物語におけるイベントの顕現性推定と物語類似性計算への応用

大竹 孝樹 [1]　横井 祥 [1,2]　井之上 直也 [2,3]　高橋 諒 [1,2]　栗林 樹生 [1,4]　乾 健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所　[3]Stony Brook University [4]Langsmith 株式会社
{takaki, yokoi, ryo.t, kuribayashi, inui}@ecei.tohoku.ac.jp
naoya.inoue.lab@gmail.com

1　はじめに

物語はあるイベントや一連のイベントの表現である [1]．物語に登場するイベントはその**顕現性**（重要さ）が異なり，物語において大きな役割を果たす重要なイベントとそうでないイベントが存在する．童話『シンデレラ』を例に挙げると，「シンデレラが王子に見初められる」というイベントは物語の進行に大きく関わる顕現性の高いイベントだが，「シンデレラが井戸で水を汲む」はそうではない．このようなイベントの顕現性推定は物語の自動生成などのタスクに役立ち，また物語分析のためのツールとしても有用である [2, 3]．物語におけるイベントの顕現性推定の有望な応用先として，**物語の類似性**計算が挙げられる．私たち人間は物語の類似性を自然に認識することができ，例えば「恵まれない境遇の女性が幸運を掴み成功を手に入れる」という物語は我々に『シンデレラ』との類似を思い起こさせる．こういった物語の類似性・類型を対象にした研究は人文学 [4, 5] 及び自然言語処理分野 [6, 7, 8] で盛んに行われてきた．とくに民俗学や物語論では，**顕現性の高いイベント**に着目して物語の類型分類・分析を行う研究が多く存在し [5, 4, 9]，これらの研究は，物語の類似性を計算機によって計算する際にイベント顕現性を考慮することの重要性を示唆している．

本研究では，物語論においてロラン・バルト（Roland Barthes）によって提案されたイベント顕現性の概念である**枢軸機能体（Cardinal Functions）** [10, 11] の定義[1)]に基づき，あるイベントの顕現性を“そのイベントを物語から削除したときに，物語全体としての首尾一貫性が損なわれる度合い”として推定する手法を提案する [13]（2 節）．人手によりイベント顕現性が付与された昔話コーパスに提案法を適用し，その有効性を実験的に検証する（3 節）．

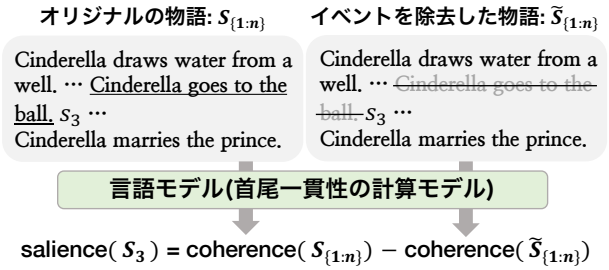

図 1　バルトの枢軸機能体の概念に基づいた物語におけるイベントの顕現性推定手法の概要．

加えて，提案法が物語の類似性計算に有用であるか検証するため，提案法を利用した物語の類似性計算方法を検討する．民俗学の専門家により物語の類型が付与された民話コーパスにおける実験で，提案法によりイベント顕現性を考慮する効果を実験的に検証する（4 節）．

2　物語におけるイベント顕現性推定

2.1　タスク設定

本研究では Ouyang ら [2] に従い，イベントそのものではなく文の顕現性を推定するタスクに取り組む．すなわち，物語を構成する各文に対して，その文が顕現性の高いイベントを含む度合いを推定する．形式的には，n 個の文から構成される物語 $S_{\{1:n\}} := \{S_1, \ldots, S_n\}$ と，ターゲット文 $S_k \in S_{\{1:n\}}$ が与えられ，タスクは S_k に対する顕現性スコア $\sigma(S_k, S_{\{1:n\}}) \in \mathbb{R}$ を予測することである．

2.2　提案手法

概要　バルトの枢軸機能体の定義に基づき，顕現性スコア $\sigma(S_k, S_{\{1:n\}})$ を“文 S_k に含まれる全てのイベントを $S_{\{1:n\}}$ から削除した際に，物語全体としての首尾一貫性が損なわれる度合い”として計算する．提案法の概要を図 1 に示す．

