社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＦＩ－ 76 (19)
2004－ＮＬ－163 (19)
2004／9／17

High-Precision Search via Question Abstraction for Japanese Question Answering

Tetsuya SAKAI　Yoshimi SAITO　Tomoharu KOKUBU
Makoto KOYAMA　Toshihiko MANABE

Knowledge Media Laboratory, Toshiba Corporate R&D Center
tetsuya.sakai@toshiba.co.jp

Abstract

This paper explores the use of Question Abstraction, i.e., Named Entity Recognition for questions input by the user, for reranking retrieved documents to enhance retrieval precision for Japanese Question Answering (QA). Question Abstraction may help improve precision because (a) As named entities are often phrases, it may have effects that are similar to phrasal or proximity search; (b) As named entity recognition is context-sensitive, the named entity tags may help disambiguate ambiguous terms and phrases. Our experiments using several Japanese "exact answer" QA test collections show that this approach significantly improves IR precision, but that this improvement is not necessarily carried over to the overall QA performance. Additionally, we conduct preliminary experiments on the use of Question Abstraction for Pseudo-Relevance Feedback using Japanese *IR* test collections, and find positive (though not statistically significant) effects. Thus the Question Abstraction approach probably deserves further investigations.

1 Introduction

Information Retrieval (IR) strategies for effective Question Answering (QA) is beginning to receive attention [1, 3, 18, 26]. This is only natural, as most QA systems use IR as a component for selecting texts from which candidate answers are extracted, and therefore the IR performance *bounds* the overall QA performance. More specifically, retrieval *precision* is important for QA, as QA systems use the top ranked documents/passages only.

This paper explores the use of Question Abstraction, i.e., Named Entity Recognition for questions input by the user, for reranking retrieved documents to enhance retrieval precision for Japanese Question Answering (QA). Question Abstraction may help improve precision because:

1. As named entities are often phrases, it may have effects that are similar to phrasal or proximity search: For example, if "*Toshiba Kenkyu Kaihatsu Sentā*" (Toshiba R&D Center)" can be recognised as an ORGANIZATION, then it may be wise to treat it as a phrase rather than as four independent terms (e.g. [4]).

2. As named entity recognition is context-sensitive, the named entity tags may help disambiguate ambiguous terms and phrases: For example, the Japanese word "*kantoku*" could mean a manager (of a baseball team) or a director (of a film): If it co-occurs with sports-related words within a question or a document, it would be tagged with POSITION_ATHLETE, but if it co-occurs with a title of a film, it would be tagged with POSITION_CELEBRITY (provided that Named Entity Recognition is accurate).

Our experiments using several Japanese "exact answer" QA test collections show that this approach significantly improves IR precision, but that this improvement is not necessarily carried over to the overall QA performance. Additionally, we conduct preliminary experiments on the use of Question Abstraction for Pseudo-Relevance Feedback (PRF) [14, 16, 17, 20] using Japanese *IR* test collections, and find positive (though not statistically significant) effects. Thus the Question Abstraction approach probably deserves further investigations.

This paper focuses on *document* retrieval for QA rather than *passage* retrieval, as we have not found a passage extraction algorithm that is significantly more effective than using the whole document in terms of the overall QA performance [19].

The remainder of this paper is organised as follows. Section 2 describes relevant features of our Japanese QA system. Section 3 describes how retrieval precision may be improved based on Question Abstraction. Section 4 describes recently-proposed performance metrics which are applicable to both IR and QA with ranked lists of exact answers, which we use in our experiments along with traditional metrics. Section 5 describes our experiments using Japanese QA test collections and Section 6 describes our preliminary PRF experiments in traditional IR tasks. Finally, Section 7 concludes this paper.

2 ASKMi

The ASKMi Japanese QA system is fully described in [19], and its first participation at the NTCIR-4 QAC (Question Answering Challenge) Task is described in [21]. Here, we briefly describe ASKMi's features that are relevant to the present study.

