Explore, Propose, and Assemble: An Interpretable Model for Multi-Hop Reading Comprehension

Yichen Jiang* Nitish Joshi* Yen-Chun Chen Mohit Bansal

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Code available at https://github.com/jiangycTarheel/EPAr

Outline

UNC NLP

1. Motivate Multi-Hop QA (Yichen)
2. Introduce our 3-module system EPAr (Yichen)
3. Datasets + Main Results (Nitish)
4. Ablations (Nitish)
5. EPAr Demo (Nitish)
6. Related Works and Conclusions (Nitish)

Multi-Hop QA

[Welbl et al. 2018]

The Haunted Castle (Dutch : Spookslot) is a haunted attraction in the amusement **park Efteling** in the Netherlands . It was designed by Ton van de Ven and ...

Query subject: *The Haunted Castle*
Query body: located_in_the_administrative_territorial_entity

Multi-Hop QA

[Welbl et al. 2018]

The Haunted Castle (Dutch : Spookslot) is a haunted attraction in the amusement **park Efteling** in the Netherlands . It was designed by Ton van de Ven and ...

Efteling is a fantasy-themed amusement park in **Kaatsheuvel** in the Netherlands. The attractions are based on elements from ancient myths and legends, fairy tales, fables, and folklore.

Query subject: *The Haunted Castle*
Query body: located_in_the_administrative_territorial_entity

Multi-Hop QA

[Welbl et al. 2018]

The Haunted Castle (Dutch : Spookslot) is a haunted attraction in the amusement **park Efteling** in the Netherlands . It was designed by Ton van de Ven and ...

Efteling is a fantasy-themed amusement park in **Kaatsheuvel** in the Netherlands. The attractions are based on elements from ancient myths and legends, fairy tales, fables, and folklore.

Kaatsheuvel is a village in the Dutch province of North Brabant, situated ... it is the largest village in and the capital of the municipality of **Loon op Zand**, which also consists ...

Query subject: *The Haunted Castle*
Query body: located_in_the_administrative_territorial_entity
Answer: Loon op Zand

Multi-Hop QA Requirements

UNC NLP

- Success on Multi-Hop Reasoning QA requires a model to:
  - Locate a reasoning chain of important/relevant documents from a large pool of documents
  - Consider evidence loosely distributed in retrieved documents to predict the answer

Multi-Hop QA Requirements

UNC NLP

- Success on Multi-Hop Reasoning QA requires a model to:
  - Locate a reasoning chain of important/relevant documents from a large pool of documents
  - Consider evidence loosely distributed in retrieved documents to predict the answer
- Is addressing these two problems enough?

Divergent Reasoning Chains

The *Polsterberg Pumphouse* (German : Polsterberger Hubhaus) is a pumping station above **the Dyke Ditch** in the **Upper Harz** in central Germany ...

Query subject: *Polsterberg Pumphouse*
Query body: located_in_the_administrative_territorial_entity

Divergent Reasoning Chains

[Welbl et al. 2018]

The *Polsterberg Pumphouse* (German : Polsterberger Hubhaus) is a pumping station above **the Dyke Ditch** in the **Upper Harz** in central Germany ...

The Dyke Ditch is the longest artificial ditch in the **Upper Harz** in central Germany.

Query subject: *Polsterberg Pumphouse*
Query body: located_in_the_administrative_territorial_entity

Divergent Reasoning Chains

[Welbl et al. 2018]

The *Polsterberg Pumphouse* (German : Polsterberger Hubhaus) is a pumping station above **the Dyke Ditch** in the **Upper Harz** in central Germany ...

The Dyke Ditch is the longest artificial ditch in the **Upper Harz** in central Germany.

The **Upper Harz** refers to ... the term Upper Harz covers the area of the seven historical mining towns (\"Bergst\u00e4dte\") - Clausthal, Zellerfeld, Andreasberg, Altenau, Lautenthal, Wildemann and Grund - in the present-day German federal state of **Lower Saxony**.

Query subject: *Polsterberg Pumphouse*
Query body: located_in_the_administrative_territorial_entity
Answer: Lower Saxony

Multi-Hop QA Requirements

UNC NLP

- Success on Multi-Hop Reasoning QA requires a model to:
  - Locate a reasoning chain of important/relevant documents from a large pool of documents
  - Consider evidence loosely distributed in all documents from a reasoning chain to predict the answer
  - Weigh and merge evidence from **MULTIPLE** reasoning chains to predict the answer

EPAr: Explore-Propose-Assemble reader

- Our EPAr model consists of three modules:
  - **Document Explorer (DE)**: Iteratively selects relevant documents and represents multiple reasoning chains in a tree structure

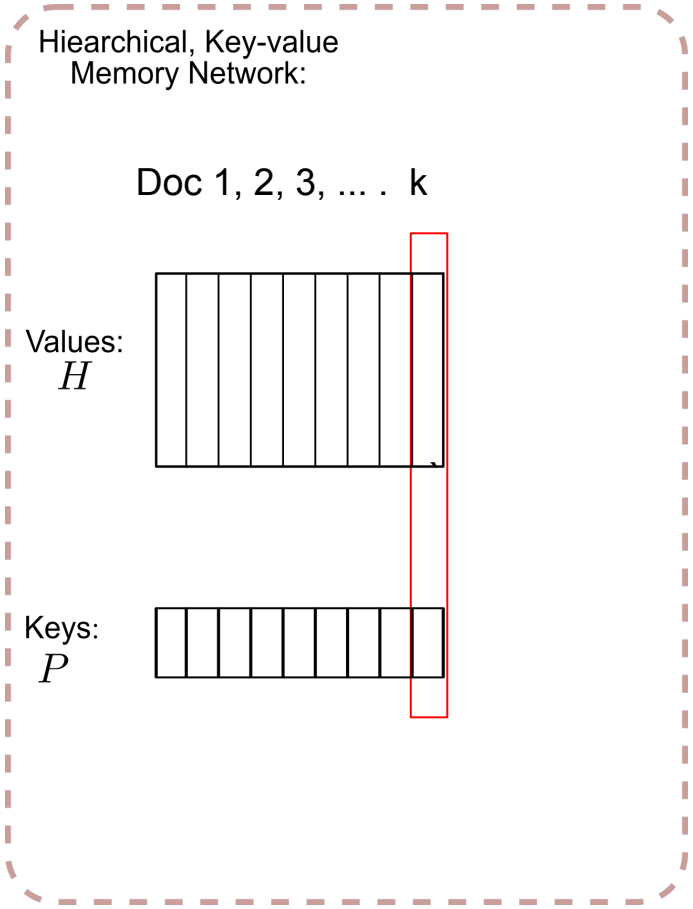
EPAr: Explore-Propose-Assemble reader

- Our EPAr model consists of three modules:
  - **Document Explorer (DE)**: Iteratively selects relevant documents and represents multiple reasoning chains in a tree structure
  - **Answer Proposer (AP)**: Proposes a candidate answer from every root-to-leaf chain in the reasoning tree

EPAr: Explore-Propose-Assemble reader

- Our EPAr model consists of three modules:
  - **Document Explorer (DE)**: Iteratively selects relevant documents and represents multiple reasoning chains in a tree structure
  - **Answer Proposer (AP)**: Proposes a candidate answer from every root-to-leaf chain in the reasoning tree
  - **Evidence Assembler (EA)**: Extracts key sentences from every reasoning chain and combines them to make a unified prediction

