動画の談話構造解析

福島 健司 † 平尾 努 § 上垣外 英剛 † 奥村 学 † 永田 昌明 §
† 東京工業大学 §NTT コミュニケーション科学基礎研究所
{fukuken@lr., kamigaito@lr., oku@}pi.titech.ac.jp
{tsutomu.hirao.kp, masaaki.nagata.et}@hco.ntt.co.jp

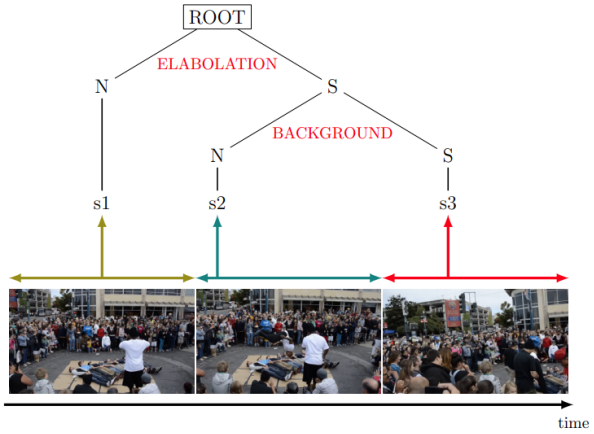
1 はじめに

テキストを構成する文や節の間には「因果」や「対比」などの何らかの関係が成り立っており，こうした関係に基づきテキストを構造化したものを談話構造という．談話構造をあらわす理論として，修辞構造理論 (Rhetorical Structure Theory: RST) [1] がよく用いられる．RST ではテキストは，終端記号が Elementary Discourse Unit (EDU) という節相当の談話単位，非終端記号が EDU によって構成されるテキストスパンの核性[1)]，枝がスパン間の修辞関係をあらわす句構造木として表現される．つまり，修辞構造はテキストがあらわすイベント間の関係を木構造として表現したものとして捉えることができる．

一方，テキストとして記述されたイベントの多くは実世界での出来事と結びついている[2)]．よって，それを記録した映像，つまり，動画にもテキストと同様に修辞構造が存在すると考えられる．動画の修辞構造を解析できるようになれば，一貫性の高い動画要約や動画内容理解の支援などアプリケーションの高度化に貢献できる．

本研究では，動画の修辞構造解析の第一歩として，動画とそれに含まれるイベントに対してキャプションとその修辞構造の注釈を与えたベンチマークデータセットを構築し，既存の修辞構造解析手法をベースとして動画の修辞構造解析がどの程度可能かを検証する．

図 1 に動画の修辞構造木の例を示す．動画は時間情報を伴ったイベントに分割され，分割されたイベントにはその内容をあらわすキャプション (文) が付与される．動画修辞構造木は，終端記号がイベント，つまり動画中の区間とそれに対応するキャプション文，非終端記号がイベントとキャプション文がなすスパンの核性，枝が 2 つのスパン間の修辞関係をあらわす木である．テキストの修辞構造木では終端記号は文よりも小さい EDU であるが，動画修辞構造の場合は，イベントを終端記号とするので，文が談話構造の最小単位になることに注意されたい．



s1. some guys are lying down and many people are assembled around them.
s2. a man speaks to the cloud with a microphone and the crowd clap their hands.
s3. another man run and jump over the guys and a cheer goes up from the crowd.

図 1 動画に対する修辞構造木．動画は https://youtu.be/coKW897eLyg より得た．

2 関連研究

動画をイベントに分割し，それぞれのイベントが支配する区間に対しキャプションを与えるという試みは Dense Video Captioning (DVC) と呼ばれ，コンピュータビジョン分野で研究が盛んに行われており，いくつかのベンチマークデータセットが整備されている．そのなかでも，ActivityNet Captions [3)] は最大規模のデータセットである．人間の動作を含む YouTube の 2 万動画に対しイベント分割とイベントに対するキャプション文を与えたデータセットで

1) 修辞関係で結びついた 2 つのスパンのうち，主となる役割を担うスパンが核 (Nucleus)，補助的な役割を担うスパンが衛星 (Satellite) となる．
2) もちろん，テキストで記述されるイベントのすべてが実世界のイベントと結びついているわけではない．
3) http://activity-net.org/challenges/2017/captioning.html

