2P5

文の構造と結束性に寄与する特徴的な語を考慮した文間依存関係に基づく文書要約手法の提案

A Coherent Text Summarization Method based on Semantic Correlations between Sentences

舘林　俊平　　原口　誠
Shumpei TATEBAYASHI　Makoto HARAGUCHI

*1北海道大学 大学院情報科学研究科 コンピュータサイエンス専攻
Graduate School of Information Science and Technology Hokkaido University

We propose an automatic text summarization system taking balances between the importance of sentences in each segment and the importance of sentences to connect several segments. The latter importance is used to extract contextual sentences involving contextual words. In order to separate the notion of importance into the two as in the above, we compute a chunk of core sentences in each segment by a clique finding algorithm, and then calculate the degree of latter importance of sentences from the chunk just in the way used in KeyGraph. Finally, the overall importance is determined by a scheme very similar to topic-sensitive PageRank. We have already made some experiments for newspaper articles, and verified its effectiveness.

1. 研究の位置づけと目的

近年、コンピューターによって処理された膨大な量の情報が存在している。その膨大な量の情報としてテキスト情報が上げられる。インターネットの発達、様々な文書の電子化によってテキスト情報はどんどん増えていく。そのような背景のもと、テキスト自動要約の研究が活発に行われている。現在の文書要約の研究では，まず最初に重要文または重要箇所抽出を行い，その後，照応解析などの自然言語的な補正処理を行って要約文を生成する手法が大半を占める．

要約の起点となる重要文抽出手法の研究において，単語の出現頻度をベースにしたモデルが多く存在する．

しかし，単語の出現頻度をベースにしたモデルでは，全体を通して頻度の高い単語の影響を強く受けすぎてしまうこと，語の出現傾向の変化によって十分な精度が得られないこと，という問題がある．この問題は，高頻出語を多数含む文の重要度が顕著に高くなる傾向を持っていると言い換えることができる．またそれぞれの文の抽出が独立に行われるため，冗長性が高いことも考えられる．上記のような問題点によって，文書内の特定の部分に抽出される重要文が固まってしまう場合があり，全体からバランスよく重要文を抽出することができない．そのような問題から，文書の全体像の把握という目的には，あまり適していないと考えることができる．

そこで，本研究では文書の全体像の把握に向けた重要文抽出のためには，文書の概要をつかむための視点を考慮する必要があると考える．1 段階大きな視点から見ることで，文書は，つながりをもつ複数の文からなる話題の集合と捉えることができる．話題を捉えることができれば，文間の関係と同じように，話題間の構造を考えることができる．話題間の構造を考えた場合に話題をつなぐ役割を果たしている文や単語は，文書の全体像を把握するために重要な概念であると考える．よって，本研究では，元の文書の話題構造を考慮することで全体像を把握するための重要文抽出を目的とする．話題毎の分割統治によって全体からバランスよく文を抽出することに加え，話題間の連結構造を捉えた重要度を提案する．

連絡先: 原口誠，北海道大学大学院情報科学研究科，〒 060-0814 札幌市北区北 14 条西 9 丁目，TEL(FAX):011-706-7106,E-mail:mh@ist.hokudai.ac.jp

