Overview of the INEX 2010 Link the Wiki Track

Andrew Trotman[1], David Alexander[1], Shlomo Geva[2]

[1]Department of Computer Science
University of Otago
Dunedin
New Zealand

[2]Faculty of Science and Technology
Queensland University of Technology
Brisbane, Australia

Abstract. The INEX 2010 Link-the-Wiki track examined link-discovery in the Te Ara collection, a previously unlinked document collection. Te Ara is structured more a digital cultural history than as a set of entities. With no links and no automatic entity identification, previous Link-the-Wiki algorithms could not be used. Assessment was also necessarily manual. In total 29 runs were submitted by 2 institutes. 70 topics were assessed, but only 52 had relevant target documents and only 45 had relevant links in the pool. This suggests that the pool was not diverse enough. The best performing run had a MAP of less than 0.1 suggesting that the algorithms tested in 2010 were not very effective.

1 Introduction and Motivation

Keeping a rich hypermedia document collection such as the Wikipedia up-to-date is problematic. Each time a new document is added new links from that document into the collection are needed (and vice versa). Each time a document is deleted links to that document must be deleted. If a document changes then the links must be updated. This served as the motivation for the Link-the-Wiki track in 2007.

Deletion of a document is a simple maintenance problem – simply remove link from documents that refer to the document being deleted. Update is analogous to deletion then reinsertion. Consequently the entire link-the-wiki problem can be examined from the perspective of document insertion. In fact, this is the approach taken by the track from 2007-2009. Choose a set of documents from within the collection; remove them from the collection; then measure the efficacy of automatic link discovery algorithms on those documents as if each were a new addition to the collection. These topic-documents were referred to as orphans.

As early as 2007 two effective algorithms were seen, Geva's algorithm [1] and Itakura's algorithm [2].

Geva's algorithm builds an index of titles of documents in the collection and searches for those titles in the topic-document. The algorithm is effective in the

Wikipedia because Wikipedia documents are about entities and their names are essentially canonical and if they appear in the text they are probably accurate.

Itakura's algorithm builds an index of links in the collection and orders these on the proportion of times the anchor-text seen as a link to the number of times it is seen as a phrase anywhere in the collection. This is essentially the strength of the anchor-text as an anchor and the most likely target. The algorithms has proven problematic to implement –several implementations were seen at INEX 2009 but Trotman's 2008 implementation remains the highest performing [3].

In 2007 the performance of algorithms was measured against the links that were in the collection before the document was orphaned. This was the same approach taken by others (for example Milne & Witten [5]). Results from evaluation in this way show extremely high performance. Otago, for example, achieved a MAP score of 0.734. Milne & Witten [5] see similar such scores using their machine learning approach.

In 2008 and 2009 TREC-style manual assessment was performed. Runs were pooled and manually assessed (to completion) for accuracy. As a twist on the experiment the links present in the Wikipedia itself were added to the pools. The evaluation showed that Geva's algorithm, Itakura's algorithm and the Wikipedia were performing comparably and that MAP scores in previous years had been inflated by non-relevant links present in Wikipedia articles. It is not known which links those are or why there are there, but it has been speculated that the links might themselves be put in by bots using similar algorithms to those of Milne & Witten, Geva, and Itakira.

During the running of the track the organizers became aware of an alternative link discovery scenario. The New Zealand Ministry for Culture and Heritage has an encyclopedia-like collection (Te Ara) that when complete "will be a comprehensive guide to the country's peoples, natural environment, history, culture, economy, institutions and society". It does not contain links between articles.

Linking Te Ara is more complex than linking the Wikipedia for many reasons. The articles are often digital narratives and in being so do not represent entities – the controlled vocabulary of the Wikipedia is not present. Geva's title-matching algorithm is unlikely to be effective. There are no links in the collection and so the machine learning algorithms of Milne & Witten[5] and of Itakura & Clarke [2] can't be used.

2 Te Ara Test Set

The document collection used in 2010 was the 2010 dump of Te Ara. It is a single 48MB XML file consisting of 36,715 documents. The task was to link each and every document. As such the topic set was the document collection itself.

