Integrating Query Interpretation Components into the Information Retrieval Experiment Platform

Marcel Gohsen, Benno Stein

Bauhaus-Universität Weimar

Abstract

We describe our contribution of query entity linking and query interpretation software as components of the Information Retrieval Experiment Platform (TIREx) to the 1st International Workshop on Open Web Search (WOWS'24). Query interpretation is the task of determining all plausible user intents behind a search query and can be used to diversify the search results of a retrieval system. As TIREx components, the query entity linking method has identified a total of 89,289 entity candidates in 2,544 queries from 31 standard information retrieval datasets. In addition, a total of 2,304 interpretations of 1,225 search queries from 18 keyword query datasets have been found, which means that on average there is more than one plausible interpretation per search query. This is an indication that most search queries are ambiguous.

Keywords

Query understanding, Query interpretation, Query entity linking, Query segmentation

1. Introduction

The Information Retrieval Experiment Platform (TIREx) [1] integrates `ir_datasets` [2], `ir_measures` [3], PyTerrier [4], and the TIRA Integrated Research Architecture [5] in order to provide an open-source platform to expriment with information retrieval datasets and evaluate retrieval systems in a reproducible fashion. A key aspect of TIREx is that information retrieval collections are static, and therefore query processors only need to be executed once per dataset, so that downstream experiments can utilize the cached and publicly available results instead of executing the processors from scratch.

To streamline further research towards query understanding, we contribute query entity linking and query interpretation components from Kasturia et al. [6] to the TIREx platform. The foundation for interpreting search queries is query segmentation, a method for grouping keywords of a search query into phrases with the aim of maximizing retrieval efficiency when these phrases are matched with the search results. Segmentations of a query form "skeletons" for query interpretations, in which segments are replaced by linked entities when a segment refers to an entity. Each different entity-linked segmentation (i.e., interpretation) represents a different user intent. For example, interpretations of the query "new york times square dance" can be either ⟨`New_York_City` | `Times_Square` | dance⟩ which refers to a dance event on the Times Square in New York City or ⟨`New_York_Times` | `Square_Dance`⟩ which references an article in the New York Times about square dancing.

As demonstrated by the example above, search engine queries can be ambiguous. Automated approaches to interpreting search queries can help a search engine understand a user's intent or diversify search results based on all possible interpretations. Moreover, prior studies have shown that extending queries with (linked) named entities can increase the effectiveness of sparse retrieval [7], entity retrieval [8, 9] and semantic search [10]. Integrating query entity linking and query interpretation components into TIREx facilitates further research in these directions.

In this paper, we describe the query entity linking and query interpretation components and report on entity and interpretation statistics of queries from standard information retrieval datasets. As part of these exemplary analytics, we find that on average queries from almost all common information retrieval datasets have more than one plausible interpretation. We also observe that the number of

WOWS'24: 1st International Workshop on Open Web Search, March 28, 2024, Glasgow, Scotland

marcel.gohsen@uni-weimar.de (M. Gohsen); benno.stein@uni-weimar.de (B. Stein)

0000-0002-1020-6745 (M. Gohsen); 0000-0001-9033-2217 (B. Stein)

interpretations of a query does not correlate with the number of relevant documents. Consequently, this opens a direction for future work to determine whether query interpretations as an intermediate step in retrieval would increase the number of relevant search results when a query is ambiguous (i.e., the number of plausible interpretations is greater than one).

2. Entity-based Query Interpretation

The query interpretation approach of Kasturia et al. [6] consists of three main phases: entity linking, query segmentation and a combination phase, in which the linked entities and query segmentations are combined into query interpretations. The result of the query interpretation approach is a list of entity-linked segmentations ranked by a relevance score.

2.1. Query Entity Linking

The entity linking approach tries to find entities for all the $O(n^2)$ potential segments of an n-term query. The entity linking module is based on titles of Wikipedia articles, redirects, and disambiguation pages. Each Wikipedia article represent an entity and its title, all redirect candidates, and disambiguation names from Wikipedia serve as plausible query segments that refer to this entity. The about 13 million distinct key–value pairs (keys are potential query segments and values are lists of entities that can be referred to by this segment) are stored in a RocksDB table[1] for fast exact-match access and in a Lucene index[2] to quickly find imperfect matches.

