Deep Neural Sequence to Sequence Lexical Substitution for the Polish Language

Michał Pogoda[0000−0002−5011−5573], Karol Gawron[0000−0002−1020−3018], Norbert Ropiak[0000−0003−3616−1298], Michał Swędrowski[0000−0001−8029−6582], and Jan Kocoń[0000−0002−7665−6896]

Wrocław University of Science and Technology, 50-370 Wrocław, Poland
{michal.pogoda, karol.gawron, norbert.ropiak, michał.swedrowski, jan.kocon}@pwr.edu.pl

Abstract. The aim of this paper is to investigate the applicability of language models to the problem of lexical substitution in a strongly inflected language. For this purpose, we focus on pre-trained models based on transformer architectures, in particular BERT and BART. We present a solution in the form of the BART-based sequence-to-sequence model. Then we propose and explore a number of approaches to generate an artificial dataset for lexical substitution, using the adapted PLEWiC dataset as a reference. During this study we focus on Polish as an example of a strongly inflected language

Keywords: Natual Lanugage Processing · Lexical Substitution · Sequence to Sequence.

1 Introduction

Lexical substitution is the task of finding alternative word substitutions that preserve a statement's meaning in context. Its applications include text simplification and paraphrasing. The main challenge is finding a word substitute that not only preserves the meaning of the original sentence but is also grammatically correct and fits the context.

To approach near-human quality in this task, models must perfectly match the relationships between synonymous expressions and their meaning in the context. Often expressions, that by definition are not synonyms can become ones through contextual clarification. For example, a context can restrict the meaning of a hyperonym so that the meaning becomes equivalent to its hyponym).

Her pet was barking loudly

Here we can - with reasonable probability - replace the word *"pet"* with a word *"dog"* even despite the fact that they do not share a synonymy relationship.

Large pre-trained models such as BERT[3] embed a large amount of information regarding word placement in a context. It has been shown [18, 22] that, despite not being trained strictly in a lexical substitution task, they exhibit

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50

state-of-the-art results in similar problems. In this paper, we evaluate a solution analogous to the one proposed in [18] in a strongly inflected language setting (the Polish Language) and compare the approach to the supervised learning of the sequence-to-sequence model on synthetic and natural sets.

This paper contains the following contributions:

- Adaptation of the PLEWiC dataset as a lexical substitution benchmark.
- Comparison of different methods for generating synthetic lexical substitution datasets.
- Evaluation of BERT-based lexical substitution for the Polish language and comparison with fine-tuned BART[12] model.

2 Related Work

One of the earliest approaches to finding word alternatives were based on lexical networks (WordNets) using synonyms. As an example, one can consider the baseline method proposed by D. McCarthy et al. [14]. This approach, however, has some limitations: they rely only on manually annotated relations and omit all words that are not marked as synonyms, but can still be good alternatives to the selected word; they do not use the context of the word being replaced.

In a highly influential work, T. Mikolov at al.[16] proposed the idea of unsupervised word embeddings becoming a source of inspiration for approaches like the usage of skip-gram-based word representations [15] to perform lexical substitution based on vector distance.

Modern approaches are mostly based on deep language models. N. Arefyev et al.[1] showed that large pretrained models can give better performance than previous unsupervised and supervised methods of lexical substitution. They also noted that for BERT-type models whose tokenization is subword, a multi-token prediction could give better results.

According to our knowledge, for SemEval[14] and CoinCo[10] benchmarks, the best models for English are based on transformer architectures. W. Zhou et al.[22] present methods for target masking: full, partial, none. They also show a positive effect on the model proposal sorting method's performance, which is based on the cosine similarity of the embeddings of the proposed sentences to the original sentences.

To address the high computational complexity of sorting, instead of using the original model, distilled versions of models can be used, additionally trained to evaluate the sentence semantic similarity (STS). N. Reimers et al. [20] show the advantage of such a trained model in the STS task and make the distilUSE model available as part of the Sentence Transformers framework[1].

In a paper on a similar lexical simplification task, J. Qiang et al. [18] present a method that does not hide part of the information from the model (as in partial masking, which can make it challenging to find suitable alternatives) but

[1] https://www.sbert.net/

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50

combines the original sentences with the sentence where the selected (target) word is masked.

