4M－5

日本語文章校正ツール“Chanterelle” －入力ミス及び表記揺らぎについて－

奥村 薫
Microsoft Corporation

1. 始めに

日本語の誤り検出では、英語のスペルチェックに相当するほど有効な検出手法は、まだ一般化していないといえよう。ある種の文法チェックが入力ミス検出に先行した面もあるが、入力ミスのほうがユーザにとっては、より切実な問題である。

今回研究開発した日本語校正ツール、愛称“Chanterelle”では、入力ミスに対してバックグラウンド・チェックを行うに十分な検出率と低い過剰警告率とを、初めて実現した。本稿では特に重要な入力ミス、及び表記揺らぎについて考察する。なおChanterelleはWord 2000に日本語校正機能として搭載される。

2. Chanterelleの特徴

（1）過剰な警告が少ない

実用化する際に最も重要な要件である。1ページあたり数個以上の過剰警告があると、そのツールを邪魔だと感じることが、ユーザビリティでも観察されている。Chanterelleでは、従来の入力ミス検出に比べて、過剰な警告が1／2～1／5に減っている。

（2）広範囲の文章に対応

従来の誤り検出では、口語的あるいは専門用語の多い文が混じると、過剰警告が極端に増えてしまう傾向があった。Chanterelleはこれらに対しても、比較的警告が少なく、安定した結果を得られる。

（3）表記揺らぎをバックグラウンドで検出する

人が見つけにくい表記の不統一を、入力と同時にチェックする機構を備えている。

（4）莫大なエラー・コーパスを元に研究・検証

校正ツールのトレーニングには、実際の誤り例が重要である。今回は、15,000件のエラー・コーパスをもとに、間違いの分析と精度向上を図った。本コーパスはEncarta（百科事典）作成時に生じた赤字から抽出したものである。

3. 入力ミスの検出

伝統的には、形態素解析に失敗した個所を、未知語としてチェックする方式がある。しかし対象を絞らない限り、辞書をかなり充実させても、「未知語＝間違い」ではないケースが多すぎた。また付属語部分に関しては、口語的表現あるいは、多少変わった言い回しなどがあると解析が失敗やすい。一方、入力ミスであっても、半数近くは何らかの形態素解析結果が成り立って、見落としてしまう。

Chanterelleではいくつかの手法を併用して、入力ミス検出を行っている。

（1）統計情報による検出

我々が形態素解析コンポーネントとして用いているT-Hammer〔1〕は、辞書・文法情報・及び統計情報を元に解析を行い、未知語に対しても、比較的安定した結果を得ている。形態素解析段階での“日本語らしさ”を用いて、辞書にはなくとも日本語らしさがある程度高いものに対しては、警告を出さない。これにより、過剰な警告を大幅に減らすことが出来た。

検出例：気づいてしまってので登録しておきます。
→しまったので

（2）特定文節パターンによる検出

明らかなミスであるのに、文節レベルの形態素解

Japanese Style Checker "Chanterelle" - Detecting of Typo-error and Inconsistent Writing Variations
Kaoru Okumura, Natural Language Group, Microsoft Corporation
One Microsoft Way, Redmond WA 98052-6399, U.S.A.　e-mail: kaoruo@microsoft.com

析が出来てしまうものも多い。ただし、複数文節を調べると、入力ミス時に現れやすいパターンがある。
検出例:書いていたわかったのですが、→いて

(3) 不自然なアルファベット

ローマ字入力の際には、母音が欠けたり、子音を打ちすぎたりすると、仮名にならないアルファベットとして残ることがある。

- 1 文字のみのアルファベットが日本語中に現れた場合に検出対象とする。
- 大文字は、略号等意識して入れている可能性が高いので除く。
- このミスでは母音“aiueo”は、発生しないので除く。また、nも多くの場合には「ん」となり、“n”自体の出現頻度も高いので除く。

検出例:対応でｋない→できない

精度

文章校正ツールの精度のデータは対象の文章や間違いの質・量によってかなり異なってくる。Chanterelle の入力ミスに関する過剰な警告は、通常の文章に対して、文字数／過剰警告数＝1,500～2,500程度。検出率については未知語によるチェックと同等との結果が出ている。

4. 表記揺らぎの検出

「コンピューター」と「コンピュータ」のように、いずれが間違っているわけではないが、統一したほうが望ましい語句が同一文章中に混在している現象を、「表記揺らぎ」と呼ぶ。局所的には正しく見えるため、人々が最も気づきにくいもののひとつである。

アルゴリズムによる揺らぎ判定

- 片仮名語　　ギリシャ & ギリシア
- 全角／半角　　ＩＰＳＪ & IPSJ
- 数字　　平成 11 年 &平成十一年

意味 ID による揺らぎ判定

- 送り仮名　　表わす & 表します
- 漢字／仮名　　山歩き & 山あるき

バックグラウンドでの揺らぎ検出

揺らぎを検出するには、文書全体にわたって対象語句を比較する必要がある。さらにバックグラウンドでユーザの入力・修正と並行してチェックするために、対象単語の集合を効率的に保持し、追加・削除に対応できるデータ構造を導入した。この構造により、次のような UI を実現している。

- 表記 A が既に文中にあり、その揺らぎである A’が入力された場合には、遡って A に対してもチェックを付与する。
- 表記揺らぎ A、A’があり、A’が修正あるいは消された場合には、A についていたチェックを消す。
- A の頻度が A’に対して極端に多い場合には、A’に対してのみチェックを付与する。このメカニズムにより、頻出単語を 1 回だけ間違えた場合に多量のチェックが出てしまうような、わずらわしさを回避することが出来る。

5. 速度及びサイズ

Chanterelle の速度及びサイズは次の通りである。

測定環境 200MHz Pentium、64MB メモリ
　　Windows98 及び Windows NT4.0
速度：　約 1,250 文字／秒
コンポーネントのサイズ：　約 1,700KB
Working set：　約 3,920KB

6. 今後の課題

入力ミス過剰警告の質

過剰警告の量はかなり少なくなったが、まれに意外な個所が入力ミスと判定されることもある。過剰検出の質についても、更なる研究をつづける予定である。

入力ミス書き換え候補の強化

入力ミスの一部に対しては今回書き換え候補を提示しているが、その率は、まださほど高くはない。高品質の日本語を生成し、評価するための形態素解析の研究も、今後考慮に値する。

参考文献

〔1〕 Patrick Halstead, 奥村薫："ロバストな日本語形態素解析－辞書依存性の低いハイブリッドアルゴリズムの提案－", 情報処理学会第５４回全国大会 P.2-55,56 (1998)