共起語を介した文間の相互依存関係に基づく重要文の多段階抽出法

四ッ谷 雅輝　　溝江 彰人　　吉岡 真治　　原口 誠

北海道大学大学院工学研究科

1 はじめに

近年の計算機の急速な普及によって，電子メールや電子図書等に加えて，従来，紙に記述されてきた文章などもテキストデータとして計算機で処理することが可能になってきている．膨大になったテキストデータ集合によって，ユーザが本当に欲しい情報が見えにくくなる一方で，これらを正確に分析や検索が可能ならば，ユーザの意図しない有用な知見を獲得できる可能性も考えられる．

電子化された膨大なテキストから，求める情報を得るための技術は，情報検索，テキスト分類，データ加工の三つに大別して考えることができる．情報検索とはユーザの情報要求に従い，求める情報を探すことであるが，検索対象が膨大なテキスト集合であると情報検索が成功することは困難である．そのため，あらかじめ情報を内容や観点，目的に応じてテキストを分類しておくことにより，求める情報を探し出す範疇を限定し，情報を探し出す作業負担を軽減する必要がある．

また，計算機が処理するテキスト量やユーザが読むテキスト量を制御することが求められている．データ加工とは，元の情報をアクセスが容易になるような情報中間物を作ることといえる．具体的には，情報の構造化や情報の要約などが挙げられる．本研究では，この情報の要約に焦点を当てる．

一般にテキストの要約とは，筆者が主張する話題を中心に，原文の大意を保持したまま，テキストの長さ，複雑さを減らす処理である．要約生成は，テキストの内容を理解し，中心的な話題を特定し，それを簡潔にまとめるという三つの作業から成る．テキストの内容の理解には，意味解析，文脈解析といった高度な言語処理が要求され，これらの解析を行ったとしても，現段階において，計算機がテキストを正確に理解できるとは言えない．また，話題を特定したり，簡潔な文を生成するという言語処理においても困難が多い．しかし，その要約の利用目的によっては，重要な情報を伝えている文をテキスト中から抜き出すことによって生成された要約文であっても，その役目を果たすことは可能である．

本研究では，テキストの要約を情報の要約と捉え，対象とするテキストの量を制御することによって，情報検索やテキスト分類を支援することを目的としている．対象とするデータは物語文を想定している．物語文の文書構成は様々であり，新聞記事や科学技術論文の要約に用いられるような手がかり語や位置情報を利用することは困難である．また，物語文の要約に必要とされる文の中のストーリ展開を示す文は，重要度の低い単語で構成されることもあり，単語の重要度の線形和による手法も困難である．そこで，本稿では，重要な文の選定に共起語を介した文間の相互関係から再帰的に文の重要度を決定する方法を提案する．文間の相互関係には，共起語の特徴とその文を構成する単語の出現数や単語が担う文中での役割等の特徴を反映させた強度を持たせる．本手法の特徴は，文の重要度が，「重要な文で使用された語を使用している文は重要である」という観点に立ち，集積と分配を繰り返しながら再帰的に伝播することである．

2 手法

2.1 本研究のアプローチ

人間が重要であると感じる文は，その一文を読んだだけでは決定することはできず，文を読み進めることに従い，文間のつながりの影響を受けながら刻々と変化するものであると考えられる．つまり文の重要度は，その一文を構成する語の情報のみによって決まるものではなく，文同士のつながりにも依存する点があると考えている．つまり，重要な文とつながりのある文は，その文もまた重要な文であるという文の重要度が伝播してゆく関係の下にあると考えられる．さらに，この文間のつながりには，強度が存在し，その強度が強ければ強いほど，重要度の伝播の度合いも強いものと考えられる．本研究では，文間のつながりに共起語を用い，文の重要度のモデル化を PageRank アルゴリズム [PBMW98] を拡張することによって試みる．

