Spoken Web Search

Nitendra Rajput
IBM Research
4, Block C, ISID Campus
Vasant Kunj, New Delhi 110070
India
rnitendra@in.ibm.com

Florian Metze
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
USA
fmetze@cs.cmu.edu

ABSTRACT

In this paper, we describe the “Spoken Web Search” Task, which was held as part of the 2011 MediaEval campaign. The purpose of this task was to perform audio search in several languages, with very little resources being available in each language. The data was taken from audio content that was created in live settings and was submitted to the “spoken web” over a mobile connection.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing.

General Terms

Algorithms, Performance, Experimentation, Languages.

Keywords

Spoken Term Detection, Web Search, Spoken Web.

1. INTRODUCTION

The “spoken web search” task of MediaEval 2011 [5] involves searching *for* audio content *within* audio content *using* an audio content query. The task required researchers to build a language-independent audio search system so that, given an audio query, it should be able to find the appropriate audio file(s) and the (approximate) location of query term within the audio file(s). Evaluation was performed using standard NIST metrics.

As a contrastive condition (i.e. a "general" run in MediaEval’s terms), participants were asked to run systems not based on an audio query, as the organizers also provided the search term in lexical form.

Note that language labels and pronunciation dictionaries were not provided. The lexical form cannot be used to deduce the language in the audio-only condition. The goal of the task was primarily to compare the performance of different methods on this type of data, not so much a performance comparison geared towards different sites.

2. MOTIVATION

Imagine you want to build a simple speech recognition system, or at least a spoken term detection (STD) system in a new dialect, for which only very few audio examples are available. Maybe there even is no written form for that dialect? Is it possible to do something useful (i.e. identify the topic of a query) by using only those very limited resources available?

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy.

3. RELATED WORK

This task has been suggested by IBM Research India, and is using data provided by this group, see [2]. Previous attempts at spoken web search have mostly focused on searching through the meta-data related to the audio content [3][4].

4. TASK DESCRIPTION

Participants received development and test utterances (audio data) as well as development and test (audio) queries, described in more detail below. Only the occurrence of development queries in development utterances was provided.

Participants were required to submit the following runs in the audio-only condition (i.e., without looking at the textual form of the queries):

- On the test utterances: identify which query (from the set of development queries) occurs in each utterance (*0-n* matches per term, i.e. not every term necessarily occurs, but multiple matches are possible)
- On the test utterances: identify which query (from the set of test queries) occurs in each utterance (*0-n* matches)
- On the development utterances: identify which test query occurs in each utterance (*0-n* matches)

The purpose of requiring these three conditions is to see how critical tuning is for the different approaches, i.e., we assume that participants already know their performance for "dev queries" on "dev utterances", so for evaluation we will evaluate the performance of unseen "test queries" on previously known "dev utterances" (which could have been used for unsupervised adaptation, etc), known queries (for which good classifiers could have been developed) on unseen data, and unseen queries on unseen utterances.

Optionally, participants were asked to submit the same runs also using the provided lexical form of the query, i.e. they could use existing speech recognition systems, etc., for comparison purposes.

Not every test query occured in the data (that’s why it is “*0-n* matches”).

Participants could submit multiple systems, but had to designate one primary system. If more then one were submitted as primary, the last one uploaded was considered "primary".

Participants were allowed to use any additional resources they might have available, as long as their use is documented in the working notes paper.

4.1 Development Data

Participants were provided with a data set that has been kindly made available by the Spoken Web team at IBM Research, India [6]. The audio content is spontaneous speech that has been created over phone in a live setting by low-literate users. While most of the audio content is related to farming practices, there are other domains as well. The data set comprises audio from four different Indian languages: English, Hindi, Gujarati and Telugu. Each data item is ca. 4-30 secs in length. However, the language labels were intentionally not provided either in the development or the evaluation data set.

As already mentioned above, participants were allowed to use any additional resources they might have available, as long as their use is documented in the working notes paper.

The development set contains 400 utterances (100 per language) and 64 queries (16 per language), all as audio files recorded in 8kHz/ 16bit, as WAV file. For each query and utterance, we also provided the lexical transcription in UTF8 encoding. The transcriptions were in a Romanized transliterated form. For each utterance, the organizers provided 0-n matching queries (but not the location of the match).

