ニュース記事の自然な音声出力のためのテキスト変換

林 由紀子† 松原 茂樹‡

†名古屋大学大学院情報科学研究科 ‡名古屋大学情報連携基盤センター

1 はじめに

ポッドキャスティングなどの普及により，自動車の運転中や街中での歩行中など，自分が聴きたい音声コンテンツを気軽に聴取可能な環境が整いつつある．そのようなコンテンツの１つにニュース音声があり，忙しい現代人が情報を効率的に入手する手段として活用されている．現在配信されているニュース音声は，人が記事を読み上げた音声を録音することにより作成されており，今後，配信するニュースの種類と規模を増大させるために，ニュース音声データを自動で作成することが望まれる．

一般に，テキストの読み上げ音声を作成するには，単に音声合成システムを使用すればよい．近年，音声合成技術の進展は著しく，音響的にかなり自然な音声の生成が可能になりつつある．しかしながら，書き言葉と話し言葉との間には，語彙や言い回しなどに違いがあり，テキストをそのまま音声に変換しただけでは，言語的に不自然な箇所が発生することになる．音声の自然さは聞き手の聞きやすさに影響を与えるため，言語的にも自然な音声を生成することが重要である．

そこで本論文では，新聞記事テキストを用いて聞きやすいニュース音声を自動生成することを目的に，書き言葉から話し言葉への変換について述べる．本研究では，特に，体言止めが新聞記事に頻出すること，また，体言止めを含む文をそのまま読み上げた音声は自然さが大きく損なわれることに着目し，省略された表現の補完方法について検討する．

本手法では，体言止め文における，文末の名詞のタイプ，係り受け関係，時制等の情報を用いた統計的な手法により省略表現の補完を実現する．新聞記事を用いた補完実験では，79.9% の正解率を達成し，本手法の利用可能性を確認した．

2 書き言葉と話し言葉への変換

2.1 書き言葉と話し言葉の違い

一般に，自然言語は書き言葉と話し言葉に大別することができる．書き言葉は文字による伝達を意図しており，話し言葉は音声による伝達を意図している．表１に，書き言葉と話し言葉の言語的特徴を示す [1, 2]．このように書き言葉と話し言葉との間には多くの違いがあり，特に，音声として読み上げた場合には，文体の違いや体言止めの存在など，文末表現の形態の違いが音声の不自然さに大きく影響する（図１参照）．

表 1: 書き言葉と話し言葉の言語的特徴

	書き言葉	話し言葉
文体	常体が主	敬体が主
文の長さ	長め，重文や複文も多い	短め
語彙	比較的難しい	比較的易しい
語調	改まった表現が多い	くだけた表現が多い
体言止め文	有り	無し

書き言葉	話し言葉
村山富市首相は年頭の記者会見で、「創造とやさしさの国造りのビジョン」と題する所感を発表**した**。今月中に首相を囲む学者グループが発表する「村山ビジョン」の基本的な考えを示した**もの**。「わが国にふさわしい国際貢献による世界平和の創造」と銘打った非軍事分野の国際貢献など「四つの創造」を打ち出して**いる**。所感は、冒頭で戦後五十周年の節目の年のキャッチフレーズを「改革から創造へ」と**表現**。	村山富市首相は年頭の記者会見で、「創造とやさしさの国造りのビジョン」と題する所感を発表**しました**。今月中に首相を囲む学者グループが発表する「村山ビジョン」の基本的な考えを示した**ものです**。「わが国にふさわしい国際貢献による世界平和の創造」と銘打った非軍事分野の国際貢献など「四つの創造」を打ち出して**います**。所感は、冒頭で戦後五十周年の節目の年のキャッチフレーズを「改革から創造へ」と**表現しています**。

図 1: 文体及び体言止めに関する書き言葉と話し言葉の比較

本論文では，このうち体言止めを変換の対象とする．体言止めは，名詞や代名詞で終わらせることにより文章を読み手に印象づける修辞技法である．

- 要旨は次の通り。
- 年末に米紙で論争を展開。
- 同署は発砲事件とみて捜査。

などは体言止めの例である．新聞記事において特に体言止めが頻出する理由として，強調，印象付けという効果の他に，可能な限り冗長性を排除し，文章を限られた文字数に収める意図もあるためと考えられる．一方，読み上げにおいて体言止めを用いると，唐突で高圧的な印象を与えることが多く，自然な音声出力のためには，文末に適切な動詞や助動詞などを補う必要がある．

