Test Collections for Spoken Document Retrieval from Lecture Audio Data

Tomoyosi Akiba[1], Kiyoaki Aikawa[2], Yoshiaki Itoh[3], Tatsuya Kawahara[4], Hiroaki Nanjo[5], Hiromitsu Nishizaki[6], Norihito Yasuda[7], Yoichi Yamashita[8], Katunobu Itou[9]

(1)Toyohashi Univ. of Technology, 1-1 Hibarigaoka, Tenpaku, Toyohashi, Aichi, JAPAN (akiba@ics.tut.ac.jp)
(2)Tokyo Univ. of Technology, (3)Iwate Prefectural Univ.,
(4)Kyoto Univ., (5)Ryukoku Univ., (6)Univ. of Yamanashi, (7)NTT, (8)Ritsumeikan Univ., (9)Hosei Univ.

Abstract

The Spoken Document Processing Working Group, which is part of the special interest group of spoken language processing of the Information Processing Society of Japan, is developing a test collection for evaluation of spoken document retrieval systems. A prototype of the test collection consists of a set of textual queries, relevant segment lists, and transcriptions by an automatic speech recognition system, allowing retrieval from the Corpus of Spontaneous Japanese (CSJ). From about 100 initial queries, application of the criteria that a query should have more than five relevant segments that consist of about one minute speech segments yielded 39 queries. Targeting the test collection, an ad hoc retrieval experiment was also conducted to assess the baseline retrieval performance by applying a standard method for spoken document retrieval.

1. Introduction

The lecture is one of the most valuable genres of audiovisual data. Previously, however, lectures have mostly been archived in the form of books or related papers. The main reason is that spoken lectures are difficult to reuse because browsing and efficient searching within spoken lectures is difficult.

Spoken document processing is a promising technology for solving these problems. Spoken document processing deals with speech data, using techniques similar to text processing. This includes transcription, translation, search, alignment to parallel materials such as slides, textbooks, and related papers, structuring, summarizing, and editing. As this technology is developed, there will be advanced applications such as computer-aided remote lecture systems and self-learning systems with efficient searching and browsing. However, spoken document processing methods are difficult to evaluate because they require subjective judgment and/or the checking of large quantities of evaluation data. In certain situations, a test collection can be used for a shareable standard of evaluation.

To date, test collections for information retrieval research have been constructed from such sources as newspaper articles (Kitani et al., 1998), Web documents (Oyama et al., 2005), and patent documents (Fujii et al., 2005). Test collections for cross-language retrieval (Gey and Oard, 2001; Kishida et al., 2005), open-domain question answering (Voorhees and Tice, 1999; Kato et al., 2005), and text summarization (Hirao et al., 2004) have also been constructed.

A test collection for Spoken Document Retrieval (SDR) is usually based on a broadcast news corpus. Compared to broadcast news, lectures are more challenging for speech recognition because the vocabulary can be technical and specialized, the speaking style can be more spontaneous, and there is a wider variety of speaking styles and structure types for lectures. Moreover, a definition of the semantic units in lectures is ambiguous because it is highly dependent on the queries. We aim to construct a test collection for ad hoc retrieval and term detection.

The rest of this paper is organized as follows. Section 2. describes how we constructed the test collection for spoken document retrieval, targeting lecture audio data. In Section 3., we evaluate the test collection by investigating its baseline retrieval performance, which was obtained by applying the conventional document retrieval method.

2. Constructing a Test Collection for SDR

A test collection for text document retrieval comprises three elements: (1) a huge document collection in a target domain, (2) a set of queries, and (3) results of relevance judgments, i.e., sets of relevant documents that are selected from the collection for each query in the query set.

In the spoken document case, the text collection should not merely be replaced with a spoken document collection. Two additional elements are necessary for an SDR test collection: (4) manual transcriptions, and (5) automatic transcriptions of the spoken document collection. The manual transcriptions are necessary for relevance judgment by the test collection constructors and can be used as a "gold standard" for automatic transcription by test collection users. The automatic transcriptions obtained by using a Large Vocabulary Continuous Speech Recognition (LVCSR) system are also desirable for supporting those researchers who do not have their own facilities for speech recognition or are interested only in aspects of text processing in SDR.

These elements of our SDR test collection are described in the following subsections.

