Evaluating an averaged perceptron morphosyntactic tagger for Polish

Piotr Pęzik[*†]
Sebastian Laskowski[†]

[*]University of Lodz
piotr.pezik@uni.lodz.pl

[†]VoiceLab
sebastian.laskowski@voicelab.pl

Abstract

This paper evaluates the accuracy of an averaged perceptron morphosyntactic tagger for Polish submitted to the PolEval 2017 competition. Trained and tested on the PolEval datasets using both the original tagset and its Universal Dependencies translation, the tagger achieves up to 92.02% accuracy of full tagset prediction and 98.2% accuracy of tagging flexemes. The accuracy of tagging out-of-vocabulary word forms (up to 67.08%) was consistently higher for different versions of the model than the results reported for currently available taggers of Polish.

1. Introduction

Part-of-speech taggers are some of the basic building blocks of natural language processing systems. Their accuracy rates reported for languages with relatively simple inflectional morphologies, such as English, have exceeded 97% (Toutanova et al., 2003), although exact performance scores may vary significantly across different registers and domains. Prospects of further improvement are currently uncertain, and it has been suggested that some advantages could be gained from a deeper linguistic analysis of tagging results (Manning, 2011). Currently available morphosyntactic taggers of Polish, which have to deal with its complex morphology, manage to correctly label approx. 91% of word tokens, with just above 60% accuracy on out-of-vocabulary word forms (Kobyliński, 2014). This paper describes an averaged perceptron implementation of a morphosyntactic tagger for Polish (called APT), whose results were submitted to the Poleval 2017 competition. We describe the feature set used to train the tagger as well as its performance on two training datasets and three tagsets: a) the full National Corpus of Polish (NCP) tagset b) its flexemic subset and c) its translation into the Universal Dependencies (PoS) scheme (De Marneffe et al., 2014).

2. Previous work

Recent morphosyntactic taggers for Polish have used Brill's algorithm (Acedański, 2010), memory-based (Radziszewski and Sniatowski, 2011) and CRF models (Waszczuk, 2012; Radziszewski, 2013) as well as voting ensembles of these systems (Kobyliński, 2014). To the best of our knowledge, there have not been any published evaluations of an Averaged Perceptron (AP)-based tagger for Polish, even though such tagging models were proposed quite some time ago by (Collins, 2002) and even adopted for some Slavic languages, cf. (Hajič et al., 2009). Given the relative simplicity of such a model and its good reputation for speed and accuracy at least for English, we considered it worthwhile to evaluate an AP-based tagger for Polish using the datasets released in the PolEval competition.

3. The data

The tagger described in this paper was trained and evaluated on PolEval training and test data sets, comprising over 1.215 million and 27 300 manually annotated tokens respectively. The morphosyntactic annotations provided in subtask 1A and 1B of PolEval[1] were already segmented and thus the results reported here refer to the accuracy of predicting the labels of a segmented stream of tokens rather than to the performance of a complete PoS tagging solution. To check the ability of our model to learn from additional data, we have also run some experiments on an extended training set including over 447 000 tokens from a recent version of the KPWr corpus (Broda et al., 2012). Detailed statistics of these collections are summarized in Table 1.

Corpus	Sentences	Tokens
PolEval TR(ain) set	85 663	1 215 514
PolEval TE(st) set	1 626	27 359
KPWr 1.2	28 680	447 575
PolEval TR + KPWr	114 343	1 663 089

Table 1: Training and test sets used in the experiment.

It should be noted that the sample used to evaluate the tagger is rather small compared with the conventional test splits used in machine learning experiments. Nevertheless, using this test set allowed for maintaining some consistency of the reported accuracy scores with the results reported for the other systems submitted to PolEval.

[1]Descriptions of these tasks are available at `http://poleval.pl/index.php/tasks/`.

4. Implementations

4.1. spacy.io

For the purposes of this paper, we evaluated two AP tagger implementations[2]. The first one is available in the `spacy.io` library and, for technical reasons, it required a translation of the NCP tagset used in both PolEval and KPWr, into the Universal Dependencies scheme. The translation resulted in an additional comparison of tagging performance on two largely equivalent tagsets, whose results are reported below. An example translation of NCP tags into the UD tagset one is shown in Table 2.