2.2 関連研究

書き言葉から話し言葉への変換に関する従来研究として，大泉らは，「構造改革する」から「構造を改革する」への変換のような，名詞化した用言の変換手法を提案している [2]．また，鍛治らは「受諾する」から「引き受ける」への変換のように，書き言葉に特有の語彙を話し言葉の語彙に言い換える手法を提案している [3]．一方で，著者らは，書き言葉から話し言葉への文体変換規則を作成している [4]．新聞記事を用いた変換実験で

は，99.8%の高い正解率を達成しており，変換規則の利用可能性を示している．

しかしながら，体言止めの補完についてはこれまで検討されていない．1995 年の毎日新聞に出現した 1,017,373 文のうち，体言止めを含む文は 300,560 文存在しており，全体の実に 29.5%を占めている．このことからも，新聞記事の読み上げ音声出力においては，体言止めの補完は不可欠な技術であるといえる．

3 体言止めの補完

本研究では，**体言止め文**を「文中で最後に出現する名詞の後に続く語句が句読点または記号のみである文」と定義する．一般的に体言止め文とは，書き手が，文末表現を原文（体言止めでない文）から削除することにより生成したものであるから，削除された表現を推定し補完することにより，元の原文に変換できる．

一般に，体言止め文の補完語句としてどのような表現が相応しいかは，文脈に大きく左右されるため，網羅性を備えた補完規則を作成することは難しい．そこで本研究では，テキストコーパスを用いた統計的な方法により，体言止めの補完を実現する．すなわち，ある体言止め文に対して，その文と同様の特徴を有する体言止め文に注目し，そこから最尤の補完語句を選定する．

3.1 補完手法の流れ

補完処理は以下に示す 3 つの段階から構成される．

(1) **体言止め文との照合**　学習データから，文中で最後に出現する名詞が体言止め文と一致する文を取り出す．

(2) **文脈に基づく絞り込み**　取り出した各文に対して，最後に出現する名詞の係り文節という文脈情報を用いて，補完の候補となる文を絞り込む．

(3) **出現頻度に基づく選択**　絞り込まれた文において，最後に出現する名詞に後続する文末表現の出現頻度を算出し，頻度が最大となる文末表現を補完語句として選択する．

図 2 に，体言止め文「保守派とリベラル派双方の論客が年末に米紙で論争を展開.」に対する補完の流れを示す．まず，(1) 文中の最後に出現する名詞が「展開」である文を学習データから取り出し，そこから，(2)「展開」に係る文節に助詞「を」が含まれる文に対象を絞り込み，最後に，文末表現ごとの頻度を算出し，最も多い「しました」が補完語句として選択される．

本章の以下では，補完処理において併せて実行する処理について説明する．

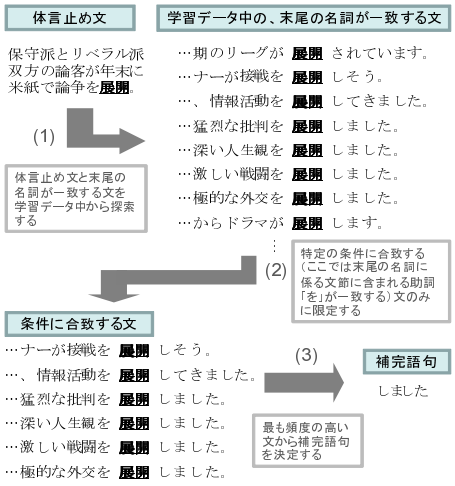

図 2: 統計的方法による補完の流れ

3.2 固有表現の置換

体言止め文との照合（処理 (1)）を促進するために，文中に現れる固有表現ならびに数値表現をカテゴリ名によって置換することにより，文を汎化する．固有表現解析には，CaboCha[5] を用いた．CaboCha による固有表現解析では、IREX(Information Retrieval and Extraction Exercise)[6] の定義による PERSON, LOCATION などの固有表現分類が使用されている．図 3 に固有表現の置換の概要を示す.「村山富市氏」を PERSON で,「二八%」を PERCENT で置き換える．置換操作は，体言止め文の末尾名詞、及び、学習データ中の文の末尾名詞に施す.

