A Test Collection for Entity Search in DBpedia

Krisztian Balog
University of Stavanger
krisztian.balog@uis.no

Robert Neumayer
NTNU Trondheim
robert.neumayer@idi.ntnu.no

ABSTRACT

We develop and make publicly available an entity search test collection based on the DBpedia knowledge base. This includes a large number of queries and corresponding relevance judgments from previous benchmarking campaigns, covering a broad range of information needs, ranging from short keyword queries to natural language questions. Further, we present baseline results for this collection with a set of retrieval models based on language modeling and BM25. Finally, we perform an initial analysis to shed light on certain characteristics that make this data set particularly challenging.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

Keywords

Entity retrieval, test collections, semantic search, DBpedia

1. INTRODUCTION

Many information needs revolve around entities as has been observed in different application domains, including question answering [14, 21], enterprise [1], and web [19] search. This is reflected by the recent emergence of a series of benchmarking campaigns focusing on entity retrieval evaluation in various settings. The INEX 2007-2009 Entity Retrieval track [8, 9] studies entity retrieval in Wikipedia. The Linked Data track at INEX 2012 also considers entities from Wikipedia, but articles are enriched with RDF properties from both DBpedia and YAGO2 [22]. The TREC 2009-2011 Entity track [1, 3] defines the related entity finding task: return homepages of entities, of a specified type, that engage in a specified relationship with a given source entity. In 2010, the Semantic Search Challenge introduced a platform for evaluating ad-hoc queries, targeting a particular entity, over a diverse collection of Linked Data sources [11]. The 2011 edition of the challenge presented a second task, list search, with more complex queries [4]. Finally, the Question Answering over Linked Data challenge focuses on natural language question-answering over selected RDF datasets, DBpedia and MusicBrainz [14].

Finding new challenges and tasks for entity search was one of the main topics of discussion at the recently held 1st Joint International Workshop on Entity-oriented and Semantic Search (JIWES) [2]. The following action points were identified as important priorities for future research and development:

(A1) Getting more representative information needs and favouring long queries over short ones.

(A2) Limiting search to a smaller, fixed set of entity types (as opposed to arbitrary types of entities).

(A3) Using test collections that integrate both structured and unstructured information about entities.

In this paper we address the above issues by proposing an entity search test collection based on DBpedia. We synthesise queries from all these previous benchmarking efforts into a single query set and map known relevant answers to DBpedia. This results in a diverse query set ranging from short keyword queries to natural language questions, thereby addressing (A1). DBpedia has a consistent ontology comprising of 320 classes, organised into a 6 levels deep hierarchy; cf. (A2). Finally, as DBpedia is extracted from Wikipedia, there is more textual content available for those who wish to combine structured and unstructured information about entities, thereby addressing (A3).

On top of all these, there is one more important, yet still open question: To what extent can methods developed for a particular test set be applied to different settings? To help answer this question we evaluate standard document retrieval models (language models and BM25) and some of their fielded extensions. We make the somewhat surprising finding that, albeit frequently used, none of these extensions is able to substantially and significantly outperform the document-based (single-field) models. Our topic-level analysis reveals that while often a large number of topics is helped, an approximately identical number of topics is negatively impacted at the same time. Developing methods that can realise improvements across the whole query set appears to be an open challenge.

Our contributions in this paper are threefold. First, we create and make publicly available a data set for entity retrieval in DBpedia.[1] Second, we evaluate and compare a set of baseline methods on this data set. Third, we perform a topic-level analysis and point out certain characteristics that make this data set particularly challenging.

The remainder of this paper is organised as follows. In Section 2 we introduce our test collection. Next, in Section 3 we present and evaluate baseline methods. This is followed by a topic-level analysis in Section 4. We summarise our findings in Section 5.

