大規模言語モデルを用いた Story Intention Graph の自動生成の精度改善

吉川祐輔 [1]　井上壮志 [1]　銭本友樹 [2]　東中竜一郎 [2]
[1] インターメディアプランニング株式会社 [2] 名古屋大学大学院情報学研究科
{y_yoshikawa,t_inoue}@ipi.co.jp
{zenimoto.yuki.u1@s.mail, higashinaka@i}.nagoya-u.ac.jp

概要

対話システムが物語を理解できれば，話者の価値観や経験などをより深く理解することができるようになると考えられる．先行研究で我々は，物語の表現として Elson によって提案された Story Intention Graph (SIG) に着目し，大規模言語モデル（LLM）を用いて，物語文から SIG の自動生成に取り組んだ．しかしながら，SIG の構造は複雑であり，物語文から SIG を自動生成する精度は高くなかった．本研究では，より高度な推論が可能と考えられる LLM を用いた SIG の自動生成の改善を行う．具体的には, LLM として GPT-4o, および，推論力に優れた OpenAI o1 を採用し, プロンプトへの shot の導入を行った．実験では，自動生成された SIG を，人手によって作成された SIG と比較した．その結果，OpenAI o1 に shot の追加を行うことで，人手作成 SIG の約 76%という比較的高い精度で SIG を自動生成可能なことが分かった．

1　はじめに

人間同士の対話において，お互いの価値観や経験は物語によって表出されることが多い [1, 2]．そのため，対話システムが物語を理解できれば，話者の価値観や経験などをより深く理解することができるようになると考えられる．

物語理解はこれまでに重要なトピックとして研究が進められてきた．物語の表現の研究 [3] や，ナラティブやニュース等から，物語構造を自動的に獲得する研究 [4, 5] もなされてきた．我々は，Elson によって提案された Story Intention Graph (SIG) [6] と呼ばれる物語の表現に着目している．SIG は，物語のイベントと，その解釈，および，登場人物の意図や目的を構造化しており，複雑な内容を表すことができる [7]．童話のような物語以外にも，ブログなどに含まれる Personal Story を表現する用途にも用いられている [8]．

先行研究で我々は，大規模言語モデル（LLM）を用いて，物語文から SIG の自動生成に取り組んだ [9]．具体的には，GPT-4 を用い，プロンプトとして，SIG の要素の定義等を与えることで，SIG の自動生成を実施した．しかしながら，SIG の構造は複雑であり，SIG を自動生成する精度は高くなかった．評価も SIG の構造としての妥当性の観点からのみ行っており，内容面からの評価ができていなかった．

本研究では，より高度な推論が可能と考えられる LLM を用い，SIG の自動生成の改善を行う．具体的には，LLM として GPT-4o, および，推論力に優れた OpenAI o1 を採用し, プロンプトへの shot の導入を行う．さらに，SIG の評価尺度を作成し，自動生成された SIG の内容面からの評価をできるようにする．実験の結果，OpenAI o1 に shot の追加を行うことで，人手作成 SIG の約 76%という比較的高い精度で SIG を自動生成可能なことが分かった．

以降では，まず SIG および LLM を用いた SIG の自動生成について述べる．そして，SIG の評価手法について述べた後，評価実験について述べる．最後に，本稿のまとめを述べる．

2　Story Intention Graph

SIG は，物語をノード（フレーム）およびエッジでグラフ化することで, 物語のイベントの時系列と，その解釈，および，登場人物の意図や目的を表現する．

SIG は，ノード，フレーム，エッジからなる有向グラフである．ノードは命題を保持し，ノード間はエッジで接続される．一部のノードには，複数のノードがネストして含まれうるため，フレームとも

みなすことができる．ノード，フレームおよびエッジのタイプとして以下が存在する．

ノード　TE (Text), P (Proposition), I (Interpreted Proposition), A (Affect)

フレーム　B (Belief), G (Goal)

エッジ　f (follows), ia (interpreted as), wc (would cause), wp (would prevent), p (provides for), d (damages) など

TE ノードが物語のテキストを表現し，P ノードが出来事を表す．I ノードは出来事の解釈を表し，A ノードが登場人物の意図や目的などを表す．詳細については [6, 9] を参照されたい．