1）　枢軸機能体は“物語の行動にとって論理的に本質的なものであって，そのひとつでも削除されれば，その因果的・年代記的な首尾一貫性は損なわれてしまう”[12] ものと定義される．

r を後述するイベント削除関数とし，$\widetilde{S}_{\{1:n\}} := \{S_{\{1:k-1\}}, r(S_k), S_{\{k+1:n\}}\}$ を，$S_{\{1:n\}}$ において S_k に含まれる全てのイベントが削除された物語とする．$c(S)$ を物語 S の首尾一貫性スコアとすれば，S_k に対する顕現性スコアは以下の式で与えられる．

$$\sigma(S_k, S_{\{1:n\}}) := c(S_{\{1:n\}}) - c(\widetilde{S}_{\{1:n\}}) \tag{1}$$

イベントの削除方法　イベント削除関数 r として以下の 3 つを検討する．

- **文削除**: 文そのものを削除する
- **動詞置換**: 文に含まれる全ての動詞を，“do”, “does”, “did” など一般的な動詞で置き換える [2)]
- **動詞・項置換**: 動詞置換と同様の処理に加え，動詞の項（主格・目的格）を不定代名詞 “someone”, “something” で置き換える [3)]

物語の首尾一貫性の計算方法　See ら [14] に従い，事前学習済み言語モデルによって計算される物語テキストの生成確率をその物語の首尾一貫性スコアと見なす．物語テキストの首尾一貫性スコア $c(S)$ は以下の式で与えられる．

$$c(S_{\{1:n\}}) := \frac{1}{|S_{\{k+1:n\}}|} \log P(S_{\{k+1:n\}} \mid S_{\{1:k-1\}}, S_k), \tag{2}$$

$$c(\widetilde{S}_{\{1:n\}}) := \frac{1}{|S_{\{k+1:n\}}|} \log P(S_{\{k+1:n\}} \mid S_{\{1:k-1\}}, r(S_k)), \tag{3}$$

ここで $|S_{\{k+1:n\}}|$ は $S_{\{k+1:n\}}$ に含まれるトークンの総数である．詳細を付録 A.1 に示す．

3　実験 1: 物語中の文の顕現性推定

本節では，提案法が物語におけるイベント顕現性を推定できるかどうかを実験的に確かめる．具体的には，人手によりイベント顕現性が付与された物語コーパスを用いて，各提案法の性能を複数のベースライン手法と比較する．

3.1　実験設定

データセット　評価用データセットとして，Finlayson [15] によって提案された ProppLearner コーパスを使用した．コーパスの基本統計量を付録 A.2 の表 3 に示す．コーパス中の各物語には，動詞に対して Propp の機能[4)]（顕現性の高いイベント）が付与されている．2.1 節のタスク設定に従い，各手法はそのような動詞を含む文を判別する．

提案法に用いる言語モデル　首尾一貫性の計算に使用する言語モデルとして，事前学習済みの GPT-2 [17] を使用した．物語ドメインへの適応の度合いが異なる 3 種類の fine-tuning の設定を検討する．

- **fine-tuning なし**: 事前学習済みの GPT-2 をそのまま使用する．
- **BookCorpus**: 一般的な物語ドメインへの適応を狙い，物語ドメインの大規模コーパスである BookCorpus [18] で fine-tuning する．
- **ProppLearner**: トランスダクティブ設定 [19, 20] のドメイン適応として，ProppLearner コーパスで fine-tuning する．

ベースライン手法　提案法の比較対象として，以下のベースライン手法を検討する:

- ランダムベースライン: 各文に対して $[0, 1)$ の範囲のランダムなスコアを与える．
- TF-IDF ベースライン: 各文に対してその文を構成する単語の TF-IDF 値の和を与える．

提案法と TF-IDF ベースラインの組み合わせ手法　加えて，各提案法と TF-IDF ベースラインを組み合わせる手法を検討する．各提案法と TF-IDF ベースラインが各文に対して計算する顕現性スコアはまず各物語内で $[0, 1]$ の値に正規化され，それら 2 つの値の和が組み合わせ手法の顕現性スコアとなる．

