Peer-reviewed & recommended by PCI Archaeo

Archaeology specific BERT models for English, German, and Dutch

Alex Brandsen[1]

[1]**Leiden University, a.brandsen@arch.leidenuniv.nl, ORCID: 0000-0003-1623-1340**

ABSTRACT

This short paper describes a collection of BERT models for the archaeology domain. We took existing language specific BERT models in English, German, and Dutch, and further pre-trained them with archaeology specific training data. We then took each of these three archaeology specific models and fine-tuned them for Named Entity Recognition.

Keywords: BERT, Archaeology, Natural Language Processing, Named Entity Recognition

1 INTRODUCTION

Archaeology deals with a deluge of data (Bevan, 2015), and this of course includes text data. To efficiently analyse large amounts of texts, such as the hundreds of thousands of grey literature reports, Natural Language Processing (NLP) techniques are needed, such as Named Entity Recognition (NER), which tries to extract relevant concepts from text. In the case of our project, the entities we are looking for are artefacts, time periods, locations, materials, contexts, and species. Traditionally, NER was often performed using Conditional Random Fields (CRF) (Brandsen et al., 2019) or rule-based methods (Vlachidis et al., 2013).

With the rise of Deep Learning, BERT (Bidirectional Encoder Representations from Transformers) models can leverage large amounts of unlabelled text to create a language model, which can then be fine-tuned to perform specific NLP tasks – such as NER – with a smaller amount of labelled data (Devlin et al., 2019). This pre-training phase gives the model linguistic context before even seeing the labelled data, which has lead to state of the art results on many tasks. In previous research we showed that a BERT model for Dutch archaeology outperforms all other methods (Brandsen et al., 2022). In this paper, we present the same model types, but also for English and German.

2 METHODS

We started with the three existing fill-mask BERT models for English, German, and Dutch. We provide the model name, HuggingFace (Wolf et al., 2020) link, and citation for each:

- English: bert-base-cased (Devlin et al., 2019)
- German: bert-base-german-cased (Chan et al., 2021)
- Dutch: bert-base-dutch-cased (De Vries et al., 2019)

We then used the HuggingFace run_mlm.py script to further pre-train these existing models with archaeology specific, unlabelled data. See the file `run_mlm_on_commandline.sh` in the code folder for the parameters used.

For English, we used roughly 44k documents (∼264 million tokens) from the ADS library, for German we used about 21k documents (∼107 million tokens) from the Heidelberg university library, and for Dutch we used around 60k Dutch excavation reports (∼650 million tokens) from the DANS data archive. Unfortunately, as a large proportion of the texts is copyrighted, under embargo, or otherwise access-protected, we can not share this training data publicly.

Model name	Language	Task	HuggingFace URL
ArchaeoBERT	English	Fill-mask	ArchaeoBERT
ArchaeoBERT-NER	English	NER	ArchaeoBERT-NER
bert-base-german-cased-archaeo	German	Fill-mask	bert-base-german-cased-archaeo
bert-base-german-cased-archaeo-ner	German	NER	bert-base-german-cased-archaeo-NER
ArcheoBERTje	Dutch	Fill-mask	ArcheoBERTje
ArcheoBERTje-NER	Dutch	NER	ArcheoBERTje-NER

Table 1. Overview of the six BERT models, including links to HuggingFace.

Method	Language	Precision	Recall	F1
CRF	Dutch	0.773	0.564	0.643
General BERT	English	0.762	0.688	0.723
Archaeo BERT	English	**0.784**	**0.731**	**0.757**
CRF	German	0.677	0.325	0.414
General BERT	German	**0.698**	0.634	0.664
Archaeo BERT	German	0.681	**0.702**	**0.692**
CRF	English	0.774	0.592	0.655
General BERT	English	0.766	0.716	0.740
Archaeo BERT	English	**0.768**	**0.730**	**0.749**

Table 2. Overview of precision, recall and F1 on the NER task, comparing the archaeology specific BERT model versus the generic BERT model and CRF. Highest scores per language are highlighted in bold.

This process creates an archaeology-specific fill-mask BERT model for each language, but these can't do NER yet. So we fine-tuned them on labelled NER data, creating NER prediction models for each language. Table 1 shows an overview of all the models we trained. The labelled NER data was annotated by students, this data and the labelling process is further described in a previous paper (Brandsen et al., 2020).

3 RESULTS

To see how well the models performed, we evaluated the recall, precision, and F1 score on a set of annotated NER data for each language. The results can be found in Table 2. We see that except for precision in German, the archaeology specific BERT models outperform the other methods. We also see that the increase in performance with the archaeology specific model is smaller for English than for the other two languages, something we will investigate in future research. For a more in depth error analysis of the Dutch BERT model entity predictions, as well as more information about the CRF methods, please see Brandsen et al. (2022).

