94-g-4

우리말 시소러스 자동구축을 위한 프레임웍

A Framework for the Automatic Construction of Hangul Thesaurus

김 현 주　　김 성 혁
숙명여자대학교 문헌정보학과

Hyun-Ju Kim　　Sung-Hyuk Kim
Dept. of Library & Information Science, Sookmyung Women's Univ.

본 논문에서는 원자력 공학분야 145개 용어를 분석하여 복합어의 합성원리와 기본용어의 위치정보를 이용한 시소러스 자동구축의 틀을 제안하였다.

1. 서론

ISO 2788 에서는 시소러스(thesaurus)를 '상위 및 하위 개념사이의 전후관계를 명백하게 하기 위하여 공식적으로 조직·통제된 색인어의 어휘.'라고 정의하였다. 이러한 시소러스는 색인작업시 색인자와 정보검색을 수행하는 최종이용자 혹은 중개자 사이에서 일관된 용어사용을 유도함으로써 정보검색시 검색효율을 증가시키는 역할을 수행하여 왔다.

시소러스를 구축하기 위하여 용어를 수집할 때 용어가 기록된 정보원 중 표준화된 용어의 정보원에는 분류표, 백과사전, 해당분야의 전문용어사전(辭典) 등이 있다. 이러한 정보원은 전자출판으로 인해 기계가독형 형태를 생산하고 있기 때문에 자동으로 시소러스를 구축하기위한 정보원으로 이용될 수 있다. 본 논문에서는 이러한 기계가독형 우리말 용어사전을 기본으로 하여 시소러스 자동구축의 틀을 제안하고자 한다.

2. 기본사항

① 대상 사전은 특정의 과학기술분야 기계가독형 사전으로 가정한다.

시소러스를 자동으로 구축할 때 인문.사회과학 분야는 용어의 다의성과 이형동의어의 처리 문제가 심각하다. 그러나 과학기술분야는 용어의 특정성이 높고 대부분 사전에 어휘통제와 상관참조가 존재하고 있으므로 어느정도 이러한 문제가 해결될 수 있다.

② 사전에 수록된 용어의 한글 표제(entry)만을 그 용어의 뜻 파악 대상으로 한다.

과학기술분야 용어는 복합어(compound word)와 혼합어(mixed word) 형태가 많다. 일본의 한 연구에 의하면 과학기술 분야 영어용어의 69%가 2단어 이상의 용어로 구성되어 있음이 나타났다.(Fujiwara et al 1987) 따라서 대상 용어들을 구성하는 단어들이 많으면 많을수록 해당 어근과 그 어근이 용어에서 나타나는 위치정보를 이용하여 용어간 상호연관성을 파악할 가능성이 높아지는 것이다.

③ 이러한 전제 아래 시소러스 자동구축에 필요한 단계를 도출하기 위하여 KIST·연구개발정보센터에서 간행한 분야별 용어 순열색인 중 원자력공학 분야 145개 용어를 분석하였다.

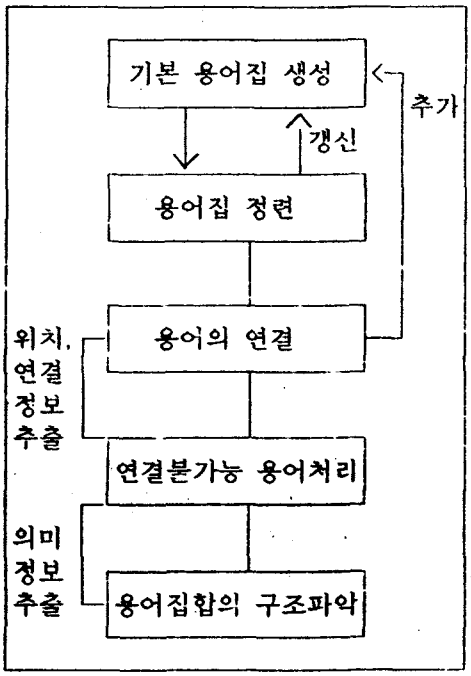
3. 시소러스 구축과정

사전에 수록된 과학기술분야의 용어는 어휘적

단어(lexical word)에 해당하며 어휘적 단어의 종류는 다음과 같다.

