A Large-Scale Multimodal Movie Dialogue Corpus

Ryu Yasuhara, Masashi Inoue*, Ikumi Suga, Tetuo Kosaka
Yamagata University
m.inoue@acm.org

Abstract

We present an outline of our newly created multimodal dialog 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 segments taken from 1,722 movies. These dialogues are automatically segmented by using deep neural network-based voice activity detection with filtering rules. Our corpus can reduce the human workload and machine-processing effort required to analyze human dialogue behavior by using movies.

Corpus Overview

- The corpus consists of two parts: movie files and annotations.
- Movie files are hosted under creative commons license by the Internet Archive. They were categorized into genres.
- Annotation files specify dialogue segments in each movie file by start time, end time, and segment label.

Statistics of movies

Source movies	1,722
Total duration	2050.85hours
Average duration	1.2hours
Movie genres	22
Single genre movies	1066
Multi genre movies	656

Statistics of annotations

Total number of dialogue segments	149,689
Total duration of dialogue segments	1168.87houes (4,207,917 sec)
Average dialogue segment duration	28.11 sec
Average number of dialogue segments per movie	87

Distribution of movie genres

Action	**Adventure**	**Animation**	**Biography**	**Comedy**	**Crime**	**Documentary**	**Drama**	**Fantasy**
16	28	11	3	166	52	28	144	7
Filmnoir	**Horror**	**Music**	**Musical**	**Romance**	**Sci-Fi**	**Thriller**	**War**	**Western**
22	146	6	44	13	79	24	16	261
Sports	**Family**	**Mistery**	**History**	**Multi genres**				
Not existed as a solo genre				656				

Method

Dialogue segments were automatically detected by a deep neural network-based voice activity detection. Problematic segments were removed by heuristics filtering rules.

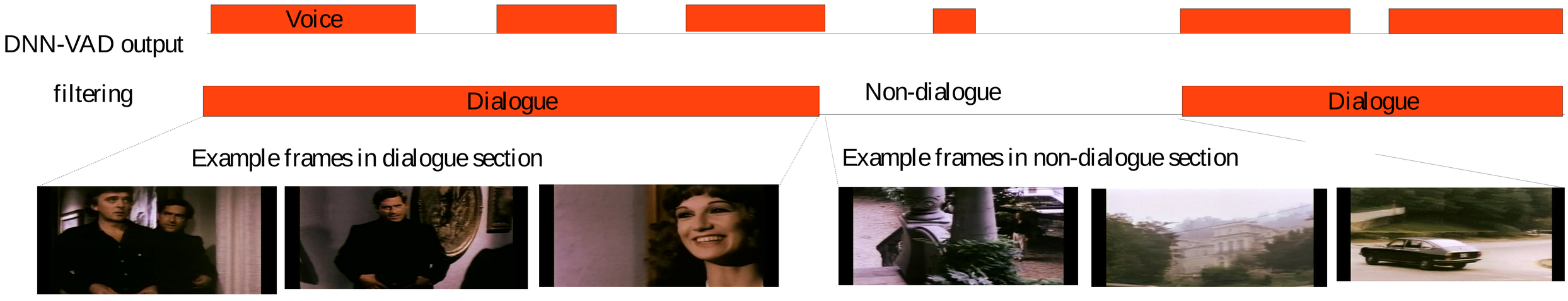
DNN-VAD

- Training data: The feed-forward DNN sound models were trained on about 2.5 hours of a variety TV program.
- Input : MFCC extracted from each frame of the audio track.
- Output : The likelihoods for speech, non-speech, and silence.
  l(speech) > {l(non-speech), l(silence)} → “voice” label
- Smoothing : combine fragmented voiced segments into one voiced segment.

Filtering rules

Two types of filtering were applied to the VAD results to focus on dialogues than voiced segments. .

- Concatenation : combine multiple voiced segments into one dialogue.
- Removing : Small voice segments were removed that are considered isolated utterances such as sigh or shout.



Accuracy of automatic dialogue detection

Two randomly chosen movie were used to measure the accuracy of automatic dialogue segment detection. Test segments were sampled at a regular interval. Ground truth is manually provided. Movies with single genre tags were considered.

