3D2-1

物語生成のためのトピックブリッジング手法の提案

Bridging Topics for Story Generation

佐藤真[*1] 赤石美奈[*1] 堀浩一[*1]
Makoto Sato Mina Akaisihi Koichi Hori

[*1]東京大学大学院工学系研究科航空宇宙工学専攻
Department of Aeronautics and Astronautics, University of Tokyo

This paper introduces a method for bridging topics designed to facilitate generating stories over documents. First, we present a method for topic extraction based on narrative structure with k-means algorithm. We then model the story generation process and present a method for finding a bridge document between two documents.

1. はじめに

近年，知識創造・伝達の潜在的なダイナミクスを扱うこと，すなわち設計や研究の情報を動的なものとして扱うことの重要性が認識されている．すなわち，情報を蓄積されたときの文脈のみで利用するのではなく，人間の置かれた状況や文脈に応じて多様な視点から柔軟に情報の捉え，新しい物語性を付与した形で情報を伝えることで，情報をより価値のあるものとするということである．そこで我々はこれらを実現するためには情報を分解・再構成する方法論の開発が有効であると考え，過去に起きたことのデータベースをそのまま使うのではなく，一度文脈から切り離した形でドキュメントを保存し，必要に応じて物語性を付与した新しい文脈で再構成した形で提示するシステムを構築しようとしている．本研究は物語生成支援システムの要素である情報の再構成を担うトピックブリッジング手法を提案する．

2. 提案手法

トピックブリッジングの目的は，２つのドキュメントのトピックのギャップを埋めることができるようなドキュメントの特徴を提示し，物語生成を支援することである．たとえば，我々の研究について考えると，「蓄積された情報を利用する」というトピックから「情報をより価値のあるものにする」というトピックに橋を渡すために「情報の分解・再構成のフレームワークの開発」をするというストーリーを描いている．本手法ではこのような橋はどんなものになりうるかということを提示することを狙う．

手順の概要は次の通りである．まず，２つのドキュメントを用意する．一つは橋の起点となりもう一つは終点となる．それぞれのドキュメントの内容の特徴を表す行列を定義する．そして，それぞれの行列を操作することにより，各々のドキュメントから複数のトピックを抽出する．トピックの連鎖のモデル化を行い，これに基づき２つのトピックのギャップを特徴づける行列を推定する．

2.1 物語構造モデル

物語構造モデルは表１のようにドキュメントの階層構造の要素と物語構造のモデルの要素を対応付けを行うモデルである．

表 1: Mapping of Narrative Components to Text Elements

Narrative component	Text element
world model	set of stories
story	sequence of scenes (documents)
scene	chunk of events
event	set of terms (sentence)
character	term

たとえば，語をキャラクターに見立て，語の集合をイベント，イベントの集合をシーン，シーンの連鎖をストーリー，ストーリーの集合をワールドモデルとする．

ナラティブ連想情報アクセス・フレームワーク [1] では，物語構造モデルにおいてキャラクタの連鎖構造を操作して，元の文脈におけるストーリーをシーンに分解し，またシーンを新しい文脈において再構成してストーリーを生成する．異なる操作を通じて異なる文脈の物語を生成することができる．キャラクタの連鎖構造は次に示す２つの基本概念，共起依存度と吸引力によって定義される．

2.2 共起依存度と吸引力

文書を構成する N 個の語 $t_i(i = 1 \sim N)$ について, 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．

$$d_s(t_i, t_j) = \frac{sentences_s(t_i, t_j)}{sentences_s(t_i)} \tag{1}$$

吸引力 $a(t_i)$ は, 語 t_i が文書中の他の語を引き付ける力を，他の語から語 t_i に対する出現依存度の総和として定義する. 吸引力の大きさは文書中の語の重要度を示す指標となる．以下の式で計算される．

$$a_s(t_j) = \sum_{t_i \subset T} d_s(t_i, t_j) \tag{2}$$

さらに共起依存度と吸引力を拡張して文脈依存吸引力 [2] を定義する．シーンをつなげて新しいストーリを作るとき，前のシーンにより重要だった語に次のシーンにおいて依存される語

連絡先: 佐藤真，東京大学大学院工学系研究科航空宇宙工学専攻，〒 113-8656 文京区本郷 7-3-1 工学部 7 号館堀・赤石研究室，03-5841-6637，satomakoto[at]ailab.t.u-tokyo.ac.jp

はより重要になるとした．時刻 τ における語 t_j に文脈依存吸引力 $c_\tau(t_j)$ は，時刻 $\tau-1$ における文脈依存吸引力 $c_{\tau-1}$ と，つなげるシーン s における共起依存度 $d_s(t_i,t_j)$ を用いて次のように定義する．

$$c_\tau(t_j) = \sum_{t_i \subset T} c_{\tau-1}(t_i) d_s(t_i,t_j) \tag{3}$$

2.3 トピック抽出

シーンの内容を特徴づけるための量的な指標を定義する．もっとも簡単な分類手法の一つであるである k-means 法を用いてクラスタリングを行いトピックの抽出を試みる．

入力は文書のトピック数と共起依存度であり，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．この出力はトピック・モデル [3] と似ている．トピック・モデルでは，トピックは語の確率分布によって表され，ドキュメントはトピックの混合で表されるという概念に基づく．ただし，トピックモデルは生成モデルであるが，我々の手法は語の出現回数と共起回数という表層的な情報のみに着目している．