以上より，提案法はイベント削除方法 (3 種類) × 言語モデルの fine-tuning 設定 (3 種類) = 9 種類のバリエーションをもつ．TF-IDF ベースラインとの組み合わせ手法を含めると計 18 種類となる．

評価指標　Liu ら [21] に従い，顕現性推定をランキング問題とみなして評価を行う．すなわち，各手法はそれぞれの物語に含まれる文を顕現性スコアの高い順にランク付けし，最終的に構成されるランキングの良さを評価する．評価尺度として，2 値適合性に基づくランキングの評価に一般的に用いられる Mean Average Precision (MAP) [22] を使用した．

3.2　実験結果

表 1 に各手法の結果を示す．**+TF-IDF** に対応する行は組み合わせ手法の結果を表している．全提案法

2)　置換先の一般的な動詞として “do”, “does”, “did”, “done”, “doing” を使用する．例えば，品詞タグが VBZ（三人称単数現在形の動詞）である場合，“does” で置き換える．

3)　ARG0 (agent) に該当するテキストのスパンを “someone” で，ARG1 (patient) に該当するスパンを “something” で置き換える．

4)　Propp の機能は “登場人物の行為で、しかも、筋＝出来事全体の展開過程にとって当の行為がもちうる意義 (位置) という観点から規定された登場人物の行為” [16] と定義される．

は一貫してランダムベースラインの性能を上回り，提案法単独では (文削除, ProppLearner) が最も高い性能を示した．

イベントを削除する方法の比較　動詞置換や動詞・項置換と比べて，文削除が概ね良い性能を示していることがわかる．動詞置換や動詞・項置換によるイベント削除は文からイベントの情報のみをより精緻に削除することが期待されたが，実験結果はこれらのイベント削除方法は効果的でないことを示している．動詞置換や動詞・項置換では，文に含まれる単語を置換する操作によって不自然な文が生成され，こうした文が言語モデルの推論に悪影響を及ぼしている可能性が考えられる．

言語モデルの fine-tuning の効果　GPT-2 を BookCorpus で fine-tuning した場合には一貫して提案法の MAP スコアが改善しており，またトランスダクティブ設定の fine-tuning をした場合には文削除と動詞・項置換による提案法の MAP スコアが改善していることが確認できる．これらの結果から，首尾一貫性の計算に使用する言語モデルを物語ドメインに適応させることが有効であることがわかる．

提案法と TF-IDF ベースラインの組み合わせ手法　組み合わせ手法は全ての場合で各提案法そのもの，あるいは TF-IDF ベースラインそのものの性能を一貫して上回る結果となり，最終的に提案法（文削除, BookCorpus）と TF-IDF ベースラインの組み合わせが全手法の中で最も高い性能を達成した．これらの結果は，バルトの枢軸機能体の概念に基づいて提案法が計算するイベント顕現性の手がかりと，単語の頻度・逆文書頻度に基づく手がかりが相補的な関係にあり，これらが統合されることでより良いイベント顕現性の尺度となっていることを示唆している．

付録 A.3 の表 5 に童話『シンデレラ』の toy example に対する提案法の実際の振る舞いを示す．

4　実験 2: 物語類似性計算への応用

本節では，提案法が計算する顕現性スコアが物語の類似性計算に有用であるか実験的な検証を行う．具体的には，まずイベント顕現性を考慮する/しないという点のみが異なる物語の類似性計算方法を検討する．次にそれらを専門家による分類が付与された物語データセットに適用し，内的/外的な評価を行うことで，提案法により顕現性を考慮することで物語の類似性計算が改善するかどうかを確かめる．

手法	+ TF-IDF	Fine-tuning	MAP
ランダム	-	-	0.213
TF-IDF	-	-	0.279†
提案法（文削除）	-	-	0.261†
		BookCorpus	0.265†
		ProppLearner	**0.280**†
	✓	-	0.294†
		BookCorpus	***0.301***†
		ProppLearner	0.295†
提案法（動詞置換）	-	-	0.245
		BookCorpus	0.258†
		ProppLearner	0.219
	✓	-	0.286†
		BookCorpus	0.287†
		ProppLearner	0.266†
提案法（動詞・項置換）	-	-	0.254†
		BookCorpus	0.258†
		ProppLearner	0.266
	✓	-	0.285†
		BookCorpus	0.295†
		ProppLearner	***0.301***†

