情報処理学会研究報告

IPSJ SIG Technical Report

Vol.2009-GN-72 No.9
2009/5/21

表現の幅を広げる文章作成支援システム

松浦 純樹† 北澤 宏文† 小林 孝典† 市村 哲†

現在，ワープロソフトを使用しての文章作成が主流である．しかし，従来技術による文字入力は入力内容をそのまま日本語変換するだけである．これでは表現の仕方が自分任せであるため，語彙が豊富な人でないと偏った言い回しになってしまう．そのため表現が偏り単調な文章になってしまう．そこで，本研究では日本語入力システム（IME）を利用して変換単語に使用されることの多い修飾語を Web 上の文章から取得，また単語の類語を表示することで表現の偏りの解消や発想支援を行うシステムを作成した．評価実験によりシステムの有用性と課題を知ることができた．

A composition support system which widens width of the expression

Junki Matsuura† Hirohumi Kitazawa†
Takanori Kobayashi† and Satoshi Ichimura†

Word-processing software is now popular. However, Input Method Editor (IME) conventionally converts input text into Japanese as it is. If a user doesn't have abundant vocabularies, translated Japanese becomes biased expression or monotonous sentences. Our Input Method Editor imports adjective words from web services that are frequently used on the Internet. The system, also cancels deflection of the expression by displaying synonyms. This may be useful for Idea processing.

† 東京工科大学

Tokyo University of Technology

1. はじめに

現在，ワープロソフトを使用しての文章作成が主流である．ワープロソフトを使用することによって文章作成は快適に行えるようになった．しかし，従来技術による文字入力は入力内容をそのまま日本語変換するだけであり，一般の人が文章を作成する場合，語彙が豊富でないと単調な文章や偏った言い回しが繰り返されてしまうのが現状である．

文の意味を変えずに表現や言葉を変えることを英語で Paraphrase（パラフレーズ）という．英語では文章中での同じ単語の繰り返しを避ける傾向があり，整えられた文章ほどその傾向が強い．日本語の文章でも表現の言い換え，言葉を変えることは文章を書くにあたって重要なことであるという意見が文章力トレーニング講座を開いている大学教授や翻訳家のブログやニュースサイトのトピック，小説/随筆の書き方サイト等に掲載されている他，文章を書く際に役に立つ辞書として類語辞書・辞典が挙げられている．

通常，日本語文字入力には日本語入力システム（IME）が用いられる．日本語用の IME としては，Microsoft 社の MS-IME の他に，ジャストシステムの ATOK やバックスの VJE などが有名である．また，普通の日本語変換の他に，過去の入力内容を反映することで予測的な変換をしてくれる予測変換と呼ばれるものもある．しかし，予測変換も過去の自分が行った入力内容をそのまま変換することに利用されることが多いために，さらに表現が偏ってしまうという問題がある．

そこで本稿では，IME を利用して変換文字列の類語の取得，または変換単語に使用されている修飾語を Web 上の文章を探索・取得し，取得した結果を表示することで表現の偏りの解消や発想支援を行うシステムを提案する．

日本語入力システムの変換文字列の取得には， Text Services Framework（以下 TSF）を利用した．また，類語の取得には Yahoo!類語辞書，Web 上の文章の探索には Google 検索を用いて実装を行った．

以下第 2 章では提案の内容について述べる．第 3 章では実装したシステムについて述べ，第４章にシステムの評価結果を示す．最後に第５章でまとめを述べる．

2. 提案

従来技術による問題点は従来の漢字変換や予測変換では表現が偏ってしまうということである．そこで，入力された文字列を日本語変換時，変換を行っている単語の類語やふさわしい修飾語をWeb上の文章より探して表示を行うシステムを提案する．取得してきた情報をデータベースに蓄積することでネットワーク負荷を軽減するように実装した．

