自然言語処理 117－17
（1997. 1. 21）

見出しを利用した 新聞・レポートからのダイジェスト情報の抽出

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

文書の選別などを目的に、見出しを補完するようなダイジェスト情報を新聞と経済関係のレポートから抽出した実験について報告する。見出しに含まれる名詞キーワードを核として、それを含む文を本稿で提案している方法で選択して抽出することにより、1カ月分の新聞記事 (約1.3万件) と6カ月分のレポート (62トピック) を対象とした実験では、新聞記事の93%、レポートの98% について1～3文程度 (元の文章の1～3割程度に相当) のダイジェスト情報が抽出できた。98のダイジェストを目視した結果、8割以上について文書選別用のダイジェストとして使えることが分かった。目視の際に見つかったキーワード照合や照応表現などに関する問題点の考察も行っている。

An Empirical Study for Digest Generation: Constituent Word Correspondence between Titles and Body Parts of Japanese Articles

Yoshio Nakao

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

This paper presents a simple method of automatic digest generation of articles, which is intended to provide an effective view of a large amount of retrieved texts. The method generates a digest using a seed list which contains constituent nouns of a title. A digest generator repeats sentence selection until the seed list is empty, in which it calculates sentence relevance based on the overlaps between the seed list and a sentence, and it picks up the most relevant sentence and eliminates its constituent nouns from the seed list. It generated one to three sentences digests from 93 % of 13,562 newspaper articles and 61 of 62 economic reports. 98 samples of them are evaluated manually and more than 82 % of them are judged to be enough understandable to recognize outline of the source articles.

1 はじめに

本稿は、見出しを核（シード）とした文書のダイジェストを生成する実験についての報告である。

近年、文書の電子化が進み、インターネット／イントラネットに代表される新たな文書流通メディアの出現によって、計算機上で閲覧可能で再利用の操作を行える文書の量が爆発的に増加している。一方で、技術の高度化に伴う技術文書のボリュームの増加・多様化が進み、大量の文書を蓄積して再利用したいという要求が高まっている。筆者は、現在このような文書の再利用の過程をトータルに支援するシステムの構築を目指して研究を進めている。ダイジェスト生成は、文書再利用過程の第一段階である文書選別の過程を支援するためのものである。

大量の文書を利用するためには、まず、個々の文書の有用性を素早く判定し、利用目的にあった文書を選別することが重要である。そのためには、文書一覧に文書内容をイメージできるような情報を合わせて提示することが求められる。このような目的にあった情報としては、文書の見出しや抄録などがあるが、見出しが十分に文書内容を表現していない場合や、抄録がない場合も多い。また、特にオンラインで文書を閲覧する場合には、表示可能な文字数が限られるため、例え抄録が作成されていても長過ぎて一覧表示に適さないこともある。

文書の見出しは、要約の核となる概念は含んでいても完全な文（述語とモダリティを備えた文）の体裁をとっていないことが多いので、何の話題が取り上げているかは分かっても、どのような観点から、どのような論調で話題を取り上げているのか分からないことがあり、それだけでは十分でない。そこで、見出しを補足するような簡潔な情報（ダイジェスト情報）を抽出してみることにした。

亀田 [1] では、このようなダイジェストに関連して、見出しに含まれる単語を多く含む部分を（見出しに関連の深い）重要な部分として認定する手法（擬似キーワード相関法）を示しているが、実際に見出しに関連の深い部分を全て抽出するとどうなるのかなどについての詳細の報告はされていない。

そこで、本稿では、まず見出しのキーワードを含む文を全て抽出した結果を示し[1]たあと、抽出する文を絞り込むための手法を提案し、提案手法に基づいて作成したダイジェストの特性や問題点などについて分析する。実験対象としては、日本経済新聞の1990年12月の1カ月分の記事（約1.3万件）と、社内で流通している経済関係のレポート[2]の1996年1～6月分（62件）を使用した。