2.1. Target Document Collection

We chose the Corpus of Spontaneous Japanese (CSJ) (Maekawa et al., 2000) as the target collection. It includes several kinds of spontaneous speech data, such as lecture speech and spoken monologues, together with their manual transcriptions. From among them, we selected two kinds of lecture speech: lectures at academic societies, and simulated lectures on a given subject. The collection contains 2702 lectures and more than 600 hours of speech. Table 1 summarizes the collection. Because its size is comparable

Table 1: Summary of the target document collection from CSJ.

	Speakers	Lectures	Data size (hours)
Academic lectures	838	1007	299.5
Simulated lectures	580	1699	324.1

to the Text Retrieval Conference (TREC) SDR test collection (Garofolo et al., 1999), it is sufficient for the purposes of retrieval research.

2.2. Queries

Queries, or information needs, for spoken lectures can be categorized into two types: those searching for a whole lecture and those asking for information described in part of a lecture. We focus on the latter type of query in our test collection, because this would seem much more likely than the former in terms of the practical use of lecture search applications. For such a query, the length of the relevant segment will vary, so a document, in Information Retrieval (IR) terms, must be a segment with variable length. In this paper, we refer to such a segment as a "passage".

Another reason why we focused on partial lectures arises from technical issues about constructing a test collection for retrieval research. If we regard each lecture in the collection as a document, the corresponding ad hoc task is defined as searching for relevant documents from among the 2702 documents. This number is much fewer than that for the TREC SDR task, which has 21,754 documents (stories) in the target collection.

Therefore, we constructed queries that ask for passages of varying lengths from lectures. We tried to control the length to about five utterances on average. Because a query tends to ask for something specific, which can be described in such a passage, the query is less like a query in document retrieval, but more like a question submitted to a question answering system. In addition to the guidelines, we constructed about 100 queries in total.

2.3. Relevance Judgment

Relevance judgment for the queries was conducted manually and performed against every variable length segment (or passage) in the target collection. One of the difficulties related to relevance judgment comes from the treatment of the supporting information. We regarded a passage as irrelevant to a given query even if it was a correct answer in itself to the query, when it had no supporting information that would convince the user who submits the query of the correctness of the answer. For example, for the query "How can we evaluate the performance of information retrieval?", the answer "11-point average precision" is not sufficient, because it does not say by itself that it is really an evaluation measure for information retrieval. The relevant passage must also include supporting information indicating that "11-point average precision" is one of the evaluation metrics used for information retrieval.

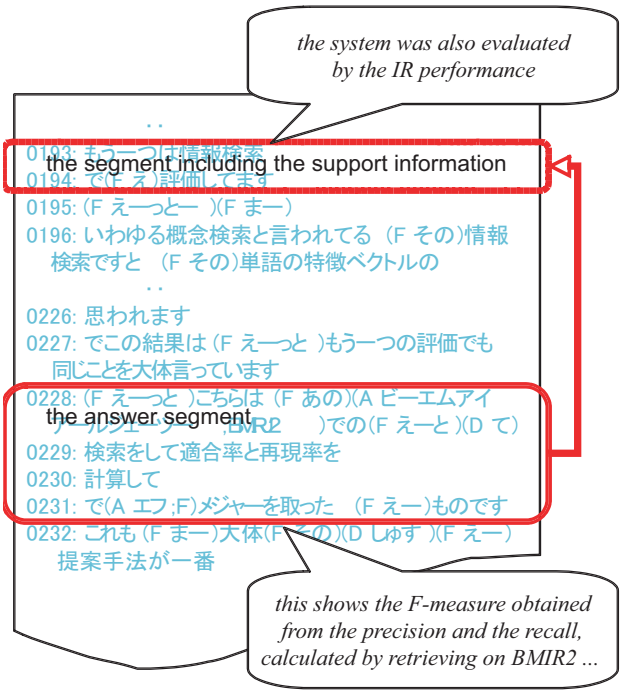

Figure 1: An example of the answer and the supporting segment.

The supporting information does not always appear together with the relevant passage, but may appear somewhere else in the same lecture. Therefore, we regarded a passage as relevant to a given query if it had supporting information in some segment of the same lecture. If a passage in a lecture was judged relevant, the range of the passage and the ranges of the supporting segments, if any, along with the lecture ID, were recorded in our "golden" file.

The relevance judgment was performed by the constructor of each query. The assessor selected the candidate passages from the target document collection by using the document search engine specifically prepared for the work, and labeled them into three classes according to the degree of their relevancy: "Relevant", "Partially relevant", and "Irrelevant".

