未知語獲得システムとその評価

亀田弘之（東京工科大学）

〒192　東京都八王子市片倉町1404-1
東京工科大学工学部情報工学科
TEL: 0426-37-2111 内線2212
E-mail: kameda@cc.teu.ac.jp

1．はじめに

自然言語は、人間相互の意思疎通のための道具であるとともに、思考を行うための道具でもある。従って、機械に自然言語を処理する能力を付与することができれば、人間と機械とが自然言語を用いて相互に円滑な意思疎通を行うことが可能となり、マルチメディア等におけるマンマシンインタフェースが格段に向上されるとともに、機械が自律的に思考を行うことも可能となることが期待されるため、真に高度な知識情報処理システムの構築が望まれる。

このような観点から筆者らは、思考過程の解明とその工学的応用を念頭に置きつつ、自然言語を素材として、自然言語処理の高度化に関する研究を行っている[1,2]。本稿ではそのうち、筆者らが作成中の「日本語漢字仮名交じり文を対象とする未知語獲得システム」をとりあげ、そのシステムの概要とその簡単な評価について述べる。

2．未知語の定義・種別

2-1. 未知語の定義

人間は、自然言語を媒体として相互の意思疎通を行う場合、状況や背景的知識等を適宜利用して、見かけ上同一の表現であっても多様な意味を表現・伝達することができるとともに、さらには必要に応じて新たな単語や表現を創造する。これに対して、その受信者は、多くの場合何の支障もなく、それらの単語や表現に担われた発話者の意図する意味を円滑に推察し理解することができる。

一方、現在の機械は、上述したような高度な処理能力を持っておらず、一般的には、単語辞書と、単語の配列を規定する文法規則（統語規則）とを主たる知識として言語処理を行っているため、多義的な表現や新しい創造的な表現は、十分に処理することができない。特に、システムの単語辞書に予め載っていない単語は、システムにとっては未知となる。本稿ではこのような観点から、「機械にとっての未知語」とは未登録単語のことであるとし、以下の議論ではこの定義によるものとする。

2-2. 未知語の種別

日本語における未知語には、大きく分けると3つの種別があり、本研究ではそれらを第一種の未知語・第二種の未知語・第三種の未知語と呼ぶこととする。以下にそれらの定義と実例とを示す。なお、この定義に関する詳しい説明は、参考文献[3]を参照されたい。また、以下の未知語の実例（下線の付されたもの）はすべて、広辞苑（第4版・岩波書店）の主見出しとして記載されていないものである。

2-2-1. 第一種の未知語

【第一種の未知語の定義】　単語自体は辞書に登録されているにもかかわらず、表記が辞書のものと異なるために、辞書検索に失敗する単語（異表記同義語）のこと。

この種の未知語は、日本語における表記の多様性によるものであり、例えば、異表記の種類により、以下のようなものがある。

(1) 漢字異表記：「喜ぶ」と「慶ぶ」
(2) 送りがな異表記：「行う」と「行なう」
(3) 混ぜ書き異表記：「飛び込む」と「飛びこむ」
(4) 片仮名異表記：
　「ソフトウェア」と「ソフトウエア」、
　「コンピューター」と「コンピュータ」、
　「バイオリン」と「ヴァイオリン」
(5) 記号の異表記：　「百」と「１００」

第一種の未知語はこのように、表記におけるゆれ・慣用的用法・学術（専門）用語の表記規約等に起

因するものの他に、「みんなでガンバロー！」のように特定の単語・表現を強調する等の特殊な用法に起因するものもある。

2-2-2. 第二種の未知語

【第二種の未知語の定義】　単語の各構成要素は辞書に登録されているが、その単語自体は辞書に登録されていない単語（既知語を用いて造語された複合語）のこと。

第二種の未知語は、最も出現頻度が高く、その例としては、以下のようなものがある。

例：「情報学」　（情報 ＋ 学）
「数学辞典」（数学 ＋ 辞典）
「再試験」　（再　＋ 試験）
「湾岸支援」（湾岸 ＋ 支援）

日本語においては、必要に応じてさまざまな複合単語が日常的に造語され利用されるので、この種の未知語の処理は重要である[4]。

2-2-3. 第三種の未知語

【第三種の未知語の定義】　単語の構成要素として、単語辞書に登録されていないものが含まれるもの。

第三種の未知語には以下のようなものがある。

(1) 辞書にない単語構成要素（強調文字で表記）を部分的に含む単語
例：「ＩＰアドレス」、
「トラブル・シューティング」
(2) 単語構成要素すべてが辞書にない単語
例：「ボリス・パンキン」、
「インターネット」
(3) その他
例：（省略により第三種の未知語となるもの）
「フ諸島」（フォークランド諸島のこと）
「東工大」（東京工業大学のこと）