Entities from perfect and imperfect matches are then ranked by commonness scores (i.e., the likelihood of an entity-mention link). To compute the commonness of a mention-entity pair, a Wikipedia dump in combination with the computation methodology from Ferragina and Scaiella [11] is used.

Table 1 displays the identified and linked entities in two queries from TREC Web Track 2009 [12] and 2012 [13] datasets. The top four entities linked to segments in the query "obama family tree" are highly relevant and are assessed as such with commonness scores greater than 0.29. Less relevant entities like the TV series or the music album "Family Tree" have received scores of 0.07 and less. In contrast, scores of entities in the query "pork tenderloin" seem to be less informative. While Tenderloin in San Francisco can still be part of a meaningful interpretation (i.e., the user is looking for pork in Tenderloin, San Francisco), the other entities have a rather weak connection to pork. However, these entities have been awarded a comparatively low score of 0.12 or less.

2.2. Query Segmentation

Query segmentation methods aim to rank the possible 2^n valid segmentations of an n-term query according to retrieval effectiveness when the segments are treated as phrases to be matched in the search results. The query interpretation approach employs query segmentation approaches from Hagen et al. [14, 15] due to the simplicity of the approaches and the associated lower runtime. These query segmentation approaches rank the possible segmentations of a query by summing up pre-computed segment weights stored in a hash table for quick access. The segment weights are occurence frequencies from the Google n-gram corpus[3] in case a segment does not represent a title of a Wikipedia article. In case this segment is the title of a Wikipedia article, the segment weight is 1 + occurrence frequency of the most frequent word-2-gram in that segment.

Most of the 2^n segmentations of an n-term query do not yield plausible interpretations. Kasturia et al. [6] have shown that often only the segmentation with the highest score is used as an interpretation skeleton and that segmentations with lower scores, which show large differences to segmentations with higher scores, are rarely used as skeletons. Therefore, respective filter heuristics are applied to not forward all segmentations to the combination phase. A first filter removes segmentations whose

[1] https://rocksdb.org/
[2] https://lucene.apache.org/
[3] https://catalog.ldc.upenn.edu/LDC2006T13

Table 1
Linked entities from Wikipedia and their associated mention segments and commonness scores for two example queries from TREC Web Track 2009 and 2012 datasets. Entities with a score of less than 0.05 have been excluded.

Rank	Mention	Entity	Score
For query "obama family tree"			
1	family	`Family_(biology)`	0.88
2	obama	`Barack_Obama`	0.63
3	family tree	`Family_Tree`	0.37
4	obama family	`Family_of_Barack_Obama`	0.29
5	family tree	`Family_Tree_(TV_series)`	0.07
6	family tree	`Family_Tree_(Nick_Drake_album)`	0.06
For query "pork tenderloin"			
1	port tenderloin	`Pork_tenderloin`	0.92
2	tenderloin	`Tenderloin,_San_Francisco`	0.38
3	tenderloin	`Beef_tenderloin`	0.18
4	tenderloin	`Tenderloin_(musical)`	0.12
5	tenderloin	`Tenderloin_(film)`	0.09
6	tenderloin	`Tenderloin,_Manhattan`	0.05

highest weighting segment is contained in a higher-ranked segmentation. A second filter removes segmentations for which the score ratio to the lowest kept higher-ranked segmentation falls below a threshold of 0.66.

2.3. Query Interpretation

To build query interpretations for a query, the approach combines segmentations with linked entities. To "fill" the segmentation skeletons forwarded by the segmentation phase, the approach collects the entities ranked by commonness for each segment (discarding entities with a commonness of 0) and adds the option of not linking a segment to an entity, but keeping it as a phrase as a fallback. The potential interpretations can then be derived by a Cartesian product of the not-0-common entities and the unlinked respective segments.

