Overview of the NTCIR-10 SpokenDoc-2 Task

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

Xinhui Hu
National Institute of Information and Communications Technology

Yoshiaki Itoh
Iwate Prefectural University
Sugo 152-52, Takizawa, Iwate, Japan

Tatsuya Kawahara
Kyoto University
Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8501, Japan

Seiichi Nakagawa
Toyohashi University of Technology
1-1 Hibarigaoka, Tohohashi-shi, Aichi, 440-8580, Japan

Hiroaki Nanjo
Ryukoku University
Yokotani 1-5, Oe-cho Seta, Otsu, Shiga, 520-2194, Japan

Yoichi Yamashita
Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu-shi, Shiga, 525-8577, Japan

ABSTRACT

This paper describes an overview of the IR for Spoken Documents Task in NTCIR-10 Workshop. In this task, the spoken term detection (STD) subtask and ad-hoc spoken content retrieval subtask (SCR) are conducted. Both of the tasks target to search terms, passages and documents included in academic oral presentations. This paper explains the data used in the tasks, how to make transcriptions by speech recognition and the details of each tasks.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Performance

Keywords

NTCIR-10, 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. Following the NTCIR-9 SpokenDoc task[1, 2], we evaluated the SDR based on a realistic ASR condition, where the target documents were spontaneous speech data with high word error rate and high out-of-vocabulary rate.

In the NTCIR-10 SpokenDoc-2 task, 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). In addition, an "inexistent Spoken Term Detection (iSTD) task" was also conducted. In the iSTD task, task participants inspect whether a queried term is existent or inexistent in a speech data collection.

Spoken Content Retrieval: Among spoken documents, find the segments including the relevant information related to the query, where a segment is either a document (resulting in document retrieval task) or a passage (passage retrieval task). This is like an ad-hoc text retrieval task, except that the target documents are speech data.

2. DOCUMENT COLLECTION

Two document collections are used for the SpokenDoc-2.

Corpus of Spontaneous Japanese (CSJ) It is released by the National Institute for Japanese Language[4]. Among CSJ, 2,702 lectures (602 hours) are used as the target documents for SpokenDoc-2. In order to participate in the subtask targetting the CSJ, the participants are required to purchase the data by themselves.

Corpus of Spoken Document Processing Workshop (SDPWS) It is released by the SpokenDoc-2 task organisers. It consists of the recordings of the first to sixth annual Spoken Document Processing Workshop, 104 oral presentations (28.6 hours).

Each lecture in the CSJ and the SDPWS 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 SCR tasks.

3. TRANSCRIPTION

Standard 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 can use the following three types of transcriptions.

1. Manual transcription

   It is mainly used for evaluating the upper-bound performance.

2. Reference automatic transcriptions

   The organizers prepared four reference automatic transcriptions for each collection. It enables that those who are interested in SDR but not in ASR can participate in our tasks. It also enables the comparison of the IR methods based on the same underlying ASR performances. The participants can also use multiple 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
   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.

   (b) Syllable-based transcription
   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 SCR can use this transcription.

   Two different kinds of language models are used to obtain these transcriptions; one of them is trained by matched lecture text and the other is by unmatched newspaper articles. Thus, there are four transcriptions for each collection: word-based with high WER, word-based with low WER, syllable-based with high WER, and syllable-based with low WER.

3. 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

4.1 Models for transcribing the CSJ

To realize open speech recognition, we used the following acoustic and language models, which were trained by using the CSJ 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.

We trained two kinds of language models. One of them were word-based trigram models with a vocabulary of 27k words and were used to make the word-based transcriptions. The others were syllable-based trigram models, which were trained by the syllable sequences of each training group, and were used to make the syllable-based transcriptions.

We used Julius [3] 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

In addition to the language models described above, which are referred to as *matched* models, we also prepared the *unmatched* language models, which are trained by the newspaper articles. They are also divided into the word-based tri-gram model and the syllable-based tri-gram model. The word-based model is the one provided from the Continuous Speech Recognition Consortium (CSRC), whose vocabulary size is 20k words. The syllable-based model was trained by the syllable sequence of the same newspaper articles as the word-based model. The transcriptions obtained by using these language models are called *unmatched* transcriptions.

4.2 Models for transcribing the SDPWS

The acoustic model for recognizing SDPWS data is same as those for the CSJ data, described in the last subsection, except that all the lecture data is used all together

for training it. The two *matched* language models, which are word-based tri-gram model and syllable-based tri-gram model, are also trained by using all the lecture transcriptions in the CSJ at the same time, while the two *unmatched* language models are identical to the *unmatched* word-based and syllable-based models for recognizing the CSJ.

4.3 ASR performance for each ASR model

Finally we provided four sorts of transcriptions for each the speech documents collections to the task participants as follows:

REF-WORD-MATCHED was produced by the ASR with the word-based trigram LM trained from CSJ

REF-SYLLABLE-MATCHED was produced by the ASR with the syllable-based trigram LM trained from CSJ syllable-represented

REF-WORD-UNMATCHED was produce by the ASR with the word-based trigram LM trained from the newspaper articles

REF-SYLLABLE-UNMATCHED was produce by the ASR with the syllable-based trigram LM trained from the newspaper articles syllable-represented

The AM described on Sec. 4.1 was commonly used for transcribing speeches.

Table 1 shows the ASR performances of the CSJ and SDPWS speech transcriptions. The performance measures are word (syllable)-based correct rate and accuracy rate.

5. SPOKEN TERM DETECTION TASK

5.1 Task Definition

Our STD task is to find all IPUs which include a specified query term in the CSJ or SDPWS. For the STD task, a term is a sequence of one or more words. This is different from the STD task produced by NIST [1]

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 Query Set

The STD task consists of two sub-tasks: **the large-size task** on CSJ and **the moderate-size task** on SDPWS. Therefore, the organizers provided two sets of the query term list, i.e. the list for CSJ lectures and the list for the SDPWS oral presentations. Each participant's submission (called "run") should choose one from the two according to their target document collection, i.e. either **CSJ** or **SDPWS**.

