

書きことばから話しことばへの変換

大泉敏貴† 鍛冶伸裕‡ 河原大輔‡ 岡本雅史‡ 黒橋慎夫‡ 西田豊明‡
 †東京大学工学部電子情報工学科 ‡東京大学大学院情報理工学系研究科

{ohizumi, kaji, kawahara, okamoto, kuro, nishida}@kc.t.u-tokyo.ac.jp

1 はじめに

計算機技術の急速な発達に伴って、計算機上での表現は劇的にその自由度を上げてきた。その中でも音声合成は、新しい情報伝達方式を提供する技術であり、近年、音声合成を用いたアプリケーションが数多く開発されている。しかし、それらのアプリケーションに用いられている現時点での音声合成には、二つの問題がある。一つは、音韻やイントネーションの不正確さなどの音響的な問題であり、もう一つは、音声化されるテキスト自体の言語的な問題である。

計算機に蓄積されているテキスト情報の多くは話しことばではなく書きことばで作成されているが、それらの知的資源を活用しようとして、音声化されることを前提としていない書きことばの表現をそのまま音声にしてしまうところに問題がある。この問題は、自然言語処理における言い換え技術で解決することが可能である。音声として使用したいテキスト情報を、事前に話しことば表現に言い換えておくことで、より自然な発話を実現することができる。

以上のような観点から本研究では、書きことば表現のテキストを話しことば表現に変換するシステムを提案する。音声として利用する際にも聞き手が違和感なく受け入れられる、理解しやすい文体への変換を実現することを目的とし、システムの構築及び評価実験を行った。

2 書きことばから話しことばへ

書きことばから話しことばへの変換を行うにあたって、双方の違いを明確にしておく必要がある。そこで、書きことばとは異なる話しことばの特徴について簡単に考察する。

2.1 話しことばの特徴

音声による話しことばの表現は、瞬間的に聞き手の聴覚を刺激しながら時間とともに流れ去ってしまう。聞き手はその刻々に意味をとらえなければならず、話し手も聞き手の制約のことを考えて表現をしていかなければならない [1]。そのため、話しことばでは書きことばのように複雑な構造をとる場合は少なく、より平易で直接的な表現が用いられることが多い。

加えて、聞き手を対象とする話しことばには、何らかの待遇表現*が用いられるのが普通である。また、音声で伝えられる話しことばでは、聞いてわかりにくい語の使用は避けられるのが自然である。書きことばでは、漢字の意味から判断がつきやすい語でも、音声だけが情報源となるとまた違ってくる。そのため、「私立」と「市立」のように、同音異義語で読み替えの工夫がなされているものもある。

さらに、俗に口調と呼ばれるような要素にも、話しことばに特有のものがある。

(1) 理事会を 開き、この問題に 関し、正式に遺憾の意を…

(1) の中に出現する連用形で止める表現は、日常的な話しことばで用いられることは少なく、「開いて、…」 「関して、…」と接続助詞のついた形が一般的である。

2.2 変換の方向性

前節で挙げた特徴から、話しことばというのは、音声の状態でも情報が的確に伝えられるように、ことばが適宜補われているものであると考えることができる。また、話しことばには、省略によって簡潔に意味を伝える場合 (2a) や、自分が喋り易いように口調を整えたり、モダリティ表現を付加する場合 (2b) も存在する。

(2) a. 家来る? ← (僕の) 家 (に) 来る?

b. えー、この場合はね…

*相手に対する敬意や親しみを示す表現。敬語や丁寧体、「よ」「ね」などの終助詞などがそれにあたる。

日常的な話しことばでは、これらの表現が頻繁に出現する [2]。共有する知識を省略したり、待遇表現を用いたりすることで、情報伝達の効率化やコミュニケーションの円滑化を図ることができるからである。