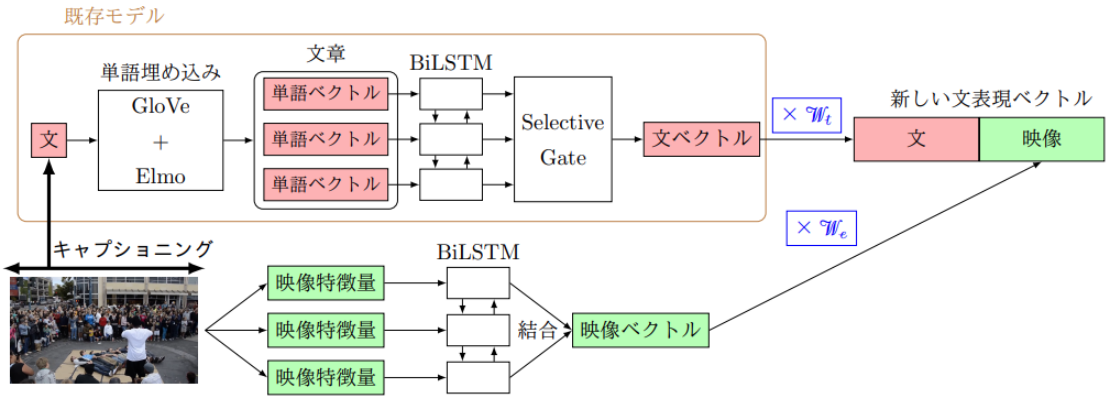


図 2　映像特徴量のモデルへの導入方法

あり，ActivityNet Challenge[4]でのシェアドタスクのデータとして用いられていることから，このデータを用いた DVC の研究が盛んに行われている．なお，1 つの動画は平均 3.67 個のイベントに分割される．つまり，1 つの動画に与えられるキャプション文は平均で 3.67 文となる．また，1 つのイベントは平均 36 秒となっている．ただし，DVC は動画中のイベント同定とキャプション付与が目的であるため，修辞構造の注釈は与えられていない．つまり，動画中のイベントに修辞構造の注釈を与えたデータセットは現状では存在しない．

一方，本研究と同様に動画の修辞構造解析を目的とした研究として [2] がある．この研究では，動画のキャプションに対して修辞構造木を推定した後，キャプション文とそれに該当する動画中の区間の割り当てを行うことで動画の修辞構造木を推定する．[2] では，このタスクのためのデータセットを整備していないため，得られた修辞構造木の妥当性は不明である．さらに，修辞構造解析を，キャプションテキストのみを対象に行っているので，技術としてはテキスト修辞構造解析そのものであり，動画から得られる特徴の利用が考慮されていない．

3　提案手法

近年，ニューラルネットワークを用いたテキスト修辞構造解析手法の研究が盛んに行われており，その性能が飛躍的に向上している．本研究では，テキスト修辞構造解析のベンチマークデータセットである RST-DT [3] を用いた評価結果で良い結果を残した Kobayashi らの解析器 [4] (Span-based Parser: SBP) をベースとして，キャプション文から得られる特徴とイベント区間から得られる特徴を同時に利用するビデオ修辞構造解析手法を提案する．SBP は，イベントが構成するスパンを入力として，それを再帰的に分割することで木を構築し，分割した 2 つのスパンの核性と関係ラベルをそれぞれ 3 クラス，18 クラスの分類問題を解くことで決定する．

3.1　イベントのベクトル表現

本研究では，SBP で利用するスパン (イベント系列) ベクトルをイベントに与えられたキャプション文とイベントに含まれる動画区間のフレームから得た画像特徴量の双方を連結することで生成する．このベクトル表現を得るためのニューラルネットワークの構造を図 2 に示す．

まず，キャプション文内の各単語において，GloVe[5]，ELMo[6] から得た単語ベクトルを結合する．次に，それらに BiLSTM と Selective Gate[7] を適用して文の表現ベクトルを得る．j 番目の単語の内部表現ベクトルは，前向き LSTM（$\overrightarrow{\mathrm{LSTM}}$）と後向き LSTM（$\overleftarrow{\mathrm{LSTM}}$）を用いて以下で定義する．

$$\overrightarrow{\mathbf{h}}^w_j = \overrightarrow{\mathrm{LSTM}}(\overrightarrow{\mathbf{h}}^w_{j-1}, \mathbf{w}_j),\quad \overleftarrow{\mathbf{h}}^w_j = \overleftarrow{\mathrm{LSTM}}(\overleftarrow{\mathbf{h}}^w_{j+1}, \mathbf{w}_j),$$
$$\mathbf{h}^w_j = [\overrightarrow{\mathbf{h}}^w_j; \overleftarrow{\mathbf{h}}^w_j]. \tag{1}$$

