Overview of INEX 2008 Link the Wiki Track

Wei Che (Darren) Huang[1], Shlomo Geva[2] and Andrew Trotman[3]

Faculty of Science and Technology, Queensland University of Technology, Brisbane, Australia [1, 2]
Department of Computer Science, University of Otago, Dunedin, New Zealand [3]
w2.huang@student.qut.edu.au [1]
s.geva@qut.edu.au [2]
andrew@cs.otago.ac.nz [3]

Abstract. The Link the Wiki track at INEX 2008 offered two tasks, file-to-file link discovery and anchor-to-BEP link discovery. In the former 6600 topics were used and in the latter 50 were used. Manual assessment of the anchor-to-BEP runs was performed using a tool developed for the purpose. Runs were evaluated using standard precision & recall measures such as MAP and precision / recall graphs. 10 groups participated and the approaches they took are discussed. Final evaluation results for all runs are presented.

Keywords: Wikipedia, Link Discovery, File-to-File, Anchor-to-BEP, Assessment, Evaluation.

1 Introduction

Trotman & Geva [1] introduced the Link the Wiki task in 2006. It ran at INEX for the first time in 2007 [2]. This contribution discusses the track as it was run in 2008. The track provides an independent evaluation forum for approaches to link discovery in the Wikipedia. In 2007 the track examined file-to-file linking in the Wikipedia, but in 2008 this was extended to include anchor to best entry point (anchor-to-BEP) link discovery. A test set including document collection, qrels, metrics, and tools for evaluating submissions [3] was constructed and is now provided for future experimenters. The document collection was the INEX Wikipedia collection, the topics (known as orphans) were documents from within the collection.

Ten groups from eight different organizations participated in the track. 25 runs were submitted to the file-to-file task and 31 runs to the anchor-to-BEP task. All runs were evaluated against a ground truth of those links already present in the collection. Anchor-to-BEP runs were additionally evaluated against a ground truth determined through manual assessment. These manual assessments allow for file-to-BEP, anchor-to-file, anchor-to-BEP and also file-to-file assessment; something that was essential because many submitted runs were file-to-file runs despite the task being defined as anchor-to-BEP; that is, the anchor texts were the document title and the best entry point was the beginning of the target document.

Anchor-to-BEP link discovery differs from traditional link discovery by pointing from anchors directly to relevant material within the target document, rather than pointing to simply the document [4][5]. The purpose of *focused link discovery* is to identify anchors together with the corresponding *best entry points* such that the link is relevant *with respect to the anchor's specific context.*

2 Document Collection and Resources

The document collection was the INEX Wikipedia collection of 659,388 articles. For the file-to-file task 6600 documents were randomly selected from the collection. For anchor-to-BEP assessment each participating group was asked to nominate 5 candidate documents, 10 groups participated which resulted in 50 documents for manual assessment. These documents are known as pre-orphans. The documents were separated from the collection by removing all outgoing links from the documents into the collection as well as all incoming links from the collection into the documents. These separated documents are known as orphans.

The orphaning process itself was performed by the track participants. The exact method was left to the participant however the requirement was that the process should be equivalent to: orphaning one document; identifying the links to and from that document; then returning the (original) document to the collection. In this way each orphan was linked against the remainder of the collection as it would have been if that orphan was presented for insertion into the collection.

Various resources were made available to participants including: a text-only version of the collection (with the XML removed) so that file-offset-lengths could be computed by counting characters from the start of the file; *XML2FOL*, a program that produces a list of all the offsets and lengths of all elements in an XML file; *XML2FOLpassage*, a program to convert any INEX XPath specification into the FOL format, *XPath2txt*, a program that extracts the text of a given element from a given file. These resources could be used by participants to validate their systems.

3 Task Specification and Submission Format

The task was specified as twofold: the identification of links from the orphan into the document collection; and the identification of links from the collection into the orphan.

In the anchor-to-BEP scenario the best 50 anchors within the orphan could be identified, and for each no more than 5 BEP destinations could be specified. Alongside these the best 250 incoming links from anchor texts in the remaining collection to BEPs within the orphan could be specified. For file-to-file evaluation the task was to identify the best 250 outgoing and best 250 incoming links.

The specification of a file-to-file link is a special case of the specification of an anchor-to-BEP link. A file-to-file link is from the start of the source to the start of the target. This reduction of the complex task to the less complex task provided a low-cost entry into the track for those who had not participated before.

