認知システムにおけるストーリーのマルチモーダルな関連付けに向けた試み
An Attempt toward Multimodal Story Association in a Cognitive System

秋元 泰介†，内海 彰‡
Taisuke Akimoto, Akira Utsumi
†九州工業大学，‡電気通信大学
Kyushu Institute of Technology, The University of Electro-Communications
akimoto@ai.kyutech.ac.jp

Abstract
The basic objective of this study is to obtain a computational model of the memory that produces complex associations between stories based on various types of similarity and relatedness. It is aimed at providing a basis for generative story cognition in a cognitive system. The present work particularly focuses on the memory structure underlying multimodal similarity between stories. In dealing with this issue, we introduce multimodal distributional semantics. In the proposed memory structure, stories are associated via the connections with conceptual and visual memory items. In this paper, we present an initial implementation of this idea and discuss future issues toward the next step.

Keywords — Story, Memory, Similarity, Cognitive System, Multimodal Distributional Semantics

1. はじめに

統合的な知能の仕組みを計算論的に探求する認知システムや認知アーキテクチャの研究[1–3]では，心・知能の様々な現象や機能を統合的に説明できるような原理や枠組みを明らかにすることが主要な問題となる．この問題に対して本研究は，物語（narrative）ないしストーリーという情報の形に着目する．人工知能や認知科学において，物語は古くから重要な問題として扱われており（例えばSchankやWinstonによる研究が挙げられる[4, 5]），本研究もその延長上にある．

物語の基本的な性質として，複数の事象を筋立てること，あるいは事象を中心として世界を時間的・言語的に構造化することが挙げられる．この性質は，個体間のコミュニケーションを媒介する本来の意味での物語に限らず，個体内に生じる心的表象としての世界（過去，未来，現在，虚構）の構造にも当てはめることができるだろう[6]．この個体内に生じる物語的な表象のことを特に「ストーリー」と呼ぶ．以上のように考えて，ストーリーの生成的認知の原理を探求することが本研究の基本的な目的である．

そのための主要な問題の一つとして，本研究が特に注目しているのは，既存のストーリー（記憶）をもとに新しいストーリーを生み出す仕組みである．Akimoto [7]はこの問題をさらに，ストーリーどうしを様々な種類の類似に基づいて関連付けることと，既存のストーリーを混ぜ合わせて新しいストーリーを生み出すことの二つに分けて整理している．本稿で扱うのは前者の，ストーリーどうしを関連付ける仕組みである．

これまでに，その全体的な枠組みと，その中の概念的な類似をもとにストーリーを関連付ける仕組みの簡単な実装を提案してきた[8]．本稿では，新たにマルチモーダル分散意味論[9, 10]を取り入れて，概念だけではなく，視覚的な類似も交えてストーリーを関連付けるための初歩的な試みについて報告する．

これ以降の構成は次のようになっている．まず，2節でストーリーの関連付けの全体的な構想を述べる．この構想の部分的な実装として，概念的及び視覚的な類似をもとにストーリーを関連付ける仕組みを3節で示し，4節でその動作例を分析する．最後に5節で今後の展望を述べる．

2. ストーリーの関連付けの基本構想

本節では，認知システムにおけるストーリーの関連付けの位置付けや仕組みに関する基本的な考え方を述べる．

2.1 ストーリーの関連付け

本研究では，認知システムの内部で，あるストーリーが活性化した時に，それに類似する複数のストーリーが自ずと（自動的・無意識的に）活性化することを，ストーリーの関連付けと呼ぶ．

ストーリーの関連付けは，特に以下に挙げるような認知プロセスの共通基盤として位置付けられる．

・ 想起：あるストーリーまたはその部分に注目して

いるときに，それに類似する他のストーリーが活性化することは，想起の予備段階として位置付けられる．

- 生成：新しいストーリーを作る際に，既存のストーリーを，ストーリーを作るための知識ないし素材として取り込む．
- 一般化：類似するストーリーどうしを関連付けること，あるいはストーリー間に類似を見出すことによって，複数のストーリーに通じる一般的な構造としてのスキーマを作り出す．
- 事物の主観的意味：ストーリー中に現れる物・者・場所（例えば「私」，家族，愛車，故郷）等に，様々なストーリーを結び付けることによって，個体にとっての主観的な意味を形成する．

これらは，個体が時間的な広がりのある主観的な世界を構築することや，それ以前の経験をもとに新しいストーリーを生成するための主要な認知プロセスとなる．

2.2 関連付けの仕組み

ストーリーの関連付けは，意味，構造，感覚運動的特徴等，様々な種類の類似の複合的な働きによって生じると考えられる．例えば，人間の記憶の振る舞いを想像（内省）してみると，前方から犬が歩いてくるのを見て過去に犬に襲われたことを思い出したり，「ウエスト・サイド・ストーリー」を観ながら「ロミオとジュリエット」を思い出したり，街中で鉄の匂いを感じて昔鋳物工場の立ち並ぶ道を通学していた日々を思い出したりする．ストーリーの関連付けは，想起のプロセスにおいては，予備的な段階として位置付けられるが，上に例示したような多様な現象を再現できるようなモデルを目指すべきであろう．

2.2.1 背景

この問題に取り組むに当たって，想起の計算モデルに関するこれまでの研究が一つの足がかりになる．まず，想起の計算モデルに関する研究の主要な系統の一つとして，類似に基づく想起の計算モデルが提案されてきた．例えば，Thagard ら[11]は，意味的類似，構造的類似，目的・用途的類似の 3 種類の並列的な制約充足に基づく想起モデル ARCS を提案している．一方，Forbus ら[12]は，表層的な類似に基づく選択と，構造的な類似に基づく選択の二段階処理による想起モデル MAC/FAC を提案している．これらは，処理の方式は異なるが，想起を複数種類の類似の複合的な働きとしてモデル化している点や，アナロジーに基づく構造的な類似を主要な要因の一つに位置付けているという点では似ている．