ASKMi's Semantic Class Recogniser extracts named entities (i.e. candidate answers) from documents offline [9, 19]. It also extracts named entities from the user's question on-line: for example, if the question is "*Toshiba no shachō* (Toshiba's president)", it is transformed into "COMPANY *no* POSITION_MISC" through named entity recognition, where *no* is the Japanese possessive particle. We refer to this process as *Question Abstraction*. ASKMi performs rule-based Answer Type Recognition (or Question Classification) *after* Question Abstraction so that many questions can be covered with relatively small number of patterns. For the above example, Answer Type Recognition outputs person name answer types such as CELEBRITY and PERSON_MISC. Currently, ASKMi maintains more than 100 answer types.

After Question Analysis, which includes not only Answer Type Recognition but also other answer type sensitive features such as query expansion and document constraint generation [19], ASKMi's retriever retrieves a given number of documents that contain strings tagged with designated answer types, e.g. CELEBRITY or PERSON_MISC. (We use the top ten documents throughout this paper.) For a given question q and each document d, the original document score ($origdscore(q,d)$) is computed based on the Okapi/BM25 algorithm [25]. Optionally, we can perform a linear transformation of the original BM25-based document score for QA as follows:

$$dscore(q,d) = \max\{0, P_D*(origdscore(q,d)-1)+1\} \quad (1)$$

where $P_D(\geq 0)$ is a constant called the *document score parameter*. Clearly, $P_D = 0$ implies $dscore = 1$ (i.e. a constant), and $P_D = 1$ implies $dscore = origdscore$ (which is the default setting). $P_D > 1$ emphasises the document score curve and $0 < P_D < 1$ does the opposite.

For each answer candidate c in a retrieved document, its candidate score ($cscore(q,c)$) is computed mainly based on proximity with query terms [19, 21]. Then, the candidates are consolidated across documents to form an answer a, whose score is calculated as:

$$ascore(q,a) = \sum_{c,A(c)=a} dscore(q,D(c)) * cscore(q,c) \quad (2)$$

where $D(c)$ and $A(c)$ represent the mappings from a candidate c to the corresponding document and to the corresponding answer, respectively.

Finally, the Answer Formulator modifies the above answer score based on several heuristics, e.g. expected numerical ranges [7]. Answer string consolidation is performed where appropriate to minimise redundancy in the answer list [19, 22].

3 High-Precision Search via Question Abstraction

We now describe our new precision-oriented search algorithm based on ASKMi's Question Abstraction. Suppose that the question is "*Toshiba no shachō* (Toshiba's president)", and that "*Toshiba*" (as a COMPANY) and "*shachō*" (as a POSITION_MISC) were obtained through Question Abstraction. (ASKMi's Semantic Class Recogniser actually assigns a confidence value to each named entity [19], but here we simply ignore those with low confidence values.) Then, after retrieving a set of documents based on Okapi/BM25, we *rerank* the documents by modifying the document scores as follows:

$$origdscore_{new}(q,d) = (1 + P_A * \log_2(1 + n_A(q,d))) * origdscore_{old}(q,d) \quad (3)$$

where $n_A(q,d)$ is the number of "hits" based on Question Abstraction: For the example question given above, if the document contains both "*Toshiba*" (as a COMPANY) and "*shachō*" (as a POSITION_MISC) then $n_A(q,d) = 2$. If it contains only one of them, then $n_A(q,d) = 1$. Thus, documents are promoted according to how many named entities they share with the question. $P_A(\geq 0)$ is a parameter for controlling the impact of Question Abstraction.

Further, if the above approach does improve retrieval precision, then it may be useful to let $P_D > 1$ in Equation 1, i.e. to emphasise the document score curve through linear transformation, so that the document score component becomes dominant when calculating the final answer score using Equation 2.

