Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop

Tomoyosi Akiba
Toyohashi University of Technology
1-1 Hibarigaoka, Tohohashi-shi
Aichi, 440-8580, Japan
akiba@cs.tut.ac.jp

Hiromitsu Nishizaki
University of Yamanashi
4-3-11 Takeda, Kofu
Yamanashi, 400-8511, Japan
hnishi@yamanashi.ac.jp

Kiyoaki Aikawa
Tokyou University of Technology
1404-1 Katakura, Hachioji
Tokyo, 192-0982, Japan
aik@media.teu.ac.jp

Tatsuya Kawahara
Kyoto University
Yoshidahonmachi, Sakyo-ku
Kyoto, 606-8501, Japan
kawahara@media.kyoto-u.ac.jp

Tomoko Matsui
The Institute of Statistical Mathematics
10-3 Midorimachi, Tachikawa
Tokyo, 190-8562, Japan
tmatsui@ism.ac.jp

ABSTRACT

This paper describes an overview of the IR for Spoken Documents Task in NTCIR-9 Workshop. In this task, the spoken term detection (STD) subtask and ad-hoc spoken document retrieval subtask (SDR) are conducted. Both of the subtasks target to search terms, passages and documents included in academic and simulated lectures of the Corpus of Spontaneous Japanese. Finally, seven and five teams participated in the STD subtask and the SDR subtask, respectively. This paper explains the data used in the subtasks, how to make transcriptions by speech recognition and the details of each subtask.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Performance

Keywords

NTCIR-9, spoken document retrieval, spoken term detection

1. INTRODUCTION

The growth of the internet and the decrease of the storage costs are resulting in the rapid increase of multimedia contents today. For retrieving these contents, available text-based tag information is limited. Spoken Document Retrieval (SDR) is a promising technology for retrieving these contents using the speech data included in them. In NTCIR-9 SpokenDoc (IR for Spoken Documents), we evaluate the SDR, especially based on a realistic ASR condition, where the target documents are spontaneous speech data with high word error rate and high out-of-vocabulary rate.

The Spoken Document Processing Working Group[1], which is part of the special interest group of spoken language processing (SIG-SLP) of the Information Processing Society of Japan, have already developed prototypes of SDR test collections; CSJ Spoken Term Detection test collection and CSJ Spoken Document Retrieval test collection[1]. The target documents of both the test collections are spoken lectures in Corpus of Spontaneous Japanese (CSJ)[8]. By using (and extending) these test collections, two subtasks were conducted.

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

Spoken Document Retrieval: Among spoken documents, find the passages including the relevant information related 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 result of STD may be used.

2. DOCUMENT COLLECTION

Our target document collection is the Corpus of Spontaneous Japanese (CSJ) released by the National Institute for Japanese Language. Among CSJ, 2702 lectures (about 600 hours) are used as the target documents for our both STD and SDR tasks (referred to as ALL). The subset 177 lectures (about 44 hours) of them, called CORE, is also used for the target for our STD subtask (referred to as CORE). The participants are required to purchase the data by themselves. Each lecture in the CSJ is segmented by the pauses that are no shorter than 200 msec. The segment is called Inter-Pausal Unit (IPU). An IPU is short enough to be used as the 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.

3. TRANSCRIPTION

Standard STD methods first transcribe the audio signal into its textual representation by using Large Vocabulary

[1] http://www.cl.ics.tut.ac.jp/~sdpwg

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 automatic reference transcriptions. These enabled participants who are interested in SDR but not in ASR to participate in these tasks. It also enables the comparison of the IR methods used by different participants based on the same underlying ASR performance. The participants were also permitted to use both transcriptions at the same time to boost the performance.

   The textual representation of them is the N-best list of the word or syllable sequence depending on the two background ASR systems, along with the lattice and confusion network representation of them.

   (a) Word-based transcription (denoted as "REF-WORD") obtained by using a word-based ASR system. In other words, a word n-gram model is used for 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 our-of-vocabulary (OOV) query terms used in our STD subtask. 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 REF-WORD of the ALL and CORE lectures.

   (b) Syllable-based transcription (denoted as "REF-SYLLABLE") obtained by using a syllable-based ASR system. The syllable n-gram model is used for the language model, where the vocabulary is the all Japanese syllables. The use of it can avoid the OOV problem of the spoken document retrieval. The participants who want to focus on the open vocabulary STD and SDR can use this transcription. Table 1 also shows the syllable-based correct rate and accuracy for REF-SYLLABLE of the ALL and CORE lectures.

2. Participant's own transcription
   The participants can use their own ASR systems for the transcription. In order to enjoy the same IV and OOV condition, their word-based ASR systems are recommended to use the same vocabulary list of our reference transcription, but not necessary. When participating with the own transcription, the participants are encouraged to provide it to the organizers for the future SpokenDoc test collections.

4. SPEECH RECOGNITION MODELS

To realize open speech recognition, we used the following acoustic and language models, which are trained under the condition as described below.

All speeches except CORE parts were divided into two groups according to the speech ID: 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.

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

The acoustic models are tri-phone based, consisting of 48 phonemes. The feature vectors consist of 38 dimensions: 12 dimensional Mel-frequency cepstrum coefficients (MFCCs), the cepstrum difference coefficients (delta MFCCs), its acceleration (delta delta MFCCs), delta power, and delta delta power, and they ere calculated every 10 msec. 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 27k vocabulary. On the other hand, syllable-based trigram models, which are trained by the syllable sequences of each training group, are used to make the syllable-based transcription.

We used Julius [7] as a decoder, with a dictionary containing 27k vocabulary. All words registered in the dictionary were appeared in the both training set. The odd group lectures are recognized by the Julius using the even acoustic model and language model. And the even group lectures are recognized by it with the odd models.

Finally, we obtained N-best speech recognition results for all spoken documents. The followings models and dictionary can be 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

5. SPOKEN TERM DETECTION SUBTASK

5.1 The task definition

Our STD task is to find all IPUs that include a specified query term in the CSJ. For the STD subtask, a term 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 of a score for an IPU. If a score of an IPU for a query term is greater than or equal to the threshold, the IPU is outputted. One of evaluation metrics is based on these outputs. However, participants can output IPUs up to 1,000 per each query. Therefore, IPUs with scores less than the threshold may be submitted.

