6P-3

言い換え処理時に発生する誤りを自動修正するための後編集に関する研究

大田 健翔 †

† 京都工芸繊維大学 情報工学課程

1. はじめに

計算機を使った自然言語の言い換え処理技術は，様々なタスクに対して活用されている．

情報検索を例にとれば，東芝ソリューションでは，企業が蓄積している大量の業務文書のデータ内から目的の文書を検索する際に，元の検索文を言い換えた表現のものも検索文として使用することで検索の精度を高めている[1]．また，文章平易化のタスクにおいては，先天的聾者のための web 情報読解支援ブラウザの研究がなされていた[*1]．以上で述べたタスクだけでなく，様々な分野へ応用されている．

生成したい言い換え文章はタスクの目的に依存するが，言い換え処理は日本語-日本語間での翻訳と捉えることができるため，翻訳のルールと辞書を差し替えることで，言い換えエンジン自体は別のタスクへ転用できると考えられる．つまり，言い換え処理は分野(タスク)を横断する技術になる可能性を秘めていると言える[2]．

本研究では，最終的な目標を，分野横断を可能にする言い換え処理のミドルウェアの実現とするが，その効果を測定するには，タスクの量を考えればコストが膨大にかかる．そのため本研究では，対象を平易化処理にタスクを絞る．部分的な扱いにとどまるが，一つのタスクに絞って性能を向上させることは，言い換え処理の持つ分野横断的側面から見れば，他のタスクでも同様の結果になる可能性があると考えるからである．したがって，本研究の目的は平易化処理のタスク内で精度を改善することとする．

2. 先行研究

平易化処理に関する研究を 3 件挙げる

鍛冶らの研究では，入力文中に辞書内の語彙が存在した場合，入力文と語彙の言い換え表現の間でマッチングを取りながら入力文章中に埋め込むことで，格の重複を避けながら平易文章を生成する手法を提案した[3]．この手法では，辞書の見出し語と，言い換え表現の間で格フレームパターンが変化した時に誤りが起こるとしている．

洞井らの研究では，新聞記事の内容を，馴染みのある表現へ言い換える手法として，単語の 1 対 1 語変換だけでなく，説明が困難な場合は 1 対 N(≥2)語変換によって，言い換え文を生成するとしている[4]．1 語変換においては単語の置換なので成功するが，掲載されている例が N 語へ変換する名詞同士の場合にとどまっており，名詞から動詞へ変換する場合は難しいと考えられる．

梶原らの研究では，文字列置換による名詞の置換と，「サ変名詞+サ行変格活用の動詞」を動詞に言い換える手法を取っている[5]．この研究での特徴は同等表現が N 語の場合は注釈をつけるというものである．これによって，埋め込みの際の活用や格助詞の変化を考慮せずに入力文の読みやすさを高めようという試みである．しかし，1 語で置換できない語が文中に頻出した場合，注釈の数が増え，文章と注釈を往復しなければならないため，別の理由から読みやすさを損なう原因となってしまうと考えられるため，改善の余地がある．

3. 提案手法

本研究では，上述の梶原らの研究[5]を参考として改善を図る．

具体的には，梶原らでは注釈としていた部分を本文内に完全に埋め込み，その後に編集処理をかけることで，活用形の修正や埋め込み箇所の前後における格助詞の訂正を行う．概略図を図 1 に示す．

3.1 言い換えエンジン

全体図を図 2 に示す．処理の流れを説明する．入力文を形態素解析(MeCab[*2] を使用)し，各形態素の原型をキーとして，言い換え表現を格納した辞書データベースにアクセスする．言い換え表現が見つかった場合は，アクセスしたキーを対応する表現で置換する．「サ変名詞+サ行変格活用の動詞」の場合は活用形を修正する．

A Study of Post-Edit for Auto-Correcting Errors in Paraphrased Sentence :
† Kensho Ota (Kyoto Institute of Technology)

