ショートノート

日本語文章推敲支援ツールにおける受身形の抽出法†

牛島和夫†† 石田真美††
尹志熙†† 高木利久††

日本語文章推敲支援ツール「推敲」は機械可読な日本語文章を解析して，推敲に役立つ情報を提供することを目的として開発したシステムである．この中の，PASSIVE と呼ぶツールは，受身形の候補を捜し出し文章中に出現位置を明示するものである．本稿では，PASSIVE で採用している，受身形抽出のための五つの判定条件とその検査方式，およびその評価について述べる．

1. まえがき

日本語文章推敲支援ツール「**推敲**」[1),2)] は機械可読な日本語文章を解析して，推敲に役立つ情報を提供することを目的として開発したシステムである．

「推敲」には現在 16 のツールが用意してあり，その主なものには，文頭，文末，文の長さを文の出現順に列挙する **SENTENCE**，「これ」，「それ」，「あれ」などの 11 個の指示詞を際立たせて出力する **KOSOA**，字種を指定して文字列を切り出し KWIC リストを作成する **KWIC**，受身形の候補を捜し出し文章中にその位置を指摘する **PASSIVE** などがある．本稿では，PASSIVE で採用している，受身形抽出のための五つの判定条件とその検査方式について述べる．

2. PASSIVE の開発方針

受身の文は一般に長く，読みにくく，**曖昧になりがちである**．そのため，正確な記述が要求される科学技術文章などでは受身の乱用は慎まなければならない[3)]．

日本語文章中から受身形を厳密に抽出するには高度な自然言語処理の手法が必要であろう．ところが，分かち書きされていない日本語文章を辞書に基づいて正攻法で文法処理や意味解析を行っていたのでは，解析時間がかかりすぎることが予想できる．そこで，PASSIVE の開発には以下のような方針を設けた．

（1） 字面解析だけで受身形を抽出することを目指す．

（2） 文章中に受身と疑わしい箇所があれば，それを指摘することができればよい；推敲をするのは計算機ではなく書き手である．

（3） しかし，受身であるにもかかわらず，それを検出できないという誤りは犯してはならない．

（4） 実用規模の長さの文章（文字数にして1万字程度）を待ち遠しくない時間内で処理してほしい．

3. 五つの判定条件

PASSIVE では以下の五つの判定条件を用いて，文章中から受身形の可能性がある箇所を抽出する．

条件1：「れ」を含まなければ，受身ではない．

根拠1：受身文では，動詞の活用語尾に助動詞「れる，られる」の活用形が接続する．

助動詞「れる，られる」は，受身だけでなく，可能，自発，尊敬を表す場合にも用いられる．これらの区別は字面解析では行えないので，PASSIVE では区別しないこととする．事実，文章中で受身か可能か読み手には判断しがたい場合もあり，書き手に推敲を求める意義がある．また，吉田は規格化日本語において可能の「れる，られる」を「ことができる」に書き換えることを奨めている[4)]．以下では，受身と書いて可能，自発，尊敬も含むものとする．

条件2：「れ」の一文字前が「か，さ，た，な，ま，ら，わ，が，ば」以外の文字の場合は受身ではない．

根拠2：助動詞「れる」は五段活用動詞の未然形のあ段とさ変動詞の未然形「さ」とにだけ接続する．助

† A Simple Method to Extract Passive Voices in the Writing Tools for Japanese Documents by KAZUO USHIJIMA, MAMI ISHIDA, JEEHEE YOON and TOSHIHISA TAKAGI (Department of Computer Science and Communication Engineering, Faculty of Engineering, Kyushu University).

†† 九州大学工学部情報工学科

表 1　判定条件 1-3 を満足する下一段活用動詞の語幹部の一覧
Table 1　A list of the stem part of the *shimo-ichidan* verbs satisfying the conditions 1 to 3.

やせかれ	あるきづかれ	しみったれ
痩せかれ	歩きづかれ	甘ったれ
いいつかれ	よみづかれ	しょぼたれ
言いつかれ	読みづかれ	しみたれ
あるきつかれ	まぬかれ	悪たれ
歩きつかれ	たちくされ	潮たれ
さがしつかれ	立ちくされ	ぶっこわれ
探しつかれ	すてくされ	ひわれ
たちつかれ	ふてくされ	ひびわれ
立ちつかれ	捨てくされ	えみわれ
あみつかれ	不てくされ	笑みわれ
のみつかれ	ふ貞くされ	干われ
よみつかれ	不貞くされ	とらわれ
飲みつかれ	うなされ	捕らわれ
読みつかれ	つまされ	顕われ
編みつかれ	しおたれ	現われ
はしりつかれ	あくたれ	著われ
おどりつかれ	へこたれ	表われ
走りつかれ	すたれ	あばれ
躍りつかれ	あまったれ	

