Evaluating Data Augmentation Techniques for the Training of Luxembourgish Language Models

Isabella Olariu[†⋆], **Cedric Lothritz**[⋆], **Tegawendé F. Bissyandé**[⋆], **Jacques Klein**[⋆]
[†] Zortify S.A.
[†] 9, Rue du Laboratoire, L-1911 Gare Luxembourg
[⋆]University of Luxembourg
[⋆]6, Rue Coudenhove-Kalergi, L-1359 Luxembourg
isabella@zortify.com
{cedric.lothritz, tegawende.bissyande, jacques.klein}@uni.lu

Abstract

Training large language models is challenging when data availability is limited, as it is the case for low-resource languages. We investigate different data augmentation techniques for the training of models on Luxembourgish, a low-resource language. We leverage various word substitution methods for artificially increasing textual data: synonym replacements, entity replacements and modal verbs replacements. We present DA BERT and LuxemBERT-v2, two BERT models for the Luxembourgish language. We evaluate our models on several downstream tasks and conduct an ablation study to assess the impact of each replacement method. Our work provides valuable insights and highlights the importance of finding solutions to training models in low-resource settings.

1 Introduction

Neural network models are data-hungry, making them challenging to exploit when resources are scarce. The development of Natural Language Processing (NLP) tools for low-resource languages is, however, important since a large number of people around the world predominantly speak a language that can be classified as under-resourced due to its shortage in available data (Feng et al., 2021). Therefore, the research community is looking for ways to get extra data for training models targeting low-resource languages. Data augmentation is a common practical way of generating synthetic data by slightly altering existing data.

Luxembourgish, the national language of Luxembourg, is an example of a low-resource language, in a country that is known as being multilingual: in addition to Luxembourgish, German, French, English, Portuguese and Italian are widely spoken among its citizens. Only about 430 000 citizens (Eberhard et al., 2022) speak Luxembourgish as their native language. Given the limited number of speakers, textual data in Luxembourgish is not abundant. LuxemBERT is an existing language model for Luxembourgish and was developed by Lothritz et al. (2022) for use cases mainly targeted to the financial technology (FinTech) domain. To address the limitation of insufficient data, the authors develop a novel data augmentation technique leveraging automatic translation of common words from a closely related language.

In this study, we investigate the effectiveness of data augmentation techniques other than the one used by Lothritz et al. (2022). We use synonym, entity, and modal verb replacements to create new data for building Luxembourgish language models.

We explore the following research questions:

RQ1: What impact on the model's performance can we observe when we modify its input data through data augmentation techniques?

RQ2: Which data augmentation technique has the highest impact on our model's performance?

The contributions of this paper are threefold: **(i)** we contribute to the community with new pre-trained models for Luxembourgish; **(ii)** we provide insights on the effectiveness of existing data augmentation techniques for low-resource language modeling; **(iii)** we assess, from a different perspective, the relevance of the data augmentation proposed in LuxemBERT by discussing the added value of traditional data augmentation techniques.

2 Related Work

One of the most common choices of language models for many low-resource languages is mBERT (Pires et al., 2019; Wu and Dredze, 2020), a multilingual BERT model (Devlin et al., 2019). mBERT was trained on 104 languages, one of which is Luxembourgish. Even though mBERT includes a range of low-resource languages, Wu and Dredze (2020) do not recommend using it as the only option for low-resource languages. It was trained solely on Wikipedia articles, therefore its ability to

Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023), pages 174–179
September 18–22, 2023. ©2023 Association for Computational Linguistics

learn and understand a language decreases notably the smaller the Wikipedia size of the respective language is.

LuxemBERT is a recent Luxembourgish BERT model (Lothritz et al., 2022). The authors implement a data augmentation technique based on partial translation to train this model. They augment the training data by incorporating text data from an auxiliary language, German, which is structurally closely related to Luxembourgish. Specifically, they translate a subset of common and unambiguous German function words (e.g. pronouns, determiners, prepositions) to Luxembourgish.

