Using Wikipedia to Translate OOV Terms on MLIR

Chen-Yu Su, Tien-Chien Lin and Shih-Hung Wu*
Department of Computer Science and Information Engineering
Chaoyang University of Technology Taichung County 41349, TAIWAN (R.O.C)
*Contact author: shwu@cyut.edu.tw

NTCIR-6 Experiment Result

Table 1. The MAP of official runs

	C-CJK-T-01		C-CJK-D-02	
Recall	Rigid	Relax	Rigid	Relax
0.0	0.4884	0.6986	0.4651	0.6463
0.1	0.2278	0.3315	0.1887	0.2767
0.2	0.1384	0.2130	0.1101	0.1533
0.3	0.0888	0.1330	0.0716	0.0826
0.4	0.0489	0.0531	0.0359	0.0470
0.5	0.0256	0.0268	0.0235	0.0185
0.6	0.0129	0.0139	0.0089	0.0047
0.7	0.0051	0.0061	0.0027	0.0043
0.8	0.0016	0.0000	0.0021	0.0036
0.9	0.0000	0.0000	0.0000	0.0000
1.0	0.0000	0.0000	0.0000	0.0000
MAP	**0.0704**	**0.0992**	**0.0584**	**0.0802**

Table 2. The R-Precision of official runs

	R-Precision	
	Rigid	Relax
C-CJK-T-01	0.1528	0.1811
C-CJK-D-02	0.1357	0.1579

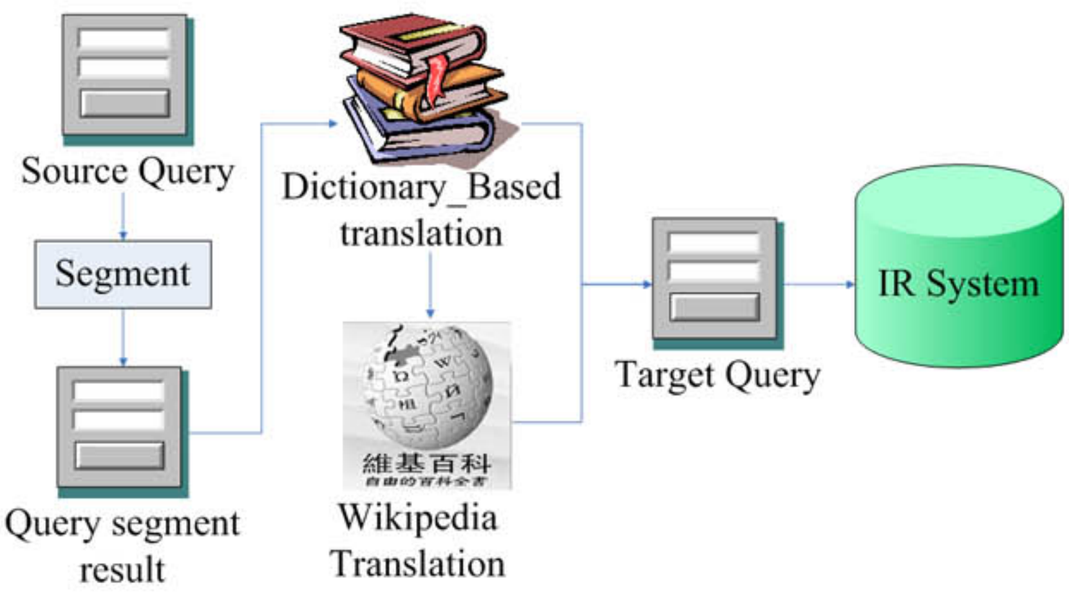
Table 3. The MAP of each run

	MAP	
	Rigid	Relax
C-C-T	0.2183	0.3194
C-C-D	0.1893	0.2784
C-J-T	0.0842	0.1140
C-J-D	0.0259	0.0409
C-K-T	0.0440	0.0660
C-K-D	0.0374	0.0579

* C-C Run : Using Chinese Query Retriveal Chinese Document
* C-J Run : Using Chinese Query Retriveal Japanese Document
* C-K Run : Using Chinese Query Retriveal Korean Document
* C-CJK Run : Using Chinese Query Retriveal Chinese, Japanese and Korean Documents

Translation Method

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



Wikipedia Translation method

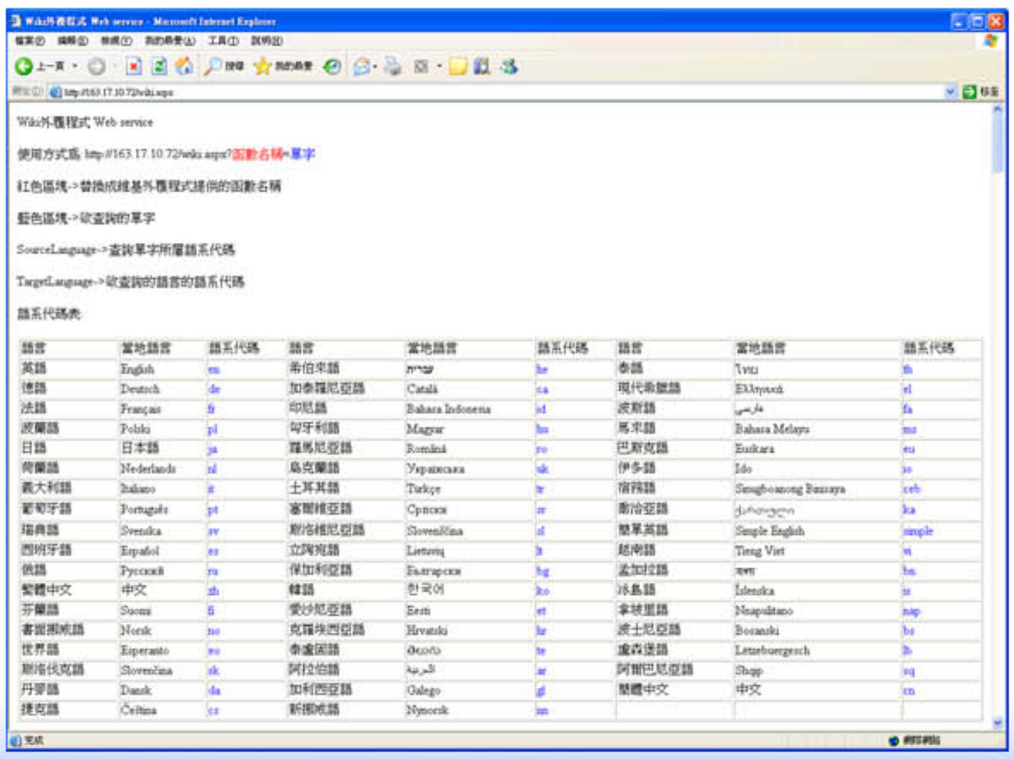
Using ”other language list” to translate OOV terms

Figure 1. term ”blog” in Chinese Wikipedia page

Figure 2. term ”blog” in Japanese Wikipedia page

Our Free Online Wikipedia Translation Web Service

URL: http://wil.csie.cyut.edu.tw/WikiWebService/



Conclusion

In MLIR (Multilingual Information Retrieval), the translation of query terms requires dictionary in various languages. The C-CJK task is using Chinese Query retrieval Chinese, Japanese and Korean documents.

In this paper, we regard Wikipedia as an additional live-dictionary to translate the OOV (Out-Of-Vocabulary) terms. No free live-dictionary contains so many languages before. We find that the Wikipedia translation is a good resource that can improve the performance of a MLIR system.

We developed a machine readable Wikipedia translation API. This API helped us using Wikipedia to translate OOV terms into various languages.