Named Entity Recognition

Manual annotation of the entire thesis archive is not only time-consuming and non-scalable, but also prone to human error. Therefore, a more efficient automatic approach based on deep learning models was explored. A preliminary experiment compared two model types: **BERT** (Bidirectional Encoder model from Transformers), a pretrained NLP model commonly used for sequence labeling tasks, and **Gemini**, a general-purpose generative AI model developed by Google. The models were evaluated on their performance in the NER task, using the same labeling scheme and rules that were used by the human annotators.

The comparison was motivated by need to find the model that is most appropriate for the available training data. While BERT and related architectures are commonly used for NER, most models assume that the training data is fully and consistently annotated (Jie et al., 2019). They struggle when a large proportion of the entities are labeled as *outside* (O) due to incomplete annotation. As specified in the annotation guidelines, human annotators only labeled the first five instances of each entity, and restricted their work to the main sections, leaving the rest of the text untouched (annotated entities were automatically propagated to other occurrences of the same entity in the text). Consequently, while most identified entities are correct, not all relevant entities were identified, and examples of the negative class are not reliable. This creates noisy training data, which undermines the effectiveness of the BERT model's fine-tuning. Training and evaluation loss are commonly used to describe model convergence during training, but because of the high ratio of O-labels relative to all other labels (0.9975%), loss approaches zero even if entity predictions are incorrect. A large language model that can reason and has access to web content may be able to rely less on training data and more on background knowledge.

A further limitation of BERT models is their context window of just 512 tokens. Give that the training theses have an average word count of more than 40 thousand, texts need to be split into smaller sequences. While this is a reasonable approach for training for sentence-level learning, it limits the opportunity for document-level learning. Gemini models can have context windows of more than a million tokens, allowing them to process entire theses. This allows them to benefit from document-level patterns such as that the *Title* and *Author* are always found near the beginning of a document, and that the *subject* label should only appear a few times in a document.

Model evaluation

A custom evaluation methodology was developed to compare the effectiveness of the models in extracting entities, with a focus their applicability to the downstream task of entity linking. Standard evaluation metrics of NER models are derived from a confusion matrix. Rather than computing metrics at a sentence level, they were calculated at a document level, based on the model's prediction across an entire thesis.

For evaluation, the set of entities in the reference data was compared against the set of entities predicted anywhere within the same thesis, regardless of their exact location or number of occurrences. For example, if *“Nieuwegein”* was labeled as *spatial* in a given thesis, and the model predicted *“Nieuwegein”* as *spatial* at any point, this was counted as a **True positive**, even if these did not correspond to the same textual occurrence of the entity. This approach compensates for the missing labels in the annotations, and reflects the practical goal of capturing the correct entities for subsequent knowledge modeling.

Confusion matrix

	Labeled in reference	Not labeled in reference
Captured by model	**True positive (TP)**: Entity was annotated and was captured by the model	**False positive (FP)**: Model predicted an entity that was not annotated (a spurious label)
Not captured by model	**False positive (FN):** Model did not predict an entity that was labeled	**True negative (TN):** Entity was not labeled and the model did not predict it (can’t be counted)

Due to the sparseness of the reference annotations, true negatives can’t be reliably counted and therefore should be excluded from evaluation metrics. It is also expected that a well-performing model may predict valid entities that were not labeled by the human annotators. Therefore, several metrics were used to evaluate the models based on binary classification from the confusion matrix.

Recall is the true positive rate (TRP), calculated as the proportion of annotated entities that were correctly identified by the model. A high recall indicates that the model was able to successfully capture the entities that had been manually labeled, even if it made additional unverified guesses. The conceptual opposite is precision, the proportion of all the model's predictions that had been in the manual annotations. F1 score is their harmonic mean. Finally, the Jaccard similarity index is a quantification of the similarity between two sets. It provides a more straightforward bag-of-words comparison between the predicted and reference sets.

$$\text{Recall (or TPR)} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{TP}{TP + FP}$$

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

$$\text{Jaccard index} = \frac{TP}{TP + FP + FN}$$

For the BERT model, metrics were computed for each evaluation thesis after each epoch and then averaged, and the best epoch was reported. For the Gemini model, they were calculated from the response produced for each evaluation thesis. Results are reported per entity label and as a weighted average based on reference label frequency.

