Digging in documents: using text mining to access the hidden knowledge in Dutch archaeological excavation reports

Brandsen, A.

7

Using BERT for Named Entity Recognition

"*Look at all the exciting new discoveries, look at all the knowledge here.*"
Bert, Sesame street ep. 1621, 'Bert and Ernie in a Pyramid'

Accepted for publication as: Brandsen, A., Verberne, S., Lambers, K., & Wansleeben, M., 2021. Can BERT Dig It? – Named Entity Recognition for Information Retrieval in the Archaeology Domain. *Journal on Computing and Cultural Heritage*

The amount of archaeological literature is growing rapidly. Until recently, these data were only accessible through metadata search. We implemented a text retrieval engine for a large archaeological text collection (∼ 658 Million words). In archaeological IR, domain-specific entities such as locations, time periods, and artefacts, play a central role. This motivated the development of a Named Entity Recognition (NER) model to annotate the full collection with archaeological named entities. In this paper, we present ArcheoBERTje, a BERT model pre-trained on Dutch archaeological texts. We compare the model's quality and output on a Named Entity Recognition task to a generic multilingual model and a generic Dutch model. We also investigate ensemble methods for combining multiple BERT models, and combining the best BERT model with a domain thesaurus using Conditional Random Fields (CRF). We find that ArcheoBERTje outperforms both the multilingual and Dutch model significantly with a smaller standard deviation between runs, reaching an average F1 score of 0.735. The model also outperforms ensemble methods combining the three models. Combining ArcheoBERTje predictions and explicit domain knowledge from the thesaurus did not increase the F1 score. We quantitatively and qualitatively analyse the differences between the vocabulary and output of the BERT models on the full collection and provide some valuable insights in the effect of fine-tuning for specific domains. Our results indicate that for a highly specific text domain such as archaeology, further pre-training on domain-specific data increases the model's quality on NER by a much larger margin than shown for other domains in the literature, and that domain-specific pre-training makes the addition of domain knowledge from a thesaurus unnecessary.

7.1 Introduction

Like in other domains, archaeologists produce large amounts of text about their research. Besides research leading to scholarly output, commercial archaeology companies survey and excavate areas before developers build there and might destroy the archaeological remains. For each of these investigations, a report is written and stored in a repository. In the Netherlands, more than 4,000 of these documents are produced every year (Rijksdienst voor het Cultureel Erfgoed, 2019a), with the total currently estimated at 70,000. These documents are used to some extent by both academic and commercial archaeologists to do further research.

Currently, this so-called 'grey literature' is underused, as the available search tools only offer metadata search, making searching through these reports time

consuming and inaccurate (Habermehl, 2019). A strong need for better search tools has been well documented in prior work (Van den Dries, 2016; Habermehl, 2019; Richards *et al.*, 2015; Brandsen *et al.*, 2019), as the information in the full text of the reports can be of great value. Archaeological information needs are often recall-oriented list questions, consisting of a combination of What, Where and When aspects, e.g. "Find all cremations from the Early Middle Ages in the Netherlands" (Brandsen *et al.*, 2019). These are difficult to satisfy as the previously available search interfaces only offer search on the title, a short description, and sometimes information about the dating and type of archaeology encountered (stored in metadata fields), but the latter two are often missing or incorrectly assigned. Archaeologists want to search in more detail, and are often interested in the so-called 'by-catch': a single find unlike the rest of an excavation. For example, on an excavation yielding mainly Bronze Age material, a single Medieval cremation most likely will not be mentioned in the metadata, making it difficult to retrieve without manually searching through all the PDFs.

To address these needs, we implemented a text retrieval engine for a large collection of archaeological reports in the Netherlands. The retrieval collection contains an export (obtained in 2017) of every PDF file in the DANS repository[1] with the label 'Archaeology'. This totals over 60 thousand documents and 658 Million tokens.

A full text search would alleviate a lot of the current challenges archaeologists face in their search of information, but as Habermehl (2019) mentions, even in the relatively structured metadata, both synonymy and polysemy are a challenge, which is likely to be even worse in the free text in the body of the documents.

- Synonymy is a challenge because it leads to a lower recall: as there are numerous ways to write concepts relevant to archaeology, a search for one of these variants will not return the others. Specifically time periods have many synonyms. For example, the 'Early Middle Ages' can also be expressed as the 'Early Medieval Period', or 'Merovingian Period', or as dates that fall within the period, such as '600 CE' and '1400 BP'.
- Polysemy on the other hand, causes precision to be lower because one word can have multiple meanings, causing irrelevant meanings to appear in the search results. A good archaeological example is *Swifterbant*, which is a location, a type of pottery, an excavation event, and a time period. This problem of polysemy causes query ambiguity, as a full-text search engine does not know which meaning the user is looking for in their query, and then also does not know which meaning to retrieve from the corpus.

[1] https://easy.dans.knaw.nl/ui/home

Automatic query expansion is often used to combat problems with synonymy, either by using thesauri or embeddings to add synonyms and similar terms to a query and increase the recall (Soto *et al.*, 2008; Carpineto & Romano, 2012). Unfortunately in the case of time periods, this is difficult, as some time periods span thousands or millions of years, and adding each year with multiple variations (AD, BC, CE, BCE, BP) would result in an extremely large query. Polysemy is usually addressed in web search engines by diversifying search results or query suggestions (Capannini *et al.*, 2011; Song *et al.*, 2011): for each possible meaning of the ambiguous query, at least one relevant result is shown. For our specific domain, this is not possible because we do not have the large amount of user traffic that generic web search engines have, to be able to learn the different relevant results for any query term.

Instead, we opt for Named Entity Recognition (NER) to automatically detect archaeological entities in the corpus, and then allow archaeologists to find these using an entity-based query interface, combined with a full text search. The entity search attempts to solve the polysemy problem, as the user specifies – in the structured query interface – which meaning of a word they are looking for, e.g. the Location[2] *Swifterbant*. In this case, only documents where the Location entity *Swifterbant* has been detected will be returned. Although this helps the user specify their query, it also means that entities that have not been correctly identified will not be returned; in other words, errors in the NER output might propagate to retrieval errors. Therefore, to give the user freedom in the query form that best suits their information need, we combine entity search with full-text search.

We have previously published a prototype of our search engine online. The search engine uses ElasticSearch (Gormley & Tong, 2015) to index the full text, and in the prototype, entities were automatically labelled with a baseline NER model based on Conditional Random Fields (CRF). The resulting entity-based full-text search was experienced as positive by a focus group of archaeologists (Brandsen *et al.*, 2019).