The format of a query term list for the large size task is as follows.

```
TERM-ID term Japanese_katakana_sequence
```

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

An example list is:

```
SpokenDoc2-STD-formal-SDPWS-001 アーティキュレーション アーティキュレーション
SpokenDoc2-STD-formal-SDPWS-002 ＩＢＭ アイビーエム
SpokenDoc2-STD-formal-SDPWS-003 アカデミックハラスメント アカデミックハラスメント
SpokenDoc2-STD-formal-SDPWS-004 Ａｄａｂｏｏｓｔ アダブースト
...
```

Here, the "Japanese kantakana sequence" is an optional information. This means a Japanese pronunciation of a term. Though the organizers do **not** assure the participants of its correctness, it may be helpful to predict the term's pronunciation. Notice that, for the judgment of the term's occurrence in the golden file, the "`term`" is searched against the manual transcriptions; i.e. the "`Japanese_katakana_sequence`" is never considered for the judgment.

We prepared the 100 query terms for each STD sub-task. For the large-size task, 54 of the all 100 query terms are OOV queries that are not included in the ASR dictionary of the MATCHED-conditioned word-based LM and the others are IV queries. On the other hand, for the moderate-size task, 53 of the all 100 query temrs are OOV queries. The average occurrences per a term is 18.0 times and 9.4 times for the large-size task and the moderate-size, 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 3 to 18 morae.

5.3 System Output

When a term is supllied 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

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.

5.4.1 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)

Table 1: ASR performances [%].

(a) For the CSJ speeches.

transcriptions	Word Corr.	Word Acc.	Syll. Corr.	Syll. Acc.
REF-WORD-MATCHED	74.1	69.2	83.0	78.1
REF-WORD-UNMATCHED	59.5	55.7	80.6	77.1
REF-SYLLABLE-MATCHED	—	—	80.5	73.3
REF-SYLLABLE-UNMATCHED	—	—	75.5	71.4

(b) For the SDPWS lectures.

transcriptions	Word Corr.	Word Acc.	Syll. Corr.	Syll. Acc.
REF-WORD-MATCHED	68.4	63.1	79.7	75.3
REF-WORD-UNMATCHED	48.4	43.7	67.8	62.8
REF-SYLLABLE-MATCHED	—	—	72.7	67.7
REF-SYLLABLE-UNMATCHED	—	—	60.3	55.2

D: Target document set:

- CSJ: 2,702 lectures from the CSJ.
- SDPWS: 104 oral presentations from the SDPWS.

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

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

5.4.2 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 "<**RESULT**>".

- <**RUN**>

  <**SUBTASK**> "STD" or "SCR". For a STD subtask submission, just say "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. "CSJ" if the target document set is the CSJ lectures. "SDPWS" if SDPWS lectures.

  <**TRANSCRIPTION**> The transcription used as the text representation of the target document set. "MANUAL" if it is the manual transcription. "REF-WORD-MATCHED" if it is the reference word-based automatic transcription obtained by using the matched-condition language model. "REF-WORD-UNMATCHED" if it is the reference word-based automatic transcription obtained by using the unmatched-condition language model. "REF-SYLLABLE-MATCHED" if it is the reference syllable-based automatic transcription obtained by using the matched-condition language model. "REF-SYLLABLE-UNMATCHED" if it is the reference syllable-based automatic transcription obtained by using the unmatched-condition language model. Note that these four transcriptions are provided by the organizers. "OWN" if it is obtained by a participant's own recognition. "NO" if no textual transcription is used. If multiple transcriptions are used, specify all of them by concatenating with the "," separator.

- <**SYSTEM**>

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

- <**RESULT**>

  <**QUERY**> Each query term has a single "QUERY" tag with an attribute "id" specified in a query term list (Section 5.2). 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 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 1 shows an example of a submission file.

5.5 Evaluation Measures

The official evaluation measure for effectiveness is F-measure at the decision point specified by the participant, based on recall and precision micro-averaged over the queries. F-measure at the maximum decision point also used for evaluation. In addition, F-measures based on macro-averaged over the queries and mean average precision (MAP) will also be used for analysis purpose.

```
<ROOT>
<RUN>
<SUBTASK>STD</SUBTASK>
<SYSTEM-ID>TUT</SYSTEM-ID>
<PRIORITY>1</PRIORITY>
<TARGET>CSJ</TARGET>
<TRANSCRIPTION>REF-WORD-UNMATCHED,
REF-SYLLABLE-UNMATCHED</TRANSCRIPTION>
</RUN>
<SYSTEM>
<OFFLINE-MACHINE-SPEC>Xeon 3GHz dual CPU, 4GB memory
</OFFLINE-MACHINE-SPEC>
<OFFLINE-TIME>18:35:23</OFFLINE-TIME>
...
</SYSTEM>
<RESULT>
<QUERY id="SpokenDoc2-STD-formal-CSJ-001">
<TERM document="A01F0005" ipu="0024" score="0.83"
 detection="YES" />
<TERM document="S00M0075" ipu="0079" score="0.32"
 detection="NO" />
...
</QUERY>
<QUERY id="SpokenDoc2-STD-formal-CSJ-002">
...
</QUERY>
</RESULT>
</ROOT>
```

Figure 1: An example of a submission file.

Mean average precision for the set of queries is the mean value of the average precision values for each query. It can be calculate as follows:

$$MAP = \frac{1}{Q}\sum_{i=1}^{Q} AveP(i) \qquad (1)$$

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.

$$AveP(i) = \frac{1}{Rel_i}\sum_{r=1}^{N_i}(\delta_r \cdot Precision_i(r)) \qquad (2)$$

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.

5.6 Evaluation Results

5.6.1 STD task participants

the eight teams participated in the STD tasks with 48 submisison runs. In addition, the six runs as the baseline results were submitted by the organizers. The team IDs are listed in Table 2. Five teams submitted the results for the large-size task and all teams submitted the results for the moderate-size task.

