様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月 １７ 日現在

機関番号：９４３０５
研究種目：基盤研究(B)（一般）
研究期間：2021～2023
課題番号：２１Ｈ０３５０５

研究課題名（和文）動画談話構造解析とそれを用いた要約生成

研究課題名（英文）Discourse parsing for videos and its application to summarization

研究代表者
平尾　努（Hirao, Tsutomu）

日本電信電話株式会社ＮＴＴコミュニケーション科学基礎研究所・協創情報研究部・主任研究員

研究者番号：４０３９６１４８

交付決定額（研究期間全体）：（直接経費）　　13,300,000円

研究成果の概要（和文）：ストーリを伝える動画にはいくつかのイベントが含まれており，それらが関連を持つことで動画全体のストーリを伝える．こうしたイベントの間に成り立つ関係を解析することは，動画の内容理解，要約や Video QAといった下流タスクの性能を向上させるために役立つ．本研究課題では，動画の背後にあるストーリ構造を修辞構造理論に基づく木としてとらえ，解析器の訓練・評価のためのデータセットを構築し，ベースライン解析器の性能を調べた．その結果，テキスト知識を解析器のエンコーダに転移することが有効であることがわかった．また，動画の修辞構造がマルチモーダル要約に役立つことを示した．

研究成果の学術的意義や社会的意義
SNSの発展に伴いインターネット上に投稿される動画は増加の一途をたどっている．しかし，テキストとは異なり，自然言語でそれらを検索することや概要を簡単に把握することは困難であり，人間の情報アクセスを支援する仕組みが必要である．動画の修辞構造を明らかにする研究成果はこうした課題の解決に貢献するという点で大きな意義がある．また，学術的にも視覚と言語の融合に基づく談話構造解析という新しい研究課題であり，その達成に向けた研究成果の意義は高い．

研究成果の概要（英文）：Videos that convey a story contain several events, and the relationships between these events contribute to the overall story of the video. Analyzing the relationships between such events helps improve video understanding and the performance of downstream tasks such as summarization and Video QA. In this research, we represent the underlying story structure of videos as trees based on Rhetorical Structure Theory, construct a dataset for training and evaluating parsers, and investigate the performance of baseline parsers. The results showed that transferring textual knowledge to the parser's encoder is effective. Furthermore, we demonstrated that the rhetorical structure of videos is beneficial for multimodal summarization.

研究分野：自然言語処理

キーワード：自然言語処理　視覚と言語　修辞構造解析

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

SNS の発展に伴い，日常的に投稿される動画が日々増加している．このため，ユーザが動画に効率的にアクセスする技術，その内容を簡単に把握する技術の開発が必要とされている．動画へのアクセスを支援するための方法の一つとして，自然言語をクエリとした動画検索があり，実際に利用されている．しかし，一般的にはユーザが動画に付与したタイトルやタグを検索対象としているため，検索漏れが起こる可能性が高い．さらに検索結果が大量にある場合，人間がすべての動画をみてその取捨選択を行わねばならないという問題がある．一方，ニューラルネットの発展により，文生成技術が大きく進展したことから，動画に対するキャプション生成技術の研究が盛んに行われている．動画中のイベント検出とそれにキャプションを与える dense video captioning (以降 DVC) タスクが提案され，ActivityNet Challenge でのシェアドタスクの一つとしても採用されている．動画中の各イベントに対するキャプションはインデキシングに利用でき，さらにイベントを代表するフレームとキャプションを要約として利用できれば動画を見ることなく内容を把握する助けとなる．しかし，ほとんどのシステムはイベント検出を過剰に行い，一つの動画に対し 100-200 文の大量かつ冗長なキャプションを与える．よって，現状の DVC 技術は動画の検索に利用できるものの，要約としては利用できない．さらに，動画中のイベントの間の関係を同定することができれば「イベント A の結果 B が起きた動画」のように高度な検索が可能となるが，こうした解析技術の研究開発はこれまで行われていない．このように現状では動画へのアクセス支援の需要は大きいもののそれを解決するための技術は発展途上であり，まだ改善の余地がある．

２．研究の目的

そこで本研究課題では，動画の談話構造解析，つまり動画中のイベント間の関係を解析する技術を開発し，人間の動画へのアクセスを手助けする下流タスクである DVC タスクや動画要約の高度化に貢献する基盤技術を確立することを目的とする．以下，個々の課題の目標について述べる．

