

FEP型校正支援システムの試作

2 J-2

竹元義美 山田洋志

(NEC 情報メディア研究所)

1 はじめに

ワープロが普及し、パーソナルな環境で文書を作成するユーザが増えるとともに、文章を推敲/校正するシステムが重要視されてきている[4]。校正支援システムは、独立のソフトやワープロソフトへのアドオン型ソフトとして提供されている。しかし、その機能や使い勝手が一般ユーザのニーズへ結びついておらず、本格的な普及には至っていない。そこで、一般ユーザ向け校正支援システムの新しいユーザインターフェースを考える必要がある。

Windowsの仮名漢字変換IMEの結果を自動的に検査し、問題があればユーザにリアルタイムで警告するFEP型の校正支援システムを試作した。問題がなければ通常の入力と変わらず、ユーザは検査が行われていることを意識する必要がない。検査機能は、形態素解析を利用した誤り検出、既存文書の類似文字列検索を利用した誤り検出、を実現した。

2 設計方針

本節では、試作したFEP型校正支援システムの設計方針を述べる。設計にあたり、(1)文章検査の自動実行、(2)システムの汎用性、に着眼した。

2.1 文章検査の自動実行

従来の校正支援システムは、書き終えた文書全体や範囲指定した部分に対して、ユーザの指示により文章検査を実行するユーザ指示型が主流であった。ユーザは文書作成を一段落したときに検査を実行して、書いた文書を丹念に校正できる。しかし、一般ユーザは常にこのような校正スタイルを望んでいない。文章を書き終えて厳密に校正するより、文章を書いている途中で誤りがあればユーザが指示しなくともリアルタイムに教えてくれる方が便利な場合も多い。

そこで、仮名漢字変換の結果を自動的に検査する自動実行型を考えた。文章中の誤りは、仮名漢字変換ミスやタイプミスによるものが多いことから、入力と同時に検査する機能の有用性は高い。一方で、自動実行型では、文章作成中に誤りを警告するので、ユーザの思考のじゃまになる点も考慮する必要がある。

2.2 システムの汎用性

従来の校正支援システムは、独立型かアドオン型が主流であった。独立型は、文書修正用にエディタは使えないが、通常は別のエディタで作成した文書を読み込んで校正する[1][2]。アドオン型は、特定のエディタに

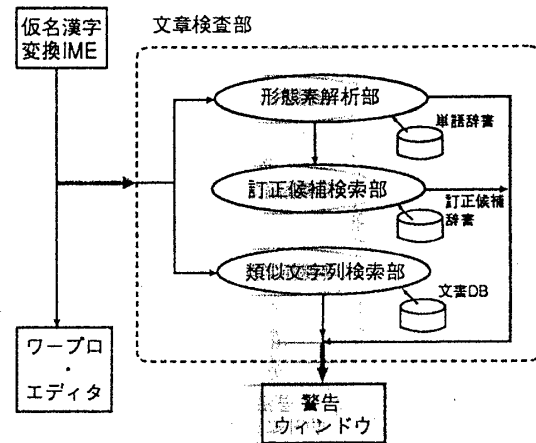


図 1: システム構成

後から組み込んで用いる[4]。また、特定の仮名漢字変換に組み込みのFEP型もある[5]。

しかし、エディタや仮名漢字変換は様々なものが存在する。そこで、任意のAP/仮名漢字変換と組み合わせ可能なようにシステムの汎用性を考えた。

3 システム構成と文章検査機能

前節の設計方針から、Windowsの仮名漢字変換IMEの結果を自動的に検査し、ユーザにリアルタイムで警告するFEP型の校正支援システムを試作した。本節では、システム構成と検査機能について述べる。

3.1 システム構成

通常のWindowsの仮名漢字変換では、仮名漢字変換IMEの確定文字列をワープロやエディタなどのAPに送る。本システムでは、この文字列をフックして文章検査部へ渡す(図1参照)。形態素解析部は、文字列を単語辞書を用いて形態素解析し、誤りを判定する。訂正候補検索部は、誤り語の訂正候補を訂正候補辞書から検索する。類似文字列検索部は、既存文書などのDBから類似文字列を検索して表記ゆれなどを検出する。

文章検査の結果、誤りがあれば警告ウインドウで警告する。ユーザの思考をじゃましないように作成中の文章に直接マークなどはしない。警告ウインドウは、警告のたびに開いたりせず常駐ウインドウとした。ユーザは好きなきときに警告ウインドウを覗けばよい。

本システムは、IMEの結果を検査するので任意のAP/仮名漢字変換との組み合わせが可能である。図2では、校正支援FEPアイコンが画面上にあり、本システムは起動している。エディタはWindowsのメモ帳、仮名漢字変換はVJEを使っている。画面端の警告ウインドウには誤り検出結果が表示されている。

