DCU at NTCIR-10 Cross-lingual Link Discovery (CrossLink-2) Task

Shu Chen
CLARITY: Centre for Sensor Web Technologies / CNGL
Dublin City University
Dublin 9, Ireland
shu.chen4@mail.dcu.ie

Gareth J. F. Jones
CNGL, School of Computing
Dublin City University
Dublin 9, Ireland
gjones@computing.dcu.ie

Noel E. O'Connor
CLARITY: Centre for Sensor Web Technologies
Dublin City University
Dublin 9, Ireland
Noel.OConnor@dcu.ie

ABSTRACT

DCU participated in the English to Chinese (C2E) and Chinese to English (C2E) subtasks of the NTCIR 10 CrossLink-2 Cross-lingual Link Discovery (CLLD) task. Our strategy for each query involved extracting potential link anchors as n-gram strings, cleaning of potential anchor strings, and anchor expansion and ranking to select a set of anchors for the query. Potential anchors were translated using Google Translate, and a standard information retrieval technique used to create links between anchors and the top 5 ranked retrieved items selected as potential links for each anchor. We submitted a total of four runs for E2C CLLD and C2E CLLD. We describe our method and results for file-to-file level and anchor-to-file level evaluation.

Team Name

DCU

Subtasks

Chinese to English, English to Chinese

Keywords

Cross-lingual Link Discovery, automatic anchor selection, information retrieval, Wikipedia,

1. INTRODUCTION

The cross-lingual link discovery (CLLD) task aims to detect potentially important semantic links between documents in different languages. The task requires participants to extract textual terms from source documents as anchors which are then linked to other documents in an alternative language. DCU participated in the Chinese to English (C2E) and English to Chinese (E2C) CrossLink-2 subtasks at NTCIR 10 [7].

The general approach taken in our participation was to extract n-grams from the query source document as potential anchors, and then to filter these to remove strings which are unlikely to be valid words or form the basis of meaningful links. Potential anchors were then expanded to include related descriptive terms using Wikipedia Miner[1]. Expanded potential anchor n-grams are then compared with the title of the source document, ranked based on semantic similarity measures using Wikipedia Miner, and up to 250 terms then selected as potential anchor points. The potential anchors were then translated into the target linking language, and used as input queries to retrieve a set of Wikipedia pages, with the top 5 ranked pages being selected as potential links for each anchor.

This remainder of this paper is organised as follows: Section 2 reviews background materials on the analysis of Wikipedia, Section 3 describes the implementation of DCU's cross-lingual link system, Section 4 describes our evaluation result, and Section 5 concludes the paper.

2. BACKGROUND MATERIAL

The online Wikipedia encyclopedia forms a tremendous resource describing a huge number of diverse topics in multiple languages. Versions of pages in different languages are not direct translations but are generally written by a native speaker of the language from their cultural perspective on the topic. The links between topics in Wikipedia mean that it forms a complex and valuable knowledge resource. Analysis of Wikipedia enables the relatedness of concepts to be measured, although this is only meaningful if the concepts are sufficiently richly described, so as to match descriptions of concepts to which they are being compared, and they have been suitability contextually disambiguated.

A number of strategies can be utilized to use Wikipedia as a knowledge resource by implementing algorithms into software to mine Wikipedia on-line content [3]. Examples of such software include: Wikipedia Miner [3] and Yago [4]. Alternatively researchers could start from scratch and build their own algorithms to mine Wikipedia directly.

Some projects, such as DBpedia[2], provide structured content from the information created as part of the Wikipedia project. In [2], the authors describe the term "text wikification" as the task of automatically extracting the most important words and phrases in a document, and for each such keyword identifying an appropriate link to a Wikipedia article. Wikipedia Miner is an open-source toolkit designed to carry out wikification that allows users to avoid the extensive effort needed to mine Wikipedia's rich content [3]. Wikipedia Miner can process an XML Wikipedia dump file,

