

# DCU's Experiments for the NTCIR-8 IR4QA Task

Jinming Min

Jie Jiang

Johannes Leveling

Gareth J. F. Jones

Andy Way

Centre for Next Generation Localisation

School of Computing

Dublin City University

Dublin 9, Ireland

{jmin, jjiang, jleveling, gjones, away}@computing.dcu.ie

## ABSTRACT

We describe DCU's participation in the NTCIR-8 IR4QA task [16]. This task is a cross-language information retrieval (CLIR) task from English to Simplified Chinese which seeks to provide relevant documents for later cross language question answering (CLQA) tasks. For the IR4QA task, we submitted 5 official runs including two monolingual runs and three CLIR runs. For the monolingual retrieval we tested two information retrieval models. The results show that the KL-Divergence language model method performs better than the Okapi BM25 model for the Simplified Chinese retrieval task. This agrees with our previous CLIR experimental results at NTCIR-5. For the CLIR task, we compare query translation and document translation methods. In the query translation based runs, we tested a method for query expansion from external resource (QEE) before query translation. Our result for this run is slightly lower than the run without QEE. Our results show that the document translation method achieves 68.24% MAP performance compared to our best query translation run. For the document translation method, we found that the main issue is the lack of named entity translation in the documents since we do not have a suitable parallel corpus for training data for the statistical machine translation system. Our best CLIR run comes from the combination of query translation using Google translate and the KL-Divergence language model retrieval method. It achieves 79.94% MAP relative to our best monolingual run.

## Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Search and Retrieval

## General Terms

Algorithms, Design, Experimentation, Performance

## Keywords

machine translation, query formulation, retrieval models, relevance feedback

## 1. INTRODUCTION

In this paper, we describe our experiments for the IR4QA subtask of the NTCIR-8 ACLIA task. We took part in the English to Simplified Chinese CLIR task. Our strategies are

from two perspectives: one is translating queries from English to Simplified Chinese (Query Translation), while the other is to translate the English corpus into Simplified Chinese (Document Translation). For query translation, we use the Google translate online service<sup>1</sup>. We also test a method for query expansion from external resources on the CLIR task. This method has already been successfully applied in monolingual task [18]. For the document translation method, we utilize the statistical machine translation system built for DCU's participation for NIST machine translation evaluation<sup>2</sup>.

This paper is structured as follows: Section 2 introduce our system in overview and outlines our basic strategies for this task, Section 3 describes our query translation method using the Google translate online service, Section 4 describes our statistical machine translation system for document translation, Section 5 describes the query expansion from external resource method as applied to the IR4QA task, Section 6 introduces our monolingual retrieval models including the Okapi BM25 model and KL-Divergence language model, Section 7 describes and analyses our official results, and finally Section 9 gives conclusions and directions for further work.

## 2. SYSTEM OVERVIEW

In this section, we introduce our system overview for the IR4QA task. In Figure 1 we present two strategies for IR4QA task: one is from the query translation perspective and another is from the document translation perspective. The main components in Figure 1 include:

**Google Translation** Translate the official English topics into Simplified Chinese using Google translate online service;

**Chinese Segmentation** Segment the Simplified Chinese sentences into words using LDC Chinese segmentation tool<sup>3</sup>;

