

On Building Benchmark Datasets for Understudied Information Retrieval Tasks: the Case of Semantic Query Labeling

Discussion Paper

Elias Bassani^{1,2}, Gabriella Pasi²

¹Consorzio per il Trasferimento Tecnologico - C2T, Milan, Italy

²University of Milano-Bicocca, Milan, Italy

Abstract

In this manuscript, we review the work we undertake to build a large-scale benchmark dataset for an understudied Information Retrieval task called Semantic Query Labeling. This task is particularly relevant for search tasks that involve structured documents, such as Vertical Search, and consists of automatically recognizing the parts that compose a query and unfolding the relations between the query terms and the documents' fields. We first motivate the importance of building novel evaluation datasets for less popular Information Retrieval tasks. Then, we give an in-depth description of the procedure we followed to build our dataset.

Keywords

Vertical search, Structured document search, Semantic query labeling, Dataset

1. Introduction

The past few years have witnessed a continuous rise of interest in the application of Deep Learning techniques to Information Retrieval (IR) tasks. As reported in a recent survey by Guo et al. [1], the IR community has mostly focused on the application of Neural Networks to *Ad-hoc Retrieval* ([2, 3, 4, 5]), *Question Answering* ([6]), *Community Question Answering* ([7, 8]), and *Automatic Conversation* ([9, 10]). However, the potential of Deep Learning in solving many other IR tasks remains mostly unexplored.


The availability of multiple large-scale datasets for models bench-marking and evaluation is one of the principal factor for raising the interest of the research community towards specific tasks. For example, for the evaluation of *Question Answering* many benchmark datasets have been developed, such as TREC QA [11], WikiQA [12], WebPA [13], InsuranceQA [14], WikiPassageQA [15], and MS MARCO [16]. Sometimes it is easy to build large-scale datasets for specific tasks with low effort by leveraging publicly available online resources, such as

IIR 2021 – 11th Italian Information Retrieval Workshop, September 13–15, 2021, Bari, Italy

✉ e.bassani3@campus.unimib.it (E. Bassani); gabriella.pasi@unimib.it (G. Pasi)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Community Question Answering platforms (e.g., Quora¹, Yahoo! Answer², Stack Overflow³, etc.). *Community Question Answering* research datasets include Quora Dataset⁴, Yahoo! Answers Dataset [7], SemEval-2017 Task3 [17], CQADupStack [18], ComQA [19], and LinkSO [20]. Moreover, some big private companies have actively contributed to provide *expensive* large-scale benchmark datasets to the research community, such as Microsoft⁵ with its MS MARCO [16] dataset. Unfortunately, other tasks appear to be research matter only for those companies that can afford to produce the datasets needed for models training and evaluation, and, unfortunately, the majority of these datasets are never made available to the research community. As well known, this situation also poses reproducibility issues that can hardly be overcome.

One of the IR sub-fields that received limited attention from academicians for the study of the application of Deep Learning techniques is Vertical Search. However, nowadays, many different kinds of vertical online platforms, such as e-commerce websites (e.g., Amazon⁶), media streaming services (e.g., Netflix⁷, Spotify⁸), job-seeking platforms (e.g., LinkedIn⁹), digital libraries (e.g., DBLP¹⁰), and several others, provide access to domain-specific information through a search engine to millions of users every day. What makes Vertical Search interesting from a research perspective and, potentially, for the application of sophisticated Machine Learning-based approaches is that vertical platforms usually organize their information in structured documents, which require to be treated appropriately during search to leverage the additional information encoded in their structure. However, search functionalities on vertical platforms are usually delivered as standard keyword-based search, or through *uncomfortable* faceted search interfaces, which require additional effort from the user. Unlike in Web Search, user queries in vertical systems often contain references to specific structured information contained in the documents. Nevertheless, Vertical Search is often managed as a traditional retrieval task, treating documents as unstructured texts and taking no advantage of the latent structure carried by the queries. Exploiting this latent information could unfold the relations between the query terms and the documents' structure, thus enabling the search engine to leverage the latter during retrieval.

2. Semantic Query Labeling

Semantic Query Labeling [21] is the task of 1) locating the constituent parts of a query (*segmentation*) and 2) assigning predefined and domain-specific semantic labels to each of them (*classification*). Conducting this task in a pre-matching phase could allow a search engine to leverage the structure and the semantics of the query terms, making it able to effectively take advantage of the structure of the documents during retrieval, thus enhancing the matching

¹<https://www.quora.com>

²<https://yahoo.com>

³<https://stackoverflow.com>

⁴<https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs>

⁵<https://www.microsoft.com>

⁶<https://www.amazon.com>

⁷<https://www.netflix.com>

⁸<https://www.spotify.com>

⁹<https://www.linkedin.com>

¹⁰<https://www.dblp.org>

process. For example, in the movie domain, the query “*alien ridley scott 1979*” carries references to structured information usually contained in the documents of a movie corpus: the title of a movie, *Alien*, the name of a movie director, *Ridley Scott*, and a date, *1979*. In this case, the query could be segmented accordingly into *alien*, *ridley scott*, and *1979* and the query segments could be tagged with the labels *Title*, *Director*, and *Year*, respectively.