Encoder-only architectures like BERT pose a significant computational complexity problem for multi-token prediction. To get around it, a Seq2Seq BART model can be used. L. Martin [13] applied it to the sentence simplification task, which leads us to believe that it would also work well for the lexical substitution task. In the following subsections, we summarize how substitutions are obtained in the models trained with the masked language modeling objective. In our work, applying ideas from the presented papers, we compare the model used in the current SOTA i.e. BERT, and an approach using the BART sequence-to-sequence model for the task of lexical substitution in the Polish Language.

2.1 Full Masking

One of the simplest approaches involves replacing a piece of text with a mask token (or multiple tokens) [22]. The problem with this approach is that the semantic meaning of the masked text fragment is completely lost. For example, if we replace the word *cat* in the sentence *She has a cat* with a mask, the linguistically correct responses will be tokens such as *car* or *computer* (see Figure 1).

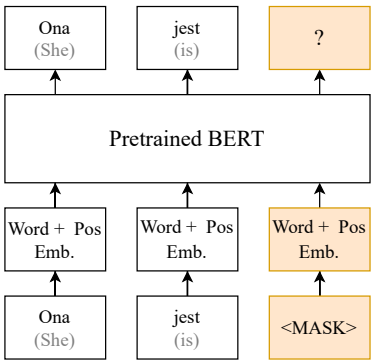

Fig. 1. Full masking approach. We replace a word with a `<MASK>` token (or tokens), and the model finds a substitution based purely on context. It does not have any information on what word was initially used, and substitutions will likely be returned based only on the frequency of occurrence in a given context.

2.2 Partial Masking

The second approach, that partially address the issue with full masking, is the so-called partial masking [22]. Instead of completely replacing the selected fragment with a mask, dropout can be applied after the initial embedding of the input

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50

tokens. The assumption is that with this approach, some semantic information is preserved, while at the same time, the model has such little information about the initial token that it can propose insertions other than the trivial copy of the input token (see Figure 2).

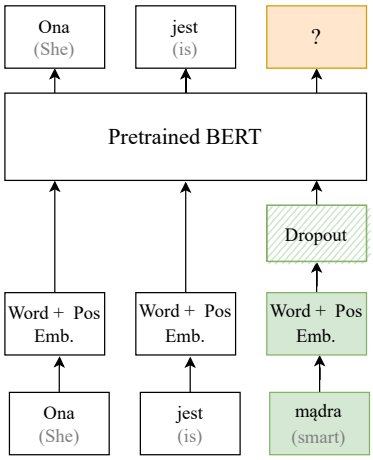

Fig. 2. Partial masking approach. Dropout is applied after initial embedding. That makes it harder for the model to just copy the input with nearly 100% confidence.

2.3 Cross-sentence Relationship

To our best knowledge, currently, the most promising approach is to use language models that have been trained on the Next Sentence-Prediction-task [18]. In particular, the original BERT models, as later approaches, often dropped this task from pretraining. J. Qiang et al. [18] exploited the fact that to perform next-sentence predictions, the model must learn some semantical relationship between the sentences. The authors empirically showed that if we prepare the input to the model in a manner analogous to the examples in the next-sentence-prediction task, but instead of two consecutive sentences, we repeat the same sentence once in its entirety and once with the target word masked, the model will suggest semantically similar words in place of the mask. This approach is shown in Figure 3. Experiments we performed in Section 8 prove, that this kind of approach works noticeably better than raw MLM and dropout methods for generating semantically similar words.

3 Multi-token Prediction

The problem of converting single-token expressions into multi-token ones (and vice versa) is critical in strongly inflected languages (such as the Polish Lan-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50

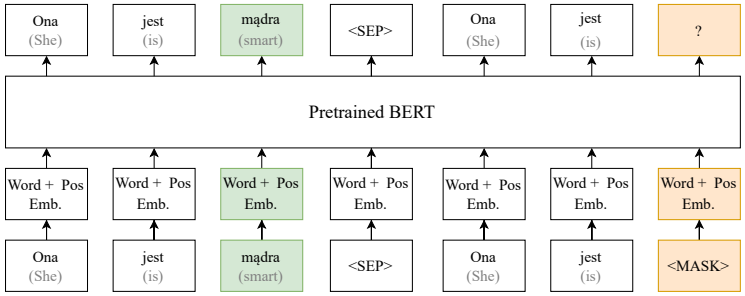

Fig. 3. BERT using cross-sentence relationship. The first sentence is provided without any changes, whereas the second one has mask tokens in place of the desired substitution.

guage) because the limited number of tokens is not able to cover all inflectional varieties even for relatively popular words and, as a result, multi-token words are a frequent phenomenon. An issue with using encoder-only models (like BERT) is that the number of predicted tokens in the output always matches the number of tokens masked in the input. This means that the model, knowing how many tokens are masked, will only predict the words that fully fit into space. As a result, when converting a single-token expression to a multi-token expression (or vice versa), multiple passes would have to be performed, at least one for each possible length of the target expression (see Figure 4).

