LE-002

# 要約事例を用例として利用したニュース記事要約

Example-based Summarization of News Articles by Using Summary Instances

牧野 恵† 山本 和英†

Megumi Makino† Kazuhide Yamamoto†

## 1 はじめに

近年、インターネット等を通じて大量の電子化文書が溢れている。そのため、ユーザが効率良く情報にアクセスするための支援技術に対して需要が高まっている。テキスト自動要約もその技術の 1 つとして注目され、盛んに研究が行われている。

我々は実際の要約事例を模倣利用して要約を行う用例利用型要約の研究を行っている [1]。これは用例利用型の要約が以下のような利点を持ち合わせていると考えるからである。

**(1) 保守が容易である**

システムにとって、修正や保守が容易であることはとても重要である。用例利用型のシステムは用例を追加することによって容易に修正をすることができる。さらにその追加された用例が他の用例に副作用を及ぼすことがない。これに対して、人手で要約ルールを作成するルールベースの要約は修正や保守に高いコストがかかる。また、修正するためのルールが他のルールと競合してしまう等の影響もあると考える。

**(2) 重要度の設定をしなくてもよい**

現在、多くの統計的要約手法が提案されている [2, 3, 4]。これらの多くは、語句に対して重要度を設けることで重要部分を同定し、要約文を作成している。しかし語句に対する重要度は記事に依存し、さらに人間が要約する際に必要とする語と相関があるような重要度を一意に決めることは難しい。用例利用型の要約では重要度を設定せず、そのかわりに類似度の計算をする。我々は重要度の計算よりも、2 つの表現間で類似度を測るほうが容易であると考える。

**(3) 日本語としてより頑健で正しい要約ができる**

日本語は膠着語であるため、文の表現がとても豊かであるという特徴を持っている。そのため、統計的に有意ではないような表現でも、実際には要約表現となっている場合がある。用例利用型の要約では実際に存在する要約事例を使用しているため、このような場合にも対応でき、より頑健で日本語として正しい要約が作成できるのではないかと考える。

これらの特徴を理由に、本論文では要約方法を用例に委ねた用例利用型の要約手法を提案する。用例には実際の要約事例を用いており、提案手法は類似用例の選択、文節の対応付け、対応文節の組合せの 3 つから構成される。

用例利用型要約の既存研究としては Nguyen らの手法 [5] がある。Nguyen らは用例として原文と要約文の対から作成したテンプレートルールを使用している。なお入力の単位は 1 文であり、用例を模倣利用することで文短縮を行っている。これに対し、本論文では入力の単位を複数文 (ニュース 1 記事) として、1 文に要約しているため、要約率は必然的に低くなる。そのため Nguyen らよりも我々が取り組む問題はさらに困難である。また我々は人手で作成された要約文のみを用例として用いているため、原文との対応を取る必要がなく、容易に用例を収集することができる。そのため原文と要約文の対応コーパスが少ない特許や医学文書の分野でも効果的に要約が作成できると考える。

またニュース記事などの複数文を要約する既存研究は、$tf \cdot idf$ などで重要文を判定し、それをさらに文圧縮するものが多い。それに対して本論文では、複数文から語句を取り出して、これらを組合せることによって文を生成する手法である。よって本手法で作成した要約文は高圧縮であり、複数文の情報をできるだけ網羅したものであると考える。

†長岡技術科学大学 電気系
E-mail: {makino,ykaz}@nlp.nagaokaut.ac.jp

## 2 提案手法

要約事例を用例として利用し、その要約表現に従うことで入力であるニュース記事を 1 文へ要約する。我々が定義する用例とは人手で作成された要約文である。本論文では日経 goo[1] からメール配信された新幹線要約文 [6] を用例として使用した。この新幹線要約の記事は本来 1〜3 文で構成されているが 2 文目以降は付加的な情報であることが多い。そのため、記事の要約としては 1 文目だけでも十分であると考えて、用例として使用するのは新幹線要約の記事 1 文目に限定した。

図 1 にシステムの流れを示す。また提案手法は以下に示す 3 つの処理から構成される。

1. **類似用例の選択:** 入力に対する類似用例を選択する
2. **文節の対応付け:** 入力と類似用例の文節を比較して、類似した文節を対応付ける
3. **対応文節の組合せ:** 対応付けられた文節を組合せて 1 文に要約する