4 Q-measure and R-measure

This section briefly describes *Q-measure* and *R-measure*, which are recently-proposed IR evaluation metrics based on multigrade relevance [20, 21, 22, 24]. In particular, Q-measure is also applicable to QA evaluation with ranked lists of exact answers, provided that the answer data contain *answer equivalence classes* and *answer correctness levels*. (In fact, existing standard Japanese QA test collections already have equivalence classes, and one only has to assign correctness levels to each answer string in order to enjoy the advantages of Q-measure: See Section 5.) In our QA experiments described in Section 5, we use Q-measure along with traditional Reciprocal Rank [27] because Q-measure can

properly handle multiple-answer questions as well as correctness levels. In our *IR* experiments described in Section 6, we use both Q-measure and R-measure along with traditional noninterpolated Average Precision.

Let X denote a relevance level, and let $gain(X)$ denote the *gain value* for successfully retrieving an X-relevant document. Further, let L denote the size of a given ranked output and let $X(r)$ denote the relevance level of the document at Rank r ($\leq L$). Then, the *gain at Rank* r is given by $g(r) = gain(X(r))$ if the document at Rank r is relevant, and $g(r) = 0$ if it is nonrelevant. The *cumulative gain at Rank* r is given by $cg(r) = g(r) + cg(r-1)$ for $r > 1$ and $cg(1) = g(1)$. In particular, let $cig(r)$ denote the cumulative gain at Rank r for an *ideal* ranked output (see later).

We now introduce the *bonused* gain at Rank r, simply given by $bg(r) = g(r) + 1$ if $g(r) > 0$ and $bg(r) = 0$ if $g(r) = 0$. Thus, the system receives an extra reward for finding a relevant document. Then, the *cumulative bonused gain at Rank* r is given by $cbg(r) = bg(r) + cbg(r-1)$ for $r > 1$ and $cbg(1) = bg(1)$. Q-measure and R-measure are defined as:

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \leq r \leq L, g(r) > 0} \frac{cbg(r)}{cig(r) + r}$$

$$= \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r) \frac{cbg(r)}{cig(r) + r} \quad (4)$$

$$R\text{-}measure = \frac{cbg(R)}{cig(R) + R} \quad (5)$$

where R is the total number of relevant documents and $isrel(r)$ is a binary flag such that $isrel(r) = 1$ if the document at Rank r is relevant and $isrel(r) = 0$ otherwise.

As have been proven in [22], Q-measure is equal to one iff a system output (s.t. $L \geq R$) is an ideal one, and R-measure is equal to one iff all the top R documents are (at least partially) relevant.

Using similar notations, traditional Average Precision (AP) and R-Precision can be expressed as:

$$AP = \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r) \frac{count(r)}{r} \quad (6)$$

$$R\text{-}Precision = \frac{count(R)}{R} \quad (7)$$

where $count(r)$ denotes the number of relevant documents within the top r documents.

By definition, $cbg(r) = cg(r) + count(r)$. Therefore, Q-measure and R-measure can alternatively be expressed as:

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r) \frac{cg(r) + count(r)}{cig(r) + r} \quad (8)$$

$$R\text{-}measure = \frac{cg(R) + count(R)}{cig(R) + R} \quad (9)$$

It can be observed that Q-measure and R-measure are "multigrade extensions" of Average Precision and R-Precision, respectively.

Q-measure is akin to Average Normalised (Discounted) Cumulative Gain (Average n(D)CG) [11] and Average Weighted Precision (AWP) [12]. However, AWP has been shown to be unreliable if relevant documents are found below Rank R, but Q-measure is free from this problem [20, 22]. Average n(D)CG is a *DCV* (Document Cutoff Value)-based metric [8] and requires additional parameters (number of documents to examine and the logarithm base for discounting). In contrast, Q-measure and R-measure are *recall*-based metrics just like Average Precision and R-Precision, and do not require these parameters.

When Q-measure is used for QA evaluation with ranked lists of answers, document relevance levels are replaced with answer correctness levels, and the answer equivalence classes are used to avoid rewarding systems that include duplicate answers in the ranked answer list. On the other hand, the use of R-measure for QA is not recommended because R (number of answer equivalence classes) is generally very small [21, 22].