(1) データセット構築：動画中のイベント間の関係を解析する技術はこれまでにない研究課題であるため，この技術を評価するためのデータセットが存在しない．よって，動画に対しイベント間の関係をあらわすアノテーションを新たに付与する必要がある．本課題では，テキストの談話構造解析の一つである修辞構造理論 (Rhetorical Structure Theory: RST) をベースとしたアノテーションを与えた信頼性の高いデータセット構築を目標とする．具体的には，動画のイベント区間を葉とし，中間ノードをそれが支配するイベントスパンの役割(核または衛星)，枝がイベントスパンの間の修飾関係をあらわす構成素木のアノテーションを与える．また，解析器の性能比較のため，各イベントにはキャプションも付与する．
(2) ベースライン解析器の実現：(1) のデータセットを用いて解析器を実現し，特に，テキスト特徴を利用した場合，映像特徴を利用した場合の解析性能を比較し，動画の修辞構造解析にとってどのような特徴が有効であるかを議論することを目標とする．また，修辞構造解析そのものについてはテキスト・動画に共通するので，テキストの修辞構造解析の高度化にも取り組む．
(3) 動画要約データセット構築：動画対し，フレーム抽出+テキストによる要約を人手で作成し，(1)で作成したデータセットとの関連を調べ，動画の談話構造が要約に与える影響を明らかにすることを目標とする．特に，人間が要約を作成する際に修辞構造を意識しているかどうかを詳しく調べる．

３．研究の方法

研究の目的で述べた 3 つの目標の達成のため，(1) 動画に対して 2 名が独立に修辞構造のアノテーションを与え，その一致度合いを過去の修辞構造のアノテーションに対する一致と比較評価することで信頼に足るデータセット構築が可能であることを示す．(2) ニューラルテキスト修辞構造解析器と同一のアーキテクチャを用いて，エンコーダが映像を入力として受け取ることができるよう改良し，テキスト特徴と映像特徴を用いた場合の違いを議論できるようにする．(3) 修辞構造のアノテーションを与えた場合と同様に，2 名が独立に動画からテキスト要約を作成し，その内容に合致するフレーム抽出を行う．2 名の間のフレームの一致度合い，テキスト要約の一致度合いを過去の研究の一致度合いと比較することで信頼に足るデータセットが構築叶であることを示し，さらに，フレームと(1)のデータの間のアラインメントをとり，抽出したフレーム間に修辞構造が成り立つかどうかを調べる．

４．研究成果

数分程度の短い動画に対し，イベント分割，修辞構造，キャプションのアノテーションを与えた

データセットを作成した．修辞関係についてはテキスト修辞構造を参考にして動画用に新たに 9 種を定義した．50 の動画に対しては異なる 2 名が独立にこれらのアノテーションを与えた．同程度の長さの動画にイベント分割とキャプションのアノテーションが与えられたデータセットである Activity Captions と比較すると動画中に同定されたイベント数が約 2 倍あり，粒度の細かいイベント同定がなされていることが明らかになった．また，作業者間のイベント区間同定，キャプション間の類似度とも我々のデータセットのほうが ActivityNet Captions よりも高く，一貫したデータあることを示した．また，修辞構造に関してはテキスト修辞構造のデータセットである RST-DT におけるアノテーションの一致と比較した．その結果，我々のデータセットでの 2 名の作業者間の一致は RST-DT のそれと同等程度であることがわかった．この結果は，動画に修辞構造を導入することが妥当であることを示す．

作成したデータセットを用いて映像特徴を用いたベースライン解析器を実現した．ボトムアップテキスト修辞構造解析を参考に，任意の動画区間を Temporally-Sensitive Pretraining(TSP) を用いてベクトル化するエンコーダにベクトルを受け取りシフトまたは還元の解析アクションを決定する分類層を加えた解析器を実装した．この解析器と (1) で作成した人手キャプションのみを用いて同様にボトムアップ解析する手法を実装し，解析性能を比較したところ，人手キャプションを用いた手法の性能は，映像特徴を利用した解析器よりも大幅に性能が高いことがわかった．一方，自動キャプションを用いると映像特徴を用いる解析よりもさらに性能が劣化した．キャプション品質を人間並に向上させることは容易ではいので，映像特徴に基づく解析器を高度化が望まれる．そこで，映像特徴に基づく解析器のエンコーダをビデオキャプショニングタスクで事前学習することでテキスト知識を転移させたところ性能向上がみられた．また，テキスト修辞構造解析の高度化にも取り組んだ．LLM を用いてシフト還元解析を模倣する手法を提案し，3 種のベンチマークデータセットで世界最高性能を達成した．具体的には，スタックとキューの内部状態をプロンプトに変換し，シフトまたは還元動作を LLM で決定する．単純な手法ながら高い解析性能が得られたことから，動画の修辞構造解析の高度化にもおそらく大規模な事前学習モデルが必要であろうという知見が得られた．