Document Explorer: Key-Value MemNet

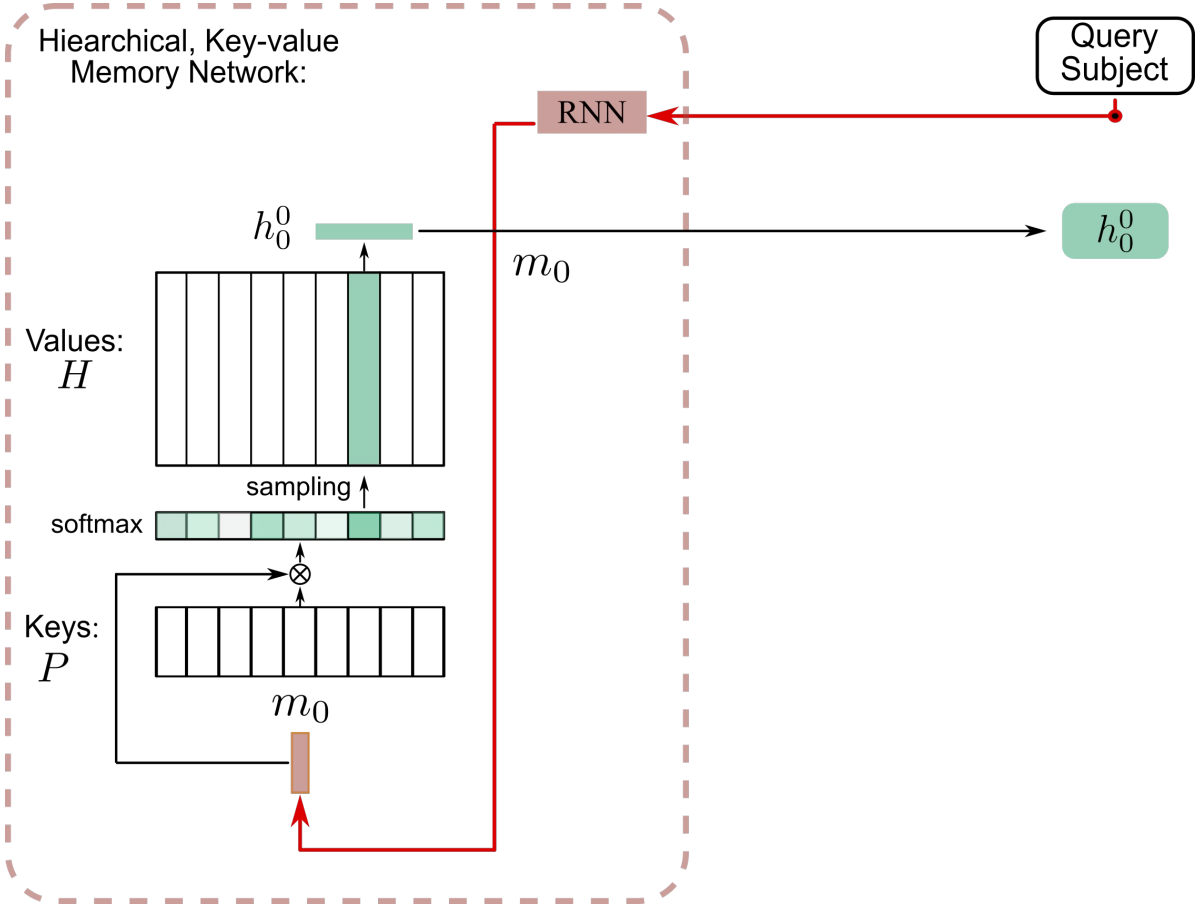


Document Explorer

Contextualized **word representations** for every document

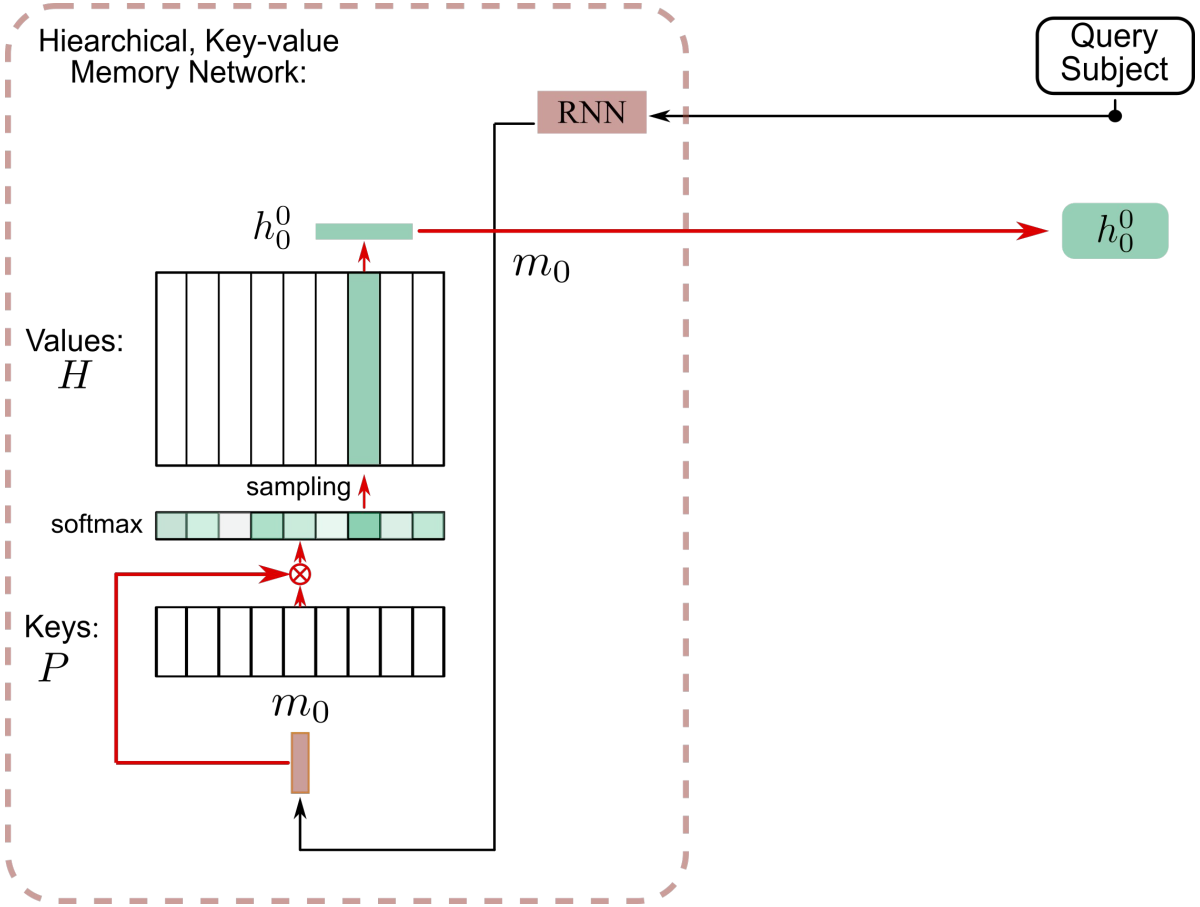
Document representation as a fixed-sized vector

