WWWを用いた書き言葉特有語彙から話し言葉語彙への言い換え

鍛治 伸裕　　岡本 雅史　　黒橋 禎夫

東京大学大学院 情報理工学系研究科
{kaji, okamoto, kuro}@kc.t.u-tokyo.ac.jp

1　はじめに

近年，音声合成技術を用いて既存の電子テキストを音声に変換するアプリケーションが関心を集めている [1]．こうしたアプリケーションには，当然，自然な音声を出力することが求められる．しかし，既存のテキストの多くは書き言葉で書かれていて，書き言葉に特有の表現を含んでいるため，出力される音声が不自然になってしまうという問題がある．

本論文では，書き言葉では使われるが話し言葉では殆んど使われない語を**書き言葉特有語彙**，話し言葉で通常使われる語を**話し言葉語彙**と呼ぶ (図 1)．図中の左円は書き言葉で使われる語，右円は話し言葉で使われる語 (=話し言葉語彙) を表す．円の重複部分は書き言葉と話し言葉の両方で使う表現なので，書き言葉特有語彙は左円の中の色がついた部分にあたる．円の外は，どちらでも使われない不自然な表現を表す．

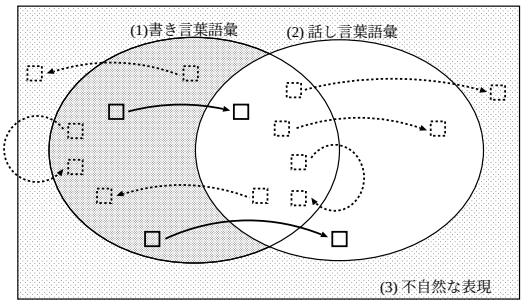

図 1: 書き言葉特有語彙と話し言葉語彙

前述の問題を解決するためには，言い換え技術を使うことができる．すなわち，音声合成の入力となる書き言葉テキストに前処理を施し，あらかじめ書き言葉特有語彙を話し言葉語彙に言い換えておけば，書き言葉テキストからでも自然な音声合成を行うことが可能になる．この言い換えは図 1 の実線矢印に対応する (それ以外の言い換えは破線矢印で表されている)．

本論文は，書き言葉特有語彙から話し言葉語彙への言い換えを学習する方法を提案する．手法の特徴は，ある表現が書き言葉特有語彙であるか話し言葉語彙であるかを，書き言葉コーパスと話し言葉コーパスでの出現確率にもとづいて決定することである．それらのコーパスは WWW から自動収集した大規模なものを使う．手法の流れは以下のようになっている．

(1) 国語辞典から用言の言い換えを学習する．ここでは，Kaji らの手法 [2] を用いる．

(2) 書き言葉コーパスと話し言葉コーパスを WWW から自動収集する．

(3) それら二つのコーパスを用いて，(1) で学習した言い換えの中から，書き言葉特有語彙から話し言葉語彙への言い換えを選び出す．

テキストを，音声合成に適したテキストに言い換えるためには，当然，用言以外の表現も言い換え対象とする必要があるが，ここでは言い換え対象を用言に限定して議論を行う．

2　用言の言い換えの学習

用言の定義文は，その見出し語の言い換えを含んでいる．このことに着目し，Kaji らは，国語辞典の定義文から用言の言い換えを学習する方法を提案している [2]．ここでは，その手法を用いて，国語辞典の定義文から用言の言い換えを学習した．使用した国語辞典は例解小学国語辞典 [4] で，5,836 パターンの言い換えが学習された．以下に,「激怒する」「相乗りする」「発汗する」の定義文と，そこから学習された言い換えを示す．

(1) a. **激怒する**　激しく怒ること
　　b. **相乗りする**　乗物などにいっしょにのること
　　c. **発汗する**　汗をかくこと

(2) a. 激怒する → 激しく怒る
b. 相乗りする → いっしょにのる
c. 発汗する → 汗をかく

こうして学習された言い換えの中から，書き言葉特有語彙から話し言葉語彙への言い換えを取り出すことが，本論文の中心的な問題となる．以下では，言い換えの左辺を source と呼び，右辺を target と呼ぶ．