上に挙げた類似に基づく想起の計算モデル[11, 12]は，何れも，記憶要素間（検索キーに相当する probe と長期記憶内の要素）の比較に基づく類似度の計算を基本としている．しかし，あるストーリーと他のストーリーとを比較するためには，予め比較対象となるストーリーが参照可能な状態になっている必要があり，ストーリーの関連付けを比較に基づく処理としてモデル化することは適切ではないように思われる．ストーリーの関連付けはむしろ，参照可能なストーリーを限定的に活性化する処理として位置付けられるだろう．

こう考えると，ストーリーどうしが自動的に関連付けられるような構造が記憶の内部に備わっている必要がある．これに関連する研究として，Schank [13]は，スクリプト理論[14]を発展させた，MOPs（memory organization packets）というスキーマ的な知識構造（目標指向的な場面系列）に基づいて，複数のエピソードを動的に組織化するモデルを提案している．また，この考え方をもとに，Kolodner [15]は，E-MOPs（episodic MOPs）という知識構造に基づく記憶システムを実装している．しかし，MOPs のような一元的な構造では，記憶の複雑な振る舞いは説明できないだろう．

2.2.2 構想

そこで本研究では，ストーリーどうしが様々な種類の類似に基づいて自ずと関連付けられるような記憶システムの構築に取り組む．その基本となる考え方は以下の通りである．

- 一つ一つのストーリーは，ある時ある場所における具体的な出来事を表す，一回的な情報である[1]．
- しかし，ストーリーを構成する要素の大部分は，一般的な水準において，複数のストーリーの間で共有される．
- このストーリー間で共有される一般的な要素を介して，ストーリーどうしが自ずと関連付けられる．

ストーリー間で共有される一般的な要素として想定されるのは，現在のところ，主に以下の 4 種類である．

- 一般的概念：単語の意味に相当する要素（例えば「犬」「社長」）．

[1] 但し，類似した複数のストーリーが一つに圧縮される場合もあるだろう（例えば「私は子供の頃に毎朝相撲の練習していた」）．この種の時間的な圧縮はストーリーの部分的な一般化や忘却として説明できるだろう．また，ストーリーグラマー[16]やスクリプト[14]のような，物語的な構造を持つが具体的・一回的な出来事を表さない表象的要素のことは「スキーマ」と呼び，一般的な水準の方に位置付ける．

- 感覚運動パターン : 事物の感覚運動（視覚，聴覚，味覚，嗅覚，触覚，運動）的な類似のもとになる要素.
- 個別的概念 : 特定の存在物に対応する，固有名詞的な要素（例えば「タマ」「太郎」「私」）．また，その特別な下位区分として，ストーリーが起きる時間及び場所に対応する概念を，それぞれ「時間概念」「場所概念」と呼ぶ.
- スキーマ : 複数のストーリー（事物）に通じる一般的かつ複合的な構造.

さらに，これらの一般的要素とストーリーを，情報の具体性及び複合性に着目して，図 1 に示す 3 つの層に分ける．外側の層(1)は，具体的かつ一回的な情報としてのストーリーからなる．内側の層(3)は，比較的断片的かつ抽象的な要素としての一般的概念と感覚運動パターンからなる．中間の層(2)には，一般的概念や感覚運動パターンが複合的に構造化された要素としてのスキーマと個別的概念が含まれる.

このような形で，個々の記憶要素が，同じ層内の他の要素や，別の層の要素と結び付くことによって，様々な記憶が全体として組織化され，このネットワーク状の記憶組織における活性伝播を通してストーリーどうしが関連付けられる．以上が本研究の基本的な考え方である.

2.3 マルチモーダルな関連付けに向けて

この枠組みの中で，本稿で焦点を合わせるのは，ストーリー間の感覚運動的な類似を生み出す，感覚運動パターンである．人間の経験に関する記憶や，人間が想像する未来や虚構の世界は，身体性に根ざした，マルチモーダルな情報であると考えられる．ストーリーをマルチモーダルな情報として扱うことは，想起や生成をはじめとする各種認知プロセスの質的な豊かさや柔軟性に通じる，重要な問題といえる.

従来の認知アーキテクチャにおけるエピソード記憶[17–19]や，前述した想起の計算モデル[11–13, 15]の研究の多くは，記号的な情報処理の枠組みの中で行われてきた．一方，近年はニューラルネットワークに基づくエピソード記憶の研究も行われている．例えば，Rothfuss ら[20]は convolutional long short-term memory に基づく記憶モデルを提案している．しかしこちらは反対に，視覚的な側面に比重が置かれており，概念的な構造はほとんど扱われていない.

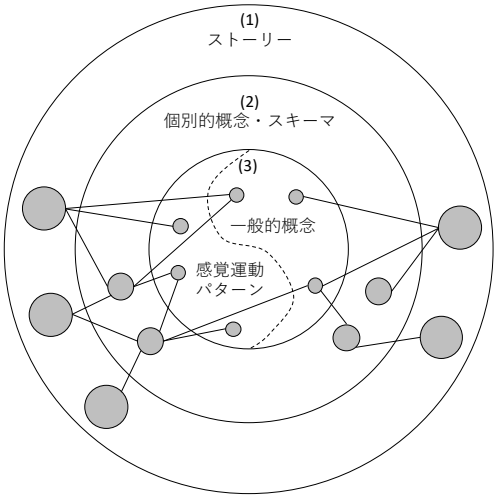

図 1　記憶組織の全体像.

ストーリーのマルチモーダルな関連付けのモデル化に向けた一つの試行として，本研究では，計算言語学において近年研究されている，マルチモーダル分散意味論[9, 10]の導入を試みる．従来の分散意味論の研究では，計算機上での単語の意味を，テキストコーパス上での単語の使われ方（単語が現れる表層的な文脈）に基づく，多次元ベクトルによって表現するモデルが提案されてきた[21, 22]．こうしたモデルは，任意の単語間の関連性を，ベクトル間の角度や距離に基づいて柔軟かつ容易に計算できるため，自然言語処理の様々なタスクにも応用されている．しかし，身体性や記号接地問題の観点から，テキストデータのみから獲得される単語の意味表現が，実世界の対象に根ざしていないという問題も指摘されてきた[23]．この問題に対して，近年，コーパス上での使われ方だけではなく，諸種の感覚運動的情報も組み合わせて単語の意味表現を構築しようとする，マルチモーダル分散意味論の研究が行われている．これまでに提案されている主な方法は，コーパスから獲得される単語のベクトルと，単語に対応する画像の特徴ベクトルとを組み合わせるというものであり（例えば[9]），本研究でもその方法を取り入れる.

