Designing an Evaluation Framework for Spoken Term Detection and Spoken Document Retrieval at the NTCIR-9 SpokenDoc Task

Tomoyosi Akiba[(1)], **Hiromitsu Nishizaki**[(2)]
Kiyoaki Aikawa[(3)], **Tatsuya Kawahara**[(4)], **Tomoko Matsui**[(5)]

[(1)]Toyohashi University of Technology, 1-1 Hibarigaoka, Tenpaku, Toyohashi, Aichi, JAPAN, akiba@cs.tut.ac.jp
[(2)] University of Yamanashi, 4-3-11, Takeda, Kofu, Yamanashi, JAPAN, hnishi@yamanashi.ac.jp
[(3)] Tokyo University of Technology [(4)] Kyoto University [(5)] The Institute of Statistical Mathematics

Abstract

We describe the evaluation framework for spoken document retrieval for the IR for the Spoken Documents Task, conducted in the ninth NTCIR Workshop. The two parts of this task were a spoken term detection (STD) subtask and an ad hoc spoken document retrieval subtask (SDR). Both subtasks target search terms, passages and documents included in academic and simulated lectures of the Corpus of Spontaneous Japanese. Seven teams participated in the STD subtask and five in the SDR subtask. The results obtained through the evaluation in the workshop are discussed.

Keywords: spoken document retrieval, spoken term detection, passage retrieval

1. Introduction

The growth of the internet and the decrease in storage costs are resulting in the rapid increase of multimedia content today. To retrieve these contents, available text-based tag information is limited. Spoken Document Retrieval (SDR) is a promising technology for retrieving content using included speech data.

The NTCIR Workshop [1] is a series of evaluation workshops designed to enhance research in information access technologies by providing large-scale test collections and a forum for researchers. We proposed a new task called "IR for Spoken Documents," shortened to "SpokenDoc," and it was accepted as one of the core tasks in the ninth NTCIR Workshop(Sakai and Joho, 2001). In the NTCIR-9 SpokenDoc(Akiba et al., 2011), we evaluate SDR, especially based on a realistic ASR condition, where the target documents are spontaneous speech data with high word error rates and high out-of-vocabulary rates.

The Spoken Document Processing Working Group, which is part of the special interest group of spoken language processing (SIG-SLP) of the Information Processing Society of Japan, had already developed prototypes of SDR test collections, the CSJ Spoken Term Detection test collection(Itoh et al., 2010) and the CSJ Spoken Document Retrieval test collection(Akiba et al., 2009). The target documents of both test collections are spoken lectures in the Corpus of Spontaneous Japanese (CSJ)(Maekawa et al., 2000). By extending these test collections, two subtasks were defined.

Spoken Term Detection (STD): Within spoken documents, find the occurrence positions of a queried term. The evaluation should consider both the efficiency (search time) and the effectiveness (precision and recall).

Spoken Document Retrieval (SDR): Among spoken documents, find passages including information relevant to the query. This is like an ad hoc text retrieval task, except that the target documents are speech data. To accomplish the task, the results of STD may be used.

The organization of this paper is as follows. Section 2. describes the data prepared for our evaluation task. In Section 3. and 4., the task definitions and the evaluation results are presented for the STD and SDR subtasks, respectively. Finally, in Section 5., we give our conclusions.

2. Data

2.1. Document Collection

Our target document collection is the CSJ released by the National Institute for Japanese Language. Within CSJ, 2702 lectures were used as the target documents for our STD and SDR tasks (referred to as ALL). A subset of 177 lectures, called CORE, was also used as the target for our STD subtask (referred to as CORE). The participants were required to purchase the data for themselves. Each lecture in the CSJ is segmented by pauses that are no shorter than 200 ms. The segments are called Inter-Pausal Units (IPUs). An IPU is short enough to be used as an alternate to the position in the lecture. Therefore, the IPUs are used as the basic unit to be searched in both our STD and SDR tasks.

2.2. Transcription

Standard STD and SDR methods first transcribe the audio signal into its textual representation by using Large Vocabulary Continuous Speech Recognition (LVCSR), followed by text-based retrieval. The participants could use the following two types of transcriptions.

1. Reference automatic transcriptions
   The organizers prepared two reference automatic transcriptions. They enabled those who were interested in

[1] http://research.nii.ac.jp/ntcir/index-en.html

Table 1: ASR performances [%].
(a) For the CORE lectures.

Transcriptions	W.Corr.	W.Acc.	S.Corr.	S.Acc.
REF-WORD	76.7	71.9	86.5	83.0
REF-SYLLABLE	—	—	81.8	77.4

(b) For the ALL lectures.

Transcriptions	W.Corr.	W.Acc.	S.Corr.	S.Acc.
REF-WORD	74.1	69.2	83.0	78.1
REF-SYLLABLE	—	—	80.5	73.3

SDR but not in ASR to participate in our tasks. They also enabled comparisons of the IR methods based on the same underlying ASR performance. The participants can also use both transcriptions at the same time to boost the performance.

