

提題表現に基づく重要段落抽出

中野 滋 徳[†] 足 立 顕^{††} 牧 野 武 則[†]

日本語には係助詞「は」「も」などを用いた強調表現がある。係助詞は語に付属し、付属した語句を文および段落の中心的話題である提題表現とする機能を持つ。本稿では、係助詞を伴う提題表現に基づいた重要段落の抽出について述べる。文章中に記述している語をベクトル化し、段落ごとに得たベクトルとの内積値を計算することで重要段落を得る。実験では、ニュース報道記事を用いた比較的大きなコーパスを対象に抽出精度の評価を行った。その結果、毎日新聞記事 61.2%、日経新聞記事 77.9%の抽出精度を得た。考察では、社説を用いた実験結果について述べる。

キーワード: 重要段落, 重要文抽出, 提題

Key-Paragraph Extraction Based on Topicalized Expression

SHIGENORI NAKAN[†] AKIRA ADACHI^{††} and TAKENORI MAKINO[†]

Japanese has the emphasized expression with a surface marker of “HA” or “MO”. The surface marker is called “*Kakari-Joshi*” which focuses words and phrases as topicalized expression in sentences. This paper describes a method of key-paragraph extraction based on topicalized expressions. The method extracts key-paragraph by inner product which is calculated between article vector and each paragraph vector. The vector is composed by frequency of words in the article. In the experiment, the accuracy was evaluated by using a large number of newspaper articles. The accuracy of 61.2% is achieved for the Mainichi, and 77.9% for the Nikkei. In consideration, we describe additional experiments about the result which is obtained by using editorials.

Keywords: *Key-paragraph, Extraction, topicalized-words*

1. はじめに

大量の文書情報を扱う現代社会において、

少ないコストで、その全容を把握できる抄録や要約を自動的に生成するシステムが求められている。

要約研究では、「要約」を原文の大意を維持したまま文書情報を圧縮する技術と位置づけ、テキストの解釈、重要部分の抽出、要約

[†] 東邦大学 大学院理工学研究科 情報科学専攻
Department of Information Sciences Toho University
^{††} 富士通株式会社 情報メディアソリューション本部
Division of Media Solution, Fujitsu Co. Limited

文としての解釈を行うことで要約を作成する研究が報告されている。これらの研究は、テキスト自動要約に関する研究動向¹⁾として整理されている。

要約を生成するためには、まず、文書中の重要部分を認識する必要がある。重要部分の認識の研究では、ベクトル空間モデルを用い文間の関連度を計算することで重要部分を取得する手法²⁾や、文より得た単語を要素とするベクトルを用い、文間の結束度をベクトルの内積値として重要文を抽出する手法³⁾の報告がある。

本稿では、「は」「も」など係助詞を伴う提題部分の表現を利用することで文書中の重要部分の抽出を行う手法について述べる。文書全体より得た語彙頻度を要素とするベクトルと段落単位に得た語彙頻度を要素とするベクトルの内積値を計算することで重要段落を得る。

文間での類似度や結束度から重要部分を得るのではなく、文書全体との比較により文書中の重要部分を認識し、比較的大きなデータを対象とする点で従来研究と異なる。

従来研究では解析精度を評価するために、人手により処理結果の判定を行っている。人手による評価では、評価者のレベルおよび関心領域の違いにより、評価結果が左右する。複数人で評価し、その評価結果を整理することでシステムの精度を測定している。そのため評価する文書量は限られ、少ないデータの中での評価結果となっている。

本実験および評価では、比較的、文章の構造が明確なニュース報道記事を用いる。記者ハンドブック⁴⁾によれば、ニュース報道記事は結論を記事の先頭に記述することを推奨している。情報の発信者が重要とする部分が記事の先頭にある可能性が高い。この特徴を用いることで、比較的大きなデータを用いて手法の評価が可能である。解析により得た重要

箇所が記事中のどの部分の記述かを判定することで、本手法の評価を行う。

本稿では2節で文中の提題表現の役割について述べ、3節で重要段落の決定方法について述べる。4節で評価方法について述べ、5節で実験ならびにその結果について述べる。6節で考察を行う。

2. 提題表現の役割

提題表現とは、係助詞「は」などを伴って文中で表現したものである。「文の主題」と呼ばれる名詞句を中心とした表現である。

提題表現が文および要約文中に与える影響について次に述べる研究報告がある。

三上の著書「象は鼻が長い⁵⁾」において、提題表現「X八」が句点を越え複数の文に影響と与えると述べている。提題表現が他の文や段落、文章全体に影響を与える可能性があることを指摘している。