表 1　各手法の MAP スコア．ランダムベースラインはシード値を変えて 10 回の実験を行った際の平均値である (標準偏差 = 0.015)．ダガーのついた値はランダムベースラインから $p < 0.05$ で統計的に有意な差があったものを示している．統計的検定は Wilcoxon signed-rank test [23] により行った．太字の値は提案法単独で最も性能が高いものを，太字かつ斜体の値は提案法と TF-IDF ベースラインの組み合わせ手法の中で最も性能が高いものを示す．

4.1　物語類似性の計算方法

物語テキスト S 及び S' の埋め込み表現をそれぞれ $\mathbf{S}, \mathbf{S}'$ とする．$\mathbf{S}, \mathbf{S}'$ はそれぞれ物語 S 及び S' を構成する文の**文ベクトル和**である．本研究では S と S' の類似度 $\mathrm{sim}(S, S')$ をコサイン類似度 $\cos(\mathbf{S}, \mathbf{S}')$ で計算する．

イベント顕現性を考慮しない場合:　文ベクトルを，文を構成する単語の単語ベクトルの**和**とする．

イベント顕現性を考慮する場合:　文ベクトルを，文を構成する単語の単語ベクトルの**顕現性重みつき和**とする．重みは提案法が文に対して計算した顕現性スコア[5)]であり，(文削除, fine-tuning なし) 及び (文削除, BookCorpus) の 2 通りを検討する．比較対象として，重みを単語の TF-IDF 値とするベースラインを検討する．また各提案法と TF-IDF ベースラ

5)　顕現性スコアは負の値をとる場合もあるため，重みは 1 + [0, 1] に正規化した顕現性スコアとした．
6)　重みを 1 + (正規化した顕現性スコアと正規化した TF-IDF 値の平均値) とした．

顕現性重み付け	+ TF-IDF	RSA Score	ARI @最下層	ARI @中間層
なし	-	.1658	.0045	.0024
TF-IDF	-	**.1780**	.0029	.0034
提案法 (文削除, fine-tuning なし)	-	.1706	.0047	.0040
	✓	.1724	.0049	.0052
提案法 (文削除, BookCorpus)	-	.1707	.0044	.0040
	✓	.1725	**.0060**	**.0054**

表 2　物語の類似性計算の実験結果．太字の値は各指標で最も性能が高いものを示す．+TF-IDF に対応する行は各提案法と TF-IDF ベースラインの組み合わせ手法を示す．

インの組み合わせ手法[6)]を検討する．全ての場合で重み付けを行う単語は文に含まれる動詞とその項（主格・目的格）とした．

4.2　実験設定

データセット　評価用データセットとして，大竹ら [24] が作成した 793 編の民話からなるデータセットを使用した．データセットの基本統計量を付録 A.2 の表 4 に示す．各物語には，民俗学において最も一般的な民話の分類体系の一つである **ATU 分類** [5] が付与されている．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「Crimes punished (罪が罰せられる)」）に基づき階層的に分類した体系である．

内的な評価方法: RSA Score　ATU 分類の分類木上での物語の距離構造が，どの程度埋め込み空間にエンコードされているかを **RSA Score** [25, 26] を用いて定量的に評価する．今回使用する民話データセットの埋め込み空間における表現を r_1，分類木上での離散的な表現を r_2 とする．Bouchacourt ら [26] に従い，r_1 内で物語ペアの全組み合わせについて類似度を計算し，その類似度列を s_1，r_2 において同様に求めた類似度列を s_2 として，s_1 と s_2 の間のスピアマンの順位相関係数を RSA Score とする．ただし，分類木における物語間の類似度は，物語 S と S' の分類木上でのパス長を $\mathrm{dist}(S, S')$ とし，$\max_{S,S'}[\mathrm{dist}(S, S')] - \mathrm{dist}(S, S')$ として計算した．RSA Score の値が高ければ，ATU 分類における物語の距離構造が，埋め込み空間によく反映されている（すなわち，r_1 で近い物語ペアは r_2 で近く，r_1 で遠い物語ペアは r_2 で遠くなっている）ことを示す．

