文書作成履歴を利用した校正支援機能

2J－3

山田洋志　竹元義美
(NEC 情報メディア研究所)

1 はじめに

コンピュータ上で作成される文章の量は年々増加している。文章の作り方も,以前の推敲済みの文章の清書用としての使われ方から,下書き・推敲段階を含めた使い方に変わってきている。それに伴い,文章の誤りを自動的に見つけ出し,あるいは,文章の読みやすさについて助言する校正支援システムの有用性も増している[1]。

現在,多くの校正支援システムでは文書中の誤りを検出するために,大量の校正知識(用例や誤り例)を利用している[2, 3]。こういった大量の知識の収集や選別には,大きなコストがかかる。しかし,ユーザや文書内容などによって対象を限定すれば,はるかに少ないデータで効果をあげられるはずである。

われわれは,個人あるいは共同作業をするグループの文書作成支援について研究を行っており,その一環として,文字の抜け,カタカナ語表記や同音語選択の誤りなどの表層の誤り検出を目的とした校正支援機能の開発を行っている。本報告では,誤り検出のために,入力されたテキストと過去に作成されたテキストとを比較し,差分が小さい場合に検出対象とする方式について述べる。

2 従来の検出方法と問題点

本節では,従来の誤り検出方法の問題点を説明する。

多くの校正支援システムは,文章の誤りを検出するために,形態素解析をベースにした誤入力判定と校正知識を利用した誤用判定を行っている。(文章の読みやすさなどの評価を行う場合は,解析や大規模な知識を使用せず,比較的少ないルールを使用するシステムもある[4]。)

形態素解析による誤入力判定は,誤入力によって日本語の文法からはずれた文字列ができるのを見つけ出すものである。また,校正知識による誤用判定は,かな漢字変換の同音語選択誤りのように,個々の単語は正しいが用法が誤っている箇所を,単語の共起関係の利用や頻出する誤りパターンの登録によって見つけ出すものである。

表1: 入力誤りの例

種類	例
脱落	アルゴリズム → アルゴズム
挿入	これからの → これからのの
置換	人口問題 → 人工問題

いずれの方法によっても,誤りを見つけるためには大量のデータが必要である。形態素解析による検出では単語辞書の語彙が少ないと正しい語を誤りと過剰に検出してしまう。また,単語の共起を利用して用法の誤りを見つける場合,単語の組み合わせについての情報を必要とするため,データ数が数十万に達することがしばしばである。

データ量の問題を解決する方法のひとつとして,検査対象となる文章を特定個人や分野で絞りこむことが考えられる。

これにより,実際に必要になる校正知識の量を減らすことができるが,どの知識が必要であるかの選択は依然として困難である。あらかじめ知識を分類しておき,ユーザあるいはシステムがいくつかの分類を選択する方法もあるが正確な分類は困難である。また,用意された分類がユーザの必要とする範囲と一致しなければ,結局不適当な校正知識を用いることになる。

3 文書作成履歴の利用

前節の問題を解決するために,校正知識をシステムが用意するのではなく,すでに作成されている文書を正解として利用し,誤り検出を行う方式を検討した。一般に,ユーザや分野を狭くすると何度も出現する単語や言い回しが増えてくる。そこで,過去に作成したテキストを校正知識の代わりに利用して誤り検出を行う。つまり,検査対象となるテキストを過去に作成したテキストと比較して,類似した文字列があればタイプミスなどの可能性があるとして警告する。"類似"の定義としては,テキストの相違(脱落,挿入,置換)が一定数以下であること,とした(表1)。文字列照合のアルゴリズムは、Wagnerと Fischerによる2次法[6]を用いている。

A Proofreading Function Using Document Writing History

Hiroshi Yamada and Yoshikazu Takemoto
Information Technology Research Labs, NEC Corp.

