Definition of SQ-STD task at NTCIR-11
SpokenQuery&Doc (Ver. 0.2)

NTCIR-11 Spoken Query and Spoken Document Retrieval Task Organizers
June 24, 2014

1 Document Collection

Corpus of 1st to 7th Spoken Document Processing Workshop (SDPWS1to7) is to be used as the document collection for the NTCIR-11 Spoken-Query&Doc task. It is distributed from the SpokenQuery&Doc task organisers. It consists of the recordings of the first to seventh annual Spoken Document Processing Workshop with slide-change annotation.

Each lecture in the SDPWS1to7 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 the STD and SCR tasks.

Unlike the previous “the corpus of Spoken Document Processing Workshop (SDPWS)” used in the NTCIR-10 SpokenDoc-2 task, SDPWS1to7 includes the additional 10 lectures from the 7th workshop held in 2013.

1.1 Component Files

The component files of the document collection grouped into two categories; those provided for each lecture and those provided for each IPU. The formers are named as its lecture ID, while the latters are named as its IPU ID, which are lecture ID followed by a series number (started with 0) of IPUs connected with a hyphen. Each of them has its own extension.

We also refer to slide IDs, which are denoted within some of the files. A slide ID is a series number (started with 1) of presentation slides.

1.1.1 Files for each lecture

VAD file It records the result of the voice activity detection applied on the audio data of the lecture. Its extension is .seg. Each line of a file, which is corresponds to an IPU, has two integers formatted as follows.

<start time> <end time>
A unit of them is 1/16000 second from the beginning of the lecture, i.e. 16000 means one second from the beginning.

**Slide group file** *(Not used for SQ-STD task.)* It describes slide groups of the lecture. Its extension is `.grp`. Each line of a file corresponds to a slide group, which is described as a sequence of contiguous slide IDs. Note that, in this file, slide IDs are never omitted so that each slide ID appears exactly once in a file.

**Time stamps of slide transitions** *(Not used for SQ-STD task.)* It records time stamps of starting each presentation slide. Its extension is `.tmg`. Each line is formatted as follows.

```
<slide ID> [<minutes> ":"] <second>
```

The second column denotes a starting time of a slide from the beginning of a lecture. Note that the first slide of each slide group must have a corresponding line, but the others have not always a line in this file, i.e. some inner slides in a slide group can be omitted.

Notice that, for most of the lectures in the collection, time stamps are recorded as second-level granularity, so that they are not accurate enough to locate the exact position in its corresponding audio file. (It was such a limitation brought by using the off-the-shell software designed for recording an oral presentation, which was used in most of our recordings.)

**Slide-to-IPU alignment file** *(Not used for SQ-STD task.)* It describes alignments between the starting time of a slide and an IPU. Its extension is `.align`. Each line is formatted as follows.

```
<slide ID> <IPU ID> ["+"
```

The lines without "+" at its end mean that the slide denoted by `<slide ID>` is started at the beginning of the IPU denoted by `<IPU ID>`, while those with "+" at its end mean that the slide is started somewhere within the IPU. This file provides an easy way to divide a transcription of a lecture into a set of documents.

**Manual transcription file** It describes text transcription of a lecture obtained by a human transcriber. Its extension is `.txt`. Each line is formatted as follows.

```
<IPU ID> ":" <text>
```

Several tags, which are explained in another document (the annotation manual), are introduced to describe nonverbal events in the text transcription. Among them, `<slide ID>` tag is used to indicate the position that the slide denoted by `<slide ID>` is shown for the first time in the lecture.
**Reference automatic transcription** The organizers prepared five automatic transcriptions. Three of them, whose file extension is "\_word.jout", are the word-based transcription made by a large vocabulary continuous speech recognizer using a word-based trigram model, while the other two, whose file extension is "\_syll.jout", are the subword-based transcriptions made by a continuous syllable speech recognizer using a syllable-based trigram model. The other differences are in the language models and the acoustic models used in the speech recognition system.

The five automatic transcriptions are referred to as following identifiers.

- REFWORD-MATCH, REF-SYLLABLE-MATCH Their file extension is \unmatchLM\_{word,syll}.jout. The acoustic model and the language model are trained by using the Corpus of Spontaneous Japanese. (same as "matched" transcriptions used in the NTCIR-10 SpokenDoc-2)

- REF-WORD-UNMATCH-LM, REF-SYLLABLE-UNMATCH-LM Their file extension is \unmatchLM\_{word,syll}.jout. The acoustic model is trained by using CSJ, while the language model is trained by newspaper articles. (same as "unmatched" transcriptions used in the NTCIR-10 SpokenDoc-2)

- REF-WORD-UNMATCH-AMLM (new in the NTCIR-11) Its file extension is \unmatchAMLM\_{word}.jout. Both the acoustic model and the language model are trained by "unmatched" condition. They are those distributed as Julius dictation kit v4.3.1 [1], whose acoustic and language models are trained by ASJ Continuous Speech Corpus (JNAS) and Balanced Corpus of Contemporary Written Japanese (BCCWJ), respectively.

