自然言語処理 26－3
（1981. 7. 2）

日本語文論文タイトルからのキーワード自動抽出システム（JAKAS）

○荒木啓介，金子明夫，高野文雄，日夏健一
（日本科学技術情報センター）

1．はじめに

文献のインデクシングの補助機能として，日本語のタイトルおよび／または抄録文からのキーワード自動抽出は重要な課題であり，相当な精度のものが実現すれば大幅な省力化になるが，技術的にはむずかしい問題があった。(文献2) つまり
①，連続した漢字かな混り日本語文を適切な単位に切断する。
②．切断された文字列がキーワードとして重要かどうかを判定する。
③．取り出されたキーワードにフリガナを与える。
の三点である。

③は，著者らが既に開発し，JICSTにおいて実用化している漢字ーカナ変換システム（K－KACS）により，99.9％の精度が保証されている。

②は，タイトルについては，英語文などで確立した手法である，不要語（of, with, that など）を除いた，残りをキーワードとする方法により容易であるが，長い抄録文については，構文解析や意味解析を要するため，英語文等についても解決されていない。

①は日本語特有の問題であり，今までいくつかの研究があるが完璧とは行かないようである。(文献2,3)

以上の状況から，当面論文タイトルを対象に，日本語文の効率良い切断方式を検討した。
著者らは既に，カナに変換された日本語文が，適度に分かち書きされている必要性から，K－KACSに分かち書き機能を与えていたが，これだけでは適切なキーワードを取り出すのに十分ではない（精度60％，文献4）

そこで，日本語文中における漢字の役割に注目し，以下のアルゴリズムを考案し，それに基づき9万タイトル以上の論文タイトルを人手で分析しながら，一見バカバカしい作業を積み重ねて用語を収集し，辞書を作成した。並行してプログラムを作成し，未知のデータについてテストしたところ，予想を上回る性能が得られた。

2. 切断アルゴリズム.

2.1. Cut処理. : 日本語文の漢字列のうち，接頭辞，接尾辞，接中辞になることのある文字(文字列)を，辞書により切断する。

(例) 水溶液中触媒反応 → 水溶液 中 触媒反応

2.2. Pass処理 : 切断対象となる文字(文字列)も，意味のある熟語を作るものについては，その調らべた辞書により切断させない。

(例) 中央 P. 中毒 P. 幸中 P. etc

以上の処理は，特に接辞のみに適用するのではなく，接辞を含まない用語の長い列についても適用される。その場合は，用語のタイプとしてその数が少ない方を主として辞書化し，Cut処理し，全体をバラバラにして行く。

(例). カブトムシ 幼虫 ミトコンドリア ATPアーゼ 活性化物質

カブトムシ: 無数にあるタイプの用語
幼虫 ミトコンドリア: 有限なタイプの用語
ATPアーゼ 活性化物質: 無数にあるタイプの用語

2.3. イタリック文字のプログラムによる切断. : 動植物名などは，イタリックシフト記号により囲まれているので，これを単にスペースに置き換えて切断する。半スペースはパスする。

(例) Escherichia coli (coli: 半スペース)