5.6.2 STD task results

First of all, Table 3 summarizes the number of transcription(s) used for each run. And the evaluation results are summarized in Table 4 for the large-size task with the 21 submitted runs and the baseline (three runs). Table 5 also shows the STD performance for the moderate-size task of the 27 submitted runs and the baseline (three runs). These tables represent the F-measures at the maximum point and specified decision point by the participant, based on both of micro-averaged and macro-averaged, and MAP values. And, the index size (memory consumption) and search speed by one query are also shown in these tables.

The baseline systems (BL-1, BL-2, and BL-3) 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 BL-1 was made of the transcriptions of REF-SYLLABLE-MATCHED. The index for the BL-2 was made of REF-WORD-MATCHED. The two indeces from the transcriptions of REF-SYLLABLE-MATCHED and REF-WORD-MATCHED were used in BL-3. In BL-3, the search engine searches a query term from the index of REF-SYLLABLE-MATCHED if the term is OOV. The decision point for calculating F-measure (spec.) was decided by the result of the NTCIR-9 formal-run query set[1]. We adjusted the threshold to be the best F-measure value on the formal-run set, which was used as a development set.

In the large-size task, runs that use only the single transcriptions “REF-SYLLABLE-MATCHED” got worse performance compared to the runs with “REF-WORD-MATCHED”. For example, “BL-1”, “NKI13-7”, “akbl-1,2,3”and “TBFD-4” did not outperform the “BL-2” that used only “REF-WORD-MATCHED.” The IV query terms can be efficiently detected from the index made of the word-based transcription. On the other hand, in the case of the OOV query term detection, the index made of the transcription produced by using the syllable-based LM worked well. Therefore, “BL-3” was better than “BL-2”.

“NKI13-1”, which got the best performance among the runs by team NKI-13, used the two transcriptions: REF-WORD-UNMATCHED and REF-SYLLABLE-UNMATCHED. The difference between “NKI13-1” and “NKI13-2” is the transcriptions. “NKI13-2” used REF-WORD-MATCHED and REF-SYLLABLE-MATCHED which were produced by the match-conditioned LMs. In addition, “TBFD-1,2,3”, output the high performance STD, also used the transcriptions made by the unmatch-conditioned LMs. “NKI13-1” and “TBFD-1,2,3” outperformed “ALPS-1” used the 10 sorts of transcription made by match-conditioned models. It is interesting because it is generally considered that match-conditioned models conduce to better STD performance. This is the opposite, however, the ASR performance between the transcriptions by the matched and unmatched model is not major difference.

The best STD performance was “TBFD-9” which used the OWN transcriptions, but it was not speech recognition result.

On the other hand, for the moderate-size task, “ALPS-1” and “IWAPU-1” got the best performance at the F-measure and MAP, respectively. They did not use any transcription by the unmatch-conditioned LM. This is because the ASR performances of REF-WORD-UNMATCHED and REF-SYLLABLE-UNMATCHED are worse than the condition-matched transcriptions.

Table 2: The STD task participants.

For the large-size task			
Team ID	Team name	Organization	# of submitted runs
akbl	Akiba Laboratory	Toyohashi University of Technology	3
ALPS	ALPS lab. at UY	University of Yamanashi	1
NKI13	NKI-Lab	Toyohashi University of Technology	6
SHZU	Kai-lab	Shizuoka University	2
TBFD	Term Big Four Dragons	Daido University	9

For the moderate-size task			
Team ID	Team name	Organization	# of submitted runs
akbl	Akiba Laboratory	Toyohashi University of Technology	3
ALPS	ALPS lab. at UY	University of Yamanashi	1
IWAPU	Iwate Prefectural University	Iwate Prefectural University	1
NKGW	Nakagawa-Lab	Toyohashi University of Technology	3
NKI13	NKI-Lab	Toyohashi University of Technology	8
SHZU	Kai-lab	Shizuoka University	2
TBFD	Term Big Four Dragons	Daido University	8
YLAB	Yamashita-lab	Ritsumeikan University	1

6. INEXISTENT SPOKEN TERM DETECTION TASK

The inexistent spoken term detection (iSTD) is the new task conducted in the NTCIR-10 SpokenDoc-2. In the iSTD task, task participants inspect whether a queried term is existent or inexistent in a spoken documents collection. Unlike the conventional STD tasks, the iSTD task has mainly two characteristics: (existent and inexistent) terms in a query set are evaluated together, and each queried term is evaluated in terms of the existence of it at least once in a spoken documents collection or not. The SDPWS is used as the target document collection.

6.1 Query

We define two classes as follows:

Class $\in$: is a set of queried terms existing at least once in the target collection.

Class $\notin$: is a set of queried terms that are inexistent in any target spoken document.

Figure 2 shows an example of a query set. The query consists of N sorts of terms and thier ID numbers. Note that task participants will be not informed which terms belong to the Class $\in$ (and the others to the Class $\notin$, although Figure 2 indicates the class of each term.

The format of a query term list that was provided to participants was the same as the STD moderate-size task. The moderate-size query set includes 100 Class $\notin$ terms, and the other terms belong to Class $\in$.

6.2 Submission

6.2.1 File Name

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.

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

iSTD-X-SDPWS-N.txt

term ID,	term,	Class
001,	A,	$\notin$
002,	B,	$\in$
003,	C,	$\in$
004,	D,	$\notin$
005,	E,	$\in$
006,	F,	$\notin$
007,	G,	$\in$
008,	H,	$\notin$
009,	I,	$\notin$
010,	J,	$\in$

Figure 2: An example of a query set for the iSTD task.

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

N: Priority of run (1, 2, 3, ...)

For example, if the group "NTC" submits two files, the names of the run files should be "iSTD-NTC-SDPWS-1.txt" and "iSTD-NTC-SDPWS-2.txt."

6.2.2 Submission Format

The submission file, which must be a well-formed XML document, is organized with the single root level tag `<ROOT>` and three second level tags `<RUN>`, `<SYSTEM>` ,and `<RESULT>`, which is the same as the submission format for the STD task described in Section 5.4.2.