The NTCIR *IR* test collections (See Section 6) have three relevance levels: S-relevant (highly relevant), A-relevant (relevant) and B-relevant (partially relevant). (Thus, an ideal ranked output for an NTCIR IR topic has all S-relevant documents at the top, then all A-relevant documents, and then all B-relevant documents). We let $gain(S) = 3$, $gain(A) = 2$, $gain(B) = 1$ throughout our IR experiments. Similarly, the QA test collections we use (See Section 5) have three correctness levels: S-correct, A-correct and B-correct. We use the same gain values as above throughout our QA experiments. Moreover, following the TREC/NTCIR traditions, we let $L' = 1000$ for the IR experiments and $L' = 5$ for the QA experiments, where $L' (\geq L)$ is the maximum ranked output size allowed.

5 QA Experiments

This section examines the effect of document reranking using Question Abstraction and/or Linear Transformation of the document score curve using Japanese QA test collections. Table 1 provides information on the QA test collections we used. The original QAC collections are from the QAC tracks of the NTCIR workshops [5, 6], while the TCQA collections were developed at Toshiba as described in [23].

For evaluating document retrieval performance in QA, we use noninterpolated Average Precision by regarding the supporting documents as relevant. For evaluating the final QA performance, we use Reciprocal Rank and Q-measure. As the original QAC collections already had explicit equivalence classes, we manually

assigned a correctness level to each of the QAC answer strings, as described in [21, 22]. As for the TCQA collections, equivalence classes and correctness levels were manually assigned at the same time. Table 2 shows some examples of answer equivalence classes (or *Answer Synsets*) and correctness levels for the QAC2 test collection, where the i-th answer synset is denoted by $AS(i)$: See [21] for more details.

Table 3 summarises the results of our experiments. The QAC1 Additional questions were used as our training data, and P_A (Equation 3) and P_D (Equation 2) were set to 0.6 and 2.5, respectively. Document reranking based on Question Abstraction is denoted by QAB, Linear Transformation of the document score curve is denoted by LT, and the combination of the two is denoted by QAB+LT. Our "baseline" does not use Question Abstraction and uses $P_D = 1$ (i.e. no linear transformation). **Boldface** values indicate higher average performance compared to the baseline, while "$*$" and "$**$" indicate statistically significant differences with the baseline in terms of the Sign Test ($\alpha = 0.05$ and $\alpha = 0.01$, respectively). Note that the IR performance of LT is the same as that of the baseline, because LT does not affect document ranking. For the same reason, the IR performance of QAB+LT is the same as that of QAB.

The results of our QA experiments can be summarised as follows:

- In terms of *IR* performance, the positive effect of reranking based on Question Abstraction is statistically significant for all test collections except (b). Thus Question Abstraction is probably effective for enhancing retrieval precision, although its improvement is rather small when averaged over the whole question set.

- In terms of *QA* performance, however, the effect of Question Abstraction is not clear: On average, QAB outperforms the baseline in (a), (b) and (e) but the differences are not statistically significant for (b) and (e).

- The effect of Linear Transformation alone on QA performance is not clear either: LT shows significant improvements over the baseline in (a), (b) and (d), but actually hurts performance for (c).

- The combination of Question Abstraction and Linear Transformation is significantly improves QA performance for (a) and (b) only. Moreover, QAB+LT is actually slightly less effective than LT in (b) and (d). Thus the combination is not altogether successful.

In short, although Question Abstraction does improve IR precision, its advantage is not necessarily reflected in the overall QA performance.

Given the above results, we formed the following hypothesis: Among the named entity tags (or *semantic classes*) obtained through Question Abstraction, some almost always improve IR performance, and others almost always hurt it.

To test the above hypothesis, we examined the results of the QAC1 Additional case more closely. For this test collection, Question Abstraction improved 98 questions and hurt 45 questions in terms of IR performance. We therefore investigated which named entity tags were actually extracted from these questions. More specifically, we automatically computed a *penalty score* for each named entity tag based on the aforementioned 45 questions: For example, if only one named entity tag was found in a particular question, we added one to its penalty score; If two named entity tags were found, we added 1/2 to the penalty score for each of the named entity tags, and so on.

