トピック語を網羅する文抽出のための TextRank 向け文間関係尺度の検討

金子 浩一[†]　渋木 英潔[§]　中野 正寛[†]　宮崎 林太郎[†]　石下 円香[§]　永井 隆広[‡]　森 辰則[§]

[†]横浜国立大学 大学院 環境情報学府　[‡]横浜国立大学 工学部　[§]横浜国立大学 大学院 環境情報研究院

E-mail: {kaneko,shib,nakano,rintaro,ishioroshi,nagadon,mori}@forest.eis.ynu.ac.jp

1　はじめに

Web 上に存在する情報は，ブロードバンド化の進展やブログ等の普及に伴い，爆発的に増加し続けている．これらの情報の中には，出所が不確かな情報や利用者に不利益をもたらす情報などが含まれており，信頼できる情報を利用者が容易に得るための技術に対する要望が高まっている．しかしながら，情報の内容の真偽や正確性を検証することは困難である上に，その情報が意見などの主観を述べるものである場合には，利用者により考え方や受け止め方が異なることから，その真偽や正確性を検証することはさらに困難なものとなる．そのため，情報の信憑性は，最終的に個々の情報利用者が判断しなければならず，利用者による信憑性の判断を支援する技術の実現が優先して解決すべき課題であると考える．我々は，主観的な意見や評価だけでなく，疑問の表明や客観的事実の記述を含めたテキスト情報を広く言明[1]と呼ぶこととし，ある言論集合における個々の言論の相対的な位置づけを提示することで情報信憑性判断を支援することを目指している．

現在，我々が Web 上の情報の信憑性を判断しようとした場合，Google などの Web 検索エンジンを用いて得られた関連文書を読んで判断することが多い．しかしながら，ある検索文書に書かれている内容が他の検索文書に書かれている内容と矛盾している場合，検索エンジンはその矛盾点を読み解くためのいかなる手掛かりも示さない．これに加えて，検索された文書全てに目を通すことは困難であるが，検索エンジンのランキングは信憑性判断の観点から行われているわけではなく，上位の文書だけを読めば良いというわけではないため，現在の情報検索技術は情報信憑性の判断支援という観点からは満足できるものではない．それゆえ，情報信憑性の判断支援のための技術が求められている．

信憑性の判断を支援する技術は大きく 3 つに分類できる．第一の技術は，利用者が着目するトピックに関連する言論群やそれらの言論の発信者[2]といった記述の抽出技術である [5]．第二の技術は，言論間の対立関係や根拠関係などを解析するための技術である [6, 10]．第三の技術は，第一，第二の技術で抽出，解析された情報を利用者が容易に判断できるように要約，整理する技術である [9]．我々は，第三の技術として，利用者の情報信憑性判断を支援するためのサーベイレポートを自動的に生成することを目標としている [3]．我々は，言論マップ生成システム [6, 10] と連携し，言論マップ生成システムによって解析された論理関係を考慮した要約の生成を目標としている．本稿では，言論マップ生成システムの前処理として，検索された Web 文書の絞り込みを行う文抽出手法について検討する．

2　情報信憑性判断支援のための要約

自動要約の技術には，重要文抽出による要約と文圧縮による要約が存在する [12]．これに対して，情報信憑性判断支援のための要約では，重要パッセージ抽出による抜粋型の要約が中心技術となる．人間が文章の信憑性を判断する場合，対立や根拠といった骨子となる文間の関係だけではなく，ニュアンスの伝わり方の違いなど，個人の感性に影響される微妙な表現も考慮されることが多い．このような微妙な表現を現在の技術で正確に処理することは困難であるため，可能な限り原文書の表現を保持したまま提示することが現時点での最適の方法であると考えられる．それゆえ，パッセージ単位による抽出を要約の中心技術とした．

一般的な目的の要約では，原文書に書かれている内容を利用者に伝えることが重要な目的である [2]．しかしながら，情報信憑性の判断を支援するという目的を達成するには十分ではない．例えば，二つの文書に「ディーゼル車は環境に良い」と「ディーゼル車は環境に悪い」という対立する言明がそれぞれ書かれていた場合，単純に両方の言明を並べて提示することは，利用者

