Overview of the INEX 2009 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. In the third year of the Link the Wiki track, the focus has been shifted to anchor-to-bep link discovery. The participants were encouraged to utilize different technologies to resolve the issue of focused link discovery. Apart from the 2009 Wikipedia collection, the Te Ara collection was introduced for the first time in INEX. For the link the wiki tasks, 5000 file-to-file topics were randomly selected and 33 anchor-to-bep topics were nominated by the participants. The Te Ara collection does not contain hyperlinks and the task was to cross link the entire collection. A GUI tool for self-verification of the linking results was distributed. This helps participants verify the location of the anchor and bep. The assessment tool and the evaluation tool were revised to improve efficiency. Submission runs were evaluated against Wikipedia ground-truth and manual result set respectively. Focus-based evaluation was undertaken using a new metric. Evaluation results are presented and link discovery approaches are described.

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

1 Introduction

The Link the Wiki track was run for the first time in 2007 [1, 2]. It aims to offer an independent evaluation forum for researchers to work together to solve the problem of anchor-to-bep link discovery. The participants are encouraged to utilize different technologies, such as data mining, natural language processing, machine learning, information retrieval, etc., to discover relevant anchors in a new article and link the anchor to best entry points in other documents.

In 2007, the file-to-file (i.e. F2F) runs were evaluated against the Wikipedia ground truth whilst the anchor-to-bep (i.e. A2B) task was introduced in 2008 [3]. High fidelity file-to-file link discovery within the Wikipedia has been achieved as an outcome in 2008, as measured in comparisons with the ground truth. The focus has now been shifted to anchor-to-bep link discovery. Several improvements, including the submission specification, the tools, evaluation methods and metrics, have been made to conduct a better experiment in focused link discovery. Apart from the Wikipedia collection, the Te Ara encyclopedia was introduced and the tasks, *Link Te*

Ara and *Link Te Ara to Wiki*, were set up for the first time. Despite its small size, there is a real challenge offered by the Te Ara collection. Since it is not extensively linked, and since page names are not necessarily as informative as Wikipedia page names, both link mining and page-name matching - the methods that work particularly well with the Wikipedia - are ineffective with the Te Ara.

Six groups from different organizations participated in the 2009 track. 16 runs were received for the file-to-file task while 13 runs for the anchor-to-bep task and 8 runs for the F2F on A2B task were submitted. Two groups were also involved in the Te Ara tasks with 7 runs contributed. All link the wiki runs were evaluated against the Wikipedia ground truth. All anchor-to-bep runs were additionally evaluated in different ways such as anchor-to-file and anchor-to-bep. The qrels are obtained through manual assessment. A set of evaluation results is depicted and a brief discussion is presented in this paper.

2 Document Collection

Two collections, the Wikipedia and the Te Ara, were used in the Link the Wiki track in 2009. The Wikipedia corpus consists of 2,666,190 articles with roughly 50GB in size. This collection is much larger than the one used in 2008. For file-to-file link discovery, 5000 articles were randomly selected, but filtered by certain criteria such as the document size and the number of anchors (i.e. links) to control the quality of the documents used in the task. For anchor-to-bep link discovery, the participants nominated 33 topics and submissions were manually assessed by the nominator who is expected to be fully acquainted with the topic content.

The Te Ara Encyclopedia was also used in the Link the Wiki track for 2 designated Te Ara tasks. At the time of writing, the collection contains 3179 articles with around 50MB in size without images. Currently there is no link in the collection and some of documents are still small. New approaches were expected to carry out focused link discovery without taking any advantage of link mining and page name match. The linking was required for the whole collection.

3 Task Specification

3.1 Tasks

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 at both file-to-file and anchor-to-bep levels. Anchor-to-bep link discovery: This task represents the main goal of the Link the Wiki track. Researchers are encouraged to develop focused link discover algorithm, produce reliable assessments and participate in the forum to discuss solutions to focused link discovery. Only 50 anchors and up to 5 beps per anchor were allowed for each topic. At most, 250 incoming links could be specified in the case of the Link the Wiki task. Each

incoming link must be from a different document. Only outgoing links were needed for the Te Ara tasks because all documents were used and so all incoming links were discovered anyway. The *Link-Te-Ara* task is to discover anchor texts and link them to best entry points within the collection. *Link-Te-Ara-to-Wiki* is designated to link the anchor text from a Te Ara topic to best entry points in the Wikipedia documents. File-to-file link discovery for the Wikipedia collection: As a special case of the anchor-to-bep task, this task has lower complexity and offers an entry level for newcomers. 5000 documents were selected for file-to-file link discovery. Up to 250 outgoing links and up to 250 incoming links were to be specified per topic. Missing topics were regarded as having a score of zero for the purpose of computing system performance.