**Indexing** Index the Simplified Chinese corpus or English corpus using the Lemur toolkit<sup>4</sup>;

**Retrieval** Retrieve relevant documents in the suitable index;

<sup>1</sup><http://translate.google.com/>

<sup>2</sup><http://www.itl.nist.gov/iad/mig/tests/mt/>

<sup>3</sup><http://www ldc.upenn.edu/>

<sup>4</sup><http://www.lemurproject.org/>

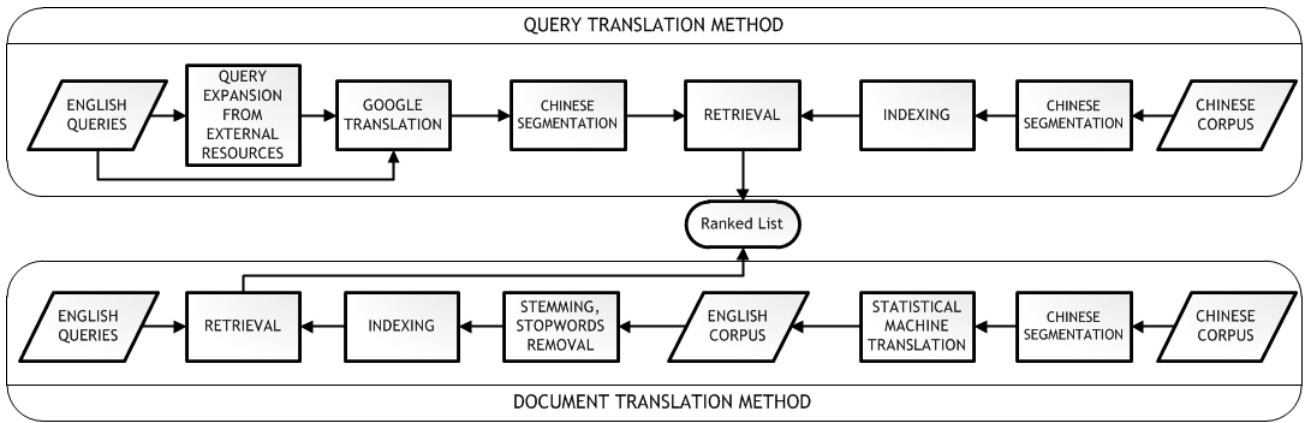


Figure 1: IR4QA CLIR Task Overview.

Query translation and document translation are the main methodologies used in the CLIR research. Through the comparison of these two strategies, we hope to gain more insight into how machine translation research can improve the CLIR task. We are using two state-of-art machine translation systems: one is the Google translate online service and another is DCU’s system used for the NIST Machine Translation evaluation system. In the query expansion method, there is an optional step called query expansion from external resource.

### 3. QUERY TRANSLATION BY GOOGLE

the Google translate online system offers the state-of-art translation quality and it currently widely used in the CLIR research [15]. In our official runs, we translated the English questions into Simplified Chinese using Google translate online service. Our results show that Google translate partly solves the out-of-vocabulary problem in CLIR research.

Google’s machine translation system uses a statistical method. The effectiveness of this approach greatly depends on the training parallel corpus, and Google has a great advantage over many other machine translation group to find more parallel corpus from their web search engine system. With its very large of parallel training datasets, Google’s translation system has acquired very good evaluation results in the NIST machine translation campaign [2]. For CLIR research, Google’s translation system has also showed very effective performance [15]. Named entities translation has been a research problem in machine translation and CLIR research for a long time. Since the named entities are usually new words emerging in recent years, the machine readable bilingual dictionary usually don’t have good coverage of these new terms. This leads to translation errors. It could be a good explanation why Google system can translate the name entities well since they are performing algorithms to align the web parallel documents into aligned translation sentences which can be used as the training corpus in statistical machine translation system [1].

### 4. DOCUMENT TRANSLATION BY DCU’S MT SYSTEM

The documents of NTCIR-8 ACLIA task come from the LDC Chinese Gigaword Third Edition [6]. The content of

the documents are mainly newswire archives from year 2002 to 2005. To accomplish document translation, we employed the DCU MATREX machine translation system to translate all the Chinese documents into English for cross lingual information retrieval task [5]. The details are presented as follows:

#### 4.1 Preprocessing

The format of Chinese documents provided in NTCIR-8 is SGML. To feed them into Statistical Machine Translation (SMT) system, we performed the following procedures to preprocess them into plain text:

- Join sentences within different lines into one line. Since the wrap breaking positions of the original documents are not aligned with the punctuation, we used heuristic rules according to document types of the SGML file to determine the way of joining sentences in multiple lines.
- Split the joined sentences into shorter ones by punctuation. Both Chinese and English punctuation with different priorities is used to split the sentences. Punctuation is preserved in the shorter sentences to conform with the training data of SMT system. Maximum sentence length was set to 90 Chinese characters and all sentences longer than this are chopped into shorter ones.
- Remove SGML tags. All SMGL tags are removed for SMT purpose, however, the correspondences between IR documents in TREC format and the generated plain texts are stored for further processing.
- Chinese word segmentation. We used Stanford Chinese word segmentation [3] to segment the generated plain texts.

