

トピックブリッジングによるストーリー生成支援

Topic Bridging for Story Generation

佐藤 真*¹ 赤石 美奈*¹ 堀 浩一*¹

Makoto Sato Mina Akaishi Koichi Horii

*¹ 東京大学大学院工学系研究科

School of Engineering, University of Tokyo

We propose a framework, topic bridging, for story generation support. A document is defined as a story fragment and a story framework is defined as a sequence of story fragments. Topic bridging suggests story fragments which can bridge between the start topic and the goal topic to generate a story framework. Providing the story framework supports the user's story generation process.

1. はじめに

ストーリーの形式で表現される知識は様々な長所をもつことが知られている。Brownら [Brown 05] は、人々が知識創造のスパイラルをまわし組織に参加し積極的な活動をするためには、知識を教科書にあるような専門的な文言の羅列で表すだけでなく、日々使うような言葉で人々がわかりやすいストーリーの形式をとるべきであるとしている。たとえばプログラミングを学ぶとき関数の説明 (References) のみではその関数を用いて何ができるかは伝わりにくい。どのように使うか (Tutorials, Sample Applications) や困ったときにどうすればよいか (FAQ) という目的に応じて、様々な関数に関する知識を連繫させて筋立てて説明するほうが関数についての知識も伝わるだろう。

人間はコンテキストに応じて経験的に関連する知識をつなぎ合わせてストーリーをつくることができる。つまりストーリーを構成する知識にはコンテキストによって異なる意味や価値を持たせることができる。それゆえ知識をつなぎ合わせてストーリーをつくらうとするとき、その知識がストーリーにおいてどのような意味や価値があるかという知識の文脈依存性を考慮する必要がある。

そこで、ストーリーの骨格にストーリー断片を連結することで変化するトピックのダイナミクスをモデル化、この逆問題としてスタートとゴールのトピックがあらかじめ分かっているときどんなストーリー断片を連結すればよいかということを推定する手法を提案する。

なお、本論文では、ストーリーは幾つかのドキュメントから構成されるとし、構成要素であるドキュメントをストーリー断片と呼ぶ。ストーリー断片を連鎖させてストーリーの骨格を自動的に生成することをストーリー生成支援と定義する。

2. 関連研究

自然言語によるストーリーを自動的に生成する手法については1960年代から盛んに研究がされてきた。人工知能をベースにしたストーリー生成は物語文法に基づくアプローチであるTALE-SPIN [Meehan 77] に遡ることができる。TALE-SPINでは、物語の世界を設定しその世界におけるエージェントが

ゴールを達成するために合理的な行動をとるとして、エージェントの行動に基づきストーリーを生成する。TALE-SPINの提案以降、物語文法による手法は物語理解のための理論とともに開発されてきている [Bringsjord 99]。スタートとゴールを設けて情報のつながりをストーリーとして提示する手法も提案されている [Shahaf 10]。

機械的にストーリーを処理しつつユーザのストーリー生成を支援する手法としては、赤石の文書の語彙連鎖構造に着目して物語を分解・再構成するナラティブ連想情報アクセス [赤石 06]、蓄積された情報断片の連鎖可能性を可視化する手法であるAmitaniらのGenerative website [Amitani 08] が挙げられる。大澤は新たなシナリオの創発の支援として、データの可視化とシステムと人間のコミュニケーションをどのようにデザインすればよいかという指針を示すチャンス発見の方法論 [大澤 06] を提案している。本研究では、ストーリーのトピックについて数学的なモデルを提案した上で、大量の文書情報を対象により機械的にストーリー生成を支援する手法の確立を狙う。

3. トピックブリッジング

たとえば環境問題に対して航空宇宙工学の見地からどんなアプローチがあるかということ議論するとき、効率的なエンジンを設計すればよい、再利用可能なコストパフォーマンスのよい材料を使えばよい、あるいは機体に関するだけでなく最適な飛行経路を設計すればよい、など様々な視点から検討するだろう。これは航空宇宙工学というトピックと環境問題というトピックがどのようなストーリー断片の連鎖によってつながることができるか、どんなストーリーをつくることができるかと考えることに相当する。トピックブリッジングはこのようなプロセスを支援するものである。