For submission purposes the orphans were identified by the triplet (topic-id, file name, title). Although each is unique for each orphan (and are thus any one could have been used) all three were used for clarity's sake. Both the INEX ad hoc XPath syntax and the INEX File-Offset-Length (FOL) formats were used for submissions. All file offsets and lengths were specified as character offsets with respect to the text content of the files; counting from zero; and ignoring all mark-up. An anchor might be specified, for example, as (*23816.xml, 1234, 8*) but a BEP has no length so it would be specified (*23816.xml, 672*). Examples of the anchor-to-BEPs formats are given in Figure 1.

```
<link>
  <anchor>
    <offset>234</offset>
    <length>24</length>
  </anchor>
  <linkto>
    <file>123.xml</file>
    <bep>334</bep>
  </linkto>
   ... <multiple links for an anchor>
</link>
```

```
<link>
    <anchor>
      <start>/article[1]/p[5]/text()[3].12</start>
      <end>/article[1]/p[5]/text()[3].32</end>
    </anchor>
    <linkto>
      <file>43768.xml</file>
      <bep>/article[1]/p[3]/text()[4].40</bep>
    </linkto>
    ... <multiple links for an anchor>
</link>
```

Fig. 1. Sample Anchor-to-BEP Submission Format

4 Preparation of qrels

For the file-to-file evaluation of the 6600 orphans the ground truth was constructed without manual assessment. The links from the pre-orphan to the remaining collection were extracted and used as the outgoing ground truth. The links from the collection into the pre-orphans was used as the incoming ground truth. For anchor-to-BEP assessment this is not possible because BEPs are rarely specified in the Wikipedia.

There are known problems with using the Wikipedia itself as the ground-truth: some Wikipedia links are topically-obsolete or have been incorrectly assigned; linking is not exhaustive; articles are unlikely to link to very recently added content; and some links are inserted by bots. As a consequence, evaluation results may be biased various ways. On the one hand, results may appear optimistic because some links are trivial to discover (such as year-links). On the other hand, results may appear pessimistic because useful links not already in the Wikipedia are considered non-relevant. However, evaluation based on the Wikipedia ground-truth does measure performance relative to what is present, and so it is reasonable to believe it is useful.

Although the Wikipedia does contain anchor-to-BEP links, in practice they are rarely used. In order to evaluate anchor-to-BEP link discovery an evaluation result-set was created through manual assessment. A special case of pooling was used in the track – all links for a given orphan were pooled, then for each anchor, all BEPs were pooled. The pool was assessed to completion.

5 Assessment

An assessment tool (see Figure 2) was used in 2008 to facilitate the assessment of link discovery in both the anchor-to-BEP and file-to-file scenarios. As assessment is laborious and time consuming, special care was taken to minimize the amount of mouse motion and clicking – a single click could, for example, be used to specify a relevant link.

Overlapping links were also addressed by the tool. Links for each anchor were grouped for easy and clear presentation (for example, the anchor *Modern Information Retrieval* may appear as the anchors: *Information Retrieval* or *Modern Information Retrieval*), but the tool also captured each sub-anchor explicitly so that the assessor could differentiate and judge with respect to each sub-anchor. File-to-file links were presented as linking from the title of the document to the beginning of the target document.

No concerns were raised about the tool. The number of links identified per topic varied from 405 to 1722. An average of about 5 hours was spent assessment an orphan. Most links were file-to-file. Less than 10% of the links identified for each orphan were judged relevant.

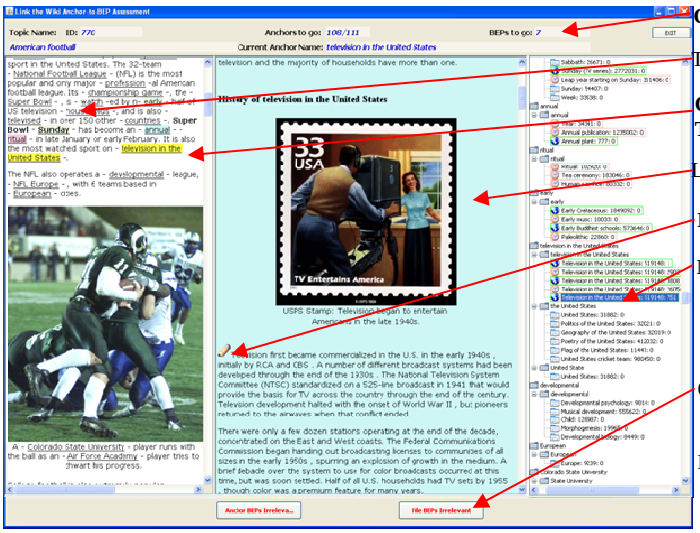

Fig. 2. The Assessment Tool

6 Evaluation

6.1 The Evaluation Tool

An evaluation tool (*ltwEval*, see Figure 3) was developed for the track. Performance measures including Mean Average Precision (MAP), precision at the number of

relevant documents (R-Prec), and precision at given retrieval cut-offs (P@5, P@10, P@20, P@30, P@50 and P@250) were computed. The tool draws Interpolated Precision / Recall plots allowing graphical comparative analysis of multiple runs. *LtwEval* is GUI driven and was written in Java for platform independence.