Document Explorer: Read Unit



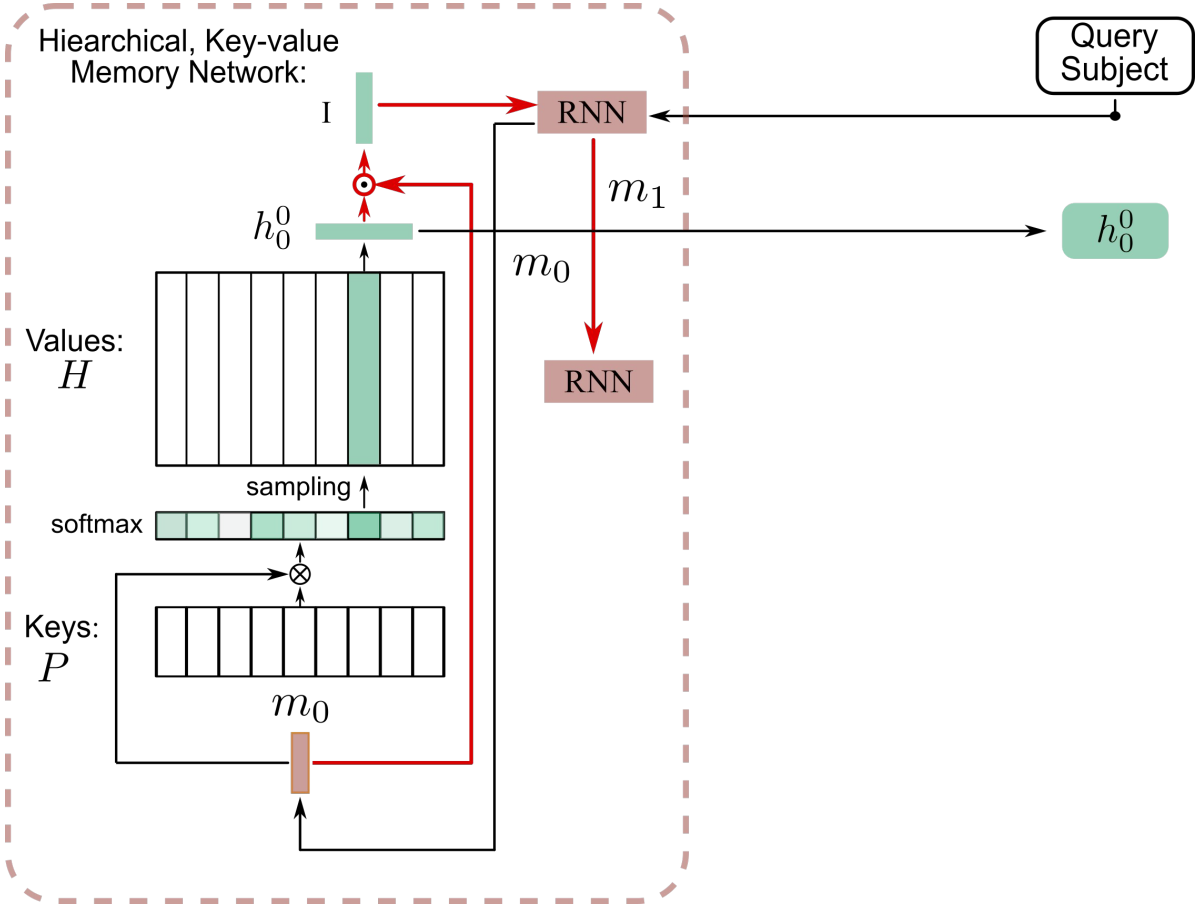
Document Explorer

Document Explorer: Read Unit



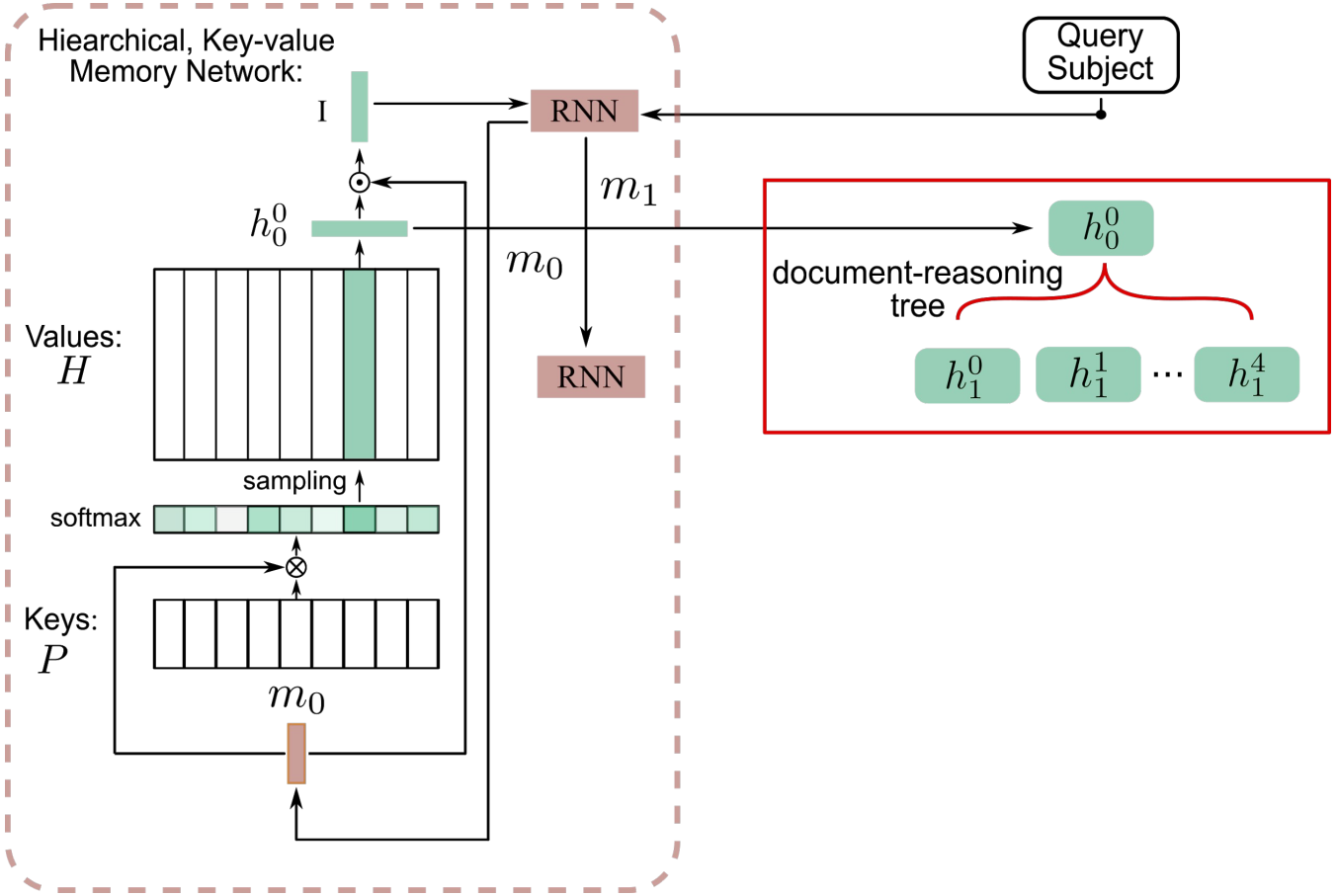
Document Explorer

Document Explorer: Write Unit



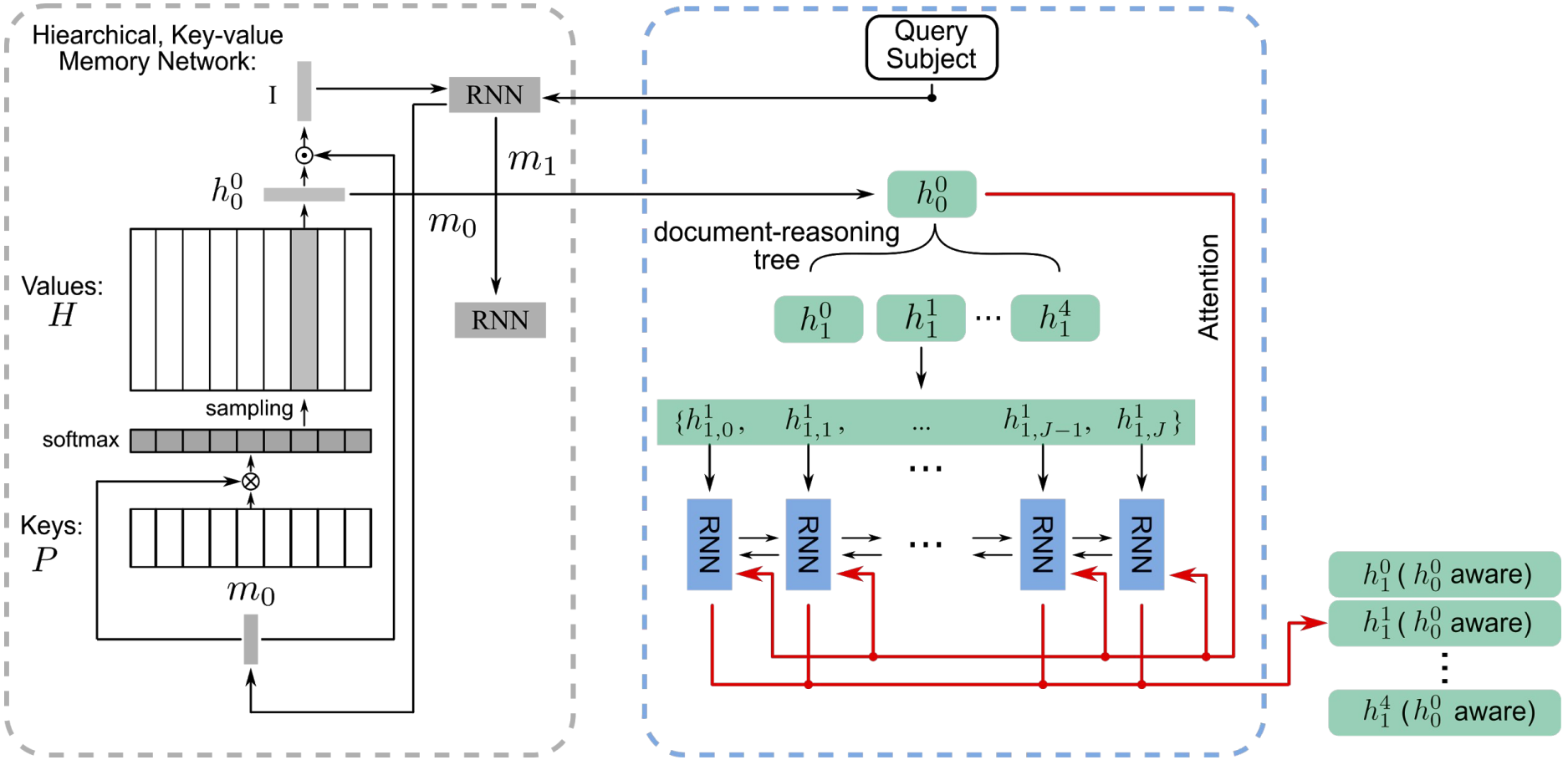
Document Explorer

Document Explorer: Reasoning Tree



Document Explorer

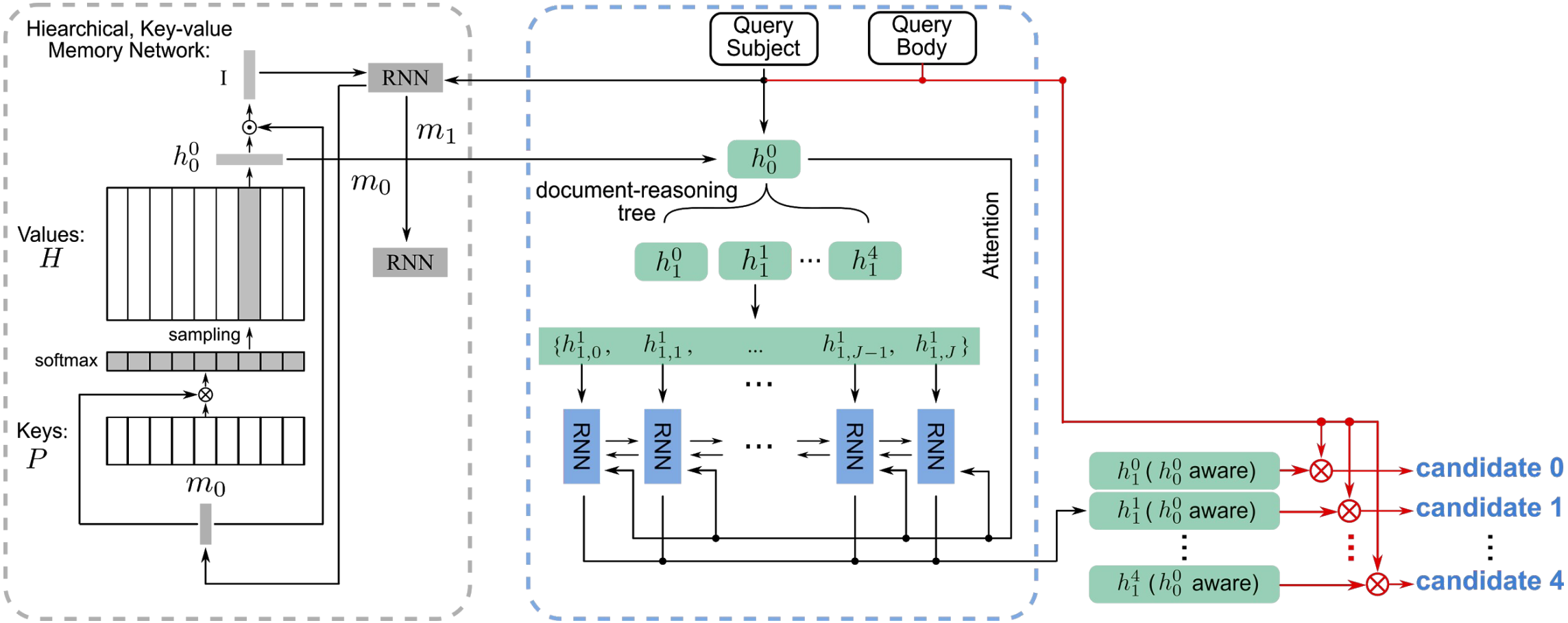
Answer Proposer



Document Explorer

Answer Proposer

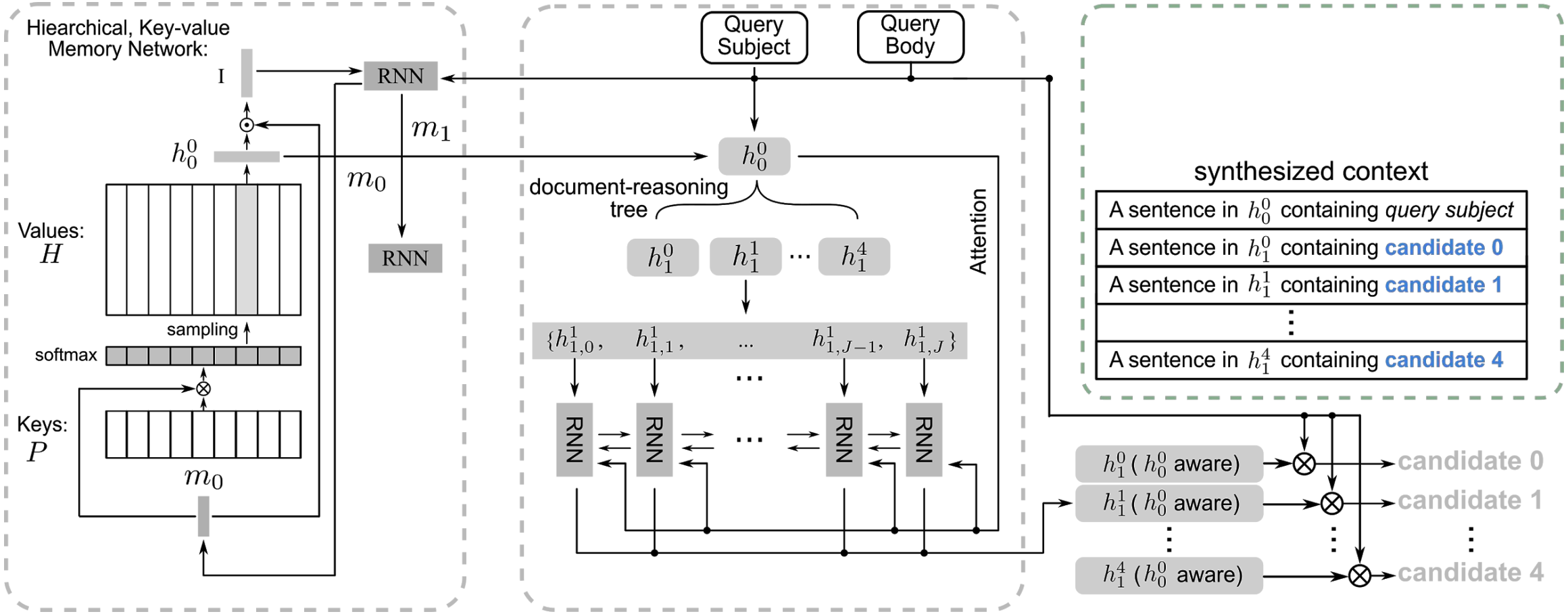
Answer Proposer



Document Explorer

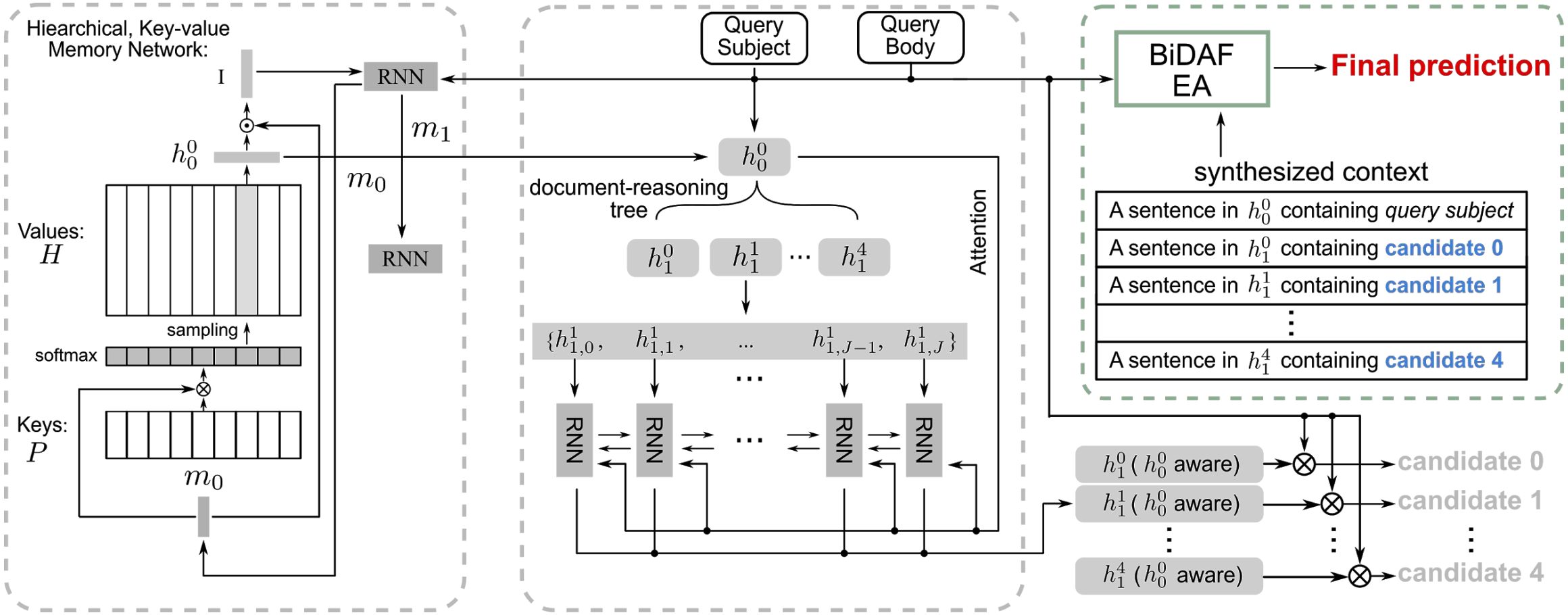
Answer Proposer

Evidence Assembler



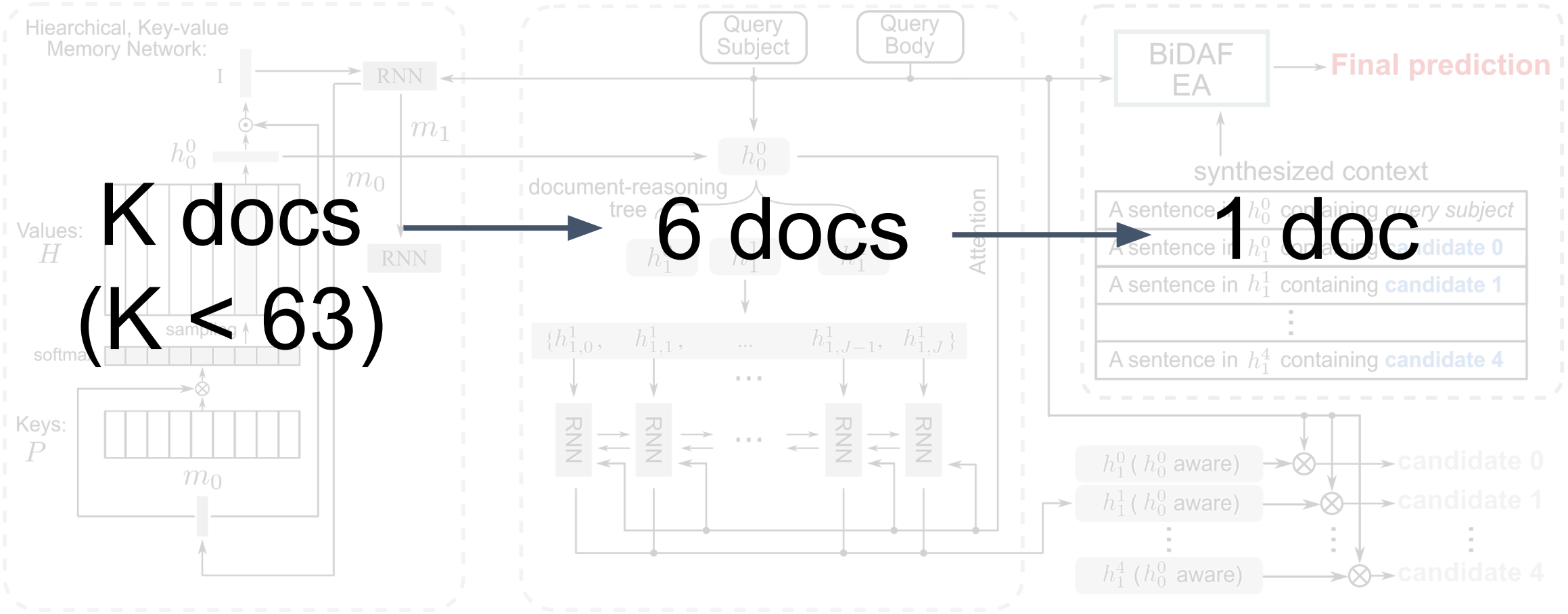
Document Explorer | Answer Proposer | Evidence Assembler

Evidence Assembler



Document Explorer | Answer Proposer | Evidence Assembler

EPAr



K docs (K < 63) → 6 docs → 1 doc

Document Explorer | Answer Proposer | Evidence Assembler

Joint Optimization

- Supervise the first hop of Document Explorer using document having shortest TF-IDF distance w.r.t. the query subject

$$\mathcal{L}_{DE,1}$$

Joint Optimization

UNC NLP