2.2 PageRank アルゴリズム

PageRank アルゴリズムとは，Web のリンク構造をページをノード，リンクをエッジとするグラフモデルで展開され，「多くの良質なページからリンクされているページはやはり良質なページである」という概念の下，Web ページの重要度を決定する．PageRank アルゴリズムでは，あるページ q からページ p へのリンクを，ペー

ジ q から p への支持投票とみなし，その票数から重要度を算出する．PageRank アルゴリズムの特徴的な点は，リンク元であるページ q の重要度によって，1票の重みが変わる点である．ページ q が重要度の高いページであったら，1票の重みが増え，それがページ p にも反映され，ページ p は重要度の高いページとなる．逆に，ページ q の重要度が低いページであった場合，1票の重みが抑えられ，ページ p は重要度の低いページとなる．無論，この場合でも，票を得られないページと比べたら，重要度は高くなる．

PageRank は，次式を繰り返し適応することによって求められる．[Hen00]

$$R(p) = \frac{\epsilon}{n} + (1-\epsilon) \cdot \sum_{(q,p)\in G} \frac{R(q)}{outdegree(q)} \qquad (1)$$

$R(p)$ はページ p の PageRank,$R(q)$ はページ p へのリンクを持つページ q の PageRank を示す．n は対象とするグラフ G（Web ページをノードとし，Web ページ間のリンクをエッジとしたグラフ）のノード総数（Web ページ数），$outdegree(q)$ はページ q からの外向きリンク数であり，$\sum_{p\in G} R(p) = 1$ としている．ϵ は通常 0.1 ~0.2 の間に設定され，「ユーザは $(1-\epsilon)$ の確率で現在の Web ページからリンクをたどり，ϵ の確率でまったく無関係な Web ページにジャンプする」というモデル化を行っている．まったく無関係な Web ページにジャンプとは，たどれるリンクが存在せず，ブラウザの機能を用いて，ジャンプ前のページに戻ったりすることも含まれる．PageRank の実計算としては，ページ間の推移確率行列の転置行列の優固有ベクトルを求めることによって，得ることができる．

2.3 PageRank アルゴリズムの拡張

PageRank アルゴリズムを文空間に適応する時，問題になる点を二点挙げる．一点目は，ノード間のリンク・被リンク関係のみに着目し，ノードの内容は考慮していない点が挙げられる．もう一点は，リンク・被リンク関係が存在するか，否かの 2 値 $(0,1)$ しか持たず，どのくらいの相互関係があるのかという強度の概念が考慮されていない点である．文空間に当てはめて考えてみると，個々の文の内容が考慮されておらず，文の重要度を決定するにあたって，無視することのできない情報である．また，相互関係がどのくらいあるのかという，強度の概念も重要な情報である．そこで本研究では，従来，各ノード間にリンク関係があるかないかによって生成されていた推移確率行列に対して，文の内容に関する情報と文の相互関係に強度という情報を反映させた重要度伝播行列を定義する．

要請 1 語の共起する数が多ければ多いほど伝播力は大きく，またその語の重要度が大きければ大きいほど伝播力は大きい．

要請 2 文を構成する語が重要な語であるほど伝播力は大きい．

上記の要請のもと，文 q から文 p への相互関係を次式によって定義する．

$$\alpha_{qp} = \left(\sum_{w_q \in p \cap q} w_q(t) \right) \log \sum_{w \in q} w_q(t) \qquad (2)$$

ただし，$\sum_p \alpha_{ip} = 1$ なる正規化を行う．(2) 右辺第 2 項の log は構成する語が多ければ多いほど伝播力が大きくなることを防ぎ，第 1 項とのスケールバランスをとる関数として用いている．単語の重要度は，要約の目的とテキストのドメインに応じて適切な重み付け手法を選択できる枠組みとしている．本研究における実験では，文 q を構成する単語 t の重み $w_q(t)$ は，

$$w_q(t) = TF(t) \times Case_q(t) \times 1/Depth_q(t) \qquad (3)$$

としている．ただし，$TF(t)$ はテキストにおける単語 t の出現頻度，$Case_q(t)$ はヒューリスティックに定めた格の偏重度，$Depth_q(t)$ は述語からの係り受けの階層の深さである．