以下、第2章で、ダイジェスト作成の実験の詳細を報告し、第3章で生成したダイジェストの観察と問題点の分析を行う。

2 見出し関連文の抽出実験

2.1 見出しキーワードを含む文の抽出

今回の実験では、まず、見出しから日本語形態素解析ツール jmor[2] のキーワード抽出機能を使って、見出しに含まれる名詞を「見出しキーワード」として抽出した。jmor によって抽出される名詞（キーワード候補）とは、文中に含まれる名詞（サ変名詞・形容動詞語幹を含む）から機能語や数字・時詞・相対名詞などを除いたもので、ほぼ概念語に相当するものである。jmor には名詞の連続を複合語としてまとめて抽出する機能もあるが、今回の実験ではこの機能は用いず、個々の名詞を別々にキーワードとして抽出している。例えば、以下の例で【】で囲まれたものが抽出された名詞である。以下ではこのようなものを「名詞」と呼ぶ。

> 【神戸】沖【ヘリ】【事故】、【操縦】【士】の【遺体】も【発見】。
> 【General Motors】は【研究開発】を一つの【研究所】に【集中】を【計画】中

次に、記事やレポートの本文からも同様に名詞を抽出し、見出しキーワードと一致する名詞を含む文を全て抽出した。ただし、見出しキーワード以外の名詞が全く含まれない文は、ダイジェスト情報としては冗長と考え、抽出しなかった。この結果、新聞記事で元の7割ほど、レポートで4～5割ほどの量[3]の文が抽出された。1記事あたりの平均抽出量は、新聞で6文（361文字）、レポートで12文（827文字）だった。なお、この方法で1つの文も抽出できなかった割合（抽出失敗率）は、新聞で6.9%、レポートで 1.6% であった。（表1：太字の値は最頻区間に対応）。

通常の要約の場合、標準的な要約の分量は原文の 1/3 ～ 1/4 程度が目安であるといわれている [3]。

[1] ただし、本稿での見出しと本文中の文との関連性の判定法は亀田のものとは異なる。

[2] 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。

[3] EUC コードによるバイト数で比較した場合。

よって、抽出された文の量は、通常の要約としても少し多過ぎ、文書選別用のダイジェストとしては多過ぎるといえる。

表 1: 見出しキーワードが出現する文の本文に占める割合

文字単位の抽出比率	新聞		レポート	
	記事数	構成比	記事数	構成比
ALL	**2,237**	**16.5 %**	2	3.2 %
90% 以上	1,083	8.0 %	2	3.2 %
80% 以上	1,758	13.0 %	6	9.7 %
70% 以上	1,642	12.1 %	5	8.1 %
60% 以上	1,441	10.6 %	7	11.3 %
50% 以上	1,250	9.2 %	3	4.8 %
40% 以上	1,027	7.6 %	11	17.7 %
30% 以上	813	6.0 %	**13**	**21.0 %**
20% 以上	654	4.8 %	8	13.0 %
10% 以上	501	3.7 %	4	6.5 %
10% 未満	218	1.6 %	0	0 %
抽出失敗	938	6.9 %	1	1.6 %
total	13,562	100 %	62	100 %
平均値	64 %		49 %	
中央値	70 %		43 %	
平均抽出量	5.9 文 (361 文字)		12 文 (827 文字)	

2.2 抽出する文の絞り込み

見出しに関連する文を全て抽出するだけでは、文書選別用のダイジェストとしては量が多過ぎるので、何らかの手段によって抽出する文を選択して絞り込む必要がある。前掲の亀田 [1] では単に関連度の高い文から順に選択しているが、同じ見出しキーワードに関連する文を複数抽出しても冗長である可能性があり、短いダイジェストを作成するには効率が悪い。また、抽出結果は処理の停止条件（関連度の閾値）に依存するので、抽出する前にそれぞれの文書の種類に応じた最適な停止条件を見つける必要がある。

そこで、出来るだけ文数が少なくなるように、また、見出しキーワードと一致する名詞の異なり数が最終的には絞り込み前と同じになるように、文の重要度の算定基準を調整しながら文を選択して抽出するアリゴリズムを考えた[4]。