「文章構造と要約文の諸相⁶⁾」では、要約文における原文の残存傾向を調査している。調査結果では、筆者の意見や主張を表す部分に、原文の提題表現が、「それが何についての内容か」という対象を示す手段として、提題表現が要約文中に残される傾向があることを指摘している。

「文章論の基礎問題⁷⁾」では、係助詞「は」の特徴は一個の文を越えて、文章の文脈の中で、題目とか、中心概念とか、執着の対象など、論者に最も気になるものを、まず掲げるための表示記号に使われる傾向があると述べている。

提題表現は、情報の発信者が発信する事柄について、受信者に強い印象を与える目的で用いる表現であると言える。

3. 重要段落の決定方法

本稿で提案する重要段落の決定手法につい

て述べる．

重要段落の決定には，文章全体から得た情報と評価単位とする部分より得た情報を比較することにより重要部分の判定を行う．

名詞を中心とした語彙の出現頻度を解析に用いるベクトル構成要素とする．文章全体のベクトルと段落単位のベクトルとの内積値を計算し，最も大きな値となる段落を重要段落と判定する．文章全体より得たベクトルを記事ベクトル，段落より得たベクトルを段落ベクトルと呼ぶ．

ベクトルの内積値を尺度として用いる場合，ベクトルを構成する各要素が独立である必要があるが，類義語などとの依存関係にあるものが存在する．意味的に依存関係にあるかの判断としてシソーラスなど用いる場合があるが，本手法では名詞を中心として，表層表現が異なれば独立した要素として扱う．

ベクトル要素の対象とする品詞は，名詞，固有名詞，サ変名詞とする．機能語など構文上の役割を持つ語および，直接，対象を示さない動詞，形容詞や，時，程度などを表す副詞類は対象としない．

段落ベクトルの算出にあたり，文中に提題表現として用いられた要素に対して重みを与えるための段落単位の対角行列を，提題行列と呼ぶ．

提題行列に重みを与えるにあたって，対象となる要素の例を表 1 に示す．

表 1 提題行列の要素

- 例 1) 太郎の娘の花子は…
- 例 2) 太郎と花子も…
- 例 3) 昨日買った車では…

表 1 の太字の部分は，提題表現とする語彙を示す．提題表現とする語彙は，用言の連体形または連体化助詞などにより示された連体修飾部分を除き，係助詞までの区間に現れた

語彙を採用する．表 1 の例 1) の「太郎の娘の」の部分は，「花子」に対する修飾部分とし，提題表現の対象としない．

重要段落を求めるための計算式を表 2 に示す．

表 2 重要段落の計算に用いる情報

記事ベクトル $S = (s_1, s_2, \dots, s_n)$
段落ベクトル $P_j = (p_{j1}, p_{j2}, \dots, p_{jn})$

$$\text{提題行列 } T_j = \begin{pmatrix} t_{j1} & 0 & \dots & 0 \\ 0 & t_{j2} & \dots & 0 \\ \vdots & 0 & \dots & \vdots \\ 0 & 0 & \dots & t_{jn} \end{pmatrix}$$

S の構成要素は，記事中出现する語彙頻度を表し， P_j の構成要素は，段落中出现する語彙頻度を表す． j は段落番号を示す．

記事ベクトルと段落ベクトルの関係は， $S = \sum P_j$ となる． T_j の対角要素は，式 (1) に示す値を指定したものである．

$$t_{ji} = \begin{cases} \alpha & \text{提題表現 (} P_j \text{ 段落内)} \\ 1.0 & \text{提題以外} \end{cases} \quad (1)$$

式 (1) の α の値は，提題表現として用いた要素に対する重みを指定する．

提題行列を段落ベクトルの各要素に式 (2) を用いて適用する．

$$P_j T_j = (p_{j1} \cdot t_{j1}, p_{j2} \cdot t_{j2}, \dots, p_{jn} \cdot t_{jn}) \quad (2)$$

提題行列適用後の段落ベクトル $P_j T_j$ と記事ベクトル S との内積値 (sim) を式 (3) により求める．

$$sim(S, P_j T_j) = \frac{\sum s_j p_{ji} t_{ji}}{\sqrt{\sum s_j^2 \sum (p_{ji} t_{ji})^2}} \quad (3)$$