2.1 機能

ユーザはIMEを使い日本語入力を行う．日本語変換を行った時に，変換文字列を形態素解析(自然言語で書かれた文を，意味を持つ最小の文字列に分割し品詞に分ける作業)して「名詞」である場合に，その名詞に使用される修飾語を表示する．未確定のうちに変換を行った「名詞」に対して使われている修飾語と思われる文字列を，Google検索を用いてWeb上より検索した結果から探し出して候補があった場合にその表示を行うようにした．

また，ユーザが変換中の文字列を取得し，変換文字列を形態素解析して「形容詞・形容動詞」・「副詞」・「動詞」に対してその類語をWeb上の類語辞書から検索して表示する．検索に用いる辞書はYahoo!類語辞書とする．基本形でなければ辞書に登録されていないため，検索する語の基本形を茶筌より取得する．

「名詞」に付属する修飾語，修飾語等の類語の候補が見つかった場合には，その単語を再度変換した場合に少しでも速く表示できるようにするために，結果をデータベースへと保存するようにした．Webでの検索を行う前にデータベースを参照して見つかればそこで表示を行う．データベースの登録の際には，基本形で登録することにより重複を防いでいる．

2.2 システム構成

以下にシステム構成について述べる．ワープロソフト等で入力された変換されている日本語文字列を取得する．取得した文字列から変換中である文字列を判断し，品詞を識別する．品詞が「名詞」である場合，Google検索を利用してWeb上の文章から変換中の単語を修飾している語を探索し取得する．取得した単語を表示する形式で専用のテキストファイルへと並べて書き込む．このとき，データベースへも書き込むことにより二度目の変換時には素早く修飾語候補を表示することができる．

品詞が「形容詞・形容動詞」，「副詞」，「動詞」の場合には，その語の基本形を茶筌を利用して取得する．基本形の単語をYahoo!類語辞書にて検索し，その類語があれば一覧を取得し表示する形式でテキストファイルへと書き込む．テキストファイルに書き込まれた単語の一覧を，検索を行った単語の要素としてデータベースへ登録を行う．登録された単語を変換されたときにその表示を行う．また，登録された時にもすぐにその表示を行う．図１にこのシステムの概要図を示す．

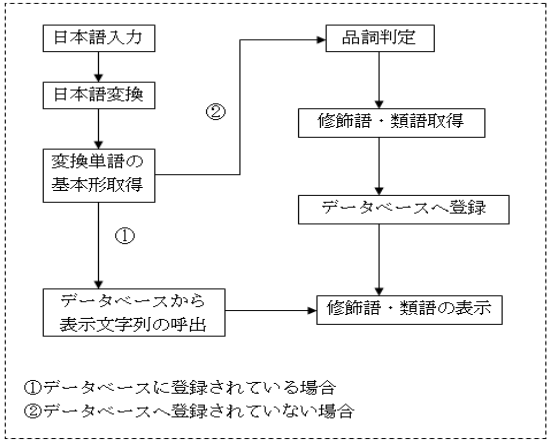

図１　システム概要

3. システム実装

3.1 日本語入力文字列処理

IME による変換中の文字列取得は TSF を使用して実装した．TSF は Windows XP 以降の Microsoft Windows オペレーティングシステムで提供されるシステムサービスである．Windows の入力システムを担当しており，先進のテキスト入力や自然言語技術を利用するためのフレームワークを提供している．TSF は，テキストの入力元の詳細を認識する必要なしに，テキスト入力を受け取ることができ，手書きや音声認識にも対応している．

単語の基本形を取得するためには，TSF を用いて取得した未確定文字列を茶筌を用いて処理する必要があるが，そのためには変換中の単語に含まれた文字列を知る必要がある．

例として，「今日は暑くなる」と入力した場合に大抵の場合は図２のように，「今日は」と「暑くなる」に分かれて変換されるはずである．このとき，「今日は」の部分を変換中には「今日」，「暑くなる」の部分を変換中には「暑」または「暑く」の部分を取得しなければならない．これは，茶筌によって「今日は暑くなる」という文字列で形態素解析を行うと図３のように「今日」「は」「暑く」「なる」と分解されるため，変換している単語の基本形を取得するためには形態素解析後の形態素を検索できる形で取得しなくてはならないからである．変換中に「今日は」と「暑くなる」に分かれているからと「暑くなる」という文字列を取得してしまうと形態素解析を行った結果である「暑く」を取得することができなくなり，基本形となる「暑い」が取得できない．