3.2 Submission

Each submission run must specify the task (i.e. *LTW_F2F*, *LTW_A2B*, *LTW_F2FonA2B*, *LTAra_A2B* and *LTAraTW_A2B*) performed. The *description* section in the submission format is used to state different link discovery approaches. A sample format in the case of the link the wiki task is presented below.

```
<outgoing>
  <anchor name="Luminiferous aether" offset="1688" length="19">
    <tobep offset="2038">123456</tobep>
    <tobep offset="971">359</tobep>
    …
  </anchor>
  …
</outgoing>
<incoming>
  <bep offset="2038">
    < fromanchor offset="799" length="9" file="654321">radiation</fromanchor>
    < fromanchor offset="1019" length="10" file="3162088">medication</fromanchor>
    …
  </bep>
  …
</incoming>
```

Fig. 1. Sample link the wiki Submission Format

An anchor text was specified in three parts; the start position of the anchor (i.e. **O**ffset), the **L**ength of the text term and the anchor text itself. The position and length were indicated in characters. The offset specified the anchor starting position within the corresponding text-only document. The anchor text itself was used to verify the specification of the offset-length. The document name could be a unique number in the Wikipedia, or a unique name in the Te Ara collection. A destination link could be specified in two parts: a unique document name and a best entry point. It is the best starting point of the content where the relevant content section starts from.

3.3 Restriction of Linking

An anchor, indicated by a combination of *Offset* and *Length*, must appear only once in a topic - although it may have multiple distinct best entry points. An anchor-text in

one document can be linked to several destinations (beps) in other distinct documents. It means that the same set of *Offset* and *Length* should not appear more than once and hence there is no duplicated anchor set for a given topic. For the evaluation purpose, the first 50 anchor sets are extracted and only the first 5 links within the instances of the same anchor offset-length are taken. Document title can also be an anchor, but like any other anchor it can be linked to at most 5 destinations.

3.4 Assistant Program

In order to facilitate the identification of the offset and length for each anchor and bep, several tools have been developed and distributed to participants. A Java program, *XML2FOL*, was created to produce a list of offset-length for all the element nodes in a given XML document. Another Java program, *XML2TXT*, was used to convert the XML document into the text-only content. Apart from the tools, a text-only version of the collection was also available so the offset could be computed by counting the characters from the beginning of the document. These two programs could be embedded into the participant's link discovery system as a parser to identify offset-length for the anchor texts and to produce text only document.

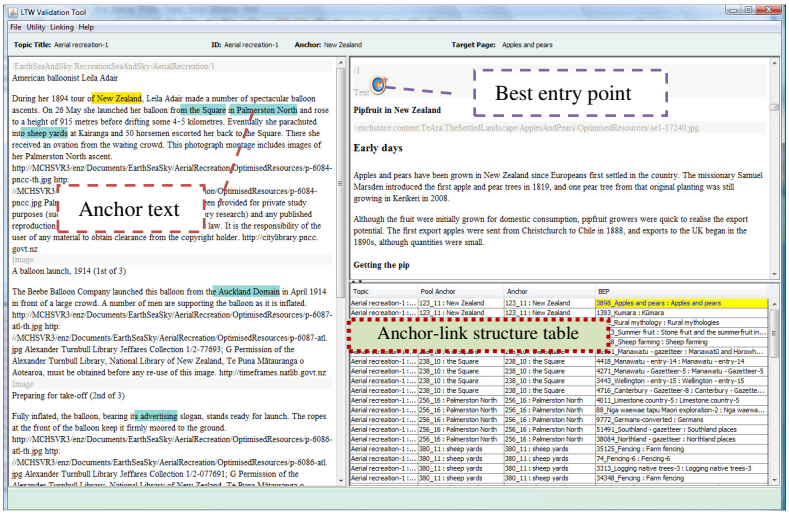

Fig. 2. The Validation Tool

The validation tool was introduced in 2009 and delivered to the participants so as to self-verify their link discovery submissions (see Figure 2). The anchors are highlighted in the left screen while the right screen shows the link content with a best entry point on it and a table recording the hierarchical structure of anchor-links for the given topic. The participants can click on a link in the table to check the particular anchor-link result. This tool intends to bring up what the link discovery application should look like and facilitate to revise the linking results. It can also be seen as a pre-assessment process.