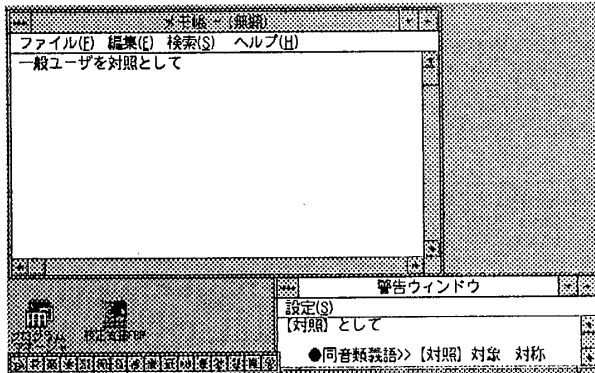


図 2: 本システムの画面例

3.2 文章検査機能

文章検査機能は、形態素解析を利用した誤り検出、類似文字列検索を利用した誤り検出、を実現した。

3.2.1 形態素解析を利用した誤り検出

フックした確定文字列に形態素解析を実行し、次の検査を行う。

誤字・脱字 単語辞書に未登録のために形態素解析に失敗した箇所を誤字・脱字として検出する。また、形態素解析に成功してもルールにより信頼性が低い箇所は誤りとして検出する。例えば、「これあでの」(正解「これまでの」)と誤入力したら下線部を誤字として検出する。

誤りやすい同音語 あらかじめ単語辞書内の誤りやすい同音語に校正用の属性を付与しておく。形態素解析の結果、この属性を持つ単語を検出し、訂正候補検索を実行する。例えば、「特徴」と「特長」は、使い分けを誤りやすい同音語であり、「特徴」を検出すると「特長」を提示する。

3.2.2 類似文字列検索を利用した誤り検出

フックした文字列から、ある長さ(4文字)以上の漢字列および片仮名列を切り出し、これと類似する文字列を以前に作成した文書などのテキストファイルから全文検索する。類似の判定は、異なり文字数に基づく。例えば、文字列「ABCD」に対して、「ABXD」「ACD」「ABCYD」はいずれも異なり文字数は1になる。現在は、異なり文字数1,2の文字列を類似文字列とする。文字列照合のアルゴリズムは、WagnerとFischerの2次法[6]を用いた。

この誤り検出は、表記のゆれや仮名漢字変換ミスの検出に有効である[3]。例えば、入力「インターフェイス」に対して、文字列「インターフェース」が文書DBにあれば、以前使用した表記とのゆれが検出できる。また、誤り入力「人口知能」は、文字列「人工知能」が文書DBにあれば検出できる。

4 使用感と今後の課題

本節では、本システムの主観的な使用感に基づき、今後の課題について述べる。

警告表示のタイミング

本システムでは、仮名漢字変換の確定を文章検査の自動実行のトリガとした。確定文字列に誤りを検出するたびに警告を表示する。ユーザは都合のよいときに警告ウィンドウを覗けばよいといっても警告が出るたびに視点が移ってしまうため煩わしい。

文章作成中のユーザのキー入力イベントを監視して、適切なタイミングで警告表示することを考える必要がある。例えば、ユーザは文章作成を一段落したとき、文書を一旦セーブする。このセーブイベントをトリガとして警告表示することも考えられる。

文章と警告との対応

本システムでは、エディタ内の文章と警告とのリンクが実現されていない。これは、警告を後で確認するのに必要となる。つまり、警告からそれに対応する文章中の箇所へ、また、その逆へと、例えばクリック操作一つで移動できる必要がある。

文章検査機能の強化

本システムは誤りをリアルタイムに警告するため、従来のユーザ指示型システムより警告内容の的確さが要求される。同音語検査は、前後単語との共起情報を用いるなどで精度向上を図る。また、使い分けの難しい同音語には用例提示機能が必要である。誤字・脱字検査では、ユーザがよく使う未知語の学習機能が必要である。

処理速度の高速化

形態素解析および類似文字列検索は、ユーザが文章検査を意識しなくてもいいように高速に実行されなければならない。形態素解析は、パソコンPC-9821An(CPU:Pentium 90MHz)上で50万語辞書をメモリにロードして400文字/秒以上[2]、ロードしなくても40文字/秒以上の速度で実行できる。仮名漢字変換の確定文字列の長さは、長くて10文字程度なので十分実用的なスピードで処理できる。一方で、類似文字列検索は、1Mのテキストファイルを参照するのに3秒弱の時間を要するのでマルチタスク化などを検討する。

5 おわりに

Windowsの仮名漢字変換IMEの結果を自動的に検査し、問題があればユーザにリアルタイムで警告するFEP型の校正支援システムを試作した。今後は、4節の検討をもとに、本システムを改良してゆく。

参考文献

- [1] 牛島 他, “日本語文章推敲支援ツール『推敲』の試作とその作成環境”, 情処研報 84-SW-35-2, 1984
- [2] 竹元 他, “ペン校正支援システム”, 情処50全大 4N-5, 1995
- [3] 山田 他, “文書作成履歴を利用した校正支援機能”, 情処52全大 2J-3, 1996
- [4] 特集, “誤った日本語に気が付き始めたワープロ”, 日経バイト, 1月号, 1995
- [5] 特集, “「一太郎7」で始まる最新オフィス・ソフト”, 日経パソコン, 新年合併号, 1996
- [6] 角田, “ファイル間の相違検査法”, 情報処理 Vol.24, No.4, 1983