データドリブンな文章構造と情報伝達の抽出手法

Data-driven methods of extracting text structure and information transfer

本那 真一 [1], 村山 太一 [2], 松井 暉 [3]
[1] 横浜国立大学 経営学部
[2] 横浜国立大学 大学院環境情報研究院
[3] 横浜国立大学 国際社会学府
honna-shinichi-cd@ynu.jp, {murayama-taichi-bs，matsui-akira-zr}@ynu.ac.jp

概要

文章構造は，読者の注意喚起や興味喚起，理解・記憶の効率など，文章を評価する要素に大きな影響を与えるにもかかわらず，文章構造と評価との関係は明らかでない．本稿では，小説，Wikipedia 記事，学術論文，映画字幕といった多様な媒体を対象に，文章構造と評価の関係を定量的に分析した．具体的には，意味依存度の低い特徴量を抽出し，文章中の構造を捉え，評価に与える影響を検証した．その結果，評価が高い文章には特定の構造が顕著に見られることが明らかとなり，情報提示の順序や関係性が読者の評価に与える影響が確認された．これにより，効果的な文章作成や情報提示の設計において，媒体ごとの構造や転換タイミングを考慮する重要性を示すことができた．

1 はじめに

文章構造は単なる形式的な体裁ではなく，人々の注意喚起や興味喚起，あるいは理解・記憶の効率を大きく左右する重要な要素である [1, 2]．特に，評価される文章がどのような文書構造を持つかという点について，様々な視点から議論がなされている [3, 4, 5]．本稿では，これまで多くの分野で探究されてきた文章構造の理論やモデルを発展させ，文章構造と評価の関係を明らかにする．

文章をモデル化し，読者との関係を掘り下げる研究は様々な視点からなされてきた [6, 7, 8, 9, 10]．先行研究では文章構造そのもののを定量的に取得することや，読者・受け手の理解度や感情変動との関係を分析したものが多い．一方で，評価される文章とそうでない文章の間に文章構造の違いが存在するのかという問いはまだ明らかでない．この問いに応えることは，情報伝達の効果測定や文章作成支援の実践にも直結し，重要な知見をもたらす．本稿では，この文章構造と評価の関係という点に焦点を当て，媒体をまたいだ大規模データを用いることで，評価が高い文章にはどのような構造パターンが潜んでいるのかを明らかにする．

具体的には，小説，Wikipedia 記事，学術論文，映画字幕といった多様な媒体を対象とする．まず，これらの媒体の文書構造を把握するために，文章から意味に依存しない特徴量を抽出する．これらの特徴量を基に，分割した各文章にクラスタを割り当て，文章がどのようにクラスタを遷移していくかの遷移パターン，これを本稿における文書構造とする．その後，捉えた文章構造が，どの順番・タイミングで登場するかを評価指標と結びつけることで，文章構造と評価がどの程度関連するかを明らかにする．これにより，本稿では既存研究が示唆する文章構造の重要性を再検証するとともに，評価に結びつきやすい文章構造の特徴を導き出すことを目指す．

2 データ

本稿では，小説，Wikipedia，論文，映画の 4 種類のデータを用いた．それぞれについて文章をどの単位で分割するか（分割単位）と，何を評価指標にするかについて述べる．各データには N 個のまとまった文章があり，それぞれの n に評価 Score_n が評価指標として割り当てられる．また，n は分割単位によって $\mathrm{Block}_{n,i}$ に分割される．この際に同一の n から分割された $\mathrm{Block}_{n,i}$ は n に紐づいた同じ評価 Score_n を持つ．

- **小説家になろうデータ** 日本国内で多数のオンライン小説が投稿されている「小説家になろう」から取得した本文 6,580 件を分析対象とする．小説は各話単位で分割し，それぞれにつけ

