4P-07

多言語展開された同一 YouTube コンテンツのコメントを対象とした言語ごとの特徴語抽出方式

西山　実希† 岡田 龍太郎† 峰松 彩子† 中西 崇文†
武蔵野大学データサイエンス学部データサイエンス学科†

1. はじめに

近年，YouTube などの動画配信サイトにおいて，多言語に翻訳された動画コンテンツが配信されており，それぞれの言語で書かれたコメントが散在している．これらの多様な言語で書かれた同一動画コンテンツのコメントを対象として，言語ごとで特徴となる単語を抽出し，比較することができれば，それぞれの言語ごとで異なる文化背景を明らかにすることが可能になると考えた．

本稿では，多言語展開された同一 YouTube コンテンツのコメントを対象とした言語ごとの特徴語抽出方式について示す．本方式では，言語ごとに分けられたコメント群を対象として，それを比較することで，同一動画コンテンツを閲覧したユーザの考え方の違いの考察を試みた．本方式により，多言語展開された同一 YouTube コンテンツにおいて，コメントから国や言語ごとの違いを比較することを可能とし，ユーザの文化や趣味嗜好に基づいたコンテンツを提供することを可能とする．

2. 関連研究

伊藤ら[1]は，Twitter データを用いて，位置情報付き日本語ツイートから単語ごとに算出された地域依存度による重みづけを行い，地域特徴語を抽出する手法を示している．

堺ら[2]は，YouTube における炎上動画のメタデータを取集し，炎上動画を分類するシステムを実現している．また，炎上動画のコメントを分析し，特徴抽出することで各ユーザーのコメントが他と区別できる特徴を持つか否かを検証している．

本稿では，YouTube データを用いて，言語が異なる同一動画コンテンツのコメントから言語ごとの特徴語を抽出することにより国や言語ごとの違いを比較することを試みている．

Language-specific feature word extraction method for comments on the same YouTube content deployed in multiple languages
Mitsuki Nishiyama†, Ryotaro Okada†, Ayako Minematsu†, Takafumi Nakanishi†
† Department of Data Science, Musashino University

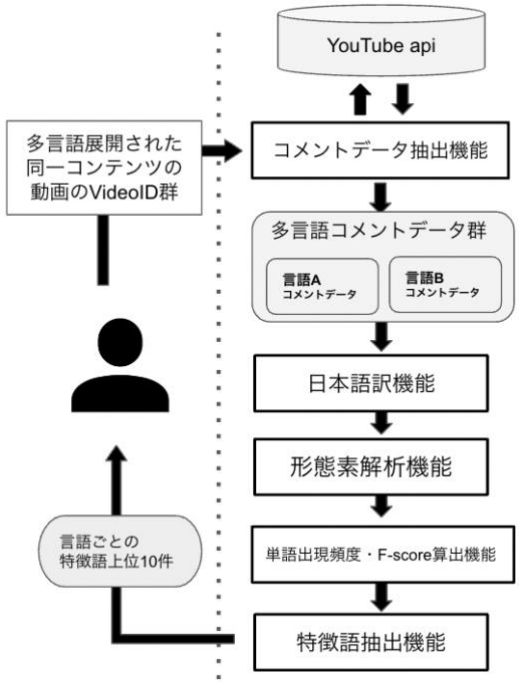

図 1. 本方式の全体像

3. 提案方式

3.1 全体像

本提案方式の全体像を図 1 に示す．本方式は，コメントデータ抽出機能，日本語訳機能，形態素解析機能，単語頻出頻度・F-score 算出機能，特徴語抽出機能からなる．

3.2 多言語展開された同一コンテンツの動画の VideoID 群

本稿では，YouTube 動画において異なる言語で吹き替え・字幕を用いた翻訳をされた，内容が同一の動画の VideoID 群を入力とする

3.3 コメントデータ抽出機能

YouTubeAPI を通じて，3.2 節の入力された VideoID 群からコメントについた返信を含めた全てのコメントデータを取得する．また，取得したデータを言語・国ごとに格納を行う．

3.4 日本語訳機能

言語を統一するため，3.3 節で収集した日本語以外の言語のデータを対象として DeepLAPI[3]を用いた日本語訳を行う．

3.5 形態素解析機能

3.4 節で日本語訳したデータを含めた全てのコメントデータを対象として形態素解析を行い，名詞，動詞，形容詞のみをそのコメントの重要な単語として抽出を行う．

3.6 単語出現頻度・F-score 算出機能

3.5 節で残った単語群を用いて算出された単語出現頻度から，Scaled F-score を用いた F-score の算出を行う．これにより，語句の重要度や特徴的な度合いを評価し，言語ごとで特徴となる単語を抽出することが可能となる．