3. 部分的な実装

これまでの研究では，ストーリー，個別的概念，一般的概念の 3 種類の要素を結合したシステムの簡易的な実装を試みた[8]．今回の実装では，新たに視覚パターンの導入を試みる．なお，この実装には，記憶を動的に形成・組織化する仕組みやスキーマは含まれていない.

3.1 記憶の構造

記憶は，図2に示すネットワーク構造により表現される．記憶の構成要素（ノード）は次の4種類である．

- ストーリーの集合: $\mathrm{S} = \{\mathrm{s}_i\}$
- 個別的概念の集合: $\mathrm{D} = \{\mathrm{d}_j\}$
- 一般的概念の集合: $\mathrm{G} = \{\mathrm{g}_k\}$
- S，D，Gの各要素に付随する視覚パターン:

$$\mathrm{V}^{(\mathrm{S})} = \left\{\mathrm{v}_i^{(\mathrm{S})}\right\}, \mathrm{V}^{(\mathrm{D})} = \left\{\mathrm{v}_j^{(\mathrm{D})}\right\}, \mathrm{V}^{(\mathrm{G})} = \left\{\mathrm{v}_k^{(\mathrm{G})}\right\},$$

$$\mathrm{V} = \mathrm{V}^{(\mathrm{G})} \cup \mathrm{V}^{(\mathrm{D})} \cup \mathrm{V}^{(\mathrm{S})}$$

ストーリーは本来何らかの構造的な要素として表現されるべきであるが，この実装では単純化のために一つのノードとしている．

これらの要素はそれぞれ以下の情報を持つ．

- ストーリーs: 固有番号，文，画像（任意）．
- 個別的概念 d: 名前．
- 一般的概念 g: 単語，単語ベクトル．
- 視覚パターン v: 付随対象の記号，視覚ベクトル．

なお，ストーリーや概念に対する視覚パターンの有無は任意であり，視覚パターンと結合されないストーリーや概念があってもよいこととする．

要素間の関係（結合の強さ）は基本的に重み付きエッジにより表現される．S–D，S–G，D–G，及びD–D（個別的概念どうし）の各領域間のエッジの集合を，それぞれ$W^{(\mathrm{SD})} = \left\{w_{ij}^{(\mathrm{SD})}\right\}, W^{(\mathrm{SG})} = \left\{w_{ik}^{(\mathrm{SG})}\right\}, W^{(\mathrm{DG})} = \left\{w_{jk}^{(\mathrm{DG})}\right\}, W^{(\mathrm{DD})} = \{w_{jl}^{(\mathrm{DD})}\}$とし，それぞれ行列形式で要素間の重みを定義する．例えば，$w_{23}^{(\mathrm{SG})} = 0.5$は，ストーリー$\mathrm{s}_2$と一般的概念$\mathrm{g}_3$の間の重みが0.5であることを意味する．結合が無い部分の値は0である．これらのエッジの重みは，本来は何らかの仕組みによって形成されるべきであるが，現在は暫定的に手作業で任意の値を設定することにしている．

視覚パターンとその付随対象（ストーリー，個別的概念，一般的概念の何れか）は，一対一的に結合される．このエッジ集合をそれぞれ$R^{(\mathrm{G})}, R^{(\mathrm{D})}, R^{(\mathrm{S})}$とする．これらの重みは心像性（イメージの浮かびやすさ）に相当するものとする．

なお，一般的概念どうし及び視覚パターンどうしの結び付きの強さは，それぞれのベクトル空間上での角度（コサイン類似度）に基づいて計算される．

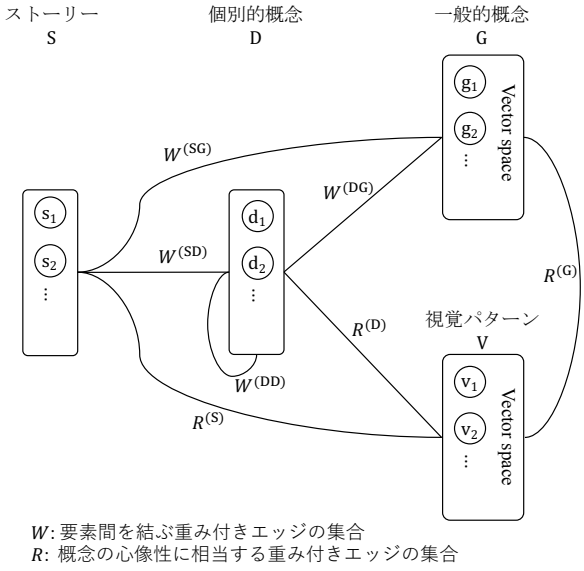

図2　記憶のネットワーク構造．

3.1.1 一般的概念のベクトル空間（単語ベクトル）

一般的概念は，単語を基本単位として，分散意味論に基づく多次元ベクトルにより表現する．このベクトル間のコサイン類似度を，一般的概念間の類似度（重み）と見なす．

今回の実装では，単語のベクトル空間の構築にはword2vec [22]のSkip-gramモデル（300次元，ウィンドウサイズ10，negative sampling）を用いた．学習には，『現代日本語書き言葉均衡コーパス』[24]を使用し，結果として得られた175,801（出現頻度10以上の単語に限定）の単語ベクトルの中から，一般的概念に対応するもののみを抽出して使用する．

3.1.2 視覚パターンのベクトル空間（視覚ベクトル）

視覚パターンは，ストーリー，個別的概念，一般的概念の何れかに対応する画像の特徴表現に相当するベクトルにより表現する．視覚パターン間の類似度（重み）もまたコサイン類似度により計算される．

マルチモーダル分散意味論の研究では，画像の特徴表現として，SHIFTやSURF等の局所特徴列を用いる方法[9, 25]や，学習済みconvolutional neural network（CNN）において，ある画像を入力した時の最終隠れ層の活性状態を特徴表現として抽出する方法[26, 27]が用いられているが，本研究では後者を用いた．実際に使用したモデルはResNet152-hybrid1365 [28][2]であり，最終隠れ層からは2048次元のベクトルが得られる．