After the previous steps, we obtained 4,642,223 Chinese sentences from the original NTCIR-8 documents. These sentences are the inputs for the SMT system. Translated documents are used for CLIR purpose. The details of our SMT system is presented in section 4.2, and corpus decoding is described in section 4.3.

## 4.2 SMT system configuration

Our SMT system was used for the DCU NIST 2009 evaluation, it is an augmented phrase-based Chinese-English SMT system. An augmented phrase table is fed into Moses decoder [7] for translation.

The corpora used for system training come from LDC resources, which are listed in Table 1:

| Type             | Resource number  |
|------------------|--|
| Parallel data    | LDC2000T46, LDC2000T50,<br>LDC2002E18, LDC2002E27,<br>LDC2002L27, LDC2002T01,<br>LDC2003E07, LDC2003E14,<br>LDC2003T17, LDC2004E12,<br>LDC2004T07, LDC2004T08,<br>LDC2005T01, LDC2005T06,<br>LDC2005T10, LDC2005T34,<br>LDC2006T04, LDC2007T09 |
| Monolingual data | LDC2007T07   |

Table 1: SMT corpora

By performing data cleaning and preprocessing, we used 3.4 million sentences in the parallel data corpora, and 12 million sentences for language model building. During the training phase, we used the GIZA++<sup>5</sup> toolkit to perform word alignment and adopt the “grow-diag-final” refinement method [8]. After word alignment, the method in [20] is used for phrase extraction. The language model in our experiments is a 5-gram language model using the SRILM<sup>6</sup> toolkit with modified Kneser-Ney smoothing [10].

We tuned the trained phrase-base SMT system on NIST 2006 Chinese-English current test set which contains 1,664 sentences. Each source sentence has 4 references.

## 4.3 Parallel decoding

In section 4.1, we obtained more than 4 million sentences for SMT inputs. However, for any state of the art SMT system, it will take a very large amount of time to process a corpus of this size. In our CLIR scenario, we take the following two measures to speed up the SMT decoding:

- A Smaller distortion limit is set to the phrase-based SMT decoder. Since for the IR scenario, word order is not as important as in the case of standard machine translation. We adopt distortion limit of 4, which speeds up the decoder by 2 times compared to the default value of 6.
- B Parallel computing scheme is taken for mass corpus decoding. The decoding corpus is split into small parts, and a cluster of servers is used to perform the parallel decoding. The size of the split corpora is calibrated to fit the memory usage of their correspondent filtered phrase tables. Several decoding tasks (determined by the CPU numbers) run in parallel to perform corpus decoding.

In our experiment, 48 CPUs are used to carry out document translation. It took approximately 8 days for our SMT system to process all the input Chinese sentences. The

<sup>5</sup><http://fjoch.com/GIZA++.html>

<sup>6</sup><http://www.speech.sri.com/projects/srilm/>

translated English sentences are used for CLIR in the following sections.

In the previous procedures, all the data are lowercased and detokenized. After translation, we use the Moses Recaser to recase all the results for further usage.

## 5. PRE-TRANSLATION QUERY EXPANSION FROM DBPEDIA

Query expansion from external resource gains lots of attention in recent IR research [18, 19] and we test this method for this CLIR task. The classical query expansion method expands the original query with feedback terms selected from the assumed top relevant documents of target corpus in the prior retrieval. Query expansion from external resource selects the feedback terms from an external resource. These resources are usually an external relevant corpus, the search engine snippets or Wikipedia related resources. In our experiment, we use the English DBpedia<sup>7</sup> as the external resource for query expansion. Here DBpedia can be viewed as the Wikipedia<sup>8</sup> abstract documents collection.