[1] http://bit.ly/dbpedia-entity

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

2. TEST COLLECTION

We consider a range of queries from various benchmarking evaluation campaigns and attempt to answer them using a large knowledge base. In our case this knowledge base is DBpedia, as described in Section 2.1. Further, we describe both queries and relevance judgements in Section 2.2. To conclude the description of the test collection, we give an overview of the evaluation metrics we use in Section 2.3

2.1 Knowledge base

We use DBpedia as our knowledge base, specifically, version 3.7. DBpedia has—apart from being one of the most comprehensive knowledge bases on the web—the advantage of using a consistent ontology to classify many of its entities via a *type* predicate. The ontology defines 320 classes, organised into a 6 levels deep hierarchy. This version of DBpedia describes more than 3.64M entities, of which 1.83M are classified in the DBpedia ontology.

2.2 Queries and relevance assessments

We consider queries from the following benchmarking evaluation campaigns (presented in temporal order):

- **INEX-XER**: The INEX 2009 Entity Ranking track seeks a list of entities (e.g., "US presidents since 1960"), where entities are represented by their Wikipedia page [9]. We map Wikipedia articles to the corresponding DBpedia entry.
- **TREC Entity**: The related entity finding task at the TREC 2009 Entity track focuses on specific relationships between entities (e.g., "Airlines that currently use Boeing 747 planes") and requests entity homepages from a Web corpus to be retrieved [1]. The Wikipedia page of the entity may also be returned in the answer record; we mapped these to the corresponding DBpedia entry.[2] We use 17 out of the original 20 queries as for the remaining 3 queries there are no relevant results from DBpedia.
- **SemSearch ES**: Queries in the ad-hoc entity search task at the 2010 and 2011 Semantic Search Challenge refer to one particular entity, albeit often an ambiguous one (e.g., "Ben Franklin," which is both a person and a ship), by means of short keyword queries. The collection is a sizeable crawl of Semantic Web data (BTC-2009) [4, 11]. DBpedia is part of this crawl, in fact, 59% of the relevant results originate from DBpedia. 130 queries (out of the total of 142) have relevant results from DBpedia.
- **SemSearch LS**: Using the same data collection as the ES task, the list search task at the 2011 Semantic Search Challenge targets a group of entities that match certain criteria (e.g., "Axis powers of World War II") [4]. Out of the original 50 queries, 43 have results from DBpedia.
- **QALD-2**: The Question Answering over Linked Data challenge aims to answer natural language questions (e.g., "Who is the mayor of Berlin?") using Linked Data sources [14]. We used the query sets that were developed for DBpedia, and collapsed both training (100) and testing (100) queries into a single set. We filtered out queries where answers are not DBpedia pages (for example, "How many students does the Free University in Amsterdam have?" where the answer is a number). This leaves us with 140 queries in total.

[2]In the 2010 edition, Wikipedia pages are not accepted as entity homepages, therefore, those results cannot be mapped to DBpedia with reasonable effort. We did not include REF 2011 queries as the quality of the pools there is found to be unsatisfactory [3].

Table 1: Queries used for experimental evaluation.

Query set	#queries	avg(|q|)	avg(#rel)
INEX-XER	55	5.5	29.8
TREC Entity	17	6.7	13.1
SemSearch ES	130	2.7	8.7
SemSearch LS	43	5.4	12.5
QALD-2	140	7.9	41.5
INEX-LD	100	4.8	37.6
Total	485	5.3	27.0

- **INEX-LD**: The ad-hoc search task at the INEX 2012 Linked Data track uses IR-style keyword queries (e.g., "England football player highest paid") over a collection of Wikipedia articles enriched with RDF properties from both DBpedia and YAGO2 [22]. We mapped relevant Wikipedia pages to DBpedia; all 100 of the original queries were usable .

The selection above covers a broad range of information needs, ranging from short keyword queries to natural language questions. In all cases, we use only the keyword part of the query and ignore any additional markup, type information, or other hints (like example entities) that may be available as part of the topic definition according to the original task setup. Also, we take relevance to be binary, that is, both relevant and primary for the TREC Entity queries, and fair and excellent for SemSearch queries count as correct. We normalised all URIs to conform with the encoding used by the official DBpedia dump, replaced redirect pages with the URIs they redirect to, and filtered out URIs that are not entity pages (e.g., categories or templates). Table 1 provides an overview.

