Automatic Article Generation via Multi-Document Summarization

Connor Lennox
connor.lennox@unh.edu
Department of Computer Science, UNH

Advisor: Laura Dietz

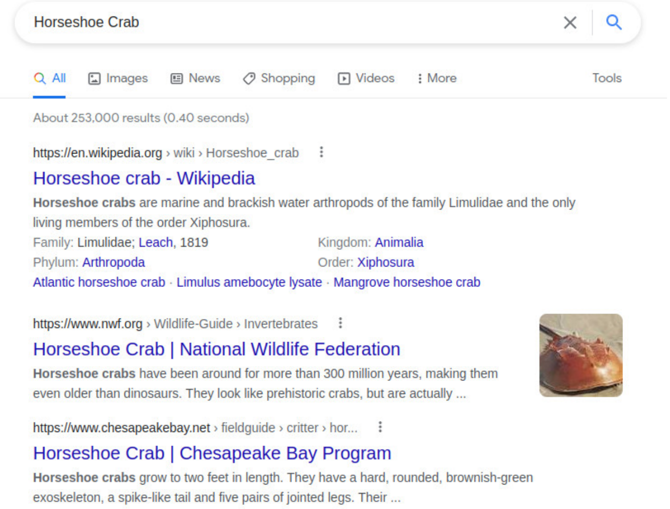
Graduate Research Conference, 2022

Information Retrieval

- When a user (you) enters a query (web search), what documents (pages) should we display to provide information?

Information Retrieval

- When a user (you) enters a query (web search), what documents (pages) should we display to provide information?



About 253,000 results (0.40 seconds)

https://en.wikipedia.org › wiki › Horseshoe_crab

Horseshoe crab - Wikipedia

Horseshoe crabs are marine and brackish water arthropods of the family Limulidae and the only living members of the order Xiphosura.

Family: Limulidae; Leach, 1819 Kingdom: Animalia

Phylum: Arthropoda Order: Xiphosura

Atlantic horseshoe crab · Limulus amebocyte lysate · Mangrove horseshoe crab

https://www.nwf.org › Wildlife-Guide › Invertebrates

Horseshoe Crab | National Wildlife Federation

Horseshoe crabs have been around for more than 300 million years, making them even older than dinosaurs. They look like prehistoric crabs, but are actually ...

https://www.chesapeakebay.net › fieldguide › critter › hor...

Horseshoe Crab | Chesapeake Bay Program

Horseshoe crabs grow to two feet in length. They have a hard, rounded, brownish-green exoskeleton, a spike-like tail and five pairs of jointed legs. Their ...

Task

- Users don't want to read an entire ranking of documents to find the information they're looking for!
- Goal: Develop a system that can effectively merge all documents of a ranking into a single, informative article which can be presented to the user
- **Given query q and ranking of documents $d_1, d_2, ..., d_n$, generate an article a that is similar to a reference article.**

Challenges

- **Text Quantity**: Document rankings contain a large amount of text, too much to process simultaneously
- **Topical Breadth**: For broad queries, many subtopics will be addressed in documents
- **Information Ordering**: It is important that information is presented to the user in a logical, “coherent” order

