情 報 学 基 礎 50－1
（1998. 7. 17）

文書の話題構成に基づく重要語の抽出

仲尾 由雄
nakao@flab.fujitsu.co.jp
富士通研究所
〒211-8588 川崎市中原区上小田中 4-1-1

文書の概要把握の支援用の見出しの自動生成を目指して行った基礎的実験について報告する。長めの報告書の前書き部を除いた3節それぞれに対し、(a) 情報検索で広く用いられている $tf \times idf$ 法による評価の高い語を抽出する場合、(b) 文書内単語出現確率に基づく単語の情報量の大きい語を抽出する場合、(c) 節内単語出現確率が文書内単語出現確率より有意に高い語を抽出する場合を比較実験し、その性質を分析した。その結果、(1) 対象節とその周辺の節との間の出現確率の異なり、(2) 対象節内における単語の出現箇所の集中度、の両者の評価を組み合わせることが効果的であるとの結論を得た。$tf \times idf$ 法と単語出現確率の統計モデルとの関係に関する考察も行っている。

Automatic Keyword Extraction based on the Topic Structure of a Text

Yoshio Nakao

Fujitsu Laboratories Ltd.
4-1-1, Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa, 211-8588 Japan

This paper reports an experiment made for automatic generation of headings, which are intended to be attached to an automatically generated summary of a text, especially of a long one. The experiment lists up words of high relative frequency in a text section in the order of significance, and examines their correlation with words taken from headings in the section. The result indicates two important factors to analyze word occurrence distribution: the difference of word density among through sections and the biased local distribution in a section. This paper also discusses the relation between an IR model of $tf \times idf$ and a stochastic model of likelihood ratio test of goodness-of-fit.

1 はじめに

本稿では、文書の概要把握の支援用の見出しの自動生成を目指して行った基礎的実験について報告する。長めの報告書の節から節内外の単語の出現分布を手がかりに自動抽出した重要語と、節中の見出しに含まれる語とを比較したものである。これは次のような問題の解決を意図している。

長い文書から文を抜粋して要約を作成する場合、作成した要約が長過ぎて読みにくくなり、要約本来の機能を果たさくなってしまう場合がある。例えば、多くの話題を同じ程度の比率で含むような長い報告書を要約しようとすると、主要な話題を拾うだけで要約が長くなり、結果として要約が素早く概要を把握するためには役に立たなくなることがある。この場合、文の抜粋に加え、要約として抜粋した内容を簡潔に示す見出しを添えることができれば、要約の拾い読みが可能になるので、素早く概要を把握できるようになる。

今回の実験はこのような問題意識から、見出しの核となる主要な単語を、文書中での単語の出現分布を手がかりに抽出することを目的としている。

実験では、長めの報告書 (80 ページ程度) の前書き部を除いた 3 節に関し、$tf \times idf$ 法を利用した手法により重要語を抽出し、(1) 文書内単語出現確率に基づく単語の情報量の大きい語を抽出する場合、(2) 文書内単語出現確率と節内単語出現確率を尤度比検定し、節内出現確率が有意に高い語を抽出する場合とを比較した。

以下、第 2 節で本実験の基本方針を示し、第 3 節で実験の詳細について述べたあと、第 4 節で実験結果に基づく考察を試みる。

2 重要語抽出の基本方針

本稿における重要語抽出の基本方針は、文書中のある話題のまとまりに関して特徴的に出現する単語を抽出するというものである。「話題のまとまり」とは、今回の実験では、文書中の論理的単位である節のことである[1]。

「特徴的に出現」の判定に関わる要素としては、今回の実験では、以下の 2 つを考慮している。

- 単語の出現確率 (出現密度)

