Automatic Detection of Polish Metaphorical Expressions

Anonymous ACL submission

Abstract

This study investigates the automatic detection of metaphorical expressions in Polish texts using a BERT-based classifier trained on a manually annotated corpus. We discuss challenges in adapting existing methods to Polish, including issues with discontinuities and multi-level metaphors. Experiments yield promising results, with an average F1 score of 0.669 in cross-validation. Analysis reveals a strong correlation between specific lemmas and metaphoricity, suggesting high lexicalization in the corpus. Further experiments on an unbalanced test set demonstrate some ability to recognize novel metaphors but highlight the need for a larger, more diverse corpus to improve performance and compete with English-language metaphor detection systems.

1 Introduction

Automatic detection of metaphorical expressions, one of the biggest problems in text comprehension, has been a very popular topic for English since the beginning of the second decade of the 21st century. Initial methods were based on selecting relevant syntactic-semantic features based on WordNet, FrameNet, VerbNet, SUMO ontologies, distributional models, etc., and then training the model using supervised learning (reviewed, e.g., by Veale et al. (2016)).

With the rapid development of deep neural networks, this learning model started to be used for metaphor detection based on token classification in a sentence (Rei et al., 2017; Gutiérrez et al., 2017). It has become so popular that almost all submissions to the collaborative task at the Proceedings of the Workshop on Figurative Language Processing in 2018 and 2019 were based on this method (Leong et al., 2018, 2020). Similar experiments for Polish on the FigAN and FigSen corpora[1] based on the word2vec distribution model were carried out by Mykowiecka and Marciniak (2020).

The development of large language models (LLMs), again led to a change in metaphor detection methodology, with RoBERTa (Liu et al., 2019) used by both MrBERT (Song et al., 2021) and the best-performing MelBERT (Choi et al., 2021) for the metaphor detection task in 2019.

Methods based on LLMs were not used for Polish due to the lack of appropriately labelled data. With the development of a baseline corpus of Polish metaphors (Hajnicz, 2022), this gap has been partially bridged. In the following section, several experiments analysing the usefulness of this corpus for the automatic detection of metaphorical expressions will be discussed.

2 Analysis of the corpus for automatic detection of metaphorical expressions

Most metaphor recognition models for English were trained and evaluated on the VUA corpus (Steen et al., 2010; Krennmayr and Steen, 2017) or its subset. The corpus of Polish metaphors used by us adopted different tagging rules. The most significant difference in terms of automatic detection of metaphorical expressions is the labelling of whole phrases, rather than just single words, as metaphorical. This change positively affects the quality of the corpus from a linguistic point of view, but for natural language processing it is often a hindrance. In most tasks (e.g. terminology and proper name detection), the tagging BIO is used (where the tag B together with all immediately following tags I form a single detected phrase). E.g. in Example 2 we have the tagged metaphorical phrase „zająć miejsce" (*take a place (in a competition)*)[2].

[1] https://zil.ipipan.waw.pl/CoDeS

[2] All metaphors occurring in this chapter have been taken from the underlying corpus of Polish metaphors, full sentences have been simplified and modified

(1)

Pierwsze	miejsce	zajął	Kowalsk*i*	.
First	*place*	*took*	*Kowalski*	.
O	***B***	***I***	***O***	***O***

2.1 Difficulties in BIO encoding

Several properties of the Polish make it difficult to apply the BIO encoding. We briefly comment on them below. (examples in the Appendix A)

Discontinuities, while almost non-existent for proper names or terminology, are not unexpected for metaphors. We can obtain discontinuity by performing an inversion or by adding a new subordinator to the verb. The problem becomes more complex when additional elements contain independent metaphors, further complicating the encoding process.

Multilevel metaphors present another challenge, occurring when part of a metaphorical expression is itself a separate metaphor. These can involve conventional metaphors with one-element carriers, metaphorical expressions, and metaphors with one-element carriers in relation to specific subjects. The BIO encoding struggles to accurately represent these layered metaphors, often resulting in loss of metaphoricity or creation of unintended metaphors due to word order variations.

Shared phrase fragments between metaphorical and non-metaphorical expressions present strong signals suggesting both metaphorical and non-metaphorical markings for the same words. This phenomenon can occur with reflexive markers associated with both literal and metaphorical verbs, or through coordination where a verb is used both literally and metaphorically. These cases highlight the complexities of accurately encoding and detecting metaphors in natural language processing tasks.

2.2 The final format of the data

Due to the aforementioned issue, we decided to simplify the corpus format for automatic detection of metaphorical expressions. In the end, we dropped the linking information between tokens and kept only the information about whether a token is an element of a metaphorical expression (and thus obtained the IO tagging).