As was discussed in the previous section on inter-annotator agreement, NER evaluation can be ambiguous since it assess not only if an entity was given the right label, but also whether the entity span matches exactly the reference or only partially overlaps with it (Seow et al., 2025). In this comparison, a string span-matching approach was used, requiring predicted entity spans to match the reference spans in order to count as correct. This was done because the inclusion of different tokens can change the intended meaning of an entity, particularly for the *Spatial* and *Subject* labels.

To account for scanning errors and minor variations that do not affect meaning, string similarity matching with Levenstein Distance was used. The Normalized Levenstein Distance quantifies the similarity between two strings by counting the number of single-character edits between them normalizing by the length of the longest string (Yujian & Bo, 2007). Strings were converted into lowercase and all punctuation was removed, and pairs with a distance of more than 0.9 were considered to be a match. In all cases, the labels type assigned by the model was required to be the same as the reference label.

A document-level splitting strategy was used to randomly divide the annotated theses into a training set of 16 and an evaluation set of 4. This approach was chosen over random sequence-level splitting chiefly to reflect real-world document processing and achieve a relatively balanced distribution of label types between training and validation. Using the same training and validation data also enables a fair comparison between the different types of models.

Model Results

BERT is pre-trained on large language corpora to understand context using bidirectional masked language modeling and next-sentence prediction (Devlin et al., 2019). Through transfer learning, it can be fine-tuned on annotated training data for specific NLP tasks. This study used BERTje, a monolingual Dutch BERT model with a token-level classification head for sequence labeling tasks (Vries et al., 2019).

The BERTje checkpoint[1] was fine-tuned on two versions of the training dataset to assess the effect of label sample size. The original BIO labeling scheme resulted in 13 possible labels, leading to a small sample size for per class. To avoid this, the documents were re-labeled with an IO (inside-outside) scheme. This reduced the number of labels, increased the sample size of each, and made explicit that that tokens that were previously labeled with B or I and the same label type are semantically related.

Label Sample Sizes

[1] https://huggingface.co/GroNLP/bert-base-dutch-cased

Label	BIO-labeling	IO-labeling
O	514852	514852
I-title	281	311
I-author	64	113
I-issued	19	36
I-spatial	152	309
I-subject	208	338
I-inGroup	116	160
B-title	30	-
B-issued	17	-
B-author	49	-
B-spatial	157	-
B-inGroup	44	-
B-subject	130	-

The models were trained on a single GPU for 10 epochs and batch size of 8. After each epoch, evaluation metrics were computed using the method discussed earlier, and the weights of the epoch with the highest F1 score were saved.

[Here is where the BERT model results will go]

Gemini Model

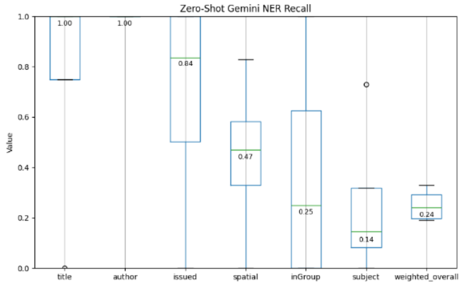
Gemini 3 Flash is a proprietary reasoning model intended for multi-model inputs. This model was chosen among commercial LLMs because it is the latest model with a free tier. Two different iterations of the model were tested with different prompts, which correspond to a few-shot and a zero-shot run. In the few-shot run, the model was provided with instructions, annotation rules, two examples of labeled texts, and an unlabeled thesis to annotate. Only two training examples could be provided (instead of 16 like in the BERT model) because of the model's limit on 250,000 input tokens per minute on the free tier. The training examples were randomly chosen for each run. The annotation rules were taken from the same annotation guide used by the human annotators. In the zero-shot run, the same prompt was used, but examples were not provided. English-language prompts were created based on the prompt engineering recommendations[2] suggested by Google, and a few different versions were tested to find the best-performing prompt. The prompts used are shown in the appendix.

