김일성종합대학학보(어문학)
주체106(2017)년 제63권 제4호

문헌자동분류에서 미등록어의 류형과 처리방법

최 영 희

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《교원, 연구사, 학생들이 책을 빨리 찾아보도록 하자면 책분류사업을 잘하여야 합니다. 책을 그 내용에 따라 정확히 분류해놓아야 독자들이 요구하는 책을 쉽게 찾아볼수 있습니다. 책을 분류하는것은 과학적내용에 따라 구분하는 사업인것만큼 쉬운 일이 아닙니다.》(《**김정일**전집》 제7권 279페지)

오늘날 지식경제시대의 요구에 맞게 전통도서관들을 전자도서관화하는 사업에서 문헌자동분류사업이 차지하는 몫은 자못 크다.

문헌자동분류에서 미등록어처리는 중요한 문제의 하나로 제기된다. 그것은 미등록어처리정도가 문헌자동분류의 완비정도를 평가하는 중요한 지표의 하나로 되기때문이다.

일반적으로 자연언어처리에서 미등록어는 사전의 완비정도에 따라 개수가 다르게 결정되지만 미등록어가 생기지 않는 경우란 없다.

미등록어의 발생은 문헌자동분류를 위한 조선어처리에서도 례외로 되지 않는다.

문헌자동분류에서 말하는 미등록어는 추출, 선정된 주제어들을 분류용어사전과 비교할때 분류기호가 얻어지지 않는 단어를 말한다.

미등록어처리를 성과적으로 하자면 미등록어들이 생기는 원인과 그 류형에 대하여 알아야 한다.

문헌자동분류에서 미등록어가 생기는 원인은 두가지로 볼수 있다.

그것은 첫째로, 새로운 단어들이 출현하는데 맞게 분류용어사전의 갱신보충이 따라서지 못하는데 있다.

시대의 발전은 새로운 단어의 출현을 동반한다.

그것은 과학과 기술, 경제와 문화, 사회제도와 사람들의 사상의식 등의 끊임없는 발전이 그 반영인 새 단어를 절실히 요구하기때문이다.

시대의 발전과 함께 새롭게 출현하는 단어들을 분류용어사전에 모두 수록한다는것은 쉬운 일이 아니다.

바로 이것이 문헌자동분류를 위한 언어처리에서 미등록어가 생기는 중요한 원인의 하나이다.

미등록어가 생기는 원인은 둘째로, 언어생활에서 같은 의미를 가진 단어에 대한 표기불일치와도 관련된다.

일반적으로 언어생활에서 사람들은 주제사상적내용의 심오성과 론리성, 과학성에 힘을 넣으면서 자기의 개성적인 문체로 글을 쓴다. 이로부터 같은 의미를 가진 단어라고 해도 집필자에 따라 표기가 서로 다른 경우가 있다.

그것을 사전적표기와 집필자의 표기에서 보기로 하자.

사전적표기와 집필자의 표기

사전적표기	집필자의 표기
○ 모음조화의 변화	○ 모음조화변화
○ 나이별심리차이	○ 나이에 따르는 심리적차이
○ 리론교육과 실천교육 결합방법	○ 리론교육과 실천교육을 결합하는 방법
○ 문학예술에 대한 당의 령도	○ 문학예술에 관한 당의 령도

사전적표기에서의 《모음조화의 변화》는 집필자의 표기에서 속격토 《의》를 빼버린 합성어 《모음조화변화》로, 사전적표기에서의 《나이별심리차이》는 집필자의 표기에서 불완전명사 《별》대신에 여격토 《에》와 동사규정형 《따르는》과 불완전명사 《적》이 삽입된 《나이에 따르는 심리적차이》로 되였다. 그리고 사전적표기에서의 《리론교육과 실천교육 결합방법》은 집필자의 표기에서 대격토 《을》과 《하다》형 동사규정형이 삽입된 《리론교육과 실천교육을 결합하는 방법》으로 되였으며 사전적표기에서의 《문학예술에 대한 당의 령도》는 집필자의 표기에서 보조동사규정형 《대한》대신에 역시 보조동사규정형인 《관한》이 교체된 《문학예술에 관한 당의 령도》로 되였다.

