DCU's Experiments in the NTCIR-8 IR4QA Task

Jinming Min Jie Jiang Johannes Leveling Gareth J.F. Jones Andy Way

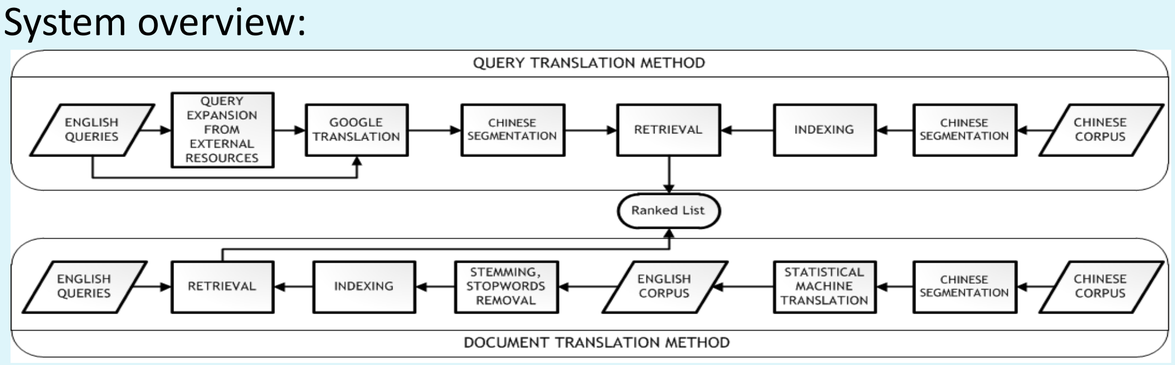
CNGL, School of Computing, Dublin City University, Ireland

OVERVIEW

Research objectives:

- Application of state-of-art machine translation system on cross language retrieval task (CLIR)
- Evaluation of query expansion from external resource on cross language retrieval task
- Comparison of KL-divergence language model and Okapi BM25 model on simplified Chinese retrieval task

System overview:



MACHINE TRANSLATION

DCU MATREX machine translation system for document translation:

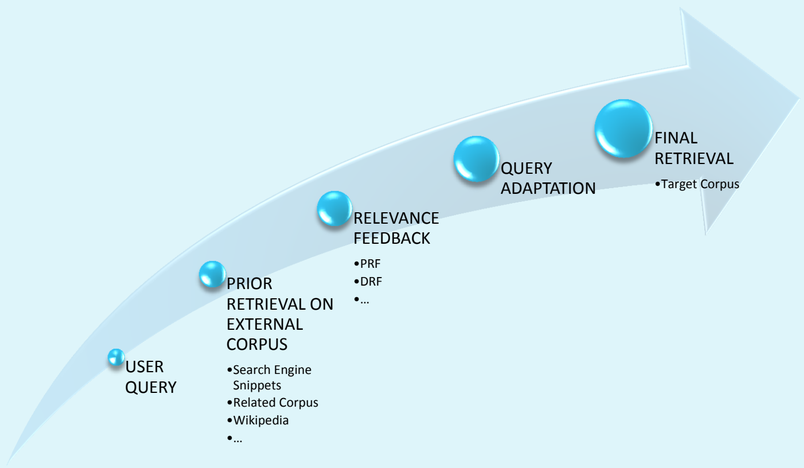
- Augmented phrase-based Chinese-English SMT system
- 3.4 million sentences for translation model training
- 12 million sentences for language model building
- 4 million Chinese sentences are translated into English

Google translation system for query translation:

- Very large parallel training corpus
- Align parallel web documents into corresponding translated sentences
- Good translation of the named entities

QUERY EXPANSION FROM EXTERANL RESOURCE

- Traditional query expansion (QE) methods are based on the document collection being searched
- External evidence for query expansion (QEE) is based on use of a separate document collection describing concepts in the query, e.g. Wikipedia
- QEE can achieve better search results for some topics



MONOLINGUAL RETRIEVAL MODEL

- KL-Divergence LM
- Language model retrieval method
- KL-Divergence is a non-symmetric measure of the difference between two probability distribution D and Q
- Okapi BM25
- Probability retrieval model, bag of words retrieval function
- Proven to be effective in lots of retrieval tasks

EXPERIMENTAL RESULTS

Runs	Methodology	MAP		NDCG
DCU-CS-CS-01-T	LM	**0.4187**	100%	**0.6545**
DCU-CS-CS-02-T	BM25	0.3260	77.86%	0.5566
DCU-EN-CS-01-T	MT+LM	0.2284	54.55%	0.4597
DCU-EN-CS-02-T	Google+LM	0.3347	**79.94%**	0.5695
DCU-EN-CS-03-T	QEE+Google+LM	0.3215	76.79%	0.5671

FAILURE OF OKAPI BM25

- Topic: 郭台铭是哪家公司的总裁?

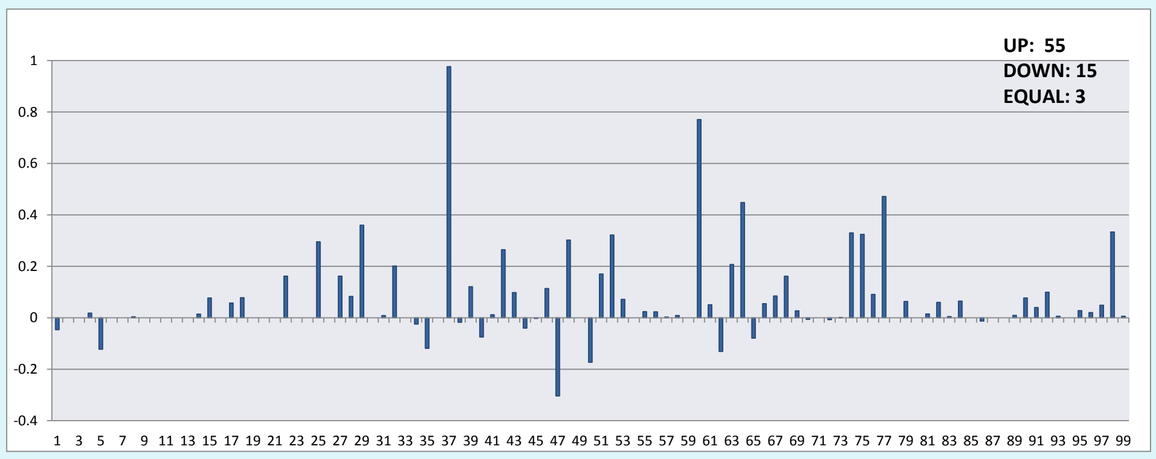
  Wrong segmentation:

  郭 台 **铭** 是 哪家 公司 的 总裁 ?

  Top ranked documents from Okapi BM25 model

  陈 德 **铭** 是 中共 十六大 代表 , 九届、十届 全国 人大 代表。
- Analysis
- Wrong segmentation for Chinese names
- Documents containing term with incorrect segmentation of Chinese characters with different meaning can rank highly in Okapi BM25 model

SIGNICANCE TEST (KL-Divergence VS. BM25)



CONCLUSION

1.Google translation works well for query translation

2.External query expansion is potentially promising for CLIR task

3.Okapi BM25 model fails for Chinese queries due to segmentation errors, particular problem for out-of-vocabulary words

This research is funded by Science Foundation Ireland (SFI) as part of the National Development Plan (NDP) 2007-2013