3．未知語獲得システム

一般に「未知語獲得」過程は、「未知語処理（未知語の検出・内部構造推定・意味推定）」と「未知語の辞書登録」とからなる。本研究では、未知語処理として、検出と（第二種の未知語に対しては）内部構造推定とを行っており、意味推定は特に行っていない。以下に、まず本研究における未知語処理の基本的考えを述べたあと、未知語獲得システムの概要、未知語獲得システムの処理の流れ、動作例について順次述べる。

3-1. 未知語処理の基本的考え

簡単のために、以下のような文法体系を考える。

```
文　　→　形容詞 ＋ 名詞
文　　→　連体詞 ＋ 名詞
形容詞　→　大きい
連体詞　→　大きな
名詞　　→　夢
```

このような文法体系に対して、以下のようなプログラムをprologにより記述する。

```
文(A,C,文(_形容詞,_名詞)):-
    形容詞(A,B,_形容詞),名詞(B,C,_名詞).
文(A,C,文(_連体詞,_名詞)):-
    連体詞(A,B,_連体詞),名詞(B,C,_名詞).
形容詞([大,き,い|T],T,形容詞(大きい)).
連体詞([大,き,な|T],T,連体詞(大きな)).
名詞([夢|T],T,名詞(夢)).
連体詞(AT,T,未知語連体詞(_連体詞):-
  append(A,T,AT), list_to_atom(A,_連体詞).
```

このプログラムの最初の７行が上記の文法体系をprologに変換したものであり、末尾の２行が連体詞のみに関する未知語処理プログラムである。

なお、実際に作成したシステムにおいては、末尾２行の本体部分に語尾チェック処理や内部構造チェック処理等の処理を含ませるとともに、名詞・副詞に関しても未知語処理するようにしている。

3-2. 未知語獲得システムの概要

筆者らは上述の種々の未知語を対象とする「未知語処理システム」を作成中であるが、そのシステムは、ノート型パーソナルコンピュータDigital HiNote CT475（主記憶20MB, ハードディスク350MB, DEC製）上に、Arity/Prolog（Version 5.1, ライフボート社製）を用いてインプリメントし、処理モジュール（主処理モジュール部・テキスト入力モジュール部・統語解析モジュール部・第一種未知語処理モジュール部・第二種未知語処理モジュール部・第三種未知語処理モジュール部・単語獲得モジュール部・補助関数モジュール部、これら全体で約２，４００行）と知識ベース（単語構成要素データベース・単語辞書データベース・造語規則データベース・統語規則データベース）とからなり、プロトタイプシステムが現在稼働中である。

3-3. 未知語獲得システムの処理の流れ

本システムの処理の流れの概要を次頁の図１に示す。処理制御の流れは、prologシステムに依存しており、トップダウン的に処理する。以下に図１に準

じてアルゴリズムの概略を述べる。

①**テキスト読込み：** 漢字仮名交じりべた書きの日本語文を、文字列としてキーボードから読込む。

②**文リスト構築：** 読込まれた文字列を文字毎に分解しリスト構造の形式に変換する。変換結果を以下、文リストと呼ぶ。

③**文法による先頭単語品詞の推定：** 文法（統語規則）を参照しながら、文リストの先頭に位置する単語の品詞を、トップダウンに推定する。

④**単語候補切出し：** 文リストの先頭側から単語候補として、部分リストを切出す。

⑤**辞書検索：** 単語候補としての部分リストが、辞書に登録されているか検索する。検索に成功すれば、これは未知語ではなく、処理は⑨に移る。検索に失敗する場合は、これを未知語候補とみなし、処理は次の⑥へ移る。