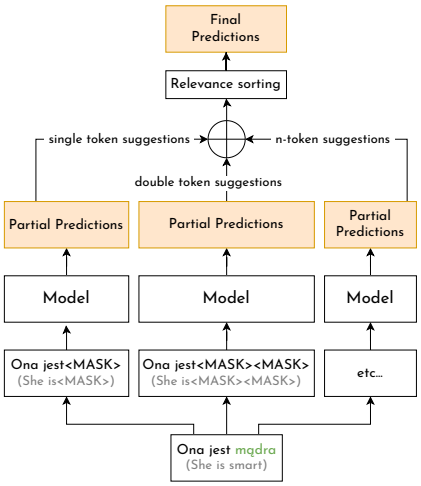

Fig. 4. Multi-token prediction with an encoder only. If token-length of possible substitution is not known in advance, we would have to perform inference for every possible substitution length.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50

For this reason, we also investigated the BART model, which does not cause this problem. The reasons for choosing BART are:

- Seq2Seq architecture: During inference, the model is not restricted to fill n-tokens space only. Because of that, we can perform a single beam-search to get predictions for multiple-tokens lengths.
- One of the tasks on which the model is pre-trained is multi-token infilling. This task is similar to the problem on which we want to perform model fine-tuning.
- Availability for the the Polish language: There exists a model pre-trained only on the Polish corpora (which can be a significant advantage over multilingual models).

4 Datasets

To perform fine-tuning and model testing, we selected the PlEWi Corpus – PLEWiC [5]. It is a collection of language errors for the Polish language, based on the editing history of the Polish version of Wikipedia. From the perspective of this work, the subset of stylistic errors is essential, especially edits based on synonyms. The error type is automatically annotated using a set of hand-crafted heuristics. The authors manually evaluated 200 samples from each category, and in the case of synonyms replacement, they claimed that their heuristics achieved 99% precision. We used a subset of examples from the dataset whose substitution is marked as a synonymous replacement. Unlike other datasets in this field, this one has, in general, only one correct substitution for each element. For training on this dataset, we used 18635 samples for training and 6212 samples for validation.

We also generated a synthetic training set based on the KPWr Corpus [17] and excerpts from the Polish version of the book *Sherlock Holmes* manually annotated with wordnet senses [6]. The exact process of generating the collection is described in Section 6.

5 Sequence-to-Sequence Substitution

The BART[12] model is a complete encoder-decoder transformer, and both the input and output are text. Originally, BART was pre-trained on a variety of denoising tasks, including multi-token masking. For example, from the text:

John really likes Anna.

Randomly selected tokens (both their number and position) were replaced by a single mask token, e.g.:

John<mask>Anna.

The purpose of this task is to reproduce the original sentence. In this way, a conditional language model is produced. Unfortunately, there is no information

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50

about the semantics of the masked part of the sentence in the original problem formulation. Therefore, the predicted substitutions' order depends only on the frequency of use in a given context. For this reason, we designed a new task that is as close as possible to the original pre-training task but includes the necessary information by prepending the masked sentence with the original content as follows:

[really likes] John<mask>Anna.

This gives the model (at least in theory) easy access to the masked fragment's semantic content. As we wanted the problem to be as similar as possible to the pre-trained definition (due to the scarcity of training data), we defined the model's output to be the whole sentence (like in a masking pre-training task) rather than just the substituted part. To extract the substituted part for evaluation, we truncate the output from the beginning and the end of the sentence by the number of non-masked characters in the input sentence. Unfortunately, here the most considerable advantage, which is the Seq2Seq architecture, is at the same time the most significant problem. We have no guarantee that the model will not change anything except the masked fragment. For example, the output of the model for the presented example could theoretically be:

John adore Kristina

Furthermore, the original BART model was pre-trained to remove words, and we could still observe such behavior after fine-tuning, making the substitution extraction a problematic task.