5.2 STD query set

We provided two sets of the query term list, i.e. the list for ALL lectures and the list for the CORE lectures. Each participant's submission (called "run") should choose one from

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

the two according to their target document collection, i.e. either ALL or CORE.

We prepared the 50 queries sets for the CORE and ALL lectures sets. For the CORE, 31 of the all 50 queries are out-of-vocabulary queries that do not included in the ASR dictionary and the others are in-vocabulary queries. On the other hand, for the ALL, 24 of the all 50 queries are out-of-vocabulary queries. The average occurrences per a 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. The range of query length distributes from 4 to 14 morae.

5.3 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 score for each occurrence of the given term are to be output.

All STD systems must output following information:

- document (lecture) ID of the term,
- IPU ID,
- a score indicating how likely the term exists with more positive values indicating more likely occurrence
- a binary decision as to whether the detection is correct or not.

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

5.4 Submission of the formal-run

Each participant is allowed to submit as many search results ("runs") as they want. Submitted runs should be prioritized by each group. Priority number should be assigned through all submissions of a participant, and smaller number has higher priority.

File Name

A single run is saved in a single file. Each submission file should have an adequate file name following the next format.

STD-*X*-*D*-*N*.txt

X: System identifier that is the same as the group ID (e.g., NTC)

D: Target document set:

- ALL: **ALL** 2702 lectures.
- CORE: **CORE** 177 lectures.

N: Priority of run (1, 2, 3, ...) for each target document set.

For example, if the group "NTC" submits two files for targeting **ALL** lectures and three files for **CORE** lectures, the names of the run files should be "STD-NTC-ALL-1.txt", "STD-NTC-ALL-2.txt", "STD-NTC-CORE-1.txt", "STD-NTC-CORE-2.txt", and "STD-NTC-CORE-3.txt".

Submission Format

The submission files are organized with the following tags. Each file must be a well-formed XML document. It has a single root level tag "<**ROOT**>". It has three main sections, "<**RUN**>", "<**SYSTEM**>", and "<**RESULTS**>".

- <**RUN**>

  <**SUBTASK**> "STD" or "SDR". For a STD subtask submission, it must be "STD".

  <**SYSTEM-ID**> System identifier that is the same as the group ID.

  <**PRIORITY**> Priority of the run.

  <**TARGET**> The target document set, or the used query term set accordingly. "ALL" if the target document set is **ALL** lectures. "CORE" if **CORE** lectures.

  <**TRANSCRIPTION**> The transcription used as the text representation of the target document set. "MANUAL" if it is the manual transcription provided by the CSJ. "REF-WORD" if it is the reference word-based automatic transcription provided by the organizers. "REF-SYLLABLE" if it is the reference syllable-based automatic transcription provided by the organizers. "OWN" if it is obtained by a participant's own recognition. "NO" if no textual transcription is used.

- <**SYSTEM**>

  <**OFFLINE-MACHINE-SPEC**>
  <**OFFLINE-TIME**>
  <**INDEX-SIZE**>
  <**ONLINE-MACHINE-SPEC**>
  <**ONLINE-TIME**>
  <**SYSTEM-DESCRIPTION**>

- <**RESULTS**>

  <**QUERY-ID**> Each query term has a single "QUERY-ID" tag with an attribute "id" specified in a query term list. Within this tag, a list of the following "TERM" tags is described.

  <**TERM**> Each potential detection of a query term has a single "TERM" tag with the following attributes.

  document The searched document (lecture) ID specified in the CSJ.

  ipu The searched Inter Pausal Unit(IPU) ID specified in the CSJ.

  score The detection score indicating the likelihood of the detection. The greater is more likely.

  detection The binary ("YES" or "NO") decision of whether or not the term should be detected to make the optimal evaluation result.

Figure 4 shows an example of a submission file.

```
<ROOT>
<RUN>
<SUBTASK>STD</SUBTASK>
<SYSTEM-ID>TUT</SYSTEM-ID>
<PRIORITY>1</PRIORITY>
<TARGET>CORE</TARGET>
<TRANSCRIPTION>REF-SYLLABLE</TRANSCRIPTION>
</RUN>
<SYSTEM>
<OFFLINE-MACHINE-SPEC>Xeon 3GHz dual CPU, 4GB memory
</OFFLINE-MACHINE-SPEC>
<OFFLINE-TIME>18:35:23</OFFLINE-TIME>
...
</SYSTEM>                                            [t]
<RESULTS>
<QUERY id="SpokenDoc1-STD-dry-CORE-001">
<TERM document="A01F0005" ipu="0024" score="0.83"
 detection="YES" />
<TERM document="S00M0075" ipu="0079" score="0.32"
 detection="NO" />
...
</QUERY>
<QUERY id="SpokenDoc1-STD-dry-CORE-002">
...
</QUERY>
</RESULTS>
</ROOT>
```

Figure 1: An example of the submission file.

5.5 Evaluation measures

Detected IPUs by each system are judged whether the IPUs include a specified term or not. The judgment is 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, Hiragana and Katakana) level.

The official evaluation measure for effectiveness is F-measure at the decision point specified by the participant, based on recall and precision averaged over queries (described as "F-measure(spec.)"). F-measure at the maximum decision point (described as "F-measure(max)"), Recall-Precision curves and mean average precision (MAP) are also used for analysis purpose.

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 whose scores are 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. F-measure at the maximum decision point is calculated at 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 calculate 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 i-th query of the query set. The average precision is calculated by averaging of the precision values computed at the point of 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 the all relevance terms of query i are found, and Rel_i is the number of the relevance terms of query i. δ_r is a binary function on the relevance of a given rank r.

Table 3: The number of transcription(s) used for each run.

Set	Run	REF-WORD	REF-SYLLABLE	OWN	total
CORE	AKBL-1	0	1	0	1
	AKBL-2	0	1	0	1
	ALPS-1	1	1	8	10
	ALPS-2	1	1	8	10
	IWAPU-1	0	0	4	4
	IWAPU-2	0	0	4	4
	NKGW-1	0	0	2	2
	NKI11-1	0	1	0	1
	NKI11-2	0	1	0	1
	RYSDT-1	1	0	0	1
	RYSDT-2	1	0	0	1
	YLAB-1	0	0	0	0
ALL	NKI11-1	0	1	0	1
	NKI11-2	0	1	0	1
	RYSDT-1	1	0	0	1
	RYSDT-1	1	0	0	1

5.6 Evaluation result

5.6.1 STD subtask participants

In NTCIR-9 SpokenDoc STD subtask, seven teams participated in the task with 18 submission runs. The term ID is listed in Table 2. All the seven teams submitted the results for the CORE query set. However, only the two teams submitted the results of the ALL query set.