3　LLM による SIG の自動生成

LLM による SIG の自動生成は，先行研究 [9] と同様，以下のステップに対応するプロンプトを LLM に与えることで行う．

1. エージェントの特定
2. TE ノードの生成
3. P ノード，f エッジの生成
4. I ノード，関連エッジの生成
5. B フレーム，G フレーム，関連エッジの生成
6. エッジの追加生成
7. A ノード，p, d エッジの生成

本稿では，自動生成の精度改善のため，SIG の自動生成において次の 3 つの改善を行う．

- LLM に, GPT-4o と GPT-4o より推論能力が高い OpenAI o1 を採用する．
- プロンプトに shot を追加する．shot は文献 [6] に含まれるものを用いる．なお，Elson の文献ではノードの内容は述語項構造によって表現されるが，本研究ではノードの内容を自然言語に置き換えた上で shot として入力する．
- プロンプトにおける，各種の定義を明確化し，特定の指示や禁止事項の追加を明示的に行う．

これらの改善により，複雑な物語であっても，高い推論能力と事例の導入により，その構造を的確に理解し，内容的に妥当な SIG が生成されることを期待する．

4　SIG の評価方法

SIG の自動生成の改善にあたって，内容的に妥当な SIG が生成されたかどうかを評価するための評価指標が必要である．ここでは，人手で作成された SIG を正解として，自動生成された SIG と比較することで精度を計測することを考える．SIG は複雑なグラフ構造であるため，その比較方法は自明ではない．そこで本稿では，以下の評価指標を導入する．

まず，松原ら [10] の方法に倣い，SIG をラベル付き有向グラフ $G = (V, E, s, t; \mathcal{A}, \mathcal{R}, \phi, \psi)$ の組とする．ここで，各変数の意味は以下のとおりである．

V　ノード集合
E　エッジ集合
$s : E \to V$　エッジに始点ノードを割り当てる関数
$t : E \to V$　エッジに終点ノードを割り当てる関数
$\mathcal{A}$　ノード種類 (P, I, B, G, A) 集合
$\mathcal{R}$　エッジ種類 (ia, a, c, in, p, d など) 集合
$\phi : V \to \mathcal{A}$　ノードに種類を割り当てるノードマッピング関数
$\psi : E \to \mathcal{R}$　エッジに種類を割り当てるエッジマッピング関数

このうちエッジ集合 E を比較する. 比較対象とするグラフ G_1 のエッジ集合 E_1 の要素 e_1 と，もう一方のグラフ G_2 のエッジ集合 E_2 の要素 e_2 について，

$$\exists e_1 \in E_1, \quad e_2 \in E_2 \quad \psi(e_1) = \psi(e_2) \cap \phi(s(e_1)) = \phi(s(e_2)) \cap \phi(t(e_1)) = \phi(t(e_0))$$

を満たす e_2 が存在すれば，e_1 と e_2 の組み合わせを「一致ペア」とする. その集合を M_0 とする．ただし，エッジ集合 E_1, E_2 の各要素は各々 1 組しか「一致ペア」になれないものとして数え上げを行う．比較対象の一方のグラフ G_1 のエッジ集合 E_1 と，他方のグラフ G_2 のエッジ集合 E_2 について，Precision (p) と Recall (r) に相当する数値を計算する．

$$p = |M_0|/|E_1|, \quad r = |M_0|/|E_2|$$

その上で，この 2 値の調和平均 F_1 をとる．

$$F_1 = 2pr/(p + r)$$

この F_1 値が大きいほど 2 つのグラフが類似していると考える．

評価においては，正解データである人手作成の SIG を G_1 とし，LLM による自動生成の SIG を G_2 として F_1 を計算する．

5　実験

実験として，所定の物語文について，人手作成による SIG と自動生成による SIG の精度を比較した．

表 1　モデル別，shot の有無による F_1 値の平均値

	shot なし	shot あり
GPT-4o	0.393	0.409
OpenAI o1	0.422	**0.456**
人手作成同士	0.597	

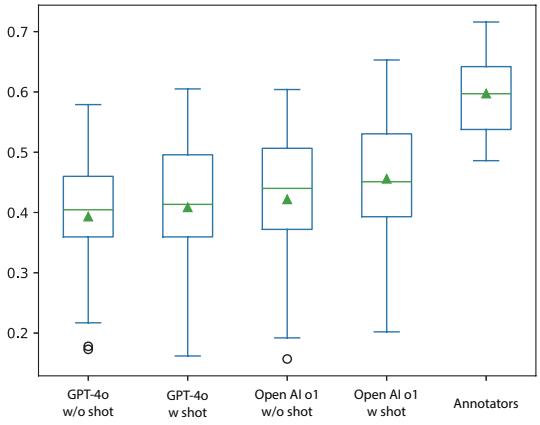

図 1　モデル別，shot 有無による F_1 値の箱ひげ図.Annotators は人手作成同士を表す.

