RALI Experiments in IR4QA at NTCIR-7

Lixin Shi, Jian-Yun Nie and Guihong Cao
Département d'informatique et de recherche opérationnelle, Université de Montréal
C.P. 6128, succursale Centre-ville, Montréal, Québec, H3C 3J7 Canada

Abstract

In this report, we examine what information retrieval techniques can help identify documents that contain answers to different types of question. In particular, we exploit different external resource according to the type of question. In particular, Wikipedia will be exploited for identifying personal names and their translation, as well as biography-related keywords. Google search engine is used to identify additional translations of personal names. Our experiments show that these techniques can significantly increase retrieval effectiveness.

Keywords: *Question answering, Information retrieval, Wikipedia, Passage retrieval, Personal name translation.*

1. Introduction

Question answering is usually composed of two main steps: passage/document retrieval that aims to identify passages/documents that may contain possible answers, and answer identification within these passages/documents. IR4QA (Information Retrieval for Question Answering) is a task of ACLIA (Advanced Cross-lingual Information Access), which focuses on complex cross-lingual questions answering (CCLQA) problems [13]. The task of IR4QA is to identify the documents which may contain an answer to the question. In this report, we describe the experiments carried out at the RALI lab of University of Montreal at IR4QA. We have explored several techniques to help identify candidate documents, in particular, by exploiting external resources such as Wikipedia and Google search engine. This report will describe the approaches as well as their effectiveness in IR4QA tasks.

In NTCIR7-ACLIA, we deal with four types of questions. Some examples for each type are shown below[1]:

- List/Event: *List major events in formation of European Union.*
- Biography: *Who is Alberto Fujimori?*
- Definition: *What is the Human Genome Project?*
- Relationship: *What is the relationship between Saddam Hussein and Jacques Chirac?*

The basic approach to identify candidate documents is to treat all types of question in the same way using a standard IR system or passage retrieval system. However, intuitively, each type of question should be answered by a specific type of answer and document. For example, for a question about a person, we can expect that the desired documents are more of biography type, i.e. documents describing all the aspects of the person such as education, birth and death, and so on. So, if we have some knowledge about each type of question and enhance the method with such specific knowledge, we can create a better IR model for retrieving the specific documents for it.

For many questions, we are faced with the problem of identification and translation of personal names. The existing tools for name recognition or translation are insufficient: several personal names in the questions are unknown. We also observe that resources such as Wikipedia [2] contain a list of persons in several languages. Such list can be used to help name identification and translation.

Many previous studies have tried to use external sources for question answering [2][3][5][10]. However, they usually try to use them in the later question identification phase, while the external knowledge is not fully exploited in the first retrieval phase. In our experiments, we intend to exploit external knowledge more extensively at the retrieval phase.

In this report, we focus on English to Chinese (both simplified and traditional Chinese) cross-lingual QA. In our experiments, we will consider the following problems:

- For biography-related question, we exploit Wikipedia to extract some biography-related keywords such as "born", "father", "studied", which are more often used in biographies than in

[1] Details can be found at http://aclia.lti.cs.cmu.edu/wiki/TaskDefinition#Format

[2] http://www.wikipedia.org/

other types of document. These keywords can help determine if a document is possibly the biography of someone. We will exam if these words are useful for retrieving biography documents.

- We extract personal names (both in English and Chinese) from Web resources, namely Wikipedia and Google search. The corresponding names in other languages are considered as their possible translations. With these names and possible translations, we hope to be equipped to recognize variations of the name of the same person, as well as their translations.
- Besides the above name translations, we will also use an existing machine translation system to perform question translation. In our case, we use Google translation.

Our final results show that integrating the knowledge extracted from Wikipedia can produce good performance. The traditional pseudo-relevance feedback (PRF) method can also significantly improve the performance.

The official runs we submitted to IR4QA are ranked in the middle among all the participants. However, we discovered a bug in our program used to produce the official results. After the correction of the bug, all the new results turn out to be better than, or similar to, the best official submissions. This indicates that our approach can be effective. In this report, we will present both the erroneous official results and the corrected results.

The remaining of the report is organized as following: Section 2 describes our specific processing on documents and questions. Section 3 describes the retrieval method we used. Section 4 describes our experiments and results. Finally, some conclusions and future work are outlined in Section 5.

2. Document and Question Processing

2.1 Question analysis and Text segmentation

A question answering system usually includes the following four modules [7]: (1) Question analysis, (2) Document retrieval, (3) Passage selection (4) Answer extraction.

Question analysis is the first step of QA, which determines the type of question and the important concepts involved, as well as their relationships. To classify questions, the approaches proposed in the literature include machine learning-based approaches such as SVM [4][15][14], rule-based approaches [6], or both of them [9]. In our implementation, we used a simple rule-based approach for both Chinese and English question classification. For example, if a question includes keywords such as 谁是, 是何人, 是哪位, 生平事迹 in Chinese or "*who*", "*biography*" in English, we classify it into biography category. This method successfully detected biography questions, and produced only 3 errors for Chinese question and 4 errors for English questions out of 40 questions. For all types of question, our classification accuracies are 92% (Chinese questions) and 89% (English questions).

