用例文を利用したニュース記事からの単文要約

牧野 恵　池田 諭史　山本 和英
長岡技術科学大学 電気系
E-mail: {makino,ikeda,ykaz}@nlp.nagaokaut.ac.jp

1 はじめに

現在、携帯電話や電光掲示板などへの利用を目的とし、ニュース記事を対象にした自動要約の研究が盛んに行われている [1, 2]。これらは表示機器の大きさや人間の読むスピード等の制限により、一般的なニュース記事に比べ記事全体をより圧縮したものでなくてはならない。そのために文中の重要箇所を抽出したり、冗長な表現を不要箇所として削除するなどの手法が提案されている。重要箇所を抽出するためには語に対して個々に重要度を設定することが多い。しかし同じ語であっても記事の内容によって重要度が異なり、また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困難である。

そこで我々は単語や文節の重要度は設定せず、要約方法を「用例文」に委ねた用例利用型の要約手法を提案する。用例文には人手で作成されたニュース記事の要約文を用いており、類似した内容の用例文は同じような表現方法や類似した語で構成されていることが多いという特徴がある。つまり記事の内容によって要約表現が決まっている。

我々は以前に用例文と入力文で類似した文節を部分的に置換することで要約を行う手法を提案した [3]。しかし用例文は単文から作成されているものが少なく複数文の情報を凝縮した形で表されていることや、置換する文節の決定に格情報のみを用いていたことが原因で、置換できずに文として出力されない場合があった。そこで本稿ではこれらを考慮して入力を “ニュース記事” へと変更し、さらに文節の類似度を用いることで記事から単文要約を作成する手法を提案する。評価実験では抽出した文節に対する評価を行い要約率約 5%で F 値 0.46 が得られた。

2 提案手法

用例文を要約事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単文へ要約する。我々が定義する用例文とは人手により作成された要約文である。この用例文は要約文と対応した原文を用いないことから容易に収集可能である。本稿では日経 goo[1)] からメール配信された新幹線要約文 [4] を用例文として使用した。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ため、用例文として使用するのは新幹線要約記事の 1 文目に限定した。

提案手法は「類似用例文の選択」「対応文節候補の抽出」「対応文節候補の組合せ」の 3 つの処理から構成される。次節以降で各処理の詳細について説明する。

2.1 類似用例文の選択

まず入力記事に対して内容が類似した用例文「類似用例文」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と用例文の類似度を算出する。類似度には両者で一致した自立語が入力記事のどの位置に出現しているかによって決まるスコアを導入する。これは内容の類似を測る際にどの単語が記事の内容を表すものとして相応しいかに単語毎にスコアを与える妥当であると考えたためである。

スコアの算出には日本経済新聞 1999 年度 [2)] から無作為に取り出した 5000 記事を用いた。データはニュース記事のタイトルと本文 (1〜218 文) から構成されている。ニュース記事のタイトルは本文の要約であり内容をよく表したものである。またニュース記事のタイトルは例 1 に示すように本稿で使用する用例文の情報と等価である。

例 1)

ニュース記事タイトル 米 AT&T、CATV 網を開放

用例文 AT&T は高速ネット接続を可能にする CATV 網を他の通信会社に開放する

ニュース記事本文の何文目に、タイトルの自立語がどの程度含まれるかを調査した。この調査結果を図 1 に示す。図の縦軸は i 文目に含まれるタイトルの自立語総数を、5000 記事中に i 文目が存在した文の数で割ったものである。30 文目程度より長い記事はそれほど多くはないため大きく変動している。

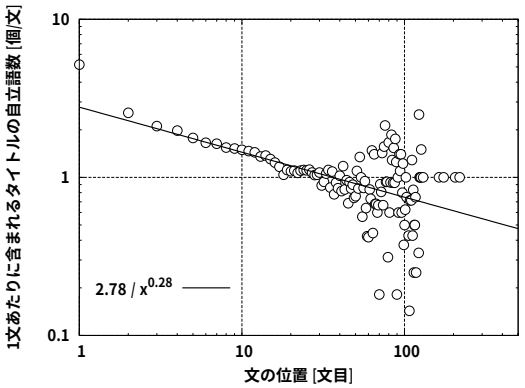

図 1 文の位置によるタイトルの自立語含有量

この図のスケールは両対数であり、図中に結果をベキ乗近似した近似曲線と式を示す。データが分散している部分でも偏りは少ないため全てのデータを近似の対象として近似式を求めている。