4 Preparation of qrels

There are two types of qrels used for the evaluation of the link discovery results. One is the Wikipedia ground truth and the other is generated from the manual assessment set. The Wikipedia ground truth is derived from the existing links in the Wikipedia collection. This is a simple way to achieve the automatic evaluation. However, the experiments undertaken in the past 2 years have shown that the comparative evaluation using automatic qrels is unsound in terms of the users' point of view. Some Wikipedia links are topically-obsolete or redundantly assigned. Many of anchors are linked to the documents with the same name. The relevant portions of the document content have not been further discovered. All relevant contents that are not in the Wikipedia are also considered non-relevant for the evaluation. As a consequence, the evaluation result might appear either optimistic or pessimistic. However, evaluation based on the Wikipedia ground-truth does measure performance relative to what is present, and so it is reasonable to use it in comparisons.

Apart from the file-to-file ground truth, Wikipedia can also produce the anchor-to-file ground truth. The offset value is set to the very beginning of the document. Although the Wikipedia does contain anchor-to-bep links, in practice they are rarely used. In order to experiment the anchor-to-bep technology, a special pooling procedure was applied to collect all anchors and links from participants' runs and Wikipedia. The pool for each topic was generated by the following three parts: anchor-to-bep (*A2B*), the file-to-file link discovery on A2B (*F2FonA2B*), and anchor-to-file Wikipedia ground truth. Since not all the offset-length sets were specified preciously and anchor texts could be indicated by different ways, overlapped anchor texts (i.e. offset-length) were merged as *a pool anchor* or *anchor representative*. For example, *quantum theory of atomic motion in solids* is an anchor in the article of *Albert Einstein*. However, *quantum theory*, *atomic* and *atomic motion* could be anchors returned by different participants. Therefore, the anchor texts shown to the assessor on the screen might not be the anchor returned by the system; instead it could be a combined anchor representative. In the case of *F2FonA2B*, the anchor was set as the topic title and linked to the beginning of the target document. The anchor-to-file set from the Wikipedia presents a one-to-one relation and the bep was set at the very beginning of the document. The pool was assessed to completion. The evaluation was expected to carry out at different levels: file-to-file, anchor-to-file, file-to-bep and anchor-to-bep.

5 Assessment and Evaluation

5.1 Manual Assessment

As the assessment is laborious and time consuming we have designed the assessment tool to maximize assessor efficiency. The assessment tool can be seen in Figure 3. Either the anchor representative or the bep link could be identified relevant (or non-relevant). Once the anchor representative was assessed as non-relevant, all

anchors and associated links inside this anchor representative became non-relevant. The relevance status could be simply assigned by mouse right or left click. If the target document of the outgoing link was assessed as relevant, the best entry point was indicated by mouse left double-click. Incoming links in the submission were not properly explored in 2009. Most of them were specified in the file-to-file manner, i.e. incoming document title to the beginning of the topic article. Assessing incoming links was achieved for the first time in 2009.

According to the survey carried out after the assessment, a lack of related anchor texts highlighted in the incoming document could be a major obstacle to efficiency. Sometimes it is difficult to identify whether the incoming document is relevant to the topic content or not. Indicating the best entry point in the target document is also a difficult task to achieve without any supplemental information (e.g. system's discovered bep). Highlighting anchor texts or related phrases on the document seems necessary. For instance, a sub-title or a paragraph paired with the linking anchor text (or related phrases) could be a best start point for reading from. Each topic contains around 1000 anchor links and 900 incoming links. A log was created to record all the activities during the assessment. Then time to completion of a topic was estimated at around 4 hours.

