조어 중심적 주제어간 관계 추출 및 분석

정한민(Hanmin Jung)† 이미경(Mi-Kyoung Lee)†† 성원경(WonKyung Sung)†††
†한국과학기술정보연구원 정보서비스연구팀 책임연구원
†† 한국과학기술정보연구원 정보서비스연구팀 연구원
†††한국과학기술정보연구원 정보서비스연구팀 팀장/책임연구원
jhm@kisti.re.kr, jerryis@ks.ac.kr, wksung@kisti.re.kr

Analyzing and Extracting Relations between Topic Keywords Based on Word Formation

요 약

본 연구는 기존에 잘 알려지고 널리 사용되고 있는 어휘 의미망이나 시소러스를 활용하기 어려운 과학기술 분야, 특히 IT 분야에서 대용량 용어간 관계를 빠른 시간 내에 구축하여 검색 브라우징, 내비게이션 용도로 활용하는 것을 목표로 한다. 시소러스 구축 절차를 따르는 경우에 분야 전문가에 의한 정교한 작업과 고비용을 필요로 하여 충분한 구축 크기를 확보하는 것에 현실적인 어려움이 있다. 시소러스 자동 구축 방법론을 사용하는 경우에도 해당 용어들이 출현하는 방대한 말뭉치를 확보해야 하며 관계 구축 결과에 대한 직관적 이해가 쉽지 않다는 단점이 있다. 본 연구는 해외 학술 논문 말뭉치와 메타데이터에서 획득한 37만 여 주제어들을 이용하여 상·하위 관계, 관련어, 형제 관계를 추출하기 위해 조어적 기준에 근거한 규칙들을 이용한다. 이들 규칙을 이용하여 추출한 관계 수는 상·하위 관계 60여 만 개, 관련어 640여 만 개, 형제 관계 2,000여 만 개 등이다. 또한, 추출 결과 중 일부를 수작업으로 분석하여 단순한 추출 규칙에서 발생하는 오류 유형을 찾아내고 향후 과제에서 해결할 수 있는 방안에 대해 논하자고 한다.

키워드: 조어, 주제어, 관계 추출, 상·하위 관계, 관련어, 형제 관계, 동의 관계

1. 서론

과학기술 분야에서 주제어는 연구 대상이나 방법이 되는 주요 용어로 정의할 수 있다. 과학기술 분야에서의 주제어는 매우 빠르게 생겨난다는 특징과 함께 복합명사 형태를 많이 가지는데, 그 이유는 신기술의 발전과 함께 새롭게 대두되는 개념을 표현하기 위해 기존 용어들을 조합하는 방식을 주로 채택하고 있기 때문이다. 주제어는 정보 서비스에서의 연구 동향 파악이나 검색 브라우징, 내비게이션 등의 용도로 활용된다. 개체 중심적 통합 검색을 표방하는 시맨틱 웹 기반 정보 서비스인 OntoFrame은 주제어, 특히 복합명사 형태, 를 하나의 개체로 정의하고 주제어를 포함한 사용자 검색 요청이 발생하는 경우에 해당 주제어와 관련된 정보들을 통합하여 검색 결과로 제시함으로써 심도 깊은 정보 제공을 가능하게 한다 [1]. 그림 1에서 보듯이 주제어 통합 검색 결과 페이지는 주제별 전문가, 주제별 전문 연구 기관, 관련 주제어, 주제 동향 추이, 주제별 분류 문서 등의 정보를 제시한다. 다만, 관련 주제어의 경우에 단순히 입력된 주제어의 일부 단어와 매칭되는 주제어들을 나열함으로써 가독성이나 의미 관계 파악에 있어 어려움이 있다 (그림 1 참조). 이에 본 연구는 주제어들 간의 상·하위 관계를 포함하여 여러 관계들을 자동으로 추출하고 이를 시각적으로 제시함으로써 서비스의 사용 편의성을 높이고자 한다.

본 논문은 먼저 2장에서 관련 연구들을 살펴봄으로써 본 연구의 필요성을 지적하고, 3장에서는 주제어간 관계

추출 방법 및 그 결과를, 4장에서는 추출된 주제어간 관계 샘플들을 분석한 결과를 제시하고자 한다. 이를 통해 향후 규칙 보완 방안과 추가적인 언어 자원 도입의 필요성을 살펴보도록 한다.