After runs were submitted a set of documents were chosen for manual assessment. First, the collection was ordered on the number of links in each document. This was then divided into 10 deciles. Finally, one work-set was built by randomly selecting one document from each decile. Seven such (non-ovelapping) work sets were assessed to completion resulting in a total of 70 assessed documents.

3 Runs

In total 29 runs were submitted by 2 institutes. QUT submitted 5 runs and Otago submitted 24 runs.

3.1 Submission Format

Results were submitted in the 2009 Link-the-Wiki Te Ara format, except that certain elements and attributes were made optional:

- The root element, <inexltw-submission>, had attributes for the participant's numeric ID, the run ID and the task (LTeAra).
- The <details> element give information about the machine on which the results were produced, and how long it took to produce them. This element was optional.
- The <description> gave an explanation of the linking algorithm.
- The <collections> element contained a list of document collections used in the run.
- Each topic was in a <topic> element which contained an <anchor> element for each anchor-text.
- One or more <tobep> elements, within each <anchor> element, gave the offset and the target document ID for the link. If the offset was specified, the target document ID was optional because the offsets were relative to the single XML file containing the collection. If no offset was specified it was assumed to be the start of the document.
- Each topic could contain up-to 50 anchors, and each anchor could contain up-to 5 BEPs (each in different target documents).

Example.
An example of a submission is:

```
<inexltw-submission participant-id="12"
 run-id="Otago_LTeAraA2B_01"
 task="LTeAra">
 <details>
 <machine>
 <cpu>Intel Celeron</cpu>
 <speed>1.06GHz</speed>
 <cores>1</cores>
 <hyperthreads>1</hyperthreads>
 <memory>128MB</memory>
 </machine>
 <time>3.04 seconds</time>
 </details>
 <description>
 Describe the approach here, NOT in the run-id.
 </description>
```

```
 <collections>
 <collection>TeAra_2010_Collection</collection>
 </collections>
 <topic file="9638" name="Matariki - Maori New Year">
 <outgoing>
 <anchor offset="7445748" length="8" name="balloons">
 <tobep offset="7952293">10151</tobep>
 <tobep offset="10553520">12991</tobep>
 <tobep offset="11686141">14270</tobep>
 <tobep offset="8016276">10208</tobep>
 <tobep offset="7226359">9363</tobep>
 </anchor>
 </outgoing>
 </topic>
</inexltw-submission>
```

DTD.
The DTD for the submission format was:

```
<!ELEMENT inexltw-submission (details, description,
collections, topic+)>
<!ATTLIST inexltw-submission
 participant-id CDATA #REQUIRED
 run-id CDATA #REQUIRED
 task (LTAra_A2B) #REQUIRED>

<!ELEMENT details (machine|time)>

<!ELEMENT machine
(cpu|speed|cores|hyperthreads|memory)>
<!ELEMENT cpu (#PCDATA)>
<!ELEMENT speed (#PCDATA)>
<!ELEMENT cores (#PCDATA)>
<!ELEMENT hyperthreads (#PCDATA)>
<!ELEMENT memory (#PCDATA)>

<!ELEMENT time (#PCDATA)>

<!ELEMENT description (#PCDATA)>

<!ELEMENT collections (collection+)>
<!ELEMENT collection (#PCDATA)>

<!ELEMENT topic (outgoing|anchor+)>
<!ATTLIST topic
```

```
  file CDATA #REQUIRED
  name CDATA #IMPLIED>

<!ELEMENT outgoing (anchor+)>

<!ELEMENT anchor (tobep+)>
<!ATTLIST anchor
  name CDATA #IMPLIED
  offset CDATA #REQUIRED
  length CDATA #REQUIRED>

<!ELEMENT tobep (#PCDATA)>
<!ATTLIST tobep
  offset CDATA #REQUIRED>
```

4 Assessment Tool

A new cross-platform assessment tool was written in C/C++ using SQLite and GTK+. This tool was written in an effort to reduce the assessment time and this increase in the number of topics that could be assessed in a "reasonable" period of time. Of course, the assessment of "incoming" links was not necessary in 2010 as the entire collection was linked.