For document retrieval, previous studies show that passage retrieval usually leads to better result than whole document retrieval [10]. In our experiments, we use the same approach. The question is how we should segment a document into passages. Callan [1] listed the following methods:

- Discourse-based model: The segmentation is performed according to structure properties, such as sentence and paragraph.
- Semantic model: It divides document into semantic pieces.
- Window model: It segments a document into text portions of fixed size.

The third method is the simplest one to use, and this is the one we choose to use in our experiments. We use overlapping windows, which guarantee that a complete sentence is included in at least one passage. In our experiments, we set the passage length to 250 Chinese characters.

A special processing is about document title. The title of a document usually contains important information, which is shared by all passages under it. Therefore, when we divide document into passages, we duplicate the title in each passage.

To build an IR system for Chinese, the next step is to choose the index unit. Different from most European languages, there is no natural word boundary in Chinese texts. According to previous studies, we can use one of the following two methods to define index units [12].

(1) Cutting a text into n-grams – unigrams or bigrams of characters

(2) Segmenting a text into words by a word segmentation tool.

The advantage of using character n-grams is that it does not require any linguistic knowledge. Previous studies showed that n-gram-based approaches can achieve a performance comparable to word-based approaches. However, using an n-gram-based approach, it would be more difficult to integrate knowledge extracted from the Web later. Therefore, we choose to use words as our index units.

Our word segmentation software is ICTCLAS[3], which is developed by the Institute of Computing Technology of the Chinese Academy of Sciences.

[3] http://ictclas.org

ICTCLAS is based on multi-layered hierarchical hidden Markov models, and integrates Chinese word segmentation, POS tagging, NE recognition, disambiguation and unknown words recognition functions. We use ICTCLAS for both words segmentation and tagging.

The ICTCLAS system (the version we used) only supports simplified Chinese. For traditional Chinese documents, we have to first convert them into simplified Chinese, then use ICTCLAS to process them.

2.2 Identification of Personal Names and their Translations from External Resources

External knowledge resources have been used for question answering in many previous studies. Wikipedia is one of the most used resources. The knowledge is usually used in QA phase or to help in cross-lingual question translation. For example, Ferrández [3] uses the Inter Lingual Index (ILI) model in CL-QA, which allows references between words in different languages. Wikipedia is used in order to overcome the drawback of ILI that contains very few proper names.

Following the same idea, we exploit Wikipedia in our work and focus on using it in the information retrieval phase. Two kinds of knowledge are used in our IR system:

- Personal names extracted from Chinese and English Wikipedia articles;
- Some keywords or sentence patterns from Chinese Wikipedia biography documents.

We detect personal name entries in Wikipedia by looking at Wikipedia Template information. For example, if we detect the template such as "{{[BD]|…}}" or "[[Category:<year>(出生|逝世)]]" in an article, it implies that the entry of this article is a personal name. Then, from redirection and cross-reference information, we can build a list which includes personal names in various forms and languages.

In Wikipedia, the corresponding entries in different languages are cross-referenced. Therefore, while we extract personal names in Chinese, we also collect the corresponding names in English, and this allows us to build up a translation table to map English names to Chinese names.

Most elaborated word segmentation approaches can achieve a segmentation accuracy of over 90%. This performance has been believed to be sufficient for IR. Nie [11] shows that a slight difference in segmentation accuracy does not have a strong impact on IR performance. However, if a personal name is wrongly segmented, the performance will be greatly deteriorated. For example, without the personal names acquired from Wikipedia, the personal name 小泉纯一郎 will be segmented into 小泉, 纯一, and 郎 by ICTCLAS. In the retrieval phase, many irrelevant passages will be retrieved for the corresponding question. To avoid this situation, we detect personal names in Wikipedia articles, and add them to the dictionary of ICTCIAS to be used in word segmentation.

For some personal names, we may not find them in Wikipedia. To identify the translation(s) of these names, we exploit a search engine (Google in our case). In fact, in many Web documents, the translation of a personal name often follows it, such as in "搜狐公司董事局主席兼首席执行官张朝阳 (Charles Zhang)"[4], where the name 张朝阳's translation is put in parentheses after it. This phenomenon is widely spread. So we request the search engine for documents written in Chinese but contain the English name. Then, we try to find translation according to the pattern "English-name (Chinese-name)" or "Chinese-name (English-name)".

Using this method, we successfully found the Chinese names of "Zeng Yitao" and "Charles Zhang" as 曾溢滔 and 张朝阳. This method is similar to the one used in [16]. The above method turns out to be a good complement to the previous name translation method.