However, the baseline NER model offers room for improvement. As prior work on archaeological NER indicated, CRF with common token-, context- and thesaurus-based features leads to relatively low F1 scores, around 0.50 to 0.70 (Brandsen *et al.*, 2019, 2020). In the last couple of years, transfer learning, and specifically BERT models (Devlin *et al.*, 2019), have been used successfully to get state-of-the-art (SotA) results for NER. On general domain benchmarks the SotA methods yield impressive F1 scores of up to 0.943 (Yamada *et al.*, 2020).

[2]Entity types will be capitalised from here on for clarity.

However, in other domains and languages the performance of NER systems is generally lower (Lee *et al.*, 2019).

BERT has not been applied to the archaeology domain yet in any language, and we believe this domain could benefit from context-dependent embeddings due to the above-mentioned polysemy. Two generic Dutch BERT models have been released (De Vries *et al.*, 2019; Delobelle *et al.*, 2020) which can help our research. Prior work on language- and domain-specific BERT models reports mixed results on the effect of pre-training on language- and domain-specific data (see Section 7.2.4). In this paper, we investigate whether BERT can improve NER in the Dutch archaeology domain, and to what extent further pre-training on domain-specific texts improves the quality of the model. We compare Google's multilingual model (Devlin *et al.*, 2019), the Dutch BERTje model (De Vries *et al.*, 2019), and our own ArcheoBERTje model that we further pre-trained on Dutch excavation reports. We do not compare the Dutch RobBERT model as it has a different training procedure and longer training times. We analyse the differences between the three models and we experiment with ensembles to combine multiple models and a domain-specific thesaurus. As there is unfortunately no test collection with relevance assessments available for the Dutch archaeology domain, we do not evaluate the performance of the information retrieval, only the performance of the NER.

We address the following research question:

1. To what extent does further pre-training a BERT model with domain-specific training data improve the model's quality in our highly specific domain?
2. Can a domain-specific BERT model be improved by adding domain knowledge from a thesaurus in a CRF ensemble model?
3. What errors are made by the models and what are the differences in predicted entities between the three models?

The contributions of our paper are three-fold: First, we propose entity-driven full-text search in which the professional user enters a structured query, and documents are filtered for the occurrence of the query entities detected by our new domain-specific BERT model. Second, we show that for a highly specific domain such as archaeology, further pre-training on domain-specific data increases the model's quality on NER by a much larger margin than shown for other domains in the literature. Third, we show that the domain-specific BERT model outperforms ensemble methods combining different BERT models, and also outperforms a CRF-based ensemble of BERT with explicit domain knowledge from the archaeological thesaurus.

We make our modified training data set, the pre-trained ArcheoBERTje model, and the fine-tuned ArcheoBERTje model for NER publicly available (Brandsen, 2021b).[3]

7.2 Related Work

In this section, we first summarise different approaches to NER (knowledge-driven and data-driven), followed by a discussion of related work on NER for document retrieval, on IR and NER in the archaeological domain, and we summarise the prior work on domain-specific BERT models.

7.2.1 Knowledge-driven and Data-driven NER

Early NER systems were knowledge-based, and relied on thesauri and handcrafted rules to detect entities (Rau, 1991). These methods are limited by the coverage of the thesaurus. Therefore, data-driven methods have become more popular, typically approaching NER as a supervised machine learning problem.

A highly effective machine learning method is Conditional Random Fields (CRF) (Lafferty *et al.*, 2001), which has become a common baseline for NER. Since 2011, word embeddings have become increasingly important as representations in NER. Especially Word2vec (Mikolov *et al.*, 2013) has been used extensively for NER (Sienčnik, 2015; Seok *et al.*, 2016). These embeddings-based methods typically feed the embeddings to CRF and/or Bi-LSTM algorithms to make NER predictions.

A big shift in NLP was introduced by Devlin *et al.* (2019), who presented their BERT (Bidirectional Encoder Representations from Transformers) architecture in 2019. BERT and other contextual embedding architectures are currently achieving SotA results with transfer learning for a large range of NLP tasks, including NER. Two major differences with previous embedding models are (1) that BERT embeddings are contextual, meaning that the same token can have a different embedding based on context, and (2) that it handles out-of-vocabulary words effectively, by dividing tokens into sub-tokens it does have in vocabulary, using the WordPiece (Devlin *et al.*, 2019) or SentencePiece (Kudo & Richardson, 2018) tokeniser.

Recent results indicate that ensemble methods that combining generic and domain-specific BERT models (Copara *et al.*, 2020), combining BERT with dic-

[3] https://doi.org/10.5281/zenodo.4739063, also available via the HuggingFace library for ease of use: https://huggingface.co/alexbrandsen

tionary features (Li *et al.*, 2020), or adding a CRF on top of BERT (Souza *et al.*, 2019) can improve NER quality. In this paper, we investigate whether addition of information from a thesaurus can improve NER in a highly specific domain.

7.2.2 NER for Document Retrieval

In the context of document retrieval, NER can play a role in better ranking or filtering documents based on entities in the query. Guo *et al.* (2009) were the first to address the task of recognising named entities in queries. They found that, despite queries in web search being short, 70% of the queries contained a named entity. They classify the entities according to a predefined taxonomy using a weakly supervised topic modelling approach on the query data. Cowan *et al.* (2015) also address NER in queries, but for the travel domain. They use CRF on the queries for extracting the relevant entities.

More recently, the relevance of NER on queries has been emphasised for the e-commerce domain. Wen *et al.* (2019) and Cheng *et al.* (2020) both implement end-to-end query analysis methods for e-commerce search; the extracted queries are then used to filter the retrieved products.

As opposed to the prior work, we do not focus on query analysis but on document analysis; our expert users prefer the use of structured queries, which makes query analysis unnecessary (see Section 7.4.4). Our documents, on the other hand, are long and unstructured (as opposed to the products in e-commerce search), making NER on the document side necessary for matching structured queries to the relevant documents.

7.2.3 IR and NER in Archaeology

As argued by Richards et al. (Richards *et al.*, 2015), archaeology has great potential for thesaurus-based IR and NER, as it has a relatively well-controlled vocabulary and there are thesauri of archaeological concepts available in multiple languages. However, unlike some other fields, archaeology terminology partly consists of common words, like 'pit', 'well' and 'post'. In addition, words can be archaeological entities or not, depending on the context in which they are used (past or present). For example, the word 'road' is not archaeologically relevant in the snippet "pit next to the main road", but is part of an archaeological entity in the snippet "a Roman road from 34 CE".

Archaeology has started experimenting with IR relatively recently. The focus of the prior work is on Information or Knowledge Extraction, mainly for automatically generating document metadata. An early study by Amrani *et al.*

(2008) aimed specifically at extracting information for archaeology professionals in a knowledge-based approach. A more data-driven approach using machine learning to detect Time Period entities was investigated in the OpenBoek project (Paijmans & Brandsen, 2010, 2009), but since then most studies have been knowledge-driven (Jeffrey *et al.*, 2009; Byrne & Klein, 2010; Vlachidis *et al.*, 2013, 2017).