4 USAGE

How to use these models depends on the task you are trying to perform. In the simplest case, the models can be used to do NER predictions on new text by importing the HuggingFace transformers package, loading the models from the HuggingFace repository, setting up a predictor, and running inference, in just a few lines of code. See the example below, using Transformers v4.37.0, where we imported the English Archaeology NER model and ran it on a sentence:

```
from transformers import pipeline

pipe = pipeline("token-classification", model="alexbrandsen/ArchaeoBERT-NER")

predictor = pipeline(
    'ner',
     model=model,
     tokenizer=tokenizer,
     device = 0,
```

```
    grouped_entities = False
)

sentence = "We have found a cup in a Medieval well."
entities = predictor(sentence)
```

Which will return the entities cup (artefact), Medieval (time period), and well (context):

```
[
    {'entity': 'B-ART', 'score': 0.9598702, 'index': 5, 'word': 'cup', 'start': 16, 'end': 19}
    {'entity': 'B-PER', 'score': 0.9939248, 'index': 8, 'word': 'Medieval', 'start': 25, 'end': 33}
    {'entity': 'B-CON', 'score': 0.58865726, 'index': 9, 'word': 'well', 'start': 34, 'end': 38}
]
```

Some possible uses for entities extracted from archaeological text are information retrieval (Brandsen, 2022), enriching metadata, and as a preprocessing step for further knowledge extraction, such as knowledge graph creation or entity linking.

If you want to use the models for a task other than NER, you will need to fine-tune the fill-mask model(s) for your chosen task. This is a bit more involved than just running inference, and is outside the scope of this paper. Full instructions on how to do this can be found in the HuggingFace documentation pages.

Another possible use for the fill-mask models is to further pre-train them even further, with even more (or more specific) training data. In this case, the `run_mlm_on_commandline.sh` script in the code folder can be pointed to your own data, and this will update the model. Note that large amounts of text are needed for this, at least 50 million tokens are recommended.

5 DISCUSSION

Of course, since the release of GPT-3 and its associated ChatGPT service (Brown et al., 2020), BERT has largely been overtaken in terms of performance by these Large Language Models (LLMs). However, for specific domains and low-resource languages, BERT can still outperform LLMs. We ran a quick trial using Meta's Llama model (Touvron et al., 2023) trying to do NER in English, and found it difficult to control the output, while the results were mediocre. For example, we used this prompt "You are an expert archaeologist, who is annotating sentences with archaeological named entities, including artefacts, places, time periods, materials, locations, contexts and species. You are given an input as plain text, and should output CONNL format. The input is: 2 pots were uncovered near London , dated to the Late Medieval period .", which returned:

```
> 2 P
pots W
U
cove
r O
near O
London B-LOC
,
dated O
to O
the L
ate M
edieval P
e
r
i
o
d B-PER
```

It is clear that Llama sort of understands the task, but the output is not correct, nor correctly formatted in the CONLL format. More research is needed to see if this new generation of language models can help us for this specific task. At the same time, it is worth noting that the BERT models are at least intelligible and appropriate for the domain specific work one might want to do with them.

Also something to keep in mind is that BERT is already a very computationally intensive model, using up much electricity and CO2 (Bender et al., 2021), and these new LLMs are even bigger and more energy consuming, and this should be kept in mind when using these methods. Is the (sometimes small) increase in performance worth the extra time, expertise, and energy required for these models?

6 CONCLUSION

In this paper, we presented six BERT models for use in NLP in archaeology, in three languages. We also provide the methods on how to further pre-train models, and hope to either see models in other languages by other researchers, or we will train models ourselves, hopefully leading to an ecosystem of language models for archaeology.

ACKNOWLEDGEMENTS

We would like to thank the DANS archive, the ADS library, and Heidelberg University for sharing the data that these models were trained on. The models were trained using the compute resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University.

FUNDING

The project in which this research has been carried out has been funded by NWO (The Dutch Research Council) under the "Future directions in Dutch archaeological research" programme.

CONFLICT OF INTEREST DISCLOSURE

The author declares that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. The author declares the following non-financial conflict of interest: the author is a recommender of PCI Archaeology.

DATA AND CODE AVAILABILITY

The code needed to further pre-train BERT models and to run inference with the models is included in the 'code-examples' folder of this Zenodo archive.

As for the data, as a large proportion of the texts is copyrighted, under embargo, or otherwise access-protected, we can not share this training data publicly.

REFERENCES

Bender, E. M., Gebru, T., Mcmillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT '21, March 3–10*, Canada. ACM.

Bevan, A. (2015). The data deluge. *Antiquity*, 89(348):1473–1484.

Brandsen, A. (2022). *Digging in documents: using text mining to access the hidden knowledge in Dutch archaeological excavation reports*. PhD thesis, Leiden University.

Brandsen, A., Lambers, K., Verberne, S., and Wansleeben, M. (2019). User Requirement Solicitation for an Information Retrieval System Applied to Dutch Grey Literature in the Archaeology Domain. *Journal of Computer Applications in Archaeology*, 2(1):21–30.

Brandsen, A., Verberne, S., Lambers, K., and Wansleeben, M. (2022). Can BERT Dig It? - Named Entity Recognition for Information Retrieval in the Archaeology Domain. *Journal on Computing and Cultural Heritage*, 15(3):1–18.

Brandsen, A., Verberne, S., Wansleeben, M., and Lambers, K. (2020). Creating a Dataset for Named Entity Recognition in the Archaeology Domain. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 2020-December.

Chan, B., Schweter, S., and Möller, T. (2021). German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain. International Committee on Computational Linguistics.

De Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., and Nissim, M. (2019). BERTje: A Dutch BERT Model. *arXiv*.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (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, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models.

Vlachidis, A., Binding, C., May, K., and Tudhope, D. (2013). Automatic metadata generation in an archaeological digital library: Semantic annotation of grey literature. *Studies in Computational Intelligence*, 458:187–202.

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., and Rush, A. (2020). 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, Stroudsburg, PA, USA. Association for Computational Linguistics.