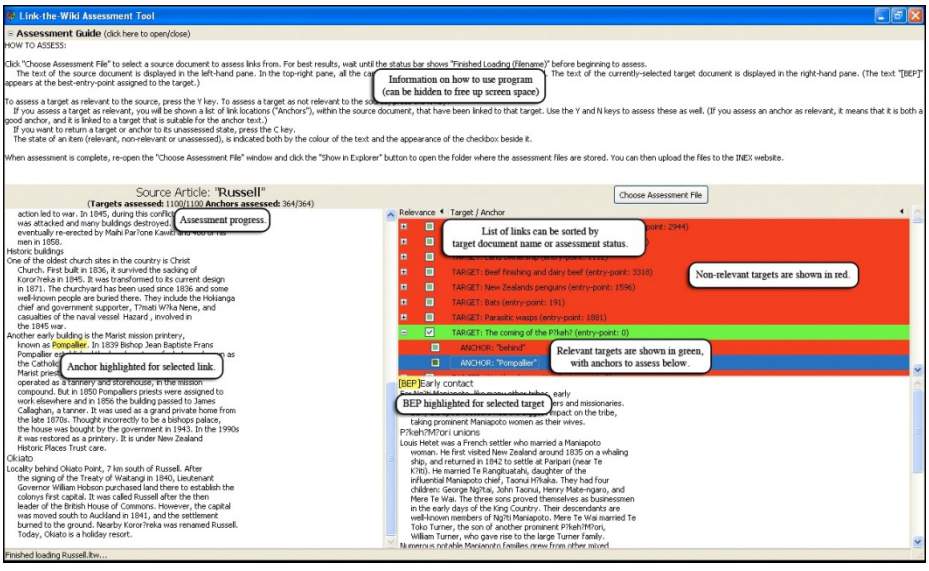

Figure 1: 2010 Link-the-Wiki assessment tool

A screenshot of the 2010 assessment tools is shown in Figure 1. Brief assessing instructions are given in the top window. On the left is source document with the

current anchor highlighted in yellow. On the right at top is the list of links from the pool with assessed-relevant links marked in green and assessed-non-relevant links marked in red (un-assessed had a white background). On the right at the bottom is the target document with the best-entry-point (BEP) marked in yellow.

Several changes were made to the assessment method. First, the assessor is shown the target document (and BEP). If this was not relevant then all anchors to that document must be non-relevant. This made it possible to assess several links in one click. If the target document was relevant then the assessor was shown all anchors for that document and asked to assess which were relevant. In this way the assessor was simply choosing whether the given anchor was an accurate way to link the two documents.

The 2010 assessment tool does not ask the assessor to choose a BEP in the case were the document is relevant but the BEP was badly placed. There were several reasons for this decision, the strongest of which was that the results from the INEX ad hoc track BEP experiments show that BEPs are usually at (or very close to) the start of the given document (311.5 characters in 2009 [4]); but also because the pool was (supposed to be) assessed to completion and so the usefulness of the BEP could be determined by the assessor.

5 Metrics

As is the informal convention at INEX, the metrics for the Link-the-Wiki track in 2010 were not published before the runs were submitted. As is also the informal convention, the metric changed in 2010.

In a Link-the-Wiki run it is possible (and correct) to identify more than one anchor targeting the same document. It is also possible and correct to identify more than one target per anchor. Consequently metrics based on recall (such as un-interpolated Mean Average Precision (MAP)) are meaningless. If there is only one relevant target document, but the link-discovery algorithm identifies two different anchors for that target then what is the recall? Exactly this happens in this very document, the reference to Itakura's algorithm and to the paper by Itakura & Clarke are different anchors for the same document. This also happens in the submitted runs. The runs were, consequently, de-duplicated by taking on the highest ranking instance of a target for a given topic and ignoring other instances of the target.

The relevance of each target was then determined using the manual assessments. Then the score for each position in the results list was 1 if any target for that anchor was relevant and 0 otherwise. Mean un-interpolated Average Precision is then computed from this.