More recently, Talboom experimented with embeddings in a Bi-LSTM model to recognise zooarchaeological entities (species and specific bones) (Talboom, 2017). A notable exception to the Information Extraction research we often see in archaeology is the work by Gibbs & Colley (2012) who created a full-text search engine on a small Australian corpus (roughly 1,000 documents) combined with facets based on manually entered metadata.

So far, NLP in the archaeology domain has not benefitted from BERT-based models. We believe it is a good candidate domain for BERT as the polysemy mentioned in the introduction and the present/past distinction mentioned above should be easier to detect with the context-dependent embeddings that BERT produces.

7.2.4 Language- and Domain-specific BERT Models

The original BERT paper (Devlin *et al.*, 2019) did not only present an English BERT model, but also a multilingual model (multiBERT) trained on data in 104 languages. This model is often used when no single-language model is available (Hakala & Pyysalo, 2019; Moon *et al.*, 2019; Kim & Lee, 2020). Research by Wu & Dredze (2020) shows that multiBERT achieved higher accuracy on NER and other NLP tasks than monolingual models trained with comparable amounts of data. Moon *et al.* (2019) also showed that fine-tuning multiBERT on a mixed language NER dataset provided better results than fine-tuning on individual languages.

However, recent work has shown that for some languages, multiBERT is outperformed by language-specific BERT models (Nozza *et al.*, 2020). For NER, this has been shown for Finnish (Virtanen *et al.*, 2019), Dutch (De Vries *et al.*, 2019), German (Chan *et al.*, 2021) and Russian (Kuratov & Arkhipov, 2019), among other languages.

For specific domains, it has been shown that further pre-training the English BERT-base model on large amounts of text from that domain increases the quality of the model on multiple tasks, although sometimes by a small margin. BioBERT in the biomedical domain shows an increase in F1 for NER of only 0.62% point (Lee *et al.*, 2019). SciBERT, trained on a large amount of scientific texts from

different domains, shows an increase in F1 for NER of 2 to 5% points, indicating that domain pre-training is useful for NER (Beltagy *et al.*, 2020). They also show that training BERT from scratch with a domain-specific vocabulary does not increase F1 substantially compared to fine-tuning an existing BERT model with an existing generic vocabulary, gaining only 0.6% points.

When we look at research done on non-English in a specialised domain like our study, there is little prior work. A study in the Russian cyber-security domain shows that the Russian model (RuBERT) outperformed multiBERT, and further pre-training RuBERT with domain-specific documents yielded the highest F1 (Tikhomirov *et al.*, 2020). In the Spanish biomedical domain, Akhtyamova (2020) shows a similar result, although their NER BERT model is trained for 30 epochs, possibly leading to over fitting.

To our knowledge, we are the first to address domain-specific NER for Dutch, and we are the first to automatically label a large archaeological document collection with our domain-specific BERT model for the purpose of professional search.

7.3 Data

The unlabelled data set we use for further pre-training the Dutch BERTje model to ArcheoBERTje consists of over 60k documents and 658 Million tokens across 16.6 Million sentences, around 2GB of data. The documents mainly consist of survey/excavation reports, but also include other documents such as research plans, appendices, maps, and data descriptions.

The labelled training data we use for NER we created previously (Brandsen *et al.*, 2020), and consists of fifteen documents that have been annotated by archaeology students. While fifteen reports is a relatively low number, these are longer than average documents, totalling 1,343 pages (average 89 pages per document), containing roughly 440,000 tokens and almost 43,000 annotated entities across six categories: Artefacts, Time Periods, Locations, Contexts, Materials and Species, see Table 7.1. The Inter Annotator Agreement reported is 95% (average pairwise F1 score), so it is of relatively high quality (Brandsen *et al.*, 2020). The data is stored in the BIO annotation schema, and is available for download.[4]

The data set has been split into 5 folds of 3 documents each. All methods are evaluated using this 5 fold split.

[4]Zenodo repository: http://doi.org/10.5281/zenodo.3544544

Entity	Description	Examples
Artefact	An archaeological object found in the ground.	Axe, pot, stake, arrow head, coin
Time Period	A defined (archaeological) period in time.	Middle Ages, Neolithic, 500 BC, 4000 BP
Location	A placename or (part of) an address.	Amsterdam, Steenstraat 1, Lutjebroek
Context	An anthropogenic, definable part of a stratigraphy. Something that can contain Artefacts	Rubbish pit, burial mound, stake hole
Material	The material an Artefact is made of.	Bronze, wood, flint, glass
Species	A species' name (in Latin or Dutch)	Cow, *Corvus Corax*, oak

Table 7.1: Descriptions and examples for each entity type. Examples are translated from Dutch. Adapted from (Brandsen *et al.*, 2020, p. 4574).

7.3.1 Pre-processing

For cross-validation, we divided the fifteen annotated documents across five folds so that each fold has a roughly equal number of tokens. The exact fold split and training data can be found on in the Zenodo repository.

We found that in the data set, sentences often exceed the maximum sequence length of 512 WordPiece tokens. This is not because sentences actually have more than 512 words, but partly because tables and OCRed maps and images create very long 'sentences' that are not cut up by the sentence detection algorithm. The other cause is that words that are uncommon outside of archaeology are cut up into many sub-tokens by the WordPiece tokeniser, as they do not exist in the vocabulary (also see Section 7.6.2).

Since sentences longer than 512 tokens will be trimmed, some of the input tokens will not get a prediction. To counteract this, we wrote a pre-processing script that attempts to break at a punctuation mark ('.', ';' or ',') between the 60th and 90th token and if there are none, it inserts a line break after the 90th token. This shortened the sentences sufficiently to have almost no instances where the sentence was longer than 512 WordPiece tokens. Only 136 tokens in the entire data set fell outside the 512 limit and received no prediction. These tokens only contained two entities, so the effect on the performance metrics will be negligible.

7.4 Methods

7.4.1 Baselines

As the first baseline, we use the method we published previously (Brandsen *et al.*, 2020), where we trained a CRF model using common word shape features (e.g. occurrence of uppercase letters, numbers), part-of-speech tags (e.g. noun, verb) and an archaeological thesaurus in a five word window, and performed hyperparameter optimisation. We used the same features, leading to a micro F1 score of 0.62. This is relatively low when comparing the score to NER in other domains, where F1 scores between 0.8 and 0.9 are common (Akhtyamova, 2020; Lee *et al.*, 2019).

The second baseline is the standard NER pipeline of `spaCy` 2.0, with default parameters (architecture: TransitionBasedParser.v2, random seed, max_steps: 20,000, Adam.v1 optimiser with learn_rate of 0.001). This method uses pre-existing Dutch word embeddings (`nl_core_news_lg`) with a deep convolutional neural network with residual connections, and a transition-based approach as the classifier (Honnibal & Montani, 2017).