[1] http://wikipedia-miner.cms.waikato.ac.nz

[2] http://dbpedia.org/

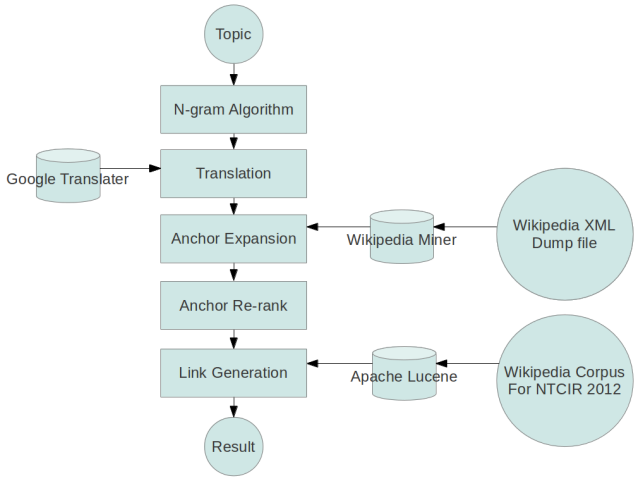

Figure 1: DCU's CLLD system design

a single XML file containing the full content of Wikipedia captured at a certain point, into a run-once extraction to generate a series of flat-file summaries of Wikipedia's structure. Developers and researchers can then process these structured summaries. Wikipedia Miner also uses a model package to simplify access to stored data by wrapping it with easy to understand classes [3]. Moreover, it provides a comparison package to measure concept relatedness and an annotation package for text annotation and disambiguation.

3. APPROACH DESCRIPTION

This section describes our method for carrying out CLLD in CrossLink-2. It begins with an overview of our n-gram extraction and filtering method to form a set of potential anchors, and then describes anchor ranking which selects a set of query anchors, anchor translation and link construction. Figure 1 summarizes the overall architecture of our system.

3.1 N-gram Entity Extraction

The n-gram entity extraction stages aim to select potential anchors which to be used as sources in CLLD. An anchor consists of a word or sequence of words extracted from a given text. The anchor can refer to a person's name, an organization, a place, a definition, etc. In NTCIR 10, the source or query document is in the form of a Wikipedia webpage. Each of these is a structured document consisting of a title, article body, and references. It is specified that an extracted anchor must not include any tag information.

The absence of word boundary markers in Chinese text means that identification of potential anchor words is non-trivial. One approach to this would be to apply a complex morphological segmentation process. However, we opted for a simpler approach based on extraction of arbitrary n-grams and subsequent filtering which seeks to eliminate n-grams which are unlikely to be meaningful or useful.

In the initial stage each word in the English documents and each character in the Chinese documents was indexed with its zero-based offset by counting the number of bytes from

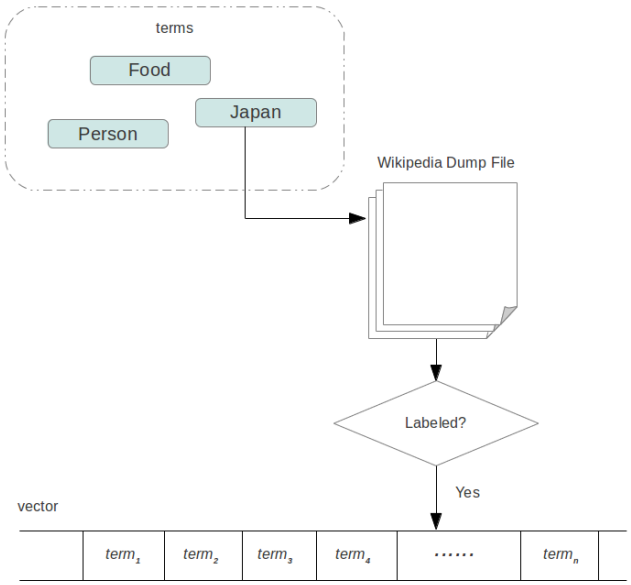

Figure 2: Anchor Selection

the beginning of the source document. The body text of the source document with the corresponding index data was next extracted and all XML markup was removed using regular expressions. A stop punctuation list was used to detect the boundary of each sentence, stop markers included: period, comma, exclamation mark, semicolon, brackets, etc. Finally, all overlapping n-grams of multiple lengths in each sentence were extracted. The number of terms in an extracted anchor, N, was assigned from 1 to 5 (words) for English documents, and from 2 to 5 (characters) for Chinese documents [1].

