Evaluating Retrieval Performance for Japanese Question Answering: What Are Best Passages?

Tetsuya Sakai and Tomoharu Kokubu
Knowledge Media Laboratory, Toshiba Corporate R&D Center, Japan

Categories and Subject Descriptors

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

General Terms

Experimentation

Keywords

Question Answering, Passage Retrieval

1. INTRODUCTION

Question Answering (QA) has recently received attention from the information retrieval, information extraction, machine learning and natural language processing communities. While traditional Information Retrieval (IR) systems return a list of documents, recent QA systems are tackling the problem of returning short, exact answers in response to open-domain, fact-based questions. TREC started the English QA track at TREC-8 (though systems were to return text snippets instead of exact answers up to TREC-10), and NTCIR started the Japanese QA track at NTCIR-3 [5].

The popular approach to QA is the combination of passage retrieval and information extraction. Passage retrieval is used for selecting texts that match the terms extracted from the input question, and information extraction is used for extracting candidate answers from the texts. An important question here is how to define a passage: Long passages (e.g. whole documents) may introduce much noise at the answer selection stage, whereas using short passages (e.g. a few sentences) may imply failure to retrieve texts that contain good answers. How a passage should be defined depends primarily on how the search terms extracted from the question are distributed over each document.

At NTCIR-3 QAC1 (Question Answering Challenge 1), a collection of Japanese newspaper articles was used as the knowledge source. Many participants treated each *paragraph* as a passage, as paragraph boundaries were explicitly given in the newspaper CD-ROM data. This paper questions this popular approach by automatically generating a document retrieval test collection from the QAC1 Question Answering test collection and comparing retrieval performances of five different passage types.

2. EXPERIMENTS

The QAC1 Japanese Question Answering Test Collection includes 195 official Task 1 ("Main Task") questions, using Mainichi Newspaper articles from 1998 and 1999 (236,664 documents) as the knowledge source. The Task 1 Answer File contains, for each question, answer strings with IDs of documents that support the answers (though supporting documents were not evaluated at NTCIR-3 QAC1) [5]. For retrieval performance evaluation, we converted the 1998 portion of the QAC1 collection into a document retrieval test collection by regarding the Task 1 questions as search requests and the supporting documents as "relevant" ones. Requests with at least three "relevant" documents from 1998 were selected: 62 requests and 377 "relevant" documents were thus obtained. We call this test collection *QAC1-D98*.

We examined five passage types:

Doc Documents. That is, each newspaper article was regarded as a passage.

Par-H Paragraphs *without* the Headlines of the original newspaper articles. This was the popular approach at NTCIR-3 QAC1. Approximately five paragraphs were extracted from each article.

Par To each paragraph, the original article headline was concatenated.

Tile Our own adaptation of Hearst's TextTiling [2] was used to break up the original articles into topical passages. We used sentences rather than word windows for *tokenization*, and 100-word blocks for cosine-based *similarity determination*. *Boundary identification* was performed as in [2]. Approximately two passages were extracted from each article, as newspaper articles are seldom multitopic. As with **Par**, each passage includes the article headline.

ETile Each Tile, obtained as described above, was Expanded "upwards and downwards" as follows: (i) To the tile, add the sentence immedietely above it, or the one immediately below; (ii) Repeat (i) until the tile length exceeds one-third of the average document length. Thus, this avoids extremely short passages, and produces *overlapping* ones.

Thus, **Par-H** is the only passage type that discarded the original article headlines.

For retrieval, we used the BRIDJE system that emloys Okapi/BM25 term weighting [6]. Pseudo-relevance feedback was not used. Using the QAC1-D98 document retrieval test collection, passage types were evaluated as follows:

1. Produce a ranked list of passages, where each Passage ID consists of the original Document ID plus a suffix.

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007.

2. For each retrieved passage *p* generated from a "relevant" document, *remove p* if it does *not* contain a correct answer string.

3. Remove all Passage ID suffixes from the list to produce a list of Document IDs.

4. Remove duplicate Document IDs from the above and evaluate this list using trec_eval.

Table 1 shows, for each passage type, Average Precisions (AveP) and Precisions at document cutoffs 30/10 (PDoc30/PDoc10) averaged across the 62 questions. AveP is a more stable measure than PDoc, but PDoc is important for QA as only the top ranked passages are used for answer selection. In addition, using **Doc** as the baseline, per-question comparisons and statistical significance information are provided: For example, **ETile** outperforms **Doc** for 38 questions while the reverse is true for only 22 questions in terms of AveP, which is a significant difference using the Sign Test (α = 0.05). Boldface values emphasize superiority over **Doc**. Table 2 compares passage types other than **Doc**: For example, **ETile** outperforms **Par-H** for 49 questions while the reverse is true for only 12 questions in terms of AveP, which is a significant difference (α = 0.01). Our significance test results show that:

(a) **ETile** significantly outperforms **Tile** (AveP), **Doc** (AveP), **Par** (AveP/PDoc10) and **Par-H** (all measures).