7.4.2 Fine-tuning BERT for Dutch Archaeology and NER

Model training for evaluation To train ArcheoBERTje, we started with the Dutch BERTje model (De Vries *et al.*, 2019) and further pre-trained the model with our complete unlabelled archaeological collection, split into a 90/10 train and validation set.[5] We used the same configuration as BERTje, with a batch size of 4. We decided not to train a model from scratch as previous research showed only a minimal increase in quality compared to further pre-training (Beltagy *et al.*, 2020) an existing model, and because our corpus is relatively small and would probably not be enough to train an effective model.

To fine-tune the BERT models for the NER task, we used the labelled data and 5-fold cross-validation as described in Section 7.3.[6] For model comparison and to investigate the stability of each model with different random seeds, we trained all three models 10 times per fold, each time using a different seed (1, 2, 4, 8, 16, 32, 64, 128, 254, 512) and report averages over all runs and folds (50 runs in total per BERT model).

[5] We used HuggingFace's (Wolf *et al.*, 2020) language modelling script version 3.0.2.
[6] We used HuggingFace's token classification script version 3.0.2.

Model for full collection labelling To create the best possible model for inference on the entire corpus, we performed a grid search across hyperparameters as suggested by (Devlin *et al.*, 2019). We optimised the hyperparameters with fold 2 as test set, fold 1 as development set, and the other folds as training set, as this combination had the median F1 score across all models and folds. The grid search yielded the following optimal parameters for our data: 2 training epochs, $5 * 10^{-5}$ learning rate and 0.1 weight decay. We then fine-tuned the inference model on all labelled data with these hyperparameters. This way we maximise the amount of training data available for training the model that we use to label the full collection.

7.4.3 Ensemble Methods

As far as we are aware, we are the first to combine a multilingual model, a language-specific model and a domain-specific model into one ensemble method. We evaluate the following ensemble methods (one run over 5 folds per ensemble):

- Majority voting on the predictions of multiBERT, BERTje and ArcheoBERTje;
- CRF which uses the prediction labels of the three models as features;
- CRF which uses the prediction labels of ArcheoBERTje only;
- CRF which uses the prediction labels of the three models as features, combined with the baseline features;
- CRF which uses the prediction labels of ArcheoBERTje only, combined with the baseline features;
- CRF which uses the embeddings produced by ArcheoBERTje as features.

The above mentioned 'baseline features' are those adopted from prior work (See Section 7.4.1) and include word shape, part-of-speech tags and thesaurus features. We optimised the hyperparameters of each CRF ensemble with gradient descent using the L-BFGS method, optimising c1 and c2 (the coefficients for L1 and L2 regularisation). The optimisation was run separately for each fold. All CRF ensembles use a 5-token window, taking into account the features from the two tokens before and after the current token.

The thesaurus we use in our CRF baseline and ensembles is the ABR (*Archeologisch Basisregister*) (Brandt *et al.*, 1992; Brandsen *et al.*, 2020), a thesaurus containing time periods (e.g. Bronze Age), artefacts (e.g. axe) and materials (e.g. flint). A token is assigned the binary feature 'occurs in period/artefact/material list' if it is part of an n-gram that occurs in the thesaurus. So the token 'Bronze' would only be assigned a positive value for the feature if the token 'age' follows it.

7.4.4 Entity-driven Document Search

Indexing Before we index the documents, we first run the inference NER model on each page to detect the entities. We then store the entities and full text in a JSON file for each document, together with the relevant metadata (authors, DOI, coordinates, document type, etc) retrieved from the DANS repository via an API.

To tackle the synonymy problem for time periods (see Section 7.1), we use a custom script that translates all extracted Time Period entities to year ranges. It uses regular expressions to convert dates (e.g. '100 BCE', 'start of the 9th century') and an extended and customised version of the PeriodO time period gazetteer (Rabinowitz *et al.*, 2016) to translate Time Periods (e.g. 'Bronze Age', 'Medieval period'). These date ranges are added to the JSON and can be used to filter results by allowing users to specify a date range in their query. These JSON files are then sent to an instance of ElasticSearch running on a webserver, which indexes them. At the moment, the retrieval unit is a page, so for any query the terms/entities must occur together on a page. We are aware this is not optimal, as search terms might be split across pages. As such, in future work we will index per document section by using a section detection algorithm.

Query Interface and Analysis Our search engine has a faceted search interface in which metadata filters are combined with entity fields and full-text search (Tunkelang, 2009). We have included facets for document type and subject (metadata fields). In addition, as requested by our target group, we added geographical search via a map functionality, which allows users to draw a rectangle or polygon to search only in a certain region.

At query time, the user can specify if they are looking for a specific entity type and/or specify a date range in which they are interested. The entities and date range are used to filter the result set and can be combined with a standard full text search. This allows for relatively complex queries such as "Artefact: urn AND Context: cremation AND startdate < -2000 AND enddate > -800 AND fulltext: upside down". This example is a real request entered by an archaeologist, who was looking for upside down urns in the Bronze Age in or around cremations. Users do not need to use complex query syntax, but can instead define their query by filling in the relevant fields in the graphical user interface, as shown in Figure 7.1.

Document Ranking Most archaeological information needs are recall-oriented tasks: the users want a complete list and do not mind having irrelevant results

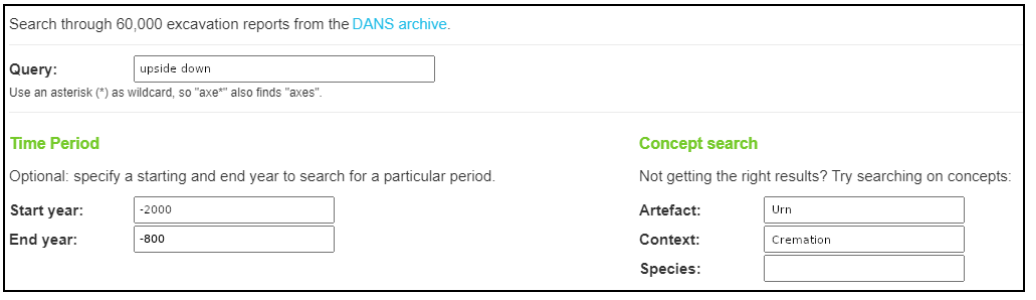

Figure 7.1: Query interface showing query for "Artefact: urn AND Context: cremation AND startdate < -2000 AND enddate > -800 AND fulltext: upside down". Interface and query translated to English for the readers' convenience.

in the (top of) the result set (Brandsen *et al.*, 2019). As the focus of our work is on entity-driven search, we opt for the default ElasticSearch ranking model, consisting of Term Frequency - Inverse Document Frequency (TF-IDF) and the field-length norm (the shorter the field, the higher the relevance) (ElasticSearch, 2018). The only field included for ranking is the page text content, other fields are only used for filtering.