このアルゴリズムは、

1. 見出しキーワードと一致する文中の名詞の異なり数
2. 見出しキーワードと一致する文中の名詞の延べ数
3. 見出しキーワードと一致しない文中の名詞の延べ数

の 3 つの値[5]をこの順で[6]比較することで文の重要度を評価し、重要度順に文を選択して抽出するという操作を繰り返すものである。文の重要度の評価は、抽出した部分にはまだ出現していない見出しキーワード（未出現キーワード）に基づいて、文を選択するたびに繰り返し行う。選択操作は、見出しキーワードを全て含むような文の集合が得られた時点で終了する (図 1参照)。

```
未出現キーワード列 ← 見出しキーワード。
抽出結果 ← 空。
while 未出現キーワード列が空でない。
 do (重要文選択)
    未出現キーワード列に基づき文の重要
    度を評価。
    重要度最大の文（複数可）を選択。
      （1つも選べなければ処理終了。）
    選択された文を抽出結果に加える。
    選択された文中の見出しキーワードを
    未出現キーワード列から除く。
 od
```

図 1: ダイジェスト情報抽出のアルゴリズム

これにより、新聞記事で原文の約 3 割、レポートで 1 割程度の量の文に絞り込めた (表 2)。例としてレポート中の長めの記事のダイジェストを示す。

- IBM の OS と Notes の戦略
  (抽出率 5%。原文は 8 段落、91 行)
  … OS/2Warp は一般のデスクトップ用だが、Merlin は企業用の OS であり、一般のデスクトップの OS では Windows に負けたのを暗黙のうちに認めたと解釈されている。…世界中の企業でのデータや情報の 70% は IBM のコンピュータ、特に大型機に存在すると見られ、Notes によってそれらを Internet や intranet を介して処理しようという IBM の戦略は妥当であろう。…

3 生成されたダイジェストの観察とそれに基づく考察

3.1 主観的な評価

今回の実験で作成しようとしている文書選別用のダイジェストと一般の要約との違いは、以下の点に

[4] 厳密には以下に示すアルゴリズムでは最小の文数が抽出されることは保証されない。

[5] 見出しキーワードとの関連性が高く、かつ、抽出することで増える情報量が多い、という評価基準に相当。見出しキーワードに一致しない文中の名詞の異なり数を考慮してもよいが、実験的にはほとんど変化が見られなかった。

[6] ただし 2.1と同様、見出しキーワード以外の名詞 (3) が 0 のものは抽出しない。

表 2: 絞り込み後の抽出比率 (文字単位)

文字単位の抽出比率	新聞		レポート	
	記事数	構成比	記事数	構成比
ALL	450	3.3%	0	0 %
90% 以上	43	0.3%	0	0 %
80% 以上	186	1.4%	0	0 %
70% 以上	359	2.7%	0	0 %
60% 以上	587	4.3%	0	0 %
50% 以上	944	7.0%	1	1.6 %
40% 以上	1,506	11.1%	3	4.8 %
30% 以上	2,061	15.2%	0	0 %
20% 以上	**2,765**	**20.4%**	5	8.0 %
10% 以上	2,673	19.7%	10	16.1 %
10% 未満	1,050	7.7%	**42**	**67.7 %**
抽出失敗	938	6.9%	1	1.6 %
total	13,562	100 %	62	100 %
平均値	32 %		11 %	
中央値	27 %		7 %	
平均抽出量	1.9 文 (135 文字)		1.5 文 (121 文字)	

ある。

- 要約は、文章の主要な内容を残らず取り上げて簡潔にまとめることが望ましい。
- 文書選別用のダイジェストは、本文で書いてありそうなことが推測できる範囲で、なるべく少い量であることが望ましい。

つまり、文書選別用のダイジェストには、文書の内容の簡潔なサンプルであればよく、必ずしも文書の主要な内容を網羅している必要はないという性質がある。例えば、いくつかの項目（類例など）が列挙されているような文書については、列挙されている一つでも抽出できれば十分な場合がある。というより、列挙されているものを全て抽出してしまうと、文書選別のためには分量が多過ぎて、選別に余計な時間がかかってしまう。

以上の考えより、ダイジェストの評価の手始めとして、今回作成したダイジェストを見出しと併せて読んだ時に文書内容の推測ができるかという観点から主観的に評価してみた。

評価対象としたのは、1990 年 12 月末分に相当する 87 記事 (記事の元データの末尾 1,000 行に含まれていたもの) と 6 月分のレポート (11 記事) である。