られた感想コメント数を評価指標とした．

- **Wikipedia データ**
  英語版 Wikipedia のダンプデータを利用し，3270,926 件の記事をセクション（小見出し）単位に分割した．評価指標として，各記事の 2023 年の 1 年間の閲覧数を用いた．
- **arXiv データ**
  物理学，数学，計算機科学などの分野のプレプリント論文が集積された arXiv にアップロードされた論文から本文 720,126 件を取得し，subsubsection 単位で分割した．発行後 10 年間の引用数（分野ごとに正規化）を評価指標とした．
- **映画字幕データ**
  opensubtitles.org から得た英語字幕データ 857,201 件を対象とし，教師なし手法である Bayesian Blocks[11] を用い，映画内の発話のタイムスタンプを利用して分割した．また，opensubtitles.org におけるダウンロード数を評価指標として用いた．

分析の前処理として，作品内の分割単位が極端に多い，あるいは少ないものを四分位範囲（IQR）で除外し，残った作品を評価指標に応じて 10 の評価グループ `group_0`〜`group_9` に分割する．`group_0` が最も評価が低く，`group_9` が最も評価が高い．また，arXiv は引用数 0 件が多いためグループ数を調整した．

3 分析手法

本章では, 評価と文章構造の関係を明らかにするために，文章からその構造的特徴を定量的に捉える手法を示す. 分析の概観を図 1 を示す.

3.1 文章の分割とクラスタリング

文章を異なる役割を持つクラスタが連続しているものとして捉える処理を行う．

準備 1. 文章のブロック化: まず，2 データで提示した分割単位で文章を分けて複数のブロック化し，それぞれのブロックから特徴量を抽出してベクトル化する．具体的には, 以下の 4 種類の特徴量を抽出する.

1. **品詞の出現頻度（POS）**
2. **ストップワードの頻度（Stopword）**
3. **係り受けラベルの出現頻度（Dep.）**
4. **感情辞書に基づく感情スコア（Emotion）**

これらはいずれも固有名詞や文脈特有の表現に左右されにくい共通点をもつ．用いる特徴量のうち上 3 つは，文章全体の構造や文法的な骨格を示す指標となる．これらは，論文・小説・字幕といったテキストの種類に依存せず，文章を形態素・文法要素レベルで捉えることで一定の数値化が可能であるとされる [12, 10]．また，感情辞書を用いた感情スコアは，基本的な感情を辞書参照で機械的に算出し，感情の傾向を捉えられる [13].

準備 2. クラスタリング: 次に，この特徴ベクトルを持つ各ブロックに対し似た役割同士でグルーピングするために k-means クラスタリングを適用した．本稿ではクラスタ数を 5 とした．加えて，各クラスタを特徴づける要因を調べるために LightGBM[14] を用いてクラスタを識別するモデルを構築し，Feature Importance を参照してどの特徴量がクラスタ判別に寄与しているかを明らかにする．これにより，それぞれのクラスタがどのような言語的・感情的性質をもつかを把握できる．

3.2 構造遷移の分析

取得したクラスタの並びを用いて,「構造遷移」を取得し，その順序と位置の観点から分析を行う．

定義 3.1（構造遷移） ある作品において，連続するクラスタが切り替わることを「構造遷移」と呼ぶ．たとえば，文章のクラスタの並びが「AAABBAACCC」のものがあった際に，構造遷移は $A \to B$，$B \to A$，$A \to C$ の 3 回起こっているとする．

手法 1. 遷移順序の分析: 文章の中で，クラスタがどのような順番で登場するかに着目する．定義 3.1 で示した構造遷移を用いて，各評価グループ内でクラスタ間の遷移確率を集計し，マルコフ遷移行列を構築する．構築される行列は，あるクラスタから別のクラスタへ移る確率を示すものであり，グループごとにどのような流れでクラスタが展開していくかを表す．その後，グループ間での遷移順序の違いを比較するために，グループ間のマルコフ遷移行列を Jensen-Shannon 距離によって比較する．値が大きいほど，クラスタの遷移の仕方が異なることを意味する．評価が高いグループと低いグループとのあいだで，クラスタの出現順序にどのような違いがあるかを把握する．

