日本語文書校正支援ツールの開発

−解析手法の検討と評価−

3L−7

納富　一宏　白石　誠　増田　進二　加藤　達矢　内山　明彦

早稲田大学理工学部

1．はじめに

日本語文書を対象とする校正・推敲支援向けソフトウェアを設計する場合、例えばパーソナルコンピュータ等の小規模な動作環境でも解析パフォーマンスを低下させないことが重要である。

これらの支援ソフトの多くは、形態素解析を中心とするパージング（parsing）により文法チェックや表記上のチェックを行なうことを目的としている。従って、ソフトウェアの性能を左右する設計の中心はパーザ部に帰着されることになるが、解析能力と実行速度の関係から動作環境を考慮した適切な解析手法を選択しなければならない。

我々は標題のソフトウェアツール「ＨＳＰ」をパーソナルコンピュータ上に試作し（Ｃ言語）、解析能力とパフォーマンスを中心に評価・検討を行なってきた。本稿では、前回の本学会全国大会で報告した「マニュアル作成向け日本語文書校正支援ツール」におけるパージング手法の検討とその評価について述べる。評価対象は、①字面情報のみ，②付属語接続行列のみ，③自立語辞書のみ，④これらの組合せ、の４つのパージング手法であり、解析能力と実行速度から各々のパフォーマンス値を算出して比較を行なった。ここでパーザは、①格解析，②推敲対象の抽出，③統語的接続の検定，の３点を行なうものとした。

以下、パージング手法の詳細と評価基準となるパフォーマンス値の算出について述べる。また、実際の文書を用いて行なった処理結果とその評価を示して、最後に結論を述べる。

2．パージング手法

2-1 データ構造

ＨＳＰでは、文章を１センテンス毎に切り分け、これを更に形式文節に分けることで、文節単位の解析を実現している。この構造を「３つ組形式文節（ＪＦＫ構造）」と呼んでいる（図１参照）。ＪＦＫ構造は、文字種別により自立部（Ｊ部）、付属部（Ｆ部）、句読点部（Ｋ部）に分割される。内部的には、①文字列格納バッファ、②スキャン用文字種別格納バッファの合計２本のバッファを用意して、これらのバッファに対するオフセット値と各パート長とを保持することで一つの形式文節を表現する。

これらの分割は統語接続検定の際の辞書引きの目安となる。ここで、「目安」というのは、形式文節の取得を文字種別により行なっている関係上、自立語がひらがな表記される場合など、必ずしもＪＦＫ構造が文節になるとは限らないため、辞書引きが１回では済まないという意味である。詳しいことは後ほど触れる。

非ひらがな列	ひらがな列	句読点・文末記号
自立部（Ｊ）	付属部（Ｆ）	句読点部（Ｋ）

図１．３つ組形式文節（ＪＦＫ構造）

2-2 自立語辞書・付属語接続行列

ＨＳＰでは、高速なチェックを実現するために、ＪＦＫ構造を用いることで最初の形態素抽出には辞書を使用せず、文字種のみでこれに対処している。

次に、文に現れる語の辞書引きが必要になるが、ここでも高速化を図っている。即ち、ＪＦＫ構造のＪ部、Ｆ部がそれぞれ辞書引きの対象となるはずであるが、Ｊ部は非ひらがな列、Ｆ部はひらがな列という条件を満たす。ここで、Ｆ部は、①助詞，②用言活用語尾，③助動詞，および④これらの組み合わせという、４つの場合であることが期待される。そこで、付属語列の解析には、①〜③の品詞を持つ平仮名列（パターン）について接続を定義し、これを隣接行列として表現したものをオンメモリで持つことにより、Ｆ部の解析を実現した。Ｊ部については、およそ５万語弱の自立語辞書を用意し、この辞書検索により解析を進めていく。自立語辞書および付属語接続行列のプロフィールを表１に示す。

表１．自立語辞書と付属語接続行列のプロフィール

自立語辞書		付属語接続行列	
登録語数	58012 語	行列サイズ	353×353 ビット
識別品詞数	30 個	識別パターン数	355 パターン
ページ数	464 ページ	識別パターンサイズ	2−8(可変) バイト
ページサイズ	2048 バイト	接続の種類	2 種類
初期ページ占有率	80 %		
インデックスサイズ	6 バイト		
データ圧縮	なし		