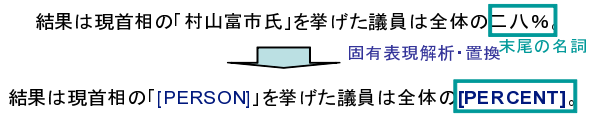

図 3: 固有表現の置換

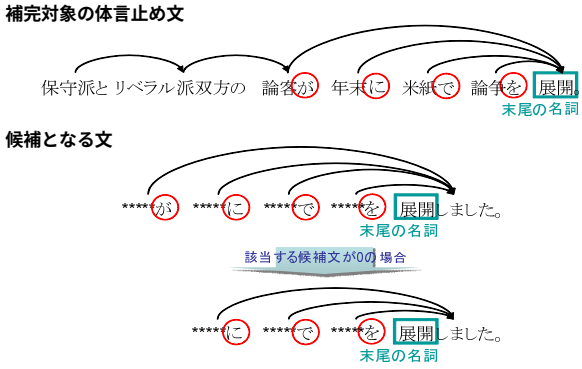

図 4: 助詞の用い方

3.3 係り文節中の助詞の利用

文脈に基づく絞り込み (処理 (2)) のために，文の最後に出現する名詞の係り文節に含まれる助詞に着目する．図 4 に，係り文節中の助詞の利用の概要を示す．まず補完対象の体言止め文と，最後に出現する名詞に係る文節に含まれる助詞が 4 種類全て一致する文に絞り込む．しかし，係り文節全てについて一致する候補文はほとんどない可能性がある．その場合，条件となっている係り文節を左から 1 つ削り，同様に候補文を絞り込む．条件となる係り文節の削除は，候補文が見つかるか，係り文節が最後の 1 つになるまで続ける．

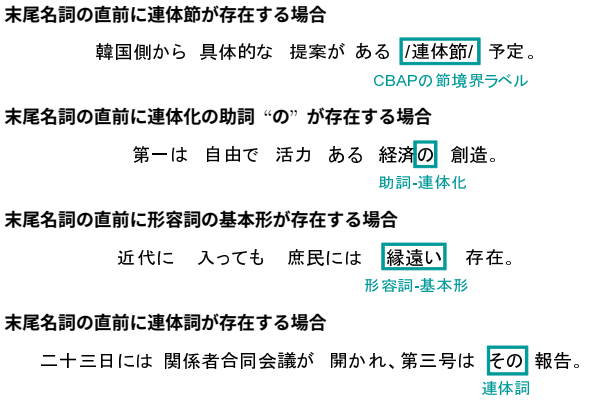

図 5: 連体修飾語の例

3.4 連体修飾語

出現頻度に基づく選択 (処理 (3)) における精度を高めるため，連体修飾語の有無によって補完語句を選択する．連体修飾語は体言を修飾する語である．文の最後に出現する名詞の直前に連体修飾語が存在する場合，名詞はその修飾語を含めて 1 つの名詞句として扱われ，後ろに「する」「できる」「なさる」「くださる」などが接続することはない．図 5 に連体修飾語の例を示す．各文の末尾名詞は全てサ変接続の名詞であるが，直前に連体節などの連体修飾語が存在し,「する」の補完は不適切である．したがって，末尾名詞の直前に以下の 4 つのいずれかが存在する場合，補完語句の候補から上述の動詞を削除する．

