テキスト解析ツール ibukiTool について

4H−4

辻子純央, 兵藤安昭, 池田尚志
{tsujiko,hyodo,ikeda}@ikd.info.gifu-u.ac.jp
岐阜大学工学部

1　はじめに

我々は日本語文節解析システム IBUKI を開発している [1]。またその応用として、自動点訳システム IBUKI-TEN[2] を開発した。このような応用システムでは、解析誤り箇所や未登録語を見い出し、辞書の修正や登録をしていく作業が欠かせない。

そこで我々は、任意のテキストデータを入力として、IBUKI で解析された結果をさまざまな角度から眺めることができ、また辞書登録、再解析が簡単に行えるなどの機能を備えたテキスト解析ツールを開発した。解析結果は市販の RDB に格納しており、RDB の諸機能を利用してツールを構成している。ibukiTool は、未登録語の処理など辞書作成ツールとして活用できるばかりでなく、小説などの著作物の語彙統計を取るなど文書解析ツールとして便利に活用できる。

2　日本語文解析システム IBUKI

IBUKI はまず文節解析を行い、ついで係り受け解析を行う [1]。文節解析では、文節として可能性のあるものをすべて求めた上で、文節単位のコスト最小法で解を求めている。文節には文節カテゴリ（構文的な観点から文節を分類したもの）を付与する。係り受け解析では文節カテゴリを元に、係り受け可能な文節を求め文節間の関係を解析している。ibukiTool は現在のところ IBUKI の文節解析を利用している段階である。

3　ibukiTool

ibukiTool の主な機能は、現在のところ次のようである。

- 入力テキストファイルを指定して文節解析
- 任意のキー入力されたテキストを文節解析
- 解析された語の一覧を表示（延べと異なり）
- 解析された複合語の一覧を表示（延べと異なり）
- 未登録語として解析された語とその出現文脈を表示（延べと異なり）
- 語の一覧を、出現順、読みの順、品詞、出現頻度順などで整列して表示
- 適当にフィルタリングして、条件に合うものだけを表示
- 解析された語が出現した文を表示
- 解析された語の辞書情報を表示
- 解析誤りの可能性ありと判定した箇所を表示
- メモリ上の辞書に語を登録
- メモリ上の辞書から語を削除
- 新しい辞書でテキストを再解析
- メモリ上の辞書をファイルに保存

解析誤りの可能性に関しては、独立文節（機能語を伴わない文節）や、終端文字がひらがな小文字の文節等、統計的に誤りの多い箇所を抽出しており、適合率は 14 %、再現率は 70 %程度である [3]。なお、IBUKI の解析精度は、EDR 日本語コーパスのデータに対して 98 %程度である [1]。また辞書としては、自立語については EDR の辞書をベースにしており、機能語については長単位の表現を採用するという方針で我々の研究室で作成したものを用いている。

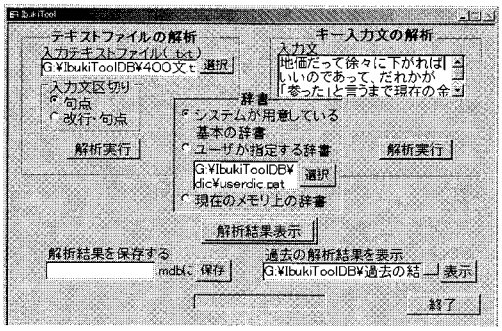

図 1: ibukiTool

4　利用例

ibukiTool でテキスト 400 文 (新聞記事 400 文、サイズ 32.4KByte) を解析した例を示す。

表 1 はシステムが用意している辞書での解析結果の統計, 表 2 はその解析結果を元に 79 語を辞書に登録した場合の解析結果である。

また、抽出例として普通名詞語彙の頻度上位 10 語を表 3 に示す。

Text analysis Tool : ibukiTool
M Tsujiko, Y Hyodo, T Ikeda
Faculty of Engineering, Gifu University
Gifu-shi, 501-1193, Japan

図 2: 解析結果

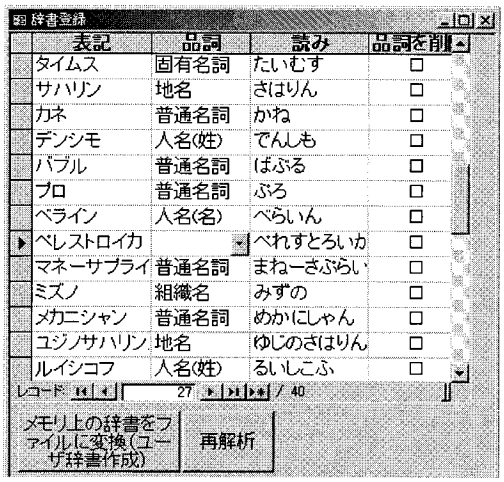

図 3: 辞書編集作業

5 おわりに

文書データを解析し、単語や複合語の一覧・語彙統計、未登録語、誤り可能性のある文節等の情報を抽出・表示するツールを開発した。

このツールを用いて、辞書や解析システムの整備を行うことができる。

さらに効果的な解析及び解析結果の活用ができるよう整備を進めたい。

表 1: 登録前の統計

出現語彙数（延べ）	9988
出現語彙数（異なり）	2213
出現複合語数（延べ）	871
出現複合語数（異なり）	605
未登録と解析された語の数（延べ）	100
未登録と解析された語の数（異なり）	66
解析誤りの可能性ありと指摘された箇所	44

表 2: 登録後の統計

出現語彙数（延べ）	9917
出現語彙数（異なり）	2201
出現複合語数（延べ）	830
出現複合語数（異なり）	579
未登録と解析された語の数（延べ）	17
未登録と解析された語の数（異なり）	13
解析誤りの可能性ありと指摘された箇所	42

表 3: 普通名詞頻度上位 10 語

頻度	表記	読み
31	女性	じょせい
26	男性	だんせい
25	人	ひと
22	経済	けいざい
20	国民	こくみん
20	問題	もんだい
19	世代	せだい
16	日本人	にほんじん
15	政治	せいじ
14	世界	せかい

参考文献

[1] 兵藤安昭, 池田尚志：文節単位のコストに基づく日本語文節解析システム, 言語処理学会 第5回年次大会 (1999)

[2] 横平貫志, 兵藤安昭, 池田尚志他：自動点字翻訳システム IBUKI-TEN, 情報処理学会 第61回全国大会 (2000)

[3] 村上裕, 神光太郎, 兵藤安昭, 池田尚志：複合語・文節解析誤り個所の検出, 言語処理学会 第7回年次大会 (2001)