Towards Proactive Information Retrieval in Noisy Text with Wikipedia Concepts

Tabish Ahmed, Sahan Bulathwela ,

Centre for Artificial Intelligence, University College London (UK)

Introduction

- **Proactive Information Retrieval (IR)**
  - Without interrupting the User-experience → Minimizing Human Effort
  - Need for Proactive IR → Queries can be short
  - State-of-the-art Neural Approaches→ Sensitivity to Noise (Jones et al. 2021)
- **Spotify Podcast Dataset**
  - ASR transcripts of ~105,000 Podcasts with 18% Word-Error Rate
  - Segment Retrieval task with ~ 3.5 Million, 2-minute segments of Podcasts with a 1-minute overlap

- **Spotify Podcast Dataset**

  - 8 training topics / 50 testing topics

  - All topics provided with descriptions → proxy for user history (in the context of Proactive IR)

- **Differences from Previous Work**

  - Wikipedia-based entity linking has been previously explored by

  - (Azad & Deepak, 2019) and (Nasir et al., 2019) albeit on non-noisy text with a focus on Query-Expansion

  - NER / POS-tagging for Noisy Retrieval was shown to be effective by DCU (Moriya & Jones, 2020)

  - Another state-of-the-art approach (Jones et al., 2021) focuses on

    - word embeddings + Sequential Dependence Model + Neural re-ranking (Galuscáková et al., 2020)

  - Our approach differs in

  - 'Wikification' of Queries + Segments

  - We choose the 8 training topics and descriptions with ~14,000 negative segments per topic

  → Down-sampled dataset called **Podcast Small → 14,179, annotated 2 min segments**

Previous Work

- **Concept Based User Modelling**

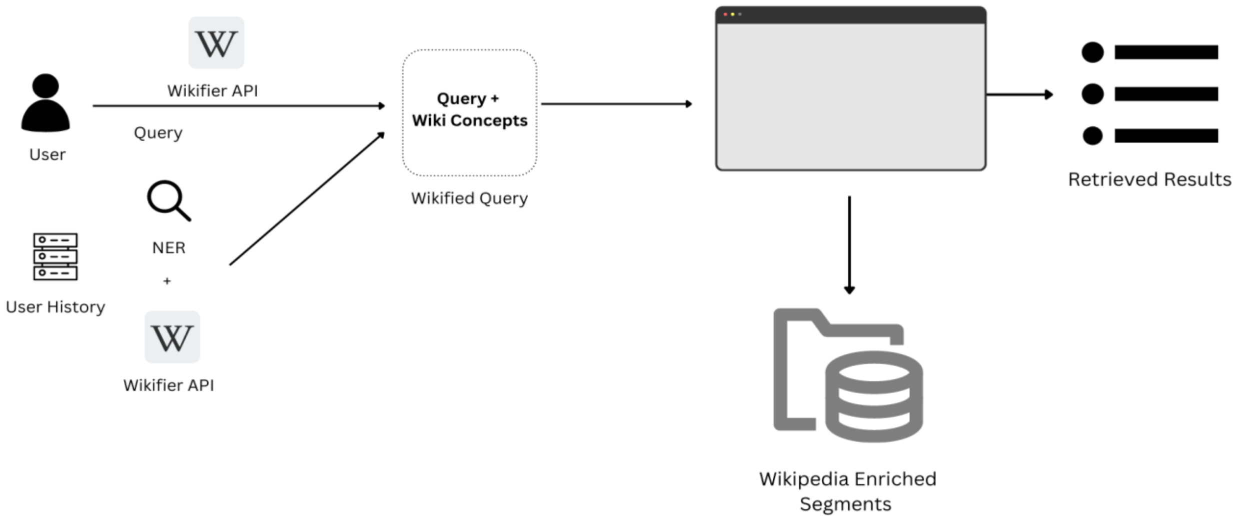
  - Short text of social media posts can be used to build a user profile (Piao and Breslin, 2016)
    - Feature space can be too vast → Challenge in increasing recall
  - Expert-based annotation to improve retrieval has been explored in Education (Corbett & Anderson, 1994)
    - But, is it scalable?
  - ‘Wikification’, Connecting natural text to Wikipedia articles (Brank et al., 2017)
  - Eventual emphasis on disambiguating the meaning of the query through key-word extraction

