Peer Community In
Archaeology

# Multilingual Named Entity Recognition in archaeology: an approach based on deep learning

***Maria Pia di Buono*** *based on peer reviews by* ***Shawn Graham*** *and 2 anonymous reviewers*

Alex Brandsen (2023) Archaeology specific BERT models for English, German, and Dutch. Zenodo, ver. 4, peer-reviewed and recommended by Peer Community in Archaeology. https://doi.org/10.5281/zenodo.10610882

Submitted: 29 August 2023, Recommended: 28 February 2024

**Cite this recommendation as:**
di Buono, M. (2024) Multilingual Named Entity Recognition in archaeology: an approach based on deep learning. *Peer Community in Archaeology*, 100394. 10.24072/pci.archaeo.100394

Published: 28 February 2024

Archaeology specific BERT models for English, German, and Dutch" (Brandsen 2024) explores the use of BERT-based models for Named Entity Recognition (NER) in archaeology across three languages: English, German, and Dutch. It introduces six models trained and fine-tuned on archaeological literature, followed by the presentation and evaluation of three models specifically tailored for NER tasks. The focus on multilingualism enhances the applicability of the research, while the meticulous evaluation using standard metrics demonstrates a rigorous methodology.

The introduction of NER for extracting concepts from literature is intriguing, while the provision of a method for others to contribute to BERT model pre-training enhances collaborative research efforts. The innovative use of BERT models to contextualize archaeological data is a notable strength, bridging the gap between digitized information and computational models.

Additionally, the paper's release of fine-tuned models and consideration of environmental implications add further value.

In summary, the paper contributes significantly to the task of NER in archaeology, filling a crucial gap and providing foundational tools for data mining and reevaluating legacy archaeological materials and archives.

***References:***

Brandsen, A. (2024). Archaeology specific BERT models for English, German, and Dutch. Zenodo, 8296920, ver. 5 peer-reviewed and recommended by Peer Community in Archaeology. https://doi.org/10.5281/zenodo.8296920

# Reviews

## Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.5281/zenodo.8300777
Version of the preprint: 2

### Authors’ reply, 02 February 2024

Dear recommender and reviewers,

thank you for your work on this, I read the reviews with great pleasure, and welcome the feedback mentioned.

First of all, I want to apologise for not making the context of this paper very clear: this submission to the CAA proceedings is an ’alternative format’ (as defined at https://2023.caaconference.org/proceedings/), specifically a collection of models combined with a cover text. The paper should be a succint description of the product, up to a 1000 words. Some of the comments in the reviews are very valid, but unfortunately outside the scope of this particular paper. However, I’ll list all the issues raised below, and explain if and how I’ve addressed them, for a full overview. The parts I haven’t addressed due to the scope will be dealt with in a future paper, where we compare the use and performance of rule-based, CRF, and BERT in much more detail.

**Reviewer 1:**

”the paper could benefit from a more extensive evaluation of the performance of the models fine-tuned for NER tasks through an error analysis of the entities”

Unfortunately out of scope, but I will deal with this in more detail in a future paper, and we also provide this for the Dutch models in a previous paper (https://doi.org/10.1145/3497842), which I have now cited in the relevant sections.

”omitting code sections and providing links to the scripts in footnotes;”

There was a bit of a clash between this and the requests of reviewer 3, who wanted to see more code.. However I’ve removed some of the less relevant code sections, and instead focused on 1 code snippet showing how to use the model.

”removing text closely tied to technical comments (e.g., lines 45-46 and 81).”

I found 45-46 to be essential to the description of the training, but line 81 has been removed.

**Reviewer 2:**

”It would be useful to expand on the traditional use of CRF and rule-base methods”

As mentioned above, this will be discussed in more detailed in a future paper, and the CRF methods are detailed in https://aclanthology.org/2020.lrec-1.562 and https://doi.org/10.1145/3497842, which I have now cited in the relevant sections.

” it would be useful to see a visualization (or comparative visualization) of archaeological wordlists, akin to the one published in Delvin et al. 2019”

This is an excellent suggestion, but also unfortunately out of scope for this short paper. I will add this to the future paper.

”The limitations of GPT-3 and ChatGPT relative to BERT could be expanded upon.”

I’ve added some more information and example output of Llama to further illustrate this.

”It would also be useful to see an example of how others might use NER in their own archaeological work.”

I’ve added a paragraph giving some suggestions of possible uses.