  ある単語が、重要語抽出対象の節で比較的高密度で出現し、その周囲の節中の出現密度との差が有意である場合に、その単語を重要語と見なす。

- 節内における単語の出現箇所の集中度

  ある単語が、重要語抽出対象の節内で出現箇所が局所に集中する傾向が有意に強い場合に、その単語を重要語と見なす。

このように、本稿では、文書を単語を盛った籠 (word basket) の並びとして扱い、それぞれの語と別の語との出現位置の関係などは無視している。すなわち、文書全体をいくつかの区間 (basket) に分割し、それぞれの区間内の出現密度の違いの統計的評価で抽出すべき重要語がどこまで認定できるか、というのが本稿の問題設定である。

3 重要語抽出実験

3.1 実験条件

実験に用いた文書と正解データ

実験文書としては、(社) 電子工業振興協会『自然言語処理システムの動向に関する調査報告書』(平成 9 年 3 月) 第 4 章「ネットワークアクセス技術専門委員会活動報告」(pp. 117–197) を用いた。この文書は、1,440 文 (延べ 17,816 内容語[2]) からなり、4.1 節から 4.4 節の 4 節に分かれている。このうち極端に短い 4.1 節 (前書きにあたる部分) を除いた 3 節が今回の重要語抽出実験の対象である (表 1)。

評価実験の重要語の正解データとしては、節中の見出しに含まれる内容語[2] を用いた。例えば、4.3 節を対象に重要語を抽出した場合には、その節自体の見出し「ネットワーク上の検索サービス」をはじめとし、4.3 節に含まれる 4.3.1 節などの見出しのいずれかに含まれる内容語が正解データとなる。図 1 に 4 レベル目までの全見出しを示す。なお、正解データ用の見出しは、これに加え 6 レベル目まで用いた。例として、4.2.2 節の下の 5〜6 レベルを示す。

(2) インタフェースを支えるネットワークプログラム技術 (a) Ｊａｖａ (b)Telescript

[1] 将来的には [1] を下敷きに開発した手法 [2] によって自動的に行う予定である。

[2] 名詞・動詞・形容詞。詳細次節。

4. ネットワークアクセス技術委員会
4.1 調査の概要
4.2 ネットワークアクセスのインタフェース
4.2.1 提言:10年後のネットワークアクセスインタフェースはこうなる
(1) ネットワーク情報への多様なアクセス
(2) 個人向けインタフェースを支えるエージェント技術
(3) セキュリティ・個人認証の今後
(4) 機械翻訳と多国語
4. 2. 2 現状と問題点
(1) アクセスインタフェースの多様化
(2) インタフェースを支えるネットワークプログラム技術
(3) セキュリティ・個人認証
(4) 機械翻訳・言語処理技術
4. 3 ネットワーク上の検索サービス
4. 3. 1 検索サービスの調査
(1) WWW検索サービスの概要
(2) 情報収集/検索方式
(3) 情報提示方式
(4) 今後の課題
4. 3. 2 検索技術の動向
(1) キーワード抽出
(2) 文書自動分類
(3) 要約・抄録技術
(4) 分散検索
4. 3. 3 電子出版及び電子図書館
(1) 電子出版
(2) 電子図書館
4. 4. 検索エンジン
4. 4. 1. 日本語の全文検索技術の動向
(1) 文字列検索アルゴリズム
(2) インデックス作成法
(3) 日本語の全文検索技術
(4) 製品化動向
(5) 今後の課題
4. 4. 2. 有限オートマトンによる自然言語処理技術の動向
(1) 有限変換器のコンパクト化
(2) 文字列パタン照合
(3) 書き換え規則,Two-level モデル
(4) 形態素解析, 構文解析
(5) まとめ
4. 4. 3 情報フィルタリング技術の動向
(1) 内容に基づくフィルタリング (content-based filtering)
(2) 協調フィルタリング (collaborative filtering)
(3) ユーザモデリング
(4) まとめ
4. 4. 4 情報抽出/統合技術の動向
(1) 検索ナビゲーション技術
(2) 情報統合技術
(3) 情報の可視化技術

図 1: 実験文書中の見出し

表 1: 実験文書中の節の構成

	節の大きさ (語数)		見出し中の語数	
	延べ	異なり	延べ	異なり
4.1 節	308	140 (48)	2	2(1)
4.2 節	4,771	1,625 (604)	95	54(52)
4.3 節	6,067	1,387 (679)	164	80(70)
4.4 節	6,670	1,853 (822)	103	64(58)
total	17,816	5,005(2,153)	364	200(181)

