索引作成支援システム「SAKUIN君」

ベンチャー6

大森久美子　東田正信
NTT情報流通プラットフォーム研究所

索引作成支援システム「SAKUIN君」*1は日本語文字列の特徴を利用して、文章から重要な単語を自動抽出するシステムです。SAKUIN君が抽出した索引語候補の中から不要な語を取り除いたり若干の修正をするだけで索引が作れます。

1 用語（専門用語、人名、会社名等）の自動抽出機能

索引作成支援システムが、文書の中から、索引語等の用語を自動的に探してくれ ます。

2 読み（フリガナ）の自動付与機能

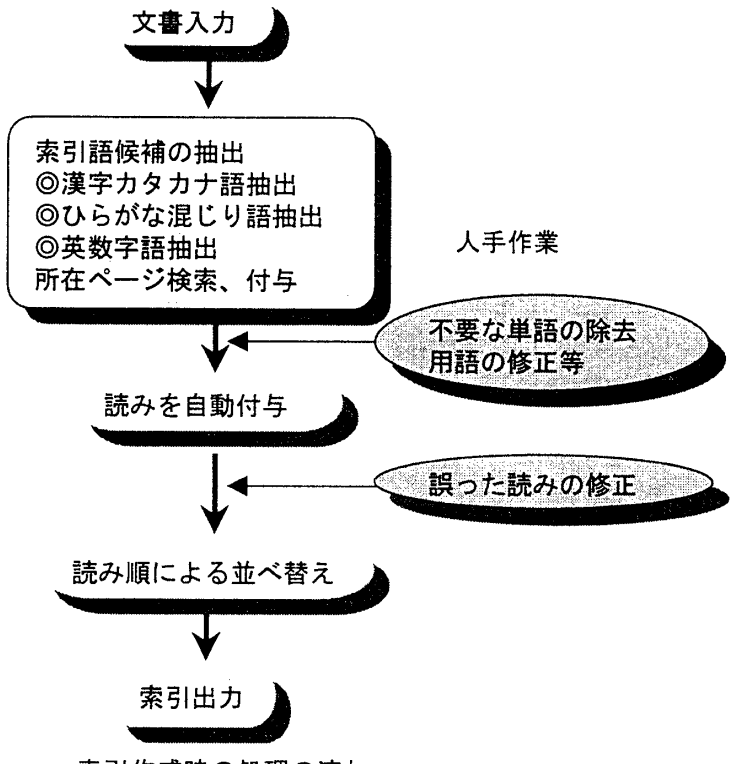
索引作成支援システムが、自動的に読みを付与してくれます。

3 用語の所在（ページ）自動付与機能

索引作成支援システムが、自動的に用語の所在ページを探してくれます。

4 編集（修正、削除等）機能

用語の編集が、画面上で簡単に行なえます。



索引作成時の処理の流れ

SAKUIN君の他の用途

★利用者辞書の作成
機械翻訳などの言語処理の前処理として利用者辞書に登録すべき単語の抽出に利用できます。

★文書の用語統一
マニュアル等で使用されている用語の一覧表を作れますので、用語の統一が容易にできます。

◇対象文書：テキストファイルならどんなものでもOKです。
◇動作環境：Windows3.1, Windows95, Windows NT4.0
ハードディスク 5MBの空領域
◇プログラムサイズ：20KS