この機能によって以下のような誤りが検出できる。

文字の誤入力 タイプミスによる誤字・脱字が検出できる。形態素解析による誤り検出とは違い,一般に辞書に載っていないような特殊な語であっても,以前に使われていれば検出することができる。漢字については,かな漢字変換時点で変換結果がおかしくなりユーザが自分で誤りに気付くことが多い。ひらがな語やカタカナ語の入力の場合に,無変換や字種変換で入力する場合は変換用辞書を参照しないので,入力誤りに気付きにくい。

単語の誤入力 同音語の選択誤りやタイプミスの結果が別の単語になった場合は形態素解析による検出ができない。本方式では,前後関係が同じテキストがあれば検出できる。

カタカナの表記ゆれ 長音や母音の使い方など複数の表記が可能なカタカナ語について,以前と違う表記を使った場合に検出できる。

いずれの誤りに対しても入力テキストに対応する既存テキストを誤り訂正用のデータとしてユーザに提示,あるいは置換することができる。

4 試作システム

本検査機能をFEP型校正支援システム[5]へ組み込んだ。

このシステムは,Windows3.1のIMEの出力文字列をフックし,文書検査を実行して結果を専用の警告ウィンドウに表示するもので形態素解析による検査機能も備えている。

比較に使用するテキストはファイル名で指定する。テキスト(確定文字列)が入力されると,そこから4文字以上の漢字列およびカタカナ列を切り出して指定ファイルから類似文字列を検索する。類似文字のしきい値は2とした。すなわち,指定ファイル中に1ヶ所か2ヶ所だけ異なる文字列があれば警告する。漢字熟語の多くは2文字なので,しきい値を2とすることで,同音語選択誤りによる単語の置換を検出することができる。

5 今後の課題

本方式による誤り検出の課題をあげる。

誤り区分の判定 誤り検出の方法が誤りの種類によらないため,誤りの種類によって対処を変えたい場合は,何らかの区分方法が必要になる。

検出精度 類似判定のためのしきい値は,大きすぎると誤り以外の箇所を多く検出してしまい,小さすぎると誤りを見逃しやすくなる。また,長いテキストをそのまま探索すれば,類似したテキストは見つかりにくくなる。評価によって最適なテキスト長やしきい値を見つける必要がある。また文字種により処理を変えることも考えられる。

テキストが類似していても誤りとは限らない。特に,2文字の漢字語が入れ替わっている場合については検出精度がよくない。テキストそのものの他に文章の読みも利用すれば,同音語の誤りをより正確に判定できる。読みを得るためには,文章解析を利用する,IMEから受け取るなどの方法がある。

正解テキストの選択 本方式が有効に働くためには正解となるテキストがすでに作成されている必要がある。同じようなテキストが,実際にはどの程度存在するかの詳細な調査が必要である。

実行速度 現在,テキストファイル1Mバイトから類似文字列を検索するのに最大2–3秒かかっており,その間IMEの反応が鈍くなってしまう。速度の向上あるいは,Windows95に移行してマルチタスクを利用するなどの方策が必要である。

6 おわりに

ユーザあるいは分野を限定した誤り検出の方法として,類似文字列を検索する方法を提案した。本方式によって,システムが大量の知識を用意しなくても誤字・誤用の検出が行える。今後,従来方式と比較しながら定量的な性能評価およびユーザによる使用感の評価を行っていく予定である。

参考文献

[1] “特集『誤った日本語に気付き始めたワープロ』”, 日経バイト1月号,pp.148-154,(1995)

[2] 奥村, 脇田, 金子: “日本語校正支援システムにおける校正知識”, 情処48回全国大会,5Q–6(1994)

[3] 山田, 福島, 竹元: “ペン入力校正支援システム”, 言語処理学会第1回年次大会 (1995)

[4] 倉田, 菅沼, 牛島: “日本語文章推敲支援ツール『推敲』のパソコン上での実用化”, コンピュータソフトウェア,Vol.6,No.4 (1989)

[5] 竹元, 山田: “FEP型校正支援システムの試作”, 情処52回全国大会,2J–2 (1996)

[6] 角田: “ファイル間の相違検査法”, 情報処理 Vol.24,No.4 (1983)