そのため，変換前の文字列と変換後の文字列を比較して変化している部分を取得することで茶筌による形態素解析を行った状態に含まれる文字列を取得するようにした．例えば「あつくなる」から「暑くなる」に変換した場合には「暑」の部分が取得され，「暑くなる」から「熱くなる」へ変換したときには「熱」の文字を取得する．

図２　文字変換

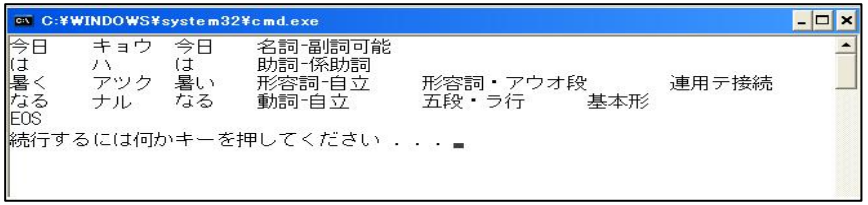

図３　茶筌による形態素解析

3.2 表示候補の取得

3.2.1 名詞変換

変換候補が名詞の場合には Google 検索により，その名詞に使われている修飾語を取得するようにした．このとき，そのまま取得した名詞だけで検索を行っても修飾語が付いていない場合が多く，固有名詞では特にその傾向が強い．この問題に対応するため，取得した単語に修飾語の語尾になる文字を付けて OR 検索を行うことで修飾語の候補が検索しやすくした．そして，Google 検索を行った結果の HTML ソースをテキストファイルとして保存する．

保存したテキストファイルから変換した名詞を検索する．このときに正規表現を用いて前方探査で文字列を検索する．正規表現ライブラリには鬼車5)を使用した．見つかった位置の 10 文字ほど前から文字列を取得し，茶筌による形態素解析を行う．この結果を取得し，品詞が修飾語になっているものを探し出してその基本形を取得する．テキスト内の名詞がなくなるまで探索し動作を繰り返す．

3.2.2 動詞・修飾語の変換

取得した変換中文字列の単語が「名詞」ではないときは Yahoo!類語辞書を用いてその単語を検索するようにした．Yahoo! 類語辞書で検索を行うためには，検索する語を URL エンコードして URL パラメタに含めて検索を行う．このとき，検索する単語は基本形にする必要がある．

取得した HTML ソースから<!--類語結果-->と <!--/類語結果-->に囲まれた部分を鬼車を用いて検索し取得するようにした．

3.3 データベースへの登録

Google 検索や Yahoo!類語辞書で取得した文字列をデータベースに登録するようにした．データベースには軽量な SQlite を使用した． 2 度目に同じ単語を変換した場合にはデータベースから呼び出すことができるため，Google 検索，Yahoo!類語辞書での検索を行う必要がなくなる．登録する単語が「寒い」という単語であれば「寒くなる」のように形が変わっていてもその基本形となる「寒い」という単語で登録するようにした．「寒い」と検索して類語を取得してきた後に「寒くなる」と入力したときでも「寒く」の基本形から「寒い」の類語をデータベースから呼び出すことができる．

3.4 修飾語・類語の表示

取得してきた類語・修飾語はポップアップ形式で表示するように実装した．類語・修飾語の一覧は，一度データベースへ登録をしてから呼び出して候補を表示する．2 度目の同じ単語の変換の場合はデータベースから呼び出すことで表示を行う．

Microsoft Word で名詞変換を使用して Web から「例」に使用されることの多い修飾語の取得例を図 4 に，動詞の変換をして「逃げる」の類語を取得している例を図 5 に示す．