3 書き言葉/話し言葉コーパスの自動収集

提案手法には，大規模な書き言葉コーパスと話し言葉コーパスが必要となるが，問題は，どのようにして大規模な話し言葉コーパスを準備するのかということである．日本語の話し言葉コーパス [3] は，近年少しずつ整備されてきているが，十分な規模とはいえない．

そこで我々は，WWW から自動収集した書き言葉コーパスと話し言葉コーパスを使う．WWW には，新聞記事などの書き言葉テキストや，友達同士のチャットなどの話し言葉テキストが混在しており，収集に用いることができる．

図 2 に収集手法の概要を示す．まず，WWW から Web ページを抽出して，そこから html タグなどを取り除く．こうして得られたコーパスを WWW コーパスと呼ぶ．そして，各ページを (1) 書き言葉 (2) 話し言葉 (3) 判断困難の三つに分けて，(1) または (2) に分類されたページだけを，書き言葉コーパス，話し言葉コーパスとして収集する．各ページを，書き言葉と話し言葉の二種類ではなく，三種類に分類するのは，WWW コーパスには人間でも書き言葉か話し言葉かの判断の難しいページがあると考えられるからである．

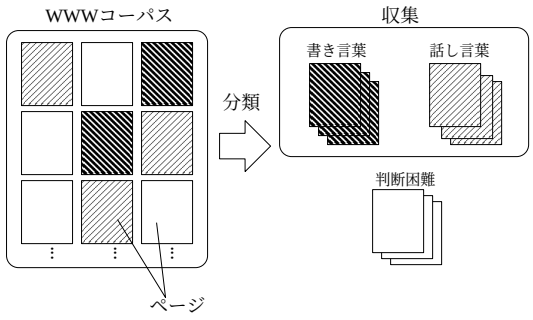

図 2: コーパスの自動収集

3.1 待遇表現に着目した収集手法

書き言葉と話し言葉の違いの一つに，待遇表現がある．待遇表現は書き言葉ではあまり使われないが，話し言葉では頻繁に用いられるという傾向がある．そこで，各 Web ページに含まれる待遇表現の割合に着目して書き言葉コーパスと話し言葉コーパスの収集を行った．ここでは待遇表現の中でも，相手に対する親愛や丁寧さを表す言い回し (親愛表現，丁寧表現と呼ぶ) に着目して，以下のような手順で収集を行った．

はじめに，親愛表現または丁寧表現を含む文を判定する．判定は次のように行った，まず WWW コーパスを Juman[1]を用いて形態素解析する．そして，助詞「ね，よ，わ，さ，ぜ，な」のどれか一つを含む文は親愛表現を含む文と判定する．次に，それ以外の文で，機能語「です，ます，ください，ございます」のどれか一つを含む文，もしくは「ですます活用」の用言を含む文は丁寧表現を含む文と判定する．

その後，各 Web ページに対して以下の二つの数値を求める．前者を**親愛表現率**，後者を**丁寧表現率**と呼ぶ．

- 親愛表現を含む文数 /Web ページの全文数
- 丁寧表現を含む文数 /Web ページの全文数

親愛表現率と丁寧表現率がともに 0 である Web ページを書き言葉コーパスとして収集する．そして，親愛表現率が 0.2 以上の Web ページと，親愛表現率が 0.1 以上で丁寧表現率が 0.2 以上の Web ページを，話し言葉コーパスとして収集する．それ以外の Web ページは判断困難とみなし，収集には用いない．表 1 に，この分類規則をまとめる．

表 1: Web ページの分類規則

親愛表現率 $= 0$	→ 書き言葉
丁寧表現率 $= 0$	
親愛表現率 > 0.2	→ 話し言葉
親愛表現率 > 0.1	→ 話し言葉
丁寧表現率 > 0.2	
上記以外	→ 判断困難