5.6.2 STD Techniques used

Here, we provide a brief overview of STD techniques used by the participants. For more details, please refer to the participant's papers. Table 3 summarizes the number of transcription(s) used for each run.

AKBL [5] submitted two runs for the CORE set. The indexing method, called Metric Subspace Indexing, was quite different from those used in the text indexing. The term detection was performed on these indices and based on the Hough Transform algorithm usually used in the image processing. The method incorporating the multiple candidates from speech recognition into their indexing was also investigated. The used transcription was REF-SYLLABLE.

Table 2: STD subtask participants. * mark indicates the organizers' team.

For CORE set			
Team ID	Team name	Organization	# of submitted runs
AKBL*	Akiba Laboratory	Toyohashi University of Technology	2
ALPS*	ALPS lab. at UY	University of Yamanashi	2
IWAPU	Iwate Prefectural University	Iwate Prefectural University	2
NKGW	NAKAGAWA LAB	Toyohashi University of Technology	1
NKI11	NKI-Lab	Toyohashi University of Technology	2
RYSDT	Ryukoku_NL-SLP_lab	Ryukoku University	3
YLAB	Yamashita laboratory	Ritsumeikan University	1

For ALL set			
Team ID	Team name	Organization	# of submitted runs
NKI11	NKI-Lab	Toyohashi University of Technology	2
RYSDT	Ryukoku_NL-SLP_lab	Ryukoku University	3

ALPS [10] submitted two runs for the CORE set. The ten (including REF-WORD, REF-SYLLABLE, and OWN) transcriptions obtained from various recognition systems were incorporated into the sausage-style lattice, called PTN, and the search was performed on it by using the DTW (Dynamic Time Warping) algorithm. To reduce the false detection, two additional scores, which roughly corresponded to the degree of consensus and ambiguity in the competing syllables in the lattice, were also incorporated into the distance score used in the DTW process.

IWAPU [11] submitted two runs for the CORE set. They used multiple OWN transcriptions of various subword units, including monophone, triphone, syllable, demiphone, and Sub-phonetic segment (SPS), by using the multiple speech recognition systems prepared for these units. The query was also converted to these subword sequences, then the detection was performed for each subword representation using the DTW algorithm. These multiple detection results were integrated into the final results by interpolating their detection scores linearly to improve the STD performance.

NKGW [4] submitted one run for the CORE set. Two OWN transcriptions were used at the same time, where the word-based transcription was used for the in-vocabulary (IV) search query and the syllable-based transcription was used for the IV and out-of-vocabulary (OOV) query. For the syllable-based transcription, some kind of inverted index based on the syllable tri-gram was used for indexing, where the substitution and insertion errors were dealt with by introducing the extra error-predicted indices, while the deletion error was dealt with by removing a syllable from the input query sequence. The index was also augmented by the distance score according to the error-prediction, which was used to reduce the false detection without applying the expensive DTW-based confirmation.

NKI11 [6] submitted two runs for the CORE set and two runs for the ALL set. The suffix array was used for indexing, which was constructed from the phoneme sequence obtained from the transcription of spoken documents. At the detection time, the suffix array was searched against the phoneme sequence of the query by using DTW algorithm. To improve the efficiency of the search on the suffix array, the phoneme sequence of the long query term was divided into subsequences, each of which was then searched against the suffix array. The detection results of the subsequences were further confirmed to form the final detection results. The used transcription was REF-SYLLABLE.

RYSDT [9] submitted three runs for the CORE set and three runs for the ALL set. The term detection was performed based on the Hough Transform, a line detection algorithm usually used in image processing area. Several filtering methods were applied to the query-document image plane to improve the line detection performance. The used transcription was REF-WORD.

YLAB [14] submitted one run for the CORE set. They used NO transcription obtained from speech recognition, but used the vector quantization (VQ) sequence of spoken document. For each document group consisted of the lectures by the same speaker, the individual VQ code set was produced and used only for it. The V-P score, which encoded the similarity between a phoneme and a VQ code, was obtained from the same document group and used for the detection guided by the DTW algorithm.

5.6.3 Results

The evaluation results are summarized in Figure 2 and Table 4 for the CORE query set of the 13 submitted runs and the baseline. Figure 3 and Table 5 shows also the STD performance for the ALL query set of the five submitted runs and the baseline. The offline processing time and index size are also shown in Table 6 only for the runs using some indexing method for efficient search.

The baseline system has a dynamic programming (DP) based word spotting, which can decide whether a query term is included in an IPU or not. The score between a query term and an IPU is calculated based on phoneme-based edit distance. The phone-based index for the baseline system is made of the transcriptions of REF-SYLLABLE. The decision point for calculating ąÈF-measure(spec.)ąÉ 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.

Fore the CORE query set, most of the runs that used the subword-based indexing and a simple matching method (DP or exact matching) outperforms the baseline performance for

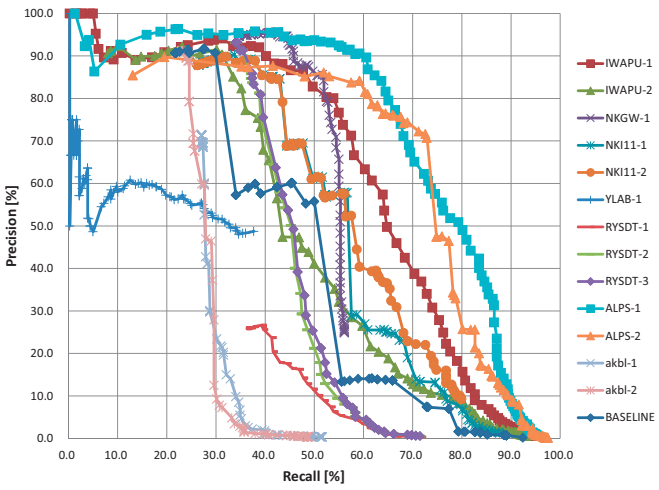


Figure 2: Recall-precision curves for the CORE query sets.

