F-011

係り受け関係を考慮した難解な語句を平易な表現へ変換する手法の提案
Proposal of a Method to Convert Difficult Words to Plain Expressions Considered Dependency Relations

財満 利希† 吉村 枝里子‡ 土屋 誠司‡ 渡部 広一‡
Toshiki Zaiman Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年，人と円滑なコミュニケーションの取れる知能ロボットの実現が求められており，そのためには常識的な会話をする能力が必要であると考えられる．人とコンピュータとの会話を実現する手法の 1 つとして，ニュース記事をコンピュータの発話に利用する手法[1]が提案されている．しかしニュース記事中には，会話の際にはあまり使用しない難解な言葉がしばしば使用されているため，難解な表現をなじみがある平易な表現に変換してから会話に利用することが必要であると考える．

そこで Web サイトに存在しているニュース記事を対象に，新聞記事中の難解語を平易な表現へ変換する手法の提案[2]が行われている．しかし，既存手法において，難解と判定された語に対しての平易な表現への変換方法では，難解と判定された語を元に変換候補語の取得を行っているため，変換候補語が取得できない場合や変換候補語の数が少なく，変換できない場合や，変換後の表現として適切でないものが存在する．そこで，変換を行う際に，文中の難解語と判定された語と係わりのある語についても考慮した変換をする必要がある．語の係り受け関係を調べ，変換を行う際に必要な技術について説明する．

2. 関連技術

2.1. 概念ベース

概念ベース[3]とは，電子化された国語辞書などから自動的に構築した知識ベースである．ある語を概念と定義し，概念の特徴を表す語（属性）と，属性の重要さを表す重みの対の集合で構成されている．ある概念 A は m 個の属性 a_i と重み w_i（>0）の対により，式 1 で表現される．

概念 $A=\{(a_1, w_1), (a_2, w_2), \cdots, (a_m, w_m)\}$ (1)

2.2. 関連度計算方式

関連度計算方式[4]とは，2 つの概念間の強さを定量的に表現する手法である．関連度は 0.0 から 1.0 の実数値で表され，概念間の関連が強いほど大きな値を示す．

2.3. EMD を用いた記事関連度計算方式

類似画像検索の分野で注目されている Earth Mover's Distance（以降 EMD とする）を用いて記事間の関連性の定量化[5]を行う．記事関連度は 0.0 から 1.0 の実数値で表され，記事間の関連が強いほど大きな値を示す．

2.4. 単語親密度

単語親密度[6]とは，単語に対するなじみの度合いについて主観的に評価した値であり，被験者 40 名が 1 から 7 までの 7 段階評定を行った結果を平均化したものである．この数値が高いほど，よりなじみがある単語であることを表す．

2.5. 南瓜（CaboCha）

南瓜[7]とは係り受け解析器の 1 つである．文法規則によって文の構造を句･文節を単位として解析する．係り受けとは語の間にある修飾－非修飾の関係を指す．

2.6. 京大格フレーム

京大格フレーム[8]とは用言と名詞を用法ごとに整理したデータベースであり，Web 上の日本語テキストをリソースとして生成されている．任意の入力された用言または名詞に対し，格フレームとリソース上での出現頻度が得られる．

3. 既存手法

入力文中に一般的にあまりなじみがない語（難解語）が含まれていた場合，その語句をなじみがある表現（平易な表現）に変換する[2]．

3.1. 変換対象語の抽出

記事中で使用される語と一般的な会話文中で使用される語[9]に難易度の差があることに着目し，難易度の境界の指標として単語親密度を用いる．確率密度関数を用い，一方に属する値が他方の分布にできるかぎり属さないような状態となる値が最適であると考え，新聞記事と一般的な会話文中で使用される語のリソースとなるコーパスからそれぞれ 2000 語を無作為に選出し，境界となる閾値を求めた．単語親密度が閾値 5.81 未満の語を変換対象語とする．

3.2. 変換手法の選別

1 語変換と N 語変換の 2 つの変換手法が用いられている．1 語変換とは 1 語の変換対象語を別の 1 語に変換する手法であり，N 語変換とは 1 語の変換対象語を別の複数の語に変換する手法である．1 語変換が行えない場合，N 語変換に移行する．