[1]本稿では，文献 [10] に倣って一つ一つの文に対する個々の記述を言明と呼び，言明の集合である言論と区別することにする．

[2]我々は，発信者の定義を，Web 上のある情報を記述した人物や組織と定義している．

が言明の信憑性を判断する際の材料とならない．そればかりか，「対立している言明のどちらかが間違っているのか」，それとも，「互いに別の面を述べていてある状況下では両立できるのか」といった，どのように解釈すべきかの判断すらできない．前者の解釈であれば，各言明の真偽を判断するために，その言明が導かれるまでの根拠などを中心に提示すべきであるし，後者の解釈であれば，両立できる状況に焦点を当てて提示すべきである．それゆえ，情報信憑性のための要約では，原文書の内容に加えて，その内容をどのように解釈すべきかといった記述の提示が必要となる．我々は，解釈のための記述も Web 文書からのパッセージ抽出により生成することを意図しており，原文書の内容に関する記述と解釈のための記述を利用者が区別しやすいように，構造化文書としてサーベイレポートを生成することを検討している．

我々は，このような要約を生成する処理の過程において，言論マップ生成システムと連携し，言論間の関係の解析結果を取得する予定である．言論マップとは，言論間の類似，対立，含意等の論理関係を解析してマップ化したものである．言論マップ生成システムに解析を依頼する前に，解析時間の短縮や後の処理の効率化のため，テキストの量を減らすための大まかな絞り込みを行う．この絞り込みでは，トピック語が網羅されたパッセージを残しつつ，不要なテキストを削除することを目指す．本稿におけるトピック語とは，4章で用いる「マイナスイオン」や「女性専用車両」といったトピック名そのものを指すのではなく，それらのトピックに関する文に現れる語のことであり，トピック語を網羅するとは一般的な要約において注目される主要トピックについての語だけでなく，関連するトピックにまつわる語についても網羅することを意味している．

3 TextRank に基づく文抽出手法

我々は，文抽出に比較的高精度である TextRank アルゴリズムを用いた．TextRank は，Mihalcea ら [4] によって提案され，PageRank[1] を自然言語処理に適用したものである．TextRank はグラフ構造に基づいたランキングアルゴリズムであり，頂点となるテキストの断片について，その局所的な情報ではなくグラフ構造全体から得られるテキスト全体に関わる大域的な情報をもとに頂点の重要度を決定する．グラフ構造において，多くの頂点からリンクされている頂点は重要だという考えに基づき文の重要度を決定する．

本章では，Mihalcea らが提案した TextRank アルゴリズムにおいて用いられている文間関係尺度及び，我々が提案する文間関係尺度について示す．

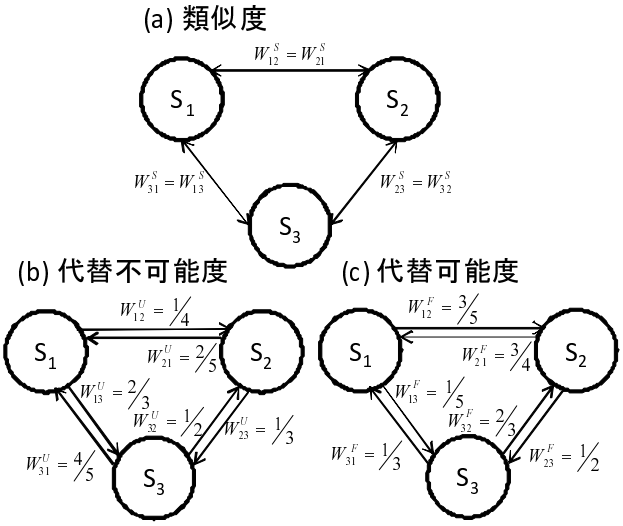

図 1: グラフ構造およびリンクの重みの計算の例

3.1 Mihalcea らが用いた文間関係尺度

TextRank アルゴリズムでは文を頂点とし，文 S_i と文 S_j の間にあるリンクの重み W_{ij}^S を以下の式で計算する．

$$\begin{aligned} W_{ij}^S &= Similarity(S_i, S_j) \\ &= \frac{|\{w_k \mid w_k \in S_i \& w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)} \end{aligned} \quad (1)$$

ここで，w_k は文中の名詞であり，$|S_i|$ は文 S_i 中の名詞の数である．図 1(a) は文を頂点としたグラフ構造の例である．

文 S_i の重要度 $TR(S_i)$ は以下の式で計算される．

$$\begin{aligned} TR(S_i) &= (1-d) \\ &+ d * \sum_{S_j \in In(S_i)} \frac{W_{ij}^S}{\sum_{S_k \in Out(S_j)} W_{jk}^S} TR(S_j) \end{aligned} \quad (2)$$