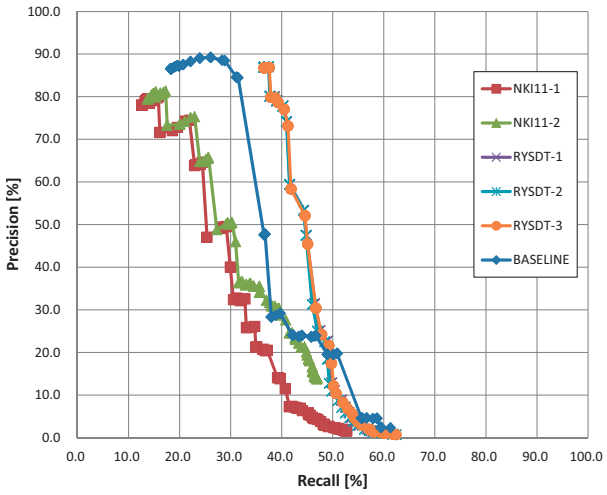


Figure 3: Recall-precision curves for the ALL query sets.

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*) are less than the baseline.

The best STD performance is "ALPS-1" which uses much more the amount of information of the speech. It used 10 kinds of transcriptions of the speech. However, the retrieval time is the worst among the all submissions. "IWAPU-1" also got good STD performance, which uses a few kinds of subword-based indices. Therefore, to combination of multiple types of index may be effective to improve STD performance. Term *NKGW* and *NKI11* got the performance little bit better than the baseline. However, they realized the fast search compared with term *ALPS* and *IWAPU*.

In the ALL query set, it may be more difficult task comparing with the CORE query set because the baseline performance of the ALL is less than the one of the CORE. Nevertheless, the only runs of term *RYSDT* outperformed the baseline at F-measure(max). These results are better than the CORE query set.

6. SPOKEN DOCUMENT RETRIEVAL SUBTASK

6.1 Task Definition

Two tasks (sub-subtasks) were conducted for the SDR subtask, both of which share the same query topic list. The participants could submit the result of either or both of the tasks. The difference was in the unit of the target document to be retrieved.

- Lecture retrieval

  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.

6.2 Query Set

We constructed queries that ask for passages of varying lengths from lectures. Five subjects are 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 are removed because their topics are not appropriate for the SDR task, and some are revised their expression to reduce their ambiguity. Finally, we obtained 86 query topics.

A query topic is represented by a natural language sentence. The format of a query topic list is as follows.

```
TERM-ID question
```

6.3 Submission

Each participant is allowed to submit as many search results ("runs") as they want. Submitted runs should be prioritized by each group, because the specific number of runs with higher priority will be used for the pooling data for the manual relevance judgment. Priority number should be assigned through all submissions of a participant, and smaller number has higher priority.

File Name

A single run is saved in a single file. Each submission file should have an adequate file name following the next format.

SDR-*X*-*T*-*N*.txt

X: System identifier that is the same as the group ID (e.g., NTC)

T: Target task

- LEC: Lecture retrieval task.
- PAS: Passage retrieval task.

N: Priority of run (1, 2, 3, ...) for each target document set.

For example, if the group "NTC" submits two files for targeting lecture retrieval task and three files for passage retrieval task, the names of the run files should be "SDR-NTC-LEC-1.txt", "SDR-NTC-LEC-2.txt", "SDR-NTC-PAS-1.txt", "SDR-NTC-PAS-2.txt", and "SDR-NTC-PAS-3.txt".

Table 4: STD evaluation results on each measurement for all submitted runs of the CORE set. "Search time" shows the average time for finishing search process for each query. * mark indicates the organizers' team.

Runs	F-measure (max) [%]	F-measure (spec.) [%]	MAP	Search time [sec.]	Machine specifications
Baseline	0.527	0.516	0.595	36.4	Core i7 975 3.33GHz, 4 core CPU, 8GB memory
AKBL-1*	0.393	0.393	0.264	0.0017	Xeon X5560 2.67GHz, 6 core x 2 CPUs, 24GB memory
AKBL-2*	0.385	0.370	0.272	0.0013	Xeon X5560 2.67GHz, 6 core x 2 CPUs, 24GB memory
ALPS-1*	0.725	0.708	0.837	13.50	Core i7 975 3.33GHz, 4 core CPU, 6GB memory
ALPS-2*	0.714	0.697	0.757	13.44	Core i7 975 3.33GHz, 4 core CPU, 6GB memory
IWAPU-1	0.644	0.628	0.772	3.5	Xeon 2.53GHz quad-core 2 pieces, 12GB memory
IWAPU-2	0.510	0.297	0.733	3.5	Xeon 2.53GHz quad-core 2 pieces, 12GB memory
NKGW-1	0.645	0.585	0.491	0.0016	Xeon 2.93GHz 24core CPU, 74GB memory
NKI11-1	0.570	0.559	0.684	0.00094	Core i7-2600 3.4GHz, 8GB memory
NKI11-2	0.569	0.556	0.672	0.00094	Core i7-2600 3.4GHz, 8GB memory
RYSDT-1	0.318	0.152	0.393	4.12	Xeon 3.20GHz 8core CPU, 4GB memory
RYSDT-2	0.526	0.287	0.468	2.96	Xeon 3.20GHz 8core CPU, 4GB memory
RYSDT-3	0.521	0.334	0.469	2.90	Xeon 3.20GHz 8core CPU, 4GB memory
YLAB-1	0.425	0.425	0.344	38.8	Core2Quad Q9650 3.0GHz, memory 4GB

Table 5: STD evaluation results on each measurement for all submitted runs of the ALL set. "Search time" shows the average time for finishing search process for each query.

Runs	F-measure (max) [%]	F-measure (spec.) [%]	MAP	Search time [sec.]	Machine specifications
Baseline	0.459	0.310	0.451	548	Core i7 975 3.33GHz, 4 core CPU, 8GB memory
NKI11-1	0.367	0.360	0.339	0.0031	Core i7-2600 3.4GHz, 8GB memory
NKI11-2	0.396	0.332	0.344	0.0031	Core i7-2600 3.4GHz, 8GB memory
RYSDT-1	0.531	0.070	0.431	53.16	Xeon 3.20GHz 8core CPU, 4GB memory
RYSDT-2	0.530	0.082	0.426	48.32	Xeon 3.20GHz 8core CPU, 4GB memory
RYSDT-3	0.531	0.119	0.434	48.30	Xeon 3.20GHz 8core CPU, 4GB memory

Submission Format

The submission files are organized with the following tags. Each file must be a well-formed XML document. It has a single root level tag "<**ROOT**>". Under the root tag, it has three main sections, "<**RUN**>", "<**SYSTEM**>", and "<**RESULTS**>".