In the interpretation ranking, three weights from the entity linking literature are combined: (1) the above described commonness CMN, (2) the likelihood of two entities to occur together (relatedness REL), and (3) the likelihood of an entity to occur with the unlinked segments (context CXT). The relatedness and context weights computations are calculated by using Wikipedia-based joint word–entity embeddings[4] provided by Yamada et al. [16]. The by the authors suggested configuration have been used: average cosine similarity of an entity's embedding vector with the other entities in an interpretation (relatedness) or with the unlinked segments in an interpretation (context).

An interpretation I's score is the averaged weighted sum of the commonness, relatedness, and context scores of the entities $e \in I$ with the weights $\alpha = \beta = \gamma = 1$ suggested by Kasturia et al. [6]:

$$score(I) = \frac{1}{|\{e \in I\}|} \cdot \sum_{e \in I} \left(\alpha \cdot \mathrm{CMN}(e) + \beta \cdot \mathrm{REL}(e) + \gamma \cdot \mathrm{CXT}(e)\right),$$

Table 2 presents the found query interpretations for the example queries from Table 1 and an additional query from TREC Web Track 2012. For the query "obama family tree" the most plausible interpretations have been identified, which is that a user is looking up the family tree of Barack Obama. These interpretations are on rank 1 and rank 2 and can be seen as equivalent because the concept of a family tree is semantically equivalent to the linked entity `Family_Tree`. The interpretation on rank 3 assumes that the segment "obama" is a concept and not an entity which is wrong, but the assigned relevance is desirably low. For the query "pork tenderloin" the only found interpretation is the dish pork tenderloin. Desirabel (but less likely) would also be the interpretation $\langle$pork | `Tenderloin,_San_Francisco`$\rangle$ as

[4] https://wikipedia2vec.github.io/wikipedia2vec/

Table 2
Entity-based query interpretations and their relevance scores for three example queries from TREC Web Track 2009 and 2012 datasets.

Rank	Interpretation	Score
For query "obama family tree"		
1	⟨Barack_Obama | family tree⟩	0.77
2	⟨Barack_Obama | Family_Tree⟩	0.50
3	⟨obama | Family_Tree⟩	0.37
For query "pork tenderloin"		
1	⟨Pork_tenderloin⟩	0.92
For query "last supper painting"		
1	⟨Last_Supper | painting⟩	1.50
2	⟨last supper | Painting⟩	1.16
3	⟨Last_Supper | Painting⟩	1.12

someone might look for meat vendors in Tenderloin in San Francisco. However, the system did not provide this plausible interpretation. Another interesting example is the query "last supper painting". The system identifies three plausible interpretations which translate to a user who is interested in (1) the painting process of the painting "Last Supper" by Leonardo da Vinci, (2) any painting that depicts the last supper, and (3) information about "Last Supper" by Leonardo da Vinci. All three interpretations are highly likely which is reflected in the assigned relevance scores.

2.4. Integration into TIREx

Both, the query entity linking and the query interpretation software, are Java-based implementations that require external data (e.g., RocksDB and Lucene Index for entity linking) to function as intended. To integrate these components into TIREx and ensure reproducible execution across all systems, we dockerize the query interpretation and entity linking software and bundle the image with all necessary external data. Both Docker images are available in our public container registry[5] to use outside TIREx. The use in combination with TIREx is documented in examplary Jupyter notebooks in the code repositories of query entity linking[6] and query interpretation[7] components.

3. Query Analytics

To gain insights into the identified entities and interpretations, we perform a statistical analysis of the entity and interpretation frequencies of queries from common information retrieval datasets available in TIREx.