Previous Work

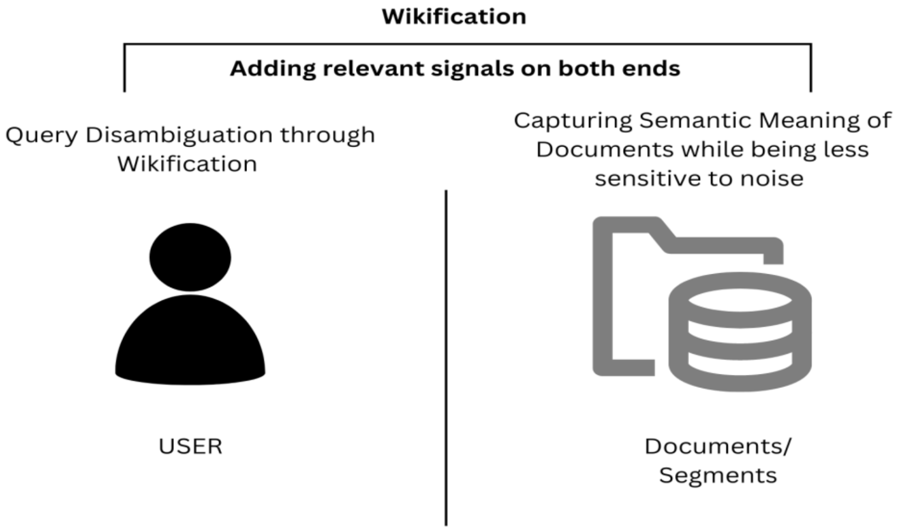
- **Noisy Information Retrieval**

  - Sophisticated probabilistic models outperform state-of-the-art Neural IR approaches (Jones, 2021)
    - A linear combination of BM25 and DPH used by DCU (Moriya & Jones, 2020)
  - Noise sensitivity of Neural approaches was further recently corroborated by (Sidiropoulos et al., 2022)
  - Combine and re-rank approach by (Galuscáková et al., 2020) uses
    - query + description (as modified queries)
    - word-embeddings and a sequential dependence model
    - a final neural re-ranking with a model trained on an orthogonal dataset

Overall Design



Will “Two-pronged” Wikification lead to a better overlap?



Research Questions

RQ1 : Do Wikipedia concepts carry a signal that indicates relevance of documents to queries?

RQ2: Can Wikipedia annotations improve noisy information retrieval?

Down-Sampled Data → Wikfier-API → Jaccard Similarity Coefficient → One-tailed Mann Whitney-U-Test

Relevant — Non-relevant

Is there a statistically significant difference between the median of Jaccard Similarity of the two groups??

Research Questions

RQ1 : Do Wikipedia concepts carry a signal that indicates relevance of documents to queries?

RQ2: Can Wikipedia annotations improve noisy information retrieval?

PODCAST SMALL → Baseline Performance → DCU Approach BM 25 + DPH → Proposed Models → Evaluation

- NDCG
- NDCG @ 30
- Precision @ 10

BM25 DPH

- **Wiki_rel**
- **Ent_wiki_rel**

Extending more on RQ2

- **Baselines**

  - Both Models (BM25, DPH) are informed by results published in (Jones et al., 2021)