図1にトピックブリッジングの概念図を示す。たとえばスタートのトピックに「航空宇宙工学」を設定し、ゴールのトピックに「地球温暖化防止」を設定したら、「航空機材の軽量化」「エンジンの効率化」「航空経路の最適化」などといったストーリー断片がスタートとゴールをつなぐブリッジとなる。トピックブリッジングでは、ユーザがスタートのトピックとゴールのトピックを設定し、これらをつなげるブリッジを、ストーリーの構成要素となるストーリー断片が蓄積されたデータベースから自動的に探索し、ストーリーの骨格を生成する。

連絡先: 佐藤 真, 東京大学大学院工学系研究科, 東京都文京区本郷 7-3-1 東京大学工学部 7号館, 03-5841-1839, satomakoto[at]ailab.t.u-tokyo.ac.jp

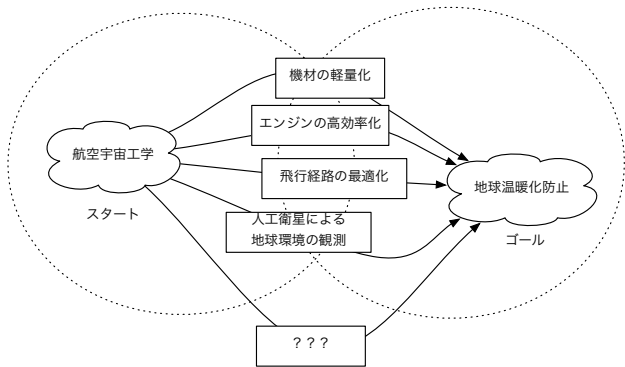


図 1: トピックブリッジングの概念図。航空宇宙工学から地球温暖化防止につながるストーリー。従来の研究を当たれば、航空機材の軽量化、エンジンの効率化、航空経路の最適化などをすれば地球温暖化を防止できるだろう、あるいは人工衛星を用いた地球観測を行えば地球温暖化の現状を把握することができるだろう。現在、他にも全く議論されていないストーリーを生成できるかもしれない。

3.1 トピック伝播モデル

ストーリー断片間の文脈依存性を考慮した関係の抽出をストーリー断片の表層的な情報のみから行うために、ストーリー断片の連鎖の仕方によってトピックが変化するとする**トピック伝播モデル**を提案する。

複数のストーリー断片を連鎖させてストーリーの骨格を生成するとき、それぞれのストーリー断片はその文脈に依存して異なる意味を持つ。たとえば原子力発電に関するストーリー断片について考えると、地球温暖化に関するストーリー断片の後に原子力発電のストーリー断片をつなげたストーリーでは原子力発電は比較的 CO2 排出量が少ないということに焦点が当てられるであろう。しかし一方で大地震に関するストーリー断片の後に原子力発電に関するストーリー断片をつなげたストーリーでは原子力発電の危険性に焦点が当てられるであろう。これを「地球温暖化」のストーリー断片に「原子力発電」のストーリー断片を連結したときに、「地球温暖化」のストーリー断片で重要であった語の重要度が「原子力発電」のストーリー断片に出現する語に伝播し重要度を増すというようになる、モデル化する。

トピック伝播モデルを数学的に定義するために活性伝播モデルを参考にする。活性伝播モデルはネットワークの挙動の解析やネットワークを構成するノードの重要度の計算などに用いられ、ドキュメント群に対する分析や応用に利用する研究も数多く行われている [Mani 99, 松村 02, 綾 05]。活性伝播モデルで扱うネットワークは活性値を持つノードとノードの間をつなげる重み付きのリンクで構成され、リンクの重みはあるノードの活性値を隣接する他のノードへ伝播する割合を示す。トピック伝播モデルでは、トピックは語の重要度を要素とするベクトルあるいは行列で表されるとして、ストーリーにストーリー断片をつなげるによりストーリー断片内の語の関係を通じて語の活性度が伝播し変化する。図 2 にトピック伝播モデルの概念図を示す。