For an English query, the documents used for feedback are retrieved from the external resource (DBpedia in our experiments). The top 30 ranked documents in the prior retrieval are chosen as the assumed relevant documents. From all the words in the top 30 documents we first remove the stop words. The stop word list was produced from the DBpedia document collection, for which we computed the term frequency in the DBpedia collection and select the top 500 words as the stop words. And for these top 30 relevant documents, we compute a word frequency list and remove the stop words and ignore the original words contained in the “query”. Equation 1 is used to rank the terms. Here the  $r(t_i)$  means the number of documents which contain term  $t_i$  in the top 30 assumed relevant documents.  $idf(t_i)$  uses the method as Equation 2. here  $t_i$  is the  $i$ th term, and  $N$  is the total number of documents in this collection;  $n(t_i)$  is the number of the documents which contain the term  $t_i$ .

$$S(t_i) = r(t_i) * idf(t_i) \quad (1)$$

$$idf(t_i) = \log \frac{N - n(t_i) + 0.5}{n(t_i) + 0.5} \quad (2)$$

We select the top 10 feedback terms to add into the original English query. These query words are sent to the Google translate online service and the Simplified Chinese query is returned. These new formulated Simplified Chinese queries are sent into the retrieval system to get the search results. This experimental run has the id DCU-EN-CS-03-T.

would normally be considered for QE because the documents in DBpedia are usually very short length. If we only used 10 or 20 as the assumed relevant documents, it was found to be difficult to get useful feedback terms from the relevant documents. For the number of feedback words, we select the top 10 words ranked using Equation 1.

## 6. MONOLINGUAL RETRIEVAL MODEL

In our official runs, we use two retrieval models in our experiments: the Okapi BM25 model and the KL-Divergence

<sup>7</sup><http://dbpedia.org/>

<sup>8</sup><http://en.wikipedia.org/wiki/>

**Table 2: Official Experimental Results for IR4QA Task.**

| Runs           | Methodology       | MAP    | NDCG   |
|----------------|-------------------|--------|--------|
| DCU-CS-CS-01-T | LM                | 0.4187 | 0.6545 |
| DCU-CS-CS-02-T | Okapi             | 0.3260 | 0.5566 |
| DCU-EN-CS-01-T | MT + LM           | 0.2284 | 0.4597 |
| DCU-EN-CS-02-T | Google + LM       | 0.3347 | 0.5695 |
| DCU-EN-CS-03-T | QEE + Google + LM | 0.3215 | 0.5671 |

Language modeling method. As suggested by NTCIR-5 paper [9], we find that the KL-Divergence language model method performs better than the Okapi BM25 model for the Simplified Chinese retrieval task. These two retrieval model are all implemented in the Lemur toolkit.

Details of the Okapi BM25 model can be found in [11]. The document term frequency (*tf*) weight used in the Okapi BM25 model is shown in Equation 3.

$$tf(t_i, D) = \frac{(k_1 + 1) \cdot f(t_i, D)}{f(t_i, D) + k_1 \cdot (1 - b + b \frac{l_d}{l_c})} \quad (3)$$

$f(t_i, D)$  is the frequency of query term  $t_i$  in Document  $D$ ,  $l_d$  is the length of document  $D$ ,  $l_c$  is the average document length of the collection, and  $k_1$  and  $b$  are parameters set to 1.0 and 0.3 respectively since our target documents are of short-length [11]. The *idf* of a term is given by  $\log(N/n(t_i))$ , where  $N$  and  $n(t_i)$  have the same definitions as before.

The query *tf* function (*qtf*) is also defined using Equation 3 where  $k_1$  and  $b$  are set to 1000 and 0, so *qtf* will usually be approximately equal to 1. The score of document  $D$  against query  $Q$  is calculated as shown in Equation 4.

$$s(D, Q) = \sum_n^{i=1} tf(t_i, D) \cdot idf(t_i) \quad (4)$$

In the retrieval process, we also test the effectiveness of query expansion (QE). The query expansion method utilizes the Okapi feedback algorithm.

$$RW(i) = \log\left[\frac{(r + 0.5)(N - n - R + r + 0.5)}{(n - r + 0.5)(R - r + 0.5)}\right] \quad (5)$$

$$Weight(t) = r * RW(i) \quad (6)$$

where  $r$  is the number of top-ranked feedback documents which contain the term  $t$ , and  $RW(i)$  is computed by Equation 5. In Equation 5,  $N$  is the total number of documents in the corpus and  $n$  is the number of documents where the term  $t$  appears, and  $R$  is the number of known relevant document for a query. The terms in the top feedback documents with higher weight are selected as the feedback terms.

