94-g-2

자동색인어를 이용한 전거파일의 구축

Construction of the Authority Files Using Automatically Indexed Terms

한 영 균
울산대학교 인문대학 국어국문학과

Han, Young-Gyun
Department of Korean Language & Literature, University of Ulsan

요 약

본 연구는 자동색인시스템의 색인기능을 통해서 추출된 색인어를 이용해서 검색시스템에서 요구되는 전거파일을 구축하는 작업의 효용성을 확인하기 위한 시험적 연구의 결과를 정리한 것이다. 언론연구원의 KINDS 서비스 시스템의 신문기사 데이터베이스에서 색인시스템을 통해 추출된 약 80만개의 색인어를 기본자료로 삼아 색인어를 하위분류하고, 그것을 이용한 전거파일 구축의 가능성을 타진해 본 것이다.

1. 서론

1.1. 전거파일은 동의어파일과 함께 정보검색시스템의 검색효율을 제고하는 데에 중요한 역할을 하며, 검색의 재현율을 높이기 위해서는 반드시 필요하다. 특히 본문검색을 위주로 하면서 다양한 계층, 다양한 요구를 가진 사용자가 이용하는 신문기사정보서비스시스템이나 생활정보시스템의 경우 그 중요성은 더욱 커진다. 이러한 전거파일의 구축에는 ① 동어이표기(원어 및 한자표기 포함) ② 약어 ③ 동음이의어 등의 목록작성이 선행되어야 하는데, 시소러스와는 달리 고유명사까지 포괄해야 하는 것이어서 수작업을 통해서 인인이 자료를 수집하는 것은 쉽지 않은 일이다. 지금까지 개발된 여러가지 형태의 문헌정보서비스시스템들이 적지 않지만, 전거파일을 지원하는 시스템은 찾아보기 어려운 것도 상대적으로 작업의 양이 방대한 데에 연유한다고 할 수 있을 것이다. 본 연구는 자동색인시스템을 통해서 추출된 색인어를 이용해서 전거파일구축에서의 용어수집문제를 처리함으로써, 단시일내에 효율적으로 전거파일을 구축하기 위한 연구의 결과 일부를 정리한 것이다.

1.2. 본 연구에서 이용한 자료는 한국언론연구원의 KINDS시스템에서 제공되는 것이다. KINDS시스템은 현재 국내에 공개되

이 있는 ON-LINE 문헌정보시스템 중에서 최대 규모의 것으로 데이콤의 천리안시스템을 통해서 일반인도 활용할 수 있으며, 색인어는 자모순으로 확인할 수 있다. 본 연구에 이용된 색인어는 용어별 빈도와 함께 제공된 것인데, 색인어의 수가 약 80만, 누적빈도는 근 5,000만에 달해서 색인어의 추출대상이 된 기사본문의 규모는 6,000만 어절 가까운 것으로 추정할 수 있다.[1] 신문기사정보서비스에 요구되는 명사류는 거의 빠진 것이 없는 것으로 보아도 좋을 것이다.[2]

2. 색인어의 분류 및 분석 결과

2.1. 자료의 선별을 위해서 1차적으로 약 80만개의 색인어를 고유명사와 보통명사로 분류하는 작업을 진행하였는데, 그 분류의 결과는 다음과 같다.[3]

보통명사 : 약 550,000어 (69.2%)
고유명사 : 약 149,500어 (18.7%)
분석불가 : 약 26,000어 (3.3%)
색인오류 : 약 69,000어 (8.7%)

분석이 불가능한 예들은 전체의 3.3%를 차지하는데, 한글로 된 것 중에서 의미를 파악하기 어려우면서 사전에도 등재되지 않은 예들과 영어 약어중에서 그 원어를 파악하기 어려운 것들이 많다. 그런데 이들 대부분은 그 빈도수가 1~2회에 불과한 것으로 그 비중이 상대적으로 낮은 것이라고 할 수 있다. 색인오류이는 전체의 8.7%를 차지하는데, 대부분 복합명사나 동사의 활용형을 잘못 절단한 것들이다.[4]

2.2. 2차분류는 고유명사를 그 특성에 따라서 (1) 인명 (2) 지명 (3) 회사명 (4) 기관/단체명 (5) 상품명 (6) 건축·시설물 (7) 작품명 (8) 사건/대회명 (9) 법률/조약 (10) 부족/종족명의 열 개 하위영역으로 나누었는데, 그 분석 결과는 다음과 같다.