だが、音声化に際しての情報伝達の効率化を目的として、書きことばテキストを話しことばに言い換えようとする場合は、聞き手の情報共有度が明確ではなく、話者の調子も考える必要がない。したがって、ニュースなど聞き手を特定しない不特定多数への情報伝達を目的とする発話が、本研究で求められているといえよう。

これらの考察に基づいて、本研究では書きことばテキスト内の情報を補い、より平易化する方向を目指す。つまり、場面や話し手、聞き手に依存しない、内容的な表現に対する言い換えを実現することを目的とする。

3 変換すべき書きことば表現

ことばとしての違いを考慮しつつ、書きことばから、自然でわかりやすい話しことば表現を生成するという観点から、変換すべき書きことば表現を分類した。

書きことば特有の表現

普通体 聞き手を想定しない脱待遇の表現

- (3) 我輩は猫である。

文語調 堅苦しく古めかしい表現

- (4) 本日 午前十時より、会議室にてミーティングを行う。

漢語調 難解な漢字を使った表現

- (5) 酸欠^{さんけつ}のために、養鱔場^{ようまんじょう}で曝気^{ばっき}をする。

複雑な構造を持つ表現

名詞化された用言を含む表現 主にサ変名詞や形容詞・副詞の語幹などを含む複合語からなり、格関係が明示されていない表現

- (6) 車両点検後、始業前点呼を行う

体言からなる複合名詞 名詞同士の意味関係が明示されていない表現

- (7) 平成十四年度公認会計士試験第三次試験

音声化に適さない表現

記号表現 視覚的な記号を用いた表現

- (8) 金属製の容器(オーバーバック)。

参照表現 前後の文章の参照を示す表現

- (9) 下記の要領で実施する。

書きことば特有の表現は、それに対応する話しことば的な表現に言い換える。複雑な構造を持つ表現のうち、名詞化された用言を含む表現は、格解析結果から格関係を補う。

しかし、漢語調の体言は研究では扱わない。また、体言からなる複合名詞と音声化に適さない表現については、適切な変換に高度の文脈解析が必要であるとともに、変換後に、かえって冗長性を増す可能性があることから、本研究では変換対象としなかった。

4 変換システムの構築

話しことばへの変換を実現するために、以下に挙げる処理を段階的に行うシステムを構築した(図1)。

1. 名詞化された用言を含む表現の変換

名詞化された用言に対する格関係を解析し、格情報などを補完する。

2. 漢語調の変換

話しことばにそぐわない漢語調の用言を、より平易なものへと言い換える。

3. 普通体と文語調の変換

普通体と文語調を話しことば的な口語調へと変換する。

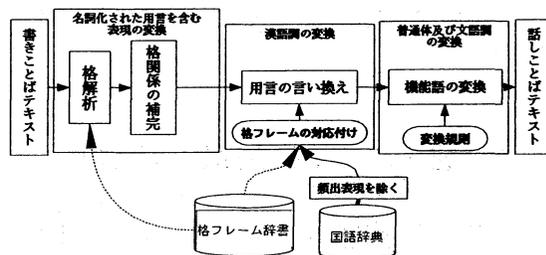


図1: システム構成

各々の変換の前処理として、JUMANとKNPによる形態素及び構文解析を行い、解析結果を利用している。

4.1 名詞化された用言を含む表現の変換

この変換では、名詞化された用言の格解析結果から、格関係を補う。格関係の解析には、河原らの提案する大規模コーパスから自動構築された格フレーム辞書に基づく格解析 [3] を用いた。

