상위어 자동추출 알고리즘 개발

Development of the Algorithm for the Automatic Extraction of Broad Term

최유미, 사공철*
만도기계(주) 중앙연구소, *숙명여자대학교 문헌정보학과
Choi Yu-Mi and Sakong Chul*
Mando Machinery Co. R&D Center, *Dept. of Lib. & Sci., Sookmyung Women's Univ.

문헌정보학분야의 용어사전을 이용한 자동시소러스 구축을 위한 첫단계로 「문헌정보학 용어사전」 MRD를 구성하고 이를 이용하여 상위어 자동 추출 알고리즘을 개발하였다. MRD구성시 전처리과정을 통하여 상위어 추출에 불필요한 정보가 수록되는 것을 방지하였다. 상위어 추출을 위한 알고리즘 개발은 무작위 표본추출을 통하여 「문헌정보학 용어사전」에 기술된 문장의 구문적 특성을 분석한후, 이 구문정보를 이용하여 수행하였다. 본 연구에서 제시된 알고리즘의 효율성 평가결과 89.4%의 정확도를 보였다.

1. 서론

학문 주제분야의 세분화 및 학제적 상호연관 현상으로 인하여 특정적이면서도 복잡한 양상을 띄는 이용자의 정보요구 및 학문분야의 용어의 변화를 수용할 수 있는 자동시소러스의 구축에 대한 연구가 활발하게 진행되어 왔다. 또한, 구미에서는 전자출판 및 정보기술의 발달로 상용 기계가독형사전(Machine Readable Dictionary : MRD)의 개발과 이를 이용한 자연언어의 어휘관계를 규명하려는 연구가 수행되어 왔다.

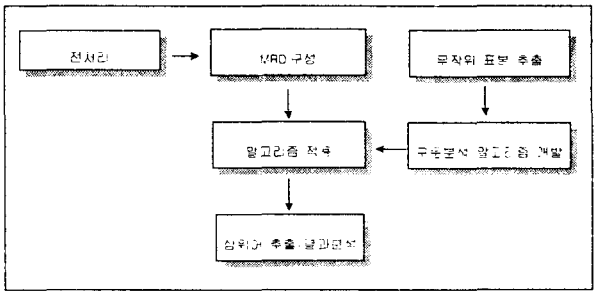
국내에서는 상용이 아닌 연구를 목적으로 국어사전 MRD를 이용하여 한국어 명사의 상위어를 추출하는 연구가 수행되었다(문, 김, 1994, 김 등 1995).

그러나, 특정학문분야의 용어사전은 비교적 간결체로 기술된 국어사전과는 달리 교과서와 같은 해설적인 문장구조를 갖고 있으며, 기술방식 역시 일관적인 형식으로 이루어지지 않았기 때문에 해당 용어사전의 문장기술방식에 맞는 구문정보의 수집 및 이를 통하여 어휘관계가 추출될 수 있도록 하여야만 한다.

따라서, 문헌정보학 분야의 MRD를 이용한 자동시소러스 구축을 위한 첫 단계로 「문헌정보학 용어사전」의 MRD 구성 및 표목에 대한 상위어 자동 추출에 관한 연구를 수행하였다.

2. 본론

상위어 자동추출 알고리즘을 개발하기 위해서는 우선 「문헌정보학 용어사전」에 기술된 문장의 구문적 특성을 파악하여야 하므로, 무작위 표본 추출을 하여 구문적 특성을 조사분석하였다. 이러한 과정에서 얻은 구문정보를 이용하여 상위어 추출을 위한 10개의 알고리즘을 개발한 후, 현대 AXIL311워크스테이션상에서 C언어로 구현하였다 <그림 1>.



<그림1>상위어 추출 알고리즘 개발 흐름도

2.1. MRD의 구성

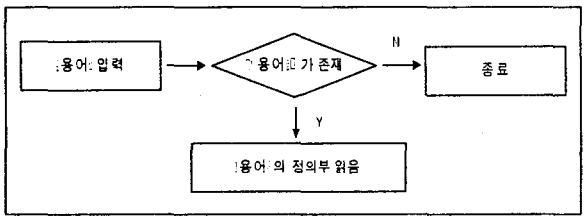
적합한 상위어를 추출하기 위한 구문분석의 소스로 사용되는 MRD는 상위어 추출 알고리즘 개발에 있어 중요한 요소라 할수 있다.
본 연구에서는 사용한 「문헌정보학 용어사전」MRD는 전자출판용 파일을 텍스트 형태로 변환한 것으로 그 구성형식은 다음과 같다.

```
1표목2 [한자표기] 혹은 영자표기
3 정의. 부연설명.

* 기호설명
1:표목의 시작 2:표목의 종료/한영자표기시작
3 : 정의 및 부연설명부의 시작.
```

