Passage-level Evidence for Cross-Language Information Retrieval

Jae-Hyun Park
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst, MA
01003
jhpark@cs.umass.edu

Bruce Croft
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst, MA
01003
croft@cs.umass.edu

ABSTRACT

Machine translation (MT) techniques can be used to generate a query in a target language from a query in a source language for the cross-language information retrieval (CLIR). Recent MT systems have advanced enough to generate translations which are human-readable, However, translation error is still a serious impediment which hurts the effectiveness of a CLIR system. To compensate for defects in a machine-translation result, we propose a method using passage-level evidence. By combining a document retrieval model with a passage retrieval model, we prevent the retrieval model from assigning a high score to a non-relevant document because of translation error. The retrieval model incorporating passage retrieval shows better results than a document retrieval model. In particular, the passage retrieval model achieves more improvement when the translation quality of queries is relatively low.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory

Keywords

Cross Language Information Retrieval, Passage-based retrieval model

1. INTRODUCTION

Cross-language information retrieval (CLIR) has been studied to satisfy users' information needs from digital content that is not written in their native language. As the world has been globalized and users need to access information in different languages more frequently for various purposes such as searching web contents, communicating through social network services, browsing expert-knowledge and so on, the necessity of the CLIR system become more crucial. While machine-readable dictionaries have been used for CLIR in the past, recent work using machine translation (MT) systems show reliable results [7]. Our preliminary experiments using machine-translated queries show competitive or better results than queries in a source language.

Although the machine translation technique has progressed enough to produce human-readable translation results, translation errors are still a serious impediment to a CLIR system. People can guess the correct meaning of a defective translation result by considering the overall context of translation results and filling in missing information with his background knowledge. On the other hand, background knowledge is inapplicable for an information retrieval (IR) system. Considering the context of a machine-translation result may introduce unexpected errors to the IR system. Thus, IR systems are affected more by translation errors than people.

To compensate for defects in machine-translation results, we use passage-level evidence for the CLIR system. Previous work shows that passage-level evidence can improve the effectiveness of an IR system when documents are long and span different subject areas [3]. Lie and Croft showed that passage retrieval can be implemented in a language modeling environment. By evaluating whether a syntactically coherent unit covers the overall topic of a query, passage-level evidence prevents the IR system from being affected by partially relevant documents. For example, Eyal et al. exploit passage-level evidence to consider inter-document similarity for a re-ranking method [2]. The passage-based information feature based on passage-level evidence outperforms other inter-document similarity features.

In this paper, we assume that incorrect translation results are not related to each other. Therefore, correct translations can dominate translation errors within a passage because correct translations are coherent with each other. By combining a passage-level evidence with a document retrieval model, we aim to make the CLIR system be robust against translation errors in queries. Moreover, we also expect the passage-level evidence can utilize imprecise but reasonable translation results. For example, the translation of "human rights" by the Google translate API is " 人类 权利 " [1] [2]. Although " 人类 " and " 权利 " are direct translation of "hu-

[1] https://developers.google.com/translate/

[2] The Google translate API has been ameliorated and output the correct translation for this case now.

man” and “rights”, “ 人权 ” (the abbreviation form of “ 人类 权利 ”) is the spontaneous expression in Chinese corresponding to “human rights”. The passage retrieval model can handle this kind of not precise but tolerable translation by assigning a high score to a document when “ 人类 ” and “ 权利” ” occur together in the syntactically coherent unit of text.

In the remainder of this paper, we describe the document retrieval model with passage-level evidence. Then, we present our experimental setup and analysis of experiments. Finally, we present our conclusions and future work.

2. PASSAGE RETRIEVAL MODEL

A passage can be defined in various ways. Callan used three approaches to define passages in a document: paragraphs, bounded-paragraphs and fixed-size window [1]. The paragraphs and bounded-paragraphs approaches use a set of collection-specific heuristics. In the window approach, a document D is divided into passages according to a number of words. Contrary to expectations, retrieval based on paragraphs and bounded-paragraphs performed poorly compared to the fixed-size window. Callan concluded that length-based criteria used for merging and dividing paragraphs could not organize text by content and more semantic information would be required to improve the effectiveness of a passage retrieval model. In other work, a passage retrieval model based on a fixed-size window shows good results [?, 3].