There are several other data augmentation techniques that prove to be useful when working with limited data (Hedderich et al., 2021; Xu et al., 2019). The idea is that because there is not enough data for low-resource languages, the existing data has to be leveraged as efficiently as possible through various augmentation techniques which makes it possible to generate more data without collecting additional samples. Hedderich et al. (2021) differentiate between approaches performed on a word or sentence level. They suggest replacing words with synonyms and named entities of the same type on a token level. On a sentence level, they propose using back-translation to create more diverse sentences. This approach translates a sentence in a source language to a sentence in a target language, before translating it back to the source language (Sennrich et al., 2016). Pellicer et al. (2023) propose paraphrasing as an efficient strategy to add lexical diversity while retaining the original meaning. Negation is another approach that creates new sentences by reversing the meaning of the original ones (Tarasov, 2020).

3 The Data

Pre-Training Data. This dataset was used in the pre-training corpus of LuxemBERT (Lothritz et al., 2022), which consists of a total of 12 million sentences, out of which six million are Luxembourgish and six million are partially translated German sentences. It was collected from different sources including news articles, chatrooms, user comments posted on Radio Television Luxembourg (RTL),[1] a Luxembourgish news station website, and the Luxembourgish Wikipedia. Lothritz et al. (2022) provide further details on the breakdown of the pre-training corpus.

Data for Data Augmentation. We use the existing six million Luxembourgish sentences from LuxemBERT to create the same number of new (augmented) sentences. Furthermore, to perform word substitutions via synonym, entity, and modal verb replacements, for our data augmentation task we collect additional data from the Luxembourgish Online Dictionary[2] consisting of Luxembourgish modal verbs, first names, surnames and locations (e.g. countries, cities, etc.). We also create a dictionary consisting of Luxembourgish words and corresponding synonyms.[3]

Data Augmentation Scheme. Our data augmentation scheme is applied to the six million Luxembourgish sentences that LuxemBERT was trained on and checks for each word whether that word is in one of our lists or dictionary. If there is a match with words from the original data, we replace those matches with random words from the corresponding lists.

The systematic substitution of words from LuxemBERT's training data with words from our lists allows us to obtain new sentences containing different words without considerably changing the meaning of the original sentences. Following these steps, we create six million new Luxembourgish sentences, for a total of 12 million, the same number of sentences used for LuxemBERT.

4 Experimental Setup

In this section, we introduce our novel models and the baselines we compare them against, describe the training and fine-tuning specifications and formulate the set of experiments consisting of five downstream tasks to evaluate our models on.

4.1 Models

As mentioned in Section 1, we compare two new BERT models to LuxemBERT to assess the impact of our data augmentation scheme. We describe our two models, DA BERT and LuxemBERT-v2, which we trained using an augmented dataset.

DA BERT: Data Augmented BERT is a model which we build and pre-train completely from scratch using the data obtained through our data augmentation scheme. The configuration specifications are the same as for Lothritz et al. (2022) and are as follows: a vocabulary size of 30 000, 12

[1] https://www.rtl.lu/

[2] https://lod.lu/

[3] Data available at https://github.com/iolariu/Data-Augmentation

attention heads, 12 hidden layers, and maximum sequence length of 512.[4]

LuxemBERT-v2: This model is also trained with augmented data. We do not pre-train this model from scratch, but continue pre-training LuxemBERT by adding more data. To the original 12 million Luxembourgish sentences, we add our new 6 million augmented sentences to obtain a final dataset of 18 million sentences.

4.2 Training Parameters

To configure our DA BERT and LuxemBERT-v2 models, we re-use the same parameters as LuxemBERT (Lothritz et al., 2022) originating from the BERT-base model (Devlin et al., 2019): 12 Transformer blocks, 768 hidden layers, 12 self-attention blocks, and a total of 110 million trainable parameters. We choose a tailored alphabet size of 120 characters as for LuxemBERT to take into account the Luxembourgish alphabet by restricting the characters to letters used in the Luxembourgish language.

