IISR Crosslink Approach at NTCIR 9 CLLD Task

Chun-Yuan Cheng
Department of Computer Science and Engineering
Yuan Ze University
Chungli, Taiwan
s1006005@mail.yzu.edu.tw

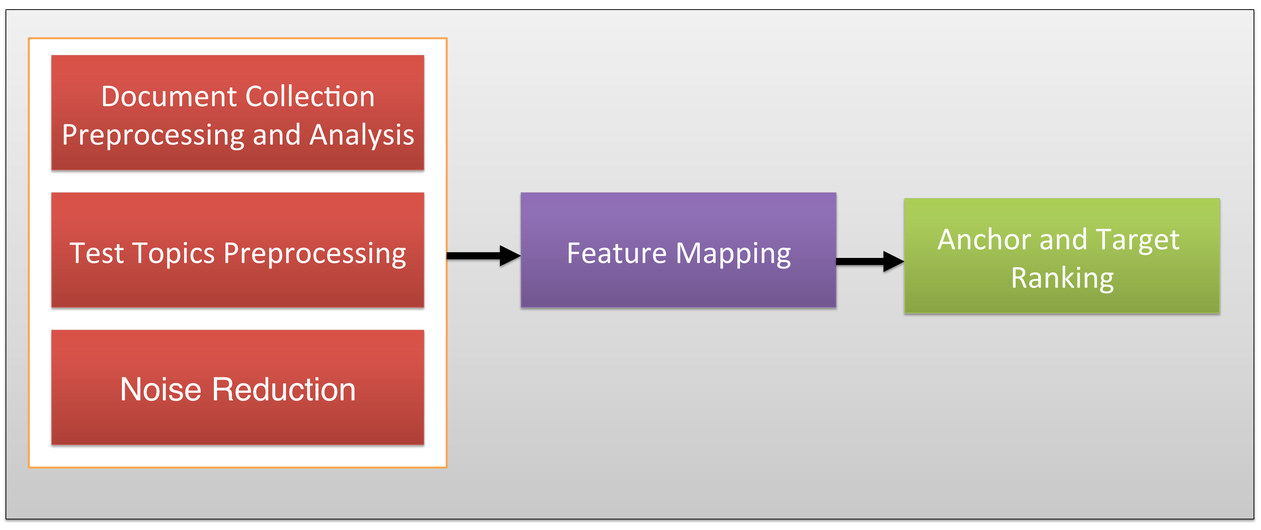
Yu-Chun Wang
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
d97023@csie.ntu.edu.tw

Richard Tzong-Han Tsai*
Department of Computer Science and Engineering
Yuan Ze University
Chungli, Taiwan
thtsai@saturn.yzu.edu.tw

*Corresponding author

ABSTRACT

We propose a simple and effective approach to discover the links. Our method comprises preprocessing steps, anchor-target link mapping and the ranking steps. We rank the anchor candidates by the Wikipedia category sets and the PageRank method, and we select the Korean target pages with the mutual information between English anchors and Korean titles of Wikipedia articles The official file-to-file evaluation with the manual assessment of our system is achieved from 0.6 to 0.7 in P10 precision, which shows that our approach can achieve satisfactory results.



System overview

A. Document Collection Preprocessing and Analysis

In our analysis, the document collection includes 39,798 independent categories and 3,044,968 anchor links. There are 129,696 anchors (39.2%) which do not link to others, and 204,305 anchors (61.7%) do not belong to any category. We establish a relational database to record the anchor mapping information.

B. Test Topics Preprocessing

We adopts ***n-gram*** algorithm to extract the possible anchor candidates. We collect all the titles of the articles in English Wikipedia by using ***Wikipedia API***. The n-gram anchor candidates are checked whether the candidates existed or not, where n is from 1 to 5.

C. Noise Reduction

1. Remove all the English stop-words with the ***NLTK*** library.
2. Remove anchors that match the time or date formats in regular expressions.
3. Remove anchors that have lower-case initials.
4. Apply the maximum matching method to locate and select the longest candidate from the n-gram results.

D. Feature Mapping

Our system selects to map the English anchor candidates in the test topics with the Korean articles in the document collection. Table shows the feature sets that our system uses to match the anchor candidates.

Feature correspond table

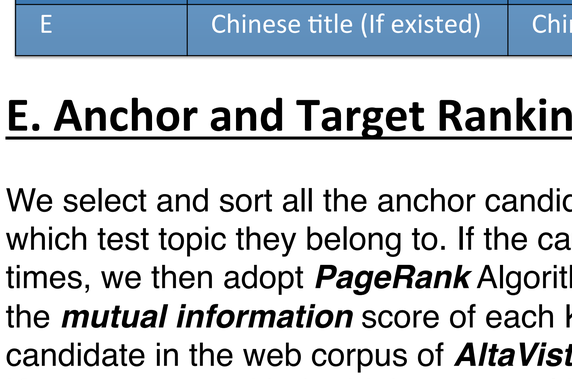
Feature	Candidate Anchor	Document Collection
A	English title	English title via Google translate API (Korean title)
B	Singular English title	English title via Google translate API (Korean title)
C	Singular English tilte	English title (If existed)
D	Korean title (If existed)	Korean title
E	Chinese title (If existed)	Chinese title (If existed)

E. Anchor and Target Ranking

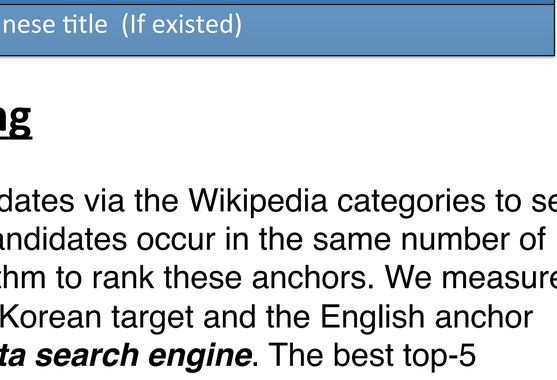
We select and sort all the anchor candidates via the Wikipedia categories to see which test topic they belong to. If the candidates occur in the same number of times, we then adopt ***PageRank*** Algorithm to rank these anchors. We measure the ***mutual information*** score of each Korean target and the English anchor candidate in the web corpus of ***AltaVista search engine***. The best top-5 Korean targets with the highest mutual information score are remained as the final results.

File-to-File Evaluation with manual assessment results: Precision-at-N

Run ID	P5	P10	P20	P30	P50	P250
IISR_singular_term_to_en_title	0.604	0.720	0.416	0.373	0.232	0.046
IISR_ko_title_to_key_name	0.692	0.712	0.606	0.532	0.402	0.086
IISR_kttkn_to_ztttzt	0.672	0.656	0.618	0.533	0.405	0.088
IISR_sttent_to_ztttzt_to_kttkn	0.660	0.648	0.610	0.527	0.401	0.088
IISR_sttent_to_kttkn	0.620	0.648	0.606	0.529	0.406	0.088



Interpolated Precision-Recall of E2K links by standalone feature



Interpolated Precision-Recall of E2K links by combined feature

IISR Lab