Finally, after we filtered out the queries that had no more than four relevant passages in the target collection, 39 queries were selected for our test collection. Table 2 shows some statistics of the result.

2.4. Automatic Transcription

A Japanese LVCSR decoder (Lee et al., 2001) was used to obtain automatic transcriptions of the target spoken documents. Because the target spoken documents of lecture speech are more spontaneous than those of broadcast news, the speech recognition accuracy was expected to be worse than for TREC SDR. To achive better recognition results, both the acoustic model and the language model were trained by using the CSJ itself (Kawahara et al., 2003). Figure 2 shows the two distributions of the word error rates of the 70 academic lectures, obtained by using the closed and open settings. They differ in their average, but have almost the same shape, which ranges between about 0.65 and 0.95.

Table 2: Statistics for the results of the relevance judgment.

Label	Passages per query	Unique lectures per query	Utterances per passage
Relevant	11.18	7.90	10.39
Relevant & Partially Relevant	12.69	9.26	10.88

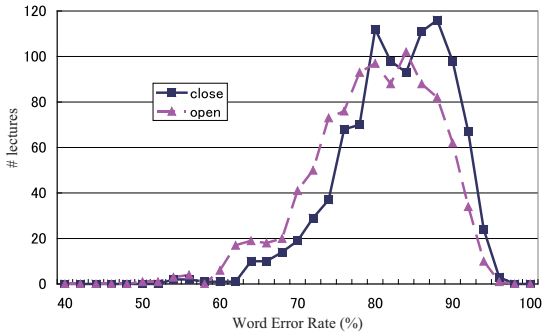

Figure 2: Distribution of word error rates in 70 academic lectures.

Table 4: Statistics of the redefined task.

Utterances per passage	15	30	60	Lecture
Target documents	60,202	30,762	16,060	2,702
Average relevant documents (**R**)	16.36	12.77	10.90	8.13
Average relevant documents (**R+P**)	19.03	14.79	12.54	9.44

For the first attempt, we decided to use the recognition results in a closed setting. The Word Error Rate (WER) was about 20%, which is comparable with that of the TREC SDR task.

2.5. Summary of the Test Collection

Table 3 shows a summary of the constructed test collection compared with the TREC-9 SDR test collection.

3. Evaluation

To evaluate the test collection and to assess the baseline retrieval performance obtained by applying a standard method for SDR, an ad hoc retrieval experiment targeting the test collection was conducted.

3.1. Task Definition

The primary task of our test collection, i.e., to find passages with variable utterance length, is not conventional. Because we wanted to evaluate the performance obtained by applying the standard method for SDR, and to compare the results with other studies in SDR and IR research, we redefined the conventional retrieval task, instead of searching for variable length segments in the collection.

Firstly, we defined pseudopassages by automatically segmenting each lecture into sequences of segments with fixed

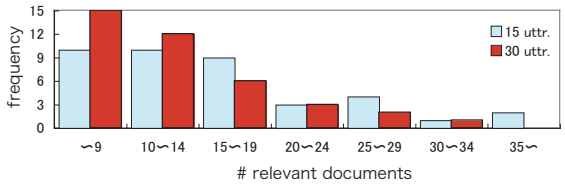

Figure 3: The distribution of the relevant documents.

numbers of sequential utterances: 15, 30, and 60. When 30 utterances are used in a segment, the number of pseudopassages is 30,762 and the number of words in a document is 204.2 on average, which are comparable numbers to those for TREC SDR.

Next, we assigned retrieved pseudopassages a relevance label as follows: if the pseudopassage shared at least one utterance that came from the relevant passage specified in the "golden file", then the pseudopassage was labeled as "relevant". Two kinds of relevance degree were used for the evaluation as follows.

R The passages labeled "Relevant" are used for deciding the relevant pseudopassages.

R+P The passages labeled either "Relevant" or "Partially relevant" are used for deciding the relevant pseudopassages.

Table 4 shows the size of the target documents (the number of pseudopassages) and the number of relevant documents for each task. Figure 3 shows the distribution of the relevant documents found in our redefined ad hoc retrieval task.

3.2. Ad hoc Retrieval Methods

All pseudopassages were then indexed by using either their words, their character bi-grams, or a combination of the two. The vector space model was used as the retrieval model and TF–IDF (Term Frequency–Inverse Document Frequency) with pivoted normalization (Singhal et al., 1996) was used for term weighting. We compared three representations of the pseudopassages: the 1-best automatically transcribed text, the union of the 10-best automatically transcribed texts, and the reference manually transcribed text.