3.3. 1 語変換

3.3.1. 変換候補語の取得

同義，上位，反意，類義関係にある語を格納した関係語辞書から変換対象語の同義語・類義語を取得し，変換候補語とする．

3.3.2. 単語親密度と関連度による変換語の選出

単語親密度 5.81 以上の変換候補語と変換対象語の関連度を算出し，最も関連度が高い語を変換語とする．

3.3.3. 1 語変換から N 語変換へ移行する条件

関連度の閾値設定は概念ベースの評価方法である *X-ABC* 評価を参考にして行う[3]．評価セットは基準概念 X と，概念 X と高関連の概念 A，中関連の概念 B，まったく関連のない概念 C の 500 組のセットから構成される．

難解語と変換語との関連度が，*X-A* 間の平均値である 0.34 より低い場合に N 語変換へ移行する．

3.4. N 語変換

3.4.1. 変換候補文の取得

変換候補文は大辞林[10]から取得する．変換対象語と一致する辞書中の見出し語の定義文を変換候補文とする．

3.4.2. 多義語の選定

見出し語が多義語である場合，1 つの見出し語に対して複数の定義文が与えられており，変換候補文が複数取

† 同志社大学大学院理工学研究科
Graduate School of Science and Engineering, Doshisha University
‡ 同志社大学 理工学部
Faculty of Science and Engineering, Doshisha University

得される場合がある．そこで多義語の意味特定手法として EMD を用いた記事関連度計算方式を用いる．変換対象語を含む入力文と変換候補文間で記事関連度計算を行い，複数ある変換候補文のうち，入力文との記事関連度が最も高くなった 1 文を選出する．

3.4.3. 不要語句の削除

変換候補文となる辞書中の定義文内には置き換えの際に不要となる語が含まれている場合がある．例として「～の別名」，「～の呼称」，「～など」，「転じて」などが挙げられ，これらを不要語句として削除する．

4. 提案手法

4.1. 複合語処理

入力文中に複合語が含まれている場合があり，変換対象語を抽出する際，複合語を構成する単語ごとに分けられてしまい，不自然な変換が行われてしまう．そこで変換対象語の抽出を行う前に，入力文中に名詞が連続していた場合，複合語とみなし，1 つの単語として処理する．

4.2. 係り受け関係を考慮した変換

変換対象語と係り受け関係がある語を考慮することにより新たな変換候補語を取得し，関連度を用いて変換する手法を提案する．まず，南瓜を用いることにより変換対象語と係り受け関係にある語を調べる．その際「名詞－用言」の関係とする．次に，変換対象語と係り受け関係にある「名詞」あるいは「用言」と格助詞を京大格フレームへの入力語とし，変換候補語を取得する．そして，取得した変換候補語と変換対象語の関連度を算出し，最も関連度が高い語を変換語とする．