<그림 2> 파일 구성 형식

본 연구에서는 전자출판용 파일을 텍스트파일로 변환하는 과정에서 자동 생성된 '1', '2', '3'이라는 숫자를 MRD의 각 필드를 구분하는 식별기호로 사용하였다. 컴퓨터를 사용하여 구축된 MRD에서 용어를 인식하는 과정은 <그림 3>과 같다



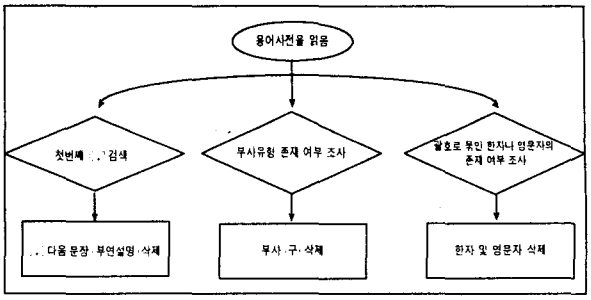
<그림 3> MRD 인식 과정

MRD구축시 MRD의 크기를 줄이고 또한 상위어 추출시 모호성을 최소화하기 위하여 문장에서 중요한 역할을 하지 않는 요소는 전처리 과정을 통하여 <표1>과 같이 제외시켰다.

<표1> MRD구성의 전처리과정

부연설명	상위어는 부연설명이나 용례에서보다는 표목에 대한 정의부에서 추출되므로 표목과 정의부 이외의 부연설명삭제
부사	부사가 상위어로 추출되는 것을 방지키 위해 MRD 정의부에 포함된 부사제거.
한자 영자	표목과 정의부에 수록된 한자 및 영자는 부정확한 상위어 추출원인이 될 수 있으므로 삭제

<그림 4>는 MRD 구성을 위한 전처리 과정을 도식화한 것이다.



<그림 4> MRD구성의 처리 과정

2.2. 상위어 추출 알고리즘 개발

「문헌정보학 용어사전」에 적합한 상위어 자동추출 알고리즘 개발을 위하여 무작위 표본추출을 통하여 분석된 구문정보를 토대로 10개의 구문 유형별 상위어 추출 알고리즘을 개발하였다.

▶유형 1 : 구별어 + 중심어
한국어의 특성상 문장의 중심이 되는 어구(중심어)는 일반적으로 구별어로 한정되어 문장의 뒷부분에 나타난다(예1).

예1)무간기본
발행년 등의 표시가 없는 책.

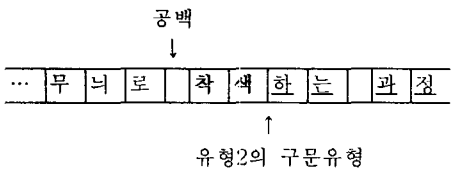
이때 주의할 점으로, 예2)에서와 같이 어떠한 구별어로도 수식되지 않는 명사형 용어만으로 정의문이 구성되었다면 이는 상위어가 아닌 동의어이므로 이러한 경우에는 추출하지 않는다.

예2)간찰 (簡札)
서간.

▶유형 2 : 중심어 + '하는 것.'
명사형 용어에 동작의 의미를 첨가한 단어로서 그 동작을 행하고 난 후의 상태나 또는 그러한 행위를 나타내는 유형이 문장의 마지막에 나타날 때 이러한 유형 앞의 용어를 상위어로 추출한다(예3).

예3)간략분류
분류표에서 … 상위의 주제로 **분류하는 것.**

본 유형을 알고리즘으로 구현하는데 있어 적용한 원리는 <그림 5>와 같다.



<그림 5> 규칙2의 알고리즘 적용 원리

▶유형3 : 중심어 + 자격을 나타내는 조사

'어떤 지위나 신분이나 자격을 가지고'의 뜻으로 쓰이는 부사격 조사의 선행 용어가 중심어가 되는 구조이다(예4)

예4)간격문자 space character
인쇄되지 않은 도형문자**로서** ….

