人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-05

キーワードのシソーラス上の位置関係にもとづく文章の話題の推敲支援

Polishment of Document Topic by Keyword Relationship in a Thesaurus

大野 祐樹 [*1]　　砂山 渡 [*1]
Yuuki Ohno　　Wataru Sunayama

[1] 広島市立大学 情報科学部
Faculty of Information Sciences, Hiroshima City University*

Abstract: 文章の推敲支援の多くは，表層的な修正を促すものが多く，文章の主題や主題に関連する話題の吟味を促すものはあまり見られない．そこで本研究では，文章からキーワード（文章の主題および話題を表す単語）を抽出した上で，それらのシソーラス上の意味のつながりをもとに，文章内で述べられている内容を推敲するための指標を計算して利用者に提示し，利用者の推敲を促すシステムを提案する．

1　はじめに

近年，インターネットの普及により Twitter や Facebook 等の SNS サイトやブログを通じて誰でも手軽に情報発信ができるようになった．インターネット上では自分の発信した情報を必ずしも親しい人だけが見ているとは限らない．名前も顔も知らない相手に対して文章だけで誤解なく意図を伝えるためには，文章を見直し改める推敲が必要となる．

近年では，フリーソフトの推敲支援ツール [1] や [2] があり，手軽に文章の推敲を行う事ができるようになった．しかし，従来のツールでは文章の文法間違いや適切ではない単語を見つけるなどの表面的な推敲はできても，自分が主に述べたいことを見直す話題自体の見直しや推敲を行う事はできない．その要因として，自分の主張が文章中でどのようなキーワードとして出現しているかわからないことや，それらをどのように推敲すれば，読み手に意図が伝わりやすくなるのか分からないということが挙げられる．

そこで本研究では文章中に現れる筆者の主張を特徴づける単語をキーワードとして抜き出した上で，それら抽出したキーワード集合が，筆者の期待する話題としてふさわしいかを検証するための指標を提示し，話題の推敲を支援するシステムの構築を目指す．本稿では，具体的な話題の推敲支援システムの構築に向けて策定した指標と，その有効性について検証した結果について述べる．

*(連絡先) 砂山渡, 731-3194, 広島市安佐南区大塚東 3-4-1, 広島市立大学大学院情報科学研究科, sunayama@hiroshima-cu.ac.jp

2　関連研究

文章の推敲を支援するための研究は，PC があまり普及していない時代から行われてきており [3]．文法の正しさや誤字の検出など，表層的な指摘を行った上で文章の体裁を整える支援をする研究や，修正を促すシステムはこれまでに開発されてきている [4]．また表層的な表現に加え，談話レベルでの推敲を促す研究 [5] もある．これらの研究とは，文章の特徴を抽出し推敲支援を行う点で類似しているが，本研究では，表層的な表現，また言い回しなどの表現の推敲ではなく，表現のおおもととなる話題の推敲を扱う点で異なる．

表層的な表現の推敲に加え，キーワードを抽出し文章の特徴づけにより推敲の支援を行う研究 [6] もあり，キーワードを抽出することで，文章の特徴を捉えるという点で類似している．本研究では，抽出した各キーワードの位置づけに基づいて，それらをどのように扱うべきかの指標を与える点で異なる．

また，話題に対して指針を与える研究 [7] もあり，文章の中から話題を抽出し話題に対する推敲を支援する点で類似している．しかし，与える指針はふさわしくない話題の削除を促すものとなっており，本研究では，ふさわしくない話題に対してだけでなく，良い話題を広げる指針としての活用も期待できる．