活性伝播モデルにおけるリンクとノードに相当する概念として、赤石が提案した語の関係を表す共起依存度と語の重要度を表す吸引力 [赤石 06] を用いる。語の活性度として**文脈依存**

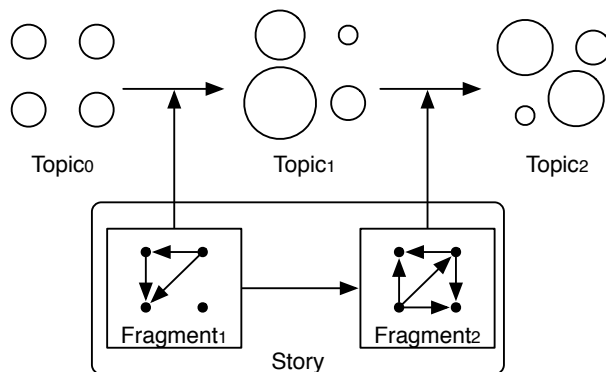


図 2: 文脈依存吸引力の概念図。トピックは文脈依存吸引力を持つ語の分布で特徴づけられる。連結するストーリー断片は語間の共起依存関係を持つ。ストーリー断片をつなげるによりその共起依存関係を通じて語の文脈依存吸引力が伝播し変化する。

吸引力を定義する。

3.2 共起依存度と吸引力

共起依存度はストーリー断片における語の関係を表し、吸引力は語の重要度を表す。

ストーリーに含まれる T 個の語 $t_i (i = 1 \sim T)$ について、あるストーリー断片における語 t_i から語 t_j への共起依存度 $d(t_i, t_j)$ は、語 t_i が出現した同じ文中に語 t_j が出現する条件付き確率で定義する。共起依存度 $d(t_i, t_j)$ は、語 t_i が出現した文の数を $sentences(t_i)$ 、語 t_i と語 t_j が同時に出現した文の数を $sentences(t_i, t_j)$ とすれば、以下の式で計算される。

$$d(t_i, t_j) = \frac{sentences(t_i, t_j)}{sentences(t_i)} \quad (1)$$

吸引力 $a(t_j)$ は、語 t_j がストーリー断片内の他の語を引き付ける力を、他の語から語 t_j に対する共起依存度の総和として定義する。吸引力の大きさはストーリー断片内の語の重要度を示す指標となる。以下の式で計算される。

$$a(t_j) = \sum_{i=1}^T d(t_i, t_j) \quad (2)$$

吸引力はストーリー断片が単独で存在するときに定義する語の重要度である。これらの概念を拡張し文脈依存吸引力を定義する。

3.3 文脈依存吸引力

あるストーリー断片 S において語 t_1 と、語 t_1 と関係ある語 t_2 と、語 t_1 と関係のない語 t_3 があるとする。このときストーリー断片 S をつなげる前のストーリーで語 t_1 が重要な語であれば、ストーリー断片 S をつなげた後のストーリーでは語 t_3 より語 t_2 のほうが重要になるであろう。

そこで共起依存度と吸引力の概念を拡張し、あるストーリーで重要な語に次につなげるストーリー断片で依存される語がより大きな吸引力を持つとする。ストーリー断片のつながりを考慮した語の重要度の指標として、 T 個の単語 $t_i (i = 1 \sim T)$ の、ストーリー断片を n 回連結したときの文脈依存吸引力 (context-dependent term attractiveness) $c_n(t_i)$ を連結前の

ストーリーにおける語の文脈依存吸引力と連結するストーリー断片における語の共起依存度の積の総和と定義する。

$$c_n(t_i) = \sum_{k=1}^T c_{n-1}(t_k) d_{S_n}(t_k, t_i) \quad (3)$$

ここで $d_{S_n}(t_i, t_j)$ は n 回目に連結したストーリー断片 S_n における語 t_i から語 t_j への共起依存度とする。図 2 に概念図を示す。なお初期文脈依存吸引力 $c_0(t)$ は適当に与える必要がある。