The tool gives the number of outgoing and incoming links in the qrels as well as in each run (duplicate links being eliminated). Performance measures can be calculated using all topics in the qrels or just the topics in the run. From the evaluation result table (that displays all metrics), the color and used in graphing can be specified.

The Wikipedia ground-truth qrels (for both the 6600 file-to-file topics and the 50 anchor-to-BEP topics) can only be used to evaluate the submission runs in *file-to-file* mode while the manual assessment results can be used to perform the evaluation in several different modes. Besides evaluating file-to-file links, the anchor-to-BEP submission runs are also evaluated at *file-to-BEP*, *anchor-to-file* and *anchor-to-BEP* modes. The file-to-BEP evaluation considers the entry point, weighting the link score by BEP proximity in a similar manner to that used in the ad hoc track: the score drops linearly to 0 over a distance of 1000 characters; an exact match is given a score of 1 while 0 is given to the BEP beyond 1000 characters. If more than one BEP is specified in the target document, the the closest is used. The evaluation in anchor-to-file and anchor-to-BEP mode considers only the first 50 anchors, and only the first BEP of each anchor.

6.1 Metrics

In the INEX use case of link discovery it is important to rank the discovered links for presentation to the page author. A typical scenario might involve a user who wishes to inspect and then accept or reject recommended links. This use case was modeled in the manual assessment evaluation where assessors did exactly this. In a realistic link discovery setting the user is unlikely to trudge through hundreds of recommended anchors, so the best anchors should be presented first. The link discovery system must also balance extensive linking against link quality.

Traditional measures such as MAP, R-Prec, P@n and Interpolated Precision-Recall plots address the problem of file-to-file link discovery well, but do not address the performance of anchor-to-BEP methods at all well (because anchor and BEP near misses are not considered); it is necessary to adapt metrics to the problem.

For evaluation purposes runs must be of a finite length (and quite short for manual assessment purposes). Often there are more known relevant links in the qrels than the assessment imposed submission length – in short, there are sometimes more than 50 relevant anchors in an orphan despite the submission requirements capping the number of anchors that can be identified at 50.

To address this problem MAP was altered so that it now corresponds to the maximum point of recall in a run or the actual number of relevant links, whichever is smaller. That is, as the run length was limited to 50, the calculation of MAP was based on a maximum recall of 50 relevant links. Because of this, a run consisting of 50 relevant links scores a MAP of 1.0 and the RP curve depicts a line-at-1.

An anchor may be defined by a user in several slightly different ways. For instance, *The Theory of Relativity*, *Theory of Relativity*, and *Relativity* may be

conceptually identical anchors. Furthermore, if the anchor text occurred several times in a document only one instance is likely to be anchored (according to the Wikipedia guidelines) and so the location of an anchor may vary without becoming semantically incorrect (we leave for further work the question of which occurrence of an anchor is *best*). During assessment anchors were explicitly assessed as either relevant or irrelevant. Only relevant anchors contributed to the score of a submission – through the score assigned to the relevant links, if exist. In a quick pass over the orphan the assessor could reject all anchors that were trivially irrelevant – even without looking at the linked documents. Year links, for instance, could be rejected outright without the need for inspection of the target.

Similarly to anchors, a BEP cannot be defined with absolute accuracy. Some reasonable proximity to a designated BEP in the assessments must be allowed. So a BEP might be considered relevant if, when viewed on a screen, it is no more than some distance (*N* characters) away from a point chosen by an assessor. The track defined the BEP score of a link as:

$$\text{bep score} = \text{file score} \times [1 - (|\text{bep\_position}_{\text{Run}} - \text{bep\_position}_{\text{qrel}}|/ N)]. \quad (1)$$

So in summary, an anchor-to-BEP link was assessed as relevant on the basis of approximately matching both the anchor and the BEP of a relevant link in the assessments. Anchors were either accepted or rejected. Having computed all individual anchor-to-BEP link scores for accepted anchors, the document score can be derived using the Average Precision in the usual manner. The MAP can then be computed over the entire set of topics.