According to the above analysis, the named entity tags with the highest penalty scores included EVENT, POSITION_CELEBRITY, POSITION_ATHLETE, POSITION_MISC among others. Using the EVENT tag appears to have hurt IR performance in some cases because of *synonymy*: For example, the question may contain "Nagano Olympic Games" (as an EVENT) but the supporting (i.e. relevant) documents may contain "Nagano Winter Olympic Games" or just "Winter Olympic Games" instead. Whereas, the POSITION-related tags appear to have hurt IR performance in some cases because of complete absence of position-related named entities within documents. For example, from a question of the form "... *no sakka wa dare desuka* (Who is the author of ...?)", Question Abstraction obtains "*sakka* (author/writer)" (as a POSITION_CELEBRITY), but the supporting documents may not contain "*sakka*" at all: Consider document contexts such as "Macbeth *by* William Shakespeare", "William Shakespeare *wrote* Macbeth".

Based on the above analysis, we have conducted additional experiments by *discarding* named entities such as those mentioned above when using Question Abstraction for document reranking. However, this was not successful. For example, discarding the POSITION-related tags for the QAC1 Additional collection gives 0.5486 in Average Precision, which is above the baseline but below the original QAB performance. In fact, further analyses suggested that the aforementioned hypothesis is probably untrue, as the same named entity tags appear to improve some questions while hurting others. For example, the EVENT tag mentioned above emerged as one of the most effective named entity tags when we analysed the 98 questions that were *improved* through Question Abstraction. Further studies are required in order to clarify *when* the document reranking based on Question Abstraction is worthwhile.

Table 1. Japanese QA Test Collections.

Name	#Questions∗	Document Type	#Documents
QAC1 Task 1 Additional	753	Mainichi newspaper	220,078
QAC1 Task 1 Formal	195	*ditto*	*ditto*
QAC2 Task 1 Formal	195	Mainichi/Yomiuri newspaper	593,636
TCQA-1 Bottom-Up	357	*ditto*	*ditto*
TCQA-1 Top-Down	268	*ditto*	*ditto*

∗Questions without answers/supporting documents excluded.

Table 2. Examples of QAC2 answer synsets and correctness levels (English translations).

QAC2-10001-01 ($R = 1$) Q: "Who is the seventh Japanese Major Leaguer?" $AS(1) = \{<$"Masato Yoshii",$S>$, <"Yoshii",$B>\}$
QAC2-10031-01 ($R = 1$) Q: "Where did Antonio Inoki's retirement match take place?" $AS(1) = \{<$"Tokyo Dome",$S>$, <"Bunkyo-ku, Tokyo",$A>\}$
QAC2-10049-01 ($R = 1$) Q: "How long did the Suharto Administration last?" $AS(1) = \{\ldots, <$"thirty-two years",$S>$, <"over thirty years",$A>$, <"thirty years",$B>\}$
QAC2-10074-01 ($R = 1$) Q: "Which Japanese person won the Nobel Peace Prize?" $AS(1) = \{<$"former prime minister Eisaku Sato", $S>$, <"Eisaku Sato", $S>$, <"Mr. Sato",$B>\}$
QAC2-10079-01 ($R = 1$) Q: "What is the abbreviation for Deoxyribonucleic Acid?" $AS(1) = \{<$"DNA", $S>$, <"DNA (Deoxyribonucleic Acid)", $B>\}$
QAC2-10124-01 ($R = 7$) Q: "What are the names of the satellites of Jupiter?" $AS(1) = \{<$"Amalthea", $S>, \ldots\}$ $AS(2) = \{<$"Adrastea", $S>\}$ $AS(3) = \{<$"Io", $S>$, <"Satellite Io", $B>\}$ $\vdots$
QAC2-10135-01 ($R = 2$) Q: "Who created the work of art "The Kiss"?" $AS(1) = \{<$"Auguste Rodin", $S>$, <"Rodin", $A>\}$ $AS(2) = \{<$"Gustav Klimt", $S>$, <"Klimt", $A>\}$
QAC2-10157-01 ($R = 10$) Q: "According to a Russian public opinion survey, who were the Top Ten Most Authoritative Politicians of the Century?" $AS(1) = \{<$"Lenin", $A>\}$ $AS(2) = \{<$"Former Prime Minister Stalin", $A>$, <"Stalin", $A>\}$ $\vdots$ $AS(7) = \{<$"Former President Gorbachev", $A>$, <"Gorbachev", $A>\}$ $\vdots$
QAC2-10177-01 ($R = 1$) Q: "What is the capital of Pakistan?" $AS(1) = \{<$"New Delhi, India", $S>$, <"New Delhi", $A>$, <"India", $A>\}$