- 단일어(simple word)
- 합성어(complex word)
  - 파생어(derivational word)
  - 복합어(compound word)
  - 혼합어(mixed word)

원칙적으로는 단어와 어근(root)을 구별하여야 하겠지만, 본 논문에서는 용어의 전체적인 의미 파악의 가능성을 어근과 어근의 위치정보에서 찾고자 하기 때문에 접사를 포함한 어근, 혹은 어근을 편의상 용어를 구성하는 단어로 가정한다. 따라서 복합어와 혼합어의 구별은 없으며 2개 이상의 단어로 구성된 용어는 복합어로 간주하였다. 시소러스 구축의 기본절차는 (그림 1)과 같다.



(그림 1)

① 기본 용어집 생성

전체 용어집합을 U = { U (i,j,n) | i = 인덱스, j = 단어수, n = 빈도수 } 라고 정의하면 기본 용어집은 B1 = { U (i,1,n) | i = 인덱스, n = 빈도수 } 으로 정의할 수 있다.

한단어로 구성된 용어를 기본 용어집에 수록한다. 기본 용어집의 역할은 용어의 연결을 위한 기초가 된다. 따라서 효과적인 용어의 연결을 위하여 단어와 어근, 용어에 대한 단어의 빈도수를 고려하는 정련과정이 필요하다. 이 기본 용어집은 전 과정을 통하여 새롭게 등장하는 키(key) 용어를 계속하여 수록함으로써 해당분야의 완전한 기본 용어집을 구축하고 핵심용어를 파악할 수 있는 출발점이 된다.

원자력 공학분야 145개 용어중, 18개 용어가 초기 기본 용어집에 수록되었으며 n ≥ 2인 기본 용어는 6개 (원자로, 연료, 핵, 임계, 용융, 원자력) 였다.

② 용어의 연결

용어의 연결과정에서는 B1에서 n ≥ 2 인 용어를 기초로 광범위하게 복합용어를 연결시킨다.

사전에 수록된 용어의 표제만을 용어의 상호관계를 파악하기 위한 대상으로 가정하였으므로, 기본 용어가 나타나는 모든 용어를 연결하여 용어집합을 생성한다. 이 단계에서는 관련어와 상, 하위어를 구별할 수 있는 위치정보가 수집되어야 하며 집합과 집합사이의 연결도 파악되어야 한다.

기본용어의 위치가 해당 용어의 의미결정에 큰 영향을 미치기 때문에 용어의 연결은 기본용어의 위치에 따라 이루어진다. 기본용어가 복합어의 첫머리에 나타난 용어집합과 마지막에 나타난 용어집합을 비교해 보면, 전자는 복합어의 핵에 대한 공통 특성이나 대상을 표현하고 있는 반면 후자의 경우는 기본용어를 핵으로하는 용어들의 집합임을 알 수 있다. (표 1 참조)

... 원자로	원자로 ...
AGR 원자로	원자로 건조
BWR 원자로	원자로 격시
고선속 원자로	원자로 노심
.	.
.	.

(표 1)

사용한 용어 연결과정은 아래와 같다. (▨ : 기본 용어)

- U (k1,j,0) = ▨ □ □ ˙ ˙ .
  -> B_2F = { U (k1,j,0) }
  U (k2,j,0) = □ ˙ ˙ □▨ .
  -> B_2L = { U (k2,j,0) }

- B_2L 의 원소 중 L (k2, j≥3, 0) 를 대상으로 (j-1) 번째 단어(키용어)가 기본 용어집에 존재하면 해당 용어집합과 연결시키고 L (i, j≥3, 0) 를 대상으로 중간연결을 실시한다. 이러한 과정은 키용어를 변화시키면서 (j-2)번 실행한다.

키용어가 기본 용어집에 존재하지 않으면 B1에 등록시키고 빈도수를 조사한 후 해당 용어집합을 동일한 과정을 거쳐 생성하도록 한다.

