Utilizing Knowledge Bases for Text Retrieval: A Wishlist

Laura Dietz
dietz@cs.unh.edu

I'll be honest: we physicists talk a big game about the theory of everything, but the truth is, we don't really understand why ice skates work, how sand flows, or where the static charge comes from when you rub your hair with a balloon.
Utilizing Knowledge Graphs in Text-centric Information Retrieval

Laura Dietz (@lauradietz)
University of New Hampshire

Alex Alexelev (@alexex)
Retrieval for Open-ended Information Needs

Requiring long, complex answers

Intended queries:
- how ice skates work
- UK leaving Europe
- cashflow important for investment
- effects of water pollution
- Diesel scandal affect Daimler AG

If yes, why? If not, why not?
What do I need to know to understand the answer?
What is the problem? ...and the solution?

Wikipedia

Web Search

Not enough / recent information

Manually sift through many web pages

Train computers to recycle Web content to write a comprehensive articles in response to a search query

Laura Dietz     dietz@cs.unh.edu    Utilizing Knowledge Bases for Text Retrieval: A Wishlist
Query

predominant facts and introduction

more details about Heading 1

more details about Heading 2
Step 1: Find Relevant/Central Entities

Query

predominant facts and introduction
more details about Heading 1
more details about Heading 2
Step 2: Find Relevant Relations

Query

predominant facts and introduction

more details about Heading 1

more details about Heading 2
Step 3: Find Relevant Text + Consolidate

Query
- predominant facts and introduction
- more details about Heading 1
- more details about Heading 2
How to Find Relevant Entities?

Q: diesel scandal affect Daimler
How to Find Relevant Entities?

(1) Entity linking the query

Q: diesel scandal affect Daimler

(2) Search in KB index

(3) Relevance Feedback

pretend these are relevant
How to Use Entities for Text Ranking?

(1) Entity linking the query

Q: diesel scandal affect Daimler

(2) Search in KB index

Q  →  [database]

(3) Relevance Feedback

Q  →  [database]

pretend these are relevant

Laura Dietz  dietz@cs.unh.edu  Utilizing Knowledge Bases for Text Retrieval: A Wishlist
Finding Relevant Entities: What Works?

(1) Entity linking the query

Q: diesel scandal affect Daimler <- Sparse

(2) Search in KB index

Q ➔ Wiki pages of relevant entities may not mention query

(3) Relevance Feedback

Q ➔ pretend these are relevant

Strongest feature! room for improvement
Identifying Relevant Relations in a KG

Q: diesel scandal affect Daimler

Naive approach: Select sub-KG of relevant entities.

So many connections in a knowledge graph
- Some are relevant!
- But many are only relevant in a certain (other?) context.
KGs started with the "most popular" facts then it grew in number of nodes and number of connections, aiming for better coverage.

Hub nodes: New York City, California, United States
Link Structure in KGs Became Unhelpful

KGs started with the "most popular" facts then it grew in number of nodes and number of connections, aiming for better coverage.

KGs in 2013

KGs in 2019

Hub nodes: New York City, California, United States
KGs started with the "most popular" facts then it grew in number of nodes and number of connections, aiming for better coverage.

Hub nodes: New York City, California, United States
ENT Rank for Entity Ranking

(1) Retrieve text + entity links and entities

(2) Build candidate graph

(3) Learn edge weights & Predict entity ranking

Laura Dietz dietz@cs.unh.edu Utilizing Knowledge Bases for Text Retrieval: A Wishlist
ENT Rank for Entity Ranking

(1) Retrieve text + entity links and entities

(2) Build candidate graph

(3) Learn edge weights & Predict entity ranking

Laura Dietz dietz@cs.unh.edu Utilizing Knowledge Bases for Text Retrieval: A Wishlist
Volkswagen had intentionally programmed turbocharged direct injection (TDI) diesel engines to activate some emissions controls only during laboratory emissions testing.

.. investor lawsuit seeking class action status ... seeking compensation for the drop in stock value due to the emissions scandal.
Volkswagen had intentionally programmed turbocharged direct injection (TDI) diesel engines to activate some emissions controls only during laboratory emissions testing. .. investor lawsuit seeking class action status ... seeking compensation for the drop in stock value due to the emissions scandal.
Extracting Relevant Relations

Relation Extraction:
[Schuhmacher, Roth, Ponzetto, Dietz 16]
[Roth et al 14]
(best at TAC KBP 13)

Research question:
relevant documents + extraction = relevant relations?

Laura Dietz dietz@cs.unh.edu Utilizing Knowledge Bases for Text Retrieval: A Wishlist
Goal: Relations need to be relevant and correct

Query: Raspberry Pi

[diagram showing relationships between entities]

no signal

rf: relation extraction

dbp: knowledge base

50% / 50%

N/A 60% queries
Goal: Relations need to be relevant and correct only considering correct extractions.

Issue 1: Correct Vs. Relevant Extractions

Schema-based: 50% relevant [Schuhmacher 16]
OpenIE-based: 50% relevant [Kadry & Dietz 17]
Human-based: 50% relevant (sentence-level)

50% / 50%
Schema-based: N/A for 60% of queries (TAC KBP 13)

Open IE: 5% sentences with correct annotations (no coref)
Leads to only marginal improvements for IR, e.g.
Ranking entity-query support sentences for relevance.

![Bar chart showing MAP for different methods: All together, TF-IDF, POS/NER, Parsing, OpenIE]
Interesting relations are a bit more complicated.

Volkswagen had intentionally programmed turbocharged direct injection (TDI) diesel engines to activate some emissions controls only during laboratory emissions testing.

.. investor lawsuit seeking class action status ... seeking compensation for the drop in stock value due to the emissions scandal.

Beyond more than one sentence. Include multiple intermediate entities. ...also not just triples + coref...
Data: Effects of Water Pollution/Eutrophication

6. The pollution often comes from non point sources such as agricultural runoff, wind-blown debris and dust. Nutrient pollution, a form of water pollution, refers to contamination by excessive inputs of nutrients. It is a primary cause of eutrophication of surface waters, in which excess nutrients, usually nitrogen or phosphorus, stimulate algae growth.

7. Nutrient pollution, a form of water pollution, refers to contamination by excessive inputs of nutrients. It is a primary cause of eutrophication of surface waters, in which excess nutrients, usually nitrogen or phosphorus, stimulate algal growth. Sources of nutrient pollution include surface runoff from farm fields and pastures, discharges from septic tanks and feedlots, and emissions from combustion. Excess nutrients have been summarized as potentially leading to:

8. Human interference in the phosphorus cycle occurs by overuse or careless use of phosphorus fertilizers. This results in increased amounts of phosphorus as pollutants in bodies of water, resulting in eutrophication. Eutrophication devastates water ecosystems by inducing anoxic conditions.
Given: query $Q$ and outline of Headings

**CAR Y1, Y2:**
Paragraph ranking per heading.
Optimize relevance

**CAR Y3:**
Paragraph ordering.
Maximize coverage, topical coherence

**Up next:**
Multi-paragraph summarization + query-KG

---

Laura Dietz     dietz@cs.unh.edu    Utilizing Knowledge Bases for Text Retrieval: A Wishlist
General purpose schema with many types

High coverage/recall (40%?)

Extraction of complex relations (not just triples + coref)

Bridging existing KGs with text

Relevant information extraction

Query-specific knowledge graphs

TREC CAR Dataset http://trec-car.cs.unh.edu/
Ask me for a data set to play around with...
Laura Dietz dietz@cs.unh.edu Utilizing Knowledge Bases for Text Retrieval: A Wishlist