調査結果より記事の内容を良く表している単語を文の位置で判断できると考え「類似用例文」を選択する際のスコアに用いた。次式に用例文 A とニュース記事 B の類似度を示す。

$$Sim(A,B) = \sum_{i=1}^{n} weight_i \cdot ||T_1(A) \cap T_i(B)|| \cdot Score(i) \quad (1)$$

ここで n はニュース記事を構成している文数を表し、$T_i(\cdot)$ は i 文目の単語 (1-gram) 集合を表す[*1]。$||T_1(A) \cap T_i(B)||$ は $T_1(A)$ と $T_i(B)$ の積集合における要素数である。また用例文とニュース記事で文末の文節に含まれる動詞が一致した場合 $weight_i = 3$ としその他は $weight_i = 1$ とする。この $weight_i$ は文末動詞が内容をより表現しているという経験則に基づき導入した。$Score(i)$ は用例文の単語とニュース記事 i 文目の単語が一致したときに加算されるスコアであり次式で表す。

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

入力したニュース記事の 1 文目の単語と用例文の単語が一致した場合は i=1 のスコアを使用する。これは図 1 においてニュース記事 1 文目は近似曲線から大きく外れているため近似式を用いず調査で得られた値 (5.15) をそのまま用いた値である。1 文目以外の場合は図 1 から得られた近似式を使用する。また入力したニュース記事に対する類似用例文は式 (1) で求める類似度が高いものである。ここで “品質管理能力などの再強化策を話し

[*1] 用例文 A は 1 文で構成されているため常に $T_1(A)$ である。

合うものづくり懇談会が 24 日、会合を開催した。(以下省略)” というニュース記事に対して得られた類似用例文を例 2 に示す。

例 2)

類似用例文 1 日本外交の基盤強化へ自民党特命委は 8 日初会合

類似用例文 2 道路公団民営化委が 24 日、首相官邸で初会合を開いた

このように互いに内容が類似した用例文であってもいくつか要約表現がある。そのため本稿では類似度 1 位の類似用例文のみを使い単文要約を行うのではなく、この時点では上位 5 位の類似用例文それぞれから単文要約を作成し、後で定義するスコアで最も高かったものを最終的に出力する。

2.2 対応文節候補の抽出

次に類似用例文の文節に当てはまる対応文節の候補をニュース記事から抽出する。まず準備として「パタンの生成」を行いその後パタンを用いて「対応文節候補の抽出」を行う。

2.2.1 パタンの生成

準備としてニュース記事と類似用例文の構文解析の結果を用いてそれぞれからパタンを生成する。構文解析には南瓜 [3] を用いた。パタンは文節内の助詞と読点以外を ID_i に汎化したものである (図 2, 3 の①)。また汎化の対象となった語についても後の対応文節候補の抽出で使用するため、その情報を保持する (図 2, 3 の②)。情報を保持する際は固有表現タグ (NEtag) も考慮し、{語句, NEtag} という 2 つ組の形で保持する。入力したニュース記事から生成するパタンを「入力パタン」、類似用例文から生成するパタンを「用例パタン」と呼ぶ。これらは対応文節候補の抽出で使用する。

用例パタンの生成では構文解析結果を用いて連体修飾部を削除し、その後文節内の助詞と読点以外を ID_i に汎化する。連体修飾部を削除するのは連体修飾部の内容や要約表現が被修飾部によって異なるためである。そのため問題の簡略化を考えニュース記事の文節を抽出する際にも類似用例文中の連体修飾部は要約事例として使用しない。またサ変名詞には動詞的用法と名詞的用法がある [5]。そのためサ変名詞全てを名詞とすることや、その連体修飾部を削除することはできない。そこでサ変名詞に直接係る文節を全て参照し文節内に格助詞が 1 つも含まれていなければ名詞的用法として用いられていると判断し連体修飾部の特定、削除を行う。ただし連体修飾部の削除はパタンが短くなりすぎることを避けるため以下の条件のいずれかに当てはまる体言に対してのみ、その連体修飾部も使用する。