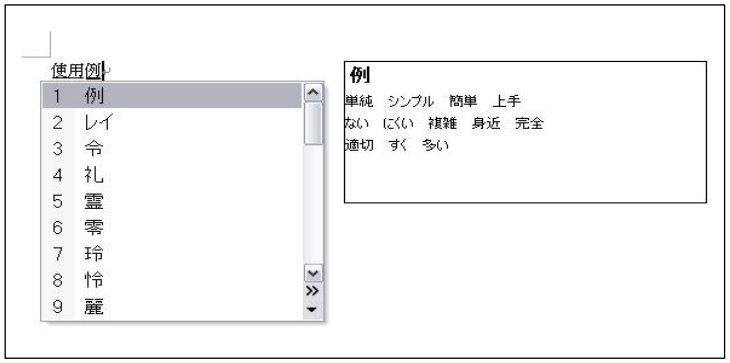

図 4 システム使用例 1（名詞変換）

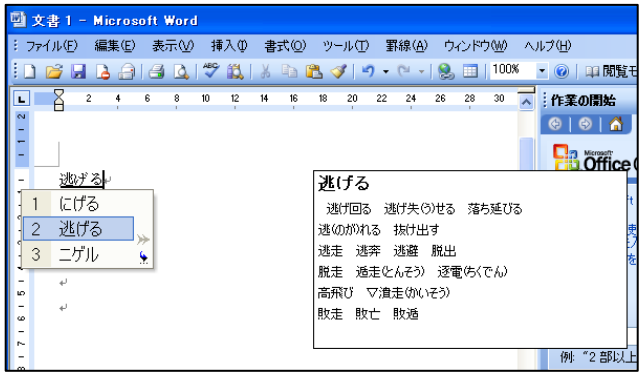

図 5 システム使用例 2（動詞変換）

4. 評価実験

4.1 実験方法

本システムの目的は，単語の言い換え，または修飾語をつけることにより，表現をより良くすることである．そこで，実際にこのシステムを使用してもらい，システムの使用に関するアンケートに答えてもらった．また，使用して作成した文と使用せずに作成した文を比較してどちらが良いかという評価を客観的に数人に対して行った．実験では，まず写真を見て 2，3 文を書いてもらい．次に，このシステムを使用した状態で写真を見て同じ文章を書いてもらった．このとき，文章を変えたいと感じたならば表現を変えて書いてもらうようにした．

4.2 実験結果

被験者 6 人に対して行った. まず, このシステムを使用し文章を入力してもらい, その後以下の各項目に 5 段階（1 ：とても悪い，5 ：とても良い）で評価してもらった．このシステムの使用に関する評価結果の平均値を表 1 に記す.

① システムの動作の速さ　② システムの使いやすさ（操作性）
③ 単語の表示の仕方　④ 使えそうな単語が表示されたか

表 1　システムの評価結果表

	①	②	③	④
平均	2.33	3.5	4	3

また，システム使用前・使用後の文を客観的に評価してもらった．以下に評価のまとめを記す．また，作成された文の例を表 2 に示す.

・システム使用後は「静かな」「壮大な」「瞬く」などの表現が良い
・「まるで素敵な宇宙だ」の「素敵な」のように無駄に思える加筆がみられる
・全体的には使用後が良い

表 2　作成された文（例）

提案システム使用なし	提案システム使用あり
水面に富士山が写っている.	静かな水面に富士山が写っている.
富士山の上に流れ星が降り注ぐ，まるで宇宙だ.	綺麗な富士山の上に，不思議な流れ星が降り注ぐ，まるで素敵な宇宙だ.
湖に富士山が写りとても奇麗な画像です.	湖に富士山が写りとても艶やかな画像です.
町の光がやたら光ってる	町の光が無茶苦茶に瞬く.
湖に映る富士山や夕焼けが対照的で綺麗だ.	湖に映る富士山や真っ赤な夕焼けが対照的で綺麗だ.

