語彙的連鎖を用いた調停要約生成手法の提案

朱 丹† 渋木 英潔‡ 森 辰則‡

†横浜国立大学 大学院 環境情報学府 ‡横浜国立大学 大学院 環境情報研究院
E-mail: {gladys,shib,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．このような考えから，我々は，利用者の信憑性判断を支援するための技術の一つとして，調停要約に関する研究を行っている [1, 2, 3].

調停要約とは，一見対立しているようにみえるが，実際はある条件や状況の下で互いの内容が両立できる場合に，両立できるようになる条件や状況を分かりやすく提示する簡潔な文章のことである．中野ら [1] は，利用者が信憑性を判断したい言明[1]（以降，**着目言明**）をクエリとして検索した Web 文書集合から，調停要約として適切な内容を含む文の連続を 1 つのパッセージとして抽出することで調停要約を自動的に生成する手法を提案している．中野らの手法では，調停要約としての尤もらしさのスコアを各文に付与した後，各文のスコアを平滑化し，平滑化されたスコアが閾値以上となる範囲の文の連続をパッセージとして抽出する．しかしながら，平滑化されたスコアに基づいてパッセージの範囲を決定する方法は状況的な推論に基づいた方法であり，意味的なまとまりに基づいて判断しているわけではない．それゆえ，意味的なまとまりがない文群が 1 つのパッセージとして切り出されてしまい，結果として，調停要約として不適切な文を含むパッセージを出力してしまう場合があった．渋木ら [2] や石下ら [3] などにおいて，中野らの手法を改善した調停要約生成手法が提案されているが，パッセージを切り出す方法に関しては中野らの手法と同一であり，パッセージの範囲に関する問題は改善されていない．我々は，意味的なまとまりを考慮したパッセージ抽出を行うことで，この問題を改善できると考えた．

意味的なまとまりを考慮するための方法として，語彙的連鎖による語彙的結束性を用いて判断する方法がある．語彙的連鎖とは，文書中で互いに意味的な関係を持つタームの連続のことであり，望月ら [4] は語彙的連鎖を用いてパッセージ検索を行う手法を提案している．望月らは，クエリ中の単語に基づいて語彙的連鎖を作成し，語彙的連鎖の範囲を切り出すことで意味的なまとまりがあるパッセージを抽出しており，我々は，このパッセージ抽出方法を調停要約生成に応用する．しかしながら，調停要約生成では，クエリとの関連性に加えて，両立できるようになる条件や状況などに関する記述を含むパッセージを出力するため，パッセージ検索とは異なった，語彙的連鎖の作成や利用を行う必要がある．以上の背景から，本稿では，語彙的連鎖を用いた調停要約生成手法を提案する．

本稿の構成は以下の通りである．2 章では，調停要約生成における語彙的連鎖の種類およびパッセージの定義について述べる．3 章では，語彙的連鎖を用いた調停要約の生成方法について述べる．4 章では，調停要約コーパスを用いた実験を行い，結果の考察を行う．5 章は，まとめである．

2 調停要約生成における語彙的連鎖

2.1 語彙的連鎖の種類

中野らの調停要約生成手法 [1] では，着目言明のトピックとの関連性を計算するためのトピック特徴語，着目言明を肯定する意見であるかを計算するための肯定側特徴語，着目言明を否定する意見であるかを計算するための否定側特徴語の 3 種類の特徴語を定義し，着目言明および Web 文書から抽出された，これらの特