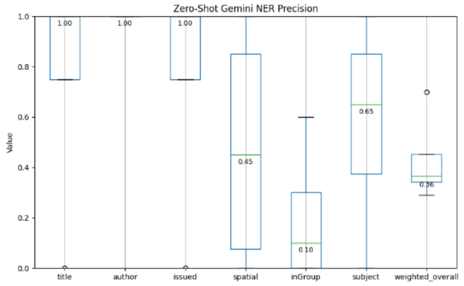
The following boxplots show the distribution of recall, precision, F1, and Jaccard index when the models were applied to the four validation theses. (Please note that I don't think that all of these

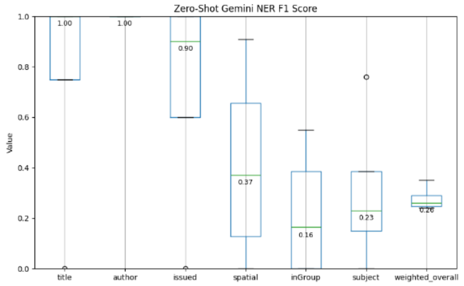
[2] https://ai.google.dev/gemini-api/docs/prompting-strategies

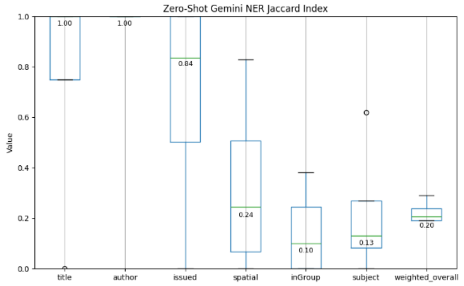
boxplots should be presented in the thesis, perhaps just the means, but I wanted to show the differences in results between the four thesis).