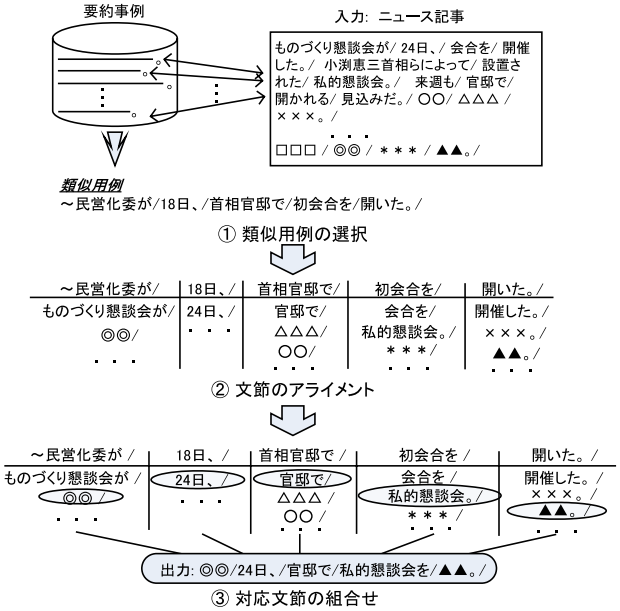


図 1 用例を利用した要約システムの概要

各処理の詳細は次節以降で述べる。

### 2.1 類似用例の選択

まず入力記事に対して内容が類似した用例 (類似用例) を選択する。内容の類似を両者に共通して出現する単語が多いことと捉え、自立語の一致を基に入力記事 $I$ と用例 $E$ の類似度を算出する。類似度は以下に示す式で定義した。

$$Sim(I,E) = \sum_{i=1}^{n} Score(i) \cdot \{weight \cdot ||Tv_1(I) \cap Tv_i(E)|| + ||To_1(I) \cap To_i(E)||\} \quad (1)$$

ここで $n$ は入力を構成している文数を表し、$Tv_i(\cdot)$ は $i$ 文目の文末文節に含まれる動詞及び、動詞+読点の文節に含まれる動詞の集合である。さらに $To_i(\cdot)$ は $i$ 文目 $Tv_i(\cdot)$ 以外の自立語集

合を表し[*1]、$||To_1(I) \cap To_i(E)||$ は $To_1(I)$ と $To_i(E)$ の積集合における要素数である。また $Score(i)$ と $weight$ は入力で内容を表している単語が一致した場合に、より高い類似度を与えるために設けたものである。単語は文の位置によって、その記事の内容を表す貢献度が異なると観察できたため、文の位置 $i$ によるスコア $Score(i)$ を導入した。このスコアは実験的に決め、評価では以下のものを用いた。

$$Score(i) = \begin{cases} 5.15 & \text{if} \quad i = 1 \\ 2.78/i^{0.28} & \text{otherwise} \end{cases} \tag{2}$$

また文末文節に含まれる動詞は内容を表すのに重要な役割を担っていることが観察されたため、文末の動詞を重視するための $weight$ も導入した。評価には $weight = 3$ を用いて実験を行っている。

### 2.2 文節の対応付け

次に得られた類似用例と入力の文節を比較し、対応した文節の対応付けを行う。ここで対応した文節とは、類似用例と入力で類似した文節のことである。アライメントの手順を次に示す。

1. 類似用例と入力の文を構文解析器し、文節に分割する。また固有表現タグも獲得する。構文解析には南瓜 [3] を用いた。
2. 類似用例の連体修飾部を削除する。連体修飾部は被修飾語によって書き方や修飾の長さが変わる。よって被修飾語が完全に一致した場合のみ連体修飾部も用例として従うべきである。本論文では体言が完全に一致する場合は少なかったため、用例の連体修飾部は模倣利用する対象として除外した。
3. 類似用例と入力で文節の対応付けを行う。対応付けでは類似用例の 1 文節に対して、類似していると判定された入力中の複数の文節が対応付けられる。本論文では対応付けを行うために (1) 格助詞の一致、(2) 固有表現タグの一致、(3) 拡張型編集距離を用いた類似度、(4) 相互情報量を用いた類似度の 4 つを用いた。(3)、(4) の尺度では類似用例の文節に対して類似度の高い入力の上位 3 文節を対応付ける。これらについて例 1 を用いて以下で説明する。

