Linking References to Documents in Parliamentary Debates

Floris Bos[1], Marc van Opijnen[2 [0000−0001−6922−3390]], and Maarten Marx[1 [0000−0003−3255−3729]]

[1] IRLab, Informatics Institute, University of Amsterdam
https://irlab.science.uva.nl
[2] Publications Office of the Netherlands (Logius|Koop), The Hague
marc.opijnen@koop.overheid.nl,maartenmarx@uva.nl

The source code for this research is publicly available at https://github.com/BluntKatana/linking-references-to-docs-in-parl-debates.

Abstract. In Dutch parliamentary debates, over 95% of references to documents are implicit and non-standardized, hindering document accessibility and analysis. To address this challenge, we introduce a two-phase approach to automatically detect and link these references. The first phase uses a Large Language Model (LLM), specifically Gemini 2.5 Flash, for reference detection and semantic enrichment, extracting features like document type, a summary, and keywords. The second phase links these references to known documents using vector similarity search. Our large-scale analysis of 281 debates confirms the scale of the problem, revealing that nearly 74% of all detected references are implicit. Evaluation on a new, manually annotated gold-standard dataset of 191 references shows our detection method achieves an F1-score of 0.49, while the LLM classifies semantic features like document type with 92-97% accuracy. For the linking task, evaluated on 1,933 references, combining LLM-generated keywords with metadata filtering proves most effective. This approach correctly identifies the target document in 35% of cases (Hit@1) and places it in the top 10 candidates of 57% of the time (MRR 0.42). This work serves as a strong baseline for resolving complex, implicit references in a parliamentary proceedings. The methodology is inherently language-agnostic and shows significant promise for adaptation to other domains, such as legal case law or historical archives.

Keywords: Document Linking · Large Language Models · Known-item Search · Semantic Search

1 Introduction

"The meaning of a document is in its use", free after Witgenstein, is a statement librarians, archivists and IR scholars can relate to. Think of bibliometrics, the use of PageRank as a quality indicator, and anchor text as concise summaries as examples in which the use of a document, witnessed by a reference to it in another document, is employed to enrich the representation of the document, leading to improved performance in e.g., retrieval tasks.

We can access this use of documents in a corpus if the corpus is structured as a (weighted) directed network in which the links indicate references between documents. Hypertext, standardization (URLs), Web 2.0 techniques like WikiLinks, but also powerful parsing tools like GROBID [13] have made these links (almost) directly available in specific document collections. In other cases, the references are implicit and work has to be done to turn a document collection into a directed network.

This task is more generally known in NLP as Named Entity Recognition (NER) and Linking. Nowadays, NER can be performed with high accuracy by LLMs, needing no or only a few training examples [24]. Linking —finding out what a certain string is referring to— is usually treated as a known-item search task: there exists exactly one correct referent (although it may not be present in the use knowledge base). The difficulty of this task depends on the amount of homonimity (same string referring to multiple entities) and the synonymity (one entity having multiple "names"). For the traditional NER types, persons, organizations, locations, these aspects are usually handled using context and semantic embeddings [11].

Examples of corpora containing documents implicitly referring to other documents are court cases referring to laws and other court cases [15], scientific articles mentioning the use of datasets [17,19] or containing archival references [21].

In this paper, we study references to parliamentary documents (legislative proposals, motions, amendments, letters by ministers, etc) made in parliamentary debates. Examples of such references, with the reference to a parliamentary document in italics, are the following:

- "Hoe denkt mevrouw Hermann om te gaan met *de motie die zij in dit verband heeft ingediend*?" [3]
- "In november 2001 heeft de Kamer *een motie aanvaard waarin de regering werd gevraagd haar een notitie toe te sturen.*" [4]
- "In *de brief* heeft u ook aangetroffen dat er gewerkt wordt aan een structurele aanpak." [5]

For Dutch parliamentary proceedings, this is indeed a problem as less than 5% of these references are made explicit using a identifier [23]. We show how a multilingual LLM can solve the recognition task very well, while generating several disambiguation and retrieval features. These features are turned into semantic embeddings and matched against embeddings of the documents for linking the reference. Our best performing system successfully identifies 44% of the references with a precision of 55%. It accurately links 35% of references to the

[3] Translation: "How does Mrs. Hermann intend to deal with *the motion she has submitted in this regard*?