- [名詞+…+サ変名詞+する] という表現は、サ変名詞の格解析結果に従って格を補う。
 - (10) 構造改革する ⇒ 構造を改革する (改革 - フ格:構造)
- [名詞+…+サ変名詞+を+ { 行う, する, 図る }] の場合, [名詞+…+サ変名詞+する] に変換し, さらに, サ変名詞の格解析結果に従って, 修飾節のノ格を変換する。
 - (11) 内閣の大幅な改造を行う ⇒ 内閣を大幅に改造する (改造 - フ格:内閣)
- [サ変名詞+名詞] で終わる複合名詞は, 末尾の名詞が, サ変名詞に対し, { 外の関係, ガ格, フ格 } の関係にあるときはサ変名詞に対し [する] を補う。また, サ変名詞の格解析結果に従って格を補う。
 - (12) 機種依存文字 ⇒ 機種に依存する文字 (依存 - ガ格:文字 - 二格:機種)
- 接頭辞は, 形容詞または副詞と対応させて変換する (13a)。接尾辞は, [サ変名詞+ { 前, 後 }] のとき, [サ変名詞+する+ { 前 }], [サ変名詞+した+ { 後 }] に変換する (13b)。形容詞の語幹は連用修飾, 連体修飾を適切に変換する (13c)。
 - (13) a. 再逮捕される ⇒ 再び逮捕される
 - b. 実験後, … ⇒ 実験した後, …
 - c. 大幅改善する ⇒ 大幅に改善する

4.2 漢語調の変換

書きことばに含まれる漢語調の用言を, 国語辞典を利用して, より平易な表現に変換する。

漢語調の用言, すなわち難解な漢字を使った用言とは, サ変名詞・ナ形容詞・ノ形容詞から一部の基本的な用言 (運転する, 簡単だ, …など) を除いたものあると考えることができる。したがって, 書きことばに含まれる「サ変名詞・ナ形容詞・ノ形容詞」で, な

おかつ「例会小学国語辞典 [4] に頻出語として登録されていない用言」を, ここでの変換対象とした。

国語辞典を利用した変換処理は, 鍛冶らにより提案された手法を用いる [5]。

本研究では, 例解小学国語辞典と岩波国語辞典 [6] の定義文を用いて変換を行った。このシステムにより, 漢語調の変換が実現された (14)。

- (14) a. 方針を決定する ⇒ 方針をはっきり決める
- b. 規則に違反する ⇒ 規則をやぶる

4.3 普通体及び文語調の変換

書きことばは特有の表現の中でも, 「だ」「である」などの機能語を, 話しことば的な語調で言い換える。これらの語は, テキスト中に出現する頻度は高いが, その種類としては多くない。そこで本研究では, それぞれの機能的な表現に対する変換規則を作成し, 適用することで, 口語調への変換を実現する。

変換規則は, 形態素情報, 構文情報, 特に, 品詞や活用形, 係り受けなどの情報を参照することができ, より柔軟で, 高精度の変換が可能である。また, 変換規則は適度に汎用化されており, 規則の絶対数を抑えることができた。

普通体は相手の存在を前提としない脱待遇の表現であり, 話しことばでは用いられることが少ない。そこで, 公の場での発話に一般的に用いられる丁寧体[†]を, 普通体のかわりに用いることで待遇表現を補う。

- (15) 不本意だが仕方ない ⇒ 不本意ですが仕方ありません

文語調と呼ばれる古めかしい言い回しは話しことばには相応しくない。その表現と意味的に対応する, 日常的に用いられる表現へと言い換える。

- (16) 例のごとく実行すべきこと ⇒ 例のように実行すべきこと

5 実験・考察

構築したシステムを実際に動作させ, 処理結果に対する検証を行った。実験対象には, WWW 上のニュース記事 (122 文) を用いた。評価は筆者らによって行われ, 文法的整合性, 話しことばとしての違和感に対して検証した。結果を表 1 に示す。

[†]一般的に「です・ます調」と呼ばれる表現。

表 1: 各システムにおける変換精度

文法的整合性 違和感	とれている		とれていない
	ない	ある	
名詞化された用言を含む表現の変換	24	24	4
漢語調の用言の変換	40	34	26
普通体及び文語調の変換	132	10	8

文法的整合性がとれているかどうかという点での精度は、名詞化した用言を含む表現の変換が 92%、漢語調の用言の変換が 74%、普通体及び文語体の変換が 94%であった。また、文法的に正しく言い換えられた中で、話しことばとして違和感のないものは、名詞化した用言を含む表現の変換が 50%、漢語調の用言の変換が 54%、普通体及び文語体の変換が 93%であった。

