修士論文提出者	舘林俊平	コンピュータサイエンス専攻 知識ベース研究室
修士論文題名	意味的構造と文間相互関係に基づく文書要約手法の提案	

1　研究の位置づけと目的

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．例として，文に含まれる単語の頻度の和によって重要度を決定する線形和モデルや共起語を介した文間関係に基づいた PageRank モデルなどが提案され，これらの手法は，高要約率での代表文の抽出にはある程度の効果があると考えらている．

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．文書は文の集合であり，その文間の関係は単語によって表すことができるが，それを 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．それによって，今までの単語と文という関係に加え，話題と話題間の構造というもう一つの関係を示すことができるようになる．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．

この新しい重要度を提案するために，文書の話題への分割と話題間の連結構造を考える必要がある．文書の話題の変化を捉えるために，テキストタイリングという手法を利用し，話題間の連結構造を捉えるために，話題の中心語彙群との共起に基づく話題連結キーワードの抽出手法を提案する．

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．

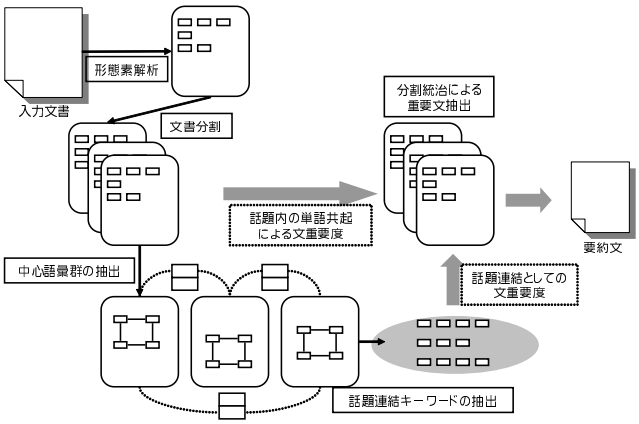

図 1　本要約システムの概要図

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．

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

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，類似度が全体的に低下する．類似度 0 の点が連続する場合には，話題の境界を認定することができない．小さな範囲の中で特定の語が繰り返し使用されるこ

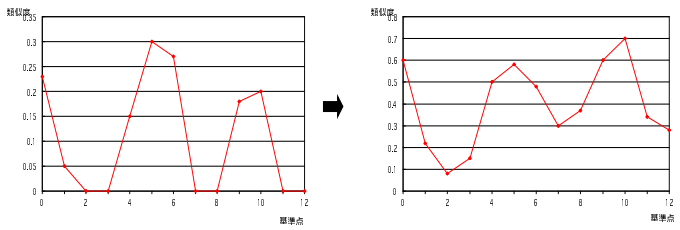

図 2　共起を考慮したことによる類似度の変化

とは稀であるが，言い換えた語やその関連語が使用されることは多いという仮定に基づき，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．[4].

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　話題連結キーワードの抽出

話題間の連結構造を捉えるために，話題連結キーワードの抽出を提案する．概念同士をつなぐ情報を扱った研究として KeyGraph[3] という手法がある．これは主張の内容を表すキーワード抽出を目的とした手法である．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．

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

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

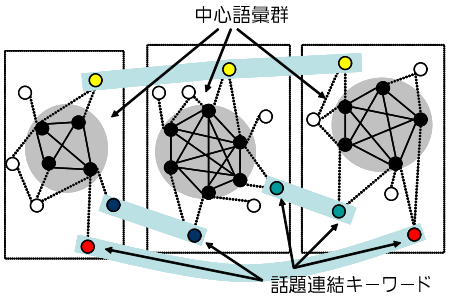

図 3　話題連結キーワード

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．

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

2.3　話題連結キーワードによる文の重要度

スコアづけされた関節語を用いて，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は，

- 関節キーワードを多く含む文ほど，関節としての重要度は高い
- 関節キーワードのスコアが大きな単語を含むほど，関節としての重要度は高い

という二つの要請を満たす必要がある．そこで，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．これによって，話題間をむすびつける話題連結キーワード，そしてそれを基にした文の重要度が定義されたことになる．

2.4　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 アルゴリズムは web 検索エンジン Google[*1] のベース技術として用いられており，「多くの良質なページか

[*1] (日本語) http://www.google.co.jp/

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.」という再帰的な関係をもとに，全てのページの重要度を判定したものである.

