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

2005－NL－167 (1)
2005－SLP－ 56 (1)
2005／5／26

話し言葉における出現位置情報を用いたキーワード抽出

福田 雅志　　延澤 志保　　太原 育夫

東京理科大学大学院　理工学研究科

テキスト内のキーワードを抽出する手法は従来から研究されており，重要文抽出や要約において数多く利用されている．キーワード抽出では，タイトルや段落や章などの文書構造を利用する手法が有効であると知られている．しかし，話し言葉には句読点など文書構造を示すものがなく，そのままでは書き言葉を対象とした手法を適用することができない．そこで本稿では，話し言葉のキーワード抽出を目的とし，テキストを適切なブロックに分割することで文書構造の利用を可能にする手法を提案する．ブロックへの区切りの推定には，ブロック毎の語句の出現傾向を利用する．本稿で提案した手法と文書全体に対して TF-IDF 重み付けを行う手法の正解率を比較した結果，本稿の手法の有効性が確認された．

Keyword Extraction for Spontaneous Speech

Masashi Fukuda　　Shiho Nobesawa　　Ikuo Tahara

Tokyo University of Science

In this paper we propose a method for keyword extraction from spontaneous speech texts. Document structure, shown by punctuation marks, chapters, etc., is considered as an effective issue for keyword extraction. However, we can not use document structure information in processing spontaneous speech. Thus we propose a method to devide a text into three blocks which can be assumed as document structure. We had experiments on keyword extraction and the results indicated that our method had better precision rate comparing to a simple TF-IDF method.

1　はじめに

テキスト内のキーワードを抽出する手法は従来から研究されており，重要文抽出や要約において数多く利用されている [1, 2]．キーワード抽出では，テキストの種類における特徴を利用する手法が有効であると知られている [3]．例えば，新聞記事や論文においては，タイトルや段落，章題などの文書構造を利用し，そのような重要な語や文に対して，意味的な繋がりや語の類似性からキーワードを抽出する手法がある [4]．これらは書き言葉を対象とした研究であり，話し言葉を対象とする場合，句読点など文書構造を示すマークがないことや，フィラー，言い間違いなどによる繰り返し表現といった特徴 [5] のため，書き言葉を対象とする手法をそのまま適用することができない．そこで本稿では，話し言葉の文書構造を示すブロックの区切りの推定手法について検討する．

2　ブロック区切りを利用したキーワード抽出

話し言葉の書き下し文には，段落や章など話し手の意図する文書構造は明示されない．本稿では，書き言葉を対象とする手法を話し言葉に適用するため，テキストをブロックに分割することで文書構造の推定を行う．ブロック区切りの推定は，テキスト中の語のブロック毎の出現頻度を基に行う．各ブロックについて，出現頻度の高い語をそのブロックのブロック特有語と呼ぶ．本稿で提案する手法は，ブロック特有語の分布の偏りを基にブロック区切りを推定するものである．

例えば科学論文を対象とする場合にはキーワードの出現が期待できる序論部と結論部の重みを増すなど，ブロックの重要度を考慮することで，キーワード抽出の精度の向上が期待できる．本稿では，テキストのブロック推定を行った上で，TF-IDF 重み付けを用いてキーワード抽出を行う．

3　ブロック区切り推定

本稿では，科学論文を対象としてキーワード抽出を行う．実験をともなう科学論文の講演にはある程度決まった話の流れがあり，表 1 に示すように，研究背景を話す前半，実際の実験及びその結果に関して話す中盤，結果からの考察と今後の課題について話す後半と，大きく三つに分類することができる (以下，それぞれ第

一ブロック，第二ブロック，第三ブロックと呼ぶ）．ここで，キーワードとして注目すべき語句は第一ブロックおよび第三ブロックに多く現れると考えられる．ブロックの区切りを単語の出現傾向を用いて推定し，それによって得た第一ブロックと第三ブロックからキーワード抽出を行うことにより，文書全体から抽出した場合に比べより高い精度のキーワード抽出が期待できる．

表 1: ブロック素性分割

第一ブロック	第二ブロック	第三ブロック
研究背景	実験条件	結論・考察
問題点	実験手法	今後の課題
提案手法	実験結果	

3.1 ブロック特有語の相対頻度

本稿では，国立国語研究所の「話し言葉コーパス」（以下 CSJ）から 133 編をトレーニングコーパスとして，3 編をテストコーパスとして用いる．テストコーパスはトレーニングコーパスに含まれない．