例 1) **入力と得られた類似用例**

**入力ニュース記事:**
品質管理能力などの再強化策を話し合うものづくり懇談会が 24 日、会合を開催した。(以下省略)
**得られた類似用例:**
道路公団民営化委が 18 日、首相官邸で初会合を開いた。

#### (1) 格助詞の一致

類似用例と入力で同じ格を持つ文節を対応付ける。例 1 で格助詞による対応付けを行うと以下のような対応文節が得られる。

例 2) **格助詞の一致による文節対応付け**

| 類似用例における文節 | 入力における文節 |
|---|---|
| 初会合 を | 再強化策 を<br>会合 を |
| 道路公団民営化委 が | ものづくり懇談会 が |

#### (2) 固有表現の一致

類似用例と入力で同じ固有表現タグが存在する場合、そのタグを含む文節を対応付ける。例 1 で固有表現の一致による対応付けを行うと以下のような対応文節が得られる。

例 3) **固有表現の一致による文節対応付け**

| 類似用例における文節 | 入力における文節 |
|---|---|
| 18 日 [*DATE*] | 24 日 [*DATE*] |
| 道路公団民営化委が [*ORG*] | ものづくり懇談会が [*ORG*] |

#### (3) 拡張型編集距離を用いた類似度

拡張型編集距離は「日銀が」と「日本銀行が」のような略記の文節を抽出できるように、山本ら [7] が提案している文字重み付きの拡張型編集距離を適用する。拡張型編集距離では相違尺度を類似尺度に変換したものであり、さらに文字によって類似度に加算するスコアを変えている。漢字で構成されるものは文字自体が意味を表しているため、本論文では漢字で一致した文字のみに重みをおいて拡張型編集距離の計算を行った。

#### (4) 相互情報量を用いた類似度

相互情報量を用いた文節類似度では「会議 を開く」や「大会 を開く」のように同じ動詞の目的格となり、統語的に同じ振る舞いをする文節を抽出するために導入した。これには Lin[8] の相互情報量を用いた類似文節の獲得手法を適用する。Lin はテキストコーパスから係り受け関係にある 2 文節とその文法関係に対して “(have, subj, I)” のように 3 つ組 $(w, r, w')$ を作成している。3 つ組には以下の式で与えられる相互情報量も付加している。

$$\begin{aligned} I(w, r, w') &= \log \frac{P(w, r, w')}{P(r) \times P(w|r) \times P(w'|r)} \\ &= \log \frac{||w, r, w'|| \times ||*, r, *||}{||w, r, *|| \times ||*, r, w'||} \end{aligned} \tag{3}$$

上式の $*$ は任意の単語、例えば $||*, r, *||$ ならば文法関係が $r$ である 3 つ組の出現頻度を表す。さらに単語 $w_1$ と単語 $w_2$ の類似度を算出するために以下の式を用いて計算を行っている。

$$\begin{aligned} &Sim_w(w_1, w_2) \\ &= \frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)} \end{aligned} \tag{4}$$

式 (4) における $T(w_i)$ は式 (3) の $I(w_i, r, w')$ が正となるような $(r, w')$ の集合を表す。本論文ではあらかじめ日本経済新聞 2 年分から係り受け関係にある 2 文節の単語 ($w$ 及び $w'$) とその間の助詞 ($r$) で 3 つ組を作成し類似度を算出した。このとき固有表現タグが付いている場合はその固有表現タグで汎化した形を用いている。

### 2.3 対応文節の組合せ

続いて、抽出した対応文節を組合せて 1 文へと要約する。要約文を作成する際、対応文節の助詞は類似用例の助詞へと変更する。図 2 を用いて説明する。
図中の数字は文節の番号を表しており、節点 (ノード $n_i$) は前節で得た類似用例の文節に対応付けられた入力中の文節である。ここでノード重み $N(n_i)$ に類似用例の文節と対応付けられた入力中の文節の類似度、エッジ重み $E(n_{i-1}, n_i)$ にノード間 (文節間) の連接の良さを導入する。これにより対応文節の組合せ問題はノード重みとエッジ重みの和を最大にするような最適経路問題に帰着することができる。そこで経路列 $W_p = \{n_0, n_1, n_2, \cdots, n_m\}$[*2] に対し、以下のスコアを最大にするような経路を動的計画法を用いて求める。このとき最適経路列 $\hat{W}_p$ は以下で与えられる。

$$\hat{W}_p = W_p \quad \text{s.t.} \quad \underset{p}{\text{argmax}}\, Path(W_p) \tag{5}$$