[4] Translation: "On november 2001 the Parliament has accepted *a motion in which the government has asked her to send a memorandum.*"

[5] Translation: "In *the letter* you also found that work is being done on a structural approach."

right document, with the correct document appearing in the top 10 candidates 57% of the time. We also release a manually curated dataset of 5 debates with 191 identified references from the Dutch House of Parliament, which serves as the gold standard for further evaluation.

Our approach is not novel, but can be seen as a strong baseline for the document linking task in a domain with novel challenging characteristics like 1) long references (log normally distributed, median=1.39, max=4.08), 2) anaphoric references ("my letter to the minister of last spring") and 3) vague and hard to resolve references.

We end this introduction with a motivation for this task based on the related case of references to datasets in scientific articles. Searching for datasets is difficult for humans [9], despite the fact that several well developed search systems [2] and an extensive body of research exists [3]. According to [9], this is due to the mismatch between the information need and the representation of the datasets. Users search for a dataset which is useful given a specific task. But the representation of the dataset is based on the metadata or the data it contains, and not on how the dataset is, or can be, used. The latter of course can be found in articles which mention the dataset. Enriching the representation of datasets with their use thus gives rise to the field of dataset mention detection [8].

2 Related work

The detection and reconcilliation of dataset mentions in scientific articles is a task similar to ours, with similar features, in particular no standardized manner of referencing and thus a large variation in references to the same entity [17]. Several open benchmarks have been created [8,19]. Tested on the DMDD benchmark, a BERT embedding based classifier worked best for detection, followed by a CRF model from [8] using carefully chosen features (POS tags and dataset cue terms), which outperformed a BiLSTM classifier on fully learned features. An initial experiment testing entity linking showed little difference between lexical (BM25) and embedding (ColBERT, not fine-tuned) methods [19].

A problem which resembles our case because of the use of numerical identifiers in references to documents is linking in legal documents, to both case law and legislation. Several systems have been proposed, from rule-based to systems learned from examples [16,1,20,22].

An interesting case in which the representation of a document cannot be based on its content is described by Suzuki and Oard [21]. They want to enable search in huge archives of physical boxes with documents which for the sheer reason of size cannot be OCRed. References to these boxes in scientific articles are thus the only way to semantically enrich their representation. This resulted in the SUSHI shared task at NTCIR 2025 where participants must create a search system returning archival records based on a set of documents with detected references to them [14].

3 Methods

Our approach is a two-phase system: the first phase extracts candidate references from a Dutch parliamentary minute and enriches them with semantic features, and the second links them to known Dutch parliamentary documents using vector similarity search. We also introduce the annotated dataset used for development and evaluation.

3.1 Recognition, description, and improving precision

The first stage of our system identifies, enriches, and validates the references. This recognition process uses a large language model (LLM), specifically Gemini 2.5 Flash [7]. This model allows us to process entire minutes in a single pass due to its high token limits (input=1,048,576 tokens, output=65,535 tokens). Using the LLM, we identify reference spans in the text. We also let the LLM extract several semantic features which are used later in the linking phase. These features include the full sentence containing the text span, a predicted document type (e.g., motion, letter, legislative proposal), the reference type as defined later in Table 1, a concise summary of the reference's content, and a set of thematic keywords, preferably aligned with the TOOI thematic thesaurus used across Dutch parliamentary documents [18].