[2] https://github.com/CSAILVision/places365

ストーリーに対応する視覚ベクトルには，ストーリーに含まれる単一の画像から得られるベクトルをそのまま用い，一般的概念または個別的概念に対応する視覚ベクトルは，当該概念に対応する複数枚の画像（例えば「犬」の画像セット）から得られたベクトルの重心とする．

前述したように，視覚パターンとその付随対象の間の重みは心像性に相当するものであり，その値は，ある概念に対応する n 枚の画像セットにおける全通りペアのコサイン類似度の平均値とする（Kiela ら[29]の image dispersion の計算式を応用して，「犬」のように概ね似たようなイメージに結び付く概念は心像性が高く，「幸せ」のように多様なイメージに結び付きそうな概念は心像性が低いと仮定）．ストーリーとそれに対応する（単一画像から得られる）視覚ベクトルの間の重みは常に 1 とする．

3.2 活性伝播の局所的な仕組み

局所的な計算の模式図を図 3 に示す．あるノードの活性度が，複数のノードから受け取る活性度に基づいて計算される．この計算式を以下のように定義する．

$$a_{\mathrm{y}} = \sum_{i}^{n} \mathrm{out}(\mathrm{x}_i) w_i \tag{1}$$

ここでout(x)は，ノード x の出力活性度を計算する関数であり，

$$\mathrm{out}(\mathrm{x}) = \begin{cases} 0, & a_{\mathrm{x}}\beta < \theta_{out} \\ a_{\mathrm{x}}\beta, & a_{\mathrm{x}}\beta \geq \theta_{out} \end{cases} \tag{2}$$

と定義する．θ_{out}は小さな活性度の出力を遮断する閾値である．βは出力活性度を抑制または増幅する係数であり，1 を基本値として，ストーリー（β_{S}），個別的概念（β_{D}），一般的概念（β_{G}），視覚パターン（β_{V}）の各領域について任意の値を設定することができる．

3.3 活性伝播の流れ

活性伝播の流れは，あるストーリーが活性化し，そこから内側の層へと伝播していき，再び外側の層へと向かうというように仮定する．初期状態として，一つのストーリーの活性度を 1 とし，それ以外の全ノードの活性度を 0 とする．その後，以下に示す 7 つのステップを経て，他のストーリーが活性化する．各ステップ冒頭の角括弧に伝播経路を略記する．

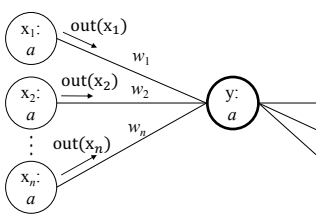

図 3　活性伝播の局所的な仕組み．

Step 1 [S → D]: ストーリーの出力活性度に基づいて，各個別的概念の活性度を計算する．

$$a_{\mathrm{d}_j} = \sum_{i}^{|\mathrm{S}|} \mathrm{out}(\mathrm{s}_i) w_{ij}^{(\mathrm{SD})} \tag{3}$$

Step 2 [S + D → D, V]: ストーリーと個別的概念の出力活性度に基づいて，一般的概念と視覚パターンの活性度をそれぞれ以下のように計算する．

$$a_{\mathrm{g}_k} = \sum_{i}^{|\mathrm{S}|} \mathrm{out}(\mathrm{s}_i) w_{ik}^{(\mathrm{SG})} + \sum_{j}^{|\mathrm{D}|} \mathrm{out}(\mathrm{d}_j) w_{jk}^{(\mathrm{DG})} \tag{4}$$

$$a_{\mathrm{v}_i^{(\mathrm{S})}} = \mathrm{out}(\mathrm{s}_i) r_i^{(\mathrm{S})} \tag{5}$$

$$a_{\mathrm{v}_j^{(\mathrm{D})}} = \mathrm{out}(\mathrm{d}_j) r_j^{(\mathrm{D})} \tag{6}$$

Step 3 [G ↔ V(G)]: G と V は連動的に活性化することとし，その関係を

$$a_{\mathrm{g}_k} = a_{\mathrm{g}_k} + \mathrm{delta}\left(\mathrm{v}_k^{(\mathrm{G})}\right) r_k^{(\mathrm{G})} \tag{7}$$

$$a_{\mathrm{v}_k^{(\mathrm{G})}} = a_{\mathrm{v}_k^{(\mathrm{G})}} + \mathrm{delta}(\mathrm{g}_k) r_k^{(\mathrm{G})} \tag{8}$$

と定義し, Step 3 と後述する Step 5 において実行する．ここで delta(x)は，当該ノードの活性度が最後に変化した際の変化分を返す関数である．Step 2 が G の最初の変化であるため，$\mathrm{delta}(\mathrm{g}_k) = a_{\mathrm{g}_k}$となる．一方，Step 2 終了時点では V(G)の活性度は全て初期値（0）のままであり，$\mathrm{delta}\left(\mathrm{v}_k^{(\mathrm{G})}\right) = 0$となるため，Step 3 においては G 側の活性度には影響を与えない．

Step 4 [G → G | V → V]: 一般的概念と視覚パターンそれぞれの内部において活性度が伝播する．

$$a_{\mathrm{g}_k} = a_{\mathrm{g}_k} + \sum_{i \neq k} \mathrm{out}(\mathrm{g}_i) \mathrm{sim}(\mathrm{g}_k, \mathrm{g}_i) \tag{9}$$

$$a_{\mathrm{v}_l} = a_{\mathrm{v}_l} + \sum_{i \neq l} \mathrm{out}(\mathrm{v}_i) \mathrm{sim}(\mathrm{v}_l, \mathrm{v}_i) \tag{10}$$

ここで，sim はノード間の類似度を返す関数であり，

$$\mathrm{sim}(\mathrm{x}, \mathrm{y}) = \begin{cases} 0, & \cos(\mathrm{x}, \mathrm{y}) < \theta_{sim} \\ \cos(\mathrm{x}, \mathrm{y}), & \cos(\mathrm{x}, \mathrm{y}) \geq \theta_{sim} \end{cases} \tag{11}$$