2.3 Evaluation metrics

We use standard IR evaluation metrics: Mean Average Precision (MAP) and Precision at rank 10 (P@10). To check for significant differences between runs, we use a two-tailed paired t-test and write $^{\triangle}/^{\triangledown}$ and $^{\blacktriangle}/^{\blacktriangledown}$ to denote significance at the 0.05 and 0.01 levels, respectively.

3. BASELINE METHODS AND RESULTS

This section presents our baseline methods (Section 3.2), followed by and experimental comparison (Section 3.3). We start out by introducing our experimental setup (Section 3.1).

3.1 Experimental setup

We indexed all entities that have a label (i.e., a "name") but filtered out redirect pages. We considered the top 1000 most frequent predicates as fields; this was done to ensure that all fields occur in sufficiently many entity descriptions. Note that this number is two magnitudes larger than what was considered in prior work (6 in [17] and 11 at most in [12, 13]). We employ a heuristic to identify *title* fields; following [16], attributes names ending in "label", "name," or "title" are considered to hold title values. For each entity, we store a content field, collapsing all its predicates. We kept both relations (i.e., links pointing to other DBpedia pages) and resolved relations (i.e., replacement of the link with the title of the page it points to) in our index.

3.2 Baseline methods

We consider two sets of baseline methods. One is based on language modeling the other is based on BM25. This particular choice

Table 2: Results and baseline comparison. Significance for rows 2-4 is tested against row 1; for rows 6-7 tested against row 5.

Model	INEX-XER		TREC Entity		SemSearch ES		SemSearch LS		QALD-2		INEX-LD		Total	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
LM	.1672	.2618	.0970	.1294	.3139	.2508	.1788	.1907	.1067	.0507	.1057	.2360	.1750	.1816
MLM-tc	.1585	.2345$^{\triangledown}$	.0855$^{\triangledown}$	.1176	.3541$^{\blacktriangle}$	.2838$^{\blacktriangle}$	.1738	.1744	.0989$^{\triangledown}$	.0507	.1044	.2320	.1813$^{\triangle}$	.1847
MLM-all	.1589	.2273	.0641	.0882	.3010	.2454	.1514	.1581	.1204	.0593	.0857$^{\triangledown}$	.1850$^{\blacktriangledown}$	.1668	.1639$^{\blacktriangledown}$
PRMS	.1897$^{\triangle}$	.2855	.1206	.1706	.3228	.2515	.1857	.2093	.1050	.0693$^{\triangle}$	.0840$^{\blacktriangledown}$	.2030$^{\blacktriangledown}$	.1764	.1862
BM25	.1830	.2891	.0882	.1000	.3262	.2562	.1785	.2116	.1184	.0657	.1178	.2470	.1856	.1936
BM25F-tc	.1720$^{\blacktriangledown}$	.2655$^{\blacktriangledown}$	.0848	.0882	.3337$^{\triangle}$	.2631$^{\triangle}$	.1718	.2163	.1067$^{\triangledown}$	.0621	.1169	.2490	.1820$^{\triangledown}$	.1922
BM25F-all	.1810	.2836	.0824$^{\triangledown}$	.0824	.3286	.2585	.1789	.2163	.1189	.0686	.1155	.2470	.1855	.1942

is made because we consider both families of methods state-of-the-art that are frequently applied in the context of various entity search tasks, see, e.g., [5–7, 10, 15, 18]. Here, we confine ourselves to a basic approach where a (fielded) document-based representation is built for each entity. This representation makes limited use of entity-specific features, such as type information and related entities; we leave these to future work.

Specifically, we use the following language modeling based methods: **LM**: the standard query likelihood approach [23]; **MLM-tc**: the Mixture of Language Models [17], with two fields: title and content. Following [16] we set the title weight to 0.2 and the content weight to 0.8; **MLM-all**: the Mixture of Language Models [17], where all fields are considered with equal weight; **PRMS**: the Probabilistic Retrieval Model for Semistructured Data. The difference to MLM-all is that field weights are determined dynamically for each query term [13]. All methods use Dirichlet smoothing with the smoothing parameter set to the average (document or field) representation length.

