PolEval 2019 — the next chapter in evaluating Natural Language Processing tools for Polish

Łukasz Kobyliński[†], Maciej Ogrodniczuk[†], Jan Kocoń[‡], Michał Marcińczuk[‡], Aleksander Smywiński-Pohl[*], Krzysztof Wołk[⊖], Danijel Koržinek[⊖], Michal Ptaszynski[$], Agata Pieciukiewicz[⊖], and Paweł Dybała[⊕]

[†] Institute of Computer Science, Polish Academy of Sciences
{lukasz.kobylinski,maciej.ogrodniczuk}@ipipan.waw.pl
[‡]Wrocław University of Science and Technology
michal.marcinczuk@pwr.edu.pl
[*]AGH University of Science and Technology
apohllo@agh.edu.pl
[$]Kitami Institute of Technology
ptaszynski@ieee.org
[⊖]Polish-Japanese Academy of Information Technology
{kwolk,danijel,agata.niescieruk}@pja.edu.pl
[⊕]Jagiellonian University in Kraków
pawel.dybala@uj.edu.pl

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.

1. Introduction

PolEval[1] 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 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 (Wawer and Ogrodniczuk, 2017; Kobyliński and Ogrodniczuk, 2017; Ogrodniczuk and Kobyliński, 2018) 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.

2. Task 1: Recognition and normalization of temporal expressions

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 (Saurí et al., 2006) is a markup language for describing timexes that has been adapted to many languages. PLIMEX (Kocoń et al., 2015) 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*.

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 (UzZaman et al., 2013). 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[2] at the level of annotating boundaries/

[1] http://poleval.pl

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

types[3] and local/global normalization[4] (Kocoń et al., 2015).

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>
```

Evaluation We utilize the same evaluation procedure as described in article (UzZaman et al., 2013). 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:

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

where, $\text{Sys}_{\text{entity}}$ contains the entities extracted by the system that we want to evaluate, and $\text{Ref}_{\text{entity}}$ 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:

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

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

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*.

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

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 (Marcińczuk, 2017). 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] http://poleval.pl/task1/plimex_annotation.pdf

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

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 (Oleksy et al., 2018). The system should generate a lemma for given set of phrases with regards to the context, in which the phrase appears.

Training data The training data consists of 1629 documents from the KPWr corpus (Broda et al., 2012) 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
[...]
```

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

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

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

The accuracy 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

Problem statement Entity linking (Moro and Navigli, 2015; Rosales-Méndez et al., 2018) 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)[5], 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. 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.

[5] `https://www.wikidata.org/`

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[6]
2. Tygodnik Podhalański[7]

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[8]. 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.

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 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 (Wróbel, 2017), categorization of articles into Wikipedia categories and WD types, Wikipedia redirections and internal links.

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.

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

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

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

5. Task 4: Machine translation

Problem statement Machine translation is a translation of text by a computer, with no human involvement. Pioneered in the 1950s, machine translation can also be referred to as automated translation, automatic or instant translation.

Currently there are three most popular types of machine translation system: rules-based, statistical and neural:

- Rules-based systems use a combination of language and grammar rules plus dictionaries for common words. Specialist dictionaries are created to focus on certain industries or disciplines. Rules-based systems typically deliver consistent translations with accurate terminology when trained with specialist dictionaries (Forcada et al., 2011).
- Statistical systems have no knowledge of language rules. Instead they "learn" to translate by analyzing large amounts of data for each language pair. They can be trained for specific industries or disciplines using additional data relevant to the sector needed. Typically, statistical systems deliver more fluent-sounding but less consistent translations (Koehn et al., 2007).
- Neural Machine Translation (NMT) is a new approach that makes machines learn to translate through one large neural network (multiple processing devices modelled on the brain). The approach has become increasingly popular amongst MT researchers and developers, as trained NMT systems have begun to show better translation performance in many language pairs compared to the phrase-based statistical approach (Wolk and Marasek, 2018).