表 2 に，自動収集された書き言葉コーパスと話し言葉コーパスの一部を示す．話し言葉コーパスでは，認識された親愛表現と丁寧表現に下線を引いている．

[1] http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

表 2: 収集されたコーパスの具体例

書き言葉コーパス
診察室で医療者がみる姿かたちだけでなく，患者の目や心に映るこうした波風を理解し，それに基づいて日々の医療を創りあげて行くことが，いわゆるＱＯＬを重視した医療であろう．一人ひとりの患者のＱＯＬを知るには，基本的には彼女に問いかけ，話をよく聴く以外に術はない．

話し言葉コーパス
美味しいキムチを食べましょ〜！最近よくお客さんから質問される事があり ます.「スーパーでよくキムチを買うんやけど スーパーで売っているキムチって何で酸っぱいの？」結構こんな印象を持ってる人がたくさんいますよ ね．いつも聞かれたらこう答えるようにしてい ます.「スーパーで売ってるキムチって大半がキムチと違う物やからやで」

3.2 収集されたコーパスの評価

収集されたコーパスの規模: 収集に使用した WWW コーパスと，収集された書き言葉コーパスと話し言葉コーパスの規模を表 3 にしめす．収集された話し言葉コーパスは，既存のものと比較しても大きなものであり，十分な量のコーパスが収集できたと考えている．既存の話し言葉コーパスの中で，我々の知る限り最も大規模なものは「日本語話し言葉コーパス [3]」だが，その規模はおよそ 700 万語である．

表 3: コーパスの規模

	ページ数	語数
WWW コーパス	336,341	391M
書き言葉コーパス	38,472	38M
話し言葉コーパス	33,186	51M

適合率による評価: 書き言葉または話し言葉として収集されたページをランダムに 240 ページ取り出し，適合率による評価を行った．240 ページの中には，手法が書き言葉と判断したページが 125 ページ，話し言葉と判断したページが 115 ページ含まれていた．

二人の被験者 (以下では被験者 1，被験者 2 と呼ぶ) の判断に基づく適合率を表 4 に示す．各被験者による適合率は 95%(=228/240) と 92%(=221/240) で，その平均は 94%あった．この結果から，収集されたコーパスが高い質を持っていることが確認できた．

表 4: 自動収集の適合率

	被験者 1	被験者 2
書き言葉コーパス	95% (119/125)	89% (110/125)
話し言葉コーパス	94% (109/115)	97% (111/115)
合計	**95%**(228/240)	**92%**(221/240)

4 言い換えの選択

収集されたコーパスを用いて，2 節で学習された言い換えの中から，source が書き言葉特有語彙で target が話し言葉語彙であるようなものを選び出す．

source/target が書き言葉特有語彙であるか話し言葉語彙であるかは，その書き言葉コーパスでの出現確率と話し言葉コーパスでの出現確率から判断できると考えられる．つまり，書き言葉特有語彙は，話し言葉コーパスより書き言葉コーパスに偏って出現すると予想される．逆に，話し言葉語彙は，話し言葉コーパスに偏っているか，もしくは偏りはないと考えられる．

source が書き言葉特有語彙で target が話し言葉語彙であるような言い換えを正例，それ以外の言い換えを負例と考れば，解くべき問題は二値分類であると考えることができる．そこで本論文では，SVM を用いた手法を提案する．素性は以下の 4 つを用いた．

(1) source の書き言葉コーパスでの出現確率

(2) source の話し言葉コーパスでの出現確率

(3) target の書き言葉コーパスでの出現確率

(4) target の話し言葉コーパスでの出現確率

データセット: 2 人の被験者が SVM で学習するためのデータセット (正例が 70 個，負例が 130 個) を作成し，提案手法の評価を行った．

データセットは，[2] の手法によって学習された言い換えの中から，無作為に抽出した 200 の言い換えで作成した．言い換えが正例であるか負例であるかは，2 人の被験者が個別に判断した．200 の言い換えは，2 人の被験者の判断が一致したものだけで構成されている．