4.3. システム全体の流れ

最初に 1 語変換，次に係り受け関係を考慮した変換，最後に *N* 語変換の順に変換を試みる．システム全体の流れを図 1 に示す．

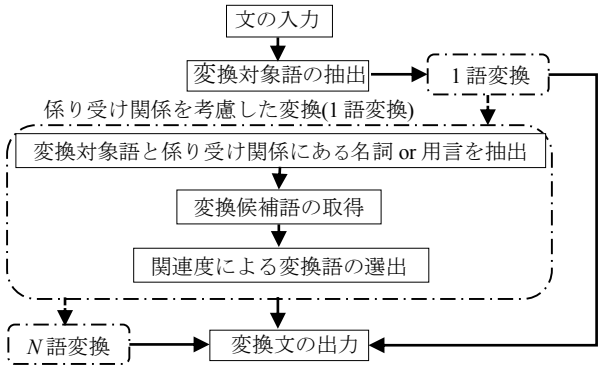

図 1　システム全体の流れ

5. 評価

評価には朝日新聞から取得した 50 記事からランダムに選んだ記事文を使用した．全単語数は 1567 語，うち単語親密度の閾値によって難解語と判断された語は 236 語である．被験者 3 名に変換前と変換後の文を比較してもらい，平易性と意味保持性について評価した．平易性に関しては，変換前の文より変換後の文の方が平易な表現となっている場合は○，変換後の文より変換前の文の方が平易な表現となっている場合は×とした．また意味保持性に関しては，変換前の文と変換語の文を比較し，意味が保持され違和感がない場合は○，少し違和感がある場合には△，意味が違っていたり，日本語表現としての意味が保持されていない場合は×とした．既存手法（1 語変換+*N* 語変換）と提案手法（1 語変換+係り受け考慮+*N* 語変換）において，平易性に関する評価結果を表 1，意味保持性に関する評価結果を表 2 に示す．

表 1 平易性に関する評価

	既存手法	提案手法
○	83.5%	85.6%
×	16.5%	14.4%

表 2 意味保持性に関する評価

	既存手法	提案手法
○	75.8%	77.1%
△	10.1%	12.3%
×	14.1%	10.6%

6. 考察

提案手法では平易性，意味保持性ともに，既存手法より精度が向上した．1 語+*N* 語変換に係り受け関係を考慮した変換をはさむことで，*N* 語変換では変換されなかった変換対象語の変換が行われたと考えられる．しかし，今回追加した手法では「名詞－用言」の係り受け関係のみを考慮しているため，変換に移れない場合も多く存在した．これは，京大格フレームの，名詞から用言，用言から名詞を取得するという仕組みが関係しており，この関係性を考慮して，変換候補語を取得する手法を取り入れたためである．今後は変換対象が 1 語の場合だけでなく，慣用句や言い回しなど，複数の語や句を別の語や句へ変換する手法を取り入れることにより，更なる精度の向上が期待できると考える．

7. おわりに

本稿では，語と語の係り受け関係を考慮することにより，難解な語句を平易な表現へ変換する手法の精度向上を目指した．その結果，平易性に関しては 2.1%，意味保持性に関しては○が 1.3%，△が 2.2%の精度を向上することができた．

謝辞

本研究の一部は，科学研究費補助金（若手研究（B）24700215）の補助を受けて行った．

参考文献

[1] 吉岡 孝治, 吉村 枝里子, 土屋 誠司, 渡部広一, “常識的連想によるニュースヘッドラインからの会話文生成”, 情報処理学会研究報告（知能と複雑系）, 2010-ICS-158, No.4 (2010).

[2] 芋野 美紗子, 吉村 枝里子, 土屋 誠司, 渡部 広一, “新聞記事中の難解語を平易な表現へ変換する手法の提案”, 自然言語処理, Vol.20, No.2, pp.105-132 (2013).

[3] 奥村 紀之, 土屋 誠司, 渡部 広一, 河岡 司, “概念間の関連度計算のための大規模概念ベースの構築”, 自然言語処理, Vol.14, No.5, pp.41-64 (2007).

[4] 渡部 広一, 奥村 紀之, 河岡 司, “概念の意味属性と共起情報を用いた関連度計算方式”, 自然言語処理, Vol.13, No.1, pp.53–74 (2006).

[5] 藤江 悠五, 渡部 広一, 河岡 司, “概念ベースと Earth Mover’s Distance を用いた文書検索”, 自然言語処理, Vol.16, No.3, pp.25-49 (2009).

[6] 天野 成昭, 近藤 公久, “NTT データベースシリーズ日本語の語彙特性（第 1 期 CD-ROM 版）”, 三省堂 (1999).

[7] 工藤 拓, “CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer”, http://taku910.github.io/cabocha/.

[8] 河原 大輔, 黒橋 禎夫, “高機能計算環境を用いた Web からの大規模格フレーム構築”, 情報処理学会自然言語処理研究会資料, 2006-NL-171-12, pp.67-73 (2006).

[9] 国立国語研究所, “日本語話し言葉コーパスの構築法”, 国立国語研究所, pp.82-91 (2006).

[10] 松村 明, “大辞林 第 2 版”, 三省堂, (2006).