1D1-4in

宇宙開発と社会との新たな連関を探る トピックブリッジング手法

Topic Bridging for Searching Unrevealed Relation between Space Development and Society

石川 雄基 Ishikawa Yuki　佐藤 真 Sato Makoto　堀 浩一 Hori Koichi　赤石 美奈 Akaishi Mina

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

In this paper, we present a computer-aided system for thinking from textual data. We propose the system applying the ideas of *Topic Bridging* and *Term attractiveness* for searching unrevealed relation between different topics. *Topic Bridging* estimates an appropriate topic as the middle of a story which connects between two topics as the start and end of the story by calculating matrixes. *Term attractiveness* deals with the value based on frequency in the use of a pair of words. A user inputs his textual data as the start of a story and the theme he treats, and our system outputs candidates of the story ending by evaluation criteria of *Term attractiveness*. The user chooses one from candidates, and our system outputs appropriate topics as the middle of this story by evaluation criteria of *Topic Bridging*. We take an example of finding unexpected topics connect topics of space development and social benefit.

1. はじめに

JAXA は長期ビジョンで、宇宙開発の「社会への定着・浸透」に取り組むとしている。このように今後の宇宙開発が抱える課題として、宇宙産業における新市場開拓、一般市民に向けた「社会への定着・浸透」が存在する。新市場を開拓するためには、これまでの宇宙産業における顧客と違うターゲットを取り込めば産業界への新たなアプローチが出来る。また、一般市民に向けた「社会への定着・浸透」を実現するためには、ある意味で遠い存在の宇宙開発を、一般市民が認識できる範囲内でも社会に役立つと感じてもらえればよい。

これらの課題に共通する解決策は、対象者にとっての宇宙開発の価値を高めることだ。その実現には、宇宙開発と対象者との意識との結びつき・関連が強まればよい。つまり、宇宙開発と対象者の間で、新しい関係や接点を見いだせばよい。

そこで本論文では、宇宙開発と社会との結びつきを強める新たな仕組みやアイデアの創出作業を想定して、あるトピックと別のトピックの連関を探るためのアイデア発想支援システムを提案する。具体的には文書情報を活用して、異なる2つのトピックを連結する新たなトピックを文書群から推定する。宇宙開発というトピックと社会の利益というトピックを結び付ける予想外のトピックを発見するという実験例を示す。

2. システム概要

前節で述べたように、この論文では文書情報を活用した発想支援システムを提案する。そこでは、文献[佐藤 10]でのトピックブリッジングという概念と、文献[赤石 06]での吸引力という概念を導入する。トピックブリッジングでは、2 つの異なるトピックをそれぞれ物語の始点・終点と解釈して、その間をつなぐトピックを推定する。吸引力では、語のペアの出現頻度を基にして、それぞれの語の重要度を吸引力という値で評価する。発想支援システムのユーザは、物語の始点となるスタートテキストおよび扱いたいテーマワードを入力する。発想支援システムは吸引力の概念を用いて、テーマに合った物語の終点の候補を提示する。候補からユーザが終点を選択すると、システムではトピックブリッジングの概念を基にして物語の中間点にふさわしい新たなトピックを提示する。