筆者の主観的評価でダイジェストとして明らかに不適切だったものは、新聞記事で 16 件 (18%)、レポートでは 1 件 (9%) だった。「明らかに不適切」というのは、

- 文の抽出に失敗したもの（新聞で 4 件）
- 見出しとダイジェストの関係が判然としないために混乱するもの
- 抽出された文の間の関係が判然としないために混乱するもの
- 見出しに書いてあること以上の情報を得られないもの

などである。直観的にいえば、本文を読まなくても雑談の話題として提供できる程度に理解できた気になれば、不適切とはしなかった。

新聞記事で明らかに不適切とした 16 件の内訳は、「私の履歴書」「交遊抄録」「トップ群像」や用語解説などのコラムが 9 件、対話形式の特集記事（「ハイテク分野…」）など 5 件、通常の事件記事 2 件である。以下に示すように、事件記事以外は、見出しが記事の要旨と直接の関係をもたないものがほとんどであった。

「不適切なダイジェスト」の見出し (新聞)

- 伊勢湾、不明の漁民、遺体で。**[文抽出失敗]**
- アメ横に７８万人、家族連れらで“大渋滞”。
- 丸紅会長春名和雄氏（３０）小さな流れ――生かされてきた実感（私の履歴書）（終）
- 来年メセナ正念場（文化往来）
- ウサギとホウレンソウ――慶応大学教授鷲見洋一氏（交遊抄）**[文抽出失敗]**
- 名より実の大使役（消息）
- 遊んで買い物、子供の街（消費の現場）
- 会社概要――キリンビール（トップ群像）**[文抽出失敗]**
- キリンビール――記者から（トップ群像）**[文抽出失敗]**
- 第３２４話、人材難の中で、社長奮闘（７）自らゆとり持ち活路（サラリーマン）終
- 電算機の速度を示す単位群。
- ハイテク分野、広がる新単位革命。
- ハイテク分野、広がる新単位革命――電算機、「ＣＵＰＳ」が登場。
- ハイテク分野、広がる新単位革命――超電導、「テスラ」が主役に。
- ハイテク分野、広がる新単位革命――半導体、「メガ」から「ギガ」へ。
- ２１世紀にこうなる、金沢工業大学経営科学研究所所長城阪俊吉氏に聞く。

レポートのダイジェストで明らかに不適切としたものを示す。

○ Windows NT4.0 の問題点。
Windows NT 4.0 は Windows95 のインターフェイスを追加したもので、企業マーケットではヒット商品になると見られており、８月に出荷すると見られていたのが、９月かそれ以後になると噂されている。…

これは、明らかに不適切とまではいえないかもしれないが、問題点が一つも現れていないので不適切とした[7]。反対に、不適切かもしれないが結局不適切とはしなかったダイジェストに次のものがあった。

○ Javaの問題点の修正
…ソフトウェア会社SymantecはCafeというJava開発環境用ソフトウェアをすでに発表したが、最近Web上でinteractiveなデータベースのアプリケーション・プログラムを書くのに便利なライブラリーを発表した。…

これにも問題点が一つも現れていないが、「便利なライブラリ」というのは修正結果の一つと読めるので不適切とはしなかった[8]。

以上のように、主観的ではあるが、8割以上は抽出された情報と見出しを併せて読むだけで、雑談で話題提供できるくらいの情報を把握できた気になるので、このままでも、ダイジェストとして利用可能と思われる。特に、経済動向のレポートや新聞の事件記事のように、世間で何が起きているかが大体把握できればよいものであれば、この程度でも十分であろう。

3.2 先頭段落のみを抽出する場合との比較

特に新聞記事では、先頭の段落が重要であるとされており、そのことをヒューリスティックとして用いて情報抽出や要約作成を行う既存研究が多い(例えば、[4]、[5])。

今回の手法では、文書における文の出現位置のヒューリスティックは用いていないので、その必要性を検討するために、今回の手法で作成したダイジェストと記事の先頭段落との関係を調べてみた。調べた項目は、次の3点である。