The textual representation of the transcriptions will be the N-best list of the word or syllable sequence depending on the two background ASR systems, along with their lattice and confusion network representations.

(a) Word-based transcription (denoted as "REF-WORD") obtained by using a word-based ASR system. In other words, a word n-gram model was used as the language model of the ASR system. With the textual representation, it also provides the vocabulary list used in the ASR, which determines the distinction between the in-vocabulary (IV) query terms and the out-of-vocabulary (OOV) query terms used in our STD subtask.

(b) Syllable-based transcription (denoted as "REF-SYLLABLE") obtained by using a syllable-based ASR system. The syllable n-gram model was used for the language model, where the vocabulary is all Japanese syllables. Using this model can avoid the OOV problem of the spoken document retrieval. Participants who want to focus on the open-vocabulary STD and SDR can use this transcription.

Table 1 shows the word-based correct rate ("W.Corr.") and accuracy ("W.Acc.") and the syllable-based correct rate ("S.Corr.") and accuracy ("S.Acc.") for these reference transcriptions.

2. Participant's own transcription
The participants could use their own ASR systems for the transcription. To enjoy the same IV and OOV conditions, we recommended that their word-based ASR systems should use the same vocabulary list as our reference transcription, but this was not necessary. When participating with their own transcriptions, the participants were encouraged to provide them to the organizers for future SpokenDoc test collections.

2.3. Speech Recognition Models

To realize open speech recognition, we used the following acoustic and language models, which were trained under the condition described below.
All speeches except the CORE parts were divided into two groups according to the speech ID number: an odd group and an even group. We constructed two sets of acoustic models and language models, and performed automatic speech recognition using the acoustic and language models trained by the other group.
The acoustic models are triphone based, with 48 phonemes. The feature vectors have 38 dimensions: 12-dimensional Mel-frequency cepstrum coefficients (MFCCs); the cepstrum difference coefficients (delta MFCCs); their acceleration (delta delta MFCCs); delta power; and delta delta power. The components were calculated every 10 ms. The distribution of the acoustic features was modeled using 32 mixtures of diagonal covariance Gaussian for the HMMs.
The language models are word-based trigram models with a vocabulary of 27k words. On the other hand, syllable-based trigram models, which were trained by the syllable sequences of each training group, were used to make the syllable-based transcription.
We used Julius (Lee and Kawahara, 2009) as a decoder, with a dictionary containing the above vocabulary. All words registered in the dictionary appeared in both training sets. The odd-group lectures were recognized by Julius using the even-group acoustic model and language model, while the even-group lectures were recognized using the odd-group models.
Finally, we obtained N-best speech recognition results for all spoken documents. The followings models and dictionary were made available to the participants of the SpokenDoc task.

- Odd acoustic models and language models
- Even acoustic models and language models
- A dictionary of the ASR

3. The Spoken Term Detection Subtask

3.1. Task Definition

Our STD task is to find all IPUs that include a specified query term in the CSJ. A term in this task is a sequence of one or more words. This is different from the STD task produced by NIST [2]
Participants can specify a suitable threshold for a score for an IPU; if the score for a query term is greater than or equal to the threshold, the IPU is output. One of the evaluation metrics is based on these outputs. However, participants can output up to 1000 IPUs for each query. Therefore, IPUs with scores less than the threshold may be submitted.

3.2. STD Query Set

We provided two sets of query term lists, one for ALL lectures and one for CORE lectures. Each participant's submission (called a "run") should choose the list corresponding to their target document collection, i.e., either ALL or CORE.

[2] "The Spoken Term Detection (STD) 2006 Evaluation Plan," http://www.nist.gov/speech/tests/std/docs/std06evalplanv10.pdf

We prepared 50 queries each for the CORE and ALL lecture sets. For the CORE, 31 of the 50 queries are OOV queries that are not included in the ASR dictionary and the others are IV queries. On the other hand, for the ALL, 24 of the 50 queries are OOV queries. The average occurrences per term is 7.1 times and 20.5 times for the CORE and ALL sets, respectively.
Each query term consists of one or more words. Because the STD performance depends on the length of the query terms, we selected queries of differing length. Query lengths range from 4 to 14 morae.

3.3. System Output

When a term is supplied to an STD system, all of the occurrences of the term in the speech data are to be found and the score for each occurrence of the given term is to be output. All STD systems must output the following information:

- document (lecture) ID containing the term;
- IPU ID;
- a score indicating the likelihood for the term's existence, with more positive values indicating its occurrence is more likely;
- a binary decision as to whether the detection is correct or not.

The score for each term occurrence can use any scale. However, the range of scores must be standardized for all terms.