Gemini model Zero-shot

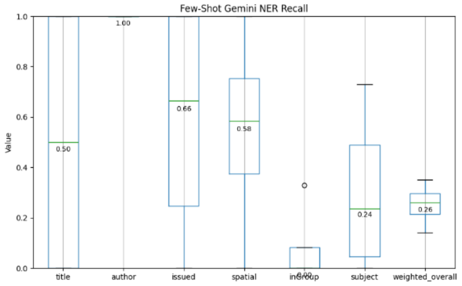

Zero-Shot Gemini NER Recall

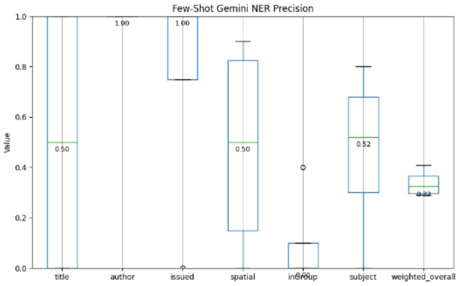

Zero-Shot Gemini NER Precision

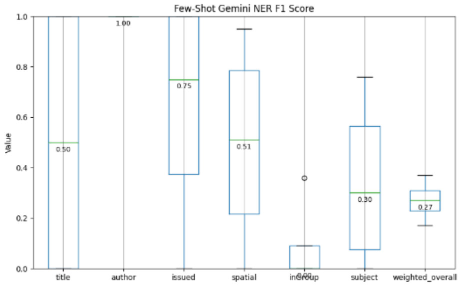

Zero-Shot Gemini NER F1 Score

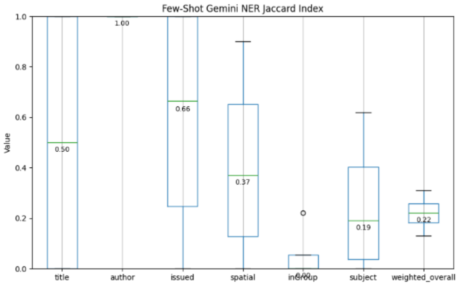

Zero-Shot Gemini NER Jaccard Index

Gemini model few-shot


Few-Shot Gemini NER Recall


Few-Shot Gemini NER Precision


Few-Shot Gemini NER F1 Score


Few-Shot Gemini NER Jaccard Index

Inference

The model determined to produce the best results will be used for inference. The inference results of form the pool of entities for each thesis that can now be used for entity linking.

Appendix: Gemini-3-flash prompts

Few-shot prompt:

```
Named Entity Recognition task in Dutch.

Below are 2 examples showing the text of human-geography theses. Each word in the
thesis is labeled as a title (Title of the thesis), author(Author of the thesis),
issued (Time of publication), spatial (The spatial extent, study area or spatial
coverage of the entire thesis, given as placenames or place descriptions),
subject (Concept that is the main subject of the thesis), inGroup (the group of
persons studied), or as O (empty label). Non-empty labeled can start with a B- to
indicate the beginning of an entity span, or as I- to indicate the inside of an
entity span. After reading the examples, extract the tags from the new thesis
text, and return it as a JSON, where each key is a label and each value is a list
of entities that have that label. Only extract entities that are in the text.
Follow the annotation rules.

Annotation rules:

• Don’t tag exhaustively, but only the first 5 mentions (e.g. a particular place,
or a particular method) of any given concept in those sections that should be
searched
• Restrict search to particular sections: Don’t use TOC, no prefaces. Focus on
the main sections (introduction/method/conclusion). Avoid using sections
literature review, background, results, or TOC or literature list
• If several different concepts are mentioned in a sentence, annotate them
separately
• inGroup: should be specific for the research design
• Subject: Leave out subjects unless there is explicitly a conceptual model (key-
concepts)
• Spatial: the largest extent of the research area related to a goal/question.
Any spatial level that links to a different research goal can appear separately.
In case there is no placename available for this, encode the information on most
specific level that is there (“plein in Amersfoort”).

Examples:

Example 1:
[Text of training thesis 1]

Example 2:

[Text of training thesis 2]
```

```
Please annotate the following student thesis:

[Text of Evaluation thesis]

"""
```

Zero-shot prompt:

```
Named Entity Recognition task in Dutch.

Label each word in this human-geography thesis as a title (Title of the thesis),
author(Author of the thesis), issued (Time of publication), spatial (The spatial
extent, study area or spatial coverage of the entire thesis, given as placenames
or place descriptions), subject (Concept that is the main subject of the thesis),
inGroup (the group of persons studied), or as O (empty label). Return it as a
JSON, where each key is a label and each value is a list of entities that have
that label. Only extract entities that are in the text. Follow the annotation
rules.

Annotation rules:

• Don’t tag exhaustively, but only the first 5 mentions (e.g. a particular place,
or a particular method) of any given concept in those sections that should be
searched
• Restrict search to particular sections: Don’t use TOC, no prefaces. Focus on
the main sections (introduction/method/conclusion). Avoid using sections
literature review, background, results, or TOC or literature list
• If several different concepts are mentioned in a sentence, annotate them
separately
• inGroup: should be specific for the research design
• Subject: Leave out subjects unless there is explicitly a conceptual model (key-
concepts)
• Spatial: the largest extent of the research area related to a goal/question.
Any spatial level that links to a different research goal can appear separately.
In case there is no placename available for this, encode the information on most
specific level that is there (“plein in Amersfoort”).

Please annotate the following student thesis:

[Text of Evaluation thesis]

"""
```