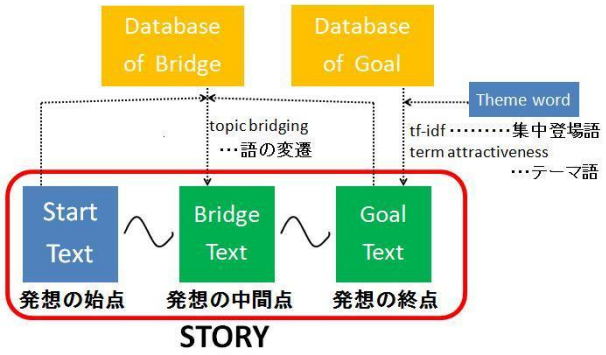

表 1　システム概要図

2.1 ユーザから観た際のシステム入出力

提案のシステムではまず、発想の始点になるスタートテキストと、発想のテーマにしたいテーマワードを、ユーザ側が入力する。その入力に応じて、発想の終点になるゴールテキスト、発想の始点と終点を結ぶブリッジテキストをシステム側で出力する。ゴールテキストは、テーマワードに関する吸引力が高いものをゴールデータベースから選び出す。その際に参考として、ゴールテキストにおいてテーマを論じる場面でだけ特に集中して登場する語をユーザは知ることができる。この集中登場語は、今回のテーマをゴールテキスト著者が論じる際の象徴的な語となり、テーマの捉え方を知るのに便利な語となりうる。以下では、この集中登場語をゴールトピックワードと呼ぶことにする。

ユーザは、スタートテキストの内容を始点に、ブリッジテキストの内容を中間点に、ゴールテキストの内容を終点にしたストーリーを作る、という制約下で発想を行う。このような制約下で発想させることで、宇宙開発と社会との新たな連関を探る・見つけ出すことが狙いである。以上の推定結果を基に、ユーザが新たなストーリーを考える。この様にシステムによる出力結果によって、宇宙開発分野における新たな仕組みやアイデアの考案を支援するシステムとなる。

連絡先：石川雄基，東京大学　工学系研究科
航空宇宙工学専攻　知能工学研究室
E-mail: yishikawa@ailab.t.u-tokyo.ac.jp

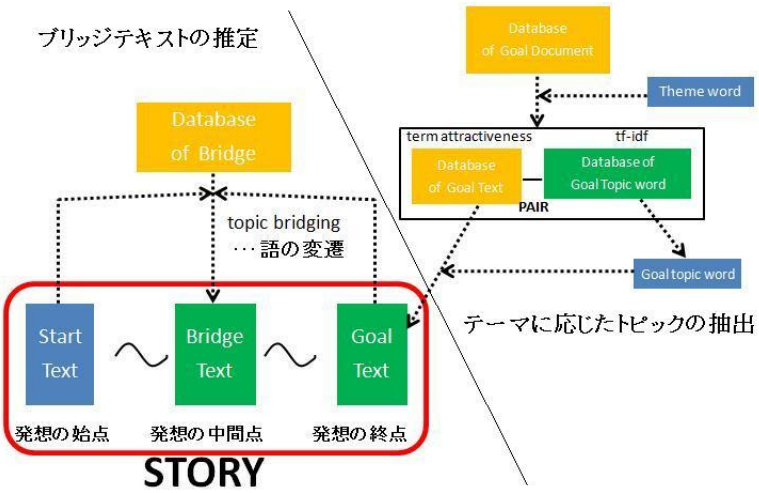

表 2　システム詳細図

ここまでは、システムを全体的に把握してもらうために、表 1 のシステム概要図を用いていた。これからは、システムの詳細な手法や計算方法を見ていきたい。そこで 2.2 節および 2.3 節では、表 1 の概要図を詳しくした表 2 のシステム詳細図を用いて説明する。システム全体は、「テーマに応じたトピックの抽出」と「ブリッジテキストの推定」に分けられる。例えばユーザが宇宙開発側なら、前者では宇宙開発の仕組みやアイデアにより取り込みたいターゲットの意識を、ターゲットが著した文書から推定する。後者では、宇宙開発側の意識とターゲットの意識を繋げる橋渡し役の事柄を、別分野の文書から推定する。

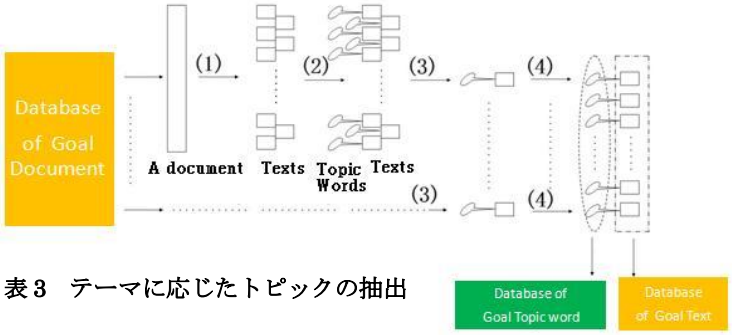

表 3　テーマに応じたトピックの抽出

2.2 テーマに応じたトピックの抽出

前者の「テーマに応じたトピックの抽出」では、ゴールデータベースの文書情報の中から、ユーザの入力に合ったゴールテキストを出力する。その過程で、tf-idf という概念と、吸引力という 2 つの異なる概念を使用する。

