Query Expansion from Wikipedia and Topic Web Crawler on CLIR

Meng-Chun Lin, Ming-Xiang Li, Chih-Chuan Hsu and Shih-Hung Wu

Department of Computer Science and Information Engineering

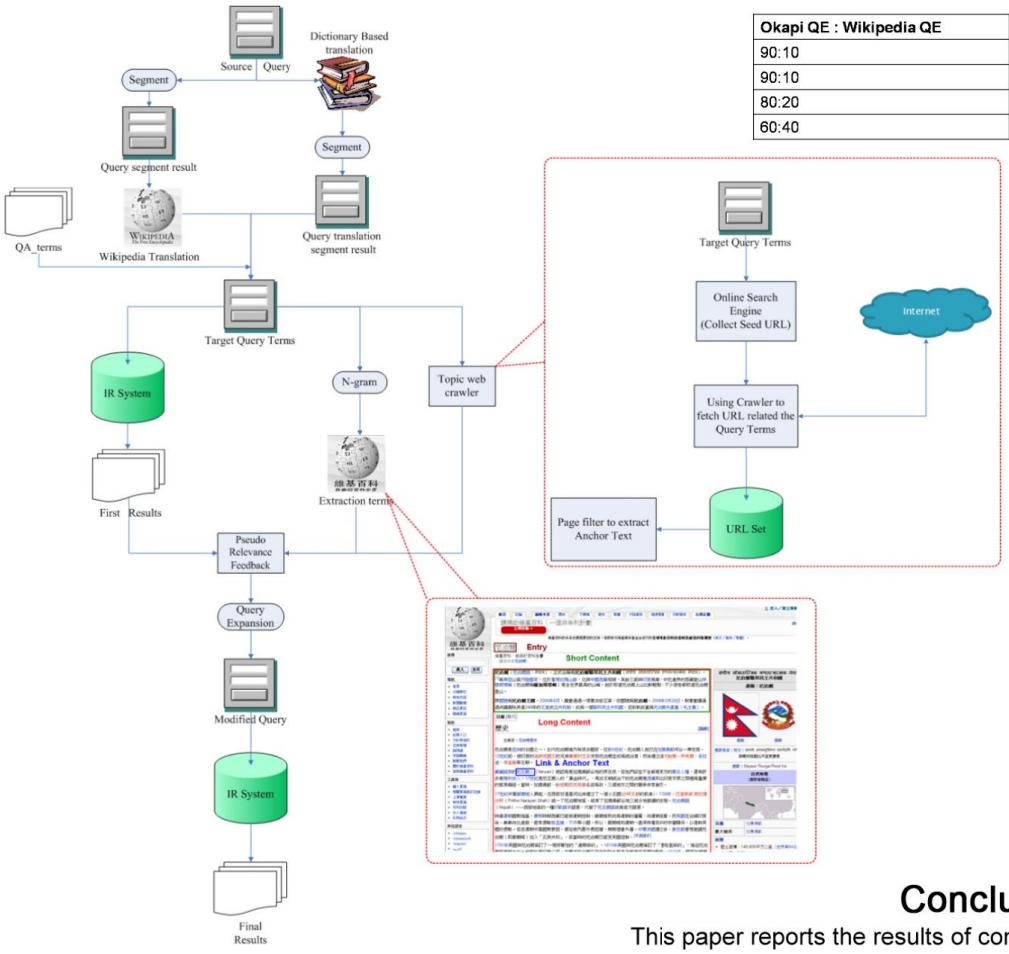
Chaoyang University of Technology Taichung County 41349, TAIWAN (R.O.C)

Translation Method

1. Using dictionary to translate Query terms.
2. Using Wikipedia to translate OOV terms.

Query Expansion Method

1. Using Okapi BM25 to expan target Query terms.
2. Using Wikipedia and Topic web crawler to expan expand target Query terms.



NTCIR_Experiments

Table 1. The MAP of CS-runs using the different proportion in QE term from Okapi and other source(Wikipedia and Topic web crawler).

Okapi QE : Wikipedia QE	Run	MAP
70:30	EN-CS-T	0.2014
80:20	EN-CS-T(QA)	0.2031
100:0	EN-CS-D	0.1601
100:0	EN-CS-DN	0.1696
Okapi QE : Topic web crawler QE	**Run**	**MAP**
70:30	EN-CS-T	0.2071
60:40	EN-CS-T(QA)	0.2084
70:30	EN-CS-D	0.1652
90:10	EN-CS-DN	0.1707

Table 2. The MAP of CT-runs using the different proportion in QE term from Okapi and other source(Wikipedia and Topic web crawler).

Okapi QE : Wikipedia QE	Run	MAP
60:40	EN-CT-T	0.1782
70:30	EN-CT-T(QA)	0.1948
80:20	EN-CT-D	0.141
80:20	EN-CT-DN	0.1571
Okapi QE : Topic web crawler QE	**Run**	**MAP**
40:60	EN-CT-T	0.1839
40:60	EN-CT-T(QA)	0.206
50:50	EN-CT-D	0.1461
40:60	EN-CT-DN	0.1676

Table 3. The MAP of JA-runs using the different proportion in QE term from Okapi and Wikipedia.

Okapi QE : Wikipedia QE	Run	MAP
90:10	EN-JA-T	0.1636
90:10	EN-JA-T(QA)	0.1625
80:20	EN-JA-D	0.0929
60:40	EN-JA-DN	0.0905

Conclusion

This paper reports the results of combining query terms from different sources on query expansion in CLIR. We tested this idea on EN-JA,EN-CT,and EN-CS pairs.The method in official runs combines the translation results from Wikipedia and Google translation. We conducted several additional runs to show that the combined QE is better than QE from a single source.In additional runs, we added a topic web crawler for further query expansion in EN-CT and EN-CS. The titles and anchor texts in related pages were treated as another source of QE. The experiment results show that this further expansion improved performance.

Acknowledgement

This research was partly supported by the National Science Council under NSC 98-2221-E-324 -025.