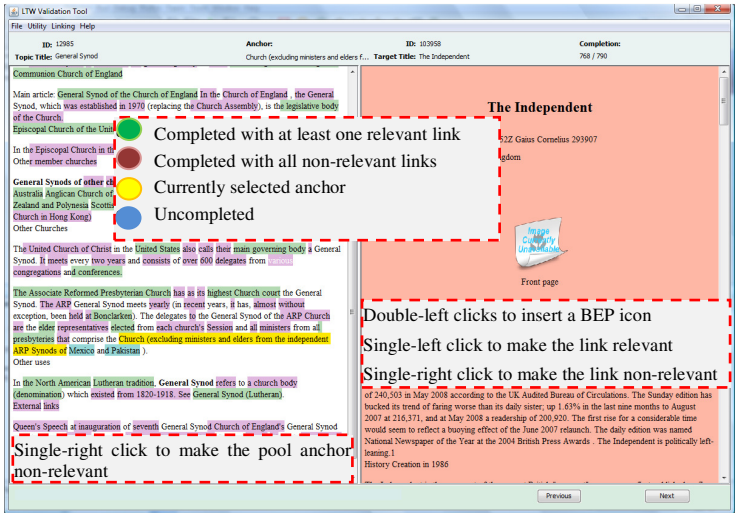

Fig. 3. The Assessment Tool

5.2 Metrics

As with all metrics, it is important to first define the use-case of the application. The assumption at INEX is that link-discovery is a recommendation tasks. The system produces a ranked list of anchors and for each a set of recommended target/bep pairs. The list should also be comprehensive because it is not clear that the document author can know a priori which links will be relevant to a reader of the document. That is, link discovery is a recall oriented task. The Mean Average Precision based metrics are

very good at taking rank into account and are recall oriented. A good metric for link discovery should, consequently, be based on MAP. The difficulty is computing the relevance of a single result in the results list. For evaluation purposes it is assumed that if the target is relevant and the anchor overlaps a relevant anchor then the anchor is relevant; $f_{anchor}(i) = 1$.

The assessor might have assessed any number of documents as relevant to the given anchor. If the target of the anchor is in the list of relevant document then it is considered relevant; $f_{doc}(i) = 1$. The contribution of the links' bep is a function of distance from the assessor's bep [4]:

$$f_{bep}(j) = \begin{cases} \dfrac{n - 0.9 \times d(x,b)}{n} & if\ 0 \le d(x,b) \le n \\ 0.1 & if\ d(x,b) > n \end{cases}$$

Where $d(x,b)$ is the distance between submission bep and result bep in character. Therefore, the score of $f_{bep}(j)$ varies between 0.1 (i.e. d is greater than n) and 1 (i.e. the submission and result beps are exactly matched). The score of 0.1 is reserved for the right target document with an indicated bep not in range of n. n typically is set up as 1000 (characters). The score of a result in the results is then:

$$P = \left[(f_{anchor}(i)) \times \frac{\left(\sum_{j=1}^{m} \left(f_{doc}^{i}(j) \times f_{bep}^{i}(j)\right)\right)}{m_i} \right]$$

Where m is the number of returned links for the anchor and m_i is the number of relevant links for the anchor in the assessments. As the result list is restricted to 5 targets per anchor m_i is capped at 5 for evaluation. A perfect run can thus score a MAP of 1.

5.3 Evaluation

Based on the portable evaluation tool, *ltwEval*, used in 2008, new functionality has been added to achieve a better interaction of the graphs and additional evaluation setup, which increase the usability of the tool. Numerous evaluation metrics including precision, recall, MAP, and precision@R were used to evaluate submission at different levels of linking. Different runs can be evaluated and easily compared to each other via the tool. Interpolated-Precision/Recall graphs can be produced for sets of run.

For the file-to-file evaluation (i.e. F2F and F2FonA2B), the number of outgoing and incoming links have been restricted by 250. Links beyond this number were truncated. The total number of relevant links is based on the ground truth, but at last 250 to make sure the measurement of Recall is meaningful. For the anchor-to-bep evaluation against ground-truth, the first 50 anchors for each topic were taken and the first link from each anchor was collected. As a result, there were 50 outgoing links per topic, used for evaluation. By contrast, first 250 incoming links were taken to do the evaluation since the discovery of bep in the topic document is not that obvious. Most incoming links belong to the same bep. Therefore, in the INEX use case of link discovery it is important to rank the discovered links for presentation to the page

author. This use case was modeled in the manual assessment 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.

