EVALITA 2011

Automatic Speech Recognition

Large Vocabulary Transcription

Fabio Brugnara, Roberto Gretter, Marco Matassoni (FBK-irst, Trento)

1. Introduction

In the Large Vocabulary Transcription task, systems are required to transcribe audio sequences of Italian parliament. Two subtasks are defined, and applicants may choose to participate in any of them:

- ***Transcription***
- ***Constrained transcription***, using the accompanying minutes

Two modalities are allowed:

- ***closed***: only distributed data are allowed for training and tuning the system
- ***open***: the participant can use any type of data for system training, declaring and describing the proposed setup in the final report

The evaluation is based on Word Accuracy, computed as Minimum Edit Distance between the recognizer output and the reference annotation. Training and development material extracted from wide-band (16kHz) corpora are provided as well as the evaluation tool.

2. Task materials

The training set consists in:

- about 30h of parliament audio sessions along with related (automatic) transcriptions
- 5-years (1 legislature) minutes of parliament sessions
- lexicon covering acoustic and partly language model data

The development set contains:

- 1 hour parliament audio session
- the minutes of the session
- the reference transcription

The test set is distributed as:

- 1 hour audio sequences from parliament sessions

Distributed data can be used only for the Evalita context, no fee is required.

3. Description of the distributed package

Training data for Acoustic Model (AM) training

The training material for AM training includes audio and transcriptions data:

- `am/data/training.list` contains the list of prefixes to be used for AM training.

For each PREFIX there are the following files:

- `am/data/PREFIX.sph` audio file in NIST sphere format, 16 kHz
- `am/data/PREFIX.txt` automatic transcription in words/sentences
- `am/data/PREFIX.wrd` automatic transcription in words
- `am/data/PREFIX.sphn` automatic transcriptions in phones

Files with suffixes `.sphn` and `.txt` share the following format:

```
file_id

t1_1 t1_2 +tr1_1+ item1_1 +tr1_2+ item1_2 +tr1_3+ item1_3  .... +tr1_n+

t2_1 t2_2 +tr2_1+ item2_1 item2_2 item2_3 +tr2_2+  .... +tr2_n+

....
```

where `file_id` identifies the audio file, t*_1 is an absolute time marker (integer, in samples) and identifies the beginning of the sentence in the audio file, t*_2 is an absolute time marker (integer, in samples) and identifies the end of the sentence in the audio file, `+tr*_*+` is a relative time marker (integer, in samples) inside the sentence, `item*_*` either word(s) or phone(s).

The files with suffix `.wrd` have the following format:

```
 file_id

 ts1_1 ts1_2 item1

 ts2_1 ts2_2 item2
```

```
        ....
```

Where `file_id` identifies the audio file, `ts*_1` is an absolute time marker (float, in seconds) identifies the beginning of the item in the audio file, `ts*_2` absolute time marker (float, in seconds) identifies the end of the item in the audio file, `item*` either word or phone

The lexicon `am/lex/amtrain.lex` contains the transcription in SAMPA phones of every word in the training. It has the following format:

```
word1 phone1_1 phone1_2 ... phone1_n

word2 phone2_1 phone2_2 ... phone2_n

word2(2) phone3_1 phone3_2 ... phone3_n

word2(3) phone4_1 phone4_2 ... phone4_n

...
```

Where `word`* is a word, `word*(n)` is the n-th possible transcription of `word`*, `phone*_*` is a SAMPA phone.

Language Model (LM) training

The files in `lm/data/leg14/sed*/*` are text data from 751 sessions (sedute), including:

- 17463 `.htm` files, original data downloaded from the web pages;
- 751 `.txt` files, concatenation of all .htm files of the same session, after removal of html tags and maintaining punctuation, page numbering, some formatting, etc.
- 751 `.ntxt` files, cleaned version of the `.txt` files, ready for LM building.

The applied text processing includes: removal of punctuation symbols, numbers normalization (`articolo quarantanove comma cinque del regolamento`), lowercase (social forum di firenze), separation of words (`la discussione sull' ordine dei lavori`), removal of formatting patterns.

The lexicon `lm/lex/lmtrain.lex` contains the transcription in SAMPA phones of every word that appears more than twice in the text data used for building the LM.

It has the following format:

```
word1 phone1_1 phone1_2 ... phone1_n

word2 phone2_1 phone2_2 ... phone2_n

word2(2) phone3_1 phone3_2 ... phone3_n
```

```
word2(3) phone4_1 phone4_2 ... phone4_n
...
```

Where `word`* is a word, `word*(n)` is the n-th possible transcription of `word*`, `phone*_*` is a SAMPA phone.

Development set

The files in dev`/*/*` are audio and text data from 2 sessions (sedute), including:

- 2 `.sph` files, audio file in NIST sphere format, 16 kHz
- 2 `.stm` files, reference transcription, checked manually;
- 71 `.htm` files, original data downloaded from the web pages;
- 2 `.txt` files, concatenation of all `.htm` files of the same session, after removal of html tags and maintaining punctuation, page numbering, some formatting, etc.
- 2 `.ntxt` files, cleaned version of the `.txt` files;

Test set

The format is the same as for the Development set, but obviously no `stm` files are provided.

The participants are expected to submit the output of the proposed systems in the `ctm` form (see next section).

4. Evaluation procedure

To perform evaluation on the output of a speech recognizer, a reference file (suffix `.stm`), provided for the development data, and a hypothesis file (`.ctm`) are required.

The evaluation tool called `sclite` (maintained by NIST) is provided in the distribution.

The format of a `ctm` file contains a word for every row, together with timing information:

```
label condition start_time duration word
```

where:

- `label` is the id of the file;
- `condition` is not used (1);
- `start_time` is the time, a float expressed in seconds;
- `duration` is the duration of the word, a float expressed in seconds;

- `word` is the recognized word.

For instance:

```
423TestSet 1 9.025 0.350 uno

423TestSet 1 9.375 0.790 cinque

423TestSet 1 10.695 0.930 contraria

423TestSet 1 16.595 0.280 allora
```

....

Scoring will be case-insensitive. To score the files, go to the `eval` directory:

```
cd eval
```

and run the following commands:

```
rm  eval_example/423TestSet.stm.filt  eval_example/423TestSet-2.ctm.filt
```

(in this way, every time files `.filt` are recreated)

```
sctk-2.3.11/bin/hubscr.pl -p sctk-2.3.11/bin/ -h rt-stt -l english \

     -g sctk-2.3.11/src/test_suite/exampleIT.glm \

     -r eval_example/423TestSet.stm eval_example/423TestSet-2.ctm
```

Several files are created, in particular the `.dtl` file contains the score expressed as "Percent Word Accuracy".

As example, the file:

```
eval_example/423TestSet-2.ctm.filt.dtl
```

reports a possible result of the tool.