() 内は頻度 2 以上の語

重要語抽出対象の単語

今回の実験の単語認定には、日本語形態素解析ツール jmor[3] を使って切り出した内容語 (名詞・動詞・形容詞) を用いた。jmor によって切り出される名詞には、形容動詞語幹が含まれ、機能語や数字・時詞・相対名詞 (左右/上下/以上/以下など) は含まれない。例えば、実験文書の先頭の 3 文から、以下の【】で囲まれたものが切りだされた。【】内の "/" の後ろは、活用語の終止形語尾である。見出し中の語と抽出した重要語の比較においては、終止形語尾つきで表記が一致するものを同一の語とみなした[3]。

4.1 【調査/する】の【概要/】

【インターネット/】は【予想/する】されて【い/る】た以上の【早さ/】で【急速/】に【普及/する】して【い/る】る。【業務/】はもちろん特に【家庭/】での【利用/する】が【急速/】に【広が/る】って【い/る】る。

3.2 予備実験: $tf \times idf$ 法による重要語抽出

本節では、情報検索や文書分類において、文書中のキーワードの重みとして広く用いられている $tf \times idf$ 法を用いた重要語抽出の実験結果を示す。$tf \times idf$ 法には色々な変種が存在するが (例えば [4])、基本的には、文書内の単語の出現頻度 (term frequency:TF) に文書頻度の逆数 (inverse document frequency: IDF) に相当する値を乗ずることで、どの文書にも出現しているような語の重みを減じて取り扱う手法である。

[3] 「い/る」は "要る" "居るのいずれの意味でも同一の語とみなすことになる。また、「い/る」と「要/る」のように表記が違う語は例え意味が同じでも別の語とみなした。

今回の最初の実験では、文書中の節を $tf{\times}idf$ 法における1つの文書として取り扱い、以下の値 v_w によって単語の重要度を評価し、重要度の高い順に重要語抽出を試みた(以下「単一ブロック $tf{\times}idf$ 法と称する)。

$$v_w = tf_{w,s} \times \log(\frac{|D|}{df_w}))$$

ここに

v_w	単語 w の評価値
$tf_{w,s}$	単語 w の節 s における出現頻度
$\|D\|$	文書中の節の数
df_w	単語 w の出現した節の数

である。

図2の＊の折れ線が、単一ブロック $tf{\times}idf$ 法の抽出精度である。抽出精度とは、抽出した重要語の見出し中の語に対する再現率と適合率であり、以下の値を用いている。

$$\text{再現率} \equiv \frac{\text{見出し中の語と一致した重要語の数}}{\text{見出し中の異なり語数}}$$

$$\text{適合率} \equiv \frac{\text{見出し中の語と一致した重要語の数}}{\text{抽出した重要語の数}}$$

図2中、点線と破線は、以下に示す再現率・適合率の基準値 (baseline) である (その他の折れ線については後述)。

- 点線: 節内頻度法 (高めの基準値)

  それぞれの節における出現頻度の高い順に重要語を抽出した場合の抽出精度値である。

- 破線: 文書内頻度法 (低めの基準値)

  文書全体における出現頻度の高い順に重要語を抽出した場合の抽出精度である。

図によれば、単一ブロック $tf{\times}idf$ 法は、再現率6%程度 (上位12語に相当) までは完全に抽出が成功している (適合率100%) が、10%程度以降 (同21語以降) で、節内の頻度順に抽出した場合より抽出精度が悪くなっている。特に再現率16%(同29語以降) 以降では、文書内の頻度順に抽出した場合と同程度以下の抽出精度しかあがっていない。

この主要な原因は、節の数が4(うち実験対象は3)と少なく、また、主題も大きくとらえれば共通であるため、文書をつらぬく主題に関わる語などが全ての節に出現していたためであると考えられる。

