日本語校正支援システム (Joyner) の研究について (2)
－誤用候補展開処理－

2J－5

伊吹 潤　中村 直人　徐 国偉 松井 くにお
富士通研究所

1　はじめに

文章中の誤りの検出、訂正を行なう方式として正しい単語情報と共に誤った単語を形態素辞書に登録する方式[1]が知られているが、この枠組では検出できる誤りが狭い範囲（登録されたもの）に限られるという問題点をもつ。カタカナ語句の表記の揺れや漢字同音語誤り等に範囲を限定して一般的な対処が可能なシステムの提案[2]も行なわれているが、表記レベルの誤り全般を統一的に処理できるような枠組は未だない。

我々は表記誤り全般を広範囲に検出できるようにするため、混同しやすい単語あるいは文字列同士をグループ化した情報（誤用候補情報）を単語情報とは独立して保持し、一旦正しい単語のみを利用してテキストを形態素解析した後で、これを用いて誤りの検出と誤り内容の推定を行なう仕組み（誤用候補展開）を実現した。ここでは、誤用候補展開部の処理目標とした誤りについて述べ、その処理のための枠組について説明する。

2　処理対象とする誤りの種類と処理方針

新聞校正作業に関する統計[1]から表記レベルでの誤りを見ると、許容できない誤りとしては同音異義語誤りや入力ミスが上位を占めることが判る。我々は同音異義語、入力ミス（文字の混同等の予測可能なもののみ）に加え、表記の揺れ（新聞校正時には誤りとして扱う）を処理対象として選択した。

まず、形態素解析時に単語境界が正しく認識されるかという基準によって対象を文脈依存単語誤り（単語境界は正しく認識される）、非単語誤り（認識されない）に分類し、各々に対する処理方針を検討した。

2.1　文脈依存単語誤りへの対処

単語自体を見ただけでは誤りかどうかが不明で、前後の文脈によって始めて判るような種類の誤りであり、下の例に示すように同音異義語が主なものである（括弧内が正しい単語）。

- 漢字の同音異義語による誤り
  ex. 差し示す（指し示す）、安全保証（安全保障）
- 漢字の類義語による誤り
  ex. 非安定（不安定）、客脚（客足）

既に[1]で指摘の通り、この場合は形態素解析結果だけで誤りであるかを正確に判断することは難しい。我々はこれらを複合名詞、複合動詞との照合（正解語探索）によって訂正を行なうという方針をとっており、このために動詞や名詞を文脈となり得る部分と共に検出して、最終的な判断は正解語探索によって行なうこととした。

2.2　非単語誤りへの対処

この種類の誤りでは形態素解析で単語の認識に失敗する。処理対象としては次のような誤りを選択した。

- 漢字の同音文字、類形文字による誤り
  ex. 不情理（不条理）、均衡（均衡）
- カタカナ表記の揺れ、誤り
  ex. テイルランプ（テールランプ）、ホットドック（ホットドッグ）
- 平仮名単語の表記の揺れ
  ex. そのとうり（そのとおり）

テキスト中に非単語誤りが誤りがあれば単語辞書とテキストとのマッチングに失敗する。形態素解析では未登録語が出現するだけでなく、単語境界の認識がうまく行なわれず、本来の境界以外の場所に境界が出現する可能性がある。

我々のシステムのもつ形態素解析部は他の処理とも共用されるために誤り部分の特定のためのチューニングを行なうことはできない。そのため、形態素解析結果に対する後処理という形で非単語誤りの処理をを行なうこととし、処理誤りの認識と正しい単語境界による抽出のためのヒューリスティックスを導入している。

Japanese spelling corrector(Joyner)(2)
Jun IBUKI, Naohito NAKAMURA, Guowei XU, Kunio MATSUI
Fujitsu Laboratories Ltd.
1015 Kamikodanaka, Nakahara-ku, Kawasaki-shi 211 Japan

