Osaka Kyoiku University at NTCIR-10 CrossLink-2

- Link Filtering by Title Tag of Corpus as a Dictionary -

SATO, Takashi
Information Processing Center
Osaka Kyoiku University
4-698-1 Asahiga-oka
Kashiwara, Osaka, JAPAN
+81-72-978-3823
sato@cc.osaka-kyoiku.ac.jp

ABSTRACT

Our group (OKSAT) submitted two types of runs named SMP and REF for every subtasks of NTCIR-10 Cross-lingual Link Discovery (CLLD). Our method uses titles in Wikipedia pages (corpus) of source language as a entries of a dictionary, so no external dictionary is required. For SMP, we aimed to discover cross-lingual links of actual Wikipedia, in other words it targets Wikipedia ground truth. For REF, on the other hand, we aimed to discover as much meaningful cross-lingual links as possible automatically.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering, Query formulation, Retrieval models, Search process, Selection Process.*

General Terms

Experimentation, Performance, Measurement.

Keywords

Wikipedia, Cross-lingual Link Discovery, CrossLink-2, Anchor text, String search, Gram.

Team Name

OKSAT

Subtasks

Chinese to English, Japanese to English, Korean to English, English to Chinese, English to Japanese, English to Korean

1. INTRODUCTION

Our group submitted two types of runs for every subtasks of NTCIR10 closslink-2 [1], in other words 12 runs in total. In this paper, we introduce our approach (how to extract candidate anchor texts, filtering them, and getting cross-lingual link). And then, we describe experimental results, comments on what we concerned, and subject of further researches.

2. OUR APPROACH

Our group submitted two types of runs named SMP and REF for every subtasks of NTCIR-10 Cross-lingual Link Discovery (CLLD). SMP corresponds to our submitted runs OKSAT-(CJK2E|E2CJK)-A2F-01-SMP. And REF corresponds to OKSAT -(CJK2E|E2CJK)-A2F-01-REF. Our method uses titles in Wikipedia pages (corpus) of source language as entries of a dictionary, so no external dictionary is required.

For SMP, we aimed to discover cross lingual links of actual Wikipedia in the net, in other words it targets Wikipedia ground truth. For REF, on the other hand, we aimed to discover as much meaningful cross lingual links as possible automatically. Figure 1 shows process to get anchor text and cross-lingual link in our experiment.

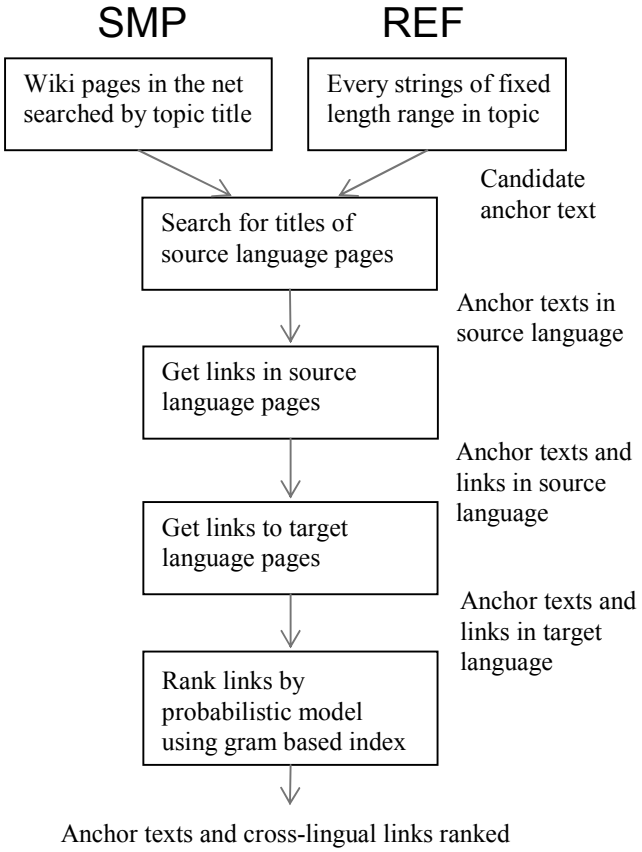


Figure 1. Process to get anchor texts and cross-lingual links.

2.1 Extracting Candidate Anchor Text

SMP and REF are different in this step.

SMP get the candidate anchor texts as anchor text of Wiki page (written in source language) in the Internet retrieved by title of each topic.

