Post-processing parameter has been changed (4 sec. -> 5 sec.) to get a better result.
Numbers in tables are changed accordingly.

Large-Scale Multimodal Movie Dialogue Corpus

Ryu Yasuhara
Yamagata Univeristy
3-16, 4 Jyonan, Yonezawa
Yamagata, Japan
tdh51810@st.yamagata-u.ac.jp

Masashi Inoue*
Yamagata Univeristy
3-16, 4 Jyonan, Yonezawa
Yamagata, Japan
mi@yz.yamagata-u.ac.jp

Ikuya Suga
Yamagata Univeristy
3-16, 4 Jyonan, Yonezawa
Yamagata, Japan
tnn65025@st.yamagata-u.ac.jp

Tetsuo Kosaka
Yamagata Univeristy
3-16, 4 Jyonan, Yonezawa
Yamagata, Japan
tkosaka@yz.yamagata-u.ac.jp

ABSTRACT

We present an outline of our newly created multimodal dialogue corpus that is constructed from public domain movies. Dialogues in movies are useful sources for analyzing human communication patterns. In addition, they can be used to train machine-learning-based dialogue processing systems. However, the movie files are processing intensive and they contain large portions of non-dialogue segments. Therefore, we created a corpus that contains only dialogue segments from movies. The corpus contains 149,689 dialogue seg-ments taken from 1, 722 movies. These dialogues are au-tomatically segmented by using deep neural network-based voice activity detection with filtering rules. Our corpus can reduce the human workload and machine-processing ef-fort required to analyze human dialogue behavior by using movies.

CCS Concepts

•**General and reference** → **Empirical studies;**

Keywords

Dialogue; Multimodal; Corpus; Movie; Film; VAD; DNN

1. INTRODUCTION

Dialogue data is essential for understanding patterns in human communication and developing dialogue-processing systems. However, collecting a large amount of dialogue data is costly and there are few publicly available multimodal corpora. Instead of collecting new dialogue data, we can use recorded dialogues such as those in movies. However, when using movie data as a dialogue corpus, we need annotations that indicate the beginning and ending time of dialogue segments. Often, manual extraction of the segments is too costly, which means that a method is required for automatic processing. We address this issue by applying deep neural network (DNN)-based voice activity detection (VAD).

Table 1: Corpus overview

Property	Value
Source movies	1, 722
Average duration of movies	1.2 hours
Movie genre	22

2. CORPUS OVERVIEW

An overview of the corpus is provided in Table 2. The corpus consists of two parts: movie files and annotations. These files are three-column CSV files that contain the following information about a segment: the starting time, ending time, and the label (dialogue or not). The annotation files and detailed information are available on our project web page[1]. The statistics of the corpus is provided in Table 2. The movie files need to be downloaded separately from the hosting site. Since the annotation files include a complete list of URLs for the movie files and a sample scraping script is provided, users can easily download the movie files automatically.

The corpus can be used either for analysis of human communication or as training data for machine-learning-based systems. The use of a video annotation tool such as ELAN enables users to browse through the dialogue segments efficiently based on the corpus annotation (Figure 1). Furthermore, by using a video processing tool, users can extract dialogue video segments based on the annotation.

*Corresponding author

ICMI'16, November 12–16, 2016, Tokyo, Japan
ACM. 978-1-4503-4556-9/16/11...$15.00
http://dx.doi.org/10.1145/2993148.2998523

[1]http://i.yz.yamagata-u.ac.jp/moviedialcorpus/index.html

Table 2: Statistics of annotations

Property	Value
Total number of non-dialogue segments	149,211
Total number of dialogue segments	149,689
Average non-dialogue segment duration	21.31 sec.
Average dialogue segment duration	28.11 sec.

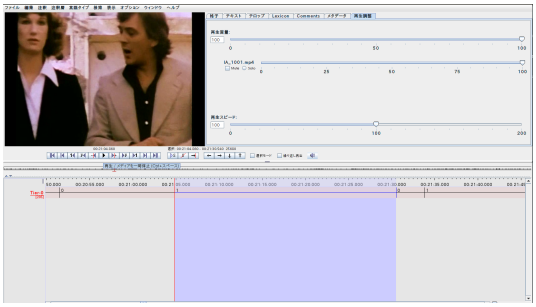

Figure 1: Example annotation displayed on ELAN.

3. METHOD

3.1 Procedure

The corpus construction was conducted by following four steps: **Crawling:** Movie files of the specified types were downloaded from the web site. **Pre-processing:** The audio data were extracted from the video files and converted into audio feature data. **VAD:** The spoken segments were identified based on the classification of audio data for each time slice. **Post-processing:** The identified speech segments were converted into dialogue segments. We explain each step below.

3.2 Crawling

We used feature films hosted in the Internet Archive[2]. The movies are provided under creative commons licenses and were targeted for collection as follows. Each movie is associated with text tags, which often refer to the genres of films. First, we selected the tags that correspond to those of the major movie genres listed in an Internet movie database[3]. As a result, 22 tags were identified. Next, all movie files associated with these tags were downloaded, and in this way we obtained 1,722 video files.

3.3 Pre-processing

The audio files were separated from the downloaded video files. The MFCC (mel-frequency cepstral coefficient) features were extracted from the audio files.

3.4 VAD

The audio data were fed into the feed-forward DNN classifier [3]. Given an audio feature for an audio frame, the classifier generates as output the likelihood of an audio file belonging to one of three classes, namely speech, non-speech noise, and silence. If the likelihood of one speech class is higher than that of the other two classes, the frame is assigned a label as the voiced segment. The classifier had been trained on about 2.5 hours of acoustic data from several episodes of a variety TV program. The data were manually annotated with voiced segments.

3.5 Post-processing

The outputs of VAD are often fragmented due to speaking behavior such as pauses in utterances. Therefore, we applied a smoothing method to combine small voiced segments into chunks of voiced segments that are slightly longer. After obtaining utterance segments, we selected usable ones. Since we are interested in dialogues, rather than isolated utterances, we removed voiced segments with a duration of less than 5 seconds.

4. DISCUSSION

4.1 Related Works

Other researchers created some movie dialogue datasets before our present work. For example, the scripts of 753 movies, from which 132,229 dialogues were extracted[1], were used. Movie scripts were also used to construct the corpus in [2]. However, both of the works are based on text scripts that do not directly correspond to the video data as a multimodal information source.

The movie files from the Internet Archive have been used as the test collection in TRECVID[4]. However, the videos that were used are short, several minutes long, and the collection is intended for the benchmarking of known item detection or semantic indexing tasks and is not usable as a dialogue corpus.

5. CONCLUSIONS

We developed a multimodal dialogue corpus that consists of dialogue segments taken from movie files in the public domain. The corpus enables us to obtain a large number of dialogue segments either for communication analysis or machine learning purposes. Although the corpus is intended to be a collection of dialogues, current data can include monologues and narrations as well. Therefore, future plans include the addition of participant information to each dialogue based on visual analysis and scene segmentation information.

6. REFERENCES

[1] R. E. Banchs. Movie-dic: a movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 203–207, Jeju, Korea, 2012.

[2] J. E. S. Marilyn A. Walker, Grace I. Lin. An annotated corpus of film dialogue for learning and characterizing character style. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012.

[3] I. Suga, R. Yasuhara, M. Inoue, and T. Kosaka. Voice activity detection in movies using multi-class deep neural networks. In *the 5th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan*, Honolulu, Hawaii, November 2016.

[2]https://archive.org/details/feature_films

[3]http://www.imdb.com/genre/

[4]http://trecvid.nist.gov/