新聞記事データを素材とする漢字KLICとKWICの作成

7B-4

亀田弘之（東京工科大学）・森田敏生（東京大学）・
倉島顕尚（東京大学）・藤崎博也（東京大学）

1．はじめに

自然言語処理技術は、近年、目ざましい進歩を遂げており、例えばワードプロセッサーは既に実用の段階に達している。しかしながら、機械による自動翻訳をはじめとする多種多様の高度な言語処理システムには、未だ解明・解決すべき問題点が数多く残されており、自然言語処理のための基礎理論・技術のより一層の発展が必要である。そのためには、大量の言語用例を収集・分析し、言語自体に関する知見を実証的に得ることが重要である。

筆者らはこのような見地から、昭和57年に計算機可読な新聞記事データを入手・加工し、大型計算機上にて利用可能な形に整備するとともに、用字・用語データベースを作成してきた[1,2]。この作業の一環として今回は、漢字KLIC(漢字 Key Letter In Context、全体で4,539,148行)と、KWIC(Key letter In Context、全体で5,471,686行)とを作成し、多くの研究者にも比較的手軽に利用することのできるようにとの配慮から、これらをマイクロフィッシュ（以下では、MFと略す）化したのでその報告をする。

2．素材とした新聞記事データの概要

漢字KLICとKWICを作成するための素材として以下のような新聞記事データを使用した。

(1)データ：昭和57年の朝日新聞84日分。

(2)サンプリング方法：各月から曜日に重複のないように7日分ずつを無作為抽出した。但し、新聞の第1面のほとんどを占めてしまうような特別に大きなニュースのあった日や、高校野球の特集の日は除いた。

3．漢字KLICの概要

3-1. 作成手順･･･新聞記事データをもとに、以下のような手順で漢字KLICのMFを作成した。

(1)漢字KLICを各日付け毎に作成。

(2)上記(1)で作成した84日分のデータをマージ・ソート。

(3)MF作成用システム(COM)の制御コードを挿入。

(4)業者によりMF化。

3-2. 漢字KLICの説明･･･漢字KLICは、図1のように、漢字1文字（キー漢字）とそれが出現した箇所の前後20文字分のテキストおよびレコードIDから構成されており、キー漢字に関してソートされている。なお、レコードIDは、その漢字が出現した新聞記事の日付、掲載面または掲載覧、テキストの種類（見出し・本文・写真の説明文・その他）から構成されている。

4．KWICの概要

4-1. 作成手順･･･新聞記事データと、筆者らにより既に作成されている形態素解析システム[3]とを用いて、以下のような手順でKWICのMFを作成した。

(1)形態素解析用の単語辞書を作成。

(2)形態素解析用の文法表を作成。

(3)新聞記事データを形態素解析処理し、単位切りデータを作成。

(4)単位切りデータを用いて、各日付け毎にKWICを作成。

(5)上記(4)で作成した84日分のデータをマージ・ソート。

(6)MF作成用システム(COM)の制御コードを挿入。

(7)業者によりMF化。

4-2. 単語単位の設定･･･KWICを作成する際には、単語単位を設定することが必要である。本研究では、単語辞書作成に新明解国語辞典（三省堂）を使用したこと、および、この辞書は国語学者らにより厳密に作成されたものであることを鑑み、新明解国語辞典の辞書見出しを単語単位とすることとした。なお、この新明解国語辞典に記載されていない複合語は、原則として語構成要素を1単位とすることとした。
（例：「東京方面」は「東京」と「方面」とに分割する。）

4-3. 単語辞書の概要･･･形態素解析用単語辞書(185,937項目)は以下の手順で作成した。

(1)計算機可読な国語辞書（新明解国語辞典、三省堂）を入手。

(2)小見出しを主見出し化。

Generation of Kanji-KLIC and KWIC from Newspaper Articles
Hiroyuki KAMEDA(Tokyo Engineering University), Toshio MORITA,
Akihisa KURASHIMA and Hiroya FUJISAKI(University of Tokyo)