To evaluate our approach, we use four prompting strategies based on the number of examples provided in the prompt (zero-shot vs. few-shot), and whether detection and validation are combined into a single prompt or separated into two prompts (single-pass vs. two-pass). A two-pass approach allows us to optimize for high recall first, then focus on precision after. In all prompts, we also let the LLM give a confidence score to each reference, using this score, we are able to filter out any vague, hypothetical, or future references.

3.2 Linking

Once candidate references are identified, the system attempts to link them to known parliamentary documents through semantic matching and narrowing down the vector search space by filtering. The problem is treated as known-item search, as each parliamentary document in our corpus is directly tied to a permanent identifier, often composed of a dossier number and a sequential document number (e.g., "34775-12"). Once an identifier can be linked to a reference, we have a match.

In this linking step we use the semantic features generated during recognition to create a query vector and/or filter the search space. This query vector is used to match against precomputed document vectors of known parliamentary documents using cosine similarity. The query vectors are built up in a modular way by concatenating them into a single string before embedding. This allows us to perform ablation experiments. For instance, we can assess the impact of

using only the sentence and keywords, or the sentence and summary, or the sentence and summary with filtering on the publication year. This modular design supports extensibility.

The content of each parliamentary document in the search space (N=14,027) is embedded using the gte-mulitlingual-base model [25], which is selected for its strong multilingual performance and 768-dimensional output vectors. Alongside the embedding the document type, date issued, and identifier are stored in an Elasticsearch index [5]. This allows us to filter the search space before performing vector similarity search using cosine similarity. As linking is performed offline and on a finite corpus of static parliamentary documents, speed is not a major issue in our current system.

3.3 Annotated Dataset

We created an annotated dataset from 5 recent debates from the Dutch Parliament consisting in total of 1.841 sentences and 30.142 words. 93 sentences (5.05%) contained one of more references to parliamentary documents. The annotation protocol was based on [12].

The following text strings were annotated as references: any sequence of words within the parliamentary minutes that refers to a parliamentary document or dossier. This includes explicit mentions of dossier or document numbers, as well as implicit references (e.g., by name, context, or document type). References to relevant third-party documents directly related to the parliamentary discussion were also annotated. For each identifier hit, annotators provided additional information. The most important information is the 'reference type', with definitions provided in Table 1. Depending on this reference type, annotators also provided

Table 1. Definitions for each reference type in the annotation protocol.

Reference Type	Definition
explicit-dossier	Explicit mention of a dossier number.
explicit-parl-doc	Explicit mention of a dossier number followed by a document number.
impl-local	Reference to a document or dossier already mentioned, without an explicit number.
impl-ext-dossier	Reference to an identifiable dossier without explicit mention of a dossier number.
impl-ext-parl-doc	Reference to a specific parliamentary document without numbers.
impl-third-party	Reference to materials not part of the parliamentary database but relevant to the meeting.

additional metadata: the type of parliamentary document (e.g., letter, motion or legislative proposal), the correct external identifier (such as the dossier and/or document number), or an identifier linking to a previously annotated reference (in case of a document reference making an anaphoric reference to an earlier (often explicit) document reference).

We evaluated inter-annotator agreement (IAA) on 194 matched references. We found substantial agreement for 'reference type' (Cohen's $\kappa = 0.87$) and

'document type' ($\kappa = 0.89$) [4]. Agreement for 'external identifier' was moderate ($\kappa = 0.59$), confirming the inherent difficulty of the linking task.

3.4 Evaluation metrics

We evaluate detection of references using precision, recall and F1. As references can be long, exact match would be too strict. Instead, we follow [10], and state that a pair (t, p) of a true and a predicted string match iff the Jaccard similarity of the two sets of tokens is strictly larger than 0.5. This yields a one-to-one mapping. We use the same matching to produce Cohen's κ. We view recognition as a known item search task and evaluate it with mean $Hit@k$ (indicating whether the desired item is among the top k ranked results) and mean reciprocal rank (MRR).