Using the Okapi feedback algorithm for QE, we set the number of feedback documents to 5, and the number of feedback terms as 20. These feedback terms are added to the query with a factor 1. All these parameters are adjusted manually to get the best result.

Another retrieval model that we are using is Language model (KL-divergence) feedback retrieval example, and it uses a collection mixture method and Dirichlet smoothing. KL-divergence is usually used to compute the “distance” of two distribution [4]. It was applied in the language model based retrieval method successfully [21]. The score to rank

the document by query in KL-divergence language model can be as:

$$-D(\Theta_Q || \Theta_D) \approx \sum_w p(w|\Theta_Q) \log p(w|\Theta_D) \quad (7)$$

In Equation 7,  $\Theta_Q$  and  $\Theta_D$  denote the parameters of the query unigram language model and the document unigram language model. Shown with a smoothing scheme, the KL-divergence scoring formula is:

$$\sum_{w:c(w;d), p(w|\Theta_Q) > 0} \log \frac{p_s(w|d)}{\alpha_d p(w|c)} + \log \alpha_d \quad (8)$$

when applying the Dirichlet smoothing with

$$p_s(w|d) = \frac{c(w, d) + \mu p(w|c)}{|d| + \mu} \quad (9)$$

and

$$\alpha_d = \frac{\mu}{\mu + |d|} \quad (10)$$

So the new KL-divergence scoring formula is:

$$\sum_{w:c(w;d), p(w|\Theta_Q) > 0} p(w|\Theta_Q) \log\left(1 + \frac{c(w, d)}{\mu p(w|c)}\right) + \log \frac{\mu}{\mu + |d|} \quad (11)$$

## 7. EXPERIMENTAL RESULTS

Here we give our official results in the NTCIR-8 ACLIA IR4QA task. In our official results, we use the following techniques as:

**LM** Language modeling retrieval model using collection mixture method and Dirichlet smoothing;

**Okapi** Classical Okapi BM25 retrieval model;

**MT** Statistical machine translation system;

**Google** Google online translation system;

**QEE** Query expansion from external resources.

Our formal results show that the language model based method performs better than the Okapi BM25 model for the Simplified Chinese retrieval task. And our best result comes from the combination of query translation by Google online and language model retrieval method. Also our query expansion from external resources shows some improvement in some topics. It can be continued to be a promising direction in CLIR research.

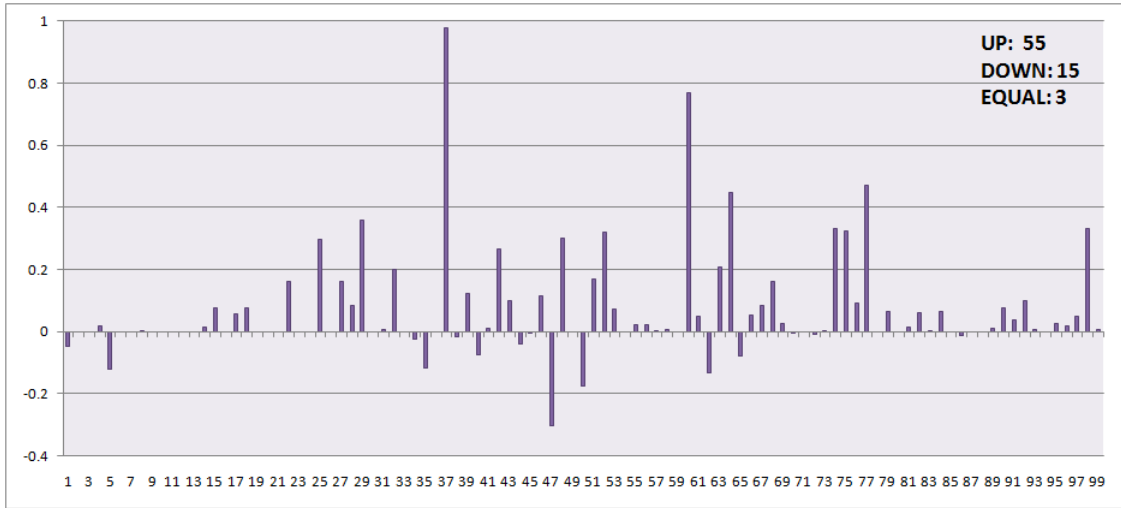


Figure 2: MAP Difference.