Task description The task is to train as good as possible machine translation system, using any technology, with limited textual resources. The competition will be done for 2 language pairs, more popular English-Polish (into Polish direction) and pair that can be called low resourced Russian-Polish (in both directions).

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.

Evaluation The participants were asked to translate with their systems test files and submit the results of the translations. The translated files should be aligned at the sentence level with the input (test) files. Submissions that were not aligned were not be 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 (Papineni et al., 2002)
- NIST (Doddington, 2002)
- TER (Snover et al., 2006)
- METEOR (Banerjee and Lavie, 2005)

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. Because of the curiosity we also translated our test samples with the Google engine. Please note that this results are not comparable because Google engine was not restricted to any data. For EN to PL translation ModernMT obtained 16.29 BLEU points whereas Google engine scored 16.83. For PL to RU we obtained 12.71 versus 15.78 of the Google, in RU to PL the scores were 11.45 and 13.54 respectively.

6. Task 5: Automatic speech recognition

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. (Harper, 2015; Vincent et al., 2016; Fiscus et al., 2006). The significance of this particular competition is the choice of language. To our knowledge, this is the first strictly Polish evaluation campaign of ASR.

$$w* = \arg\max_i P(w_i|O) = \arg\max_i P(O|w_i) \cdot P(w_i) \tag{2}$$

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 (Young et al., 2002). Currently, this concept was expanded to a more useful implementation based on Weighted Finite-State Transducers (Mohri et al., 2002). Some of the most recent solutions try to bypass the individual sub-steps by modeling the whole

	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.

process in a single end-to-end model (Graves and Jaitly, 2014), however knowledge of the mentioned disciplines is still essential to successfully perform the tuning of such a solution.

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.

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 (Koržinek et al., 2017)
- PELCRA parliamentary corpus (Pęzik, 2018)
- A collection of 97 hours of parliamentary speeches published on the ClarinPL website (Marasek et al., 2014)
- Polish Sejm Corpus for language modeling (Ogrodniczuk, 2012)

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

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 (Fiscus, 1998).

7. Task 6: Automatic cyberbullying detection

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 (Ptaszynski and Masui, 2018), 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.

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 (Ptaszynski et al., 2017), 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 (Ptaszynski and Masui, 2018). 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).

Training data To collect the data, we used the Standard Twitter API[9]. 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[10]: @tvn24, @MTVPolska, @lewy_official, @sikorskiradek, @Pontifex_pl, @donaldtusk, @BoniekZibi, @NewsweekPolska, @AndrzejDuda, @lis_tomasz, @tvp_info, @pisorgpl, @K_Stanowski, @R_A_Ziemkiewicz, @Platforma_org, @RyszardPetru, @RadioMaryja, @rzeczpospolita, @PR24_pl.

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.

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.

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.

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

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

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

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

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

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

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.

9. 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.

10. References

Banerjee, Satanjeev and Alon Lavie, 2005. 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*.

Broda, Bartosz, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński, 2012. KPWr: Towards a Free Corpus of Polish. In (Calzolari et al., 2012).

Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (eds.), 2012. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey: European Language Resource Association.

Doddington, George, 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc.

Fiscus, Jon, 1998. Sclite scoring package version 1.5. *US National Institute of Standard Technology (NIST), URL:* `http://www.itl.nist.gov/iaui/894.01/tools`.

Fiscus, Jonathan G, Jerome Ajot, Martial Michel, and John S Garofolo, 2006. The Rich Transcription 2006 Spring Meeting Recognition Evaluation. In *International Workshop on Machine Learning for Multimodal Interaction*. Springer.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers, 2011. Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Graves, Alex and Navdeep Jaitly, 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning*.

Harper, Mary, 2015. The Automatic Speech Recognition in Reverberant Environments (ASpIRE) challenge. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE.

