顕現的要素の出現順序に基づく物語の類似性尺度

大竹 孝樹 [1]　横井 祥 [1,2]　井之上 直也 [1,2]　乾 健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所 AIP センター
{takaki, yokoi, naoya-i, inui}@ecei.tohoku.ac.jp

1 はじめに

1.1 物語の類型・類似性

物語を語り共有するという営みは我々人間の最も基本的な社会的活動の一つである [6]．我々が語る物語の中には地域や時代が異なるにも関わらず共通の類型を持ったものが存在し，我々はそうした**物語の類型・類似性**を自然に認識することができる．例えば「恵まれない境遇の主人公がなんらかの幸運を得て幸せな人生を歩む」タイプの物語はしばしば「シンデレラストーリー」と呼ばれる．このような物語の類型・類似性を理解し捉えようとする試みは人文学でも [11] 自然言語処理分野でも [2, 5, 8, 9] 盛んに取り組まれてきた．

1.2 物語におけるイベントの顕現性と出現順序

物語の類型・類似性を考える際には，物語中に現れる**イベント**（できごと）がひとつの重要な鍵となる [6]．実際，自然言語処理における物語の類似性の研究の多くが，語彙の分布などに加え物語中のイベントに注目している [2, 5, 8]．

本研究ではとくに，物語中のイベントの**顕現性**（顕著さや重要性）と**出現順序**に着目する．これらの観点が有用であることを示す例として童話「赤ずきん」を考えよう．「赤ずきん」と同様の類型を持つ物語は複数の国や地域を跨いで存在することが知られており[*1]，これらの物語は共通して以下のイベント群を持つ．

1. 女の子が（主には）祖母の元を訪れるよう頼まれる
2. 祖母の元を訪れる
3. 狼/怪物に飲み込まれる
4. （多くの場合）最終的に救出される

顕現性： 物語テキスト中のイベントには，顕現性の高い（物語において重要なでストーリー展開に影響を及ぼす）イベントと，そうでないイベントが存在する．「赤ずきん」の例では，「狼/怪物に飲み込まれる」などのイベントは「赤ずきん」型物語の類型を考える際に重要なイベントと考えられる．一方で，「赤ずきん」と共通と類型を成す物語の中には，「赤ずきんが美しい花で覆われた大地を見る」というイベントも存在するが，これは大筋のストーリー展開に影響せず，「赤ずきん」の類型を特徴付けるイベントではない．このように，物語中の要素の顕現性を考慮して物語の類似性を計算する研究としては [8, 9] がある[*2]．

出現順序： イベントの出現順序も物語の類型を特徴付ける重要な側面である．例えば「赤ずきん」のイベントの出現順序を入れ替えると

1. 女の子が狼/怪物に飲み込まれる
2. 救出される
3. 女の子が祖母の元を訪れるよう頼まれる
4. 女の子が祖母の元を訪れる

という物語になるが，これは明らかに「赤ずきん」型の物語ではない．イベントの出現順序を考慮する研究としては [5] がある．

1.3 貢献

以上より，物語の類似性を考える際には，イベントの顕現性と出現順序がそれぞれ重要であると考えられる．しかし，これらがそれぞれどの程度有効なのか，またこれら両方を組み合わせることがどの程度有効なのかについてはこれまで検証がなされてこなかった．本研究の貢献は以下の通り．

- 各物語テキストに現れるイベント群の**知識表現**とその**類似性計算**の手法を，「**顕現性**」「**出現順序**」の捉え方にのみに違いが生じるよう複数種類設計した（概要は表1の通り）．

[*1] https://www.pitt.edu/~dash/type0333.html

[*2] これらの研究はイベントのみでなく，より一般的な要素について顕現性を考慮しているものである

表1: イベントの顕現性と出現順序を考慮した物語の表現．2 段目と 4 段目の表現ではイベントの顕現性が考慮されている．濃い色になっているイベントが顕現性の大きなイベントを表す．

顕現性	出現順序	物語表現
		{arrive, go, look, look, see, swallow}
✓		{arrive, go, look, look, see, swallow}
	✓	go → look → see → arrive → look → swallow
✓	✓	go → look → see → arrive → look → swallow

- これらの手法を用いて物語をクラスタリングし，これが人手で分類されたデータにどの程度よく当てはまっているかを検証することで，「顕現性」「出現順序」「その組み合わせ」の物語の類型・類似性計算への貢献度を確かめた．

