Which entities are relevant to the story?

Pooja H. Oza and Laura Dietz

University of New Hampshire

Overview

- Information is spread across various articles
- Construct a summarized story
- Capture complex relations between articles
- We focus on retrieving relevant entities for a summarized story
- Resembles entity retrieval task but different

Overview

President of United States

Joe Biden

Radiocarbon Dating

In 1939, **Martin Kamen** and **Samuel Ruben** of the **Radiation Laboratory at Berkeley** began experiments to determine if any of the elements common in organic matter had isotopes with half-lives long enough to be of value in biomedical research. They ...

Task

- Input - two types of input
  - A short text such as “2020 US presidential elections nominations”, “Radiocarbon dating background history” which we call as a **story request**
  - A list of story requests such as [“Radiocarbon dating samples”, “Radiocarbon dating background history”], which we define as **multi-story query request**

- Output - a list of relevant entities

- We assume access to a text corpus and a general knowledge base such as Wikipedia

Motivation

- Previous work based on keyword matching or entity attributes - entity names, entity types, graph walks
- Previous work focused on high-precision while our work focuses on retrieving entities of varying degree
- Our work uses text based signals to retrieve relevant entities
- We hypothesize that the text surrounding the entity mentions provide a strong indicator about entity relevance

In 1939, **Martin Kamen** and **Samuel Ruben** of the **Radiation Laboratory at Berkeley** began experiments to determine if any of the elements common in organic matter had isotopes with half-lives long enough to be of value in biomedical research. They synthesized ^{14}C using the laboratory's cyclotron accelerator and soon discovered that the atom's **half-life** was far longer than had been previously thought. This was followed by a prediction by **Serge A. Korff**, then employed at the **Franklin Institute** in **Philadelphia**, that the interaction of **thermal neutrons** with ^{14}N in the upper atmosphere would create ^{14}C. It had previously been thought that ^{14}C would be more likely to be created by **deuterons** interacting with ^{13}C. At some time during World War II, **Willard Libby**, who was then at Berkeley, learned of Korff's research and conceived the idea that it might be possible to use radiocarbon for dating.

Motivation

- The common approach retrieves entities by keyword matching in knowledge base (**KG-Entity**)

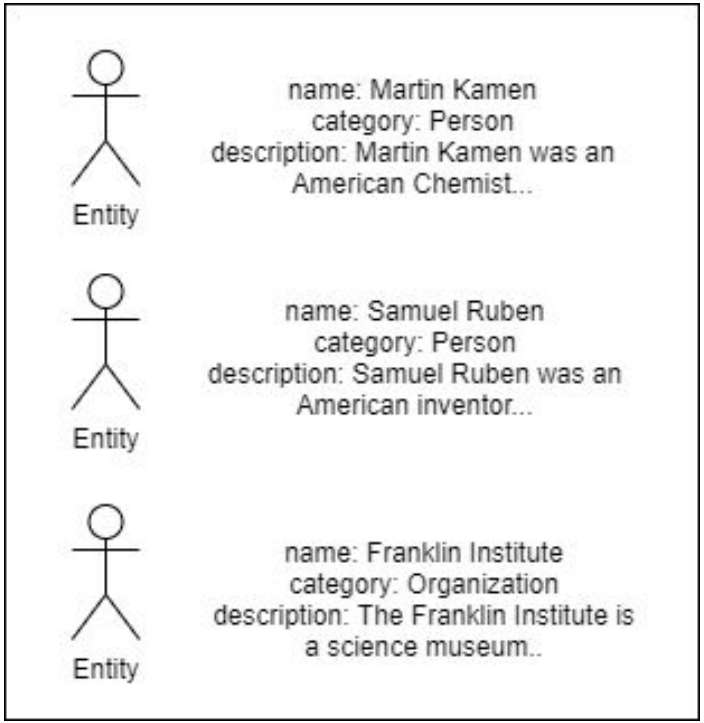
- Our approach is based on pseudo-relevance feedback on entity links in relevant text
  - Retrieve text passages relevant to the query
  - Use occurrence of entities or co-occurrence of two entities

