構造化文書検索のための自然言語クエリから問い合わせ言語への変換

林 由紀子 *，松原 茂樹 (名古屋大学)

Query Transformation from Natural Language to Formal Language for Structured Document Retrieval

Yukiko Hayashi, Shigeki Matsubara (Nagoya University)

1 はじめに

近年，XML の普及にともない，構造化文書を対象とした情報検索が重要になりつつある．構造化文書が対象の情報検索では，XPath[2] などの形式的な問い合わせ言語が多く用いられる．しかし，問い合わせ言語の利用には専門的な知識を必要とし，一般の利用者が用いるには適さない．

本稿では，自然言語で記述されたクエリを問い合わせ言語に変換する手法を提案する．これにより，利用者は自然言語文を入力しさえすれば構造化文書の検索が可能となる．

2 問い合わせ言語 NEXI

本研究では，問い合わせ言語の 1 つである Narrowed Extended XPath I (NEXI)[3] を使用する．NEXI は，XPath[2] を簡略化した問い合わせ言語であり，検索すべき要素とそれに含まれる内容を指定できる．例えば，

```
//sec[about(.//p, database)]
```

は，結果として出力する要素に sec 要素を指定し，その絞り込み条件として sec 要素の子孫要素 p に database というキーワードを含むことを指定している．

3 自然言語クエリの変換

自然言語で記述されたクエリを等価な問い合わせ言語に変換する手順を以下に示す．なお，本研究では，入力は英語文であるとする．

1. **単語のタグ付け** 入力された自然言語の各単語に対し，特殊タグあるいは品詞タグを付与する．特殊タグとして 3 種類を設けた．表 1 に特殊タグ名とその意味，及び付与される単語の例を示す．特殊タグが付与されなかった単語には，品詞タグを付与する．
2. **パターンマッチングによる情報要求の抽出** タグの並びに関するパターンのマッチングによって，クエリ中に表現された情報要求を抽出し，中間言語で表現する．表 2 に中間言語とその意味を示す．パターンマッチングのテンプレートは，実データの分析に基づき人手で作成した．例えば，“RTN ELM” というパターンにマッチした場合 return(ELM) という中間言語を生成する，というテンプレートがある．テンプレートは合計 10 個作成した．
3. **問い合わせ言語の生成** 生成した複数の中間言語から，等価な問い合わせ言語を生成する．

上記の手順により，例えば “Find sections containing the paragraphs about databases. ” という自然言語クエリに対しては，まず図 1 のようにタグ付けする．次に，パターンマッチングにより，return(sec), include(sec,p), about(p,database) という中間言語で表現される情報要求を抽出する．それぞれの中間言語から`//sec`, `sec//p`, `p[about(.,database)]` という NEXI を生成し，出力する要素に sec 要素を指定するように結合すると，2 節に示した NEXI を生成できる．

Table 1: 特殊タグの意味と付与される単語の例

タグ名	意味	例
ELM	構造化文書の要素にあたる語	section, paragraph
BND	要素とそれが含む内容とのつなぎ語	about, containing
RTN	結果として出力する要素を指定する語	find, search

Table 2: 中間言語とその意味

中間言語	意味
about(X,A)	要素 X はキーワード A を含む
includes(X,Y)	要素 X は要素 Y を子孫に持つ
return(X)	X は出力要素である

Table 3: 評価実験の結果

評価基準	個数	割合 (%)
正解	26	55.3
不正解	21	44.7

```
Find/RTN sections/ELM containing/BND the/AT
paragraphs/ELM about/BND databases/NN
```

Fig. 1: 自然言語へのタグ付け

4 評価実験

本手法の評価のために，INEX[1] が提供している XML 文書検索のためのクエリセットを用いて，変換実験を行った．INEX のクエリでは，ある情報要求を自然言語 1〜2 文で記述した description と，問い合わせ言語 NEXI で記述した title がセットになっている．そこで本実験では，description を入力として NEXI 形式に変換し，生成した NEXI について，title を正解データとして評価を行った．結果を，正解（オリジナルの NEXI を完全に再現する，または description の構造を再現する）と，不正解（description の構造と一致しない，または NEXI を生成できない）の 2 種類で評価した．表 3 に，評価実験の結果を示す．正解率は 55.3%であり，今回のような単純な方法でも多くのクエリを正しく処理できた．

5 おわりに

本稿では，自然言語クエリを変換して問い合わせ言語を生成する手法を提案した．今後は，パターン数の増加やあいまい性解消の導入により，正解率の向上を目指す．

文献

(1) INEX http://inex.is.informatik.uni-duisburg.de/
(2) XPath http://www.w3.org/TR/xpath
(3) A. Trotman, et al.: INEX 2004 Workshop Proceedings, pp. 16-40, 2004.