日本語文書校正支援ツールの開発

5C−6

ーマニュアル作成支援についてー

納富　一宏　白石　誠　内山　明彦
早稲田大学理工学部

1．はじめに

近年，テクニカル・コミュニケーション(TC)への関心が高まり，マニュアルの作成や開発手法の提案をはじめ，広くは，文書処理を応用した各種支援システムに関する研究[1]が盛んである．

実際，従来の研究では，文章の統計情報（1文当たりの平均文節数や漢字混入率など）を利用した推敲支援システムや，形態素情報を利用した校正支援システムが開発されている[2][3][4]．

これらのシステムでは，一般に，高速性と信頼性とが同時に満たされなければならない．

このためには，入力テキストの校正を目的とした解析の前段で，さまざまな文法情報を取得しなければならない．また，構文規則に照らした文法的な接続関係だけでなく，誤った同音異字語（かな漢字変換における誤変換）の検出なども，校正のための機能として求められる．更に，テクニカルライティングにおける要求として，文の読みやすさやわかりやすさなど，推敲レベルの支援も必要となる．特に，ワードプロセッサの一般化・大衆化により，エンドユーザにとっても，英文スペルチェッカと同程度，あるいはそれ以上の能力をもった日本語文書の校正・推敲システムへの期待は大きい．

しかしながら，汎用的な日本語文書の校正支援では，テキストの多様性への対応のため，パーソナルコンピュータ以下の小規模システムでは先の要求に充分に答えることは難しい．

そこで我々は，マニュアル程度の記述表現に対応でき，かつ小規模システムでも動作可能な日本語文書校正・推敲支援システムについて研究を進めている[1]．

本稿では，マニュアル作成時の日本語文書校正を支援するソフトウェア・ツールについて述べる．本ツールは，①形態素レベルの情報を字面から抽出し，②自立語(キーワード)を検出すると共に，③付属語の文法接続の検定，および④主要動詞の格フレーム検定を行なう．また解析の高速性を重視して，辞書を極力使用しないことを方針とし，格文法解析レベルの情報を収集する．校正・推敲支援では，ユーザの指定したチェック項目，およびその閾値に従って，被解析文書の統計情報から文書校正の示唆（メッセージ）をユーザに与えることができる．

以下，本ツールが採用している解析手法を示し，ツールの構成と動作例について紹介する．

2．字面情報の利用

ここでは，誤字・脱字の検出，あるいは文字列の文法的な接続検定を行なうことを主目的とする．誤字・脱字の種類としては，①付属語列の誤り，②自立語列の誤りの2つのレベルを想定する．

先に述べたように，字面情報(文字種別)を最大限に利用し，形態素解析における辞書検索を極力排除するために，辞書探索による自立語チェックを解析の後段に置く．更に，構文解析に代えて，文節の連続性を調べるために，述語に対する格の共起情報(格フレーム)のみを利用する．また，マニュアル作成を前提とした校正支援であるため，自立語(名詞類)の限定を同時に行なう．

以下，文節区切り，形式文節の接続判定と合成，係り受けについて述べる．

2.1 文節区切り

解析の単位を文節とし，以下のような3つのパートからなる形式文節（図1）を用いる．この形式文節を「3つ組形式文節」と呼ぶことにする．形式文節の成立可否は，各部の文字列の省略状態により分類された組み合わせ表（表1）に従う．

非ひらがな列	ひらがな列	句読点・文末記号
自立部	付属部	句読点部

図1．3つ組形式文節の構造

表1．3つ組形式文節判定と接続可否

自立部	付属部	句読点部	判定	接続	例
○	○	○	◎	×	ファイルに，
○	○		◎	①	記録した
○		○	◎	×	予約，
	○	○	❶	×	これが，
○			❷	②	リモコン
	○		❸	③	この
		○	×	×	．

表中の空欄は文字列の省略（NULL文字列）を意味する．判定とは切出し文字列が形式文節となるか否かを表す（◎：無条件，❶〜❸：条件付き）．接続とは後続の形式文節との接続可否を表す（①〜③：条件付き）．判定および接続の条件は述べる．

