A Brief of the MATBN Corpus

MATBN Mandarin Chinese broadcast news corpus

The MATBN Mandarin Chinese broadcast news corpus is a product of a joint project sponsored by the National Science Council, Taiwan. It contains a total of 198 one-hour news shows from the Public Television Service Foundation, Taiwan with corresponding transcripts. The primary purpose of this collection is to provide training and testing data for continuous speech recognition evaluation in the broadcast news domain.

The MATBN corpus spanned the period November 17, 2001 through April 3, 2003. Each one-hour broadcast news episode recording was first made in stereo with a 44.1kHz sampling rate and 16 bit resolution by a DAT recorder set up in the TV broadcasting studio. Then, each DAT recording was converted into a single Microsoft Windows wave file. Finally, the signal was down-sampled to 16 kHz with a resolution of 16 bits. During this operation, only the left channel was selected.

The MATBN corpus has been segmented, labeled, and transcribed manually using the DGA&LDC Transcriber[1] [Barras *et al.* 2001]. The transcripts are in Big5-encoded form, with SGML tagging to annotate acoustic conditions, background conditions, story boundaries, speaker turn boundaries, and audible acoustic events, such as hesitations, repetitions, vocal non-speech events, and external noises. These tags include time stamps that are used to align the text with the speech data. In the 198-hour broadcast news corpus, based on hand-segmentation results, there are 4,100 stories, 581 headlines, 197 weather forecasts, and 197 ending sections. Around 143 hours of speech from 10 weather forecasts and all the stories, headlines, and ending sections were carefully transcribed, while the remaining weather forecasts and segments containing advertising or pure music were just annotated with time stamps without orthographic transcripts. There are 7 anchor reporters, 386 field reporters, and 5,900 interviewees. The identities of some field reporters and interviewees could not be determined. Since the unidentified field reporters and interviewees could correspond to the same person, the true numbers of field reporters and interviewees could be lower than the above numbers. The transcripts contain around 2.3 million Chinese characters in total.

A development set and an evaluation set have been defined for the benchmark test. The development set consisted of five shows recorded on 2003/01/24, 2003/01/27, 2003/02/07, 2003/03/05, and 2003/03/06, while the evaluation set consisted of five shows recorded on 2003/01/28, 2003/01/29, 2003/02/11, 2003/03/07, and 2003/04/03. The basic guidelines for making selections are as follows: First, we wanted to include as many studio anchors as possible. Second, the test shows had to be broadcast after January 1st, 2003 so that we could use the newswire text before January 1st, 2003 to train the language models.

[1] Transcriber can be downloaded at http://www.ldc.upenn.edu/mirror/Transcriber.old/en/menu_web.html.

The corpus is distributed on 5 DVDs:
MATBN_1: TRAIN_1 (36 speech files and the corresponding transcription files)
MATBN_2: TRAIN_2 (36 speech files and the corresponding transcription files)
MATBN_3: TRAIN_3 (36 speech files and the corresponding transcription files)
MATBN_4: TRAIN_4 (41 speech files and the corresponding transcription files)
MATBN_5: TRAIN_5 (39 speech files and the corresponding transcription files),
DEVELOPMENT (5 speech files and the corresponding transcription files),
EVALUATION (5 speech files and the corresponding transcription files)

For details of the MATBN corpus, please refer to [Wang *et al*. 2005] or http://sovideo.iis.sinica.edu.tw/SLG/corpus/MATBN-corpus.htm.

References

Barras, C., E. Geoffrois, Z. B. Wu and M. Liberman, "Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production," *Speech Communication*, 33, 2001, pp. 5-22.

Wang, Hsin-min, Berlin Chen, Jen-Wei Kuo, and Shih-Sian Cheng, "MATBN: A Mandarin Chinese Broadcast News Corpus," *International Journal of Computational Linguistics and Chinese Language Processing*, 10(2), June 2005, pp. 219-236.