- Supervise the first hop of Document Explorer using document having shortest TF-IDF distance w.r.t. the query subject
- Supervise the last hop of Document Explorer using documents which contain at least one mention of the answer

$$\mathcal{L}_{DE,1} + \mathcal{L}_{DE,2}$$

Joint Optimization

UNC NLP

- Supervise the first hop of Document Explorer using document having shortest TF-IDF distance w.r.t. the query subject
- Supervise the last hop of Document Explorer using documents which contain at least one mention of the answer
- Use cross-entropy loss from the answer selection process of the Answer Proposer and the Evidence Assembler

$$\mathcal{L}_{DE,1} + \mathcal{L}_{DE,2} + \mathcal{L}_{AP} + \mathcal{L}_{EA}$$

Datasets

UNC NLP

[Welbl et al. 2018]

- WikiHop:
  - Based on Wikipedia articles
  - Size is 51k instances : 44k train, 5k dev, 2.5k test

Datasets

UNC NLP

[Welbl et al. 2018]

- WikiHop:
  - Based on Wikipedia articles
  - Size is 51k instances : 44k train, 5k dev, 2.5k test
- MedHop:
  - Based on the domain of molecular biology
  - Much smaller dataset : 1.6k train, 342 dev, 546 test

Results - WikiHop

	Dev	Test
BiDAF Welbl et al., 2017*	-	42.9
Coref-GRU (Dhingra et al., 2018)	56.0	59.3
WEAVER (Raison et al., 2018)	64.1	65.3
MHQA-GRN (Song et al., 2018)	62.8	65.4
Entity-GCN (De Cao et al., 2018)	64.8	67.6
BAG (Cao et al., 2019)	66.5	69.0
CFC (Zhong et al., 2019)	66.4	70.6
EPAr (Ours)	**67.2**	69.1

Results - MedHop

	Test (Masked)	Test
FastQA (Weissenborn et al., 2017)	23.1	31.3
BiDAF (Seo et al., 2017)	33.7	47.8
CoAttention	-	58.1
Most Frequent Candidate	10.4	58.4
EPAr (Ours)	**41.6**	**60.3**

- A simple frequency statistic is able to achieve strong results in the unmasked setting

Results - MedHop

	Test (Masked)	Test
FastQA (Weissenborn et al., 2017)	23.1	31.3
BiDAF (Seo et al., 2017)	33.7	47.8
CoAttention	-	58.1
Most Frequent Candidate	10.4	58.4
EPAr (Ours)	**41.6**	**60.3**

- Masking candidate expressions removes such answer frequency cues

Analysis & Ablations

UNC NLP

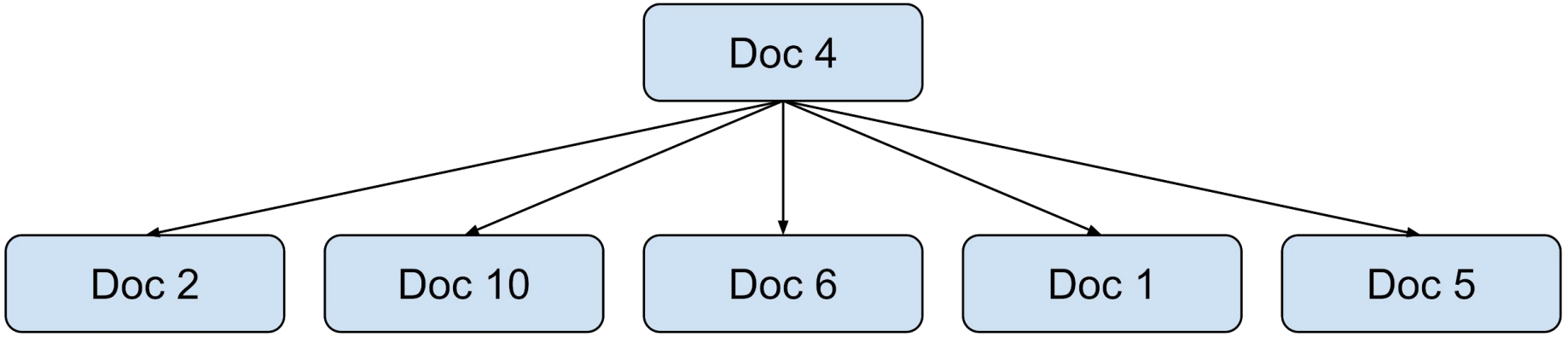
How to evaluate the quality of the reasoning tree retrieved by the Document Explorer?