2.3 Extracting Biography-Related Keywords

For biography questions, the desired documents should preferably describe the biography of a person. A typical biography document would contain words such as date of birth (death), education, occupations, etc. of the person. The occurrence of such words in a document can distinguish a biography from a simple document about a person. In our approach, we exploit such special words for biography documents.

To extract biography-related words, we try to determine the most distinctive words in biographies on Wikipedia. We consider all Wikipedia documents whose title is a personal name as a biography document. Therefore, we can separate Wikipedia documents into two groups: biography and other. We compare the frequencies of words appearing in the two groups. The most distinctive words which have a high frequency in biography documents and a low frequency in other documents, are considered to be biography-related. In our experiment, we select 110 words according to term's document frequency. Below are some examples of biography-related words extracted:

出生 (born), 毕业 (graduate), 大学 (college, university), 岁(age, year).

The extracted biography-related terms are used to expand biography questions, so that biography-like

[4] http://corp.sohu.com/20060507/n243126051.shtml

passages can be ranked higher. See Section 3.2 for more details.

2.4 Question Translation

Most implementations of current CL-QA systems are based on the use of on-line translation services [8]. Similarly, we also use an existing translation tool - Google on-line translation - as our basic question translation method. However, the online translation tool cannot produce the correct translation for personal names. To solve this problem, we use the name list extracted from Wikipedia to provide additional personal name translations. Moreover, this approach allows us to obtain name variants for the same person. For example, the translations of *Clinton* can be 柯林顿 or 克林顿. All the possible translations are thus added into the query and this is equivalent to query expansion.

We use a two-stage translation approach: (1) We first use Google to translate the whole question with the personal name replaced by a variable – this indeed produces the translation of a question pattern; (2) Then we add into the translated pattern the translations of the personal name obtained from Wikipedia.

Let us use an example to show how a question is translated: "*What is the relationship between Yang Liping and dance?*"

An English POS tagger, OpenEphyra[5], is used to tag the question. The POS tagging result is:

What<WP> is<VBZ> the<DT> relationship<NN> between<IN> Yang Liping<NEperson> and<CC> dance<NEmusicType|NEsport|NEstyle>

Here, *"Yang Liping"* is identified as a personal name. To find more translations for this personal name, we replace it by *x* in the original question and make it translated by Google. This produces the following translation for all the words except for the personal name:

之间的关系是什么 *x* 和跳舞

Then, we replace *x* by the Chinese name(s) of "Yang Liping" from the list of personal names and translations extracted from Wikipedia. We obtain the following final translation:

之间的关系是什么杨丽萍和跳舞

3. Passage and Document Retrieval

We use Indri[6] from Lemur Toolkit as our basic IR system. Indri combines an inference network with a language modeling approach, and supports structured query operators.

3.1 Combining Words and Word-Bigrams

As words have been segmented from texts and questions, we can use them as the basic indexing units. However, we will encounter the problem of low precision for proper nouns such as 移动梦网 (ONTERNET), which is the name of a web site. Using word segmentation, this proper noun is segmented into three words 移动(move/mobile), 梦(dream), 网(net), and they are then considered to be three independent words during the retrieval process. In order to take into account the possible relationships between different segmented words, we use word bigrams as complementary indexing units. For the above example, we will have 移动+梦 and 梦+网 as additional indexes. They are more closely related to the word 移动梦网 than the individual words 移动, 梦, and 网.

In Indri retrieval system, we combine the original query term and bigrams as follows:

$$\#weight(1.0\ \#combine(t_1,...,t_n)\ w_{bi}\ \#combine(\#1(t_1,t_2),\#1(t_2,t_3),...))$$

where $t_1, \ldots t_n$ are the segmented words; 1.0 and w_{bi} are the respective weight set for words and bigrams. In our case, w_{bi} is empirically set at 0.1.

3.2 Expending Biography Questions

For biography questions, we expand them by the extracted biography-related words.

More specifically, a biography question will be rewritten as follows in Indri:

$$\#\ weight\ \ (1.0\ \#\ combine\ \ (t_1,...,\ t_n)\ w_{bio}\ \#\ wsum\ \ (w_1,...\ w_m))$$

where $t_1,\ldots, t_n$ are original terms in the question, $w_1,\ldots, w_m$ are the extracted biography-related keywords, and w_{bio} is a weight expressing the importance of the biography words in query. Our test results show that a weight set at around 0.2 is the most appropriate.

3.3 Expanding Personal Names in Monolingual Retrieval

For questions involving a personal name, the personal name is a very important key. However, the name often has various forms, especially for translated foreign names. The following examples show some name variants of the same person:

- Chiang Kai-she: 蒋中正, 蒋介石

[5] http://www.ephyra.info/

[6] http://www.lemurproject.org/indri/

- Clinton: 柯林顿, 克林顿
- Ryutaro Hashimoto: 桥本龙太郎, 桥本

Therefore, finding various expressions of a personal name is useful for retrieving more relevant documents.