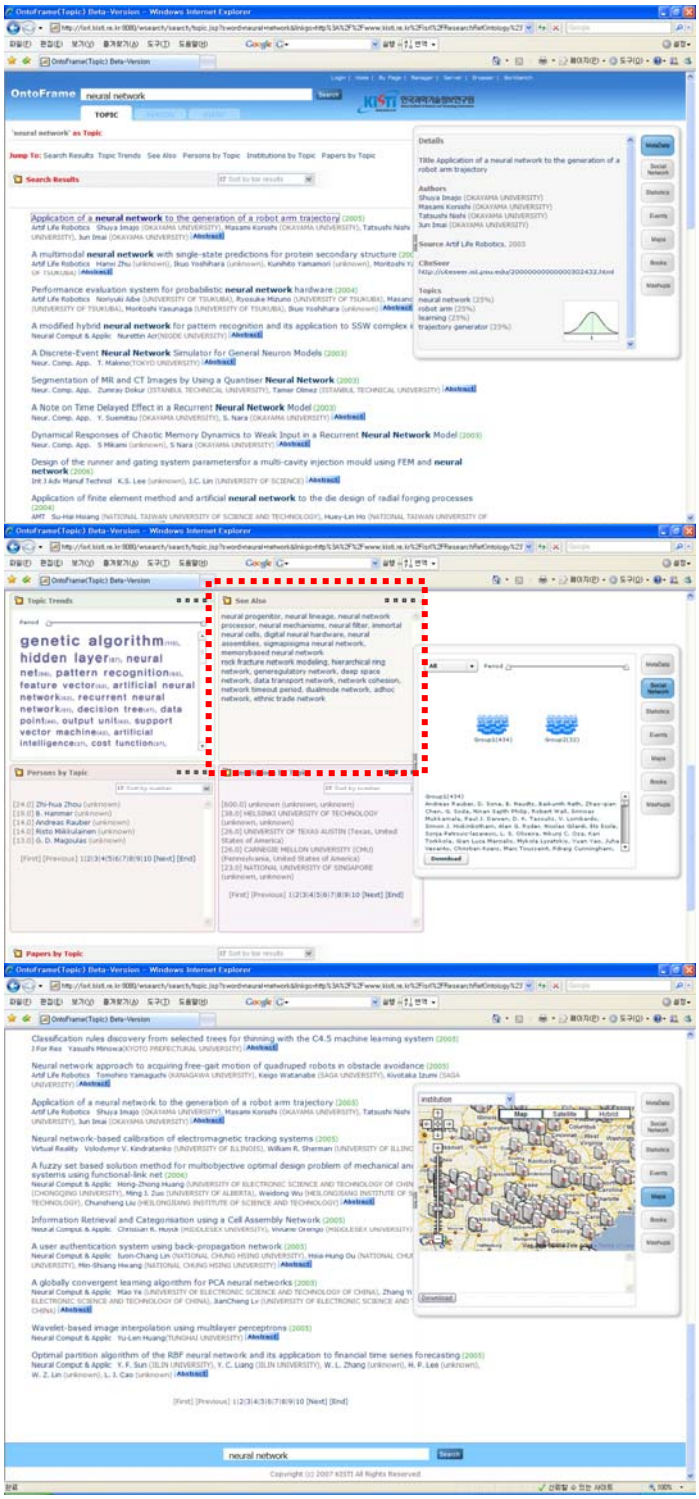

그림 1. 개체 중심적 통합 검색을 제공하는 OntoFrame 서비스 (붉은 사각형 내의 정보가 관련 주제어 목록)

2. 관련 연구

과학기술 분야에서 용어간 계층 관계를 구축하는 데 있어, 전문성을 기준으로 하는 연구들이 있다. [2], [3]은 구성 정보, 문맥 정보 등을 이용하여 용어간 전문성을 측정하고 이를 계층 관계 구축에 이용한다. 특히, [3]은 전문 용어와 함께 출현하는 수식 어구가 일반 명사의 그것보다 제한적인 형태로 나타난다는 사실에 근거하고 있다. 그렇지만, 이들 방법은 수식 어구나 문맥 정보가 충분히 확보될 수 있는 대용량 말뭉치를 이용해야 한다는 측면에서 단일어 중심의 용어간 계층 관계 구축에 보다 적합하다.