と定義する．ベクトル間のコサイン類似度（cos）が閾

値θ_{sim}未満のノード間では伝播が生じないようになっている．G と V の領域ごとに閾値を設ける場合，それぞれ$\theta_{simG}, \theta_{simV}$と表記する．

Step 5 [G ↔ V(G)]: Step 4 によって，G と V の各ノードの活性度が更新されるため，その変化分を式 7 と式 8 により双方に反映させる．

Step 6 [G + D + V(D) → D]: 個別的概念の活性度を，一般的概念，視覚パターン，及び隣接する個別的概念の出力活性度に基づいて計算する．

$$a_{\mathrm{d}_j} = \sum_{k}^{|\mathrm{G}|} \mathrm{out}(\mathrm{g}_k) w_{jk}^{(\mathrm{DG})} + \sum_{l}^{|\mathrm{D}|} \mathrm{out}(\mathrm{d}_l) w_{lj}^{(\mathrm{DD})} + \mathrm{out}(\mathrm{v}_j^{(\mathrm{D})}) r_j^{(\mathrm{D})} \quad (12)$$

Step 7 [G + D + V(S) → S]: 最後に，各ストーリーの活性度を，内側から来る出力活性度をもとに計算する．

$$a_{\mathrm{s}_i} = \sum_{k}^{|\mathrm{G}|} \mathrm{out}(\mathrm{g}_k) w_{ik}^{(\mathrm{SG})} + \sum_{j}^{|\mathrm{D}|} \mathrm{out}(\mathrm{d}_j) w_{ij}^{(\mathrm{SD})} + \mathrm{out}(\mathrm{v}_i^{(\mathrm{S})}) r_i^{(\mathrm{S})} \quad (13)$$

4. 動作例と分析

上記のプログラムの動作例を示し，主に以下の二つの観点から動作を分析する．

(1) 概念的にはあまり似ていないが，視覚的には似ているストーリーどうしを関連付けられる．
(2) 各要素の関与の度合いを出力係数βによって調整することで関連付けの振る舞いが変化する．

4.1 記憶の内容

前述したように，現在の実装には記憶を自律的に形成するアルゴリズムがないため，記憶を構成する要素や要素間の結合の大部分は手作業で用意する．

まず，表 1 に示す 6 つのストーリーを用意した．なお，#5 以外のストーリーにはそれぞれ 1 枚の画像が付与されている（ストーリー#5 は，遠い昔のことなので具体的なイメージが消失していると想定する）．これらのストーリーは，A)牧場にいる牛の話，B)肉を食べる話，C)ゴルフの話という 3 つのグループに分けられる．（なお，ここで「グループ」というのは説明のための概念であり，プログラム内にはこれに対応する構造はない．）これらのグループ間には図 4 に示すような関係がある．まず，A–B 間は概念的には似ているが視覚的には似ていない（「食べる」という動詞が同じであり，「牛」と「焼き肉」や「ステーキ」も概念的に近い）．反対に，A–C 間は概念的には似ていないが視覚的には似ている（牧場に牛がいる光景とゴルフ場で人がゴルフをしている光景が似ている）．そして B–C 間は，概念的にも視覚的にもあまり似ていない．

さらに，上記の 6 つのストーリーに関連する個別的概念，一般的概念，視覚パターンを用意した．表 2 にそれらをまとめる．ストーリー及びこれらの要素の結合構造の詳しい説明は省略するが，要素間の結合がある場合の重みは一律で 0.5 としている．

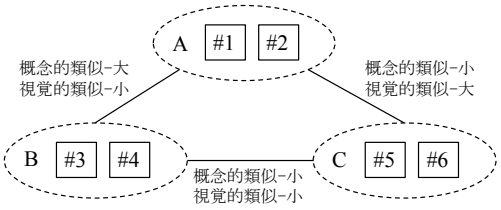

図 4 ストーリー間の関係.

表 1 記憶要素として用意した 6 つのストーリー.

		内容（文）	画像
A	#1	先月、私は K 牧場に行った。牛たちは牧草を食べたり、「モーモー」と鳴いたりしていた。	[緑の牧場風景. 遠くに牛たち.]
	#2	先週、私は M 牧場に行った。牛たちは牧草を食べたり、寝そべったりしていた。	[黄味がかった牧場風景. やや遠くに牛たち.]
B	#3	一昨日、私は、友人の太郎と焼き肉屋 U に行って、焼き肉をたらふく食べた。	[網の上に並んだ肉]
	#4	昨日、私は、ステーキ屋 Y に行って、大きなステーキを食べたが、不味かった。	[皿の上のステーキ]
C	#5	2001 年の夏、どこかのゴルフ場で、私ははじめてゴルフをして、ナイスショットを連発した。	無し
	#6	今朝、テレビを付けたら、ゴルフの大会を中継していた。それをぼんやりと観ていたら、S 選手がミスショットを連発して苛立っていた。	[ゴルフ場（グリーン）でパターを打つ人]

表 2 ストーリー以外の記憶要素（ノード）.

種類	ノード	数
一般的概念	牧場, 牛, 牧草, 食べる, 寝そべる, 友人, 焼き肉, ステーキ, 不味い, 夏, ゴルフ, テレビ, 失敗, 苛立つ, レストラン, アスリート, 人間	18
個別的概念	私, K 牧場, M 牧場, 太郎, 焼き肉屋 U, ステーキ屋 Y, S 選手	7
視覚パターン	(VG) 牧場, 牛, 牧草, ,焼き肉, ステーキ, ゴルフ, テレビ, レストラン, アスリート, (VS) #1, #2, #3, #4, #6	14

各視覚パターンに対応する特徴ベクトルを構築する際に用いた画像は次の方法で用意した．一般的概念に対応する画像は，Flickr[3]のキーワード検索（API による自動抽出）により，一般的概念（を英単語に置き換えたもの）をキーとして上位 10 枚の画像を取得した．ストーリーに対応する画像は，一般的概念に対応する画像の中から手作業で適当な 1 枚を選んだ．なお，個別的概念には視覚パターンが付与されていない（視覚パターンの働きに焦点を合わせるために個別的概念の関与が小さくなるようにしている）．