(1) 인명 : 약 62,000어
(2) 지명 : 약 21,000어
(3) 회사명 : 약 18,000어
(4) 기관/단체명 : 약 27,000어
(5) 상품명 : 약 2,800어
(6) 건축/시설물명 : 약 9,500어
(7) 작품명 : 약 4,400어
(8) 사건/사고명 : 약 2,200어
(9) 법률/조약/회의명 약 2,300어
(10) 종족/부족명 : 약 300어

2.3. KINDS시스템에서의 자동색인이 중 고유명사를 대상으로 한 분석의 결과는 자동색인이를 이용해서 전거파일 구축에 사용될 자료를 축적하는 작업이 상당히 효율적임을 보여준다고 할 수 있다.

우선 양적인 면에서 국내에서 가장 충실한 내용을 담고 있는 것으로 알려진 중앙일보 시소러스DB와 비교할 때 크게 뒤지지

1) 자동색인시스템은 기본적으로 명사류만을 추출한다. 그런데 한국어 문어 텍스트에서 명사류는 대체로 전체의 90% 정도를 차지한다. 6,000만 어절이라는 숫자는 이를 바탕으로 산출한 것인데, 원고지 1매당 50어절이 들어간다고 보면 보통 사람이 시간당 원고지 100매씩 하루에 8시간을 읽을 때 10명이 150일이 소요되는 분량이다.

2) 이는 후술할 중앙일보시소러스DB와의 비교를 통해서도 어느 정도 확인할 수 있다.

3) 분류 및 분류 결과의 입력에는 약 25MM가 투입되었다.

4) 이들은 자동색인시스템의 성능보강을 위한 기초자료로 활용될 수 있다.

않는다는 점을 지적할 수 있다.[5]

본 연구에서 얻어진 고유명사의 총수는 149,500개이다. 그런데 중앙일보 시소러스DB에서 이들 고유명사를 포괄하는 보조키워드파일·기관/단체명파일·회사명파일·인명파일의 총항목수는 228,098개이다. 양자의 차이가 약 78,500 항목에 달하는 것이다. 중요한 것은 이러한 차이가 대체로 중앙일보 시소러스DB의 인명파일 및 회사명 파일의 항목수와 본 연구 결과로 얻어진 자료의 인명항목수 및 회사 항목수의 합계와의 차이에서 기인한다는 점이다. 회사명 파일의 항목수는 중앙일보DB의 경우 총 74,575개인데 본 연구에서의 분석 결과는 18,000개에 불과해서 약 56,600 항목이 적으며, 인명파일의 경우 중앙일보 시소러스DB가 86,401 항목인데 비해서 본 연구 결과 얻어진 것은 약 62,000 항목으로 24,000항목 정도의 차이가 있는 것이다. 이들 두 항목의 차이를 합하면 약 80,000 항목이 되는바, 이는 앞에서 본 78,500이라는 고유명사 항목수의 차이를 넘는 것이다.

이런 관점에서 특히 주목되는 것은 보조키워드파일과 기관/단체명 파일의 규모다. 중앙일보 시소러스DB에서는 국내지명 중 동명/읍면리명, 외국지명, 작품명, 상품명, 사건명, 유명건축물명 등을 보조키워드파일에 담고 있는데, 그 총 항목수는 디스크립터 13,638개 비디스크립터 20,426개로 34,108개이다. 그런데 본 연구의 분석 결과 얻어진 지명, 상품명, 작품명, 사건/사고명, 건축물명을 합하면 약 4만개에 달하는 바, 국내 지명 및 외국 지명이 중앙일보의 시소러스DB에서는 시소러스와 보조키워드파일에 나뉘어 있음을 감안하면 비슷한 규모라고 할 수 있는 것이다. 기관/단체명의 경우도 중앙일보 시소러스DB가 33,014개이고 본 연구에서의 분석 결과는 약 27,000개로 그리 큰 차이를 보이지 않는다. 이러한 결과는 본 연구에서 기초자료로 택한 신문기사 자동색인어가 대체로 신문기사 검색에서 요구되는 용어를 충실하게 담고 있음을 의미하는 것으로 해석해도 무리가 없음을 의미하는 것으로 판단된다.

