# タイトルパタンによる文書の一文概要生成

長安義夫, 山本和英
長岡技術科学大学 電気系
E-mail:{nagayasu,ykaz}@nlp.nagaokaut.ac.jp

## 1 はじめに

Web の発展に伴い検索技術の重要性が高まってきている。検索技術に要求される条件として「検索の網羅性」と「結果の効果的な提示」がある。しかし後者の要求は十分に満たされているとは言えない。Web 検索エンジンで主に使われている文書内容の抜粋（スニペット）は効果的提示を満たすための重要な方法のひとつであるが、Web 検索はいまだ利用者の勘や慣れが必要な場面も多い。

スニペット以外の情報としては検索文書に付与されているタイトルがある。しかし Web 文書のタイトルは内容と一致していないものが非常に多く有用でない。我々は文書の概要を端的に示す文が検索結果をより効果的に提示できるのではないかと考えた。本稿では、検索結果を効果的に提示するための文書の一文概要を自動生成する方法を提案する。

関連する研究として、鈴木ら [1] は要約を目的とし Support Vector Machine(SVM) を用いて重要文節を抽出している。この研究では文節を抽出しているが SVM による二値分類は単語の抽出にも応用可能である。廣島ら [2] は Web 文書を対象に SVM を用いて単語抽出を行い実際に 3-gram 確率を指標としてヘッドライン生成を行っている。この研究では単語抽出を全て SVM で行っている。西村ら [3] は機械翻訳を目的とした、コーパスから作成した表現パタンを用いた文生成を行っている。

## 2 一文概要とタイトルパタン

### 2.1 一文概要

スニペットでは、それが文書の核心であるのか、その文書にとって重要でない箇所であるかは判断できない。後者であればそのスニペットは文書の要約としてふさわしくない。我々はスニペットのような局所的な検索指標だけではなく、文書全体を表現するような大局的な検索指標があることが望ましいと考える。この条件を満たすものの例として文書のタイトルを挙げる。ではどのようなタイトルが望ましい検索指標となるのか。

廣島ら [2] はヘッドラインの必須条件として「（1）内容網羅性」、「（2）可読性」、「（3）高圧縮性」を挙げている。しかし Web 検索の場面においては、内容を網羅したものでなくとも内容を示唆する程度の短文で十分検索の指標となりうると考える。よって我々は上記の（2）と（3）に加えて「内容指示性」の 3 つを揃えた文を一文概要と定義する。またこの一文概要を自動で生成することを “概要生成” と呼ぶ。

### 2.2 タイトルパタン

本来、文の自動生成は非常に困難な課題である。しかし本稿で目的とする “概要生成” は必要最小限の構成要素（名詞、動詞、それらをつなぐ助詞）さえあればよい。そのため一般の文生成より簡易に行える。このような単純な表現は、生成するよりもコーパス中から抽出するほうが簡単である。よってこれらを汎化して “骨組み” とし、これに名詞や動詞などの内容語の “肉付け” を行うことで簡単に一文概要となるのではないかと考えた。タイトルパタンとはつまり一文概要における “生成の骨組み” である。これによって文生成が簡易に行え、かつ可読性や高圧縮性といった条件も満たすことができる。タイトルパタンの具体例は 3.1.3 節で示す。

## 3 手法

手法は大まかに事前準備と概要生成に分けることができる。事前の準備では名詞のクラスタリング、SVM による重要語分類モデルの学習、タイトルパタン辞書の構築の三つを行う。クラスタリングはタイトルパタンの汎化を目的としている。SVM による重要語分類モデルの学習は名詞の IDF 値とともに名詞の重要度計算に用いる。ここで重要語の定義とは「文書の主題を最もよく表す語」とする。

概要生成はまず SVM と IDF の両方のスコアを用いて重要語を抽出する。抽出した単語を基にタイトルパタンを決定し、タイトルパタンを用いて概要候補を生成する。生成した複数の概要候補を決定するため一文概要の順位付けを行う。

### 3.1 事前準備

#### 3.1.1 名詞のクラスタリング

名詞の分類にはクラスタリングツール「GETA(1)」とシソーラスの 2 種類を用意する。これらはタイトルパタンの汎化に使用するためである。「GETA」ではコーパスから名詞の頻度を素性にクラスタリングを行った。シソーラスでは木構造であることを利用し葉の数が閾値以下となるようにクラスタを形成した。クラスタにはそれぞれ識別 ID が付与されている。以降これをクラスタ ID と呼ぶ。