[*1] http://www.jst.go.jp/kisoken/presto/complete/jyohou/seika/2ki/04.pdf

[*2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

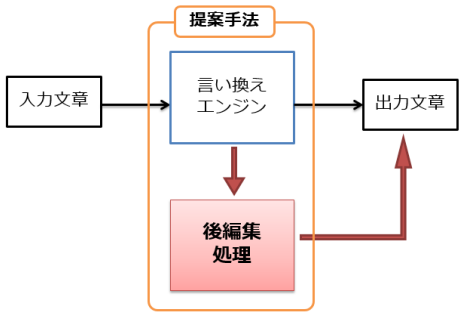

図 1 : 提案手法の概略図

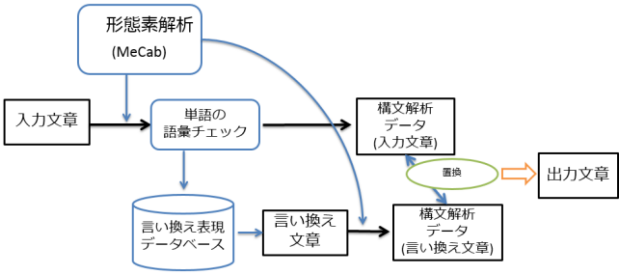

図 2 : 言い換えエンジンの全体図

3.2 後編集処理

後編集処理には統計的機械翻訳(SMT)の技術を用いる．学習したデータの異なる，SMT を 2 つ用意した．

一つは，3.1 のエンジンに生成させた誤りのある文とそれを人手で修正したデータ対で学習させた翻訳モデルを使用した SMT であり，もう一つは，語学学習 SNS「Lang-8」の添削ログから生成したコーパスを用いた SMT である．

当初の予定では，ルールベースによる後編集処理も考えていたが，3.1 のエンジンにおいて文字列を埋め込むときに考慮すべきルールに組み込むことと同等の結果であったので，扱わないこととする．

4. 実験

言い換え表現の為の辞書は梶原らの研究[5]と同様，小学生向けの国語辞典[6]を利用した．実験の前段階として，語釈文から見出し語と同等の内容の部分のみを手作業で抽出し，辞書ファイルを作った．実験用データは「asahi.com[*1]」のヘッドライン記事を 3 日分，計 100 文取得したものを用いた．この 100 文を言い換えエンジンにかけて，誤りの生じた文 60 文を収集し，手作業で修正した．学習データ 50 文，ディベロップメントデータを 10 文として SMT モデルを学習させた．ツールは Moses[*2] と GIZA++[*3] を利用した．テストデータは梶原らの研究[5]の付録で公開している計 50 文のうち，言い換えに失敗しているもの及び，注釈部分を本文に組み込んだ時に誤る例合計 21 文を使用した．

5. 結果

想定していたことではあるが．修正すべき箇所以外の部分も翻訳するため，文章として破壊される傾向が見られる．現在，どのような評価尺度導入するか検討中の段階である．

6. おわりに

本研究では，言い換えエンジンのミドルウェア実現を達成するための第一歩として，文章の平易化をタスクに選択し，後編集処理を挟むことで誤りのない日本語文へ訂正する手法を提案した．

今後は，学習データ数を増やすことで精度が上昇するか共に，入力文を新聞記事以外の場合にどのように修正されるかを確認したい．

謝辞

本研究では奈良先端科学技術大学院大学の水本智也氏から翻訳モデルを頂いた．
水本氏及び，Lang-8 の関係者各位にこの場を借りて感謝致します．

参考文献

[1]齋藤佳美, 倉田早織, 加納敏行. パラフレーズ技術を利用した情報・知識利活用ソリューション. 東芝レビュー, Vol. 64, No. 8, pp. 66-69, aug 2009.
[2] 乾健太郎, 藤田篤. 言い換え技術に関する研究動向. 自然言語処理= Journal of natural language processing, Vol. 11, No. 5, pp. 151-198, oct 2004.
[3] 鍛冶伸裕, 黒橋禎夫, 佐藤理史. 国語辞典に基づく平易文へのパラフレーズ. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2001, No. 69, pp. 167-174, jul 2001.
[4] 洞井知彦, 吉村枝里子, 土屋誠司, 渡部広一. 語句変換による難解文から平易文への言い換え手法. 情報処理学会研究報告. Vol. 2011, No. 1, pp. 1-6, mar 2011.
[5] 梶原智之, 山本和英. 小学生の読解支援に向けた語釈文による換言. NLP 若手の会第 7 回シンポジウム, 2012.
[6] 湊吉正監修. チャレンジ小学国語辞典第五版. 株式会社ベネッセコーポレーション. 2011

[*1]:http://www.asahi.com/
[*2]:http://www.statmt.org/moses/
[*3]:http://code.google.com/p/giza-pp/