Word	NCP	UD
Proces	subst:sg:acc:m3	NOUN Animacy=Inan Case=Nom Gender=Masc Number=Sing
budzi	fin:sg:ter:imperf	VERB Aspect=Imp Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin
wielkie	adj:pl:acc:f:pos	ADJ Degree=Pos Case=Acc Gender=Fem Number=Plur
emocje	subst:pl:acc:f	NOUN Case=Acc Gender=Fem Number=Plur

Table 2: Example translations of NCP tags into their Universal Dependencies equivalents.

Some of the NCP flexemes (such as the `depr` tag used to explicitly mark depreciative noun forms) and categories (such as) were lost in the process of translation due to the lack of their UD equivalents. Perhaps the most significant difference between the two tagsets is the absence of an explicit tag for proper nouns in the NCP scheme. It should also be noted that some of the general equivalence relations had lexeme- or even form-specific exceptions. For example, forms of words such as *kilka* and *niejaki* were translated as determiners (`DET`) into the UD tagset, even though they are marked as adjectives in NCP, which in turn were by default translated as `ADJ`s in UD. The complete set of tagset translation rules is available in the public repository of our tagger (see the *Availability* section of this paper).

[2]Both of which were originally written by Mathew Honnibal, see `https://explosion.ai/blog/part-of-speech-pos-tagger-in-python`.

4.2. Adapted NLTK

The second implementation used in this experiment was based on the AP tagger source code available in the NLTK package[3]. We modified this implementation to have an option of prioritizing the highest scoring interpretations which at the same time happen to be listed in a reference morphological dictionary, such as the SGJP edition of the Morfeusz dictionary (Woliński, 2014)[4].

The full set of features used in both implementations of the AP taggers is shown in Table 3. Most of them are self-explanatory lexical and contextual features and feature combinations. The last feature listed is simply one of a thousand Brown cluster identifiers (Brown et al., 1992) computed for word forms found in the balanced subset of the National Corpus of Polish using a minimum frequency threshold of 10 occurrences. The basic intuition behind using Brown clusters as a feature when predicting a word form's part of speech is illustrated in Table 4, where a set of relatively rare verb forms are grouped into one cluster. Such labels may be a potentially useful indication indicating the affinity of words which are not found in the training set or even in the reference morphological dictionary to the grammatical class of a known word from from the same cluster. For full morphosyntactic tagging of Polish, such clusters seem to be of limited use, as they rarely reflect distinctions between detailed grammatical categories such as gender or number [5].

Tagging feature
i-th word's suffix
i-th word's prefix
i-th word
(i-1)-th tag
(i-2)-th tag
(i-1)-th + (i-2)-th tags
(i-1)-th word
(i-2)-th word
(i+1)-th word
(i+2)-th word
(i-1)-th tag + i-th word
(i+1)-th word's suffix
(i-1)-th word's suffix
i-th word's Brown cluster

Table 3: Features used to train the averaged perceptron tagger.

5. Evaluation

5.0.1. A flexeme tagger

As a first experiment, we evaluated our tagger's capacity to predict flexeme tags in the Poleval test

[3]`http://www.nltk.org/_modules/nltk/tag/perceptron.html`.

[4]`http://sgjp.pl/morfeusz/morfeusz.html`.

[5]The NCP Brown clusters are available in the project's repository.

Cluster	Word form
000010001	przełożywszy
000010001	sfotografuj
000010001	zaprojektujesz
000010001	oczyściwszy
000010001	zawieźcie
000010001	formatując
000010001	dorysuj
000010001	pielęgnujmy
000010001	wezwijmy
000010001	rozszerzmy

Table 4: An example Brown cluster generated from NCP.

set. The flexeme has been defined as "a morphosyntactically homogeneous set of forms belonging to the same lexeme" (Woliński, 2014). In Polish flexemes are somewhat more specific than what could be described as main part-of-speech categories. For example, there are 36 National Corpus of Polish (NCP) flexeme categories[6] (including the `ign` tag used to mark unknown forms in the process of automatic tagging), compared with only 17 main part-of-speech categories used in the UD scheme, although, as indicated above, some UD distinctions are in a sense more specific than NCP ones and vice versa.