Kobyliński, Łukasz and Maciej Ogrodniczuk, 2017. Results of the PolEval 2017 competition: Part-of-speech tagging shared task. In (Vetulani and Paroubek, 2017), pages 362–366.

Kocoń, Jan, Michał Marcińczuk, Marcin Oleksy, Tomasz Bernaś, and Michał Wolski, 2015. Temporal Expressions in Polish Corpus KPWr. *Cognitive Studies — Études Cognitives*, 15.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al., 2007. 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*.

Koržinek, Danijel, Krzysztof Marasek, Łukasz Brocki, and Krzysztof Wołk, 2017. Polish Read Speech Corpus for Speech Tools and Services. *arXiv preprint arXiv:1706.00245*.

Marasek, Krzysztof, Danijel Koržinek, and Łukasz Brocki, 2014. System for Automatic Transcription of Sessions of the Polish Senate. *Archives of Acoustics*, 39(4):501–509.

Marcińczuk, Michał, 2017. Lemmatization of Multi-word Common Noun Phrases and Named Entities in Polish. In Ruslan Mitkov and Galia Angelova (eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2017)*. INCOMA Ltd.

Mohri, Mehryar, Fernando Pereira, and Michael Riley, 2002. Weighted Finite-State Transducers in Speech Recognition. *Computer Speech & Language*, 16(1):69–88.

Moro, Andrea and Roberto Navigli, 2015. Semeval-2015 Task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Ogrodniczuk, Maciej, 2012. The Polish Sejm Corpus. In (Calzolari et al., 2012), pages 2219–2223.

Ogrodniczuk, Maciej and Łukasz Kobyliński (eds.), 2018. *Proceedings of the PolEval 2018 Workshop*. Warsaw: Institute of Computer Science, Polish Academy of Sciences.

Oleksy, Marcin, Adam Radziszewski, and Jan Wieczorek,

2018. KPWr annotation guidelines – phrase lemmatization. CLARIN-PL digital repository.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of Association for Computational Linguistics*. Association for Computational Linguistics.

Pęzik, Piotr, 2018. Increasing the Accessibility of Time-Aligned Speech Corpora with Spokes Mix. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis, Takenobu Tokunaga, Sara Goggi, and Hélène Mazo (eds.), *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Languages Resources Association.

Ptaszynski, Michal, Juuso Kalevi Kristian Eronen, and Fumito Masui, 2017. 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*.

Ptaszynski, Michal and Fumito Masui, 2018. *Automatic Cyberbullying Detection: Emerging Research and Opportunities*. IGI Global Publishing, 1st edition.

Rosales-Méndez, Henry, Aidan Hogan, and Barbara Poblete, 2018. VoxEL: A Benchmark Dataset for Multilingual Entity Linking. In *International Semantic Web Conference*. Springer.

Saurí, Roser, Jessica Littman, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky, 2006. TimeML Annotation Guidelines, Version 1.2.1.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*, volume 200.

UzZaman, Naushad, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky, 2013. 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)*, volume 2.

Vetulani, Zygmunt and Patrick Paroubek (eds.), 2017. *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań, Poland: Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Vincent, Emmanuel, S Watanabe, Jon Barker, and Ricard Marxer, 2016. The 4th CHiME speech separation and recognition challenge. *URL: `http://spandh.dcs.shef.ac.uk/chime_challenge/` (last accessed on 1 August, 2018)*.

Wawer, Aleksander and Maciej Ogrodniczuk, 2017. Results of the PolEval 2017 competition: Sentiment Analysis shared task. In (Vetulani and Paroubek, 2017), pages 406–409.

Wolk, Krzysztof and Krzysztof Marasek, 2018. Survey on Neural Machine Translation into Polish. In *International Conference on Multimedia and Network Information Systems*. Springer.

Wróbel, Krzysztof, 2017. KRNNT: Polish Recurrent Neural Network Tagger. In (Vetulani and Paroubek, 2017).

Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al., 2002. The HTK book. *Cambridge University Engineering Department*, 3:175.