- 나머지 용어들에 대한 정련과정 : 기본 용어집이 불충분할 때 반드시 필요하다.

연결결과 145개 용어중 109개의 용어가 연결되었고 정련 후 7개 용어가 연결되어 총 116개 용어가 2 용어 이상 연결되었다. (연결율 80%) 생성된 용어집합의 예는 표 2, 3과 같다.

(: 키용어)

... 원자로
.
.
가스 연료형 원자로
연료F ┐
연료L ├ 용어집합과
핵F ┘ 연결
사용후 연료 요소
파손 연료 검출 시스템

(표 2) 키용어가 B1에 있는경우

... 원자로
.
.
고속 중성자 원자로
중성자 감속재 ┐ 중성자F
중성자 반사체 ├ 용어집합
중성자 흡수체 ┘ 생성
펄스 중성자 기법
펄스 중성자형 원자로
열 중성자형 원자로

(표 3) 키용어가 B1에 없는 경우

③ 연결 불가능 용어에 대한 처리

앞서 116개의 연결된 용어 외에 연결되지 않은 29개 용어의 유형은 다음과 같다.

- 우리말 단일어 (경수로, 재관수, 증배율, 증식비, 탈피복 : 5 용어)
- 기본 용어와 일치하지 않는 우리말 복합어 (가속기 증식로, 농축 우라늄, 반응도 사고, 배가 시간, 소리 떨림 연소, 안전 조치, 압력 범위, 열 수로 인자, 유량 상실 : 9 용어)
- 외래어 표기법에 의한 우리말 용어 (버클링, 블랭킷, 빔 구멍, 서프레션 챔버, 소스 텀, 오클로 현상, 인 파일 루프, 채널 박스, 코어 캐처, 터빈 트립, 파이프 휘프 : 11 용어)

영어 (약자) 용어 (ATWS, ECCS, MLF, Furex법 : 4 용어)

이러한 용어들은 표제만으로는 그 뜻을 파악할 수 없지만 사전의 용어 정의 부분과 완전형 부분, 참조사항 정보 등을 이용한다면 연결이 가능해 진다. 예를 들어 아래에서 보여지듯 사전의 정의부분에는 보통 그 용어의 상위어가 나타나므로 연결에서 제외된 용어의 연결가능성을 찾을 수 있다.

예 (: 상위어)

- AET : 방사선 방어물질의 일종.
- APT : 방사선 방어물질의 한 가지.
- 시스테아민 : 방사선 장해에 대한 화학적 방어제의 하나.

④ 연결된 용어집합에 대한 구조파악

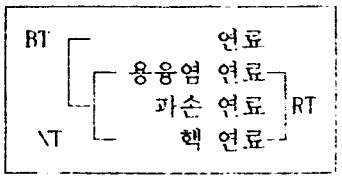
생성된 용어집합을 대상으로 하여 용어의 연결 정보와 기본용어의 위치, 우리말 단어의 합성법에 의한 단어 합성 원리를 이용하면 관련어와 상위어, 하위어를 추출할 수 있다.

본 연구에서 분석한 원자력 공학분야 145개 용어의 특징 및 단어합성 원리는 다음과 같다.

- 용어를 이루는 단어들의 대부분이 명사이다.
- 한자어가 많다. (원자력 공학 분야 145 용어중 131개 용어 : 90.34 %)
- 용어에 나타나는 접사는 대부분 한정적 접사이다.
- 유속 복합어가 많다.
- 어휘핵(lexical head)은 구조의 오른쪽에 위치한다.
- 핵은 자매항을 갖는다.
- 핵의 1차 투사만이 가능하다.