3.1. Datasets

TIREx accesses datasets through the Python package of `ir_datasets` [2]. Therefore, all the processed datasets are well documented in the online catalog[8] of `ir_datasets`. Most of the 31 considered datasets originate from shared tasks like TREC or Touché. We distinguish these datasets by query type which are *keyword queries* (e.g., "french lick resort and casion" from TREC Web Track 2009 [12]) and *natural language queries* which can either be a question (e.g., "Are gas prices too high?" from the Touche 2021 Argument Retrieval task [17]) or a description of an information need (e.g., "Provide information about the genes Ret and GDNF in kidney development." from TREC 2005 Genomics Track [18]). The

[5] `registry.webis.de/code-lib/public-images/query-interpretation:1.0`
`registry.webis.de/code-lib/public-images/query-entity-linking:1.0`
[6] https://github.com/webis-de/query-entity-linking
[7] https://github.com/webis-de/query-interpretation
[8] https://ir-datasets.com/

Table 3
Statistics about the number of queries, number of queries with at least one entity, average number of terms, average number of entity mentions per query, and the total number of entity candidates suggested by the query entity linker for each dataset in TIREx. The first column of this table represents the dataset identifier of each respective dataset as specified in `ir_datasets`.

Dataset	Ref.	Queries			Entities	
		Count	w. Entities	Terms	Mentions	Count
Keyword Queries						
clueweb09/en/trec-web-2009	[12]	50	45 (90%)	2.1	1.8	891
clueweb09/en/trec-web-2010	[21]	50	43 (86%)	2.1	1.5	1,031
clueweb09/en/trec-web-2011	[22]	50	46 (92%)	3.4	2.8	1,309
clueweb09/en/trec-web-2012	[13]	50	45 (90%)	2.3	1.9	1,038
clueweb12/trec-web-2013	[23]	50	49 (98%)	3.3	3.0	1,717
clueweb12/trec-web-2014	[24]	50	50 (100%)	3.3	3.2	1,375
cord19/fulltext/trec-covid	[25, 26]	50	50 (100%)	3.2	2.4	744
disks45/nocr/trec-robust-2004	[27, 28, 29]	250	249 (99%)	2.7	2.7	5,626
disks45/nocr/trec7	[29, 30]	50	50 (100%)	2.4	2.7	1,087
disks45/nocr/trec8	[29, 31]	50	50 (100%)	2.4	2.6	970
gov/trec-web-2002	[32]	50	50 (100%)	3.2	3.0	1,459
gov/trec-web-2003	[33]	50	48 (96%)	2.2	2.3	677
gov/trec-web-2004	[34]	225	221 (98%)	3.4	3.2	7,897
gov2/trec-tb-2004	[35]	50	49 (98%)	3.2	2.8	1,302
gov2/trec-tb-2005	[36]	50	50 (100%)	3.1	2.8	1,166
gov2/trec-tb-2006	[37]	50	48 (96%)	3.0	3.1	1,544
nfcorpus/test	[38]	325	316 (97%)	3.5	2.6	5,528
wapo/v2/trec-core-2018	-	50	49 (98%)	3.1	2.9	1,296
Natural Language Queries						
antique/test	[39]	200	196 (98%)	9.3	4.7	11,298
argsme/2020-04-01/touche-2020-task-1	[40, 41]	49	48 (98%)	6.6	3.9	1,202
argsme/2020-04-01/touche-2021-task-1	[17]	50	49 (98%)	5.4	3.1	1,293
clueweb12/touche-2020-task-2	[40, 42, 43]	50	50 (100%)	8.4	4.8	2,370
clueweb12/touche-2021-task-2	[17]	50	50 (100%)	8.4	4.8	2,232
cranfield	-	225	225 (100%)	18.0	9.0	18,301
medline/2004/trec-genomics-2004	[44]	50	47 (94%)	4.9	3.4	1,341
medline/2004/trec-genomics-2005	[18]	50	50 (100%)	16.2	8.9	2,980
msmarco-passage/trec-dl-2019/judged	[45, 46]	43	42 (97%)	5.4	3.1	985
msmarco-passage/trec-dl-2020/judged	[47, 46]	54	53 (98%)	6.0	3.5	1,609
vaswani	-	93	93 (100%)	10.9	6.5	5,756
Other						
medline/2017/trec-pm-2017	[19]	30	30 (100%)	7.5	7.7	1,502
medline/2017/trec-pm-2018	[20]	50	50 (100%)	5.8	5.4	1,763
Sum		2,544	2,491 (98%)	-	-	89,289

datasets of the TREC Precision Medicine Track of 2017 [19] and 2018 [20] do not fit into either of these query type categories since the queries from these datasets are compounds of a disease name, a gene, and a demographic (e.g., "melanoma BRAF (V600E) 64-year-old male").