Therefore, we use the fixed-size window approach. A document D is divided into passages by a fixed-size window, w; $Passage(D, w) = \{m_0, ..., m_J.\}$ by a fixed-size window w. The first passage m_0 starts from the first word in a document and contains w words. Then, a window moves $w/2$ words for the next passage m_1. Thus, all passages overlap the half of words with its adjacent passages.

We define a passage retrieval model by substituting a document with a passage. For example, the potential function of the Markov random field model [4] with a passage m_j can be defined as follows:

$$\begin{aligned}\psi_{T,m_j}(c) &= \lambda_T logP(q_i|m_j) \\ &= \lambda_T log\left[(1-\alpha_{m_j})\frac{tf_{q_i,m_j}}{|m_j|} + \alpha_{m_j}\frac{cf_{q_i}}{|C|}\right] \quad (1)\end{aligned}$$

, in which a passage m_j is used instead of a document. Then, we interpolated the passage retrieval model with a document-level retrieval model. For example, the potential function in Eq 2 can be interpolated with the document-level potential function as follows:

$$\psi_T(c) = \lambda_{passage}\psi_{T,m_j}(c) + (1-\lambda_{passage})\psi_{T,D}(c) \qquad (2)$$

Because our approach of passage-level evidence substitutes a document with a passage, a passage retrieval model can be applied to any kind of a document-level retrieval model. In this paper, we use the Markov random field model with sequential dependency as a baseline model.

3. EXPERIMENTS

3.1 Experimental Setting

Table 1: Overview of TREC collections and topics. *Aver. Length* is the average length of documents in a collection.

	# Doc	Aver. Length of docs	Topic
TREC5	164,779	337	1–28
TREC6			29–54
TREC9	127,938	560	55–79

We conducted experiments using TREC English-Chinese CLIR collections. The collections consist of Chinese news articles collected from Xinhwa, People’s Daily, Hong Kong Commercial Daily, Hong Kong Daily News and Takungpao [?]. Table 1 shows the statistics of the collections. There are a total of 79 topics for these collections. Average differences in the lengths of these queries is shown in Table 2. Each topic consists of a title, description and a narrative part. Each of these three topic parts are expressed in both Chinese and in English. In this paper, we use title and description parts. Table 2 shows the statistics of the topics of three TREC collections.

The retrieval experiments are implemented using Indri toolkit [5]. We generate translated Chinese queries from English queries using Google translation API. We compare two versions of the Chinese topics, original Chinese queries and machine-translated Chinese queries. We use the sequential dependency variant of the Markov Random Field model as a baseline retrieval model. This model is parameterized with three weights, one for each of the potential functions. We use $\lambda_T = 0.85$, $\lambda_O = 0.10$ and $\lambda_U = 0.05$ for all experiments in this paper.

3.2 Experimental Results

Table 3 shows experimental results of the baseline system with the native Chinese queries and translated Chinese queries. Translated queries show competitive results with original Chinese queries. When we compare title queries and description queries, Title queries show better results than description queries. These results are consistent with the experimental results of [7]. Previous machine translation research has shown that additional context improves machine translation. However, we observe that longer queries actually contain more translation errors.

Table 2: Statistics of queries. *MT* means Chinese queries translated from English queries.

	Topics	Aver. length of <desc> query		Average length of <title> queries	
		Chinese	MT	Chinese	MT
TREC5	1-28	15.75	17.5	12.64	11.39
TREC6	29-54	18.30	20.11	12.00	12.07
TREC9	55-79	21.08	22.32	6.32	6.32

The effectiveness of translated description queries of TREC

Table 3: The mean average precision(MAP) of title and description queries. *Chinese* and *Translation* denote original Chinese queries and automatically translated queries, respectively. Numbers in parentheses represent improvement over original Chinese Queries.