We also use **BM25**: with standard parameter settings ($k_1 = 1.2$, $b = 0.8$) [20]; **BM25F-tc**: the fielded extension of BM25 [20], we consider title and content fields, the title weight is set to 0.2 and the content weight to 0.8 [16]; **BM25F-all**: all fields are considered with equal weight. We use the same b value for all fields in the fielded variant BM25F, analogous to [18].

3.3 Results

Table 2 reports the results. We observe that the various query sets exhibit different levels of difficulty; this is indeed what we would have liked to achieve by considering different types of information needs. SemSearch ES queries (that look for particular entities by their name) are the easiest ones, while natural language queries (TREC Entity, QALD-2, and INEX-LD) represent the difficult end of the spectrum. List-type queries (INEX-XER and SemSearch LS) stand halfway in between, both in terms of query formulation (mixture of keyword and natural language) and retrieval difficulty. While a direct comparison of the scores to the official results of these benchmarks is not possible (due to the different collection used and/or that only a subset of the original queries is used here), based on manual inspection of a randomly selected subset, these results appear to be very reasonable.

When looking for significant differences in Table 2, we cannot find many. MLM-tc represents the only case when a significant improvement is observed on the whole query set; the absolute score difference compared to LM, however, is less then 5% and most likely it is a consequence of the improvements on a particular subset of queries (SemSearch ES). In all other cases, there is either no significant improvement or only a given subset of queries are significantly helped while another subset is significantly hurt.

4. ANALYSIS

In this section we perform a topic-level analysis in order to gain some insights into the differences between the various methods (or lack of thereof). Given the space limitations, we focus on (some of) the LM-based approaches; also, according to Table 2 these exhibit more differences than their BM25-based counterparts.

We compare the MLM-all (fielded language models, with equal field weights) to the baseline (single-field) LM method in Figure 1 and to a more sophisticated PRMS method (with query term-specific field weighting) in Figure 2. In both figures the X-axis represents individual query topics, ordered by AP differences (shown on the Y-axis). MLM-all is taken to be the baseline, that is, positive values indicate that the other method outperforms MLM-all and negative values mean the advantage of MLM-all on that particular topic.

First, we observe that a large number of topics is affected, esp. on the easier query subsets (Figures 1(a)–1(d) and 2(a)–2(d)). These improvements, however, do not add up; many of the topics that are improved by moving from LM to MLM-all are hurt when a transition from MLM-all to PRMS is made. When looking into the individual topics with little to no performance differences (i.e., the ones "in the middle" of Figures 1(e)–1(f) and 2(e)–2(f)) we find that both methods that are being considered do equally bad on these topics—in many cases they fail to find any relevant results.

5. SUMMARY

In this paper we made several contributions to three main topics that were identified as important priorities for future research and development for the field of entity search [2]: (A1) getting more representative information needs and favouring long queries over short ones, (A2) limiting search to a smaller, fixed set of entity types (as opposed to arbitrary types of entities), and (A3) using test collections that integrate both structured and unstructured information about entities.

We developed and made publicly available a test collection based on DBpedia and synthesised queries from a number of previous benchmarking evaluation efforts, resulting in a set of nearly 500 queries and corresponding relevance judgments. To initiate further research, we provided baseline results and showed some of the limitations of existing methods based on language models and BM25. Additionally, we provided topic-level analysis and insights on how the choice of retrieval models is bound to the characteristics of different query sub-sets.

The resources developed as part of this study are made available at `http://bit.ly/dbpedia-entity`. It is our plan to maintain "verified" experimental results, a list of papers using this test collection, and pointers to additional related resources (e.g., source code) at the same website.

