2H5-OS-35c-1in1

映画からのマルチモーダル対話コーパスの作成

Building multimodal dialogue corpus from movies

井上雅史 Masashi Inoue　安原龍 Ryu Yasuhara　菅郁巳 Ikumi Suga　小坂哲夫 Testuo Kosaka

山形大学
Yamagata University

We present an outline of our 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 dialogue segments taken from 1,722 movies. These dialogues are automatically segmented by using deep neural network-based voice activity detection with filtering rules. The total duration of the original movie files is 2,065 hours. Our corpus can reduce the human workload and machine-processing effort required to analyze human dialogue behavior by using movies.

1. はじめに

人間のコミュニケーションにおけるパターンを一般的に理解するためには，大量の対話データが不可欠である．しかし，対話データを収集するにはコストがかかり，自由に利用可能なマルチモーダル対話コーパスはほとんどない．また，実験室環境ではなく日常の様々な場面での対話を収集することも課題となる．対話収集コストを回避しつつ多様な対話場面を対象とするために，新たに対話データを収集するのではなく，映画などの記録された対話を使用することも考えられる．しかし，映画の大部分は対話と無関係であり，データを対話コーパスとして使用する場合，対話区間の開始時間と終了時間を示すアノテーションが必要である．対話区間の手動抽出はコストがかかりすぎるため，自動処理が必要になる． 本研究では，ディープニューラルネットワーク（DNN）ベースの音声区間検出（VAD）を適用することで，対話区間をある程度の精度で自動抽出し，コーパスを構築することができた．

2. コーパスの概要

コーパスの概要は，表 2. 上部に示されている．コーパスは，映像ファイルとアノテーションの 2 つの部分で構成され，映像ファイルは元のアーカイブサイトからダウンロードでき，アノテーションファイルはプロジェクト Web ページから提供されている *1．アノテーションファイルは，映像区間の開始時間，終了時間，およびラベル（対話であるか否か）からなる 3 列の CSV ファイルである．

コーパスの元となった映画の統計情報は，表 2. 下部に示されている．配布データには映像ファイルの URL リストが含まれており，スクレイピングスクリプトのサンプルも提供されているので，利用者は必要な映像ファイルを，自動的にダウンロードすることができる．

コーパスは，人間のコミュニケーションの分析のため，または機械学習ベースのシステムの訓練データとして使用することができる．ELAN などのビデオアノテーションツールを使用すると，コーパスのアノテーションに基づいて対話区間を効率的にブラウズできる（図 1）．

表 1: コーパスの概要

映画総数	1,722
映画の平均時間	1.2 (h)
映画の合計時間	2,065 (h)
映画のジャンル	22
会話区間数	502
平均会話時間	14.8 (s)
平均非会話区間時間	9.18 (s)

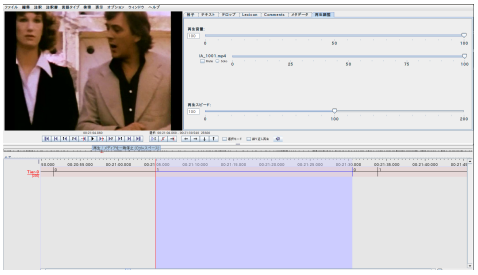

図 1: ELAN によるアノテーション閲覧例

3. 手法

3.1 手順

コーパスの構築は，以下の 4 つの手順で行われた．

クローリング：動画ファイルを Web サイトからダウンロード．
前処理：音声データをビデオファイルから抽出．
VAD：有声区間を音声データの各時刻において識別．
後処理：識別された有声区間を対話区間に変換．

以下に手順の各ステップを説明する．

連絡先: m.inoue@acm.org

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

3.2 クローリング

Internet Archive [*2] でホストされている，パブリックドメインに属す長編映画を使用した．各映画はタグ付けされており，いくつかのタグは，ジャンル情報である．そこでまず，インターネットムービーデータベース [*3] に挙げられている 22 の主要な映画ジャンルに対応するタグを選択した．次に，これらのタグに関連付けられたすべての映像ファイルをダウンロードし，1,722 の映像ファイルを取得した．