  - DPH is based on the Divergence From Randomness Framework (Amati, 2006)

$$rel(q, s) = f(q_{\mathrm{txt}}, s_{\mathrm{txt}})$$

- **DCU Approach**

  - A linear combination of BM25 + DPH models

  - The approach we picked from DCU approaches was

  - *Topics → Topics + Entities (Description) → IR on (BM25 + DPH) → Evaluation*

$$rel(q, d, s) = f(q_{\mathrm{txt}} + d_{\mathrm{ent}}, s_{\mathrm{txt}})$$

Proposed Models

- **Wiki_rel**

  - The Model Differs from the DCU approach by

  → Using Wikipedia concepts extracted from the entire description rather than just the entities

$$rel(q, d, s) = f(q_{\text{txt}} + q_{\text{wiki}} + d_{\text{wiki}}, s_{\text{txt}} + s_{\text{wiki}})$$

- **Ent_wiki_rel**

  - The model differs from **Wiki_rel** approach by

  → Entities are also added to query from description

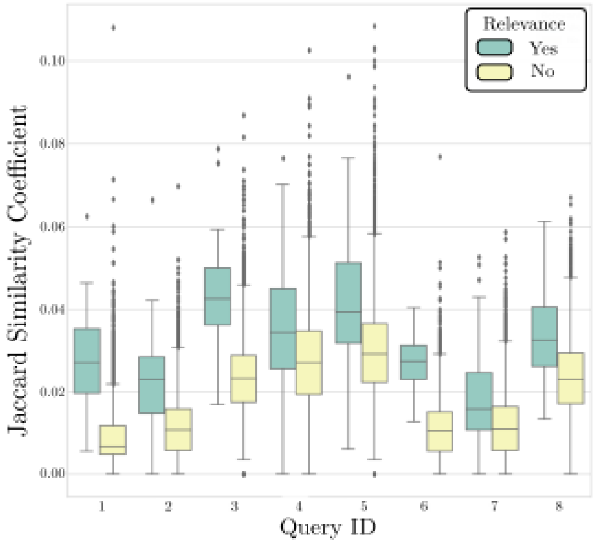
$$rel(q, d, s) = f(q_{\text{txt}} + d_{\text{ent}} + q_{\text{wiki}} + d_{\text{wiki}}, s_{\text{txt}} + s_{\text{wiki}})$$

Results RQ1

	Number of Documents		Median Jaccard Similarity			MannWhitney U test	
Query ID	Relevant	Non-Relevant	Relevant	Non-Relevant	Difference	Test Statistic	p value
1	70	14109	0.027	0.007	0.020	25680	1.03E-49
2	63	14116	0.023	0.011	0.012	47466	8.84E-46
3	78	14101	0.043	0.023	0.019	264688	1.22E-16
4	78	14101	0.034	0.027	0.007	393033	1.41E-06
5	80	14099	0.039	0.029	0.010	455015	1.42E-03
6	37	14142	0.027	0.010	0.017	35748	8.37E-48
7	77	14102	0.016	0.011	0.005	56322	2.80E-44
8	80	14099	0.032	0.023	0.010	271961	6.31E-16

- We Reject the Null Hypothesis. The median of Jaccard Similarity of the relevant set is > the median on non-relevant set

Results RQ1



- We Reject the Null Hypothesis. The median of Jaccard Similarity of the relevant set is > the median on non-relevant set

Results RQ2

Model	Features: Query Text	Features: NER Entities	Features: Wiki Concepts	Metrics: NDCG	Metrics: NDCG at 30	Metrics: Precision at 10
Baselines						
DPH	×	o	o	0.48	*0.30*	*0.32*
BM25	×	o	o	0.48	0.28	0.31
DCU	×	×	o	**0.51**	**0.32**	0.30
New Proposals						
Wiki_rel	×	o	×	*0.49*	0.29	0.31
Ent_Wiki_rel	×	×	×	**0.51**	*0.30*	**0.36**

- Results show there is promise in using Wikipedia concepts
- Gains in early precision, but NDCG performance is similar ?

