東北工業大学地域連携センター・研究支援センター紀要 EOS
Vol.33 No.1 p.15 ～ p.18, 2021

令和元年度学内公募研究（萌芽型）
〔研究論文〕

ビデオ内会話シーンアノテーション基準の作成

井上　雅史 [1)]

A coding scheme for the annotation of conversation scenes in videos

INOUE, Masashi [1)]

Abstract

映画映像に対して，会話の形式のアノテーションを実施するための作業手順及び作業内容について，現実的な作業時間で作成可能な基準を決定した。会話形式として既存研究で設定されている４種類に加えて，６種類の会話区分と，会話と紛らわしい３種類の音声活動区分を設計した。映画２作品に対する試行の結果，事前に設定した区分に収まらない「その他」と判定される事例を，１～２割程度に抑えることができた。

１　はじめに

人間の会話を理解するためには, 会話事例を集積したデータベースの構築が欠かせない。特に，映像として音声や身体動作，表情などのさまざまな情報を利用可能な，マルチモーダルコーパスを整備することが重要となる。しかし，会話の収録にはコストがかかり，多様な状況下での会話を映像として収集することは容易ではない。本研究では，大量の映画データから多様な会話を収集するため, 大量のパブリックドメインの映画をコーパスとし，会話シーンを抜き出して，その種類をラベル付けするアノテーション基準を導入する。

２　アノテーション内容

アノテータ（作業者）は，１分ごとに区間を区切られた映画を最初から最後まで見て，登場人物が会話している（言葉でコミュニケーションを取っている）場面について，映像や声色，しぐさなどから，その会話に合うラベルを選択する。音声を聞き取って内容を理解する必要はなく，映像からの印象で判定する。区間は時間によって機械的に分けられているので, ストーリー上中途半端なところから始まったり終わったりしていることがある。劇中劇のような，登場人物が別の役になりきって行われる会話も，通常の会話と同様にラベル付けする。

１）工学部　情報通信工学科
Department of Information and Communication Engineering

３　ラベルの種類

３．１　基本手続き

ラベルは表１の13種類である。どれか１つだけ選択するのではなく，同一区間で当てはまるもの全てを選択する。個々の発話に対してではなく，会話を構成する一連のやり取り全体に対して付与する。１から４のラベルの判断基準を図１に示す。５から９のラベルにおいては，情報の出し手（話し手）の発話だけでなく，それに対する情報の受け手（聞き手）の反応（例えば返事や相槌など）も，会話を構成する要素となる。したがって，直前の区間で行われた発話への返事や相槌などの反応のみを含む区間に対しても，それ以外のラベルと合わせて，５から９のラベルを付与する。

３．２「13. その他」について

明らかに会話だが１～１２のどのラベルにも当てはまらないものや，そもそも会話かどうか判断に迷う場合は，このラベルを付与したうえで，スプレッドシートのメモ欄にどのような内容であるか具体的に書く。判断に迷う理由も可能な限り記述する。

３．３　ラベルを複数付ける場合

判断に迷う場合や複数のラベルどちらにも当てはまる場合は，複数のラベルを付けることができる。ただし，同じように複数のラベルが付いていても，１分間の中で会話の種別が切り替わる場合と，会話の分類としてどちらもあり得る場合とがありえるので，そのどちらの理由で複数のラベルが付いているのかをメモ欄に記載する。

表１　会話形式のラベル

番号	形式	説明
1	雑談	会話の目的や話題などがあらかじめ定められていない会話。
2	用談・相談	会話の目的はある程度決まっているが時間や場所などは定められていない会話（５から９にあてはまる場合を除く）。
3	会議・会合	会話の目的が決まっていて，なおかつ時間や場所などが定められている会話（５から９にあてはまる場合を除く）。
4	授業・レッスン・講演	先生や講演者など会話の流れを導く人物がいる場での会話。
5	主張・独白	自身の意見や内面を伝えようとして，情報伝達が一方向的に行われる会話。ただし４ではないもの。相手のいない独言は11のモノローグとなる。
6	喧嘩・叱責	相手に対する否定的な感情をぶつける会話。
7	挨拶	挨拶の発話やそれに引き続く形式的な会話。
8	依頼・懇願・謝罪	話し手の利益になるような聞き手の好意的な反応を期待して行われる会話。聞き手や第三者のためを思った説得などは含まない。
9	指示・命令	聞き手が話し手の伝えた内容に従って行動することを期待して，一方向的な情報伝達を行っている会話（○○を作れ，○○へ行って，など）。
10	ナレーション	映画の登場人物以外による語りで，視聴者を聞き手に想定したもの。ただし，登場人物が話していても，内容がナレーション的ならこのラベルを付ける。語り手が不明であってもよい。

11	モノローグ	映画の登場人物による語りで，聞き手がその場にいないもの。登場人物の内言（頭の中で考えていること）の音声化や手紙の読上げを含む（10 のナレーションと区別する）。
12	歌	映像中の登場人物によるその場での歌唱あるいは BGM としての歌。ただし，ミュージカルやオペラなどでの歌に乗せた会話は，12 以外のラベルを付与する。
13	その他	上記以外。