3つ組形式文節は，文字種別（ひらがな，カタカナ，漢字，英数字，記号など）により入力文字列を部分文字列に分解し，3つ組となるようにこれを再構成することで得られる．形式文節の区切りは，非ひらがな列・非句読点文字集合の直前で切るようにする．例えば，例文「フォーマットした文書ファイルをドライブAに挿入してください．」の形式文節は，以下のようになる．

表2．形式文節の例

文節No	自立部	文字種別	付属部	句読点部
1	フォーマット	カタカナ	した	
2	文書	漢　字		
3	ファイル	カタカナ	を	
4	ドライブ	カタカナ		
5	A	ｱﾙﾌｧﾍﾞｯﾄ	に	
6	挿入	漢　字	してください	．

3つ組形式文節の導出過程で，後段の辞書探索へのエントリを確定することができる．文節判定条件❶では，接続詞，副詞，あるいは「ひらがな表記の指示代名詞＋格助詞」である可能性が高いため，これ以外は探索をキャンセルしてエラーとして警告すればよい．また，❷では，次節の複合名詞化処理を施すことで探索を保留する．そして，❸ではひらがな表記の連体詞である可能性が高く，この場合は後述の係り受け処理を施すことで探索をキャンセルできる．

2.2 形式文節の接続判定と合成

3つ組形式文節では，表2の文節№2および5のように，自立部のみで孤立した文節が現れる場合がある．このような場合は，後続の文節と合成して形式文節数を少なく抑える必要がある．この処理を「複合名詞化」と呼ぶことにする．

句点“。”，“．”（前後がアルファベット文字列の場合は除く）は文の区切り記号，読点“、”，“，”は文節の区切り記号であるとみなすことができる．従って，形式文節間の連結は句読点部の省略如何により判定することができる．

表1の文節合成のための接続可否条件①

Development of Japanese Document Proofreading Tool for Manual Generation
Kazuhiro NOTOMI, Makoto SHIRAISHI, Akihiko UCHIYAMA
School of Science and Engineering, Waseda Univresity

②③について説明する.①は，係り受け（連体修飾語句）を構成する場合の条件であり，付属部の終端が動詞・助動詞・補助動詞・形容詞・形容動詞・連体詞の活用語尾あるいはそれに準ずるひらがな表記文字条件列の場合，接続可能，それ以外は接続不可能となる．ただし，接続可能な場合でも，被解析文が単文形式でないときは後述する格フレーム解析を行なうまで，合成を保留する必要がある．

②は，主として複合名詞を構成する場合の条件であり，いくつかの例外を除いてほぼ無条件に後続の形式文節に接続可能である．

③は，主として連体詞による修飾句を構成する場合の条件であり，ひらがな表記される連体修飾語句の場合，接続可能，それ以外は接続不可能となる．

以上の条件を満たす場合に限り，形式文節の合成を行なう．

2.3 係り受け

係り受けの解析では，主として文節の接続条件　によるものを対象として，連体修飾文節の判定を行なう．判定結果により文節の合成を行なうか，格解析まで処理を保留するかが決まる．

文節間の係り受けには，連用修飾と連体修飾の２つの場合があるが，これを判定するために文節の機能別分類を表３のようにする．連体修飾文節が用言である場合，文は複文である可能性が高いため，連用修飾文節，すなわち格の判定が同時になされなければならない．

単文であることが格解析により保証される場合は，係り受けの非交差条件に則って，隣接する形式文節の修飾・被修飾関係のみを調べる．

表３．形式文節の機能別分類

機能No	形式文節種別		対象品詞	例
1	独立		接続詞	しかし
2	連体修飾	格	用言連体形／格助詞「の」	入力した／ファイルの
		その他	連体詞	この
3	連用修飾	格	用言連用形／その他の格助詞	削除し／データを
		その他	副詞	必ず
4	述語		用言終止形	変更する

3．接続と共起関係の検定

3.1 付属語の文法接続の検定

付属語列の接続可否を判定するために，以下のような付属語列の接続行列を用いる．これらの接続行列は，付属語解析テーブルとしてあらかじめ用意しておく．

表４．接続行列の例

後置↓ ＼ 前置→	自立部	が	し	られ	なかっ	た
が	×		×	×	×	○
し	○	×		×	×	×
られ	×	×	×		×	×
なかっ	×	×	○	○		×
た	×	×	○	○	○	
終了	○	○	○	○	×	○