## 8. ANALYSIS

In this section, we mainly compare the performance of two retrieval model in Chinese monolingual retrieval - Okapi BM25 and KL language model. From the official evaluation results, the KL language model outperforms the Okapi BM25 with MAP 0.4187 against 0.3260. From Figure 2, the MAP difference between these two RUNs are showed. In the overall 73 topics which the official judgement has been provided, there are 45 topics with increased MAP and 15 topics with decreased MAP and 3 topics with unchanged MAP in KL-LM retrieval model comparing to Okapi BM25 model. By per-topic analysis, topic 38 was identified since it gains MAP 1 in the KL-LM retrieval model and MAP 0.0236 in Okapi BM25. The title part of the topic 38 is “夏雨和袁泉的关系是? ”. After Chinese segmentation, it is transferred into “夏雨和袁泉的关系是? ”. The P@10 for topic 38 in Okapi BM25 run is zero which means no relevant document is found in the top 10 documents. Checking the top document, several documents contain the individual Chinese character “夏” or “袁” which is a frequently used family name in China, but these documents are not relevant to “夏雨” or “袁泉”. In the top documents from KL-LM RUN, usually the Chinese name “夏雨” and “袁泉” appears together as “夏雨的女友、著名演员袁泉前不久在香港艺术节上演出音乐话剧《琥珀》”.

Another example is from topic 61 where the title part is “郭台铭是哪家公司的总裁(董事长)? ”. For this topic, the MAP is 0.1807 in Okapi BM25 Run and 0.9512 in KL-LM Run. Checking the first document XIN-CMN-20050205.0091 in the ranking list of Okapi BM25 Run, it contains the Chinese character “铭” several times like “陈德铭是中共十六大代表, 九届、十届全国人大代表。”. “铭” is not a frequently used character in Chinese which means it has high BM25 score in this document. It can explain why document XIN-CMN-20050205.0091 has higher rank in Okapi BM25 Run since “铭” appears in the query also.

Comparing the Okapi BM25 model and KL-LM retrieval model, Okapi BM25 treats the whole document as a bag of word and KL-LM treats the document as a language model. From the perspective of bag of words, usually the documents

containing more terms in query with high BM25 score have higher ranks; from the perspective of language model, documents whose language model produce the query with higher probability have higher ranks. This difference explains why the KL-LM model favors the document containing query terms in sequence.

In past research experiments in TREC [14, 12, 13], the Okapi BM25 has been regarded as a robust state-of-art algorithm in IR research. In the ACLIA2 ir4qa task, it does not perform well. Through our observation, for those queries containing Chinese name usually the Okapi BM25 model can get good results. The reason is due to the failure of the Chinese segmentation, the Chinese names are segmented into individual character. These Chinese characters in Chinese names are usually used in the document as different meaning, then documents containing these characters in Chinese name with different meaning have higher ranks in Okapi BM25 model. In English, usually the word in names do not have different meaning which is the big difference with Chinese. This can explain the failure cases in our Okapi BM25 Run.

The two paired t-test [17] is conducted on the Okapi BM25 Run and the KL-LM Run and the P value is less than 0.0001 and the difference of these two RUNs is considered to be significant by conventional criteria.

## 9. CONCLUSION

CLIR has been researched for a long time and the results comparing with the monolingual retrieval task has increased much in recent years. In NTCIR-5, out best RUN only get 35% of the monolingual retrieval effectiveness. But today, 79.94% performance has been gained in this task.

In this paper, we mainly described our translation system and retrieval method for IR4QA task. Comparing to the query translation method by Google translate online service, the document translation method does not perform well. And the query expansion from external resource method which has got good performance in monolingual task does not get better result in CLIR task. Also we demonstrate and analyse that the KL-Divergence language modeling method

performs better than Okapi BM25 model in a Simplified Chinese news retrieval task.