ここで，式 (1) を文空間へ適応を考えると，Web 空間では ϵ を規定していたが，文空間の場合，他のどの文とも相互関係を持たないことは少ない．仮にこのような状況が生じた場合は，シソーラスによって，強度を考慮に入れた相互関係を作ることを想定しているので，$\epsilon = 0$ とし，すべての文は他の文と何らかの相互関係を持つモデルを考える．よって，式 (1) の代わりとして，

$$R(p) = \sum_q \alpha_{qp} R(q) \qquad (4)$$

を考えればよく，文の重要度は式 (2) で生成された重要度伝播行列の転置行列の優固有ベクトルを求めることで得ることができる．

2.4 手順

重要文抽出による要約文獲得の処理の大まかな流れは以下のようになる．

1. 自然言語データから要約に用いる格情報や位置情報を抽出する．
2. 重要度伝播行列を作成し，文の重要度を算出する．
3. ユーザの質問式に応じた要約文の提示する．

1. について，本研究では，テキストに書かれた文字列を情報として計算機に格納するため，形態素解析・構文解析を行った．この解析には解析ツールである南瓜 [工藤 02] を利用した．解析ツールの出力結果は，日本語文法に厳格に則っており，要約のための情報としては，非常に詳細である．そこで解析ツールから得られた表層情報から，語の役割を示す格を定義した．方法としては，動詞を中心とした格の役割を考えるとき，助詞が果たす影響は大きいことから，助詞の働きによる分類を行った．表層情報のみから助詞の働きを同定することは困難であるが，重要文抽出の一つの情報としての格であるという

位置付けから助詞の品詞体系によって大まかな助詞の働きを分類できるとした．

2. については，前節で説明した重要度伝播行列の転置行列の優固有値を求めることによって，文の重要度を決定する．

3. については，ユーザは質問式によって，要約の利用目的に応じて要約率を調整できる設計になっている．また，文の重要度の決定に用いる単語の重要度の決定方法についても，対象テキストのドメインやユーザの目的に応じて，組替えることができる設計になっている．本システムでデフォルトとして用いている単語の重要度の式 (3) についても，ユーザが与えたキーワードに対する重み付けも可能であり，また，格の偏重度に対しても変更することができるので，格に対する考察や実験に従って，ユーザの要求を与えることができる．

一般的にテキストの主題は複数ある場合が多く，個々の重要度は読者の観点によって変わってくるものであり，同一読者であっても，状況に応じて観点は異なるものであるとの考えから，ユーザの情報要求がシステムの精度に直接的に繋がるものであると考えられる．

3 実験と考察

3.1 実験方法

本研究で提案する手法の有効性を確かめるため，重要文抽出手法を用いた要約システム Posum[望月 02] と比較実験を行う．本研究と Posum の共通点は，共に共起語情報を用いることであり，Posum では語彙的連鎖として文間のつながりを考慮している．相違点としては，Posum は，その情報を語に反映させ，文の重要度を語の重要度の線形和によって決定しているのに対し，本研究では，文の重要度を集積と分配の再帰的伝播によって決定している点である．文の重要度を決定している．

また，実験データセットとして，物語文を選択した．その理由として，物語文は登場キャラクターのイベントによるストーリ展開が明確であり，単語による文間のつながりの特性が顕著であると考えられるからである．

3.2 評価方法

各々のシステムを評価するためには，正解データセットが必要になる．正解データとして，大学生，大学院生の被験者 7 名に実際に物語文から重要な文を抽出して頂いた．抽出の方法として，物語の前文 46 文から重要な文だと思われる文を 25 文を選んでもらい，その 25 文からさらに，20 文，その 20 文からさらに 15 文と，同様にして 5 文まで選んでいただいた．選ばれた文には一票を与え，その多数決により重要文を決定した．

