A Japanese KWIC Indexing System

SYUNSUKE UEMURA*

0. Preface

This report describes an experiment on Japanese KWIC indexing system. Major basic characteristics of the research are as follows.

- (1) The system is perfectly Japanese language oriented. It can process Kanzi, hiragana, katakana and so on as its data. A kanzi-teletypewriter is used as an off-line IO device for a computer.
- (2) A simple method of automatic segmentation has been developed, programmed and proved to be practical.
- (3) Some topics are investigated concerning index format of Japanese sentences (vertical or horizontal line outputting, and their conversion from each other).

1. Japanese character string

Many kind of characters are used in Japanese sentences, kanzi (Chinese characters), hiragana, katakana, and even alphabets.

There is no custom of segmentation between words.

These special features cause much trouble in processing Japanese by computers. It is necessary to have an appropriate IO device and also to develope method of automatic segmentation (if not by hands).

2. Segmentation for KWIC index

To make Japanese KWIC index, we have to detect the positions of keywords (or phrases). Though this is a kind of segmentation the process here is rather simple. We don't have to segment Japanese sentences into separate words, but only the detection of character position from which the keyword begins (we call it 'keyword position') will be requested for indexing.

3. Automatic segmentation

The method we have developed here for automatic segmentation is based upon the analyses of character classes and transition from one class to another. The rules are roughly described as follows.

Suppose we have a Japanese title,

国産の COBOL コンパイラ研究

We can easily obtain a string of character classes from it, kanzi, kanzi,

This paper first appeared in Japanese in Joho-Shori (the Journal of the Information Processing Society of Japan), Vol. 10, No. 5 (1970), pp. 270–278.

* Software Division, Electrotechnical Laboratory, MITI, Tokyo.

hiragana, alphabet, ..., kanzi.

(a) The head position of kanzi string has high possibility to be a keyword position.

A long kanzi string usually consists of the series of 2 kanzi phrases.

情報|処理|研究

There are also some groups of kanzi that are used as prefixes or suffixes.

超|高速,計算|機

Any long kanzi strings can be divided into meaningful units by the rule above, and they show how we should segment kanzi strings (i.e. keyword positions).

超|高速|電子|計算|機

(b) Also the head of alphabetic string has high possibility to be a keyword position, especially when they consist of more than 1 character.

(c) The head of *katakana* string is sure to be a keyword position in usual Japanese sentences.

(d) The transition point to hiragana from others (kanzi, katakana,) has to be carefully examined, because it has high possibility to be zyosi or zyodosi which should not be a keyword. Japanese titles of 148 scientific reports were investigated and the number of combination of actually used zyosi and zyodosi was proved to be less than 30. Some of them are:

These rules require only small tables for segmentation. Processing speed

These rules require only small tables for segmentation. Processing speed is expected to be much shorter in comparison with longest-matching method, and the experiment proved it.

4. Page formatting

Since this indexing system is for Japanese, not only usual horizontal output form but also vertical line outputting has been designed and programmed. Transformation from outputting in lateral lines into vertical style (or vice versa) requires some character relacements. '('or')' in horizontal style should be 'c' or 'c' in vertical style, and so on. Fig. 1 (shown next) is an example of vertical style KWIC index.