4.3 考察

実験の結果から，作成された文を客観的に評価したところシステム使用後の方が良いという意見が多かった．しかし，システム使用前の方が良いと感じる場合もあることがわかった.

また，問題点として動作が重いということが挙げられた．これは実験の際に，前の人の使用したデータベースを残してしまっていると公平な評価ができないために実験の前にデータベースを毎回消していたため，すべての単語に対して yahoo!類語辞書や，Google を用いた修飾語の取得を行っていることが原因の一つであるといえる．他の要因としては，回線の混雑，セキュリティソフトの動作が挙げられる.

システムに関するアンケートにおいて④の設問は人によって「使える単語が出た」という意見と「あまり出なかった」という意見が半々であった．評価が低かった人の感想として「もっと多くの表示が欲しい」との意見があった．Google による検索の件数を変更できるようにするべきである．また，Google 検索による修飾語の取得は結果が変化することがある．同じ単語の変換時にいつもデータベースから結果を表示するのではなく定期的に Google 検索による修飾語の取得を実行すると結果も変わり表現が増えると考えられる.

また，数人が意見を書いているが，表示するだけでなく選択したいという感想があった．現在のシステムでは，文を書き変えるためには自分で表示された文字列を入力する必要がある．これを選択できるようにするため，単純に変換候補の中に含

めてしまうと同じ読みで違う単語が存在する場合に変換候補が多くなりすぎて元々の単語を変換するのが困難になってしまう．そのため，この機能を実装するのであれば変換文字列の前に挿入してしまうという方法が考えられる．

また，表２の一文目は「水面に映る」が正しく，本システムでは「映る」を変換してみた際に「(鏡に）映る」といった類語候補が表示される．そのため回避できる間違いであるがシステム使用後も直されていない．このことから，もっと自然に意識せずとも使用できるようなインターフェースにする必要があると考えられる．

5. まとめ

本稿では，日本語変換時にその単語の類語，または名詞に使われている修飾語をWeb より取得することで候補を表示するシステムを提案した．日本語入力を行いその変換中に形容詞・形容動詞・動詞・副詞の場合には類語を，名詞の場合には付属する修飾語を表示することが可能となった．変換している単語の抽出には「茶筌」を使用することで基本形の取得が可能である．提案したシステムにより，文の表現を変える手助けをすることが可能となった．しかし，表示するだけでなく選択したいという意見や，もっと多くの表示が欲しい等の意見があった．

評価実験により考えられる今後の課題を以下に示す．

１．表示する候補を多くする
２．同じ名詞でも Google 検索を定期的に実行
３．動作の高速化
４．表示だけでなく選択を行えるようにする

今後はこれらの課題を解決し新たな機能や自然なインターフェースになるように検討してゆきたい．

参考文献

1) 茶筌（形態素解析ツール）
http://chasen.naist.jp/hiki/Chasen/

2) Microsoft IME (マイクロソフトが開発した日本語入力システム)
http://office.microsoft.com/ja-jp/ime/FX101486491041.aspx

3) PoBox (ソニーが開発した文章入力補助機能)
Toshiyuki Masui Sony Computer Science Laboratories,
Inc.3-14-13 Higashi-GotandaShinagawa,Tokyo 141-0022, Japan
http://pitecan.com/papers/HUC99/HUC99.pdf

4) 電子辞書を用いた比喩による文章作成支援システム
Transactions of Information Processing Society of Japan 42(5) pp,
1232-1241 20010515

5) 鬼車 （正規表現ライブラリ）
http://www.geocities.jp/kosako3/oniguruma/index_ja.html

6) SQlite （軽量データベース）
http://msdn.microsoft.com/ja-jp/library/default.aspx

7) iconv （文字コード変換）
http://www.gnu.org/software/libiconv/

8) Social IME ～ みんなで育てる日本語入力 ～
慶應義塾大学 理工学研究科 萩原研究室 奥野 陽
http://www.social-ime.com/