2.4. Quenah処理 : 通常の単純な切断では，切口に不要な文字列が付着することがあるので，これを消去(切断)する。

(例). 濃厚 trans-デカリン —のデリミッターにより切断したら，前後に逆のぼり，文字の変り目でまた切断する。

↓

濃厚 +trans デカリン

2.5. Alter処理 : 以上の切断機能がなかなか適切なあらゆる場合に自由に対処し，具体的に切断結果を指示する。

(例). 上 Cut, 陸上 Pass とすると，大陸上高気圧が切れないので，大陸上 Alter 大陸_上_ と指示しておく。

この機能により，ハイフンやカンマを含むあらゆるケースにつき，用語，用語の標準化を可能な限り行うことが出来る。

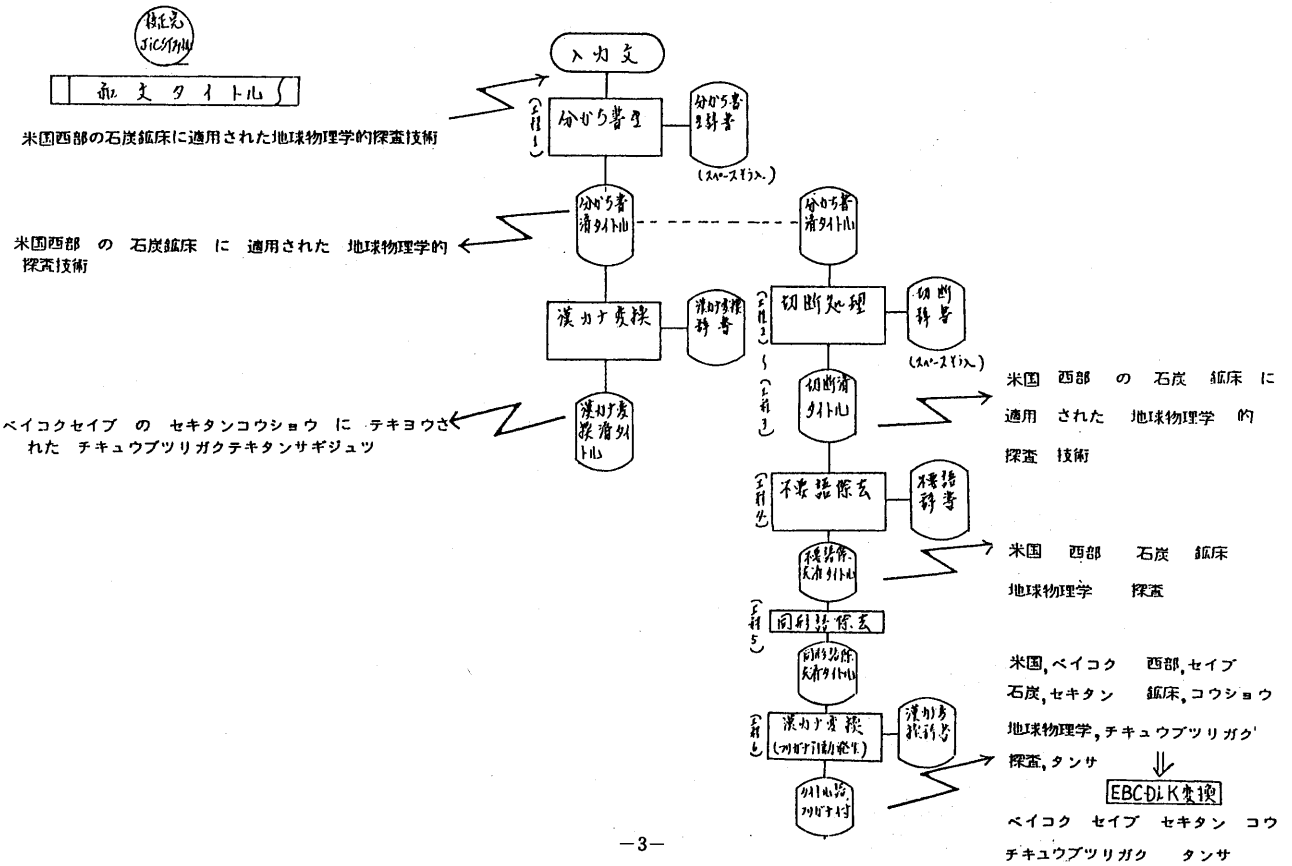
（例）
A・B変換
A／B変換
A－B変換
A—B変換
A↔B変換
} Alter　A B変換

辞書は人間がメンテナンスしやすいように，使用した文字毎に（ただし接辞以外の用語については，用語の種類毎に）集めて辞書マスターとしているが，プログラムの実行形式においては，先頭の文字の下に，長い順に配列し，タイトル文と後方に向けて最長マッチングする。

なお，不要語は英語等の場合に比してかなり多く，助詞，助動詞，接続詞，副詞　等のほか，一般的な，意味の少ない名詞（推薦，検討，etc）等にわたる。

3．システムの概要

3.1. 入出力および処理フロー，：　ディスクベースにより以下のように連続処理する。



3.2. システム規模 : 辞書（Version 1）、切断辞書 4,305語
不要語辞書、 216 語.