表中の○は許容可，×は不可を表す．終了とは，３つ組形式文節の付属部の終了を意味する．「が」は接続助詞，「し」は名詞用法．

3.2 主要動詞の格フレーム検定

マニュアル作成に代表されるテクニカルライティングでは，特に動詞（述語）の限定と格の完備性が要求される．このためには，主要動詞集合をあらかじめ定義し，この集合に含まれない動詞は，エラーとして警告する．また，文の意味の曖昧性を排除するためには，ある述語に対する格（連用修飾文節）の省略を抑えなければならない．そこで先の主要動詞について格フレームテーブルを用意し，必須格の省略が起こっている場合にエラーとする．

4．校正支援ツールの構成と機能

以上の手法を用いて，日本語文書校正支援ツールをパーソナルコンピュータ上に作成した．以下，ツールの構成について述べる．

本ツールの構成を図２に，処理の概要を図３に，そして動作画面例を図４に示す．

校正支援のための処理は図３のように４つのステージに分かれており，それぞれ①第１形式文節抽出，②第２形式文節抽出，③格文節接続判定，④格文節抽出と呼ぶ．各ステージは更にいくつかのフェーズに分割される．また，第５ステージとしてオプショナルな自立語辞書探索がこれに続く．

主な校正支援機能は，①付属語列チェック，②主要動詞チェック，③主要動詞格フレームチェック，④主要名詞チェック，⑤自立語辞書探索である．また，文書の統計情報として，①数(文,文節,文字)，②平均長(文,文節)，③文字混入率(漢字,ひらがな,カタカナ,英数字など)を提示することができる．統計情報は，ユーザの定めた許容範囲(例えば１文中の最大文節数を７とするなど)を超えた場合に警告を発するのに用いられる．更に，文節単位の解析結果をフラグとして詳細表示させることができる．

ユーザは本ツールのメッセージに従って，入力文書を１文単位で編集することができる．

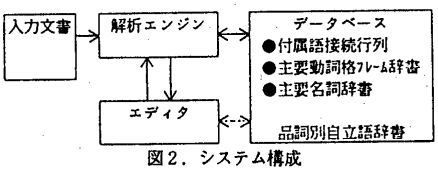

図２．システム構成

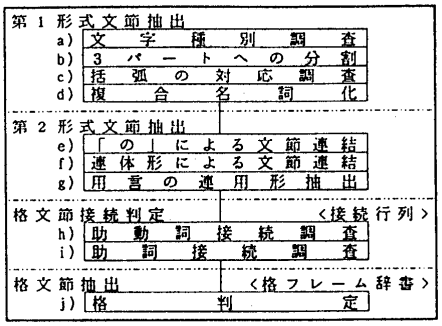

図３．校正支援のための格解析処理

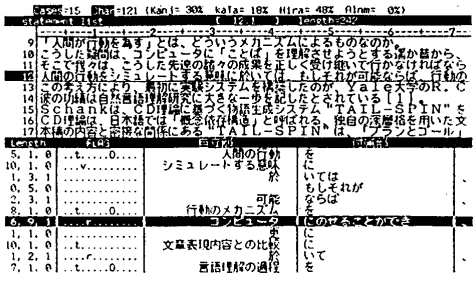

図４．動作画面例

5．おわりに

日本語構成支援ツールについて述べた．今後の課題としては，解析精度を上げることと，同音異字への対応として意味に関する校正機能について検討する必要がある．また，現在，推敲支援機能に関してシステムの拡張を行なっている．

【参考文献】
[1]納富，内山:自然言語処理を応用したマニュアル作成支援システムーマニュアル推敲支援に関してー,情処自然言語処理研究会85-12，(1991.09).
[2]高橋，吉田:計算機マニュアル推敲・査読支援システムMAPLEの開発と運用，情処学論，Vol.31,No.7,pp.1051-1062(1990).
[3]菅沼，牛島:日本語文章推敲支援ツール「推敲」におけるとりたて詞「は」の抽出法とその評価，情処学論，Vol.32，No.11，pp.1392-1400(1991).
[4]納富，内山:知的ワードプロセッサにおける文脈情報の利用，第43回情処全大，(1991.10).