実験結果: 20 分割の交差検定によって評価を行った．実装には学習パッケージ TinySVM[2] を用いた．

[2] http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/

表 5: SVM による分類の具体例

分類結果	言い換え	素性			
		source		target	
		書き言葉	話し言葉	書き言葉	話し言葉
正例	激化する → はげしくなる	151/3,169,253	17/5,186,414	152/3,169,253	88/5,186,414
	受諾する → 引き受ける	50/3,169,253	5/5,186,414	280/3,169,253	259/5,186,414
	×きざだ → 気取っている	1/3,169,253	0/5,186,414	40/3,169,253	129/5,186,414
負例	食事する → 食べる	27/3,169,253	170/5,186,414	3,114/3,169,253	17,239/5,186,414
	引越しする → 転居する	3/3,169,253	89/5,186,414	35/3,169,253	14/5,186,414
	×軟化する → 軟らかくなる	25/3,169,253	5/5,186,414	11/3,169,253	2/5,186,414

カーネル関数を用いない場合 (linear) と，カーネル関数に 2 次，3 次の多項式関数を使った場合 (poly2, poly3) で実験を行った．表 6 に，それぞれの場合の正例と負例の分類精度，正例に分類された言い換えの適合率，正しく正例に分類された言い換えの再現率の値を示す．カーネル関数に 2 次の多項式を使った場合の精度が最も高く，79%であった．

表 6: 分類の精度，適合率，再現率

	linear	poly2	poly3
精度	72%	**79%**	76%
適合率	60%	**69%**	63%
再現率	59%	**72%**	62%

議論: SVM に与える素性として使った出現確率は，自動収集された書き言葉コーパスと話し言葉コーパスから求めている．さらに，今回用意したデータセットも比較的小規模なものである．このことを踏まえると，実験結果の 79%は十分に高い精度であり，使用した 4 つの素性は有効に働いていると考えることができる．

求められた出現確率の中には，人間の直感と反する不適切な値が計算されたものがあった．例えば「観劇する」は，書き言葉特有表現であると考えることができるが，書き言葉コーパスで 4 回，話し言葉コーパスで 43 回出現していた．収集されたコーパスを調査したところ，たまたま「劇」に関する話題が話し言葉コーパスに多かったことが分かった．WWW コーパスの規模をさらに大きくすれば，コーパス間の話題の差が小さくなり，こうした問題を緩和できると考えられる．

表 5 に，分類された言い換えの具体例と，素性の値を示す．素性の列は左から，source の書き言葉コーパスでの出現確率，source の話し言葉コーパスでの出現確率，target の書き言葉コーパスでの出現確率，target の話し言葉コーパスでの出現確率を表す．出現確率は「出現頻度/コーパスの全表現数」という表記になっている．×印は分類結果が誤っていることを表す．

5 まとめ

本論文では，まず，書き言葉コーパスと話し言葉コーパスを WWW から自動収集する手法を述べた．次に，それらを利用して，書き言葉特有語彙から話し言葉語彙への言い換えを学習する手法を提案した．そして実験を行い，いずれの手法も有効であることを確認した．今後は，提案手法を，用言以外の言い換えにも適用していきたいと考えている．

参考文献

[1] Tomohiro Fukuhara, Toyoaki Nishida, and Shunsuke Uemura. Public opinion channel: A system for augmenting social intelligence of a community. In *Workshop notes of the JSAI-Synsophy International Conference on Social Intelligence Design*, pp. 22–25, 2001.

[2] Nobuhiro Kaji, Daisuke Kawahara, Sadao Kurohashi, and Satoshi Sato. Verb paraphrase based on case frame alignment. In *Proceedings of ACL 2002*, pp. 215–222, 2002.

[3] Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. Spontaneous speech corpus of japanese. In *Proceedings of LREC 2000*, pp. 947–952, 2000.

[4] 田近洵一（編）. 例解小学国語辞典. 三省堂, 1997.