- We argue that the entities that co-occur in relevant text passages, likely have a relevant connection

Claim 1

- Retrieving entities through relevant passage is more effective than knowledge base attributes

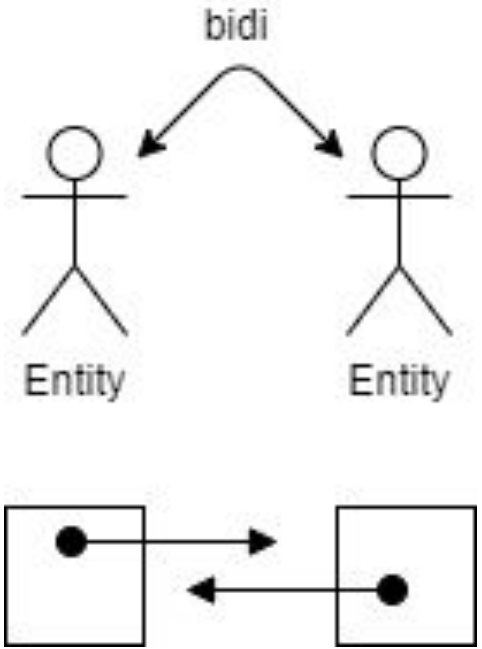
In 1939, **Martin Kamen** and **Samuel Ruben** of the **Radiation Laboratory at Berkeley** began experiments to determine if any of the elements common in organic matter had isotopes with half-lives long enough to be of value in biomedical research. They synthesized ^{14}C using the laboratory's cyclotron accelerator and soon discovered that the atom's **half-life** was far longer than had been previously thought. This was followed by a prediction by **Serge A. Korff**, then employed at the **Franklin Institute** in **Philadelphia**, that the interaction of **thermal neutrons** with ^{14}N in the upper atmosphere would create ^{14}C. It had previously been thought that ^{14}C would be more likely to be created by **deuterons** interacting with ^{13}C. At some time during World War II, **Willard Libby**, who was then at Berkeley, learned of Korff's research and conceived the idea that it might be possible to use radiocarbon for dating.



Claim 2

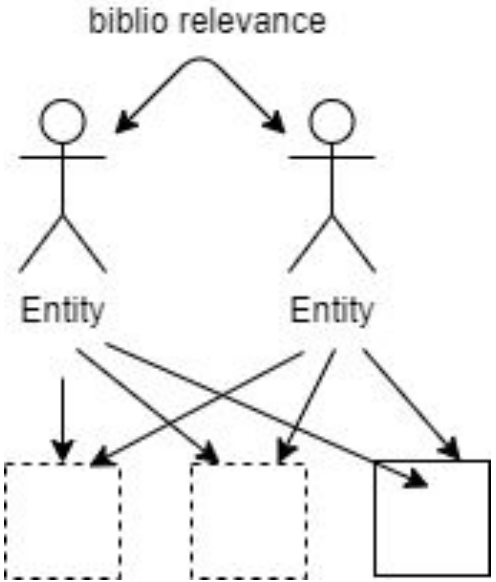
- Relevance through relevant passages is more effective than knowledge base links

In 1939, **Martin Kamen** and **Samuel Ruben** of the **Radiation Laboratory at Berkeley** began experiments to determine if any of the elements common in organic matter had isotopes with half-lives long enough to be of value in biomedical research. They synthesized ^{14}C using the laboratory's cyclotron accelerator and soon discovered that the atom's **half-life** was far longer than had been previously thought. This was followed by a prediction by **Serge A. Korff**, then employed at the **Franklin Institute** in **Philadelphia**, that the interaction of **thermal neutrons** ...