4 Results

We present the evaluation of our system in two stages. First, we determine the best prompting strategy for reference detection using our annotated dataset. Second, we conduct a large-scale experiment on a full parliamentary year to demonstrate the system's performance on detection and linking in a real-world scenario.

4.1 Reference Detection

Table 2 contains the detection results for the four prompting strategies. The results show clearly that the agentic setup with two passes performs better and that providing a few examples is also beneficial. In the linking experiments, we will work with the correctly detected references by the fewshot-two-pass system.

We also asked the LLM to classify each detected reference along two dimensions: the document type (choice of 9) classes, like letter, motion, law, etc) and the type of reference (implicit, explicit, in total 6 classes). These tasks were easy for the LLMs, with accuracies ranging between 92 and 97 percent except for document type prediction with the zero-shot-single-pass LLM which only scored 43%.

Table 2. Detection performance for each prompt strategy on the gold-standard dataset (N=191).

Strategy	Precision	Recall	F1
zeroshot-single	0.47	0.38	0.42
fewshot-single	0.48	0.40	0.44
zeroshot-two-pass	0.53	0.37	0.44
fewshot-two-pass	**0.55**	**0.44**	**0.49**

4.2 System in Action: A Large-Scale Experiment

To assess the practical applicability of our system, we applied our fewshot-two-pass strategy to 281 plenary debates from the 2019-2020 parliamentary year, totaling 3,318,439 words in

Table 3. Distribution of detected reference types in large-scale experiment (N=14,976).

Reference Type	Count	Perc.
Implicit Local	6,432	42.3%
Explicit	3,825	26.2%
Implicit Parl. Doc.	2,618	17.5%
Implicit Dossier	1,084	7.0%
Third Party	1,014	7.0%

191,009 sentences. This large-scale analysis allows us to first understand the prevalence and nature of references in parliamentary proceedings and then to create a high-confidence dataset for the evaluation of our linking method. In these 281 minutes, the system detected 14,976 references. The detected references amount to an average of 53.3 references per minute (median=49). On average every 13th sentence contains a reference to a document. The distribution of references over the 5 types, summarized in Table 3, shows that the vast majority (73.8%) is implicit, highlighting the need of a semantic approach next to simple rule-based methods. The length in number of words of the found references is lognormally distributed with a median of 3, a mean of 4, and a long tail of long (max is 58 words) and often linguistically complex references.

4.3 Reference Linking

We can use the 3,825 explicit references found in the large-scale detection experiment (see Section 4.2) to evaluate the performance of our linker. For these, we have the correct ground truth (the explicit link), which after removal creates a realistic implicit query. For example, the explicit reference "De motie-Moorlag (31532, nr. 248)" becomes the implicit query "De motie-Moorlag ()", with "31532-248" as the (correct) identifier. Of these 3,825 explicit references, 1,892 consisted of only a reference number and thus were not usable, leaving us with 1,933 test instances. The system had to find the correct document in a search space of 14,027 parliamentary documents sourced from the 2019-2020 parliamentary year.

The results of this linking experiment are shown Table 4. They demonstrate that enriching the query with LLM-generated semantic features is crucial. The best performing combination of an enriched query (text+sentence+keywords) and metadata filtering (year+doctype) correctly links the document 35% of the time and places it in the top 10 candidates 57% of the time.

Table 4. Linking performance for different feature combinations (N=1,933). The average 95% CI across all measurements is approximately ±0.03.

Features		Hit@1	Hit@10	MRR
Query	Filtering			
text+sentence	-	0.03	0.06	0.04
text+sentence+summary	-	0.14	0.30	0.19
text+sentence+keywords	-	0.26	0.51	0.34
text+sentence+keywords	doctype	0.30	0.55	0.39
text+sentence+keywords	year	0.30	0.55	0.38
text+sentence+keywords	year+doctype	**0.35**	**0.57**	**0.42**

When no filtering is applied, using keywords (text+sentence+keywords) is much more effective than using a summary (text+sentence+summary). Both the Hit@1 score and MRR nearly double from 0.14 to 0.26, and 0.19 to 0.34, respectively. We hypothesize that keywords provide a more discriminative and concise representation for known-item search. While a summary captures the general topic,

it can introduce narrative language and noisy vocabulary, leading to topic-drift. In contrast, keywords, especially when aligned with the thesaurus, function as key descriptors that are more likely to match the characteristics of the target document.