3.4. Evaluation Measures

IPUs detected by each system were judged by whether or not the IPUs included the specified term. The judgment was based on a "correct IPUs list" for each specified term. The definition of correct IPUs for a specified term is based on perfect matching to the manual transcriptions of the CSJ in Japanese representation (Kanji, Hiragata and Katakana).
The official evaluation measure for effectiveness is the F-measure at the decision point specified by the participant, based on recall and precision averaged over queries (described as "F-measure (spec.)"). The F-measure at the maximum decision point (described as "F-measure (max)"), Recall–Precision curves and mean average precision (MAP) were also used for analysis purposes. They are defined as follows:

$$Recall = \frac{N_{corr}}{N_{true}} \tag{1}$$

$$Precision = \frac{N_{corr}}{N_{corr} + N_{spurious}} \tag{2}$$

$$F - measure = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}, \tag{3}$$

where N_{corr} and $N_{spurious}$ are the total number of correct and spurious (false) term (IPU) detections having scores greater than or equal to the threshold, and N_{true} is the total number of true term occurrences in the speech data. Recall–precision curves can be plotted by changing the threshold value. In the evaluation, the threshold value is varied in 100 steps. The F-measure at the maximum decision point is calculated as the optimal balance of Recall and Precision values from the recall–precision curve.
MAP for the set of queries is the mean value of the average precision values for each query. It can be calculated as follows:

$$MAP = \frac{1}{Q} \sum_{i=1}^{Q} AveP(i), \tag{4}$$

where Q is the number of queries and $AveP(i)$ means the average precision of the ith query of the query set. The average precision is calculated as the average of the precision values computed for each of the relevant terms in the list in which retrieved terms are ranked by a relevance measure:

$$AveP(i) = \frac{1}{Rel_i} \sum_{r=1}^{N_i} (\delta_r \cdot Precision_i(r)), \tag{5}$$

where r is the rank, N_i is the rank number at which all relevant terms of query i are found, and Rel_i is the number of relevant terms of query i. δ_r is a binary function on the relevance of a given rank r.

Table 2: STD evaluation results on each measurement for all submitted runs of the CORE set. The * mark indicates the organizers' team.

Runs	F-measure (max) [%]	F-measure (spec.) [%]	MAP
Baseline	0.527	0.516	0.595
AKBL-1*	0.393	0.393	0.264
AKBL-2*	0.385	0.370	0.272
ALPS-1*	0.725	0.708	0.837
ALPS-2*	0.714	0.697	0.757
IWAPU-1	0.644	0.628	0.772
IWAPU-2	0.510	0.297	0.733
NKGW-1	0.645	0.585	0.491
NKI11-1	0.570	0.559	0.684
NKI11-2	0.569	0.556	0.672
RYSDT-1	0.318	0.152	0.393
RYSDT-2	0.526	0.287	0.468
RYSDT-3	0.521	0.334	0.469
YLAB-1	0.425	0.425	0.344

3.5. Evaluation Results

3.5.1. STD Subtask Participants

In the NTCIR-9 SpokenDoc STD subtask, seven teams participated in the task (Kaneko et al., 2011; Nishizaki et al., 2011; Saito et al., 2011; Iwami and Nakagawa, 2011; Katsurada et al., 2011; Nanjo et al., 2011; Yamashita et al., 2011). Eighteen runs were submitted. All seven teams submitted results for the CORE query set. However, only two teams submitted results for the ALL query set.

3.5.2. Results

The evaluation results are summarized in Figure 1 and Table 2 for the CORE query set of the 13 submitted runs and the baseline. Figure 2 and Table 3 also show the STD performance for the ALL query set of the five submitted runs and the baseline. The offline processing time and index size

Table 4: System information related to the offline and online processing for those runs using indexing method.

Set	Runs	Offline Time [s.]	Index Size [K byte]	Search Time [ms.]	Machine Specifications
CORE	AKBL-1*	1147.716	3400000.00	1.7	Xeon X5560 2.67 GHz, 6 core × 2 CPUs, 24 GB mem.
	AKBL-2*	692.875	3400000.00	1.3	Xeon X5560 2.67 GHz, 6 core × 2 CPUs, 24 GB mem.
	NKGW-1	1420.700	3590000.00	1.6	Xeon 2.93 GHz 24 core CPU, 74 GB memory
	NKI11-1	626.497	5715.55	0.94	Core i7-2600 3.4 GHz, 8 GB memory
	NKI11-2	626.497	5715.55	0.94	Core i7-2600 3.4 GHz, 8 GB memory
ALL	NKI11-1	9009.770	83570.50	3.1	Core i7-2600 3.4 GHz, 8 GB memory
	NKI11-2	9009.770	83570.50	3.1	Core i7-2600 3.4 GHz, 8 GB memory

Table 3: STD evaluation results on each measurement for all submitted runs of the ALL set.