1. 先頭段落中の文がどれだけ抽出されたか(表3)
2. 今回作成のダイジェストに先頭段落中の文がどれくらいの割合で含まれるか(表4)
3. 長さはどちらが短いか(表5)

表3、4では、抽出が失敗した記事(見出しキーワードを含む文がなかったもの)以外の実験対象記事全てについて、個々の記事ごとに求めた1、2の値を文数単位の比率で集計してある。

[7]本文に書いてありそうなことのイメージはつかめるが、「NTって…というのが問題なんだってね」のような形で別の人との会話の話題にするレベルにはない。

[8]「Javaってdbのアプリ書くのが大変だったらしいけど、Symantecという会社が便利なライブラリーを発表したらしいよ」と話題にできる。

表3: 先頭段落中の文の抽出比率

	新聞		レポート	
	全体	絞込後	全体	絞込後
平均値	73 %	46 %	49 %	16 %
中央値	75 %	40 %	45 %	9 %
最頻区間	ALL (46 %)	[50%, 60%) (22 %)	[40%, 50%) (16 %)	(0%, 10%) (30 %)

下段の()内は最頻区間に含まれる記事数の全体に占める割合

表4: ダイジェストに先頭段落起源の文が占める比率

	新聞		レポート	
	全体	絞込後	全体	絞込後
平均値	54 %	72 %	49 %	68 %
中央値	45 %	ALL	40 %	ALL
最頻区間	ALL (29 %)	ALL (56 %)	ALL (20 %)	ALL (61 %)

下段の()内は最頻区間に含まれる記事数の全体に占める割合

表5: 今回作成のダイジェストと先頭段落の比較表

	新聞		レポート	
	ダイジェスト	先頭段落	ダイジェスト	先頭段落
内容適切	71 記事	76 記事	10 記事	10 記事
より短い	40 記事	19 記事	9 記事	0 記事
全く同一	7 記事		0 記事	

長さ(「より短い」以下)は両方の内容が適切な場合にのみ比較

表3によると、先頭段落中の半数程度以上の文に見出しキーワードが出現していることが分かる(「全体」の抽出比率)。今回の手法で絞り込んだ後でも、先頭段落中の文の抽出比率は、全体の平均抽出率(新聞約3割、レポート約1割: 表2参照)の1.5倍程度ある。表4によると、今回作成のダイジェスト(「絞込後」)に先頭段落起源の文の占める割合は平均約7割と高率で、また、先頭段落起源の文だけからなるものが大半を占めている。複数の段落からなる記事に限っても[9]、ダイジェストの半数程度(新聞42%、レポート53%)は先頭段落起源の文しか含んでいない。これらは、先頭段落を特別に重視しなくても、ある程度先頭段落の重要性が反映された結果が得られそうなことを示唆している。

表6: 原データにおける段落数の分布

	新聞		レポート	
母集団	記事数	構成比	記事数	構成比
単一段落	3,062	24%	10	16%
複数段落あり	9,562	76%	51	84%
全体	12,624	100%	61	100%
平均値	4.0 段落		5.1 段落	
中央値	3 段落		5 段落	

表5では、主観的な評価の対象とした記事を対象に内容の適切性と長さの比較結果を集計しある。長さについては両者の内容が不適切でない場合にかぎり

[9]表4の「最頻区間」の値から表6の「単一段落」の構成比の値を引いて、「複数段落あり」の構成比と比較する。

文字数 (バイト数) で比較した。これから、新聞、レポートとも今回作成のダイジェストの方が簡潔である（短い）という傾向が見てとれる。

先頭段落をダイジェストとして不適切としたものは、新聞で11件、レポートで1件だった。新聞では、コラムなどで見出しと同じものが第一段落となっているものなど、ほとんどデータ処理上の問題であった。レポートで不適切としたのは、以下のものである (記事は2.2のダイジェストの例と同一）。