We pre-train our model on the Masked Language Modeling task and leave out Next Sentence Prediction due to the largely unordered nature of our dataset. We pre-train our model for 10 epochs using a masking probability of 15%.

4.3 Baselines

We examine two baseline models for comparison purposes: mBERT and the original LuxemBERT.

mBERT: The multilingual BERT model was trained on a mixture of high- and low-resource languages. Luxembourgish is one of the included languages and this part was trained on the Luxembourgish Wikipedia data, which contained 59 000 articles at the time of release of the model. The architecture consists of 12 Transformer blocks, 768 hidden layers, 12 self-attention blocks, and 110 million trainable parameters (Devlin et al., 2019).

LuxemBERT: We consider LuxemBERT as another baseline model, which is one of the currently existing BERT-based models for the Luxembourgish language. LuxemBERT and DA BERT use the same configurations in terms of model architecture, training parameters, and dataset size.

4.4 Downstream Tasks

To evaluate the performance of our language models, we fine-tune them on the same five downstream tasks as in Lothritz et al. (2022).

POS Tagging. This sequence labelling task consists of assigning to each word in a given sequence of words a specific grammatical class (Jurafsky and Martin, 2008). We use the dataset provided by Lothritz et al. (2022), which consists of 450 Luxembourgish news articles and 5500 sentences. It is labelled with 15 POS tags including verbs, pronouns, adjectives, and adverbs.

Named Entity Recognition. This sequence-to-sequence task extracts key information in a given piece of text. It assigns a label to each word in a sentence by locating and classifying proper names in the sentence. We use the same dataset as for POS tagging,for which we have five labels: person, organisation, location, geopolitical entity, and miscellaneous.

Intent Classification.[5] Sometimes also referred to as intent recognition, this task tries to find an author's intention given an extract of text, where the labels of the intents are determined in advance. We use the Banking Client Support dataset created by Lothritz et al. (2021), which consists of 28 intents associated to various banking requests, such as checking bank account balances, opening and closing bank accounts, or ordering a new credit card.

News Classification. This task consists of correctly classifying news articles into various topics such as politics or sports. The dataset was created by Lothritz et al. (2022) and consists of 10 052 Luxembourgish news articles, which can be classified into eight topics.

Winograd Natural Language Inference. This task consists of a pair of texts A and B, where text A contains one or several pronouns and text B contains a substring of text A, where the pronoun in text B is replaced by either a word or a name. The label is 1 if the pronoun was replaced with the correct token from text A, or 0 otherwise. We use the original WNLI dataset (Levesque et al., 2012) translated into Luxembourgish by Lothritz et al. (2022).

[4] Models available at https://huggingface.co/iolariu/DA_BERT and https://huggingface.co/iolariu/LuxemBERT-v2

[5] We distinguish between IC a and IC b, where we use all labels for IC a, but leave out trivial intents (e.g. *greeting*, *thanking*, *goodbyes*) for IC b.

Fine-tuning Parameters. To allow for a fair comparison, we choose the values of the fine-tuning hyperparameters identical to those used for LuxemBERT. Details for the chosen values can be found in Lothritz et al. (2022).

5 Experimental Results

In this section, we present the results from our experiments across six downstream tasks and address the research questions introduced in Section 1. For each task, we fine-tune the pre-trained models over five runs and take the average of the performance of each run as our final evaluation measure. The F1 scores for each model on each task are reported in Table 1.

5.1 RQ1: What impact on the model's performance can we observe when we modify its input data through data augmentation techniques?