On the other hand, REF makes the candidate anchor texts of every string (grams [2]) in the fixed range length from each topic body. Figure 2 shows example grams when the length ranges from one to three.

topic body : ABCD

candidate anchor text : A, AB, ABC, B, BC,
BCD, C, CD, D

Figure 2. Example of topic body and candidate anchor texts generated.

2.2 Filtering Candidate Anchor Text

For both type of runs, the candidate anchor texts are filtered by the following two steps.

[step1] They should mach a title of a Wikipedia pages in source language corpus.

[step2] Then there are links from the Wikipedia page of step1 to pages of target language corpus.

2.3 Getting Cross-lingual Link

In step1 of **2.2**, links in matched page are extracted also. Then in step2, we get links to the pages of target language corpus as cross-lingual links.

2.4 Ranking Link

Our REF runs find many cross-lingual links. So we now consider ranking retrieved links. In order to rank anchor text (word) of links by probabilistic model [5], the indices to retrieve anchor text not only in topics but also in corpus quickly are inevitable. Indices for corpus of each language are made by 'variable length gram encoding in fixed byte' method [2-4].

In preparation we examine distribution of UTF-8 characters [6] in each corpus. Set of characters (alphabet) which appear more than 6 times (for CJK) or 11 times (for English) in each corpus is coded by Huffman coding [7]. In other words, very rare characters are ignored because they do not seemed to be retrieved as anchor text of topics. Then we made grams in fixed 6 byte length from characters coded.

3. EXPERIMENTAL RESULTS

The programs for SMP and REF are constituted along with the flow shown in Figure 1. The parameter of fixed length range in REF is set between 6 to 60bytes, i.e. 2 to 20 characters of UTF-8 3byte code. We omit length one character strings from candidate anchor texts because many meaningless texts for anchors are extracted at present.

The environment of experiment is shown in Table 1, and the execution time for ‘extracting candidate anchor text’, ‘filtering candidate anchor text’, and ‘getting cross-lingual link’ (**2.1-2.3**) about SMP and REF is shown in Table 2.

Table 1. Execution environment.

CPU	Intel Corei5-3570@3.40GHz 4C/4T
Memory	DDR3-1600 4Gx2
HDD	SATA300 500GB 7200rpm 16MBbuf
OS	FreeBSD 8.3
Programming Language	C, Perl 5.12

Table 2. Execution time (sec).

subtask	SMP	REF
C2E	0.22	2.9
J2E	0.38	5.6
K2E	0.13	1.6
E2C	0.38	8.8
E2J	0.40	8.8
E2K	0.38	8.8

Concerning **2.4**, indexing is shown in Table 3. It shows alphabet size, index size, indexing time, and average/max gram length for each language. The gram length is short when Huffman code for rare characters is included, on the contrary it is long when every characters in gram are popular in corpus.

Compared to other language (CJK), the indexing time for English is much longer. Because, in addition to the difference of corpus size, English characters are expressed in 1 byte and most CJK characters are expressed in 3 byte in UTB-8, so the number of English characters is 3 times more than that of CJK per byte.

Table 3. Statistics of indices.

language	alphabet size	index size (GB)	time (min.)	gram length average/max
Chinese	18,065	2.13	98	3.58 / 12
English	11,845	24.81	1156	3.73 / 15
Japanese	12,116	5.66	250	3.17 / 11
Korean	12,105	1.10	58	3.61 / 11

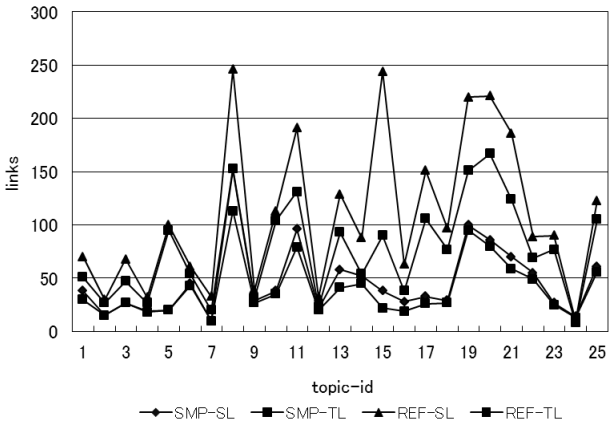
4. LINKS AND ANCHORS

Submitted run REF (OKSAT-(CJK2E|E2CJK)-A2F-01-REF) is aimed to discover as much meaningful cross-lingual links as possible automatically. So, we are interested in the number of links and anchor texts.

