M2 Lexical Resources Report

Karolin BOCZOŃ Omar CHERIF Maxime MÉLOUX

February 2023

1 Introduction

Word embeddings are the core of transformer-based language models such as BERT, GPT-3, and RoBERTa [1]. These models have shown impressive performance in various NLP tasks, including language translation, sentiment analysis, and text generation [2]. The transformer architecture relies on the use of high-quality word embeddings to efficiently process and extract meaningful representations of textual data.

Embedding values are inferred from the examples seen by the model during pre-training. There is therefore an issue with non-frequent words, as their embeddings may be inaccurate or incomplete. This can further propagate poor performance on NLP tasks that involve those.

In this report, we will explore a method proposed in the paper titled "Rare Words: A Major Problem for Contextualized Embeddings and How to Fix it by Attentive Mimicking" [3], in which the authors propose a way to improve embeddings for non-frequent words using a form-context model. This approach considers both the surface form of the word and the context it appears in in order to compute a new embedding.

We aim to provide an overview of the method described in the paper and evaluate its results on another language. Specifically, the goal of this report is to reproduce experiments for the Polish language using the HerBERT language model [4]. Our main focus is to see if the method is easily translatable onto other models and languages.

2 Method

In this section, we shortly explain the contributions of the authors of the original paper.

In order to compute new embeddings for rare words, the authors use adjusted Attentive Mimicking (AM) [5], a method that generates high-quality embeddings based on the surface form and context of a token, along with one-token approximation (OTA), which generates the embedding that a multi-token word would have if it was treated as a single token.

To evaluate the performance of a base model and the same model after updating embeddings, the authors create WNLaMPro, a dataset consisting of two components:

- A list of words from the English edition of Wikipedia along with their frequencies, which was then crossed with English WordNet in order to obtain the list of antonyms, hypernyms and cohyponyms, as well as corrupted versions of frequent words.
- A list of English patterns such as "<X> is a [MASK]", used to test how well a given model understands a target word <X> before and after its embeddings are replaced with the generated ones.

3 Experiments

3.1 Language choice

Our initial goal was to run the experiments on the French language using Free Wordnet for French (WOLF) [1] and the CamemBERT model [6]. However, after some work, we discovered that the quality of the WordNet is below our expectations. This is mainly due to the fact that it was built from the Princeton WorldNet (PWN)[2] WordNet, using English as a pivot language. As a result, when reading a synset for a French word, if the English equivalent happens to be polysemous, we may obtain unrelated words in our dataset, as seen in the example below:

```
>>> wn.synsets('gonzesse', lang='fra')[0].hypernym_paths()[0]
[..., Synset('bird.n.01'),
      Synset('gallinaceous_bird.n.01'),
      Synset('domestic_fowl.n.01'),
      Synset('chicken.n.02'),
      Synset('chick.n.01')]

>>> [x.lemma_names(lang='fra') for x in _[-3:]]
[['oiseau', 'volaille'],
 ['poulet'],
 ['gonzesse', 'nana', 'poussin']]
```

Here, *gonzesse* (slang for "woman") is being translated as *chick*, which can also refer to a young bird, which leads to the incorrect hypernyms *chicken*, *bird* and *poultry*.

Due to this issue, we decided to change the language to Polish. As one of the few WordNet resources developed independently from PWN [7], plWordNet[3] does not have that polysemic translation problem. In addition, our group luckily contains two Polish speakers which were able to write patterns and evaluate model outputs.

3.2 Dataset

We created the Polish version of WordNet Language Model Probing (WNLaMPro) using a pre-generated word frequency list[4] from Wikipedia, as we had insufficient computing resources to process the Wikipedia dump on our own. We only kept nouns and adjectives present in the plWordNet, and implemented antonym, hypernym, cohyponym and corruption extraction. Since the authors' code did not include this part, we tried to replicate the algorithm given in the original paper to the best of our abilities. In total, we obtained 4,252 triplets of the form (word, relation, [target words]). Statistics on the complete dataset are shown in Table 1.

Rel.	Subset size			Mean targets		
	Rare	Medium	Frequent	Rare	Medium	Frequent
Antonyms	4	8	30	1.0	1.0	1.0
Hypernyms	14	49	115	3.1	3.1	3.2
Cohyponyms	359	1184	1662	31.0	28.7	28.8
Corruptions	—	—	776	—	—	1.0

Table 1: Distribution of triples based on frequency and relation

[1] http://almanach.inria.fr/software_and_resources/WOLF-en.html
[2] https://wordnet.princeton.edu/
[3] http://plwordnet.pwr.wroc.pl/wordnet/
[4] https://github.com/IlyaSemenov/wikipedia-word-frequency

A challenge that appeared while generating the dataset was part-of-speech tagging. Since we were unable to process the full Wikipedia corpus, we could not POS tag it and therefore lost important data in the process. For example, the most common word in the dataset is *w* (preposition - *in*; Polish does not have articles), but the only plWordNet subsets available for *w* are *tungsten.n.01* and *watt.n.01*. The dataset therefore contains an entry for *w* with cohyponyms being other chemical element names, but the associated corpus frequency is that of the preposition *w*. We believe that this issue is most commonly encountered for short words, but it is possible that this may cause dataset frequencies to be higher than they should be overall.