▶유형4 : 수단·방법의 부사격조사 처리

수단·방법·이유 등을 뜻하는 부사격 조사와 알고리즘 3의 부사격 조사의 형태가 동일하므로, 문장에서 부사격 조사의 성격을 구분할 수 있는 구문정보가 필요하다. 표본분석결과, 일반적으로 이용 및 사용의 의미를 지닌 동사와 병행하여 나타나는 부사격 조사는 대부분 수단·방법·이유 등을 뜻하므로 부사격 조사 다음의 동사의 유형에 따라 상위어 추출여부를 결정하도록 하였다.

▶유형5 : 중심어+목적격 조사+지칭형 동사

"~을(를) + 지칭형 동사"와 같이 어떤 사물을 지칭하는 의미를 지닌 동사가 취하는 목적어는 문장의 중심어이므로 이를 상위어로 추출한다(예5).

예5) 감광유제
사진 필름이나 종이에 입히는 화학약품**을 뜻함**.

▶유형6 : 중심어+속관계를 나타내는 구문

표목이 어떤 개념의 일부분임을 나타나는 구문유형이 정의부에 존재할 때, 이 유형에 선행되는 중심어를 상위어로 추출한다(예6)

예6) 컴퓨터과학 computer science
정보과학**의 한 분야**. …

▶유형7 : 중심어+서술형 어미

서술형 어미 "~이다" 앞의 용어를 상위어로 추출한다(예7)

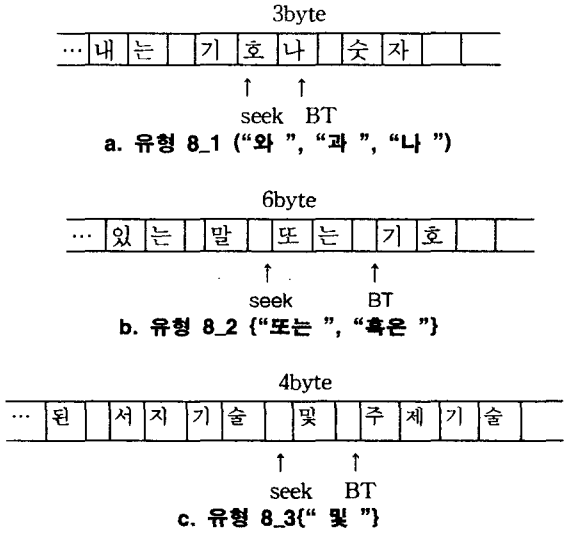
예7) 마이크로스태트
프리즘을 장치한 … 만든 음화마이크로사진**이다**.

▶유형8 : 용어의 대등 연결

등위연결조사나 등위연결 부사로 대등하게 연결된 두 중심어를 상위어로 추출한다(예8).

예8) 구분문자
데이터 항목을 … 문자**나** 기호.

<그림 6>은 대등연결조사의 유형에 따라 상위어가 추출되는 과정을 도식화한 것이다.



* seek : patt8을가리키는 pointer
BT : 상위어를 가리키는 pointer

<그림 6> 유형8의 상위어 추출 원리

▶유형 9 : 표목 유형별 상위어 추출

본 알고리즘은 「문헌정보학 용어사전」에서 나타나는 기술방식의 특성에 따라 상위어 추출과 관련 없는 표목에 대한 처리 및 표목자체가 상위어로 추출되는 경우를 방지하기 위한 것이다. 예9)와 같이 "~의 해", "~회의", "~협약", "~주간" 등으로 끝나는 표목은 상위어추출 대상에서 제외시켰다. 또한, 예10)과 같이 표목자체가 상위어로 추출될 경우에는 이를 상위어로 추출되지 않도록 하였다.

예9)세계도서의 해 international book year
현대사회에 있어서 … 세계도서의 해였다.

예10)본문 body matter
인쇄시 서문이나 삽화와 구별되는 본문.

▶유형10 : 불용어 처리
본 연구는 일반 단어가 특정학문분야인 문헌정보학 분야의 용어에 대한 상위어를 추출하고자 하는 것이므로, 상위어로서 추출되는 용어의 특정성이 결여된 경우에는 상위어로 추출되지 않도록 불용어로 간주하였다.