要約には様々な観点があり，その評価は困難であるが，本研究の評価は，システムによる要約文と正解データセットとの適合率と再現率を要約率毎に調べることによって一つの評価を与える．一つの評価が与えられるとする理由は，本研究の目的が，情報の要約であり，情報の欠損をできるだけ抑えて，テキストのサイズを圧縮することに主眼が置かれているからである．

本研究において適合率と再現率は，要約率毎の正解データセットの文数とシステムによって出力される文数は一致するため，適合率と再現率は一致するので，これを改めて正解率とする．また，要約率とは対象テキストの全文数に対するシステムの出力文数を要約率している．

3.3 結果

実験において求められた，各々のシステムの要約率毎の正解率を図 1 に示す．Posum1 は単語共起のみで語彙的連鎖を結成しており，Posum2 はシソーラスによる類義を含めて語彙的連鎖を結成している．シソーラスには角川類語新辞典が用いられている．Posum の要約率が 50% までなのは，システムの仕様による．システムは，これらの表で与えられた要約率に最も近い文数だけ抽出する．また，票数が同じで複数の正解文が存在し，抽出文数をまたがる場合，両者とも正解文としている．

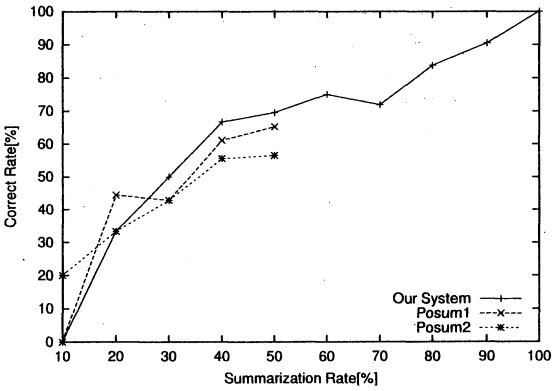

図 1: 物語文の要約率と正解率の関係

3.4 考察

図 1 より本システムは，要約率 10%～20% の区間で Posum2 に劣るものの，他の区間では，よい結果が出ているといえる．情報の要約として，高い要約率よりも信頼できる要約が期待される場面を想定すると実験対象が物語文のみではあるが，有意な結果であるといえる．

正解率の低かった 10%，20%，30% の要約について，正解文と本システムによる要約文の内容の比較を行うと，正解文は少ない文数の中にも物語の起承転結のストーリが展開されていることが分かった．人間の要約モデルは大要を把握し，物語の大まかな流れに沿って文を抽出しており，その抽出の際，冗長性を排除し，ストーリ性を失わない文の取捨選択を行っている．計算機による要約文であるとこの物語の重要であると思われる局部は捉え

ているが，その周辺の文を再び探し出してしまう結果となっていた．つまり，物語の重要であると判定した箇所を意味的冗長性を考慮せず抽出した結果となり，その冗長にスペースを奪われた分，他の情報が書かれたセンテンスを抽出できなくなっていると考えられる．

要約率が緩くなるにつれて，正解データは，ストーリ展開を満足させた後でも，文のゆとりがあるため物語の主人公に対する描写が許され，逆に本システムでは，主要部分から他の情報が書かれた個所を抽出するゆとりが出てきて，正解文の要約文が一致するようになってきたものと考えられる．

以上の考察から重要文抽出の改善策を考える．本システムでは，一つのクラスタから重要だと思われる文集合は，ある程度，目安をつけることができるといえるので，テキストを何らかの基準でいくつかのクラスタに分割しておき，そのクラスタごとに本手法を適応することによって，要約文の質的向上を図ることが考えられる．

このクラスタには，文書を構成する段落を利用することが考えられる．クラスタを形成する段落数は，テキスト量に準じて複数であることも考えられる．そして，このクラスタ毎に本手法を適応することより，文書全体から均一的に文抽出された要約文の生成が可能となる．結果として，要約文の内容が局部的な情報に偏ることを防ぎ，物語のストーリ展開を捉えられる可能性がある．また，この分割統治による戦略は，大規模データに対する計算量抑制の効果も期待できる．