1.1.2 Files for each IPU

**Audio file** The audio files of lectures are stored in WAV format for each IPU. The file names are formatted as follows.

-  

-  

-  

2 **Transcription**

Standard SCR 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 are going to prepare four reference automatic transcriptions. It enables that those who are interested in SDR but not in ASR can participate our tasks. It also enables the comparison of the IR methods based on the same underlying ASR performances. The participants can also use both transcriptions at the same time to boost the performance. The textual representation of them will be 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 out-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: word-based with low WER, word-based with high WER, syllable-based with low WER, and syllable-based with high 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.

3 Query

The query terms used for the SQ-STD task are put together into a single file written in an XML format, called query term list. It has a single root level tag "<QUERY-TERM-LIST>". Under the root tag, there are a sequence of tag "<QUERY>", each of which corresponds to a single query term.
A “<QUERY>” tag has one attribute named “id”, where its own query term id is denoted as its value. Under the “<QUERY>” tag, it has two sections specified by the two tags named “<SPK>” and “<TXT>”.

- <TXT>

It is used for describing the materials used for the STD task from text queries. It has two attributes “text” and “yomi”. The value of the “text” tag is the manually transcribed text of the query term, while that of the “yomi” tag is the Japanese pronunciation of the query term written in a Japanese KATAKANA sequence.

Notice that, for the judgment of the term’s occurrence in the golden file, “text” is searched against the manual transcriptions, while the “yomi” is never considered for the judgment. Furthermore, the organizers do not assure the participants of the correctness of what described in the “yomi” fields, so the participants should take the responsible for using it. Nevertheless, the organizers believes it should help participants to predict the term’s pronunciation.

- <SPK>

Under this tag, the materials used for the STD task from spoken queries are described. They consist of a set of “<SEGMENT>” tags.

A “<SEGMENT>” specifies an speech segment where a query term is uttered in a spoken query. It has three attributes, “query-topic-id”, “time-from”, and “time-to”. A value of a “query-topic-id” attribute is one of the query topic IDs provided from the task organizers. A pair of the attributes “time-from” and “time-to” denotes the time interval that the query term in question is uttered in the query topic specified by the “query-topic-id”. Their values are real numbers denoted in second from the begining of the WAV format file of the spoken query topics.

Some query term may have several “<SEGMENT>” tags, just because it appears several times spread over the query topics. Participants can make use of these segments all together for searching it.

Note that the files of the spoken query topics are found in the query set for the SQ-SCR task. Please refer to the SQ-SCR task definition for more details.

Figure 1 shows an example of a query term list file.

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.
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.

SQSTD-X-T-I-N.txt

X: System identifier that is the same as the group ID (e.g., NTC)
T: Target task.
  - IPU: IPU retrieval task.
    - For SQ-STD task submission, just say “IPU”.
I: Input modality.
  - SPK: Spoken Query.
  - TXT: Text Query.
N: Priority of run (1, 2, 3, ...) for each target document set.

For example, if the group “NTC” submits two files and three files by using spoken queries and text queries, respectively, then the names of the run files should be “SQSTD-NTC-IPU-SPK-1.txt”, “SQSTD-NTC-IPU-SPK-2.txt”, “SQSTD-NTC-IPU-TXT-1.txt”, “SQSTD-NTC-IPU-TXT-2.txt”, and “SQSTD-NTC-IPU-TXT-3.txt”.

Figure 1: An example of a query term list file.
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> “SQ-STD” or “SQ-STD”`. For a SQ-STD subtask submission, just say “SQ-STD”.
  - `<SYSTEM-ID> System identifier that is the same as the group ID.
  - `<PRIORITY> Priority of the run.
  - `<TRANSCRIPTION> The transcription used as the text representation of the target document set. “MANUAL” if it is the manual transcription. “REF-WORD-MATCH”, “REF-WORD-UNMATCH-LM”, “REF-WORD-UNMATCH-AMLM”, “REF-SYLLABLE-MATCH”, or “REF-SYLLABLE-UNMATCH-LM”, if it is one of the reference automatic transcription provided from the task 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.
  - `<QUERY-TRANSCRIPTION> The transcription used as the text representation of the spoken queries. “MANUAL” if text queries are used instead of spoken queries. “REF-*” (“*” should be replaced by a transcription Identifier) if one of the reference transcription provided from the task organizers is used. “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-TERM>` Each query term has a single “QUERY” tag with an attribute “id” specified in a query term list (Section 3). 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.
Figure 2: An example of a submission file.

lecture The searched lecture ID.
ipu The searched Inter Pausal Unit ID.
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 2 shows an example of a submission file.

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 averaged over queries. F-measure at the maximum decision point, Recall-Precision curves and mean average precision (MAP) will also be used for analysis purpose.
Mean average precision for the set of queries is the mean value of the average precision values for each query. It can be calculated as follows:

\[
MAP = \frac{1}{Q} \sum_{i=1}^{Q} AveP(i)
\]  

(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.

\[
AveP(i) = \frac{1}{Rel_i} \sum_{r=1}^{N_i} (\delta_r \cdot Precision_i(r))
\]  

(2)

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