2 関連研究

2.1 物語の類型の定義づけ — ATU 分類

民俗学では，異なる国や文化圏に存在する多数の民話を分類・分析する目的で，様々な民話の類型分類が検討されてきた [1, 12]．とりわけ広く知られている類型が，Aarne–Thomson–Uther type index（以下 **ATU 分類**）である [12]：

```
Aarne-Thomson-Uther type index
├── ANIMAL TALES（動物物語）
│   └── Wild Animals（野生動物）
│       └── The Clever Fox（賢い狐）
│           └── The theft of fish（魚を盗む）
└── TALES OF MAGIC（魔法物語）
    └── Supernatural Adversaries（超自然的な敵）
        ├── The Dragon-Slayer（竜殺しの英雄）
        ├── Little Red Riding Hood（赤ずきん）
        └── ...
```

ATU 分類は，物語をそのモチーフ（物語中に出現する特徴的な要素，たとえば「Tabu: eating food of certain person (誰かの食べ物を食べてしまうという禁忌を犯す)」や「Crimes punished (罪が罰せられる)」）に基づき階層的に分類した体系で，最下層では約 2400 のクラスに分かれる．分類の上位では「Animal Tales（動物の出てくる物語）」や「Magic Tales（魔法に関わる物語）」のように登場人物や大きなモチーフで粗く分類され，その後「The Sleeping Beauty（眠れる森の美女に類する物語）や「Outcast Animals Find a New Home（ブレーメンの音楽隊に類する物語）」」のようにより詳細な分類がなされる．

2.2 物語の類型・類似性の計算

ひとことで物語の「類似性」と言っても，その中には，登場人物の類似性，プロットの類似性（粗筋），スタイルの類似性（例: 語調）など様々な側面があることが指摘され，モデル化されている [3, 9]. 本研究ではとくに物語テキスト中のイベント（できごと）の類似性に着目する．

3 イベントの顕現性と出現順序に着目した物語の表現・類似尺度

本節では，物語テキスト s の**表現** $\phi(s)$ および別の物語テキスト s' との**類似度** $\mathrm{sim}(\phi(s), \phi(s'))$ を計算する手法を合計 8 種類設計する（表2）．これらの手法は，

- イベントをシンボルとして扱うか分散表現を用いるか
- イベントの「顕現性」を考慮するか否か
- イベントの「出現順序」を考慮するか否か

（のみ）に違いを与えている．

3.1 物語における「イベント」

各物語 s の表現 $\phi(s)$ を構築する際，まずはじめに，物語テキストの**動詞**をイベントとして抽出する [5][*3]．以後，本稿での「物語」はすべてこのようにして取り出されたイベント群である．

3.2 イベントをシンボルで扱う場合

手法 a,b イベントの出現順序を考慮しない手法 a,b では，物語はイベントの頻度ベクトルで表現する．特に，顕現性を考慮する手法 b では，高頻度語を除去し[*4]，さらに TF-IDF で重み付けをおこなう．物語同士の類似度は，こうして構築されたベクトルのコサインで測る．

手法 c,d イベントの出現順序を考慮する手法 c,d では，物語はイベントの列で表現する．物語同士の類似度は，以下で説明する「編集類似度」を用いる．

編集類似度 イベントの系列 $\phi(s)$ と $\phi(s')$ 間の類似度は，Needleman–Wunsch アルゴリズム [7] に基づく系列アラインメントを介して計算される「編集類似度」を

[*3] イベントの表現として，動詞の項（主語や目的語など）をも考慮する研究もあるが [8, 9]，本研究では複数の尺度間の違いをできるだけ統制できるよう，もっとも簡単に動詞のみをイベントの表現とした．

[*4] 8 割以上の物語に出現するイベントとストップワードを除去．

表2: 手法概略及び実験結果

	イベント	顕現性	出現順序	物語の表現 – イベント群の集約	類似度計算	イベントのマッチング	ARI	AMI
a	シンボル			頻度ベクトル	コサイン	-	.020	.065
b	シンボル	✓		TF-IDFベクトル ※高頻度語除去	コサイン	-	.064	.174
c	シンボル		✓	系列	編集類似度	完全一致	.025	.071
d	シンボル	✓	✓	系列 ※高頻度語除去	編集類似度	完全一致 × TF-IDF	.046	.139
e	分散表現			平均	コサイン	-	.030	.086
f	分散表現	✓		TF-IDF重み平均 ※高頻度語除去	コサイン	-	.022	.065
g	分散表現		✓	単語ベクトルの系列	編集類似度	コサイン	.043	.107
h	分散表現	✓	✓	単語ベクトルの系列 ※高頻度語除去	編集類似度	コサイン × TF-IDF	.019	.056