5. Sorting

Sorting of key phrases is based upon kanzi-teletypewriter code of each

```
D
i
f
f
                                                                                                                                                                                                                                                                      系統故障過渡現象の計算プログラムと等価回路定数の
                                                                                                                                                                                           ė
                                                                                                                                                                                                                           発中性子崩壊定数 α 1 mmから中性子の実効増倍
                                                                                                                                                                                                                                                                                              高度並列演算装置による図形の記述(IV)型電子の平均自由
                                                                                                                                                                                         entialによる感情分析 その1 実験の高度並列演算装置による図形の記述(IV)
                                      フィルタ(吸収板)法による γ線源の純度のICの現状と
                                                                                                                                                                                 軸対称矩形断面空心ソレノイドの磁場
                                                                                                                              最近のアナログ計算機とハイブリッド
                                                                   モデル系統における過渡安定度の電気試験所におけるMHD発電の
       マクラウド真空計における毛管
                                                                                                             半導体磁電部品の進歩と新しい
                                 電子計算機による。はなし
                                                                                                                                                                                                                                                                                                                        ケーブルの充放電による
                                                                                                                                                                                                                                                     一様巡回構造を持つ線形
                                                                                                                                                                         安河内モデルを使って
                                                                                   MXを用いたSi単
                                                                                                                                                        系統故障過渡現象の
                                                                                                                                                                                                                                            軸対称矩形断面
                                                                                                                                                                                                                                                                                      M
I
S
C-V
                                                                                                                                               最近のアナログ
                                                          電気絶縁材料
系統故障過渡
                                                                                                                                                                過渡安定度
                                                                                                                                                                                                                                                                                                                                  フィルタ(
                                                                                                                                                                                                                                                                              有理関数
                多領域
域却
                                                                                                                                                                                                                                                              軸対称
                                                                                                                                                                                                                   電力
                              早 結晶と電子

中 結晶切断による残留と粒の直接観察(速報)

中 結晶切断による残留と粒の直接観察(速報)

の 研究の展開

と 研究方針

と 研究の展開
                                                                                                                                     計算機による。はなし言葉、の合成計算機による。はなし言葉、の合成計算でログラムと等価回路定数の近似性計算でログラムと等価回路定数の近似性計算における動揺方程式の新解法計算した第2種超電導体のヒステリシス損失計算した第2種超電導体のヒステリシス損失
                                                                                                                                                                                計算
                                                                                                                                                                                        計画と実施
型行列一致演算の一般性について
型行列一致演算の一般性について
型行列一致演算の一般性について
多色配色のSemal
乳統における電圧および無効電力等の微小変化量簡易測定法
                                                                                                                                                                                                                                                                     近似性
                                                                                                    軽水動力炉一次冷却系破損事故時の出力変化に関する理論的考察計測・制御技術
計測・制御技術
                                                                                                                               計算機について
                                                                                                                                                                                                                           系統における過渡安定度の研究
                                                                                                                                                                                                                                           空心ソレノイドの磁場計算
                                                                                                                                                                                                                                                    空間回路網の基礎理論
矩形断面空心ソレノイドの磁場計算
                                                                                                                                                                                                                                                                              近似を作製するプログラムについて(1)奇関数の場合
                                                                                                                                                                                                                                                                                                                 行程とドリフト速度の測定
急しゅんな高電圧多パルスの発生(マスタ・ギャップ)
吸収板)法による7線源の純度の検査
               原子炉における出力平担化の解析(2)離散モデル原子炉における一次冷却系破損事故の実験と解析(第1報)
                                                                                                                                                                                                                                     系数Keffへの変換
                                                                                                                                                                                                                                                                                              形電力計を用いる新高磁束密度鉄損測定法
                                                                                                                                                                                                                                                                                                        行列一致演算の一般性について
                                                                                                                                                                                                                                                                                                                                          及び半連続式蒸着装置
現象の計算プログラムと等価回路定数の近似性
                                                                                                                                                                                                                                                                                       曲線測定器(資料)
                                                                                                                                                                                          3色配色のSemantic
                                                                                                                                                 ◇電試彙報
                                                                                                                                                                         ◇電試彙報
                                                                                              ◆電試彙報
                                                                                                                       ◇電試彙報
                                                                                                                                                                                                                                                                                                                          ◆電試彙報
                           ◇電試彙報
                                                                                                               ◆電試彙報
                                                                                                                                                                                                                                             ◆電試彙報
                                                                                                                                                                                                                                                      ◆電烒彙器
                   ◇電試彙報
                                   ◇電試彙報
                                            ◇電試彙報
                                                    ◇電試彙報
                                                                     ◆電試彙報
                                                                                      ◇電試彙報
                                                                                                                                        ◆電試彙辑
                                                                                                                                                                                                                                     ◇電試彙辑
                                                                                                                                                                                                                                                              ◆電試彙躁
                                                                                                                                                                                                                                                                       ◇電試彙報
                                                                                                                                                                                                                                                                                        ◎電試彙報
                                                                                                                                                                                                                                                                                                                 ◆電試彙報
           ◎電試彙報
                                                              ◇電試彙報
                                                                               ◎電試彙報
                                                                                                                                                                                                                                                                                ◎電試彙報
                                                                                                                                                                                                                                                                                                 ◎電試彙報
                                                                                                                                                                                                                                                                                                          ◎電試彙報
                                                                                                                                                                                                                                                                                        ,
3
                                                             '3
1
                                                                                                                                                  '3
1
                                                                                                                                                                                                    ,
3
1
                                    '3
1
                                             '3
1
                                                     '3
1
                                                                      '3
1
                                                                               '3
1
                                                                                      '3
1
                                                                                               '3
1
                                                                                                       '3
1
                                                                                                                '3
1
                                                                                                                        '3
1
                                                                                                                                '3
1
                                                                                                                                         '3
1
                                                                                                                                                         '3
1
                                                                                                                                                                  '3
1
                                                                                                                                                                          '3
1
                                                                                                                                                                                   '3
1
                                                                                                                                                                                            '3
1