動詞「られる」は「ら」＋「れる」と解釈できる．

条件 3：「れ」の一文字前が「な」であっても「な」の一文字前が「し」または「死」以外の場合は受身ではない．

根拠 3：九州大学大型計算機センターの公用データベースである日本語辞書[5),6)]* によると，な行五段活用動詞は「死ぬ」とその複合語しかない．

条件 4：「れ」の直前の文字列が，**表 1** のいずれかと一致すれば受身ではない．

根拠 4：ら行下一段活用動詞の活用が助動詞「れる」の活用と同じなので，それを受身と区別する必要がある．表 1 は，条件 1-3 を満足するけれども受身ではない，ら行下一段動詞の語幹部の一覧である．ただし，開発方針の(3)に従い，本当の受身を排除する可能性があるものはこの表から取り除いてある．なお，この表も先述の日本語辞書に基づいて作成した．

条件 5：「われわれ」は受身ではない．

根拠 5：これは，明らかに受身ではない．

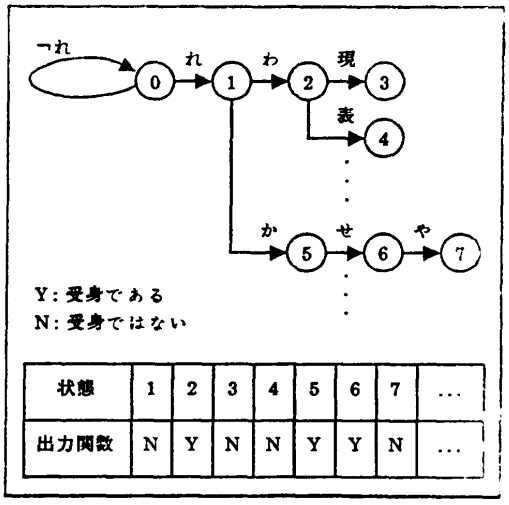
4. 判定条件の検査

五つの条件の判定問題は，日本語テキストに対するパターンマッチングの問題に帰着できる．例えば，条件 1 はパターン「れ」の，条件 2 は複数個のパターン「かれ」，「され」，…の，出現を調べることに対応する．条件 3-5 についても同様である．受身であるか否かは，条件 1-5 に対応するすべてのパターンの出現の位置関係を調べることによって判定できる．

日本語テキストにおいて複数個のパターンを検索するには，Aho-Corasick のアルゴリズムを字種の多い日本語テキスト用に改良したアルゴリズムが有効である[7)]．このアルゴリズムでは，複数個のパターンに対して有限オートマトンを作成し，テキストを先頭から一回走査するだけで，すべてのパターンのすべての出現位置を検出できる．さらに，有限オートマトンの各状態に出力関数が定義できるので，これを利用することにより，受身の判定に必要な出現の位置関係の検査も同時に行えるという特徴を持っている．

そのため，五つの条件の判定に，このアルゴリズムをそのまま適用するだけでも開発方針(4)を満足する効率が得られる．しかし，条件 1-5 に対応するすべてのパターンの最後の文字がすべて「れ」であることに着目すると，まずテキストの先頭から走査して「れ」を検出し，そのあとテキストを逆向きに走査して各条件を検査する方がより効率的である．

図 1 に，「れ」を検出したらテキストを逆向きに走査して，受身形を判定する有限オートマトンとその出力関数との一部を示す．この有限オートマトンにおいて，状態遷移に失敗したときに，その状態の出力関数の値を見れば受身か否か判断できる．例えば，テキストが「…Xわれ…」（X≠現，表，…）の場合は，状態



状態	1	2	3	4	5	6	7	…
出力関数	N	Y	N	N	Y	Y	N	…

図 1　受身を抽出する有限オートマトンと出力関数
Fig. 1　A finite automaton extracting passive voices and its output function.