用いる．アラインメントに際し，イベント $e \in \phi(s)$ と $e' \in \phi(s')$ のアラインメントスコアは次の通り：

$$\begin{cases} 1 & e = e' \quad \text{（置換）} \\ 0 & \text{（挿入，削除）} \end{cases} \tag{1}$$

顕現性を考慮する手法の場合は，置換のスコアに e の TF-IDF 値と e' の TF-IDF 値を掛け合わせる．

3.3 イベントを分散表現で扱う場合

手法 e,f イベントの出現順序を考慮しない手法 a,b では，物語はイベントの分散表現の平均で表現する．イベントの分散表現として，今回は GloVe [10] による訓練済みベクトルを利用した[*5]．物語同士の類似度は，こうして構築されたベクトルのコサインで測る．

手法 g,h イベントの出現順序を考慮する手法 g,h では，物語はイベントの分散表現の列で表現する．物語同士の類似度を「編集類似度」で計算する際は，置換スコアにイベントの分散表現間のコサインを掛け合わせる．

4 実験

物語の類似性を計算する際，物語中のイベントの顕現性・出現順序・及びその組み合わせがどの程度有効であるかを検証する．先に述べた複数の「物語表現・類似尺度」のそれぞれを用い物語の集合をクラスタリングし，クラスタリング結果と ATU 分類に基づいた人手での分類結果との一致度を確認する．

4.1 データセット・正解ラベル

クラスタリングの対象となる物語データセットとして，Ashliman 氏の Web サイトから収集した 144 編の動物に関わる物語を用いた．正解の分類は，各物語に対して Ashliman 氏によって付与された ATU 分類（22 クラス）をそのまま用いた．データセットの基本統計量は表3の「ANIMAL TAILS」の通り．

[*5] https://nlp.stanford.edu/projects/glove/

表3: 実験で用いた物語データセットの基本統計量

		全体	ANIMAL TALES
# クラス		102	22
# 物語		793	144
# 物語 / クラス	最大	22	20
	最小	2	2
	平均	7.8	6.5
	標準偏差	4.4	4.4
# 単語 / 物語	最大	8968	3837
	最小	39	63
	平均	1112	885
	標準偏差	1089	734

データセット構築手順 民俗学研究者である D. L. Ashliman 氏の管理する Web サイト Folklore and Mythology Electronic Text[*6]には世界各地の民話がその類型に基づいて収集・英訳・分類されている．まず，この Web サイトから ATU 分類がアノテートされている 793 編の民話を抽出した．とくに今回の実験では，ATU 分類の最上位の分類の一つである `1-299: ANIMAL TALES` に属するデータ（22 クラス，144 物語）のみを用いた[*7]．

4.2 クラスタリングアルゴリズム

物語表現と**非**類似尺度を元に物語の集合をクラスタリングするアルゴリズムとして，今回は階層的凝集型クラスタリング（群平均法）を採用した[*8]．各手法の類似度 $\mathrm{sim}(s, s')$ から非類似度 $\mathrm{dist}(s, s')$ への変換は，

[*6] https://www.pitt.edu/~dash/folktexts.html

[*7] ATU 分類は，「まず登場人物で粗く物語を分類し，次により詳細な情報（たとえばイベント）を使って細分化する」分類方法である．したがって，全く別の登場人物（たとえば魔法使いとキツネ）が似たイベントをおこなうふたつの物語は，ATU 分類では異なるクラスに分類されてしまう．本研究では，登場人物についてはある程度統一されているサブクラス，すなわちイベント群の一致とクラスの一致に強い相関があると考えられるサブクラスの中での細分類を試みる．

[*8] k-平均法も一般的に用いられるが，今回は，距離の公理を満たさない非類似度を統一的に扱えることと，初期値に依存しない結果を得られることから，階層的凝集型クラスタリングを採用した．

$\mathrm{dist}(s, s') := \max_{s,s'} [\mathrm{sim}(s, s')] - \mathrm{sim}(s, s')$ による．

4.3 クラスタリング結果の評価方法

物語の集合を各類似尺度でクラスタリングした結果（分割）$\widehat{C}$ と，ATU 分類に基づき人手で分類した結果（分割）C がどの程度一致しているかを評価する尺度として，adjusted Rand index（以下 **ARI**）[4] および adjusted mutual information（以下 **AMI**）[13] を用いる．ARI は「物語のペア」の全ての組み合わせについて，それらが $\widehat{C}$ および C で同じクラスタに分類されたか異なるクラスタに分類されたかの一致率である．AMI は，分割を物語集合上の離散分布だと思った際の，離散分布同士の相互情報量である．いずれも，ランダムに分割した場合に 0，完全に分割が一致していた場合に 1 となるよう調整されている．