Claim 3

- Co-coupling patterns (bibliographic count, co-coupling count) can be improved by using relevance of shared entities



Link Types

1) Relevant Co-occurrence Graph
2) Unweighted Link Patterns
3) Relevance-weighted Coupling

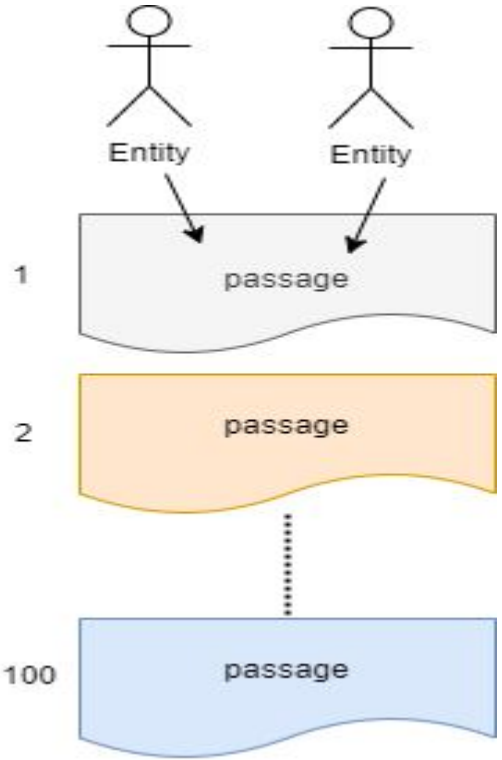
Relevant Co-occurrence Graph Link Type

- Based on thought experiment that if entities co-occur in the relevant text passages, then a relevant link exists between entities.
- We retrieve the entity-linked text passages and create entity pairs (e_i, e_j). For e.g., (Martin Kamen, Samuel Ruben)

In 1939, **Martin Kamen** and **Samuel Ruben** of the **Radiation Laboratory at Berkeley** began experiments to determine if any of the elements common in organic matter had isotopes with half-lives long enough to be of value in biomedical research. They synthesized ^{14}C using the laboratory's cyclotron accelerator and soon discovered that the atom's **half-life** was far longer than had been previously thought. This was followed by a prediction by **Serge A. Korff**, then employed at the **Franklin Institute** in **Philadelphia**, that the interaction of **thermal neutrons** with ^{14}N in the upper atmosphere would create ^{14}C. It had previously been thought that ^{14}C would be more likely to be created by **deuterons** interacting with ^{13}C. At some time during World War II, **Willard Libby**, who was then at Berkeley, learned of Korff's research and conceived the idea that it might be possible to use radiocarbon for dating.

Relevant Co-occurrence Graph Indicators

1. **Co-occ Relevance:** reciprocal rank of co-occurring entities
2. **Co-occ Count:** frequency of co-occurring entities
3. **Mention Freq:** frequency of each entity in the text passages



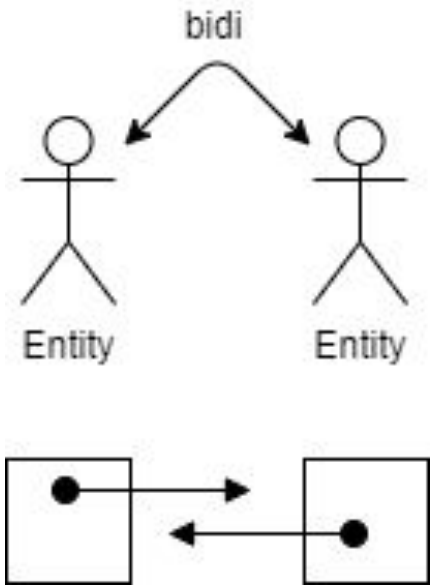
Unweighted Link Patterns Link Type

- Based on Knowledge Base links i.e. inlinks and outlinks of Wikipedia
- We determine whether link exists between the entity pairs (e_i, e_j)

Unweighted Link Patterns Indicators

Direct links: For entity pair (e_i, e_j)