- <**RUN**>

  <**SUBTASK**> "STD" or "SDR". For a SDR subtask submission, it must be "SDR".

  <**SYSTEM-ID**> System identifier that is the same as the group ID.

  <**PRIORITY**> Priority of the run.

  <**UNIT**> The retrieval unit to be retrieved. "LECTURE" if the unit is a lecture, or the sub-subtask is the lecture retrieval. "PASSAGE" if the unit is a passage, or the sub-subtask is the passage retrieval.

  <**TRANSCRIPTION**> The transcription used as the text representation of the target document set. "MANUAL" if it is the manual transcription provided by the CSJ. "REF-WORD" if it is the reference word-based automatic transcription provided by the organizers. "REF-SYLLABLE" if it is the reference syllable-based automatic transcription provided by the organizers. "OWN" if it is obtained by a participant's own recognition. "NO" if no textual transcription is used.

- <**SYSTEM**>

  <**OFFLINE-MACHINE-SPEC**>
  <**OFFLINE-TIME**>
  <**INDEX-SIZE**>
  <**ONLINE-MACHINE-SPEC**>
  <**ONLINE-TIME**>
  <**SYSTEM-DESCRIPTION**>

- <**RESULTS**>

  <**QUERY-ID**> Each query topic has a single "QUERY-ID" tag with an attribute "id" specified in a query topic list. Within this tag, a list of the following "DOCUMENT" tags is described.

  <**CANDIDATE**> Each potential candidate of a retrieval result has a single "CANDIDATE" tag with the following attributes. The CANDIDATE tags must be sorted in descending order of likelihood.

  document The searched document (lecture) ID specified in the CSJ.

  ipu-from *Used only for the passage retrieval task.* The Inter Pausal Unit ID, specified in the CSJ, of the first IPU of the retrieved passage (an IPU sequence).

  ipu-to *Used only for the passage retrieval task.* The Inter Pausal Unit ID, specified in the

Table 6: System information related to the offline processing for those runs using indexing method.

Set	Runs	Offline Time [sec.]	Index Size [K byte]	Machine specifications
CORE	AKBL-1*	1147.716	3400000.00	Xeon X5560 2.67GHz, 6 core x 2 CPUs, 24GB memory
	AKBL-2*	692.875	3400000.00	Xeon X5560 2.67GHz, 6 core x 2 CPUs, 24GB memory
	NKGW-1	1420.700	3590000.00	Xeon 2.93GHz 24core CPU, 74GB memory
	NKI11-1	626.497	5715.55	Core i7-2600 3.4GHz, 8GB memory
	NKI11-2	626.497	5715.55	Core i7-2600 3.4GHz, 8GB memory
ALL	NKI11-1	9009.770	83570.50	Core i7-2600 3.4GHz, 8GB memory
	NKI11-2	9009.770	83570.50	Core i7-2600 3.4GHz, 8GB memory

```
<ROOT>
<RUN>
<SUBTASK>SDR</SUBTASK>
<SYSTEM-ID>TUT</SYSTEM-ID>
<PRIORITY>1</PRIORITY>
<UNIT>PASSAGE</UNIT>
<TRANSCRIPTION>REF-SYLLABLE</TRANSCRIPTION>
</RUN>
<SYSTEM>
<OFFLINE-MACHINE-SPEC>Xeon 3GHz dual CPU, 4GB memory
</OFFLINE-MACHINE-SPEC>
<OFFLINE-TIME>18:35:23</OFFLINE-TIME>
...
</SYSTEM>
<RESULTS>
<QUERY id="SpokenDoc1-SDR-dry-001">
<CANDIDATE document="A01F0005" ipu-from="0024"
 ipu-to="0027" />
<CANDIDATE document="S00M0075" ipu-from="0079"
 ipu-to="0079" />
...
</QUERY>
<QUERY id="SpokenDoc1-SDR-dry-002">
...
</QUERY>
</RESULTS>
</ROOT>
```

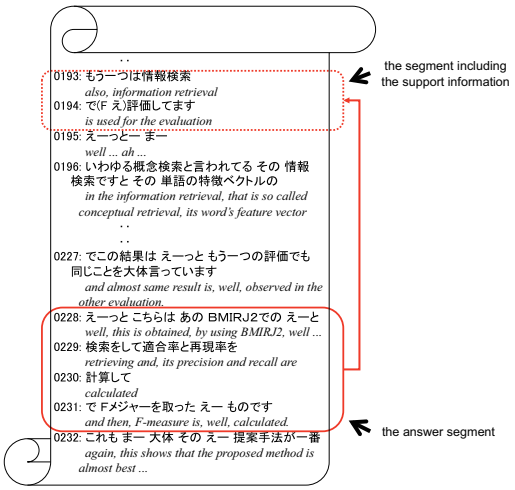
Figure 4: An example of a submission file.

CSJ, of the last IPU of the retrieved passage (an IPU sequence).

Figure 4 shows an example of a submission file.

6.4 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 in itself to the query, when it had no supporting information that would convince the user who submitted the query of the correctness of the answer. For example, for 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 5 shows an example of an answer and its supporting information for the query "How can we evaluate the performance of information retrieval?"



. .
0193: もう一つは情報検索
also, information retrieval
0194: で(F え)評価してます
is used for the evaluation
0195: えーっとー まー
well ... ah ...
0196: いわゆる概念検索と言われてる その 情報検索ですと その 単語の特徴ベクトルの
in the information retrieval, that is so called conceptual retrieval, its word's feature vector
. .
. .
0227: でこの結果は えーっと もう一つの評価でも同じことを大体言っています
and almost same result is, well, observed in the other evaluation.
0228: えーっと こちらは あの BMIRJ2での えーと
well, this is obtained, by using BMIRJ2, well ...
0229: 検索をして適合率と再現率を
retrieving and, its precision and recall are
0230: 計算して
calculated
0231: で Fメジャーを取った えー ものです
and then, F-measure is, well, calculated.
0232: これも まー 大体 その えー 提案手法が一番
again, this shows that the proposed method is almost best ...