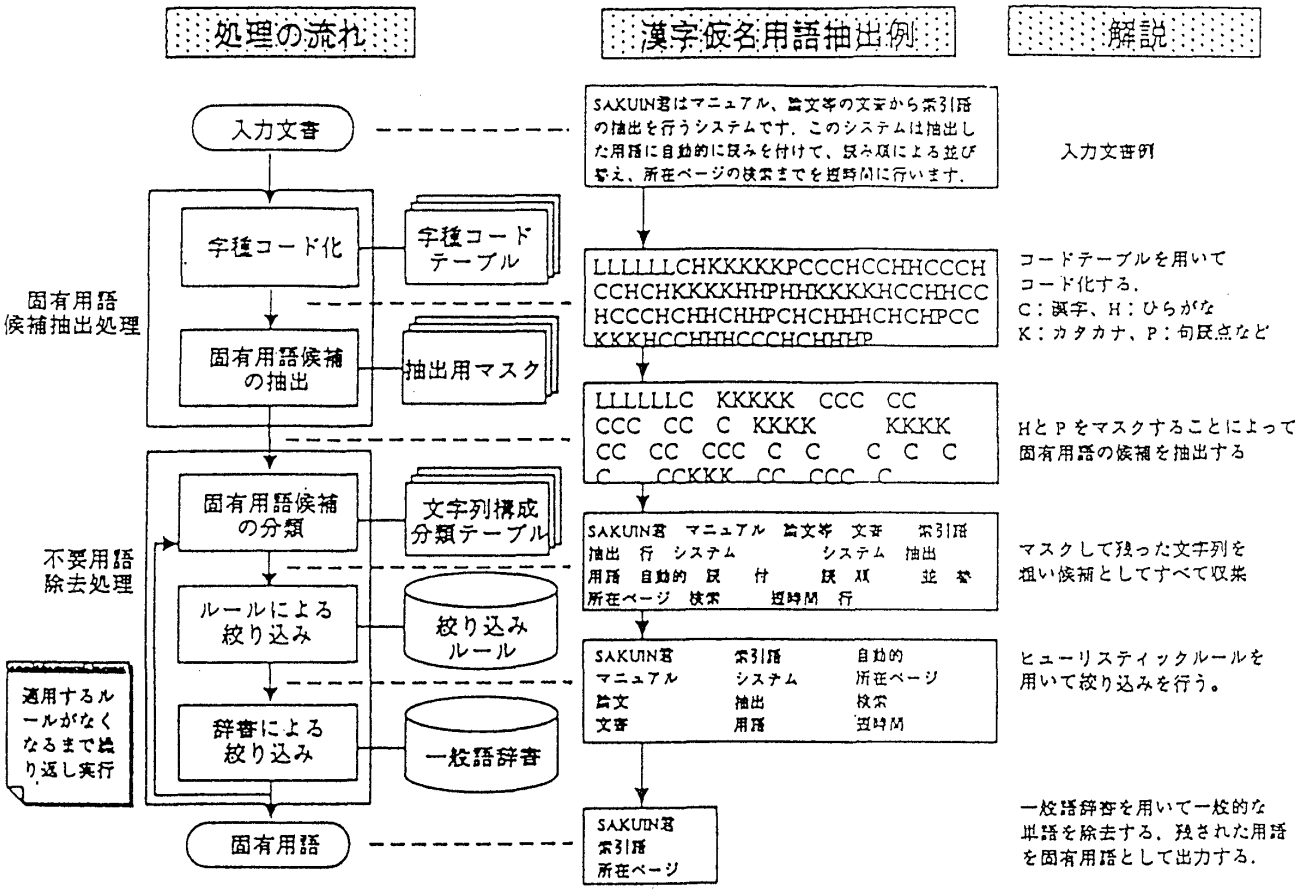
SAKUIN君に関するお問い合わせ
NTT情報流通プラットフォーム研究所
0468-59-3745　大森
kumiko@isl.ntt.co.jp

*1　SAKUIN: Semi-Automatic Knowledge-based Utility for INdexing

SAKUIN君における新規技術

以下、索引作成支援システム「SAKUIN君」開発の際に用いた新規技術について述べる。

1. 索引語抽出方法【特許公開番号 4-188364】



2. 索引語に対する読み付与方法【特許出願番号 4-23001】

基本方針： 抽出した索引語に対して、前から順に最長一致で、辞書に掲載されている語と未掲載の語に最長一致で分割する。

- 辞書に掲載されている語に対しては、辞書の読みを適用
- 辞書に未掲載の単語に対しては、
  - 漢字のみからなる単語 → 各漢字に音読みを適用
    例）統語規則（統／とう 語／ご 規則／きそく，規則は辞書掲載単語）
  - 漢字仮名交じりの単語 → 各漢字に対する送り仮名からその読みを判断
    例）取り組み（取り／と-り 組み／く-み）

参考文献

[1] Masanobu, H., Masahiro, O.: "Automated Indexing for Japanese Text"; PRICAI '94
[2] 東田 正信, 奥 雅博：「高速用語抽出技術-形態素解析技術を使用しない自然言語処理技術」; 情報処理学会研究会, 1992.
[3] 東田 正信, 奥 雅博：「日本語文書に含まれる固有用語の自動抽出方式」; 情報処理学会全国大会, 1990.