Further investigation including adjust of the parameters of relevance feedback and query expansion after query translation should be done in our future research. Our future work on CLIR research will focus on the judgement of query expansion on external resources since the results show that this method will improve the retrieval results and others don't., We will apply machine learning method to determine whether we should apply the query expansion for a particular query.

## 10. ACKNOWLEDGMENTS

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project at DCU. We also thank Dr. Jinhua Du for providing us with the details of the Google translation system and the DCU's statistical machine translation system.

## 11. REFERENCES

- [1] The Google Statistical Machine Translation System for the 2008 NIST MT Evaluation. World Wide Web. <http://www.itl.nist.gov/iad/894.01/tests/mt/2008/>.
- [2] NIST 2008 Open Machine Translation Evaluation - (MT08) Official Evaluation Results. World Wide Web. <http://www.itl.nist.gov/iad/mig/tests/mt/2008/>.
- [3] P.-C. Chang, M. Galley, and C. D. Manning. Optimizing chinese word segmentation for machine translation performance. In *StatMT '08: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Morristown, NJ, USA, 2008. Association for Computational Linguistics.
- [4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.
- [5] J. Du, Y. He, S. Penkale, and A. Way. MATREX: the DCU MT system for WMT 2009. In *WMT 2009 - Fourth Workshop on Statistical Machine Translation*, March 2009.
- [6] D. Graff. Chinese gigaword third edition. LDC, 2007.
- [7] H. Hoang, A. Birch, C. Callison-burch, R. Zens, R. Aachen, A. Constantin, M. Federico, N. Bertoldi, C. Dyer, B. Cowan, W. Shen, C. Moran, and O. Bojar. Moses: Open source toolkit for statistical machine translation. pages 177–180, 2007.
- [8] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA, 2003. Association for Computational Linguistics.
- [9] J. Min, L. Sun, and J. Zhang. ISCAS in English-Chinese CLIR at NTCIR-5. In *Proceedings of the 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, pages 123–129, Tokyo, Japan, 2005.
- [10] H. N. Reinhard Kneser. Improved backing-off for m-gram language modeling. In *In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 181–184. IEEE, 1995.
- [11] S. Robertson and K. Spärck Jones. Simple, proven approaches to text retrieval. Technical Report UCAM-CL-TR-356, University of Cambridge, Computer Laboratory, Dec. 1994.
- [12] S. Robertson, S. Walker, M. Beaulieu, M. Gatford, and A. Payne. Okapi at trec-4. In *In Proceedings of the 4th Text REtrieval Conference (TREC-4)*, pages 73–96, 1996.
- [13] S. Robertson, S. Walker, M. Beaulieu, and P. Willett. Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive track. *In*, 21:253–264, 1999.
- [14] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. pages 109–126, 1996.
- [15] T. Sakai, N. Kando, C.-J. Lin, T. Mitamura, H. Shima, D. Ji, K.-H. Chen, and E. Nyberg. Overview of the ntcir-7 aqlia ir4qa task. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, Tokyo, Japan, 2008.
- [16] T. Sakai, H. Shima, N. Kando, R. Song, C.-J. Lin, T. Mitamura, M. Sugimoto, and C.-W. Lee. Overview of ntcir-8 aqlia ir4qa. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, Tokyo, Japan, 2010.
- [17] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632, New York, NY, USA, 2007. ACM.
- [18] X. Yang, G. J. F. Jones, and B. Wang. Query dependent pseudo-relevance feedback based on wikipedia. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 59–66, New York, NY, USA, 2009. ACM.
- [19] Z. Yin, M. Shokouhi, and N. Craswell. Query expansion using external evidence. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 362–374, Berlin, Heidelberg, 2009. Springer-Verlag.
- [20] R. Zens, F. J. Och, H. Ney, and L. F. I. Vi. Phrase-based statistical machine translation. pages 18–32. Springer Verlag, 2002.
- [21] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410, New York, NY, USA, 2001. ACM.