式 (3) をすべての段落に対し適用し， $max(sim(S, P_j T_j))$ となる j を重要段落とする．

4. 評価方法

本実験では、ニュース報道記事を用いる。新聞記事の制作基準に、新聞記者が用いている資料として記者ハンドブック⁴⁾がある。

そこには新聞記事のニュース報道記事の書き方として「逆三角形」構造で記事を作成することを推奨している。「逆三角形」構造とは、第一段落に結論を置き、続いて経過的に重要な事柄を記述することを指す。

このことは、情報の発信者である記者が、重要と判断する部分を記事の第一段落に記述していることを意味する。

つまり、第一段落を重要部分と認識することができるかどうかで検出精度を測定することができる。

本実験では、文章の構造情報を用いず、3節で述べた手法により重要段落を決定し、得た重要段落が記事中の第一段落であるかを判定し手法の評価を行う。なお、本手法で重要段落と判定した結果が第一段落に所属する割合を、以後正解率と呼ぶ。

5. 実験

5.1 実験対象データ

毎日新聞コーパス(1998-99年)および日経新聞コーパス(1998年)から、ニュース報道記事に該当するものを表3に示す基準により選択し実験データとする。

表3 記事選択基準

- 朝刊一面の記事本文で段落数5段から10段の記事を選択。
- タイトル部分に[社説][余録]等の記事種別の記載があるものを除去。
- 複数の記事を含む恐れがある小見出し付きの記事を除去。
- 文中にスペースが存在する記事は表形式の記事として除去。

毎日新聞コーパスと日経新聞コーパスから得た実験データの詳細をそれぞれ表4、表5に示す。なお、毎日新聞コーパスと日経新聞コーパスより得たデータはそれぞれ毎日新聞データ、日経新聞データと呼ぶ。

表4 実験データの詳細(毎日)

項目	件数	記事平均	段落平均	文平均
記事数	979	-	-	-
段落数	5,809	5.9	-	-
文数	11,766	12.0	2.0	-
文字数	1,393,350	1423.2	239.9	118.4

表5 実験データの詳細(日経)

項目	件数	記事平均	段落平均	文平均
記事数	606	-	-	-
段落数	4,124	6.8	-	-
文数	11,151	18.4	2.7	-
文字数	1,114,058	1838.4	270.1	99.9

毎日新聞データと日経新聞データでは、1記事に含む段落数の平均値と1文あたりの長さの平均値に特徴がある。毎日新聞データでは、1文あたりの長さが日経新聞データよりも長い。一方、1記事に含まれる段落数は日経新聞データの方が多い傾向を得た。

5.2 実験システム

実験には、形態素および品詞の情報を得るため、奈良先端科学技術大学院大学松本研究室が公開している形態素解析器「茶筌」⁹⁾を用いる。

「茶筌」の解析結果より、品詞情報として「名詞-一般」「名詞-固有名詞」「名詞-サ変接続」を抽出し、記事ベクトル、段落ベクトルを求めるための要素とした。

提題行列には、表記が「は」「も」で「茶筌」が「助詞-係助詞」とした形態素を伴った語彙を対象とした。形態素解析結果では、提題部分が複数の形態素に分割した場合、形態素解析結果に従いそれぞれ独立したものと

して提題行列の要素として採用した。

これらの情報をもとに，第一段落（形式段落）を重要段落と判定した割合（正解率）を自動的に取得し，提題行列に与える重み（ α ）を変えることでどのように値が変化するかを測定する実験システムを構築した。

5.3 実験結果

提題表現が与える影響を評価するため，毎日新聞データから重要段落を抽出する実験を行った。

提題表現が情報の発信者が，発信する事柄について受信者に強く印象づけるために用いられる点に注目し，重要段落の計算方式に基づいて，提題表現を用いた要素の段落に対してのみ重み（ α ）を適用する実験を行った。この実験をケース 1 と呼ぶ。

次に，文章全体に影響を与える可能性をもつという提題表現の特徴に注目し，式 (1) の提題行列部分の定義を表 6 に示すように差し替えて，提題行列 T を再定義した。つまり一度でも用いられた提題表現の要素に対して，同一の語句を用いた段落にも共通に重み（ α ）を適用する実験を行った。この実験をケース 2 と呼ぶ。