Action	**Adventure**	**Animation**	**Biography**	**Comedy**	**Crime**	**Documentary**	**Drama**	**Fantasy**
0.93	0.86	0.91	0.82	0.86	0.93	0.89	0.89	0.81
Filmnoir	**Horror**	**Music**	**Musical**	**Romance**	**Sci-Fi**	**Thriller**	**War**	**Western**
0.84	0.81	0.50	0.46	0.82	0.86	0.91	0.92	0.85

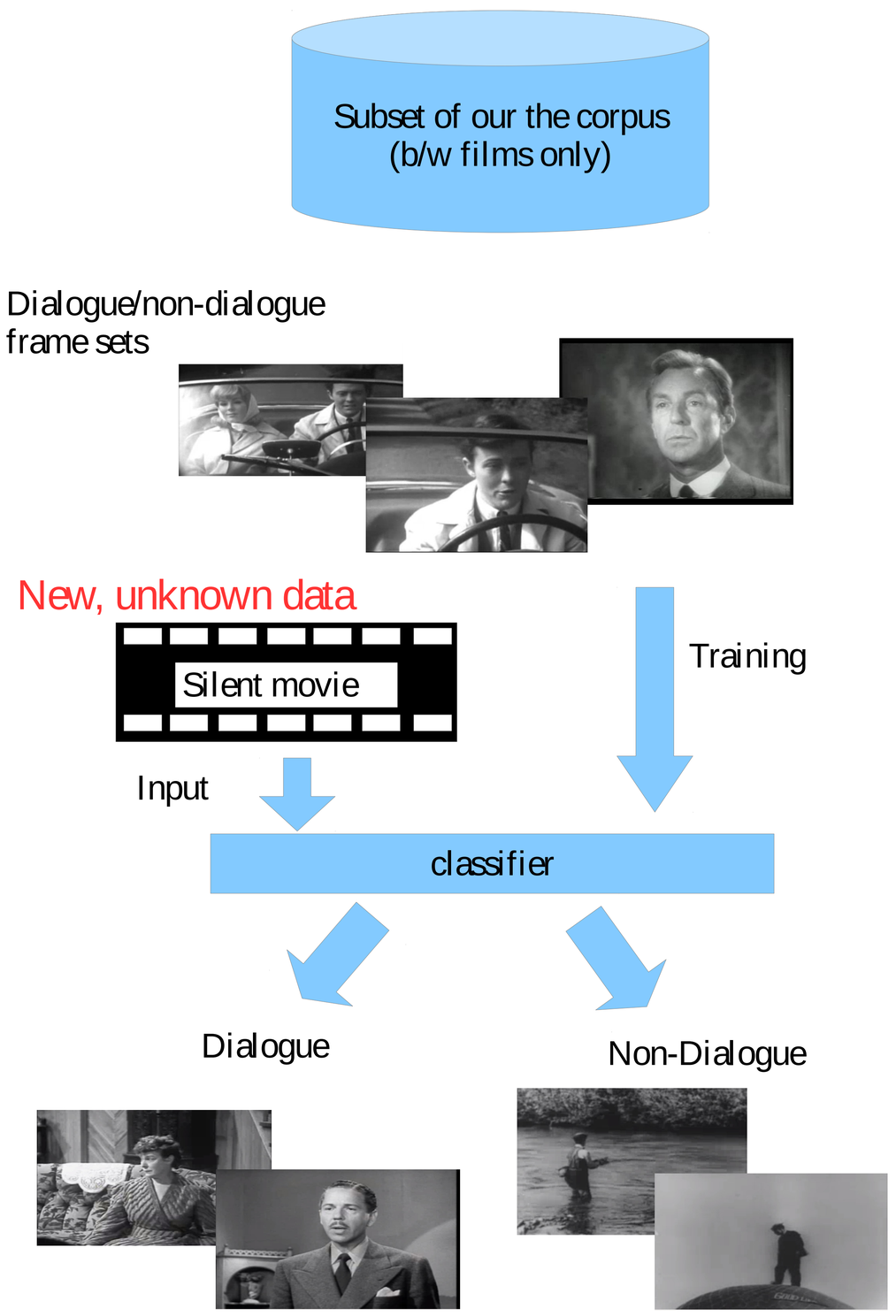
A Large-Scale Multimodal Movie Dialogue Corpus

Ryu Yasuhara, Masashi Inoue*, Ikumi Suga, Tetuo Kosaka
Yamagata University
m.inoue@acm.org