条件 1 文末の体言

条件 2 品詞が非自立であるもの (“こと, もの, ほか” 等)

条件 3 次のいずれかの語

“見方, 見通し, 方向, 予定, 考え, 見込み, 狙い, 意向, 計画”

図 2 に類似用例文とそこから生成した用例パタンについて示す。

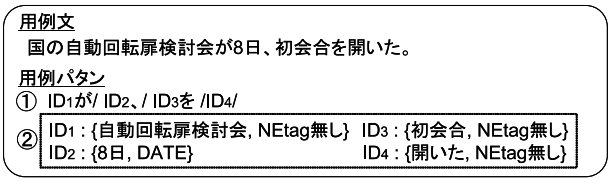

図 2 用例パタンの例

入力パタンは用例パタンと同様に生成するが連体修飾部の削除は行わない。これは対応文節候補の抽出を行う際に連体修飾部の文節が候補として用いられる場合もあることを考慮したためである。図 3 に入力したニュース記事のある 1 文から生成した入力パタンの一部の例を示す。実際は入力したニュース記事全体からこの入力パタンを生成する。

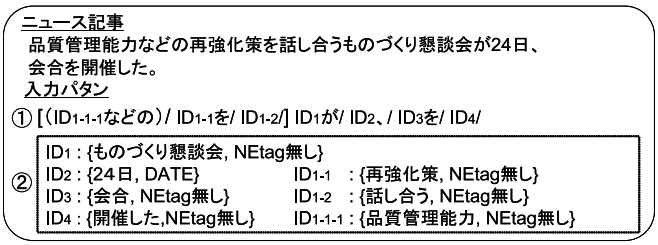

図 3 入力パタン (一部) の例

2.2.2 対応文節候補の抽出

次に入力パタンと用例パタンを用いて類似用例文の文節に当てはまる対応文節候補をニュース記事から抽出する。本稿では「格情報」「固有表現タグ情報」「拡張型編集距離」「相互情報量を用いた文節類似度」の 4 つを用いてニュース記事の文節と類似用例文の文節を比較し対応文節候補の抽出を行う。

「格情報」については用例パタンに対し入力パタンで ID_i に連接する格助詞または読点が一致した場合にそれを対応文節候補として抽出する。例えば図 2 における類似用例文の用例パタンに対し、図 3 における入力記事の入力パタンの一部から対応文節候補を抽出する。この例では用例パタンの “初会合を” に対して “会合を” や “再強化策を” が抽出される。

「固有表現タグ情報」についてはパタン作成の際保持した固有表現タグが一致した場合、対応文節の候補として抽出する。例えば用例パタンの “8 日” に対して入力パタンの “24 日” が抽出される。

さらに格や固有表現タグが一致していない文節も候補となることが考えられるため「拡張型編集距離」と「相互情報量を用いた文節類似度」も導入し、パタン作成時に保持したデータ同士を比較することで対応文節候補の抽出を行う。

「拡張型編集距離」は “日銀が” と “日本銀行が” のような文節を類似した文節として抽出できるように山本 [6] ら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、本稿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。そして用例パタンの文節に対して類似度の高い入力パタンの文節上位 3 つを対応文節候補として抽出する。

「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[7] 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。

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

上式の $*$ は文節の汎化であり、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。さらに文節 w_1 と文節 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。

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

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。本稿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 (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。これを使用し用例パタンの文節に対して類似度高い入力パタンの文節上位 3 つを対応文節候補として抽出する。