語彙数の少なかったシソーラスの分類に「GETA」によるクラスタリングを同時に用いて語彙不足を補った。語の分類が人手で行われているという点でシソーラスのクラスタは GETA のクラスタより良質であると考える。よってクラスタリングを利用する場面では常にシソーラスのクラスタを優先する。

#### 3.1.2 SVM による重要語分類モデルの学習

本文中の重要語候補に与えるスコアのひとつとして SVM を用いる。SVM の学習に用いた素性は我々の先行研究 [4] を参考に、本文中での名詞の出現頻度、対象名詞の表層形及び品詞、対象名詞の前後二形態素の表層形及び品詞、本文中での対象名詞の出現位置などである。この素性を用いて社説記事のタイトルに含まれる名詞を正例、それ以外の全ての名詞を負例として学習を行った。カーネルは線形カーネルを使用した。

#### 3.1.3 タイトルパタンの生成

**Step 1** タイトルの選別

本稿が目的とする一文概要に最も近い形式は社説のタイトルであると考えた。社説のタイトルは一文概要の条件を全て兼ね揃えている。

社説記事のタイトルの観察をさらに進めると社説記事のタイトルは概ね「意見を表明する文」と「事実を端的に表す文」の 2 つに分けられることが分かった。ここでは「意見を表明する文」を “意見文”、「事実を端的に表す文」を “事実文” と呼び、以下にその例を示す。

1. 意見文「こんな人物がなぜ次官になれたのか」
2. 事実文「よりよい防衛協力指針をつくる視点」

意見文は口語的言い回しや倒置が多く、パタンとするには複雑である。逆に事実文は単純な表現が多い。我々が必要とする概要も事実文に近い。よって我々は事実文のタイトルをタイトルパタ

ンに使用する。ただし事実文と意見文を自動で分けることは困難であるため、人手で区別し収集することにする。

以下、事実文「よりよい防衛協力指針をつくる視点」を生成例としてパタンを生成していく。

**Step 2** 文節のタグ付け

収集した社説記事のタイトルを構文解析器にかけ、文節ごとに「名詞節」「動詞節」「その他」というタグを付与する。名詞節は名詞と助詞のみを含む文節、動詞節は名詞を含まず動詞か助動詞を含む文節、その他はそれ以外の文節とした。


図 1: タグ付与と係り受け構造

**Step 3** 不必要な部分の削除

構文解析器の解析結果を元に「その他」とその節にかかる構文木の葉を全て削除する。

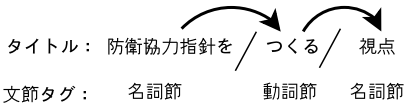


図 2: 文節の削除

**Step 4** 重要文節の決定

我々は社説のタイトル中に含まれる名詞にも、主題を表す度合に違いがあると考えた。図 2 の例で見れば、最も重要なのは「防衛協力指針」であり、「視点」は記事の主題を表しているとは言えない。そこで最も主題を表す名詞を含む名詞節をタイトルの “重要文節” と定義する。3.2 節ではこの重要文節を手がかりとして概要生成を行っていく。

重要文節はタイトルの名詞節の中で、IDF 値が最も高い名詞を含む名詞節とする。最も高い IDF 値を持つ名詞節が複数ある場合は、最も高い IDF 値を持つ文節の中から最も文頭に近い文節を重要文節とする。

抽出する名詞は、名詞節中で品詞が名詞となっている形態素を抽出する。名詞が連続している場合はそのまま複合名詞として抽出する[1]。

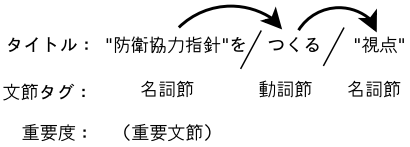


図 3: 重要文節の決定

**Step 5** パタンの汎化

タイトルに含まれる名詞をそのまま使うことはデータの過疎性に繋がるので名詞を汎化する。3.1.1 節で作成したクラスタリング結果を使用する。

タイトルから「名詞節」に含まれる名詞（複合名詞）を抽出し、クラスタリング結果と比較してその名詞が属する GETA 及びシソーラスのクラスタ ID を名詞節に付与する。それぞれのクラスタ ID は複数になることもある。名詞節の機能語は表層形のまま残す。機能語が存在しなかった場合は、空文字 “ε” を付与する。シソーラスと GETA の両方のクラスタのいずれにも抽出した名詞が属さない場合はそのタイトルを不適切であるとみなす。不適切と判断されたタイトルからはパタンを生成しない。