外的な評価方法: ARI によるクラスタリング評価　先に述べた複数の類似性計算方法により物語の集合をクラスタリングし，結果が専門家による分類である ATU 分類とどの程度一致するかを評価する．クラスタリング結果の分割 $\widehat{C}$ と正解の分割 C の一致の尺度として，一般的に用いられる Adjusted Rand Index（以下 **ARI**）[27] を用いた．また，クラスタリングアルゴリズムとして階層的凝集型クラスタリング（群平均法）を使用した．クラスタリングとその結果の評価は，ATU 分類の中間層における分類 (例: ATU type1-99『野生動物』)，及び最下層における分類 (例: ATU type333『赤ずきん』) の 2 通りで行った．

内的/外的な評価実験共に，単語ベクトルとして学習済みの GloVe [28] を使用した．

4.3　実験結果

結果を表 2 に示す．RSA Score による内的な評価に注目すると，イベント顕現性を考慮しない場合と比較して，提案法によってイベントの顕現性を考慮することで物語の類似性計算が改善していることがわかる．クラスタリングによる外的な評価に注目すると，提案法単独の重み付けをおこなった場合，ATU 分類の最下層におけるクラスタリングでは効果が見られなかったものの，中間層においては提案法のイベント顕現性スコアを考慮することでより正解の分類と近いクラスタが形成されていることがわかる．また実験 1 の結果（3 節）と同様，評価指標によらず提案法と TF-IDF ベースラインを組み合わせることで提案法単体よりも性能が向上していることが確認できる．これらの結果は，バルトの枢軸機能体の概念に基づく提案法が物語の類似性計算に役立つ可能性を示唆するものである．

5　おわりに

本論文では，物語論におけるイベント顕現性の概念であるロラン・バルトの枢軸機能体の定義を足掛かりに，物語におけるイベントの顕現性を言語モデルを利用して推定する教師なし手法を提案した．人手によりイベント顕現性が付与された昔話データセットにおける実験で，提案法はベースライン手法の性能を上回ることを示した．加えて，物語におけるイベント顕現性推定の有望な応用先として，提案法が物語の類似性計算に有用であるか検証を行った．専門家による類型分類が付与された民話データセットにおける実験で，提案法が物語の類似性計算に役立つ可能性が示唆された．

謝辞　本研究は JSPS 科研費 JP19H04425 の助成を受けたものである．

参考文献

[1] H Porter Abbott. *The Cambridge Introduction to Narrative*. Cambridge Introductions to Literature. Cambridge University Press, 2 edition, 2008.

[2] Jessica Ouyang and Kathleen McKeown. Modeling Reportable Events as Turning Points in Narrative. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2149–2158, Lisbon, Portugal, 9 2015. Association for Computational Linguistics.

[3] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Movie Plot Analysis via Turning Point Identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1707–1717, Hong Kong, China, 11 2019. Association for Computational Linguistics.

[4] Vladimir Propp. *Morphology of the Folktale*. University of Texas Press, 1968.

[5] Hans-Jörg Uther. *The Types of International Folktales: A Classification and Bibliography, based on the system of Antti Aarne and Stith Thompson*, Vol. 1-3. Suomalainen Tiedeakatemia, Academia Scientiarum Fennica, Helsinki, 2004.

[6] Lisa Friedland and James Allan. Joke retrieval: recognizing the same joke told differently. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pp. 883–892, Napa Valley, California, USA, 2008. Association for Computing Machinery.

[7] Dong Nguyen, Dolf Trieschnigg, and Mariët Theune. Using Crowdsourcing to Investigate Perception of Narrative Similarity. In *CIKM*, CIKM '14, pp. 321–330, New York, NY, USA, 2014. ACM.