As we have extracted personal names and their translations from Web resources, we can use them to determine the possible name variants: All the possible translations of a given name in English are considered to be variants. Therefore, one Chinese translation (e.g. 蒋中正) can be used to expand another translation (e.g. 蒋介石) in the monolingual retrieval process.

The expansion is implemented via the following Indri query:

$$t_1,...,t_n,\#max(p_{11},..p_{1m_1}),\#max(p_{21,..,}p_{2m_2})...$$

where p_{ij} is one of variant forms of a personal name t_i.

3.4 Determining Candidate Documents

In our approach, we first perform passage retrieval, then try to identify the candidate documents, as is required in IR4QA. We have tested two approaches.

Suppose that we have retrieved a set of passages and some of them are from the same document. The passages from the same documents are sorted and these according to their scores. Suppose that v_i is the value of *i-th* best retrieved passage of a document, and v_{doc} is the document relevant value to be assigned. v_{doc} is determined in the two following ways:

(1) Best Passage (Best-P): The document is directly assigned the score of the best retrieved passage from it, i.e., $v_{doc} = v_1$.

(2) All Passage (All-P): We consider all retrieved passages from a document. It is based on the following assumption: the more passages retrieved from a document, the more useful the document. However, if we just sum up the value of each passage, long documents will tend to have higher values. To avoid this problem, we use a degraded sum as follows:

$$v_{doc} = v_1 + v_2/4 + v_3/8 + ... + v_n/2^{n+1}$$

That is, the passages added later on will have degraded additional value for the whole document.

In addition, we apply some filtering to remove noisy passages before creating document rank list. A passage is discarded if it does not satisfy the rules or patterns of the given question type. We learn or defined manually some patterns for biography, relationship, and definition questions as follows:

- Relationship question. To find the relationships between *x* and *y* (keywords), a passage must include both of them.
- Definition question. The term to be defined should have definition words following it. Here, we simply consider all verbs as possible definition words. This means that the word to be defined should be followed by a verb in a passage for the passage to be kept.
- Biography question. We try to find some sentence patterns related to biographies. We consider some typical three-word patterns (triples) for biographies. In a triple, the first term is personal name or personal pronoun; the second is biography keyword or verb; and the third is biography keyword, verb, or empty. Such general patterns are established manually. Then using regular expressions, we can identify instances of such patterns in texts. For example "name/年(year)/出生(born)", "name/加入(join)/null" correspond to instances of such patterns. This kind of triples is often observed in biographies on Wikipedia. We have extracted about 600 triples from Wikipedia biographies.
- Event. No special rule is used.

4 Experiments

Our experiments are performed on both simplified (CS) and traditional (CT) Chinese collections. Table 1 shows some characteristics of these collections.

Table 1: Characteristics of IR4QA Chinese collections

	Collection	#doc	#question
CS	Xinhua (1998-2001) Zaobao (1998-2001)	854,645 1,071,597	97
CT	CIRB020 (1998-1999) CIRB040(2000-2001)	761,835 2,501,251	95

We submitted four sets of IR4QA results: CS-CS (Simplified Chinese monolingual), CT-CT (Traditional Chinese monolingual), EN-CS (English to Simplified Chinese), and EN-CT (English to Traditional Chinese). We used personal name expansion, biography-related words, and combination of words with bigrams in all of our official runs. Pseudo-relevant feedback is used in some of the submitted runs: RALI-*-02, -04, and -05; and passage filtering is used in RALI-*-01, -02 and -03.

However, after the submission, we discovered a bug in our program for document re-ranking – all of our submissions are based on "All-P" method (see Section 3.4), but we assigned the best passage the lowest weight instead of the highest weight. Therefore, our official submission is not good and it does not reflect the usefulness of the proposed methods.

In addition to the official runs, we also report the corrected results to provide a better idea of the approaches we proposed.

We use the evaluation package from NTCIR web site[7]. It includes three evaluation metrics [13]. The latter two metrics can handle grade relevance.

- Average Precision (AP)
- Q-measure
- a version of normalized Discounted Cumulative Gain (MSnDCG1K)

In the following sections, we will discuss about different experiments as well as the impact of different techniques.

4.1 Monolingual results

In Table 2, we compare our baseline (after bug correction) to the best two official results and our official result (RALI official - with a bug in our program). We also show the results once we fixed the bug after the official submission. In the corrected baseline method, we use none of enhancements described earlier, and only use segmented words as index units. The enhancements will be discussed later.

As we can see, our baselines are slight better (for CS) or close to (for CT) the best official results. We test two document ranking method: according to best-passage and all-passage as described in Section 3.4. They are very close to each other. Therefore, either of them is a reasonable baseline.