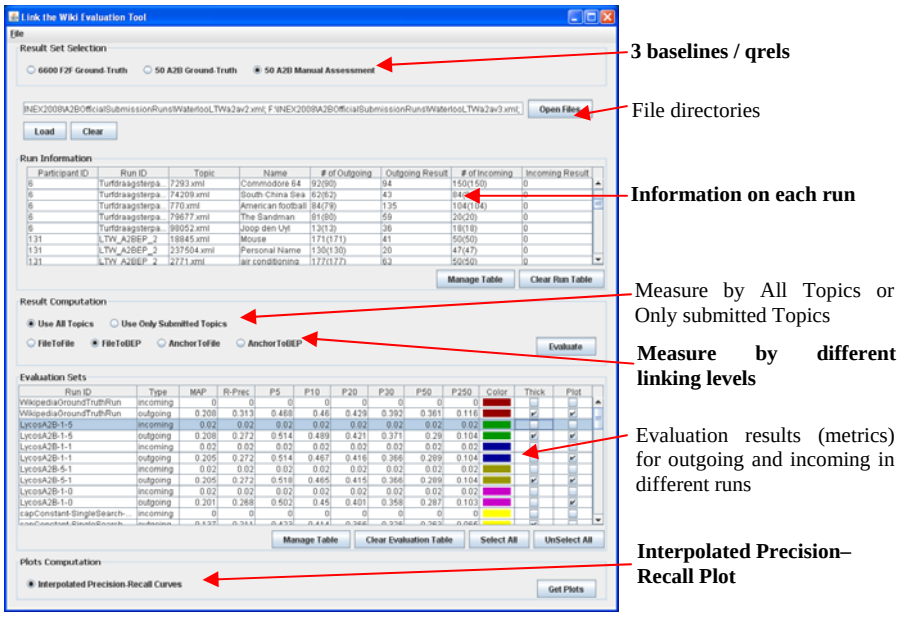

Fig. 3. The Evaluation Tool

7 Approaches to Link Discovery

This section describes some of approaches taken by track participants. In all, there were 10 participating groups (including 2 independent groups from each of: the University of Amsterdam and from Queensland University of Technology).

The University of Amsterdam (de Rijke) submitted 3 runs for the file-to-file link discovery. For the outgoing links, they selected anchors with LLR (Link Likelihood Ratio) > 1 and used the anchor text as a query to retrieve target pages (searching in the title field). For the incoming links, the topic title was used as a query to retrieve the top 250 source pages within the language modeling framework. In anchor-to-BEP link discovery, outgoing links were discovered by selecting anchors with LLR > 1 and then retrieving the target page whose title matched (exact or partial) the anchor text. The target pages were ranked according to the likelihood of the target title in the topic page (p (*Title* | D)). Incoming links were retrieved by using the topic title to find exact matches in the collection. In their third submission, the topic title was used as a query to retrieve 250 candidate target pages (ranked by cosine similarity) and the pages whose rsv was greater than 0.15 were selected as the outgoing links. Incoming links used the same strategy to select the source pages whose similarity is greater than 0.026.

Lycos Europe GmbH submitted 2 runs for the file-to-file link discovery and 5 runs for the anchor-to-BEP task. The approach used by Lycos is derived from Itakura & Clarke's approach in 2007 [14]. The difference is that Lycos dynamically selected the best-matching target for a given anchor text based on content similarity. For example, in a text about computers, the anchor "Apple" is more likely to refer to the page "Apple Computers" than to the page "Apple Records". Moreover, the system also analyzed the links between the potential targets for all anchor texts so that they could see which set of links were related (for example, the anchor "Apple" in a text that also links to "The Beatles" should most likely link to "Apple Records" and not "Apple Computers").

Know-Centre Graz submitted 2 runs for the file-to-file link discovery and 6 runs for the anchor-to-BEP task. The outgoing links were identified using gazetteer matching of page titles. The identified outgoing links were ranked using cosine similarity based on noun vectors. The incoming links were identified similarly by searching for the title and using the orphan documents nouns for calculating the cosine similarity. The difference between the two runs (here referred to as run1 and run2) was the ranking scheme. The outgoing links in run2 were ranked by the IDF frequency of the occurring text in the corpus. Differently to the incoming links in run2, the nouns for every sentence in the orphan document were used for calculating the cosine similarity to the incoming link source, wherefrom the maximum cosine similarity on each sentence was taken.

The University of Waterloo submitted 3 runs for the file-to-file link discovery and 3 runs for the anchor-to-BEP task. For the file-to-file link discovery, thir first run, they utilized the same approach they used last year (which placed first). In their second run, outgoing links were discovered using the same method as thr first, except with the cut-off for the number of links to return according to the size of topic files. Incoming links were selected using an element retrieval approach using BM25. For their third run, Outgoing links were done using page rank while incoming links are

done using topic oriented page rank assuming that what was found for the outgoing links was correct. The algorithms used for the anchor-to-BEP task was the same, except for finer granularity in specifying sources and destinations.

The Queensland University of Technology submitted 5 runs for the file-to-file task and 6 runs for the anchor-to-BEP task. Several runs used the GPX search engine using the same approach they used in 2007. Several runs used the Terrier search engine out of the box to find document to document links was also tried. Finally, several runs used frequent phrase mining to identify suitable anchors and links.

The University of Otago submitted 3 runs for the file-to-file link discovery and 3 runs for the anchor-to-BEP task. These runs were based on the Itakura & Clarke approach from 2007, but with particular attention paid to parsing issues.

8 Results and Conclusion

The tables 1 and 2 present the final assessment results using Mean Average Precision (MAP). The figures 4-11 present the Interpolated Precision / Recall graphs of each run.