2.3 対応文節候補の組合せ

続いて抽出した対応文節候補を組合せて単文へと要約する。単文要約を作成する際、対応文節候補の助詞は用例パタンの助詞へと変更する。図 4 を用いて説明する。

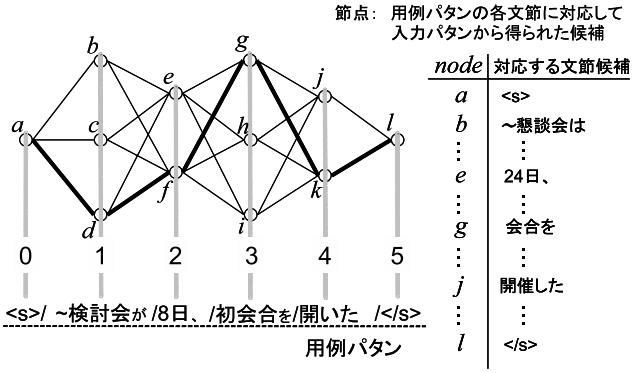

図 4 用例パタンに対応した文節の組合せによる最適経路問題

図中の数字は文節の番号を表しており節点 (ノード w_i) は前節で得た用例パタンの文節に類似した文節候補である。なお文節の番号は用例パタンの ID_i に対応しており初期状態と最終状態を明確にするため文頭記号 ⟨s⟩ と文末記号 ⟨/s⟩ を挿入した。ここでノード重み $N(w_i)$ に用例パタンの文節との類似度、エッジ重み $E(w_{i-1}, w_i)$ にノード間 (文節間) の連接の良さを導入する。これにより類似用例文に類似した文節で構成され、さらに要約として連接の良い単文要約を作成する問題は、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 $W_p = \{w_0, w_1, w_2, \cdots, w_m\}$[*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。

$$\hat{W}_p = W_p \quad \text{s.t.} \quad \operatorname*{argmax}_p Score_p(W_p) \tag{5}$$

またスコア $Score_p(W_p)$ を次式で表す。

$$Score_p(W_p) = \alpha \sum_{i=0}^{m} N(w_i) + (1-\alpha) \sum_{i=1}^{m} E(w_{i-1}, w_i) \tag{6}$$

ここで α 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 はパタンにおける文節の最終番号を表す[*3] 。以下にノード重みを定義する。

$$N(w_i) = \max \begin{cases} 0.5 & \text{助詞か NEtag が一致} \\ 1/rank & \text{それ以外} \end{cases} \tag{7}$$

これは類似用例文の文節に対応文節候補がどの程度類似しているか表すスコアである。前節で得られた候補が助詞の一致または固有表現タグ (NEtag) 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 位まで文節候補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。

$$E(w_{i-1}, w_i) = \frac{1}{|loc(w_{i-1}) - loc(w_i) + 1|} \tag{8}$$

式 (8) における $loc(w_i)$ はノードつまり文節候補 w_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。本稿では連接する対応文節候補 (w_{i-1}, w_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。連接する対応文節候補 (w_{i-1}, w_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候補の組合せは防ぐことができる。

単文要約は類似用例文上位 5 文から作成した要約で動的計画法で得られたスコア (式 (6)) が最も高かったものを最終的に出力する。

3 評価実験及び考察

3.1 実験データ

用例文には 2001～2006 年の期間に日経 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単文要約に対し客観評価及び主観評価を行った。式 (6) におけるパラメータ α には訓練で決定した 0.6 を用いた。

3.2 作成された単文要約

実際に作成された単文要約の例を図 5 に示す。

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文：
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

出力した単文要約：
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文：
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] [*4] 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

出力した単文要約：
横浜地裁は二十一日、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

図 5 作成した単文要約の例

図 5 では類似用例文も有効に働き、入力した記事の要約として適切な内容の単文要約が出力されていることが分かる。

また入力したニュース記事に対してシステムが出力した単文

[*2] 図 4 における太線ならば $W_p = \{a, d, f, g, k, l\}$ を通る経路。
[*3] 図 4 ならば $m = 5$ である。
[*4] "[...]" は類似用例文内の連体修飾部を表す。本稿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)。

要約の文字要約率は 5%(削除率 95%) であった。得られた要約率は低いものとなっている。しかし本稿では用例文の連体修飾部は要約事例として従っていないため、実際に連体修飾部も考慮すると、若干要約率が高くなる。

3.3 評価結果

本稿では客観評価と主観評価によりシステムの出力した単文要約を評価した。

3.3.1 客観評価

客観評価では正解データとシステムが出力した単文要約の文節を比較し再現率と適合率、F 値で評価を行った。本稿では用例文における連体修飾部の要約表現には従っていないため正解データの連体修飾部も削除を行った。表 1 に客観評価結果を示す。比較対象は入力したニュース記事の 1 文目である。この比較対象のデータに対しても連体修飾部の削除処理を行っている。