6 Synthetic Dataset Generation

The PLEWiC collection - despite its relatively good quality - includes edits of mainly single words. Its language domain is also limited due to the presence of Wikipedia texts and the fact that there are only examples classified as synonyms by specific heuristics. For those reasons, in our work, we explore the possibility of using synthetically generated datasets. The idea behind this approach is to take any corpus, then perform part-of-speech tagging using WCRFT2 [19], run a word sense disambiguation WoSeDon tool [7], use plWordNet to find potential synonyms, and finally use morphosyntactic dictionary Morfeusz 2 [8] to apply original word form to a synonym (see Figure 5). We formulated three methods to generate the training set.

6.1 Single-word replacement (SWR)

After performing disambiguation, we check the synset size of each word in the sentence. If the synset's size to which a word belongs is greater than one, we generate input and output sentences for each combination of two lexical units from the synset (see Figure 6). To preserve the correct form of the words, for each combination of lexical units, we use the synthetic dataset generation pipeline described above (Figure 5). The dataset was split into training (88182 entries) and validation (33766 entries) part.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50

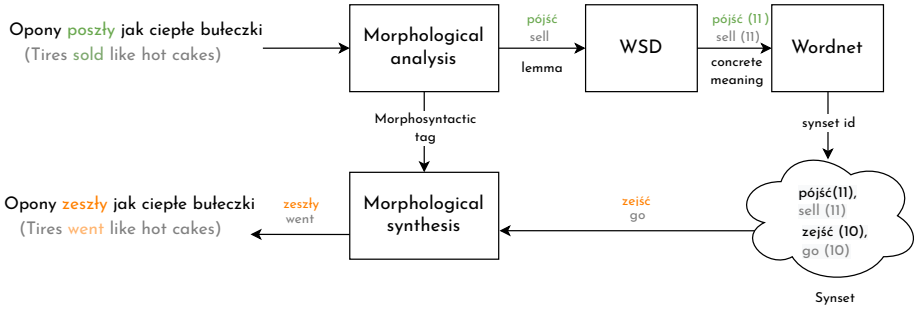

Fig. 5. General synthetic dataset generation schema.

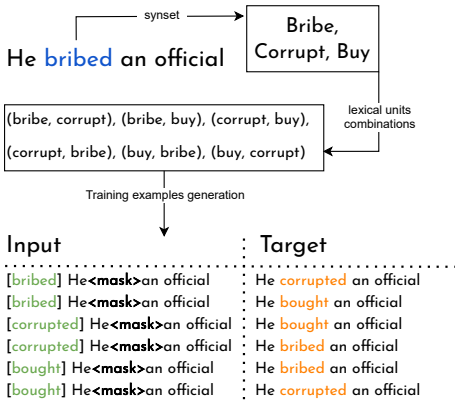

Fig. 6. Single-word replacement dataset generation. After finding a word suitable for substitution, we create all possible ordered combinations of lexical units from the synset to generate training pairs.

6.2 Single word and context replacement (SWCR)

In the SWR method, there are many examples with identical contexts in the dataset (e.g., if we have a synset of size 3 for a single word in a given sentence, there are six training examples with identical contexts). We apply a second approach to increase the context diversity and model robustness against irrelevant context changes, where words that are not masked are also replaced when generating the dataset. The words substituted outside of the masked area are the same in the input sequence and the output sequence. The process is illustrated in Figure 7. For the SCWR dataset, the training part had 3205239 samples, and the validation part had 1359584 samples.

6.3 Substitution of a longer sentence fragment (SLSF)

In a sentence corpus, we randomly select the beginning and end of the fragment we are replacing. We then search for words with synset greater than one in

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50

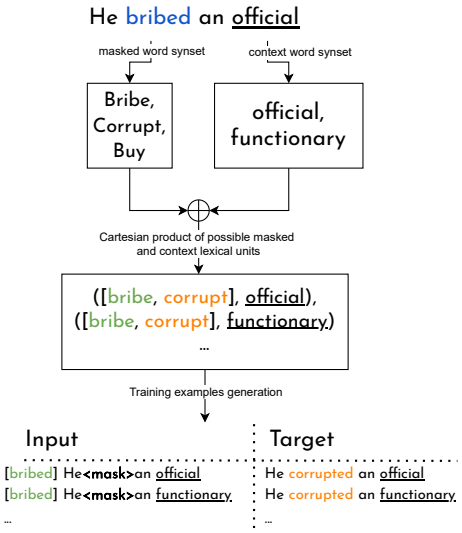

Fig. 7. Single word and context replacement (SWCR). We create an ordered combination of possible substitutions of masked-word synset and words outside of synset. Then we combine both sets (also as a cartesian product) to create a set of training examples.