**Reviewer 3:**

”In terms of the training, I might like to know a bit more about the process, in particular, the origin of the labelled NER data for the process”

Good point, I have now cited the relevant paper that describes the annotation process for the Dutch dataset.

”was the process to take an existing fill-mask BERT model (which one?) and then use a fill-mask approach on labelled archaeological texts - which in English was 44k documents? (I imagine it was a much smaller subset?) ”

I’m sorry this wasn’t completely clear, we indeed took an existing fill-mask BERT model, and then retrained it on again a fill-mask approach, but on unlabelled data. The labelled data is only used when fine-tuning for the NER task. I’ve added some clarifications in the text to make that more clear.

”How long did it take to produce the annotated training data? Did Brandsen do that work himself, or with a team? ”

It took a team of students 90 hours to annotate the data for 1 language. This is described in https://aclanthology.org/2020.lrec-1.562, which is now cited in the relevant sections

”How feasible is it for someone else to continue to improve these archaeological models - how much annotated data might one need to provide?”

I’ve added a sentence explaining this in the Usage section. You can take the fill-mask model and make it even more specific by giving it at least 50 million tokens of unlabelled text from a more specific domain, or you can create your own NER model by taking the fill-mask model and fine-tuning it on your own labelled NER data.

”I found that I could not run the supplied code example”

Thanks for checking the code, I think something has updated in the Transformers library since I wrote this paper originally.. I have updated the code examples and specified the Transformers version for clarity.

”I think it might be helpful for the reader if the actual output were presented”

Very good point, I have added the output.

Kind regards,

Alex

## Decision by Maria Pia di Buono, posted 08 January 2024, validated 08 January 2024

**Revision needed**

The paper should be accepted after a review process to address the reviewers’ suggestions.

## Reviewed by anonymous reviewer 1, 11 November 2023

Review of Archaeology specific BERT models for English, German, and DutchThe paper is an exploration of employing BERT-based models for Named Entity Recognition (NER) within the archaeological domain across three languages: English, German, and Dutch. The author introduces six models based on the BERT architecture, trained and fine-tuned on data related to the archaeological domain in three languages (English, German, and Dutch). The author first outlines the pre-training phase using data from archaeological literature. Subsequently, the paper presents three models fine-tuned for Named Entity Recognition (NER) tasks. These three NER fine-tuned models are then evaluated using performance metrics such as Precision, Recall, and F1. A strength of the paper lies in its focus on multilingualism. By training models in three distinct languages, the author demonstrates a commitment to broadening the applicability and relevance of the research.The heart of the paper lies in the presentation and evaluation of the three fine-tuned models dedicated to NER tasks. The meticulous evaluation, conducted using performance metrics such as Precision, Recall, and F1, reflects a rigorous methodology. The author’s choice of these metrics underlines a commitment to comprehensive model assessment.Furthermore, additional strengths of the paper lie in releasing the fine-tuned models and the final reflection about the environmental implications of using even larger generative LLMs. However, the paper could benefit from a more extensive evaluation of the performance of the models fine-tuned for NER tasks through an error analysis of the entities involved in the experiments. Additionally, the paper could benefit from improved readability by:

- omitting code sections and providing links to the scripts in footnotes;
- removing text closely tied to technical comments (e.g., lines 45-46 and 81).

In conclusion, the paper stands out as a valid contribution to the field of Natural Language Processing, specifically in the context of archaeological Named Entity Recognition. The approach to model development, coupled with a robust evaluation methodology, positions this paper as a valuable resource.

## Reviewed by anonymous reviewer 2, 26 November 2023

Dear authors,

This paper offers a concise but intriguing paper look at an exciting area of recent advancement in archaeology. The use of a NER for extracting relevant concepts from abundant literature to find mention of artefacts, time periods, etc. has several useful implications. Can this be expanded on in the introduction?

What stands out in this paper, is that the authors provide a method by which others can contribute to the pre-training of the BERT models. This important aim and its potential broader implications should be mentioned in the introduction.

It would be useful to expand on the traditional use of CRF and rule-base methods to in turn, help explain why the BERT model is such an advance and to demonstrate how it is different. This would help to elevate the significance of using the BERT models particularly when GPT-3 is mentioned.