We also translated the patterns present in the original English WNLaMPro, which happened to be quite a linguistic challenge, as Polish is a heavily inflected language. In the queries, we tried to prompt the masked word in the nominative case (the one used in dictionary entries), which is sometimes hard to achieve without losing naturalness. For adjectives, we also tried to prompt the masculine form as the default. Selected examples of challenging patterns can be found below:

- **Antonym pattern:** `<W> is the opposite of [MASK]`
  Polish translation: `<W> jest przeciwieństwem do [MASK]`
  Issue: The translation requires the genitive form of [MASK], which is different from the dictionary (nominative) form (e.g. *yellow:* m. nom. *żółty*, m. gen. *żółtego*).
  Updated translation: `<W> to przeciwieństwo do słowa [MASK]` (`<W> is the opposite of the word [MASK]`).

- **Antonym pattern:** `something that is <W> is not [MASK]`
  Polish translation: `coś <W> nie jest [MASK]`
  Issue: The translation requires the genitive neuter form of <W> and the nominative neuter form of [MASK], which are both different from the dictionary form (e.g. *small:* m. nom. *mały*, n. nom. *małe*, n. gen. *małego*).
  Updated translation: `<W> obiekt nie jest [MASK]` (`a <W> object is not [MASK]`).

- **Hypernym pattern:** `<W> is a kind of [MASK]`
  Polish translation: `<W> jest rodzajem [MASK]`
  Issue: The translation requires the genitive form of [MASK]. It is very difficult to find a translation not using it, leading to the awkward sentence below.
  Updated translation: `Element kategorii do której należy <W> to [MASK]` (`an element of the category to which belongs <W> is a [MASK]`).

The chosen translations were manually tested on a small number of examples to make sure that the model typically outputs the right predictions for [MASK] on frequent words.

3.3 Model and parameters

We edited the authors' original code to add support for CamemBERT (in our initial attempt) and HerBERT, which large language models for respectively French and Polish. Several versions of each are available. We decided to pick the model with the closest number of parameters, architecture and training set size to BERT base that was used in the authors' experiment. For HerBERT, the two possible choices were `herbert-base-cased`, with a similar number of parameters as BERT base but a smaller trainset, and `herbert-large-cased`, with a much higher number of parameter but equivalent trainset size. Since both of those two models are cased, we also lowercased all patterns and predictions during evaluation.

To choose the optimal model and number of iterations, we ran OTA on a small subset of the first 50 words from our dataset in alphabetical order. The evolution of the distance between OTA embeddings and original HerBERT embeddings is presented in Figure 3.3.

Based on these results, it seems that the optimal number of iterations is higher than 8,000 for both models, unlike in the original paper in which the authors found a minimum at around 4,000

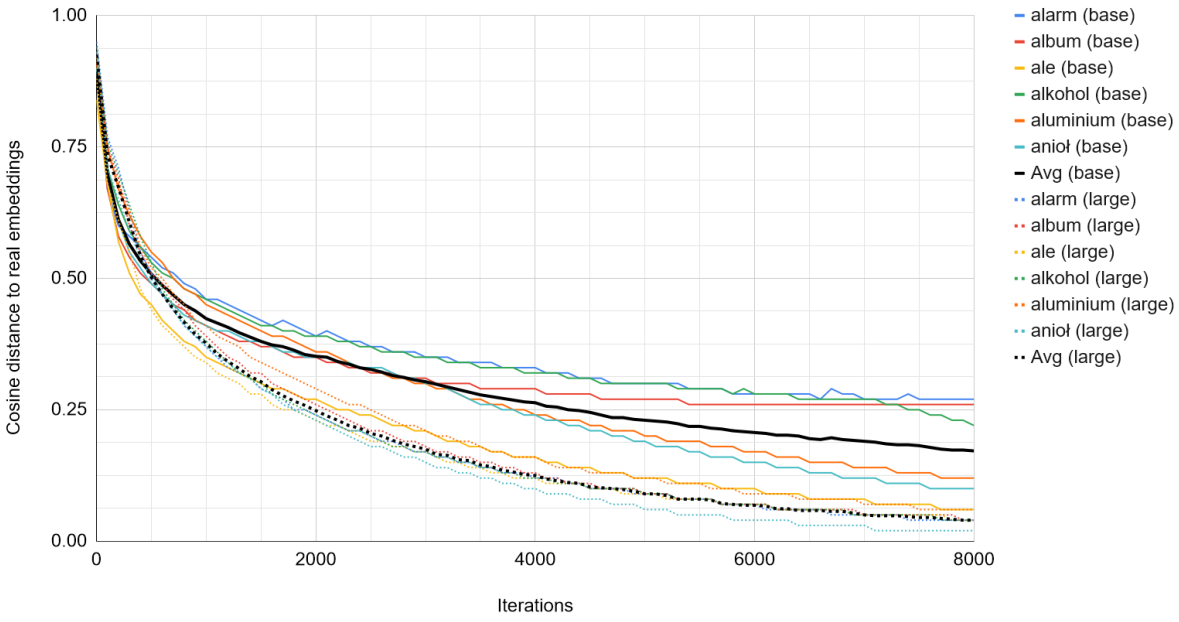

Figure 1: Impact of training duration on OTA performance (for selected words and average) on the base and large models.

iterations. We also see that the large model allows OTA to produce better embedding approximations. Our computing resources allowed us to use either model (which have a similar inference time), but forced us to limit the duration of OTA computation, which is very slow. As a result, we chose to use 'herbert-cased-large', but to stop at 4,000 iterations (OTA computing time: ≈12 hours).