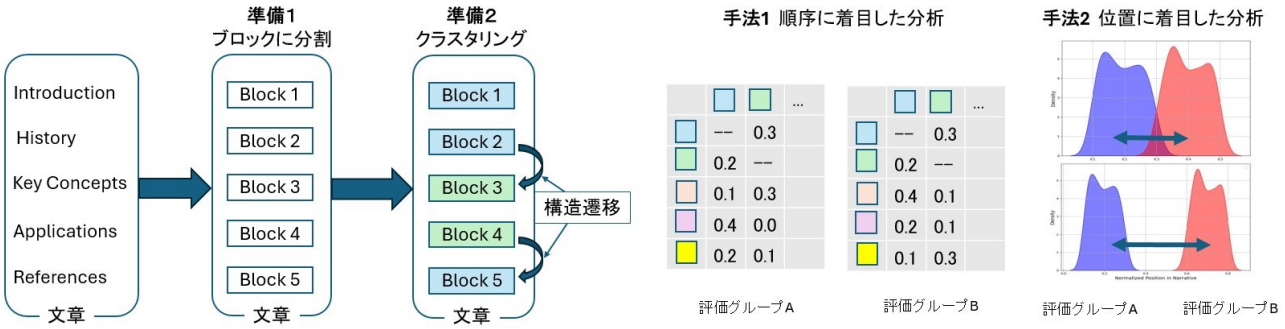


図 1: 本手法の概要図．文章を構造遷移として捉えるための前処理として分割単位で文章をブロック化した (準備 1)，それ用いてクラスタリングを実行する (準備 2)．その後，順序 (手法 1) と位置 (手法 2) に着目し，構造遷移を評価グループ間で比較する．

	(a) 小説家になろう				(b) Wikipedia				(c) arXiv				(d) 映画字幕			
Clus.	**POS**	**Stop**	**Dep.**	**Emo**	**POS**	**Stop**	**Dep.**	**Emo**	**POS**	**Stop**	**Dep.**	**Emo**	**POS**	**Stop**	**Dep.**	**Emo**
0	PROPN	–	nmod	–	VERB	of	ccomp	Joy	NOUN	the	nsubj	–	PRON	you	nsubj	Anger
1	NOUN	を	mark	Joy	NOUN	the	pcomp	Anger	NUM	–	advcl	Joy	NOUN	the	comp.	Joy
2	PRON	ない	dep	–	PRON	–	dative	Trust	ADJ	–	dep	Trust	VERB	–	acomp	–
3	NUM	–	nummod	Trust	ADP	in	aux	–	AUX	–	relcl	–	ADJ	that	advcl	Trust
4	ADJ	–	compd	Anger	ADJ	that	predet	Joy	ADV	and	compd	Anger	ADV	–	dep	Sadness

表 1: 抽出した特徴量をもとに，各クラスタで顕著だった要素をまとめた．

手法 2. 遷移位置の分析: 定義 3.1 で示した構造遷移が作品内のどの位置で起こるかを調べる．そのために，遷移位置をカーネル密度推定によって分布として扱い，その確率密度に基づいて評価グループ同士の分布差を Wasserstein 距離で計算する．構造遷移の位置はクラスタの推移が「AAABBAACCC」のような作品において全章数に対する割合から計算され，$A \to B : 0.3, B \to A : 0.5, A \to C : 0.7$ となる．Wasserstein 距離が大きいほど，クラスタの遷移の仕方が異なることを意味する．これにより，各グループにおける構造遷移のタイミングの違いを把握する．

4 結果

各媒体で推定した遷移の順番と位置の両面から，評価との関連を検討した結果を示す．まず各クラスタの特徴を概観し，そのうえで遷移順序 (参照 3.2) と遷移位置 (参照 3.2) の傾向を明らかにする．最終的に，これらの結果から，各媒体において評価が高い文章には共通の構造が存在するかどうかを考察する．