Semantic Query Labelling is a challenging task that can add context and structure to keyword-based queries, usually composed of a few terms that may be ambiguous. The main challenges of this task are related to the vocabulary overlap among different semantic classes, which could require the use of contextual information and disambiguation techniques, and vocabulary mismatch [22] between the vocabulary employed by the users to express their information need and the vocabulary used to describe the corresponding answers in the document collection. Unfortunately, the production of an appropriate dataset to evaluate the effectiveness of automatic query tagging approaches is costly, and actually, there is a lack of publicly available datasets for this task.

Despite semantic query labelling could play an important role in Vertical Search, very little work has been done in this regard. The majority of past efforts in this context come from private companies, such as Microsoft ([21, 23, 24, 25, 26]) and Yahoo! ([27]). Due to privacy issues, companies cannot release the datasets used in their studies. As well known, this makes it hard to reproduce their approaches and comparatively evaluate them. Moreover, the lack of public datasets makes it difficult for academic researchers to propose novel Semantic Query Labeling models, and evaluate their effectiveness.

As we strongly believe in the utility of advancing in Vertical Search, we have recently undertaken a step towards the definition of a benchmark dataset for this task¹¹.

3. Building a Benchmark Dataset for Semantic Query Labeling

In this section, we describe the dataset we have defined and shared [28], as well as the process we followed for manually annotating each query term. Our dataset is composed of thousands of manually-labeled real-world queries in the movie domain for training and evaluating novel methods for Semantic Query Labeling.

The choice of working in the movie domain is motivated by the fact that movie streaming platforms are popular nowadays, but they still provide a sub-optimal search experience to their users. Moreover, structured search is fundamental in this context: as we assessed during our work described here, users tend to compose their queries referring to specific movie-related information, such as the name of an actor or a director, a movie genre, a topic, and others, which are usually available as metadata. By conducting a qualitative evaluation of the top 10 results returned by the search engine of one of the most popular movie streaming services, we assessed that it is not able to correctly retrieve movies even for simple queries. For example, “*horror 2015*” retrieved only one horror movie from 2015, many other results were neither horror movies nor movies from 2015. “*2015 horror*” did not retrieve any result at all. Neither “*leone eastwood*” nor “*sergio leone clint eastwood*” retrieved any result despite the presence on the platform of all the movies directed by *Sergio Leone* and starring *Clint Eastwood* at the time of the experiment.

¹¹<https://github.com/AmenRa/semantic-query-tagging-dataset>

3.1. Query Gathering

The first step in building a dataset suitable for studying Semantic Query Labeling is the query gathering. To collect the queries that are part of our dataset, we relied on a publicly available large-scale query log of the AOL Web search engine¹², which was shared by Pass et al. [29]. This query set comprises queries issued by real users between March 1, 2006, and May 31, 2006. First of all, we defined a list of seed-terms for identifying movie-related queries: *movie*, *movies*, *film*, and *films*. Leveraging these terms, we extracted 39 635 unique queries. Then, we *manually* filtered out all the queries that did not fall into our category of interest: keyword-based queries that resemble those used by users for searching movies on movie streaming platforms. As the large majority of the initially extracted queries were related to theaters’ movie listings — note that AOL offers a general-purpose Web search engine — we ended up collecting 9 752 candidate queries. After removing the seed-terms used for gathering the queries, *manually* correcting misspellings, normalizing strings, removing the stop-words, and applying lemmatization, our dataset counts 6 749 unique queries.

3.2. Semantic Labels Assessment

The second step in the building process of our dataset was to define 1) the semantic label set to use for the creation of the ground truth and 2) the procedure to follow to assign the semantic labels to the query terms, ensuring the quality of the proposed dataset.

3.2.1. Semantic Labels

After an initial analysis of the harvested queries, we defined the following semantic classes to assign to each query term: *Title*, *Country*, *Year*, *Genre*, *Director*, *Actor*, *Production company*, *Tag* (mainly topics and plot features), *Sort* (e.g., *new*, *best*, *popular*, etc.). Following previews works in Natural Language Processing and Sequence Labeling [30], we used the IOB2 labeling format [31, 32] for *manually* assigning both semantic labels and segmentation delimiters. For example, the query “*alien by ridley scott 1979*” is labeled as follows: “*alien B-TITLE by O ridley B-DIRECTOR scott I-DIRECTOR 1979 B-YEAR*”, where the prefix **B-** indicates the beginning of a segment, the prefix **I-** indicates that the term is *inside* a segment, and the tag **O** is used to label terms with no semantic values, such as the preposition *by* in our example.

