Results of the PolEval 2017 Competition: Part-of-Speech Tagging Shared Task

Łukasz Kobyliński and Maciej Ogrodniczuk

Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warszawa, Poland
{lkobylinski, maciej.ogrodniczuk}@ipipan.waw.pl

Abstract

PolEval is a new 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. PolEval 2017 — the first edition of this competition — has included two shared tasks in the area of Part of Speech Tagging and Sentiment Analysis and has gathered 16 submissions from 9 distinct teams. The paper presents the motivation for organizing PolEval, description of the POS tagging task, data and measures used for evaluation of systems and their detailed results.

1. Introduction

The abundance of publicly available natural language processing tools and resources for Polish is a fact; the current size of the Computational Linguistics in Poland LRT list (http://clip.ipipan.waw.pl/LRT) already exceeded 200 entries and every year brings new improvements, both in terms of analyzed linguistic layers and, presumably, the quality of automatic annotation. Still, for some linguistic tasks there is no clear measure of the quality of the tools neither strict general evaluation metrics which could be used to compare algorithms and methods.

These observations motivated PolEval – the contest for evaluating tools for processing Polish inspired by SemEval (Semantic Evaluation campaign[1], already repeated in several local settings such as GermEval[2]). Such model allows language processing tools to compete against some baseline and one another to extend the current state of the art and provide a forum for the researchers to solve challenging computational linguistic problems. What is equally important, the comparison of systems requires creating high quality annotated datasets and defining evaluation metrics.

For the first edition of PolEval we decided to focus on two tasks: POS tagging and sentiment analysis for Polish. This first summary paper describes the tagging task, summarized in detail in Section 3.. Even though the scope of the competition was Polish, we aimed to engage both local and international participants. PolEval 2017 was advertised among the NLP community in Poland and worldwide by distributing calls for papers to widely recognized discussion groups and mailing lists. Training and test data (see Section 4.) has been released and evaluation measures (see Section 5.) made available two months prior to systems submission date. 16 submissions from 9 distinct teams have been gathered (see Section 6.) and system results have been made announced at PolEval site (http://poleval.pl, see also Section 7.). What we are particularly proud of, all winners managed to surpass the state-of-the-art which seems to prove usefulness of our efforts.

2. Previous Work

The first tagger for Polish, proposed by (Dębowski, 2004), has never been publicly released and is not included in further discussion. TaKIPI tagger, described in (Piasecki, 2007), assumes a heterogeneous approach to tagging, combining hand-crafted rules with decision trees. TaKIPI is tied to the original, now obsolete IPI PAN corpus tagset and is also excluded from further experiments.

Currently available taggers, using the latest version of the tagset, include: Pantera (Acedański, 2010) (an adaptation of the Brill's algorithm to morphologically rich languages), WMBT (Radziszewski and Śniatowski, 2011) (a memory based tagger), WCRFT (Radziszewski, 2013) (a tagger based on Conditional Random Fields) and Concraft (Waszczuk, 2012) (another approach to adaptation of CRFs to the problem of POS tagging). Evaluation of performance of a combination of these taggers has been presented in (Kobyliński, 2014) and further discussion on the issues of POS tagging in Polish has been presented in (Kobyliński and Kieraś, 2016).

3. Task Description

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

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

Subtask (A): Morphosyntactic disambiguation and guessing Given a sequence of segments, each with a set of possible morphosyntactic interpretations, the goal of the task is to select the correct interpretation for each of the segments and provide an interpretation for segments for

[1] See e.g. http://alt.qcri.org/semeval2017/.

[2] https://sites.google.com/view/germeval2017-absa/.

which only 'ign' interpretation has been given (segments unknown to the morphosyntactic dictionary).

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

Complete system (C): POS tagging Given a raw text in Polish, the goal of the task is to segment the text by separating individual flexemes and provide the correct lemma and POS tag for each of the segments.