Human Evaluation - Metric

Golden Reasoning Chain labeled by human annotators:

- Doc 4 → Doc 6

Reasoning Tree constructed by Document Explorer:

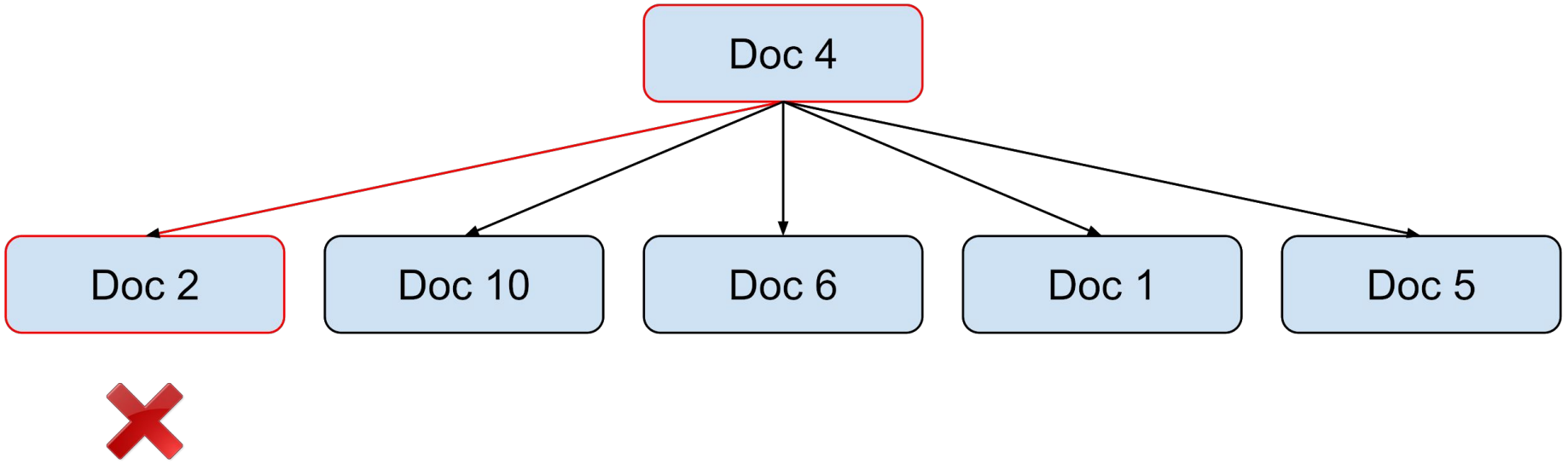


Human Evaluation - Recall@1

Golden Reasoning Chain labeled by human annotators:

- Doc 4 → Doc 6

Reasoning Tree constructed by Document Explorer:

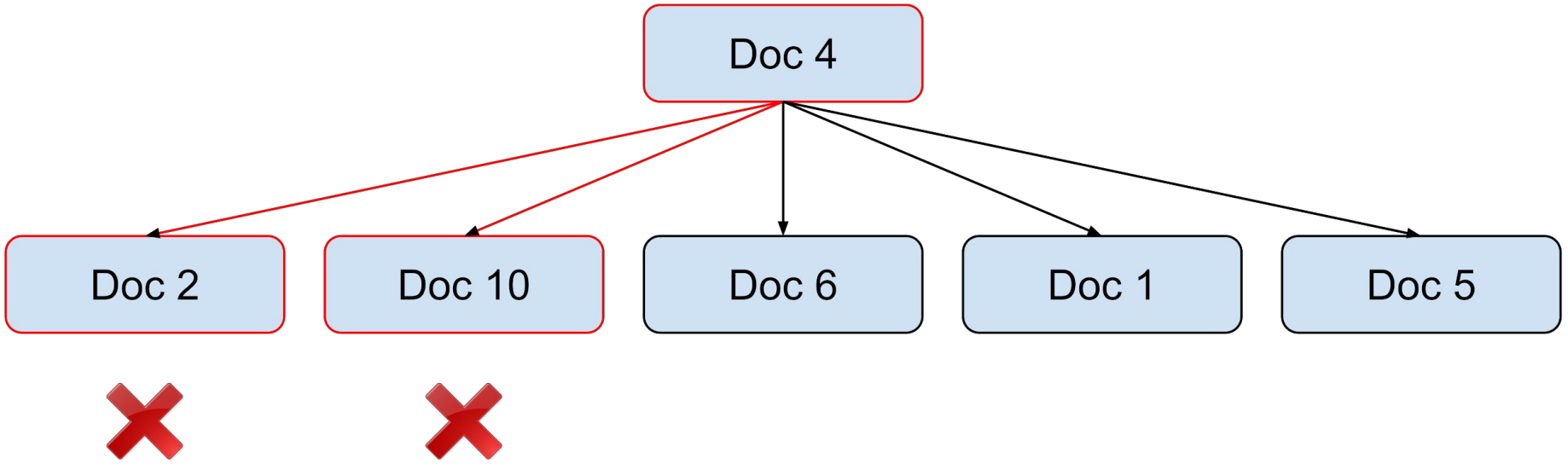


Human Evaluation - Recall@2

Golden Reasoning Chain labeled by human annotators:

- Doc 4 → Doc 6

Reasoning Tree constructed by Document Explorer:

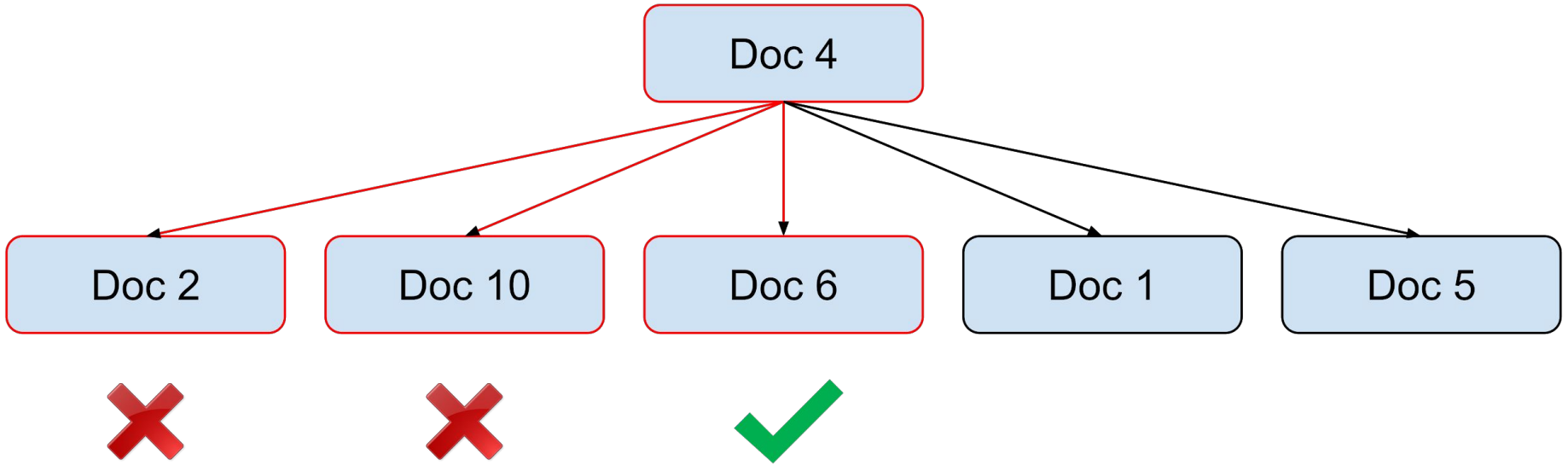


Human Evaluation - Recall@3

Golden Reasoning Chain labeled by human annotators:

- Doc 4 → Doc 6

Reasoning Tree constructed by Document Explorer:



Human Evaluation

UNC NLP

- Recall-k score is the % of examples where one of the human-annotated reasoning chains is recovered in the top-k root-to-leaf paths in the reasoning tree

	R@1	R@2	R@3	R@4	R@5
Random	11.2	17.3	27.6	40.8	50.0
1-hop TFIDF	32.7	48.0	56.1	63.3	70.4
2-hop TFIDF	42.9	56.1	70.4	78.6	82.7
DE	38.8	50.0	65.3	73.5	83.7
TFIDF+DE	**44.9**	**64.3**	**77.6**	**82.7**	**90.8**

- 2-hop TF-IDF performs much better than simple 1-hop TF-IDF retrieval

Human Evaluation

UNC NLP

- Recall-k score is the % of examples where one of the human-annotated reasoning chains is recovered in the top-k root-to-leaf paths in the reasoning tree

	R@1	R@2	R@3	R@4	R@5
Random	11.2	17.3	27.6	40.8	50.0
1-hop TFIDF	32.7	48.0	56.1	63.3	70.4
2-hop TFIDF	42.9	56.1	70.4	78.6	82.7
DE	38.8	50.0	65.3	73.5	83.7
TFIDF+DE	**44.9**	**64.3**	**77.6**	**82.7**	**90.8**

- DE without any TF-IDF retrieval pre-processing performs worse than 2-hop TF-IDF

Human Evaluation

UNC NLP

- Recall-k score is the % of examples where one of the human-annotated reasoning chains is recovered in the top-k root-to-leaf paths in the reasoning tree

	R@1	R@2	R@3	R@4	R@5
Random	11.2	17.3	27.6	40.8	50.0
1-hop TFIDF	32.7	48.0	56.1	63.3	70.4
2-hop TFIDF	42.9	56.1	70.4	78.6	82.7
DE	38.8	50.0	65.3	73.5	83.7
TFIDF+DE	**44.9**	**64.3**	**77.6**	**82.7**	**90.8**

- Combination of TF-IDF retrieval and DE performs better than each one of them alone

Answer Span Test

UNC NLP

- Recall-k score is the percentage of examples where the ground truth answer is present in the top-k root-to-leaf paths in the reasoning tree

	R@1	R@2	R@3	R@4	R@5
Random	39.9	51.4	60.2	67.8	73.5
1-hop TFIDF	38.4	48.5	58.6	67.4	73.7
2-hop TFIDF	38.4	58.7	70.2	77.2	81.6
DE	**52.5**	**70.2**	**80.3**	**85.8**	**89.0**
TFIDF+DE	52.2	69.0	77.8	82.2	85.2