目的や話題が決まっていない　→　**1．雑談**
↓　決まっている
話し合う前提で集合していない　→　**2．用談・相談**
↓　集合している
司会的人物が主にしゃべっていない　→　**3．会議・会合**
司会的人物が主にしゃべっている　→　**4．授業・レッスン・講演**

図１　ラベル１～４判定チャート

４　アノテーション手順

作業はアノテーションソフトウェアである ELAN [1] を中心にして行う。アノテータが ELAN の操作に不慣れな場合は，ELAN を動画再生に用い，ラベリングはスプレッドシート上で行うこととする。その場合は，区間調整や注釈の編集などを ELAN 上でする必要がなくなる。

ELAN の注釈層に，１分ごとに「Ｌ１」「Ｌ２」など「Ｌ○○」という区間番号が表示されるよう，事前に分割したファイルを用意しておく。アノテータは各区間を視聴し，スプレッドシートで同じ番号が書かれている行を探して，当てはまるラベルの欄に会話形式の番号を入力する。

アノテーション会話区間を図２下段の注釈層の最上位のように１分ごとに分割しているのは，会話区間を切り出す作業負荷が高いことから，作業時間短縮を意図した処置である。このような一定間隔の分割の事例として，身体動作のアノテーションにおいて，より短い１秒ごとの分割が用いられているが [2]，会話形式の判定には１分程度のまとまりが適当であろうと考えた。理想的には図２下段の注釈層の最下位層のように，会話の区間を切り出しておくのがよい。これらの区切り方の違いによる時間の差分を 20 分弱の映像について算出したところ，合計で 7.8 秒ほどのずれとなっていた。

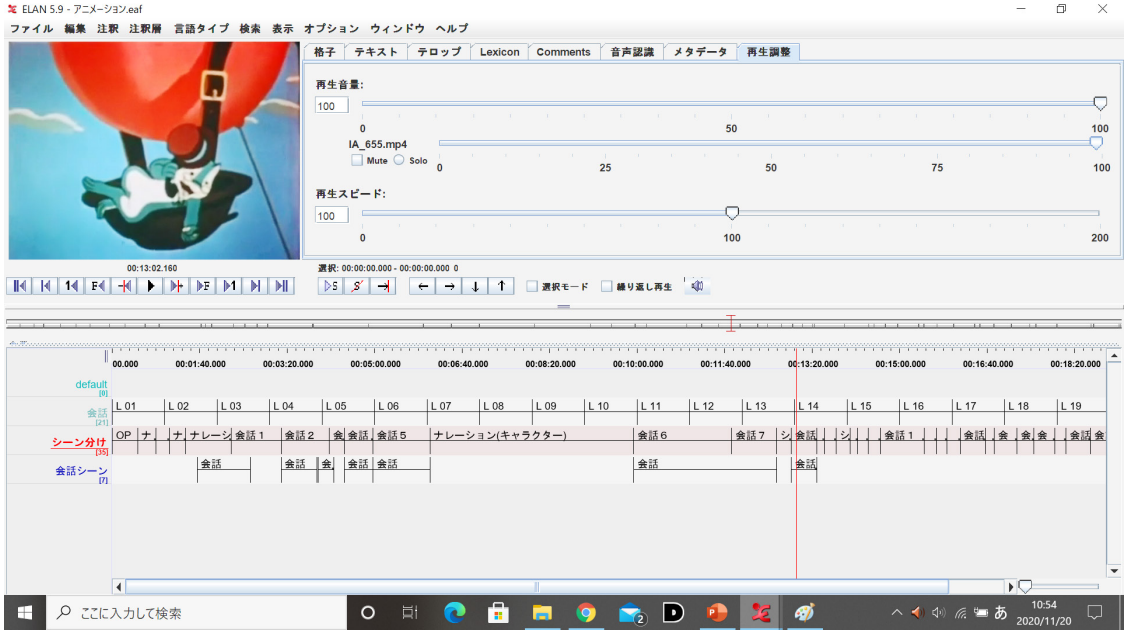

図２　ELAN 上でのアノテーション例。左上の映像に対して下のパネルで注釈を付与。

5　おわりに

大量の映画データから多様な会話を収集するため，会話シーンを抜き出して，その種類をラベル付けするアノテーション基準を導入した。この基準を未経験のアノテータ３名において，映画２作品に対して試行した結果，事前に設定した区分に収まらない「その他」と判定される事例を，１～２割程度に抑えることができた。今後は，より多くのデータでの利用可能性を確認することと，得られたラベル分布の性質の分析が望まれる。

6　参考文献

[1] Brugman, H., Russel, A. (2004). Annotating Multimedia/Multi-modal resources with ELAN. In: Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation.

[2] 牧野 遼作　ら,「実世界における身体動作のコーディング・セグメンテーション手法の提案」，第 29 回人工知能学会全国大会（2015），セッション ID 3O1-2in