이처럼 사전적표기와 집필자의 표기는 의미상의 견지에서 서로 같다고 할수 있지만 콤퓨터에서는 인공지능까지 도입하여 사전적표기와 집필자의 표기에 대한 의미분석을 하지 않는 이상 서로 다른 단어나 단어결합으로 인식한다.

이것은 바로 문헌자동분류를 위한 조선어정보처리에서 미등록어가 생기는 또 하나의 중요한 원인으로 된다.

미등록어들은 대상본문에서 추출, 선정된 단어들로서 집필자가 문헌의 주제를 반영하는데 필요한 단어로 인정하고 사용하는것만큼 문헌분류에 영향을 주는 단어이다.

대상문헌에 대한 분류기호는 주제어들을 분류용어사전과 비교하여 얻기때문에 미등록어에 대한 처리를 바로하지 않는다면 문헌자동분류의 정확도를 떨구는 결과를 가져온다.

미등록어가운데는 표기형태는 다르지만 의미적으로 사전적단어와 같은 단어도 있을수 있고 해당 문헌의 주제를 표현하는데서 중요한 역할을 하는 새로운 단어도 있을수 있다. 그러므로 미등록어들을 무시할것이 아니라 그에 대응하는 분류기호를 얻을수 있도록 해당한 처리를 하여야 한다. 다시말하여 사전적단어와 의미는 같지만 외적형태가 다른 단어들은 분해 및 합성기술을 리용하여 사전적단어를 얻어내야 하고 주제적가치가 있는 새로운 단어들은 보충하여야 하며 그렇지 못한 단어들은 삭제하는 작업을 하여야 한다. 그래야 문헌자동분류의 정확도를 높일수 있다.

문헌자동분류과정에 생기는 미등록어는 무엇을 기준으로 하는가에 따라 여러가지 류형으로 구분할수 있다.

우선 사전적표기와의 일치성정도에 따라 완전형미등록어와 부분형미등록어로 나눌수 있다.

완전형미등록어는 미등록어를 분해하였을 때 그 어떤 형태부도 사전적표기와 전혀 일치하지 않는 미등록어를 말한다.

완전형미등록어에는 주로 시대와 과학기술의 발전으로 새롭게 출현한 단어들이 속한다.

○ 선군

○ 불보라

○《광명성－3》호 2호기

부분형미등록어는 미등록어를 분해하였을 때 일부 형태부가 사전적표기와 일치하지 않는 미등록어를 말한다.

부분형미등록어는 다시 접두형미등록어, 접미형미등록어, 접두접미형미등록어로 구분할수 있다.

접두형미등록어는 미등록어에서 앞부분의 일부 형태부가 사전적표기와 일치하지 않는 미등록어를 말한다.

○ 고추: 풋고추, 절임고추, 생고추

○ 눈길: 생눈길, 찬눈길

○ 관계: 동등관계, 반순서관계

접미형미등록어는 미등록어에서 뒤부분의 일부 형태부가 사전적표기와 일치하지 않는 미등록어를 말한다.

○ 단어: 단어문장, 단어구분, 단어구조화, 단어사전

○ 사업: 사업방법, 사업부담, 사업조직

○ 문서관리: 문서관리기구, 문서관리방법

접두접미형미등록어는 미등록어에서 앞뒤의 형태부가 사전적표기와 일치하지 않는 미등록어를 말한다.

○ 교양기사: 혁명적교양기사집필, 계급교양기사집필

○ 인식방법: 과학적인식방법론

○ 응용과학: 현대응용과학적방법

미등록어는 또한 언어의 어휘론적형태에 따라 단순어와 합성어, 단어결합으로 된 미등록어로 나누기도 한다.

단순어로 된 미등록어는 한개의 말뿌리로 이루어진 단어로서《도서》,《혁명》,《언어》등을 실례로 들수 있다.

언어학적으로 합성어로 된 미등록어는 두개이상의 말뿌리들이 합쳐져서 쓰이는 과정에 공고한 하나의 단위로 굳어진 단어이고 단어결합으로 된 미등록어는 문장의 구성재료로서 두개이상의 단어들의 의미-문법적인 맞물림이다.

편의상 두개이상의 단어들을 공백이 없이 붙여쓴것은 합성어로 된 미등록어범주에, 띄여쓴것은 단어결합으로 된 미등록어의 범주에 넣고 취급하기로 한다.

