Partha Talukdar

HP Labs, India
partha.talukdar@hp.com

Optimal Text Selection Module Version 0.2

Keywords: *Complete diphone coverage, Festival, Speech Synthesis, Unit selection*

Contents

1. Usage

To use the system, run the following command:

```
$ ./text_sel <Corpus Filename> <Complete Coverage (Y/N)>
```

The following two files are accessed by the program and hence they should be present in the same directory from which the program is run.

units: This file contains the units and their frequencies. A sample "units" file is present in the "sample" directory. It is important to arrnge the units in the correct format, which is:

```
unit_name_1 unit_frequency_1
.                           .
.                           .
unit_name_n unit_frequency_n
```

original_text: The same set of sentences should be there in the file input by the user and this file. This file just contains a more readable version of the sentences contained in the input file. It is **IMPORTANT** to have one to one correspondence in terms of sentences between this file and the file input by the user. This file is useful because in most of the cases the input file contains the phonetic representation of sentences while it is desirable to have the selected sentences stored in orthographic text.

The program generates the following two files:

selected_sentences: This file contains the set of selected sentences and they are picked up from the original_text file.

uncovered_units: This file contains the set of uncovered units which could not be covered by the set of sentences. This file is NOT generated if complete coverage is obtained.

2. Mini-FAQ

>> What does complete coverage mean ?

Complete coverage refers to complete coverage of the units one is trying to cover in the text file input to the text selection system. The system uses this information to assign weights to the units and accordingly weigh sentences. If you are specifying weight of the units to be covered in the "units" file then you can use the option "N". This will make the system use the weight as it is. But if you use "Y", then the inverse of the specified weight is used in calculating weight of sentences. If the user specifies that the units are all present in the input text (specifying it through option "Y"), then inversing the frequency (assuming frequency is listed as weight in the "units" file) makes the system focus on the less frequent units first, the high frequency ones will anyway get covered en-route. So if the user is specifying the unit weights himself, and not relying on frequency based assignment, then he should use the "N" option irrespective of whether complete coverage of units is possible or not. The system will always output the best it can do and hence whatever can be covered in the text will be covered. Though using these two options might result in different number of selected sentences (since unit weighting schemes are different), but coverage of units in both the cases are going to be same. Whatever is not uncovered, will be stored in both the cases in a file "uncovered_units".

>> How are units defined here - are they simply diphones?

They can be anything. The system uses string matching and hence any size of unit is fine: i.e. it can be word, diphone, phone etc. The only constraint is that it should be a continuous string. Also units in the phonetised text should be separated by delimiters (space, - etc.).

>> Is it correct to assume that "original_text" file is simply the orthographic representation, and that it is related to the input file by a G2P tool?

Yeah exactly. This is done so that the final selected sentences are stored in normal orthography. Care should be taken so that there is a one-to-one sentence correspondence between the "original_text" file and the phonetised version of this file which is input to the text selection system.

>> Is it necessary to use the orthography ?

No, orthographic text is not necessary. However, the presence of the file "original_text" is. So if you want the selected sentences also in transcription, then you can make a copy of the input phonetic file as "original_text" and run the system. The system basically selects the final sentences from this file and hence its presence is necessary, though it doesn't matter whether it is in orthographic text or in phonetic.

REFERENCES

[1] Thomas H. Cormen, Charles E. Leiserson and Ronald L. Rivest. Introduction to Algorithms. MIT Press, Massachusetts, USA.