まず、ゴールデータベース内の各文書を、文字数毎の切断により複数のテキストにする。これが表 3 の(1)に当たる。続いて、ゴール文書データベース内の各文書（表 4 では d と記す）において、文字数毎の切断により複数のテキスト（表 4 では t と記す）に分割した際に登場する各言葉（w と記す）の特徴を表 4 の計算式により数値化する。テキスト t に出現する語のうち、次のような式で定義された tf-idf の値が最も高いものを、テキスト t にタグ付けする。これが表 3 の(2)に当たる。なお tf-idf を使ったのは、特定のテキストに集中して登場する単語を抽出する狙いがある。

続いて、単語のペア(w1,w2)を考えた際の共起依存度および吸引力を表 5 の式で定義して、システムに入力したテーマワードが w1 である際の吸引力が最も高いテキスト t を文書 d のなかから抽出する。これが表 3 の(3)に当たる。なお吸引力を使ったのは、テキストでのテーマ語を抽出する狙いがある。最後に、各ゴール文書から抽出されたテキストおよびタグを、ゴール文書データベース全体でまとめる。これが表 3 の(4)に当たる。ここで得られたゴールトピックワード群を、ユーザに向けてゴールトピックワード選択肢として出力する。それぞれのゴールトピックワード選択肢は、ゴールテキストデータベースにタグ付けされている。ユーザは、ゴールトピックワード選択肢のなかから、ゴールトピックワードを選択して、システムに入力する。ユーザによるゴールトピックワードの選択に応じて、ゴールテキストを呼びだしてユーザに出力する。

$$tf(t,w) = \sum_i 1\ (w = w_i \in t)$$

$$idf(d,w) = \log\left(\sum_j 1\ (\forall i \mid w \neq w_i \in t_j \in d) \right)$$

$$tf\text{-}idf(t,w) = tf(t,w) \times idf(d,w)$$

d: 文書

t: 文書 d の一部から生成したテキスト

t_j: 文書 d から j 番目に生成したテキスト

w: 単語

w_i: テキスト内で i 番目に登場する単語

表 4　本システムでの tf-idf の定義式

$$共起依存度(w1 \rightarrow w2, t) = \frac{\sum_i 1\ (w_1, w_2 \in l_i \in t)}{\sum_i 1\ (w_1 \in l_i \in t)}$$

$$吸引力(w1, t) = \sum_{w2} 共起依存度(w1 \rightarrow w2, t)$$

w_1, w_2: ペアとなる各単語

t: テキスト

l_i: テキスト内で i 番目に登場する行

表 5　本システムでの吸引力の定義式

2.3 ブリッジテキストの推定

発想の始点になるスタートテキストと発想の終点になるゴールテキストとの橋渡しとなる内容を、ブリッジテキストとして出力する。その際、トピックブリッジングという手法を使用する。その手法を以下に述べる。

まず、スタートテキストおよびゴールテキストにおいて、各単語の吸引力をベクトルにした吸引力ベクトルを各テキストで求める。ブリッジテキストデータベースにある各テキストにおいては、スタートテキストおよびゴールテキストに登場する単語での共起依存度による行列を作る。