各々のテキストを 3 つに仮ブロック分けし，ブロック毎の語の出現頻度を計った．抽出された約 3 万語のうち，出現文書数が多いブロック特有語の一部を表 2 に挙げる．表 2 の数値は全テキスト中の出現頻度を表す．

表 2　ブロック別出現頻度の例

特有語	第一ブロック	第二ブロック	第三ブロック
場合	227	433	476
結果	131	223	436
研究	276	67	67
例	89	189	122
先程	57	145	140

トレーニングコーパス中の各語の出現頻度を基にブロック特有語を選び，(1) 式を用いてブロック特有語の各ブロックの出現相対比を求める．(1) 式において $tf_i(t)$ は第 i ブロックのブロック特有語 t の出現頻度であり，$m_i(t)$ はブロック特有語 t について最も少ないブロックの $tf_i(t)$ を 1 としたときの相対値である．

$$m_i(t) = \frac{tf_i(t)}{min(tf_1(t), tf_2(t), tf_3(t))} \quad (tf_i(t) \neq 0) \quad (1)$$

表 2 の各ブロック特有語に (1) 式を適用した結果を表 3 に示す．

表 3　語の相対頻度 m_i 値の例

t	$m_1(t)$	$m_2(t)$	$m_3(t)$
場合	1.00	1.91	2.10
結果	1.00	1.70	3.33
研究	4.12	1.00	1.00
例	1.00	2.12	1.37
先程	1.00	2.54	2.46

本稿では m_i を用いてブロック特有語の偏りを算出し，ブロック区切りの推定を行う．コーパスの全てのブロック特有語に対して $m_i(t)$ を計算し，これをブロック区切り推定の参照データとして保管する．

3.2 ブロック特有語の相対頻度の累積和を用いた推定

ブロック推定の対象とするテキストを一定の単位で機械的に仮分割し，ブロック特有語の出現頻度を調べることで，ブロック特有語の分布の偏りからブロック区切りの推定が可能である．話し言葉書き下し文では文の区切りすら明確ではないため，本手法では文字単位でテキストを分割し，テキスト先頭からある文字までの間のブロック特有語の出現頻度を利用する．

対象文書に対して $m_i(t)$ を用いて先頭から j 番目に出現するブロック特有語 t_j がブロック i に属する可能性を示す評価値 $B_i(t_j)$ を (2) 式で定義し，これを用いて算出することでブロック区切りの推定を行う．これをブロック特有語の相対頻度の累積和と呼ぶ．

$$B_i(t_j) = \sum_{k=1}^{j} m_i(t_k) \quad (2)$$

CSJ のあるテキストについて，文書中に出現する全ての語に対して参照データ中の語との完全一致での照合を行い，ブロックごとの $B_i(t_j)$ を算出した結果を図 1 に示す．図 1 の横軸は参照ブロック特有語数 j を表し，縦軸は $B_i(t_j)$ を表す．

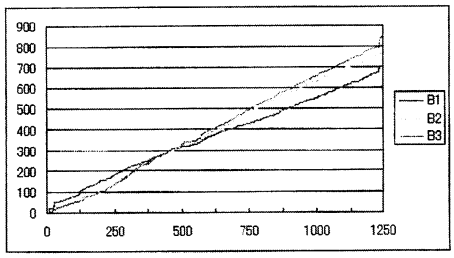

図 1　ブロック特有語の相対頻度の累積和の例 1

このグラフで横軸 423 と 746 の箇所で折れ線グラフの上下が入れ替わっている．つまり先頭から 423 番目のブロック特有語と 746 ブロック特有語の付近でそれぞれブロックが変更されていると推定できる．