[*1] 用例 $E$ は 1 文で構成されているため $i$ によらず、常に $Tv_1(E)$、$To_1(E)$ である。

[*2] 図 2 における太線ならば $W_p = \{a, c, e, g, k, m, n\}$ を通る経路。

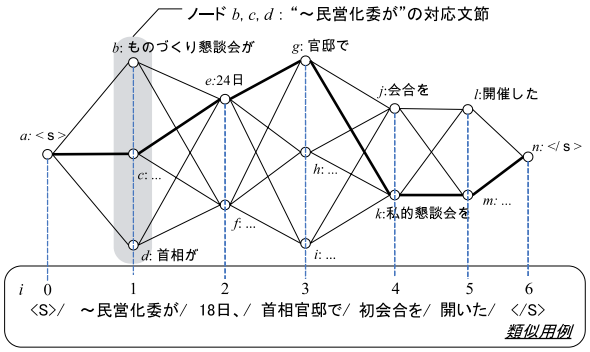


図 2 対応付けられた文節の組合せによる最適経路問題

またスコア $Path(W_p)$ を次式で表す。

$$Path(W_p) = \alpha \sum_{i=0}^{m} N(n_i) + (1-\alpha) \sum_{i=1}^{m} E(n_{i-1}, n_i) \qquad (6)$$

ここで $\alpha$ はノード重みとエッジ重みに対して与えるパラメータを表し、$m$ は類似用例の文節の最終番号を表す[*3]。評価では、パラメータ $\alpha$ に訓練で決定した 0.6 を用いた。続いて、以下にノード重みを定義する。

$$N(n_i) = \max \begin{cases} 0.5 & \text{格助詞か固有表現タグが一致} \\ 1/rank & \text{それ以外} \end{cases} \qquad (7)$$

これは類似用例の文節に対応文節がどの程度類似しているか表すスコアである。前節で得られた対応文節が格助詞の一致または固有表現タグの一致である場合、実験的に 0.5 とした。拡張型編集距離や相互情報量による類似度では上位 3 位まで対応文節を出力しているためその順位 ($rank$) の逆数をスコアに導入した。次にエッジ重みを以下に定義する。

$$E(n_{i-1}, n_i) = \frac{1}{|loc(n_{i-1}) - loc(n_i) + 1|} \qquad (8)$$

式 (8) における $loc(n_i)$ はノードつまり対応文節 $n_i$ が入力したニュース記事の何文目に存在しているかという情報である。本論文では連接する対応文節 $(n_{i-1}, n_i)$ がどれだけ離れているかということを $loc(\cdot)$ の差の絶対値を取ることで測っている。連接する対応文節 $(n_{i-1}, n_i)$ が入力したニュース記事で同一の文に存在する場合には要約として連接が良いと考えて高いスコアを与えた。逆にこのスコアを導入することで様々な文の位置から文節を取ってくる連接の悪い対応文節の組合せは防ぐことができる。

# 3 評価実験

## 3.1 実験データ

用例には 2001～2006 年に日経 goo メールから収集した新幹線要約文 1 文目 26784 文を用いた。またノード重みとエッジ重みに対するパラメータを調整する訓練データとして 150 記事 (333[形態素/記事], 112[文節/記事]) を用意し、さらに得られたパラメータを使用してテストを行うため 134 記事 (339[形態素/記事], 116[文節/記事]) のテストデータを用意した。訓練データは日本経済新聞 1999 年の記事を用いており、テストデータには日本経済新聞 2000 年を用いている。これらの記事には日経 goo メールニュースの 1 文目と人手で対応を取り正解データとしたものが存在する。また相互情報量を用いた文節類似度については日本経済新聞 2 年分 (1999～2000 年) を用いて構築した。評価実験ではシステムが出力した要約文に対し主観評価を行った。

[*3]図 2 ならば $m = 6$ である。

## 3.2 実験結果

入力したニュース記事に対してシステムが出力した要約文の文字要約率は 5% (削除率 95%) であった。入力の単位が記事であることから、得られた要約率は非常に低いものとなっており、高圧縮な要約が作成できたことが分かる。