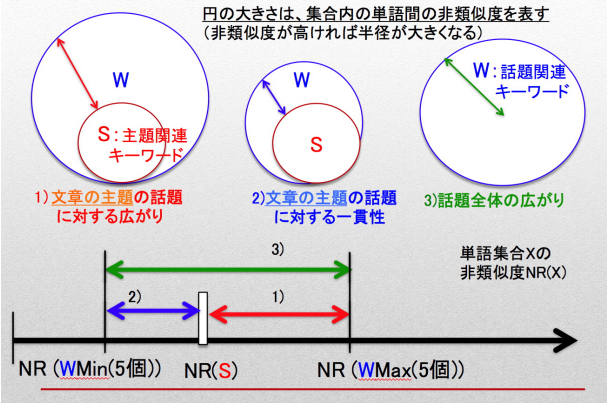

図 1: 話題推敲のための指標とキーワードとの関係

3 文章の話題の推敲支援のための指標

文章の話題の推敲に役立てられる指標として，以下の 3 つを用意する．

1 文章の主題の話題に対する広がり

2 文章の主題の話題に対する一貫性

3 話題全体の広がり

1 は，文章の主題に関連した話題が幅広く述べられているかを確認するため，2 は，文章の主題に関連した話題のみでまとめられているかを確認するため，3 は，文章が話題の全体が幅広い内容を取り扱っているかを確認するための指標となる．

これらの指標の計算のために，以下の 3 種類のキーワードを文章から抽出する．

a 主題キーワード：文章のテーマとなるキーワード

b 主題関連キーワード：文章に現れる筆者の主張を表すキーワード

c 話題関連キーワード:文章の特徴を表すキーワード

この 3 種類のキーワードは，文章内の名詞の出現頻度を用いて抽出する．これは，文章の主題、話題に関わる単語ほど多く出現し，また読み手にそう解釈される可能性が高いと考えたことによる．すなわち,「主題キーワード」は再頻出語,「主題関連キーワード」は頻度上位 5 単語,「話題関連キーワード」は頻度上位 10 単語とする．

1 に，文章の話題の推敲に役立てられる指標とキーワードとの関係を表した図を示す．すなわち，単語集合が与えられたときに，その単語集合内の単語がどれだけ似ているか似ていないかを表す非類似度を定義し，非類似度が高いほど，その単語集合が表す話題の広がりが大きいと考える．その上で，話題の広がりや一貫性を表す指標を計算する．以下で，この各指標の計算方法について述べる．

3.1 単語集合の非類似度

単語集合 X の非類似度 $NR(X)$ は，シソーラス（ある概念に沿って上位-下位のリンクでつながれた木構造のデータベース）内の単語集合の位置関係に基づいて計算する．本研究ではシソーラスに日本語 WordNet[8] を用い，その中の上位語と下位語のリンクでつながれた，全ての名詞のシソーラス内の位置情報を利用することとした．

まず，単語 w の深さ Depth(w) を，ルートノードから単語 w のノードまでの階層（リンク）の数，単語 w の高さ高さ Height(w) を，単語 w のボトムノード (w_b) からの階層の数として，式 (1) で表す．

$$Height(w) = Depth(w_b) - Depth(w) \qquad (1)$$

ただし，ノードはシソーラス上の一つの単語，ルートノードはシソーラス内の最上位語，ボトムノード w_b はシソーラス内で一番深い（ルートノードから最も遠い）ノードとする．

次に単語の非類似度について，単語 $W = \{w_1, w_2, ..., w_n\}$ の非類似度 $NR(W)$ を，式 (2) で与える．すなわち，式 (3) で表される単語間の相対的な距離の遠さと，式 (3) で表されるシソーラスの構造によるお互いの類似性の積によって，単語集合の非類似度を表す．式 (3) は，各単語がシソーラス内で深い位置にあるほどお互いの相対的な距離が遠くなることにより定めた．また式 (3) の sh_k は，シソーラス内で各単語をリンクに沿って上位にたどったときに，各単語がシソーラス内で交わるノードの高さを表す（n 個の単語があったとき，それらはシソーラス上で最大 $n-1$ 箇所で交わる）．これにより，シソーラス内で各単語が上の方で交わるほど，お互いの類似性が低いと考えたことにより定めた．

$$NR(W) = RD(W) \times RH(W) \qquad (2)$$

$$RD(W) = \prod_{i=1}^{n} Depth(w_i) \qquad (3)$$

$$RH(W) = \prod_{k=1}^{n-1} sh_k \qquad (4)$$

3.2 文章推敲のための指標

1 に示す 3 つの指標を非類似度を用いて計算する方法について述べる．

3.2.1 文章の主題の話題に対する広がり

文章の主題の話題に対する広がりの指標 C_1 は，主題関連キーワード集合 S の非類似度 $NR(S)$ と，話題関連キーワード集合 W の部分集合として作られる 5 個の単語による非類似度のうち，非類似度が最大になる単語の部分集合 W_{max} の非類似度 $NR(W_{max})$ の差として式 (5) で与える．これにより，主題を表す単語に対して，どの程度広い話題が取り扱われているかを測ることで，主題に関連してどの程度話題が広げられているかを確認できる．