this fragment and generate all possible combinations from them. The sentences formed from a pair of two different combinations are training examples. The process is depicted in Figure 8. The size of the training part of the SLSF dataset has 97562 entries and the validation had 35086 entries.

6.4 Wordnet-based Dataset Generation

During the manual evaluation of the synthetic substitution examples, we noticed that, despite using WoSeDon, words are still empirically often assigned to incorrect meanings - e.g. sentence "Plane is a flat, two-dimensional surface." could lead to generation of "Aircraft is a flat, two-dimensional surface.". Thus, especially when several words in one sentence are replaced, it is sometimes difficult to guess the original semantic meaning of the sentence. To overcome the limitations of WoSeDon, we took usage example word usage sentences from plWordNet (Polish Wordnet, [4]) as base sentences. In this case, there was no problem with the word sense disambiguation task because we can assume with high probability that the occurrence of the lemma of the analyzed word in the usage example can be considered as the meaning of the lexical unit for which the example was created (see Figure 9). For this dataset, the training part consists of 37436 examples and the validation part consists of 13012 examples.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50

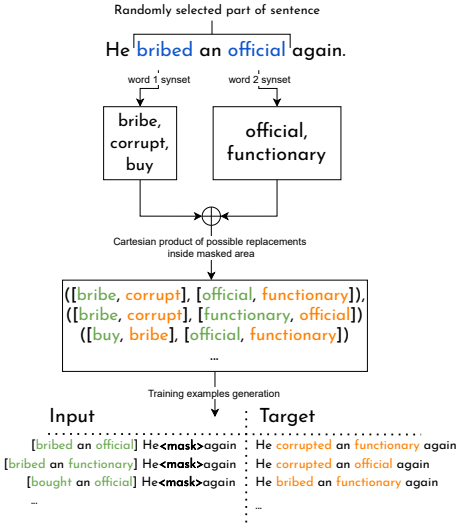

Fig. 8. Substitution of a longer sentence fragment (SLSF). Multiple words are replaced at once so that the model can learn multi-word substitutions.

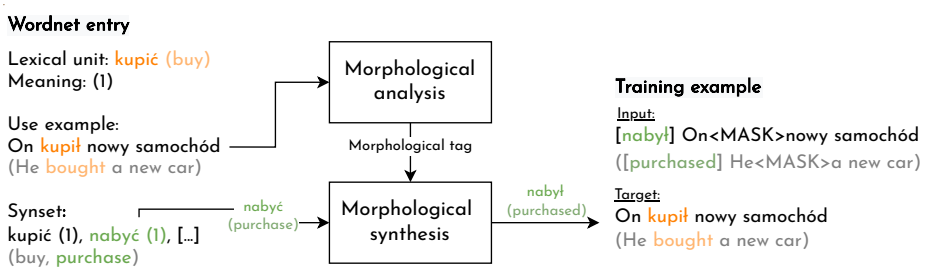

Fig. 9. Generating artificial dataset from plWordNet. The approach is similar to SWR approach. However, WSD stage is omitted as the example is already assigned to the concrete meaning.

7 Experiment setup and initial results

We used pre-trained Polish BERT[11] and pretrained Polish BART*"A Repository of Polish NLP Resources"*[2] models compatible with the Huggingface Transformers framework [21]. We trained BART model with Adam[9] optimizer for three epochs, with a learning rate 3×10^{-5}, batch size of 2 with two gradient accumulation steps.