3.3 前処理

ダウンロードした映像ファイルから音声ファイルを分離し，MFCC（メル周波数ケプストラム係数）特徴を抽出した．

3.4 VAD

音声データは，フィードフォワード DNN 分類器 [4] に供給された．音声データの各フレームでの音声特徴が与えられると，音声区間検出（VAD）ように構築されたフィードフォワード DNN 分類器 [4] は，3 つのクラス，すなわち音声，非音声ノイズ，および無音の尤度を算出する．音声クラスの尤度が他の 2 つのクラスの尤度よりも高い場合，フレームには有声区間としてのクラスラベルが割り当てられる．今回使用した分類器の訓練には，日本語によるバラエティ番組のいくつかのエピソード約 2.5 時間分の音響データを用いた．

3.5 後処理

VAD の出力は，発話の休止などの行動によって小区間に断片化されることが多い．そこで，スムージングを適用して小さな音声区間候補を，大きな音声区間のチャンクである発話区間として結合した．また，発話区間を得た後，使用可能な区間を選択した．今回は，孤立した発話ではなく対話に興味があるので，持続時間が 5 秒未満の発話区間を削除した．

4. 考察

4.1 関連研究

映画からの対話データセット作成の試みは，これまでにもいくつかある．例えば，753 本の映画の脚本を使用して，132,229 の対話を集めたコーパスがある [1]．脚本を用いたコーパスには，登場人物に焦点を当てたものもあり，862 本の映画から 7,400 の登場人物についての対話を集めている [3]．これらのどちらも，脚本が情報源であるため映像データと対応付いておらず，マルチモーダルコーパスとはなっていない．

インターネットアーカイブ上の映画データは，TRECVID[*4] のテストコレクションとして使用されている．しかしこのテストコレクションは，数分程度と短い映像からの既知のアイテム検出または意味的索引付けタスクのベンチマークを目的としており，対話コーパスとしては使用できない．

4.2 限界

現在のコーパスには以下の限界がある．まず，映画内での使用言語の情報が存在していない．対象となったほとんどの映画は英語で会話が行われていると思われるが，使用言語が英語ではない映画がいくつかあり，映画中で複数言語を使用している可能性もある．従って，ユーザが言語固有の現象に注目したい場合には，言語情報を取得するために音声による言語判定または映画情報データベースの検索による使用言語同定などの前処理ステップを追加する必要がある．

∗2 https://archive.org/details/feature_films
∗3 http://www.imdb.com/genre/
∗4 http://www-nlpir.nist.gov/projects/trecvid/trecvid.data.html

表 2: 映画ジャンルごとの対話アノテーションの精度

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

第 2 の限界は，対話区間検出の精度である．対話区間のアノテーションの付与は精度志向で行っているため，いくつかの対話を取り漏らしている可能性がある．さらに，使用された VAD 法の特性のために，歌の音楽音が音声音と混同されがちである．表 2 に，サンプリングした評価データに基づく，各ジャンルごとの判定精度を示す．音楽映画（Music）やミュージカル映画（Musical）において，ある程度の数の非対話区間が，対話としてアノテーションに含まれてしまっている．この評価に用いた評定データも，コーパスとは別に提供している．

第 3 の限界は，コーパス中の対話が演技に基づくものであることである．多くの映画において会話は俳優によってなされているため，実際の対話と似ているものの，自発会話と異なる性質が存在する可能性がある．利用目的によっては，日常場面での実際の会話を収録したデータ [2] などとの比較によって，対話の性質の検証も必要となる可能性がある．

5. おわりに

パブリックドメインの映画から抽出した対話区間からなるマルチモーダル対話コーパスを構築した．このコーパスを，大規模な対話データが必要となる，コミュニケーション分析や機械学習ために使用できると考えている．今後の計画として，映像の視覚的分析とシーン分割情報に基づいて，各対話区間に参加者情報を追加することを考えている．

参考文献

[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, July 2012.

[2] H. Koiso, T. Tsuchiya, R. Watanabe, D. Yokomori, M. Aizawa, and Y. Den. Survey of conversational behavior: Towards the design of a balanced corpus of everyday japanese conversation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, May 2016.

[3] 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.

[4] 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.