4.2 結果と分析

上記のネットワーク構造を用いたプログラムの動作例を示す．概念系のノードと視覚系のノードそれぞれの関与の度合いの増減によるプログラムの振る舞いの違いを観察するために，出力係数（式 2 におけるβ）の調整により，基本（Base），概念系重視（CE），視覚系重視（VE），概念系のみ（CO），視覚系のみ（VO）という 5 種類のモードを用意した（表 3）．ここで概念系というのは個別的概念（D）と一般的概念（G），視覚系は視覚パターン（V）である．

その他のパラメータは，何れのモードにおいても，活性度の出力閾値$\theta_{out} = 0.2$，一般的概念間の伝播閾値$\theta_{simG} = 0.2$，視覚パターン間の伝播閾値$\theta_{simV} = 0.5$とした．θ_{simV}がθ_{simG}よりも大きいのは，視覚パターン間のコサイン類似度の値が，一般的概念間のそれと比べて全般的に大きな値になる傾向が顕著であったためである．

上記の条件で，各モードについて，各ストーリーを焦点として実行した結果を図 5 にまとめる．グラフには，Step 7 後の各ストーリーの活性度を，最大値が 1 になるように正規化した値が示されている．例えば，最上段の Focus #1 は，各モードについて，初期状態でストーリー#1 の活性度を 1 として実行した後の，#1–#6 までの各ストーリーの活性度を表している．

表 3　5 種類のモード．

	Base	CE	VE	CO	VO
β_S	1.0	1.0	1.0	1.0	1.0
β_D	1.0	1.5	0.5	1.0	0.0
β_G	1.0	1.5	0.5	1.0	0.0
β_V	1.0	0.5	1.5	0.0	1.0

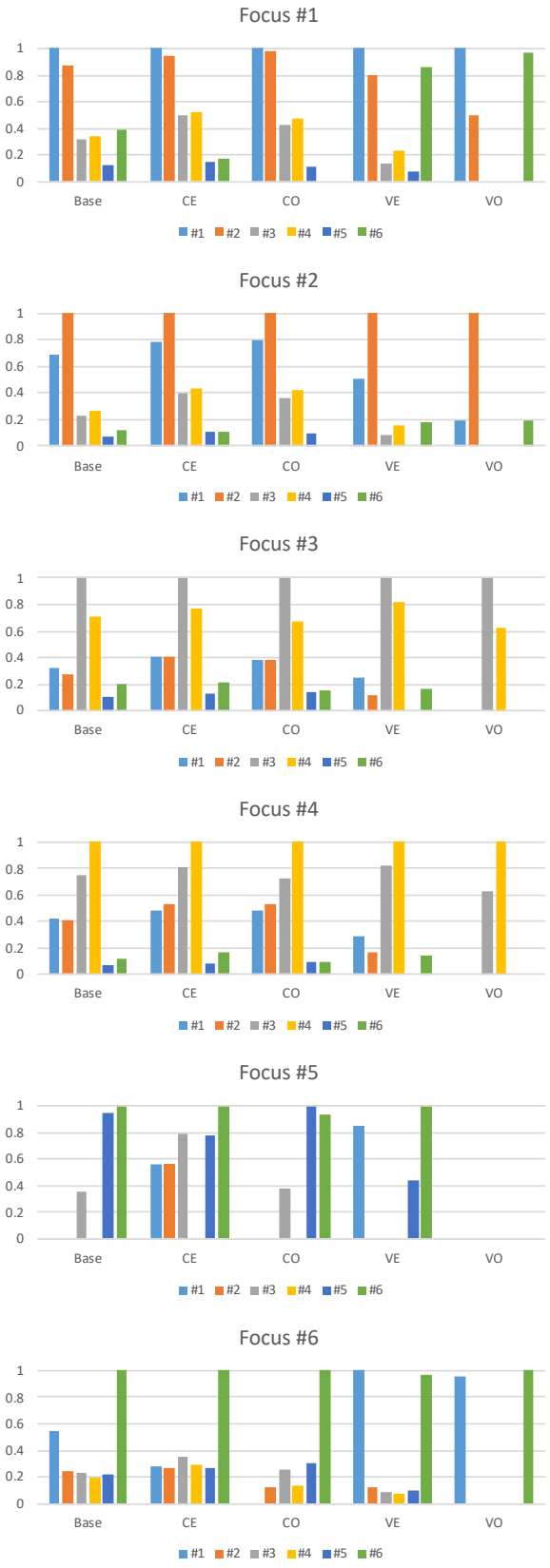

図 5　各ストーリーを焦点とした実行結果．モードごとに，最終ステップ後の各ストーリーの活性度（正規化後の値）を表す．

[3] https://www.flickr.com

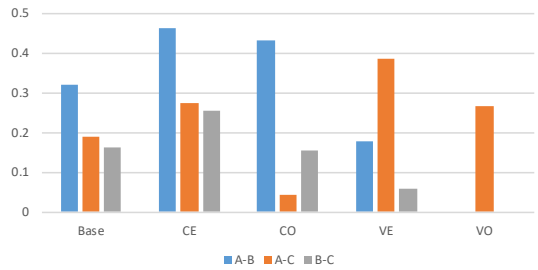

図 6　各モードにおけるグループ間の関連の強さ．モードごとに，グループ間のストーリー対全通りの活性度（正規化後）の平均を表す．

それから図 6 は，3 つのストーリーグループそれぞれの間の関連の強さを見るために，図 5 のデータにおける A–B 間，A–C 間，B–C 間それぞれに対応する部分の値を平均したものである．各グループに 2 つのストーリーがあり，実行後の活性度は非対称（例えば Focus #1 における#6 と，Focus #6 における#1 の値は少し異なる）であるため，2 つのグループの間には 8 通りの値がある．

以上のデータをもとにプログラムの動きを細かく見ていく．まず図 6 を見ると，概ね想定（図 4）通りに，概念系に比重をおいた CE と CO では A–B 間の関連が強まり，視覚系に比重をおいた VE と VO では A–C 間の関連が強まっている．基本となる Base は，これらの中間的な振る舞いと捉えてよいだろう．B–C 間の値は，何れのモードにおいても比較的小さいが，概念系のノードによる関連は生じているようである．