また動詞節は一律に “動詞” として汎化している。

表 1 にタイトルパタンに付与されているタグやクラスタ ID な

[1] これ以降で用いる複合名詞は全てこの方法で抽出している。

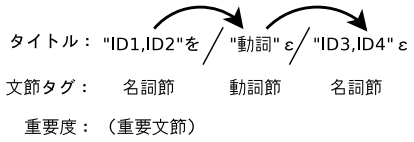


図 4: 名詞の汎化

どの構成を示した。実際のタイトルパタンはこの構成表の項目が上から順に並んでいる。クラスタ ID が見付からない箇所は機能語と同様に “ε” となる。

表 1: タイトルパタンの構成

| 名詞節 | 動詞節 |
|---|---|
| “名詞節” タグ | “動詞節” タグ |
| シソーラスのクラスタ ID | 係り先文節番号 |
| GETA のクラスタ ID | 機能語 |
| 重要文節フラグ (0 or 1) | |
| 係り先文節番号 | |
| 機能語 | |

以上の手順で作られたタイトルパタンの例を表 2 に示す。

表 2: タイトルパタンの例

| 文節 | 該当する文節のパタン |
|---|---|
| “より” | （削除） |
| “よい” | （削除） |
| “防衛協力指針を” | 名詞節/1f9eb2/87/1/1/を |
| “つくる” | 動詞節/2/ ε |
| “視点” | 名詞節/0ed7c7::0faa7a::30f78e/0/0/-1/ ε |

こうしたタイトルパタンを辞書として保持しておく。

### 3.2 一文概要の生成

“内容指示性” を考えると、一文概要に必要なのは文書の主題をよく表す単語とその内容を支える補助的な説明である。概要に使用する単語全てが主題を表している必要はない。ここでは文書の主題を最もよく表す語を “重要語”、その重要語を補助的に説明する語を “付加語” と呼ぶことにし、これらとタイトルパタンを用いて一文概要を生成する。重要語は IDF 値と SVM の学習結果によって抽出される。付加語はタイトルパタンに付与されているクラスタ ID を基に複数抽出される。

付加語の選び方は概要生成の手法に依存するためここでは重要語の抽出と付加語の抽出を分けて説明する。

#### 3.2.1 重要語抽出

まずタイトルパタン生成時と同様に本文を構文解析器にかけ文節に「名詞節」「動詞節」「その他」のタグを付与し、名詞節から名詞（複合名詞）を抽出する。このとき抽出した名詞同士が連続する文節から取り出され、かつ文節同士が「の」で繋がれている場合はそれを大きな名詞節ととらえ名詞同士を「の」で連結しひとつの名詞として扱う。これらの名詞に文生成のためのクラスタ ID を付与しておく。これらの名詞は付加語の候補でもある。

複合名詞の意味や使われかたの性質は概ね複合名詞の最後の形態素によって決まると考える。本稿では複合名詞と単名詞を同列に扱うことがしばしばあるため、その場合は最終形態素の名詞を複合名詞の代わりに用いることにする。本稿では複合名詞の最終形態素の名詞をその複合名詞の “代表名詞” と呼ぶことにする。本文で出現した名詞に対する名詞抽出例とその代表名詞を表 3 に示す。

こうして取り出された名詞にコーパスから取得した IDF 値を付与する。ただしコーパス中でゼロ頻度となる名詞にはある一定の高い値を付与することにする。IDF 値の低い名詞は社説の主題に無関係であることがほとんどであり、名詞の重要度計算におけるノイズとなることが予想できる。よってここでは IDF 値で順位付けを行い上位半分のみを取り出して重要語候補として扱う。

表 3: 名詞の抽出例

| 出現例 | 抽出名詞 | 代表名詞 |
|---|---|---|
| …が国債を… | 国債 | 国債 |
| …に社会不安が… | 社会不安 | 不安 |
| …日本の外交は… | 日本の外交 | 外交 |

