DEVELOPMENT OF BASIC PRACTICAL TECHNIQUES FOR JAPANESE LETTER STRING PROCESSING - AUTOMATIC KEYWORD EXTRACTION AND AUTOMATIC READING

K. Araki, K. Hinatsu, K. Itayama, T. Sahara, Y. Sakagami and F. Takano

The Japan Information Center of Science and Technology (JICST)
2-5-2 Nagatacho Chiyodaku Tokyo 100 Japan

Japanese is a peculiar language among the thousands of languages in the world. There exist only two of the same class; Japanese and Korean. Japanese is written both in Chinese characters (ideograph) and in Kana (Katakana and Hiragana - phonetic symbols) in mixture without any space. Moreover, Chinese characters in Japanese have, in most cases, several readings and play several roles depending on the context and letter string characteristics. So for written Japanese, it was very difficult to segment letter string and extract adequate terms from sentence and to give them correct readings automatically, which are indispensable for terminology, automatic reading, automatic indexing, key-boarding from on-line terminals otherwise more than 2,000 character key-board is necessary.

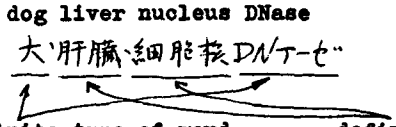
The authors invented efficient algorithm and developed computer programms and dictionaries for successful solution of the problems above for the first time in Japan.

The system consists of two subsystems called K-KACS (Kanji-Kana Automatic Conversion System) and JAKAS (Japanese Keyword Automatic Selection).

Some Chinese characters act both as suffix, preffix or preposition and as parts of meaningful words. We comprehensive-

ly collected such characters (about 500) and those terms in which the characters are included not as fixes or prepositions but as important part (about 8000 words). Letter string which is matched with dictionary term is passed but the letter remained and coincides with the special character itself is cut.

In case of long letter string without such special letter, sentence is cut by those terms of dictionary which are thought to be definite within reasonable amount. That is:



indefinite type of word. definite type of word.

Equally, among the variety of readings - in some cases more than 8 - some are special and definite and others are indefinite but obey to rules. We collected these special readings (about 25,000) for about 2,000 Chinese characters and developed algorithm and programme to select the correct reading for each Chinese character with the precision higher than 99.94 %.

As the dictionary is small enough and logic is simple, implementation and maintenance are relatively easy and the speed is very high.

JICST adopted this system for its information file production and services of more than 400,000 citations per year and save costs.

By the development of the techniques, processing of Japanese has become to be able to cope with western languages. We were awarded for the work The Prize of Learning of Japan Association of Information and Documentation in 1980, and have applied patent (Japan Patent Kokai Showa 55 (1980) - 102074).

000167 81/11/05

A010 7314550010Ⓐ P108428104258105060Ⓐ1101447
B010 1BB03030GⒶ534.2-7/-8
B020 2DC06020VⒶ551.463
B510 1E05072G0806IP0303
C010 a105 0100EN 800978514E759BDZ USA80 OceansⒶOceans
C020 2Ⓐ ⒶⒶ1980
C030 ⒶⒶⒶ127-131
D010 1Univ. New HampshireⒶUniv. New Hampshire
E010 1COX PⒶCOX P
E020 1HARVEY PⒶHARVEY P
E030 1RENTIS PⒶRENTIS P
E040 1SIVAPRASAD KⒶSIVAPRASAD K
E050 1YILDIZ AⒶYILDIZ A
E060 1YILDIZ MⒶYILDIZ M
F010 Sound propagation in a shallow water region overlying a viscoelastic halfspace.
G010 粘弾性半空間上の浅海領域の音波伝搬
H010 (ネンダンセイハンクウカンジョウ ノ センカイリョウイキ ノ オンパデンパン)
I010 Green関数と基準モード展開法とを用いて浅海の音波伝搬を調べた。理想流体の境界が上面の圧力零，下面は粘弾性固体で，流体の音速は深さに依存するとして解析した。バルト海の実験海域のパラメータにより数値計算し，送受波器間水平距離と伝搬損失との関係をプロットした。結果は実験と定性的に一致し，海底の組成が音波伝搬に重要なことがわかった

J010 000313Ⓐ* 音波伝搬Ⓐオンパデンパン
J020 000840Ⓐ* 水中音響Ⓐスイチュウオンキョウ
J030 035336Ⓐ* 海底Ⓐカイテイ
J040 000999Ⓐ 粘弾性Ⓐネンダンセイ
J050 000302Ⓐ 音速Ⓐオンソク
J060 021666Ⓐ 水深Ⓐスイシン【フカサ】
J070 010455Ⓐ 依存性Ⓐイゾンセイ
J080 007916Ⓐ バルト海Ⓐバルトカイ

K001 015216Ⓐ01Ⓐ 波動伝搬Ⓐハドウデンパン
K002 046038Ⓐ02Ⓐ 伝搬Ⓐデンパン
K003 000292Ⓐ01Ⓐ 音響学Ⓐオンキョウガク
K004 002882Ⓐ02Ⓐ 物理学Ⓐブツリガク
K005 004281Ⓐ03Ⓐ 科学Ⓐカガク【science】
K006 013895Ⓐ01Ⓐ 位置Ⓐイチ
K007 000427Ⓐ01Ⓐ 機械的性質Ⓐキカイテキセイシツ
K008 035699Ⓐ01Ⓐ 弾性波速度Ⓐダンセイハソクド
K009 025003Ⓐ02Ⓐ 波の速度Ⓐナミノソクド
K010 000951Ⓐ03Ⓐ 速度Ⓐソクド【スピード】
K011 014185Ⓐ01Ⓐ 深さⒶフカサ
K012 014135Ⓐ02Ⓐ 長さⒶナガサ
K013 042740Ⓐ03Ⓐ 幾何学量Ⓐキカガクリョウ
K014 031902Ⓐ01Ⓐ 北東大西洋Ⓐホクトウタイセイヨウ
K015 031901Ⓐ02Ⓐ 北大西洋Ⓐキタタイセイヨウ
K016 031900Ⓐ03Ⓐ 大西洋Ⓐタイセイヨウ
K017 020834Ⓐ04Ⓐ 海洋名Ⓐカイヨウメイ

L010 浅海Ⓐセンカイ

M010 *粘弾性半空間Ⓐネンダンセイハンクウカン
M020 浅海Ⓐセンカイ
M030 領域Ⓐリョウイキ
M040 音波伝搬Ⓐオンパデンパン

original title
translated title
automatically transl. title (reading)
human indexing
up-word pasting by thesaurus
free human indexing
automatic indexing from Japanese title
reading

000041 81/11/05