表 6 ケース 2 の提題ベクトル（再定義）

$$\text{提題行列 } T = \begin{pmatrix} t_1 & 0 & \cdots & 0 \\ 0 & t_2 & \cdots & 0 \\ \vdots & 0 & \cdots & \vdots \\ 0 & 0 & \cdots & t_n \end{pmatrix}$$

$$t_i = \begin{cases} \alpha & \text{提題表現 (記事内)} \\ 1.0 & \text{提題以外} \end{cases}$$

ケース 1，ケース 2 に示した基準を段落（形式段落）単位に対して適用した実験結果を得た。

図 1 に毎日新聞データを用い段落単位に得た実験結果を示す。縦軸に第一段落を重要段

落と判定した割合を正解率として示し，横軸に重み（ $1.0 \leq \alpha \leq 3.0$ ），変化量 0.05 で示す。

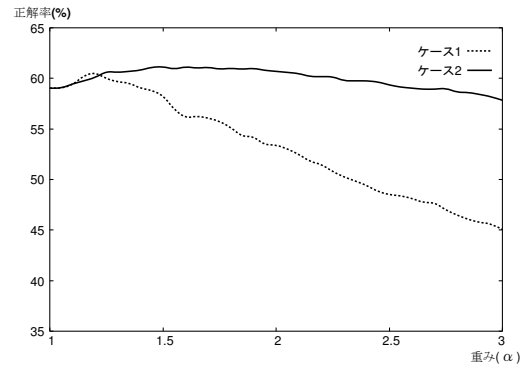


図 1 段落単位の正解率（毎日）

ケース 1 では重み（ α ）を増加したとき，わずかに上昇する傾向が認められたが，すぐに下降する傾向が観測された。ケース 2 では，重み（ α ）を増加することで，安定して正解率が増加する傾向を得た。 $\alpha = 1.5$ の時，ケース 2 で最も高い正解率（61.2%）を示した。 α をさらに増加すると，正解率は低下した。

図 2 はケース 2 で $\alpha = 1.5$ を設定した場合の各段落の選択比率を示す。なお，集計対象とするデータは段落数が 5 の 461 記事を対象とした。

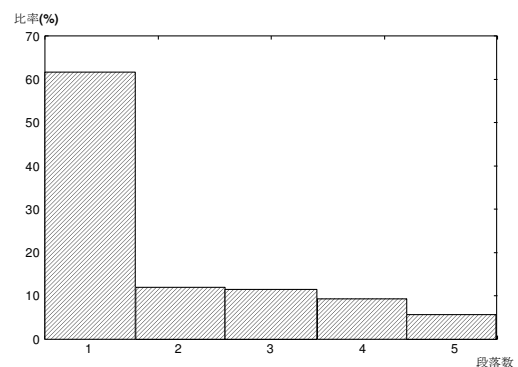


図 2 重要段落判定結果の分布（ $\alpha = 1.5$ ）

図 2 では，重み（ $\alpha = 1.5$ ）において，重要段落と判定した段落の分布を示している。第

一段落を選択した比率がもっとも高い。

したがって、本手法が記事中の重要段落とする第一段落を選択することに対して有効であることを示す結果を得た。また、図1より、重み(α)を変えることで正解率を上げることができることを示した。

同様の実験を文単位で行った場合の結果を図3に示す。

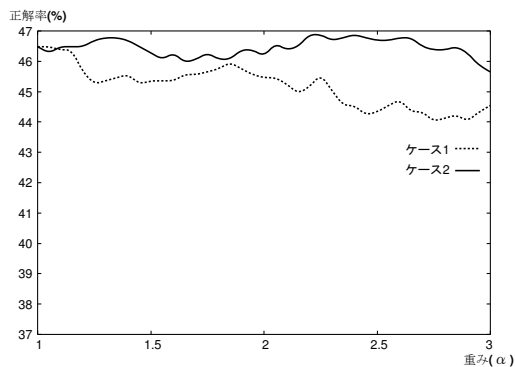


図3 文単位の正解率(毎日)

図3に示した文単位の実験では、記事を文単位に分割し、本手法により重要部分と判定した文が、記事のどの段落に所属しているかを調べ、第一段落に所属する割合を正解率として評価した。

解析単位として文を用いた場合、正解率は段落単位と比較し低い値となった。重み(α)を変化することによって得た値も安定していない。

毎日新聞データと同様にして日経新聞データを用いた実験結果について述べる。図4に解析の単位に段落を用いた場合の結果を示す。

図4では、 $\alpha = 2.2$ の時、ケース2でもっとも高い値(77.9%)を示した。毎日新聞データを用いた場合と同様に、ケース2が安定して高い正解率を示した。 α を変化することにより得る正解率も、毎日新聞データの場合と同様の傾向を観測した。