3.7 特徴語抽出機能

3.6 節で算出した単語出現頻度，F-score を用いて言語ごとに特徴語抽出を行う．これにより，言語ごとの特徴語の比較が可能となる．

4. 実験

4.1 実験目的

本システムの有効性を検証するため，多言語対応している YouTube 動画から国ごとのコメントによる反応の違いを抽出できることを確認する．

実験として，入力する動画として YouTube 上に公開されている多言語展開しているユニバーサル・ピクチャーズ公式チャンネルの「ジュラシックワールド/新たなる支配者」という映画の日本，アメリカ，オーストラリア，インド，フランスのそれぞれの国に対応した言語のオフィシャルトレーラーを用いた．各動画の VideoID を表 1 に示す．

4.2 実験結果

表 2 に，国ごとの動画から抽出された特徴的な単語を F-score の高い順に 10 件示す．

表 2 より，日本の言語に対応した動画では，「アツ」「面白い」「w」という感情を表す単語が抽出されていることがわかる．これにより，日本では感情を込めたコメントをする傾向にあると考えられる．アメリカの言語に対応した動画では，「ジェフ」「サム」「キャラクター」という登場人物に関係する単語や，「フランチャイズ」「フィナーレ」などの作品情報に関する単語が抽出されていることがわかる．これにより，アメリカでは作品の情報についてのコメントをする傾向があると考えられる．オーストラリアの言語に対応した動画では，「イエス」「神代わり」「信仰」などの宗教に関する単語が抽出されていることがわかる，これにより，オーストラリアでは宗教に関するコメントをする傾向にあると考えられる．インドの言語に対応した動画では，「マラヤーラム語」「カンナダ語」などの言語に関する単語が抽出されていることがわかる．これにより，インドでは公用語として使用されている英語やヒンディー語の他にも多くの言語が使用されており，実験として用いた動画では英語が使われているため，マラヤーラム語やカンナダ語を普段使用する人たちからの吹き替え言語についての要望がコメントされているのだと考えられる．フランスの言語に対応した動画では，「スピノサウルス」「ギガノトサウルス」という恐竜に関する単語が抽出されていることがわかる．これにより，フランスでは登場する恐竜に着目したコメントをする傾向があると考えられる．このように，異なる言語を持つ国によって異なる考え方を持つことがわかる．また，アメリカとオーストラリア，インドの結果から，同一言語であっても使われている国や地域によって異なる文化背景を持つことがわかる．

表 1：各動画の VideoID

	JP	US	AU	IN	FR
VideoID	vjXsoPKBvoY	fb5ELWi-ekk	jseFW09claA	qhxYb2O0xtk	x0Gf4OkxeDs

表 2：各国の特徴となる単語

JP	US	AU	IN	FR
アツ	正直	イエス	原発事故	恬撹
初代	懐かしい	落ちる	indea	リチャード
哺乳類	フランチャイズ	神代わり	ピクチャーズ	可能
劇	フィナーレ	信仰	マラヤーラム語	ボア
予告	以来	犠牲	カンナダ語	ダル
w	ジェフ	私たち	炎	スピノサウルス
変わる	キャラクター	死ぬ	ユニバーサル	見つける
vs	サム	生命	吹き替え	ギガノトサウルス
チャンネル	パーク	救い	鳥肌立つ	Yes
面白い	物語	罪	アレン	生きる

5. おわりに

本稿では，多言語展開された同一 YouTube コンテンツのコメントを対象とした言語ごとの特徴語抽出方式について示した．本方式を用いることにより，多言語展開された同一 YouTube コンテンツにおいて，コメントから国や言語ごとの違いを比較することを可能とする．また，ユーザの文化や趣味嗜好に基づいたコンテンツを提供する一助となる可能性がある．

参考文献

[1] 伊藤 晶，荒川 豊，田頭 茂明，福田 晃，Twitter からの地域特徴語の自動抽出に関する一検討，情報処理学会第 75 回全国大会講演論文集，pp. 101-102，2013.

[2] 堺 雄之助，伊藤 栄典，動画サイトにおける視聴者コメントの特徴抽出，人工知能学会研究会資料 知識ベースシステム研究会，pp. 17-22，2021

[3]DeepLAPI ， https://www.deepl.com/ja/pro-api?cta=header-pro-api/