An additional advantage of this solution is that it effectively aligns the labelling of the Polish corpus with the labelling of the VUA corpus, which will allow a better comparison of the results of the initial experiments.

3 Experiments

3.1 Classifier architecture

Given the final format of the corpus data, the problem of metaphor detection can be reduced to the classification of tokens in a sentence. For this task, the architecture of `TFBertForTokenClassification`[3] was used.

It consists of a simple dense layer on top of a pre-trained BERT model, designed to predict a tag for each token in the input. Such a classifier is applied to the context vectors of consecutive input tokens generated by the BERT model. It predicts a vector of logit probabilities of assigning specific tags to an input token. This entire structure is then trained, the weights of the individual neurons of the network are tuned to the problem. The original weights of the BERT model used are also tuned in the same process. The natural choice for Polish is HerBERT (Mroczkowski et al., 2021).

3.2 Classifier training

Since the underlying corpus of metaphorical expressions is relatively small, we decided to perform a 10-fold cross-validation on its random segmentation. The size of the corpus posed a problem: the loss function for the validation data started to increase after only 7 epochs of matching weights, so the neural network very quickly stopped generalizing reasonably from the data and started to match individual examples within them. Still, the results of this experiment presented in Table 1 proved to be significantly better than we initially expected. For comparison, the F1 measure for MelBERT trained on the VUA corpus is 0.723.

3.3 Results vs. lexicalization of metaphors

In order to verify the quality of the results obtained, we decided to look at the corpus in terms of the frequency of metaphor attribution to individual lemmas. The 10 lemmas with the highest number of metaphorical occurrences in the corpus are presented in Table 2. The metaphorical occurrences of the 25 most common lemmas account for as much as 15.7% of the metaphor occurrences in the corpus. Such a high number indicates a strong link between lemmas and metaphoricity marking. Therefore, we decided to analyse what part of the metaphors tagged in the corpus are subject to lexicalisation.

[3] https://huggingface.co/docs/transformers/model_doc/bert

Run	Precision	Recall	F1
1	0.722	0.606	0.659
2	0.745	0.599	0.664
3	0.754	0.610	0.674
4	0.724	0.620	0.668
5	0.755	0.619	0.680
6	0.708	0.632	0.669
7	0.746	0.610	0.671
8	0.751	0.608	0.672
9	0.737	0.634	0.682
10	0.720	0.602	0.656
Avg.	0.735	0.614	0.669

Table 1: Cross-validation results.

Word	Occurrences		
	metaph.	literal	% metaph.
się	618	2504	0.198
cały	164	32	0.837
nowy	137	93	0.596
dobry	127	47	0.730
wielki	124	49	0.717
duży	114	44	0.722
życie	88	82	0.518
miejsce	83	142	0.369
wysoki	78	26	0.750
prowadzić	78	7	0.918
w	47	5047	0.009

Table 2: 10 words with the highest number of metaphorical occurrences and a problematic lemma.

Table 3 shows the minimum number of lemmas needed to label the given part of the metaphors occurring in the corpus. The percentage of occurrences of selected lemmas in the metaphorical sense in the corpus has been denoted as precision[4].

The low precision values are mainly due to the existence of words in which the high number of occurrences labelled as metaphorical is due to the very high frequency of these words in the corpus. Such examples that stand out in Table 2 are the reflexive particle „się", whose metaphoricity depends on the verb with which it is associated, and the preposition „w" (*in*), marked primarily in a metaphor that is a multi-word phrase or a continuous text.

[4]This is an approximation of the accuracy of the classifier, which assigns metaphoricality to all occurrences of selected words

% metaphor	No. lemmas	Precision
0.1	10	0.347
0.2	44	0.255
0.3	102	0.170
0.4	190	0.130
0.5	321	0.118
0.6	526	0.115
0.65	669	0.121
0.7	846	0.126
0.75	1082	0.125
0.8	1390	0.126
0.9	2456	0.124
1	4002	0.121

Table 3: Minimum number of lemmas needed to cover part of the corpus.

In order to obtain a better estimate of the level of lexicalisation of metaphor in the corpus, we repeated the analysis limiting ourselves to lemmas that occurred at least 5 times in the corpus (regardless of whether they were marked as metaphorical or not) and that were marked as metaphorical in at least half of the occurrences.

% metaphor	No. lemmas	Precision
0.1	21	0.720
0.2	77	0.723
0.3	178	0.720
0.4	384	0.720
0.5	862	0.725
0.6	2180	0.713
0.634	2710	0.704

Table 4: Minimum number of lemmas used mostly metaphorically needed to cover parts of the corpus.

