話し言葉への言い換えのための話題を考慮した単語選択

鍛治伸裕　　黒橋禎夫

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

1　はじめに

我々が日常的に用いる言語表現は，書き言葉と話し言葉に大別できる．書き言葉と話し言葉の違いを厳密に定義することは難しいが，その一つは表現の難解さであるということができる．難解な表現は書き言葉に特徴的であり，話し言葉ではさほど使われない．本論文では，書き言葉と話し言葉の違いの中でも，難解さに着目して議論を行う．

書き言葉と話し言葉の違いは，音声合成を行うときにしばしば問題となる．書き言葉テキストから音声を合成する場合には，難解な単語がテキストに含まれていて，出力音声が聞き取りにくくなる可能性がある．以下では，話し言葉で殆んど使われないような難解な単語を書き言葉特有語，それ以外の単語を話し言葉語と呼ぶ．

上記の問題を解決するため，我々は書き言葉から話し言葉への言い換えを提案している [2]．ここでいう話し言葉への言い換えとは，入力テキスト中の書き言葉特有語を話し言葉語に変換する処理をさす．話し言葉への言い換えには様々な処理が必要となるが，本論文では，入力テキスト中の書き言葉特有語を検出する処理を取り上げる．この処理を単語選択と呼ぶ．

単語選択は，単純には，次のような二値分類の問題と言える．すなわち，単語を入力として，その単語が書き言葉特有語か話し言葉語かを出力する問題と考えることができる．しかし，厳密にいうならば，単語の難解さは単語単独で規定されるものではない．難解さを規定するには，その単語が使用されている文脈や聞き手の年齢など，様々な要因を考慮する必要がある．例えば,「顧客」や「マネジメント」といった単語は，一般的には難解であるため，書き言葉特有語と考えることができる．しかし，ビジネスについて話をしているときならば，それらは話し言葉語であると言える．この例は，単語選択を行うには，単語がどのような話題で使われているかを考慮する必要があることを示している．

以上の議論を踏まえ，本論文では，単語選択の入力を，単語とその単語が使われているテキスト (=話題) であると規定する．これによって，単語が使われている話題を考慮することが可能となる．話題を考慮することは，特に，ある話題に特有の単語 (話題語と呼ぶ) を扱うときに重要になると考えられる．本来ならば，話題以外にも，聞き手の年齢なども考慮すべきであるが，本論文では扱わない．

本論文で提案する単語選択手法は，従来手法 [3] と同様に，書き言葉コーパスと話し言葉コーパスでの出現確率を使ったものである．従来手法と異なるのは，コーパスを話題ごとに分類しておき，出現確率を求めるときには入力テキストと同じ話題のコーパスのみを使う点である．これによって，入力テキストの話題に適応することが可能になる．この提案手法を二人の被験者が評価したところ，話題適応の効果を確認することができた．

2　提案手法の概要

提案する手法には，話題ごとに分類された書き言葉コーパスと話し言葉コーパスが必要となる．コーパスは，以下の二つの手続きで自動作成される．(1) まず，Web ページを書き言葉と話し言葉に分類する．こうして収集された Web ページを，それぞれ書き言葉コーパス，話し言葉コーパスと呼ぶ．(2) そして次に，書き言葉コーパスと話し言葉コーパスを話題ごとに分類する．コーパス作成については，次節で説明する．

話題ごとに分類された書き言葉コーパスと話し言葉コーパスを用いて，単語選択は以下のような手順で行われる．

- 入力として，単語とその単語が出現するテキスト

が与えられる．これらを入力単語，入力テキストと呼ぶ．

- 入力テキストの話題を判定する．これには，書き言葉コーパスと話し言葉コーパスを話題に分類する処理と同じ手法を使う．

- 書き言葉コーパスと話し言葉コーパスを使って，入力単語の出現確率を求める．このとき，コーパスは，入力テキストと同じ話題のコーパスだけを使う．

- 出現確率をもとにして，入力単語が書き言葉特有語であるか話し言葉語であるかを判定する．4 節で，この手法を説明する．

3　コーパスの作成

コーパス作成の手続きは，(1) Web ページの書き言葉と話し言葉への分類と，(2) 分類された Web ページ (=書き言葉コーパスと話し言葉コーパス) の話題への分類，の二つで構成される．