4 Results and discussion

The results of the evaluation of 'herbert-large-cased' on our dataset before and after OTA and AM can be found in Table 2. Since no hyperparameter tuning was performed for OTA and AM, we used the entire dataset (development and test set) for evaluation.

Set	Model	Rare			Medium			Frequent		
		MRR	P@3	P@10	MRR	P@3	P@10	MRR	P@3	P@10
Ant.	large	—	—	—	—	—	—	0.245	0.133	0.047
	OTA + AM	—	—	—	—	—	—	**0.291**	0.133	0.047
Hyp.	large	**0.132**	0.053	**0.037**	**0.140**	**0.047**	0.024	0.253	0.081	**0.049**
	OTA + AM	0.125	0.053	0.032	0.137	0.040	**0.026**	**0.261**	**0.084**	0.046
Coh.	large	**0.143**	**0.055**	**0.049**	**0.173**	**0.070**	**0.056**	**0.279**	0.118	0.078
	OTA + AM	0.132	0.043	0.047	0.163	0.063	0.052	0.278	**0.120**	0.078
Cor.	large	**0.355**	0.127	0.050	—	—	—	—	—	—
	OTA + AM	0.354	0.127	0.050	—	—	—	—	—	—
All	large	**0.280**	**0.102**	**0.049**	**0.173**	**0.070**	**0.054**	0.277	0.116	0.075
	OTA + AM	0.275	0.097	0.048	0.162	0.062	0.051	0.277	**0.118**	0.075

Table 2: Results of the evaluation performed on the entire Polish WNLaMPro dataset.

We observe that overall, the embeddings obtained through OTA and AM do not perform significantly better than raw embeddings. Furthermore, the scores are generally lower than the ones obtained by the authors of the original paper. We attribute these discrepancies to a number of reasons:

- OTA was trained for a total of 4,000 iterations, which, as seen above, is not sufficient to reach convergence. Better performance might have been achieved with 8,000 iterations or more.
- Despite our improvements, a number of issues remain in the patterns. For example, the cohyponym pattern `<W> i [MASK]` (`<W> and [MASK]`) tends to produce many inflected forms of the word *inny* (*other(s)*), which does not happen in English since only two forms of that word exist. The unnaturalness of some patterns might also produce lower quality output in some cases, although we have not noticed this in practice.
- plWordNet contains fewer words than PWN, and words from plWordNet tend to be more domain-specific rather than generic. This results in a lot of rare words from our dataset having the same hypernyms and cohyponyms. Furthermore, plWordNet contains mostly cohyponym relations and few hypernym ones, with the former typically giving a lower output precision than the latter.
- OTA and AM may simply not perform as well on Polish as it does on English, possibly due to morphological constraints and inflection. In particular, we have not attempted to tune the hyperparameters of these methods due to a lack of computing resources, which could have helped with performance.
- While it is hard to compare their performance since they operate on different languages and have therefore been evaluated on different benchmarks, it is possible that 'herbert-large-cased' achieves a lower performance than RoBERTa models.
- As mentioned above, there is a frequency mismatch for some words in the dataset due to incorrect POS attribution. It is hard to evaluate how widespread this mismatch is, and how much it may affect model performance.
- A bug was found at a late stage of the project, which caused corruptions to be stored wrongly in the dataset. As a result, the trained embeddings did not include corrupted version of words. One should therefore not consider the scores in Table 2, which also affects the overall scores. This issue did not affect other categories (antonyms, hypernyms and cohyponyms), but we did not have enough time to re-train embeddings.

5 Conclusion

In this report, we have described Attentive Mimicking and one-token approximation, two methods for improving the quality of embeddings for rare words. We have evaluated this method on the Polish language model HerBERT, using a novel dataset extracted from plWordNet and the Polish Wikipedia corpus. Although performed on a limited scale, our experiments do not show significant improvement over the baseline embeddings.

The entirety of our code and datasets is available on GitHub[5].

[5] `https://github.com/pie3636/LRExperiments`

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[2] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[3] Timo Schick and Hinrich Schütze. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8766–8774, Apr 2020.

[4] Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 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, April 2021. Association for Computational Linguistics.

[5] Timo Schick and Hinrich Schütze. Attentive mimicking: Better word embeddings by attending to informative contexts. 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 489–494, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[7] Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wroclawskiej, Wroclaw, 2009.