3.2. Analytics of Entities in Queries

Table 3 presents query and entity statistics from 31 different information retrieval datasets. All datasets contain between 30 (TREC Precision Medicine Track 2017 dataset) and 325 (NFCorpus test collection) queries of which almost all queries contain at least one entity according to the query entity linking approach. Across all datasets, almost 98% of the queries contain at least one entity. The highest ratio of queries without any entity has the dataset of TREC Web Track 2010. The seven queries from that dataset

Table 4
Statistics about the number of queries, average number of query terms, average number of interpretations per query, total number of interpretations identified by the query interpretation component, Spearman ρ correlation between the number of interpretations and the number of relevant documents (Corr. with Int.) and the total number of relevant documents for keyword query datasets in TIREx. The first column of this table represents the dataset identifier of a respective dataset as specified in `ir_datasets`.

Dataset	Ref	Queries		Interpretations		Rel. Documents	
		Count	Terms	per Query	Count	Corr. with Int.	Count
Keyword Queries							
clueweb09/en/trec-web-2009	[12]	50	2.1	1.4	72	-0.17	6,858
clueweb09/en/trec-web-2010	[21]	50	2.1	1.2	61	-0.02	5,233
clueweb09/en/trec-web-2011	[22]	50	3.4	2.2	108	0.03	3,157
clueweb09/en/trec-web-2012	[13]	50	2.3	1.4	69	-0.13	3,523
clueweb12/trec-web-2013	[23]	50	3.3	2.0	101	0.03	4,150
clueweb12/trec-web-2014	[24]	50	3.3	2.0	98	-0.04	5,665
cord19/fulltext/trec-covid	[25, 26]	50	3.2	2.2	112	0.01	26,664
disks45/nocr/trec-robust-2004	[27, 28, 29]	250	2.7	1.8	453	-0.04	17,412
disks45/nocr/trec7	[29, 30]	50	2.4	1.8	92	0.13	4,674
disks45/nocr/trec8	[29, 31]	50	2.4	1.5	77	0.02	4,728
gov/trec-web-2002	[32]	50	3.2	2.1	104	0.10	1,574
gov/trec-web-2003	[33]	50	2.2	1.4	72	-0.11	516
gov/trec-web-2004	[34]	225	3.4	2.2	489	-0.35	1,763
gov2/trec-tb-2004	[35]	50	3.2	2.0	99	0.03	10,617
gov2/trec-tb-2005	[36]	50	3.1	1.9	93	-0.13	10,407
gov2/trec-tb-2006	[37]	50	3.0	2.3	114	0.10	5,893
nfcorpus/test	[38]	325	3.5	1.9	628	0.02	15,820
wapo/v2/trec-core-2018	-	50	3.1	1.8	90	-0.03	3,948
Sum		1,225	-	-	2,304	-	116,782

for which no entity have been identified are “horse hooves”, “iron”, “vldl levels”, “kiwi”, “tornadoes”, “raised gardens”, and “ocd”. Except for Obessive-compulsive disorder (OCD), no other likely entity would be expected in these queries. Although there are songs called “kiwi” or “iron”, neither seems to be common enough to be included as a plausible entity in an interpretation of these queries.

A fairly obvious observation is that natural language queries are on average much longer than keyword queries. While keyword queries consist of approximately three query terms, the average length of natural language queries vary between 5 and 18 terms depending on the dataset. Intuitively, the number of mentions of entities correlates with the length of the query. We calculated a Spearman’s $\rho = 0.63$ between these two values over all 2,544 queries, which means that the number of plausible entities increases monotonically as the length of a query increases. The query with the most entity mentions is “is it possible to determine rates of forced convective heat transfer from heated cylinders of non-circular cross-section, (the fluid flow being along the generators)” from the Cranfield dataset. The query entity linking approach identified 22 different segments that might refer to entities in that query.