$\mathbf{w}_j$ は j 番目の単語に対する ELMo と GloVe のベクトルを結合したものである．キャプション文の単語数を n とすると，Selective Gate は j 番目の単語の内部表現 $\mathbf{h}^w_j$ と文脈ベクトル $\mathbf{s} = [\mathbf{h}^w_n; \mathbf{h}^w_1]$ を受け取り，新たな $\mathbf{h}^{w'}_j$ を生成する．

$$\mathbf{sGate}_j = \sigma\left(\mathbf{W}_s\mathbf{h}^w_j + \mathbf{U}_s\mathbf{s} + \mathbf{b}_s\right), \tag{2}$$
$$\mathbf{h}^{w'}_j = \mathbf{h}^w_j \odot \mathbf{sGate}_j. \tag{3}$$

$\mathbf{W}_s$ と $\mathbf{U}_s$ は重み行列，$\mathbf{b}_s$ はバイアスベクトル，σ はシグモイド関数，$\odot$ は要素積をそれぞれあらわす．そして，i 番目のイベントの文ベクトル $\mathbf{t}_i$ を以下の

4）http://activity-net.org

式で定義する．

$$\mathbf{t}_i = \frac{1}{\mathrm{n}} \sum_{j=1}^{n} \mathbf{h}_j^{w'}. \tag{4}$$

映像特徴量としては，動画から一定時間ごとに切り出したフレームの特徴量を利用する．イベントの動画区間に含まれる特徴量の総数を m としたとき，前向き LSTM と後向き LSTM を用い，イベントの映像ベクトル $\mathbf{e}_i$ を以下の式で定義する．

$$\begin{aligned}
&\overrightarrow{\mathbf{h}_j^v} = \overrightarrow{\mathrm{LSTM}}(\overrightarrow{\mathbf{h}}_{j-1}^v, \mathbf{v}_j), \quad \overleftarrow{\mathbf{h}_j^v} = \overleftarrow{\mathrm{LSTM}}(\overleftarrow{\mathbf{h}}_{j+1}^v, \mathbf{v}_j), \\
&\mathbf{h}_j^v = [\overrightarrow{\mathbf{h}_j^v}; \overleftarrow{\mathbf{h}_j^v}], \\
&\mathbf{e}_i = [\mathbf{h}_m^v; \mathbf{h}_1^v].
\end{aligned} \tag{5}$$

ここで，$\mathbf{v}_j$ は j 番目の映像特徴量である．$\mathbf{t}$ と $\mathbf{e}$ にそれぞれスカラー重み $\mathcal{W}_t$，$\mathcal{W}_e$ を乗じて結合することで 1 つのイベントに対するベクトル $\mathbf{u}_i = [\mathcal{W}_t \mathbf{t}_i; \mathcal{W}_e \mathbf{e}_i]$ を得る．次に，これを再度 BiLSTM に入力し，以下のイベントに対する隠れ状態ベクトル $\mathbf{f}_j$，$\mathbf{b}_j$ を得る．

$$\mathbf{f}_j = \overrightarrow{\mathrm{LSTM}}(\mathbf{f}_{j-1}, \mathbf{u}_j), \quad \mathbf{b}_j = \overleftarrow{\mathrm{LSTM}}(\mathbf{b}_{j+1}, \mathbf{u}_j). \tag{6}$$

最終的に i 番目のイベントから j 番目のイベントのスパンをあらわすベクトル $\mathbf{u}_{i:j}$ を次式で定義する．

$$\mathbf{u}_{i:j} = [\mathbf{f}_j - \mathbf{f}_{i-1}; \mathbf{b}_{i-1} - \mathbf{b}_j]. \tag{7}$$

3.2 解析モデル

スパンを $k (i < k < j)$ 番目のイベント区間で分割するかを判定するスコアは以下の式で定義される．

$$\mathrm{s}_{\mathrm{split}}(i, j, k) = \mathbf{h}_{i:k}^\top \mathbf{W}_u \mathbf{h}_{k+1:j} + \mathbf{d}_l^\top \mathbf{h}_{i:k} + \mathbf{d}_r^\top \mathbf{h}_{k+1:j}. \tag{8}$$