Note that we do not evaluate the ranking, because there is no test collection available yet for Dutch archaeological document retrieval. Therefore, the scope of this paper is limited to the NER and the evaluation thereof.

7.5 Results

7.5.1 Model Stability and Quality

Table 7.2 shows the micro average precision, recall, and F1 score for the three BERT models, compared to the CRF and spaCy baselines. We find that the multilingual BERT model does not outperform the baselines, but the more specialised BERTje and ArcheoBERTje models do, with ArcheoBERTje achieving the highest F1 score.

We also show the average standard deviation over 10 runs with different seeds for 5 folds. The standard deviation between runs is very low, between 0.015 and 0.004. The recent work by Tikhomirov *et al.* reports a standard deviation of 0.015 to 0.008, similar to our results (Tikhomirov *et al.*, 2020). When comparing the predicted labels of each of the models in a pairwise manner, the differences are significant according to McNemar's test (χ^2 between 650 and 4276, $p < 0.00001$).

Model	Precision	Recall	F1 (Std.)	Fails
CRF Baseline	**0.785**	0.526	0.630	n/a
spaCy Baseline	0.717	0.602	0.654	n/a
multiBERT	0.623	0.550	0.583 (0.015)	4
BERTje	0.718	0.682	0.699 (0.005)	0
ArcheoBERTje	0.743	**0.729**	**0.735 (0.004)**	0

Table 7.2: Micro average precision, recall and F1 score at token level (B and I labels), over 10 runs with different seeds, for each of the 5 folds (50 runs total). Standard deviation of F1 over the 10 runs is added in brackets for the BERT models. Standard deviation of precision and recall lies between 0.006 and 0.020. The 'Fails' column indicates the number of times the model failed to learn (F1 = 0).

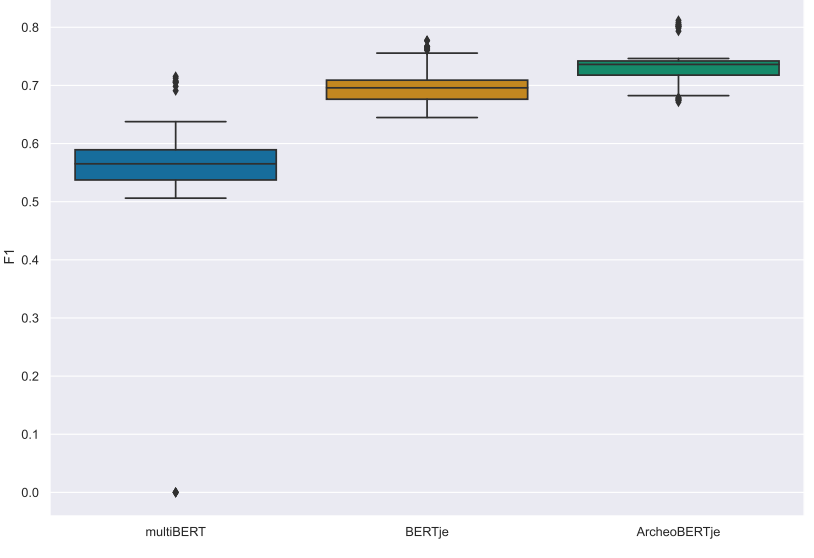

Figure 7.2: Distribution of F1 scores over ten runs with different seeds, for each of the 5 folds (50 runs per model). The zero scores for multiBERT are runs where the model failed to learn.

Figure 7.2 shows the distribution of F1 scores over the 50 runs per model in a boxplot. Here we again see that the standard deviation is low, and that ArcheoBERTje consistently outperforms the other two models. The F1 scores of 0 for multiBERT are outliers, and we assume these are caused by the ADAM optimiser getting stuck in a local minimum where the loss does not decrease. In this local minimum, predicting the majority class (O) seems to yield the highest accuracy, but of course O labels are not taken into account when calculating an F1 score for NER, so we get a score of zero. This can be solved by changing the learning rate, but this would not change the overall view that BERTje and ArcheoBERTje outperform multiBERT, so we did not investigate further on fixing this for multiBERT.

The low standard deviations for ArcheoBERTje indicate that further pre-training with domain-specific data does not only increase the model quality on average, but also makes the model more stable, reducing the chance of getting a sub-optimal model in a run.

Another way to compare the models is by looking at differences between the errors made. In Table 7.3 we report the top 10 most frequent error combinations for the three models. Here we can see that quite often, BERTje and ArcheoBERTje have similar predictions (whether correct or not), while multiBERT predicted a different label. We see that multiBERT often misses Locations (LOC), Artefacts (ART) and Species (SPE), and sometimes predicts entities that are not there. The first error combination where ArcheoBERTje outperforms BERTje is number 9, having correctly predicted B-ARTs while the other 2 models do not. In Sections 7.5.3 and 7.6.1 we further analyse the output and errors made by the ArcheoBERTje model to provide insight into the model's behaviour.

7.5.2 Ensembles

Table 7.4 shows the results of the ensemble methods.[7] The highest F1 (0.757) is obtained by he optimised production ArcheoBERTje model.

The highest precision is obtained by the CRF ensemble with the baseline features combined with the predicted labels from all three models. The highest recall is achieved by ArcheoBERTje solo.[8] Using a CRF with BERT embeddings

[7] As the standard deviation between multiple runs is low, combining multiple runs of the same model in an ensemble model is very unlikely to increase the F1 score, at the expense of a vastly increased computing time and cost. Hence we do not apply this approach.

[8] For general domain Portuguese NER, Souza *et al.* show the same pattern: Portuguese BERT has the highest recall, while combining BERT with CRF yields the highest precision and F1 (Souza *et al.*, 2019).

Freq.	True	multiBERT	BERTje	ArcheoBERTje
1137	B-LOC	O	B-LOC	B-LOC
1122	B-ART	O	B-ART	B-ART
1015	O	B-ART	O	O
575	B-SPE	O	B-SPE	B-SPE
561	O	B-LOC	O	O
466	B-PER	O	B-PER	B-PER
429	O	O	B-ART	B-ART
425	I-PER	O	I-PER	I-PER
402	B-ART	O	O	B-ART
373	O	I-PER	O	O

Table 7.3: The 10 most frequent error combinations between the 3 models for which at least one model has the correct prediction. Errors are marked in red.