評価実験では主観評価により、システム各部の評価を行った。まず類似用例選択部の評価では、システムに入力したニュース記事と選択された類似用例から、著者が以下の基準 1 に従い 4 段階で評価を行った。

基準 1. 入力であるニュース記事から自分が作成した要約文が類似用例の内容と似ているか
1) 類似している 2) やや類似している
3) あまり類似していない 4) 類似していない

主観評価の結果、1) 類似している、または 2) やや類似していると評価された類似用例文は 134 記事に対して、77 記事であった。よって類似用例の選択部の精度は 57%であり、我々が設定した類似度尺度が有効に働いていることを確認した。

続いて、文節の対応付けとその組合せの最適化を行うシステム、つまり要約文生成部の評価では、著者がニュース記事とシステムが出力した要約文から、以下の基準 2 に従い、4 段階で評価を行った。

基準 2. システムが作成した要約文はそのニュース記事の要約として適切な内容であるか
1) 適切である 2) やや適切である
3) あまり適切ではない 4) 適切ではない

正しい類似用例が選択できた 77 記事に対して、システムが出力した要約文の評価を行ったところ、77 記事中 48 記事が 1) 適切である、もしくは 2) やや適切であるという評価が得られた。

統計的要約である Knight ら [2] や用例利用型要約である Nguyen ら [5] は 1 文を約 60%～70%の要約率で圧縮した要約文を評価し、10 段階で 7～8 点の精度を得ている。それに対し本論文では、複数文から 1 文の要約文を作成するタスクであり、5%という非常に低い要約率で 62%(48/77 記事) の精度が得られた。以上の結果から直接の比較は困難だが、本手法は彼らの手法と同等もしくはそれ以上の精度が得られると考えることができる。

# 4 考察

## 4.1 類似用例の評価結果について

類似用例の評価で被験者によって類似していると判定された用例と、類似していないとされたものを比較した。類似していないと判定された用例の多くは重文構造であり、それぞれの動詞節が入力の内容と一致しているわけではなかった。すなわち、以下に示す用例のように「· · · 会見する」と「· · · 提唱」という重文構造で、入力に「会見する」という内容は存在するが、「提唱」という内容が存在しない場合があるということである。このような類似用例は入力の情報と対応がうまく取れないため、要約文を作成することが困難である。

例 4) **重文構造をもつ用例**
世界銀行総裁は２９日、日経新聞と 会見し、イラクの自立支援に向け新たな国際会議の開催を 提唱

本論文では入力と用例の類似度を算出する際に、文末及び読点の直前にくる動詞が一致した場合に重みを与えている。よって上述の例のように片方の動詞のみが一致した場合でも類似度が高くなる場合がある。また我々は類似度を計算するときに形態素の一致をみている。これにより長い複合名詞や局所的に類似する部分があると、不当にスコアが上がってしまったというような例も見られた。

これらの問題は類似度の設定で用例の構造を考慮することや自立語の全てを用いて比較するのではなく、注目すべき語を選んで比較を行うことにより解決できると考える。

### 4.2 形式的な対応のずれ

本論文での文節の対応付けは、類似用例の 1 文節に対して入力記事 1 文節を対応付けるものであった。しかし、以下に示す例では類似用例の 1 文節に対して入力記事 2 文節が対応しており、形式的な対応のずれが生じた。