The relevance of testing the accuracy of flexeme or main part-of-speech tagging for Polish can be justified by the fact that some language processing applications only require main part-of-speech annotation rather than a more complete morphosyntactic analysis. For example, a great majority of morphosyntactic queries submitted to the PELCRA 2 search engine[7] only make use of main part-of-speech categories to specify search terms. This is because such categories are often sufficient to provide a satisfactory level of disambiguation of homonymous word forms, such as *miał* (as a participle verb form) vs. *miał* (as a noun). Also, in some recall-oriented positional collocation extraction systems, only main part of speech categories are used to model phraseological relations between nodes and collocates and thus a main part of speech tagger would be sufficient to annotate the corpus data used to compile collocation databases such as HASK (Pęzik, 2012).

Table 5 shows the results of training an AP-based tagger on the PolEval data. After just 5 iterations, the model achieved 98.2% accuracy on this task[8]. These results seem to warrant two conclusions. Firstly, an AP tagger flexeme tagger can be sufficiently fast and accurate for some language processing applications. Secondly, however, as we will see in the subsequent sections the real challenge of tagging Polish lies in the finer-grained categories of ambiguous forms, such as case for nouns or gender for adjectives.

Number of tags (without *ign*)	35
Token accuracy	98.2%
Training set	PolEval TR
Test set	PolEval TE
Training time	5m 40s (5 iterations)
Tagging time	1.42s

Table 5: Speed and accuracy evaluation of an AP-based 'flexeme' tagger.

5.1. Full morphosyntactic tagging

The results submitted for Task 1A of the PolEval competition were obtained using the spacy.io implementation of the proposed AP tagger. The overall accuracy of the full NCP tag assignments was 90.91% (91.41% for known word forms). Interestingly, a relatively high accuracy score of 67.08% was obtained for unknown word forms (marked as `ign` in the test set. The impact of a tagger's ability to correctly predict out-of-vocabulary words on its overall performance may vary across different domains and registers. As summarized in Table 6, in the PolEval test set, there were 559 tokens marked as `ign`, making up just 2.04% of the total number of 27 359 tokens.

Collection	Total tokens	Unknown	Ratio
Web News 2017	127 741 000	4 225 000	3.31%
PolEval TE	27 359	559	2.04%

Table 6: Unknown words in two corpora of Polish.

In a collection of Polish 200 000 web-based news articles published between January and August 2017, which comprises over 127.7 million word tokens (874 425 types), more than 4 225 000 word tokens (3.31%) were not matched against the Morfeusz dictionary. It is interesting to notice that these tokens accounted for 404 597 or 46.26% of all distinct word form types. This can be explained, at least partly, by the fact that news reports are rather formulaic in their use of common and known words, but at the same time they are characterized by a high incidence of proper nouns and other components of named entities, which are either too rare or too recent to be covered by morphological dictionaries (e.g. *Macron*, *Instagram*). Another major factor is misspellings and latinized spellings of Polish diacritics found in comments to news articles, but also in many other genres of Polish web-based discourse. Finally, there are also neologisms and loan words which have yet to be included in the reference morphological dictionary. Examples of such forms which were correctly predicted by our tagger are shown in Table 7. Although neither the form *tweetnij*

[6] `http://nkjp.pl/poliqarp/help/ense2.html`.

[7] `http://pelcra.clarin-pl.eu/NKJP/`.

[8] Speed tests were performed on a single core of an Intel Core i7 2820QM / 2.3 GHz processor machine.

nor *focię* is included in the dictionary, they are both correctly annotated [9].

Word	Tag
Tweetnij	impt:sg:sec:perf
tę	adj:sg:acc:f:pos
focię	subst:sg:acc:f
.	interp

Table 7: Examples of correct predictions of out-of-dictionary word forms.

The results summarized in Table 8 were obtained for the NLTK implementation of the AP tagger. The main improvement in this implementation involved using a score-based ranking of dictionary-attested morphological interpretations of ambiguous tokens. In other words, when offered a set of dictionary-attested tags for a token, the tagger selected the highest-ranked prediction from this set instead of proposing a dictionary-unattested one. The best result of 92.02% accuracy was obtained for a model trained on 20 iterations over the PolEval data. The addition of some 400 000 tokens from the KPWr corpus to the training set did not produce any improvements on the PolEval test set. The accuracy of tagging after a UD translation was very similar to the one on the NCP tagset (91.08% after 5 iterations and 92.03 after 20 iterations), despite the subtler distinctions between adjectives, numerals and determiners in the UD tagset, which may have partly increased the level of morphological ambigutity.