구글 등 인터넷 포털들은 사용자 검색 로그나 내비게이션 패턴, 공기 정보 등을 이용하여 관련 검색어를 제시한다. 예를 들어, "인공지능"으로 구글에서 검색하는 경우에 "kaist 인공지능", "스타크래프트 인공지능", "인공지능 개발", "지능 전망" 등 관련 검색어들을 제시하는데, 그 연관 관계 유형이나 정확도는 예측하기 힘들다는 단점이 있다.

시소러스나 어휘 의미망은 분야 전문가에 전적으로 의존하고 있다. 이는 개념어간 계층 관계를 의미적 기준을 근거로 판단해야 하기 때문인데, 그렇지만 과학기술 분야에서는 의미적 기준뿐만 아니라 조어적 기준도 중요한 영향을 미친다. 이러한 사실에 근거하여 [4]는 2005년부터 구축되어 온 과학기술 시소러스 구축에 조어적 기준도 적용하기 시작하였다. 시소러스의 계층 관계를 조어적 기준과 의미적 기준으로 동시에 기술함으로써 다양한 관점에서의 접근이 가능할 수 있도록 한 것이다.

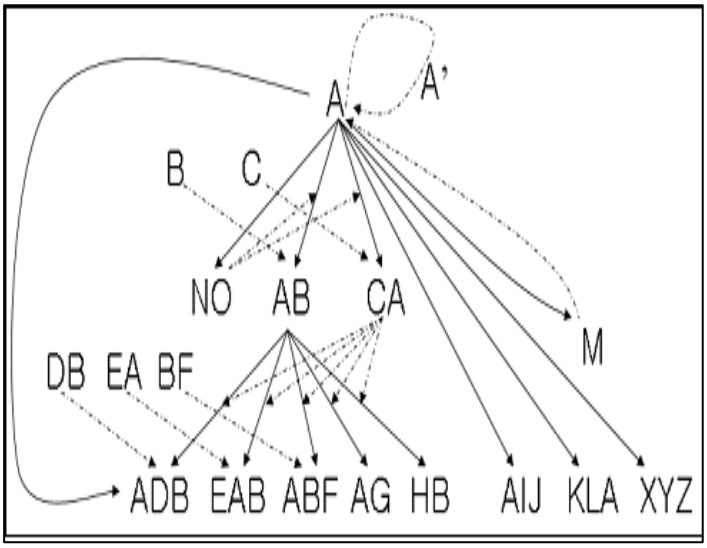

그림 2. [4]에서 사용한 조어적 유형에 따른 상·하위 관계도

[4]에서는 의미적 기준을 비합성어 [1] 에 한하여 적용하며 (A→M), 조어적 기준은 복합어 (A→CA, A→Ca, A→KLA, AB→EAB, AB→ADB)에 적용한다 (그림 2 참조). 조어적 기준 적용에서 계층 관계 설정과 차집합은 중심어를 기준으로 판단한다. 즉, 중심어가 달라지는 개념어간에는 계층 관계가 성립하지 않는다는 의미이다. 일반적으로 중심어는 마지막 용어 (A→AB는 B, AB→ABF는 F)이지만 예외적으로 A→AB, C→CA에서 A와 C가, AB→EAB에서 A가 중심어일 경우가 있으므로 분야 전문가에 의한 판단이 필요하다.

상기의 사례와 같이 조어적 유형을 활용하여 구축 효율성을 높였음에도 불구하고 여전히 분야 전문가가 높은 비중으로 개입할 수 밖에 없어, 비용적 측면에서나 구축 속도 측면에서 현실적인 한계를 가질 수 밖에 없다. OntoFrame의 경우에 37만 여 주제어를 가지고 있으며, 이 숫자는 서비스 대상 확장과 함께 급격히 커질 전망이다[2]. [4]에서 사용된 개념어가 4만 7천여 개인 점에서 상대적 괴리는 더욱 클 수 밖에 없다. 반면에, 사용자에게 제시되는 주제어간 관계는 검색 편의적인 목적으로 제공되는 것으로 정확도가 절대적인 기준으로 작용하지는 않는다.

