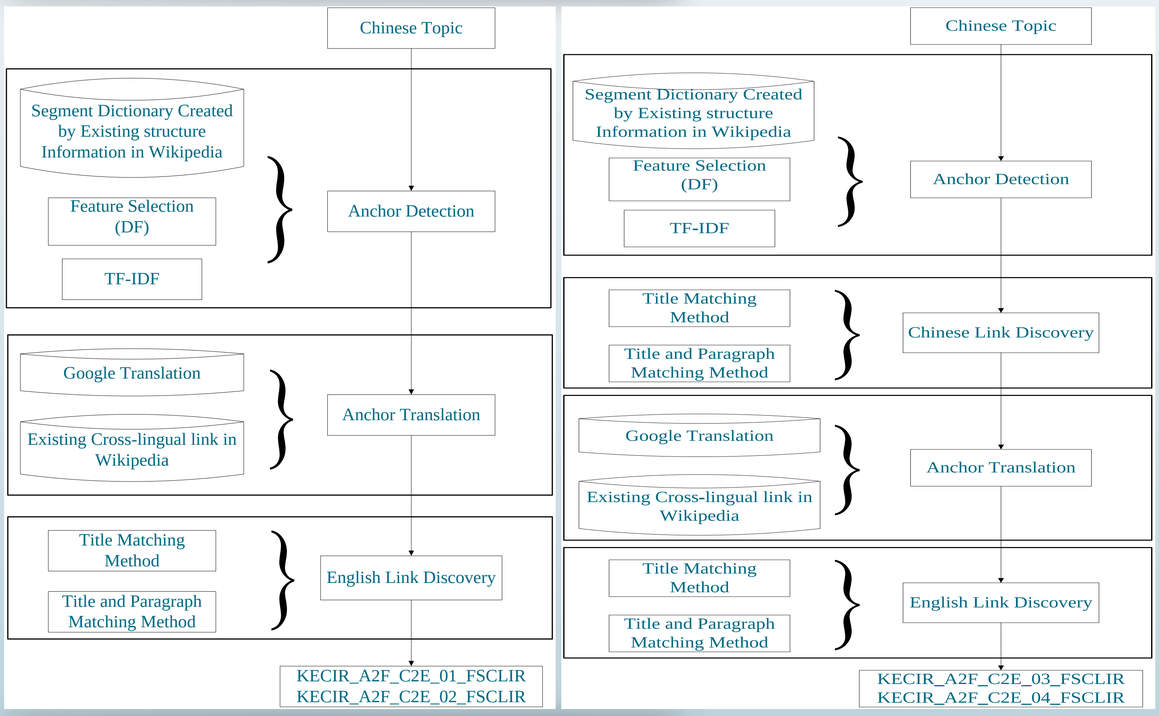
KECIR at NTCIR-10 Cross-Lingual Link Discovery Task

Jianxi Zheng, Yu Bai, Cheng Guo, Dongfeng Cai
Research Center for Knowledge Engineering,
Shenyang Aerospace University, Shenyang 110136, China

ABSTRACT

This paper presents the methods of KECIR at NTCIR-10 Cross-Lingual Link Discovery Task. Two architectures of systems are designed, both of which consist of three common modules such as anchor detection, anchor translation and link discovery. In KECIR_A2F_C2E_03_FSCLIR and KECIR_A2F_C2E_04_FSCLIR, monolingual link discovery module is considered. In order to detect anchor, feature selection method is used. In the processing of anchor translation, we use a method combined with existing cross language link and Google translation web service. For discovering link, both title and paragraph matching methods are used to retrieve the relevant link corresponding to each anchor. Four runs were submitted, and in the A2F evaluation with Manual Assessment results, the KECIR_A2F_C2E_01_FSCLIR achieved the highest score of LMAP and R-Prec in Chinese to English task. The experiment shows that CrossLink based on the first architecture of system can retrieve higher precision links for an anchor than the second one, and anchors with noisiness will result in lower values of metrics in F2F evaluation.

1. SYSTEM ARCHITECTURE



2. ANCHOR DETECTION

Feature Selection methods including DF (Document Frequency), MI (Mutual Information), CHI, and IG (Information Gain), etc have been discussed on text classification in recent years . In this short paper, we choose DF method to identify the anchor for each topic. After breaking Chinese text into separate words, the Document Frequency (DF) of each separate word in both Chinese test topic and training corpus can be obtained and then we use the TF*IDF schema to weight the importance of separate words in each topic.

3. ANCHOR TRANSLATION

We first extract the existing cross-language link from Chinese document collection as translation dictionary. And if there is no translation for an anchor in the dictionary, Google translation web service is used.

4. LINK DISCOVERY

For creating index, we first split the English raw corpora into several short paragraph texts by tag <sec> using regular expressions. Each of texts is represented as the predefined XML styles shown in Table 1, which consists of four elements such as title, id, category and body. The first element represents title of raw corpora. The second element is id of texts which is represented as id-k, where id equals with the number between tag <id> and </id> in raw corpora and k varies from 1 to the total number of tag <sec> in one. The third element represents categories of texts, which combines all categories of raw corpora using tag "|" as separate symbol. The final element is the body parts of short text, which is represented as several paragraphs of raw corpora.