書き言葉と話し言葉への分類には鍛治らの手法を用いた [3]．手法の詳細は文献を参照されたい．書き言葉コーパスと話し言葉コーパスの話題への分類は，いわゆるテキスト分類と同じタスクである．これには SVM に基づく手法を使った．訓練データは Yahoo! Japan から収集したデータを使った．Yahoo! Japan では，話題が 14 カテゴリに分類されており，各カテゴリの Web ページにリンクがはられている．これらの Web ページを訓練データに使い，書き言葉コーパスと話し言葉コーパスを 14 の話題カテゴリに分類した．SVM の素性は，名詞の Web ページにおける出現確率，カーネル関数は線形カーネルを用いた．

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 1 に示す．また，両コーパスを 14 カテゴリの話題に分類した結果を図 1 に示す．横軸は 14 のカテゴリ，縦軸は各カテゴリに分類された書き言葉・話し言葉コーパスに含まれる Web ページ数である．

表 1: 書き言葉コーパスと話し言葉コーパスの規模

	ページ数	単語数
書き言葉コーパス	989k	432M
話し言葉コーパス	1,337k	907M

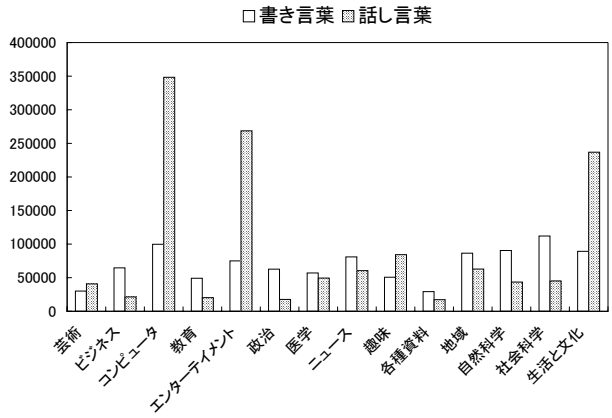

図 1: 書き言葉コーパスと話し言葉コーパスの分類結果

4　単語選択の学習

次に，書き言葉コーパスと話し言葉コーパスでの出現確率から，入力単語を書き言葉特有語と話し言葉語に分類する方法について述べる．これは二値分類問題と考えることができる．そこで本論文では，決定木と SVM の二種類の手法を試した．以下では，学習に必要な正解データ作成の手順，決定木と SVM に与えた素性を説明する．

4.1　正解データの作成

正解データは，三人の被験者によって作成された．正解データはカテゴリごとに作成される．例えば，ビジネスカテゴリの正解データは，以下のような手順で作成される．

- ビジネスカテゴリのテキストを Yahoo! Japan からダウンロードして，そのテキストから単語を無作為抽出する．

- 被験者は個別に，各単語を (1) 書き言葉特有語 (2) 話し言葉語 (3) 判断困難，に分類する．

- 以下のような単語を，書き言葉特有語として正解データに加える．

  – 三人の被験者全員が (1) に分類した単語．

  – 二人の被験者が (1) に分類して，残りの被験者が (3) に分類した単語．

  話し言葉語として正解データに加える単語も同様にして決定する．残りの単語は使用しない．

4.2 素性

決定木と SVM には次の三つの素性を与えた．(1) 入力単語の書き言葉コーパスでの出現確率，(2) 入力単語の話し言葉コーパスでの出現確率，(3) 二つの確率の比率．出現確率を求めるときには，入力テキストと同じカテゴリに分類されているコーパスのみを用いる．