4. Evaluation Data

In the initial phase of the competition the training data has been released to allow the participants to train and evaluate their systems using pre-annotated resources. In the case of Task 1 (POS tagging) we have released the publicly available part of the National Corpus of Polish (Przepiórkowski et al., 2011), manually annotated by qualified linguists, as the training data. This collection of texts has been randomly shuffled and converted from TEI P5 format to a much simpler XCES XML format, which contains information only about the morphosyntactic layer of annotation, discarding other layers, present in the original version of the corpus. The training data has been released in three versions: raw text file, morphosyntactically analyzed and segmented XCES XML and the gold-standard, containing the reference disambiguation. The training data contained 1 215 513 human-annotated segments.

The final test data has not been publicly released before and has been hand-annotated by two qualified linguists specifically for the purpose of the PolEval 2017 competition. The annotation has been conducted in two phases: in the first phase the source raw text — coming from the Polish Coreference Corpus (Ogrodniczuk et al., 2015) — has been annotated in parallel a) manually, by two qualified linguists; b) automatically, using the most recent version of the Concraft tagger, trained on the hand-annotated portion of the NCP. In the second phase, the differences between human annotators and the tagger have been found and cross-corrected by the annotator, which has not previously worked on this text part.

Due to changes in segmentation made in the second phase of annotation, we have used the result of the first phase as the test data for the competition and divided it into two parts: the first part has been released as Task 1 (A) and (B) test data and provided with gold-standard segmentation and morphosyntactic analysis created using the Morfeusz analyzer. This collection contained 27 359 morphosyntactically analyzed segments. The second part has been released as Task 1 (C) test data as a raw text file and contained 27 563 segments. [3]

5. Evaluation Measures and Baselines

We focused on keeping the evaluation procedures as similar as possible to the approaches used previously by the authors of POS taggers to report their work. This (apart from different training and testing data) makes the results comparable with previous work and helps to establish long-term best practices in this field.

Specifically, in the case of POS tagging we followed the guidelines proposed by (Radziszewski and Acedański, 2012) and have clearly separated the evaluation tasks of morphosyntactic disambiguation and end-to-end POS tagging. We have also acknowledged the importance of lemmatization as a separate (sub-)task, which is not always tackled during tagger development.

We have evaluated each of the subtasks according to the following guidelines.

Task 1. Subtask (A) Given a file with the text segmentation already provided the participant should provide a corpus in XCES format, which contains disambiguated POS tags for each of the segments. For the system evaluation we have calculated 3 key statistics: the accuracy of the system in selecting the correct tag for known words (segments, for which some interpretations have been provided in the given test file), the accuracy of the system in guessing the tags for unknown words (segments, for which only 'ign' interpretation has been given in the given test file) and the overall system accuracy.

Task 1. Subtask (B) Given a file with the text segmentation already provided the participant should provide a corpus in XCES format, which contains disambiguated lemmas for each of the segments. For the system evaluation we have calculated 3 key statistics: the accuracy of the system in selecting the correct lemma for known words (segments, for which some interpretations have been provided in the given test file), the accuracy of the system in guessing the lemmas for unknown words (segments, for which only 'ign' interpretation has been given in the given test file) and the overall system accuracy.

Task 1. Subtask (C) Given raw text the participant should perform text segmentation and provide a corpus in XCES format, which contains disambiguated POS tags and lemmas for each of the segments.

For the system evaluation we have calculated the following statistics: the accuracy of the system in selecting the correct lemma and tag and the overall system accuracy (as the weighted average of these values).

In the case of a segmentation error (a particular word has been segmented differently in the gold standard and by the participant's system), we have counted that word as a tagging mistake, both in the case of a POS tag and the lemma.