3.4 複数トピックの抽出

後述する逆問題解析を用いたトピックブリッジの計算量を抑えるために、ストーリー断片に存在する複数トピックの抽出を行う。本論文では共起依存度を要素とするベクトルを k-means 法 [Russell 09] を用いてクラスタリングを行い、各クラスタに属するベクトルの総和を各トピックの特徴量とみなす。

k-means 法への入力はストーリー断片のトピック数と共起依存度、出力は各トピックの重要度と各トピック内での語の重要度である。あるドキュメントはトピックの数を K 、語の数を T としたとき、 $K \times T$ の行列と K 次元のベクトルで特徴づける。

共起依存度を要素とするベクトルを

$$\mathbf{d}(t_i) = [d(t_1, t_i) \dots d(t_T, t_i)]^T \quad (4)$$

T 個の語に関する共起依存度ベクトル $(\mathbf{d}(t_1), \dots, \mathbf{d}(t_T))$ を K 個の ($K < N$) のクラスタ $C = \{C_1, \dots, C_K\}$ に分けるために以下を求める。

$$\arg \min \sum_{i=1}^K \sum_{\mathbf{d}(t_j) \in C_i} \|\mathbf{d}(t_j) - \boldsymbol{\mu}_i\|^2 \quad (5)$$

中心ベクトル $\boldsymbol{\mu}_i$ はクラスタ C_i に属するベクトルの平均値を表し、クラスタを特徴づけるベクトルとなる。このようにクラスタリングした結果、似たような語に依存している語が同じクラスタに属することとなる。

中心ベクトルからの距離に基づきクラスタリングされた共起依存度ベクトルにより、 k 番目のトピックにおける語 t_j の重要度の指標としてトピック吸引力 $a^{(k)}(t_j)$ を定義する。トピック吸引力はトピック内における共起依存度の和とする。トピック吸引力を要素とするトピック吸引力ベクトル $\mathbf{a}^{(k)}$ は以下のように定義される。

$$\mathbf{a}^{(k)} = \sum_{\mathbf{d}(t_j) \in S_k} \mathbf{d}(t_j) \quad (6)$$

あるトピックでトピック吸引力が大きい語は、そのトピック内で重要であるとみなせる。

また、トピック吸引力と文脈依存吸引力を合わせて、トピック文脈依存吸引力を次のように定義する。概念図を図 3 に示す。

$$c_n^{(k)}(t_j) = \sum_{i=1}^T c_{n-1}^{(k)}(t_i) d_s(t_i, t_j) \quad (7)$$

トピック文脈依存吸引力は複数トピックに対して文脈依存性を考慮した語の重み付け手法である。トピック数を 1 としたとき、文脈依存吸引力の値と同じになる。

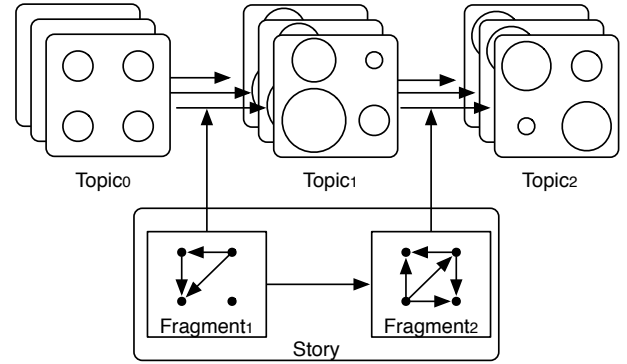


図 3: 複数トピックの概念図。ストーリーにストーリー断片を連結すると、連結したストーリー断片内の語の関係のネットワークを伝わり、複数のトピックがそれぞれ伝播する。

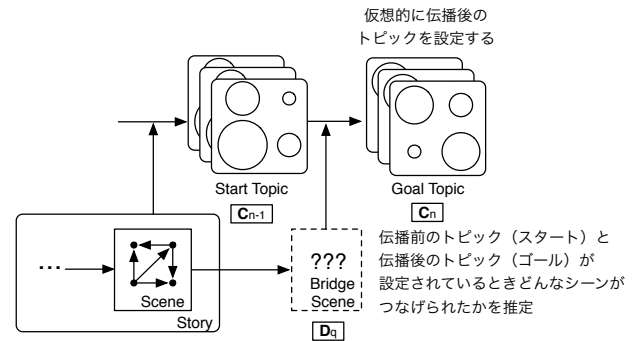


図 4: トピック伝播モデルの逆問題解析によるトピックブリッジの概念図。

共起依存度行列 \mathbf{D}_{S_n} を用いて $c_n^{(k)}$ をトピック k におけるトピック文脈依存吸引力ベクトル、 \mathbf{C}_n をトピック文脈依存吸引力行列、