                                                                                                                                                                                                            '3
1
                                                                                                                                                                                                                     '3
1
                                                                                                                                                                                                                              '3
1
                                                                                                                                                                                                                                      '3
1
                                                                                                                                                                                                                                              '3
1
                                                                                                                                                                                                                                                       '3
1
                                                                                                                                                                                                                                                               '3
1
                                                                                                                                                                                                                                                                        '3
1
                                                                                                                                                                                                                                                                                ,
3
1
                                                                                                                                                                                                                                                                                                 '3
1
                                                                                                                                                                                                                                                                                                         '3
1
                                                                                                                                                                                                           ,
1
0
                                                                                                               ,
0
1
                                                                                                                                                                                                   ,
1
2
           '0
6
                                   '0
1
                                             '0
5
                                                            '0
1
                                                                                                      ,
1
1
                                                                                                                                                         ,
1
0
                                                                                                                                                                         ,
0
9
                                                                                                                                                                                  ,
0
5
                                                                                                                                                                                                                    '0
9
                                                                                                                                                                                                                                     '0
9
                                                                                                                                                                                                                                                              '0
5
                                                                                                                                                                                                                                                                                        '0
4
                                                                                                                                                                                                                                                                                                         ,
1
2
                                                                                                                                                                                                                                                                                                                          '0
3
                                                     '0
4
                                                                              .
0
1
                                                                                      '0
9
                                                                                               '0
1
                                                                                                                        '0
1
                                                                                                                                '0
1
                                                                                                                                         '0
1
                                                                                                                                                  '0
1
                                                                                                                                                                  '1
2
                                                                                                                                                                                            '1
1
                                                                                                                                                                                                                             '0
7
                                                                                                                                                                                                                                              '0
5
                                                                                                                                                                                                                                                      '0
5
                                                                                                                                                                                                                                                                        ,
1
0
                                                                                                                                                                                                                                                                               '0
3
                                                                                                                                                                                                                                                                                                  '0
8
                                                                                                                                                                                                                                                                                                                                    '0
5
                                                                     8
6
8
                                                                                                                                                                                  6 5 3
                                                                                                                                                                                                                                             6 5 3
                                                                                                                                                                                                                                                                                       6
2
4
                                                                                                                                                                                                           .
1
0
7
3
                                                                                                                                                                                                                                     .
1
0
5
1
                                                                                                                                                                                                                                                                                                                 .
1
0
9
           .
7
5
1
                                                     .
4
5
2
                                                             .
1
4
4
                                                                               .
1
2
9
                                                                                              .
1
2
3
                                                                                                                                                  .
5
3
                                                                                                                                                                           .
1
0
0
4
                                                                                                                                                                                                    .
1
2
9
4
                                                                                                                                                                                                                     .
1
0
3
7
                                                                                                                                                                                                                                                                                .
4
2
2
                                                                                                                                                                                                                                                                                                 .
9
2
3
                                                                                                                                                                                                                                                                                                          .
1
2
9
4
                                                                                                                                                                                                                                                                                                                          .
9
8
                                                                                                                                                                                                                                                                                                                                   .
6
7
                                                                                      .
0
6
4
                                                                                                                                                                  1
2
5
9
                                                                                                                                                                                            .
1
5
3
                                                                                                                                                                                                                                                                                                                                            773
    1
0
7
3
                     1
0
4
1
                             1 2 3 3
                                     0
6
8
                                              6
6
7
                                                                                                         1 6 9
                                                                                                                 0
7
6
                                                                                                                        0
4
7
                                                                                                                                  0
5
3
                                                                                                                                          0
6
8
                                                                                                                                                           1
0
7
3
                                                                                                                                                                                                                                                        6
7
5
                                                                                                                                                                                                                                                                6
5
3
                                                                                                                                                                                                                                                                         1
0
7
3
                                                                                                                                                                                                                              8
6
8
```

Fig. 1. A Sample Output.

character, which usually reflects its 'on-yomi' order of kanzi. As is shown in Fig. 1, the result is almost natural but strange at some points.

6. Experiments

A program was written using COBOL and experimental KWIC index was obtained successfully. A sample page is shown in Fig. 1.

Input paper tape is prepared manually by kanzi-teletypewriter and computer output is punched on paper tape. Kanzi-teletypewriter prints it with average speed 2 characters/second.

Only 2 errors out of 1031 KWIC lines was found in the result of our automatic segmentation method. The indexing required 3.5 min for the computer (FACOM 230-50).

7. Conclusion

The proposed segmentation method is considered to be simple and practical. So far as the titles of technical papers concern, the method has enough accuracy for KWIC indexing. Further investigation for automatic arrangement (sorting) of Japanese words will be necessary to get more conventional indices.