Runs	F-measure (max) [%]	F-measure (spec.) [%]	MAP
Baseline	0.459	0.310	0.451
NKI11-1	0.367	0.360	0.339
NKI11-2	0.396	0.332	0.344
RYSDT-1	0.531	0.070	0.431
RYSDT-2	0.530	0.082	0.426
RYSDT-3	0.531	0.119	0.434

are also shown in Table 4 only for the runs using some indexing method for efficient search.

The baseline system used dynamic programming (DP)-based word spotting, which could decide whether or not a query term is included in an IPU. The score between a query term and an IPU was calculated using the phoneme-based edit distance. The phoneme-based index for the baseline system was made of the transcriptions of REF-SYLLABLE. The decision point for calculating F-measure (spec.) was decided by the result of the dry-run query set. We adjusted the threshold to be the best F-measure value on the dry-run set, which was used as a development set.

For the CORE query set, most of the runs that used subword-based indexing and a simple matching method (DP or exact matching) outperformed the baseline performance for F-measure (max) and F-measure (spec.). On the other hand, the runs based on the Hough Transform algorithm (*AKBL and RYSDT*) and the VQ code book (*YLAB*) performed below the baseline.

The best STD performance was "ALPS-1," which uses much more of the information in the speech. It used 10 kinds of transcriptions of the speech. However, the retrieval time was the worst among all the submissions. "IWAPU-1" also obtained good STD performance by using a few subword-based indices. Therefore, combinations of multiple indexes may be effective in improving STD performance. Teams *NKGW* and *NKI11* achieved performance a little better than the baseline. However, their searches were far faster than those of teams *ALPS* and *IWAPU*.

The tasks using the ALL query set may be more difficult than those using the CORE query set because the baseline performance for ALL is less than that for CORE. Nevertheless, the only runs of team *RYSDT* outperformed the baseline for F-measure (max). These results are better than the CORE query set.

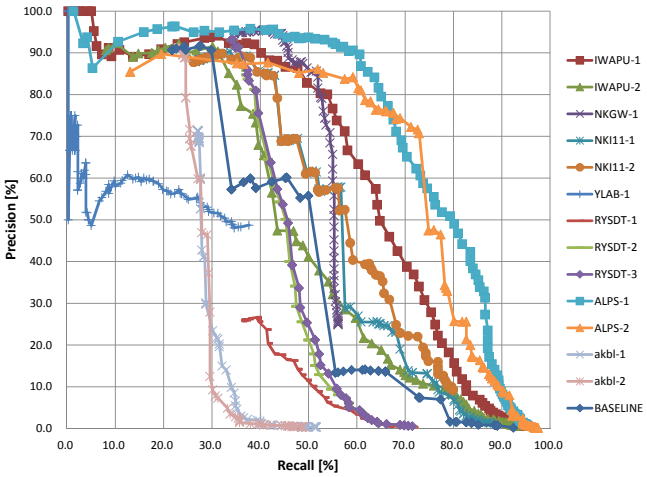

Figure 1: Recall–precision curves for the CORE query sets.

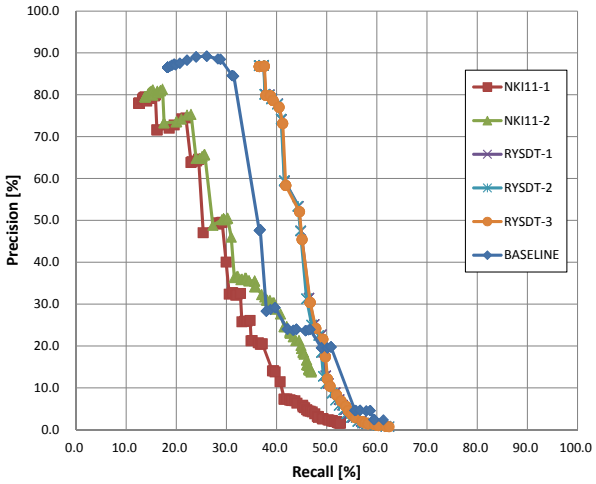

Figure 2: Recall–precision curves for the ALL query sets.

4. Spoken Document Retrieval Subtask

4.1. Task Definition

Two tasks (sub-subtasks) made up the SDR subtask; they share the same query topic list. The participants could submit the result of either or both tasks. The difference was in the unit of the target document to be retrieved.

- Lecture retrieval

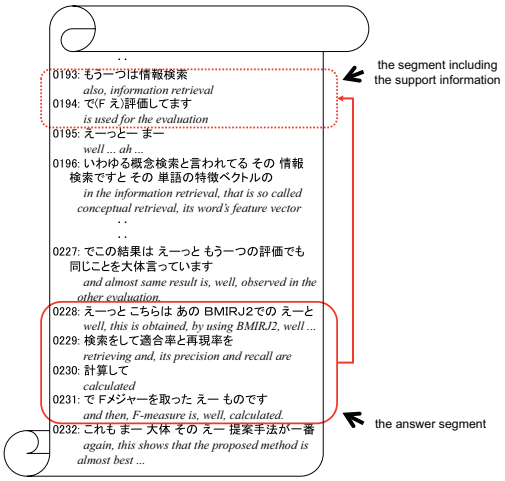

Figure 3: An example of an answer and its supporting segment.