This approach gives a highly optimistic evaluation because the run has 5 trials at each point in the results list and if any one trial is correct the run scores a relevant hit. It is also optimistic because the anchor and BEP are not considered (if multiple BEPs are seen in the assessments then if any-one is relevant the document is relevant). It also guarantees that recall cannot exceed 1.

6 Results

In total 70 topics were assessed. 18 of these had no relevant target documents in the pool. The mean number of relevant targets per topic was 8.8 and the mean number of non-relevant targets per topic was 274.6. Topic 2919 had the most relevant targets (97). Figure 2 shows the distribution of the number of relevant targets documents per topic (solid line) and the number of relevant anchor-target pairs in the pool (dotted) ordered from most relevant targets to least. In some cases the number of relevant targets far exceeds the number of relevant links (for example, topic 25591 has 27 relevant targets but only one relevant anchor-target pair was in the pool). In cases where the number of relevant anchor-target pairs exceeds the number of relevant target documents (such as topic 15765) the assessor has identified more than one anchor as relevant to the same target document (in this case 66 relevant links to 62 relevant documents). On average there are 5.5 relevant links per document and only 11 topics have more 10 relevant links in them.

Figure 3 shows the performance of all the runs submitted to the track. The best Otago run was Otago_LTeAraA2B_11 with a MAP of 0.0906. The best QUT run was QUT_LTeAraA2B_DH3 with a MAP of 0.0863. In the figure the MAP of each run is shown in brackets after the run name. As can be seen, the performance of the runs is not comparable to the high scores seen in the link-the-wiki track in previous years. Early precision is low and drops quickly – recall also that the evaluation is overly optimistic because the anchors are not evaluated (it is equivalent to previous year's file-to-file evaluation) and that the score at each point in the results list is the best of the 5 possible targets seen there.