この他にもクラスタの獲得には，文脈解析などを用いた意味的クラスタ分割や出現語の分布によるクラスタ分割などが考えられる．意味的クラスタ分割については，現段階では，計算機が自然言語の意味を理解するということが困難であることから，その近似を行うことになるがその処理自体コストがかかる作業となる．出現語の分布によるクラスタは，語における統計処理のみで実現することができ，意味的クラスタ分割より少ないコストで達成されるが，十分有効であると思われる．

その理由として，本手法では共起語によってネットワークが構成されていて，そのリンクが密な部分ほどスコアが高くなる傾向があるためである．具体的には，スコアは文の重要度に応じて，分配と集積を繰り返すという再帰的伝播をするのだが，文に集積関係が多くあれば，それが低い重要度でも累積することによって無視できなくなるからである．

同等の集積関係という条件の下で文の重要度の再帰的伝播を用いることができれば，本手法はより効力を発揮するものと考えられる．以上の理由から，集積関係を相対的に均一にするクラスタ設定が必要となり，それは語の分布を見るだけで，ある程度達成できるといえる．

重要文抽出の質的向上についての言語処理的な見解は，重要文抽出と意味的重複文削除を交互に繰り返すことによって研ぎ澄まされた重要文の抽出が期待できる．具体的には，システムによって抽出されようとしている文に対して以前に抽出された文との類似度を計算することにより，類似度がある閾値以上であると冗長とみなして排除するプロセスを組み込むことによって実現される．また，本実験よりこの類似度の閾値は要約率とともに緩めていくほうが正解率は上がるものと考えられる．

4 結び

本研究では，文の重要度の再帰的伝播による重要文抽出の方法を提案し，その有効性を物語文における実験によって確かめた．また，実験結果の考察より，精度の改善策として語の分布によるクラスタ分割の提案を行った．

本研究の展望として，文間の相互関係を考えるとき，単語の直接的な共起だけではなく，シソーラスを用いた間接的な共起についての考慮によってシステムの改善が期待できると考えている．シソーラスの概念階層に従い，共起の程度を決定して，文間の相互関係の強度に反映させることなども考えられる．さらに，単語の共起にとどまらず，指示詞や接続詞などによる文関係も考慮し，リンクを張ることも考えられる．単語の共起にとどまらない指示語や接続詞を考慮する考え方はプリミティブな文脈解析に迫るものと考えられ，重要文抽出で懸念されている文脈の首尾一貫性への改善に繋がるものと考えられる．

また，ユーザの観点を考慮することで，システムの改善を図ることが考えられる．ユーザの観点の考慮により，求められる要約の幅が狭まり，一般に求められるものよりも，「そのユーザにとって重要な」要約文を提供することができる．具体的には，ユーザクエリによって与えられたキーワードに対し，文間の相互関係の強度を増加させることなどが考えられる．この時，一度のユーザからの検索式より，必要とする情報をすべて得られることは稀であるため，ユーザに要約結果を提示し，ユーザがその結果を見て検索式を再度与え，出力に変化を与えるようなシステム設計が必要になる．

参考文献

[Hen00] Monika Henzinger. "Link Analysis Web Information Retrieval". *IEEE Bull.of the Tech. Committee on Data Engineering*, Vol. 23, No. 3, pp. 3–8, 2000.

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. "The PageRank Citation Ranking: Bringing Order to the Web". *Stanford Digital Library Technologies Project*, 1998.

[工藤 02] 工藤拓, 松本裕治. "チャンキングの段階適用による日本語係り受け解析". 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[長尾 96] 長尾真. "自然言語処理". 岩波書店, 1996.

[望月 02] 望月源. "テキスト簡易要約器 Posum version 1.50.2 マニュアル". *JAIST Technical Memorandum*, Vol. IS-TM-2002-002, , Jan. 2002.