5 評価

この節では，まず，正解データ作成の結果，決定木と SVM による分類精度を報告する．そして最後に，話題適応の効果を検証した結果を報告する．

正解データの作成: Yahoo! Japan の 14 カテゴリ中，ビジネスと医学カテゴリの正解データを作成した (表 2)．また，表 3 に，各カテゴリの正解データに含まれる書き言葉特有語と話し言葉語の例を示す．

決定木と SVM による分類精度: 正解データ中の書き言葉特有語と話し言葉語を，決定木 (C4.5)[1] と SVM[2] で分類して，その精度を調べた (表 4)．分類の際には，leave-one-out 法を使って学習を行った．この結果から，いずれの手法も，書き言葉特有語と話し言葉語をうまく分類できていることが分かる．

図 2 に，ビジネスカテゴリの正解データを決定木で分類した様子を示す．図の縦軸と横軸は，単語の書き言葉コーパスと話し言葉コーパスでの出現確率を表す．そして，図中の □ と × は，正解データに含まれる話し言葉語と書き言葉特有語を表す．直線は，決定木によって学習された分類規則である．直線の上部分に位置する単語は話し言葉語，下部分に位置する単語は書き言葉特有語に分類される．

話題適応の効果の検証: 決定木を用いて，正解データからビジネスと医学カテゴリにおける分類規則をそれぞれ学習した．学習は全ての正解データを使って行われた．そして，それらの分類規則を用いて，以下のような比較実験を行った．

まず，比較のための二つの単語選択手法を決めた．一つめの手法による単語選択は次のような手順で行われる．まず，入力単語を素性に変換する．このとき，入力テキストと同じカテゴリのコーパスだけを使って確率を計算する．次に，その素性をもとに，学習された分類規則を使って，入力単語を書き言葉特有語と話し言葉語に分類する．二つめの手法も同様の手順で行われるが，単語を素性に変換するときに全コーパスを使って出現確率を求める点が一つめの手法と異なる．分類規則は同一のものを使う．前者の手法を適応手法，後者を非適応手法を呼ぶ．

これら二つの手法の比較を行った．まずビジネスカテゴリと医学カテゴリの話題語を無作為抽出して，適応手法と非適応手法で分類した．この時，二つの手法の分類結果が食い違った単語を，ビジネスカテゴリと医学カテゴリから 50 単語ずつ，合計 100 単語を取り出した．そして，その分類結果を二人の被験者が個別に判断した．ここで話題語とは，有意に高い確率でその話題に出現している単語であるとした．検定には対数尤度比検定を使い，有意水準は 5%に設定した．また，実験に使った 100 単語は，正解データと重複しないように選んだ．

結果を表 5 に示す．ただし，被験者が判断に迷った

表 2: 正解データ

	ビジネス	医学
書き言葉特有語	49	38
話し言葉語	267	340
合計	316	378

表 3: 正解データの例

	ビジネス	医学
書き言葉特有語	施工する，製剤	軽微だ，選定
話し言葉語	企業，リサーチ	ケア，点滴

表 4: 決定木と SVM の分類精度

話題	手法	精度
ビジネス	決定木	91.1% (288/316)
	SVM	88.9% (281/316)
医学	決定木	91.3% (345/378)
	SVM	92.6% (350/378)

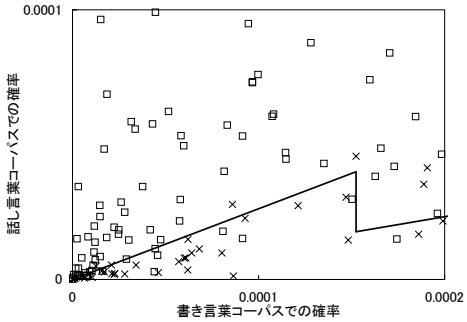

図 2: 決定木による分類

[1] http://www.rulequest.com/Personal/
[2] http://chasen.org/~taku/software/TinySVM/