3.2.2. Creation of the Ground Truth

One of the main reasons for choosing to work in the movie domain is the public availability of movie-related information. We relied on this information to ensure the quality of the ground truth labels we *manually* assigned to the query terms. In this regard, we consulted many websites that contain movie-related information while labeling the queries, such as Wikipedia¹³, IMDb¹⁴, and many others. Furthermore, particular attention was paid in discerning actors from directors, as sometimes a single person is both an actor and a director, such as *Ron Howard*.

¹²<https://www.aol.com>

¹³<https://www.wikipedia.org>

¹⁴<https://www.imdb.com>

In these cases, we followed a simple rule: if the query contains elements pointing towards a specific interpretation of the query, we labeled the query accordingly (e.g., in the query “1999 ron howard”, *Ron Howard* has been labeled as a *Director* as in 1999 he directed the movie *EDtv* and did not star in any movie), otherwise we assigned the most likely label based on the number of movies the person has directed or starred. Therefore, we can state that, where meaningful, we applied a *contextual* labeling.

3.3. Building a Fine-grained Evaluation Setting

To promote a realistic evaluation setting, we split the dataset into *train*, *dev*, and *test* sets temporally, using the queries issued in the first two months as *train set*, and those from the two subsequent two-weeks periods as *dev set* and *test set*. Temporal splitting also reduces query term overlaps between the splits: we noticed that queries issued by users in the same search session often share several terms. We also observed that *not* taking care of this aspect could yield unrealistic results when training with real-world data.

To build a fine-grained evaluation setting, we created three different scenarios of increasing difficulty by subsetting our benchmark dataset. The first scenario we built, *Basic*, comprises only queries containing the following semantic components: *Actor*, *Country*, *Genre*, *Title*, *Year*, and *O*. We then added the semantic components *Director* and *Sort* to create the *Advanced* scenario. Finally, we added *Production Company* and *Tag* to create the *Hard* scenario. The rationale behind these choices is as follows: the *Basic* scenario is composed of semantic components whose vocabularies are disjoint; the *Advanced* scenario introduces vocabulary overlaps (actors/directors), and a semantic class with few manually defined values; the *Hard* scenario introduces a semantic class often subject to omissions, e.g., *Walt Disney Pictures* \rightarrow *disney*, and a class, *Tag*, affected by vocabulary overlaps with the others and vocabulary mismatch between queries and documents. Table 1 reports some statistics regarding the proposed scenarios.

Table 1
Statistics of the proposed scenarios.

	Basic	Advanced	Hard
# train queries	3938	4292	5131
# dev queries	601	672	822
# test queries	538	610	796
Total	5077	5574	6749

4. Conclusion

In this manuscript, we described the building process of a novel benchmark dataset we have recently proposed. We hope our effort can stimulate research for the understudied task of Semantic Query Labeling and encourage other researchers in building datasets for other *not very popular* Information Retrieval tasks that could greatly benefit from the recent advancements in Deep Learning.

References

- [1] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, X. Cheng, A deep look into neural ranking models for information retrieval, *Inf. Process. Manag.* 57 (2020) 102067.
- [2] B. Mitra, N. Craswell, Neural models for information retrieval, *CoRR* abs/1705.01509 (2017).
- [3] Z. Yang, Q. Lan, J. Guo, Y. Fan, X. Zhu, Y. Lan, Y. Wang, X. Cheng, A deep top-k relevance matching model for ad-hoc retrieval, in: *Information Retrieval - 24th China Conference, CCIR 2018, Guilin, China, September 27-29, 2018, Proceedings*, volume 11168 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 16–27.
- [4] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, X. Cheng, Text matching as image recognition, in: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, February 12-17, 2016, Phoenix, Arizona, USA, AAAI Press, 2016, pp. 2793–2799.
- [5] L. Pang, Y. Lan, J. Guo, J. Xu, J. Xu, X. Cheng, Deeprank: A new deep architecture for relevance ranking in information retrieval, in: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, ACM, 2017, pp. 257–266.
- [6] A. Severyn, A. Moschitti, Learning to rank short text pairs with convolutional deep neural networks, in: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, ACM, 2015, pp. 373–382.
- [7] X. Qiu, X. Huang, Convolutional neural tensor network architecture for community-based question answering, in: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, AAAI Press, 2015, pp. 1305–1311.
- [8] Z. Wang, W. Hamza, R. Florian, Bilateral multi-perspective matching for natural language sentences, in: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, ijcai.org, 2017, pp. 4144–4150.
- [9] L. Yang, H. Zamani, Y. Zhang, J. Guo, W. B. Croft, Neural matching models for question retrieval and next question prediction in conversation, *CoRR* abs/1707.05409 (2017).
- [10] R. Yan, D. Zhao, W. E, Joint learning of response ranking and next utterance suggestion in human-computer conversation system, in: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, ACM, 2017, pp. 685–694.
- [11] E. M. Voorhees, D. M. Tice, Building a question answering test collection, in: *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2000, Athens, Greece*, ACM, 2000, pp. 200–207.
- [12] Y. Yang, W. Yih, C. Meek, Wikiqa: A challenge dataset for open-domain question answering, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, The Association for Computational Linguistics, 2015, pp. 2013–2018.