Table 3. Reranking via Question Abstraction and/or Linear Transformation ($P_A = 0.6$, $P_D = 2.5$).

	IR	QA	
	Average Precision	Reciprocal Rank	Q-measure
(a) QAC1 Additional (training data: P_A and P_D were tuned for this collection.)			
baseline	.5425	.5088	.5460
QAB	**.5508**$**$	**.5204**$**$	**.5579**$**$
LT	.5425	**.5307**$**$	**.5637**$**$
QAB+LT	**.5508**$**$	**.5311**$**$	**.5669**$**$
(b) QAC1 Formal (test data)			
baseline	.5097	.6736	.6862
QAB	**.5183**	**.6796**	**.6886**
LT	.5097	**.6949**$**$	**.6989**$**$
QAB+LT	**.5183**	**.6932**$**$	**.6972**$*$
(c) QAC2 Formal (test data)			
baseline	.4433	.4540	.3955
QAB	**.4496**$**$	.4498	.3947
LT	.4433	.4326	.3872$*$
QAB+LT	**.4496**$**$	.4434	.3915
(d) TCQA-1 Bottom-Up (test data)			
baseline	.2409	.4184	.3672
QAB	**.2423**$*$	**.4190**	.3666
LT	.2409	**.4246**	**.3738**$*$
QAB+LT	**.2423**$*$	**.4225**	**.3693**
(e) TCQA-1 Top-Down (test data)			
baseline	.1173	.4014	.2796
QAB	**.1179**$*$	**.4081**$*$	**.2844**
LT	.1173	**.4114**	**.2847**
QAB+LT	**.1179**$*$	**.4210**$*$	**.2894**

Table 4. Japanese IR Test Collections.

Name	#Topics	Document Type	#Documents
NTCIR-3	42	Mainichi newspaper from 1998 and 1999	220,078
NTCIR-4	55	Mainichi/Yomiuri newspaper from 1998 and 1999	593,636

Table 5. PRF with initial search based on Question Abstraction ($P_A = 0.6$).

	Relaxed Average Precision	Rigid Average Precision	Q-measure	R-measure
(a) NTCIR-3 (training data: P_A was tuned for this collection.)				
baseline	.4558	.3779	.4813	.4558
QAB	**.4684**	**.3859**	**.4930**	**.4705**
(b) NTCIR-4 (test data)				
baseline	.4759	.3716	.4823	.4997
QAB	**.4850**	.3667	**.4911**	**.5049**

6 Preliminary IR experiments

The QA experiments described in the previous section suggest that Question Abstraction may be useful not only for the retrieval stage of QA but also for precision-oriented document retrieval in general. As a preliminary study for testing this hypothesis, we consider a specific IR problem: Pseudo-Relevance Feedback (PRF).

PRF is known to be effective for relatively broad search topics with many relevant documents. PRF consists of three steps: (1) Initial search; (2) Query expansion by assuming that the top ranked documents are relevant; (3) Final search. Hence the retrieval *precision* at the initial search stage is very important. We therefore conducted Japanese *IR* experiments using the Japanese test collections briefly described in Table 4, by *reranking* the initial ranked output using Equation 3 in order to obtain a new set of pseudo-relevant documents. The NTCIR-3 IR test collection [2] was used to tune P_A, and we let $P_A = 0.4$. Then the NTCIR-4 IR collection [13] was used as our test data. Only DESCRIPTION runs were considered.