動画修辞構造解析を活かす下流タスクとしてマルチモーダル要約を想定し，テキスト要約とフレーム抽出によるマルチモーダル要約データセットを構築した．2 名の作業者のアノテーションの一致を調べたところ，フレーム抽出では 7 割，テキスト要約では ROUGE が 0.4 程度であり作業者間の一致の高い要約であることがわかった．簡単なベースライン要約手法として，(1) のデータからランダムにイベントを抽出することによりイベント区間中央のフレームと与えられたキャプションを抽出した要約，動画の先頭，末尾から同様に要約を作成する手法，修辞構造木を依存構造木に変換してから，深さ優先の順にイベントをたどり同様に作成した要約を評価した結果，依存構造木を用いると，フレームの一致，ROUGE スコアとも向上することを確認した．つまり，マルチモーダル要約にとって動画修辞構造解析が役立つことが明らかとなった．

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 小林 尚輝，平尾 努，上垣外 英剛，奥村 学，永田 昌明	４．巻 29
２．論文標題 疑似正解データを活用したニューラル修辞構造解析	５．発行年 2022年
３．雑誌名 自然言語処理	６．最初と最後の頁 875～900
掲載論文のDOI（デジタルオブジェクト識別子） 10.5715/jnlp.29.875	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 平尾 努，小林 尚輝，上垣外 英剛，奥村 学，木村 昭悟
２．発表標題 動画談話構造解析：ベースライン解析器とその分析
３．学会等名 第26回 画像の認識・理解シンポジウム
４．発表年 2023年

１．発表者名 Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura
２．発表標題 Can we obtain significant success in RST discourse parsing by using Large Language Models?
３．学会等名 Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics（国際学会）
４．発表年 2024年

１．発表者名 前川在，平尾努，上垣外英剛，奥村学
２．発表標題 大規模言語モデルによるシフト還元修辞構造解析の模倣
３．学会等名 言語処理学会第30回年次大会
４．発表年 2024年

１．発表者名
平尾 努，小林 尚輝，上垣外 英剛，奥村 学，木村 昭悟

２．発表標題
動画談話構造解析へ向けたデータセット構築

３．学会等名
第25回 画像の認識・理解シンポジウム

４．発表年
2022年

１．発表者名
Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, Masaaki Nagara

２．発表標題
A Simple and Strong Baseline for End-to-End Neural RST-style Discourse Parsing

３．学会等名
Findings of the Association for Computational Linguistics: EMNLP 2022（国際学会）

４．発表年
2022年

１．発表者名
前川在，小林尚輝，平尾努，上垣外英剛，奥村学

２．発表標題
逆翻訳を利用したデータ拡張による文間の修辞構造解析の改善

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名
Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, Masaaki Nagata

２．発表標題
Improving Neural RST Parsing Model with Silver Agreement Subtrees

３．学会等名
Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies（国際学会）

４．発表年
2021年

１．発表者名
Ying Zhang, Hidetaka Kamigaito, Manabu Okumura

２．発表標題
A Language Model-based Generative Classifier for Sentence-level Discourse Parsing

３．学会等名
Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing（国際学会）

４．発表年
2021年

１．発表者名
小林尚輝，平尾努，上垣外英剛，奥村学，永田昌明

２．発表標題
言語モデルと解析戦略の観点からの修辞構造解析器の比較

３．学会等名
言語処理学会第28回年次大会

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	木村　昭悟 (Kimura Akisato) (10396202)	日本電信電話株式会社ＮＴＴコミュニケーション科学基礎研究所・メディア情報研究部・主幹研究員 (94305)	
研究分担者	奥村　学 (Okumura Manabu) (60214079)	東京工業大学・科学技術創成研究院・教授 (12608)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関