In order to eliminate n-gram strings which are unlikely to be useful, the following strategies were applied:

1. Regular expressions were used to remove all entities in date and time format.
2. English and Chinese stop words were removed. The English stop word list was taken from the University of Glasgow[3] and the Chinese stop word list was taken from Baidu[4].

3.2 Selection of Potential Anchors

The remaining strings output by the n-gram extraction algorithm are mostly not meaningful as cross-lingual link anchors. The next anchor selection process aims to choose meaningful n-gram terms which can then be regarded as potential anchors for further processing.

In order to eliminate non-meaningful n-gram strings, we assumed that meaningful strings would be present in existing Wikipedia pages. The anchor selection phase created a vector containing all meaningful terms extracted from

[3] http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words
[4] http://wenku.baidu.com/view/b8b30382e53a580216fcfeb7.html

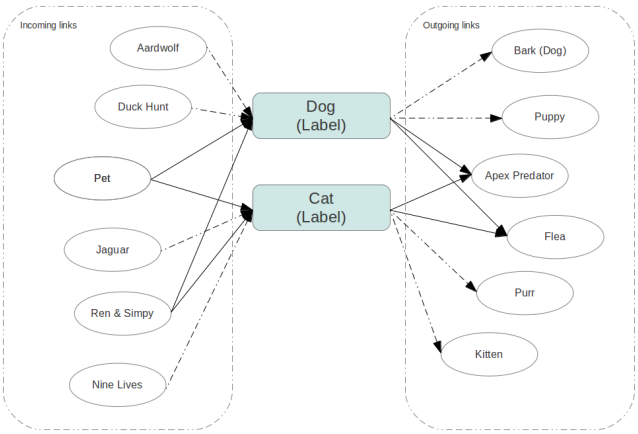

Figure 3: Incoming and outgoing links related to labels in Wikipedia Miner

the Wikipedia query pages, as shown in Figure 2. The Wikipedia Miner toolkit was used to determine whether an n-gram term was available from the Wikipedia pages by applying the Wikipedia Miner toolkit. Wikipedia Miner uses an indexed data structure created from Wikipedia XML dump files[5]. This enables several structured elements associated with Wikipedia pages to be checked [3]. Since checking all elements in the Wikipedia pages for the presence of each of our potential anchor n-gram strings was computationally demanding, we checked only structured elements, defined as *Label*, which are the titles of Wikipedia pages from the dump file. All n-grams containing positive label elements matching results were added to the output vector for the query Wikipedia page.

3.3 Filtering of Potential Anchors

The anchor filtering phase selects anchors from the vector created in the potential anchor selection phase, to be used as the link anchor set for this Wikipedia query page. In accordance with the task definition [7], the maximum number of anchors for each Wikipedia page was set to 250. Where there are more than 250 potential anchors, the anchors in the vector were ranked based on their semantic relatedness to the title of the query Wikipedia page. Since many of the title and potential anchors can be ambiguous, and calculating semantic relatedness between single words or short phrase is difficult, the anchor filtering process included two stages: anchor expansion and anchor re-ranking.

3.3.1 *Anchor expansion*

Anchor expansion aims to provide a more detailed description of the current anchor to enrich the semantic definition. For each potential anchor, Wikipedia Miner can be used to identify incoming links and outgoing links associated with each label element (Wikipedia document title).

Wikipedia Miner was used to integrate the label element with its incoming and outgoing links. An incoming link is a link which a user could follow to reach the current Wikipedia document, while an outgoing link is a link which points to another document related to an anchor in the current document. Figure 3 shows an example of incoming and outgoing links for two simple labels. We use the label element in Wikipedia Miner to expand the semantic meaning of a potential anchor. This is done by selecting the top 5 incoming and outgoing links within the Wikipedia dump file according to their frequencies.