[8] Snigdha Chaturvedi, Shashank Srivastava, and Dan Roth. Where Have I Heard This Story Before? Identifying Narrative Similarity in Movie Remakes. In *NAACL*, pp. 673–678, New Orleans, Louisiana, 6 2018. Association for Computational Linguistics.

[9] A Aarne and S Thompson. *The types of the folktale: a classification and bibliography*. FF communications. Suomalainen Tiedeakatemia, Academia Scientarum Fennica, Helsinki, 1961.

[10] Roland Barthes. Introduction à l'analyse structurale des récits. *Communications, 8, 1966. Recherches sémiologiques : l'analyse structurale du récit.*, 1966.

[11] Roland Barthes and Lionel Duisit. An Introduction to the Structural Analysis of Narrative. *New Literary History*, Vol. 6, No. 2, p. 237, 1975.

[12] Gerald Prince, 遠藤健一. 物語論辞典. 松柏社叢書, 言語科学の冒険 4. 松柏社, 改訂, 2015.

[13] Takaki Otake, Sho Yokoi, Naoya Inoue, Ryo Takahashi, Tatsuki Kuribayashi, and Kentaro Inui. Modeling Event Salience in Narratives via Barthes' Cardinal Functions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1784–1794, Barcelona, Spain (Online), 12 2020. International Committee on Computational Linguistics.

[14] Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. Do Massively Pretrained Language Models Make Better Storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 843–861, Hong Kong, China, 11 2019. Association for Computational Linguistics.

[15] Mark A Finlayson. ProppLearner: Deeply annotating a corpus of Russian folktales to enable the machine learning of a Russian formalist theory. *Digital Scholarship in the Humanities*, Vol. 32, No. 2, pp. 284–300, 2015.

[16] Vladimir Propp, 北岡誠司, 福田美智代. 昔話の形態学. 叢書記号学的実践, No. 10. 白馬書房, 1987.

[17] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, Vol. 1, No. 8, 2019.

[18] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *The IEEE International Conference on Computer Vision (ICCV)*, 12 2015.

[19] Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998.

[20] Hiroki Ouchi, Jun Suzuki, and Kentaro Inui. Transductive Learning of Neural Language Models for Syntactic and Semantic Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3665–3671, Hong Kong, China, 11 2019. Association for Computational Linguistics.

[21] Zhengzhong Liu, Chenyan Xiong, Teruko Mitamura, and Eduard Hovy. Automatic Event Salience Identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1226–1236, Brussels, Belgium, 2018. Association for Computational Linguistics.

[22] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. Evaluation in information retrieval. In *Introduction to Information Retrieval*, p. 139–161. Cambridge University Press, 2008.

[23] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, Vol. 1, No. 6, pp. 80–83, 1945.

[24] Recurrent Neural and Point-wise Mutual. 顕現的要素の出現順序に基づく物語の類似性尺度. 言語処理学会 第 25 回年次大会 発表論文集, pp. 304–307, 2019.

[25] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, Vol. 2, No. NOV, pp. 1–28, 2008.

[26] Diane Bouchacourt and Marco Baroni. How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 981–985, Brussels, Belgium, 2018. Association for Computational Linguistics.

[27] Lawrence Hubert and Phipps Arabie. Comparing Partitions. *Journal of Classification*, Vol. 2, No. 1, pp. 193–218, 1985.

[28] Jeffrey Pennington, Richard Socher, and Christopher Manning. {G}love: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, 10 2014. Association for Computational Linguistics.

[29] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771*, 2019.

[30] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

[31] Chikashi Nobata, Satoshi Sekine, and Hitoshi Isahara. Evaluation of Features for Sentence Extraction on Different Types of Corpora. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pp. 29–36, Sapporo, Japan, 7 2003. Association for Computational Linguistics.

[32] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke S Zettlemoyer. AllenNLP: A Deep Semantic Natural Language Processing Platform. 2017.