3.3. Analytics of Query Interpretations

The query interpretation approach depends on query segmentation, which requires keyword queries and does not scale well for long natural language queries. We therefore only report on query interpretation statistics for keyword query datasets. Table 4 presents query interpretation statistics for 18 keyword query datasets in TIREx. As most statistics are influenced by them, the number of queries and the query length are copied from Table 3 to facilitate comparisons.

Overall, the query interpretation approach has identified 2,304 plausible interpretations from 1,225 analyzed queries. Therefore, queries across all datasets have an average of about 1.8 plausible interpretations using the query interpretation approach, with queries from the TREC Terabyte Track 2006

dataset having the most interpretations of 2.3. One of two queries which have the most interpretations is "how has african american music influence history" from the TREC Web Track 2014 dataset with 20 different interpretations. The highest ranked intepretations with a score close to one are ⟨how | has | `African-American` | `Music` | influence | `History`⟩ and equivalent variations of that in which the entities `African-American`, `Music`, and `History` have not been linked, and thus have been considered concepts. Since the concepts music and history are equivalent to their linked entities, the interpretation variations can be considered semantically equivalent. Another interesting interpretation of that query is ⟨how | has | `African-American_Music` | influence | `History`⟩ which is probably a more relevant interpretation because it correctly identifies the relation between "african-american" and "music". Unfortunately, this interpretation has been ranked lower.

Ideally, the more ambiguous a query is (i.e., the more plausible interpretations it has), the more documents that fulfill all the different information needs become relevant. To analyze if this is given in the datasets, we compute correlation coefficients as Spearman's ρ between the number of automatically identified interpretations and the number of relevant documents (relevance > 0) in the datasets. We have found no correlation ($\rho \approx 0$) for most datasets or in the worst case a negative correlation ($\rho = -0.35$). This points into an interesting future direction to use query interpretation as an intermeditate step to diversify search result to increase this correlation. A related research question would be to analyze whether each different interpretation results in a similar number of relevant documents, which would lead to a linear growth of relevant documents as the number of interpretations increases.

4. Conclusions

To summarize, we contributed query entity linking and query interpretation components to TIREx. A total of 89,289 Wikipedia-linked entities and 2,304 segmentation-based interpretations were automatically identified, which can be reproducibly used for future research with the help of TIREx. As part of a preliminary analysis of the query interpretations, we found that the number of relevant documents for a query does not correlate with the number of plausible interpretations. This fact points in the direction of future research in which query interpretation can be used to diversify search results and what other effects query interpretation has as an intermediate step on an information retrieval pipeline.

5. Limitations

Although working with a static entity collection and precomputed commonness scores brings the advantage of reproducibility, entities that are relevant now may not be as relevant in the future. Therefore, the identified linked entities included in our query interpretations may become outdated, or their relevance score may become inadequate. A method to automatically update the entity index and the associated commonness scores from an up-to-date knowledge base can compensate for this limitation and will be implemented in the future.

Wikipedia is a well maintained knowledge base for general knowledge. However, entities from specialized areas may be inadequately represented or simply not exist on Wikipedia, and consequently the entity linking method will fail to identify these entities. A mechanism for exchanging (or extending) the knowledge base that is used for query entity linking can help to find more relevant entities. For example, the addition of a knowledge base such as PubMed[9] could increase the discoverability of medical-related entities. We aim to implement an easy method to modify and extend the knowledge base for entity linking and the interpretation of queries.

[9] http://er.tacc.utexas.edu/datasets/ped

References

[1] M. Fröbe, J. H. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, The Information Retrieval Experiment Platform, in: H.-H. Chen, W.-J. E. Duh, H.-H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023), ACM, 2023, pp. 2826–2836.

[2] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, N. Goharian, Simplified Data Wrangling with ir_datasets, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Virtual Event Canada, 2021, pp. 2429–2436.