Generation Pipeline

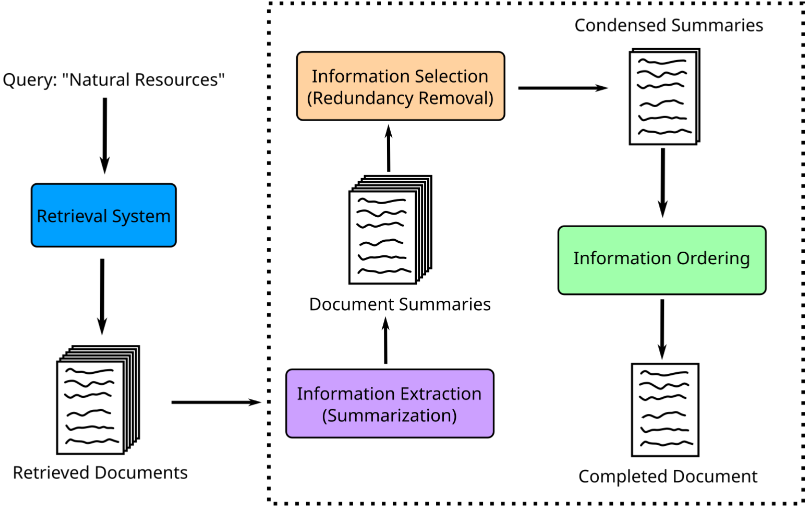
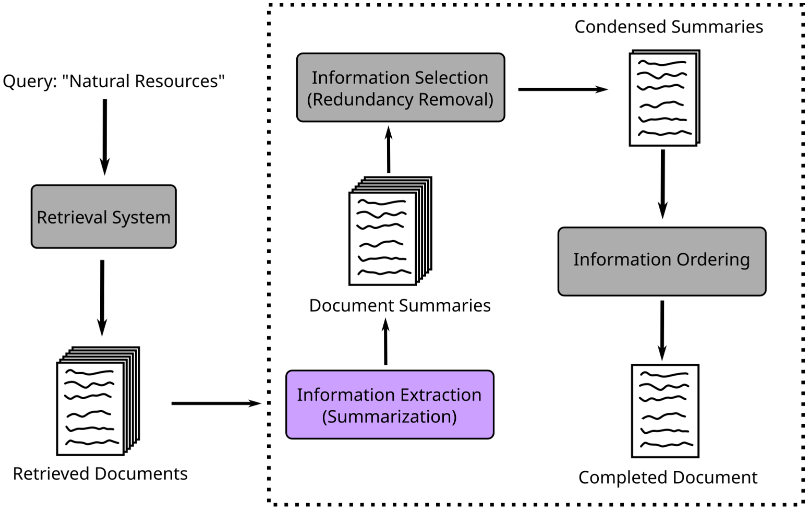


Figure: The full article generation pipeline. Elements within the dashed box represent those we will address in this work.

Information Extraction



Query: "Natural Resources"

Retrieval System

Retrieved Documents

Information Extraction (Summarization)

Document Summaries

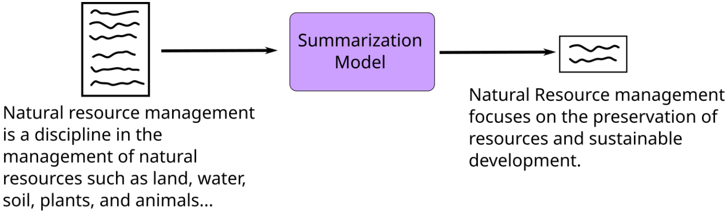
Information Selection (Redundancy Removal)

Condensed Summaries

Information Ordering

Completed Document

Information Extraction



Summarization Model

Natural resource management is a discipline in the management of natural resources such as land, water, soil, plants, and animals...

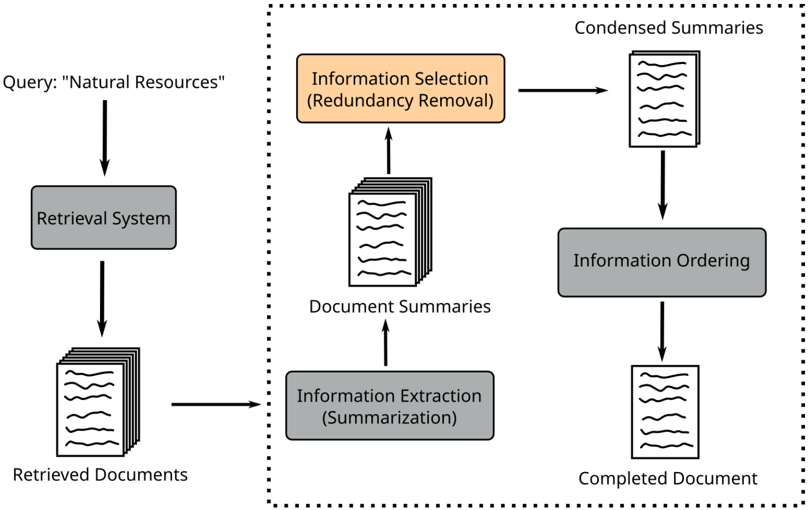
Natural Resource management focuses on the preservation of resources and sustainable development.

Figure: From an input document, information can be extracted through the use of a single-document summarization model. Here, some facts about the management of Natural Resources are found in source text to produce a shorter summary.