[1]本論文では，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**と呼ぶこととする．

表 1: 定型表現の一覧

すなわち, たとえば, 例えば, だって, つまり, なぜだ, 何故だ, 要するに, ようするに, いいかえれば, 言い換えれば, 言いかえれば, いう換えれば, 換言すれば, 一口にいえば, 一口に言えば, いわば, 言わば, 結局, 言ってみれば, いってみれば, 言って見れば, したがって, 従って, すると, そうすると, そうなると, そこで, それだから, それゆえ, それ故, それで, だから, ますから, だとすると, ついては, ですから, ゆえに, 故に, よって, それゆえに, それ故に, おかげで, そのために, それだけに, その結果, この結果, わけです, わけで, けれども, けれど, されど, しかしながら, しかし, しかるに, 然るに, それなのに, それでも, たほう, 一方, 他方, だが, だけど, だけれども, だのに, ですが, ですけれど, でも, ところが, どころか, 反面, はんめん, これに反し, これに対し, それどころか, そればかりでなく, 但し, ところで

徴語を用いて調停要約の生成を行った．また，渋木ら [2] は，パッセージ中の対比構造を把握するために「しかし」などの逆接表現を利用することで調停要約生成の精度が向上したことを示した．

本手法では，これらの従来研究を参考に，**トピック連鎖**，**肯定側連鎖**，**否定側連鎖**，**特殊表現連鎖**の 4 種類の語彙的連鎖を設定する．トピック連鎖は，着目言明のトピックに関する意味的なまとまりを示すための語彙的連鎖であり，例えば,「ディーゼル車は環境に良い」という着目言明であれば,「ディーゼル車」と「環境」に基づいてトピック連鎖が作成される．肯定側連鎖と否定側連鎖は，それぞれ，着目言明を肯定または否定する意見に関する意味的なまとまりを示すための語彙的連鎖であり，ディーゼル車の例では，肯定側連鎖は「良い」に基づいて，否定側連鎖は「良い」の対義語である「悪い」に基づいて作成される．特殊表現連鎖は，逆接や限定などの文章構造を把握するための表現を示すための語彙的連鎖であり，着目言明に関係なく，表 1 に示す事前に設定された定型表現に基づいて作成される．

2.2 パッセージの定義

望月らの手法 [4] では，文書中の各語彙的連鎖の内，出現位置に重なりのある連鎖同士を 1 つのパッセージとしてマージを繰り返していき，パッセージ内の最初の語彙的連鎖が始まるタームから最後の語彙的連鎖が終了するタームまでを 1 つのパッセージとしている．しかしながら，報知的要約の一種である調停要約生成では，抽出されたパッセージのみで内容を完全に伝える必要があるため，文境界をパッセージ境界とする必要がある．それゆえ，本手法では，最初の語彙的連鎖が始まるタームを含む文から最後の語彙的連鎖が終了するタームを含む文までを 1 つのパッセージとした．

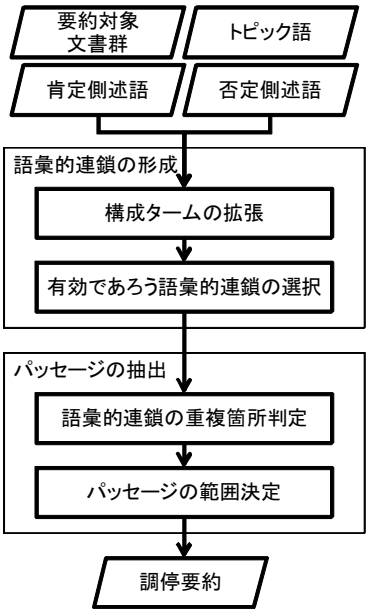

図 1: 全体の流れ

3 語彙的連鎖を用いた調停要約手法

本手法の全体の流れを図 1 に示す．入力として，要約対象となる文書群に加えて，着目言明の対立関係を表すための**トピック語**，**肯定側述語**，**否定側述語**を与える．着目言明が「ディーゼル車は環境に良い」であれば，トピック語は「ディーゼル車」と「環境」，肯定側述語は「良い」，否定側述語は「悪い」となる[2]．