The `<RUN>` and `<SYSTEM>` parts for the iSTD task are described similarly as those for the STD task. On the other hand in the `<RESULT>` part, task participants is required to submit the query list in which the queried terms are sorted in descending order based on their iSTD scores. "iSTD score" is a kind of confidence score which indicates that a term is likely to be inexistent in the target speech collection. The score is preferred to get a range from 0.0 to 1.0. For example, if a term is considered to be inexistent, the iSTD score will close to 1.0.

Figure 3 shows a format of query list that a participants is required to submit. "rank" means the position number on the query list. The numbers of "rank" have to be totally ordered; i.e, if there are some terms which have the same

Table 3: The number of transcription(s) used for each run on the STD task.

Set	Run	REF-WORD-MATCHED	REF-SYLLABLE-MATCHED	REF-WORD-UNMATCHED	REF-SYLLABLE-UNMATCHED	OWN trans.	total
large-size	BL-1	0	1	0	0	0	1
	BL-2	1	0	0	0	0	1
	BL-3	1	1	0	0	0	2
	akbl-1,2,3	0	1	0	0	0	1
	ALPS-1	1	1	0	0	8	10
	NKI13-1	0	0	1	1	0	2
	NKI13-2	1	1	0	0	0	2
	NKI13-3	0	0	1	0	0	1
	NKI13-4	1	0	0	0	0	1
	NKI13-5	0	0	0	1	0	1
	NKI13-6	0	1	0	0	0	1
	SHZU-1,2	1	1	0	0	0	2
	TBFD-1,2,3,7	1	1	1	1	0	4
	TBFD-4	0	1	0	0	0	1
	TBFD-5,6,8	1	1	0	0	0	2
	TBFD-9	1	1	0	0	1	3
moderate-size	BL-1	0	1	0	0	0	1
	BL-2	1	0	0	0	0	1
	BL-3	1	1	0	0	0	2
	akbl-1,2,3	0	1	0	0	0	1
	ALPS-1	1	1	0	0	8	10
	IWAPU-1	0	0	0	0	4	4
	NKGW-1,2,3	0	0	0	0	1	1
	NKI13-1	0	0	1	1	1	3
	NKI13-2	1	1	0	0	1	3
	NKI13-3	0	0	1	1	0	2
	NKI13-4	1	1	0	0	0	2
	NKI13-5	0	0	1	0	1	2
	NKI13-6	1	0	0	0	1	2
	NKI13-7	0	0	0	1	1	2
	NKI13-8	0	1	0	0	1	2
	SHZU-1,2	1	1	0	0	0	2
	TBFD-1,2,3	1	1	1	1	0	4
	TBFD-4	0	1	0	0	0	1
	TBFD-5,6	1	1	0	0	0	2
	TBFD-7	0	0	1	1	0	2
	TBFD-8	1	1	0	0	0	2
	YLAB-1	0	0	0	1	0	1

iSTD score, a participant should order them according to another criterion. "detection" needs either "yes" or "no" as its argument. If a participant's STD engine determines that a term should be inexistent, "detection" gets "no." This should be performed by the participant's criterion.

6.3 Evaluation Metrics

Evaluation metric we used in this task are as follows:

- Recall-Precision curve,
- Maximum F-measure (= the balanced point on Recall-Precision curve),
- F-measure calculated by top-100-ranked,
- F-measure limiting the terms which have detection="no."

Recall and Precision rates for terms positioned rank r and more than r are calculated as following functions:

$$Recall_r = \frac{T_{\notin,r}}{N_{\notin}} \times 100(\%)$$

$$Precision_r = \frac{T_{\notin,r}}{r} \times 100(\%)$$

, where $T_{\notin,r}$ means the number of $\notin$ terms positioned rank r and more than r, $N_{\notin}$ is the total number of terms belong to class $\notin$. By changing r from 1 to N, a recall-precision curve can be drawn. A maximum F-measure that is from the best balanced point in the curve will also be used for evaluation. Figure 4 shows the recall-precision curve of the iSTD result (Figure 3) using the query list shown in Figure 2. The maximum F-measure is 72.9%.

6.4 Evaluation Results

6.4.1 iSTD task participants

Table 4: STD performances of each submission on the large-size task.

run	micro ave.		macro ave.			index	search
	max. F [%]	spec. F [%]	max. F [%]	spec. F [%]	MAP	size [MB]	speed [s]
BL-1	42.32	40.71	43.91	36.70	0.500	58	560
BL-2	52.52	48.22	47.13	42.21	0.507	58	560
BL-3	54.25	50.46	46.79	43.95	0.532	116	560
akbl-1	39.74	33.76	39.09	37.34	0.490	17250	0.0633
akbl-2	38.11	27.56	38.99	38.53	0.452	18210	0.0719
akbl-3	38.12	26.88	35.35	35.54	0.390	17250	0.0587
ALPS-1	58.19	57.38	62.24	50.39	0.717	60	226.4
nki13-1	60.90	57.00	60.79	59.58	0.673	183.3	0.00296
nki13-2	56.09	52.87	52.79	50.75	0.608	168.1	0.00249
nki13-3	52.10	49.71	50.61	48.12	0.574	92.3	0.00188
nki13-4	50.58	48.88	46.83	43.61	0.511	83.0	0.00123
nki13-5	50.56	48.26	49.69	47.21	0.566	91.0	0.00187
nki13-6	45.17	43.37	45.57	40.26	0.525	85.1	0.00171
SHZU-1	49.44	47.56	44.40	44.46	0.423	118	13.70
SHZU-2	51.14	44.20	48.27	46.93	0.510	118	13.59
TBFD-1	63.33	60.26	60.33	60.33	0.553	3400	0.0848
TBFD-2	65.62	65.62	63.63	63.63	0.551	3400	0.0881
TBFD-3	64.07	61.49	60.43	60.39	0.548	1700	0.0439
TBFD-4	45.65	45.65	41.38	41.38	0.324	1700	0.0128
TBFD-5	54.24	53.63	47.27	47.27	0.391	1700	0.0131
TBFD-6	55.36	54.28	48.07	48.07	0.408	3400	0.0283
TBFD-7	54.49	54.49	42.26	42.26	0.357	1700	0.000791
TBFD-8	42.88	42.88	28.06	28.06	0.224	753	0.00154
TBFD-9	81.01	79.44	85.39	72.54	0.690	3400	0.0164

```
<RESULT>
<TERM rank="1" termid="004" score="1.00"
    detection="no" />
<TERM rank="2" termid="002" score="0.98"
    detection="no" />
<TERM rank="3" termid="001" score="0.90"
    detection="no" />
<TERM rank="4" termid="008" score="0.89"
    detection="no" />
<TERM rank="5" termid="005" score="0.85"
    detection="no" />
<TERM rank="6" termid="009" score="0.80"
    detection="no" />
<TERM rank="7" termid="003" score="0.50"
    detection="yes" />
<TERM rank="8" termid="007" score="0.45"
    detection="yes" />
<TERM rank="9" termid="006" score="0.40"
    detection="yes" />
<TERM rank="10" termid="010" score="0.10"
    detection="yes" />
</RESULT>
```