$$\mathbf{c}_n^{(k)} = [c_n^{(k)}(t_1) \dots c_n^{(k)}(t_T)]^T \quad (8)$$

$$\mathbf{C}_n = [\mathbf{c}_n^{(1)} \dots \mathbf{c}_n^{(K)}] \quad (9)$$

とすれば、トピック文脈依存吸引力と共起依存度行列の関係は

$$\mathbf{C}_n = \mathbf{D}_{S_n} \mathbf{C}_{n-1} \quad (10)$$

と表される。なお、トピック文脈依存吸引力行列の各列ベクトルが 1 つのトピックの特徴に相当するが、この並びはユーザが指定するものとする。

3.5 トピック伝播モデルの逆問題解析

トピックブリッジを実現するためにトピック伝播モデルの逆問題解析を行う。図 4 に示すように、ストーリー断片をつなげる前とつながった後のトピックがわかっているとき、どのようなストーリー断片をつなげたか、すなわちどのようなブリッジによってトピックの遷移が起こったかというダイナミクスを同定する。このとき推定されたブリッジの特徴量に類似したストーリー断片がストーリーの骨格の一部となる。

ここでは便宜的にストーリー断片から求めた特徴をトピックとして扱い、ブリッジの起点となるトピックを求めるためのスタートのストーリー断片 S_{n-1} とブリッジの終点となるトピッ

クを求めるためのゴールのストーリー断片 S_n を用意する。これらのストーリー断片からトピック吸引力ベクトルを求める。そしてスタートのストーリー断片 S_{n-1} におけるトピック吸引力を伝播前のトピック文脈依存吸引力と見なし、ゴールのストーリー断片 S_n におけるトピック吸引力を伝播後のトピック文脈依存吸引力とみなす。

トピック文脈依存吸引力行列 \mathbf{C}_{n-1} とトピック文脈依存吸引力行列 \mathbf{C}_n の関係は、連結するストーリー断片 q の共起依存度行列 \mathbf{D}_q を用いて、式 10 と表される。これは共起依存度行列 \mathbf{D}_q によるストーリー断片 S_{n-1} からストーリー断片 S_n へのトピックの伝播を表す。ここで式 10 を共起依存度行列 \mathbf{D}_q を求めたい。ただし一般的にその疎性から一意に解を持たない。そこで逆行行列を一般化した擬似逆行行列 [久保 92] を用いる。

$$\hat{\mathbf{D}}_q = \mathbf{C}_n \mathbf{C}_{n-1}^+ \quad (11)$$

ただし \mathbf{C}_{n-1}^+ は \mathbf{C}_{n-1} の擬似逆行行列とする。擬似逆行行列は解のフロベニウスノルムが最小になる解を与える。ノルム最小解はネットワークの伝播において最も伝播の係数が小さくなる状態であり、これは語の重要度の変化の最も小さい遷移行列を求めることに相当する。

ここで推定された $\hat{\mathbf{D}}_q$ がスタートのトピックとゴールのトピックをつなげるブリッジの特徴を表す行列とみなせる。

3.6 ブリッジ候補のスコアの計算

ブリッジの共起依存度行列 $\hat{\mathbf{D}}_q$ とブリッジ候補となるストーリー断片の共起依存度行列 \mathbf{D}_c の類似度をスコアとして計算する。ここでは類似度は行列の要素の積和とする。

$$\text{score}(\hat{\mathbf{D}}_q, \mathbf{D}_c) = \sum_{i,j} \hat{\mathbf{D}}_{qij} \mathbf{D}_{cij} \quad (12)$$