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 [2]. 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 実験と考察

提案した手法の有用性の確認のために実験を行った．ベースシステムとの比較により，話題連結キーワードと話題連結による重要度が，文書の全体像把握の目的に対し有用なつながりを捉えているかを確認した.

3.1 実験準備

本研究では実験データとして NTCIR3[*2] の TSC-2[1][*3] 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 テストデータには人手で作成された重要文抽出結果があるため，これを正解文とした．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.

実験環境は，PC ワークステーション（CPU:Pentium4 1.5GHz, Memory:896MB, OS:TurboLinux10）を用いた．プログラミング言語は Perl．PageRank の計算には MATLAB を利用した.

3.2 α による正解率の変化

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.

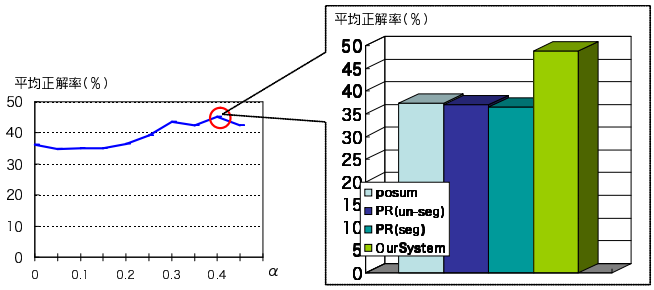

図 4　α の変化による正解率の変化

3.3 話題連結キーワードの内容評価

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，関節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

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

表す単語が関節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．文書分割を行うことで各話題内での閉じた単語出現頻度に基づいた重要文抽出が可能になり，話題内での頻出語である「当事者」「韓国」「北朝鮮」を含む各文が抽出範囲に上がってきている．

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．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．

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
PageRank+Seg+Joint	7（58 %）	01.02.08.10.11.12.28

表 1　実験内容評価（精度向上例）

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
PageRank+Seg+Joint	3（25 %）	01.24.35

表 2　実験内容評価（精度低下例）

3.4　システムの考察

まず抽出された関節キーワードに対しては，全体での頻度はそれほど高くないが，元の文書の主題の概要把握に役立つ単語を抽出できていると考えられる．また他の実験データにおいても，タイトルには含まれるが，文書中の頻度が高くない単語を抽出することができている．

ベースラインシステムとの比較においても，関節としての文の重要度を重要文抽出に組み込むことは，精度改善という面において，一定の効果が見ることができた．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，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．

4　まとめと今後の課題

本研究では文書の全体像把握のためには，話題構造と話題の連結が重要であると考え，話題連結キーワードとそれに基づく話題連結としての文の重要度を提案した．話題連結としての文の重要度を用いるために拡張した PageRank モデルによる重要文抽出の精度向上を実験において確認した．

本研究では，話題の中心語彙群として頻度和最大クリークを用いた．しかしこの制限のために，話題内での高頻度語でも中心語彙群に含まれない場合がある．このような場合には，話題内での重要度と話題連結としての重要度が加算され，文の重要度に対してその単語による影響が非常に強くなってしまう．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，再検討する必要があると考えられる．中心語彙群の検討と同時に，話題連結キーワードの定義が変わることが考えられ，その場合には抽出手法自体の改善を検討する必要がある．

今回実験を行った社説の文書は，文書を考える上では一つのジャンルでしかない．そのため，他ジャンルの文書や文数の多い文書に対しての，追加実験を行い有用性を検証する必要がある．

5　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

[1] 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, pp. 143–150, 2001.

[2] 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. Master’s thesis, 北海道大学大学院工学研究科, 2003.

[3] 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, Vol. J82-D-1, No. 2, pp. 391–400, 1999.

[4] 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. SIG03, pp. 24–36, 2000.