		TREC5	TREC6	TREC9
Title	Chinese	0.2950	0.3572	0.2852
	Translated	0.2711 (-8.10%)	0.3765 (5.40%)	0.2927 (2.63%)
Desc	Chinese	0.2875	0.4195	0.2899
	Translated	0.2274 (-20.90%)	0.3927 (-6.39%)	0.2800 (-3.41%)

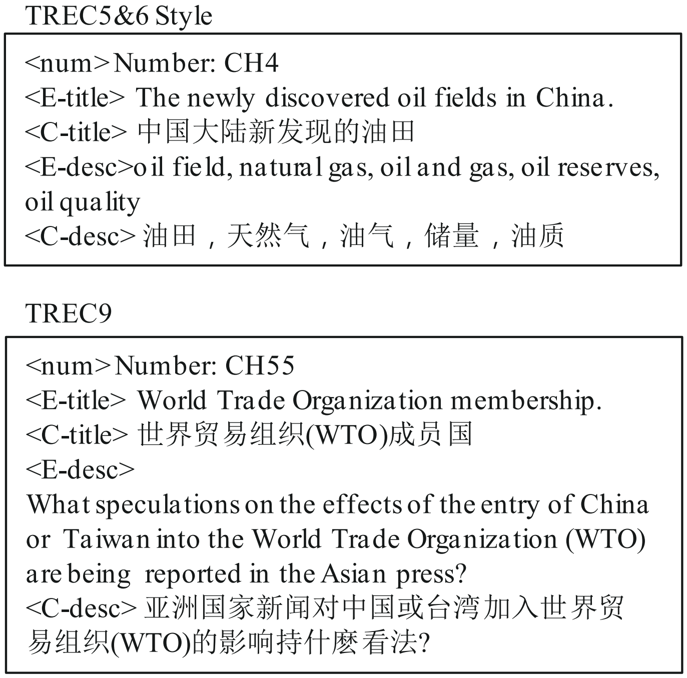
5 and TREC 6 decrease more substantially compared with TREC 9. It is important to note the differences in the style of title and description queries between TREC 5&6 and TREC9. In the TREC9 collection, a title query is a short phrase about a topic and a description query is a grammatically well-formed sentence about the topic as shown. On the other hand, in the TREC 5&6 collections, each description query is a list of phrases that is related to a topic. Title query is a phrase describing the topic in detail compared to phrases in the description query. They tend to be longer than TREC9 title queries, as shown in Table 2. Generally, they are grammatically well-formed but they are missed the subject and main verb components of a sentence. Figure 1 shows an example queries. Given that the description queries of TREC 5 and 6 do not provide good contexts to the MT system and it is reasonable to see introduce more translation errors. Our experiments show that this characteristic leads to a performance decrease in retrieval experiments, when using the translated description queries of TREC 5 and 6.

Recall that we defined a passage as a fixed window, this makes the size of the window an important parameter for any passage retrieval model. We have also seen that the average lengths of documents and topics varies in each collection, see Table 1 and Table 2. Therefore, it is reasonable to expect that the appropriate passage size would also vary between collections.

We show experimental evaluation of the passage retrieval model over different passage sizes in Table 4. We use the two tailed paired t-test at $P - value < 0.05$ for statistical significance test. This table shows that longer passages produce improved results for TREC 9, while shorter passages improve results for TREC 5&6. Documents in the TREC 9 collection are around twice the length of documents in the TREC 5 & 6 collections. This implies that a effective passage size could be a set fraction of the document length.

We analyze experimental results according to the translation quality. We use correlation in the number of words that occur in the original Chinese queries and in the translated queries as a measure of the quality of translation results as follows.

Figure 1: An example of title and description queries in the TREC 5&6 and TREC9 collections.



TREC5&6 Style

<num>Number: CH4
<E-title> The newly discovered oil fields in China.
<C-title> 中国大陆新发现的油田
<E-desc>oil field, natural gas, oil and gas, oil reserves, oil quality
<C-desc> 油田，天然气，油气，储量，油质

TREC9

<num>Number: CH55
<E-title> World Trade Organization membership.
<C-title> 世界贸易组织(WTO)成员国
<E-desc>
What speculations on the effects of the entry of China or Taiwan into the World Trade Organization (WTO) are being reported in the Asian press?
<C-desc> 亚洲国家新闻对中国或台湾加入世界贸易组织(WTO)的影响持什麼看法?

Table 4: Experimental results using translated queries. *Passage* denote experimental results with the passage retrieval model with a fixed-size window. Numbers in parentheses represent improvement over using only the document-level retrieval model.