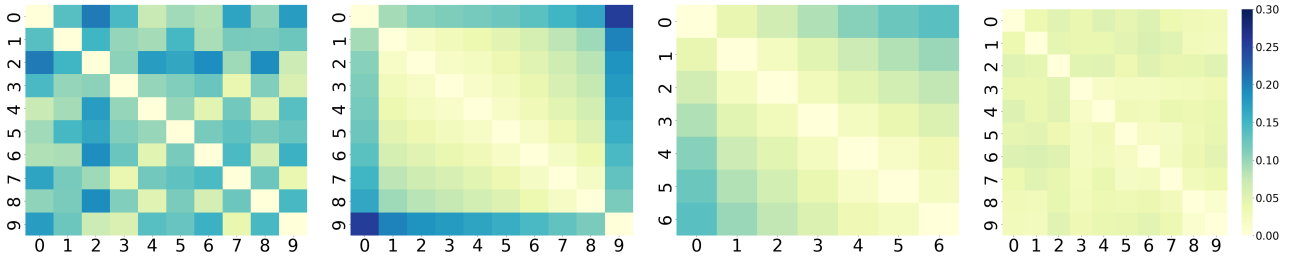
4.1 LightGBM によるクラスタ解釈

表 1 に，LightGBM による各クラスタの特徴を示す．小説家になろうではクラスタ 1 が情緒的核心を担い，クラスタ 4 は感情的対立が強い場面に対応する．Wikipedia ではクラスタ 0 が動詞中心で事実提示とわずかな喜びを示し，クラスタ 1 は名詞主体で議論の対立も含む．arXiv のクラスタ 0 は名詞主語構造が多く定義や説明が集まり，クラスタ 1 は数字や論証過程が多いブロックで肯定的ニュアンスも含む．映画字幕ではクラスタ 0 が二人称呼びかけや憤りを伴うセリフ，クラスタ 1 が明るい場面や肯定的発言を含む会話を示す．いずれも言語的・感情的役割を担うブロックが形成されており，媒体ごとに論述重視や感情の盛り上がりといった特徴が見られる．

4.2 構造遷移順序の分析

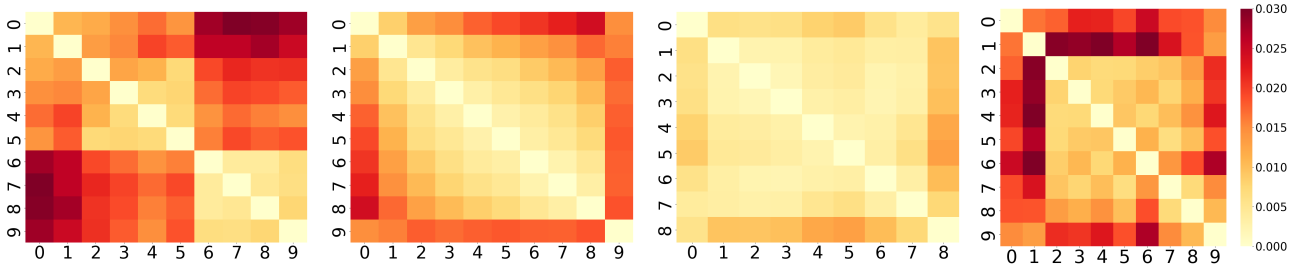
図 2 に，3.2 節の手法 1 で示した手法に基づき，構造遷移順序の評価グループ間の Jensen-Shannon 距離をヒートマップに示す．

Wikipedia と arXiv においては，評価が高いグループと低いグループのあいだで構造遷移の順序に大きな差が見られる．図 2(b)(c) のヒートマップ上では，右上や左下に行くほど該当する評価グループ間の距離が大きいことを示しており，評価が離れるほど遷移順序に顕著な違いがある可能性を示唆している．Wikipedia は情報を簡潔かつ階層的に整理する必要があり，見出し構成やセクション分けが読みやすさや利用価値に直結する．arXiv の論文は IMRAD 形式などの形式化された論理展開を踏襲することで論



(a) 小説家になろうデータ　(b) Wikipedia データ　(c) arXiv データ　(d) 映画字幕データ

図 2: 各データセットにおける評価グループ間の構造遷移順序の Jensen-Shannon 距離ヒートマップ．色が濃いほど距離が離れていることを示す．



(a) 小説家になろうデータ　(b) Wikipedia データ　(c) arXiv データ　(d) 映画字幕データ

図 3: 各データセットにおける評価グループ間の構造遷移位置分布の Wasserstein 距離ヒートマップ．色が濃いほど距離が離れていることを示す．

理展開をわかりやすく示し [2]，評価を高めやすいと考えられる．一方，小説家になろうと映画字幕では，グループ間に一貫した構造遷移順序が見られない．ストーリー性が主体となるこれらの媒体では，制作者の作風や物語の独創性が評価に直結し，クラスタの登場順序よりも内容やキャラクターの魅力が重視される可能性が高い．総じて，情報を伝えることを主目的に置いた媒体ほど順序の一貫性が評価要因となり，物語性の強い媒体ほど順序以外の要素が評価を左右すると推測される．

