel, K.: KRNNT: Polish Recurrent Neural Network Tagger. In: Vetulani and Paroubek [33]
38. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al.: The HTK book. Cambridge University Engineering Department **3**, 175 (2002)

Table 4. Results of participants for Task 6-1.

Submission author(s)	Affiliation	Submission name	Precision	Recall	F-score	Accuracy
Piotr Czapla, Marcin Kardas	**n-waves**	**n-waves ULMFiT**	**66.67%**	**52.24%**	**58.58%**	**90.10%**
Marcin Ciura	Plex, Inc.	Przetak	66.35%	51.49%	57.98%	90.00%
Tomasz Pietruszka	Warsaw University of Technology	ULMFiT + SentencePiece + BranchingAttention	52.90%	54.48%	53.68%	87.40%
Sigmoidal Team (Renard Korzeniowski, Przemyslaw Sadowski, Rafal Rolczynski, Tomasz Korbak, Marcin Mozejko, Krystyna Gajczyk)	Sigmoidal	ensamble spacy + tpot + BERT	52.71%	50.75%	51.71%	87.30%
Sigmoidal Team (Renard Korzeniowski, Przemyslaw Sadowski, Rafal Rolczynski, Tomasz Korbak, Marcin Mozejko, Krystyna Gajczyk)	Sigmoidal	ensamble + fastai	52.71%	50.75%	51.71%	87.30%
Sigmoidal Team (Renard Korzeniowski, Przemysław Sadownik, Rafał Rolczyński, Tomasz Korbak, Marcin Możejko, Krystyna Gajczyk)	Sigmoidal	ensenble spacy + tpot	43.09%	58.21%	49.52%	84.10%
Rafal Pronko	CVTimeline	Rafal	41.08%	56.72%	47.65%	83.30%
Rafal Pronko	CVTimeline	Rafal	41.38%	53.73%	46.75%	83.60%
Maciej Biesek		model1-svm	60.49%	36.57%	45.58%	88.30%
Krzysztof Wróbel	AGH, UJ	fasttext	58.11%	32.09%	41.35%	87.80%
Katarzyna Krasnowska, Alina Wróblewska	IPI PAN	SCWAD-CB	51.90%	30.60%	38.50%	86.90%
Maciej Biesek		model2-gru	63.83%	22.39%	33.15%	87.90%
Maciej Biesek		model3-flair	81.82%	13.43%	23.08%	88.00%
Jakub Kuczkowiak	UWr	Task 6: Automatic cyberbullying detection	17.41%	32.09%	22.57%	70.50%

Table 5. Results of participants for Task 6-2.

Submission author(s)	**Affiliation**	**Submission name**	**F-score**	
			Micro-Average	**Macro-Average**
Maciej Biesek		**model1-svm**	**87.60%**	**51.75%**
Sigmoidal Team (Renard Korzeniowski, Przemyslaw Sadowski, Rafal Rolczynski, Tomasz Korbak, Marcin Mozejko, Krystyna Gajczyk)	Sigmoidal	ensamble spacy + tpot + BERT	87.10%	46.45%
Krzysztof Wróbel	AGH, UJ	fasttext	86.80%	47.22%
Maciej Biesek		model3-flair	86.80%	45.05%
Katarzyna Krasnowska, Alina Wróblewska	IPI PAN	SCWAD-CB	83.70%	49.47%
Maciej Biesek		model2-gru	78.80%	49.15%
Jakub Kuczkowiak	UWr	Task 6: Automatic cyberbullying detection	70.40%	37.59%
Sigmoidal Team (Renard Korzeniowski, Przemyslaw Sadowski, Rafal Rolczynski, Tomasz Korbak, Marcin Mozejko, Krystyna Gajczyk)	Sigmoidal	ensamble + fastai	61.60%	39.64%