A Virtual Evaluation Forum for Cross Language Link Discovery

Wei Che (Darren) Huang
Faculty of Science and Technology
Queensland University of Technology
Brisbane, Australia
w2.huang@student.qut.edu.au

Andrew Trotman
Department of Computer Science
University of Otago
Dunedin, New Zealand
andrew@cs.otago.ac.nz

Shlomo Geva
Faculty of Science and Technology
Queensland University of Technology
Brisbane, Australia
s.geva@qut.edu.au

ABSTRACT
In this position paper we propose to extend the current INEX Link-the-Wiki track, based on the English Wikipedia, to a Cross Language Link Discovery (CLLD) track using the multi-lingual Wikipedia. We observe that the existing automatic evaluation methods used at INEX do not necessitate manual assessment as assessments are extracted directly from the collection and performance is measured relative to this ground-truth. The proposed track can therefore run online with continuous evaluation, free from the difficulties of scheduling and synchronizing groups of collaborating researchers. We also propose to continually publish peer-reviewed evaluation results online with open access.

Categories and Subject Descriptors
D.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval – *Search process.*

General Terms
Measurement, Performance, Experimentation.

Keywords
Link-Discovery, Cross Language Information Retrieval.

1. INTRODUCTION
Since the inception of TREC in 1992 interest in IR evaluation has increased rapidly and today there are numerous active and popular evaluation forums. It is now possible to evaluate a diverse range of information retrieval methods including: ad-hoc retrieval, passage retrieval, XML retrieval, multimedia retrieval, question answering, cross language retrieval, link discovery, and learning to rank, amongst others. Most forums are tied to a long evaluation cycle which includes a workshop, symposium, or conference at the end of each cycle.

In this position paper we propose a new *virtual evaluation track*: Cross Language Link Discovery (CLLD). The track will initially examine cross language linking of Wikipedia articles. This virtual track will not be tied to any one forum; instead we hope it can be tied to each of (at least): CLEF, NTCIR, and INEX as it will cover ground currently examined at each.

We suggest automatic as well as collaborative manual assessment of submissions. With automatic and manual assessment a continual evaluation and publication forum for CLLD is possible.

2. MOTIVATION
On the welcome page of the Wikipedia we see that different language versions of the Wikipedia have different numbers of articles. At the time of writing the Maori version has about 6,500 articles whereas the English version has about 2,800,000 articles. In the cases where an article exists in both languages a bilingual reader might prefer a Maori article to an English one. This preference should "travel" with the user as they navigate around the Wikipedia, and links to articles should be given in the user's preferred language. To achieve this it is necessary to support cross lingual links in the Wikipedia (and profiles, multiple links per anchor, and so on).

Overell [3] shows that the geographic coverage of the Wikipedia very much depends on the language version – places in the UK are best covered by the English language Wikipedia while places in Spain are best covered by the Spanish language version. Despite the geographic proximity of Spain to England, a search for the village *Wylam* in the Spanish version reports *No hay coincidencias de título de artículo.* The English language version, however, informs us that the village is the birth place of George Stephenson, the inventor of the Stephensonian locomotive (the "modern" steam train). Wylam is historically interesting to railway enthusiasts, regardless of nationality – so much so that the Spanish Wikipedia article on George Stephenson shows a link for Wylam in red (the page does not yet exist). Perhaps it should link to the English article.

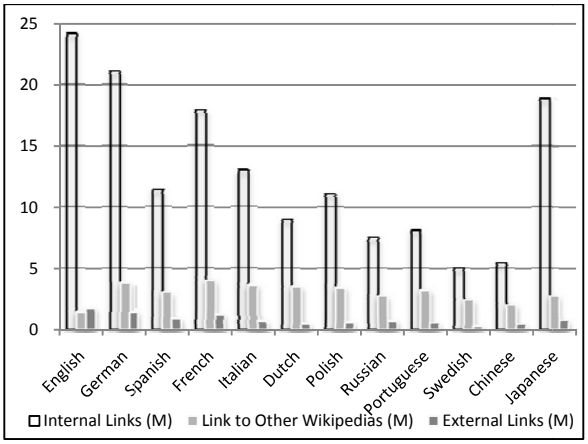

Figure 1: Cross-lingual link structure of the Wikipedia. MediaWiki provide English stats from Oct 2006 while others are from Dec 2008.

These two use-cases demonstrate a need for cross-language links within the Wikipedia. WikiMedia provide some statistics showing that there are already many cross language links. The statistics are summarized in Figure 1 where it can be seen that about a quarter of the Chinese links are to other languages (many Chinese articles

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

link to English pages, 诺森伯兰郡, for instance, links to "List of United Kingdom Parliament constituencies"). English articles, however, are not well linked to other languages.

3. TASK DEFINITION

We propose a Cross Language Link Discovery (CLLD) track run as a collaboration between INEX, CLEF, and NTCIR. Initially two linking experiments will be examined:

MULTILINGUAL topical linking is a form of document clustering – the aim is to identify (regardless of language) all the documents in all languages that are *on the same topic*. The Wikipedia currently shows these links in a box on the left hand side of a page.