Figure 5: An example of the answer and the supporting segment.

As shown in Figure 5, 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, searched its relevant passages and judged their degrees of relevancy. The assessor labeled them into three classes according to the degree of their relevancy: "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 engine against the manual transcription of the target document collection, are checked by the assessor.

6.5 Evaluation Measures

6.5.1 Lecture Retrieval

Mean Average Precision (MAP) is used for our official evaluation measure for lecture retrieval. For each query topic, top 1000 documents are 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.

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

where

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

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.

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

where $rank(r)$ is the rank that the document r is retrieved.

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

$$\mathbf{MAP} = \frac{1}{|Q|} \sum_{q \in Q} AveP_q \tag{9}$$

6.5.2 Passage Retrieval

In our passage retrieval task, the relevancy of each arbitrary length segment (passage) rather than each whole lecture (document) must be evaluated. Three measures are designed for the task; the one is utterance-based and the other two are passage-based. For each query topic, top 1000 passages are evaluated by these measures.

6.5.3 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 given 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 get 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.

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

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 the oracle ordering mapping function, which orders the utterances in the given passage p as the relevant utterances come first. For example, given a passage $p = u_1 u_2 u_3 u_4 u_5$ and suppose the relevant utterances are $u_3 u_4$, it returns as $u_3 u_4 u_1 u_2 u_5$.

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

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

6.5.4 Passage-based Measure

Our passage retrieval needs two tasks to be achieved; one is to determine the boundary of the passages to be retrieved and the other is to rank the relevancy of the passages. The first passage-based measure focuses only on the latter task and the second measure focuses both of the 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 relevancy. If the center utterance is included in some relevant passage described in the golden file, basically the returned passage is deemed relevant with respect to the relevant passage 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 relevancy. 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 $r_1 r_2 \cdots r_M (M \leq |R_q|)$, where their relevancy 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 that the passage r is placed at 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 relevancy 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 that it covers some relevant passage(s), 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 in 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.

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

Table 7: SDR subtask participants. * mark indicates the organizers' team.

Lecture retrieval task		
Team ID	Team name	Organization
AKBL*	Akiba Laboratory	Toyohashi University of Technology
ASR	team ASR	Gifu University
RYSDT	Ryukoku NL-SLP lab	Ryukoku University
TBFD	Team Big Four Dragons	Daido University, The University of Tokushima, Nagoya University
Passage retrieval task		
Team ID	Team name	Organization
AKBL*	Akiba Laboratory	Toyohashi University of Technology
DCU	DCU	Dublin City University
RYSDT	Ryukoku NL-SLP lab	Ryukoku University

Table 8: Summary of the transcriptions used for each run.

task	group	run	transcription	candidate(s)
lecture	AKBL	1	REF-WORD	1-best
		2		
		3	REF-SYLLABLE	
	ASR	1	REF-WORD	1-best
		2		
		3	MANUAL	
	RYSDT	1	OWN	1-best
	TBFD	1	REF-WORD	10-best
		2	& REF-SYLLABLE	
		3		
passage	DCU	1	MANUAL	1-best
		2	REF-WORD	
		3	MANUAL	
		4	REF-WORD	
		5		
		6		
	RYSDT	1	OWN	1-best
	AKBL	1	REF-WORD	1-best
		2		
		3	REF-SYLLABLE	

6.6 Evaluation Results

Five groups with total 21 runs have submitted the results for the formal run. Among them, four groups participated the lecture retrieval task and three groups participated the passage retrieval task. The term IDs are listed in Table 7.

6.6.1 Transcriptions

All participants used textual transcription, to which some retrieval method was applied. One participant group used their own transcription, while the other used the transcriptions provided by the organizers. Among 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, one for lecture and one for passage retrieval, by one group 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. Table 8 summarizes the transcriptions used for each run.

6.6.2 Baseline Methods

We implemented and evaluated the baseline methods for our SDR tasks, which consisted of only conventional methods for IR and applied to the 1-best REF-WORD or MANUAL transcription. Only nouns were used for indexing, which were extracted from the transcription by applying the Japanese morphological analysis tool. The vector space model was used as the retrieval model, and TF–IDF (Term Frequency–Inverse Document Frequency) with pivoted normalization [12] was used for term weighting. We used *GETA* [3] as the IR engine for the baselines.

For the lecture retrieval task, each lectures in the CSJ is 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 the rough estimate of the passage lengths of the dry run test data.

Figure 6 shows the average precisions of the baseline lecture retrieval system for each query. It indicates that the variance in the difficulties among queries is high. It also indicates that the variance in the performance difference between using manual and automatic transcription is also high; for some queries, the retrieval on the manual transcription was perfect, while that on the automatic transcription did not work at all. It suggests that how to deal 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, is one of the main challenges of SDR, though the OOV rate on the formal run queries are not high, where only three queries includes the OOV words against the REF-WORD transcription.

6.6.3 SDR Techniques Used

Here, we provide a brief overview of SDR techniques used by the participants. For more details, please refer to the participant's papers.

AKBL [5] submitted four runs for the lecture retrieval task and three runs for the passage retrieval task. Two approaches, word-based and STD-based approaches, were investigated for both the tasks. The word-based approach applied the conventional word-based IR models against the REF-WORD transcription, among which some runs applied query expansion technique based on the relevance models. The STD-based approach applied STD on the REF-SYLLABLE transcription as the preprocessing and the detection results were used as the term appearances in the following document retrieval process. For the passage retrieval task, same