[3] S. MacAvaney, C. Macdonald, I. Ounis, Streamlining Evaluation with ir-measures, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, volume 13186, Springer International Publishing, Cham, 2022, pp. 305–310.

[4] C. Macdonald, N. Tonellotto, S. MacAvaney, I. Ounis, PyTerrier: Declarative Experimentation in Python from BM25 to Dense Retrieval, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, ACM, Virtual Event Queensland Australia, 2021, pp. 4526–4533.

[5] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous integration for reproducible shared tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241.

[6] V. Kasturia, M. Gohsen, M. Hagen, Query Interpretations from Entity-Linked Segmentations, in: 15th ACM International Conference on Web Search and Data Mining (WSDM 2022), ACM, 2022.

[7] D. Shehata, N. Arabzadeh, C. L. A. Clarke, Early Stage Sparse Retrieval with Entity Linking, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, ACM, Atlanta GA USA, 2022, pp. 4464–4469.

[8] F. Hasibi, K. Balog, S. E. Bratsberg, Exploiting Entity Linking in Queries for Entity Retrieval, in: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval, ACM, Newark Delaware USA, 2016, pp. 209–218.

[9] S. Chatterjee, L. Dietz, Entity Retrieval Using Fine-Grained Entity Aspects, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Virtual Event Canada, 2021, pp. 1662–1666.

[10] M. L. Sidi, S. Gunal, A Purely Entity-Based Semantic Search Approach for Document Retrieval, Applied Sciences 13 (2023) 10285.

[11] P. Ferragina, U. Scaiella, TAGME: On-the-fly annotation of short text fragments (by wikipedia entities), in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, Toronto ON Canada, 2010, pp. 1625–1628.

[12] C. Clarke, N. Craswell, I. Soboroff, Overview of the TREC 2009 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of the Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009, volume 500–278 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2009.

[13] C. Clarke, N. Craswell, E. M. Voorhees, Overview of the TREC 2012 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of the Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012, volume 500–298 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2012.

[14] M. Hagen, M. Potthast, B. Stein, C. Bräutigam, Query segmentation revisited, in: S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, R. Kumar (Eds.), 20th International Conference on World Wide Web (WWW 2011), ACM, 2011, pp. 97–106.

[15] M. Hagen, M. Potthast, A. Beyer, B. Stein, Towards optimum query segmentation: In doubt without, in: X. Chen, G. Lebanon, H. Wang, M. J. Zaki (Eds.), 21st ACM International Conference on Information and Knowledge Management (CIKM 2012), ACM, 2012, pp. 1015–1024.

[16] I. Yamada, H. Shindo, H. Takeda, Y. Takefuji, Joint Learning of the Embedding of Words and