3.3 ブロック特有語の出現頻度の周辺和を用いた推定

累積和の手法では，$max(m_i(t))$ が大きいとき，すなわちブロック間の出現傾向に大きな偏りがあるブロッ

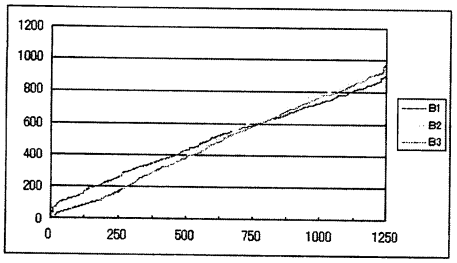

図 2　ブロック特有語の相対頻度の累積和の例 2

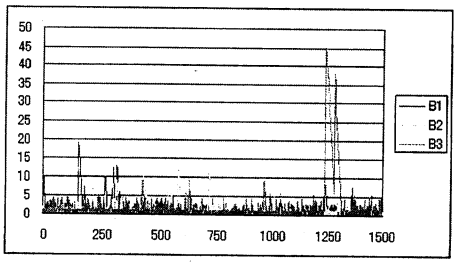

図 3　ブロック特有語の相対頻度の周辺和の例 1

ク特有語が特定の箇所に集中的に出現するときに，ブロック同士の $B_i(t_j)$ 値の差が大きくなり，折れ線グラフ同士が極端に離れてしまい，期待する区切りが得られないことがある．図 2 は累積和の手法を CSJ の別のテキストに適用して得たグラフだが，文書の第一ブロックである 1 番目から 132 番目までのブロック特有語の中で $m_1(t_j)$ が $m_2(t_j)$ と $m_3(t_j)$ より大きいブロック特有語が多く出現しているため，$B_1(t_j)$ が大きくなり $B_1(t_j)$ と $B_2(t_j)$ の折れ線グラフとの交点が中盤以降に現れる．さらには 787 番目のブロック特有語の周辺で $m_2(t_j)$ が $m_3(t_j)$ より大きいブロック特有語が出現しているため，$B_2(t_j)$ と $B_3(t_j)$ の折れ線グラフの交点が文書の終盤に出現することになるため第三ブロックがほとんど存在しないかのような結果となってしまっている．

このような結果が現れるのは，(2) 式が加算のみを行う定義式であるため $B_p(t_j)$ と $B_q(t_j)(p \neq q)$ 間の差が大きく現れたときその差を埋めるだけの $m_i(t_j)$ が現れないことが原因と考えられる．そのため，各ブロック間の $B_i(t_j)$ に大きく差がある場合にはグラフの交点が存在しないことが起こり得る．これは $B_i(t_j)$ が影響する範囲を考慮した周辺和に修正する事により解決できると考えられる．

そこで，$m_i(t_j)$ の値を累積するのではなく，ブロック特有語 t_j が影響を及ぼす範囲を考え，周辺和の算出に用いる．ブロック特有語 t_j の相対頻度 $m_i(t_j)$ は，そのブロック特有語から $m_i(t_j)$ 個後に出現するブロック特有語まで影響するものとし，1 個離れる毎に値を 1 ずつ減衰させる．例えば，$m_i(t_j)$ が 3.50 のブロック特有語 t_j はその後 3 単語において影響を与える．ブロック特有語 t_j の次のブロック特有語の相対頻度の周辺和 $C_i(t_j)$ にはブロック特有語 t_j の作用として 2.50 というスコアを加算する．そこから 1 ずつ作用値を減らしながら作用値が 0 以上の間，各ブロック特有語の $C_i(t_j)$ に作用する．ブロック特有語の相対頻度の周辺和 $C_i(t_j)$ は 3 式で表される．

$$C_i(t_j) = \sum_{k=1}^{j-1} a(m_i(t_k) - j) + m_i(t_j) \qquad (3)$$

ここで

$$a = \begin{cases} 1 & (m_i(t_k) - j \geq 0) \\ 0 & (m_i(t_k) - j < 0) \end{cases} \qquad (4)$$

とする。

この手法を用いた場合には，閾値 $max(C_i(t_j))/4$ を与え，その閾値を超える変更候補点の中から $C_2(t_j)$ が $C_1(t_j)$，$C_3(t_j)$ より高い点を見つけ，その最左点を第一ブロックから第二ブロックへの区切り，最右点を第二ブロックから第三ブロックへの区切りとする．

さて，図 2 と同様の文書に対して (3) 式を用いて算出することでブロック区切り推定を試みる．文書に出現する全ての語に対して参照データ中の語との完全一致での照合を行い，ブロックごとの $C_i(t_j)$ を算出した結果が図 3 である．図 3 の横軸は参照ブロック特有語数 j を表し，縦軸は $C_i(t_j)$ を表す．

このグラフで横軸 307，319，433，589，726 の箇所とその前後で $C_2(t_j)$ が閾値を越えており，その中でも 433 と 589 と 726 の箇所で $C_1(t_j)$ と $C_3(t_j)$ よりも大きな値が検出されていることがわかる．その最左点を第一ブロックから第二ブロックへの区切り，最右点を第二ブロックから第三ブロックへの区切りとして推定できる．つまり先頭から 433 番目のブロック特有語までが第一ブロックに属し，434 番目から 726 番目のブロック特有語までを第二ブロック，727 番目から最後のブロック特有語までを第三ブロックに属すると推定できる．