入力されたトピック語，肯定側述語，否定側述語に基づいて，トピック連鎖，肯定側連鎖，否定側連鎖をそれぞれ形成する．望月らの手法 [4] では，同一タームの反復，シソーラス上の同一概念に属するタームの連続，共起しやすいタームの連続の 3 通りの基準により語彙的連鎖を形成している．本手法でも，この基準に倣い，入力された単語のみ，シソーラスによる類義語，Web 文書からの特徴語の 3 通りのタームの拡張を行い，拡張されたそれぞれのタームにより 3 通りの語彙的連鎖を形成した．Web 文書からの特徴語とは，着目言明またはその対立言明をクエリとして検索された Web 文書集合における出現頻度に基づいて抽出された特徴語であり，中野ら [1] における肯定側特徴語と否定側特徴語の抽出方法と同一の方法で抽出を行った[3]．4 章の実験では，タームの拡張方法による精度への影響を調査する．特殊表現連鎖は，表 1 に示す表現をタームとして形成する．

[2]本稿では,「よい」,「いい」などの異表記も入力として与えている．

[3]中野らの手法ではトピック特徴語の拡張は行っていない．そのため，本手法でもトピック連鎖のタームの拡張は行わなかった．また，シソーラスによる類義語を用いる場合，トピック連鎖のタームを拡張することが可能であるが，Web 文書からの特徴語を用いた場合との比較を行うため，シソーラスによる類義語を用いる場合も肯定側連鎖と否定側連鎖のタームに対してのみ拡張を行った．

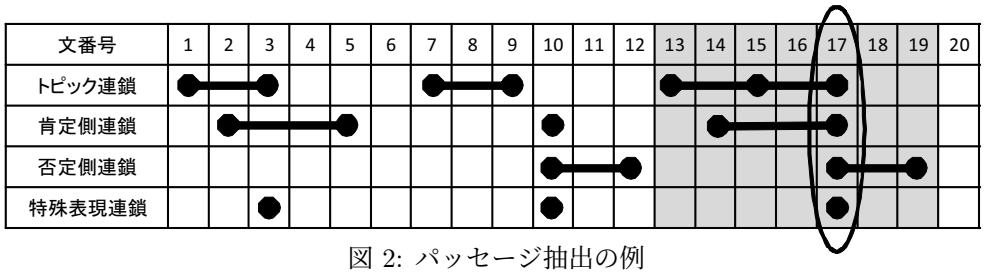

図 2: パッセージ抽出の例

形成された語彙的連鎖から有効であろう連鎖を選択するために，望月らの手法では，語彙的連鎖を構成するタームの出現位置が互いに離れている場合，その連鎖は有効な連鎖ではないだろうとして，離れている区間（ギャップ）で連鎖を切り離して別の連鎖としている．本手法でも，同様に，閾値以上のギャップがある連鎖は分割して別の連鎖とした．4 章の実験では，ギャップの閾値を変化させた場合の精度への影響を調査する．

調停要約生成の場合には，着目言明のトピックとの関連性に加えて，着目言明を肯定する意見と否定する意見の両方に公平に言及しているパッセージを抽出する必要がある．中野ら [1] や渋木ら [2] の手法では，トピック特徴語，肯定側特徴語，否定側特徴語の全ての種類の特徴語を含んでいる文を，調停要約として適切な文として高いスコアを与えている．このことから，トピック連鎖，肯定側連鎖，否定側連鎖，特殊表現連鎖の 4 種類の連鎖の出現位置が重なっている文が調停要約として適切な文であるとした．

望月らの手法では，クエリと強く関連するパッセージを計算するための条件の一つとして，各語彙的連鎖の出現位置の重複部分の多さを考慮しており，タームが存在しなくとも有効であろう連鎖が存在していれば重複していると判断していた．しかしながら，調停要約生成においては，明確に調停している文を発見した後に，その文の内容と意味的にまとまる文の範囲を決定した方が精度が高くなると考えられる．それゆえ，調停要約として適切な文を判定する際には，タームの出現位置が重なっているかどうかで判定することとし，タームの出現位置が重なっている文を起点に，そのタームを含む有効であろう連鎖が続く範囲を 1 つのパッセージとした．