본 연구는 검색 환경 하에서 대용량 주제어의 신속한 활용을 가능하게 하는 첫번째 시도로서 조어적 기준을 이용하여 대용량 주제어들로부터 상·하위 관계, 관련어, 형제 관계를 추출하고자 한다. 또한, 추출 결과 중 일부를 수작업으로 분석하여 단순한 추출 규칙에서 발생하는 오류 유형을 찾아내고 향후 과제에서 해결할 수 있는 방안에 대해 논하자고 한다.

3. 주제어간 관계 추출

주제어간 관계 추출을 위해 준비된 주제어들은 367,985개이며, 이들은 2000년 ~ 2007년 사이의 CiteSeer OAI Metadata [3] 와 Springer Journal Metadata에서 자동 추출된 것들이다. OntoFrame은 복합 명사 형태의 주제어들 위주로 개체 기반 통합 검색을 제공하므로, 본 실험에서 사용된 주제어들 역시 복합 명사들로 한정하고 있다.

4장에서 기술된 9개의 관계 추출 규칙은 상·하위 관계, 관련어, 형제 관계를 추출하기 위한 것으로, 조어적 형태만을 기준으로 작성되었다.

[1] 한 개의 실질 형태소로 되어 있는 용어, 실질 형태소에 접사가 결합된 용어, 합성어라 하더라도 사물의 명칭으로 단일어 단위로 개념을 분리하였을 경우 개념 패싯이 중심어와 달라지는 용어, 상품명이나 제품명, 생물의 고유명칭, 소프트웨어 고유명칭, 또는 두문자어.

[2] 서비스 대상 논문 수와 주제어 수는 1:1 비율에 가깝다.

[3] http://citeseer.ist.psu.edu/oai.html

표 1. 추출된 관계 유형별 빈도수

관계 유형	빈도수
상·하위 관계	59,250
관련어	6,391,082
형제 관계	20,109,265
고립어	67,225

추출된 관계에 포함되지 않은 주제어들은 고립어로 분류했으며, 그 비율은 전체 주제어 대비 약 18.3%이다 (표 1 참조). 이들은 기술된 9개의 추출 규칙에 의해 다른 주제어와 관계를 맺지 못한 것이므로, 추출 규칙들을 추가하면 이 비율은 줄어들 수 있다.

4. 주제어간 관계 분석

추출된 주제어간 관계 분석을 위해 2,660여 만 개의 관계들을 모두 분석하는 것은 현실적으로 어렵기 때문에 상대적으로 높은 빈도를 보인 "automatic"을 포함하는 주제어들과 관련된 관계들만 분석하였다 (표 2 참조).

표 2. "automatic"을 포함하는 주제어들과 관련된 관계 유형별, 추출 규칙별 빈도수

관계 유형	추출 규칙 번호	빈도수
상·하위 관계	1	14
상·하위 관계	2	2
형제 관계	3	159
형제 관계	6	2,585
관련어	4	58
관련어	5	193
관련어	7	1,540
관련어	8	131
관련어	9	73
합계		4,755

다음은 각 추출 규칙에 대한 설명, 추출 결과에 대한 설명, 추출 오류에 해당하는 예외 패턴 및 예제들을 보여준다. 조어적 기준에 의한 주제어간 관계 추출의 첫번째 시도이기 때문에 추출 규칙을 간단히 작성하였으며, 분석 또한 심도가 깊지는 않다.

추출 규칙 1 (상·하위 관계): 왼쪽 주제어의 첫번째 단어와 오른쪽 주제어의 첫번째 단어만 다르고 나머지 단어들은 동일, 왼쪽 주제어의 단어 개수가 오른쪽 주제어의 단어 개수보다 하나가 더 많음 (예. [A B C] [B C]).