IDF 値だけではデータの過疎性の問題が起き、また重要でない名詞も上位にきてしまう恐れがあることを考慮し別の指標を併用して重要語選択を行う。本稿では鈴木ら [1] や廣島ら [2] の研究を参考に SVM を使用した。SVM は各分類項目に対して識別平面からの距離を計算することができるため、この距離で名詞の順位付けを行う。この順位と IDF の順位を基にスコア計算を行い名詞の重要度を計算する。

ここで SVM が形態素単位で分類を行っているのに対し IDF 値を付与したものは複合名詞も含む。さらに IDF 値と SVM による識別平面からの距離は単位が異なる。よってスコアをまとめるときにこれらの整合性を取らなければならない。整合性を取るため、複合名詞の代表名詞が SVM で分類された名詞と一致した場合、SVM 順位の逆数と IDF 順位の逆数の和をその名詞の重要度とする。

$$Score(W_i) = \frac{1}{Rank_{IDF}(W_i)} + \frac{1}{Rank_{SVM}(R_i)} \quad (1)$$

$W_i$ は複合名詞を含む名詞、$R_i$ は $W_i$ の代表名詞、$Rank_{IDF}(W_i)$ や $Rank_{SVM}(R_i)$ はそれぞれ IDF 順位と SVM 順位を表す。ただしスコア $Score(W_i)$ は、$Rank_{SVM}(R_i)$ が別のスコア $Score(W_j)$ に既に加算されていた場合、$Score(W_i)$ に加算されることはない。計算例を表 4 に示す。

表 4: 重要語のスコア例

| $W_i$（$R_i$） | $Rank_{IDF}(W_i)$ | $Rank_{SVM}(R_i)$ | $Score(W_i)$ |
|---|---|---|---|
| 新食糧法（法） | 1 位 | 7 位 | 1.1 |
| 食管法（法） | 2 位 | 7 位 | 0.50 |
| 過剰米（米） | 3 位 | 5 位 | 0.53 |
| 減反推進（推進） | 4 位 | 24 位 | 0.29 |
| 適正量（量） | 5 位 | 90 位 | 0.21 |

この例では “法” の SVM 順位が既に “新食糧法” に加算されているので “食管法” のスコアに反映されていない。

#### 3.2.2 生成

タイトルパタンを重要語を基に選択する。

まず重要語をスコアの高い順に取り出す。取り出した名詞のクラスタ ID と同じ ID を重要文節に持つタイトルパタンを抽出する。タイトルパタンが抽出できない場合は次の重要語で同様にタイトルパタンを探す。パタンが見付かった場合、重要文節以外の文節と同じクラスタ ID を持つ名詞を付加語として保持する。動詞節にはパタンと同じ語尾を持つ本文中の全ての動詞節が候補となる。このときパタン中に名詞や動詞の候補を持たない文節が 1 つでもある場合はそのパタンによる生成を失敗とみなし別のパタンでの生成に移行する。抽出した全てのパタンで生成できなかった場合は次の重要語で生成を行う。

タイトルパタンの全ての文節に対して名詞あるいは動詞の候補が抽出された場合、それらの単語全ての組合せを出力する。

### 3.3 一文概要の順位付け

文生成の際に単語の全ての組合せを出力した。この出力結果から最も良いものを選ぶために出力結果の順位付けを行う。順位付けには “動詞を中心とした連接確率” と単語 2-gram 確率を総合したスコアを用いた。ただしスコアを求める際は複合名詞を代表名詞へと置き換えて他候補と文の長さを揃えた。

例 1 日本経済を覆う雇用不安　→　経済を覆う不安

自然な文かどうかは名詞と動詞の自然な結び付きが重要であると考える。そこで「動詞節にかかる名詞節の助詞」「動詞節」「動詞節の係先の名詞」の連接確率をコーパスから求めた。これを “動詞を中心とした連接確率” と定義する。ただし長い文では連接が多いために確率が 0 に収束してしまう。これを避けるために確率値の対数の絶対値を取り連接確率をコストとした。

重要語抽出と同様に連接確率の低い文はノイズとなる恐れがある。コスト化された連接確率で順位付けを行い、上位半分の文のみを概要候補として扱う。概要候補中で最も大きい連接コスト（＝最も低い連接確率）で正規化したものを動詞連接スコアとした。

動詞連接の順位で上位半分とされた概要候補に新たに単語 2-gram 確率を付与する。この場合も同様に確率をコスト化して単語 2-gram 確率の順位付けを行い、最も大きいコストで正規化したものを 2-gram スコアとした。