Information Extraction

- Each dinput document will have some information we want to extract and add to our final article
- Modern summarization models achieve great performance on *single-document summarization*, leverage this to extract core information of each input document
- By applying single-document summarization to all input documents, we are left with a collection of facts that we want to include in our final article

Information Selection



Information Selection

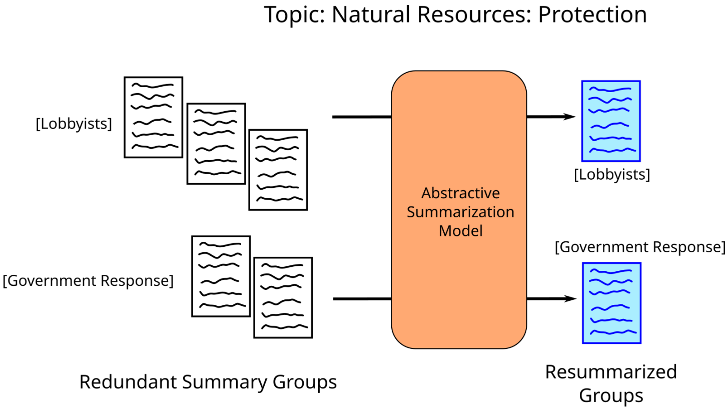


Figure: By grouping together redundant information and resummarizing, a more concise output article can be constructed

Information Selection

- By clustering information, we can identify redundant facts and combine them into a single summary
- Eliminating redundancy allows the final output to be more concise, without risk of completely removing critical information from the text
- Redundant groups are merged by using the same summarization model that performs information extraction

Discussion

- Our method of article generation avoids common pitfalls by splitting the input into smaller subproblems
- Components are modular and interchangeable: the system is agnostic to the exact summarization model used, etc.
- A side effect of our method is the ability to derive provenance of every statement in the output document

Evaluation

- Comparisons are made between a heuristic baseline method as well as two methods from recent literature
  - **Baseline** No summarization, input is output
  - **Hierarchical Transformer** (Liu et al., 2019)
  - **LoBART** (Pasunuru et al., 2021)

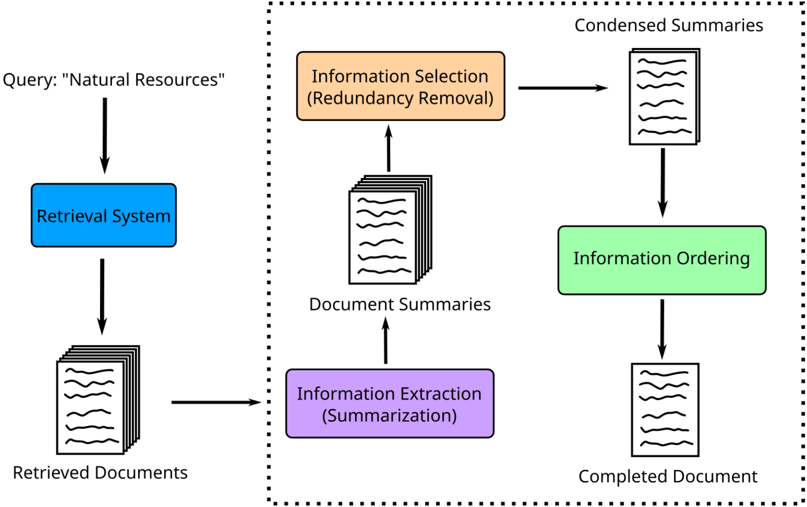
Results

Model	ROUGE-1	ROUGE-2	Manual Evaluation
Heuristic Baseline	0.165 ± 0.004	0.027 ± 0.002	1.44 ± 0.17
Hierarchical Transformer	0.074 ± 0.005	0.013 ± 0.001	-
LoBART	0.211 ± 0.005	0.052 ± 0.002	1.80 ± 0.14
Multistage (Ours)	0.172 ± 0.003	0.028 ± 0.001	1.54 ± 0.16

Table: Results of evaluation with both automatic and manual metrics. Manual scores are on a range of 0-3. For all metrics, higher is better.

- **ROUGE**: Evaluates *linguistic* overlap
- **BERTScore**: Evaluates *semantic* overlap
- **Manual Evaluation**: Annotators asked to evaluate presence of “important information”

Conclusion


Query: "Natural Resources"
Retrieval System
Retrieved Documents
Information Extraction (Summarization)
Document Summaries
Information Selection (Redundancy Removal)
Condensed Summaries
Information Ordering
Completed Document

Conclusion

- We have proposed a system for summarizing a set of documents obtained via an Information Retrieval system.
- Our method, Multi-Stage Cluster Summarization, avoids possible challenges by breaking down the summarization problem into subtasks.
- The system is capable of producing informative yet concise documents that encompass the full breadth of information about a topic.