Entities for Named Entity Disambiguation, in: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 250–259.
[17] A. Bondarenko, L. Gienapp, M. Fröbe, M. Beloucif, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of touché 2021: Argument retrieval, in: K. S. Candan, B. Ionescu, L. Goeuriot, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 12th International Conference of the CLEF Association (CLEF 2021), volume 12880 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2021, pp. 450–467.
[18] W. Hersh, A. Cohen, J. Yang, P. Roberts, M. Hearst, TREC 2005 Genomics Track Overview, in: Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005), 2005.
[19] K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bedrick, A. J. Lazar, S. Pant, Overview of the TREC 2017 Precision Medicine Track, TREC 26 (2017) https://trec.nist.gov/pubs/trec26/papers/Overview–PM.pdf.
[20] K. Roberts, D. Demner-Fushman, E. Vorhees, W. R. Hersh, S. Bedrick, A. J. Lazar, Overview of the TREC 2018 Precision Medicine Track, TREC (2018).
[21] C. Clarke, N. Craswell, I. Soboroff, G. V. Cormack, Overview of the TREC 2010 web track, in: TREC, 2010.
[22] C. Clarke, N. Craswell, I. Soboroff, E. M. Voorhees, Overview of the TREC 2011 web track, in: TREC, 2011.
[23] K. Collins-Thompson, P. Bennett, F. Diaz, C. L. A. Clarke, E. M. Voorhees, TREC 2013 web track overview, in: TREC, 2013.
[24] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, E. M. Voorhees, TREC 2014 web track overview, in: TREC, 2014.
[25] E. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. Hersh, K. Lo, K. Roberts, I. Soboroff, L. L. Wang, TREC-COVID: Constructing a pandemic information retrieval test collection, ArXiv abs/2005.04474 (2020).
[26] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. M. Kinney, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. Wade, K. Wang, C. Wilhelm, B. Xie, D. Raymond, D. S. Weld, O. Etzioni, S. Kohlmeier, CORD-19: The covid-19 open research dataset, ArXiv (2020).
[27] S. Huston, W. B. Croft, A comparison of retrieval models using term dependencies, in: CIKM, 2014.
[28] E. Voorhees, Overview of the TREC 2004 robust retrieval track, in: TREC, 2004.
[29] E. M. Voorhees, NIST TREC disks 4 and 5: Retrieval test collections document set, 1996.
[30] E. M. Voorhees, D. Harman, Overview of the seventh text retrieval conference (TREC-7), in: TREC, 1998.
[31] E. M. Voorhees, D. Harman, Overview of the eight text retrieval conference (TREC-8), in: TREC, 1999.
[32] N. Craswell, D. Hawking, Overview of the TREC-2002 web track, in: TREC, 2002.
[33] N. Craswell, D. Hawking, R. Wilkinson, M. Wu, Overview of the TREC 2003 web track, in: TREC, 2003.
[34] N. Craswell, D. Hawking, Overview of the TREC-2004 web track, in: TREC, 2004.
[35] C. Clarke, N. Craswell, I. Soboroff, Overview of the TREC 2004 terabyte track, in: TREC, 2004.
[36] C. L. A. Clarke, F. Scholer, I. Soboroff, The TREC 2005 terabyte track, in: TREC, 2005.
[37] S. Büttcher, C. L. A. Clarke, I. Soboroff, The TREC 2006 terabyte track, in: TREC, 2006.
[38] V. Boteva, D. Gholipour, A. Sokolov, S. Riezler, A full-text learning to rank dataset for medical information retrieval, in: Proceedings of the European Conference on Information Retrieval (ECIR), Springer, Padova, Italy, 2016.
[39] H. Hashemi, M. Aliannejadi, H. Zamani, B. Croft, ANTIQUE: A non-factoid question answering benchmark, in: ECIR, 2020.
[40] A. Bondarenko, M. Fröbe, M. Beloucif, L. Gienapp, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein,

H. Wachsmuth, M. Potthast, M. Hagen, Overview of touché 2020: Argument retrieval, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 11th International Conference of the CLEF Association (CLEF 2020), volume 12260 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2020, pp. 384–395.

[41] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, B. Stein, Computational argumentation quality assessment in natural language, in: P. Blunsom, A. Koller, M. Lapata (Eds.), 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017), 2017, pp. 176–187.

[42] L. Braunstain, O. Kurland, D. Carmel, I. Szpektor, A. Shtok, Supporting human answers for advice-seeking questions in CQA sites, in: N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. D. Nunzio, C. Hauff, G. Silvello (Eds.), Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings, volume 9626 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 129–141.

[43] M. Rafalak, K. Abramczuk, A. Wierzbicki, Incredible: Is (almost) all web content trustworthy? analysis of psychological factors related to website credibility evaluation, in: C.-W. Chung, A. Z. Broder, K. Shim, T. Suel (Eds.), 23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume, ACM, 2014, pp. 1117–1122.

[44] W. R. Hersh, R. T. Bhuptiraju, L. Ross, P. Johnson, A. M. Cohen, D. F. Kraemer, TREC 2004 genomics track overview, in: TREC, 2004.

[45] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, E. Voorhees, Overview of the TREC 2019 deep learning track, in: TREC 2019, 2019.

[46] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, 2018. arXiv:1611.09268.

[47] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, Overview of the TREC 2020 deep learning track, in: TREC, 2020.