This is the second year of the Link-the-Wiki track at INEX, and the year the anchor-to-BEP link discovery task was introduced. Since the anchor-to-BEP link discovery can be applied in different scenarios to enhance the efficiency of the interaction it is important to build a standard procedure to measure the performance and tools to facilitate the evaluation and assessment. This attempt has opened a door for participants to share their suggestions and opinions for the track, which will improve the capability of the track to facilitate further the link discovery research. Several qrels sets for evaluating runs at different granularity levels were produced and used to measure the performance of various approaches. The GUI-based tools balance the time-consuming assessment and evaluation processes for investigating the approaches.

Because this was the first year for the anchor-to-BEP link discovery task it was expected (and seen) that some runs would contain invalid positions for anchors and BEPs (some contained file-to-file links). Because of this, the relative comparison of runs may be biased towards correctly formatted runs (at the expense of better but incorrectly formatted runs).

Table 1. MAP of 6600 File-to-File topics link discovery evaluated by Wikipedia Ground Truth.

Outgoing Runs	MAP	Incoming Runs	MAP
Otago_nonCap-FirstPara	0.7343	QUT_GPXF2FnameInOut	0.5713
AmsterdamDeRijke_ltw02	0.3475	Waterloo_f2f#3	0.5563
Waterloo_f2f#1	0.3345	Waterloo_f2f#1	0.5540
Otago_capConst-SingleSearchWeight	0.3045	KnowCenterGraz_globalTFIDFSen	0.5369
Otago_capConst-TitleOnly	0.3045	Waterloo_f2f#2	0.5350
AmsterdamDeRijke_ltw01	0.2924	KnowCenterGraz_WordLvldisambig.	0.5299
Waterloo_f2f#2	0.2920	AmsterdamDeRijke_ltw02	0.5249
LycosF2F-1-5	0.2379	CMIC_F2F_02	0.5116
LycosF2F-1-1	0.2360	Otago_capConstant-TitleOnly	0.4869
Waterloo_f2f#3	0.2053	AmsterdamDeRijke_ltw01	0.4800

QUT_GPXF2FnameInOut	0.1440	CMIC_F2F_01	0.4579
KnowCenterGraz_globalTFIDFSen	0.1407	QUT_LTW_F2F_01	0.4322
KnowCenterGraz_WordLvldisambig	0.1129	Otago_capConst-SingleSearchWeight	0.4314
Amsterdam_a2a_2	0.1088	Amsterdam_a2a_3	0.3575
Amsterdam_a2a_1	0.1071	Amsterdam_a2a_1	0.3392
AmsterdamDeRijke_ltw03	0.1041	AmsterdamDeRijke_ltw03	0.3345
QUT_F-F_1	0.1026	LycosF2F-1-1	0.3266
QUT_F-F_2	0.1026	LycosF2F-1-5	0.3266
Amsterdam_a2a_3	0.1017	CSIR_LTW_F2F_2	0.2940
QUT_GPXF2Ftitle	0.0566	QUT_F-F_2	0.2915
CSIR_LTW_F2F_2	0.0082	Amsterdam_a2a_2	0.2879
		Otago_nonCap-FirstPara	0.2228
		CSIR_LTW_F2F_1	0.1645
		QUT_F-F_1	0.0925

Table 2. MAP of 50 Anchor-to-BEP topics evaluated by manual and Wikipedia ground-truths.