Figure 3: Format of a query list on the iSTD task.

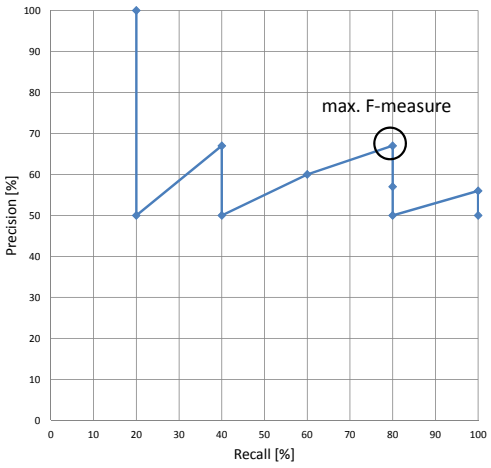

Figure 4: An example of a Recall-Precision curve

The four teams participated in the iSTD task with 15 submisison runs. In addition, the three runs as the baseline results were submitted by the organizers. The team IDs are listed in Table 6.

6.4.2 iSTD task results

Table 7 summarizes the number of transcription(s) used for each run. And the evaluation results are summarized in Table 8.

The baseline system used the DP-based word spotting which was the same as the STD tasks. And the indices were also the same as the STD tasks. In the iSTD task, first of all, the baseline system searches and detects candidates for a query term. And the detected candidate with the lowest score is used as the score of the query term. Next, the system ranks the candidates of each query term.

"ALPS-1" got the best performance at the all measures. This used the 10 sorts of transcriptions that are likely to induct false detection errors. However, "ALPS-1" excellently

Table 5: STD performances of each submission on the moderate-size task.

run	micro ave.		macro ave.			index	search
	max. F [%]	spec. F [%]	max. F [%]	spec. F [%]	MAP	size [MB]	speed [s]
BL-1	25.08	24.70	25.72	20.07	0.317	3.3	30.8
BL-2	37.58	37.46	31.43	30.42	0.358	3.3	31.9
BL-3	39.36	39.16	33.73	32.46	0.393	6.6	30.8
akbl-1	20.71	13.48	25.79	21.29	0.343	1120	0.00399
akbl-2	20.00	13.50	22.61	18.26	0.293	1120	0.00324
akbl-3	19.95	13.40	21.24	18.07	0.244	1120	0.00212
ALPS-1	46.33	42.83	52.33	39.20	0.606	45	6.06
IWAPU-1	31.37	17.27	44.49	43.74	0.675	657	2.0
NKGW-1	36.46	34.44	40.09	35.55	0.518	—	1.265
NKGW-2	33.33	27.92	32.33	23.23	0.382	2900	0.165
NKGW-3	30.98	14.09	25.70	11.43	0.284	2900	0.165
nki13-1	33.81	32.85	36.34	32.02	0.442	15.9	0.001250
nki13-2	40.24	39.73	39.97	38.29	0.456	15.6	0.000860
nki13-3	34.62	33.73	36.30	31.65	0.434	10.7	0.000785
nki13-4	41.15	40.76	39.42	38.13	0.446	10.3	0.000700
nki13-5	28.41	27.20	30.23	23.63	0.348	10.6	0.000620
nki13-6	37.56	36.71	34.77	32.57	0.390	10.5	0.000545
nki13-7	26.24	25.10	31.60	22.70	0.382	10.6	0.000705
nki13-8	27.24	26.47	29.77	23.88	0.350	10.3	0.000310
SHZU-1	28.62	27.75	29.25	27.44	0.337	6	0.525
SHZU-2	27.40	23.55	28.31	27.70	0.319	6	0.530
TBFD-1	39.69	39.15	40.70	40.70	0.336	218	0.0425
TBFD-2	39.98	38.49	39.11	39.02	0.318	218	0.0430
TBFD-3	39.83	39.40	39.14	39.14	0.321	105	0.0218
TBFD-4	25.78	25.78	23.23	23.23	0.170	105	0.0087
TBFD-5	36.27	35.83	33.27	33.27	0.264	105	0.0090
TBFD-6	36.75	36.05	34.11	34.11	0.273	218	0.0179
TBFD-7	32.60	32.60	30.53	30.53	0.239	234	0.0175
TBFD-8	31.48	31.48	24.23	24.23	0.183	43	0.0010
YLAB-1	24.10	24.04	21.57	19.93	0.221	—	569.6

inhibits the errors using their false detection control parameters.

7. SPOKEN CONTENT RETRIEVAL TASK

7.1 Task Definition

Two sub-tasks were conducted for the SCR task. The participants could submit the result of either or both of the tasks. The unit of the target document to be retrieved and the target collection are different between the sub-tasks.

- Lecture retrieval

  Find the lectures that include the information described by the given query topic. The CSJ is used as the target collection.

- 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. The SDPWS is used as the target collection.

7.2 Query Set

The organizers prepared two query topic lists; one for the passage retrieval task and the other for the lecture retrieval task. A query topic is represented by natural language sentences.