Find the lectures that include the information described by the given query topic.

- Passage retrieval

  Find the passages that exactly include the information described by the given query topic. A passage is an IPU sequence of arbitrary length in a lecture.

4.2. Query Set

We constructed queries that asked for passages of varying lengths from lectures. Five people were relied upon to invent such queries by investigating the target documents and we obtained about 90 initial queries in total. Then, we checked their appropriateness. Some queries were removed because their topics were not appropriate for the SDR task, and the expressions of some were revised to reduce their ambiguity. Finally, we obtained 86 query topics.

4.3. Relevance Judgment

The relevance judgment for the queries was performed against every variable-length segment (or passage) in the target collection. One of the difficulties related to the 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 to the query in itself, when it had no supporting information that would convince the user who submitted the query of the correctness of the answer. For example, consider the query "How can we evaluate the performance of information retrieval?" The answer "F-measure" 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 "F-measure" is one of the evaluation metrics used for information retrieval. Figure 3 shows an example of an answer and its supporting information for the query "How can we evaluate the performance of information retrieval?"

As shown in Figure 3, 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 some 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.

For each query, one assessor, i.e., its constructor, read the relevant passages and judged their degrees of relevance. The assessor classified them according to the degree of their relevance: "Relevant," "Partially relevant," and "Irrelevant." Both the pooled passages or documents submitted from the participant groups and the search results using conventional word-based document search engines against the manual transcription of the target document collection were checked by the assessor.

4.4. Evaluation Measures

4.4.1. Lecture Retrieval

Mean Average Precision (MAP) was used for our official evaluation measure for lecture retrieval. For each query topic, the top 1000 documents were evaluated.

Given a question q, suppose the ordered list of documents $d_1 d_2 \cdots d_{|D|} \in D_q$ is submitted as the retrieval result. Then, $AveP_q$ is calculated as follows:

$$AveP_q = \frac{1}{|R_q|} \sum_{i=1}^{|D_q|} \delta(d_i \in R_q) \frac{\sum_{j=1}^{i} \delta(d_j \in R_q)}{i}, \quad (6)$$

where

$$\delta(a \in A) = \begin{cases} 1 & \cdots \quad a \in A \\ 0 & \cdots \quad a \notin A. \end{cases} \quad (7)$$

Alternatively, given the ordered list of correctly retrieved documents $r_1 r_2 \cdots r_M (M \leq |R_q|)$, $AveP_q$ is calculated as follows:

$$AveP_q = \frac{1}{|R_q|} \sum_{k=1}^{M} \frac{k}{rank(r_k)}, \quad (8)$$

where $rank(r)$ is the rank of document r.

MAP is the mean of the $AveP$ over all query topics Q:

$$\mathbf{MAP} = \frac{1}{|Q|} \sum_{q \in Q} AveP_q. \quad (9)$$

4.4.2. Passage Retrieval

In our passage retrieval task, the relevance of each arbitrary length segment (passage) rather than each whole lecture (document) must be evaluated. Three measures were designed for the task; one (uMAP) is utterance based and the other two (pwMAP and fMAP) are passage based.

4.4.3. Utterance-based Measure

uMAP

By expanding a passage into a set of utterances (IPUs) and by using an utterance (IPU) as a unit of evaluation like a document, we can use any conventional measures used for evaluating document retrieval.

Suppose the ordered list of passages $P_q = p_1 p_2 \cdots p_{|P_q|}$ is submitted as the retrieval result for a query q. Suppose we have a mapping function $O(p)$ from a (retrieved) passage p to an ordered list of utterances

$u_{p,1}u_{p,2}\cdots u_{p,|p|}$. We can obtain the ordered list of utterances $U = u_{p_1,1}u_{p_1,2}\cdots u_{p_1,|p_1|}u_{p_2,1}\ \cdots u_{p_{|P_q|},1}\cdots u_{p_{|P_q|},|p_{|P_q|}|}$. Then $uAveP_q$ is calculated as follows:

$$uAveP_q = \frac{1}{|\tilde{R}_q|}\sum_{i=1}^{|U|}\delta(u_i \in \tilde{R}_q)\frac{\sum_{j=1}^{i}\delta(u_j \in \tilde{R}_q)}{i}, \tag{10}$$

where $U = u_1\cdots u_{|U|}(|U| = \sum_{p\in P}|p|)$ is the renumbered ordered list of U and $\tilde{R}_q = \bigcup_{r\in R_q}\{u|u \in r\}$ is the set of relevant utterances extracted from the set of relevant passages R_q.