$\mathbf{W}_u$ は重み行列，$\mathbf{d}_l$ と $\mathbf{d}_r$ は重みベクトルである．また，$\mathbf{h}_{i:k}$ と $\mathbf{h}_{k+1,j}$ は以下の式で定義される．

$$\mathbf{h}_{i:k} = \mathrm{MLP}_{\mathrm{left}}(\mathbf{u}_{i:k}), \tag{9}$$

$$\mathbf{h}_{k+1:j} = \mathrm{MLP}_{\mathrm{right}}(\mathbf{u}_{k+1:j}). \tag{10}$$

MLP_* は多層パーセプトロンを表し，一層の順伝播型ニューラルネットワークと，活性化関数に ReLU 関数を用いる．そして，関数 $\mathrm{s}_{\mathrm{split}}$ を最大とする $\hat{k}$ でスパンを分割する．

$$\hat{k} = \underset{\mathrm{k} \in \{i, \ldots, j-1\}}{\arg\max} \left[\mathrm{s}_{\mathrm{split}}(i, j, k) \right]. \tag{11}$$

スパン (i, j) を k で分割した 2 つのスパン間に付与する核性ラベルと関係ラベルのスコア $\mathrm{s}_{\mathrm{label}}(i, j, k, \ell)$ は以下の式で定義される．

$$\mathrm{s}_{\mathrm{label}}(i, j, k, \ell) = \mathbf{W}_\ell \mathrm{MLP}\left(\left[\mathbf{u}_{i:k}; \mathbf{u}_{k+1:j}; \mathbf{u}_{1:i}; \mathbf{u}_{j:n} \right] \right). \tag{12}$$

$\mathbf{W}_\ell$ は重み行列である．そして，以下の式でラベルが選択される．

$$\hat{\ell} = \underset{\ell \in L}{\arg\max} \left[\mathrm{s}_{\mathrm{label}}(i, j, l, \ell) \right]. \tag{13}$$

L は，核性ラベル推定の際には {N-S,S-N,N-N}，関係ラベル推定の際には 18 種類の修辞関係ラベルの集合をそれぞれをあらわす．

そして，すべてのパラメタは正解の分割 k^* と正解ラベル ℓ^* に対し，以下の損失の最小化により最適化される．

$$\begin{aligned}
&\max(0, 1 + \mathrm{s}_{\mathrm{split}}(i, j, k^*) - \mathrm{s}_{\mathrm{split}}(i, j, \hat{k})) \\
&+ \max(0, 1 + \mathrm{s}_{\mathrm{split}}(i, j, k^*, \ell^*) - \mathrm{s}_{\mathrm{split}}(i, j, k^*, \hat{\ell})).
\end{aligned} \tag{14}$$

4 実験

4.1 データセット

上で述べた ActivityNet Captions の検証用データセット中のデータ 339 件に対し，人手で修辞構造の注釈を与えたデータセット（Activity-RST）を作成した．なお，注釈付けは RST-DT 作成時に利用されたマニュアル[5]に従った．表 1 に，1 文書あたりの文数，修辞構造木の深さを，RST-DT との比較で示す．Activity-RST の木は RST-DT の木と比べノード数が少なく，階層が浅い．RST-DT の核性ラベルの頻度は，N-S が 5079，N-N が 1851，S-N が 532 であり，関係ラベルは Elaboration，Joint，Explanation の順で多かった．一方，Activity-RST の核性ラベルの頻度は，N-S が 44，N-N が 204，S-N が 1111 であり，関係ラベルは Elaboration，Background，Textual-organization の順で多かった．2 つのスパンが RST-DT では S-N に分類されることが多く，Activity-RST では N-S に分類されることが多い．この理由は，RST-DT は新聞記事が対象であるため記事の重要部が主に前に現れることに対し，Activity-RST は動画が対象であるため重要部は後に現れることにある．これらの比較から，Activity-RST は RST-DT とは性質の異なるデータセットであることが分かる．

4.2 映像特徴量

本実験で使用した映像特徴量は，C3D[8] および，RGB ResNet-200(RES)[9] と Optical flow BN-Inception(BN)[10] の組み合わせである．C3D は，3D

5） https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.201.9677&rep=rep1&type=pdf

表 1　データセットの比較

データセット	文数	木の深さ	
	平均	平均	分散
RST-DT	22.8	7.1	12.1
Activity-RST	5.0	3.0	2.9

ConvNetwork を用いて 8 フレーム毎に映像を 1 つの 500 次元ベクトルへと畳み込んだもので，ActivityNet Challenge で公式に配布されている特徴量である．RES と BN の組み合わせは，Zhou らの DVC システム [11] で利用された特徴量である．RES は映像の RGB 値を 0.5 フレーム毎に取得した特徴量，BN はオプティカルフローを 0.5 フレーム毎に取得した特徴量である．