3.3. Evaluation Metric

We used 11-point average precision (Teufel, 2007) as our evaluation metric, which is obtained by averaging the following AP over the queries.

$$IP(x) = \max_{x \leq R_i} P_i$$

Table 3: A comparison between TREC-9 SDR and our CSJ SDR test collections.

	TREC9 SDR	CSJ
Target documents	Broadcast news	Lecture speech
Quantity	557 hours	623.6 hours
Documents	21,754	2702 (30,762 seg. *)
Words per document	169	2324.9 (204.2 per seg. *)
Queries	50	39
Transcription	Low grade (WER 10.3%)	High grade
WER	26.7%	21.4%

* A succession of 30 utterances is considered to be a segment.

Table 5: 11-points average precisions using 15 utterances as a pseudopassage.

		Indexing unit		
Relevance degree	Transcription	Word	Char. 2-gram	Word + char. 2-gram
R	Reference	0.180	0.165	0.185
	10-best	0.177	0.145	0.167
	1-best	0.155	0.135	0.146
R+P	Reference	0.181	0.166	0.188
	10-best	0.179	0.150	0.171
	1-best	0.159	0.143	0.152

Table 6: 11-points average precisions using 30 utterances as a pseudopassage.

		Indexing unit		
Relevance degree	Transcription	Word	Char. 2-gram	Word + Char. 2-gram
R	Reference	0.249	0.216	0.240
	10-best	0.225	0.205	0.232
	1-best	0.213	0.188	0.207
R+P	Reference	0.249	0.220	0.242
	10-best	0.227	0.210	0.234
	1-best	0.211	0.194	0.211

$$AP = \frac{1}{11}\sum_{i=0}^{10} IP(\frac{i}{10}),$$

where R_i and P_i are the recall and the precision up to the i-th retrieved documents, respectively. In practice, we retrieved 1000 documents for each query to calculate the AP.

3.4. Results

Figure 4 shows the 11-point average precision for each query, where 30 utterances were used as a pseudo-passage and the reference transcriptions were used for indexing. It

Table 7: 11-points average precisions using 60 utterances as a pseudopassage.

		Indexing unit		
Relevance degree	Transcription	Word	Char. 2-gram	Word + Char. 2-gram
R	Reference	0.294	0.269	0.297
	10-best	0.256	0.236	0.265
	1-best	0.251	0.227	0.253
R+P	Reference	0.305	0.278	0.308
	10-best	0.261	0.243	0.271
	1-best	0.256	0.235	0.263

Table 8: 11-points average precisions using th whole lecture as a pseudopassage.

		Indexing unit		
Relevance degree	Transcription	Word	Char. 2-gram	Word + Char. 2-gram
R	Reference	0.453	0.443	0.468
	10-best	0.399	0.384	0.414
	1-best	0.411	0.397	0.426
R+P	Reference	0.473	0.454	0.489
	10-best	0.413	0.400	0.428
	1-best	0.423	0.409	0.441

indicates that the variance of the difficulty is high. For example, the hardest query can find only one (**R** degree) relevant passage in the 100-best candidates. On the other hand, the easiest query can find eight (**R** degree) relevant passages in the 10-best candidates.

Table 5, 6, 7, and 8 show the all evaluation results obtained by combining the four kinds of passage length (15, 30, 60 utterances, or a whole lecture), two kinds of relevance degree (**R** or **R+P**), three kinds of transcription (reference, 1-best or 10-best recognition candidates), and three kinds of indexing unit (word, character 2-gram, or a combination