For the passage retrieval sub-task, we constructed query topics that ask for passages of varying lengths described in some presentation in the SDPWS set. Six subjects are relied upon to invent such query topics. Each subject was asked to create 20 topics so that the first half of them should be invented after looking only at the proceedings of the workshop and the latter half might be invented by looking also at the transcriptions of the presentations. Finally, we obtained 120 query topics, where 80 of them were created only from the proceedings and the rest 40 were created by investigating also the oral presentations.

For the lecture retrieval sub-task, we re-used and revised the query topics used for the SpokenDoc-1, whose target was the CSJ. While the original topics had been constructed for the passage retrieval task so that they had asked for relatively short unit of information, e.g. named entity, they were extended to search for a lecture as a whole. The length of the new queries were also extended to include their narratives, so many of them consists of more than one sentence as a result. From the 39 and 86 query topics that were used for dry and formal run of the SpokenDoc-1 respectively, we obtained 125 query topics, where the Five of them were used for the dry run and the rest 120 were used for the formal run in the SpokenDoc-2.

The format of a query topic list is as follows.

```
TERM-ID question
```

Table 6: The iSTD task participants.

For the large-size task			
Team ID	Team name	Organization	# of submitted runs
akbl	Akiba Laboratory	Toyohashi University of Technology	3
ALPS	ALPS lab. at UY	University of Yamanashi	2
TBFD	Term Big Four Dragons	Daido University	9
YLAB	Yamashita Lab.	Ritsumeikan University	1

Table 7: The number of transcription(s) used for each run on the iSTD task.

Run	REF-WORD-MATCHED	REF-SYLLABLE-MATCHED	REF-WORD-UNMATCHED	REF-SYLLABLE-UNMATCHED	OWN trans.	total
BL-1	0	1	0	0	0	1
BL-2	1	0	0	0	0	1
BL-3	1	1	0	0	0	2
akbl-1,2,3	0	1	0	0	0	1
ALPS-1,2	1	1	0	0	8	10
TBFD-1∼9	1	1	0	0	0	2
YLAB-1	0	0	0	1	0	1

An example list is:

`SpokenDoc1-dry-PASS-0001` 話者認識の学習データのサイズが知りたい
`SpokenDoc1-dry-PASS-0002` オークションにおける自動入札戦略を知りたい
`SpokenDoc1-dry-PASS-0003` 日本語話し言葉コーパスを用いている研究を教えてください。
`SpokenDoc1-dry-PASS-0004` 情報検索性能を評価するにはどのような方法があるか知りたい
...

7.3 Submission

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.

7.4 File Name

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

SCR-*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 "SCR-NTC-LEC-1.txt", "SCR-NTC-LEC-2.txt", "SCR-NTC-PAS-1.txt", "SCR-NTC-PAS-2.txt", and "SCR-NTC-PAS-3.txt".

7.5 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 "<**RESULT**>".

- <**RUN**>

  <**SUBTASK**> "STD" or "SCR". For a SCR subtask submission, just say "SCR".

  <**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.

  The other three tags "<**SYSTEM-ID**>", "<**PRIORITY**>", and "<**TRANSCRIPTION**>" in the "<**RUN**>" section are the same as in the submission format for STD task. See Section 5.4.2

- <**SYSTEM**> Same as in the submission format for STD task.

- <**RESULT**>

  <**QUERY**> Each query topic has a single "QUERY" tag with an attribute "id" specified in a query topic list (Section 7.2). Within this tag, a list of the following "CANDIDATE" tags is described.

  <**CANDIDATE**> Each potential candidate of a retrieval result has a single "CANDIDATE" tag with the following attributes. The CANDIDATE tags should, but do not necessary to, be sorted in descending order of likelihood.

  rank The rank in the result list. "1" for the most likely candidate, incleased one at a time. Required to be totally ordered in a single "QUERY" tag.

Table 8: iSTD performances. (*1) Recall, precision and F-measure rates calculated by top-100-ranked outputs. (*2) Recall, precision and F-measure rates calculated by using outputs with "detection=no" tag which is specified by each participant. (*3) Recall, precision and F-measure rates calculated by top-N-ranked outputs. N is set to obtain the muximum F-measure.

run	Rank 100*1			Specified*2				Maximum*3			
	R [%]	P [%]	F [%]	R [%]	P [%]	F [%]	rank	R [%]	P [%]	F [%]	rank
BL-1	73.00	73.00	73.00	81.00	65.85	72.65	123	73.00	76.04	74.49	96
BL-2	74.00	74.00	74.00	81.00	71.05	75.70	114	88.00	69.84	77.88	126
BL-3	75.00	75.00	75.00	81.00	70.43	75.35	115	90.00	68.18	77.59	132
akbl-1	72.00	72.00	72.00	89.00	66.92	76.39	133	95.00	65.97	77.87	144
akbl-2	67.00	67.00	67.00	87.00	65.41	74.68	133	95.00	63.33	76.00	150
akbl-3	68.00	68.00	68.00	90.00	65.69	75.95	137	94.00	65.28	77.05	144
ALPS-1	82.00	82.00	82.00	82.00	82.00	82.00	100	85.00	80.19	82.52	106
ALPS-2	79.00	79.00	79.00	79.00	79.00	79.00	100	84.00	78.50	81.16	107
TBFD-1	70.00	70.00	70.00	78.00	72.22	75.00	108	88.00	73.33	80.00	120
TBFD-2	70.00	70.00	70.00	80.00	72.73	76.19	110	88.00	73.33	80.00	120
TBFD-3	72.00	72.00	72.00	88.00	73.33	80.00	120	88.00	73.33	80.00	120
TBFD-4	73.00	73.00	73.00	77.00	74.04	75.49	104	88.00	73.33	80.00	120
TBFD-5	70.00	70.00	70.00	88.00	70.40	78.22	125	90.00	70.31	78.95	128
TBFD-6	74.00	74.00	74.00	70.00	74.47	72.16	94	88.00	73.33	80.00	120
TBFD-7	74.00	74.00	74.00	66.00	73.33	69.47	90	88.00	73.33	80.00	120
TBFD-8	74.00	74.00	74.00	53.00	71.62	60.92	74	88.00	73.33	80.00	120
TBFD-9	74.00	74.00	74.00	45.00	69.23	54.55	65	88.00	73.33	80.00	120
YLAB-1	62.00	62.00	62.00	48.00	67.61	56.14	71	89.00	61.38	72.65	145

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 CSJ, of the last IPU of the retrieved passage (an IPU sequence).
NOTE: The IPU sequences specified in a single "QUERY" tag are required to be *exclusive* each other; i.e. no two intervals in a "QUERY", each of which is specified by "CANDIDATE" tag, are not allowed to have a common IPU.