2-3 パージングアルゴリズム

校正支援では、統語的な接続の検定を行なうことが目的の一つである。既存のパージングアルゴリズムを利用した解析手法は、パフォーマンスの低下を来すという理由から適当ではないと考えている。特に、最初に述べたように小規模システムへのインプリメントを考えた場合、問題点は少なくない。また、本ツールのようなソフトウェアの利用形態は、英文ワープロにおけるスペルチェッカ，グラマチェッカ，スタイルチェッカなどと同様、パーソナルユースにつきるであろうという観点からすると、やはり深い解析実現を期待することは困難である。パージングの目的が、文法レベルでの語構成のチェックであるならば、日本語の有する膠着性，格の可換性(屈折)，に適する解法はいくつかの制約を設ける必要がでてくる。

ＨＳＰでは、パージングに２パス（pass）を要する。第１パスではスキャンニング（scanning）に続く字面から得られる情報により表２に示す文法情報を取得する。第２パスでは、自立語および付属語の接続検定を行なう。そのためには、第１パスで切り出された形式文節が正しい場合は、先に述べた自立語辞書および付属語接続行列を検索することで済むが、形式文節が正しく切り出されなかった場合は、他の方法を用いなければならない。そこで、辞書検索のための文字列候補の切出しにいくつかのバリエーションをつくり、それぞれに対して評価値を設定する。つまり、どの切出し方法を選択したかによって、検定の確信度、即ち検定結果の信頼度が決まるようにする。これにより、単純な真偽のみの判定ではなく、幅を持った解答が得られることになる。切出し方法の概略を図２にまとめる。

表２．取得できる文法情報（文節単位）

group	bit	flag	説明	
文構造係り受け	0	c	文節中に複合名詞化された語が存在する	
	1	i	引用などの括弧書き語句が存在する	
	2	t	格助詞「の」、形容詞・形容動詞による連体修飾が存在する	
	3	v	動詞・助動詞による連体修飾が存在する	
	4	r	連用中止文節である	
	5	g	接続助詞「が」が存在する	
格判定	6	T	副助詞（提題の助詞）「は」が存在する	＝主題格
	7	S	格助詞「が」が存在する	＝上格
	8	O	格助詞「を」が存在する	＝対象格
	9	V	用言終止形である	＝述語
助動詞文体	10	U	助動詞「れる・られる」の活用形が存在する	＝受身,etc
	11	H	助動詞「せる・させる」の活用形が存在する	＝使役
	12	K	接頭語「お・ご」が存在する	＝敬語
	13	Y	助動詞「ようだ」の活用形が存在する	＝比況
	14	N	助動詞「ない・ぬ」の活用形が存在する	＝否定
	15	M	助動詞「ます」の活用形が存在する	＝丁寧
	16	D	助動詞「だ・です」の活用形が存在する	＝断定
箇条書き	17	B	最後の文節が句点で終了していない	＝箇条書き

Development of Japanese Document Proofreading Tool −Examination and Evaluation of Parsing Methods−

Kazuhiro NOTOMI, Makoto SHIRAISHI, Shinji MASUDA, Tatsuya KATOH, Akihiko UCHIYAMA

School of Science and Engineering, WASEDA University

このアプローチの狙いは、図２の１２種類の解析方法を自由に組み合わせて、ユーザの要求するチェックのみを行なえるようにすることである。解析全体の処理時間がパフォーマンスに影響してくる関係上、本手法は有効である。

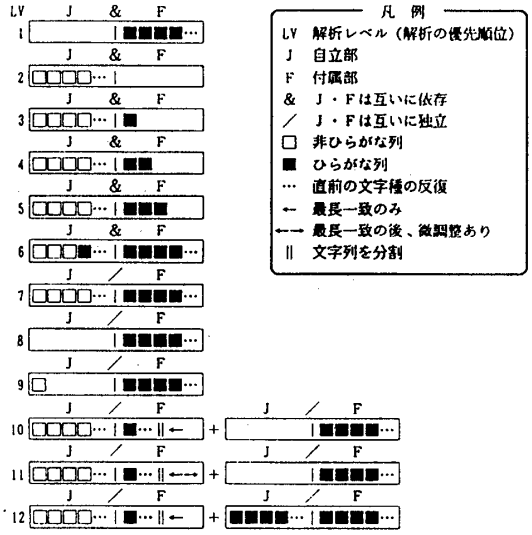

図２．パージングアルゴリズムの概要

2-4 パフォーマンス値の算出

図２に示したように、３つ組形式文節（ＪＦＫ構造）における自立部および付属部の文字列の取扱いにレベル（優先順位）を設け、これをひとつの評価値としてパフォーマンス値の算出に利用している。

校正のための統語的接続検定では、入力が正しいか誤りであるかを決定する。ＨＳＰの場合、ＪＦＫ構造を採用しているので、文節よりも更に細かい単位で入力の正誤を決定できる。そこで、ある文節が正しいと判定する条件をおよそ表３のように設定している。