2. 提案する文書要約システム

本研究で提案するシステムを，図 1 に示す．文書分割・話題連結キーワードの抽出・PageRank アルゴリズムによる重要文抽出から成るシステムを提案する．

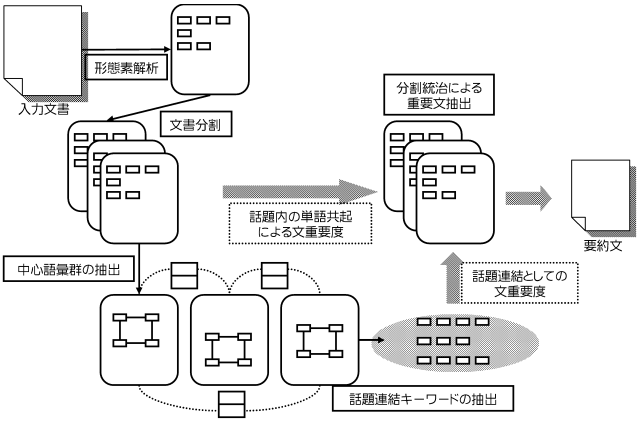

図 1: 本要約システムの概要図

2.1 文書分割

テキストタイリングは，文書の意味的に関連の深い部分には，同一の語が繰り返し出現するという性質を利用している手法である．

句読点や文末など各基準点において，その左右に同数の単語を包含する窓を設け，左右の窓間の類似度を求める．順に基準点をずらしながら類似度の変化に着目し，グラフにおける類似度の極小点を話題の境界として認定する．そしてこの話題の境界に沿って文書を分割する．窓間の類似度は式 (1) の cosine measure で表される．

$$sim(wl, wr) = \frac{\sum_t f(t_{wl}) f(t_{wr})}{\sqrt{\sum_t f(t_{wl})^2 \sum_t f(t_{wr})^2}} \quad (1)$$

しかし，短い文書の場合には，窓のサイズを小さく設定することになり，類似度 0 の点が連続する場合があり、話題の境界

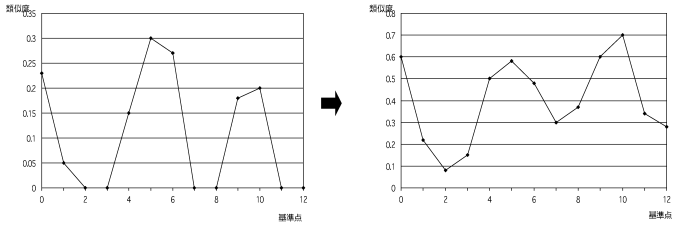

図 2: 共起を考慮したことによる類似度の変化

を認定することができない．そのような場合に、語彙的連鎖の認定のために同一の語のみを対象とするのではなく，文書内の共起情報を利用する手法も提案されている．[平尾 00].

図 2 の左のグラフのように，類似度 0 が連続する場合にも，文内の共起語を用いることで類似度の底上げを図り，話題の境界を認定することができる．

2.2 話題連結としての文の重要度

概念同士をつなぐ情報を扱った研究として KeyGraph[大澤 99] という手法がある．KeyGraph では，重要な内容というのは筆者独特の主張である，そして，筆者はその主張を示すために内容を構成しているという 2 点を前提としている．つまり，文書の基礎となる概念は，筆者の主張を導き出すために関連し合っていると考え，基礎概念同士によって支えられているものが主張点であるとしている．文書全体の単語共起グラフをベースに，基礎概念間と共起する単語を重要視し，主張点を抽出している．

本研究で抽出したいつなぐ情報とは，話題の中心語彙群をつなぐ単語である．この単語の抽出のためには，話題の中心語彙群の抽出と，複数の中心的語彙群との共起を考えることが必要となる．

まず、話題毎の単語間共起グラフを基にした話題連結キーワードの抽出手法を提案する．

1. 話題の中心語彙群として頻度和最大クリークを抽出
2. 話題連結キーワード候補として，中心語彙群と共起する単語を列挙
3. 複数の中心語彙群と共起する候補を話題連結キーワードとして抽出

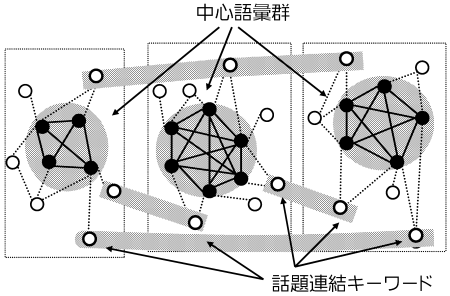

図 3: 話題連結キーワード

中心語彙群は，話題を代表する単語群であり，また，結びつきの強い単語群である必要がある．そこで，話題内での頻度上位単語による頻度和最大クリークを抽出する．最大クリークは，互いに共起する最大の語集合であるため，要請を満たしている．