The innovative use of BERT models to contextualize the model before NLP tasks are carried out is a critical strength of this paper. The data are linked on HuggingFace, but it would be useful to see a visualization (or comparative visualization) of archaeological wordlists, akin to the one published in Delvin et al. 2019 (https://production-media.paperswithcode.com/methods/new_BERT_Overall.jpg) which will allow readers to immediately recognize the way word/concept selections are streamlined.

The limitations of GPT-3 and ChatGPT relative to BERT could be expanded upon. Here it would be useful to see a part of the trial run on the Llama model, if possible, to again give readers a more specific idea of what the differences are between results achieved using the various approaches.

It would also be useful to see an example of how others might use NER in their own archaeological work.

Thank you.

## Reviewed by Shawn Graham, 13 November 2023

How do we wring new insights from legacy archaeological data? And what new issues does that act of 'data wrangling' create? Hugget (2022) for instance explores in detail the 'many characters of data' and 'data imaginaries' that the fact of digital data capture, representation, and storage create. But nevertheless, a lot of these issues can feel remote when one is face to face with what sometimes passes for archaeological 'data': badly scanned and ocr'd pdfs of poorly laid out recording templates or desk reports. The high-level issues can sometimes feel quite remote.

In which case, this welcome contribution by Alex Brandsen fills a gap in the middle between digitized data and the higher level issues. He takes existing computational language models for English, German, and Dutch and fine tunes these models on the particularities of archaeological reports, providing a tool for extracting data at scale from digitized data.

In terms of the training, I might like to know a bit more about the process, in particular, the origin of the labelled NER data for the process. Also, more detail perhaps on how the training logic works might reassure the reader of the validity of the fine-tuning process. I know than Brandsen and team have written in more detail about the intricacies of this kind of work in other publications, but if we imagine the present article as a kind of paradata document for the creation and further elaboration of this process, some more nuts-and-bolts details might be welcomed by the reader. For instance, was the process to take an existing fill-mask BERT model (which one?) and then use a fill-mask approach on labelled archaeological texts - which in English was 44k documents? (I imagine it was a much smaller subset?) How long did it take to produce the annotated training data? Did Brandsen do that work himself, or with a team? How feasible is it for someone else to continue to improve these archaeological models - how much annotated data might one need to provide? I ask all this from the point of view of imagining writing a proposal to create a fine-tuned model on a more limited archaeological domain, having to think through labour requirements, computation requirements, budget requirements and so on. Brandsen's experience would be invaluable.

I found that I could not run the supplied code example in a Google Colab notebook (which gives me access to a GPU) until I set the runtime to gpu, and then used the pipeline as a high level helper (see code below) which is slightly different than what Brandsen reports:

```
‘‘‘python
!pip install transformers
import transformers
# Use a pipeline as a high-level helper
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 returns:

```
‘‘‘json
[\’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’: 0.5886574,
’index’: 9,
’word’: ’well’,
’start’: 34,
’end’: 38\]
```

“‘

I think it might be helpful for the reader if the actual output were presented in the piece and the meaning of the different keys and values was discussed. The coding literacy of archaeologists in general seems to be higher than other humanistic disciplines, perhaps, but it would be generous to the reader to expand a bit on the necessary setup and what the output looks like and how one might reshape this data into for instance a dataframe or csv file for further analysis - even just a small script to achieve that.

The piece raises the issue of Large Language Models being used for NER tasks; if anything here I think Brandsen undersells the value of the models he has created. The internal workings of a large language model are inscrutable, whereas Brandsen’s models are at least intelligible and appropriate for the domain specific work one might want to do with them. The discussion might want to focus more on the positives of these models rather than framing them defensively in the light of LLMs. To my mind, Brandsen has performed an enormous service to the archaeological community by providing these models, and he might usefully expand on ways the models might be deployed. Further pre-training the models is gestured towards by linking to the HuggingFace documentation, but if Brandsen knows of or has available archaeological-specific documentation to achieve this it might be better to point the reader in that direction as well (in that some of the tacit things that organizations like HuggingFace assume their readers already know might not necessarily be known by the archaeologist interested in taking Brandsen’s work further).

I am glad Brandsen has done this work and made it available to the archaeological world; I feel these models will become foundational for data mining and reconsiderations of legacy archaeological materials and archives.

Huggett, Jeremy. 2022. ”Data Legacies, Epistemic Anxieties, and Digital Imaginaries in Archaeology” _Digital_ 2, no. 2: 267-295. https://doi.org/10.3390/digital2020016