複数ブロック $tf{\times}idf$ 法による重要語抽出

前節の実験により、節を単位としてIDFを計算するしたのではよい精度が得られないことがわかったので、次にIDFの計算単位を節より細かくした実験を試みた[4]。

節 (文書) を複数のブロックに分割する方法は色々な方法が考えられるが、今回は、単純に一定語数で文書を区切って実験した。具体的には文書の先頭から80語[5] 刻みで文書をブロックに分割し、当該単語の出現したブロックの数を数え文書頻度 (df_w) とした。

単語の重要度の評価値としては、次の2種類を用いた。いずれも、節が1ブロックからなる場合には、前節の評価値と一致する。それぞれの評価法につけた名前の由来については次節を参照のこと。

- 情報量型複数ブロック $tf{\times}idf$ 法

  IDFを数える単位だけを変更したもの。具体的には以下の通り。

$$v_w = tf_{w,s} \times \log(\frac{|D|}{df_w}))$$

ここに

v_w	単語 w の評価値
$tf_{w,s}$	単語 w の節 s における出現頻度
$\|D\|$	文書中のブロックの数
df_w	文書全体の単語 w の出現ブロック数
$df_{w,s}$	節 s 中の単語 w の出現ブロック数

である。

- 尤度比検定型複数ブロック $tf{\times}idf$ 法

  IDFを数える単位の変更だけでなく、節内と文書内のIDFの比をとるようにしたもの。具体的には以下の通り。

$$v_w = tf_{w,s} \times \log(\frac{df_{w,s}|D|}{|D_s|df_w}))$$

ここに

v_w	単語 w の評価値
$tf_{w,s}$	単語 w の節 s における出現頻度

[4] IDFの計算単位とするブロックを小さくすれば、たとえ文書をつらぬく主題を示す語であっても、全てのブロックに出現することはなくなるはずである。

[5] 40語～640語の範囲で等比級数的にブロックの大きさを変更して実験してみたが大差なかった。あえて比較すれば40語～160語が比較的よさそうな結果を示していた。

$\|D\|$	文書中のブロックの数
df_w	文書全体の単語 w の出現ブロック数
$df_{w,s}$	節 s 中の単語 w の出現ブロック数
$\|D_s\|$	節 s 中の総ブロック数

である。

図 2 の×の折れ線が、情報量型複数ブロック $tf \times idf$ 法の抽出精度であり、＋の折れ線が、尤度比検定型複数ブロック $tf \times idf$ 法の抽出精度である。

いずれも再現率の低い部分では、節内頻度法 (高めの基準値) よりも高い適合率を示している。そして、特に低再現率の部分では (0〜20%程度) 尤度比検定型複数ブロック $tf \times idf$ 法の方が、情報量型複数ブロック $tf \times idf$ 法より適合率がよい傾向にある。

再現率の高い部分では、いずれの方法も急激に適合率の減少する地点がある。情報量型複数ブロック $tf \times idf$ 法は、再現率 38%以降 (上位 69 語以降) で急激に適合率が低下し、その後は漸近的に節内頻度法の精度まで低下している。一方、尤度比検定型複数ブロック $tf \times idf$ 法は、再現率 30%以降 (同 56 語以降) で急激に適合率が減少し、最終的には文書頻度法 (低めの基準値) も下回る結果になっている。

急激に適合率が低下する大きな原因は、複数箇所で何らかの共通する (主題) 語が使われていることにある。両者の方法とも、文書全体の平均出現密度に比べ、節内の出現密度の高い単語を抽出する性質がある。そのため、1 箇所で集中して論じられている事柄と関わる語は優先的に抽出されるが、(別々の節に属する) 複数箇所で共通の事柄が取り扱われていると、互いの出現密度が打ち消しあって優先順位が下がってしまう。例えば、実験文書の 4.3 節中の「検索/する」(節内頻度法では第 1 位の重要語) と 4.4 節の「検索/する」(節内頻度法では第 2 位の重要語) の優先順位を見ると、情報量型複数ブロック $tf \times idf$ 法ではそれぞれ第 4 位・第 15 位、尤度比検定型複数ブロック $tf \times idf$ 法では第 4 位・第 27 位となっていた。この例に示されるように、尤度比検定型複数ブロック $tf \times idf$ 法の方が、密度の差異に敏感なため、高再現率部分での適合率の低下も激しくなるものと考えられる。