We also notice that the effectiveness in simplified Chinese (CS) is higher than in traditional Chinese (CT). This may be attributable to the utilization of segmentation tool ICTCLAS, which is a system designed for simplified Chinese. In fact, for POS-tagging and question analysis, traditional Chinese may have some subtle differences with simplified Chinese. For example, in the question of ACLIA1-CT-T174, the phrase 芮氏规模 (the Richter scale) is segmented into 芮, 氏 and 规模. "芮氏" cannot be recognized as a word because "里氏规模" is usually used in simplified Chinese. We expect that a system built for traditional Chinese could produce better results.

Table 2: comparing our baseline to the best official results

title-only		Best-1 official	Best-2 official	*RALI* official*	Baseline	
					Best-P	All-P
CS	AP	.6337	.5930	*.4684*	.6399	.6428
	Q-measure	.6490	.6055	*.4812*	.6486	.6521
	MSnDCG1K	.8270	.7951	*.7242*	.8235	.8269
CT	AP	.5839	.5521	*.3952*	.5530	.5596
	Q-measure	.6018	.5724	*.4096*	.5678	.5747
	MSnDCG1K	.7873	.7656	*.6516*	.7582	.7629

[7] http://research.nii.ac.jp/ntcir/tools/ir4qa_eval.tar.gz

Our official runs have obtained the highest score of novelty (unique relevant documents found by one team but not by others) among the participants. This score is not affected by the reversed ranking of the documents of our submissions. The novelty (unique documents) is measured using the results of both monolingual and cross-lingual retrievals together. In our official run results, 63 novel documents are found for all the questions in CS and 32 for CT (see Table 3).

The high score of novelty is largely attributable to question expansion using variants of personal names and pseudo-relevance feedback. For example, for question T42 - 谁是本拉登 (Who is Osama bin Laden), we have found 55 unique relevant documents, because:

(1) Personal name expansion by Wikipedia (PNE) helped identify the following variants of the name 本拉登: 宾拉登, 乌萨玛, 拉登, 本拉丹, 拉丹, 奥撒玛本拉登, ...

(2) Pseudo-relevant feedback (PRF) helped identify several strongly related terms such 塔利班 (Taliban), 基地 (base), 恐怖主义 (terrorism),...

Among the 256 relevant documents for this question, we found 221. This is much more then the next system, which found 157 (MITEL-EN-CS-03T). Although these unique documents are found only for a small number of topics, it shows that our method can determine more relevant results than the other participating methods.

Table 3: Maximum number of unique relevant documents retrieved for a single run of each team (From Table 15 and 16 of [13])

Simplified Chinese		Traditional Chinese	
run name	num	run name	num
RALI-EN-CS-04-T	63	RALI-EN-CT-05-T	32
OT-CS-CS-01-T	18	OT-CT-CT-05-T	5
CYUT-EN-CS-03-DN	17	CYUT-EN-CT-03-DN	3
HIT-EN-CS-02-T	10	NTUBROWS -CT-CT-05-T	2
CMUJAV-EN-CS-01-T	7	MITEL-CT-CT-04-T	1

Impact of the biography words (Bio)

From Wikipedia data, we extracted 110 words which are related to biography document, and added them into the original biography question. There are 40 biography questions for CS and CT. After adding biography terms, 18 topics have been improved by more than 0.5% in MAP, and only 3 topics decreased by more than 0.5%.

The following questions are examples for which MAP has been improved (MAP of baseline → MAP after adding biography words):

- CS-T376 谁是奥尼尔(Who is Shaquille O'Neal): 0.6215 → 0.6659 (+7.1%)

Table 4: Impact of using biography keywords, filtering, bigrams, PRF, and PNE (*:t-test<0.05, **: t-test<0.01)

		Baseline	Filter		Bio(0.2)		Bigram(0.1)		PRF(0.5)		PNE		Bio+Bi+PRF+PNE	
				% B		% B		% B		% B		% B		% B
CS-CS	AP	.6399	.6390	-0.1	.6412	+0.2*	.6429	+0.5	.6763	+5.7**	.6495	+1.5*	**.6888**	+7.6**
	Q-measure	.6485	.6486	+0.0	.6499	+0.2*	.6519	+0.5	.6848	+5.6**	.6569	+1.3*	**.6972**	+7.5**
	MSnDCG1K	.8235	.8243	+0.1	.8238	+0.0	.8265	+0.4	.8507	+3.3**	.8303	+0.8	**.8603**	+4.5**
CT-CT	AP	.5530	.5530	-0.1	.5538	+0.1	.5517	-0.2	.5896	+6.6**	.5670	+2.5**	**.6002**	+8.5**
	Q-measure	.5678	.5561	-0.3	.5688	+0.2*	.5658	-0.3	.6076	+7.0**	.5824	+2.6**	**.6150**	+8.3**
	MSnDCG1K	.7582	.7863	-0.2	.7591	+0.1*	.7582	0.0	.7883	+4.0**	.7692	+1.5*	**.7923**	+4.5**