プログラム. アセンブラ. 1本. 1200ステップ
（切断処理用）.
マスター 20本 3,500ステップ
コボル 1本 670ステップ
（共に、I/O、ファイル作成
等用）

4. 性能

4.1. 処理速度

使用計算機 : 日立 M-170
対象. : JICST科学技術文献タイトル. 3,500件
速度（処理時間） : 40分16秒
（内訳）分かち書き 切断 抽出 不要語除去、重複語除去
35秒 34分42秒 1分50秒. 2分32秒. 1分16秒

これは、1万件当りに換算すると1.91時間（設計時予測、3.3時間）

4.2、精度

出力の1/10サンプリングにより推算

タイトル数 283
正しく抽出された キーワード. 1288 （平均タイトルキーワード 4.6 個）
抽出されるべき全キーワード. 1321

$$精度 = \frac{1288}{1321} \times 100 = 97.5\%$$ （設計時予測、95%）

正しくない部分は、切りすぎと切り不足である。

5. 出力例.

(1) （原文タイトル）
G010 「地震対策特集」 大規模地震の情報伝達と応急復旧
（切断済タイトル）
M000 「 地震対策 特集 」 大規模 地震 の 情報伝達 と 応急 復旧
（自動抽出キーワード）
M020 地震対策
M050 大規模
M060 地震
M080 情報伝達
M100 応急復旧

(2)　（原文タイトル）
G010　音声信号に対する適応型反響阻止
（切断済タイトル）
M000　音声信号 に 対 する 適応型 反響阻止
（自動抽出キーワード）
M010　音声信号
M050　適応型
M060　反響阻止

(3)　（原文タイトル）
G010　改善された予測法による言語のエントロピー推定とそのヘブライ語への応用
（切断済タイトル）
M000　改善 された 予測法 に よる 言語 の エントロピー 推定 と その ヘブライ語 への 応用
（自動抽出キーワード）
M030　予測法
M060　言語
M080　エントロピー
M120　ヘブライ語
M140　応用

(4)　（原文タイトル）
G010　スピノーダルCu－10wt％Ni－6wt％Snにおける低サイクル疲労
（切断済タイトル）
M000　スピノーダル Cu － 10wt ％ Ni － 6wt ％ Sn に おける 低サイクル疲労
（自動抽出キーワード）
M010　スピノーダル
M020　Cu
M040　10wt
M060　Ni
M080　6wt
M100　Sn
M130　低サイクル疲労

(5)　（原文タイトル）
G010　鉱山総労働者数の変動状況調査
（切断済タイトル）
M000　鉱山 総 労働者数 の 変動 状況調査
（自動抽出キーワード）
M010　鉱山
M030　労働者数
M050　変動

6. JICSTにおける応用

(1) 今年10月に向けて辞書メンテナンスを重ねて精度向上を図り，57年4月からのサービスに使用する他，56年分の蓄積ファイルにも適用し，タイトルキーワードのJOISオンラインにおける一次タグとして使用できるようにする。

(2) 55年以前の旧ファイル，約200万件については，同様にタイトルキーワードが使用できるよう検討する（主として計算機時間の問題）

(3) 切断のスピードアップのため，全切断辞書が大幅に増加しないことを見通した上で，オンコア処理することを検討（1/10の時間に短縮される見込）。

7. 文献

(1) 菊池敏典．自動索引法と自動分類法 展望．情報管理 Vol 8 No 13 pp 3－9，(1965)，Vol 8. No 5. pp 9－14. (1965) Vol 9. No 1 pp 21－28 (1966)，Vol 9. No 4. pp 185－189 (1966)

(2) 長尾真，小谷幹男，池田浩之．日本語文献における重要語の自動抽出．情報処理 Vol 17. No 2. pp 110－117 (1976)

(3) 植村俊亮．日本語KWIC索引の試み．第5回情報科学技術研究集会発表論文集 pp 143－152. (1968)

(4) 荒木啓介．漢字－カナ変換システムのその後の展開．日本語論文タイトルからのキーワード自動抽出システム（JAKAS）の開発．計算言語学 24－4 (190. 8. 21)