9B−6

電子化辞書管理のための自然言語インターフェース
ー 質問文コーパスの機能分類 ー

小林 美佳　村木 一至　*桧山 努

(株)日本電子化辞書研究所　*NEC技術情報システム開発

1．はじめに

ＥＤＲ電子化辞書の開発には、新語の登録やエラー修正、品質検査等の作業が必須であり、これらの作業を計算機を用いてサポートすることは効率的な開発に有効である[1]。これらの作業を行なう辞書作業者の負担を軽くするために作業者が普段使い慣れている自然言語を用いて辞書の管理業務を行なうシステムが最適であると考えた。このシステムを使いやすいものにするには、まず辞書管理システムとしてどのような機能が必要であるかを把握する必要がある。そこで、辞書管理業務のためのコーパスを調査し、自然言語を用いた辞書管理システムに必要な機能を明らかにしようとした。

本稿ではＥＤＲ電子化辞書を管理するためのコーパスを要求別に分類することによって明らかになったシステムの機能をあげるとともにシステムを実現する際に生じる自然言語解析上の問題点とそれを解決するための方策を述べる。

2．例文の収集

実際に管理業務を行なっている２人の担当者に管理業務内容を自然言語で記述してもらうことによって例文の第１次収集を行なった。この第１次データの中には「形態素解析する」等のさらにプリミティヴな記述を必要とする語彙が含まれていた。そこで、これらの語彙を書き下す作業も行ない例文を増やした。

3．要求別分類

電子化辞書を管理するためのコーパスは属性値を検索するような逐次処理と体系的チェック、一括更新等のバッチ処理の２つに分けることが可能であることがわかった。

3．１逐次処理

1）見出し語に対する項目値の検索
- ・書くの品詞を教えて下さい。
- ・処理の概念見出しを検索して。

2）辞書の個別情報に対する更新、追加、削除
- ・処理の品詞にＪＮ１を追加して。
- ・読むで概念ＩＤが016ab0のレコードを削除。

3．２バッチ処理

1）項目間の整合性の検証
- ・品詞が動詞の見出し語を教えて下さい。
- ・動詞で左連接属性がＪＬＶ１の見出し語を検索して。
- ・品詞が動詞で左連接属性がＪＬＶ１で右連接属性がＪＲＶ２の見出し語を昇順にファイルＡに出力しなさい。

2）辞書情報に対する一括修正、追加、削除
- ・ファイルＡの内容を辞書に追加して。

3）辞書全体に対する数量調査
- ・概念見出しの平均数を教えて。
- ・最も登録数の多い品詞をあげて。

4）辞書自体に対する処理
- ・単語辞書のバックアップをとって下さい。

辞書管理システムとして上記にあげたような要求を満たす機能が必要とされる。

4．自然言語解析上の課題

自然言語を用いて辞書管理を行なうシステムを実現するうえで、自然言語解析の点で以下のような問題が起こると思われる。

1）読みの項目値の指定と用言の見出し語を指定する場合に解析上問題が起こる。

「読みがあいの見出し語を検索」

この場合ユーザの意図として下線部が読みの項目値である。したがって、「読み／が／あい／の／見出し語」と解析する必要がある。しかし、ユーザが入力する項目値を全て解析用辞書にいれることは難しい。このため、「あい」が未登録語

Natural Language Interface for Electronic Dictionary System Manager
- A classification of query corpus -

Mika KOBAYASHI, Kazunori MURAKI
Japan Electronic Dictionary Research Institute,Ltd
*Tsutomu HIYAMA
NEC Scientific Information System Development,Ltd

*本研究はEDR在任中に行なったものである。

となって処理されるか形態素解析が失敗する可能性がある。

「読むの品詞を検索」

ユーザの意図としては「読む」は項目値であり、辞書管理のためのコーパスとしては正しい構文である。しかし、この「読む」を動詞として解釈すると誤った構文となる（*書くの論文 -> 書く論文）。したがって、構文解析が失敗する可能性がある。

2） 検索文が長くなると解析精度が低くなる。

係り先の曖昧性が増え、構文解析が正しい解析結果を出力できない可能性がある。

5．入力形式の規約による課題の解決

ユーザが安心して使えるシステムを実現するには、システムの動作が安定していることが最も重要である。特に電子化辞書のような大量のデータを扱う時に途中でシステムがストップし、最初からやり直すようでは管理のためのコストが増大する。そこで、5であげたような課題を入力形式に簡単な規約を加えることによって解決する方法を提案する。

5．1 項目値を「」で括る入力

読みの項目値や用言の見出し語を「」で括って指定する。しかし、読みの項目値と用言の見出し語のみにこの規約を設定することは「」で括るための条件判断が必要となる。この判断をさけるためにこの規約を項目値の全てに対応させることにする。この程度の入力規約であれば自然言語の自然さが損なわれないと考える。

5．2 個別条件の束による入力

自然言語解析の技術では一文が長くなればなるほど、構文的曖昧性が増し解析精度は悪くなり、処理時間も長くなる傾向にある。このため一文はできるだけ短いほうがよい。長い検索文であっても項目名、項目値に対する条件、処理に対する条件に個別に分割することが可能であることが分析の結果明らかになった。

属性名（属性値）に対する条件の例

a）Nのみ 見出し語（を検索してください）
b）NのN「読む」の品詞
c）NがNの（である）N 品詞がJVEの見出し語
d）NでNをV＋N 動詞でJLV1を持つ見出し語
e）NとN 品詞と表層格情報
f）NかN 形容詞か形容動詞

処理に対する条件の例

a）出力形式に対する条件 昇順に出力
b）出力先の条件 ファイルAに出力

このように分割される各部分が条件の並列になってひとつの長い検索文になっている。そのため、分割した文を指示代名詞を用いて入力することによって、同じ検索結果を導き出すことが可能である。この個別条件の束をユーザの管理要求ととらえることができる。

一文入力の例）

「読みが「あい」で、左連接属性が「ＪＬＶ１」の見出し語と品詞を見出し語毎に出して下さい。」

個別条件の束による入力の例）

読みが「あい」である見出し語と品詞を教えて下さい。
それの左連接属性は「ＪＬＶ１」です。
それを見出し語毎に出して下さい。

このように単文レベルに分割すれば、一文内の構文の曖昧性がほとんど無くなる。したがって、十分に安定した解析結果が得られることが期待できる。

また、この個別条件の束で記述することによって、作業者自身にとっても一文で記述するより各条件を認識しやすいというメリットもある。

6．おわりに

今回、辞書管理業務に用いられるコーパスを分析した結果、入力形式に簡単な規約を加えることで自然言語を用いた辞書管理システムの実用性が高まることがわかった。そして、このような入力文を対象とする辞書管理業務のための辞書管理システムを試作を行なった。

最後に本研究の機会を与えて下さった横井所長、ならびに例文を作成していただいた森本裕子女史に深く感謝致します。

参考文献：

[1] 有岡 他「大規模英語単語辞書の開発」４５回情処全大

（株）日本電子化辞書研究所 TR-018 日本語単語辞書