まず，物語文として，Elson らの研究 [7] で使用されたイソップ童話のデータから 15 話を選定した．これらには，2 人の異なるアノテータによる SIG が収録されており，人間同士の精度も算出可能である．なお，物語文の言語は英語である．

選定された 15 話について，2 種類の大規模言語モデル（GTP-4o, OpenAI o1），shot の有無の組み合わせについて, 3 節の手法に従って SIG を自動生成した．評価尺度には，4 節で述べた F_1 値を用いた．

具体的な評価の流れであるが，ある組み合わせについて，15 の物語に対して 15 の SIG が自動生成される．ここで，それぞれの物語について 2 つの人手作成の SIG がある．自動生成された SIG のそれぞれと人手作成の 2 つの SIG を比較し，2 つの評価値（F_1）が得られる．これを 15 の物語について実施することで，30 の評価値が得られる．これらの F_1 の平均を今回は組み合わせに対する評価値として用いる．

5.1　結果

表 1 に実験結果を示す．表から，shot の追加，および，OpenAI o1 を用いる場合の評価値が最も高いことが分かる．基本的に，OpenAI o1 モデルのほうが性能が高く，shot 無しよりも有りのほうが性能が高い．人手作成同士の評価値は 0.597 であり，これはアッパーバウンドだと考えられる．OpenAI o1 使用，shot ありの場合の F_1 値は, 人手作成同士の値の約 76%の値であることから，比較的高い精度で SIG の自動生成が実現できたことが確認できる．図 1 は，表 1 を箱ひげ図として可視化したものである．

手法間での差が有意であるかを検証するため，各条件から得られた F_1 値に対して二項検定を行った．その結果，GPT-4o の場合, shot がある場合の方が，ない場合より SIG 作成能力が高いことが分かった．これは，GPT-4o が推論能力を shot によって補っていることを示唆している．また，shot ありで比較した場合, o1 の方が GPT-4o より SIG 作成能力が高いことが分かった．このことは，o1 のほうがその高い推論能力により，shot を効果的に使っていることを示唆している．

自動生成事例について取り上げると，図 2 は入力とした物語文の例である．そして，この入力に対して，OpenAI o1 に shot を与えて得られた SIG が図 3 である．適切なノードが生成されているとともに，妥当なエッジでそれらが紐づけられていることが見て取れる．

> A hungry Fox saw some fine bunches of Grapes hanging from a vine that was trained along a high trellis, and did his best to reach them by jumping as high as he could into the air. But it was all in vain, for they were just out of reach: so he gave up trying, and walked away with an air of dignity and unconcern, remarking, "I thought those Grapes were ripe, but I see now they are quite sour.

図 2　入力とした物語文の例: The Fox and The Grapes

6　おわりに

本研究では，物語の表現として Story Intention Graph (SIG) に着目し，大規模言語モデルを用いて，物語文から SIG の自動生成の改善に取り組んだ．具体的には, GPT-4o, および，推論力に優れた OpenAI o1 を採用し, プロンプトへの shot の導入を行った．その結果，OpenAI o1 に shot の追加を行うことで，人手作成 SIG の約 76%という比較的高い精度で SIG を自動生成可能なことが分かった．