それらの吸引力を基にしたベクトルおよび共起依存度を基にした行列から、表 6 の式で定義された類似度を出す。ブリッジテ

キストデータベース内にある各テキストを、類似度が高い順に、ブリッジテキスト候補としてユーザに出力する。このように 2 章で説明したような入力・出力を基にして、ユーザにストーリーを作成させることで発想支援を行う、という一連の流れである。

$$S = [S1 \cdots Sn]^T$$

$$G = [G1 \cdots Gm]^T$$

$$B'ij = \text{共起依存度}(wi \rightarrow wj, BT)$$

$$\text{類似度}(S, G, B') = \sum_{i=1}^{m} \sum_{j=1}^{n} Bij \cdot B'ij$$

ST:スタートテキスト

GT:ゴールテキスト

BT:ブリッジテキスト候補

S_i: スタートテキストにおいて i種類目の単語での吸引力

G_j: ゴールテキストにおいて　j種類目の単語での吸引力

B: G =BSを満たし

　フロベニウスノルム最小解を与える疑似逆行列

B′: ブリッジテキスト候補の共起依存度を基にした行列

B′ij:　B′の各成分

表 6　本システムでの類似度の定義式

3. 実験と評価

提案したシステムによる実験として、「JAXA が抱く『社会への定着・浸透』という意識」と「消費者相手の企業が『社会』に抱く意識」との連関を探るための、システム使用例を示す。スタートテキストとして、JAXA が文書中で「社会への定着・浸透」について述べた 1000 字程度の部分を使用する。テーマワードに「社会」を使用する。ゴール文書データベースには、東証 1 部上場企業のうち 109 社のアニュアルレポート(もしくは相当文書)を使用した。なお解析にあたり、一般的な語はストップワードとする。

また 2.2 節で述べた「テーマに応じたトピックの抽出」においては、ゴールデータベース内の各文書を文字数毎の切断により複数のテキストにする。今回の実験では、テキスト自体の文字数の目安は 1000 文字（数字・記号除く）、前後のテキストでののりしろを 50 文字（数字・記号除く）と設定する。文書の文字数によって異なるが多くの文書では、各文書から数十から数百のテキストにした。

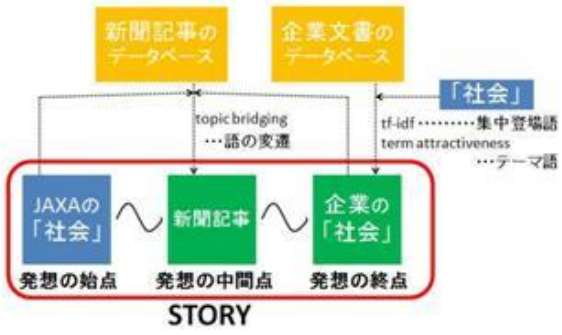

表 7　実験設定概要図

「テーマに応じたトピックの抽出」の結果として、各テキストにタグ付けされたゴールトピックワードが出力され、それらの集まりはゴールトピックワードのデータベースとなる。本システムでは、ゴールトピックワードのデータベースにある候補からユーザが選択し入力できるようにしているが、今回は最上位候補を選択し入力する。本実験でのゴールトピックワードのうち最上位候補は「CSR」で、全 109 のゴールテキストのうち 10 に対応していた(それ以外のゴールトピックワードは、3 つ以下のゴールテキストにしか対応していなかった)。以下、「CSR」をゴールトピックワードとして選択した実験結果について述べる。

「CSR」をゴールトピックワードに持つ 10 のゴールテキストから任意に 1 つ選んで発想の終点、設定済のスタートテキストを発想の始点として、橋渡しの役割を担うブリッジテキストを推定する。ブリッジテキスト候補を集めたブリッジテキストデータベースとして、毎日新聞のニュースサイト『毎日 jp』での『ニュースセレクト』内から1カ月分の新聞記事をサイト内でのジャンル(「サイエンス」「経済」「話題」)別にデータベース化したものを使用した。

「サイエンス」は研究開発側に近く、「経済」は企業側に近い。つまり、「サイエンス」はスタートテキストの内容に近く、「経済」はゴールテキストの内容に近い。「話題」は他 2 つの分類と比べて、スタートテキストの内容やゴールテキストの内容と離れている。ブリッジテキストデータベースによる実験の影響を考察するため、このようなブリッジデータベースで実験する。各ゴールテキストおよび各ブリッジテキストデータベースの組合せにおいて、1 位に推定されたブリッジテキストを表 8~10 に示す。