Table 1 shows the results of the fine-tuned models. We observe an improvement in performance of our data-augmented DA BERT and LuxemBERT-v2 models on certain downstream tasks. DA BERT outperforms all models on NER and IC b tasks. For IC a, it outperforms mBERT as well as LuxemBERT-v2. For NC, the performance of mBERT, DA BERT, and LuxemBERT-v2 are equivalent; all of them perform just slightly worse than LuxemBERT. For POS tagging, LuxemBERT-v2 reaches the same performance as LuxemBERT, outperforming both mBERT and DA BERT. Furthermore, LuxemBERT-v2 outperforms mBERT on NER, IC a, and IC b. Finally, on WNLI which can be considered as the hardest task, LuxemBERT-v2 outperforms DA BERT, but none of the models perform better than mBERT on that task.

5.2 RQ2: Which data augmentation technique has the highest impact on our model's performance?

We perform an ablation study to answer this research question which allows us to identify the effects of individual augmentation techniques. We compare the difference between applying only synonym replacements or entity replacements to the data. For this purpose, we pre-train two smaller models that we compare against a baseline model described below.

BASELINE-BERT This is a smaller BERT model that is trained only on the Luxembourgish Wikipedia data, which consists of half a million sentences. We use this model as a baseline to compare two same-sized models against for which we separately perform synonym and entity replacements.

BERT-SYNS This model is trained on a synonym-augmented Luxembourgish Wikipedia data. We generate a total of 465 070 sentences to double the corpus size compared to the one of BASELINE-BERT.

BERT-ENTS This model is also only trained on Wikipedia data, this time augmented with entity replacements. For the dataset for this model, we generate 494 241 new sentences.

As shown in Table 2, BERT-ENTS outperforms BASELINE-BERT and BERT-SYNS on four out of six downstream tasks. In contrast, BERT-SYNS outperforms BASELINE-BERT and BERT-ENTS only on one task, suggesting a tendency towards using entity replacements for better outcomes.

6 Discussion

Overall, we believe that data augmentation for our Luxembourgish language models is beneficial despite the mixed conslusions of results. DA BERT and LuxemBERT-v2 consistenly outperform mBERT on most tasks except WNLI. This could be because mBERT lacks training on augmented text data and relies merely on Wikipedia articles for each language. Low-resource languages with small Wikipedia articles perform significantly worse with mBERT. DA BERT and LuxemBERT-v2 perform better due to various data augmentation techniques, which provide more training data.

Nevertheless, mBERT performs best in the challenging WNLI task. Training data for this task is relatively small, potentially hindering the learning ability of DA BERT and LuxemBERT-v2. LuxemBERT also fails to outperform mBERT on this task. We suppose that more training examples or considering some task-specific architectural modifications could help better capture the information required for WNLI.

Lastly, inconsistent findings from our ablation study suggest that several factors could influence why a certain technique is more suitable for a specific task. For instance, entity replacements seem to help NER, whereas other techniques might fall short on properly understanding context or lack in entity diversity for that task.

Models	POS	NER	IC a	IC b	NC	WNLI
mBERT	88.6 ± 0.1	68.9 ± 1.0	46.0 ± 5.6	48.3 ± 9.4	90.0 ± 0.5	**57.3** ± **0.0**
LuxemBERT	89.0 ± 0.1	70.0 ± 0.8	**72.5** ± **1.1**	70.9 ± 1.8	**91.8** ± **0.2**	54.6 ± 1.6
LuxemBERT-v2	**89.0** ± **0.0**	69.4 ± 0.0	67.6 ± 2.5	68.0 ± 1.0	90.0 ± 2.2	55.0 ± 0.0
DA BERT	88.7 ± 0.0	**70.8** ± **0.0**	71.7 ± 2.0	**73.8** ± **2.2**	90.0 ± 2.8	52.0 ± 0.0

Table 1: Comparison of results of our fine-tuned models on downstream tasks.