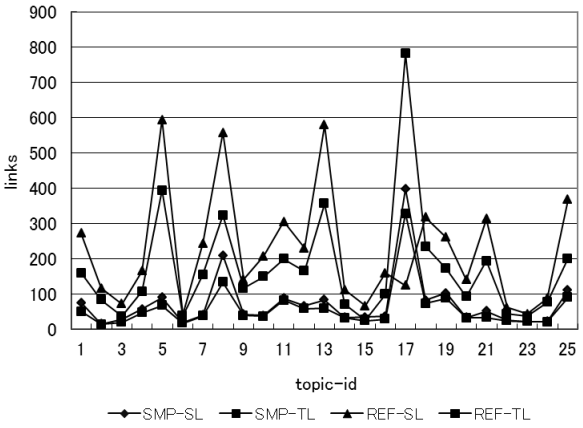
4.1 Topic by Topic

Figure 3(a), (b), and (c) show the number of links of SMP and REF in CJK2E subtask. In this figure, topic-id (column '#' in Table 3 of [1]) is in horizontal, and SL (the number of links of before 'Get links to target language pages' in Figure 1) and TL (that of after) are in vertical. It turns out that the number of links

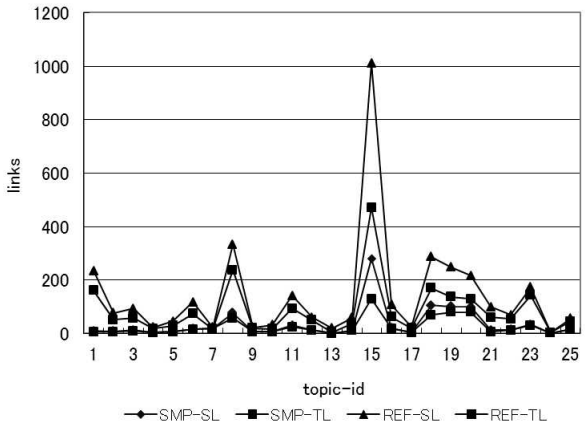
in REF is more than that of SMP in any topic. From SL (links in source language) to TL (links in target language), we know how these links are decreased.



(a) C2E



(b) J2E



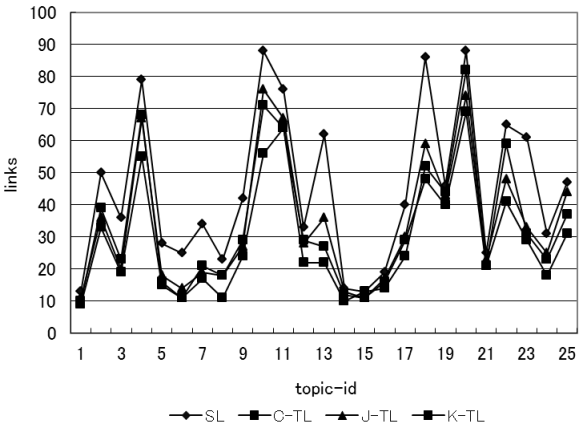
(c) K2E

Figure 3. Links run by run.

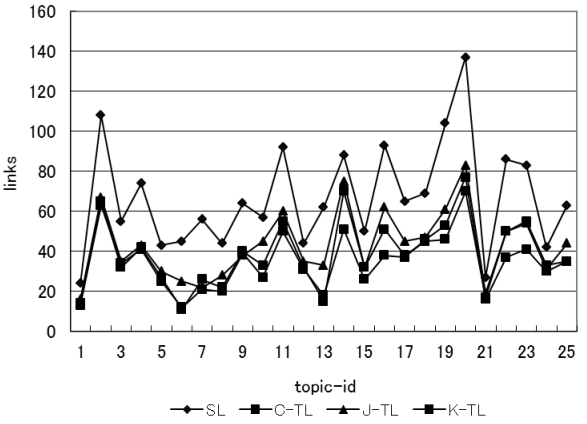
Peaks of the number of links are different by source languages. For example, the peaks are recognized at #8, #11, #19, #20 of topic-id in C2E, however, they are at #5, #8, #13, #17 in J2E and at #8, #13, #18 in K2E. Because, even if topic-id is the same, the length of wiki page for topic is different by source languages.

Figure 4(a) and (b) show the number of links of SMP and REF in E2CJK subtask. Because SL is only for English, TLs of CJK are put in a same figure for SMP and REF.