Initial tests of the model show that it can perform well with replacing one or more words while preserving the sense of the sentence, at least in a few

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50

propositions. However, it could be noticed that the proposed alternatives are quite often antonyms. For example:

Input:	Było dziś bardzo **[zimno]**, mimo że pora roku na to nie wskazuje *It was very **[cold]** today, even though the time of year doesn't indicate it.*
Substitutions:	ciepło, chłodne, wilgotno, gorąco *warm, chilly, humid, hot*

This is a group of errors that could be expected when using models based on MLM training, as both antonyms and synonyms will often have similar context and this behavior still remains even after fine-tuning. Another problem among the errors spotted in generated sets are phrases, such as *red card* or *broke a leg*, which are not correctly converted into words that maintain the sense of the whole phrase, but only the sense of the individual words of the phrase. For example:

Input:	Zawodnik otrzymał **[czerwoną kartkę]** z powodu zbyt agresywnej gry. *The player received a **[red card]** due to playing too aggressively.*
Substitutions:	czerwoną/żółtą/brązową/fioletową kartkę *red/yellow/brown/purple card*

The last difficulty, particularly in the technical aspect, is a problem arising from the nature of the seq2seq model. Although we provide a mask in the place where the model's suggestion should appear, the model can replace the rest of the sentence as well. If words outside the masked area are changed, it can be difficult to extract only the desired part of the substitution:

Input:	Reżyser chciał pokazać, że **[potrafi]** nakręcić coś nieoczekiwanego. *The director wanted to show that he **[knew how to]** shoot something* unexpected.
Substitutions:	umie, e umie, mie *could, e could, coul*

8 Evaluation

To examine the substitutions' semantic quality and grammatical correctness, we used a test sample from the PLEWiC dataset. To evaluate the quality of the models, we selected a quality measure of Recall at 10 and Recall at 5. There is always only 1 value assigned as correct in the PLEWiC dataset, so Recall at K for a single sample can only reach values of 0 and 1. In the rest of this paper, by Recall at K we mean the Recall at k averaged over all samples in the PLEWiC test set, so it tells us how often the target phrase appeared in the K best predictions of the model. We used the pre-trained (i.e.: without any fine-tuning) BERT as the base model, where only one token is masked, and only one

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50

token can be substituted. Other models used for the evaluation are BART-based models fine-tuned using: PLEWiC dataset (model: B-PLEWiC), SWR, SWCR, SLSF generated synthetic datasets described in Section 6 (models: B-SWR, B-SWCR, B-SLSF) and wordnet-based dataset described in Section 6.4 (model: B-WN).

Table 1 shows the final results on the PLEWiC test set of various models, differing mainly in the set on which they were trained. The evaluation consisted of comparing the model proposal with the word that was actually used. The first 500 records from the PLEWiC collection were used for testing. The results show the percentage of cases in which the correct word appeared in the first N model proposals. Prediction is counted as correct only if both lemma and form were correct (i.e., only "perfect" match counted).

The best models turned out to be those learned on mixed datasets, and among these, the one trained without adding the Wordnet dataset turned out to be better according to Recall at 1. However, for Recall at 10, the better one was the model trained on the mixed dataset with all synthetic sets: Wordnet and PLEWiC.

Table 1. Result [Recall@N] on PLEWiC test set of BERT baseline and BART models with different fine-tuning datasets. *The BERT model in most cases gives the original occurring word as the first proposition, hence the low Recall@1 score. In comparison, the Recall@2 for BERT is 28.8%.

Model / dataset	**Recall@1**	**Recall@5**	**Recall@10**
baseline w/o fine-tuning			
BERT	2.9*	42.3	58.9
PLEWiC only			
B-PLEWiC	56.4	73.6	76.6
synthetic only			
B-SLSF	11.4	33.8	43.8
B-SWR	13.6	35.4	44.6
B-SWCR	14.6	35.6	43.8
B-WN	7.2	36.2	47.4
mixed datasests			
B-PLEWiC+SLSF	**56.8**	**74.4**	77.6
B-PLEWiC+SLSF+WN	55.0	**74.4**	**78.0**

9 Results and Future Work

The models based on synthetically generated datasets were inferior to the dataset developed from actual Wikipedia editions (PLEWiC dataset) and the BERT-based baseline. The following causes may have contributed to this result:

– Introduction of grammatical structure errors. In the presented algorithm, we did not take context editing into account. Thus, when examining exact

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50

substitutions, a reduction in the quality of inflectional forms could have significantly reduced the results obtained.
– A difference in the domain of texts. BERT was pre-trained on a dataset consisting of Wikipedia (which is the base for PLEWiC). On the other hand, BART was fine-tuned on KPWr and Sherlock Holmes novel.