Models	POS	NER	IC a	IC b	NC	WNLI
BASELINE-BERT	88.0 ± 0.0	59.4 ± 0.0	56.9 ± 5.2	55.8 ± 3.8	85.7 ± 0.0	51.8 ± 0.0
BERT-SYNS	**88.0** ± **0.0**	61.8 ± 0.0	55.8 ± 2.4	55.4 ± 0.9	**87.8** ± **2.2**	50.0 ± 0.0
BERT-ENTS	87.0 ± 0.0	**62.0** ± **0.0**	**57.2** ± **2.3**	**59.6** ± **1.5**	84.8 ± 3.3	**54.0** ± **0.0**

Table 2: Ablation study results on downstream tasks.

7 Conclusion

In this paper we investigate the effectiveness of data augmentation techniques for low-resource language modeling, focusing on Luxembourgish. We compare two new BERT models, DA BERT and LuxemBERT-v2, to LuxemBERT and mBERT as baselines. Results show that data augmentation can improve the performance of models on certain downstream tasks and that one approach is more effective than another depending on the task.

While this study focused on synonym, entity, and modal verb replacements, we would like to see future work investigate additional techniques such as paraphrasing, back-translation or negation. We would also suggest gathering more diverse and representative data for Luxembourgish as well as exploring different model architectures such as Generative Pre-Trained Transformer (GPT; (Radford et al., 2018)) or RoBERTa (Liu et al., 2019) that are designed to capture the semantics and context of words.

8 Limitations

We argue that our study has some limitations. The choice of not training LuxemBERT-v2 from scratch due to time constraints might have affected its rather average performance compared to our expectations. We assume that during the continued pre-training of LuxemBERT, the model might have overfitted to the added portion of the data or forgotten what it had learned before.

We take into account that the slightly higher number of sentences for BERT-ENTS might result in favouring the entity replacement technique over synonym replacements.

Lastly, our study is limited to the BERT architecture. There is a risk that after data augmentation the meaning of sentences might change and that the data is not true anymore, especially after replacing entities. Using data augmentation with other models such GPT (Radford et al., 2018) could be risky as these generative models rely solely on the provided data to learn linguistic and commonsense reasoning.

9 Ethical Considerations

For this study, we trained our models on a text corpus that includes comments on news articles and chats from a chatroom. While this data originally included usernames, they were anonymised in order to comply with data privacy laws (Lothritz et al., 2022). Furthermore, we do not publish this text corpus, merely the models that were pre-trained using the corpus.

10 Acknowledgements

This work was partially supported by the Fonds National de la Recherche (FNR), Luxembourg, under project LUXEMBERT BRIDGES2021/IS/16229163.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David Eberhard, Gary F Simons, and Charles D Fenning. 2022. *Ethnologue: Languages of Africa and Europe*. SIL International Publications.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Daniel Jurafsky and James Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 2.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

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. *arXiv preprint arXiv:1907.11692*.

Cedric Lothritz, Kevin Allix, Bertrand Lebichot, Lisa Veiber, Tegawendé F Bissyandé, and Jacques Klein. 2021. Comparing multilingual and multiple monolingual models for intent classification and slot filling. In *International Conference on Applications of Natural Language to Information Systems*, pages 367–375. Springer.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and practical data augmentation in language model pre-training for Luxembourgish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5080–5089, Marseille, France. European Language Resources Association.

Lucas Francisco Amaral Orosco Pellicer, Taynan Maier Ferreira, and Anna Helena Reali Costa. 2023. Data augmentation techniques in natural language processing. *Applied Soft Computing*, 132:109803.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. OpenAI.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Alexey Tarasov. 2020. Towards reversal-based textual data augmentation for NLI problems with opposable classes. In *Proceedings of the First Workshop on Natural Language Interfaces*, pages 11–19, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Nuo Xu, Yinqiao Li, Chen Xu, Yanyang Li, Bei Li, Tong Xiao, and Jingbo Zhu. 2019. Analysis of back-translation methods for low-resource neural machine translation. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II*, page 466–475, Berlin, Heidelberg. Springer-Verlag.