		TREC5	TREC6	TREC9
Title	Translated	0.2711	0.3765	0.2927
	Passage(30)	0.2851* (5.16%)	0.3888* (3.27%)	0.2889 (-1.30%)
	Passage(100)	0.2833* (4.50%)	0.3907* (3.77%)	0.2916 (-0.38%)
	Passage(200)	0.2821* (4.06%)	0.3901* (3.61%)	0.2929 (0.07%)
Desc	Translated	0.2274	0.3927	0.28
	Passage(30)	0.2349* (3.30%)	0.4082* (3.95%)	0.2822 (0.79%)
	Passage(100)	0.2341* (2.95%)	0.4129* (5.14%)	0.2853 (1.89%)
	Passage(200)	0.2335 (2.68%)	0.4126* (5.07%)	0.2824 (0.86%)

** indicates statistically significant differences with the document retrieval model.*

Table 5: The comparison of MAP with and without the passage retrieval model. *WordRecall* denotes the ratio of matched words between Chinese and Translated queries as shwon Eq 3. *(T)* and *(D)* mean title and description queries, respectively.

	WordRecall > 0.7		WordRecall < 0.7	
	w/o Passage	With Passage	w/o Passage	With Passage
TREC5(T)	0.2884	0.2945 (2.1%)	0.2599	0.2741 (5.5%)
TREC6(T)	0.4058	0.4179 (3.0%)	0.3365	0.3523 (4.7%)
TREC9(T)	0.4193	0.4071 (-2.9%)	0.2332	0.2391 (2.5%)
TREC5(D)	0.2942	0.3027 (2.9%)	0.1696	0.1735 (2.3%)
TREC6(D)	0.4704	0.4943 (5.1%)	0.3020	0.3174 (5.1%)
TREC9(D)	06.4827	0.4779 (-1%)	0.2523	0.2558 (1.4%)

$$WordRecall(Q,T) = \frac{|Q \cap T|}{|Q|} \tag{3}$$

,in which Q and T are chinese words in a original Chinese query and a translated query, respectively. We used Stanford Word Segmenter [6] to extract Chinese words from queries. Based on this metric, we split translated queries into a high quality group and a low quality group according to the ratio of matching words in Chinese queries. A high quality group consists of translated queries, where at least 70% of the query words occur in the original Chinese queries. The other queries belong to a low quality group. Table 5 shows the comparison of experimental results between two groups. The passage retrieval model demonstrates a large improvement for queries in the low quality group.

Retrieval effectiveness for queries in the high quality group is not substantially affected by the passage model. We believe that the small proportion of translation errors does not affect the effectiveness of the document retrieval model. Queries in the low quality group contain more translation errors and the document retrieval model assign high scores to non-relevant documents based on these translation error. The passage retrieval model is able to ensure correctly translated query terms are able to identify useful documents, while reducing noise from incorrectly translated terms. In this way, the passage retrieval model prevent the IR system from assigning high scores to non-relevant documents, based on evidence from query.

4. CONCLUSION

Due to recent significant advances, machine translation systems produce human-readable outputs. Incorrect translation and unusual expression can be understand by people because we can fill in missing or imprecise meaning using the text context and our background knowledge. An information retrieval system cannot distinguish good and bad expressions in translation results. To solve this problem, we propose a method using a passage retrieval model. By considering all terms in a query together, the passage retrieval model aims to prevent the CLIR system from assigning a high score to non-relevant documents based on translation errors. In particular, experimental results show that the passage retrieval model improves the effectiveness more when the quality of translation results is low.

In this paper, we construct the passage retrieval model using a fixed window. However, an ideal passage differ according to the characteristics of documents and queries. We conducted experiments using the passage retrieval model based on three different sizes of a fixed window and the best size of a fixed window varies between queries. Therefore, our future work will involve developing a passage retrieval model based on variable-length arbitrary windows or other semantic and syntactic information.

5. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part under subcontract #19-000208 from SRI International, prime contractor to DARPA contract #HR0011-12-C-0016. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] J. P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 302–310. Springer-Verlag New York, Inc., 1994.

[2] E. Krikon, O. Kurland, and M. Bendersky. Utilizing inter-passage and inter-document similarities for re-ranking search results. *ACM Transaction on Information System (TOIS)*, 29(1):1–27.

[3] X. Liu and W. Croft. Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 375–382. ACM, 2002.

[4] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR*, pages 472–479, 2005.

[5] T. Strohman, D. Metzler, H. Turtle, and W. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, 2005.

[6] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. A conditional random field word segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing*, 2005.

[7] D. Wu and D. He. A study of query translation using google machine translation system. In *Computational Intelligence and Software Engineering (CiSE)*, pages 1–4. IEEE, 2010.