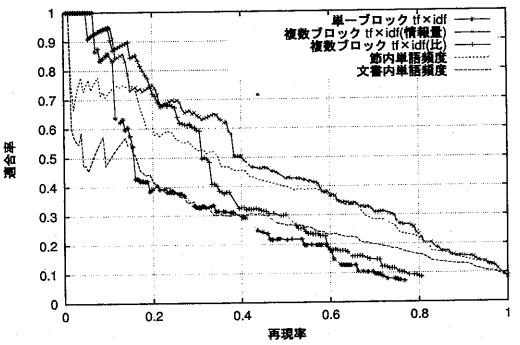

図 2: $tf \times idf$ 法の変種による重要語抽出結果

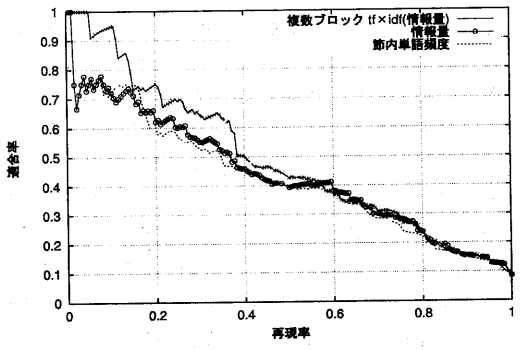

図 3: 親区間の出現確率による情報量との比較

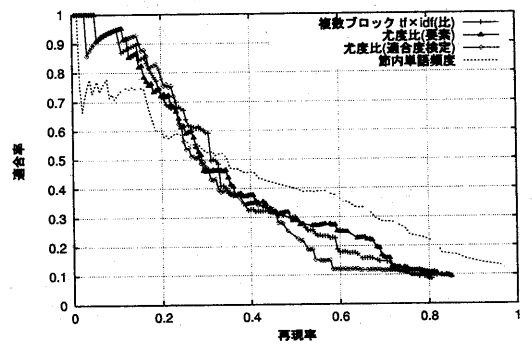

図 4: 適合度検定による重要語抽出との比較

4 考察

4.1 節内頻度法の抽出精度の低下の要因

図 8〜10 は前記の実験全てを、抽出対象を名詞のみに絞って (形容詞・動詞を抽出対象からはずして) 行った結果である。これによれば、名詞のみに対象を絞るのであれば、節内頻度法は今回用いたいずれの方法ともほとんど遜色ない抽出精度を挙げている。

今回用いた形態素解析ツール (jmor) が名詞キーワード抽出用によくチューニングされていたためと考えられる。すなわち、「以上/以下」のような機能的な名詞は (いわゆるストップワードとして) 最初から (jmor の出力) 除外されているため、単純に節内頻度を調べるだけでも精度よく重要語を抽出できたものと考えられる。

逆にいえば、今回の実験では、動詞も抽出対象としており[6]、また、形態素解析ツールが重要動詞の抽出用にはほとんどチューニングされていないにも関わらず、不要な動詞の抽出がほとんど抑えられていた点が注目される (次節の「い/る」の例参照)。すなわち、ストップワードリストを用いることなく、単一の文書の分析だけで効率的に重要語を選別できた点である。ただし、機能的な名詞の出現分布は、機能的な補助用言の出現分布今後の課題としたい。

図 5〜7 は、社内で流通している経済関係のレポート[7] の 1997 年 1〜12 月分 (111 記事) を使用した場合の実験結果である (表 2)。こちらは本実験と同様、動詞・形容詞も重要語の抽出対象としている[8]。これにおいても、節内頻度法は今回用いたいずれの方法ともほとんど遜色ない抽出精度を挙げている。このレポートでは、記事の間の話題の異なりが大きかったためと思われるが、詳細の分析は今後の課題である。