The differences of the number of cross-lingual links by target languages are not so large as the size of corpus of their languages.



(a) SMP



(b) REF

Figure 4. Links run by run for E2CJK.

4.2 Subtask by Subtask

Figure 5 shows the total number of links of qrels assessed by organizer, that is A2BWikiManualResultSet-(CJK2E|E2CJK).xml (MAN in this fugure) and REF over topics for each subtasks.

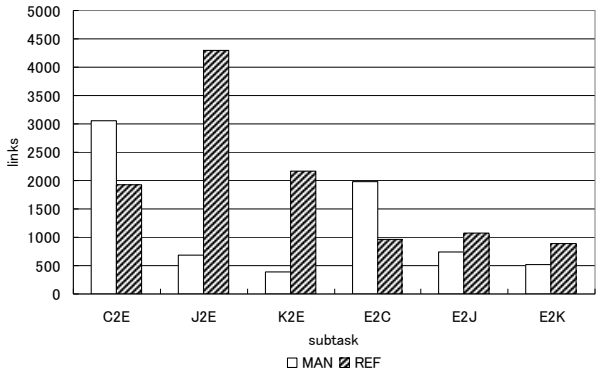


Figure 5. Links subtask by subtask.

REF is about 6 times more than MAN in J2E subtask. For ja-topic 9543.xml (title is 猪八戒), the anchor text of REF is following 75 words (with no duplication).

文字,四大奇書,小説,西遊記,キャラクター,台湾,中国語,ブタ,イノシシ,意味,逸話,朝鮮,近似,発音,名字,皇帝,天の川,設定,摩利支天,将軍,刑罰,人間,妖怪,群れ,殺し,妖怪,彼女,天竺,経典,観音菩薩,菩薩,慈悲,意志,精進料理,家業,世間,玄奘三蔵,孫悟空,一行,三蔵,普通,沙悟浄,仏教,煩悩,般若,飲む,シーン,義務,コミ,カル,性格,太鼓,怪力,食欲,楽天,武器,熊手,農具,太上老君,速度,自由,性格,長寿,果実,理解,日本,ラム,イスラム教,存在,名前,天帝,婿養子,玄奘三蔵,新聞,随筆家

On the other hand, that of MAN is following 16 words.

西遊記,フィクション,中,登場キャラ,ブ,イノ,天,天の川,摩利支天,黒,天竺,経,精進,化け物,玄奘,沙悟浄

The number of links of MAN is more than that of REF in C2E and E2C subtask. The main difference we observed is that the number of links per anchor text is one in REF, however, it is one or more in MAN. We think that the notation and synonym expansion of anchor text are effective.

5. CONCLUSIONS

Our group (OKSAT) submitted two types of runs named SMP and REF for every subtasks of NTCIR-10 Cross-lingual Link Discovery (CLLD). Our method uses titles in Wikipedia pages (corpus) of source language as a entries of a dictionary, so no external dictionary is required. For SMP, we aimed to discover cross-lingual links of actual Wikipedia, in other words it targets Wikipedia ground truth. For REF, on the other hand, we aimed to discover as much meaningful cross-lingual links as possible automatically. SMP work well although there is room for improvement. About REF, we recognized that continuous improvement is required.

6. ACKNOWLEDGMENTS

Our thanks to the task organizers who prepared experiment environment and evaluated runs.

7. REFERENCES

[1] 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 the NTCIR-10*, Tokyo, Japan, 2013.

[2] T. Sato, Fast full text retrieval using gram based tree structure, in *Proceedings of the ICCPOL '97*, Vol.2, pp.572-577, 1997.

[3] T. Sato and K. Han, NTCIR-3 CLIR Experiments at Osaka Kyoiku University - Compression of Gram-based Indices -, , in *Proceedings of the NTCIR-3*, Tokyo, December 2002.

[4] T. Sato, T. Satomoto, and K. Han, NTCIR-3 PAT Experiments at Osaka Kyoiku University -Long Gram-based Index and Essential Words -, in *Proceedings of the NTCIR-3*, Tokyo, December 2002.

[5] S.E. Robertson and S. Walker, Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval, in *Proceedings of the 17th International Conference Research and Development in Information Retrieval*, pp. 232-241, 1994.

[6] UTF-8 - Wikipedia, http://en.wikipedia.org/wiki/UTF-8.

[7] D.A Huffman, A method for the construction of minimum-redundancy codes, in *Proceedings of the I.R.E.*, Vol. 40, pp. 1098-1101, 1952.