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

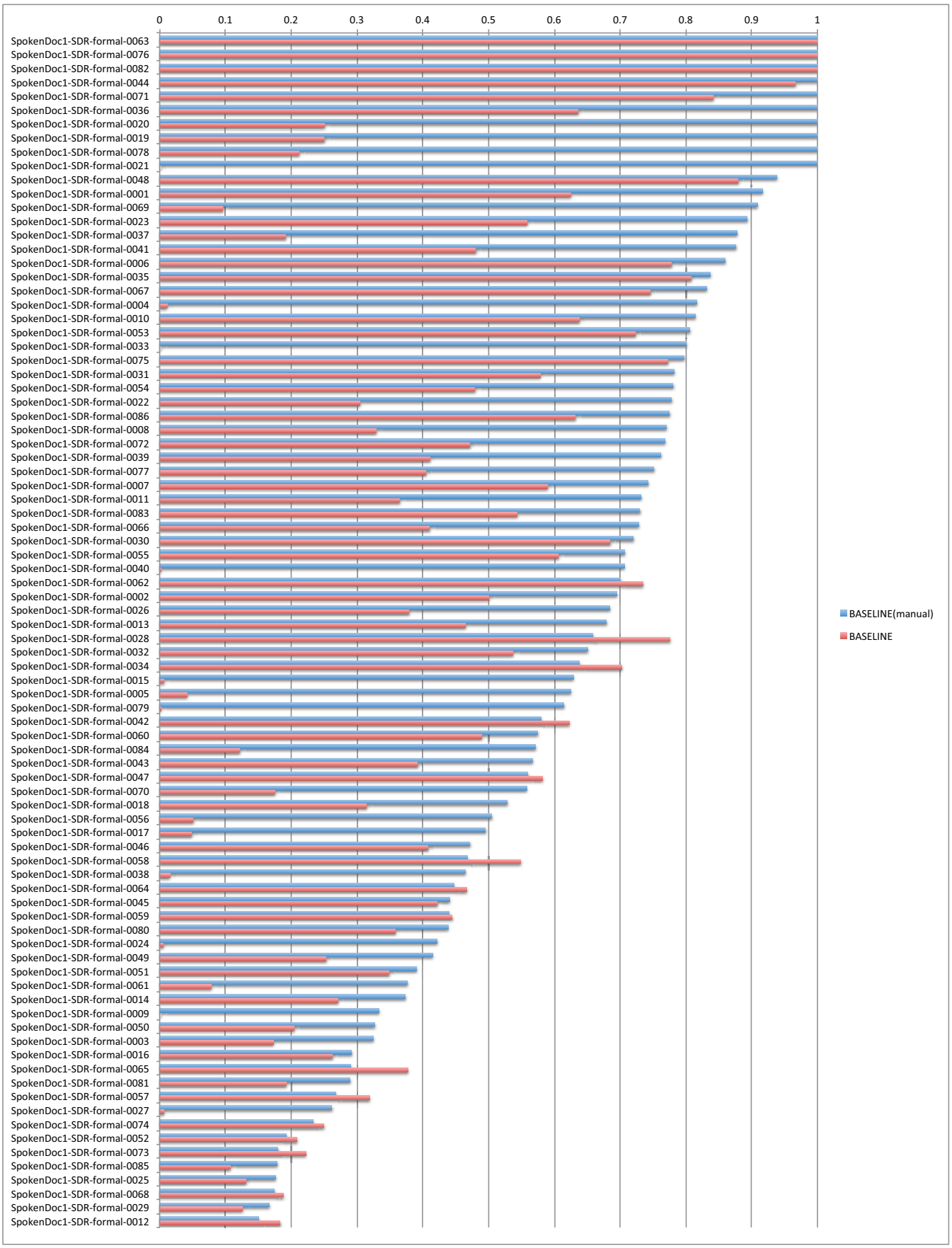


Figure 6: Average precisions of BASELINE for each query.

Table 9: Evaluation results for the lecture retrieval task.

Group ID	run	MAP	
baseline		0.393	
		0.624	(manual)
AKBL	1	0.426	
	2	0.260	
	3	0.085	
	4	0.252	
ASR	1	0.319	
	2	0.204	
	3	0.458	(manual)
RYSDT	1	0.539	
TBFD	1	0.427	
	2	0.405	
	3	0.406	

(manual) indicates that the run uses the manual transcription.

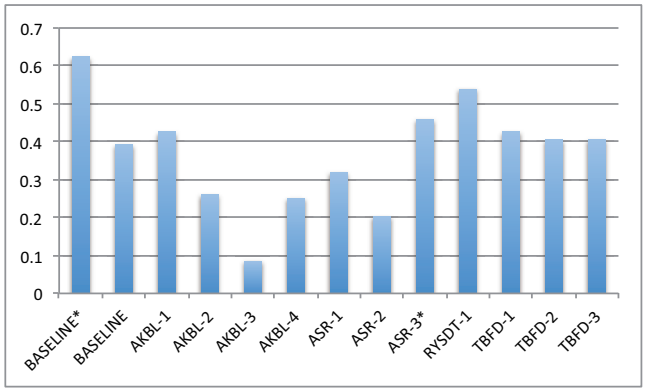


Figure 7: Evaluation results for the lecture retrieval task. * mark indicates the run uses the manual transcription.

techniques were applied to the fixed length sliding windows, which were deemed as passages. The passage retrieval results were further repressed to reduce the redundant results.

ASR [3] submitted three runs for the lecture retrieval task. The used IR method was word-based vector space model. Both document and query expansion techniques were applied at the same time using the Web search engine. The expanded documents and the expanded query were compared on the vector space with TF-IDF term weighting. The IR method that combined the results by Boolean AND retrieval and those by the cosine similarity was also introduced.

DCU [2] submitted six runs for the passage retrieval task. They were only group among the task participants who applied the segmentation methods to determine the extent of the retrieved passage. Two segmentation algorithms, TextTiling and C99, were investigated and compared to see which one was effective for the passage retrieval. The retrieval model was word-based language modeling methods for IR.

RYSDT [9] submitted one run for the lecture retrieval task and one run for the passage retrieval task. Both the nouns and verb's base form were used as their indexing unit. The retrieval model was the word-based vector space model using TF-IDF term weighting with pivoted normalization, which was same as the baseline method. For the passage retrieval task, the pre-defined fixed length segments was used as the passage, which was also same as the baseline method.

Table 10: Evaluation results for the passage retrieval task.

Group ID	run	uMAP	pwMAP	fMAP	
baseline		0.0671	0.0520	0.0536	
		0.1179	0.0915	0.0965	(manual)
AKBL	1	0.0747	0.1581	0.0686	
	2	0.0756	0.1440	0.0672	
	3	0.0323	0.0283	0.0262	
DCU	1	0.0859	0.0429	0.0500	(manual)
	2	0.0491	0.0329	0.0308	
	3	0.0713	0.0209	0.0168	(manual)
	4	0.0469	0.0166	0.0123	
	5	0.0316	0.0138	0.0120	
	6	0.0313	0.0141	0.0174	
RYSDT	1	0.0751	0.0725	0.0650	

(manual) indicates that the run uses the manual transcription.