ゴールテキスト(著者名)	ブリッジテキスト1位(新聞記事の見出し)
カシオ計算機	特集：ＩＳＳ長期滞在、野口聡一さん飛行士が京都で報告会 宇宙への夢、無限大
ダイキン	特集：ＩＳＳ長期滞在、野口聡一さん飛行士が京都で報告会 宇宙への夢、無限大
マツダ	特集：ＩＳＳ長期滞在、野口聡一さん飛行士が京都で報告会 宇宙への夢、無限大
丸紅	特集：アジア環境フォーラムｉｎ秋田 環境問題に国境なし
住友信託銀行	義題：海の汚染防止、外来種持ち込まず…エコ時代　１２年から認証制度、取得へ奮闘中
新日鉱	特集：アジア環境フォーラムｉｎ秋田 環境問題に国境なし
大林組	特集：ＩＳＳ長期滞在、野口聡一さん飛行士が京都で報告会 宇宙への夢、無限大
帝人 ※1	義題：海の汚染防止、外来種持ち込まず…エコ時代　１２年から認証制度、取得へ奮闘中
NTT	特集：アジア環境フォーラムｉｎ秋田 環境問題に国境なし
電通	義題：海の汚染防止、外来種持ち込まず…エコ時代　１２年から認証制度、取得へ奮闘中

表 8　ブリッジテキストデータベース「サイエンス」での推定

ゴールテキスト(著者名)	ブリッジテキスト1位(新聞記事の見出し)
カシオ計算機	レアアース：調達分散化 米、豪州など有力視－－経産省方針
ダイキン	インタビュー・環境戦略を語る：住友信託銀行・大久保哲夫常務執行役員
マツダ	インタビュー・環境戦略を語る：ソニー・中鉢良治副会長
丸紅※2	毎日起業家新聞：社員の海外志向を強める建築は。　「英語を社内公用語に」１５．５%
住友信託銀行※3	イラン・アザデガン油田開発：日本撤退 資源外交、見通し甘く
新日鉱	レアアース：中国、欧米向けも停止 資源外交の武器、ハイテク製造見据え
大林組	イラン・アザデガン油田開発：日本撤退 資源外交、見通し甘く
帝人	イラン・アザデガン油田開発：日本撤退 資源外交、見通し甘く
NTT	インタビュー・環境戦略を語る：ソニー・中鉢良治副会長
電通	イラン・アザデガン油田開発：日本撤退 資源外交、見通し甘く

表 9　ブリッジテキストデータベース「経済」での推定

ゴールテキスト(著者名)	ブリッジテキスト1位(新聞記事の見出し)
カシオ計算機	ＮＨＫ受信料：引き下げ「１０%還元」の行方－－検討本格化
ダイキン※4	点字毎日文化賞：「手づくりの治療」次世代へ受け継ぐ－－（財）杉山眼科過徳顕彰会
マツダ	クローズアップ２０１０：鈴氏ノーベル平和賞（その2止）　日本、評価遅ける
丸紅	ロータリー交差点：事故防止効果、検証へ　長野・飯田でデータ収集
住友信託銀行	義題：海の汚染防止、外来種持ち込まず…エコ時代　１２年から認証制度、取得へ奮闘中
新日鉱	クローズアップ２０１０：鈴氏ノーベル平和賞（その2止）　日本、評価遅ける
大林組	羽田空港：不法侵入者に備え 新国際線ターミナルで訓練
帝人	義題：海の汚染防止、外来種持ち込まず…エコ時代　１２年から認証制度、取得へ奮闘中
NTT※5	ロータリー交差点：事故防止効果、検証へ　長野・飯田でデータ収集
電通	義題：海の汚染防止、外来種持ち込まず…エコ時代　１２年から認証制度、取得へ奮闘中

表 10　ブリッジテキストデータベース「話題」での推定

それらの入力情報および出力情報から、スタートテキストの内容をストーリーの始点、ブリッジテキストをストーリーの中間、ゴールトピックワード・ゴールテキストの内容をストーリーの終点とした、ストーリーを作成する。実際に作成例として、ストーリー作成例 1~5 を表 11 に示す。