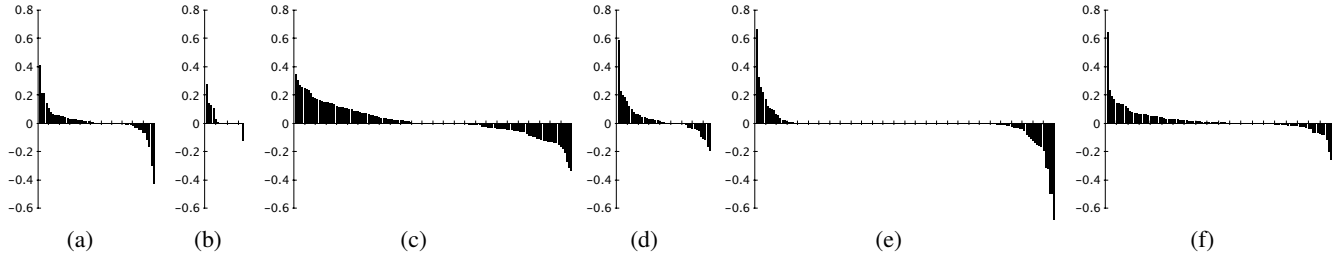

Figure 1: Topic-level differences for LM vs. MLM-all. Y-axis displays AP differences. Positive values indicate LM is better. 1(a) INEX-XER, 1(b) TREC Entity, 1(c) SemSearch ES, 1(d) SemSearch LS, 1(e) QALD-2, 1(f) INEX-LD.

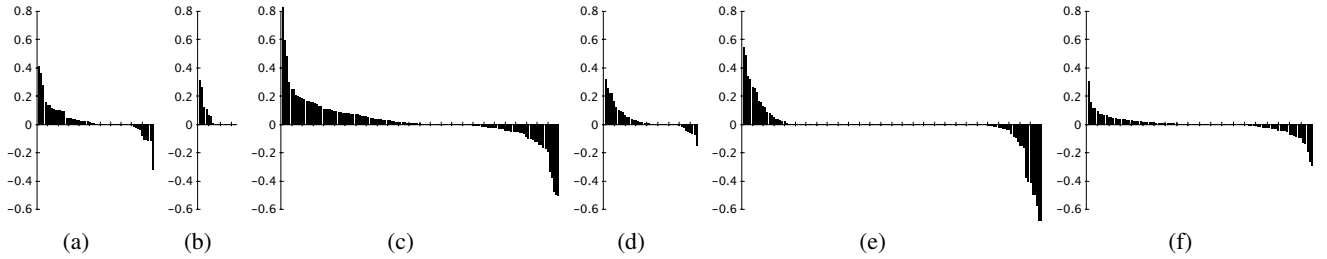

Figure 2: Topic-level differences for PRMS vs. MLM-all. Y-axis displays AP differences. Positive values indicate PRMS is better. 2(a) INEX-XER, 2(b) TREC Entity, 2(c) SemSearch ES, 2(d) SemSearch LS, 2(e) QALD-2, 2(f) INEX-LD.

References

[1] K. Balog, A. P. de Vries, P. Serdyukov, P. Thomas, and T. Westerveld. Overview of the TREC 2009 entity track. In *Proc. of the 18th Text REtrieval Conference (TREC'09)*. NIST, 2010.

[2] K. Balog, D. Carmel, A. P. de Vries, D. M. Herzig, P. Mika, H. Roitman, R. Schenkel, P. Serdyukov, and T. Tran Duc. The first joint international workshop on entity-oriented and semantic search (JIWES). *SIGIR Forum*, 46(2), December 2012.

[3] K. Balog, P. Serdyukov, and A. P. de Vries. Overview of the TREC 2011 entity track. In *Proc. of the Twentieth Text REtrieval Conference (TREC'11)*. NIST, 2012.

[4] R. Blanco, H. Halpin, D. M. Herzig, P. Mika, J. Pound, H. S. Thompson, and T. T. Duc. Entity search evaluation over structured web data. In *Proc. of the 1st International Workshop on Entity-Oriented Search (EOS'11)*, pages 65–71, 2011.