4.2 情報量と $tf \times idf$ 値との関係

情報量型複数ブロック $tf \times idf$ 法の評価値を求める式を眺めると、IDF の部分 (log をとった部分) は、文書を分割したブロック単位で集計した出現確率の

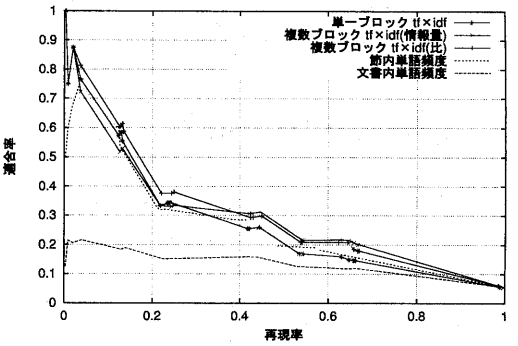

図 5: $tf \times idf$ 法の変種による重要語抽出結果 (追試)

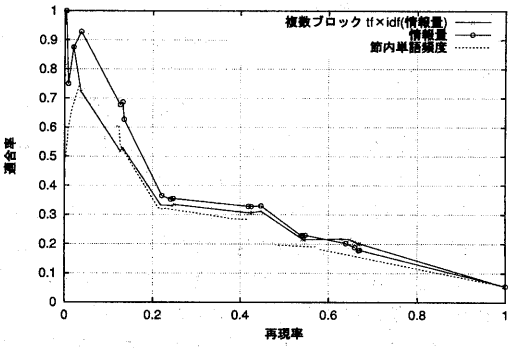

図 6: 親区間の出現確率による情報量との比較 (追試)

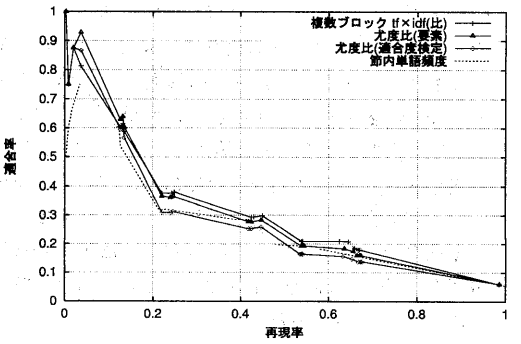

図 7: 適合度検定による重要語抽出との比較 (追試)

[6] 正解データ—実験文書の見出し中の語–にはサ変名詞以外の動詞はほとんど含まれていないことにも注意。

[7] イリノイ大学の室賀教授による計算機関連の市場動向などをまとめたもの。スタイルとしては雑誌の経済記事と同様。

[8] 前述の実験文書に比べ見出し中に動詞が出てくることが多い。ただし絶対数は多くない。

表 2: 経済レポートの数値的概要

	記事の大きさ (語数)		見出し中の語数	
	延べ	異なり	延べ	異なり
平均値	364	174 (57)	3.9	3.9 (2.3)
中央値	152	152 (48)	4	4 (2)

() 内は頻度 2 以上の語

形になっている。よって、以下のように計算される単語の情報量 I_w と比較すると、(何らかの意味で補正した) 情報量の一種としてとらえることができる。

$$\begin{aligned} I_w &= -tf_w \times \log(\text{単語の出現確率}) \\ &= -tf_w \times \log(\frac{tf_w}{|W|}) \end{aligned}$$

ここに

tf_w 単語 w の文書中での頻度

$|W|$ 文書中の単語の延べ数

である。

図 3 は、上記の式による情報量順に重要語を抽出した場合と情報量型複数ブロック $tf{\times}idf$ 法との抽出精度を比較したものである。これによれば、低再現率の部分 (再現率 37%、上位 68 語以前) で、ブロック単位に確率を計算した情報量型複数ブロック $tf{\times}idf$ 法の方が適合率が高くなっていることがわかる。