場合は，手法が単語をどちらに分類しても正解に数えた．この表から，適応手法のほうが，より高い精度で分類できていることが分かる．マクネマー検定を行ったところ，有意水準 5%で有意差が確認できた．

図 3 に，ビジネスカテゴリの「顧客」と「マネジメント」を分類する様子を示す．被験者は二人とも，これらの単語はビジネスカテゴリでの話し言葉語と判断した．図中の △ は，適応手法が求めた各単語の素性を表す．いずれも，直線 (分類規則) の上部分に位置するので，正しく話し言葉語に分類されたことが分かる．一方，非適応手法で求めたときの素性は ▲ となり，誤って書き言葉特有語と分類された．医学カテゴリでも,「肺癌」や「転移」など，同様の例を観察できた．

表 5: 話題語の分類精度

被験者	手法	精度
被験者 1	適応手法	72% (72/100)
	非適応手法	53% (53/100)
被験者 2	適応手法	75% (75/100)
	非適応手法	52% (52/100)

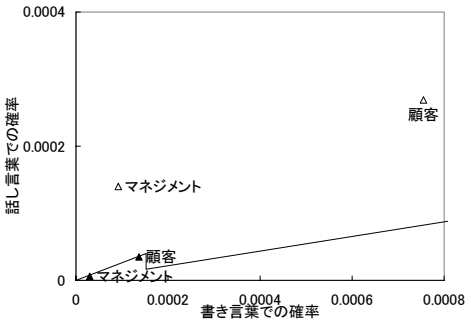

図 3: ビジネスカテゴリの例

6 議論

実験結果を観察して以下のようなことが分かった．第一に，当然のことであるが，話題語には話し言葉語が多かった．例えば，被験者 1 の判断によると，100 単語のうち，68 単語が話し言葉語，7 単語が書き言葉特有語と判断されていて，残りは判断困難とされていた．第二に，全コーパスを使う非適応手法は，話題語を書き言葉特有語に分類しやすいという傾向があった．上に示した具体例は，その典型的なケースである．この理由として，話題語は特定の話題のコーパスでしか使われないため，コーパス全体での出現確率が低くなる傾向があり，難解な書き言葉特有語との区別ができなくなるからであると考えることができる．以上二つの観察結果より，非適応手法の精度が低くなった理由を説明できると考えている．

話題適応の効果を調べるために，上記の実験以外にも次のような二つの実験を行った．一つめの実験として，全コーパスでの確率に基づく素性を使って，leave-one-out 法で正解データの分類精度を求めた．二つめに，話題語以外の単語に関しても，前節と同様の比較実験を行った．しかし，いずれの実験からも，適応手法の有効性を示すような結果は現れなかった．これには次のような理由が考えられる．まず，話題語以外の単語は，どのような話題のコーパスにでも出現するため，話題適応の効果がほとんど無いと考えられる．そして，さらに，素性を求めるために使用しているコーパスの大きさを考えると，適応手法は非適応手法よりも不利な手法であるため，分類精度が低くなったと考えられる．これらの結果と前節での結果から，適応手法と非適応手法をうまく組み合わせることが必要であると結論づけることができる．実際，言語モデルの研究では同様の試みが多数行われており，参考にすることができると考えている [1].

7 おわりに

本論文では，話題を考慮した単語選択手法を提案した．そして，実験により，その有効性を示した．今後の課題としては，聞き手などにも適応できる枠組みを構築することを考えている．

参考文献

[1] Kristie Seymore, Stanley Chen, and Ronald Rosenfeld. Nonlinear interpolation of topic models for language model adaptation. In *Proceedings ICSLP*, 1998.

[2] 大泉敏貴, 鍜治伸裕, 河原大輔, 岡本雅司, 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, pp. 93–96, 2003.

[3] 鍜治伸裕, 岡本雅司, 黒橋禎夫. WWW を用いた書き言葉特有語彙から話し言葉語彙への用言の言い換え. 自然言語処理, Vol. 11, No. 5, pp. 19–38, 2004.