そのうちストーリー作成例 4 が、宇宙開発の技術開発以外での利用や民間企業などの利用を進めていこうとする ISS 日本実

験棟「きぼう」における「きぼう利用フォーラム」での研究会の実例「宇宙鍼灸科学研究会」と一致する部分が多い。鍼灸の事例は小さな芽の段階にある施策だが、本システムを用いて現在の施策を自然な流れで発想することができた。

ストーリー作成例 1（表 8 内　※1における入力・出力情報よりストーリー作成）

JAXA が宇宙開発事業者に対する認証制度を設け、環境に配慮した宇宙開発を行う事業者に御墨つきを与える。その結果、認証された事業者は一般消費者に対して、CSR（＝企業の社会的責任）を果たしていると印象づけることができる。

ストーリー作成例 2（表 9 内　※2における入力・出力情報よりストーリー作成）

世界的な視野を持つ宇宙開発機関での機関内公用語として英語を導入する試験を行うことで、数十年先のグローバル社会を想定した英語公用語の長所・短所を知覚することができる。その結果、私企業が CSR（＝企業の社会的責任）として世界的な視野で活動する場合に、英語を社内公用語として用いるどうかの判断材料となる。

ストーリー作成例 3（表 9 内　※3における入力・出力情報よりストーリー作成）

宇宙開発事業者は、宇宙開発によって日本の資源開発における自主権益を確保する可能性を探る。その結果、私企業が宇宙開発に携わるか判断する際に、CSR（＝企業の社会的責任）として「国家を利し、かつ社会を利する底の事業」を展開するという具体的な動機づけが生まれる。

ストーリー作成例 4（表 10 内　※4における入力・出力情報よりストーリー作成）

患者に癒しを与えるとされる鍼灸治療の科学的研究を、極限の緊張を強いられる宇宙空間で行う。そうすると、伝統的な日本の鍼灸を世界にアピールする機会となると共に日本で再注目されるきっかけとなる。その結果、私企業が CSR（＝企業の社会的責任）として「グローバルに社会的責任を果たし」ながらも「それぞれの地域の役に立つ」社会貢献活動を行う際の新たな選択肢として、鍼灸治療が注目されるようになる。

ストーリー作成例 5（表 10 内　※5における入力・出力情報よりストーリー作成）

中心に進入できない部分がある輪の形をした交差点であるロータリーの特長を、人工衛星の軌道設計に活かすことができないか探る。その結果、情報通信企業が CSR（＝企業の社会的責任）として「より豊かで便利なコミュニケーション環境を実現」するために貢献することができるかもしれない。

表 11　ストーリー作成例 1~5

また、ブリッジテキストデータベースが「サイエンス」「経済」「話題」の 3 種類それぞれの場合で実験を行った。その際、スタートテキストとゴールテキストを橋渡しするブリッジテキストの推定結果（1 位、2 位、3 位、4 位、5 位、……）を、「CSR」をゴールトピックワードに持つ 10 種類の各ゴールテキストで行った。ブリッジテキストの推定結果として上位 N 位以内に表示されるブリッジテキストの数は N×10 個である。しかし、表示されたブリッジテキストに被りがあると種類数は N×10 より少なくなる。今回の実験ではブリッジテキスト候補の規模が異なるため、「種類数」および「種類数を総種類数で割った値」の両方で比較した。表 12 に示す。各項目において、3 つのうち最も数値が大きいものを赤く、最も数値が小さいものを青く塗っている。