- 連体節（名詞を修飾する節）による修飾
- 連体化の助詞「の」による修飾
- 形容詞の基本形による修飾
- 連体詞による修飾

連体節の検出には，節境界検出プログラム CBAP[7] を用いる．

3.5 時制

出現頻度に基づく選択 (処理 (3)) において，補完語句の時制が適切になるように，末尾名詞に直接かかる文節について,「現在」「昨年」などの時を示す語句の有無を調べる．図 6 に時制の付与の例を示す．図 6 上においては，末尾の名詞「上演」に「前年に」が係っているので，補完語句も過去時制に変更する．語句が存在しない場合にも図 6 下のように時を示す語句を含む文節が見つかるまで係り受け関係をたどる．ただし，複文（述語が 2 つ以上存在する文）については，従属節中の時を示す語句が主節に影響しないように，述語（動詞や名詞＋助動詞「だ」など）が存在した場合には，それ以降係り受け関係をたどることを終える．図 7 に例を示す．従属節中の語句「3 年前の」は，主節「日本は、〜が焦点。」に影響しないため，係り受け関係をたどる処理は従属節中の述語「出場できなかった」までで終える．この文には時制は付与されない．

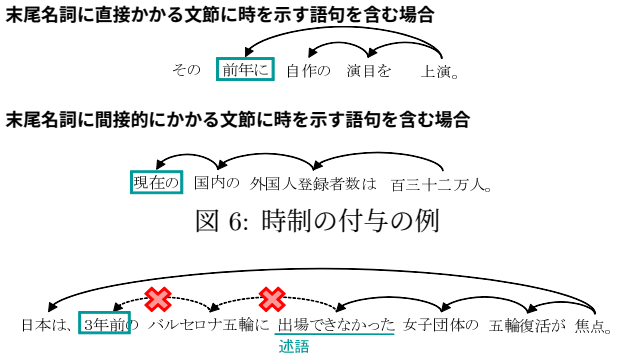

図 6: 時制の付与の例

図 7: 複文への時制の付与の例

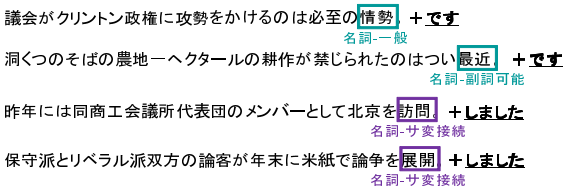

図 8: 名詞の細分類のみを利用した補完の例

3.6 補完できなかった場合の処理

処理 (1) において候補文の数が 0 であったり，処理 (2) において係り文節中の助詞の条件を緩めても補完語句が見つからなかった場合，名詞の細分類のみを用いた補完を行う．すなわち，末尾名詞がサ変接続の場合は「しました」を，それ以外は「です」を補完する．図 8 に単純補完の例を示す。

4 評価実験

4.1 実験の概要

実験には，毎日新聞 1995 年 1 月 3 日の記事 687 文のうち，体言止め文 164 文を用いた．提案手法の有効性を比較検証するため，以下の 4 つの手法を設けた．

(1) 名詞細分類に基づく単純補完
(2) 名詞細分類・時制・連体節に基づく単純補完
(3) 末尾名詞のみを用いた統計的補完
(4) 提案手法

ここで，(1) 名詞細分類に基づく単純補完とは，3.6 節で述べた末尾の名詞の細分類に基づいて補完する方法である．また，(2) は単純補完に加え連体修飾語・時制を考慮する方法，(3) は提案手法において係り文節中の助詞・連体修飾語・時制情報を使用しない方法である．

評価は以下の 3 段階で行った．

- 完全一致：正解データと全く同じ
- 適用可能：正解データとは異なるが適切
- 不一致：正解データと異なり不適切

適用可能とはあらかじめ付与した正解データとは異なるものの，適切である補完語句に対する評価であり，本研究では完全一致と適用可能を合わせて正解とした．

表 2: 体言止めの補完実験の結果

手法	完全一致	適用可能	不一致
(1)	96 (58.5%)	26 (15.9%)	42 (25.6%)
		122 (74.4%)	
(2)	103 (62.8%)	25 (15.2%)	36 (22.0%)
		128 (78.0%)	
(3)	89 (54.3%)	31 (18.9%)	44 (26.8%)
		120 (73.2%)	
(4)	102 (62.2%)	29 (17.7%)	33 (20.1%)
		131 (79.9%)	