Ensemble	Precision	Recall	F1
ArcheoBERTje (50 runs avg)	0.743	0.729	0.735
ArcheoBERTje (optimised production model)	0.784	**0.731**	**0.757**
Majority Voting	0.784	0.695	0.737
CRF with 3 BERT model prediction labels as features	0.786	0.683	0.731
CRF with only production ArcheoBERTje predictions as features	0.786	0.717	0.750
CRF with 3 BERT model prediction labels + baseline features	**0.795**	0.644	0.712
CRF with production ArcheoBERTje prediction labels + baseline features	0.793	0.649	0.714
CRF with only production ArcheoBERTje embeddings as features	0.767	0.604	0.676

Table 7.4: Micro F1 score, precision and recall for the six ensemble methods, for one run over five folds. ArcheoBERTje results averaged over 50 runs and the optimised production model are added for comparison. The ArcheoBERTje predictions used as features for CRF are from the production model. The baseline features are the word- and context-based features used for CRF in prior work.

Entity	Total	Unique	Top 5
Artefacts	2,520,492	53,675	pottery, charcoal, flint, bone, brick
Contexts	1,602,124	21,319	pit, ditch, posthole, well, house
Materials	457,031	6,146	wooden, flint, wood, metal, bronze
Locations	3,488,698	147,077	nederland, ' , groningen, noord - brabant, gelderland
Species	928,437	34,540	cow, hazel, sheep, goat, pig
Time Periods	4,698,323	98,445	roman period, iron age, 150 - 210, late medieval, modern
Total	13,695,105	361,202	

Table 7.5: Overview of entities detected in the entirecorpus, showing total and unique counts, plus the top 5 for each entity (translated from Dutch where relevant).

as features instead of the default BERT classifier (softmax), does not increase performance. Given the recall-oriented nature of professional search tasks like ours, we prioritise recall over precision for the NER labelling, and use ArcheoBERTje for labelling the full collection.

7.5.3 Analysis of the Retrieval Collection

After labelling the full retrieval collection with ArcheoBERTje, we analyse the extracted entities. Table 7.5 shows for each entity type the total frequency and the amount of unique entities. We also show the top 5 entities extracted for each type (translated from Dutch to English).

As we already mentioned in the introduction, archaeologists are interested in the What, Where and When of excavations. And so we see that Artefacts, Locations and Time Periods are the most common entities.

- For Artefacts, we see that pottery and flint are common, which we expected, but apparently also charcoal, which we did not expect, but could be explained by the use of carbon dating, which often uses charcoal as a sample.
- In the Locations category, we see that the second most common entity is an apostrophe ('). While this is clearly not a location, luckily it will not affect retrieval as it is not something users would search for, and ElasticSearch does not include apostrophes in its index, so it would not match any documents. We speculate that ArcheoBERTje mislabels apostrophes

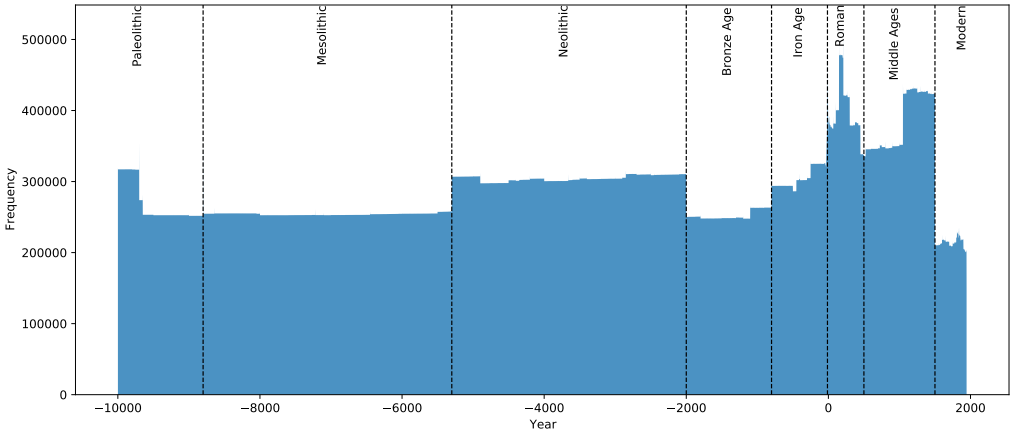

Figure 7.3: Graph showing for each year in each detected time period, how often it occurs in our data set, labelled by ArcheoBERTje. For clarity, years before 10,000 BCE are not included. Major time periods are denoted with dashed lines.

as locations because of the occurrence of apostrophes in some Dutch place names (e.g. *'s Hertogenbosch*).

- For Time Periods, the only unexpected entry in the top 5 is "150 - 210". When we investigated this further, we found this is actually a soil grain size used in coring reports, which have been incorrectly labelled as a time period by ArcheoBERTje. 150-210 µmm is the grain size for medium course sand, apparently the most common grain size in the Netherlands. When we look further down the Time Period top 100, we also see other common grain sizes: 210-300, 105-150 and 105-210. This is an issue when searching for archaeology between 105 and 300 CE, as these irrelevant coring reports will also be returned. We believe that these errors are made because these numbers come from tables, and as such do not have any sentence context, making them difficult to predict correctly. The most likely way to fix this is by making a post-processing correction on the extracted entities. This is something we will improve in the next version of our NER method.

The grain sizes are also clearly visible in Figure 7.3, in which we have plotted the frequency of years found in entities in the corpus. The figure shows a number of plateaus, indicating the use of time periods instead of single dates, i.e. the last plateau is the Late Middle Ages ending in 1500 CE. These plateaus are not completely flat as single dates and subperiods can cause spikes and smaller

sub-plateaus.

The thin spike just after the year 0 can probably be attributed to misclassified entities, i.e. the '10' in '10-02-2006' being labelled by ArcheoBERTje as a Time Period and translated to 10 CE. Other than this we see a big plateau in the middle (5300–2000 BCE), which represents the Neolithic. This indicates that a large amount of data is available describing this period in the Stone Age.

7.6 Discussion

7.6.1 Error Analysis

Figure 7.4 shows the confusion matrix between labels predicted by ArcheoBERTje and the true labels. The diagonal line and the first row and column are typical for NER. The diagonal shows the true positives, the top row is where the model predicted an entity where there isn't one, and the first column is where the model predicted O where there should be an entity. We also see the I / B label confusion quite clearly, mainly for Time Periods and Locations, where the model predicts an I instead of a B, or the other way around.

A more interesting error is the confusion between Materials and Artefacts. This is caused by words like "flint", which can be both an Artefact ("a piece of flint") or a Material ("a flint axe"). In Dutch, "pottery" has the same issue. Even archaeologists struggle with distinguishing between the two (Brandsen *et al.*, 2020), so it is unsurprising that ArcheoBERTje finds this difficult as well. As there is a lot of ambiguity in this entity category, perhaps merging the two categories into one entity type would increase the overall performance. We have seen in previous research that archaeologists will also confuse the two categories when creating queries, so having them both in one search field might not even cause any problems at search time.