3.4　ブロック区切り推定を用いた重み付け

キーワード抽出に用いるブロック特有語の重み $w(t,d)$ を以下の方法により求める．まず，(2),(3) 式を用いた方法によりブロック区切りの推定を行う．そして，推定された点によって分割されたブロックのうち，第一ブロックと第三ブロックに出現するブロック特有語のみを対象として重み付けを行う．これは第二ブロックの推定素性である実験条件や実験方法に現れると考えられるブロック特有語は高頻出または繰り返し出現する語でもキーワードには不適格であると考えるからである．重み $w(t,d)$ を，ブロック特有語 t のブロック別の出現頻度 $tf_i(t)(i = 1,2,3)$ と，文書の数 N の中でブロック特有語 t が一回以上出現する文書の数によって

求められる idf 値との積によって，以下のように定義する．

$$w(t,d) = \{tf_1(t) + tf_3(t)\}idf(t) \quad (5)$$

第一ブロックと第三ブロックのみを TF-IDF 重み付けの適用対象にするということは，$tf_2(t)$ を除いた部分に重み付けを行うということであるから，ブロック区切り推定によって得る必要があるのは，文書中の第二ブロック位置ということになる．

4 実験

キーワード抽出実験には CSJ より 3 編を試験テキストとし，訓練テキストとして同コーパスより 133 編を用いた．実験対象としたコーパスは話し手の会話の間(ま) や言葉の読み，感嘆詞の検出などを含めたデータ形式になっているため，テキスト (日本語) と特殊記号の形式を適切に修正する必要がある．以下の点に注意し，文書修正を行った．

- 時間的空白の置換

  「0001 00000.073-00006.865 L:」といった形式の箇所は講演における空白，つまり講演者が沈黙した時間であるので文書化するときには読点への変換が適当であると判断した．句点と読点の見極めは難しく，表層的なデータだけでなく内容の吟味が必要となるので講演の空白部分はすべて読点で置き換えた．本論文の実験においてキーワードを抽出する場合には読点であっても句点であっても処理および結果に違いはない．

- 読みの削除

  文書化するにあたって読み方は情報として必要ないと判断し，全て削除した．キーワードを抽出する際にも活用しないので削除が妥当であると判断した．

- 感嘆詞の削除

  「(F え)」といった形式の箇所は感嘆詞または話し手によるものではない声や音の情報である．この情報も必要ないと判断し，全て削除した．キーワードを抽出する際にも活用しないので削除が妥当であると判断した．

- 文法上の誤りの修正

  文法的に明らかに間違っている箇所や，言い直しによって繰り返されている箇所は，話し言葉書き下し文の特徴であると考えられるので修正は行わない．

形態素解析には茶筌 [6] を用いた．この解析によって対象とする文章を形態素に分解し，各語の品詞と活用形を得た．さらに，講演音声を書き下した文書のため，被験者にとっては非常に読みにくいものとなっている．本研究では，人手での正解語の抽出に際して，最小限の修正を加えた．

- 読点→句点

  文章の内容を吟味することによって句点である方が適当であると判断した場合には読点から句点への置換を行った．ただし，話題の転換を連想させると考え，改行の挿入は行わない．

- 雑音の修正

  未知語と解釈されているものの中には言葉の間に雑音が入っている語があるため，文章内から他の箇所から推定できる語を当てはめた．

4.1 人手による正解語の作成

各試験テキストについて被験者 (5 名) によってキーワードとして適当と判断された語をそれぞれ複数個抽出し，その中から重複の多かった語をそれぞれの文書の正解語とした．人手による正解語の選択ではシステムが行っているような各文についての情報や訓練テキストから得られる対象文書の情報などは公開せず，テキスト中の各語についてキーワードと感じられる語または対象文書を検索によって導きたいときに利用する語を抽出してもらった．文書中に出現する単語または隣接して出現する語であることを抽出の条件であるとし，造語や文中の離れた位置にある単語の組み合わせによる複合語は条件に当てはまらないものとした．