ベクトルの要素を共起依存度とするベクトルを定義，語 t_j に関する共起依存度ベクトル $\mathbf{d}(t_j)$ を定義する．

$$\mathbf{d}(t_j) = [d(t_1,t_j) \dots d(t_N,t_j)] \tag{4}$$

共起依存度ベクトルを k-means 法 [4] を用いてクラスタリングする．K-means 法はクラスタリング手法のなかでももっとも簡単な手法の一つであり，N 個の観測データを K 個のクラスターに分類する手法である．ここでは出現した N 個の語に関する共起依存度ベクトル $(\mathbf{d}(t_1),\dots,\mathbf{d}(t_N))$ を K 個の $(K<N)$ のクラスタ $S=S_1,\dots,S_K$ に分けるために以下を求める．

$$\arg\min \sum_{i=1}^{K} \sum_{\mathbf{d}_j \subset S_i} ||\mathbf{d}(t_j) - \mu_i||^2 \tag{5}$$

中心ベクトル μ_i はクラスタ S_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．

クラスタリングされた共起依存度ベクトルを用いて，トピック k の中における語 t_j の重要度の指標として話題内吸引力 $a^{(k)}(t_j)$ を定義する．話題内吸引力はトピック内における共起依存度の和とする．また話題内吸引力を要素とする話題内吸引力ベクトル $\mathbf{a}^{(k)}$ を次のように定義する．

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

複数のトピックであることを考慮して，文脈依存吸引力の拡張として話題内文脈依存吸引力を次のように定義する．

$$c_\tau(t_j)^{(k)} = \sum_{t_i \subset T} c^{(k)}_{\tau-1}(t_i) d_s(t_i,t_j) \tag{7}$$

話題内文脈依存吸引力はぞれぞれの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．

2.4 トピックブリッジング

トピックの連鎖をモデル化し，二つのトピックのギャップを推定し，平滑化するような橋となるトピックの特徴を表す．我々はシーンの並べ方によって各シーン内での語の重要度が変化するという文脈依存吸引力を提案したがこの逆計算を行う．つまり，あるシーン内である語がある重要度を持っていたとしたら，そのシーンの並びはどうなるかということをモデル化する．

まず，起点 $S_{\tau-1}$ と終点 S_τ から前節で述べた手法を用いてトピックの抽出を行い話題内吸引力ベクトルを求める．起点 $S_{\tau-1}$ における話題内吸引力を変化前の話題内文脈依存吸引力と見なし，終点 S_τ における話題内吸引力を変化後の話題内文脈依存吸引力と見なす．

そして，式 (7) で定義した文脈依存吸引力の逆計算を行う．今 $\mathbf{D}$ を共起依存度行列，

$$\mathbf{D} = [\mathbf{d}_1 \dots \mathbf{d}_N] \tag{8}$$

として，$\mathbf{c}^{(k)}_\tau$ をトピック k における話題内文脈依存吸引力ベクトル，$\mathbf{C}_\tau$ を話題内共起依存度行列，

$$\mathbf{c}^{(k)}_\tau = \left[c^{(k)}_\tau(t_1) \dots c^{(k)}_\tau(t_N)\right] \tag{9}$$

$$\mathbf{C}_\tau = \left[\mathbf{c}^{(1)}_\tau \dots \mathbf{c}^{(K)}_\tau\right] \tag{10}$$

とすれば，話題内共起依存度吸引力と共起依存度行列の関係は単純に

$$\mathbf{C}_\tau = \mathbf{D}_q \mathbf{C}_{\tau-1}. \tag{11}$$

となる．つまり，共起依存度行列 $\mathbf{D}_q$ が「橋」になるべく特徴を持っていると考えられる．ただ，上式は一般的にその疎性から一意に解を持たない．そこで逆行列を一般化した擬似逆行列を用いることとする．

$$\widehat{\mathbf{D}}_q = \mathbf{C}_\tau \mathbf{C}^+_{\tau-1}. \tag{12}$$

ただし，$\mathbf{C}^+_{\tau-1}$ は $\mathbf{C}_{\tau-1}$ の擬似逆行列とする．擬似逆行列は解のノルムが最小になる解を与える．このように求められた行列 $\widehat{\mathbf{D}}_q$ は「橋」となるドキュメントの特徴を示すと考えられる．

3. まとめ

本論文では複数のトピックを抽出する手法と物語生成のためのトピックブリッジング手法を提案した．トピック抽出のために物語構造モデルに基づき，話題内吸引力を定義した．トピックブリッジングのために文脈依存吸引力の逆問題を解くことを提案した．今後，提案した手法を用いた実験，アプリケーション，評価について進めていく予定である．本手法で提案した手法は非常に簡単なモデル化に基づいているので更なる拡張を考える．

参考文献

[1] 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).

[2] M. Sato, M. Akaihisi, K. Hori, “Analyzing Topic Transitions using Term Context-dependent Attractiveness,” Information Modelling and Knowledge Bases XXI, 2010.

[3] M. Steyvers and T. Griffiths, “Probabilistic topic models,” In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, 2006.

[4] C. Ding and X. He. “K-means Clustering via Principal Component Analysis,” Proc. of Int’l Conf. Machine Learning (ICML 2004), pp 225-232. July 2004.