6. Submitted Systems

Toygger (Krasnowska, 2017) Morphological disambiguation is performed by a bi-directional recurrent LSTM neural network (2 bi-LSTM layers of size 2x384). The network was implemented in Python, using the Keras library [4]

[3] We have also released the corrected version of this corpus (resulting from the second phase of the annotation) at the following address: `http://clip.ipipan.waw.pl/PolEval`

[4] `https://keras.io/`

with TensorFlow backend [5]. For each sentence, each token is represented as a concatenation of two vectors:

1. Morphological vector, i.e. the set of possible tags for each token. The set is represented as a 0-1 "bag-of-values" (BOV) vector, with the one bit representing each distinct POS/category value in the tagset and additional bits for each category representing its absence. The presence of given category value in any of the possible tags is reflected by the corresponding vector entry being 1. If no possible tag contains given category, its 'absence' bit is set to 1.

2. Word embedding (vector of length 300; Aleksander Wawer's model trained on full NKJP corpus and Polish Wikipedia [6]. For words unknown to the embedding model, a heuristic is used to approximate its embedding using vectors for textually similar words.

The output of the second bi-LSTM layer serves as common input to a number of individual dense layers with softmax output activation, each predicting a probability distribution over values of a different part of the correct tag (one layer dedicated to the POS, and one for each morphosyntactic category in the tagset). Based on the values assigned the highest probability and the initially provided set of possible tags (or the whole tagset for 'ign' tokens), a morphosyntactic tag is selected for each token.

KRNNT (Wróbel, 2017) KRNNT uses bidirectional recurrent neural networks (i.e. Gated Recurrent Units) for morphological tagging. Morphological analysis is performed by Maca. The tagger was trained on provided training data. The feature set includes potential tags, suffixes, and prefixes of tokens. The whole token form is not used. Each full tag is represented as a separate output. Collected triples from training data: word form, tag, and lemma are used in lemmatization by choosing most frequent lemma for a pair: word form and tag.

Neuroparser (Rychlikowski et al., 2017) For the training data we have used data given by the competition organizers (and Morfeusz analyser). We have split the data into train and dev (90%—10%) and converted it into *conllu* format (splitting tags into categories i.e. subst:sg:gen:f => case=gen|gender=f|number=sg|pos=subst).

As for the tagging approach, we have modified our dependency parser network. We removed part of the network that was responsible for computing word dependencies (we used previous network hyperparameters). We haven't used any hand-crafted linguistic features. The network works on character-level. After obtaining probability distribution for all the POS tags categories we rate every possibility given by the Morfeusz analyzer (or, in the case of subtask A and B — given by the authors) and choose the best. For unknown words we return the most probable POS tag.

MorphoDiTaPL (Walentynowicz, 2017) MorphoDiTa-PL is a morphosyntactic tagger based on the MorphoDiTa package developed by Czech scientists. An adaptation of this solution for Polish language was made because of linguistic similarities between Czech and Polish language. The tagger architecture is based on the Voted Perceptron algorithm. One of the advantages of MorphoDiTa-PL is that there is no need for pre-tokenization of forms in sentences (for example with the Morfeusz analyzer), what makes installation and use simple. We have used the 1-million word subcorpus of the National Corpus of Polish as learning data in the PolEval 2017 competition.

AvgPer (Pęzik and Laskowski, 2017) The proposed part-of-speech tagger is based on the spacy.io averaged perceptron implementation, which uses basic lexical and contextual feature combinations to predict morphological tags of word tokens. The motivation behind this model was to develop a baseline tagger implementation for Polish, which would be relatively robust for various domains and fast enough to be used for big-data applications. The adaptation of the spacy.io modules required some linguistic work on the translation of the NCP tagset into the Universal Dependencies scheme. The model was trained exclusively on the training data provided for subtask A of the Poleval competition. Its overall accuracy as evaluated on the test set was 90.91 (per cent of correct assignments). The tagger's accuracy for word forms unattested in the Morfeusz morphological dictionary was 67.08 per cent. These results were obtained for the better of the two models submitted to Poleval (AvgPer_Forced), in which the tagger was limited to choose from the morphological dictionary entries provided in the test set.

7. Evaluation Results

To evaluate the performance of the submitted systems, we have calculated their accuracy against the provided test data and compared them with existing systems, proposed to date. The results of evaluation of Task 1 have been presented in Tables 1 – 3.