Submission Runs	F2F	F2B	A2F	A2B	Out Wiki	In Wiki
WikipediaGroundTruthRun	0.2765	0.2079	0.3945	0.3888	1	1
LycosA2B-1-5	0.2463	0.2078	0.4973	0.4918	0.1193	0.1753
LycosA2B-1-1	0.2431	0.2050	0.4930	0.4876	0.1172	0.1753
LycosA2B-5-1	0.2427	0.2050	0.4931	0.4876	0.1169	0.1753
LycosA2B-1-0	0.2387	0.2008	0.4708	0.4656	0.1148	0.1753
Otago_capConst-SingleSearch	0.1745	0.1365	0.3952	0.3910	0.3810	0.2389
Otago_capConst-TitleOnly	0.1745	0.1365	0.3952	0.3910	0.3810	0.2408
Otago_nCapConst-WholeDoc	0.1724	0.1352	0.3896	0.3853	0.3769	0.0745
KnowCenterGrazdisamDocNoneSen	0.1546	0.1077	0.1764	0.1453	0.2370	0.1435
KnowCenterGrazdisamDocNoneTopic	0.1546	0.0603	0.2131	0.1968	0.2370	0.1429
KnowCenterGrazdisamTopicNonSen	0.1522	0.1058	0.2076	0.1662	0.2091	0.1695
KnowCenterGrazdisamTopicNonTopic	0.1522	0.0620	0.2643	0.2384	0.2091	0.1676
KnowCenterGrazglobalIDFSentence	0.1371	0.1222	0.2309	0.1895	0.2200	0.1725
KnowCenterGrazglobalIDFTopic	0.1371	0.0688	0.2873	0.2619	0.2200	0.1725
Waterloo_a2a#1	0.1282	0.1004	0.4111	0.4071	0.2191	0.2165
LycosA2B-0-1	0.1200	0.1051	0.3291	0.3249	0.0432	0.1753
QUT_LTWA2BnameRerank	0.1196	0.0946	0.3042	0.3012	0.1816	0.4615
Amsterdam_a2bep_5	0.1127	0.0847	0.2079	0.2058	0.1426	0.2349
QUT_GPXA2Bname	0.1110	0.0882	0.2912	0.2882	0.1522	0.4236
Waterloo_a2a#2	0.1071	0.0823	0.3355	0.3325	0.1854	0.1804
Waterloo_a2a#3	0.0882	0.0656	0.3874	0.3835	0.1710	0.2044
CMIC_LTW_01	0.0763	0.0576	0.1760	0.1740	0.1004	-
CSIR_LTW_A2BEP_2	0.0760	0.0478	0.1307	0.1237	0.0647	0.1577
Amsterdam_a2bep_1	0.0746	0.0556	0.1271	0.1261	0.0973	0.2349
Amsterdam_a2bep_3	0.0685	0.0518	0.0983	0.0975	0.0911	0.1566
Amsterdam_a2bep_2	0.0671	0.0491	0.1127	0.1115	0.0872	0.2349
QUT_Anchor-BEP_1	0.0524	0.0424	0.1149	0.1141	0.0729	0.0710
QUT_P9_GPXA2Btitle)	0.0487	0.0388	0.1725	0.1712	0.0533	0.4511

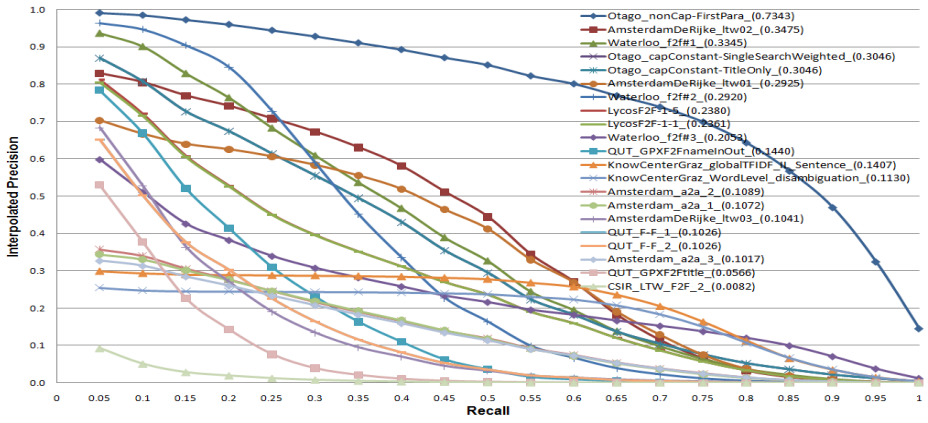

Fig. 4. 6600 File-to-File Topics Outgoing link discovery evaluated by Wikipedia Ground Truth

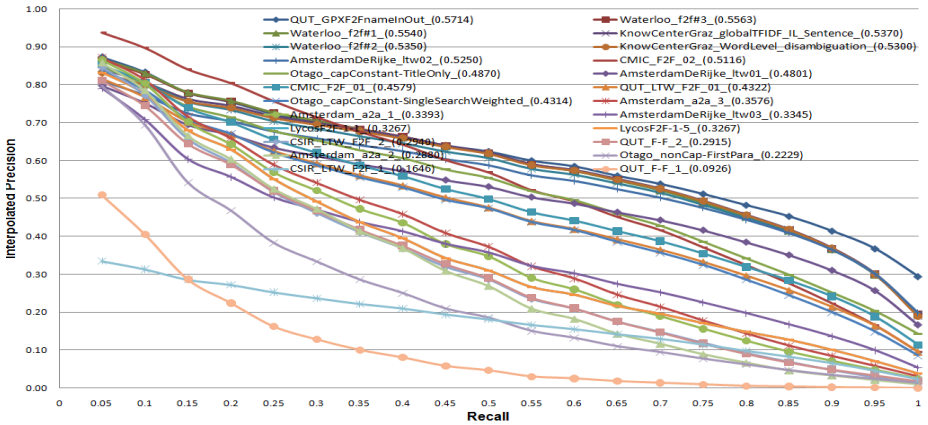

Fig. 5. 6600 File-to-File Topics Incoming link discovery evaluated by Wikipedia Ground Truth

