UNIVERSITEIT VAN AMSTERDAM

Linking References to Documents in Parliamentary Debates

Floris Bos[1] Marc van Opijnen[2] Maarten Marx[1]

[1]IRLab, University of Amsterdam
[2]Publications Office of the Netherlands (KOOP)

Why It Matters

Less than 5% of references in Dutch debates are explicit. Around 74% are implicit, hindering access, analysis, and linking across proceedings. Resolving these links improves transparency and powers IR/analytics. We provide a strong baseline, a gold dataset, and code.

Data at a glance

- 281 debates (2019-2020), 191.000 sentences.
- 14.976 references detected; ~74% implicit.
- Gold standard: 5 debates, 191 references.
- Linking eval: 1.933 implicitized queries
- Search space: 14.027 parliamentary documents

Examples

Implicit
  - "In the letter …",
  - "the motion we submitted …"

Explicit
  - "The motion-Moorlag (31524, nr. 248)"

Implicit references dominate Dutch debates; a two-phase LLM + vector search baseline links 35% @1 and 57% top-10 (N=1.933)

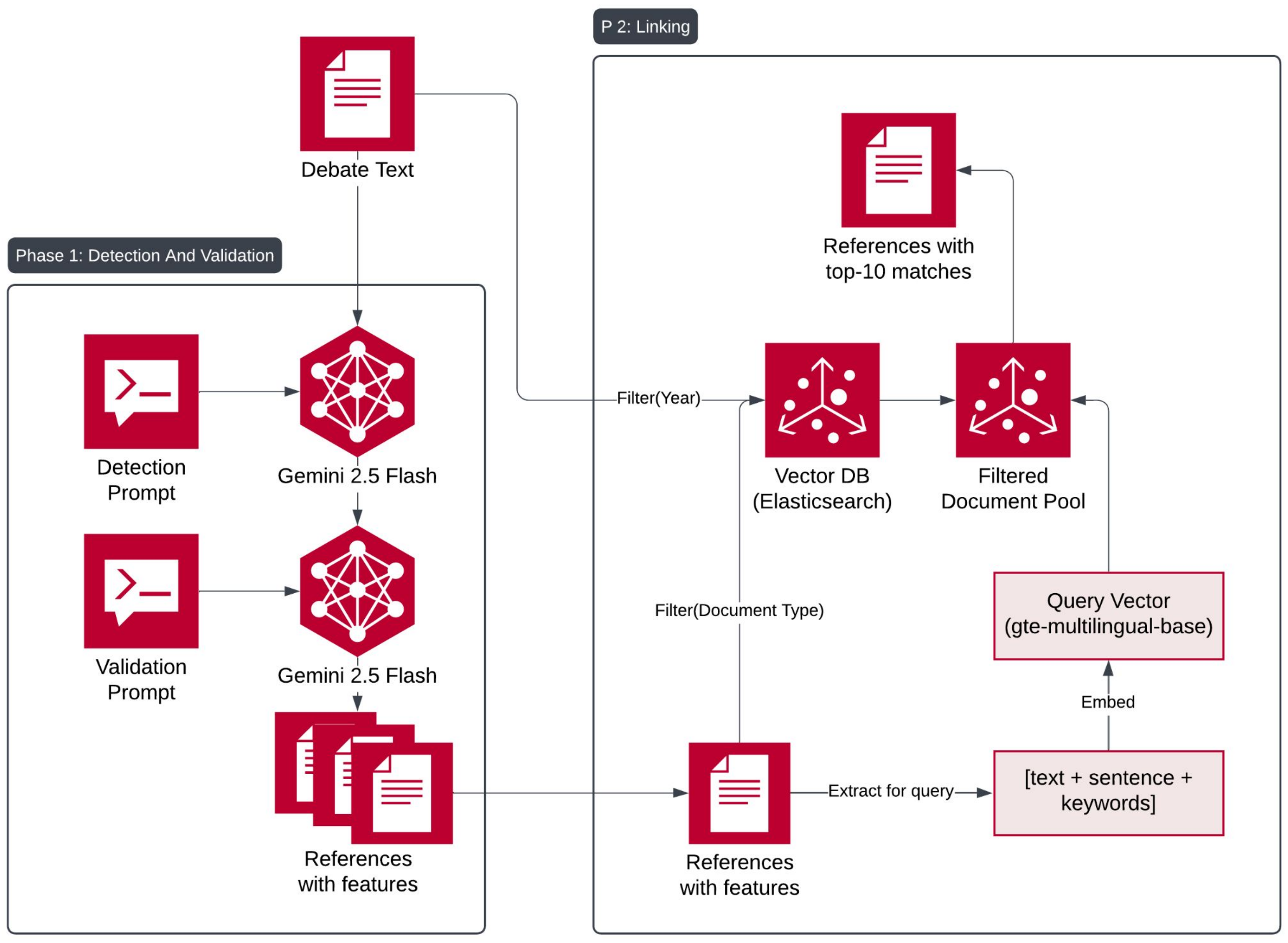
Key results

0.49
Detection F1 (few-shot, two-pass)

0.35
Linking Hit@1 (N=1.933)

0.57
Hit@10; MRR =0.42

Pipeline (high-level)



Scan for paper and code

https://thesis.florisbos.com

How we did it

- **Detect & enrich**: Few-shot, two-pass LLM (Gemini 2.5 Flash) selects spans and predics document/reference type, sentence, summary, TOOI-aligned keywords.
- **Link**: Build a modular query vector (sentence + keywords). Filter candidates by year/doctype; cosine similarity over gte-miltilingual-base embeddings in Elasticsearch

Limitations

- Anaphora and vague mentions remain hard
- Hit@1 and recall leave room for re-ranking

What is next?

- Candidate re-ranking for top-k
- Richer features/metadata
- Smaller fine-tuned models
- Human-in-the-loop