Table 4 shows the minimum number of such lemmas to label the indicated parts of the corpus. Thus, all lemmas that occurred at least 5 times in the corpus and that were labelled as metaphorical in at least half of the cases cover 63.4% of all metaphors occurring in the corpus. A classifier marking all occurrences of such lemmas as metaphorical would achieve an accuracy of 0.704 and a completeness of 0.634, resulting in an F1 measure of 0.667. Since the results of a very simple classifier obtained in this way are not statistically different from those of a classifier based on a neural network, it is reasonable to ask whether, given the current size of the corpus, the network has managed to make any

other generalisation.

4 Experiments with an unbalanced test set

The analysis in Section 3.3 showed a significant problem – the association of metaphor with specific lemmas (probably resulting from the distribution of occurrences of different types of metaphor in the texts) is large enough that the neural network may have problems generalising from the corpus data. To be able to verify the quality of the classifier for other words, a dataset containing a large number of metaphors with lemmas not present in the training data is needed.

4.1 Corpus breakdown

We divided our corpus into training, validation and test sets with relatively divergent lexica using the greedy method:

1. We assigned a density to all lemmas, i.e. the number of metaphorical occurrences divided by the square of the number of samples in which these metaphors occur (in training set at any point, whole corpus at the start)

2. iteratively, we select the lemma with the highest density, move all associated samples to the test set and update the density value for the remaining lemmas

3. once the planned size of the test set was reached (i.e., ca. 10% of the corpus size), the same procedure was used to create a validation set from the rest of the corpus.

As Table 5 shows, we achieved our goal – 36% of the words labelled as metaphors in the test part of the corpus do not occur in metaphorical meaning in the training part.

corpus part	new lemmata	new occurr.	% metaph.
test	340	863	36.0%
validation	415	688	26.8%

Table 5: Lexis of metaphorical occurrences

4.2 Results

The architecture and training method of the new classifier were identical to the original experiments (see Sections 3.1 and 3.2). The loss function on the validation data started to increase again after 7 epochs of training, so the resulting classifier achieved a similar level of generalisation. The experiment was repeated ten times.

Run	P	R	F1	New lemmas	
				Count	R
1	0.730	0.479	0.578	340	0.588
2	0.748	0.514	0.609	321	0.611
3	0.757	0.435	0.552	206	0.597
4	0.767	0.436	0.556	237	0.633
5	0.772	0.426	0.550	216	0.625
6	0.744	0.448	0.559	256	0.563
7	0.786	0.416	0.544	200	0.645
8	0.751	0.483	0.588	265	0.604
9	0.810	0.409	0.543	188	0.702
10	0.753	0.491	0.594	279	0.595

Table 6: Results on unbalanced test set.

Table 6 shows that, as expected, the completeness of the classifier decreased. A large scatter in the results obtained, clearly showing the lack of sufficient data to achieve stabilisation.

For new lemmas (i.e. lemmas that were never labelled as metaphorical in the training data), the accuracy of the classifier obtained in this way exceeded 55% every time. The neural network managed to learn some patterns even on small and heavily loaded data. Unfortunately, the number of correctly detected metaphors with new lemmas is very low – approx. 100 out of 863 were correctly detected.

5 Summary

The results of the experiments on the current version of the base corpus of Polish metaphors are less promising than expected. The small size of the data and the predominant presence of the lexicalised metaphors hinder learning and largely reduce it to the recognition of specific words in the text (see Section 3.3).

On the other hand, even for small data, it was possible to make some generalisations and draw correct inferences in situations it had not encountered before (see Section 4). Catching up with the English-language results would still require a significant increase in the size of the available data.

Limitations

The most apparent limitation in the study is the size of the underlying corpus of Polish metaphors. The rapid overfitting observed during training (after only 7 epochs) suggests that a larger dataset is needed for more robust model performance.

Secondly, the high degree of lexicalization in the corpus may have biased the classifier towards recognizing specific words rather than learning broader patterns of metaphorical usage. This limits the model's ability to detect novel or creative metaphors not present in the training data.

Thirdly, the current approach does not account for the contextual nature of metaphors, which can vary depending on the broader discourse or domain. This may result in false positives or negatives when applying the model to texts from different genres or subject areas.

Lastly, the study focuses solely on Polish, and the findings may not generalize to other languages with different linguistic structures or metaphorical conventions. Cross-lingual comparisons and adaptations of the model would require additional research and potentially different approaches.

References

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Elkin Darío Gutiérrez, Guillermo Cecchi, Cheryl Corcoran, and Philip Corlett. 2017. Using Automated Metaphor Identification to Aid in Detection and Prediction of First-Episode Schizophrenia. In (Palmer et al., 2017), pages 2923–2930.