대표어의 정확한 총수를 추출하지 못한 상태여서 단정하기는 어렵지만, 질적인 면에서도 본 연구결과 얻어진 자료는 중앙일보 시소러스 DB의 자료에 비해서 크게 뒤떨어지지 않을 것으로 판단된다. 우선 시소러스파일 못지않게 중요한 역할을 하는 보조키워드 파일 및 시소러스 보조파일을 실제 사용된 어형을 중심으로 구성할 수 있어서 효율적인 시스템구축이 가능하다고 할 수 있으며, 실제 문헌에 나타난 동형이표기의 애들이나 약어를 충실하게 반영한 전거파일의 구축이 가능하기 때문이다.

3. 마무리

본 연구는 완성된 결과를 제시한 것이기보다는 효율적 전거파일 구축을 위한 방법론에 대한 검토 결과를 보고한 것이다. 연구개발의 주안점은 신문기사DB의 검색시스템 개발을 위한 전거파일의 개발에 있는바, 여기서는 색인어의 분석 과정에서 도출된 문제점을 중심으로 자동색인어를 이용한 전거파일의 구축에 앞서 처리되어야 사항 몇 가지를 기관/단체명의 경우를 중심으로 정

5) 중앙일보 시소러스DB의 경우는 (1) 시소러스 (2) 보조키워드 (3) 인명 (4) 회사명 (5) 기관단체명 파일의 다섯으로 구성되어 있어서 본 연구에서의 분류와 일치하지는 않지만[6], 구체적인 내용을 참조하면 대비가 가능하다.

리하는 것으로 결론에 대신하고자 한다.

1) 정식명칭 및 원어의 확보

본 연구에서 사용한 자료는 간략성을 중시하는 신문기사의 속성을 그대로 반영하고 있다. 따라서 완전한 명칭보다는 약어 및 통상어가 중심이 된다. 따라서 보다 충실한 전거파일의 구축을 위해서는 이를 극복한 장치가 마련되어야 한다.

2) 기관의 하위부서명의 처리

이는 일반명사의 복합어 처리기준과 관련된 것인데, 기관명 전거파일의 표제항을 어디까지로 할 것인가 하는 문제로 집약된다. 단적으로 대학의 학과명과 대학명이 한 어절로 색인된 경우, 이를 따로떼어서 처리한 것인가 아니면 하나의 단위로 다룰 것인가가 결정되어야 하며, 후자의 경우 upword -posting의 사용 여부도 결정되어야 한다. 이런 유형으로 "민자당강남을구지구당, 광주YMCA" 등 정당·기관의 지부명, "서울경찰청강력과, 부산지검수사과" 등과 같은 기관의 하위부서명 등이 있다.

3) 한시적 단체명의 처리

신문기사의 속성상, 색인어에는 정치·종교·학술 모임 등의 계보 등 한시적 성격의 단체명이 다수 등장한다. 그러나 시스템의 규모를 어느 정도로 한정하고자 하는 한, 이들 모두를 전거파일의 표제어로 삼을 수는 없다. 따라서 이들 한시적 단체의 성격을 나누고, 그를 처리할 기준이 설정되어야 한다. 이러한 유형의 것으로는 "감시단, 고문단, 교섭단, 기자단, 기획단, 대표단, 발굴단, 발굴조사단, 방문단, 변호인단, 봉사단, 비리조사단, 사찰단, 선수단, 수사단, 시찰단, 역학조사단, 연수단, 유적지조사단, 응원단, 조사단, 지원단, 진상조사단, 학술조사단, 후원단" 등이 집미되는 용어들이 대표적이다.

참고문헌

[1] Boguraev, Bran and Ted Briscoe ed. (1989a) *Computational Lexicography for Natural Language Processing,* Longman.

[2] Boguraev, Bran and Ted Briscoe (1989b) 'Meaning and Structure in Dictionary Definitions' in Boguraev, Bran and Ted Briscoe (1989a)

[3] Boguraev, Branimir and Beth Levin (1993) 'Models for Lexical Knowledge Bases' in *Semantics and the Lexicon,* edited by Pustejovsky, Kluwer Academmic Publshers.

[4] Calzolari, Nicoletta (1982) 'The dictionary and the thesaurus can be combined' in Martha Walton Evens eds. (1982a)

[5] Evens, Martha Walton eds. (1982a) *Relational Models of the Lexicon,* Cambridge

[6] 체신부(1994), ≪제조업 경쟁력 강화사업 - 시소러스 사전구축 보고서≫