ここで，$In(S_i)$ は S_i をリンク先とする文の集合であり，$Out(S_j)$ は文 S_j がリンク元となっているリンクの先にある文の集合である．また，d は 0 から 1 の間の値をとる定数であり，実験では 0.85 とした．これは，Brin ら [1] が使った値と同じである．式 (3) の条件を満たすまで，式 (2) の計算が繰り返される．

$$\sum_{S_i \in G} (TR^{t+1}(S_i) - TR^t(S_i)) < \varepsilon \quad (3)$$

ここで，G はグラフ構造内の文の集合であり，$TR^t(S_i)$ は t 回目の計算での S_i の重要度である．ε は収束条件であり，実験では 0.0001 とした．この値は Mihalcea ら [4] が使った値と同じである．

表 1: 文と文に含まれる語の関係

	w_1	w_2	w_3	w_4	w_5	w_6	w_7
S_1	√	√	√	√	√		
S_2	√	√	√			√	
S_3	√					√	√

3.2 提案する文間関係尺度

Mihalcea ら [4] は式 (1) に示すように，文間評価尺度として，二文間に共通する語の数から決定される類似度を用いた．しかしながら，このような類似度を用いた手法では，トピック語を網羅するパッセージ抽出には不向きである可能性がある．なぜなら，類似度を用いると，他と類似する語が多い主要なトピック語を含んだ文しか抽出されないからである．それゆえ，我々は，様々なトピック語を網羅するために文間評価尺度に代替不可能度を導入した．代替不可能度とは，相手にない情報をどれだけ含んでいるかを表す尺度である．高村らは文間の含意関係を考慮して内容全体を被覆する文を選択するモデルを提案している [8]．このモデルでは，より多くの内容を含意している文が選択されやすいのに対し，代替不可能度では，トピック語を網羅するために含意していない量を文間評価尺度をとして用いる点で異なる．高村らはクエリによる制限のない generic な文書要約に注目しているが，我々が取り組んでいる情報信憑性判断支援のための要約タスクは，着目言明をクエリとした文書検索を行う前提であるため，検索された文書は query-biased である． query-biased な文書要約では，積極的に他と違う内容や単語を探しにいった方が網羅できると考えた．それゆえ，高村らとは異なる文間評価尺度を用いた．文 S_i の文 S_j に対する代替不可能度 W_{ij}^U は次の式で計算する．

$$W_{ij}^U = \frac{\mid \{w_k \mid w_k \in S_i \& w_k \notin S_j\} \mid}{\mid S_j \mid} \qquad (4)$$

代替不可能度によるリンクの重みは類似度とは異なり有向であることに注意されたい．表 1のような語 w_n を含む文 S_i が与えられたとき，代替不可能度は図 1(b) のような値になる．

また，代替不可能度を用いた場合，頂点から出ていく値と入ってくる値の差が 0 にならないため，各頂点が持つ重要度は式 (3) の計算だけでは発散してしまう．そこで，各頂点の重要度をグラフ全体の頂点の重要度の和で除することで正規化している．

表 2: 各トピックの全文数および正解文数

トピック	全文数	正解文数
女性専用車両	780	170
メタボリック[3]	392	69
マイナスイオン	743	198
還元水	651	104
裁判員制度	504	166

4 評価実験

代替不可能度の有効性を確認するために実験を行った．

4.1 実験に用いるデータ

実験では，中野ら [11] が構築したコーパスを用いた．中野らはいくつかのトピックに対して Web 文書を収集し，人手で情報信憑性判断のための抜粋要約を作成した．実験では，女性専用車両，メタボリックシンドローム，マイナスイオン，還元水，裁判員制度の 5 トピックについての文書集合と要約を用いた．それぞれのトピックには 3 人また 4 人が作業を行い，作業者が要約を作成するにあたって重要だと判断した文が記録されている．本実験では，作業者全員が重要だと判断した文を正解として用いた．表 2はそれぞれのトピックにおける全文数および正解文の数である．

4.2 比較する手法

我々は，比較対象として，二つのベースライン手法を準備した．一つは類似度を用いる元の TextRank の手法である．もう一つは，代替可能度を用いた手法である．代替可能度とは，代替不可能度とは逆に，相手の情報をどれだけ含んでいるかを表す尺度である．式 (4) の代わりに以下の式を用いる．

$$W_{ij}^F = \frac{\mid \{w_k \mid w_k \in S_i \& w_k \in S_j\} \mid}{\mid S_i \mid} \qquad (5)$$

この式は Rus ら [7] が文間の含意関係認識においてベースラインとして用いた尺度と同じであり，高村ら [8] も文間係数として用いている．表 1のような語 w_n を含む文 S_i が与えられたとき，代替可能度は図 1(c) のような値になる．

4.3 実験方法

提案手法及び 4.2節で示した比較手法について， 4.1節で示した文書集合を用いて文のランキングを行った．評価には R 精度を用いた． R 精度とは，出力された重要度ランキングの上位から正解の数だけ文を抽出し，それ

表 3: それぞれの手法における文抽出の R 精度

トピック	類似度	代替可能度	代替不可能度
女性専用車両	0.529	0.488	0.541
メタボリック[3]	0.594	0.652	0.580
マイナスイオン	0.465	0.409	0.556
還元水	0.356	0.288	0.356
裁判員制度	0.452	0.349	0.530

らの中に含まれた正解数の割合によって順位づけの質を評価するものである.