解析のパフォーマンス値は以下の式(4)のように定義した。

表３．条件設定の違い

条件レベル		自立部	相互接続	付属部
厳	1	○	○	○
↑	2	○	×	○
	3	○	×	×
↓	4	×	×	○
緩	5	×	×	×

○：チェックあり
×：チェックなし

N	１文中の文節数
L	平均解析レベルポイント
S	平均探索ポイント
B	確信度
T_G	J-F接続正当数
T_J	自立部正当数
T_F	付属部正当数
ϕ_i	解析レベル評価値（定数）
α, β, γ	探索評価値（定数）
t	全解析時間

$$L = \frac{\sum \phi_i}{N} \quad (1)$$

$$S = \frac{\alpha \sum (T_G \mid \beta T_J + \gamma T_F)}{N} \quad (2)$$

$$B = \frac{S}{L} \quad (3)$$

$$P = \frac{B}{t} \quad (4)$$

３．動作例と評価

ＨＳＰを用いて、実際の文書を処理した結果から算出した解析手法の違いによるパフォーマンス値の違いを図３に示す。また、動作画面例を図４に示す。

解析には４文書を使用し、それぞれ①コンパイラのマニュアル，②認知心理学の入門書，③コンピュータアーキテクチャの解説書，④文書処理関係の論文である。文書はあらかじめ図表，式などを削除し、制御文字を除く全ての半角文字は全角文字に変換後、ＨＳＰにおいて解析した。これら文書の文字種頻度等を表４にまとめる。

図３を見ると、解析すべき文書により若干の差はあるものの、深い解析を行なえば確信度は上がるが、全体の処理時間は遅くなるため、パフォーマンス値は低くなることがわかる。自立語の解析については辞書に登録されていない未知語の混入率が高いと、やはり処理時間はかかってしまう。自立部および付属部両方を解析する場合、相互接続を調べた方がそうでない場合に比してパフォーマンスは向上する。

表４．解析に使用した文書の文字頻度

文書名 / 文書種類	SMP001.DOC マニュアル	SMP002.DOC 専門書	SMP003.DOC 専門書	SMP004.DOC 論文
制御文字	64	51	388	396
カタカナ	716	197	4238	2569
ひらがな	1304	1410	6242	8187
ｱﾙﾌｧﾍﾞｯﾄ	320	43	1255	979
数字	17	8	243	168
空白	9	2	77	14
句読点	146	140	1086	1239
記号	72	34	429	657
漢字	472	849	6114	6216
全合計	3122	2734	20072	20425

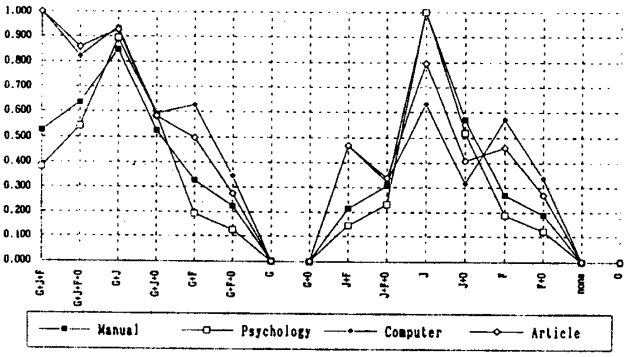

図３．解析手法とパフォーマンス値の関係

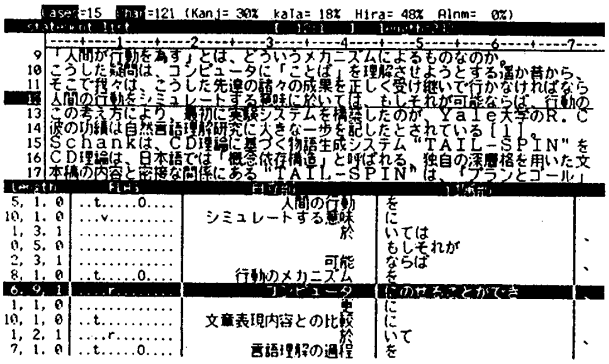

図４．動作画面例

４．おわりに

本稿では、日本語文書校正支援ツールの開発における、解析手法の検討とその評価について述べた。

今後の課題は、自立部を構成する複合名詞のより確かな解析手法の検討と付属部に混入するひらがな表記の自立語への対処である。

【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]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).

[3]納富，内山:マニュアル作成における日本語文書校正支援ツール,1992年電子情報通信学会秋季大会,(1992.09).

[4]納富，白石,内山:日本語文書校正支援ツールの開発ーマニュアル作成支援についてー,第45回情処全大,(1992.10).