6 Results and Discussion

The Queensland University of Technology (i.e. QUT) submitted 6 runs for the file-to-file (F2F) task, 4 runs for the anchor-to-bep (A2B) task and 1 run for the F2FonA2B task. University of Waterloo contributed 2 runs on the A2B task and 5 run for the F2FonA2B task. University of Amsterdam had 5 runs for the A2B task. University of Otago submitted 1 runs for the F2F task, 2 runs for the A2B task and 2 runs for the F2FonA2B task. University of Wollongong submitted 4 runs for the F2F task. Technische Universität Darmstadt contributed 4 runs on the F2F task. Apart from the Link the Wiki tasks, QUT also participated in the Link the Te Ara and Link Te Ara to Wiki tasks by submitting 1 run each. Technische Universität Darmstadt also contributed 5 runs on the Link the Te Ara task. These runs were generated by the anchor-to-bep link discovery technology.

The University of Waterloo (UW) had two approaches, one baseline and the other link-based, to undertake the experiment. For a baseline, UW produced the statistics of the phrase frequency. These phrases were located in the topic files and the most frequent links were returned. For incoming links, we scored the corpus using topic titles as query terms and returned the top documents. The link-based approach computes *PageRank* and *Topical PageRank* values for each file in the corpus for each topic, and returned the top scoring pages according to the contribution of *K-L divergence*. For incoming links, UW reversed the graph to get new *PageRank* values and returned the top pages according to the contribution of *K-L divergence* with the new *PageRank* values and the old *Topical PageRank* values.

The Queensland University of Technology (QUT) used the statistical link information of Wikipedia corpus to calculate the probability of anchors and their corresponding target documents for a list of sortable outgoing links. A hybrid approach that combines the results of link analysis method and title matching algorithm for the prediction of potential outgoing links was also undertaken. For the incoming links, the top ranking search results with topic title as the query terms retrieved from a BM25 ranking search engine were chosen as source documents that can be linked to the topics. In finding the beps for either outgoing or incoming links, QUT tried two different methods: one is that the bep is the position of the phrase in the target document where the terms of the anchor, either the entire words or part of which, appear; the other is that the best entry point is the beginning of a text block which has similar terms features with that of the passage which is extracted from the surrounding text of the anchor in source document.

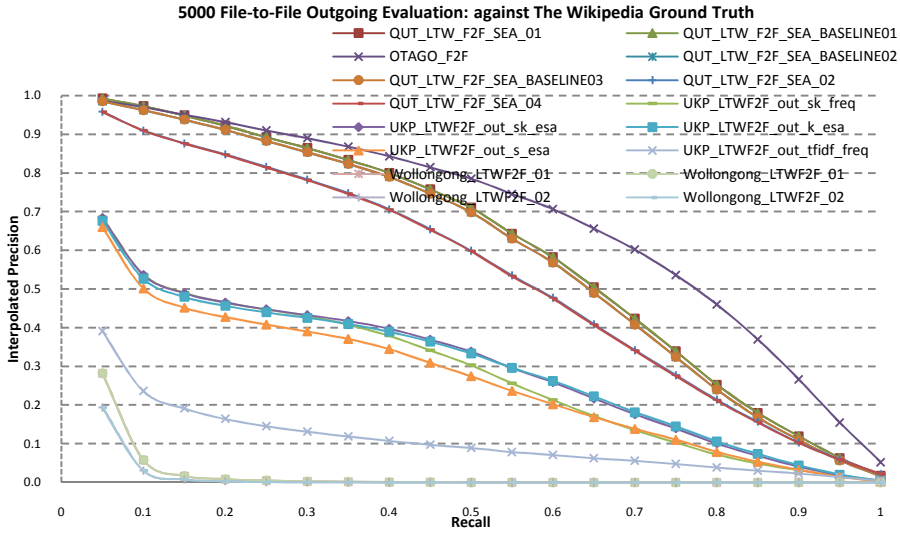

Fig. 4. 5000 F2F Topics Outgoing link discovery evaluated against Wikipedia Ground Truth