Table 7.6 shows the evaluation per entity type. In general, the I labels are more difficult to predict, and Materials are more difficult than the other entities. In fact, Materials are currently not included in the search engine, as archaeologists find it difficult to differentiate between Materials and Artefacts in their queries, so this will not affect retrieval quality. When we remove Materials from the overall micro F1 score calculation, we get an increase of only around 0.01, as there are only a small number in our training data, around 3000.

When we look at some of the errors made by ArcheoBERTje in more depth, we find some interesting patterns. For example, for missing B-ART labels, many errors are adjectives that were assigned the O label, e.g. for "big axe" or "complete pot", the adjectives are labelled O, and axe / pot are labelled B-ART. This

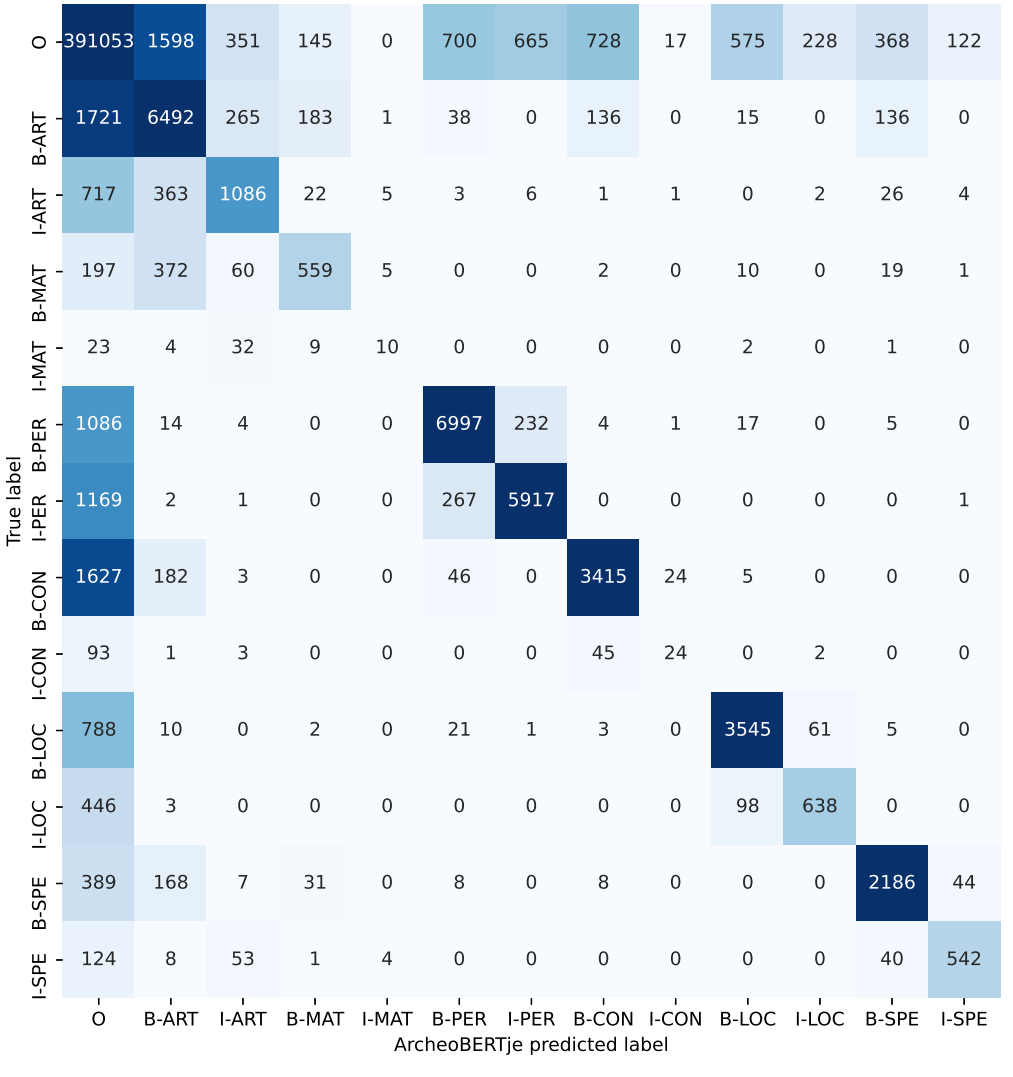

Figure 7.4: Confusion matrix between true labels and ArcheoBERTje predictions.

	Precision	Recall	F1
B-ART (Artefacts)	0.704	0.722	0.713
I-ART	0.582	0.486	0.530
B-CON (Contexts)	0.787	0.644	0.708
I-CON	0.358	0.143	0.204
B-MAT (Materials)	0.587	0.456	0.514
I-MAT	0.400	0.123	0.189
B-LOC (Locations)	0.831	0.799	0.815
I-LOC	0.685	0.538	0.603
B-SPE (Species)	0.785	0.769	0.777
I-SPE	0.759	0.702	0.729
B-PER (Time Periods)	0.866	0.837	0.851
I-PER	0.867	0.804	0.835
Macro Average	0.684	0.585	0.622
Micro Average	0.784	0.731	0.757

Table 7.6: ArcheoBERTje precision, recall and F1 score for each label.

error is not surprising as most archaeologists would probably find it difficult to define these entities as well. In addition, users are more likely to only search for the base artefact and not include an adjective, so they would search for "pot" not "complete pot". In a pilot study evaluating our archaeological search engine, we analysed users' search behaviour and found that of the 148 issued queries, none included an adjective.[9]

For Time Periods, we again see that adjectives are missed from the start of an entity, but also prepositions. Some examples include "from", "between" and "start of". Also we find that connecting words between Time Periods are missed, such as "and", "or" and "Âś" (used to denote the standard deviation of a carbon dating). While this does cause some noise, missing adjectives/prepositions or connecting words are not a considerable issue if the main period has been detected. I.e. for "start of 10th century", if we miss "start of" this means the year range is 900 to 1000 CE, instead of 900 to 925 CE. Again, as archaeologists care more about recall than precision, this should not hinder their search.

The predicted Context[10] entities also have some interesting anomalies. In particular, we analysed the top 10 most misclassified tokens and we found that

[9]Extension and publication of this user study is part of our future work.

[10]For clarity, Contexts are defined as an anthropogenic structures or objects that can contain Artefacts, i.e. rubbish pits, burials, houses, and so on.

these are all words that can denote contemporary objects (and thus not a Context) or actual (pre-)historical Contexts. An example is "*put*", which can mean a trench dug by archaeologists, or a water well found in an excavation, and both instances of *put* can contain an artefact, leading to similar contexts around these words. Other examples are "house", "church", "ditch", "mine" and "settlement". It seems that even with the context-dependent embeddings BERT produces, these ambiguous words are still a challenge. Perhaps future language models are more refined and might be able to distinguish between these types of ambiguous terms.