The NTCIR IR test collections have *multigrade* relevance judgements. Traditionally, NTCIR uses both "Relaxed" Average Precision (treating S/A/B-relevant documents as relevant) and "Rigid" Average Precision (treating S/A-relevant documents as relevant) by using two separate "qrels" files with the trec_eval program. However, Relaxed AP ignores the relevance levels completely, while Rigid AP ignores the B-relevant documents in addition. We therefore use Q-measure and R-measure along with these traditional measures.

For our IR experiments, we used the BRIDJE retrieval system [16, 17, 20] as well as the Question Analyser component of the ASKMi QA system. The BRIDJE system also uses Okapi/BM25, and its default PRF algorithm is the same as that described in [14]. The term selection criterion we used is the traditional Offer Weight [14, 16]. Based on a preliminary experiment with the NTCIR-3 collection, we used 10 Pseudo-Relevant documents to add 40 expansion terms for each topic.

Table 5 summarises the results of our PRF experiments for the IR tasks. It can be observed that reranking the initial ranked output based on Question Abstraction does improve IR performance on average (except when Rigid Average Precision is used). However, its positive effect is not statistically significant, and our results should be regarded as preliminary.

7 Conclusions and Future Work

This paper explored the use of Question Abstraction for reranking retrieved documents to enhance retrieval precision for Japanese Question Answering. Our investigation using several Japanese QA test collections suggest that this approach improves IR performance. However, its impact on the overall QA performance is not clear, even when it is combined with Linear Transformation of the document score curve. Moreover, the hypothesis that we can separate named entity tags that should be used for document reranking from those that should not be appears to be untrue. More sophisticated approaches for enhancing our Question Abstraction approach would include examining the *sequence* of named entity tags found within a given question in order to select which named entities should be used for document reranking.

We also conducted preliminary experiments on the use of Question Abstraction for improved initial search for Pseudo-Relevance Feedback (PRF) in traditional IR tasks. The results are positive, though not statistically significant. In general, there are at least two strategies for improving PRF [15]:

Document Refinement Improving the precision of the pilot search;

Term Refinement Filtering out noisy expansion terms.

As Section 6 takes exactly the Document Refinement approach, the corresponding Term Refinement approach may also be worth investigating: That is, named entity recognition may be used as term selection filters. However, as this is beyond the scope of Question Abstraction, it will be pursued elsewhere.

References

[1] *ACM SIGIR 2004 Workshop: Information Retrieval for Question Answering*, 2004.

[2] Chen, K.-H. *et al.*: Overview of CLIR Task at the Third NTCIR Workshop, *NTCIR-3 Proceedings*, 2003.

[3] Clarke. C. L. A. and Terra, E. L.: Passage Retrieval vs. Document Retrieval for Factoid Question Answering, *ACM SIGIR 2003 Proceedings*, pp. 427–428, 2003.

[4] Chen, H.-H., Ding, Y.-W. and Tsai, S.-C.: Named Entity Extraction for Information Retrieval, *Computer Processing of Oriental Languages, ,* Vol. 12, No. 1, pp. 75-85, 1998.

[5] Fukumoto, J., Kato, T. and Masui, F.: Question Answering Challenge (QAC-1): An Evalution of Question Answering Tasks at the NTCIR Workshop 3, *AAAI Spring Symposium: New Directions in Question Answering*, pp. 122-133, 2003.

[6] Fukumoto, J., Kato, T. and Masui, F.: Question Answering Challenge for Five Ranked Answers and List Answers – Overview of NTCIR4 QAC2 Subtask 1 and 2 –. *NTCIR-4 Working Notes*, 283-290.

[7] Hovy, E. *et al.*: Using Knowledge to Facilitate Factoid Answer Pinpointing, *COLING 2002 Proceedings*, 2002.