The predominance of taggers based on neural network architecture is striking in this comparison and the gain in tagging accuracy with regards to previous state-of-the-art methods is considerable. There is a 3 percentage point difference between the previously best-in-class taggers (Concraft) and Toygger, the winner of Task 1 (A). The situation is similar in the remaining subtasks of Task 1 — in the case of lemmatization the best submission has gained more than 2 percentage points in accuracy, as compared to Concraft. The overall tagging accuracy (calculated as the average accuracy of POS disambiguation and lemmatization, see Section 5.) has improved again more than 2 percentage points in the case of the KRNNT_voted system as compared to Concraft.

8. Summary

PolEval 2017 was the first edition of a NLP shared task aimed at encouraging work on tools and resources focusing on Polish language and disseminating it in the form of scientific publications and open-source tools. PolEval 2017 has resulted in providing 2 new language resources (tasks evaluation data) and has attracted 16 submissions from 9 teams in total. The systems submitted for the competition

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

[6] `http://mozart.ipipan.waw.pl/~axw/models/orth/w2v_allwiki_nkjpfull_300.model`

Table 1: Evaluation of Task 1 (A): Morphosyntactic disambiguation accuracy. Acc^{K}_{POS} — POS disambiguation accuracy for known words, Acc^{U}_{POS} — POS disambiguation accuracy for unknown words, Acc_{POS} — overall POS disambiguation accuracy.

System name	$\boldsymbol{Acc^{K}_{POS}}$	$\boldsymbol{Acc^{U}_{POS}}$	$\boldsymbol{Acc_{POS}}$ **(Subtask A score)**
Toygger	95.2425	65.4741	94.6343
KRNNT_AB	94.4888	61.1807	93.8083
KRNNT_ABv	94.3022	62.0751	93.6438
NeuroParser	94.209	64.9374	93.6109
KRNNT_AB_morf1	93.6716	59.7496	92.9785
AvgPer_Forced	91.4104	67.0841	90.9134
AvgPer_RAW	89.4776	62.4329	88.925
Morphosyntactic disambiguer	N/A	N/A	N/A (*)
Concraft	92.3060	58.3184	91.6115
WCRFT	92.0112	50.8050	91.1693
WMBT	91.3843	56.5295	90.6722

(*) — the submitted file did not comply to standard.

Table 2: Evaluation of Task 1 (B): Lemmatization accuracy. Acc^{K}_{L} — lemmatization accuracy for known words, Acc^{U}_{L} — lemmatization accuracy for unknown words, Acc_{L} — overall lemmatization accuracy.

System name	$\boldsymbol{Acc^{K}_{L}}$	$\boldsymbol{Acc^{U}_{L}}$	$\boldsymbol{Acc_{L}}$ **(Subtask B score)**
KRNNT_AB	98.194	80.8587	97.8398
KRNNT_ABv	97.959	79.7853	97.5876
KRNNT_AB_morf1	97.7724	80.1431	97.4122
NeuroParser	97.3731	84.6154	97.1125
Concraft	96.0149	78.1753	95.6504
WCRFT	95.5299	34.8837	94.2907
WMBT	95.2575	37.0304	94.0678

Table 3: Evaluation of Task 1 (C): Overall tagging accuracy. Acc_{POS} — POS disambiguation accuracy, Acc_{L} — lemmatization accuracy, Acc — overall system accuracy.

System name	$\boldsymbol{Acc_{POS}}$	$\boldsymbol{Acc_{L}}$	$\boldsymbol{Acc}$ **(Subtask C score)**
KRNNT_voted	92.9833	96.9089	94.9461
KRNNT_raw	92.6242	96.8581	94.7411
NeuroParser	91.5865	97.0032	94.2949
MorphoDiTaPL	89.6746	95.7769	92.7258
Concraft	90.0773	94.7212	92.3992
WCRFT	89.6927	93.7888	91.7407
WMBT	89.0614	93.6183	91.3398

are currently the best-performing state-of-the-art methods in Polish.