次に，この中心語彙群に含まれる単語と共起する単語を候補として列挙する．この時に，中心語彙群 g 中の単語との共起度を式 (2) として与えておく．

$$f(w,g) = \sum_{s \in SEG} |w|_s |g|_s \tag{2}$$

ここで，$|w|_s$，$|g|_s$ はそれぞれ文 s での単語 w の出現頻度と中心語彙群中の単語の出現頻度を表す．

最後に，各話題中で列挙された候補のうち，複数の話題間に存在する単語を話題連結キーワードとして抽出する．抽出した話題連結キーワードには，式 (3) によってスコアを与える．

$$key(w) = \sum_{g_i \in D} f(w, g_i) \tag{3}$$

スコアづけされた話題連結キーワードを用いて，話題連結としての文の重要度を決定する．話題構造に対して，その文がどの程度重要な役割を担っているかを示した重要度とするために，話題連結キーワードの総和を考え，文の長さによる影響を削減するために包含単語数で正規化したものを文の話題連結としての重要度とする．

2.3 分割統治による話題毎の重要文抽出システム

本研究ではベースとする重要文抽出システムとして PageRank モデルを用いた．

PageRank アルゴリズムを文書要約に応用する場合，文と文の相互関係に着目して文の重要度を決定する手法がある [四ツ谷 03]．文-文ベクトル空間は行，列をセンテンスとし，関係は文書中の語の共起のによって表される．各文の間で内容語の共起関係があれば，あらかじめ定義されている内容語の重みで表現する．

$$\vec{q} = M\vec{q} \tag{4}$$

ただし，ページの重要度を q，推移確率行列 q とする．再帰的に上の式の R を求めることで，文の重要度を求めることが出来る．

推移確率行列 M を求めるために文間の構造を文-文ベクトル空間で表す．各文に含まれる単語を要素として，式 (5) は余弦尺度を用いて文-文ベクトル空間の強度を求める．

$$sim(x,y) = \frac{\sum_{i=1}^{n}(x_i \cdot y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2 \cdot \sum_{i=1}^{n}(y_i)^2}} \tag{5}$$

この PageRank モデルを 2 つの重要度を組み合わせて用いるために拡張する．ここでの 2 つの重要度とは，話題連結としての重要度と話題内での重要度である．

$$\vec{q} = (1-\alpha)M \times \vec{q} + \alpha\vec{p} \tag{6}$$

右辺第一項の M は，話題内における推移確率行列であり，話題内での語の共起による文間の依存関係から重要度を決定する．これに対して，第二項の $\vec{p}$ はバイアスベクトルであり，これには文の話題連結としての重要度を全確率化したものを与える．α によって 2 つの重要度のバランスを制御し，文の重要度を決定する．この手法を話題毎に同じ要約率で適用し，重要度に基づいて文を抽出する．その後，話題毎の抽出文を和集合を全体での重要文抽出としての出力とする．

3. 実験

本研究では実験データとして NTCIR3[*1]の TSC-2[難波 01][*2]のデータ（98 年度毎日新聞記事）の中から，文数が 30 文を超える社説のデータを利用した．TSC-2 テストデータの重要文抽出結果を正解文とした．各々のシステムで元文書から重要と思われる文を抽出し，正解文との比較を持って客観的評価を行う．全体像の把握という目的を評価するために，文要約率は 30 %とする．

まず 2 つの重要度の配分パラメータである α による平均正解率の変化を確認した．左のグラフに，α を 0 から 0.5 まで 0.05 刻みで変化させた場合の重要文抽出の平均正解率を示す．このグラフから α を増加させる，つまり話題連結構造による文の重要度を考慮することによって平均正解率の向上が見られた．話題連結を考えた文の重要度には有用性があると考えられる．

右のグラフに，今回平均正解率が最大となった $\alpha = 0.4$ の時点での各手法の平均正解率を示す．テキスト簡易要約器 posum，そして文書分割を行わない状態での PageRank モデルによる抽出，分割のみをおこなった場合の PageRank による抽出の正解率と比較して，本研究で提案した手法によって平均正解率は約 10 %の向上が見られた．