[8] Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments, *ACM SIGIR '93 Proceedings*, pp. 329–338, 1993.

[9] Ichimura, Y. *et al.*: A Study of the Relations among Question Answering, Japanese Named Entity Extraction, and Named Entity Taxonomy (in Japanese), *IPSJ SIG Notes*, NL–161–3, 2004.

[10] Järvelin, K. and Kekäläinen, J.: IR Evaluation Methods for Retrieving Highly Relevant Documents, *ACM SIGIR 2000 Proceedings*, pp. 41-48, 2000.

[11] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM Transactions on Information Systems*, Vol. 20, No. 4, pp. 422-446, 2002.

[12] Kando, N., Kuriyama, K. and Yoshioka, M.: Information Retrieval System Evaluation using Multi-Grade Relevance Judgments - Discussion on Averageable Single-Numbered Measures (in Japanese), *IPSJ SIG Notes*, FI–63–12, pp. 105-112, 2001.

[13] Kishida, K. *et al.*: Overview of CLIR task at the fourth NTCIR workshop, *NTCIR-4 Working Notes*, pp. 1-59, 2004.

[14] Sakai, T.: Japanese-English Cross-Language Information Retrieval using Machine Translation and Pseudo-Relevance Feedback, *International Journal of Computer Processing of Oriental Languages*, Vol. 14, No. 2, pp.83-107, 2001.

[15] Sakai, T. and Sparck Jones, K.: Generic Summaries for Indexing in Information Retrieval, *ACM SIGIR 2001 Proceedings*, pp. 190-198, 2001.

[16] Sakai, T., Koyama, M., Suzuki, M. and Manabe, T.: Toshiba KIDS at NTCIR-3: Japanese and English-Japanese IR, NTCIR-3 Proceedings, 2003.
`http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-CLIR-SakaiT`

[17] Sakai, T. *et al.*: BRIDJE over a Language Barrier: Cross-Language Information Access by Integrating Translation and Retrieval, *IRAL 2003 Proceedings*, pp.65-76, 2003. `http://acl.ldc.upenn.edu/W/W03/W03-1109.pdf`

[18] Sakai, T. and Kokubu, T.: Evaluating Retrieval Performance for Japanese Question Answering: What Are Best Passages? *ACM SIGIR 2003 Proceedings*, pp. 429-430, 2003.

[19] Sakai, T. *et al.*: ASKMi: A Japanese Question Answering System based on Semantic Role Analysis, *RIAO 2004 Proceedings*, pp. 215-231, 2004.

[20] Sakai, T. *et al.*: Toshiba BRIDJE at NTCIR-4 CLIR: Monolingual/Bilingual IR and Flexible Feedback, *NTCIR-4 CLIR Proceedings*, to appear, 2004.

[21] Sakai, T. *et al.*: Toshiba ASKMi at NTCIR-4 QAC2, *NTCIR-4 Proceedings*, to appear, 2004.

[22] Sakai, T.: New Performance Metrics based on Multigrade Relevance: Their Application to Question Answering, *NTCIR-4 Proceedings*, to appear, 2004.

[23] Sakai, T. *et al.*: The Effect of Back-Formulating Questions in Question Answering Evaluation, *ACM SIGIR 2004 Proceedings*, pp. 474-475, 2004.

[24] Sakai, T.: Ranking the NTCIR Systems based on Multigrade Relevance, *AIRS 2004 Proceedings*, to appear, 2004.

[25] Sparck Jones, K., Walker, S. and Robertson, S. E.: A Probabilistic Model of Information Retrieval: Development and Comparative Experiments, *Information Processing and Management* 36, pp. 779-808 (Part I) and pp. 809-840 (Part II), 2000.

[26] Tellex, S. *et al.*: Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering, *ACM SIGIR 2003 Proceedings*, pp. 41–47, 2003.

[27] Voorhees, E. M.: Building A Question Answering Test Collection, *ACM SIGIR 2000 Proceedings*, pp. 200–207, 2000.