- IBM の OS と Notes の戦略
  IBM の OS/2Warp は Windows3.1 と Windows95 に対抗する一般のデスク・トップ用 OS であり、Internet にアクセスできるものであったが、その後継としての Merlin のベータ版を発表した。Merlin は Windows NT に対抗するものであり、Java の applets やアプリケーション・プログラムを web browser に関係なく native mode で走らせうる。Java を取り入れた OS としては最初のものであり、Microsoft は Windows95 の次の version の Nashville で初めて Java をサポートする。Microsoft は Java に対抗する Microsoft の独自の ActiveX も全力をあげて開発しており、その最新版を6月に ActiveX Control Pad という Web ページ作成用ツールと共に発表した。また Direct3D というインターフェイスも発表したが、これはゲーム、娯楽用ソフトウェア、Web ページに高度のグラフィックスを追加するためのものである。Microsoft はクリスマスまでにこれを使ったゲームを30位発表する。また Microsoft は 64 ビットの Windows NT を開発中とみられている。

内容の相対的優劣の比較も試みたが微妙で数値では示せない。しかしながら、レポートは、長さの差が顕著であり[10]、今回の手法によるダイジェストの方が有効であると判断される。新聞については、先頭段落の方が内容的には若干勝っていた。傾向としては、冒頭の文がダイジェストに含まれていない場合に先頭段落の方が理解しやすいことが多く、逆に今回作成のダイジェストの方が優れていると判定できたのはかろうじて以下の1例だけであった。

見出し	上場企業の中途採用依然活発。
先頭段落	リクルートリサーチ（本社東京、社長森村稔氏）は「上場企業における中途採用調査」の結果をまとめた。
ダイジェスト	リクルートリサーチ（本社東京、社長森村稔氏）は「上場企業における中途採用調査」の結果をまとめた。¶…八九年は六五・一%で、人手不足の深刻化により中途採用がより活発になっている。…¶…¶…

これらをまとめると以下のようになる。

- 短いダイジェストを生成するという意味では、今回の手法の方が先頭段落を全て抽出するより優れている。
- 内容の理解しやすさからすると、新聞では、先頭段落を全て抽出する方が若干優れている。
- 冒頭の一文は多くの場合ダイジェストに含めた方が理解しやすくなる。
- 今回の観察に基づく暫定的結論

  新聞のダイジェストについては、先頭段落だけを対象に今回の手法を適用するのがよく、種々の文書のダイジェストを同じ方法で生成するのであれば、冒頭の一文だけは別扱いで抽出し、その他を今回の手法で抽出するのがよい。

[10] レポートの先頭段落の長さは、実験対象全体の平均で10文。

3.3 その他の観察と考察

新聞87記事とレポート11記事を主観的に評価した際に気がついた点などについて考察する。

文抽出失敗の原因

まず、今回の手法では全く文を抽出できなかった場合について考察する。

表 7: 見出しキーワードの中で本文にも出現するものの割合

新聞

母集団	抽出数 平均	抽出数 最大	出現数 平均	出現数 最大	平均出現率 (出現数 / 抽出数)
抽出成功	7.1	17	5.5	15	77 %
抽出失敗	3.9	14	.43	9	9 %
全体	6.9	17	5.2	15	73 %

レポート

母集団	抽出数 平均	抽出数 最大	出現数 平均	出現数 最大	平均出現率 (出現数 / 抽出数)
抽出成功	3.4	6	2.7	5	79 %
抽出失敗	3.0	0	0	0	0 %
全体	3.4	6	2.6	5	78 %

表7は、各記事の

- 見出しから抽出したキーワードの異なり数（抽出数）
- 見出しから抽出したキーワードで本文にも出現したものの異なり数（出現数）

を、1文でも抽出できた記事（抽出成功）、1文も抽出できなかった記事（抽出失敗）にわけて集計したものである。これによると、抽出が失敗した記事の方が抽出数は少いものの、平均で3～4個程度は見出しキーワードが抽出されており、特にレポートでは文の抽出が成功した記事と大きな違いはない。

つまり、何らかの方法で本文中の名詞と見出しキーワードが照合できれば、文が抽出できない記事を減らすことが可能であろうと推定される。

これとは別に、今回は見出しキーワード以外の名詞を含まない文を抽出していないので、そのために文の抽出が失敗した記事もある。これについては、その前後の文も合わせて抽出することも考えられる。しかしながら、このようなものはわずか[11]であったので、あまり有効でないと考えられる。