- CS-T379 谁是邓肯(Who is Duncan): 0.6510 → 0.6647 (+2.1%)
- CS-T380 莫扎特是谁(Who is Mozart): 0.4639 → 0.4872 (+3.8%)
- CT-T194 谁是小泉纯一郎(Who is Junichiro Koizumi): 0.5630 → 0.5940 (+5.5%)
- CT-T386 请告诉我杨振宁的生平事迹 (Tell me the biography of Chen Ning Yang): 0.8266 → 0.8412 (+1.8%)

Below are some examples for which MAP is decreased:

- CS-T68 谁是张瑞敏 (Who is Zhang Ruimin): 0.9111 → 0.9049 (-0.7%). Some documents describe other people but mention 张瑞敏. These documents are kept and some of them further enhanced. This is why MAP is slightly decreased.
- CT-T397 谁是李宁(Who is Li Ning) : 0.8499 → 0.8228 (-3.2%). The degradation of MAP is mainly due to a segmentation problem of personal names: another personal name 李宁远 is wrongly segmented into 李宁 and 远 in some contexts. Therefore, the retrieved biography documents for 李宁远 degraded the result.

Overall, the improvement in MAP is relatively small (see Table 4), however, some of the improvements are statistically significant. An analysis reveals that one reason for such a limited improvement is that the relevant documents for biography questions are not exactly of the same type of biography documents on Wikipedia. They can be a document containing some description about a person, without the description can be considered as a biography of the person. Therefore, the assumption on which our approach was built, that biography questions aim to find biographies, is too strong for this the IR4QA task. Biography questions should be interpreted more broadly.

Impact of the personal name expansion (PNE)

We test the impact of performing expansion on personal names via the translations extracted from the Web. The result (Column PNE in table 4) shows that personal name expansion is helpful. Most of the improvements are statistically significant.

Personal name expansion solves a key problem in Chinese IR, i.e. name variations at different locations (mainland, Taiwan, Hong Kong, and overseas). By expanding the question by related name variants, we can retrieve documents using different variants, thus increase recall.

In addition, as our method for extracting the possible translations of a personal name is also very selective (i.e. they should correspond to some specific patterns), it does not produce much wrong translations. So, the personal name expansion process will not introduce much noise into the question.

We found 34 topics contain various personal names by named entity processing. After personal name expansion, 12 topics are improved more than 1% in MAP, examples are listed below. Others topics are only changed within ±0.5%. Below are some questions which have been improved by PNE:

- CS-T42 谁是本拉登 (Who is Osama bin Laden). PNE: 本拉登 → 宾拉登, 乌萨玛 (Osama), 拉登, …, 拉丹. MAP: 0.0508 → 0.2112 (+315%).
- CS-T324 谁是小泉纯一郎 (Who is Junichiro Koizumi). PNE: 小泉纯一郎 → 小泉 (Koizumi). MAP: 0.6147 → 0.7196 (+17%)
- CS-T197 藤森是谁 (Who is Alberto Fujimori). PNE: 藤森 → 藤森谦也. MAP: 0.8093 → 0.9347 (+16%)
- CS-T198 谁是舒马克 (Who is Michael Schumacher). PNE: 舒马克 → 舒马赫, 拉夫舒马克. PNE: 0.0468 → 0.0986 (+111%)

In all the above examples, we can see that some common variants of the personal names are included. This allowed us to identify additional relevant documents about these persons.

Impact of using bigram (Bi)

Using bigram can solve some ambiguities due to independent words. For example the separate terms 法国 (France) and 世界杯 (World Cup) can describe both "France World Cup" and "France football team in any World Cups". However, using the word bigram 法国+世界杯 (France World Cup), the token becomes

unambiguous. This example shows the possible advantage of using bigrams.

When we combine words and bigrams for monolingual Chinese retrieval, we have to set a weight w_{bi} for bigrams. This parameter is quite sensitive. For simplified Chinese, small values (0.1) of this weight can lead to some improvements. The following are some results for using bigrams:

- In topic CS-T78, 列举法国世界杯引发的争议 (List the disputes triggered by the France World Cup), the key words are 法国 (France), 世界杯 (World Cup), 引发 (trigger), 争议 (dispute). In the baseline method, the MAP is only 0.0305. The main reason is that most top ranked documents talk about the football teams or players of France in World Cup, but not about the France World Cup. By adding the bi-gram terms, the MAP is improved to 0.2180.
- In topic CS-T381, 什么是太阳风 (What is the solar wind?), we have two keywords 太阳 (sun), 风 (wind). Adding the bigram 太阳+风, the MAP is increased from 0.8252 to 0.8560, because the bigram is much less ambiguous than the two separate words.

However, bigrams do not always lead to positive results. As bigrams have much lower document frequencies and term frequencies than single words in the collection, even if we only give them a small weight, they can still dominate the overall document score in some cases, especially for longer topics. The following two questions show such examples.