* 基本語部分（自立語約 8 万 5 千語），付属語辞書（形式名詞，補助用言，自立語の活用語尾も含む），文節内における二単語間の接続検定用テーブルからなる．

2の出力関数の値から受身と判断できる．また，テキストが「…現われ…」の場合は，状態3の出力関数の値から受身ではないことが分かる．なお，「…現われ…」が受身でないのは，条件4による．

5. 評　価

五つの判定条件は，我々の研究室に蓄えられている機械可読な日本語文章142編（総文字数 1,025,911）を調査した結果決定したものである．百万字の中に「れ」を11,344個含み，うち条件2を満たすものは5,810，本当の受身は5,412であった．したがって，条件2だけでも開発方針を満たす受身の抽出法としてかなり使いものになる．一方，受身でなかったもの398のうち「な＋れ」103，ら行下一段活用動詞242（うち「現われ」235）であった．前者を条件3で，後者を条件4で判定していることになる．そのため，この142編に限れば，五つの条件による受身検出の精度（本当の受身/五つの条件を満足するもの）は0.99であった．本当に受身であるか否かの判断は視察によって行った．精度が1にならない原因，つまり五つの条件をすべて満足するが，受身ではないものは，タイプミスを除くと，「か，さ，た，な，ま，ら，わ，が，ば」＋「る」で終わるら行五段活用動詞の仮定形（「分かれば」，「変われば」など），形容動詞「まれだ」であった．これらに対して新たな条件を追加して検査することも可能であろう．しかし，推敲支援ツールの開発方針上，出現頻度を考慮してそこまでは対処していない．

なお，上述の142編とは別の25編に対して受身形抽出の精度を測定したところ，1の精度が得られ，五つの判定条件の有効性が確認できた．

一方，受身形の抽出に要するCPU時間は，有限オートマトンの作成時間を除けば，一万文字当たり約6.5 msecであった．CPU時間の測定には九州大学大型計算機センターのFACOM M380S OS Ⅳ/F4を用いた．

これらの精度やCPU時間の測定結果から，本稿で述べた五つの判定条件とその検査方式は，PASSIVEの開発方針を十分満足していると判断できる．

6. むすび

字面だけの簡単な解析で，受身形を高精度で抽出できることを示した．この高精度は，解析の対象となる文章が通常の漢字かな混じり文であることが前提となっている．なお，前述の142編の調査によれば，実際に条件4によって受身ではないと判断されたもののうちの97%は「現われ」であった．このことは，条件4として「現われ」だけを検査しても十分実用的な精度が得られることを示している．この場合，検査に要するCPU時間を短縮できるだけでなく，判定条件検査のためのプログラムを小さくできるので，パソコン上のソフトウェアにこの抽出方式を組み込む場合など，この簡便な方式がふさわしいかもしれない．

参考文献

1) 牛島和夫，日並順二，尹　志熙：日本語文章推敲支援ツール「推敲」の使用について，九州大学大型計算機センター広報，Vol. 18, No. 1, pp. 9-37 (1985).
2) 牛島和夫，日並順二，尹　志熙，高木利久：日本語文章推敲支援ツールのプロトタイピング，コンピュータソフトウェア，Vol. 3, No. 1, pp. 35-46 (1986).
3) 木下是雄：理科系の作文技術，p. 224，中央公論社，東京 (1983).
4) 吉田　将：日本語の規格化に関する基礎的研究，昭和58年度科学研究費補助金一般研究（B）研究成果報告書 (1984).
5) 吉田　将，日高　達，稲永紘之，田中武美，吉村賢治：公用データベース日本語辞書の使用について，九州大学大型計算機センター広報，Vol. 16, No. 4, pp. 335-361 (1983).
6) 稲永紘之，吉田　将：日本語処理のための機械辞書，情報処理，Vol. 23, No. 2, pp. 140-146 (1982).
7) Yoon, J., Takagi, T. and Ushijima, K.: An Experimental Study of String Matching Algorithms for Japanese Texts, *Proc. of 1986 Int. Conf. on Chinese Computing*, pp. 297-304 (1986).

（昭和61年12月2日受付）
（昭和62年6月11日採録）

牛島　和夫（正会員）

1937年生．1961年東京大学工学部応用物理学科（数理工学）卒業．1963年同大学院修士課程修了．同年九州大学中央計数施設勤務．1977年九州大学工学部情報工学科教授（計算機ソフトウェア講座担当），現在に至る．1986年4月から九州大学情報処理教育センター長を兼務．工学博士．著書「Fortranプログラミングツール」（産業図書）ほか．日本ソフトウェア科学会，電子情報通信学会，ACM各会員．

石田 真美（正会員）

昭和36年生．昭和59年九州大学工学部情報工学科卒業．昭和61年同大学院修士課程修了．同年，（株）富士ゼロックスに入社，システム技術研究所に勤務．

尹 志熙（正会員）

1959年生．1982年韓国漢陽大学工科大学電子工学科卒業．1985年九州大学大学院工学研究科情報工学専攻修士課程修了．現在同大学大学院博士後期課程在学中．字種の多い言語のテキスト処理の研究に従事．

高木 利久（正会員）

昭和29年生．昭和51年東京大学工学部計数工学科卒業．現在九州大学工学部情報工学科助手，工学博士．電子情報通信学会，人工知能学会各会員．