Apart from the accuracy of the systems, some of the authors have declared that they focused on the speed and robustness of their methods in the context of Big Data and distributed computing. In the current edition of PolEval we have focused on evaluating accuracy exclusively, but this aspect in future editions of the competition remains to be considered.

Many of the competing and winning systems were based on emerging deep learning techniques and algorithms. The PolEval competition had turned out to be an impulse for the development of this kind of NLP tools and practicing their usage on the Polish language data sets. In many cases it resulted in out-performing older solutions, usually based on machine learning.

Hopefully, future editions of PolEval will include at least the same topics as the current one. However, the list may include other important areas such as syntactic parsing (e.g. dependency) and perhaps cover other syntactic and semantic tasks.

Acknowledgments

The organizers would like to thank Dorota Komosińska for providing a software platform for manual POS annotation of evaluation data by annotators and for further processing of this data. We would also like to thank the annotators: Paulina Rosalska, Sebastian Żurowski, Monika Jędras and Estera Małek, as well as Michał Wasiluk who has helped in evaluating taggers against the test data.

We also would like to thank all participants who have made PolEval 2017 a very successful event and LTC 2017 (Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, `http://ltc.amu.edu.pl`) conference organizers for welcoming PolEval 2017 results in a special session.

Finally, we gratefully acknowledge the support of Sages, the sponsor of the prizes for the four winners of each of the subtasks.

This work has also been financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

9. References

Acedański, Szymon, 2010. A morphosyntactic Brill tagger for inflectional languages. In *Advances in Natural Language Processing*.

Dębowski, Łukasz, 2004. Trigram morphosyntactic tagger for Polish. In *In Proceedings of the International IIS:IIPWM'04 Conference*. Springer-Verlag.

Kobyliński, Łukasz, 2014. PoliTa: A multitagger for Polish. Reykjavík, Iceland: ELRA.

Kobyliński, Łukasz and Witold Kieraś, 2016. Part of Speech Tagging for Polish: State of the art and future perspectives. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2016)*.

Krasnowska, Katarzyna, 2017. Morphosyntactic disambiguation for Polish with bi-LSTM neural networks. In (Vetulani, 2017).

Piasecki, Maciej, 2007. Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, 11(1–2):151–167.

Ogrodniczuk, Maciej, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska, 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Marek Łaziński, and Piotr Pęzik, 2011. National Corpus of Polish. In Zygmunt Vetulani (ed.), *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań, Poland.

Pęzik, Piotr and Sebastian Laskowski, 2017. Evaluating an averaged perceptron morphosyntactic tagger for Polish. In (Vetulani, 2017).

Radziszewski, Adam, 2013. A tiered CRF tagger for Polish. In Robert Bembenik, Ł. Skonieczny, H. Rybiński, M. Kryszkiewicz, and M. Niezgódka (eds.), *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Springer Verlag, pages 215–230.

Radziszewski, Adam and Szymon Acedański, 2012. Taggers gonna tag: an argument against evaluating disambiguation capacities of morphosyntactic taggers. In *Proceedings of TSD 2012*, LNCS. Springer-Verlag.

Radziszewski, Adam and Tomasz Śniatowski, 2011. A Memory-Based Tagger for Polish. In *Proceedings of the LTC 2011*.

Rychlikowski, Paweł, Michał Zapotoczny, and Jan Chorowski, 2017. Character-based neural POS tagger. In (Vetulani, 2017).

Vetulani, Zygmunt (ed.), 2017. *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań, Poland.

Walentynowicz, Wiktor, 2017. MorphoDiTa-based tagger addapted to the Polish language technology. In (Vetulani, 2017).

Waszczuk, Jakub, 2012. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*. Mumbai, India.

Wróbel, Krzysztof, 2017. KRNNT: Polish recurrent neural network tagger. In (Vetulani, 2017).