Compared with the speed of the flexeme tagger, the full tagger is considerably slower with about 1000 token classifications per second on a single core of the above-mentioned machine. This is mainly due to the fact that as many as 926 distinct combinations of full morphosyntactic tags are acquired by the full tagger from the training dataset, compared with only 35 flexeme tags in the NCP tagset.

Train set	Tagset	Iter.	Acc	Acc^U
PolEval TR	NCP	5	91.95%	63.1%
PolEval TR	NCP	20	**92.02%**	**64.0%**
PolEval TR + KPWr	NCP	5	91.9%	62.6%
PolEval TR + KPWr	NCP	20	91.86%	62.43%
PolEval TR	UD	5	91.08%	61.3%
PolEval TR	UD	5	**92.03%**	**63.96%**

Table 8: Full NCP tagset results.

[9]The phrase *Tweetnij tę focię* is a rather informal way of saying *Tweet this photo!*, where *Tweetnij* can be described as a loanword and *focia* is dictionary-unattested informal derivative of *fotografia*. Such phrases may quickly become quite recurrent in web-based registers, thus causing potential problem with the accuracy of PoS taggers.

6. Conclusions and future work

The results reported in this paper suggest that a very 'lean' implementation of an averaged perceptron morphosyntactic tagger can achieve state-of-the-art accuracy for Polish and even outperform existing systems on the task of predicting tags of out-of-vocabulary word forms. Also, the lexeme tagger based on the AP model is both fast and accurate enough to be used in areas where only main part-of-speech tagging is required. It should be noted, however, that in order to ensure some consistency with the results reported by authors of other systems submitted to the PolEval competition, the evaluation of the APT tagger was performed on the 'official' test set, which is rather small. Cross-validated evaluation on larger test sets could shed more light on how stable the results reported in this paper are. The fact that no gains were observed after adding a considerable amount of training data from the KPWr corpus could mean that we have reached the model's upper bound of accuracy for the test set evaluated. Finally, we conclude that using a good reference morphological dictionary such as Morfeusz increases the overall accuracy of tagging Polish texts.

7. Availability

The source code of the tagger described in this paper, together with the trained models, datasets, tagset translations and other resources such as the NCP Brown clusters are publicly available at `https://gitlab.com/piotr.pezik/apt_pl`.

8. Acknowledgments

Research described in this paper was carreid out within grant funded by the European Union Regional Operation Program for the Pomeranian Voivodship for the years 2014-2020 based on the Resolution no 190/214/17 of the Pomeranian Voivodship Board, project RPPM.01.01.01-22-0026/16.

9. References

Acedański, Szymon, 2010. A morphosyntactic Brill tagger for inflectional languages. In *International Conference on Natural Language Processing*. Springer.

Broda, Bartosz, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński, 2012. Kpwr: Towards a free corpus of Polish. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Brown, Peter F., Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai, 1992. Class-based N-gram models of natural language. *Comput. Linguist.*, 18(4):467–479.

Collins, Michael, 2002. Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP

'02. Stroudsburg, PA, USA: Association for Computational Linguistics.

De Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning, 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *LREC*, volume 14.

Hajič, Jan, Jan Raab, Miroslav Spousta, et al., 2009. Semi-supervised training for the averaged perceptron POS tagger. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kobyliński, Łukasz, 2014. PoliTa: a multitagger for Polish. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Manning, Christopher D., 2011. *Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?*. Berlin, Heidelberg: Springer Berlin Heidelberg, pages 171–189.

Pęzik, Piotr, 2012. Graph-based analysis of collocational profiles. *Phraseologie im Wörterbuch und Korpus, Proceedings of Europhras*:227–243.

Radziszewski, Adam, 2013. A tiered CRF tagger for Polish. *Intelligent tools for building a scientific information platform*, 467:215–230.

Radziszewski, Adam and Tomasz Sniatowski, 2011. Maca-a configurable tool to integrate Polish morphological data.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer, 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics.

Waszczuk, Jakub, 2012. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *COLING*.

Woliński, Marcin, 2014. Morfeusz reloaded. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*. Reykjavík, Iceland: ELRA.