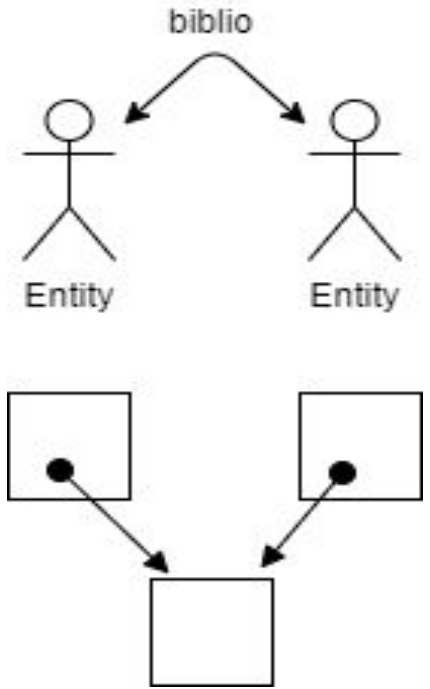
1. **Outlinks:** link from e_i to e_j
2. **Inlinks:** link from e_j to e_i
3. **Bidirectional:** outlink and inlink between e_i and e_j
4. **Undirected:** outlink or inlink between e_i and e_j



Unweighted Link Patterns Indicators

Coupling measures: For entity pair (e_i, e_j)

1. **Bibliographic:** shared outlinks between e_i and e_j
2. **Co-coupling:** common inlinks between e_i and e_j

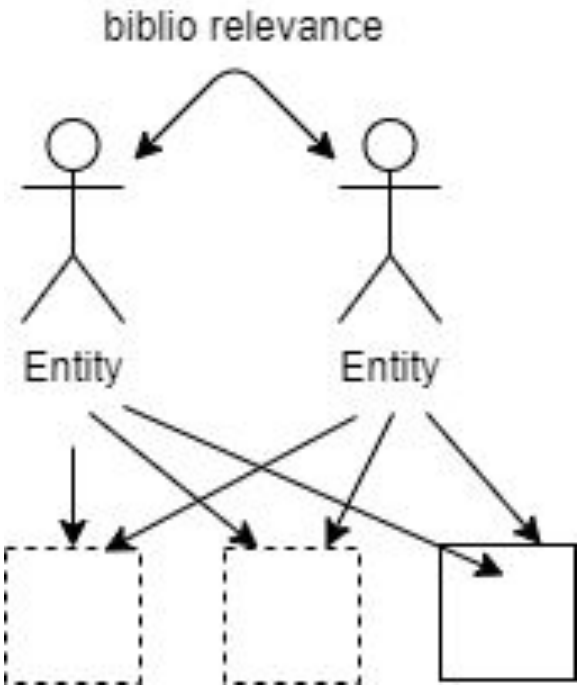


Relevance-weighted Coupling Link Type

- Hybrid of knowledge base links patterns approach and relevance-based approach
- Scores of ranked knowledge base entries is used as relevance factor

Relevance-weighted Coupling Indicators

1. **Bibliographic Relevance:** shared outlinks weighted by relevance score of ranked knowledge base entries
2. **Co-coupling Relevance:** shared inlinks weighted by relevance score of ranked knowledge base entries



Entity Ranking

- Final entity e_i score - for each indicator (except outlinks, inlinks, bidirectional), accumulate the scores of all entity pairs in which the entity e_i is present
- Learning-to-rank framework to study the combined effect of the indicators
- Top 4 best performing indicators when combined performs best

Evaluation

- TREC Complex Answer Retrieval Y1 benchmark

- 3 evaluation metrics
  - Mean Average Precision (MAP)
  - Rprecision (Rprec)
  - F1

Queries: Story request

- Story Requests
  - “2020 US presidential elections nominations”
  - “Radiocarbon dating samples”
  - “Radiocarbon dating background history”

- 1952 story requests

Results - Claim 1

- Retrieving entities through relevant passage is more effective than knowledge base attributes