○ 합성어로 된 미등록어: 선물관, 주체사상탑, 너도밤나무, 된장

○ 단어결합으로 된 미등록어: 작품의 기본주제, 대중운동을 통한 교양

일반적으로 문헌자동분류에서 리용되는 단어결합형식의 주제어들은 많지 않다.

단어결합형식의 주제어들은 본문에 대한 예비처리를 통하여 확정된 형태단어모임에 단어결합형식의 사전적표기가 있는가를 판정하는 방법으로 추출한다.

단어결합형식의 주제어들을 추출한 다음 나머지형태단어를 단위로 하여 단순어와 합성어로 된 주제어들을 추출한다. 그러므로 문헌자동분류과정에 생기는 미등록어들은 사전적표기와의 일치성정도에 따라 구분된 단순어로 된 미등록어와 합성어로 된 미등록어들이다.

문헌자동분류에서 미등록어는 그것을 이미 구축된 분류용어사전과 비교하여 새로 등록하거나 삭제하며 의미가 같거나 류사한 사전적단어로 변환하는 방법으로 처리한다.

편리상 미등록어모임을 RB, 분류용어사전의 단어모임을 M 이라고 하면 다음과 같이 표시할수 있다.

$$RB = \{RB_i | 1 \le i \le n\}$$
$$M = \{M_j | 1 \le j \le q\}$$

식에서 n은 미등록어모임에서 미등록어들의 개수, q는 분류용어사전의 단어모임의 단어개수이다.

형태단어는 형태부들의 모임이기때문에 어떤 형태단어 즉 i번째 형태단어는 구체적으로 다음과 같이 표시할수 있다.

$$RB_i = \{RB_i^k | 1 \le k \le m\}$$

식에서 RB_i^k는 i번째 형태단어의 k번째 형태부를 의미하며 m은 형태부의 개수이다.

그러면 류형에 따르는 미등록어처리방법을 구체적으로 보기로 하자.

완전형미등록어는 말그대로 해당 주제분야에서 새롭게 출현한 단어라고 볼수 있다.

그러므로 완전형형태의 미등록어모임과 분류용어사전의 단어모임사이에는 배반관계가 존재한다.

완전형미등록어처리에서는 자체의 고유한 특성으로부터 자름과 삽입연산이 아니라 등록 또는 삭제연산만이 허용된다. 그러므로 완전형미등록어는 대화의 방법으로 처리한다.

이 방법에서는 먼저 미등록어를 리용자에게 제시하여 문법적오유가 있는가를 확인하고 오유가 없을 때에는 그 단어가 분류에 영향을 주는가, 주지 않는가를 대화의 방법으로 확인한다. 미등록어가 분류에 영향을 준다고 확인되면 분류용어사전의 해당한 항목의 단어로 등록하고 그렇지 않으면 삭제한다.

부분형미등록어처리는 매우 복잡하다. 그것은 부분형미등록어형태가 다양하기때문이다.

례를 들어 부분형미등록어와 사전적표기의 차이점을 종합분석하여 보자.

부류	부분형미등록어	사전적표기	차이점
1	고등수학적기초지식교육	고등수학기초지식교육	중간위치에 있는 글자자름
2	공업생산전문화	공업적생산전문화	중간위치에 글자삽입
3	형태단어처리	형태단어	마지막위치의 글자자름
4	국제체력교예	체력교예	맨앞위치의 글자자름
5	현행교육계획작성	현행교육계획의 작성	합성어로부터 단어결합생성

보는바와 같이 부분형미등록어들을 분류용어사전의 단어와 대비하여 볼 때 뜻이 같거나 비슷한것이 있는가 하면 지어 뜻폭이 넓거나 좁은것도 있다. 그러므로 부분형미등록어를 분류용어사전의 단어와 의미상 일치하거나 류사한 단어로 변환하자면 위치에 따르는 자름과 삽입

연산을 적용하여야 한다.

부분형미등록어처리를 위해 합성어형식의 미등록어를 $H^{본}$으로, 분류용어사전의 합성어를 $H^{사}$로, 단어결합을 $U^{사}$로, 글자 또는 글자렬을 X로, $H^{본}$에 적용되는 연산규칙을 O라고 하자.

연산규칙은 다음과 같다.

첫째로, 분류용어사전의 단어를 기준으로 하여 미등록어를 비교하여 서로 다른 글자의 위치를 검색한다.