⑥**第二種の処理：** 未知語候補が、第二種の未知語かどうか調べる。第二種の未知語とみなし得る場合には、処理は⑨へ、そうでなければ⑦へ移る。

⑦**第一種の処理：** 未知語候補が、第一種の未知語かどうか調べる。第一種の未知語とみなし得る場合には、処理は⑨へ、そうでなければ⑧へ移る。

⑧**第三種の処理：** 未知語候補が、第三種の未知語かどうか調べる。第三種の未知語とみなし得る場合には、処理は⑨へ、そうでなければ④へ移る。なお、上記⑥～⑧が未知語処理の中核部分である。

⑨**文リストからの先頭単語の削除：** 文リストの先頭に位置する単語を削除し、残りのリストを新たな文リストとする。

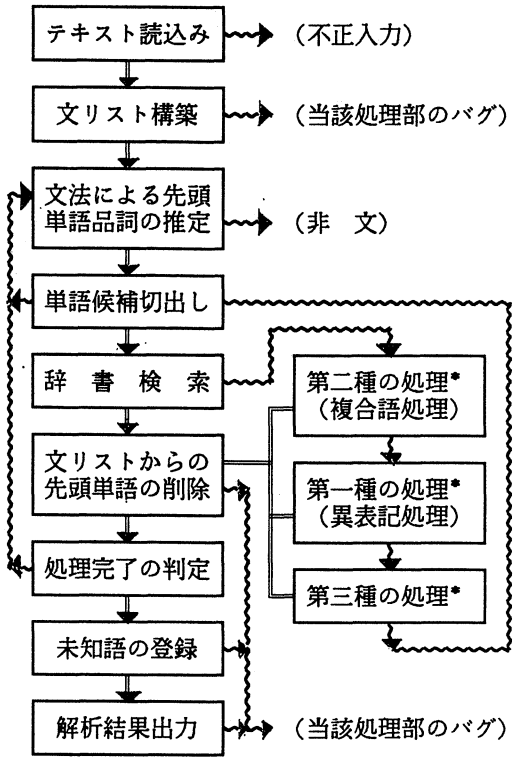
⑩**処理完了の判定：** 文リストが空リストか調べる。空リストならば、統語解析の処理は完了しているので⑪へ、そうでなければ③へ移る。

⑪**未知語の登録：** 統語解析の際に、未知語が検出されていれば、推定品詞等の情報も統合して、新たな辞書項目として辞書に登録する。

⑫**解析結果出力：** 統語解析結果をディスプレイ上に表示する。

3-4. 動作例

例えば、“イタリヤ”（第一種の未知語）、“見学旅行”（第二種の未知語）、“シェーンな”（第三種の未知語）を含む文として、「シェーンなイタリヤの見学旅行に行った」を入力すると出力として、



<<注>>　⟶ ： 処理成功時の流れ

⇝ ： 処理失敗時の流れ

* ： 未知語処理のモジュール部分

図1. 未知語獲得システムの処理の流れの概要

『文(主部(名詞句(名詞句no(連体詞(プチな，第三種未知語)，名詞句(名詞(第一種未知語(名詞(イタリア)))、助詞(の)))，名詞句(未知複合語(見学旅行)、助詞(に)))))，述部(動詞句(動詞(行く)))))』が表示される。この例では、統語規則の知識とともに、“シェーンな”が連体詞の語尾「な」を持っていること等から連体詞と推定され、“イタリヤ”が“イタリア”の異表記単語として照合され、さらに、“見学旅行”は“見学”と“旅行”が複合語の構成要素になり得るとの知識および造語規則から、第二種の未知語と推定されている。

4．システムの評価

システムの基本的性能を評価するために、単語辞書と統語規則とを以下の手順で作成し、それに基づき評価実験を行ない、動作の妥当性を確認した。

4-1. 単語辞書と統語規則の作成

以下の手順・方法により単語辞書と統語規則とを順次作成した。

(1) 印字テキストの収集…未知語獲得システムに付与するための単語・統語構造に関するものとして、比較的基本的なものが多く記載されている書籍を収集した。具体的には、「スペイン語基本文2000」(大学書林)に記載されている日本語文を収集の対象とした。(例：「明日私はゼゴビアに参ります。」)