図 2 にパッセージ抽出の例を示す．図中の黒丸はタームが出現していることを表し，黒丸間をつなぐ線は有効であろう連鎖が存在していることを表している．図 2 の例では，17 文目に 4 種類の連鎖のタームが全て出現しており，17 文目を起点として有効であろう連鎖が続く範囲である 13 文目から 19 文目までが 1 つのパッセージとして抽出される．タームの出現位置の重複判定を行った後にパッセージの範囲を決定するという流れが，従来の語彙的連鎖を用いたパッセージ抽出と大きく異なる点である．

タームの出現位置の重複により判定する場合，4 種類の連鎖のタームが 1 文内に全て出現することは稀であると考えられるため，本手法では，3 種類または 2 種類の連鎖におけるタームの出現位置の重複により近似する．しかしながら，最低限，着目言明のトピックとの関連性は保証されるべきであるため，重複判定対象の文には必ずトピック連鎖のタームが出現していることという制約を設けることとした．4 章の実験では，タームが重複する文を起点とすることの有効性と，重複判定に必要な連鎖の種類数を変化させた場合の精度への影響を調査する．

4 実験

本実験の目的は，語彙的連鎖を用いた調停要約生成において，タームが重複した文を起点とすることの有効性の確認と，重複判定に必要な連鎖の種類数，ギャップの閾値，タームの拡張方法をそれぞれ変化させた場合の精度への影響を調査することである．タームが重複した文を起点とすることの有効性を確認するため，タームが存在しなくとも有効であろう連鎖が存在していれば重複していると判断してパッセージ抽出を行う望月ら [4] の手法をベースライン手法として提案手法との比較を行う．

本実験では，調停要約コーパス [2] に収録された Web 文書集合を用いた．収録 Web 文書集合の内,「飲酒は健康に良い」,「炭酸飲料はからだに悪い」,「原発は地震でも安全である」の 3 つの着目言明に関する調停要約が記述されている文書群を要約対象文書群とした．要約対象文書群は,「飲酒は健康に良い」64 文書,「炭酸飲料はからだに悪い」15 文書,「原発は地震でも安全である」34 文書の計 113 文書となった．調停要約コーパスでは，1 つの着目言明に 4 名の作業者による調停要約が収録されており，文ごとに何名が調停要約として適切であると判断したかを計算することができる．したがって，調停要約として適切であると 1 人以上が判断した文を正解文とし，判断した人数をその文の重みとした．ただし，不正解の文の重みに関しては，不適切と判断した人数が 4 人であるとみなし，重みを 4 と

表 2: 実験結果 (着目言明ごとの F 値の平均値)

	ギャップ	入力された単語のみ			シソーラスによる類義語			Web 文書からの特徴語		
		2 種類	3 種類	4 種類	2 種類	3 種類	4 種類	2 種類	3 種類	4 種類
提案手法	5	0.284	0.153	0.000	0.281	0.209	0.023	0.213	0.221	0.073
	4	**0.322**	0.144	0.000	0.317	0.211	0.018	0.230	0.238	0.082
	3	0.291	0.129	0.000	0.292	0.190	0.013	0.248	0.253	0.077
	2	0.279	0.120	0.000	0.283	0.180	0.010	0.262	0.269	0.066
ベースライン	5	0.282	0.214	0.000	0.278	0.287	0.023	0.141	0.151	0.145
	4	0.319	0.206	0.000	0.314	0.274	0.018	0.176	0.195	0.127
	3	0.291	0.136	0.000	0.292	0.194	0.013	0.209	0.227	0.081
	2	0.279	0.120	0.000	0.283	0.180	0.010	0.262	0.269	0.066