Indicators	**MAP**	**Rprec**	**F1**
Co-occ Relevance	**0.14 ±0.005**	0.14 ±0.005	0.06 ±0.001
Co-occ Count	0.09 ±0.003	0.10 ±0.003	0.06 ±0.001
Mention-Freq	0.11 ±0.003	0.11 ±0.003	0.07 ±0.001
KG-Entity	0.03 ±0.002	0.03 ±0.002	0.01 ±0.000

Results - Claim 2

- Relevance through relevant passages is more effective than knowledge base links

Indicators	**MAP**	**Rprec**	**F1**
Co-occ Relevance	**0.14 ±0.005**	0.14 ±0.005	0.06 ±0.001
Co-occ Count	0.09 ±0.003	0.10 ±0.003	0.06 ±0.001
Mention-Freq	0.11 ±0.003	0.11 ±0.003	0.07 ±0.001
Undirected	0.07 ±0.002	0.08 ±0.003	0.06 ±0.001
Biblio Count	0.07 ±0.002	0.07 ±0.002	0.06 ±0.001
Co-coupling Count	0.03 ±0.001	0.03 ±0.001	0.05 ±0.001

Results - Claim 3

- Co-coupling patterns (bibliographic count, co-coupling count) can be improved by using relevance of shared entities

Indicators	**MAP**	**Rprec**	**F1**
Biblio Count	**0.07 ±0.002**	0.07 ±0.002	0.06 ±0.001
Co-coupling Count	0.03 ±0.001	0.03 ±0.001	0.05 ±0.001
Biblio Relevance	0.05 ±0.002	0.04 ±0.002	0.05 ±0.001
Co-coupling Relevance	**0.07 ±0.002**	0.07 ±0.003	0.06 ±0.001

Queries: Multi-story query request

- Multi-story Query Requests
  - Consists of multiple story requests connected by a larger theme
  - Example: “Radiocarbon dating”
    - Sub-stories: [“Radiocarbon dating samples”, “Radiocarbon dating background history”]
  - Results are combination of results of multiple story requests

- 132 multi-story query requests

Results - Claim 1

- Retrieving entities through relevant passage is more effective than knowledge base attributes

Indicators	**MAP**	**Rprec**	**F1**
Co-occ Relevance	**0.21 ±0.011**	0.30 ±0.010	0.31 ±0.009
Co-occ Count	0.16 ±0.008	0.26 ±0.009	0.27 ±0.008
Mention-Freq	0.19 ±0.008	0.30 ±0.008	0.31 ±0.008
KG-Entity	0.01 ±0.002	0.05 ±0.003	0.04 ±0.003

Results - Claim 2

- Relevance through relevant passages is more effective than knowledge base links

Indicators	**MAP**	**Rprec**	**F1**
Co-occ Relevance	**0.21 ±0.011**	0.30 ±0.010	0.31 ±0.009
Co-occ Count	0.16 ±0.008	0.26 ±0.009	0.27 ±0.008
Mention-Freq	0.19 ±0.008	0.30 ±0.008	0.31 ±0.008
Undirected	0.16 ±0.007	0.27 ±0.007	0.28 ±0.008
Biblio Count	0.14 ±0.006	0.25 ±0.008	0.27 ±0.007
Co-coupling Count	0.08 ±0.005	0.18 ±0.007	0.21 ±0.007

Results - Claim 3

- Co-coupling patterns (bibliographic count, co-coupling count) can be improved by using relevance of shared entities

Indicators	**MAP**	**Rprec**	**F1**
Biblio Count	**0.14 ±0.006**	0.25 ±0.008	0.27 ±0.007
Co-coupling Count	0.08 ±0.005	0.18 ±0.007	0.21 ±0.007
Biblio Relevance	0.10 ±0.006	0.21 ±0.007	0.23 ±0.008
Co-coupling Relevance	**0.13 ±0.006**	0.23 ±0.008	0.24 ±0.008

Conclusion

- Relevant text proves to be a strong indicator - between 80% and 30% - than the knowledge base links
- Relevant text based signals are more effective for retrieval of relevant entities

NSF CAREER, Grant No. 1846017

Thank you!