(2) 電子化テキストの生成…書籍等に印字されている日本語漢字仮名交じり文を、OCR(富士電機製)によりパーソナルコンピュータ(FMR,富士通製)に読込み、シフトＪＩＳコード形式のMS-DOS用テキストファイルに格納する。なお、OCRにおける読込み誤りは、エディタにより人的作業により発見・修正したが、読み誤りには比較的規則性があるため、人間の手作業による入力文の場合よりも系統的に処理することができた。

(3) タグ付きテキストの生成…上記の電子化テキストを素材として、事前に規定されている文法体系の枠組みに基づき、電子化テキストの単位切りを行うとともに、その単語の品詞情報をタグの形式で追加記入した。(例：「明日(副詞)私(名詞)は(係助詞)ゼゴビア(名詞)に(格助詞)参り(動詞)ます(助動詞)。(句点)」)

(4) 単語辞書データの生成…タグ付きテキストを入力として、例えば、データ「学校(名詞)へ(格助詞)行く(動詞)」から、学校(名詞)、へ(格助詞)、行く(動詞)の３つの単語とその品詞情報を生成する。

(5) prolog形式の単語辞書の生成…上記(4)で生成されたデータをprolog形式に変換する。(例：単語辞書データ「学校(名詞)」から、「名詞([学,校|T], T, 名詞($学校$), $意味$, $種別$).」を生成)

(6) 統語規則データの生成…タグ付きテキストを入力として、例えば上記(4)と同じデータ「学校(名詞)へ(格助詞)行く(動詞)」から、「文 -> 名詞,格助詞, 動詞」という統語規則データを生成する。

(7) prolog形式の統語規則の生成…上記(6)で生成されたデータをprolog形式に変換する。(例：統語規則データ「文 -> 名詞　動詞」から、prolog形式

```
文(A, C, 文(_名詞1, _動詞2), _未知語)
  :- 名詞1(A,B, _名詞1, _未知語1),
     動詞2(B, C, _動詞2, _未知語2),
     packing(_未知語1, _未知語2, _未知語).
```

を生成)

(8) 重複データの削除…上記の(5)と(7)で生成されたprolog形式のデータは、そのまま未知語獲得システムで利用することのできるが、重複するデータが存在する場合があるので、全データをソートして重複したものを１つに縮退させる。

以上の作業により、現在単語および統語構造ともに約２０００個程のデータを生成した。なお、上記作業のうち、(2)は３０人日、(4)は３００人日を要したが、(5)以降は、プログラム言語jgawkにより記述したユーティリティにより実行した。

4-2. システムの評価

上記の単語辞書と統語規則とをシステムに付与して、動作実験を行ったところ、人間と対話的に動作させることにより、適切な処理を実時間で行わせることが可能であることを確認した。なお、単語辞書等を大規模化するとかなりの処理時間を要することが予想されるので、今後は意味処理等の点の拡張とともに、この点に関する改良も必要である。

5．おわりに

本稿では、筆者が作成中の未知語獲得システムの概要とその簡単な評価について述べた。

なお、本研究の一部は、文部省科学研究費補助金試験研究(B)(1)（課題番号：07558274，研究代表者藤崎博也）により行われた。

<<謝辞>> 筆者に本研究テーマをお与え下さるとともに、本研究推進にあたり多くのご助言を下さった藤崎博也教授（東京理科大学教授）に感謝する。

<<参考文献>>

[1] 亀田・桜井：“べた書き日本語文からの未知語獲得システムの作成”，電子情報通信学会「思考と言語」研究会技報TL94-11，pp.17-24(1994).

[2] 亀田・藤崎：“高次辞書ﾃﾞｰﾀﾍﾞｰｽのための語彙知識自動獲得システム”，公開ｼﾝﾎﾟｼﾞｳﾑ「人文科学とﾃﾞｰﾀﾍﾞｰｽ」，pp.75-82(1995).

[3] 亀田・藤崎・森田・倉島：“未知語の分類とその処理に関する考察”，情報処理学会第36回全国大会講演論文集，5T-5，pp.1195-1196(1988).

[4] 荻野：“名詞辞書に含まれるべき見出しの範囲－特に複合名詞の扱いをめぐって－”，情報処理振興事業協会，61技-072，pp.207-221(1987).