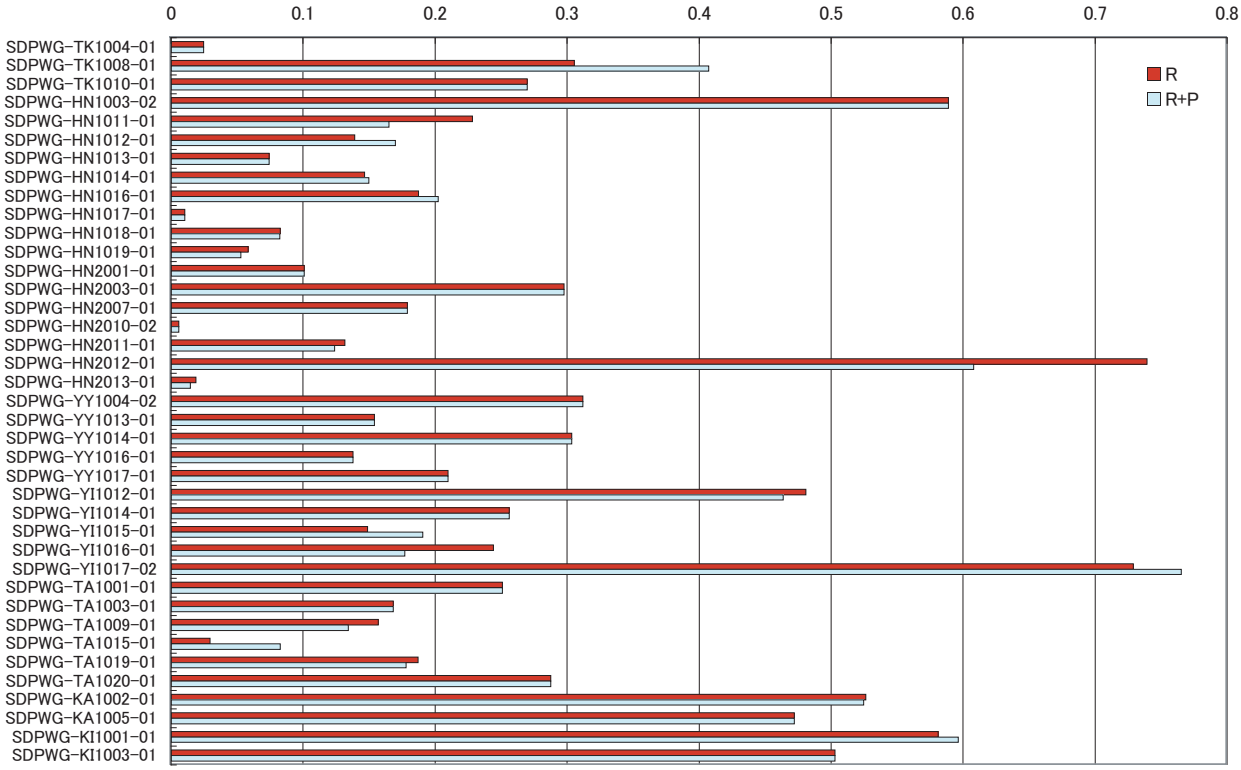

Figure 4: 11-point average precision for each query (using 30 utterances as a document, and manual transcription for the indexing.)

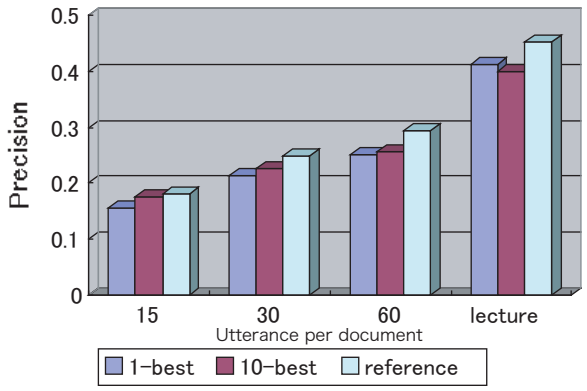

Figure 5: 11-point average precision using 1-best, 10-best, and reference transcriptions for indexing documents.

of the two).

Comparing the indexing units, using words is more effective than using character 2-grams. Using both words and character 2-grams slightly improves the retrieval performance, especially for the longer target document lengths, i.e., using 60 utterances or a whole lecture as a document. Comparing the two kinds of relevant degree, **R+P** consistently gives better results than **R**, but the difference is not large.

Figure 5 summarizes the results using the word as indexing unit and R degree for the relevancy, to compare the three kinds of representation of the target documents. It shows that using the 1-best automatically transcribed text decreases the IR performance by 10% to 15% compared with using the reference transcription. We also found that the use of 10-best candidates was effective for tasks with shorter passages, namely 15 and 30 utterances, but is less effective for those with longer passages, namely 60 utterances and whole lectures.