둘째로, 검색된 위치에 글자 또는 글자렬을 삭제 또는 삽입하여 사전적표기로 넘긴다.

앞위치연산자를 $O_{앞}$, 뒤위치연산자를 $O_{뒤}$, 중간위치연산자를 $O_{중}$ 그리고 삽입연산자를 $O_{삽}$, 삭제연산자를 $O_{삭}$이라고 하면 연산규칙 O는

$$O \rightarrow \begin{Bmatrix} O_{앞} \\ O_{중} \\ O_{뒤} \end{Bmatrix} \begin{Bmatrix} O_{삭} \\ O_{삽} \end{Bmatrix} X$$

로 표시된다.

그러면 부분형미등록어들에 연산규칙 O를 적용하여 사전적표기로 변환하는 과정은

$$(H^{본},\ O) \rightarrow \begin{Bmatrix} H^{사} \\ U^{사} \end{Bmatrix}$$

로 표시된다.

그러면 우의 표에 제시된 5개 부류에 해당한 미등록어들의 처리과정은 다음과 같이 표시된다.

즉

$$\text{1부류} \quad (H^{본},\ O_{중}O_{삭}X) \rightarrow H^{사}$$
$$\text{2부류} \quad (H^{본},\ O_{중}O_{삽}X) \rightarrow H^{사}$$
$$\text{3부류} \quad (H^{본},\ O_{뒤}O_{삭}X) \rightarrow H^{사}$$
$$\text{4부류} \quad (H^{본},\ O_{앞}O_{삭}X) \rightarrow H^{사}$$
$$\text{5부류} \quad (H^{본},\ O_{중}O_{삽}X) \rightarrow H^{사}$$

이다.

부분형미등록어들을 처리하자면 미등록어의 형태부들을 분해하고 매 형태부들을 분류용어사전의 단어와 비교하면서 후보형태부를 기준으로 하여 삽입, 삭제과정을 거쳐 분류용어사전의 단어로 변환하여야 한다.

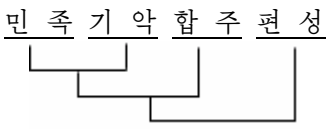
부분형미등록어처리에서는 형태부분해를 전제로 하는것만큼 어느 형태부를 기준으로 하여 분석처리하겠는가 하는 문제가 반드시 론의되여야 한다.

단어결합에서와 마찬가지로 부분형미등록어도 형태부들사이에 종속과 지배의 관계가 이루어진다.

례를 들어 《민족기악합주편성》이라는 미등록어에는 《민족》, 《기악》, 《합주》, 《편성》이

라는 형태부들이 서로 독립이 아니라 종속관계를 가지고 결합되여있다.

즉



이다.

보는바와 같이《민족》은《기악》에 종속하여 어떤 기악인가를 규정하며《민족기악》은《합주》에 종속하여 어떤 합주인가를 규정하며《민족기악합주》는《편성》에 종속하여 어떤 편성인가를 규정한다. 다시말하여 부분형미등록어 RB_i의 형태부 RB_i^1, RB_i^2, RB_i^3, …, RB_i^p에서 형태부 RB_i^p는 그앞의 형태부 RB_i^{p-1}에 대한 지배적형태부이며 반대로 RB_i^{p-1}는 RB_i^p에 대한 종속적형태부로 된다. 마찬가지로 다른 모든 형태부들도 보다 앞에 놓이는 형태부에 대하여 지배적인 형태부로 되며 앞에 놓이는 형태부는 그뒤에 놓이는 형태부에 대한 종속적인 형태부로 된다. 그러므로 미등록어에서 종속적위치에 있는 형태부들은 개념의 견지에서 보면 지배적위치에 있는 형태부에 대한 하위개념이거나 련관개념을 나타낸다고 말할 수 있다.

부분형미등록어에 내재하고있는 종속과 지배의 관계로부터 미등록어처리에서는 마지막형태부를 기준으로 하는것을 원칙으로 한다. 만일 미등록어의 마지막형태부가 일반용어이면 그것을 삭제하고 그앞의 형태부를 기준으로 하며 기준으로 된 형태부와 의미가 같은 단어로 치환하는 방법으로 미등록어를 처리한다.

미등록어처리알고리듬을 작성하기 위해 분류용어사전의 단어모임 M에서 추출된 합성어모임을 MU, MU에서 미등록어의 마지막형태부와 일치하는 합성어모임을 SU라고 표시하면 미등록어처리알고리듬은 다음과 같다.