次に図 5 を見ると，A–C 間に該当する部分の値は，ストーリー対ごとに大きく異なっている．上では VE や VO において A–C 間が強く関連付けられていると述べたが，実際のところ，VE や VO による関連が生じているのは#1–#6 間だけであり，#2–#6，#1–#5，#2–#5 の間には関連がほとんど生じていない．

#2–#6 間の関連が弱いのは，単純に，#2–#6 間において，ストーリーに付随する視覚ベクトルのコサイン類似度が$\theta_{simV} = 0.5$を下回っていたからである．#1 の画像と#2 の画像は，何れも牧場風景に相当するが，色味の違い（#1 は緑の牧草，#2 はやや黄色味がかった牧草）が，類似度に大きく影響しているようである．

一方，#1–#5 間及び#2–#5 間の関連が弱いのは，#5 に画像（の視覚パターン）が付与されていないことが主な理由である．今回提案したネットワーク構造（図 2）では，ストーリーそのものが視覚パターンを持たない場合にも，一般的概念と視覚パターンの結び付き（$R^{(G)}$）を通して，ストーリー間に視覚的な類似を生み出すこともできそうである．しかし，今回定義した活性伝播の仕組みでは，ある一般的概念から別の一般的概念に，視覚パターンを介して活性度が伝播する過程（$g_x \rightarrow v_x^{(G)} \rightarrow v_y^{(G)} \rightarrow g_y$）で，ノード間を伝わる活性度が徐々に弱まっていく．そのため，ストーリー間の視覚的な特徴による関連の強さは，結局の所，ストーリーに直接結び付いた視覚パターン（式 13 における$v_i^{(S)}$）が，Step 7 の時点でどの程度活性化しているかに強く依存している．

5.　展望

本稿では，ストーリーの関連付けに視覚的特徴を介在させる最初の試みについて述べた．本研究はまだ全体的に初期段階にあり，手付かずのままになっている問題も多数ある．例えば，記憶の動的な側面はまだ扱っていないし，記憶の構造についても検討すべき問題が多数ある．また，記憶のシステム全体としての認知的な妥当性を如何に検証するかも難しい問題である．その中の幾つかの問題に関する今後の展望を述べる．

5.1 マルチモーダルな関連付けについて

今回の実装において，ストーリーの関連付けに視覚的特徴を介在させることはできたが，その仕組みに関してはまだ検討や拡張の余地がある．例えば，ストーリーに対して単一の視覚パターンを付与するだけではなく，一つのストーリーに様々なイメージが結び付いてもよいだろう．また，現在のモデルでは，ストーリーと視覚パターンとが固定的に結合されているが，ストーリーは視覚的なイメージも含めて動的に生成される情報であると考えることもできる．なお，視覚的な特徴表現自体にもまだ課題はあるだろうが，そこを改良することは本研究の本筋からは少し外れる．

一方，ストーリーの認知においては，視覚よりもむしろ，時間軸のある運動的な特徴が重要であるとも考えられる．この問題に関連する理論として，事象や概念の運動的な特徴を抽象的な水準（例えば物理的・心理的な移動，何かを内部に取り込む，容器を満たす）で表現する Schank の概念依存理論[30]や，認知言語学におけるイメージ・スキーマ[31, 32]が挙げられる．このような要素を，マルチモーダル分散意味論の枠組みを基本とした上で，どのような形でシステムに取り入れるかを考えていく必要がある．人工知能の研究では，これらに類する知識が記号的に表現されている場合が

多いが，このような身体的な特徴を記号的に表現することにはやや疑問があるし，それを如何に学習するかという問題もある．

5.2 ストーリーの内部構造の導入に向けて

ストーリーが時間軸を持つ複合的な情報であるとすると，ストーリーは複数の部分（事象・場面や実体）からなる構造を内部に持つ必要があるだろう[4]．ストーリーが内部構造を持っていれば，関連付けを，部分構造の水準で関連ないし類似を生み出すことへと拡張することができる．例えば，ストーリー中のある事物や場面に注目した時に，それに類似する他の事象や場面（他のストーリーの部分）が活性化する，というように．この方が，ストーリーという時間軸のある情報の動きとしては妥当であるように思われる．

さらに，ストーリーどうしを内部構造の水準で関連付ける仕組みは，ストーリー間に構造的な対応関係（図7）ないしアナロジー的なマッピングを形成するための基盤として，ストーリーの生成や一般化においても重要な役割を担うだろう．これについては次の5.3節で詳しく説明する．

5.3 記憶に基づく創造性の基盤として

記憶（ストーリー）に基づいて新しい情報（ストーリー）を生み出す仕組みに相当する主要な理論やモデルとして，アナロジー，事例ベース推論，conceptual blending [33]の3つが挙げられる．アナロジーは，ある慣れ親しんだ領域（base）と新奇な領域（target）の間に構造的な対応関係を作り，それを通してbase側の知識をtargetへと転移するプロセスとして説明される．事例ベース推論は，ある問題の解決策を，過去の類似する事例（問題の解決）を再利用して導き出すような方法である．これらは何れも目標（targetや問題解決）指向的なプロセスと見なせる．一方，conceptual blendingは，（生成の観点から見ると）複数の概念構造（input spaces）から取捨選択された情報が統合されて，別の概念構造（blend）が生じるという枠組み（図8）になっている．その際，input spaces間に対応関係や共通構造