4.3　構造遷移位置の分析

図 3 に，3.2 節の手法 2 で示した手法に基づき，構造遷移位置の評価グループ間の Wasserstein 距離をヒートマップで示す．小説家になろうでは，上位・下位グループ間の距離が明確に分かれ，特定のタイミングで転換を行う作品ほど評価が高い傾向がうかがえる．映画データにも同様の傾向が見られ，低評価グループと中・高評価グループとの間に顕著な差が認められた．一方で，Wikipedia や arXiv は全体的にグループ間の類似度が高く，遷移位置という点では多くの作品において共通の位置で遷移が起きており，このことから，遷移の位置は評価に大きな影響を与えていないと推察される．これは，Wikipedia や論文がある程度形式化された構造によって書かれているために，差がうまれにくいことも推察できる．以上から，とりわけストーリー性の強い媒体では遷移の位置が高評価を獲得する上で重要な要因となっていることが示された．

5　おわりに

本稿では，小説家になろう，Wikipedia，arXiv，映画字幕といった異なる媒体を対象に，文章を役割のまとまりでクラスタリングし，その順序や位置を評価指標に基づき比較した．これにより，Wikipedia や arXiv などの定型的な構造を持つ媒体ほど一貫した順序や章立てが高評価と結びつきやすく，小説や映画などの物語性が強い媒体ではストーリーの独自性や登場人物の魅力などの構造以外の要因が大きく影響する傾向が示唆された．また，遷移位置に着目した分析では，物語性の強い媒体など，明確に形式化されていないフォーマットでは遷移位置の工夫が評価を左右する可能性があることが確認された．本稿の知見は，効果的な文章作成や情報提示の設計において，媒体ごとの構造や転換タイミングを考慮する重要性を示すものである．

謝辞

参考文献

[1] Jean M Mandler and Nancy S Johnson. Remembrance of things parsed: Story structure and recall. *Cognitive psychology*, Vol. 9, No. 1, pp. 111–151, 1977.

[2] Luciana B Sollaci and Mauricio G Pereira. The introduction, methods, results, and discussion (imrad) structure: a fifty-year survey. *Journal of the medical library association*, Vol. 92, No. 3, p. 364, 2004.

[3] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, Vol. 34, No. 1, pp. 1–34, 2008.

[4] Peter W Foltz, Walter Kintsch, and Thomas K Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, Vol. 25, No. 2-3, pp. 285–307, 1998.

[5] Camille Guinaudeau and Michael Strube. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 93–103, 2013.

[6] John M. Swales. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, 1990.

[7] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, Vol. 8, No. 3, pp. 243–281, 1988.

[8] Ryan Reagan, Florian Mairesse, and Mary Anna Walker. Emotional arcs of stories are dominated by six basic shapes. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 3234–3244, 2016.

[9] Olivier Toubia, Jonah Berger, and Jehoshua Eliashberg. How quantifying the shape of stories predicts their success. *Proceedings of the National Academy of Sciences*, Vol. 118, No. 26, p. e2011695118, 2021.

[10] Ryan L Boyd, Kate G Blackburn, and James W Pennebaker. The narrative arc: Revealing core narrative structures through text analysis. *Science advances*, Vol. 6, No. 32, p. eaba2196, 2020.

[11] Jeffrey D Scargle, Jay P Norris, Brad Jackson, and James Chiang. The bayesian block algorithm. *arXiv preprint arXiv:1304.2818*, 2013.

[12] James W Pennebaker. The secret life of pronouns. *New Scientist*, Vol. 211, No. 2828, pp. 42–45, 2011.

[13] Xiang Chen, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1–10, 2017.

[14] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, Vol. 30, , 2017.