日経新聞データを用いて文単位で行った同

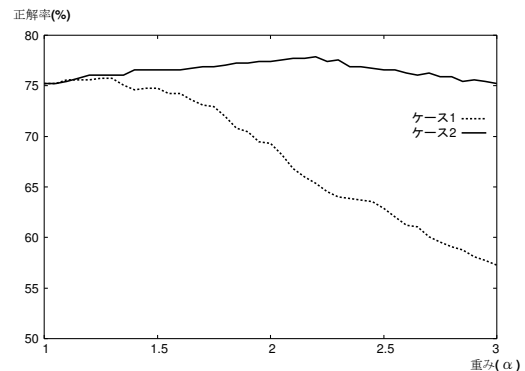


図4 段落単位の正解率(日経)

様の実験結果を図5に示す。

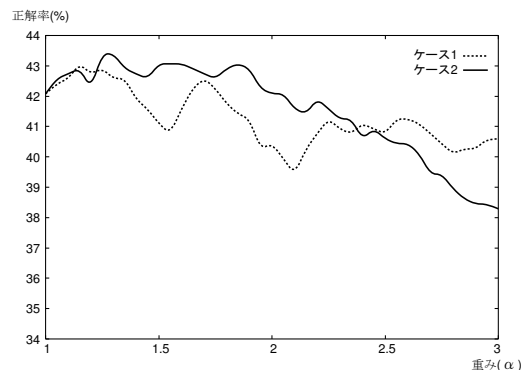


図5 文単位の正解率(日経)

日経新聞データにおいて文を解析の単位とした場合、毎日新聞データの場合と同様に、重み(α)の増加に対し安定した値を示していない。また、重み(α)を大きくすることで全体的に正解率が低くなるという結果を得た。

毎日新聞データならびに日経新聞データを用いた場合のいずれも、ケース2を用いて測定した場合の正解率が高いという結果を得た。

ケース1とケース2では、重み(α)を大きくすることで正解率に大きな開きを得た。これは、提題として用いられた部分にのみ重み(α)を適用したケース1では、話題の展開に用いた提題部分が大きな値を取り、本来の重要部分にではなく、話題の展開部分に偏った

ためである。

これは文章内で提題として用いた表現は、提題として用いた部分にのみ重要語（注目語）としての影響を与えるのではなく、文章中で用いた部分に対しても影響を与えている。

したがって、ケース2を用いた場合の方が安定して重要段落を選択する能力が高いと言える。

6. 考 察

本実験は、比較的文章の構造が明確なニュース報道記事を対象に実験を行った。ニュース報道記事が「逆三角形」構造で論旨を展開しているため、情報の発信者が重要とする部分を記事の先頭（第一段落）に記述している特徴に基づいて評価するためである。

一方、要約の実験などでは「社説」などの記事を対象にすることが多いが、実際にどの部分（段落）を情報の発信者が重要と判断して記述しているのかは不明である。

ニュース報道記事を用いた実験において有効性を示した手法を、そのまま毎日新聞「社説」に適用した結果を図6に示す。記事そのものを段落単位に5段階に正規化し、記事全体のどの部分を本手法（ケース2）が重要と判定したかを示している。なお、縦軸は記事全体に占める比率を示し、横軸は重み（ $1.0 \leq \alpha \leq 3.0$ ）、変化量 0.05 で示す。

比較のため図7に、社説同様に、記事そのものを5段階に正規化し得た結果を示す。

図6および図7では、扱うデータの違いが数値上明確に現れている。毎日新聞ニュース報道記事の場合、60%以上の記事に対して、記事の先頭から20%の範囲を重要部分と判定した。それに対し毎日新聞社説での重要と判断した比率は、先頭から20%の範囲が他の部分に比べて高いが、記事の場合ほど明確な違いはない。

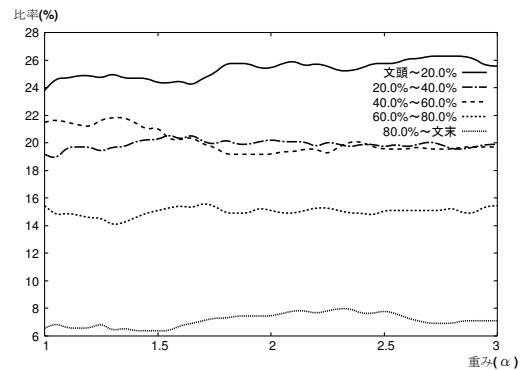


図6 社説における重要段落選択結果の分布 ($1.0 \leq \alpha \leq 3.0$)