キーワード照合失敗の原因

見出しキーワードが本文に全く出現していない事例を観察した限りでは、同じ内容を示すのに見出しと本文で異なる語が使われていたことによるものが多かった。このようなものを分類すると以下のようになる。

- 複合語 – 構成語 の関係にある語:
  「不明」 (見出し) と「行方不明」 (本文)
- 類語:
  「遺体」 (見出し) と「死体」 (本文)
  「漁民」 (見出し) と「漁業、X さん」 (本文)
- 略称 – 正式名称 の関係にある (固有) 名詞:
  「公取委」 (見出し) と「公正取引委員会」 (本文: 初出)、「公取委」 (本文：2番目以降)
  →より重要な冒頭の文が抽出されない
- その他の関連語:
  「経済 / 動向」 (見出し) と経済の変化を表す語 (「暴落」「安定基調」等)(本文)

その多くは、複合語 – 構成語の関係を記述してある辞書や、類語や上位語 – 下位語の関係を記述したシソーラスを利用することで解決可能に見える。ただし、もっと多くの事例を分析し、どのような関係を扱わなければならないのかについて、体系的に整理する必要がある。また、これらの情報を利用するためには、例えば、構成語が照合した場合、類語が照合した場合にどういう重みをつけて解釈するのか、という難しい課題があり、処理が重くなった割にはあまり効果があがらないという事態も予想される。見出しや抽出対象部分の情報量と出現単語の関係などに関するよいモデルが必要である。

[11] 新聞の「抽出失敗」の部分。記事あたり平均では1個にも満たない。

複合語や略称の問題に関しては、別の解決方法として、亀田 [1] の擬似キーワード相関法のように単語表記を構成する文字の一致を使ってキーワードを照合したり、角田 [6] が TV ニュースのキャプションと新聞記事の見出しなどとの照合に用いているような文字レベルの照合でキーワードや文の関連性を評価することも可能である。この方が上記の解決法に比べ処理が軽くてすみ、特に、検索システムと組み合わせて用いる場合などでは、速度性能の面などで有望である。

ただし、筆者の長期的目標は文書の再利用過程のトータルな支援にあり、例えば、複数の文書から同じ事柄に関する部分を抽出して文書を比較する処理などへの拡張を意図しているので、前者の解決策を追求したいと考えている。

可読性の低下の原因: 未解決の照応先

ダイジェスト中の照応表現の照応先や接続表現の接続先など、読解に重要な要素がダイジェストに含まれていないために、理解が難しくなっているものが見られた。これらのうち、照応表現が名詞を修飾しているもの (「この～」「このような～」など) については、その名詞が初めて登場する文も抽出するとうまくいく例が多く見られた。また、接続表現については、接続表現を落してしまえば問題ないような例もみられたが、そのような操作が安全かどうか (誤解を生まないかなど) を判定するのは難しそうである。この問題についても、もっと多くの事例を集めて、体系的に整理する必要がある。

以下、レポートより事例を挙げる。

例1

○ Network Computer の将来

…こういう端末機が現在約30,000,000台も使用されているが、このマーケットを活性化するためにIBM は社内で Client と呼ばれている Network Computer によって今迄 AS/400 に接続していたターミナルを置き換えると 5 月に発表したし、Wyse Technology や SunRiver Corp. などのターミナル製造会社は今までの端末機を新しいハードウェアやソフトウェアによって Network Computer に変えようとしている。…

この例では、「端末機」の抽象性が高いので冒頭の「こういう」を削除してしまうわけにはいかない。この場合は、この前の文に「端末機」が出ていた (初出でもある) ので、それを上例の冒頭の

「…」の後ろに追加すればほぼ完全なダイジェストになる（次例）。

> … Network Computer を最初に取り上げるのは、現在中型機や大型機をデスクトップや座席予約や倉庫に接続する端末機を販売している会社であろう。こういう端末機が現在約30,000,000台も使用されているが、[以下上例に同じ]