- Topic CS-T385: 列出与北京大学百年校庆相关的大事 (List events related to the Centenary Celebration of Peking University). The baseline method includes terms 北京大学 (Peking University), 百年 (centenary), 校庆 (celebration), 大事 (event), and MAP is 0.7083. After adding bigrams 北京大学+百年, 百年+ 校庆, MAP drops to 0.6741. This is because some documents which only contain 百年+ 校庆 (Centenary Celebration) are then ranked higher than the documents containing 北京大学 and 校庆.
- Topic CS-T79: 举出桥本龙太郎辞职对日本经济的影响 (List the impact of Ryutaro Hashimoto's resignation on Japan's economy) The baseline include terms 桥本龙太郎 (Ryutaro Hashimoto), 辞职 (resignation), 日本 (Japan), 经济 (economy), 影响 (impact), and MAP is 0.5834. After adding bigrams 桥本龙太郎+辞职, 日本+经济, the MAP is decreased to 0.5130. In fact, documents talking about the resignation of Ryutaro Hashimoto do not necessarily contain the bigram 桥本龙太郎+辞职. The two words can well appear separately. Therefore, it would be necessary to determine the bigrams that are useful for enhancing retrieval, instead of using all of them.

Overall, for simplified Chinese collection, using bigrams leads to slightly improved MAP. However, we failed to obtain improvements for traditional Chinese collection.

Impact of the pseudo-relevant feedback (PRF)

Pseudo-relevant feedback usually produces better result for IR, especially for short queries. We empirically set number of feedback documents to 50 and number of terms to 80. We tried several PRF weights between 0.1 and 1, and the best result is obtained at around 0.5. Table 4 shows the performance we obtain with 0.5. We can see that PRF can greatly improve MAP and all the improvements are statistically significant.

Impact of passage filtering

As described in Section 3.4, we extracted 600 triples for biography question form Chinese Wikipedia. If the passage does not match any pattern we apply a penalty to its score instead of filtering it out. For example a passage with relevance score v will have a final contribution of $v*p$ to document score where $0<p<1$ is determined manually. In our experiments, we set it to 0.5 empirically.

For some topics, the results are improved; but for some other topics, they declined. The following are some improvement examples of applying patterns of biography:

- CS-T64 谁是比尔盖茨 (Who is Bill Gates): 0.4317 → 0.4867 (+13%)
- CT-T194 藤森是谁 (Who is Alberto Fujimori): 0.5630 → 0.5817 (+3%)

However, for some documents, they mainly describe the biography of other people but can also match the triple patterns. These documents are thus ranked higher, and this leads to a worse result. Some examples are shown below:

- CS-T380 莫扎特是谁 (Who is Mozart): 0.4693 → 0.425 (-9%)
- CS-T52 谁是杨振宁 (Who is Chen Ning Yang): 0.7994 → 0.7429 (-7%)

The overall result (Column Filter in Table 4) is only slightly different from baseline. So, we do not use this technique in our final integration (the last column of Table 4). To make filtering works more efficiently, we need to find a better way to define patterns and rules.

Final combination: Bio+Bi+PRF+PNE

Now we combine all useful techniques together, and exam the final performance (see Table 4). All weights are tuned in monolingual simplified Chinese QA. We use them in traditional Chinese and cross-lingual IR for QA. Our results show that performance is better than using any single technique alone. It outperforms largely the best official Chinese monolingual result. This result shows that by enhancing the baseline methods with different means, we can produce significantly higher performance in identifying potential documents containing answers to a question.

4.2 Cross-lingual results

Our baseline method uses Google online translation to translate questions directly. In Table 5, we show the best official result, our official submissions (with the bug), the corrected runs with Google translation and a combined run using both Google and Wikipedia.

As we can see, the corrected results with Google translation alone is still lower than the best submission. However, once Google translation is combined with the translation based on Wikipedia, the performance is largely increased, and in the case of EN-CS, it becomes closer to the best submission.

Now, let us examine different enhancements using filter, biography, bigram, and pseudo relevant feedback techniques.

Table 5: Comparing our question translation method to Google translation.

		Best-1 official	*RALI* official*	Google Trans.	Google+Wiki	
						% GT
EN-CS	AP	.5959	*.4033*	.5526	.5801	+5.0
	Q-measure	.6124	*.4191*	.5606	.5927	+5.7
	MSnDCG1K	.7947	*.6701*	.7304	.7768	+6.4*
EN-CT	AP	-	*.2723*	.3067	.3561	+16.1*
	Q-measure	-	*.2868*	.3258	.3769	+15.7*
	MSnDCG1K	-	*.4845*	.4914	.5512	+12.2*

Table 6: Applying different techniques

		G+W Baseline	Filter		Bio+Bi+PRF	
				% B		% B
EN-CS	AP	.5801	.5830	+0.5	.6214	+7.1**
	Q-measure	.5927	.5957	+0.5	.6355	+7.2**
	MSnDCG1K	.7768	.7798	+0.4	.8084	+4.1**
EN-CT	AP	.3561	.3605	+1.2*	.3883	+9.0**
	Q-measure	.3769	.3813	+1.2*	.4108	+9.0**
	MSnDCG1K	.5512	.5524	+0.2	.5945	+7.9**

Impact of Filtering and other Techniques to cross-lingual QA

Now, we set the new baseline to the Google+Wikipedia method (G+W). Table 6 shows the results of using passage filtering, bigram, biography keywords and PRF on top of the baseline method.