The data also highlights the significant benefit of metadata filtering. Applying a filter for document type or year to the keyword-based query leads to a clear improvement in scores. This shows that narrowing the search space is a critical step. By pre-filtering on metadata extracted by the LLM, the system can eliminate a large number of semantically similar but incorrect documents, which inherently improves the precision of the final vector search.

5 Conclusion and Future Work

We presented a two-phase approach for detecting and linking references to documents within Dutch parliamentary debates, moving beyond simple text matching to address the challenge of implicit, non-standardized references. Our method uses a large language model for the initial detection and semantic feature extraction, followed by vector similarity search for linking.

Evaluation against a manually created gold standard dataset of 191 references showed that few-shot, two-pass prompting strategy is most effective for reference detection. A subsequent large-scale experiment on 281 debates revealed that the vast majority of references (roughly 74%) are implicit. For the linking task, we demonstrated that enriching queries with LLM-generated semantic features and applying metadata filters is critical, achieving a Hit@1 of 35% and placing the correct document in the top 10 candidates 57% of the time.

Our approach shows strong potential for generalization. First, its core components being a multilingual LLM and a multilingual embedding model make the approach largely language-agnostic. Future work could validate performance across different languages using similar corpora, such as ParlaMint [6], which contains parliamentary debates from more than 20 European countries. Second, the task of resolving implicit, non-standardized references is not unique to parliamentary proceedings. The same implementations could be adapted to other domains like legal case law, corporate financial reports, or historical archives, where linking to supporting documents is useful for comprehension and analysis.

For future work, we plan to improve linking accuracy by adding a dedicated re-ranking step for the top candidates. We will also explore fine-tuning smaller, specialized models for the detection phase to create a more efficient and accurate system. Furthermore, we aim to investigate the robustness of our system by quantifying the impact of variance in the AI's non-deterministic outputs on overall performance. Finally, implementing a human-in-the-loop framework could support continuous improvement through active learning, creating a system that adapts over time.

Acknowledgments. Thanks to Wietske Boersma. This research was supported in part by the Netherlands Organization for Scientific Research (NWO) through the ACCESS

project grant CISC.CC.016 and an Open Science Fund grant nr 01607400. Maarten Marx is partly funded by ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