ブリッジテキストデータベースのジャンル	サイエンス	経済	話題
ブリッジテキスト候補の総種類数	279 種類	447 種類	360 種類
ブリッジテキスト各 1 位の種類数	3 種類	6 種類	6 種類
各 1 位がブリッジテキスト全候補に占める割合	0.01075	0.01342	0.01667
同様にブリッジテキスト各 2 位以内の種類数	9 種類	11 種類	14 種類
各 2 位以内がブリッジテキスト全候補に占める割合	0.03226	0.02461	0.03889
ブリッジテキスト各 3 位以内の種類数	16 種類	16 種類	18 種類
各 3 位以内がブリッジテキスト全候補に占める割合	0.05735	0.03579	0.05000
ブリッジテキスト各 4 位以内の種類数	17 種類	18 種類	22 種類
各 4 位以内がブリッジテキスト全候補に占める割合	0.06093	0.04027	0.06111
ブリッジテキスト各 5 位以内の種類数	20 種類	21 種類	24 種類
各 5 位以内がブリッジテキスト全候補に占める割合	0.07168	0.04698	0.06667

表12　ブリッジテキストデータベースのジャンルによるブリッジテキスト推定結果の多様性比較

「サイエンス」「経済」「話題」のうち赤色が目立つのは、スタートテキストの内容やゴールテキストの内容と離れている「話題」である。スタートテキストの内容に近い「サイエンス」や、ゴールテキストの内容に近い「経済」では、青色が目立つ。この結果からは、スタートテキストおよびゴールテキストと離れたブリッジテキストデータベースを用いると、推定結果が拡散的になると予想できる。発想の目的によって、最適なブリッジテキストデータベースの取り方が考えられるかもしれない。

今後、ユーザを交えた実験、関連研究との比較を行って、提案した発想支援システムの評価や改良を行っていきたい。

4. 関連研究

我々に近い研究として、1 つには新規注目語の傾向検出手法[Qiaozhu 05]が挙げられる。そこでは時系列情報としてテキストを扱い、潜在的なテーマの発見を目指している。新たな関係の発見を目指す点は、我々の研究と共通している。だが、その文献ではユーザの趣向や狙いについては反映されていない。社会に対する宇宙開発の新たなアプローチを念頭に置くと、アプローチする側とされる側に合わせた出力を得たい。そこで我々の研究では、ユーザが入力したテキストを物語の一部分と見立て、ストーリーの遷移に着目している。また 文献[James 09]はストーリー生成なのだが、我々とは研究対象が異なる。その文献では、単純な 1 つの行動もしくは関係性などを表した短文を 1 つの単位として、その後の展開を推定する。一方、新たな連関を探りたい我々は、長文のテキストを 1 つの単位として、ストーリーの始点・終点を橋渡しするテキストを推定する。

他にも、新聞記事同士での線形計画法を使ったストーリー生成手法[Dafna 10]が挙げられる。そこでは関連性(Relevance)、一貫性(Coherence)、冗長性(Redundancy)および熟知度(familiarity)をユーザによる評価項目としている。だが、発想を行う目的によって最適解は異なりうる。例えば、本論文のタイトルのように宇宙開発と社会との新たな連関を探る場合には、目新しさ・新規性が重要な要素になる。今後、対象とするユーザおよび目的に合う適切な評価関数を置いたうえで、このような関連研究との比較を通してシステムの評価を行っていきたい。

5. まとめ

あるトピックと別のトピックの連関を探るために、文書情報を活用した発想支援システムを提案した。その際に、2 つの異なるトピックをそれぞれ物語の始点・終点と捉えた「トピックブリッジング」、語のペアの出現頻度に着目した「吸引力」という概念を用いた。実験として、宇宙開発というトピックと社会の利益というトピックを結び付ける予想外のトピックを発見する例を示した。

参考文献

[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.

[佐藤 10] 佐藤真 赤石美奈 堀浩一: 物語生成のためのトピックブリッジング手法の提案, 2010 年度人工知能学会全国大会, 2010.

[Qiaozhu 05] Qiaozhu Mei: Discovering Evolutionary Theme Patterns from Text, KDD'05, 2005.

[James 09] James Niehaus & R. Michael Young: A Computational Model of Inferencing in Narrative, AAAI 2009 Spring Symposium, 2009.

[Dafna 10] Dafna Shahaf & Carlos Guestrin: Connecting the Dots Between News Articles, KDD'10, 2010.