- DE alone performs best in this test

Ablations: Answer Proposer

UNC NLP

- ‘Follows’ is the subset of dev set for which the answer can be inferred from the given documents according to human annotation
- ‘Single’ or ‘Multiple’ indicates whether the complete reasoning chain comprises of single or multiple documents

	full	follows + multiple	follows + single
Full-doc	63.1	68.4	69.0
Lead-1	63.6	68.7	70.2
AP w.o. attn	63.3	68.3	69.6
AP	**64.7**	**69.4**	**70.6**

Ablations: Evidence Assembler

UNC NLP

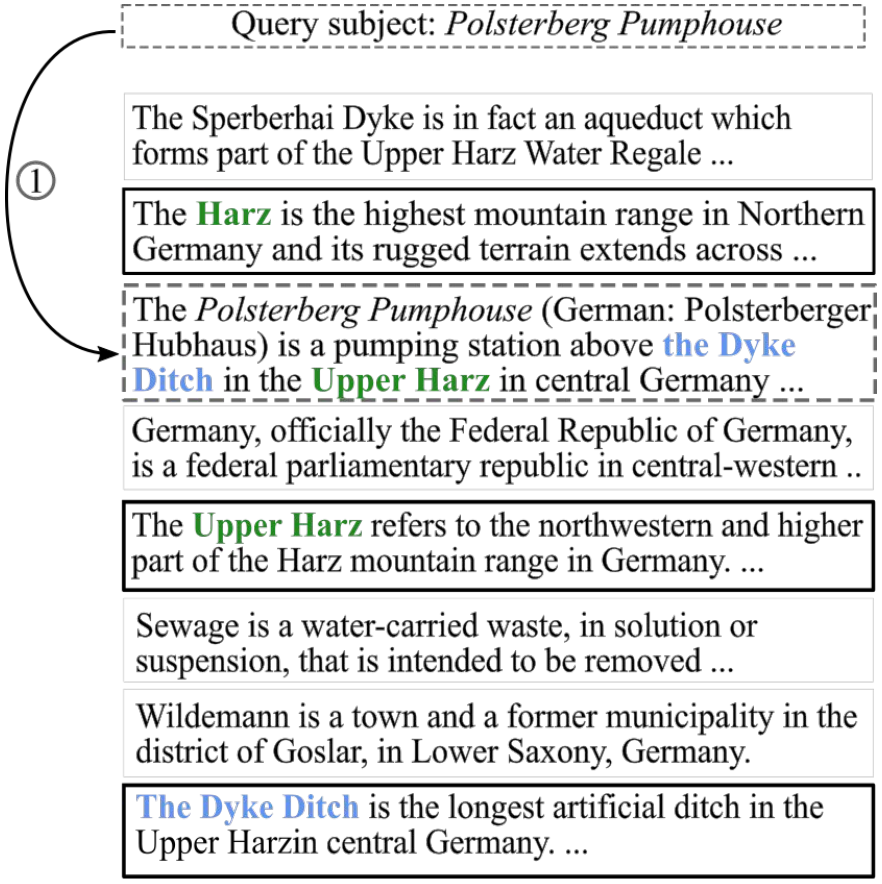
- ‘Follows’ is the subset of dev set for which the answer can be inferred from the given documents according to human annotation
- ‘Single’ or ‘Multiple’ indicates whether the complete reasoning chain comprises of single or multiple documents

	full	follows + multiple	follows + single
Single-chain	59.9	64.3	63.8
Avg-vote	54.6	56.3	55.6
Max-vote	51.5	53.9	53.3
w. Reranker	60.6	65.1	65.5
w. Assembler	**64.7**	**69.4**	**70.6**