Figure 5 shows an example of a submission file.

7.6 Evaluation Measures

7.6.1 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.

$$AveP_q = \frac{1}{|R_q|} \sum_{i=1}^{|D_q|} include(d_i, R_q) \frac{\sum_{j=1}^{i} include(d_j, R_q)}{i} \qquad (3)$$

where

$$include(a, A) = \begin{cases} 1 & \cdots \quad a \in A \\ 0 & \cdots \quad a \notin A \end{cases} \qquad (4)$$

```
<ROOT>
<RUN>
<SUBTASK>SCR</SUBTASK>
<SYSTEM-ID>TUT</SYSTEM-ID>
<PRIORITY>1</PRIORITY>
<UNIT>PASSAGE</UNIT>
<TRANSCRIPTION>REF-WORD-UNMATCHED,
REF-SYLLABLE-UNMATCHED</TRANSCRIPTION>
</RUN>
<SYSTEM>
<OFFLINE-MACHINE-SPEC>Xeon 3GHz dual CPU, 4GB memory
</OFFLINE-MACHINE-SPEC>
<OFFLINE-TIME>18:35:23</OFFLINE-TIME>
...
</SYSTEM>
<RESULT>
<QUERY id="SpokenDoc1-SCR-dry-PAS-001">
<CANDIDATE rank="1" document="10-09"
 ipu-from="0024" ipu-to="0027" />
<CANDIDATE rank="2" document="12-12"
 ipu-from="0079" ipu-to="0079" />
...
</QUERY>
<QUERY id="SpokenDoc1-SCR-dry-PAS-002">
...
</QUERY>
</RESULT>
</ROOT>
```

Figure 5: An example of a submission file.

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.

$$AveP_q = \frac{1}{|R_q|} \sum_{k=1}^{M} \frac{k}{rank(r_k)} \quad (5)$$

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

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

$$\mathbf{MAP} = \frac{1}{|Q|} \sum_{q \in Q} AveP_q \quad (6)$$

7.6.2 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.

7.6.3 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.

$$uAveP_q = \frac{1}{|\tilde{R}_q|} \sum_{i=1}^{|U|} include(u_i, \tilde{R}_q) \frac{\sum_{j=1}^{i} include(u_j, \tilde{R}_q)}{i} \quad (7)$$

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.

$$\mathbf{uMAP} = \frac{1}{|Q|} \sum_{q \in Q} uAveP_q \quad (8)$$

7.6.4 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.

$$pwAveP_q = \frac{1}{|R_q|} \sum_{k=1}^{M} \frac{k}{rank(r_k)} \quad (9)$$

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.

$$\mathbf{pwMAP} = \frac{1}{|Q|} \sum_{q \in Q} pwAveP_q \quad (10)$$

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.

$$rel(p, R_q) = \max_{r \in R_q} \frac{|r \cap p|}{|r|} \quad (11)$$

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.

$$prec(p, R_q) = \max_{r \in R_q} \frac{|p \cap r|}{|p|} \quad (12)$$

Then, $fAveP_q$ is calculated as follows.

$$fAveP_q = \frac{1}{|R_q|} \sum_{i=1}^{|P_q|} rel(p_i, R_q) \frac{\sum_{j=1}^{i} prec(p_j, R_q)}{i} \quad (13)$$

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

$$\mathbf{fMAP} = \frac{1}{|Q|} \sum_{q \in Q} fAveP_q \quad (14)$$

7.7 Evaluation Results

Seven groups with total 69 runs have submitted the results for the formal run. Among them, six groups participated the lecture retrieval task and five groups participated the passage retrieval task. The team IDs are listed in Table 9.

7.7.1 Transcriptions

Table 9: SCR subtask participants.

Lecture retrieval task		
Team ID	Team name	Organization
AKBL	TUT Akiba Laboratory	Toyohashi University of Technology
ALPS	ALPS-Lab.	University of Yamanashi
HYM	Hayamiz Lab	Gifu University
INCT	kane_lab	Ishikawa National College of Technology
RYSDT	RYukoku SpokenDoc Team	Ryukoku University
TBFD	Team Big Four Dragons	Daido University

Passage retrieval task		
Team ID	Team name	Organization
AKBL	TUT Akiba Laboratory	Toyohashi University of Technology
ALPS	ALPS-Lab.	University of Yamanashi
DCU	DCU	Dublin City University
INCT	kane_lab	Ishikawa National College of Technology
RYSDT	RYukoku SpokenDoc Team	Ryukoku University

Table 10: Summary of the transcriptions used for each run.

task	run	REF-WORD-MATCHED	REF-SYLLABLE-MATCHED	REF-WORD-UNMATCHED	REF-SYLLABLE-UNMATCHED	MANUAL	total
lecture	(baseline-1,2)	√					1
	(baseline-3,4)			√			1
	AKBL-1,7	√	√				2
	AKBL-2,8			√	√		2
	AKBL-4,5		√				1
	AKBL-3,6				√		1
	ALPS-1,2	√					1
	HYM-1,2,3	√					1
	INCT-1,2,3			√			1
	RYSDT-1,$\cdots$,9	√					1
	TBFD-1,$\cdots$,9	√	√				2
passage	(baseline-1,2)	√					1
	(baseline-3,4)			√			1
	AKBL-1,$\cdots$,6	√					1
	ALPS-1,2	√					1
	DCU-1,2					√	1
	DCU-3,4,7,$\cdots$,12	√					1
	DCU-5,6,13,$\cdots$,18			√			1
	INCT-1			√			1
	RYSDT-1,$\cdots$,8	√					1

Table 10 summarizes the transcriptions used for each run. All runs used the reference automatic transcriptions provided from the organizers except that two runs for the passage retrieval used the manual transcription.