3 誤用候補展開処理の基本的枠組

設計方針を実現するための内部構成を示し、主に誤りの検出の仕組みについて説明する。

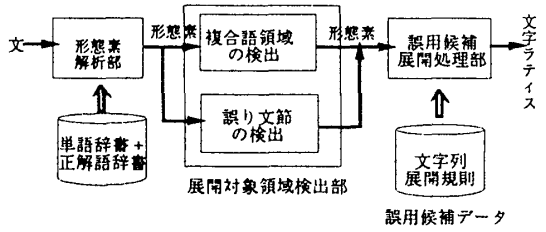

図 1: 誤用候補展開処理部の構成

3.1 形態素解析

入力文字列に対する一般的な形態素解析を行なう。ただし正解語探索の対象とする複合語句は単語辞書内に登録しておくことを前提としている。こうすれば、これらの単語は文脈を持たない単単語と認識され、（不必要な）文脈依存誤りに対する処理は起動されない。

3.2 展開対象領域検出部

先に述べたように文脈依存単語誤りの処理のために、複合語領域の抽出を行なう。又非単語誤りの処理のために誤り区画の抽出を行なう。

複合語領域の抽出　ここでは名詞の連続部、動詞の連続部を複合品詞の存在可能な領域として抽出する。

ex. 安全 保証 への 影響 を 差し 示す。（下線部が抽出対象）

誤り区画の抽出　ここでは正しい単位での切り出しのためにまず確実な境界によって区切られた区画への分割を行ない、その後、区画毎に単語の認識誤りをチェックする。

1. 確実な単語境界による分割

   活用語句の認識に失敗した場合、語幹部分が分離されて後続の平仮名部分全体に影響が及ぶことが多い（下例を参照のこと）。このような境界を避けるために、単語境界中から確実な境界の基準として次のものを選択することとした。

   - 助詞「を」、「、」等の記号の前後
   - 平仮名から漢字に字種が変わる箇所

   ex. ／彼 (名) は (助) ／黙 (動)って (尾) うなづ (未知語) い (動) た (尾) ／（／が確実な境界）

2. 誤り区画のチェック

   形態素解析では漢字の一語単語の連続やカタカナ領域の細分化等に対して全体を未登録語として処理失敗と認定するような処理が行なわれているが、ここでは処理失敗と認定する対象をより一般化するために次のような基準を設定している。

   - 未登録語を含む
   - 形態素解析結果が字種毎に設定した想定単語長よりも細かく分割された。

     （ex. カタカナ領域の想定単語長：4 漢字領域の想定単語長：2）

   ex. ／テイル ランプ／を／点灯する／（下線部が誤り区画）

3.3 誤用候補展開部

検出された複合語領域、誤り区画の各々に対して正解語（綴り）を含む解候補の生成を行なう。複合語に対しては形態素辞書中に記述した同音異義語データによる展開処理を行なう。誤り区画に対する展開処理では更に外部の規則データを利用して文字レベルのマッチングに基づく展開を行なっている。

同音異義語データについては、単語辞書から品詞によって名詞類、各活用タイプ毎の用言類等に分類した後で同一の読みをもつ単語グループを抽出した。これによって誤る可能性の低い候補への展開の危険性は増すが、候補を広く求めることを優先した。

文字レベルの展開規則は現在、漢字、カタカナ、平仮名の各字種に対応した規則を持っている。これらのデータは次の情報を利用して整備を行なった。

- 社内の校正システム用の誤用と正解のペアデータ（約 4 万例）
- 部首索引付きの漢字辞書データ（類形漢字データの抽出に利用）

4 まとめ

現在はプロトタイプの動作確認が済んだ段階であり、今後は大規模データに対する評価によって解の網羅性、複数解のコスト付けに関する評価を行なう予定である。

最後に、新聞の校正過程や原稿の誤り例についての詳細なデータや助言を頂いた福岡克氏に感謝する。

参考文献

[1] 松井他：" 日本語校正支援システム (Joyner) の研究について (1)"，情報処理学会第 52 回全国大会 2J-4(1996)

[2] 野崎：" かな漢字変換と漢字かな変換を共に用いる同音語誤りの検出方式"，情報処理学会第 45 回全国大会 4C-2(1992)