例2

○ Apple Computer が Windows 乗り入れの強化により再建中

> G.Amelio は社内の機構を改革し、Macintosh の機種を半分に減らして開発費を減らし、約3,000人の社員をレイオフしつつ、Apple Computer を建て直している。…しかし Hancock はソフトウェアをよく知っており、Apple 再建の成否は開発が遅れ続けている Copland にかかっているので、Hancock は妥当な人事と見られている。…交渉がまとまれば Microsoft は QuickTime を Internet Explorer に組み込めるようにし、同時に Apple は Windows のマルチメディア技術のサポートを強化することになろう。…

この例では、「…しかし Hancock は」の部分で文の主題（既知情報）として扱われている要素、すなわち Hancock という人が、文章全体の主題とどういう関わりがあるのかが分からない。また、「しかし」によって対比されている事態が不明であることも問題である。

1つ目の問題については、例1と同様「Hancock」という語が初めて出現した文の主文だけを取り出せば以下のように表示することが可能である。

> … Amelio は…、研究開発の最高の担当者 chief technology officer として53才の Ellen Hancock を任命した。…しかし Hancock はソフトウェアをよく知っており [以下上例に同じ]

2番目の問題は、少し深刻である。この場合は「しかし」を落してしまえばすみそうだが、どういう場合に接続詞を落せるのかを判定する方法が必要である。筆者の判断で選んだ「しかし」による関係先は以下のものであり、これを自動的に抽出するためには事態の好悪などについての深い理解が必要である[12]。

> Hancock は IBM という巨大な会社で育っているので、直ちに6,000人の血気盛んな若いエンジニアやプログラマと管理スタイルの上でぶつかり合うのではないかと見られている。

[12] この場合は文末の表現「見られている」の一致によって抽出できる可能性もある。

4 まとめ

本稿では、見出しに含まれる名詞キーワードを核としてダイジェストを生成する方法を提案した。また、名詞キーワードの照合に関する問題点や照応表現を扱う必要がある場合などを提示した。

前者に関しては、見出しに含まれる名詞キーワードを含む文を全て抽出するだけでは十分に短くならないが、本稿で提案したような手法で抽出する文を絞り込むことで、文書選別に役立つようなダイジェストが作成できることを示した。

ただし、今回行った評価は筆者の主観によるものであり、客観性に乏しいという問題がある。しかしながら、3.1節で述べたように、文書選別用のダイジェストには必ずしも文書の主要な内容を網羅している必要はないという性質があり、本来的に絶対的な評価が困難である。今後は、例えば文書選別作業の効率をどの程度改善できるのか、というような効用の面からの評価してみたいと考えている。

後者に関しては、語彙的結束性を使った文書処理とも密接に関わりあう問題である。特に文章で取り上げられた話題を特徴づける一連のキーワードを抽出するために、人間用のシソーラスを利用した文章中の関連語の連鎖を抽出を試みた Morris ら [7] の研究には、見出しキーワード照合の問題点の整理や提案手法の拡張のために有益な示唆が見られる。今後、こういった研究も参考に、問題点の整理とアルゴリズムの改良を行っていきたいと考えている。

参考文献

[1] 亀田雅之：擬似キーワード相関法による重要キーワードと重要文の抽出, 第2回年次大会, pp. 97–100 言語処理学会 (1996).

[2] 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).

[3] 佐久間まゆみ：序章, 佐久間まゆみ（編）, 文章構造と要約文の諸相, pp. 7–17, くろしお出版, 東京 (1989).

[4] 松尾比呂志：抽出パターンの階層的照合に基づく内容抽出法, 情処研報 NL-99-2, 情報処理学会 (1994).

[5] 船坂貴浩, 山本和英, 増山繁：冗長度削減による関連記事の要約, 情処研報 NL-114-7, 情報処理学会 (1996).

[6] 角田達彦, 大石巧, 渡辺靖彦, 長尾眞：キャプションと記事テキストの最長一致文字列照合による報道番組と新聞記事との対応づけの自動化, 情処研報 NL-115-11/FI-43-3, 情報処理学会 (1996).

[7] Morris, J. and Hirst, G.: Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, *Computational Linguistics*, Vol. 17, No. 1, pp. 21–48 (1991).