4.2 評価方法

本稿で提案した手法と TF-IDF 重み付けを単独で用いた場合との比較を行うため，以下の 3 種の評価スコアを用いる．

4.2.1 キーワード順位を数値化した評価

重み付けにより順位付けされた語の中で，正解語がより上位にあるほうが抽出手法として有効であると考えられる．そこで，まず文書内の全ての語に対して重み付けを行い，その重み順に語をソートする．比較手法の中で重み付けされた語の数が最も少ない手法における、その語の数を n 個とし，ソート結果中に出現する人手による正解語 t の順位を $r(t)$ としたときのスコア s_1 を

$$s_1 = \sum_t (n - r(t)) \quad (6)$$

とした．

表5 ブロック区切り推定手法別のキーワード順位

順位	なし	累積和	周辺和
1	予測	予測	定位
2	基準方向	未測定	移動音源定位
3	八本	予測係数	音像定位
4	四人分	受	定位システム
5	基準	測定	音像定位システム
6	測定	両耳間	受聴者
7	予測係数	線形予測	未測定
8	受	受聴者	移動音源
9	誤差	両耳受	音源定位
10	方向	両耳受聴	音像
11	両耳間	平均予測	忠実
12	分	基準行列	聴者
13	四人	線形	音場
14	最小	定位	受
15	線形予測	データベース	再現
16	四十五度	誤差	技術
17	受聴者	聴者	測定
18	未測定	スペクトル情報	近年コンピューター
19	両耳受	四人分	コンピューターネットワーク
20	両耳受聴	最小	近年コンピューターネットワーク

4.2.2 キーワード順位に幅を持たせた評価

文書のキーワードとしては通常複数の単語が挙げられる．このことから，重み付けされた語の順位ではなく，ある程度上位に重み付けされた語であれば同等の評価をしてよいと考えられる．つまり，順位に幅を持たせて一つのカテゴリとし，そのカテゴリ毎で評価することも有効であると考えられる．そこで，ソートした結果の上位5単語以内に含まれる正解語数を ss_5 とし，以下順位を5つずつカテゴリ分けし，ss_5 と同様に ss_{10}，ss_{15}，ss_{20} としたときのそれぞれのスコアを4,3,2,1とし，評価スコア s_2 を

$$s_2 = \sum_{i=1}^{4} ss_{i*5} * (5 - i) \qquad (7)$$

とした．

4.2.3 派生語を含めた評価

キーワードの評価として，正解語を含む複合語や正解語の一部となっている語も評価する．そこで，重み付けされた語が完全一致する場合には2，正解語がキーワードの複合語であるかキーワードが正解語の一部である場合（例えば「動的特徴量」と「特徴量」）は1として，重み付けソート結果の上位20語に対して演算を行った合計を評価スコア s_3 とした．

4.3 実験結果

実験文書1に抽出対象として本論文手法によるブロック区切り推定を行った．文書1は総単語数が1070個(読点を含めると1277個)であり，コーパスデータに用いた文書と同形式の文書である．文書化は読み手にとって読みにくくない程度まで最小限の修正を行った．ブロック区切り推定に関して累積和と周辺和の手法を用いた場合について比較を行った．結果を表4に示す．表内の数字はブロック特有語の番号を表す．それぞれの手法によって得たブロック区切りによって第一ブロックと第三ブロックの推定位置を得ることができたので，これを基にTF-IDF重み付けを用いて語の重み付けを行う．

表4 実験文書1のブロック区切り推定結果

ブロック区切り推定手法	累積和	周辺和
第一ブロック	1〜423	1〜270
第二ブロック	424〜746	271〜564
第三ブロック	747〜1277	565〜1277

この情報を基に第一ブロックと第三ブロックのみにTF-IDF重み付けを行った結果を表5に示す．ブロック分割を行わない場合，「八本」，「四人分」，「四人」，「四十五度」などの，第二ブロックに頻出した実験の説明のための語が上位にあるが，ブロック分割を行うことにより第二ブロックの影響をおさえられたことでキーワードの上位に現れなくなった．

表5に評価スコアを計算した結果が表6である．ブロック分けに累積和と周辺和を用いた場合の抽出結果と，ブロック分けをせずに抽出した結果を示している．表内の数字はそれぞれのスコアを表し，s_3 におけるカッコ内の数字は s_3 のスコアに占める派生語によるスコアを表す．

表 6　ブロック区切り推定手法別の抽出評価スコア

ブロック区切り推定手法	s_1	s_2	s_3
なし	59	2	4(2)
累積和	105	1	5(3)
周辺和	85	8	9(5)

表 6 では，累積和を用いた場合が s_1 スコアにおいて最も良い結果を示しており，次いで周辺和の場合，ブロック分けをしなかった場合となっている．しかし，s_2 と s_3 に関しては周辺和が最も優れており，TF-IDF 重み付けと比べて有効な手段であるといえる．また，累積和は文書によってはブロック区切りを得るのに必ずしも有効とは言えないことが実験によって示された．