From Figure 3, we can see that given two labels representing two anchors, there is some overlap between incoming and going links. Wikipedia Miner uses these sets of common and distinct links to generate features, and combines them by using a classifier trained over a manually defined ground truth. Two features are used in the calculation of the semantic similarity between the two labels, intersection size and union size. The similarity measure is calculated as shown in Equation 1 [3].

$$similarity(a,b) = \frac{\log max(|A|,|B|) - \log(|A \cdot B|)}{\log |W| - \log min(|A|,|B|)} \quad (1)$$

where a and b are two labels (anchors) of the Wikipedia resource, A and B are the sets of incoming and outgoing links from label a and b.
Using more than 5 links in each case would make the next stage of the calculation too computationally demanding. The expanded label (potential anchor) was taken as an enriched description of the potential anchor.

3.3.2 *Anchor Re-ranking*

The final stage of the anchor selection process of the anchor selection process was to re-rank anchors by computing the similarity between the title of the current query Wikipedia page and each element in the vector of expanded potential anchors using Wikipedia Miner again using Equation 1. The top 250 ranked n-gram terms were then selected as the anchors of the current query document.

3.4 Link Construction

Each extracted anchor was then individually translated into the target language for linking using Google Translate[6]. Links for translated anchors were then calculated as follows.

The Apache Lucene[7] information retrieval toolkit was used for link construction. This involved two steps: indexing and searching. The target Wikipedia document set was indexed into Lucene with standard *tf-idf* ranking. Each translated anchor was then applied to the indexed Wikipedia set producing a ranked output of Wikipedia pages. The top 5 ranked Wikipedia pages for each anchor were then taken as the potential links for this anchor.

4. EVALUATION

NTCIR 10 CLLD evaluated the participating systems using two benchmarks: file-to-file (F2F) level based on Wikipedia ground-truth and anchor-to-file (A2F) level based on manual

[5] `http://wikipedia-miner.cms.waikato.ac.nz/wiki/Wiki.jsp`

[6] `https://translate.google.ie/`

[7] `http://lucene.apache.org/core/`

Table 1: F2F evaluation results with Wikipedia ground truth: LMAP, R-Prec, P@N

RUN	LMAP	R-Prec	P@5	P@10	P@20	P@30	P@50	P@250
DCU-E2C-A2F-001-NW	0.045	0.129	0.112	0.156	0.168	0.145	0.118	0.043
DCU-E2C-A2F-002-NWE	0.045	0.129	0.112	0.156	0.168	0.145	0.118	0.043
DCU-C2E-A2F-003-NW	0.011	0.049	0.096	0.104	0.082	0.064	0.047	0.012
DCU-C2E-A2F-004-NWE	0.011	0.049	0.096	0.104	0.082	0.064	0.047	0.012

Table 2: F2F evaluation results with manual assessment results: LMAP, R-Prec, P@N

RUN	LMAP	R-Prec	P@5	P@10	P@20	P@30	P@50	P@250
DCU-E2C-A2F-01-NW	0.045	0.129	0.016	0.024	0.028	0.029	0.023	0.020
DCU-E2C-A2F-02-NWE	0.045	0.129	0.016	0.024	0.028	0.029	0.023	0.020
DCU-C2E-A2F-03-NW	0.005	0.019	0.024	0.016	0.024	0.023	0.021	0.019
DCU-C2E-A2F-04-NWE	0.005	0.019	0.024	0.016	0.024	0.023	0.021	0.019

Table 3: A2F evaluation results with manual assessment results: LMAP, R-Prec, P@N

RUN	LMAP	R-Prec	P@5	P@10	P@20	P@30	P@50	P@250
DCU-E2C-A2F-01-NW	0.006	0.002	0.000	0.000	0.000	0.001	0.004	0.007
DCU-E2C-A2F-02-NWE	0.006	0.002	0.000	0.000	0.000	0.001	0.004	0.007
DCU-C2E-A2F-03-NW	0.005	0.004	0.000	0.008	0.008	0.008	0.005	0.008
DCU-C2E-A2F-04-NWE	0.005	0.004	0.000	0.008	0.008	0.008	0.005	0.008