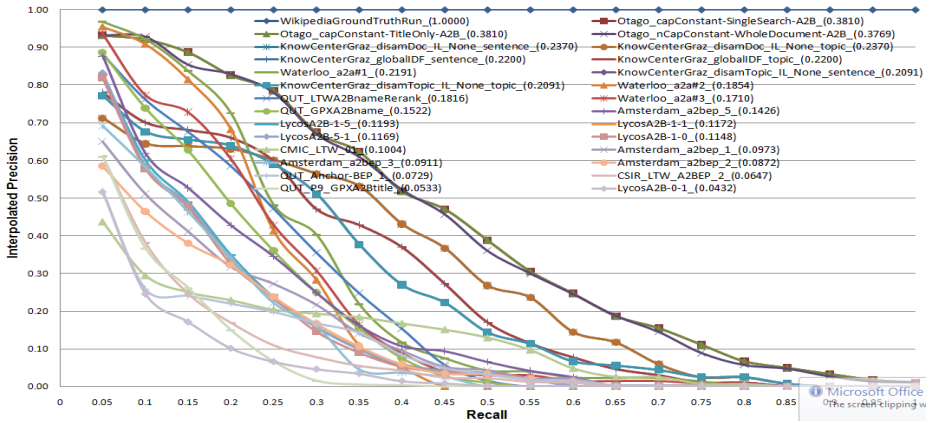

Fig. 6. 50 Anchor-to-BEP Outgoing link discovery evaluated by Wikipedia Ground Truth

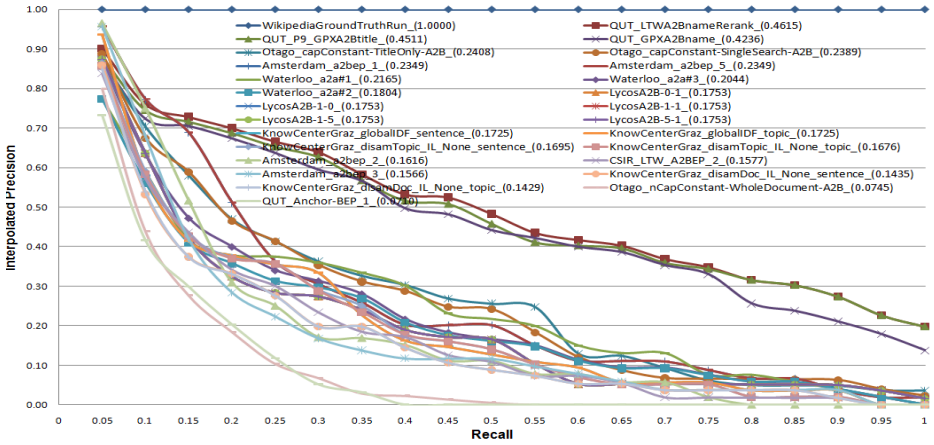

Fig. 7. 50 Anchor-to-BEP Incoming link discovery evaluated by Wikipedia Ground Truth

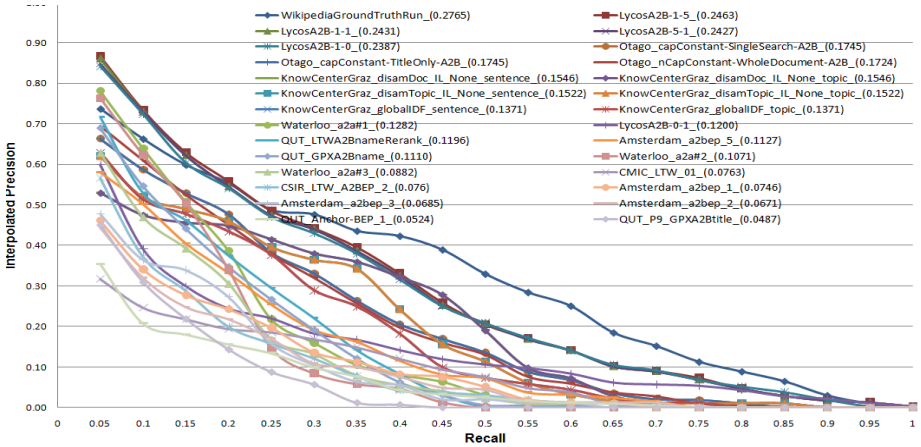

Fig. 8. 50 Anchor-to-BEP Outgoing links: File2File Evaluation by Manual Ground Truth