Table 1. the XML Style for English document preprocessing

<title>title of raw topic</title>
<id>id-k</id>
<category>category$_1$|category$_2$|...|category$_n$</category>
<bdy>text of raw test topic between tag <bdy>(or</sec>) and <sec></bdy>

For discovering link, we exploit two methods to retrieve at most 5 relevant English documents for a translated anchor. The first method is a title matching approach which aims at finding relevant titles of page for a translated anchor. In this method, Vector Space Model (VSM) API in Lucene is used to retrieve top5-ranked English documents for a translated anchor. The second method is a title and paragraph matching method which refers to retrieve title and paragraph at the same time for a translated anchor, the goal of which is to explore which page and its paragraph are linked to the corresponding anchor. This method mainly contains three steps: Step1: VSM is used to retrieve top100-ranked English paragraphs for a translated anchor. Step2: for an English document, the relevance with corresponding translated anchor is calculated by summing up weights of all paragraphs it contains. Step3: sort the relevance between the English document and its corresponding anchor by descending, at most 5 relevant English documents for an anchor are returned.

5. EXPERIMENTATION

Table 2. the descriptive of four runs

RunID	descriptive
KECIR_A2F_C2E_01_FSCLIR	FMM+TF-IDF+anchor translation+link discovery (title matching).
KECIR_A2F_C2E_02_FSCLIR	FMM+TF-IDF+anchor translation+link discovery (title and paragraph matching).
KECIR_A2F_C2E_03_FSCLIR	FMM+TF-IDF+Chinese link discovery (title matching) and anchor translation+link discovery (title matching).
KECIR_A2F_C2E_04_FSCLIR	FMM+TF-IDF+Chinese link discovery (title and paragraph matching) and anchor translation+link discovery (title and paragraph matching).

Table 3. the performance of four runs

RunID	LMAP	R-Prec	P5	P10	P20	P30	P50	P250
The performance of four runs in F2F evaluation with Wikipedia ground-truth								
KECIR_A2F_C2E_01_FSCLIR	0.046	0.105	0.136	0.128	0.138	0.132	0.110	0.055
KECIR_A2F_C2E_02_FSCLIR	**0.054**	**0.119**	**0.200**	**0.164**	**0.150**	**0.145**	**0.126**	**0.060**
KECIR_A2F_C2E_03_FSCLIR	0.036	0.097	0.080	0.108	0.106	0.107	0.102	0.052
KECIR_A2F_C2E_04_FSCLIR	0.036	0.105	0.080	0.112	0.110	0.116	0.110	0.054
The performance of four runs in F2F evaluation with manual assessment results								
KECIR_A2F_C2E_01_FSCLIR	**0.044**	**0.081**	0.080	0.088	0.094	**0.100**	0.090	**0.084**
KECIR_A2F_C2E_02_FSCLIR	0.037	0.077	0.072	0.092	**0.096**	0.095	**0.091**	0.072
KECIR_A2F_C2E_03_FSCLIR	0.031	0.076	**0.096**	**0.100**	0.088	0.081	0.083	0.071
KECIR_A2F_C2E_04_FSCLIR	0.028	0.076	0.056	0.096	0.088	0.084	0.086	0.066
The performance of four runs in A2F evaluation with manual assessment results								
KECIR_A2F_C2E_01_FSCLIR	**0.087**	**0.054**	**0.024**	**0.036**	**0.046**	**0.061**	**0.074**	**0.064**
KECIR_A2F_C2E_02_FSCLIR	0.050	0.039	**0.024**	**0.036**	0.042	0.047	0.055	0.047
KECIR_A2F_C2E_03_FSCLIR	0.044	0.032	0.016	0.020	0.024	0.036	0.046	0.040
KECIR_A2F_C2E_04_FSCLIR	0.029	0.025	0.016	0.020	0.028	0.037	0.038	0.032

In order to facilitate the description, we call four submissions including KECIR_A2F_C2E_01_FSCLIR, KECIR_A2F_C2E_02_FSCLIR, KECIR_A2F_C2E_03_FSCLIR, and KECIR_A2F_C2E_04_FSCLIR as Run 1, Run 2, Run 3 and Run 4, respectively. When runs are measured in A2F level with Manual Assessment, Run 1 gets the top score of LMAP and R-Prec outperformed other three runs. It is out of our expectation that the performance of the second run with content information is lower than one of the first run. When evaluated with Wikipedia Ground Truth in F2F level, Run 2 achieves higher performance than Run 1, which shows that content information helps discover articles of same topic and find more existing cross language links in Wikipedia. In F2F evaluation with Manual Assessment, Run 3 performs better than the remainder of runs in metric of Precision-at-5 while it ranks third in four runs in scores of LMAP and R-prec. From table 3, we conclude that the former two runs across all evaluation perform better than the rest of runs, which proves the performance of system I is higher than one of system II.

CONCLUSION

This paper describes the methods of KECIR at NTCIR-10 CrossLink-2 Task. The KECIR_A2F_C2E_01_FSCLIR achieved the best score of LMAP and R-prec with Manual Assessment in A-2-F evaluation in Chinese to English task. This demonstrates our method can effectively recommend relevant links for each anchor per test topic. However, our four runs do not perform so well in both of F2F evaluation with Wikipedia Ground Truth and Manual Assessment results. It proves that our method needs to promote the ability of finding articles of same topic from multilingual Wikipedia articles. The experiment shows that CrossLink based on the first architecture of system can retrieve higher precision links for an anchor than the second one, and anchors with noisiness will result in lower values of metrics in F2F evaluation. We plan to build a united framework for Cross-lingual link between Chinese articles and English ones in future. Other feature selection approaches to anchor detection and anchor translation based on Web will be explored in the united framework of CrossLink.