動詞連接スコアと 2-gram スコアの和をその概要候補の可読性のスコアとする。コスト化しているのでスコアが小さい程良いスコアとなる。このスコアで順位付けを行うことによって任意の数の概要候補を取得することができる。

## 4 評価実験

### 4.1 使用したツール及びデータ

本研究の目的は Web を対象とした概要生成であるが、目的を明確にするため Web 文書でなく新聞の社説記事を対象とした。Web 文書に対して社説記事は概ね「主題がひとつに限定されている」「文章が（Web と比較して）整っている」という点が異なっている。

社説記事を使用したことから、SVM の学習、IDF 値や連接確率の計算、GETA のクラスタリングは全て日本経済新聞 (2)9 年分を用いた。シソーラスには EDR 電子化辞書 (3) の概念辞書を用いた。タイトルパタン辞書は、社説タイトルから事実文を 200 記事収集し、そのタイトル群から 135 パタン生成した。

評価対象は日本経済新聞から学習データとは別の社説を無作為に 50 記事抽出して使用した。

形態素解析器、構文解析器にはそれぞれ茶筌 (4) と南瓜 (5) を使用した。SVM は TinySVM(6) を使用した。

### 4.2 評価方法

社説記事ひとつに対して重要語を 3 つ抽出し、その重要語に対してそれぞれ上位 10 位までの概要の候補を出力する。それぞれの出力候補に対して「日本語として自然な文か」と「社説の概要としてふさわしいか」のふたつの指標で被験者 3 人が独立に評価した。

テストデータ 50 記事に対して一文概要を生成したところ、システムが出力した一文概要の候補数は全部で 967 候補だった。また一記事あたりの平均出力数は 18.3 候補で最多出力数が 30 候補、最少出力数が 2 候補だった。

## 5 評価結果

### 5.1 全体評価

3 人の被験者の評価結果を表 5 に示す。評価 1 は被験者が自然な日本語であると判断した候補数の割合であり、評価 2 は記事の概要としてふさわしいと判断した候補数の割合である。

表 5: 可読性と内容一致性の評価

| 正解とした被験者数 | ≧ 1 | ≧ 2 | =3 |
|---|---|---|---|
| 評価 1（可読性） | 524/967 (54.2%) | 262/967 (27.1%) | 125/967 (12.9%) |
| 評価 2（内容一致性） | 74/967 (7.9%) | 8/967 (0.8%) | 2/967 (0.2%) |

表 5 より内容一致の評価が可読性に比べて極めて低いことが分かる。

被験者 3 人のうち 2 人以上が正解とした候補（以降正解候補）の評価は 27.1%、内容一致性の評価は 0.8%である。

### 5.2 記事別での評価

出力された結果を記事別で見た結果を表 6 と表 7 に示す。表 6 は可読性、内容一致性、ともに「一記事内に正解候補が少なくともひとつ含まれること」という条件を満たす記事数と割合を表している。表 7 は 1 記事中に含まれる正解候補の割合を示している。

表 6: 記事別の可読性と内容一致性の評価

| | 記事数 |
|---|---|
| 可読性 | 46/50 (92%) |
| 内容一致性 | 7/50 (14%) |

表 7: 1 記事中の正解候補の割合

| | 1 記事当たりの割合 |
|---|---|
| 可読な候補 | 33.2% |
| 内容が一致している候補 | 1.7% |

表 6 と表 7 から、評価に使用した記事の 92%に少なくともひとつ以上可読な候補が含まれており、かつ一記事平均では 30%強の割合の候補が可読な文であることが分かる。内容一致はほとんどないが高い可読性を示している。

## 6 考察

### 6.1 可読性

評価実験では大量の出力候補数のうち上位 10 位のみを評価の対象としている。非常に限られた上位候補を用いたことによって可読性の良い概要が多くできたと予想できる。可読性のスコアが候補の順位付けに有効であったことを示している。

また評価に使用した社説 50 記事の 1 候補当たりの平均文字数が約 45 文字であったのに対して、概要候補の平均文字数が 23 文字と比較的短い文となっていた。これはタイトルパタンの利用によって得られた最もわかりやすい効果である。

一方で長いタイトルパタンで作られた概要候補の可読性は低くなる傾向がある。正解候補の平均文字数が約 20 文字であるのに対し、どの被験者にも可読でないと判断された文の平均文字数は約 25 文字であった。文の長さに依存しない可読性のスコアを考慮する必要がある。