表 3: 不適当な補完の原因の内訳

原因	数
時制	14
補完必要な文を補完不要と判定	8
語句	7
態	3
不要な文に補完	1

(3),(4) の学習データには，毎日新聞 1995 年 1 月 4 日〜12 月 31 日の記事を用いた．1,015,556 文のうち非体言止め文 715,429 文を補完語句の候補として用いた．

4.2 実験結果

表 2 に体言止めの補完実験の結果を示す．提案手法が最も結果がよく，79.9%の正解率を達成しており，本手法の有効性を確認した．また，(1) よりも (2) の方が，(3) よりも (4) の方が正解率が高く，時制・連体修飾語・係り文節中の助詞情報を用いることの有用性が示された．

4.3 考察

4.3.1 補完語句の品質

単純補完では，補完される語句が「です」「しました」に限られることになる，一方，統計的方法においては，非体言止め文の文末表現を用いるため，それらに限らない語句の補完が可能である．そのため，以下のように，末尾名詞の細分類のみを用いた補完よりも，より文意に沿った補完を行えた場合があった．

> (細分類のみ) ピークの午後五時半ごろには大阪、京都府境の天王山トンネルを先頭に栗東インター付近まで約四十キロの車の列。+でした

> (統計的方法) ピークの午後五時半ごろには大阪、京都府境の天王山トンネルを先頭に栗東インター付近まで約四十キロの車の列。+ができました

4.3.2 不適当な補完

表 3 に，不適当な補完の原因の内訳を示す．主な原因について詳細を以下に述べる．

時制の問題

> 頼りにしていた佐保も区間四位の平凡な走り。+です (誤)

文中に明示的に時を示す語句が含まれていないため，統計的方法によって選択された語句「です」が補完されているが，文脈上過去形の「でした」が適切である．こうした文に対して適切な時制情報を得るには，前後の文の時制なども考慮する必要がある．

語句の問題

> デジタル放送は一九九四年から米国で実用化されているほか、欧州連合、香港、韓国なども準備を進めているマルチメディア時代の放送。+を始めました (誤)

文法的誤りはないが，さらに語句の意味を考慮しなければこのような誤った補完語句を除外できない．

態の問題

> 普及率はＮＨＫのＢＳ放送が全世帯の五〇%に達するのをはじめ、他のＢＳ、ＣＳ局も二、三割の世帯に普及すると予測。+しています (誤)

「BS、CS 局」は末尾名詞「予測」の主語ではない．「されます」という受動態の語句の補完が適切である．格関係などを考慮し，補完語句が受動態になる場合を判定する必要がある．

5 おわりに

本論文では，テキストを音声合成ソフトウェアを用いて読み上げる場合に不自然でない音声を出力することを目的に，書き言葉を話し言葉に変換する方法について検討した．本研究では，体言止めの補完について検討し，統計的方法による補完実験を通して手法の有用性を確認した．今後の課題は，体言止め文の補完における時制や態に頑健な手法を考案することである．

参考文献

[1] 山本雅子, 大西五郎. 話し言葉と書き言葉の相互関係ー日本語教育のために一. 言語と文化（愛知大学語学教育研究室紀要）, Vol. 8, pp. 73–90, 2003.

[2] 大泉敏貴, 鍜治伸裕, 河原大輔, 岡本雅史, 黒橋禎夫, 西田豊明. 書きことばから話しことばへの変換. 言語処理学会第 9 回年次大会発表論文集, pp. 93–96, 2003.

[3] Nobuhiro Kaji, Masashi Okamoto, and Sadao Kurohashi. Paraphrasing Predicates from Written Language to Spoken Language Using the Web. In *Proceedings of the Human Language Technology Conference (HLT/NAACL-2004)*, pp. 241–248, 2004.

[4] 林由紀子, 松原茂樹. 聞きやすい読み上げ音声出力のためのテキスト変換の検討. 情報処理学会第 69 回全国大会講演論文集, Vol. 2, pp. 581–582, 2007.

[5] 工藤拓. CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer. http://chasen.org/~taku/software/cabocha/.

[6] IREX Home Page. http://nlp.cs.nyu.edu/irex/.

[7] 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝. 節境界自動検出ルールの作成と評価. 言語処理学会第 9 回年次大会発表論文集, pp. 517–520, 2003.