EPAr Demo

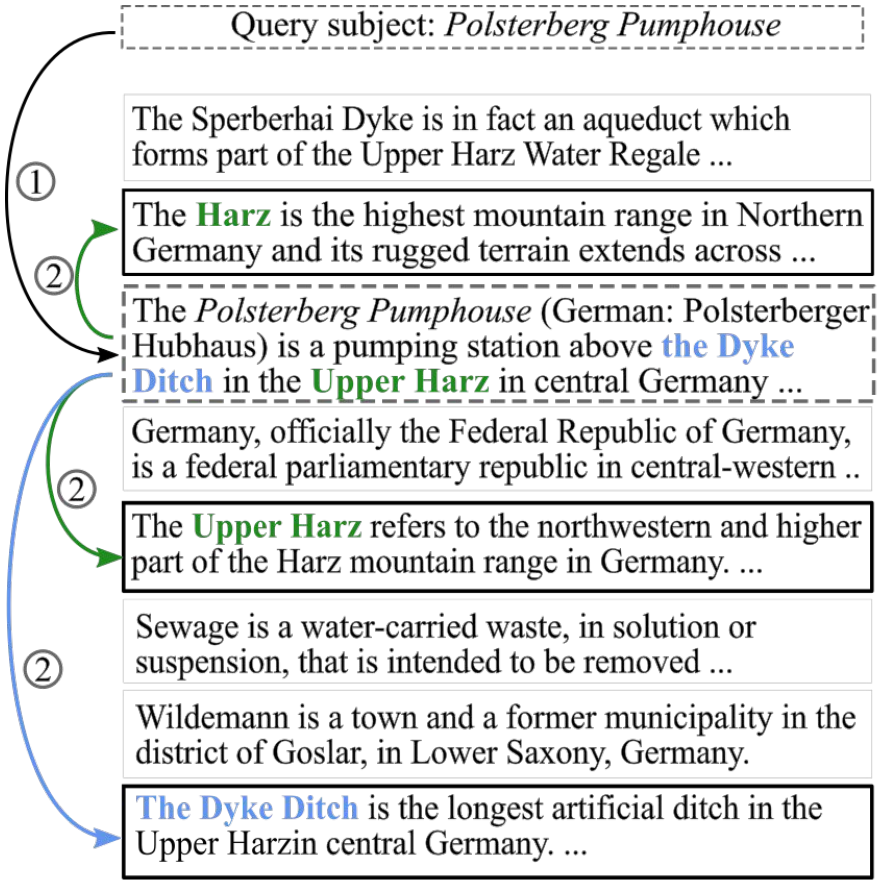
Document Explorer

EPAr Demo



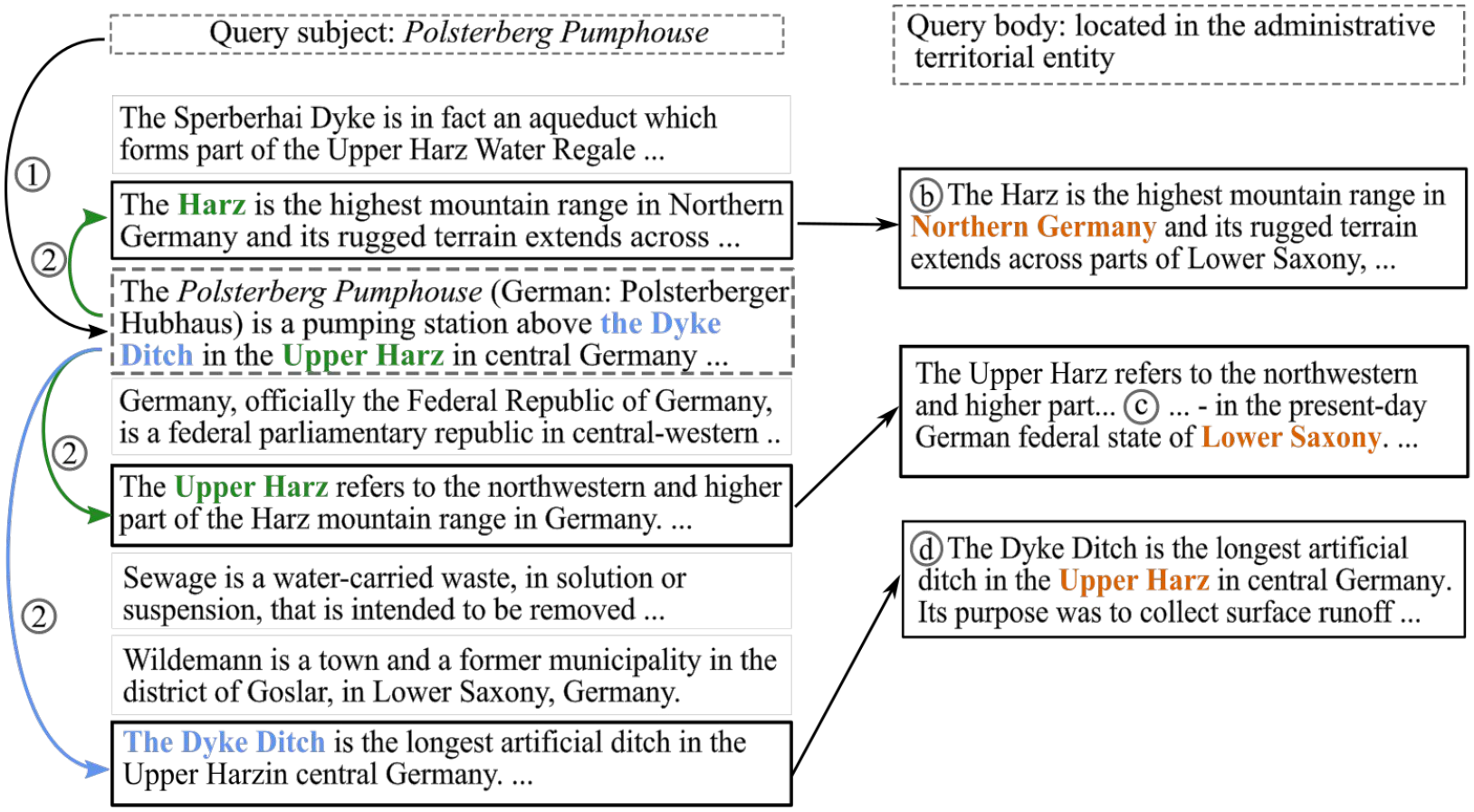
Document Explorer

EPAr Demo

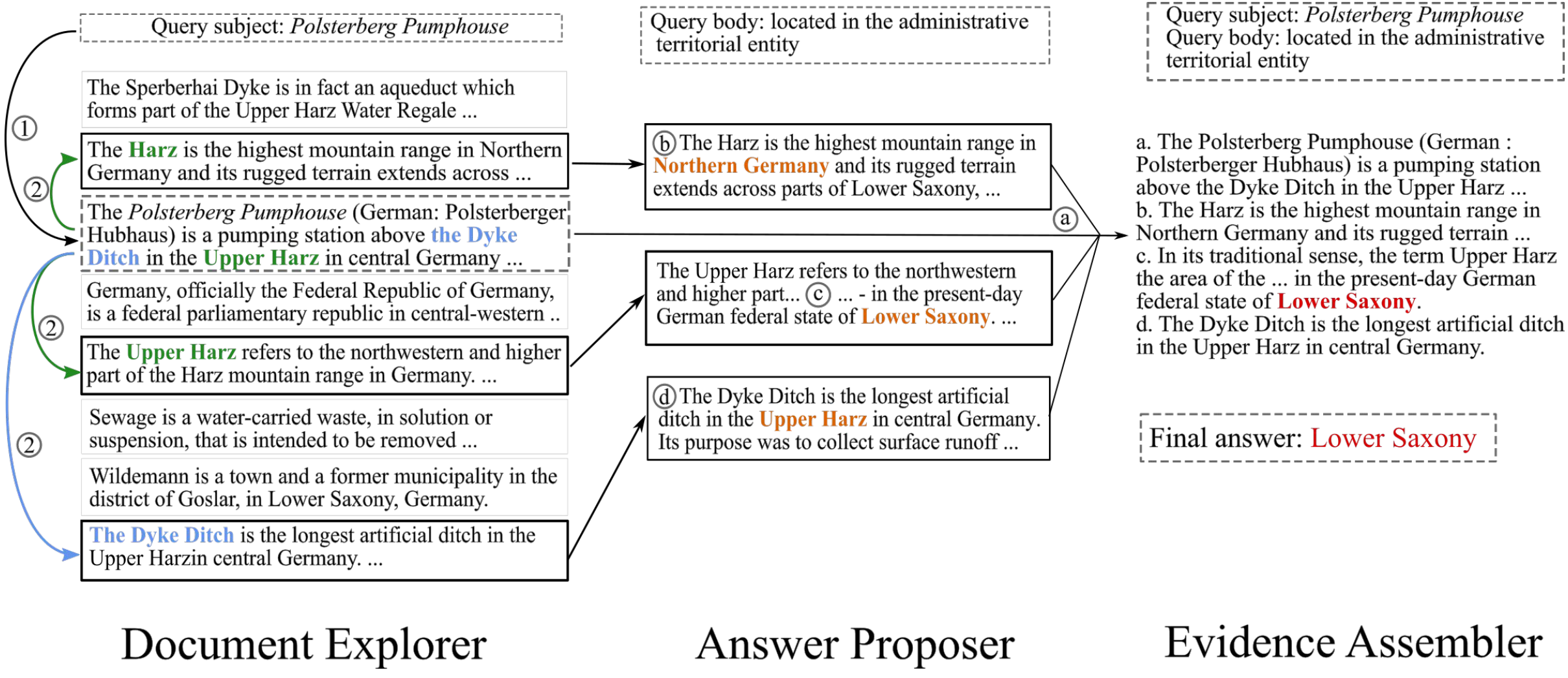


Document Explorer

EPAr Demo


Query subject: *Polsterberg Pumphouse*
Query body: located in the administrative territorial entity
The Sperberhai Dyke is in fact an aqueduct which forms part of the Upper Harz Water Regale ...
The **Harz** is the highest mountain range in Northern Germany and its rugged terrain extends across ...
The *Polsterberg Pumphouse* (German: Polsterberger Hubhaus) is a pumping station above **the Dyke Ditch** in the **Upper Harz** in central Germany ...
Germany, officially the Federal Republic of Germany, is a federal parliamentary republic in central-western ..
The **Upper Harz** refers to the northwestern and higher part of the Harz mountain range in Germany. ...
Sewage is a water-carried waste, in solution or suspension, that is intended to be removed ...
Wildemann is a town and a former municipality in the district of Goslar, in Lower Saxony, Germany.
The Dyke Ditch is the longest artificial ditch in the Upper Harzin central Germany. ...
① ② ② ②
ⓑ The Harz is the highest mountain range in **Northern Germany** and its rugged terrain extends across parts of Lower Saxony, ...
The Upper Harz refers to the northwestern and higher part... ⓒ ... - in the present-day German federal state of **Lower Saxony**. ...
ⓓ The Dyke Ditch is the longest artificial ditch in the **Upper Harz** in central Germany. Its purpose was to collect surface runoff ...
Document Explorer
Answer Proposer

EPAr Demo


Query subject: *Polsterberg Pumphouse*
The Sperberhai Dyke is in fact an aqueduct which forms part of the Upper Harz Water Regale ...
The **Harz** is the highest mountain range in Northern Germany and its rugged terrain extends across ...
The *Polsterberg Pumphouse* (German: Polsterberger Hubhaus) is a pumping station above **the Dyke Ditch** in the **Upper Harz** in central Germany ...
Germany, officially the Federal Republic of Germany, is a federal parliamentary republic in central-western ..
The **Upper Harz** refers to the northwestern and higher part of the Harz mountain range in Germany. ...
Sewage is a water-carried waste, in solution or suspension, that is intended to be removed ...
Wildemann is a town and a former municipality in the district of Goslar, in Lower Saxony, Germany.
The Dyke Ditch is the longest artificial ditch in the Upper Harzin central Germany. ...
Query body: located in the administrative territorial entity
ⓑ The Harz is the highest mountain range in **Northern Germany** and its rugged terrain extends across parts of Lower Saxony, ...
The Upper Harz refers to the northwestern and higher part... ⓒ ... - in the present-day German federal state of **Lower Saxony**. ...
ⓓ The Dyke Ditch is the longest artificial ditch in the **Upper Harz** in central Germany. Its purpose was to collect surface runoff ...
Query subject: *Polsterberg Pumphouse*
Query body: located in the administrative territorial entity
a. The Polsterberg Pumphouse (German : Polsterberger Hubhaus) is a pumping station above the Dyke Ditch in the Upper Harz ...
b. The Harz is the highest mountain range in Northern Germany and its rugged terrain ...
c. In its traditional sense, the term Upper Harz the area of the ... in the present-day German federal state of **Lower Saxony**.
d. The Dyke Ditch is the longest artificial ditch in the Upper Harz in central Germany.
Final answer: Lower Saxony
Document Explorer
Answer Proposer
Evidence Assembler

Related Works

UNC NLP

- WikiHop
  - Zhong et al., 2019 used hierarchies of co-attention and self-attention to combine evidence from scattered documents
  - Lot of work using graph networks to model multi-hop reasoning : Song et al. 2018, De Cao et al. 2018, Cao et al. 2019
  - Our model learns the relation between entities implicitly
- Architecture
  - Choi et al., 2017 and Wang et al., 2017 use similar two module systems for retrieval and answer span extraction
  - Restricted for single-hop QA
- More details in the paper

Conclusions

- In this work, we…
  - Proposed an interpretable model EPAr for multi-hop QA
  - Achieved strong results on WikiHop and MedHop datasets
  - Presented several analyses, ablations and human evaluation of our model's reasoning capabilities

Code available at https://github.com/jiangycTarheel/EPAr

Thank you for listening!
Questions?

Acknowledgement: DARPA (YFA17-D17AP00022), Google, Bloomberg, NVidia, Salesforce, Amazon AWS