以上の理由から以下の実験では周辺和によるブロック区切り推定を用いる．ブロック特有語の取得には最多取得法と最長一致法の両方を行い，比較を行った．その結果を表 7 に示す．

表 7　実験文書 1 に対する抽出評価スコア

	s_1	s_2	s_3
TF-IDF 重み付け	59	2	4(2)
本論分手法＋最多取得法	85	8	9(5)
本論分手法＋最長一致法	87	8	4(0)

また，他の文書にも本論文手法を適用し，その結果の平均を表 8 に示す．

表 8　平均評価スコア

	s_1	s_2	s_3
本論分手法＋最多取得法	1.30	2.59	1.68
本論分手法＋最長一致法	1.46	2.64	0.95

表 8 では，3 種類の評価スコア全てで TF-IDF 重み付けの精度を提案手法の精度を上回っており，提案手法の有効性が示されている．最長一致法とともに用いた場合は，派生語の取得評価スコアの数値は TF-IDF 重み付けを下回っているものの，評価スコア s_1 と s_2 に関しては最多取得法に比べてさらなる精度向上が見られ，最長一致法との併用による有効性が示されている．

5　考察

本稿では，ブロック区切り推定を用いた重み付け定義式として式 (5) を用いて，文書の中で実験を話題としている箇所を除去し，その他の箇所のみに TF-IDF 重み付けを行った．ここで，(5) はより一般に

$$w(t,d) = \{\alpha tf_1(t) + \beta tf_2(t) + \gamma tf_3(t)\} idf(t) \quad (8)$$

と書くこともできる．ここで，α, β, γ は各ブロックの適当なパラメータである．本論文ではこの式を定義式として $\alpha = 1, \beta = 0, \gamma = 1$ として実験を行ったとも言える．本論文では第二ブロックにはキーワードが出現することは稀であるという仮定と，第一ブロックと第三ブロックを同じ程度の重要性を持つという仮定の二つの仮定をもとにパラメータを決定したが，科学講演の中には，実験の条件や実験結果における注目点をキーワードとして挙げることができることもある．つまり第 2 パラメータを 0 ではない正の値を与えて考えるのがより汎用的ともいえる．抽出評価スコアがより高くなるようそれぞれのブロックのパラメータを変えて実験を重ねる必要がある．

また，ブロック区切り推定の情報として，ブロック特有語の頻度だけではなく，対象としたコーパス，今回は論文における特殊なブロック特有語，「背景」「まとめ」「実験」などの検出を行い，他のブロック特有語よりも大きな重みを与えることでブロック区切り推定の妥当性の向上も図れると考えられる．さらに，他のコーパスに対して本手法のブロック分割を適用するために，そのコーパスが幾つのブロックに分割する定義式，すなわち分割する必要がないという閾値の定義が必要であると考えられる．

6　まとめ

本稿ではブロック特有語の出現位置を元にしたブロック分割を提案した．そして，講演者が意図した重要語の出現が期待できる前半ブロックと後半ブロックに対してのみ TF-IDF 重み付けを行う手法を提案した．複数の文書に対して実験を行った結果，重み付け順位を用いたスコアと，派生語を含めたスコアで，TF-IDF 重み付けと比べて抽出精度の向上がみられ，本手法の有効性を確認した．今回は特定の一つの分野に基づいたコーパスデータで実験を行ったが，さらに異なる分野のコーパスデータを用いての追実験を行い，本手法の有効性の検証が必要である．

参考文献

[1] Klaus Zechner, "Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences," Proceedings of the 16th International Conference on Computational Linguistics(COLING-96), pp.986-989, 1996.

[2] 砂山渡, 竹内田雅彦, "文章の特徴を表すキーワードを発見して重要文を抽出する展望台システム," 電子情報通信学会誌, vol.J84-D-I, no.2, pp.146-154, 2001.

[3] Chin-Yew Lin, Eduard Hovy, "Identifying Topics by Position," In Proceedings of the 5th ACL Conference on Applied Natural Language Processing, pp.283–290, 1997.

[4] 吉見毅彦, 奥西稔幸, 山路孝浩, 福持陽士, "表題へのつながりに基づく文の重要度評価," 自然言語処理, vol.6, no.1, pp.43-57, 1999.

[5] 下岡和也, 南條浩輝, 河原達也, "講演の書き起こしに対する統計的手法を用いた文体の整形," 自然言語処理, vol.11, no.2, pp.67–83, 2004.

[6] 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座『形態素解析システム《茶筌》 version 2.3.3 使用説明書』, 2003.