表 1 正解データとの文節比較による評価結果

	比較対象	本手法
再現率	0.542	0.463
適合率	0.418	0.451
F 値	0.471	0.457

この結果より適合率に関しては本手法の方が優位であったものの再現率を考慮するとあまり良い結果は得られなかった。そこでシステムの詳細な評価を行うためシステムが出力した単文要約と選択した類似用例文について主観評価を行った。

3.3.2 主観評価

主観評価では被験者 1 人に入力したニュース記事と類似用例文、システムが出力した単文要約を与え、以下を基準としてそれぞれ 4 段階で評価を行った。

1) 入力のニュース記事から被験者が考えた要約が類似用例文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

2) システムが作成した単文要約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

この評価結果を表 2 に示す。

表 2 類似用例文と単文要約の適切さの評価結果

1) 類似用例文 \ 2) 単文要約	1	2	3	4	総計
1	19	6	11	4	40
2	14	9	11	3	37
3	5	2	8	14	29
4	1	4	4	19	28
総計	39	21	34	40	134

主観評価結果の表 2 より次の 2 点が分かる。まず類似用例文を正しく選択できていない場合はそこから生成する単文要約は適切ではないことが分かる。本稿では用例文と入力したニュース記事の両者で一致する単語に重みを付けて類似用例文を選択した。重みは一致した単語が入力したニュース記事のどの文で出現しているかという情報を用いたがそれではまだ不十分であった。今後、文の位置だけではなくどの単語に注目し類似度を測るかや、一致を考えるときの複合語の処理など検討しなくてはならない。

さらに類似用例文を正しく選択できた場合でもそこから生成する単文要約は適切ではない場合があるということが分かる。これは対応文節候補の抽出や候補の組合せ方法に問題があると考える。本稿では類似用例文と入力したニュース記事の文節類似度を求め候補を抽出した。しかし例 3 のように入力したニュース記事の 2 文節が短縮され類似用例文の 1 文節に対応する例が見られた。

例 3)
前年同月に/[*5] 比べ → 前年同月比
5 月を/メドに → 5 月メドに

用例文に要約文を使用していることから、入力したニュース記事の文節を抽出することだけでは対応できない例がこの他にも見られた。よって文節をより要約に近い形に凝縮させて類似度を求めることや、そもそも文節での抽出ではなく形態素など他の単位で抽出するなどの検討が必要である。また単語類似度や文節類似度などの類似尺度にも多くの手法が提案されているため、本タスクにより合致する類似度の考察も今後の課題として挙げられる。

4 おわりに

要約事例を用例文として使用しその表現に従うことでニュース記事から単文へ要約する手法を提案した。客観評価ではニュース記事 1 文目を要約とした比較対象よりも良い結果は得られなかった。そこでシステムの問題点をより明確にするため主観評価を行った。主観評価の結果から類似用例文の選択でより良い用例を得ることが比較的良い単文要約を得ることに直結するという知見が得られた。今後は良い類似用例文を選択する際にどのような単語に注目すべきかなど厳密な調査が必要である。

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] 諸岡祐平, 江嵜誠, 高木一幸, 尾関和彦. 重要文抽出と文簡約を併用した新聞記事の自動要約. 言語処理学会第 10 回年次大会, pp.436-439(2004).

2] 大森岳史, 増田英孝, 中川浩志. Web 新聞記事の要約とその携帯端末向け記事による評価. 情報処理学会研究報告, NL153-1, pp.9-16(2003).

3] 牧野恵, 池田諭史, 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).

4] 山本和英, 池田諭史, 大橋一輝. 新幹線要約のための文末整形. 言語処理学会論文誌, Vol.12, No.6, pp.85-112(2005)

5] 山本和英, 大橋一輝. 「サ変名詞+名詞」の複合名詞への換言. 言語処理学会論文誌, Vol.12, No.3, pp.19-42(2005)

6] 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, 言語処理学会論文誌, Vol.10, No.1, pp.63-80(2003)

7] Dekang Lin. Automatic Retrieval and Clustering of Similar Words, *Proc. of COLING-ACL98*, pp.768-773(1998).

[*5] “/” は文節区切りを表す。