[5] R. Blanco, P. Mika, and S. Vigna. Effective and efficient entity search in rdf data. In *Proceedings of the 10th international conference on The semantic web - Volume Part I*, ISWC'11, pages 83–97. Springer-Verlag, 2011.

[6] M. Bron, K. Balog, and M. de Rijke. Ranking related entities: components and analyses. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1079–1088. ACM, 2010.

[7] M. Bron, K. Balog, and M. de Rijke. Example based entity search in the web of data. In *Proceedings of the 35th European conference on Advances in Information Retrieval*, ECIR'13, pages 392–403. Springer-Verlag, 2013.

[8] A. P. de Vries, A.-M. Vercoustre, J. A. Thom, N. Craswell, and M. Lalmas. Overview of the INEX 2007 entity ranking track. In *INEX*, volume 4862, pages 245–251, 2008.

[9] G. Demartini, T. Iofciu, and A. P. De Vries. Overview of the inex 2009 entity ranking track. In *Proc. of the Focused retrieval and evaluation, and 8th international conference on Initiative for the evaluation of XML retrieval (INEX'09)*, pages 254–264. Springer-Verlag, 2010.

[10] S. Elbassuoni, M. Ramanath, R. Schenkel, M. Sydow, and G. Weikum. Language-model-based ranking for queries on RDF-graphs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 977–986. ACM, 2009.

[11] H. Halpin, D. M. Herzig, P. Mika, R. Blanco, J. Pound, H. S. Thompson, and D. T. Tran. Evaluating ad-hoc object retrieval. In *Proc. of the International Workshop on Evaluation of Semantic Technologies (IWEST'10)*, 2010.

[12] J. Kim and W. B. Croft. A field relevance model for structured document retrieval. In *Proc. of the 34th European conference on Information Retrieval (ECIR'12)*, pages 97–108. Springer, 2012.

[13] J. Kim, X. Xue, and W. Croft. A probabilistic retrieval model for semistructured data. In *Proc. of the 31st European Conference on Information Retrieval (ECIR'09)*, pages 228–239. Springer, 2009.

[14] V. Lopez, C. Unger, P. Cimiano, and E. Motta. Evaluating question answering over Linked Data. *Journal of Web Semantics*, to appear.

[15] R. Neumayer, K. Balog, and K. Nørvåg. On the modeling of entities for ad-hoc entity search in the web of data. In *Proc. of the 34th European Conference on Information Retrieval (ECIR'12)*, pages 133–145. Springer, 2012.

[16] R. Neumayer, K. Balog, and K. Nørvåg. When simple is (more than) good enough: Effective semantic search with (almost) no semantics. In *Proc. of the 34th European Conference on Information Retrieval (ECIR'12)*, pages 540–543. Springer, 2012.

[17] P. Ogilvie and J. Callan. Combining document representations for known-item search. In *Proc. of the 26th annual international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR'03)*, pages 143–150. ACM, 2003.

[18] J. R. Pérez-Agüera, J. Arroyo, J. Greenberg, J. P. Iglesias, and V. Fresno. Using BM25F for semantic search. In *Proc. of the 3rd International Semantic Search Workshop (SemSearch'10)*, pages 1–8, 2010.

[19] J. Pound, P. Mika, and H. Zaragoza. Ad-hoc object retrieval in the web of data. In *Proc. of the 19th international conference on World wide web (WWW'10)*, pages 771–780. ACM, 2010.

[20] S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389, 2009.

[21] E. Voorhees. Overview of the TREC 2004 question answering track. In *Proc. of the 13th Text Retrieval Conference (TREC'04)*. NIST, 2005.

[22] Q. Wang, J. Kamps, G. Ramirez Camps, M. Marx, A. Schuth, M. Theobald, S. Gurajada, and A. Mishra. Overview of the INEX 2012 linked data track. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*, 2012.

[23] C. Zhai. Statistical language models for information retrieval a critical review. *Found. Trends Inf. Retr.*, 2:137–213, 2008.