### 6.2 内容一致性

可読性と比べて内容一致性は非常に悪い結果を示した。重要語抽出について見てみる。図 5 は被験者がそれぞれ内容一致性について正解とした一文概要候補の重要語の SVM 順位と IDF 順位の相関図である。ただし被験者が 2 人以上正解とした候補ではデータが少なすぎるため、被験者が 1 人でも正解とした場合の候補についてのデータを用いている。

図 5 を見ると、IDF 順位で 1 位の重要語ばかりが候補になっていることが分かる。重要語は順位付けの際に IDF 順位の下位を足切りしている。このように重要語抽出のスコアが IDF 順位に偏っていることが原因である。SVM 順位には 2 つの傾向が見て取れる。SVM 順位と候補数がある程度比例している点と、40 位までの範囲で候補数が分散している点である。SVM 順位によって重要語候補の多くが上位にきているにも関わらず内容一致性の評価が低いのはなぜだろうか。

式 1 を見ると、この式はどちらか一方が上位であると点数の差が大きく開くため、上位スコアが常に優先されることになる。SVM 順位によってある程度は重要語が上位に来ている一方で、式 1 によるスコアリングが悪影響を及ぼし、内容一致性の評価を下げていると考えられる。IDF 順位と SVM 順位を等価的に扱う式が必要である。また IDF 順位の影響を見るため、IDF 順位の低い候補での評価もすべきである。

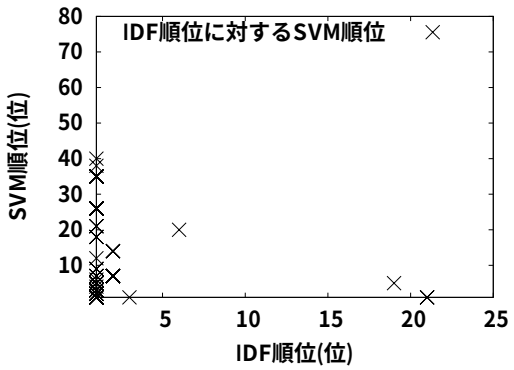


図 5: IDF 順位と SVM 順位の関係

## 7 おわりに

タイトルパタンを用いて文書の一文概要を生成する手法を検討した。社説記事に対する評価実験では可読性の評価が 27.1%、内容一致性の評価は 0.8%を得た。対象社説 50 記事の 92%にひとつ以上の可読な候補が生成されており、社説 1 記事当たりの可読な候補の割合は 33.2%となった。可読性の高い結果となったが内容一致の面では大きな課題を残した。内容一致の精度が低かった原因には、SVM によって選別された重要語を活かせるようなスコアが設定できなかったことが挙げられる。

### 使用した言語資源及びツール

(1) 汎用連想計算エンジン（GETA）, 第 2 版, 情報処理振興事業協会（IPA）, http://geta.ex.nii.ac.jp/

(2) 日本経済新聞全記事データベース 2000 年度版, 日本経済新聞社.

(3) EDR 概念辞書, 情報通信研究機構（NiCT）, http://www2.nict.go.jp/r/r312/EDR/

(4) 形態素解析器「茶筌」,Ver.2.3.3, 奈良先端科学技術大学院大学 松本研究室, http://chasen.naist.jp/hiki/ChaSen/

(5) 構文解析器「南瓜」 ,Ver.0.52, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/~taku/software/cabocha/

(6) SVM 学習ツール “TinySVM”,Ver.0.09, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/~taku/software/TinySVM/

### 参考文献

[1] 鈴木 大介, 内海 彰: Support Vector Machine を用いた文書の重要文節抽出, 人工知能学会 論文誌,21 巻 4 号 B,pp330-339,2006.

[2] 廣島 伸彰, 長谷川 隆明, 山崎 毅文: 統計的手法に基づく Web ページからのヘッドライン生成, 情報処理学会 研究報告,NL149-7,pp45-50,2002.

[3] 西村 仁志, 坂本 仁: コーパスから自動抽出した表現パターンを用いる日本語文生成, 情報処理学会 研究報告,NL148-5,pp31-36,2002.

[4] 池田 諭史, 牧野 恵, 山本 和英: 濃縮還元型文要約モデルの検討, 情報処理学会 研究報告,NL174-13,pp71-76,2006.