Elżbieta Hajnicz. 2022. Annotation of metaphorical expressions in the basic corpus of Polish metaphors. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 5648–5653, Marseille, France. European Language Resources Association.

Tina Krennmayr and Gerard J. Steen. 2017. VU Amsterdam Metaphor Corpus. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 1053–1072. Springer-Verlag, Dordrecht, Holland.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29. Association for Computational Linguistics.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, LA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Agnieszka Mykowiecka and Malgorzata Marciniak. 2020. Are white ravens ever white? — non-literal adjective-noun phrases in Polish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5871–5877, Marseille, France. European Language Resources Association.

Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors. 2017. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2017)*. Association for Computational Linguistics, Stroudsburg, PA.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. In (Palmer et al., 2017), pages 1537–1546.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. Verb metaphor detection via contextual relation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4240–4251, Online. Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Tryntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Number 14 in Converging Evidence in Language and Communication Research. John Benjamins Publishing Company, Amsterdam, Holland.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. *Metaphor: A Computational Perspective*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

A Appendix

Some properties of the Polish make it difficult to apply BIO encoding. We discuss them in more detail below.

A.1 Discontinuities

We will discuss discontinuities using the original example for BIO encoding:

(2)

Pierwsze	miejsce	zajął	Kowalsk*i*	.
First	*place*	*took*	*Kowalski*	.
O	***B***	***I***	***O***	***O***

Example 3 presents an inversion of Example 2:

(3)

Miejsce	pierwsze	zajął	Kowalsk*i*	.
Place	*first*	*took*	*Kowalski*	.
B	***O***	***I***	***O***	***O***

Example 4 adds a new subordinator to the verb:

(4)

Pierwsze	miejsce	w	konkursie
Place	*first*	*in*	*the contest*
O	***B***	***O***	***O***

zajął	Kowalski	.
took	*Kowalski*	.
I	***O***	***O***

while in both cases it can be argued that the phrase can be reconstructed by combining the word marked B with the following min words marked I (ignoring the words marked O), it is easy to imagine an example where the added element will contain a completely independent metaphor.

Example 5 contains an additional conventional metaphor with a one-element carrier – *huge*. The link between the elements of the expression *take place* is completely lost.

(5)

Pierwsze	miejsce	z	dużą	przewagą
First	*place*	*with*	*huge*	*advantage*
O	***B***	***O***	***B***	***O***

zajął	Kowalski	.
took	*Kowalski*	.
I	***O***	***O***

A.2 Multilevel metaphors

Example 6 illustrates the situation when a certain part of a metaphorical expression is a separate metaphor at the same time:

(6) Największą karierę zrobił likier.
Liqueur has had the greatest career.

It contains three metaphors:

1. a conventional metaphor with a one-element carrier – *great* (in its superlative form);
2. metaphorical expression *have a career*;
3. a metaphor with a one-element carrier *career* (in relation to liquor).

Examples 7 and 8 present the same sentence using a different word order:

(7)

Największą	karierę	zrobił	likier	.
Greatest	*carreer*	*had*	*liqueur*	.
B	***B***	***I***	***O***	***O***

(8)

Likier	największą	zrobił	karierę	.
Liqueur	*greatest*	*had*	*carreer*	.
O	***B***	***B***	***B***	***O***

In Example 7, *career* is tagged by BIO for two different reasons and as a result its metaphoricity is partilly lost, whereas in Example 8 the link between the elements of *have a career* is lost and an unintended metaphor of *have* is potentially created.

A.3 Shared phrase fragments

Omitting one of two adjacent linguistic units (haplology) can also influence automatic detection of metaphors due to strong signals suggesting both the marking of words as metaphorical and as non-metaphorical:

(9) Po fatalnej pierwszej połowie zawodnicy Górnika nawet nie starali się podnieść.
After a disastrous first half, Górnik's players did not even try to pick themselves up.

The problematic word is the reflexive marker „się" associated with both the literal verb „starać" (*to try*) and the metaphorically used „podnieść" (*to pick up*). From a practical point of view, it should be labelled as metaphorical, but since the word „starać" cannot be used without the particle „się", while it is possible to use the word „podnieść" without it, the signals of its metaphoricality may be disturbed.

The same kind of phenomenon can be introduced by coordination:

(10) Rycerz pokonał smoka i pozostałe stojące przed nim przeszkody.
The knight overcame the dragon and other obstacles before him.

In Example 10 we are dealing with both the literal meaning of the verb for *to overcome the dragon* and the metaphorical expression *to overcome obstacles*.