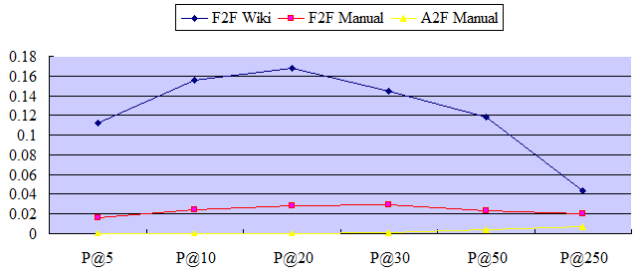

Figure 4: *DCU-E2C-A2F-02-NWE* results on F2F with Wikipedia ground truth, F2F with manual assessment results, and A2F with manual assessment results

assessment results [7]. The evaluation metrics were *LMAP*, *P@N*, and *R-Prec* [6] [5].

DCU submitted 4 runs for the CLLD task: *DCU-E2C-A2F-01-NW* and *DCU-E2C-A2F-02-NWE* for E2C, and *DCU-C2E-A2F-03-NW* and *DCU-C2E-A2F-04-NWE* for C2E. Table 1 and Table 2 show F2F evaluation results with Wikipedia ground truth and manual assessment results. Table 3 shows A2F evaluation results with manual assessment results. Figure 4 shows the performance of *DCU-E2C-A2F-02-NWE* on different evaluation benchmarks.

The official evaluation results show that our results are good on the F2F evaluation benchmark with Wikipedia ground truth, but lower for the A2F benchmark with manual assessment results. According to [7], the evaluation benchmark defines the relevance of an anchor as semantically related to a document. Our strategy uses Wikipedia Miner to disambiguate a term by involving the incoming and outgoing links of corresponding term. The disambiguation phase may ignore a term's context. The lack of semantic analysis used in the anchor expansion stage could reduce performance on manual benchmarks which can reflect human perception of relatedness and relevance.

Figure 4 shows that the *P@N* value increases with *N*. The best value is around P@20 and P@30. The result shows that our ranking algorithm gives high scores to general terms, which are less relevant or irrelevant to the current topic. For example, *sushi*, as a Japanese food, may be regarded as related to the word *Japan*. However, [7] defines that a good anchor should be something not just strongly related conceptually, but also interesting and semantically related. Our strategy which focuses only on potential anchors which are available in Wikipedia resources may miss anchors which are not present as the title of another Wikipedia page.

5. CONCLUSIONS

This paper has described details of DCU's participation in the NTCIR 10 CLLD task. We submitted results for the E2C and C2E subtasks. The system comprised n-gram entity extraction, anchor selection, translation and link construction. An n-gram algorithm was used to create potential anchors. Using Wikipedia Miner, our system sought to reduce potential problems of anchor ambiguity and to extract anchors related to the current Wikipedia query page. In future work, we plan to seek improvement in CLLD performance by analysing anchor context to disambiguate terms and selecting anchor expansion terms more judiciously.

6. ACKNOWLEDGMENTS

This work is funded by the European Commission's Seventh Framework Programme (FP7) as part of the AXES project (ICT-269980).

7. REFERENCES

[1] C.-Y. Chang, Y.-C. Wang, and R.T.-H.Tsai. IISR Crosslink Approach at NTCIR 9 CLLD task. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 469–472, Tokyo, Japan, 2011.

[2] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings*

of the sixteenth ACM conference on Conference on information and knowledge management, CIKM '07, pages 233–242, New York, NY, USA, 2007. ACM.

[3] D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 194(0):222 – 239, 2013. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

[4] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203 – 217, 2008. World Wide Web Conference 2007Semantic Web Track.

[5] L.-X. Tang, S. Geva, A. Trotman, Y. Xu, and K. Itakura. In *9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retreival, Question Answering and Cross-Lingual Information Access*.

[6] L.-X. Tang, K. Itakura, S. Geva, A. Trotman, and Y. Xu. In *The Fourth International Workshop on Evaluating Information Access*.

[7] L.-X. Tang, I.-S. Kang, F. Kimura, Y.-H. Lee, A. Trotman, S. Geva, and Y. Xu. Overview of the ntcir-10 cross-lingual link discovery task. In *Proceedings of NTCIR-10*, 2012.