A 付録

A.1 提案法詳細

首尾一貫性の計算に使用した GPT-2 は，一度に入力できるテキストの長さに制限があるため，実装上はターゲット文の前後の文脈を適当に捨てることで首尾一貫性スコア $c(S_{\{1:n\}})$ を計算した:

$$c(S) = \frac{1}{|S_{\{k+1:n-\ell'\}}|} \log P(S_{\{k+1:n-\ell'\}} \mid S_{\{1+\ell:k-1\}}, S_k), \tag{4}$$

$|S_{\{i:j\}}|$ は $(S_i, \ldots, S_j)$ に含まれるトークンの総数を示す．ℓ' 及び ℓ は言語モデルの最大入力長 L によって決まる閾値であり，$|S_{\{k+1:n-\ell'\}}| + |S_{\{1+\ell:k\}}|$ が L 以下かつ最大の値をとるように定める．$c(\widetilde{S})$ についても同様である．ただし $P(S_i)$ はトークンの生成確率の積である:

$$\log P(S_i|\text{context}) = \log P(w_1^{(i)}|\text{context}) + \sum_{j=2}^{|S_i|} \log P(w_j^{(i)}|\text{context}, w_1^{(i)}, \ldots, w_{j-1}^{(i)}). \tag{5}$$

また，各物語の最後の文の顕現性スコアを計算するため，各物語の末尾に文書の区切りを示す特殊トークンを添加した．提案法はこの特殊トークンの生成確率を利用することで，各物語の最後の文の顕現性スコアを他の文と同様に計算することが可能である．

A.2 実験に使用したデータセットの基本統計量

物語の総数	15
文の総数	1302
単語の総数	18862
機能（顕現性が高いイベント）の総数	170
1 物語あたりの平均文数	86.8
1 物語あたりの平均単語数	1257.5
1 物語あたりの平均機能数	11.3

表 3 ProppLearner コーパスの基本統計量

	物語の総数	793
	物語の平均単語数	1112
@最下層	クラス数	102
	1 クラスあたりの物語数（平均）	7.8
@中間層	クラス数	27
	1 クラスあたりの物語数（平均）	29.4

表 4 大竹ら [24] が作成した民話コーパスの基本統計量

A.3 提案法の toy example に対する振る舞い

顕現性の高いイベントを含むか否か		文	顕現性スコア
-	S_1	Cinderella draws water from a well.	0.193
✓	S_2	A fairy godmother appears and provides Cinderella with clothes, a carriage, and a coachman.	0.309
✓	S_3	Cinderella goes to the ball.	0.214
-	S_4	Cinderella greets her stepsisters at the venue , but they do not notice.	−0.014
✓	S_5	The prince falls in love with Cinderella.	0.394
✓	S_6	Cinderella marries the prince.	−0.112

表 5 童話『シンデレラ』の toy example に対して提案手法 (文削除,fine-tuning なし) を適用した際の振る舞い．

提案法の実際の物語に対する振る舞いを定性的に分析するため，6 文から構成される童話『シンデレラ』の toy example に対して提案法を適用した．表 5 に，各文に対して提案法（文削除，fine-tuning なし）が計算した顕現性スコアを示す．「魔法使いが現れ，シンデレラにドレスや馬車，御者を与える」や「シンデレラが王子に見初められる」といった顕現性の高いイベントを表現している文に対しては，「シンデレラが井戸で水を汲む」といった顕現性が低いイベントを表す文よりも高い顕現性スコアを与えていることがわかる．

A.4 実装の詳細

GPT-2 の実装は transformers [29] の事前学習済みモデル (12-layer, 768-hidden, 12-heads, 117M parameters) を使用した．実験 1, 2 で行った顕現性スコアのスケーリングには，*scikit-learn* [30] の MinMaxScaler を使用した．実験 1, 2 の TF-IDF ベースラインでは野畑ら [31] の研究において T3 と呼ばれる修正 TF-IDF 値を使用した．イベントを表す単語（動詞とその項）の抽出について，実験 1 では ProppLearner コーパスが品詞タグや述語項構造のアノテーションを含んでいるためこれらをそのまま利用した．実験 2 で使用した大竹ら [24] の民話コーパスはそのようなアノテーションを含まないため，Allen NLP [32] が提供している述語項構造解析器の実装[5)] を用い，その解析結果を利用した．

5） https://demo.allennlp.org/semantic-role-labeling