$$C_1 = NR(W_{max}) - NR(S) \tag{5}$$

3.2.2 文章の主題の話題に対する広がり

文章の主題の話題に対する一貫性の指標 C_2 は，主題関連キーワード集合 S の非類似度 $NR(S)$ と，話題関連キーワード集合 W の部分集合として作られる 5 個の単語による非類似度のうち，非類似度が最小になる単語の部分集合 W_{min} の非類似度 $NR(W_{min})$ の差として式 (6) 計算する．これにより，主題を表す単語に対して，取り扱われている話題の狭さ，すなわち一貫性の程度を確認できる．

$$C_2 = NR(S) - NR(W_{min}) \tag{6}$$

3.2.3 話題全体の広がり

話題全体の広がりの指標 C_3 は，先の述べた非類似度 $NR(W_{max})$ と非類似度 $NR(W_{min})$ との差として式 (7) で与える．すなわち，やみくもに単語の類似性がないことを話題の広がりと呼ぶのではなく，話題集合の中でも，核となる類似性が高い単語集合の非類似度 $NR(W_{min})$ に対して，どの程度話題が広げられているかを図る．

$$C_3 = NR(W_{max}) - NR(W_{min}) \tag{7}$$

4 文章の話題の推敲支援に用いる指標の有意性検証実験

4.1 単語集合内の単語の非類似度とストーリーとの関係の調査

5 個の主題関連キーワード集合 S 内の単語間の非類似度と，文章のストーリーの想像のしやすさとの関係を調査する実験を行った．実験は，40 人の男女に対して，類似性のパターンが異なる 7 種類のキーワード集合 3 セットに対して，文章のストーリーを想像しやすい順に並べてもらった．用意したパターンの単語の類似性について，シソーラスとして用いている WordNet 内において，共通の親ノードをもつ単語間には類似性がある，またそうでない単語には類似性がないとして，単語のパターンを生成した．実験に用いた単語の例を 1 に示す．

表 1: 用意したキーワード集合と類似性パターンの例 (括弧でまとられた単語間には類似性がある)

パターン	使用したキーワード
5	(戦争，敵，反逆，知恵比べ，縄張り争い)
1,1,1,1,1	(戦争)(選挙)(出席)(仲裁)(同盟)
2,2,1	(戦争，敵)(選挙，市長)(出席)
3,1,1	(戦争，敵，反逆)(選挙)(出席)
4,1	(戦争，敵，反逆，知恵比べ)(選挙)
2,1,1,1	(戦争，敵)(選挙)(出席)(仲裁)
3,2	(戦争，敵，反逆)(選挙，市長)

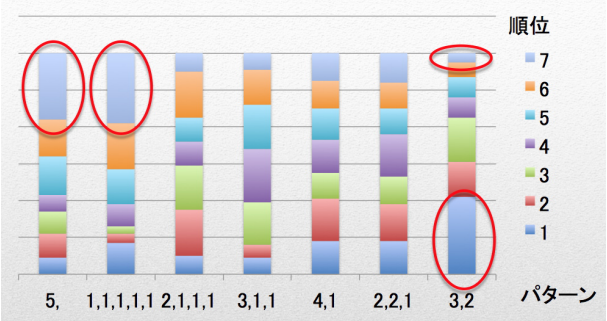

図 2: キーワードの類似度とストーリーの想像のしやすさ

図 2 に，キーワードの類似度とストーリーの想像のしやすさの関係の結果を示す．この結果から，類義語がパターン (3,2)，(2,2,1) や (4,1) のように，バランスよく含まれているほどストーリーを想像しやすかったことがわかる．特にパターン (5) のように，すべての単語が類似している場合，単語が表す範囲が狭すぎて，ストーリーが想像しにくい，またパターン (1,1,1,1,1) のように，すべての単語が類似していない場合は，単語間の関連によるストーリーが想像しにくくなったと考えられる．そのため，単語集合には一定の類似性と非類似性を併せ持つことが，ストーリーの想像には有効となることがわかった．このことから，本研究で提案した主題関連キーワード集合 S の非類似度 $NR(S)$ が，$NR(W_{max})$ と $NR(W_{min})$ の中間の値に近い，すなわち C_1 と C_2 の値が近いほど，文章のストーリーがわかりやすいと考えられ，これらをそのための指標として用いられる可能性を確認した．