例 5) **形式的な対応のずれ**
前年同月に/[*4]比べ ↔ 前年同月比
5 月を/メドに ↔ 5 月メドに

この問題は複数の文節や形態素で対応を取ることや換言によって解消することができるのではないかと考え、今後の課題にしたい。

### 4.3 作成された要約文

実際に作成された要約文の例を例 6 に示す。

例 6) **出力した要約例**
**入力したニュース記事：**
十四日の東京株式市場でソフトバンク株が急伸し、株式時価総額でトヨタ自動車を抜いて第三位に浮上した。インターネット関連の中核銘柄として、国内外の機関投資家や個人投資家の買いが集まった結果だ。日本を代表するメーカーであるトヨタの時価総額を抜いたことについて、市場では日本の産業構造の変化を象徴しているとの声も出ている。(以下省略)
**選択された類似用例：**
株式時価総額でキヤノンが９日、ソニーを抜いて電気機器業界トップに
**出力した要約文：**
株式時価総額でソフトバンク株が十四日、トヨタ自動車を抜いて第三位に
.....
**入力したニュース記事：**
神奈川県警の一連の不祥事のうち、厚木署集団警ら隊の集団暴行事件で起訴された元巡査部長、川野優被告の論告求刑公判が二十一日、横浜地裁で開かれた。検察側はひまを持て余して部下に短銃を突き付けるなど、組織における地位の高さに乗じた悪質な行為などと理不尽な暴力を指弾し、川野被告に懲役一年六月を求刑した。判決は一月十一日に言い渡される。(以下省略)
**選択された類似用例：**
大阪地裁で 22 日、[8 人が犠牲となった池田小児童殺傷事件の] [*5]論告求刑公判が開かれ、検察側は宅間被告に死刑を求刑した
**出力した要約文：**
横浜地裁で二十一日、論告求刑公判が開かれ、検察側は川野被告に懲役一年六月を求刑した

例 6 では類似用例も有効に働き、入力した記事の要約として適切な内容の要約文が出力されていることが分かる。例えば、2 つ目の出力例で「· · · 公判が開かれた。· · · 求刑した。」という入力記事にある 2 文が、類似用例に従うことで出力では「· · · 公判が開かれ、· · · 求刑した。」という要約文となった。このように既存の文抽出や文短縮手法では作成することが難しいと考えられる、複数文の文節を組合せた要約文の作成が可能になった。

## 5 結論

本論文では、要約事例を用例として使用し、その表現を模倣利用する要約手法を述べた。我々の提案する用例利用型要約の利点は保守が容易であること、語の重要度は設定しないこと、日本語としてより頑健な要約が作成できることである。なおシステムは類似用例の選択、文節の対応付け、対応文節の組合せの 3 つの処理から構成される。

システム各部の評価では類似用例の選択部と要約文の生成部をそれぞれ評価し、各々約 6 割の精度が得られた。また実際に出力した要約文を観察から、既存の手法で実現するのは難しいと考えられる複数文の情報を 1 文に圧縮した要約文が作成可能になった。

### 謝辞

本研究の一部は, 科学研究費補助金 基盤 (A)「円滑な情報伝達を支援する言語規格と言語変換技術」課題番号 16200009 によって実施した。

### 使用したツール及び言語資源

1) 日経ニュースメール, NIKKEI-goo, http://nikkeimail.goo.ne.jp/
2) 日本経済新聞全記事データベース 1999–2000 年度版, 日本経済新聞社
3) 構文解析器 CaboCha, Ver.0.53, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/~taku/software/cabocha/

## 参考文献

1] 牧野恵, 池田論史, 山本和英. 類似用例文の部分的置換による文短縮. 情報処理学会研究報告 NL173-4, pp. 21–28, 2006.

2] Kevin Knight and Daniel Marcu. Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. *Artificial Intelligence*, Vol. 139, No. 1, pp. 91–107, 2002.

3] Hongyan Jing. Sentence Reduction for Automatic Text Summarization. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pp. 310–315, 2000.

4] Joel Larocca Neto, Alexandre D. Santos, Celso A.A. Kaestner, and Alex A. Freitas. Document Clustering and Text Summarization. In *Proceedings of the 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining*, pp. 41–55, 2000.

5] Minh Le Nguyen, Susumu Horiguchi, Akira Shimazu, and Bao Tu Ho. Example-Based Sentence Reduction Using the Hidden Markov Model. *ACM Transactions on Asian Language Information Processing*, Vol. 3, No. 2, pp. 146–158, 2004.

6] 山本和英, 池田論史, 大橋一輝. 「新幹線要約」のための文末の整形. 自然言語処理, Vol. 12, No. 6, pp. 85–112, 2005.

7] 山本英子, 武田善行, 梅村恭司. 情報検索のための表記の揺れに寛容な類似尺度. 自然言語処理, Vol. 10, No. 1, pp. 63–80, 2003.

8] Dekang Lin. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of COLING-ACL98*, pp. 768–773, 1998.

[*4] "/" は文節区切りを表す。
[*5] "[...]" は類似用例内の連体修飾部を表す。本論文ではこの連体修飾部の要約表現には従っていない (2.2 節参照)。