▶ 각 알고리즘의 조합
임의의 표목에 적용될 수 있는 상위어 추출 알고리즘은 한 문장에서 복합적으로 나타날 가능성이 있다. 따라서, 다음의 논리식과 같이 임의의 용어에 10가지 알고리즘을 모두 대등하게 적용시켜 상위어가 추출되도록 하였다.

상위어 추출 알고리즘 = <9>∧(<1>∨<2>∨(<3>∧<4>)∨<5>∨<6>∨<7>∨<8>)∧<10>

3. 알고리즘 성능평가

본 알고리즘의 효율성을 입증하기 위하여 두 가지 방식의 검증작업을 수행하였다. 우선, 「문헌정보학 용어사전」의 임의의 100개의 연속된 표목을 대상으로 본 연구에서 제안된 알고리즘(A1)과 국어사전 MRD를 대상으로 하였던 선행연구(김 등, 1995)의 알고리즘(A2)을 비교하여 상대적인 효율성을 비교해보았다. A2의 수행결과 126개의 상위어가 추출되었으며 이중 22%는 A1의 불용어 리스트에 포함된 용어였다. 또한 서술형 동사가 상위어로 추출된 경우도 10%를 차지하여 전체적으로 적합한 상위어는 44%가 추출되었다.

이러한 결과로서 선행연구가 범용 국어사전에 수록된 보편적이고 일반적인 단어 및 간결하고 정형화된 기술방식을 대상으로 알고리즘을 개발하였으므로 이러한 방식은 「문헌정보학 용어사전」에는 적합하지 않음을 알 수 있다.

두 번째 방식으로 「문헌정보학 용어사전」의 연속적인 표목 200개를 대상으로 수작업으로 추출한 상위어와 본 알고리즘을 통하여 추출된 상위어의 비교를 통하여 상위어의 정확도를 비교하였다. 상위어 추출 결과, 134개의 표목에서 149개의 상위어가 추출되었다. 이중 적합한 상위어는 81.2%였고, 개념이 부분적으로 적합한 상위어는 7.4%, 부적합한 상위어는 11.4%로 추출되었다. 또한, 상위어가 추출되지 않은 표목 66개 중에서도 74.2%의 표목이 유형 4, 9, 10에 의하여 부적절한 상위어 추출이 방지되었다. 따라서, 상위어 추출 알고리즘의 전반적인 정확도는 89.4%로 비교적 높은 결과를 보인다고 할 수 있다.

4. 결론

본 논문에서는 문헌정보학 분야에서의 기계가독형사전을 이용한 자동시소러스 구축을 위한 첫 단계로 「문헌정보학 용어사전」을 이용한 MRD 구성 및 표목에 대한 상위어 자동 추출에 관하여 연구하였다. 무작위 표본 추출을 통해 「문헌정보학 용어사전」에 기술된 문장의 구문적 특성을 조사 분석하였고, 표본조사를 통하여 얻은 구문 정보를 일반화하여 상위어 추출을 위한 10개의 알고리즘을 제시하였다. 알고리즘의 성능 평가 결과 89.4%의 정확도를 보였다. 향후, 관련어 및 동의어 등의 어휘관계를 추출할 수 있는 추가적인 연구가 이루어지면 한국어 시소러스의 자동 구축에 기여할 수 있으리라 생각된다.

참고문헌

1. 김민수, 김태연, 노봉남. 1995. "국어사전을 이용한 한국어 명사에 대한 상위어 자동추출 및 WorNet의 프로토타입 개발." 「정보처리논문지」 2(6):847-856.
2. 문유진, 김영택. 1994. "한국어 명사의 Hypernym 자동추출 방법." 「한국정보과학회 학술발표논문집」 21(2):613-616.
3. 박병수,1989. "기계번역에서 본 한국어의 특징", 「정보과학회지」, 7(6):31-39.
4. 사공철 등. 1996, 「문헌정보학 용어사전」 서울:한국도서관협회.
5. 이병모, 1995. 「의존명사의 형태론적 연구」 서울:학문사.
6. 이상조,1989."기계번역 시스템을 위한 사전 구상." 「정보과학회지」 7(6):25-30.
7. 장석진. 1993. 「정보기반 한국어문법」 서울:언어와 정보.
8. Amsler, Robert A. 1984. "Machine-Readable Dictionaries." *Annual Review of Information Science and Technology* 19:161-209.