There are 4 directories in the Spoken Web data. `Audio` has the 400 audio files in four languages. `Transcripts` has the corresponding word level transcriptions in roman characters of the audio file. `QueryAudio` has the 64 query audio terms. `QueryTranscripts` has the corresponding word level roman transcription of the `QueryAudio` files. The file `Mapping.txt` shows which query is present in which audio file.

4.2 Evaluation Data

The test set consists of 200 utterances (50 per language) and 36 queries (9 per language) as audio files, with the same characteristics. As with the development data, the lexical form of the query was provided, but not the matching utterances.

The evaluation data consists of two directories. `EvaluationAudio` directory contains 200 audio files that are the utterance. This has 50 audio files for each of the four languages. `EvaluationQueryAudio` contains the 36 audio files that are the query audio terms. This has 9 audio queries from each of the four languages.

The written form of the search queries is also provided in the directory `EvaluationQueryTranscripts` (`EvaluationTranscripts` may be made available later).

Data was provided as a "termlist" XML file, in which the "termid" corresponds to the filename of the audio query. This information was packaged together with the scoring software (see below), for example:

```
<?xml version="1.0" encoding="UTF-8"?>

<termlist
ecf_filename="expt_06_std_dryrun06_eng_all_spch_ex
pt_1.ecf.xml" ... >

<term
termid="DryRun06_eng_0001"><termtext>"years"</term
text></term>

</termlist>
```

5. EVALUATION OF RESULTS

The ground truth was created manually by native speakers, and provided by the task organizers, following the principles of NIST's Spoken Term Detection (STD) evaluations.

The primary evaluation metric was ATWV (Average Term Weighted Value), as used in the NIST 2006 Spoken Term Detection (STD) evaluation [1].

The systems can be scored using the software provided in `me2011-scoring-beta2.tar.bz2` available on the FTP server. This software allowed participants to verify themselves that the organizers can process their output for scoring, and reports the respective figure of merit plus graphs. It also contains the reference ECF files (these were generated automatically from the text-only files described above).

The NIST-compatible files have been generated automatically from simple files that do not contain time information of the occurrences. Generous time windows for matches should allow for correct detections.

6. OUTLOOK

Spoken Web is primarily targeted at communities that currently do not have access to Internet. Most of such target users speak in non-traditional languages for which good speech recognition systems do not exist. Low resource speech recognition is currently receiving a lot of attention, as exhibited by research efforts such as IARPA's Babel program. We will discuss future directions at the evaluation workshop and present the outcome in future publications.

7. ACKNOWLEDGMENTS

The organizers would like to thank Martha Larson for organizing this event, and the participants for putting in a lot of hard work. Most of the content is from low-literate farmers across various parts in India – the organizers would like to thank them for creating this content. The data was kindly collected and provided by IBM Research India.

8. REFERENCES

[1] Fiscus, J., Ajot, J., Garofolo, J., Doddington, G., 2007, "*Results of the 2006 Spoken Term Detection Evaluation*," Proceedings of the ACM SIGIR 2007, Workshop in Searching Spontaneous Conversational Speech (SSCS 2007), pp. 51-56.

[2] Kumar, A., Rajput, N., Chakraborty, D., Agarwal, S. K., Nanavati, A. A., "*WWTW: The World Wide Telecom Web*," NSDR 2007 (SIGCOMM workshop), Kyoto, Japan, 27 August, 2007.

[3] J. Ajmera, A. Joshi, S. Mukherjee, N. Rajput, S. Sahay, M. Shrivastava, K. Shrivastava, *"Two Stream Indexing for Spoken Web Search,"* 20th International World Wide Web Conference, WWW 2011

[4] Mamadou Diao, Sougata Mukherjea, Nitendra Rajput, Kundan Srivastava, *"Faceted Search and Browsing of Audio Content on Spoken Web,"* CIKM 2010.

[5] http://www.multimediaeval.org/mediaeval2011/index.html

[6] http://domino.research.ibm.com/comm/research_projects.nsf/pages/pyrmeait.index.html