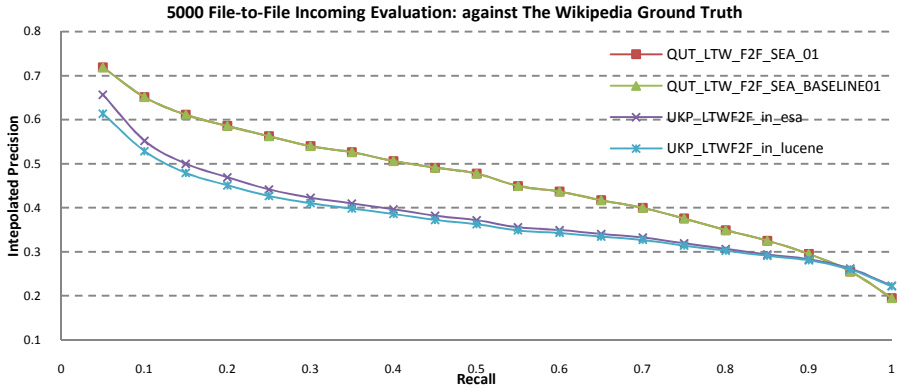

Fig. 5. 5000 F2F Topics Incoming link discovery evaluated against Wikipedia Ground Truth

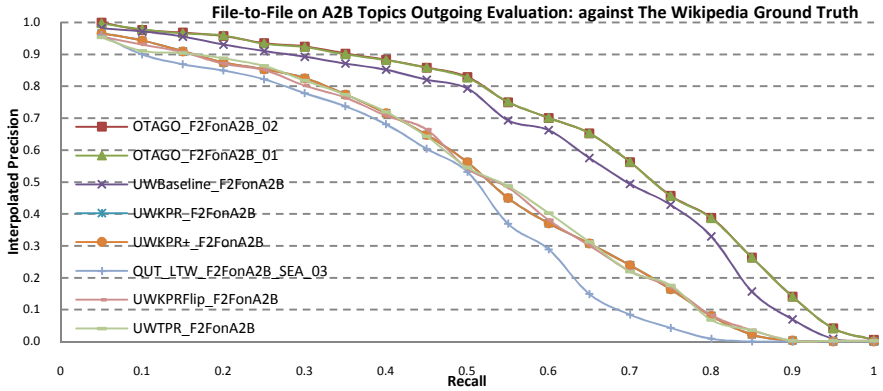

Fig. 6. F2F on A2B Topics Outgoing links evaluated against Wikipedia Ground Truth

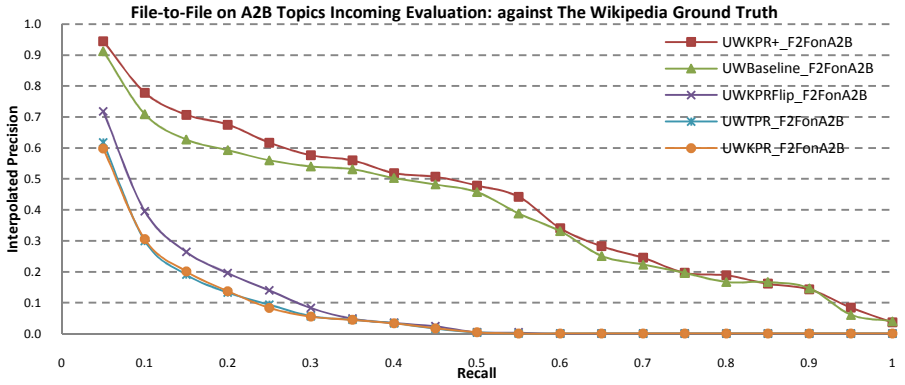

Fig. 7. F2F on A2B Topics Incoming links evaluated against Wikipedia Ground Truth