이러한 특징 및 원리와 기본용어의 위치정보

는 관련어와 상, 하위어의 추출을 가능하게 한다. 과학기술 용어는 유속 복합어가 대부분이고 어휘핵이 구조의 오른쪽에 위치하므로 2단어 용어는 자매항과 핵으로 구성되어 있다. 이때 기본용어가 해당용어의 핵에 나타나면 기본용어와 해당용어는 상위어, 하위어 관계이며 기본용어가 핵에 공통으로 위치한 용어들은 서로 자매항이 다른 관련어 관계이다. (표 4 참조)



(표 4)

기본적으로 상, 하위어의 추출은 기본용어와 그 기본용어를 포함하는 복합어 사이에서 이루어지며 용어가 3단어 이상으로 구성되어 있을 때에는 핵의 1차 투사 특성을 고려하여 자매항과 어휘핵이 구별되어야 한다. 관련어는 자매항이 서로 다르면서 동일한 핵을 가진 용어들 사이에서 추출할 수 있다.

4. 결론 및 제언

과학기술 분야는 생명주기가 짧으므로 새로운 용어의 출현과 기존 용어의 퇴보가 빈번하다. 기존의 수작업에 의한 시소러스 개발은 이러한 용어변화를 동적으로 수용할 수 없으며 비경제적이라는 이유에서 더이상 연구자들의 관심을 끌지 못하고 있다. 이에 시소러스를 자동으로 작성하려는 연구가 꾸준히 진행되어 왔으나 우리말 시소러스를 자동으로 작성하고자 하는 연구는 아직 미흡하다고 할 수 있다.

본 연구에서는 원자력 공학분야 용어 145개를 대상으로 용어의 특징을 분석하고 복합어의 합성원리를 적용하여 용어 자체에서 그 용어의 의미를 파악하여 시소러스를 구축할 수 있는 가능성을 확인하였다. 향후 기계가독형 우리말 용어사전이 구축된다면 본 논문에서 제안한 절차가 효과적으로 적용될 수 있을 것이다.

우리말 고유의 특성과 해당분야의 학문 세분화 정도, 용어의 특수성, 구축하고자 하는 시소러스의 특정성과 망라성 등은 상황에 따라 시소러스 구축 결과에 매우 큰 영향을 미칠 수 있다. 실제적인 시소러스 구축을 위해서는 이러한 사항을 고려한 우리말 단어합성 원리를 이용, 좀 더 정밀한 상,하위어와 관련어의 추출방법이 필요하다.

시소러스 자동 구축시 해결해야 할 또 다른 과제로는 동의어와 관련어를 구분하여 추출하기 위한 기준이 필요하다는 것이다. ISO 2788-1986(E)에서는 동일한 개념이 두가지 이상의 용어로 표현될 수 있을때 두 용어를 동의어로 규정하고 있다. 시소러스 자동구축시 이형동의어를 동의어로 추출하는 문제가 어렵다는 것은 앞에서도 지적한 바 있으나 특히 표제만으로 용어의 의미를 파악하는 경우에는 이형동의어의 동의어 추출이 불가능하다. 따라서 시소러스를 자동으로 구축할 때 적용할 수 있는 동의어 규정기준이 새롭게 제시되어야 할 것이다.

이러한 문제외에도 완전한 시소러스 자동구축을 위해서는 해당분야의 코퍼스 구축과 용어의 표준화및 표기법문제, 인문,사회과학분야 용어에서 많이 나타나는 용어의 중의성에 관한 연구들이 앞으로 행해져야 할 것이다.

참고문헌

성광수. 1988. "합성어 구성에 대한 검토 국어 어휘 구조와 어형성 규칙 (2) - " 한글. 201-202. 57-82.

시정곤. 1994. 국어의 단어형성 원리. 서울 국학자료원.

Aitichison, J. & Gilchrist, A. 1991. 시소러스의 작성법. 서울: 산업기술정보원

Fujiwara, Y. et al. 1987. "Analysis of Scientific and Technical Terms for Multilingual Database." Proc. of ICIK. 3-4.

Fujiwara, Y. et al. 1988. "A Dynamic Thesaurus for Intelligent Access to Research Database." Proc. of 44th FID Conference. 173-181.

Fujiwara, Y. & Lai, J. 1993. "Management and Advanced Utilization of Semantically Organized Terminology and Knowledge." Proc. of TKE'93. 141-151.

ISO 2788-1986(E). Documentation - Guidelines for the Establishment and Development of Monolingual Thesauri.