社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ＦＩ－ 84 (20)
2006－ＮＬ－175 (20)
2006／9／13

セグメント間の接続関係を考慮した文書要約に関する一考察

舘林俊平　　原口 誠
北海道大学情報科学研究科

概要： 文書全体の構造を反映した文書要約のために、セグメント内重要イベントとセグメント間重要イベントのバランスをとる一方式を提案する。セグメント間重要イベントとは、複数の異なるセグメントの結束性を高める効果を持つ語を含むイベントであり、文章全体において、各部を繋ぐ役割を持つ。こうした文書要約を実現するために、実際のアルゴリズムとしては、テキストタイリング、クリーク検出、キーグラフ、および、マルコフ連鎖に基づいたものを提案する。

A Coherent Text Summarization Method based on Semantic Correlations between Sentences

Shunpei TATEBAYASHI and Makoto HARAGUCHI
Division of Computer Science, Hokkaido University

Abstract: 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

本研究においては、文章の全体構造に照らして重要な文を抽出するための一つの試みについて述べる。

大規模なテキスト情報を処理するための様々な手法、例えば、クラスタリング、テキストマイニング、要約などに関する精力的な研究が遂行されている。おおまかに述べれば、近さ・遠さの尺度、語の頻度や共起に基づいて、同様な文や文書のクラスタを作り分類する、あるいは、頻度、共起および相関に基づいて重要であると思われる部分を列挙する、等々の手法が標準であると思われる。

ここで論じるまでもなく、何をもって近い、あるいは遠いと思うのか、あるいは、何が頻出で目立っていると考えるのかの問題は、我々が意図した背景や目的にも依存し、一般的に論じるのはなかなか難しい。だからこそ、最も素朴な経験則、すなわち、『出現回数を数えて頻出するもの、あるいは、相関の高いものは重要そうである』というルールが有効なのであろう。

この頻出性の常識をブレイクする試みとして著名なものとしては大澤のキーグラフの研究（[大澤 99]）がある。キーグラフにおいては、高頻度語を文書全体に関わる基礎的な概念として認める。ただし、そうした高頻度語のみならず、「まとまりのある」高頻度語群（具体的には共起性を用いた連結成分）により支持される語を、例え低頻度語であったとしても、基礎的なものに支持される主張を与えるキーワードだと認識し、それなりの重要度を付与する。この手法は文章全体を特徴づけるキーワード抽出法として提案されていることに注意しよう。

一方、文書要約において重要なものを抽出したい

としよう。新聞記事のような特殊な構造（先頭部分が要約になっている）を持つものは別として、多くの文章は、起承転結の構造を持つ。文章によっては、起と結はさりげなく書かれ、承と転に多くの記述がさかれる場合もある。こうした場合、単純な頻出性ではうまくいかないのは明白である。

にも関わらず、文章の全体構造を理解したうえで、適切に要約しなさいと言われれば、人はそれなりの要約をすることができる。その理由としては、

(Segment) 部分（ここでは段落等、一般には連続した文がなすセグメント）の認識を適切な粒度でできること、および

(Connection) 部分と部分の結束性・関連性の認識に基づいて、部分を繋ぐ情報に配慮できる

からだと思われる。特に後者は、文章全体を貫く主題や背景といった、文脈的なものに該当することも多いと思われる。

本稿では、上記の Segment と Connection のバランスをとり、両者を勘案したできるだけ単純な文書要約のための計算モデルを提案する。一般には、一つの文章は複数の観点から解釈でき、よって、Segment も Connection もユニークに定まるわけではない。こうした複数の解釈可能性問題は特に、複数文書を対象にしたときの要約や構造類似性抽出（[Yoshioka05, Haraguchi02]）の際に重要となるが、本稿では、とりあえず、単一文書要約を目標にしたときの、Segment と Connection のバランスのとり方に関する一つの方法を提案する。

そのための基本的な考え方は下記でまとめることができる：

テキストタイリング：まず、文書全体を意味的にまとまりのある複数のセグメントに分解する。本稿では、テキストタイリング（[Hearst94, 平尾 00]）を用いる。

クリーク形成：次に、各セグメントを代表できる語彙グループを形成する。セグメントの意味を表現できる語彙集合の定義（重みづけた語彙集合としての軸、語のコミュニティ、等々）はいろいろあると考えられるが、ここでは、共起性において互いに関連する語のクリーク（[Okubo05, Tomita03]）だと定義し、中心語彙群と称する。

セグメント間連結性：各セグメント毎に求めた中心語彙群を繋ぐ機能を持つ語のスコアを、キーグラフ同様に定め、そうした語を多く含む文に対し Connection の度合いを表すスコアーを付与する。

Topic-Sensitive PageRank：重要度付与の基本モデルとして、共起性に基づいて確率的に文を「訪問する」行為を表すマルコフモデルを用いる。ただし、文の重要度は、各セグメントにおける重要度と、Connection の度合いに応じた重要度の平均値として評価される。結果的に、オープンディレクトリが定める文脈でページの定常状態確率を求める Topic-Sensitive PageRank（[Haveliwala03]）と同じ計算式を用いる。