- [13] M. Keikha, J. H. Park, W. B. Croft, Evaluating answer passages using summarization measures, in: The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014, ACM, 2014, pp. 963–966.
- [14] M. Feng, B. Xiang, M. R. Glass, L. Wang, B. Zhou, Applying deep learning to answer selection: A study and an open task, in: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015, Scottsdale, AZ, USA, December 13-17, 2015, IEEE, 2015, pp. 813–820.
- [15] D. Cohen, L. Yang, W. B. Croft, Wikipassageqa: A benchmark collection for research on non-factoid answer passage retrieval, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, ACM, 2018, pp. 1165–1168.
- [16] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A human generated machine reading comprehension dataset, in: Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016, volume 1773 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016.
- [17] P. Nakov, D. Hoogeveen, L. Màrquez, A. Moschitti, H. Mubarak, T. Baldwin, K. Verspoor, Semeval-2017 task 3: Community question answering, CoRR abs/1912.00730 (2019).
- [18] D. Hoogeveen, K. M. Verspoor, T. Baldwin, Cqadupstack: A benchmark data set for community question-answering research, in: Proceedings of the 20th Australasian Document Computing Symposium, ADCS 2015, Parramatta, NSW, Australia, December 8-9, 2015, ACM, 2015, pp. 3:1–3:8.
- [19] A. Abujabal, R. S. Roy, M. Yahya, G. Weikum, Comqa: A community-sourced dataset for complex factoid question answering with paraphrase clusters, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 307–317.
- [20] X. Liu, C. Wang, Y. Leng, C. Zhai, Linkso: a dataset for learning to retrieve similar question answer pairs on software development forums, in: Proceedings of the 4th ACM SIGSOFT International Workshop on NLP for Software Engineering, NL4SE@ESEC/SIGSOFT FSE 2018, Lake Buena Vista, FL, USA, November 4, 2018, ACM, 2018, pp. 2–5.
- [21] M. Manshadi, X. Li, Semantic tagging of web search queries, in: ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, The Association for Computer Linguistics, 2009, pp. 861–869.
- [22] G. W. Furnas, T. K. Landauer, L. M. Gomez, S. T. Dumais, The vocabulary problem in human-system communication, *Commun. ACM* 30 (1987) 964–971.
- [23] X. Li, Y. Wang, A. Acero, Extracting structured information from user queries with semi-supervised conditional random fields, in: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009, ACM, 2009, pp. 572–579.

- [24] X. Li, Understanding the semantic structure of noun phrase queries, in: ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, The Association for Computer Linguistics, 2010, pp. 1337–1345.
- [25] N. Sarkas, S. Pappas, P. Tsapras, Structured annotations of web queries, in: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 2010, pp. 771–782.
- [26] J. Liu, X. Li, A. Acero, Y. Wang, Lexicon modeling for query understanding, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic, IEEE, 2011, pp. 5604–5607.
- [27] Z. Kozareva, Q. Li, K. Zhai, W. Guo, Recognizing salient entities in shopping queries, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers, The Association for Computer Linguistics, 2016.
- [28] E. Bassani, G. Pasi, Semantic query labeling through synthetic query generation, in: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 2278–2282.
- [29] G. Pass, A. Chowdhury, C. Torgeson, A picture of search, in: Proceedings of the 1st International Conference on Scalable Information Systems, Infoscale 2006, Hong Kong, May 30-June 1, 2006, volume 152 of *ACM International Conference Proceeding Series*, ACM, 2006, p. 1.
- [30] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 142–147.
- [31] A. Ratnaparkh, Maximum entropy models for natural language ambiguity resolution, in: Ph.D. Dissertation in Computer and Information Science, University of Pennsylvania, 1998.
- [32] E. F. T. K. Sang, J. Veenstra, Representing text chunks, in: EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway, The Association for Computer Linguistics, 1999, pp. 173–179.