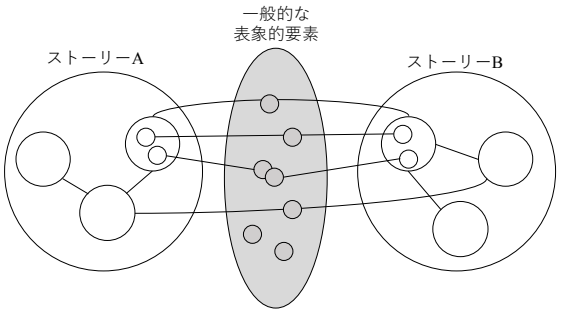

図7　ストーリーの部分構造を単位とする関連付け．

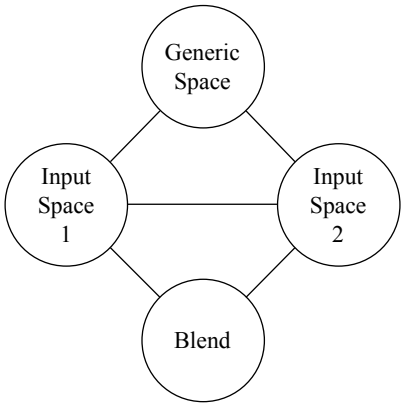

図8　Conceptual blending [33]の基本図式．

（generic space）を構成することが，blendの生成や理解の鍵になるとされる．なお，アナロジーや事例ベース推論の計算モデルに関する研究は古くから行われているが，近年はconceptual blendingに基づく創造性の計算モデルに関する研究も行われている（例えば[34]）．

上に挙げた3種類は何れも類似に基づく認知であり，ストーリー間に部分構造の水準で関連を生み出す仕組みは，類似するストーリーどうしを結び付けることと，ストーリー間に構造的な対応関係を生み出すことという二つの面で，記憶に基づく創造性の基盤になる．本研究が現在特に注目しているのはconceptual blendingであり，Akimoto [35]は，二つのストーリーを混ぜ合わせて新しいストーリーを作り出すstory blendingの計算モデルに向けた理論的な整理を行っている．

謝辞

本研究はJSPS科研費JP18K18344の支援を受けた．

[4] ストーリーがどのような形や大きさで保持されるかという問題については今後さらなる検討が必要である．ストーリーが断片化された形で保持されており，それが参照される際に，まとまりのある構造が動的に生じると考えることもできる．あるいは，ストーリーはある程度まとまった形で保持されていて，それを参照する際の文脈に応じて有用な情報が抽象されるという考え方も有り得るだろう．しかし，ストーリーに内部構造が無ければこの種の動的な側面も扱いにくい．

参考文献

[1] Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10, 141-160.

[2] Samsonovich, A. V. (2012). On a roadmap for the BICA Challenge. *Biologically Inspired Cognitive Architectures*, 1, 100-107.

[3] Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, 38(4), 13-26.

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

[5] Winston, P. H. (2012). The right way. *Advances in Cognitive Systems*, 1, 23-36.

[6] Akimoto, T. (2018). Stories as mental representations of an agent's subjective world: A structural overview. *Biologically Inspired Cognitive Architectures*, 25, 107-112.

[7] Akimoto, T. (2019). Key issues for generative narrative cognition in a cognitive system: Association and blending of stories. *Story-enabled Intelligence, AAAI 2019 Spring Symposium*.

[8] Akimoto, T. (2019). Toward complex story association in a cognitive system: A holistic framework and partial implementation. *7th Annual Conference on Advances in Cognitive Systems, Poster Collection*.

[9] Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49, 1-47.

[10] Baroni, M. (2016). Grounding distributional semantics in the visual world. *Linguistics and Language Compass*, 10(1), 3-13.

[11] Thagard, P., Holyoak, K. J., Nelson, G., & Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence*, 46(3), 259-310.

[12] Forbus, K. D., Gentner, D., & Law, K. (1994). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19, 141-205.

[13] Schank, R. C. (1982). *Dynamic memory: A theory of reminding and learning in computers and people*. Cambridge University Press.

[14] Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum.

[15] Kolodner, J. K. (1983). Maintaining organization in a dynamic long-term memory. *Cognitive Science*, 7, 243-280.

[16] Rumelhart, D. E. (1975). Notes on a schema for stories. In Bobrow, D. G. & Collins, A. (Eds.), *Representation and understanding: Studies in cognitive science*. Academic Press.

[17] Nuxoll, A. M., Laird, J. E. (2012). Enhancing intelligent agents with episodic memory. *Cognitive Systems Research*, 17-18, 34-48.

[18] Menager, D. H., & Choi, D. (2016). A robust implementation of episodic memory for a cognitive architecture. *Proc. CogSci 2016*, pp. 620-625.

[19] León, C. (2016). An architecture of narrative memory. *Biologically Inspired Cognitive Architectures*, 16, 19-33.

[20] Rothfuss, J. et al. (2018). Deep episodic memory: Encoding, recalling, and predicting episodic experiences for robot action execution. *IEEE Robotics and Automation Letters*, 3(4), 4007-4014.

[21] Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211-240.

[22] Mikolov, T. et al. (2013). Distributed representations of words and phrases and their compositionality. *Proc. 26th International Conference on Neural Information Processing Systems*, pp. 3111-3119).

[23] Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, 43(3), 379-401.

[24] Maekawa, K. et al. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2), 345-371.

[25] Takano, K., & Utsumi, A. (2016). Grounded Distributional Semantics for Abstract Words. *Proc. CogSci 2016*, pp. 2171-2176.

[26] Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. *Proc. 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 36-45.

[27] Utsumi, A. (2018). A distributional semantic model of visually indirect grounding for abstract words. *Proc. NeurIPS 2018 Workshop on Visually Grounded Interaction and Language*.

[28] Zhou, B. et al. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6), 1452-1464.

[29] Kiela, D., Hill, F., Korhonen, A., & Clark, S. (2014). Improving multi-modal representations using image dispersion: Why less is sometimes more. *Proc. 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 835-841.

[30] Schank, R. C. (1975). *Conceptual information processing*. Amsterdam: Elsevier.

[31] Lakoff, G. 著, 池上 他 訳 (1993). 認知意味論. 紀伊國屋書店. (原著 1987)

[32] Johnson, M. 著, 菅野・中村 訳 (1991). 心の中の身体—想像力へのパラダイム転換. 紀伊國屋書店. (原著 1987)

[33] Fauconnier, G., & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.

[34] Eppe, M. et al. (2018). A computational framework for conceptual blending. *Artificial Intelligence*, 256, 105-129.

[35] Akimoto, T. (2019). Theoretical framework for computational story blending: From a cognitive system perspective. *Proc. 10th International Conference on Computational Creativity*, pp. 49-56.