今後の課題として，本研究で得られた LLM による自動生成 SIG を対話システムへ組み込み, より高度な対話の実現可能性を探りたい．

特に，高度な対話として, 我々が想定している 1 つが「共想法」[11] である．共想法は, 認知機能（記

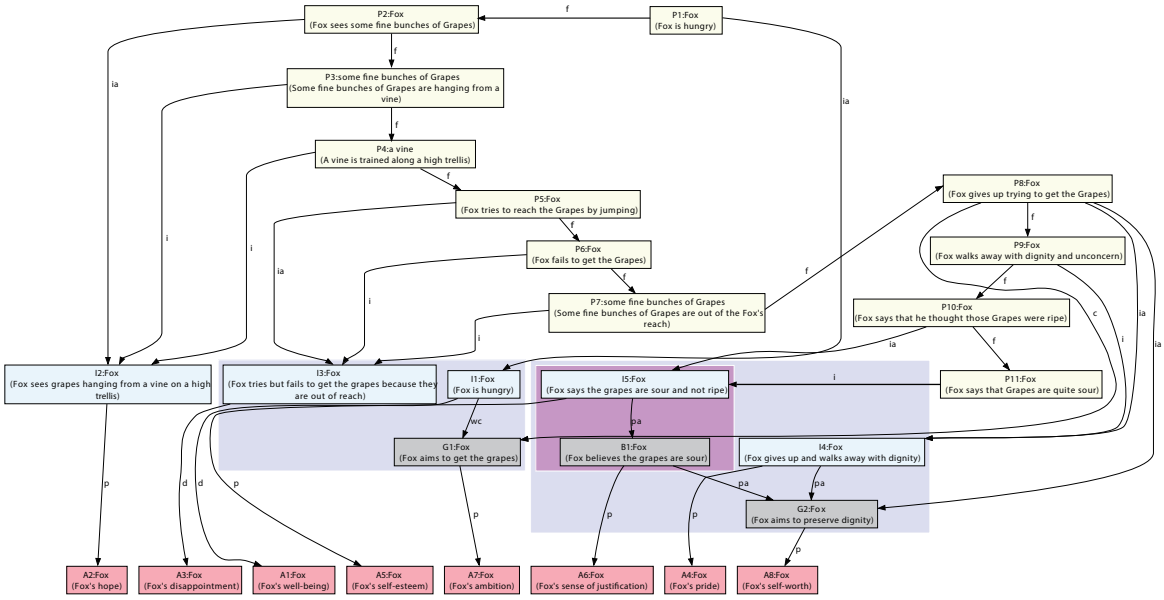

図 3　図 2 の物語文を入力として自動生成した SIG

憶体験・注意分割・計画力) を自然に活用する対話を引き出す対話支援手法であり, 話者から物語を聞き出すことが主要な課題である．自動化されたシステムの研究も進められている [12] が，SIG を用いて，より人間の発話を理解し人間に近い対話が可能なシステムが実現できれば，さらなる高齢者支援が実現できると考えている．

参考文献

[1] Roger C Schank. **Tell me a story: A new look at real and artificial memory.** Charles Scribner's Sons, 1990.

[2] Yuya Chiba and Ryuichiro Higashinaka. Analyzing variations of everyday Japanese conversations based on semantic labels of functional expressions. **ACM Trans. Asian Low-Resour. Lang. Inf. Process.**, Vol. 22, No. 2, 2023.

[3] Wendy G Lehnert. Plot units and narrative summarization. **Cognitive science**, Vol. 5, No. 4, pp. 293–331, 1981.

[4] Amit Goyal, Ellen Riloff, and Hal Daumé III. Automatically producing plot unit representations for narrative text. In **Proc. EMNLP**, pp. 77–86, 2010.

[5] Jessica Ouyang and Kathleen McKeown. Modeling reportable events as turning points in narrative. In **Proc. EMNLP**, pp. 2149–2158, 2015.

[6] David K. Elson. **Modeling Narrative Discourse**. PhD thesis, Columbia University, 2012.

[7] David K. Elson. Dramabank: Annotating agency in narrative discourse. **Proceedings of the Eighth International Conference on Language Resources and Evaluation**, pp. 2813–2819, 2012.

[8] Stephanie Lukin, Kevin Bowden, Casey Barackman, and Marilyn Walker. Personabank: A corpus of personal narratives and their story intention graphs. **Proceedings of the Tenth International Conference on Language Resources and Evaluation**, pp. 1026–1033, 2016.

[9] 吉川祐輔, 井上壮志, 東中竜一郎. 語り直しを目的とした大規模言語モデルを用いた story intention graph の作成とその評価. 言語処理学会 第 30 回年次大会 発表論文集, pp. 1423–1426, 2024.

[10] 松原徳秀, 山本章博. 類似度指標のラベル付き有向グラフへの拡張. 第 121 回 人工知能基本問題研究会, pp. 24–29, 2022.

[11] 大武美保子. 認知症予防回復支援サービスの開発と忘却の科学—共想法により社会的交流の場を生成する会話支援サービス—. 人工知能学会論文誌, pp. 568–575, 2009.

[12] Seiki Tokunaga, Kazuhiro Tamura, and Mihoko Otake-Matsuura. A dialogue-based system with photo and storytelling for older adults: Toward daily cognitive training. **Frontiers in Robotics and AI**, Vol. 8, p. 644964, 2021.