このスコアに基づきブリッジ候補としての順位を付ける。

トピックブリッジングを実現する手法として、たとえばスタートとゴールからの距離を利用する方法も考えられる。すなわちスタートとゴールをつなぐために、スタートとゴールの両方に類似しているストーリー断片をブリッジとして、スタート、ブリッジ、ゴールという順でストーリー断片の連鎖を構成してストーリーを生成するという手法である。ただ、この手法では、ブリッジ内の離れた箇所にスタートのトピックに類似したトピックとゴールのトピックに類似したトピックが存在している候補の上位になる可能性がある。

これに対して提案手法では、スタートに出現する語とゴールの出現する語がストーリー断片で共起しているときに候補の上位になるので、候補の上位のストーリー断片ほどスタートとゴールの内容そのものを結びつけることができると考えられる。すなわちスタートのストーリー断片に含まれる語とゴールのストーリー断片に含まれる語が離れた箇所に記述されているようなストーリー断片よりも、同じ文に記述されているストーリー断片の方が、スタートとゴールをつなげるブリッジとしては適当であろう。

4. むすび

知識連繋のためにはストーリーを用いることが有効であると考えトピックブリッジング手法を提案した。トピックブリッジングは、スタートとゴールの2つのトピックをつなぐためにはどのようなブリッジが適当であるかということ推定し、スタートからブリッジ、ゴールと連なるストーリーの骨格を生成するものである。

トピックブリッジングのためにトピック伝播モデルを提案した。トピック伝播モデルはストーリー断片を連鎖させてストーリーを生成するとき、連結するストーリー断片によってトピックが伝播するモデルである。これはトピックを語の重要度のベクトルあるいは行列としてみなし、ストーリー断片を語の依存関係を表す遷移行列としてみなし。このトピック伝播モデルを逆問題解析して、トピックブリッジングを実現した。あらかじめ伝播する前のトピックと伝播した後のトピックを与えたとき、どのようなストーリー断片をつなげたかということ推定する。実際に計算されるのはブリッジのストーリー断片の特徴量を表す共起依存度行列である。トピック伝播モデルにおいて推定されたブリッジのストーリー断片の共起依存度行列はスタートのトピックからゴールのトピックへ変化するための推移行列とみなすものである。

参考文献

- [Amitani 08] Amitani, S. and Edmonds, E.: Generative Website: Visualising Possible Stories (2008)
- [Bringsjord 99] Bringsjord, S. and Ferrucci, D.: *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine*, L. Erlbaum Associates Inc., Hillsdale, NJ, USA (1999)
- [Brown 05] Brown, J.: *Storytelling in organizations: Why storytelling is transforming 21st century organizations and management*, Butterworth-Heinemann (2005)
- [Mani 99] Mani, I. and Bloedorn, E.: Summarizing Similarities and Differences Among Related Documents, *Inf. Retr.*, Vol. 1, No. 1-2, pp. 35-67 (1999)
- [Meehan 77] Meehan, J.: Tale-spin, an interactive program that writes stories, in *Proceedings of the fifth international joint conference on artificial intelligence*, pp. 91-98 Citeseer (1977)
- [Russell 09] Russell, S. and Norvig, P.: *Artificial intelligence: a modern approach*, Prentice hall (2009)
- [Shahaf 10] Shahaf, D. and Guestrin, C.: Connecting the dots between news articles, in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 623-632 ACM (2010)
- [綾 05] 綾 聡平, 松尾 豊, 岡崎 直観, 橋田 浩一, 石塚 満: 修辭構造のアノテーションに基づく要約生成, 人工知能学会論文誌, Vol. 20, No. 3, pp. 149-158 (2005)
- [久保 92] 久保司郎: 逆問題, 培風館 (1992)
- [松村 02] 松村 真宏, 大澤 幸生, 石塚 満: 語の活性度に基づくキーワード抽出法, 人工知能学会論文誌, Vol. 17, pp. 398-406 (2002)
- [赤石 06] 赤石 美奈: 文書群に対する物語構造の動的分解・再構成フレームワーク, 人工知能学会論文誌, Vol. 21, No. 5, pp. 428-438 (2006)
- [大澤 06] 大澤幸生: チャンス発見のデータ分析: モデル化+可視化+ コミュニケーション→シナリオ創発, 東京電機大学出版局 (2006)