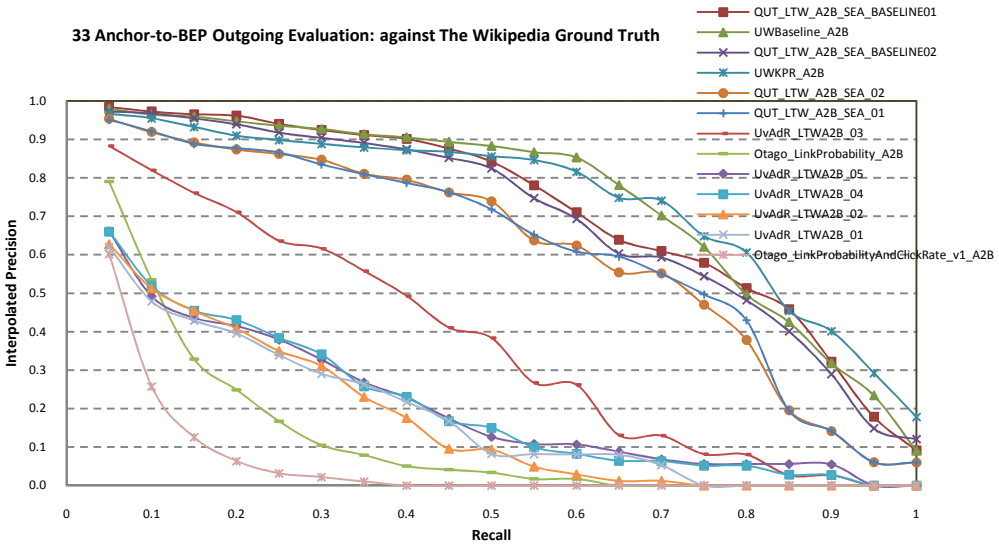

Fig. 8. 33 A2B Topics Outgoing links evaluated against Wikipedia Ground Truth

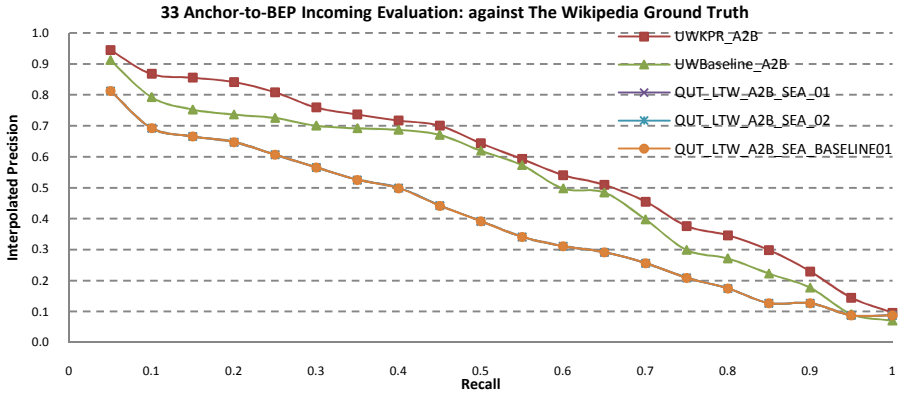

Fig. 9. 33 A2B Topics Incoming links evaluated against Wikipedia Ground Truth

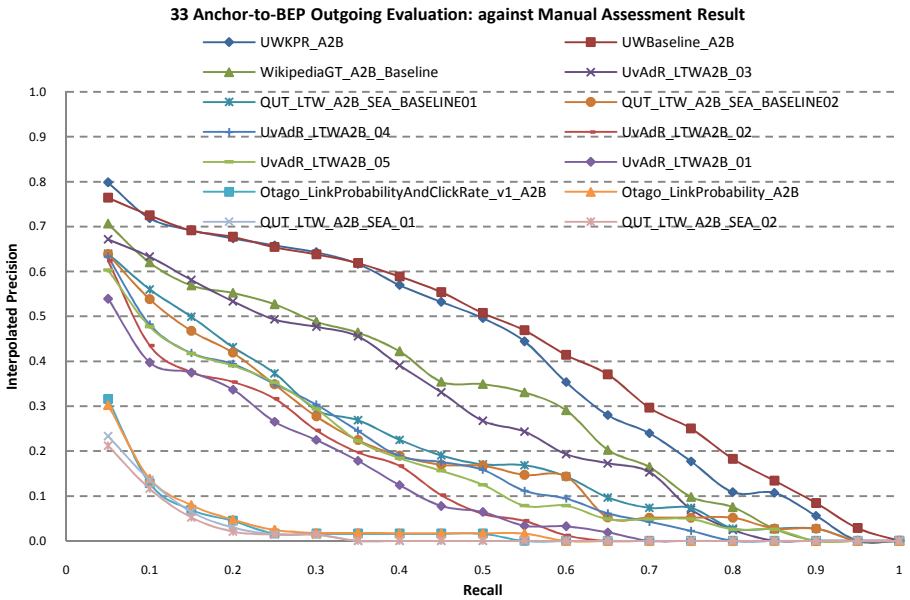

Fig. 10. 33 A2B Topics Outgoing links evaluated against Manual Assessment Set

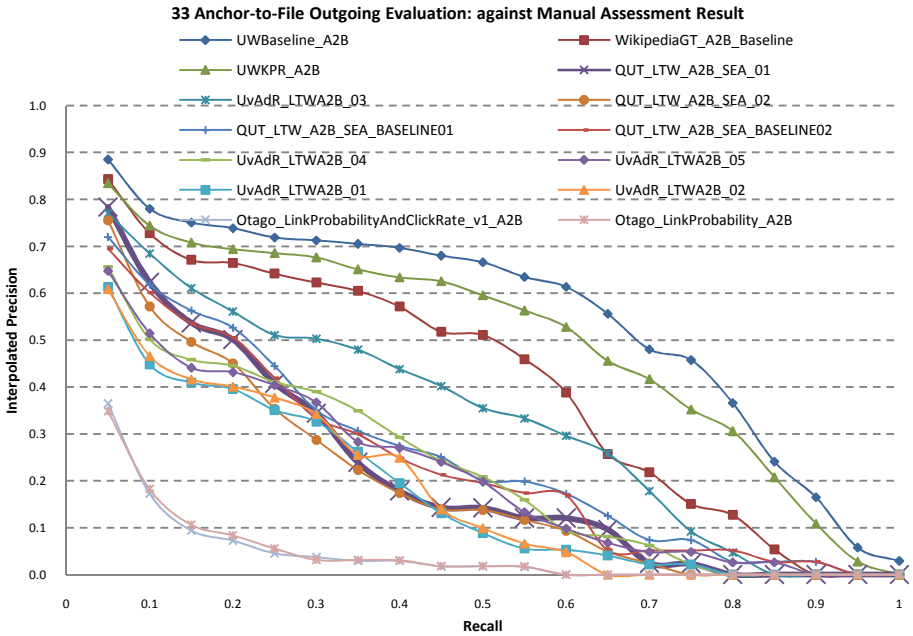

Fig. 11. 33 A2F Topics Outgoing links evaluated against Manual Assessment Set

7 Conclusion and Outlook

This is the third year of the Link-the-Wiki track at INEX. According to the file-to-file experiment, producing Wikipedia links could be achieved by current approaches. In 2009, the focus has been shifted to the anchor-to-bep link discovery and several changes have been made to improve the evaluation procedure. Assistant tools were prepared to self-examine the status of submission. The outcome is twofold: self-verification of the submission to revise the offset-length parser and pre-assessment to improve the link discovery engine. Further experiments were undertaken on the anchor-to-bep runs. The submission was evaluated on anchor-to-file, and anchor-to-bep level to test the usability of approaches provided. This aims to classify the performance of each approach on the contribution of linking for the given topic. The Te Ara collection is introduced for the first time at INEX to bring up the new concept of cross collection link discovery. Through the focus link discovery, the Wikipedia content could be fully explored. Anchors indicated for the given document could be linked to the most relevant content in the collection. Every piece of content discovered in the Wikipedia can be used to provide links for anchors from other document collections. Going through this process, a well defined knowledge network can be constructed. Based on participants' comments and ideas via survey, customization can be made, and the enhancement of evaluation procedure and efficiency is expected. According to the experiment, the contribution of each approach can be classified and future direction of anchor-to-bep link discovery can be possibly pointed out.

References

1. Trotman, A. and Geva, S. (2006) 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., Geva, S. and Trotman, A. (2009) Overview of INEX 2008 Link the Wiki Track, INEX 2008, LNCS 5631, N. Fuhr et al. (Eds.), pp. 314-325.
4. Huang, W. C., Xu, Y., Trotman, A. and Geva, S. (2009) The Methodology of Manual Assessment in the Evaluation of Link Discovery, In Proceedings of the 14th Australian Document Computing Symposium.