For the mapping function $O(p)$, we will use as our oracle the ordering mapping function, which orders the utterances in the given passage p so the relevant utterances come first. For example, given a passage $p = u_1u_2u_3u_4u_5$ and suppose the relevant utterances are u_3 and u_4, it returns the passage as $u_3u_4u_1u_2u_5$.

uMAP (utterance-based MAP) is defined as the mean of the $uAveP$ over all query topics Q:

$$\mathbf{uMAP} = \frac{1}{|Q|}\sum_{q\in Q}uAveP_q. \tag{11}$$

4.4.4. Passage-based Measure

For our passage retrieval, two tasks must be completed: one is to determine the boundary of the passages to be retrieved and the other is to rank the relevance of the passages. The first passage-based measure focuses only on the latter task and the second measure focuses on both tasks.

pwMAP

For a given query, a system returns an ordered list of passages. For each returned passage, only utterances located in the center of it are considered for relevance. If the center utterance is included in some relevant passage described in the golden file, then the returned passage is deemed relevant and the relevant passage is considered to be retrieved correctly. However, if there exists at least one formerly listed passage that is also deemed relevant with respect to the same relevant passage, the returned passage is deemed not relevant as the relevant passage has been retrieved already. In this way, all the passages in the returned list are labeled by their relevance. Now, any conventional evaluation metric designed for document retrieval can be applied to the returned list.

Suppose we have the ordered list of correctly retrieved passages $r_1r_2\cdots r_M(M \le |R_q|)$, where their relevances are judged according to the process mentioned above. $pwAveP_q$ is calculated as follows:

$$pwAveP_q = \frac{1}{|R_q|}\sum_{k=1}^{M}\frac{k}{rank(r_k)}, \tag{12}$$

where $rank(r)$ is the rank of passage r in the original ordered list of retrieved passages.

pwMAP (pointwise MAP) is defined as the mean of the $pwAveP$ over all query topics Q:

$$\mathbf{pwMAP} = \frac{1}{|Q|}\sum_{q\in Q}pwAveP_q. \tag{13}$$

fMAP

This measure evaluates the relevance of a retrieved passage fractionally against the relevant passage in the golden files. Given a retrieved passage $p \in P_q$ for a given query q, its relevance level $rel(p, R_q)$ is defined as the fraction of some relevant passage that it covers, as follows:

$$rel(p, R_q) = \max_{r\in R_q}\frac{|r \cap p|}{|r|}. \tag{14}$$

Here r and p are regarded as sets of utterances. rel can be seen as measuring the recall of p at the utterance level. Accordingly, we can define the precision of p as follows:

$$prec(p, R_q) = \max_{r\in R_q}\frac{|p \cap r|}{|p|}. \tag{15}$$

Then, $fAveP_q$ is calculated as follows:

$$fAveP_q = \frac{1}{|R_q|}\sum_{i=1}^{|P_q|}rel(p_i, R_q)\frac{\sum_{j=1}^{i}prec(p_j, R_q)}{i}. \tag{16}$$

fMAP (fractional MAP) is defined as the mean of the $fAveP_q$ over all query topics Q.

$$\mathbf{fMAP} = \frac{1}{|Q|}\sum_{q\in Q}fAveP_q \tag{17}$$

4.5. Evaluation Results

Five groups(Kaneko et al., 2011; Hasegawa et al., 2011; Eskevich and Jones, 2011; Nanjo et al., 2011; Tsuge et al., 2011) submitted a total of 21 runs. Four of the groups participated in the lecture retrieval task and three participated in the passage retrieval task.

4.5.1. Transcriptions

All participants used textual transcription, to which some retrieval method was applied. One participant group used their own transcription, while the others used the transcriptions provided by the organizers. Of the organizer's automatic transcriptions, most runs used the word-based transcription, while three runs for lecture retrieval by one group used both the word and syllable transcriptions at the same time, and two runs by one group, one for lecture and one for passage retrieval, used only the syllable transcriptions. Looking into the usage of the automatic transcription, one group used multiple (10-best) recognition candidates, while the other used only a single (1-best) candidate.

4.5.2. Baseline Methods

We implemented and evaluated the baseline methods for our SDR tasks, which consisted of only conventional methods for IR applied to the 1-best REF-WORD or MANUAL transcription. Only nouns were used for indexing; they were extracted from the transcription by applying a Japanese morphological analysis tool. The vector-space model was used as the retrieval model, and Term Frequency–Inverse Document Frequency (TF–IDF) with pivoted normalization (Singhal et al., 1996) was used for term weighting. We used *GETA* [3] as the IR engine for the baselines.