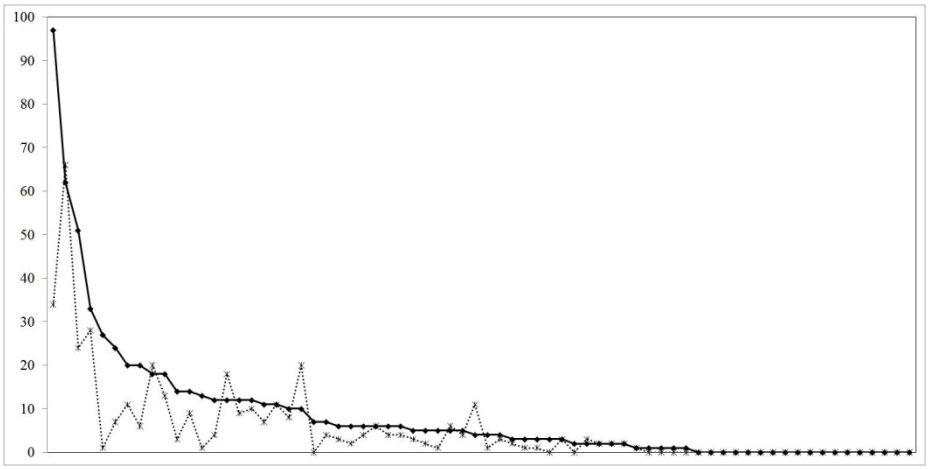

Figure 2: Relevant targets per topic

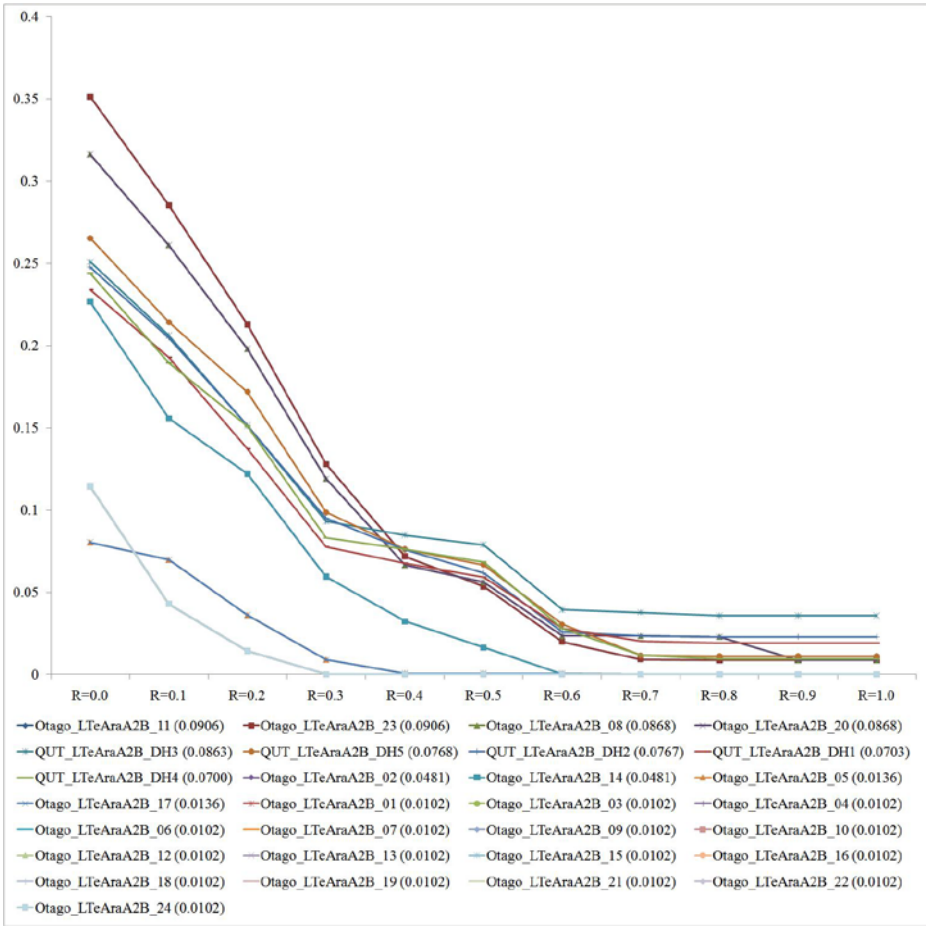

Figure 3: Precision / Recall of all submitted runs

7 Conclusions and Further Work

The INEX 2010 Link-the-Wiki track changed document collection from the Wikipedia to Te Ara. Te Ara is a far more difficult document collection because there are no prior links (so Itakura's algorithms cannot be used), and the document titles are not canonical entity names (so Geva's algorithm cannot be used). As a consequence of the increased difficulty the number of groups participating dropped to just two (Otago and QUT).

In total 70 topics were manually assessed, but of those the assessors found only 52 with relevant target documents and only 45 with relevant links in the pool. This suggests that both the pool was not diverse enough and that the algorithms tested were not effective enough. This is echoed in the evaluation where the best scoring optimistic MAP is less than 0.1 and the runs perform poorly at all recall points.

Further work on Link-the-Wiki has already started at NTCIR with the crosslink track. That track is using a multi-lingual dump of the Wikipedia to examine algorithms that discover links from orphans in one language to targets in another language. This interested in further link-discovery research are referred to NTCIR.

8 References

[1] Geva, S., *GPX: Ad-Hoc Queries and Automated Link Discovery in the Wikipedia*, in *Focused Access to XML Documents*. 2007, Springer-Verlag. pp. 404-416.

[2] Itakura, K.Y. and C.L. Clarke, *University of Waterloo at INEX2007: Adhoc and Link-the-Wiki Tracks*, in *Focused Access to XML Document*. 2007, Springer-Verlag. pp. 417-425.

[3] Jenkinson, D., K.-C. Leung, and A. Trotman, *Wikisearching and Wikilinking*, in *Advances in Focused Retrieval*. 2009, Springer-Verlag. pp. 374-388.

[4] Kamps, J., S. Geva, and A. Trotman, *Analysis of the INEX 2009 ad hoc track results*, in *Focused retrieval and evaluation*. 2009, Springer-Verlag.

[5] Milne, D. and I.H. Witten, *Learning to link with wikipedia*, in *Proceeding of the 17th ACM conference on Information and knowledge management*. 2008, ACM: Napa Valley, California, USA. p. 509-518.