Discussion

- **No Significant NDCG gains because**
- Slightly less-relevant topics contained in the segments have slightly higher scores
- Irrelevant concepts add noise (not complete noise but irrelevant topics get added to segments/topics)
- Results can go up if we only use Wikipedia topics which have anchors but have we haven't run that experiment

Discussion

- Entities from Description, similarly can also add such noise through slightly irrelevant concepts
- Our future work involves

 - **A full-scale study on the annotated Spotify Podcast Dataset (Wikified) on Testing topics**

- A user study to if the approach is useful in a real-world scenario

Discussion

Query	Query Only	Query + Description
coronavirus spread	wiki/Coronavirus	wiki/Novel_coronavirus
greta thunberg cross atlantic	-	wiki/Greta_Thunberg
black hole image	wiki/Black_hole	wiki/Black_hole
daniel ek interview	-	wiki/Daniel_Ek
michelle obama becoming	-	wiki/Michelle_Obama
	wiki/Becoming_(philosophy)	wiki/Becoming_(book)
anna delvey	wiki/Indian_anna	wiki/Anna_Sorokin
facebook stock prediction	wiki/Facebook	wiki/Facebook
	wiki/Stock	wiki/Stock

- Query Disambiguation using Wikipedia Concepts

Conclusion

- The overlap between Wikified segments and queries is statistically significant in terms of medians of Jaccard Similarity Coefficient with the relevant set having a significantly higher median
- Using Wikipedia Concepts in IR shows promising results and invites a full-scale discussion with anchor-topics and refining the approach in how only relevant Wikipedia Topics are added
- A “two-pronged” Wikification approach can facilitate a higher-degree ‘human-in-the-loop’ operation

Department of Computer Science

Towards Proactive Information Retrieval in Noisy Text with Wikipedia Concepts

Tabish Ahmed, Sahan Bulathwela ,

Centre for Artificial Intelligence, University College London (UK)

*Contact: tabish.ahmed.21@ucl.ac.uk , **m.bulathwela@ucl.ac.uk***

References

- H. K. Azad, A. Deepak, A new approach for query expansion using wikipedia and wordnet, Information sciences 492 (2019) 147–163..
- J. A. Nasir, I. Varlamis, S. Ishfaq, A knowledge-based semantic framework for query expansion, Information processing & management 56 (2019) 1605–1617
- G. Amati, Frequentist and bayesian approach to information retrieval, in: European Conference on Information Retrieval, Springer, 2006, pp. 13–24.
- P. Galušcáková, S. Nair, D. W. Oard, Combine and re-rank: The university of maryland at the trec 2020 podcasts track (2020)
- Y. Moriya, G. J. Jones, Dcu-adapt at the trec 2020 podcasts track., in: TREC, 2020
- P. Sen, Proactive information retrieval, Ph.D. thesis, Dublin City University, 2021.
- G. Piao, J. G. Breslin, Analyzing mooc entries of professionals on linkedin for user modeling and personalized mooc recommendations, in: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP ’16, 2016.
- F. Zarrinkalam, H. Fani, E. Bagheri, M. Kahani, Predicting users’ future interests on twitter,in: European Conference on Information Retrieval, Springer, 2017, pp. 464–476.

References

- A. T. Corbett, J. R. Anderson, Knowledge tracing: Modeling the acquisition of procedural knowledge, User Modeling and User-Adapted Interaction 4 (1994).
- J. Brank, G. Leban, M. Grobelnik, Annotating documents with relevant wikipedia concepts, in: Proc. of Slovenian KDD Conf. on Data Mining and Data Warehouses (SiKDD), 2017
- R. Jones, B. Carterette, A. Clifton, M. Eskevich, G. J. Jones, J. Karlgren, A. Pappu, S. Reddy, Y. Yu, Trec 2020 podcasts track overview, arXiv preprint arXiv:2103.15953 (2021).
- G. Sidiropoulos, S. Vakulenko, E. Kanoulas, On the impact of speech recognition errors in passage retrieval for spoken question answering, arXiv preprint arXiv:2209.12944 (2022).