(b) **Tile** significantly outperforms **Par** (AveP/PDoc10) and **Par-H** (all measures).

(c) **Doc** significantly outperforms **Par-H** (all measures).

(d) **Par** significantly outperforms **Par-H** (all measures).

Thus, our findings can be summarised as follows:

1. **ETile** is the best choice among our five passage types for QA. We have shown that it may be more effective than **Doc** in terms of *retrieval performance*, and it is probably superior to **Doc** for *answer selection* in QA, as an **ETile** passage is typically half as long as an entire document. Moreover, from (a), overlapping passages are better than tiles.

2. Although **Par-H** was widely used at NTCIR-3 QAC1, this is actually the worst choice. It is clear that, when breaking up a newspaper article into passages, the article headline should be added to each passage.

A QAC1 participant [3] have reported some post-submission results that are related to our Finding 2. In their experiments, document retrieval appeared to slightly outperform paragraph retrieval both in terms of retrieval performance *and* Mean Reciprocal Rank in QA, although the differences were probably not statistically significant. Interestingly, they too used paragraphs without headlines for their official submissions.

3. CONCLUDING REMARKS

The experiments described in this paper were limited to the use of *static* passages that can be generated prior to question input. However, some TREC participants argue that *dynamic* or question-specific passage selection is more suitable for QA (e.g. [1]). As for Japanese QA, the top system at NTCIR-3 QAC1 used dynamically selected three consecutive sentences as passages [4], based on their TREC QA experience. As dynamic passages are inherently overlapping, our Finding 1 does not contradict with their claims. We are currently developing our own Japanese question answering system, and would like to clarify what are best passages for Japanese QA and the relationship between retrieval and question answering performances in the near future.

Table 1: QAC1-D98 retrieval performances.

	Doc	**Par-H**	**Par**	**Tile**	**ETile**
AveP	0.3643	0.2835	0.3465	**0.3762**	**0.4071**
	-	43/18**	38/23	**27/33**	**22/38***
PDoc30	0.1333	0.1177	0.1269	**0.1366**	**0.1419**
	-	23/8**	18/17	**13/15**	**11/18**
PDoc10	0.2597	0.2097	0.2419	**0.2774**	**0.2871**
	-	26/12*	21/16	**11/22**	**11/22**

Table 2: QAC1-D98 per-question comparisons.

		Par	**Tile**	**ETile**
Par-H vs	AveP	21/40*	14/47**	12/49**
	PDoc30	6/22**	6/24**	58/28**
	PDoc10	7/23**	5/36**	5/35**
Par vs	AveP	-	22/39*	19/42**
	PDoc30	-	10/16	7/16
	PDoc10	-	8/24**	6/23**
Tile vs	AveP	-	-	20/38*
	PDoc30	-	-	3/10
	PDoc10	-	-	8/12

4. REFERENCES

[1] Clarke, C. L. A. *et al.*: Exploiting Redundancy in Question Answering, *ACM SIGIR 2001 Proceedings*, pp. 358–365 (2001).

[2] Hearst, M. A.: Multi-Paragraph Segmentation of Expository Text, *ACL '94 Proceedings*, pp. 9–16 (1994).

[3] Nomoto, M. *et al.*: NTCIR-3 QAC Experiments at Matsushita, *NTCIR-3 Working Notes*, Part IV, pp. 55–62 (2002).

[4] Lee, S. and Lee, G. G.: SiteQ/J: A Question Answering System for Japanese, *NTCIR-3 Working Notes*, Part IV, pp. 31–38 (2002).

[5] NTCIR-3 Working Notes, Part IV: Question-Answering Challenge (QAC1) (2002).

[6] Sakai, T. *et al.*: Toshiba KIDS at NTCIR-3, *NTCIR-3 Proceedings* (2003).
http://research.nii.ac.jp/ntcir/workshop/
OnlineProceedings3/NTCIR3-CLIR-SakaiT