名詞化した用言を含む表現の変換と漢語調の用言の変換では、話しことばの表現への言い換え精度があまりよくなかった。文法的整合性がとれなかった変換ミスは、格解析の失敗や格フレームの対応付けの失敗などがほとんどである。以下では、話しことばとして違和感の生じた変換の考察をする。

固有名詞や、慣用的に使われる複合語の過剰な変換が多々見られた。

- (17) a. 冬季アジア競技大会 ⇒ 冬季アジア競技 する 大会
 b. 司法解剖を行う ⇒ 司法が 解剖 する

中でも、末尾が漢字 1 字の複合名詞に対する過剰な変換が多い。

- (18) a. 表彰台 ⇒ 表彰する 台
 b. 実行犯 ⇒ 実行 する 犯

これらは、話しことばとしても変換しないで抑制すべきである。

変換時に、テンス・アスペクトの情報が補っていないために失敗してしまったものがある。

- (19) a. 不足状態 ⇒ 不足する状態△
 (不足している状態○)
 b. 分析結果 ⇒ 分析する結果△
 (分析した結果○)

これは、本システムが、外の関係を変換する際に考慮していないため、この問題に対する考慮が必要である。

漢語調の用言を意味的な同等句に言い換えたとしても、それが話しことば的に自然なものにならないものがあった。場合によってはより簡潔に言い換えたほうがしっくりくる。

- (20) a. 救済の必要性を痛感する ⇒
 救済の必要性を 心に感じる △
 (救済の必要性を感じる○)
 b. 入管難民法の改善案を提出する ⇒
 入管難民法を改めて なおす案をさし出します △
 (入管難民法を改めるための案を出します○)

これらの表現に違和感を感じるか感じないかは、評価者に大きく依存する問題であり、一意に基準を設定することは難しい。だが、コーパス上での出現頻度などは、その基準を検討するのに有効だと思われる。今後は、コーパス上での出現頻度を用いて、変換するか否かを判断する枠組を作成することで、より高精度な言い換え手法の実現を目指す。

6 おわりに

本研究における名詞化した用言や文語調の用言の言い換えは、話しことば的なものへ変換するというよりも、平易な表現へ変換するという方向で行った。だが、本来ならばそれぞれの平易な表現に対しても、話しことば的な言い換え方が存在しているはずである。今後、それらを大規模コーパス上から機械的に取得することができれば、より高精度の変換が望めるはずである。

話しことばはその性質上、電子的なコーパス資源に乏しかったが、近年大規模コーパスの構築も進められている [7]。それらを利用することで、本研究のように話しことばを扱うシステムが、より実用的で魅力的な機能を提供できるようになることを期待する。

参考文献

- [1] 大石初太郎. 話し言葉とは何か、「ことば」シリーズ 12 話し言葉, pp.36-51 (文化庁, 1970).
- [2] 村田真樹, 井佐原均. 話しことばと書きことばの diff, ワークショップ「話し言葉の科学と工学」, (2001).
- [3] 河原大輔, 黒橋禎夫. 頑健な格解析を実現する格フレーム辞書の自動構築, 言語処理学会 第 8 回年次大会, pp.515-518 (2002).
- [4] 田近海一編, 例解小学国語辞典, (三省堂).
- [5] Nobuhiro Kaji, Daisuke Kawahara, Sadao Kurohashi, and Satoshi Sato. Verb Paraphrase based on Case Frame Alignment, In *ACL02*, (2002).
- [6] 西尾実他編, 岩波国語辞典, (岩波書店).
- [7] 古井貞熙, 前川喜久雄, 井佐原均, 科学技術振興調整費 開放的融合研究推進制度 一大規模コーパスに基づく「話し言葉工学」の構築一, 日本音響学会誌, vol.56, no.11, pp.752-755 (2000).