Definition of STD subtask at NTCIR-9 SpokenDoc (Ver. 0.3)

NTCIR-9 IR for Spoken Documents Task Organizers

June 24, 2011

1 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 are used as the target documents for our both STD and SDR tasks (referred to as **ALL**). The subset 177 lectures 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.

2 Transcription

Standard STD 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

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

2. Reference Automatic Transcriptions

   The organizers are going to prepare two 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 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 SDR can use this transcription.

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 organizers will provide two sets of the query term list, i.e. the list for **ALL** lectures ant the list for the **CORE** lectures. Each participant's submission (called "run") should choose one from the two according to their target document collection, i.e. either **ALL** or **CORE**.

The format of a query term lists is as follows.

```
TERM-ID term Japanese_katakana_sequence
```

An example list is:

```
SpokenDoc1-STD-dry-ALL-0001 国立国語研究所 コクリツコクゴケンキュージョ
SpokenDoc1-STD-dry-ALL-0002 統計数理研究所 トーケイスーリケンキュージョ
SpokenDoc1-STD-dry-ALL-0003 大語彙音声認識 ダイゴイオンセーニンシキ
SpokenDoc1-STD-dry-ALL-0004 談話セグメント境界 ダンワセグメントキョーカイ
...
```

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.

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.

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 docuemnt set.

For example, if the group "NTC" submits two files for targetting **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", "STD-NTC-CORE-3.txt".

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 "<**RESULTS**>".

- <**RUN**>

  <**SUBTASK**> "STD" or "SDR". 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. "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 (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.

  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 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.

```
<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>
<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 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:

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

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.

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

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.