BILINGUAL anchor linking is exemplified by the Chinese article 诺森伯兰郡, having a link from the anchor 国会选区 to the English article "List of United Kingdom Parliament constituencies". The link discovery system must identify the anchor text in one language version of the Wikipedia and the destination article within any other language version of the Wikipedia.

4. STATIC EVALUATION

When Trotman & Geva [4] introduced the Link-the-Wiki track at INEX they noted that, technically at least, the evaluation required no human assessment. The same is true with cross-language link discovery.

Topics in the INEX Link-the-Wiki track are chosen directly from the document collection. All links in those documents are removed (the documents are orphaned). The task is to identify links for the orphans (both to and from the collection). Performance is measured relative to the pre-existing links.

For MULTILINGUAL linking the links on the left hand side of the Wikipedia page could be used as the ground truth. The performance could be measured relative to the alternate language versions of the page already known to exist.

BILINGUAL anchor linking from one document to another could also be automatically evaluated. Links from the pre-orphan to a destination page in an alternate language would be used as the ground truth – but there are unlikely to be many such links.

A same-language link from a pre-orphan to a target provides circumstantial evidence that should the target exist in multiple languages then the alternate language versions are relevant. This is essentially a triangulation: $A \xrightarrow{t} B \xrightarrow{l} C \Rightarrow A \xrightarrow{tl} C$ where *A*, *B*, and *C* are documents; and *t* designates a topical link, *l* a cross language link, and *tl* a topical cross language link.

Static assessment requires no human interaction. A web site with orphan sets, assessment sets (extracted from the pre-orphans), and evaluation software, can support a sound evaluation methodology which does not necessitate any fixed deadlines.

5. CONTINUAL EVALUATION

Huang *et al*. [1] question automatic evaluation. Their investigation suggests that many of the links in the Wikipedia are not topical, but are trivial (such as dates), and that users do not find them useful. Manual assessment is, consequently, necessary. This challenges cross language link discovery because finding assessors fluent in multiple languages is difficult – especially for a track with a relatively small number of participants but in a large number of languages (the Wikipedia has 266 languages).

We propose a novel form of evaluation called *continual evaluation* in which participants can download topics and submit runs at any time; and in which manual assessment is an on-going concern. The document collection will, initially, be static. Topics will either be chosen at random from the collection, or nominated by participants. For any given run a participant will download a selection of topics and submit a run. The evaluation will be based on metrics that consider the un-assessed document problem (such as a variant on rank-biased precision [2]), and comparative analysis will be relative to an incomplete, but growing, assessment set.

To collect manual assessments two methods are proposed: first, in order to submit a run the participant will be required to assess some anchor-target pairs in languages familiar to them; second, we will run an assessment Game With A Purpose (GWAP). Kazai et al. used a GWAP for the INEX Book track; Von Ahn & Dabbish [5] discuss GWAPS in other contexts (including the Google Image Labeler). Regardless of the method of assessment collection, we are trying to validate the minimum number of links necessary to disambiguate the relative rank order of the runs (within some known error).

6. PUBLICATION

Both automatic and manual assessment of cross language link discovery can be performed on a continual rolling basis; there is no need for topic submission deadlines, run deadlines, assessment deadlines, or even publication deadlines. At INEX the time difference between run-submission and the workshop paper submission date is long (6 July – 23 Nov). With automatic assessment it is possible to achieve a result, write, and publish a paper with a short turn around. As part of the virtual track we propose an open-access virtual CLLD workbook to which registered participants can submit their papers for peer review and publication.

7. CONCLUSIONS

We put the case for an online virtual track that examines Cross Language Link Discovery in the Wikipedia. Such a track can be *continual* because the assessments are drawn from the collection itself. To facilitate the exchange of results we propose a virtual open-access workbook to which participants can submit papers. We believe this virtual forum will better serve the link-discovery community than the existing calendar based evaluation forums.

REFERENCES

[1] Huang, W.C., A. Trotman, and S. Geva, *The Importance of Manual Assessment in Link Discovery*, in *SIGIR 2009*. 2009, ACM Press: Boston, USA.

[2] Moffat, A. and J. Zobel, *Rank-biased precision for measurement of retrieval effectiveness.* ACM Trans. Inf. Syst., 2008. 27(1):1-27.

[3] Overell, S.E., *Geographic Information Retrieval: Classification, Disambiguation and Modelling*, in *Department of Computing*. 2009, Imperial College London: London. p. 175.

[4] Trotman, A. and S. Geva. *Passage Retrieval and other XML-Retrieval Tasks*. in SIGIR 2006 Workshop on XML Element Retrieval Methodology. 2006. Seattle, USA.pp. 43-50

[5] von Ahn, L. and L. Dabbish, *Designing games with a purpose.* Commun. ACM, 2008. 51(8):58-67.