[3] `http://geta.ex.nii.ac.jp`

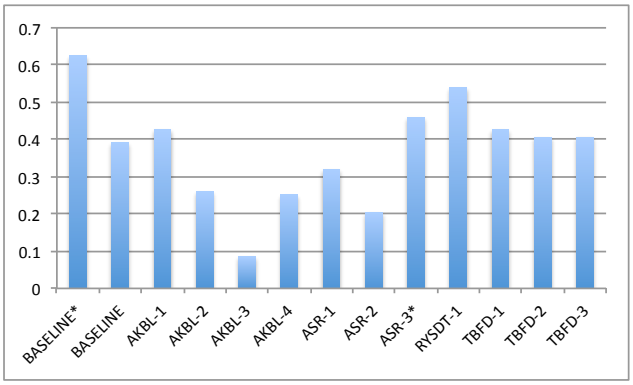

Figure 4: Evaluation results for the lecture retrieval task. The * mark indicates that the run uses the manual transcription.

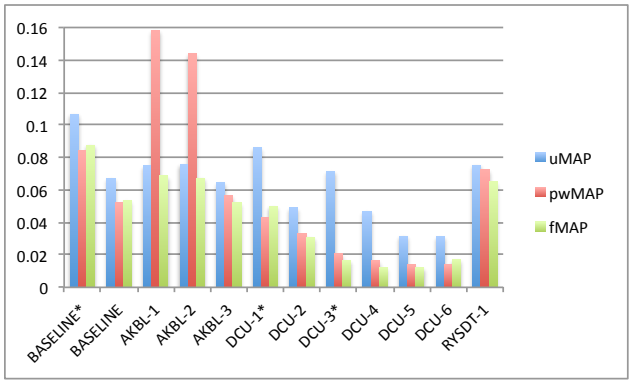

Figure 5: Evaluation results for the passage retrieval task. The * mark indicates that the run uses the manual transcription.

For the lecture retrieval task, each lecture in the CSJ was indexed and retrieved by the IR engine. For the passage retrieval task, we created pseudopassages by automatically dividing each lecture into a sequence of segments, with N utterances per segment. We set $N = 15$ according to a rough estimate of the passage lengths of the dry-run test data.

The average precisions of the baseline lecture retrieval system for each query indicated that the variance in difficulty among queries was high. It also indicated that the variance in the performance difference between using manual and automatic transcriptions was also high; for some queries, the retrieval on the manual transcriptions was perfect, while that on the automatic transcriptions did not work at all. This suggested that dealing with the mismatch between the query topic and the transcription, which is mainly caused by the OOV on the query and the recognition errors on the transcription, was one of the main challenges of SDR, although the OOV rate on the formal run queries was not high; only three queries included the OOV words against the REF-WORD transcription.

4.5.3. Results

For the lecture retrieval task, the evaluation results of all the submissions are summarized in Figure 4. It was obvious from the results that the runs using manual transcription outperformed their counterparts using automatic transcription. Among the runs using automatic transcription, RYSDT-1 outperformed the baseline significantly and TBFD-1 did so weakly (the p-value was 0.062 for the two-sided paired t-test), while AKBL-1, TBFD-2, and TBFD-3 also outperformed the baseline but not significantly. Because the retrieval technique used in RYSDT-1 was almost the same as the baseline, its use of its own transcription seemed to be the major factor of the improvement.

For the passage retrieval task, the three evaluation measures, uMAP, pwMAP, and fMAP were used. The correlation coefficients between uMAP and pwMAP, between pwMAP and fMAP, and between uMAP and fMAP, calculated using all the submitted runs, are 0.750, 0.869, and 0.884, respectively. This shows that these measures correlate well with each other, and that those measuring the same aspects (i.e., uMAP and fMAP, which measure both the accuracy of the boundaries and the relevance, while pwMAP measures only the latter) and those based on the same unit (i.e., pwMAP and fMAP, which are passage based, while uMAP is utterance based) correlate better than the others (i.e., uMAP and fMAP). The evaluation results are summarized in Figure 5. In the passage results, as in the lecture retrieval results, the runs using manual transcription outperformed their counterparts using automatic transcriptions. We will investigate below only the results using automatic transcription.

In terms of uMAP, AKBL-1, AKBL-2 and RYSDT-1 outperformed the baseline, but the differences were not significant. However, in terms of pwMAP, AKBL-1, AKBL-2 and RYSDT-1 outperformed the baseline significantly. In particular, the pwMAP values of AKBL-1 and AKBL-2 were best among all the runs including those using manual transcription. This seemed to be because their methods for reducing the redundant results worked effectively, especially for the pwMAP measure. In terms of fMAP, AKBL-1, AKBL-2 and RYSDT-1 also outperformed the baseline significantly but weakly (at the 0.05 level for the two-sided paired t-test).