Possible Use Cases

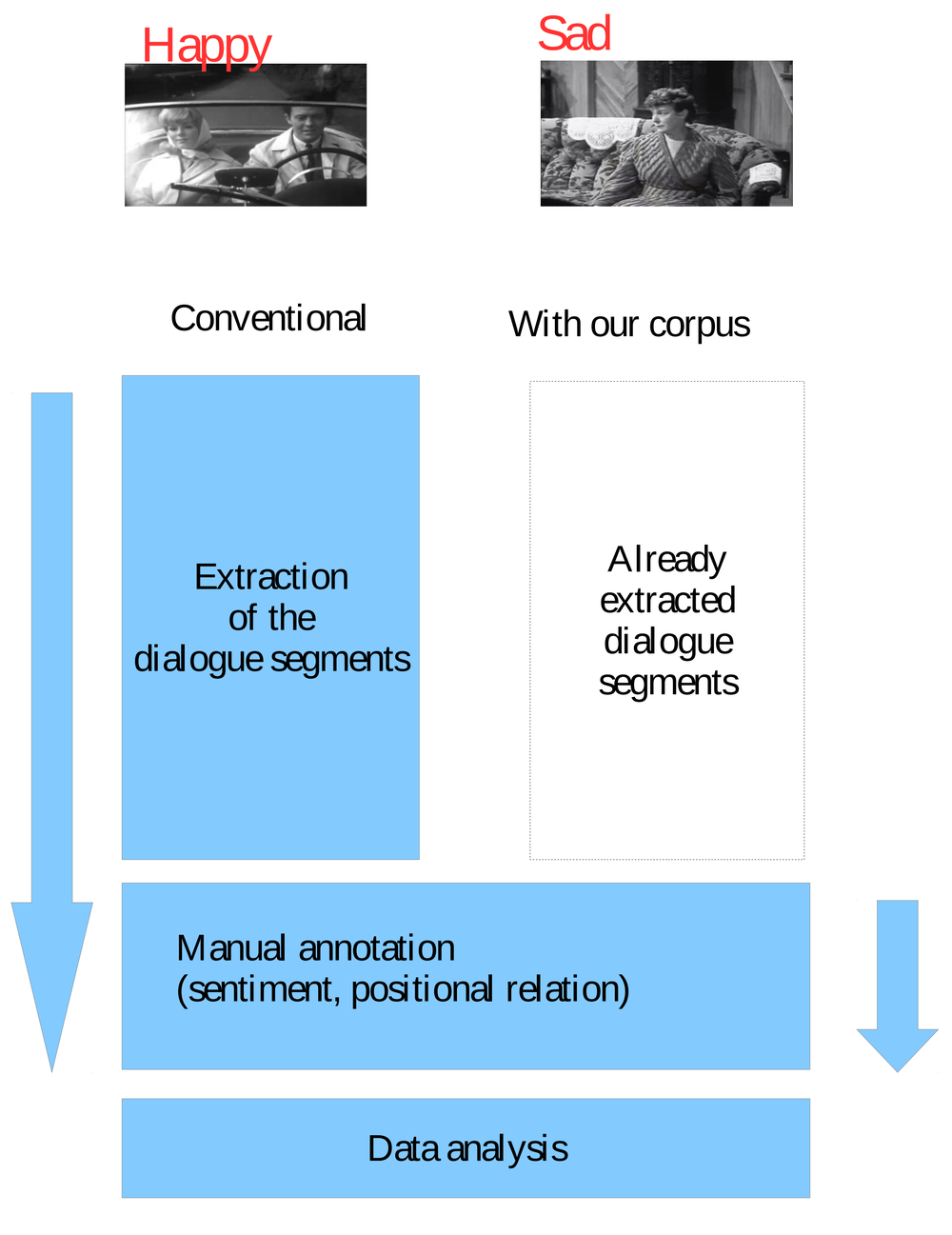
Training data for dialogue detection from silent movies (Digital Humanities)

Dialogue or speech segments can be extracted using audio information relatively easier than using visual information when the task is conducted in unsupervised form. However, many historical films are in silent format. For the computational analysis of them, dialogue segment extraction should be done automatically by using visual features. For the purpose our corpus can be used as the training data for the classifier.



Examples for analyzing dialogue sentiment and position (Computational Psycholinguistics)

When we want to understand the dialogue data, automatic analysis is not always possible especially when the needed annotation is semantic ones. For example, when we want to understand the relationship between physical positions during conversation and the atmosphere of the dialogue, we may need to annotate positional relation between interlocutors and dialogue sentiment either by hand or by additional algorithms. Such work can be carried out efficiently when we have to deal with only dialogue segments since the extraction of the dialogue segments is computationally intensive and data preparation is costly.



Data distribution

- We provide full annotation data, a python script to download public domain movie files, and example dialogue videos extracted from the films. Please visit the following link below for details.
- Corpus improvement after the publication is described and the new version of the corpus is provided there. Please see the revised proceedings paper for the changes.
- We are planning to enhance the current corpus with additional annotation and increased segmentation accuracy. If you have any suggestions for the improvement or the utility in your application, we appreciate your input.

http://i.yz.yamagata-u.ac.jp/moviedialcorpus/

YASUHARA
Ryu

INOUE
Masashi