Fine-tuning on the PLEWiC set significantly improved the results relative to baseline in the form of pre-trained BERT. The already obtained Recall at 1 value was higher than the baseline Recall at 10. Additionally, the use of the Seq2Seq model significantly simplified the problem in terms of computational complexity, as variable-length output is supported. Adding a synthetic dataset did improve overall results, however the impact was not significant.

Further research will include, in particular, study of the impact of sequence-to-sequence query definition on the final performance of the system and explore the possibility of incorporating active learning into the process of fine-tuning deep lexical substitution models.

Acknowledgements This work was financed by (1) the National Science Centre, Poland, project no. 2019/33/B/HS2/02814; (2) the Polish Ministry of Education and Science, CLARIN-PL; (3) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19; (4) the statutory funds of the Department of Artificial Intelligence, Wrocław University of Science and Technology.

References

1. Arefyev, N., Sheludko, B., Podolskiy, A., Panchenko, A.: Always keep your target in mind: Studying semantics and improving performance of neural lexical substitution. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 1242–1255 (2020)
2. Dadas, S.: A repository of polish NLP resources. Github (2019), https://github.com/sdadas/polish-nlp-resources/
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Dziob, A., Piasecki, M., Rudnicka, E.: plwordnet 4.1–a linguistically motivated, corpus-based bilingual resource. In: Fellbaum, C., P., V., E., R., M., M., M., P. (eds.) Proceedings of the 10th Global WordNet Conference: July 23-27, 2019, Wroclaw (Poland). pp. 353–362. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław (2019)
5. Grundkiewicz, R.: Automatic extraction of polish language errors from text edition history. vol. 8082 (09 2013). https://doi.org/10.1007/978-3-642-40585-3_17
6. Janz, A., Chlebus, J., Dziob, A., Piasecki, M.: Results of the PolEval 2020 Shared Task 3: Word Sense Disambiguation. In: Proceedings of the PolEval 2020 Workshop. Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland (2020), http://poleval.pl/files/poleval2020.pdf

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50

7. Kedzia, P., Piasecki, M., Orlinska, M.: WoSeDon (2016), http://hdl.handle.net/11321/290, CLARIN-PL digital repository
8. Kieraś, W., Woliński, M.: Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego. Język Polski **XCVII**(1), 75–83 (2017)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2015)
10. Kremer, G., Erk, K., Padó, S., Thater, S.: What substitutes tell us-analysis of an "all-words" lexical substitution corpus. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 540–549 (2014)
11. Kłeczek, D.: Polbert: Attacking polish nlp tasks with transformers. In: Ogrodniczuk, M., Łukasz Kobyliński (eds.) Proceedings of the PolEval 2020 Workshop. Institute of Computer Science, Polish Academy of Sciences (2020)
12. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. pp. 7871–7880 (01 2020). https://doi.org/10.18653/v1/2020.acl-main.703
13. Martin, L., Fan, A., de la Clergerie, É., Bordes, A., Sagot, B.: Multilingual unsupervised sentence simplification. arXiv preprint arXiv:2005.00352 (2020)
14. McCarthy, D., Navigli, R.: The english lexical substitution task. Language resources and evaluation **43**(2), 139–159 (2009)
15. Melamud, O., Levy, O., Dagan, I.: A simple word embedding model for lexical substitution. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. pp. 1–7 (2015)
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)
17. Oleksy, M., Marcińczuk, M., Maziarz, M., Bernaś, T., Wieczorek, J., Turek, A., Fikus, D., Wolski, M., Pustowaruk, M., Kocoń, J., Kędzia, P.: Polish corpus of wrocław university of technology 1.3 (2019), http://hdl.handle.net/11321/722, CLARIN-PL digital repository
18. Qiang, J., Li, Y., Zhu, Y., Yuan, Y., Wu, X.: Lsbert: A simple framework for lexical simplification. arXiv preprint arXiv:2006.14939 (2020)
19. Radziszewski, A.: A tiered crf tagger for polish. In: Intelligent tools for building a scientific information platform, pp. 215–230. Springer (2013)
20. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4512–4525 (2020)
21. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://aclanthology.org/2020.emnlp-demos.6
22. Zhou, W., Ge, T., Xu, K., Wei, F., Zhou, M.: Bert-based lexical substitution. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3368–3373 (2019)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_50