

6B-7

意味情報を用いた校正支援

清水 富門 柏木 貞一 披田野 陽一 石井 敬子

(株)日本電子化辞書研究所

1. はじめに

(株)日本電子化辞書研究所(略称EDR)では、大規模な電子化辞書(言葉に関する情報を計算機が利用できるようにデータ化したもの)を開発中である。EDRの電子化辞書の特徴は、実際に使われている言葉に関して、その用法や意味との対応関係を記述した単語辞書と、概念辞書(人間が言葉によって表現する意味(概念)の内容を表現する枠組)とが別になっていることである。これは、主に概念の内容に関する記述が、複雑かつ膨大で単語辞書の枠内に納まらないことによる。我々は、この膨大な意味情報を含んだ電子化辞書の利用技術の開発の一環として、日本語文校正支援システムを開発中であるが、ここでは校正支援における意味情報の利用に関して検討した結果を述べる。

2. 日本語文校正支援システムの目的・機能

2.1 目的

本システムの用途としては、一般の日本語文の校正の支援、日本語から他の言語への機械翻訳における前編集の支援、他の言語から日本語への機械翻訳における後編集の支援等を予定している。校正支援システムは、色々な人が色々な目的で使う可能性があるので、校正支援の項目を細かく分け、意図する項目を選択的に適用できるようにしている。

2.2 機能

本システムの主だった機能は次の通りである。

① 原文入力機能

文章作成者はワープロ感覚で、日本語文を入力することができる。

② 校正支援機能

校正文に対して表記上の誤りや不統一を検出したり、表記上、表現上の、誤りとは言えないが場合によっては問題となりうる特徴の検出を行なう。校正支援機能は、「送り仮名の誤り検出」、「指示語の使用頻度の検出」等、校正項目を細かく分けて選択的に適用できるように

してある。将来的には、校正者が日本語文の書き換えをするための手助け情報を提供することも予定している。

③ 原文・校正文対照保持表示機能

本システムでは、校正者が原文と校正文とを見比べる必要がある時のことを考え、原文とそれを校正してできた校正文とを一文づつ対応を取りながらファイルとして保存している。従って、校正者は校正の途中で、対応する原文を並べて見られるので、原文にどのような誤りや問題点があったのか、校正をして文章として良い物になったのか、他のもっと良い直し方はなかったのか等について再度考えることができる。

④ 印刷機能

当然の事ではあるが、原文及び校正文をラインプリンターに出力することができる。

3. 校正項目とその実現方法に関する検討

ここでは、我々が予定している校正支援の項目と、それを実現する為の手段、特に、意味に関する情報の利用の必要性、方法に関して検討した結果を述べる。ここで述べる校正支援の項目、及びそれを実現する手段には、計画中のものも含めている。全体の把握が容易となるので、その一部をも合わせて述べた。

3.1 表記上の誤りおよび不統一の検出

(1) 表記上の誤りの検出

表記上の誤りには、漢字の誤り、送り仮名の誤り等がある。

(a) 漢字の誤り

完全に誤った漢字を使った場合には、あり得ない言葉が文中にあったとして、多くは検出できる。漢字誤りの典型例に、例1のような同訓異綴り語の使い方の誤りがある。この種の誤りを検出するために、我々は、概念間の繋がりの可能性を評価し、概念辞書に記述する予定である。

例1 { 人に会う。
 { 意見が合う。
 { 事故に遭う。

Support of proofreading by using semantic information

Fumon Shimizu, Teiichi Kashiwagi, Youichi Hidano, Noriko Ishii

Japan Electronic Dictionary Research Institute, Ltd.

(b) 送り仮名の誤り

送り仮名を完全に誤った場合には、あり得ない言葉が文中にあったとして、多くは検出できる。しかし、例2、例3、例4に示すように、動詞および動詞の派生形は、複数の送り仮名の振り方が可能なことがあり、正誤の境も、必ずしも明確ではない。

例2 〔表わす 例3 〔浮かぶ 例4 〔田植え
〔表す 〔浮ぶ 〔田植

(2) 表記上の不統一の検出

表記の不統一には、漢字書き、ひらがな書きの混在、外来語の表記の不統一、送り仮名の不統一などがある。表記の不統一は、文章全体を解析し、どの語がどこでどの様な表記をされたかに関する表を作成することにより検出できる。しかし、例5に示すように同じ語であっても、使われる場合の意味の微妙な違いによって好ましい表記が異なるのではないかと思われる場合もあり、厳密には、文章中での語の意味を詳細に分析し、意味と表記の対応関係を細かく考慮する必要がある。

例5 〔君の答え
〔この問題の答

3. 2 表記上の特徴の検出

ここでは、間違いではないが、場合によっては問題となりうる表記上の特徴の検出に関して述べる。

(1) 漢字書きかひらがな書きか

極端なひらがな書き、難解な漢字の使用、宛て字の使用等が問題となりうる。これらの検出は適当なテーブルを用意し実行している。

(2) 記号類の用い方

算用数字、漢数字の使用、特殊記号の使用等が問題となりうる。これらの検出は容易である。

3. 3 表現上の特徴の検出

ここでは、間違いではないが、場合によっては問題となりうる表現上の特徴の検出に関して述べる。

(1) 難解な用語の検出

難解な用語は、人が読んで理解する場合に問題となる。難解な用語の検出には、その語をしっている人の割合、内容を表現する上でのその語の必要性といったものを評価して辞書中にもたなければならない。こういった事も語の重要な意味的側面の一つであると考えられる。難解というのとは少し異なるが、見慣れぬ略語の使用等も問題となりうる。

(2) くだけた用語の検出

くだけた用語とは、フォーマルでない語、すなわちどちらかというと自己中心的に物を見た時の呼称である。

例6 フォーマルな表現 くだけた表現
御飯<----->飯
食べる<----->食う
母<----->ママ

くだけた表現は、正式な文書等では問題となりうるが、これは言葉のにごり、すなわち言葉の中への感情の混入といったものを評価し辞書中に記載することにより検出可能になる。

(3) 簡潔な表現と説明的な表現の選択

例7に示すように、同じ事であっても簡潔に表現したり説明的に表現したりする事ができることがある。新聞などでは簡潔な表現が、会話などでは説明的な表現がそれぞれ望ましいと思われる。そこで、校正項目としても簡潔な表現の提供、説明的な表現の提供等が考えられるが、これは、語、句で表わされた概念が概念辞書の中にあるかどうかを検索する機能、概念辞書の中の概念を言葉として表現する方法をすべて出力する機能を設けることにより、可能となる。

例7 簡潔な表現 説明的な表現
炎上する<----->燃え上がる
登山する<----->山に登る
激怒する<----->激しく怒る

(4) 文構造の複雑さの検出

文構造が複雑な文は、日本語から他の言語への機械翻訳において特に問題となるが、これは文の長さを計ったり、文中の連体修飾、接続表現の数を調べたりする等、比較的表層的なチェックで検出することができる。

4. まとめ

日本語文の校正支援として考えられる項目、およびその実現における意味情報の利用の必要性・方法について検討した結果を述べた。語の意味に関する情報を用いれば、誤りのない文章への校正支援より一歩進んで、上手な文章への校正支援が可能になることが分かった。

参考文献

- 1) 内田他,「自然言語処理のための電子化辞書の構成法」,情報処理学会第35回全国大会,1S-4(1987)
- 2) 角川書店編,「新しい常用漢字の書き表し方」,(1981)