1. 변수들을 초기화한다.
2. 미등록어모임 RB의 단어개수를 변수 n에 등록하고 RB의 단어번호를 표시하는 변수 i의 초기값을 1로 설정한다.
3. 미등록어모임 RB의 마지막단어에 이를 때까지 아래의 조작을 수행한다.
   1) 미등록어모임 RB의 i번째 단어 RB_i의 형태부를 분해하여 RB_i^k에 등록한다.
   2) 분류용어사전 M에서 추출된 합성어모임 MU에서 마지막형태부가 RB_i의 마지막형태부와 같은 합성어들로 모임 SU를 형성한다.
   3) 모임 SU가 빈모임인가 아닌가를 검사한다.
      (1) 모임 SU가 빈모임이 아니라면 아래의 조작을 수행한다.
         ① 모임 SU에 미등록어 RB_i와 같은 단어로 치환한다.
         ② 모임 SU에 미등록어 RB_i와 같거나 류사한 단어가 없다면 미등록어를 분류용어사전의 단어로 등록하겠는가 아니면 삭제하겠는가를 결심하고 해당한 처리를 진행한다.
      (2) 모임 SU가 빈모임이라면 마지막형태부가 일반용어인가를 확인한다.
         ① 만일 일반용어라면 그 형태부를 삭제한 다음 그앞의 형태부를 선택하고

3－2)의 조작수행에로 넘어간다.

② 일반용어가 아니라면 미등록어를 분류용어사전의 단어로 등록하겠는가 아니면 삭제하겠는가를 결심하고 해당한 처리를 진행한다.

4) 미등록어모임 RB의 단어위치를 하나 증가시킨다.

4. 작업을 완료한다.

[알고리듬 끝]

우의 알고리듬에서 분류용어사전 M에서 추출된 합성어모임 MU에서 마지막형태부가 RB_i의 마지막형태부와 같은 합성어모임 SU에는 해당 올림말과 련관된 단어들, 상하위관계에 있는 단어들까지 모두 있다고 본다.

문헌자동분류체계에서는 부분형미등록어처리를 위해 구체적인 표기가 다르지만 의미가 같은 단어들은 검색어사전과 같이 동의어관계를 밝혀주어 처리의 복잡성을 피할수도 있다.

실례로 뒤붙이《적》과 결합된 합성어와 단어결합을 들수 있다.

뒤붙이《적》뒤에는 바꿈토《이》와 규정토《ㄴ》이 결합된 합성토가 올수 있는 가능성이 많다. 뒤붙이《적》뒤에 바꿈토《이》와 규정토《ㄴ》이 결합된 합성토가 붙으면 합성어는 단어결합으로 된다. 실례로《부르죠아적영화리론연구》는 합성어이지만《부르죠아적인 영화리론연구》는 단어결합이다. 이때 단어의 외적인 형태는 다르지만 의미는 같다고 말할수 있다.

엄밀한 의미에서 단어결합《부르죠아적인 영화리론연구》는 합성어《부르죠아적영화리론연구》와 의미는 같지만 단어결합인것으로 하여 합성어처리단계에서의 변환은 불가능하다.

분류용어사전의 올림말들을 보면 뒤붙이《적》이 붙은 합성어는 자연과학부문에는 거의 없고 대체로 사회과학부문에만 있는데 그 수는 많지 않다.

그러므로 분류용어사전에서는 같은 의미를 가지는 합성어와 단어결합사이에 동의관계를 밝혀주어 미등록어처리에 크게 품을 들이지 않고도 의미가 같은 합성어나 단어결합을 사전적단어인 합성어 또는 단어결합으로 변환하여 사전정보를 얻을수 있다.

미등록어처리를 통하여서도 사전적단어로 변환할수 없는 단어들은 하나하나 따져보고 정보적가치가 있다고 인정되면 새로운 단어로 등록하고 그렇지 못한 단어들은 삭제한다.

우리는 문헌자동분류에서 제기되는 세부적인 문제들에 대한 연구를 심화시켜 자동분류의 효률을 부단히 개선해나감으로써 전통도서관들을 전자도서관화하는데 적극 이바지해나가야 할것이다.

실마리어 미등록어, 문헌자동분류