これは、文章の主題とは関わりはないが局所的に集中して出現しているような単語に関する出現頻度がうまく補正されたためと考えられる。例えば、節内出現頻度では第 3 位 (節内出現頻度 113 回) だった 4.2 節の「い/る」[9] は、情報量型複数ブロック $tf{\times}idf$ 法では第 106 位 (節内ブロック頻度 45 回) まで評価が下がっていた。

Bookstein ら [6] は、内容語の出現は局所に偏りやすい傾向があることを示しているが、まとまりのある文章を書くためには補助的に説明を加えている語などであっても繰り返すことがある。Bookstein ら [6] では、局所的に高密度になるという性質だけでなく、もっと大きい単位で連続してあらわれる (出現ブロックの連続パタンが偏る) という性質も示している。

[9] 「ている」の補助用言がほとんど。通常の名詞抽出などと同様にチューニングするなら当然ストップワードとなるはずのもの。

今回の実験結果からすると、段落程度の大きさのブロック内の単語出現密度をその部分の主役となる語の指標として用い、より大きな周期の語の出現の繰り返し (語彙的結束性 [5]) を語と文章の主題との関連性の指標とするのがよいように思われる。これに関しては、Yarowski[7, 1992 年] の単語の前後の 100 語を文脈を使った語義の曖昧性実験や、Rosenfeld[8] が報告している自己トリガの優位性[10] などを参考に、今後検討する予定である。

4.3 尤度比検定と $tf{\times}idf$ 値との関係

図 7 の◇の折れ線は、単語の出現密度 (出現確率) に関し、節内の出現密度と文書全体における出現密度を統計的に検定し、有意に出現密度の高い語を有意水準の大きい順に抽出した場合のの抽出精度である。具体的には、以下に示す適合度の尤度比検定 (likelihood test of goodness of fit) の値[11] の大きい順に単語を抽出した。

$$\begin{aligned} -2\log\lambda &= 2 \times tf_{w,s} \times \log(\frac{tf_{w,s}}{\tilde{tf}_{w,s}})) \\ &+ 2 \times (tf_{\bar{w},s}) \times \log(\frac{tf_{\bar{w},s}}{\tilde{tf}_{\bar{w},s}})) \end{aligned}$$

ここに

v_w 単語 w の評価値

$tf_{w,s}$ 単語 w の節 s 中の出現頻度

$tf_{\bar{w},s}$ w 以外の単語の節 s 中の出現頻度

$\tilde{tf}_{w,s}$ 節 s 中の単語 w の出現頻度の期待値

である。

図 7 の△の折れ線は、前記の式の第 1 項のみの値の大きい順位重要語を抽出した場合の抽出精度である。情報量型複数ブロック $tf{\times}idf$ 法の評価式は、この第 1 項のみの形とよく似ている。

◇△の折れ線は、情報量型複数ブロック $tf{\times}idf$ 法の折れ線 (×) とよくにた挙動を示しており、IDF を単語の出現密度 (確率) の補正値とする解釈の妥当性を示す一例と考えられる。