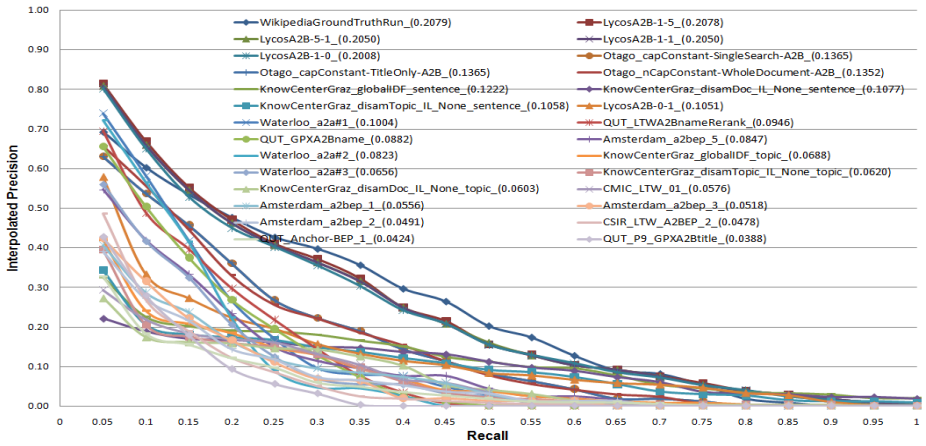

Fig. 9. 50 Anchor-to-BEP Outgoing links: File2BEP Evaluation by Manual Ground Truth

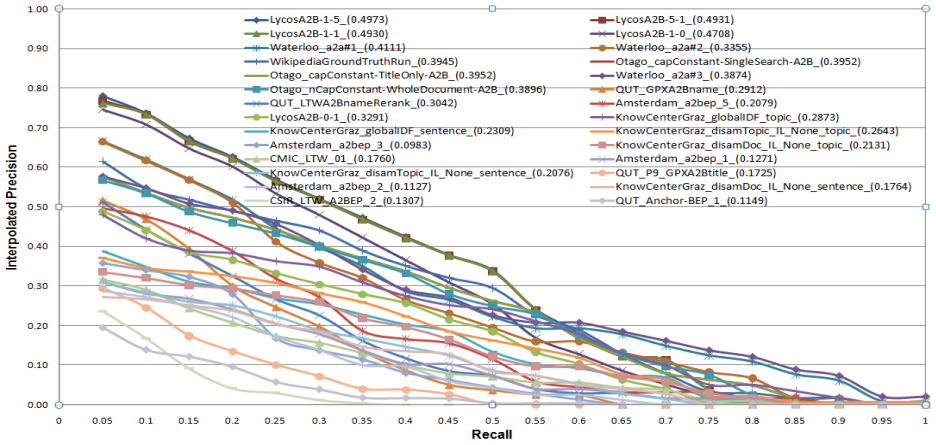

Fig. 10. 50 Anchor-to-BEP Outgoing links: Anchor2File Evaluation by Manual Ground Truth

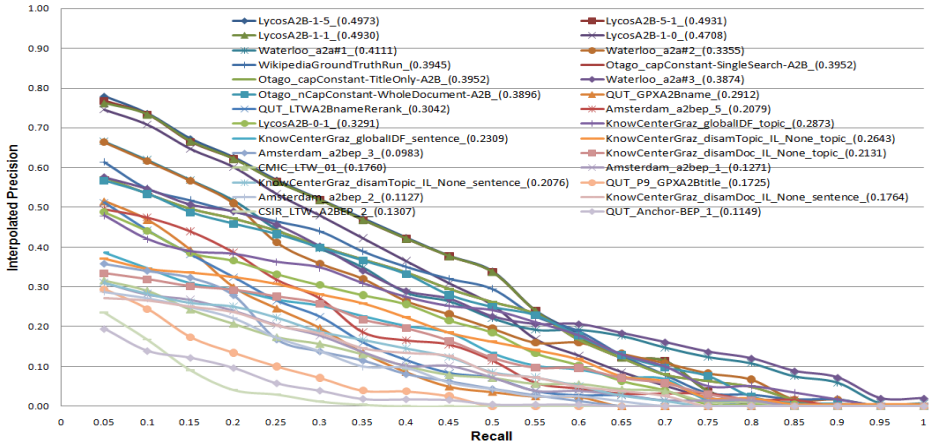

Fig. 11. 50 Anchor-to-BEP Outgoing links: Anchor2BEP Evaluation by Manual Ground Truth

References

1. Trotman, A. and Geva, S. Passage Retrieval and other XML-Retrieval Tasks, In: the SIGIR 2006 Workshop on XML Element Retrieval Methodology, pp. 48-50.
2. Huang, W. C., Xu, Y., Trotman, A. and Geva, S. (2008) Overview of INEX 2007 Link the Wiki Track, INEX 2007, LNCS 4862, N. Fuhr et al. (Eds.), pp. 373-387.
3. Huang, W. C., Xu, Y., Trotman, A. and Geva, S. (2008) Experiments and Evaluation of Link Discovery in the Wikipedia, In: the SIGIR 2008 Focused Retrieval Workshop, Singapore.
4. Voss, J. Measuring Wikipedia, In: the 10th International Conference of the International Society for Scientometrics and Informetrics (ISSI 2005).
5. Adafre, S. F. and de Rijke, M. Discovering missing links in Wikipedia, In: the SIGIR 2005 Workshop on Link Discovery: Issues, Approaches and Applications.