話題連結による重要度を組み込んだ結果，8 データ中 5 データで正解率の向上が見られ，正解率が低下するものは 1 データ，変化なしが 3 データとなった．

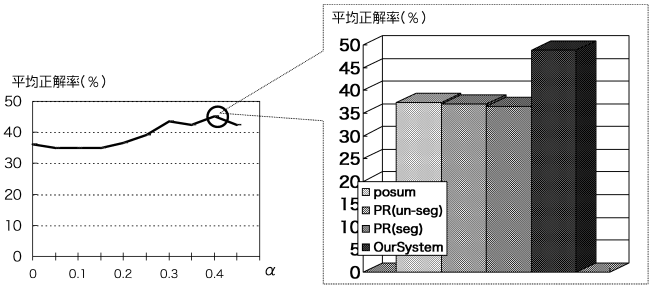

図 4: α の変化による正解率の変化

3.1 考察

抽出された内容を評価することで，話題連結キーワードの有用性を確認する．正解率が向上した場合（表 1）では，話題連結キーワードを考慮することで，正解となった 2 文は，それぞれの話題内で，話題連結としての重要度が最大の文であり，それぞれ，(朝鮮半島，平和，議題)，(平和）という話題連結キーワードを含んでいる．

今回実験にもちいた文書は，4 者会談における朝鮮半島問題に対する平和協定が主題になっており，そのテーマを表す単語が話題連結キーワードとして抽出できており，本研究で提案した手法の効果を示すことができている．posum や分割を行わない状況での PageRank との比較から，全体で単語頻度が高い単語による影響過多というものを排除することが出来ていると考えられる．これは話題連結としての重要度による効果，

対象：4 者会談 米朝の外交努力まだ不足（話題数 3）		
SYSTEM	正解数	抽出正解文
posum	2（17 %）	1.11
PageRank	2（17 %）	1.11
PageRank+Seg	5（42 %）	01.10.11.12.28
OurSystem	7（58 %）	01.02.08.10.11.12.28

表 1: 実験内容評価（精度向上例）

そして，分割統治によって，語の出現傾向の変化による精度低下を避けることができたためと思われる．

また，正解率が低下した場合（表 2）では，話題連結キーワードに関する重要性を考えることで，重要文抽出精度が下がっている．

対象：ビデオ公開米社会の変化見落とすな（話題数 3）		
SYSTEM	正解数	抽出正解文
posum	6（50 %）	01.08.10.24.26.35
PageRank	6（50 %）	01.02.10.14.24.35
PageRank+Seg	5（42 %）	01.02.10.24.35
OurSystem	3（25 %）	01.24.35

表 2: 実験内容評価（精度低下例）

このデータでは，スコアの高い話題連結キーワードとして，「大統領」という単語が抽出されている．この単語は，話題内での高頻度語であるが中心語彙群に含まれなかった単語である．このような単語が，話題連結キーワードとなった場合，話題内での文の重要度，話題連結としての文の重要度がどちらでも高くなってしまう．結果として，文の重要度に対して非常に大きな影響を持つ単語となってしまう．そのため，そのような単語を持つもののみが重要文として抽出されてしまっている．今後の課題として，話題内中心語彙群の定義に関しては制限の緩和を含め，再検討する余地があると考えられる．

4. 謝辞

NTCIR TSC-2 コレクションは国立情報学研究所の許諾を得て使用した．

参考文献

[平尾 00] 平尾努, 北内啓, 木谷 強, 語彙的結束性と単語重要度に基づくテキストセグメンテーション, 情報処理学会論文誌,Vol.41,No.SIG3(TOD6).

[四ツ谷 03] 四ツ谷雅輝, 共起語を介した文間の相互依存関係に基づく重要文抽出法の提案, 北海道大学大学院工学研究科,master's thesis (2003).

[大澤 99] 大澤幸生, ネルス E. ベンソン, 谷内田正彦.KeyGrpah：語の共起グラフの分割・統合によるキーワード抽出, 電子情報通信学会論文誌, J82-D-1, No.2, pp.391–400.1999.

[難波 01] 難波 英嗣, 奥村 学, 第 2 回 NTCIR ワークショップ自動要約タスク (TSC) の結果および評価法の分析, 情報処理学会研究報告, NL-144, pp.143–150, 2001.

∗1 情報検索システム評価用テストコレクション構築プロジェクト http://research.nii.ac.jp/ntcir/index-ja.html を参照.

∗2 テキスト自動要約タスク http://lr-www.pi.titech.ac.jp/tsc/を参照