[10] ある単語の出現によりその後に続く単語の出現する確率が大きくなる現象において、その単語自身が後続する確率の上昇が著しいということ。

[11] 単語の共起性 (出現確率の相関性) の尺度として安定であるといわれている検定手法 (Dunning[9]、久光ら [10])。

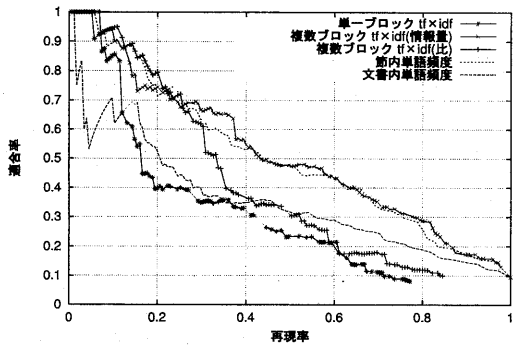

図 8: $tf \times idf$ 法の変種による名詞重要語抽出結果

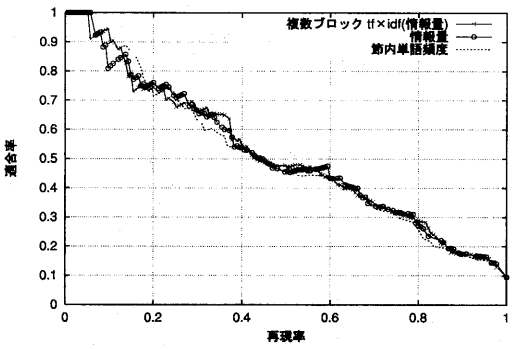

図 9: 親区間の出現確率による名詞の情報量との比較

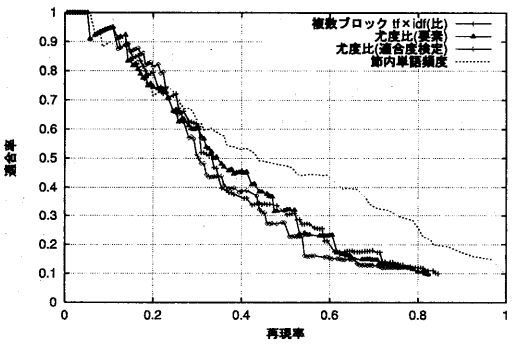

図 10: 適合度検定による名詞重要語抽出との比較)

5 まとめ

本稿では、単語の出現分布に基づき見出しの核となりうる語を抽出する手法を数種類示し、実験を通じてそれらの性質を分析した。そして、単一の文書だけを手がかりに重要語を取り出す場合でも、文書を段落程度のブロックに区切り、$tf \times idf$ 法を適用すると重要語の抽出精度が向上することを示した。また、$tf \times idf$ の値は、単語の情報量と関連が深く、IDF を単語の局所的出現を少なめに補正した出現確率の一種として機能していることを示した。

重要語を効率的に抽出するには、節などの大きな話題のまとまりの間の単語出現密度の差と、単語の局所的分布の偏りの両者を考慮する必要がある。IDF の扱いを工夫すれば、この両者を整合的に評価可能な尺度が作成できるかもしれない。今後の課題である。また、今回の知見が別の事例についても広く成り立ちうるのかを確かめることも課題である。

謝辞 実験データを提供して下さった電子情報ネットワークアクセス実行委員の方々に感謝致します。

参考文献

[1] Hearst, M. A.: Multi-paragraph segmentation of expository text, in *Proceedings of the 32nd Annual Meeting Annual Meeting of Association for Computational Linguistics*, pp. 9–16 (1994).

[2] 仲尾由雄：文書の意味的階層構造の自動認定に基づく要約作成, 第 4 回年次大会併設ワークショップ「テキスト要約の現状と将来」言語処理学会 (1998).

[3] 西野文人：日本語テキスト分類における特徴素抽出, 情処研報 NL-112-14, 情報処理学会 (1996).

[4] Zobel, J. and Moffat, A.: Similarity Measures Explored, Technical Report TR-95-3, Collaborative Information Technology Research Institute (1995).

[5] Haliday, M. A. and Hassan, R.: *Cohesion in English*, Longman, London (1976).

[6] Bookstein, A., Klein, S. T. and Raita, T.: Clumping Properties of Content-Bearing Words, *Journal of the American Society for Information Science*, Vol. 49, No. 2, pp. 102–114 (1998).

[7] Yarowsky, D.: Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, in *Proceedings of COLING '92*, pp. 454–460 (1992).

[8] Rosenfeld, R.: Adaptive Statistical Language Modeling: A Maximum Entropy Approach, Research Report CMU-CS-94-138, CMU (1994).

[9] Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics*, Vol. 19, No. 1, pp. 61–74 (1993).

[10] 久光徹, 丹羽芳樹：統計量とルールを組み合わせて有用な括弧表現を抽出する手法, 情処研報 NL-122-17, 情報処理学会 (1997).