した．すなわち，適切と判断した人数が 1〜4 人の時は重みは 1〜4，0 人の時は重みは 4 となる．

評価指標として，適合率 P，再現率 R，F 値を用い，それぞれ以下の式に従って計算した．

$$\mathrm{P} = \frac{\sum_{i \in S} w(i) answer(i) output(i)}{\sum_{i \in S} w(i) output(i)} \quad (1)$$

$$\mathrm{R} = \frac{\sum_{i \in S} answer(i) output(i)}{\sum_{i \in S} answer(i)} \quad (2)$$

$$\text{F 値} = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}} \quad (3)$$

S は要約対象文書中の全ての文，$w(i)$ は i 番目の文の重みである．また，$answer(i)$ は i 番目の文が正解文である場合に 1，そうでなければ 0 を返す関数，$output(i)$ は i 番目の文が出力されたパッセージ中に含まれている場合に 1，そうでなければ 0 を返す関数である．不正解の文を出力した場合，適合率の計算において分母に不正解の文の重みが加算されることになるため，不正解の文の重みがペナルティとして作用する．本稿では，不正解の文の重みを 4 としており，最も厳しいペナルティの下で評価している．

名詞，動詞，形容詞を内容語と定義し，各着目言明の形態素解析結果から述部の内容語を肯定側述語，述部以外の内容語をトピック語とした．また，対義語辞書を用いて，肯定側述語の対義語を否定側述語とした．語彙的連鎖のタームの拡張に用いるシソーラスには角川類語新辞典 [5] を用いた．

表 2 に結果を示す．紙面の都合により 3 つの着目言明における F 値の平均値のみを示す．太字の数字が全ての条件での最大値を示し，下線の数値がタームの拡張方法ごとの最大値を示している．全体の傾向として，同一の条件であれば，提案手法の方がベースライン手法よりも高い値を示している．このことから，調停要約生成においては，タームが重複した文を起点とすることが有効であるといえる．条件を変化させた場合の傾向として，入力された単語のみ，シソーラスによる類義語，Web 文書からの特徴語の順で，タームの数が増加し，再現率の上昇と適合率の低下が見られた．また，重複判定に必要な連鎖の種類数が少なく，ギャップの閾値が大きいほど，再現率の上昇と適合率の低下が見られた．結果として，表 2 に示されるように，シソーラスによる類義語で拡張したタームで語彙的連鎖を形成し，ギャップを 4 として有効であろう連鎖を選択し，2 種類の連鎖による重複判定を行った場合が最も高い値を示すこととなった．

5 まとめ

本稿では，語彙的連鎖を用いた調停要約生成手法を提案した．本手法では，従来の語彙的連鎖によるパッセージ抽出手法と異なり，タームの出現位置が重なっている文を起点に，そのタームを含む有効であろう連鎖が続く範囲を 1 つのパッセージとして抽出する．調停要約コーパスを用いた従来のパッセージ抽出手法との比較実験により，提案手法の有効性を確認した．

参考文献

[1] 中野正寛, 渋木英潔, 宮崎林太郎, 石下円香, 金子浩一, 永井隆広, 森辰則: 情報信憑性判断支援のための直接調停要約生成手法, 電子情報通信学会論文誌 (D), Vol.J94-D, No.11, pp.1019–1030, 2011.

[2] 渋木英潔, 中野正寛, 石下円香, 永井隆広, 森辰則: 調停要約生成手法の改善と調停要約コーパスを用いた評価. 第 10 回情報科学技術フォーラム (FIT 2011) 講演論文集, pp. RE–003, 2011.

[3] 石下円香, 渋木英潔, 中野正寛, 宮崎林太郎, 永井隆広, 森辰則: 直接調停要約自動生成システム HERMeS の言論マップとの連携. 言語処理学会第 17 回年次大会発表論文集, pp. P1–13, 2011.

[4] 望月源, 岩山真, 奥村学: 語彙的連鎖に基づくパッセージ検索, 自然言語処理, Vol.6, No.3, pp.101–126, 1999.

[5] 大野晋, 浜西正人: 角川類語新辞典, 角川書店, 1981.