4.4 実験結果

表 3に実験の結果を示す. 代替不可能度を用いた手法がメタボリックシンドローム以外のトピックにおいて他の手法よりも同等かそれ以上の精度で重要文を抽出することができている. 代替不可能度を用いたとき, メタボリックシンドロームのトピックにおいて, 他の手法より R 精度が低くなったのは, トピック語に多様性がなかったためである. つまり, 他と類似している重要文が他のトピックよりも多かったためである. 一方で, このようなトピック語に多様性がないトピックでは, 代替可能度を用いた手法が最も精度が良いことから, 代替可能度と不可能度を組み合わせることにより, このようなトピックにおいても高精度で重要文を抽出できると考えられる. 代替可能度と不可能度を組み合わせる手法の検討については今後の課題である.

5 まとめ

本稿では, 情報信憑性判断を支援する要約を生成する過程において, 通常の要約で注目される主要トピックに加えて, 関連するトピックにまつわる語を網羅する文抽出の手法について検討を行った. TextRank における文間関係尺度として, ある文が相手の文にない情報をどれだけ含んでいるかを表す尺度である代替不可能度を導入した結果, 類似度を用いたベースラインよりも高い精度で, 人手によって重要だと判断された文を抽出できることを確認した.

今後の課題としては, 代替可能度と代替不可能度を組み合わせる手法を検討することである.

謝辞

本研究は, 独立行政法人情報通信研究機構の委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」プロジェクトの成果である.

[3]正式にはメタボリックシンドローム

参考文献

[1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, Vol. 30(1-7), pp. 107–117, 1998.

[2] Aria Haghighi and Lucy Vanderwende. Exploring Content Models for Multi-Document Summarization. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2009)*, pp. 362–370, 2009.

[3] Koichi Kaneko, Hideyuki Shibuki, Masahiro Nakano, Rintaro Miyazaki, Madoka Ishioroshi, and Tatsunori Mori. Mediatory Summary Generation: Summary-Passage Extraction for Information Credibility on the Web. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 23)*, pp. 240–249, 2009.

[4] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP2004)*, pp. 404–411, 2004.

[5] Rintaro Miyazaki, Ryo Momose, Hideyuki Shibuki, and Tatsunori Mori. Using Web Page Layout for Extraction of Sender Names. In *Proceedings of the 3rd International Universal Communication Symposium (IUCS 2009)*, pp. 181–186, 2009.

[6] Koji Murakami, Eric Nichols, Suguru Matsuyoshi, Asuka Sumida, Shouko Masuda, Kentaro Inui, and Yuji Matsumoto. Statement Map: Assisting Information Credibility Analysis by Visualizing Arguments. In *Proceedings of the 3rd Workshop on Information Credibility on the Web (WICOW2009)*, pp. 43–50, 2009.

[7] Vasile Rus, Philip M. McCarthy, Danielle S. McNamara, and Arthur C. Graesser. A Study on Texual Entailment. In *Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligenece (ICTAI'05)*, pp. 326–333, 2005.

[8] 高村大也, 奥村学. 施設配置問題による文書要約のモデル化. 人工知能学会論文誌 Vol.25, No.1, 人工知能学会, 2010.

[9] 渋木英潔, 中野正寛, 宮崎林太郎, 石下円香, 鈴木貴子, 森辰則. 情報信憑性判断のための要約に関する基礎的検討. 言語処理学会 15 回年次大会発表論文集 pp.136–139, 言語処理学会, 2009.

[10] 村上浩司, 増田祥子, 松吉俊, 乾健太郎, 松本裕治. 言明間の意味的関係の体系化とコーパス構築. 言語処理学会 15 回年次大会発表論文集 pp.602–605, 言語処理学会, 2009.

[11] 中野正寛, 渋木英潔, 宮崎林太郎, 石下円香, 森辰則. 情報信憑性判断のための自動要約に向けた人手による要約作成実験とその分析. 自然言語処理研究会報告 2008-NL-187, pp.107–114, 情報処理学会, 2008.

[12] 富田紘平, 高村大也, 奥村学. 重要文抽出と文圧縮を組み合わせた新たな抽出的要約手法. 自然言語処理研究会報告 2009-NL-189, pp.13–20, 情報処理学会, 2009.