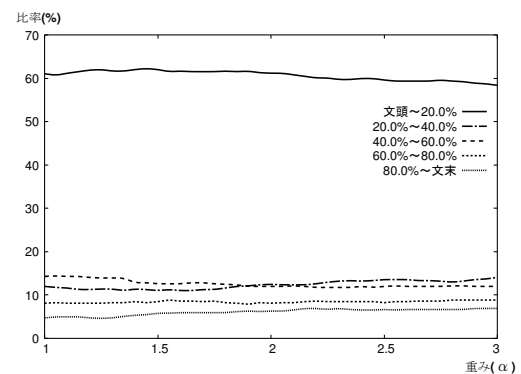


図7 記事における重要段落選択結果の分布 ($1.0 \leq \alpha \leq 3.0$)

社説の場合、ニュース報道記事の記述スタイルと異なる形式を採用していると考えられることができる。ニュース報道記事では、第一段落に記事の重要部分をまとめて配置する記述方針をとっているが、社説などでは、実際にどの部分に重要部分を配置するのか明確な取り決めはない。従って、図6に示した分布を得ることは当然の結果であると言える。また、社説では重要部分を1つの段落で記述していない可能性も否定できない。

本考察においては、社説とニュース報道記事との記事スタイルの違いを実験結果として得たことにとどめ、社説における重要部分判定精度の評価については今後の課題として扱う。

7. ま と め

本稿では、文中の名詞を中心とした出現頻度をもとに、文章全体の記事ベクトルと段落ベクトルとの内積を重要段落の判定基準に置いた。

実験データは、記事の先頭に情報の発信者が重要とする部分を記述しているニュース報道記事を対象とし、比較的大きなデータを用いた実験を行った。提題表現として用いた語に対して、重み (α) を与えることで、記事中の重要部分を毎日新聞データの場合 61.2%、日経新聞データの場合 77.9%の正解率で検出できることを示した。

解析単位を段落とした場合、毎日新聞データより得た正解率と日経新聞データより得た正解率では、大きな開きがある。記事中に占める第一段落の割合は、毎日新聞データで 25.2%、日経新聞データで 27.7%であり、数値上の大きな違いはない。記事の記述スタイルに違いがあると予測できるが、詳細な分析については今後の課題とする。

また、実験で用いた条件ケース 1、ケース 2 では、重み (α) を大きくすることで、正解率に大きな開きを得た。これは、提題表現は提題表現として用いた文にのみ影響を与えるだけでなく、提題表現として用いられた語彙が、提題として表現していない他の部分に対しても影響を与えていることを実験により示した。

考察では、一般に要約等の実験で用いることの多い社説にも本手法を適用した実験結果を示した。社説とニュース報道記事との間には重要部分の存在位置に関して明確な違いを得た。社説に対しての重要部分の判定がどの程度正確になされたかという評価については、今後人手による評価等を行うことで本手法の有効性を示す予定である。

参 考 文 献

- 1) 奥村学, 難波英嗣. “テキスト自動要約に関する研究動向”. 自然言語処理, Vol.6 No.6, (2002)
- 2) 福本惇一: “文間関連度に基づく内容抽出手法”. 言語処理学会, 第 3 回年次大会, (1997)
- 3) 福本文代, 福本惇一, 鈴木良弥: “文脈依存の度合いを考慮した重要パラグラフの抽出”. 自然言語処理, pp.89-109, (1997)
- 4) “記者ハンドブック”. 株式会社共同出版社, pp.10-11. (2001).
- 5) 三上章. “象は鼻が長い”. くろしお出版, (1960).
- 6) 佐久間まゆみ. “文章構造と要約文の諸相”. くろしお出版, (1989).
- 7) 林四郎: “文章論の基礎問題”. 三省堂, (1998)
- 8) 永野賢. “文章論総説”. 株式会社朝倉書店, (1986).
- 9) 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原 正幸: “日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書, 奈良先端科学技術大学院大学”. (2000)
- 10) “CD-毎日新聞’98-99”. 毎日新聞社, (1998-9)
- 11) “日経全文記事データベース CD-ROM 版’98”. 日本経済新聞社, (1998)