A special case is the word "*poel*" (pond). We see that this token is always labelled as O while it is in fact a Context. When we checked the sentences this word occurs in, we see they are all very typical of Contexts, i.e. "we found pottery in the pond", which is similar to sentence structures of other Contexts that are classified correctly. The only possible explanation we can find is that the word *poel* only occurs in one of the documents, so when this document is in the test set, the word does not occur at all in the train or dev set. This confirms the importance of creating train-test splits on the document level, to avoid overfitting. At the same time, this might be an issue that could be potentially alleviated by increasing the size of the training data.

More generally speaking, we see that the BERT models make impossible B and I predictions, i.e. an I label without a B label for the previous token. Unlike CRF, which learns the probabilities of two labels occurring after one another, BERT sees every token as an individual classification task without taking into account the predicted label of the previous token. This might explain why the CRF model with ArcheoBERTje labels as features (see Table 7.4) outperforms ArcheoBERTje on precision, as it corrects some of these mistakes. Perhaps another approach to correct this is a rule-based postprocessing step that checks the validity of I labels following B labels, and corrects impossible combinations.

During the annotation process, we used a test document of a hundred sentences (1,962 tokens) to calculate the Inter Annotator Agreement (Brandsen *et al.*, 2020). We added ArcheoBERTje predictions to this data, to see if ArcheoBERTje predictions are more often wrong when humans also have disagreement, indicating that the model mimics human confusion. We disregard tokens where everyone (including ArcheoBERTje) predicts an O label, leaving 292 tokens. In 57.5% of these tokens, all annotators and ArcheoBERTje predict the same label. In 31.5% of tokens, there is some disagreement between annotators, but ArcheoBERTje predicts the same label as the majority, and in 4.4% of tokens, ArcheoBERTje predicts a label different from the majority. In 6.5% of tokens, ArcheoBERTje predicts one label, while annotators all predict the same different label. This is only a small sample, but the above suggests that BERT models

are decently equipped to learn from the majority where there is inter-annotator disagreement.

7.6.2 Tokenisation Issues

The vocabulary of a BERT model is determined by the collection used for pre-training. The WordPiece tokeniser optimises the set of (sub-word) tokens to maximise the coverage of the collection's vocabulary. The same tokenisation is applied to the input sentences at inference. An example is shown below, where we compare tokenisation with the multiBERT and BERTje vocabularies. We see that target entities ("*Swifterbant*", "*aardewerkscherven*" and "*Midden Neolithicum*") are split up into three or more sub-tokens by the multiBERT and BERTje tokenisers.

Original sentence:
"*In put twee werden 3 Swifterbant aardewerkscherven aangetroffen uit het Midden Neolithicum.*" ("In trench two, 3 Swifterbant pottery shards from the Middle Neolithic were found.")

multiBERT tokenisation (23 tokens):
In put twee werden 3 Swift ##er ##bant aarde ##werks ##cher ##ven aan ##get ##roffen uit het Midden Neo ##lit ##hic ##um .

BERTje tokenisation (20 tokens), also used for ArcheoBERTje:
In put twee werden 3 Swift ##er ##ban ##t aardewerk ##scher ##ven aangetroffen uit het Midden Neo ##lith ##icum .

As an additional analysis, we trained a SentencePiece tokeniser on our archaeological collection, with the same vocabulary size as the BERTje model (30k).

Archaeology tokenisation (14 tokens):
In put twee werden 3 Swifterbant aardewerk ##scherven aangetroffen uit het Midden Neolithicum .

The examples show that a more specific pre-training corpus would lead to more complete domain words. However, our collection is small for such from-scratch pre-training and the experiments in the sciBERT paper have shown that even a much larger pre-training collection only gives a +0.6% point F1 increase compared to further pre-training the generic model (Beltagy *et al.*, 2020).

Understandably, the problem of input sequences longer than 512 tokens was occurring more often with the multilingual model, as the vocabulary (with fixed

size) is not solely Dutch. This means that many less common Dutch words are not in the vocabulary, and are cut into many sub-tokens by the WordPiece tokeniser. This effect is aggravated by the Dutch language having a lot of compound words and a much longer average word length (4.8 in English (Norvig, 2013) vs. 8 in Dutch (Corstius, 1981)).

For our experiments comparing the different BERT models, it was sufficient to split up long sentences in the training and test data as a data preprocessing step. However, for the inference described in Section 7.5.3, we did not preprocess the text, and as such, entities found in long sentences after 512 SentencePiece tokens will have been assigned the incorrect "O" label, skewing the results. In future research, we will implement an automatic sentence splitting module, similar to the one implemented in FLAIR (Akbik *et al.*, 2019).

7.7 Conclusion

In this paper, we have evaluated BERT models for Named Entity Recognition in the Dutch archaeological domain, with the purpose of improving our archaeological search engine. We implemented the search engine for a large archaeological text collection, with a structured query interface that allows the specification of entity types. The document collection is automatically annotated with archaeological named entities such as Location, Time Period, and Artefact.

In response to our research questions, first, we found that fine-tuning a BERT model with domain-specific training data improves the model's quality by a large margin for the archaeological domain, larger than in related work addressing domain-specific BERT models. We achieve an average F1 of 0.735 after hyperparameter optimisation, and very small standard deviations over runs with different random seeds.

Second, the domain-specific BERT model was superior in F1 and recall than an ensemble combining multiple BERT models, and could not be further improved by adding domain knowledge from a thesaurus in a CRF ensemble model. This indicates that after pre-training and fine-tuning on a domain-specific collection, the BERT model already covers the relevant information from the domain thesaurus. We did find a higher *precision* when we combined all three BERT models in a CRF model and added domain knowledge. However, as almost all information needs in archaeology are recall-oriented, and combining models is computationally expensive and environmentally taxing Strubell *et al.* (2020), we opt for the ArcheoBERTje model for labelling the full retrieval collection.

Third, our error analysis shows that there is confusion between the Artefact

and Material entities, similar to what humans experienced in the annotation process. For Artefacts and Time Periods, a common error is missing the adjective or preposition in an entity. The detection of Time Periods is a bit noisy, with other non-year numbers erroneously labelled as time ranges. Context entities such as "house" and "ditch" are difficult for the models to distinguish from non-entity words. Creating train-test splits on the document level is important to avoid overfitting, as the consistently misclassified Context "*poel*" shows, which only occurs in one document. An analysis of tokenisation by each of the models indicates that the multiBERT model is hampered by the rough tokenisation, splitting many relevant terms in sub-words.

In the near future, we will evaluate the entity-driven search engine with users, both in a controlled experiment and in natural search contexts. We will also investigate entity-based query suggestion. Once entities are mapped to a thesaurus or embedded in a semantic space, this allows for query improvement by suggesting parent or sibling entities in the thesaurus or nearest-neighbours in the embedding space.