4.4 実験結果

実験結果を表2の右側に示す．

顕現性の効果　「手法 a→ 手法 b」及び「手法 c→ 手法 d」ではスコアが向上しており，シンボルを用いる手法においては，イベントの顕現性を考慮することが有効であった．一方，「手法 e→ 手法 f」及び「手法 g→ 手法 h」ではスコアの低下が見られた．分散表現の和を用いる手法においては，TF-IDF や高頻度語除去を適用するだけでは顕現性の考慮を活かせないことが分かる．

出現順序の効果　「手法 a→ 手法 c」及び「手法 e→ 手法 g」ではスコアが向上しており，「手法 b→d」及び「手法 f→ 手法 h」ではスコアの低下が見られた．実験結果からは，顕現性を考慮していない場合に限り出現順序を考慮することが有効であることがわかった．

分散表現の効果　手法 a，b，c，d のスコアは，手法 e，f，g，h よりも概して高い結果となった．意外なことであるが，今回の手法群に関しては，イベントの表現としてシンボルを用いる方が分散表現を用いるよりも物語の類似性尺度としては有用という結果であった．

4.5 考察

分散表現の和を用いる手法においては，顕現性の効果は見られなかった（「手法 e→ 手法 f」及び「手法 g→ 手法 h」でスコアが低下した）．その理由として，顕現性を考慮する方法に TF-IDF を使用していたことが考えられる．TF（単語頻度）を考慮することは，物語テキスト中に頻出するイベントを顕現的であるとみなすことになるが，「赤ずきん」の例でも分かる通り，重要で顕著なイベントは一度しか起きない可能性がある．

今回イベントとして物語テキストから抽出してきた動詞のみを使用した．イベントの表現としては（主語，動詞）や（主語，動詞，目的語）なども考えられる．イベントによっては，注目している項が主語にも目的語にもなり得，さらにそれによりイベントの意味が大きく変わる場合がある（たとえば “kill” という動詞がこれにあたる）．今回はもっとも簡単な例として動詞のみを用いたが，項を考慮することでイベントの顕現性や出現順序を考慮する効果が変わることも考えられる．

謝辞　本研究の一部は JST CREST(課題番号: JPMJCR1513), 及び株式会社日立製作所の支援を受けて行った．

参考文献

[1] Jan Harold Brunvand. *A type index of urban legends. Encyclopedia of Urban Legends*. ABC-CLIO, 2012.

[2] Snigdha Chaturvedi, Shashank Srivastava, and Dan Roth. “Where Have I Heard This Story Before? Identifying Narrative Similarity in Movie Remakes”. In: *NAACL-HLT*. 2018.

[3] Bernhard Fisseni and Benedikt Löwe. “Which dimensions of narratives are relevant for human judgments of story equivalence ?” In: 2012.

[4] Lawrence Hubert and Phipps Arabie. “Comparing partitions”. In: *Journal of Classification* (1985).

[5] Victoria Anugrah Lestari and Ruli Manurung. “Measuring the Structural and Conceptual Similarity of Folktales using Plot Graphs”. In: *LaTeCH@ACL*. 2015.

[6] Inderjeet Mani. *Synthesis Lectures on Human Language Technologies. Computational Modeling of Narrative*. Morgan Claypool Publishers, 2012.

[7] Saul B. Needleman and C D Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins.” In: *Journal of molecular biology* 48 3 (1970), pp. 443–53.

[8] Dong Nguyen, Dolf Trieschnigg, and Mariët Theune. “Folktale Classification Using Learning to Rank”. In: *ECIR*. 2013.

[9] Dong Nguyen, Dolf Trieschnigg, and Mariët Theune. “Using Crowdsourcing to Investigate Perception of Narrative Similarity”. In: *CIKM*. 2014.

[10] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “GloVe: Global Vectors for Word Representation”. In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543.

[11] Vladimir Iakovlevich Propp. *Morphology of the Folktale*. The University of Texas Press, 1968.

[12] Hans-Jörg Uther. *The types of international folktales : a classification and bibliography : based on the system of Antti Aarne and Stith Thompson*. Helsinki : Suomalainen Tiedeakatemia, 2004.

[13] Nguyen Xuan Vinh, Unsweduau Julien Epps, and James Bailey. “Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance”. In: *Journal of Machine Learning Research* (2010).