5. Conclusion

We have presented an overview of the IR for Spoken Documents (SpokenDoc) task in the NTCIR-9 Workshop. Our task consisted of the spoken term detection (STD) subtask and the ad hoc spoken document retrieval subtask (SDR). Both subtasks targeted search terms, passages and documents included in academic and simulated lectures of the Corpus of Spontaneous Japanese. Seven teams participated in the STD subtask and five participated in the SDR subtask.

We will have the second round of the SpokenDoc task in the next NTCIR-10 Workshop. The details of the task will be found in the NTCIR-10 Web page[4]. Please consider joining us, if you are interested in our task presented in this paper.

6. References

Tomoyosi Akiba, Kiyoaki Aikawa, Yoshiaki Itoh, Tatsuya Kawahara, Hiroaki Nanjo, Hiromitsu Nishizaki, Norihito

[4] http://research.nii.ac.jp/ntcir/ntcir-10/index.html

Yasuda, Yoichi Yamashita, and Katunobu Itou. 2009. Developing an sdr test collection from japanese lecture audio data. In *Proceedings of the 1st Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC2009)*, pages 324–330.

Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyoaki Aikawa, Tatsuya Kawahara, and Tomoko Matsui. 2011. Overview of the IR for spoken documents task in NTCIR-9 workshop. In *Proceedings of The Ninth NTCIR Workshop Meeting*, pages 223–235.

Maria Eskevich and Gareth J. F. Jones. 2011. DCU at the NTCIR-9 SpokenDoc passage retrieval task. In *Proceedings of The Ninth NTCIR Workshop Meeting*.

Kiichi Hasegawa, Hideki Sekiya, Masanori Takehara, Taro Niinomi, Satoshi Tamura, and Satoru Hayamizu. 2011. Toward improvement of SDR accuracy using LDA and query expansion for SpokenDoc. In *Proceedings of The Ninth NTCIR Workshop Meeting*.

Yoshiaki Itoh, Hiromitsu Nishizaki, Xinhui Hu, Hiroaki Nanjo, Tomoyosi Akiba, Tatsuya Kawahara, Seiichi Nakagawa, Tomoko Matsui, Yoichi Yamashita, and Kiyoaki Aikawa. 2010. Constructing japanese test collections for spoken term detection. In *Proceedings of International Conference on Speech Communication and Technology*, pages 667–680.

Keisuke Iwami and Seiichi Nakagawa. 2011. High speed spoken term detection by combination of n-gram array of a syllable lattice and LVCSR result for NTCIR-SpokenDoc. In *Proceedings of The Ninth NTCIR Workshop Meeting*.

Taisuke Kaneko, Tomoko Takigami, and Tomoyosi Akiba. 2011. STD based on hough transform and SDR using STD results: Experiments at NTCIR-9 SpokenDoc. In *Proceedings of The Ninth NTCIR Workshop Meeting*.

Kouichi Katsurada, Koudai Katsuura, Yurie Iribe, and Tsuneo Nitta. 2011. Utilization of suffix array for quick STD and its evaluation on the NTCIR-9 SpokenDoc task. In *Proceedings of The Ninth NTCIR Workshop Meeting*.

A. Lee and T. Kawahara. 2009. Recent development of open-source speech recognition engine julius. In *Proceedings of the 1st Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC2009)*, page 6 pages.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of Japanese. In *Proceedings of International Conference on Language Resources and Evaluation*, pages 947–952.

Hiroaki Nanjo, Kazuyuki Noritake, and Takehiko Yoshimi. 2011. Spoken document retrieval experiments for spokendoc at ryukoku university (RYSDT). In *Proceedings of The Ninth NTCIR Workshop Meeting*.

Hiromitsu Nishizaki, Hiroto Furuya, Satoshi Natori, and Yoshihiro Sekiguchi. 2011. Spoken term detection using multiple speech recognizers' outputs at NTCIR-9 SpokenDoc STD subtask. In *Proceedings of The Ninth NTCIR Workshop Meeting*.

Hiroyuki Saito, Takuya Nakano, Shirou Narumi, Toshiki Chiba, Kazuma Kon'No, and Yoshiaki Itoh. 2011. An STD system for OOV query terms using various subword units. In *Proceedings of The Ninth NTCIR Workshop Meeting*.

Tetsuya Sakai and Hideo Joho. 2001. Overview of NTCIR-9. In *Proceedings of The Ninth NTCIR Workshop Meeting*, pages 1–7. `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-OV-SakaiT.pdf`.

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

Satoru Tsuge, Hiromasa Ohashi, Norihide Kitaoka, Kazuya Takeda, and Kenji Kita. 2011. Spoken document retrieval method combining query expansion with continuous syllable recognition for NTCIR-SpokenDoc. In *Proceedings of The Ninth NTCIR Workshop Meeting*.

Yoichi Yamashita, Toru Matsunaga, and Kook Cho. 2011. YLAB@RU at spoken term detection task in NTCIR9-SpokenDoc. In *Proceedings of The Ninth NTCIR Workshop Meeting*.