(3)用言を活用させ、各活用形を１つの見出しとして登録。
(4)２分木探索用にソート。
(5)予備的に84日分の新聞記事データを形態素解析し、その結果をもとに未登録語と処理結果に悪影響を及ぼす語を抽出。
(6)パソコンを用いて未登録語とその品詞・活用形情報とを入力し、大型計算機上の辞書とマージ。追加した語は、固有名詞を主と　して25,923個であった。
(7)形態素解析処理に悪影響を及ぼす語を　辞書から削除。削除した語は、図２に示したような語66個であった。

4-4. 文法表の概要・・・いわゆる学校文法をもとに、品詞間の接続表を作成・利用した。この表では、助動詞の各活用形と助詞は１項目として取り扱っており、全体で86行×59行のマトリックスとなった。

4-5. 形態素解析の概要・・・形態素解析は、最長一致法に基づく解析方法により行ったが、上記の単語辞書と文法表とを利用して、品詞の接続条件をチェックするとともに、文節内の構造に関する情報にもとづき品詞の予測を行ったため、高速・高精度の処理が実現された。

4-6. 形態素解析結果・・・図３は、形態素解析結果である。品詞の予測ミスのために、同一の語でも出現箇所により異なった処理結果となっているものがあるが、処理精度は参考文献[2]の式によると約99.8%であった。また処理速度は、CPU時間にして新聞記事１日分あたり約70～80秒であった。

4-7. KWICの説明・・・KWICは、図４のように、単語（キー単語）とそれが出現した箇所の前後20文字分のテキストおよびレコードIDから構成されており、キー単語に関してソートされている。なお、レコードIDの構成は、漢字KLICのそれと同様である。

い、いな、うと、えない、がい、きた、
ことに、ござ、して、しな、しまった、
すれば、たこ、たとい、たばかり、ために、
たわけ、だと、つとめて、てくれ、ていた、
となっ、につい、のす、はこ、はな、みて

図２．辞書から削除した語の例

／サンフランシスコ／で／の／交渉／が／決裂／し／た／日米／航空／問題／が／，／二十二／日／の／桜内／外相／と／レーガン／大統領／，／リーガン／財務／長官／の／各／会談／で／取り上げ／られ／，／交渉／打開／に／向け／努力／し／て／行く／と／の／基本／姿勢／で／一致／し／た／.

図３．形態素解析結果の例

5．おわりに

新聞記事を素材とする漢字KLICとKWICとのマイクロフィッシュ作成について報告した。

謝辞　本研究は、昭和62年度文部省科学研究費特定言語「言語情報処理の高度化のための基礎的研究」（代表者長尾真）課題番号63101004の援助のもとに行われたものであり、データを供与された朝日新聞社、計算機可読な単語辞書を提供された三省堂、形態素解析システムに関して助言された筑波大学荻野綱男助教授及び東京大学藤崎・広瀬研究室の諸氏に深く感謝する。

〈〈参考文献〉〉

[1]藤崎・亀田：“新聞記事を対象とする自動単位切り処理とそれに基づく語彙調査”、情処学会、NL研究会、Vol.51-2(1985)
[2]藤崎・亀田：“自動単位切りによる新聞記事の語彙調査”、文部省特定研究『情報化社会における言語の標準化』成果報告書、pp.661-675(1986)
[3]藤崎・亀田・森田・倉島：“高検索機能データベース作成のための形態素解析と統語解析”、情処学会第36回全国大会(1988)

お巡りさんが何度か振り回された。三鷹署の　話
同試験所の岩田一夫主任研究員の　話
清水教授の　話
の学習机もでている。メーカーやデパートの　話
せている。日本のプロ野球界では、よくある　話

では、どこかで飼われているうち、逃げ出し
では、オキゴンドウが海岸に打ち上げられる
では、穀倉地に囲まれた同遺跡には、卒塔婆
では、値段も固定式のものとほとんど変わら
なんですが……。

図１　漢字KLICの一部

結している。レーガン米政権はこんど新たな　攻撃
手はいないが、俊足ぞろいで、機動力のある　攻撃
場規律の乱れの温床になっている」といった　攻撃
う）網によって日本軍の真珠湾などへの先制　攻撃
大を中心にした関東学生選抜は、得意のパス　攻撃

が開始されればソ連がイランを支援するだろ
が期待できる。また、沖縄大会の話題は、甲
が強まり、当局は七月に改訂案を提案した。
が近いことを知らされ、ルーズベルト大統領
が決まらず苦戦。一度は関学主力の関西学生

図４　KWICの一部