4.3　実験設定

10 分割交差検証 (訓練:検証:テストデータ=8:1:1) により解析器の評価を行った．また，5 モデルアンサンブルを用い，異なる初期値の 3 回の試行の平均値を最終的な評価結果とした．

解析に用いるキャプションは，ActivityNet Captions で与えられた正解キャプションと Zhou らのシステム [11] に正解イベントの区間を与えて生成したキャプションを試した．C3D，RES+BN とも ActivityNet Captions の正解イベントの区間から得たものを利用した．そして，テキストから得た特徴と映像から得た特徴のそれぞれの組み合わせも評価した．

評価指標は RST-Parseval[3] に従い，ラベルなしスパン (Span)，核性ラベル付きスパン (Nuc)，関係ラベル付きスパン (Rel)，すべてのラベル付きスパンの一致 (Full) の micro-averaged F_1 値を用いた．

4.4　結果と考察

表 2 に評価結果を示す．テキスト特徴量，映像特徴量を単独で用いた場合を比較すると，正解キャプションを利用した場合のスコアが最も高い．正解キャプションを用いた場合よりもスコアは大きく下がるが，C3D，RES+BN を用いた場合がほぼ同等のスコアであり，システムが生成したキャプションを用いた場合が最もスコアが低く，C3D，RST+BN よりも 2 ポイント程度劣化している．修辞構造の注釈は正解キャプションに基づき行われたため，正解キャプションを利用した場合が最も良い結果を得たことは妥当であろう．しかし，C3D，RES+BN よりもシステムが生成したキャプションを用いた場合に性能が劣化していることは自動生成されたキャプション文の質が低いことを示唆している．

表 2　Micro-averaged F_1 による性能比較

Model	Span	Nuc	Rel	Full
正解文のみ	84.9	72.9	63.0	62.9
生成文のみ	80.4	66.4	55.7	55.7
正解文+C3D	**85.5**	**74.0**	**63.9**	**63.8**
生成文+C3D	81.3	67.5	56.5	56.4
C3D のみ	82.4	68.9	58.0	57.9
正解文+RES+BN	**85.6**	**73.8**	**63.9**	**62.8**
生成文+RES+BN	82.3	68.9	57.9	57.8
RES+BN のみ	82.7	69.2	57.8	57.7

次にテキスト特徴量と映像特徴量を組み合わせた場合をみると，正解キャプションに対して C3D，RES+BN の双方を組み合わせるとすべての指標においてスコアが改善されていることがわかる．この結果は，映像特徴量が修辞構造解析に役立っていることを示している．しかし，システムが生成したキャプションを組み合わせると，C3D，RES+BN をそれぞれ単独で利用する場合よりもスコアは劣化しており，システムが生成したキャプションの質の低さが悪影響を与える結果となった．

5　おわりに

本稿では，動画の修辞構造解析のため，ActivityNet Captions データセットの一部に修辞構造の注釈を与えたベンチマークデータセットを構築した．テキストと映像特徴量が解析性能に与える影響を調べた結果，正解キャプションから得た特徴量を用いる場合の性能が最も良く，正解キャプションと正解イベントの区間から得られる映像特徴量を組み合わせることでさらに性能が向上することを確認した．今回のテキストと映像の特徴量の組み合わせ方は非常に単純なものであるため，これをさらに洗練されたものに改良することで性能向上をはかりたい．また，未知のデータの解析のためにはイベントの区間推定，キャプションの生成をともに自動的に行わねばならない．今回の実験結果では自動推定したキャプションで大きく性能が劣化したことからキャプション生成における改善の必要性が明白となった．また，イベント区間の自動推定についても手法を開発し，修辞構造解析の性能に与える影響を調べる必要があるだろう．

参考文献

[1]W.C. Mann and S.A Thompson. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, USC/ISI, 1987.

[2]Arjun Reddy Akula and S. Zhu. Visual discourse parsing. *ArXiv*, Vol. abs/1903.02252, , 2019.

[3]Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001.

[4]Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. Top-down rst parsing utilizing granularity levels in documents. In *Proceedings of the 2020 Conference on Artificial Intelligence for the American*, pp. 8099–8106, New York, America, September 2020.

[5]Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[6]Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[7]Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1095–1104, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[8]Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3D: generic features for video analysis. *CoRR*, Vol. abs/1412.0767, , 2014.

[9]K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[10]Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37 of *Proceedings of Machine Learning Research*, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[11]Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.