For the lecture retrieval task, most runs (27 runs) used the transcriptions on the matched condition, while the other seven runs by two groups used those on the unmatched condition. Looking into the type of transcriptions, 13 runs by two groups used both the word-based and syllable-based transcriptions, 17 runs used only the word-based transcription, and four runs by one group used only the syllable-based transcription.

For the passage retrieval task, except for the two runs using manual transcription, all runs used only the word-based transcription. Among them, most runs (24 runs) used those on the matched condition, while nine runs by two groups used those on the unmached condition.

7.7.2 Baseline Methods

We implemented and evaluated the baseline methods for our SCR tasks, which consisted of only conventional methods for IR and applied to either the 1-best REF-WORD-MATCHED or REF-WORD-UNMATCHED. Run ID baseline-1 and baseline-2 used the REF-WORD-MATCHED, while the baseline-3 and baseline-4 used the REF-WORD-UNMATCHED. 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 either TF–IDF (Term Frequency–Inverse Document Frequency) or TF–IDF with pivoted normalization [5] was used for term weighting, which are referred to as run 2 (4) and 1 (3), respectively. We used *GETA* [2] 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 $N = 15$

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

Table 11: Evaluation results for the lecture retrieval task.

Quality of transcription	run	MAP
MATCHED	(baseline-1)	0.268
	(baseline-2)	0.231
	AKBL-1	0.365
	AKBL-4	0.212
	AKBL-5	0.223
	AKBL-7	0.401
	ALPS-1	0.384
	ALPS-2	0.381
	HYM-1	0.408
	HYM-2	0.399
	HYM-3	0.372
	RYSDT-1	0.378
	RYSDT-2	0.370
	RYSDT-3	0.376
	RYSDT-4	0.367
	RYSDT-5	0.355
	RYSDT-6	0.348
	RYSDT-7	0.340
	RYSDT-8	0.292
	RYSDT-9	0.364
	TBFD-1	0.368
	TBFD-2	0.368
	TBFD-3	0.392
	TBFD-4	0.372
	TBFD-5	0.375
	TBFD-6	0.370
	TBFD-7	0.363
	TBFD-8	0.383
	TBFD-9	0.381
UNMATCHED	(baseline-3)	0.238
	(baseline-4)	0.225
	AKBL-2	0.341
	AKBL-3	0.208
	AKBL-6	0.223
	AKBL-8	0.367
	INCT-1	0.324
	INCT-2	0.320
	INCT-3	0.320

according to the rough estimate of the passage lengths of the dry run test data.

We also investigated the human retrieval performance for the passage retrieval task. We asked the human subjects to find the arbitrary length passages relevant to the given query topic from the document collections and to rank the retrieved passage according to the degree of their relevancy, as the participant's system did. We employed again the six subjects who had created the query topics, but we assigned each subject the other topics than those he had created. Finally, we obtained the ranked lists of all the 120 query topics used for the formal run of the passage retrieval task.

7.7.3 Results

For the lecture retrieval task, the evaluation results of all the submissions are summarized in Table 11.

For the passage retrieval task, the evaluation results are summarized in Table 12.

8. CONCLUSION

This paper introduced the overview of the 2nd round of the IR for Spoken Documents (SpokenDoc-2) Task in NTCIR-10 Workshop.

Table 12: Evaluation results for the passage retrieval task.

Quality of transcription	run	uMAP	pwMAP	fMAP
MANUAL	(human)	0.232	0.273	0.199
	DCU-1	0.125	0.088	0.079
	DCU-2	0.128	0.072	0.076
MATCHED	(baseline-1)	0.133	0.100	0.087
	(baseline-2)	0.092	0.082	0.068
	AKBL-1	0.102	0.088	0.063
	AKBL-2	0.129	0.125	0.086
	AKBL-3	0.131	0.137	0.093
	AKBL-4	0.131	0.123	0.087
	AKBL-5	0.122	0.132	0.083
	AKBL-6	0.126	0.139	0.089
	ALPS-1	0.075	0.046	0.033
	ALPS-2	0.075	0.046	0.033
	DCU-3	0.075	0.048	0.043
	DCU-4	0.091	0.059	0.041
	DCU-7	0.079	0.078	0.050
	DCU-8	0.098	0.096	0.063
	DCU-9	0.089	0.074	0.043
	DCU-10	0.102	0.071	0.044
	DCU-11	0.102	0.049	0.025
	DCU-12	0.100	0.048	0.027
	RYSDT-1	0.108	0.081	0.079
	RYSDT-2	0.108	0.080	0.078
	RYSDT-3	0.106	0.080	0.077
	RYSDT-4	0.099	0.078	0.074
	RYSDT-5	0.100	0.074	0.069
	RYSDT-6	0.098	0.105	0.079
	RYSDT-7	0.093	0.103	0.077
	RYSDT-8	0.097	0.096	0.074
UNMATCHED	(baseline-3)	0.071	0.059	0.044
	(baseline-4)	0.048	0.049	0.037
	DCU-5	0.044	0.039	0.027
	DCU-6	0.062	0.028	0.025
	DCU-13	0.052	0.031	0.015
	DCU-14	0.070	0.029	0.018
	DCU-15	0.054	0.046	0.022
	DCU-16	0.063	0.042	0.025
	DCU-17	0.035	0.038	0.026
	DCU-18	0.043	0.050	0.028
	INCT-1	0.075	0.045	0.038

9. REFERENCES

[1] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui. Overview of the IR for spoken documents task in NTCIR-9 workshop. In *Proceedings of The Ninth NTCIR Workshop Meeting*, pages 223–235, 2011.

[2] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui. Designing an evaluation framework for spoken term detection and spoken document retrieval at the NTCIR-9 SpokenDoc task. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[3] 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.

[4] K. Maekawa, H. Koiso, S. Furui, and H. Isahara.

Spontaneous speech corpus of Japanese. In *Proceedings of International Conference on Language Resources and Evaluation*, pages 947–952, 2000.

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