表 2: 実験に用いた文章の主題キーワードと W_{max} と W_{min} との関係

文章	W_{max} が含む	W_{min} が含む
1	○	×
2	×	○
3	○	○
4	×	×
5	×	○
6	○	○
7	×	×
8	○	×

表 3: 話題の広がりの評価結果

文章	W_{max} が含む	評価の平均
6	○	7.1
3	○	7.0
1	○	6.9
8	○	6.4
5	×	5.5
2	×	5.3
4	×	5.1
7	×	4.3

4.2 話題の幅広さと一貫性の指標の検証

主題キーワードが単語集合 W_{max} に含まれていると，主題に関連した話題が幅広く述べられていると言えるか，また主題キーワードが W_{min} に含まれていると，主題に関連した話題のみでまとめられていると言えるかを検証する実験を行った．実験は 40 人の男女に 800 字程度の 8 つの文章を被験者に読んでもらい，各文章について,「文章の話題の広がり」と「文章の一貫性」を 10 段階で評価してもらうことで行った．用意した各文章の W_{max} と W_{min} が，主題キーワードを含むか否かについてまとめたものを表 2 に示す．

表 3 に話題の広がりの結果を示す．主題キーワードが W_{max} に含まれていると文章の話題の広がりの評価結果は高くなった．このことから，話題の広がりを大きくしたい時は、W_{max} に主題キーワードが含まれるように修正を促すことができると考えられる．

表 4 に話題の広がりの結果を示す．主題キーワードが W_{min} に含まれていると文章の一貫性の評価結果は高くなった．このことから，文章に一貫性をもたせたい時は，W_{min} に主題キーワードが含まれるように修正を促すことができると考えられる．

5 おわりに

本稿では，キーワードのシソーラス上の位置関係にもとづいて，文章の話題の推敲に用いられる指標を提案した．評価実験により，指標を推敲に有効に役立てられる可能性を検証した．

今後は，話題の推敲支援システムとして具体的に実装と評価を行っていきたい．

表 4: 話題の一貫性の評価結果

文章	W_{min} が含む	評価の平均
6	○	7.9
5	○	7.1
2	○	7.1
3	○	6.4
7	×	5.4
1	×	5.4
8	×	5.3
4	×	4.6

参考文献

[1] 日本語小論文評価採点システム Jess：(URL) http://coca.rd.dnc.ac.jp/jess (2015/3/6 access)

[2] 森リン，日本語の文章解析ソフト：(URL) http://www.mori7.info/moririn/ (2015/3/6 access)

[3] 倉田昌典，菅沼明，牛島和夫：日本語文章推敲支援ツール『推敲』のパソコン上での実用化，コンピュータソフトウェア，Vol.6, No.4, pp.373-385 (1989)

[4] 奥村有希，大野博之，稲積宏誠：技術文章作成支援ツールの推敲支援機能の拡張?長い修飾節に起因する悪文の検出手法の提案，教育システム情報学会研究報告，Vol.22, No.6, pp.186–191 (2008)

[5] 飯田龍，徳永健伸：談話レベルの推敲支援のための人手修正基準，言語処理学会第 19 回年次大会発表論文集，pp.830–833 (2013)

[6] 石岡恒憲．亀田 雅之：コンピュータによる小論文の自動採点システム Jess の試作，計算機統計学，Vol.16, No.1, pp.3–19 (2003)

[7] 菅沼明，小野貴博：文章推敲支援における読み手に誤解される文の抽出，情報処理学会研究報告，DD, Vol.2007, No.50, pp.31–38 (2007)

[8] 日本語 WordNet：(URL) http://nlpwww.nict.go.jp/wn-ja/ (2015/3/6 access)

[9] 砂山渡，高間康史，西原陽子，梶並知記，串間宗夫，徳永秀和：統合環境 TETDM を用いたマイニングツールの開発と利用の実践, 人工知能学会論文誌，Vol.29, No.1, pp.100 – 112 (2014)