References

1. Agnoloni, T., Bacci, L., Peruginelli, G., van Opijnen, M., et al.: Linking european case law: BO-ECLI parser, an open framework for the automatic extraction of legal links. In: Proc. JURIX '17, pp. 113–118. IOS Press (2017)
2. Brickley, D., Burgess, M., Noy, N.: Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In: Proc. WWW '19. pp. 1365–1375 (2019). https://doi.org/10.1145/3308558.3313685
3. Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., et al.: Dataset search: a survey. The VLDB Journal **29**(1), 251–272 (2020)
4. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
5. Elasticsearch: https://www.elastic.co/elasticsearch (2025), accessed: 2025-06-20
6. Erjavec, T., Kopp, M., Ljubešić, N., Kuzman, T., et al.: ParlaMint II: advancing comparable parliamentary corpora across Europe. Language Resources and Evaluation (Dec 2024). https://doi.org/10.1007/s10579-024-09798-w
7. Google Cloud: Gemini 2.5 Flash | Generative AI on Vertex AI. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash (2025), accessed: 2025-05-27
8. Heddes, J., Meerdink, P., Pieters, M., Marx, M.: The automatic detection of dataset names in scientific articles. Data **6**(8), 84 (2021)
9. Hulsebos, M., Lin, W., Shankar, S., et al.: It took longer than I was expecting: Why is dataset search still so hard? In: Proc. HILDA '24. pp. 1–4 (2024). https://doi.org/10.1145/3665939.3665959
10. Kirillov, A., He, K., Girshick, R., Rother, C., et al.: Panoptic Segmentation. In: Proc. CVPR '19. pp. 9404–9413. IEEE, Long Beach, CA, USA (2019)
11. Kolitsas, N., Ganea, O.E., Hofmann, T.: End-to-end neural entity linking. arXiv preprint arXiv:1808.07699 (2018)
12. Lee, S., DeLucia, A., Nangia, N., Ganedi, P., et al.: Common law annotations: Investigating the stability of dialog system output annotations. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the ACL 2023. pp. 12315–12349 (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.780, https://aclanthology.org/2023.findings-acl.780/
13. Lopez, P.: GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In: Proc. TPDL '09. pp. 473–474. Springer (2009)
14. Oard, D.W., Suzuki, T., Ishita, E., Kando, N.: Searching unseen sources for historical information: Evaluation design for the NTCIR-18 SUSHI pilot task. In: Proc. EMTCIR '24. vol. 3854 (2024)
15. van Opijnen, M.: Citation analysis and beyond: in search of indicators measuring case law importance. In: Proc. JURIX '12, pp. 95–104. IOS Press (2012)
16. van Opijnen, M., Verwer, N., Meijer, J.: Beyond the experiment: the eXtendable legal link eXtractor. In: Proc. Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts (ASAIL), ICAIL'15 (2015)

17. Otto, W., Zloch, M., Gan, L., Karmakar, S., et al.: GSAP-NER: A novel task, corpus, and baseline for scholarly entity extraction focused on machine learning models and datasets. In: Proc. EMNLP '23. pp. 8166–8176 (Dec 2023), https://aclanthology.org/2023.findings-emnlp.548/
18. Thema-indeling voor Officiële Publicaties (TOP-lijst). https://standaarden.overheid.nl/tooi/waardelijsten/work?work_uri=https%3A%2F%2Fidentifier.overheid.nl%2Ftooi%2Fset%2Fscw_toplijst, accessed: 2025-06-20
19. Pan, H., Zhang, Q., Dragut, E., Caragea, C., et al.: DMDD: A large-scale dataset for dataset mentions detection. Trans. Assoc. Comput. Linguistics **11**, 1132–1146 (2023)
20. Savelka, J., Ashley, K.D.: Using conditional random fields to detect different functional types of content in decisions of united states courts with example application to sentence boundary detection. In: Proc. Workshop on Automated Semantic Analysis of Information in Legal Texts (ASAIL), ICAIL'17. vol. 10 (2017)
21. Suzuki, T., Oard, D.W., Ishita, E., Tomiura, Y.: Automatically detecting references from the scholarly literature to records in archives. In: Proc. Int. Conf. on Asian Digital Libraries. pp. 100–107. Springer (2023)
22. Varga, D., Gojdic, M., Szoplák, Z., Gurskỳ, P., et al.: Extraction of legal references from court decisions. In: Proc. ITAT '23. pp. 89–95 (2023)
23. Venema, P.L.P.: Improving Links to Referenced Documents in Dutch Parliamentary Meeting Notes. Bachelor's thesis, University of Amsterdam, Amsterdam, The Netherlands (2024), https://scripties.uba.uva.nl/search?id=record_54697
24. Wang, S., Sun, X., Li, X., Ouyang, R., et al.: GPT-NER: Named entity recognition via large language models. arXiv preprint arXiv:2304.10428 (2023)
25. Zhang, X., Zhang, Y., Long, D., Xie, W., et al.: mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In: Proc. EMNLP (Industry Track) '24. pp. 1393–1412 (2024)