For passage filtering, we notice that the performance is much better than using it in mono-lingual QA. We obtain improvements for both for CS and CT, and some of them are statistically significant. The main reason is possibly that in cross-lingual QA, the retrieved passages are much noisier than in monolingual case. Therefore, filtering becomes necessary.

We can also see that these techniques together can significantly improve the baseline method. When the baseline method is enhanced by bigrams, biography keywords and PRF, the results are also better than the best submission.

We still observe some remaining problems. The translated questions usually are not very accurate. For example, in the question "*Tell me Guo Jingjing's achievements in divin*g", the word "*diving*" is translated as 潜水 (often used to mean scuba diving), while the correct translation is 跳水. The wrong choice of translation word greatly deteriorated the retrieval effectiveness of this question. So there is still room for further improvements in determining the correct translation words.

5 Conclusion

In this report, we described our experiments in IR4QA. To improve the effectiveness, we used a module for detecting personal names, using both an existing word segmenter and Wikipedia. To deal with biography questions, we extracted biography-related keywords, and this helped identify biography documents for biography questions. However, the desired documents are not all of the same biography-style documents as in Wikipedia and many relevant documents do not contain the frequent biography-related keywords. So, our method failed to improve much the baseline method. Only a slight change is observed.

The traditional pseudo-relevant feedback showed consistent improvements.

By combining all the enhancements, we obtained very good results and the improvements over the baseline methods are statistically significant for both mono-lingual and cross-lingual QA.

Due to a bug in our program, our official submission is not well ranked among all the participating runs. However, once the bug is corrected, we can obtain effectiveness higher than or close to the best submissions. The specific processing we added after the official submission can further improve the effectiveness.

This series of experiments showed that specific knowledge for each type of question is highly useful. In our future work, we will try to determine more such knowledge and for more types of question.

References

[1] Callan, J. Passage-level evidence in document retrieval. *SIGIR,* pp.302-310, 1994.

[2] Chan, Y., K. Chen, and W. Lu. Extracting and ranking Question-Focused terms using the titles of Wikipedia articles. *NTCIR-6*, pp.210-215, 2007.

[3] Ferrández, S., et al. Applying Wikipedia's multilingual knowledge to cross-lingual question answering. *NLD*, pp.352-363, 2007.

[4] Hacioglu, K. and W. Ward. Question classification with support vector machines and error correcting codes. *HLT-NAACL,* pp.28-30, 2003.

[5] Kata, B., et al. External knowledge sources for question answering. *TREC,* 2005.

[6] Kwok, K.L., P. Deng and N. Dinstl. NTCIR-6 monolingual Chinese and English-Chinese cross-lingual question answering experiments using PIRCS. *NTCIR-6*, pp.190-197, 2007.

[7] Llopis, F, J. Luis and A. Ferrández. Using a passage retrieval system to support question answering process. *ICCS,* pp.61-69, 2002.

[8] Magnini, B., et al. Overview of the CLEF 2006 Multilingual Question Answering Track. *CLEF*, pp.223-256, 2006.

[9] Mitamura, T., et al. JAVELIN III: Cross-lingual question answering from Japanese and Chinese documents. *NTCIR-6*, pp.202-209, 2007

[10] Mori, T. and K. Takahashi. A method of cross-lingual question-answering based on machine translation and noun phrase translation using Web documents. *NTCIR-6*, pp.182-189, 2007.

[11] Nie, J.-Y., Brisebois, M. and Ren, X. On Chinese text retrieval. Conference on Research and Development in Information Retrieval. *SIGIR*, pp.225-233, 1996.

[12] Nie, J.-Y., et al. On the Use of Words and N-grams for Chinese Information Retrieval. *IRAL,* pp.141-148, 2000.

[13] Sakai, T., Kando, N., Lin, C.-J., Mitamura, T., Ji, D., Chen, K.-H., Nyberg, E. Overview of the NTCIR-7 ACLIA IR4QA Task, *Proceedings of NTCIR-7*, to appear, 2008.

[14] Wu, Y., K. Tsai, and J. Yang. Two-pass named entity classification for cross language question answering. *NTCIR-6*, pp.168-174, 2007.

[15] Zhang, D. and W. Lee. Question classification using support vector machines. *SIGIR*, pp.26-32, 2003.

[16] Zhang, Y. and Vines, P. Using the web for automated translation extraction in cross-language information retrieval. *SIGIR '04*. pp. 162-169, 2004.