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．

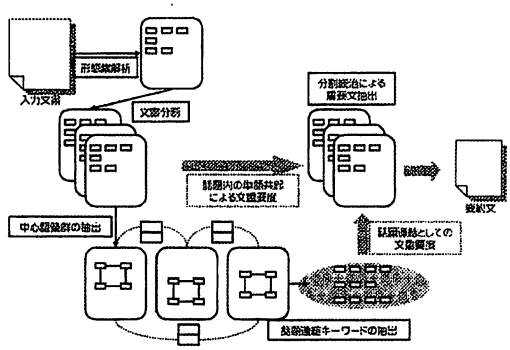

図 1: 本要約システムの概要図

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

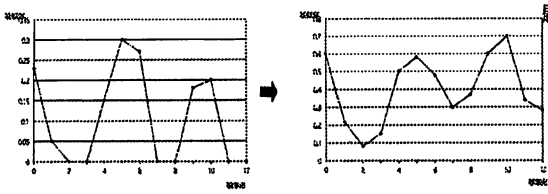

図 2: 共起を考慮したことによる類似度の変化

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) の余弦尺度で表される．

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

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 0 の点が連続する場合があり、話題の境界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]) の手法がある．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（土台）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（屋根）であるとしている．ただし、キーグラフにおいては土台とは文章全体から抽出される高頻度語間の共起性に基づく連結成分であり、本研究において抽出したい、各セグメントにおける話題を表現しているとは限らない。そこで、各セグメントにおける話題を表現するものとして、相互に共起しあう語群（共起性グラフにおけるクリーク）をセグメント毎の中心語彙群として定め、さらに、複数の中心語彙群との共起により、話題（セグメント）を結合する語を抽出する：

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

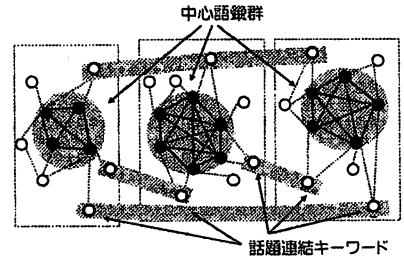

図 3: 話題連結キーワード

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．

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

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．

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

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、文間の共起性のみに着目したマルコフモデルに、複数のセグメントを繋ぐ機能を持つ文に、一定のバイアスを付与する確率計算モデルを採用する。結果的に計算式は topic-sensitive PageRank と同一である。

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最も基本的である (例えば、[四ツ谷 03])．これは、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によって表される．良く知られた事実として、共起度に基づいて、文間の遷移確率を定めた場合、高頻度語をより多く含む文のスコアー（定常状態における確率）が高くなる傾向を持つ。この性質により、

(D1) 文章全体で高頻度語を含む文を訪れる確率が増加し、

(D2) 例えば、導入部や結論部など、主要部と比較すると、文全体における高頻度語の出現度合が比較的小さな文の重要度の低下をもたらす。

この特に、(D2) の欠点を解消するために、本研究においては、

> 異なる複数のセグメントを連結する語彙を含む文Sに、その度合いに応じたバイアスをかけ、Sの訪問確率の低下を防ぎ、それと同時に、Sにより主要部分と結合された、比較的マイナーなセグメント内の重要文を訪れる確率を増加させる。

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*の TSC-2[難波 01]†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

*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.

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

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．

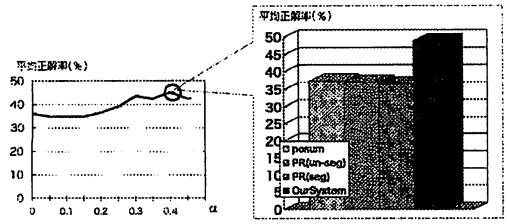

図 4: α の変化による正解率の変化

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．

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）

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，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, 北海道大学工学研究科修士論文 (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.

[舘林 06] 舘林俊平、意味的構造と文間の相互依存関係に基づく文書要約手法の提案、平成 17 年度修士論文、北海道大学情報科学研究科 (2006).

[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.

[上田他 02] 上田、小山 : 共通意味断片の抽出による複数文書要約、言語処理学会弟 6 回年次大会、360–363, 2000.

[Ikeda98] T.Ikeda, A.Okumura, K.Muraki: Information Classification and Navigation Based on 5W1H of the Target Information. COLING-ACL 1998. 571-577

[Haveliwala03] Taher H. Haveliwala: Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 4, pp. 784-96, 2003.

[Hearst94] M. A. Hearst: Multi-Paragraph Segmentation of Expository Text, Proc. of the 32nd Meeting of the Association for Computational Linguistics, pp. 9 - 16, 1994.

[Haraguchi02] M. Haraguchi, S. Nakano and M. Yoshioka: Discovery of Maximal Analogies between Stories, Springer-LNAI 2534, pp. 324 - 331, 2002.

[Okubo05] Y.Okubo and M.Haraguchi: Finding Significant Web Pages with Lower Ranks by Pseudo-Clique Search, Springer-LNAI 3735, pp. 346 - 353, 2005.

[Tomita03] E. Tomita and T. Seki, An Efficient Branch-and-Bound Algorithm for Finding a Maximum Clique, DMTCS'03, Springer-LNCS 2731, pp. 278 - 289, 2003.

[Yoshioka05] M. Yoshioka, M. Haraguchi and A. Mizoe: Towards Constructing Story Databases Using Maximal Analogies between Stories, Springer-LNAI 3359, pp. 243 - 255, 2005.