As a whole, the evaluation results show that the ad hoc retrieval task for lecture audio data is much more difficult than that for broadcast news, where the precision was reported to be around 0.45 for a task condition comparable to our 30-utterances condition. Except when the whole lecture is used as a passage, the retrieval performance is very low. This is partly because a relevant passage often has its supporting segments separated from it in the same document, meaning that the relevant passage does not always have self-contained information.

4. Conclusion and Future Work

A test collection for spoken lecture ad hoc retrieval was constructed. We chose the Corpus of Spontaneous Japanese (CSJ) as the target collection and constructed 39 queries designed to ask for information described in a partial lecture rather than a whole lecture. Relevance judgments for these queries were conducted manually and performed against every variable length segment in the target collection. The automatic transcriptions of the target collection were also constructed by applying a Large Vocabulary Continuous Speech Recognition (LVCSR) decoder, to support

researchers in various fields.

To evaluate the test collection and to assess the baseline retrieval performance obtained by applying a standard method for SDR, an ad hoc retrieval experiment targeting the test collection was conducted. It revealed that the ad hoc retrieval task for lecture audio data was much more difficult than that for broadcast news.

We are now constructing another test collection for the term detection task. We will also prepare another automatic transcription with moderate WER by using an acoustic model and a language model trained in open conditions.

5. References

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2005. Overview of patent retrieval task at NTCIR-5. In *Proceedings of the Fifth NTCIR Workshop Meeting*, pages 269–277.

John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees. 1999. The TREC spoken document retrieval track: A success story. In *Proceedings of TREC-9*, pages 107–129.

Fredric C. Gey and Douglas W. Oard. 2001. The TREC-2001 cross-language information retrieval track: Searching arabic using english, french or arabic queries. In *Proceedings of TREC-10*, pages 16–25.

Tsutomu Hirao, Manabu Okumura, Takahiro Fukusima, and Hidetsugu Nanba. 2004. Text summarization challenge 3 – text summarization evaluation at NTCIR workshop 4. In *Proceedings of the Fourth NTCIR Workshop*.

Tsuneaki Kato, Jun'ichi Fukumoto, and Fumito Masui. 2005. An overview of NTCIR-5 QAC3. In *Proceedings of the Fifth NTCIR Workshop Meeting*, pages 361–372.

Tatsuya Kawahara, Hiroaki Nanjo, Takahiro Shinozaki, and Sadaoki Furui. 2003. Benchmark test for speech recognition using the corpus of spontaneous Japanese. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 135–138.

Kazuaki Kishida, Kuang hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, Hsin-Hsi Chen, and Sung Hyon Myaeng. 2005. Overview of CLIR task at the fifth NTCIR workshop. In *Proceedings of the Fifth NTCIR Workshop Meeting*, pages 1–38.

Tsuyoshi Kitani, Yasushi Ogawa, Tetsuya Ishikawa, Haruo Kimoto, Ikuo Keshi, Jun Toyoura, Toshikazu Fukushima, Kunio Matsui, Yoshihiro Ueda, Tetsuya Sakai, Takenobu Tokunaga, Hiroshi Tsuruoka, Hidekazu Nakawatase, and Teru Agata. 1998. Lessons from BMIR-J2: A test collection for Japanese IR systems. In *Proceedings of ACM SIGIR*, pages 345–346.

Akinobu Lee, Tatsuya Kawahara, and K. Shikano. 2001. Julius — an open source real-time large vocabulary recognition engine. In *Proceedings of European Conference on Speech Communication and Technology*, pages 1691–1694, Sept.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of Japanese. In *Proceedings of LREC*, pages 947–952.

Keizo Oyama, Masao Takaku, Haruko Ishikawa, Akiko Aizawa, and Hayato Yamana. 2005. Overview of the NTCIR-5 WEB navigational retrieval subtask 2. In *Proceedings of the Fifth NTCIR Workshop Meeting*, pages 423–442.

Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Pivoted document length normalization. In *Proceedings of ACM SIGIR*, pages 21–29.

Simone Teufel. 2007. An overview of evaluation methods in TREC ad hoc information retrieval and TREC question answering. In Laila Dybkjär, Holmer Hemsen, and Wolfgang Minker, editors, *Evaluation of Text and Speech Systems*, number 37 in Text, Speech and Language Technology, pages 163–186. Springer.

Ellen M. Voorhees and Dawn M. Tice. 1999. The TREC-8 question answering track evaluation. In *Proceedings of the 8th Text Retrieval Conference*, pages 83–106, Gaithersburg, Maryland.