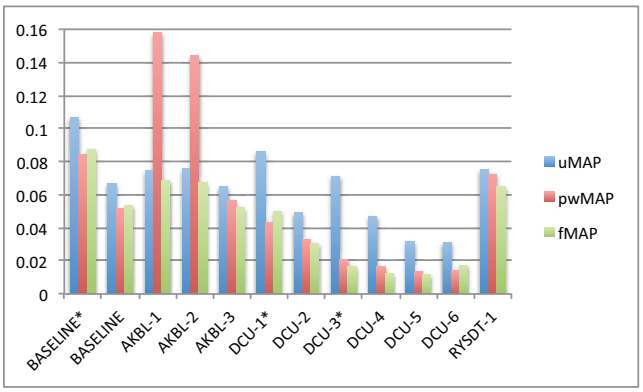


Figure 8: Evaluation results for the passage retrieval task. * mark indicates the run uses the manual transcription.

TBFD [13] submitted three runs for the lecture retrieval task. Both the word indices obtained from the REF-WORD transcription and the syllable n-gram indices obtained from the REF-SYLLABLE transcription were used. Query expansion technique was applied, where the vector of the query terms and the vector from the search results against Wikipedia articles were interpolated to form the expanded vector. Some of their runs determined the interpolation weight dynamically, while the others used static weight. They were only group among the participants of SDR subtask who used the multiple candidates from speech recognition. Each candidate was weighted by its reciprocal rank to reward more the better-ranked one.

6.6.4 Results

For the lecture retrieval task, the evaluation results of all the submissions are summarized in Table 9 and Figure 7. 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 weekly (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 same as the baseline, its use of the own transcription seemed to be the major factor of the improvement.

For the passage retrieval task, the three evaluation measures, uMAP, pwMAP, and fMAP are used. The correlation coefficients between uMAP and pwMAP, between pwMAP and fMAP, and between uMAP and fMAP, calculated by using all the submitted runs, are 0.750, 0.869, 0.884, respectively. It shows that these measures correlate each other well, and that those measuring the same aspects (i.e. uMAP and fMAP, which measures both the accuracy of the boundaries and the relevancy, while the 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 other (i.e. uMAP and fMAP).

The evaluation results are summarized in Table 9 and 10. It was also observed in the passage results that the runs using manual transcription outperformed their counterparts using automatic transcriptions. We will investigate only the results using automatic transcription below.

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. Especially, the pwMAP values of AKBL-1 and AKBL-2 were surpassed among all the runs including those using manual transcription. It seemed because of their methods for reducing the redundant results, which worked effectively especially for the pwMAP measure. In terms of fMAP, AKBL-1, AKBL-2 and RYSDT-1 also outperformed the baseline significantly but weekly (at 0.05 level for the two-sided paired t-test).

7. CONCLUSION

This paper introduced the overview of the IR for Spoken Documents (SpokenDoc) task in NTCIR-9 Workshop. Our task had the spoken term detection (STD) subtask and ad-hoc spoken document retrieval subtask (SDR). Both of the subtasks targeted to search terms, passages and documents included in academic and simulated lectures of the Corpus of Spontaneous Japanese. Finally, seven and five teams participated in the STD subtask and the SDR subtask, respectively.

This paper described the detail task definitions of each subtask, and introduced the outlines of STD and SDR methods of each participant. In addition to this, all the participants' STD and SDR performances were shown.

8. REFERENCES

[1] T. Akiba, K. Aikawa, Y. Itoh, T. Kawahara, H. Nanjo, H. Nishizaki, N. Yasuda, Y. Yamashita, and K. Itou. 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, 2009.

[2] M. Eskevich and G. J. F. Jones. DCU at the NTCIR-9 SpokenDoc passage retrieval task. In *Proceedings of the Ninth NTCIR Workshop Meeting*, 2011.

[3] K. Hasegawa, H. Sekiya, M. Takehara, T. Niinomi, S. Tamura, and S. Hayamizu. Toward improvement of SDR accuracy using LDA and query expansion for SpokenDoc. In *Proceedings of the Ninth NTCIR Workshop Meeting*, 2011.

[4] K. Iwami and S. Nakagawa. 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*, 2011.

[5] T. Kaneko, T. Takigami, and T. Akiba. STD based on hough transform and SDR using STD results: Experiments at NTCIR-9 SpokenDoc. In *Proceedings of the Ninth NTCIR Workshop Meeting*, 2011.

[6] K. Katsurada, K. Katsuura, Y. Iribe, and T. Nitta. Utilization of suffix array for quick STD and its evaluation on the NTCIR-9 SpokenDoc task. In *Proceedings of the Ninth NTCIR Workshop Meeting*, 2011.

[7] A. Lee and T. Kawahara. 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, 2009.

[8] K. Maekawa, H. Koiso, S. Furui, and H. Isahara. Spontaneous speech corpus of Japanese. In *Proceedings of LREC*, pages 947–952, 2000.

[9] H. Nanjo, K. Noritake, and T. Yoshimi. Spoken document retrieval experiments for spokendoc at ryukoku university (RYSDT). In *Proceedings of the Ninth NTCIR Workshop Meeting*, 2011.

[10] H. Nishizaki, H. Furuya, S. Natori, and Y. Sekiguchi. Spoken term detection using multiple speech recognizers' outputs at NTCIR-9 SpokenDoc STD subtask. In *Proceedings of the Ninth NTCIR Workshop Meeting*, 2011.

[11] H. Saito, T. Nakano, S. Narumi, T. Chiba, K. Kon'No, and Y. Itoh. An STD system for OOV query terms using various subword units. In *Proceedings of the Ninth NTCIR Workshop Meeting*, 2011.

[12] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of ACM SIGIR*, pages 21–29, 1996.

[13] S. Tsuge, H. Ohashi, N. Kitaoka, K. Takeda, and K. Kita. Spoken document retrieval method combining query expansion with continuous syllable recognition for NTCIR-SpokenDoc. In *Proceedings of the Ninth NTCIR Workshop Meeting*, 2011.

[14] Y. Yamashita, T. Matsunaga, and K. Cho. YLAB@RU at spoken term detection task in NTCIR9-SpokenDoc. In *Proceedings of the Ninth NTCIR Workshop Meeting*, 2011.