추출 결과 샘플에서 예외를 발견하지 못함.

추출 규칙 2 (상·하위 관계): 왼쪽 주제어의 첫번째 단어와 오른쪽 주제어의 첫번째 단어만 다르고 나머지 단어들은 동일, 오른쪽 주제어의 단어 개수가 왼쪽 주제어의 단어 개수보다 하나가 더 많음 (예. [A B] [C A B]).

추출 결과 샘플에서 예외를 발견하지 못함.

추출 규칙 3 (형제 관계): 두 주제어의 길이가 3이상이고, 왼쪽 주제어의 첫번째 단어와 오른쪽 주제어의 첫번째 단어만 다르고 나머지 단어들은 동일, 오른쪽 주제어의 단어 개수가 왼쪽 주제어의 단어 개수와 같음 (예. [A B C]와 [D B C], 표 3 참조).

두 주제어간 서로 다른 단어가 상·하위 관계인 경우에 상·하위 관계도 나타날 수 있으나 추출 결과 샘플에서는 발견하지 못함. 추출 규칙 6의 경우와 유사함.

표 3. 추출 규칙 3에 의해 추출된 주제어간 관계 중 예외 예

1	A B C	A′ B C	동의 관계
	Automatic test generation	Computer-aided test generation	

추출 규칙 4 (관련어): 왼쪽 주제어의 마지막 단어와 오른쪽 주제어의 마지막 단어만 다르고 나머지 단어들은 동일, 구성 단어 개수는 3이상임 (예. [A B C]와 [A B D], 표 4 참조).

대부분의 관계가 관련어에 속하나, 일부 관계는 동의 관계나 상·하위 관계를 가짐. 동의 관계의 경우에는 두 주제어간 서로 다른 단어가 유의어인 경우이며, 상·하위 관계의 경우에는 두 주제어간 서로 다른 단어 중 하나가 추상적 수준 (어휘 의미망에서 상위 계층에 속할 수 있는)에 위치함.

표 4. 추출 규칙 4에 의해 추출된 주제어간 관계 중 예외 예

1	A B C	A B D	상·하위 관계
	Automatic text representation	Automatic text summarization	
2	A B C	A B D	동의 관계
	Automatic topic identification	Automatic topic retrieval	
3	A B C D	A B C E	상·하위 관계
	Automatic word sense clustering	Automatic word sense disambiguation	
4	A B C	A B C′	동의 관계
	Automatic term acquisition	Automatic term extraction	

추출 규칙 5 (관련어): 왼쪽 주제어의 첫번째와 마지막 단어가 오른쪽 주제어의 첫번째와 마지막 단어와 일치하고, 왼쪽 주제어의 단어 개수가 오른쪽 주제어의 단어 개수와 같음 (예. [A B C]와 [A D C], 표 5 참조).

두 주제어간 서로 다른 단어가 유의어 또는 상·하위 관계를 가지는 경우에 형제 관계, 동의 관계, 상·하위 관계 등 구체적인 관계를 가짐.

표 5. 추출 규칙 5에 의해 추출된 주제어간 관계 중 예외 예

1	A B C	A D C	형제 관계
	Automatic rule learning	Automatic grammar learning	
2	A B C	A B′ C	동의 관계
	Automatic context switching	Automatic mode switching	
3	A B C	A D C	상·하위 관계
	Automatic term extraction	Automatic feature extraction	
4	A B C D	A B E D	상·하위 관계
	Automatic test data generation	Automatic test pattern generation	
5	A B C D	A B C′ D	동의 관계
	Automatic test pattern generation	Automatic test sequence generation	

추출 규칙 6 (형제 관계): 두 주제어의 길이가 2이고, 왼쪽 주제어의 첫번째 단어와 오른쪽 주제어의 첫번째 단어가 다르고, 왼쪽 주제어의 마지막 단어와 오른쪽 주제어의 마지막 단어가 일치함 (예. [A B]와 [C B], 표 6 참조).

대부분 형제 관계에 속하나 의미적 기준으로 그 내에서 클러스터링이 추가적으로 가능하거나, 상·하위 관계를 맺을 수 있는 경우가 많음.

표 6. 추출 규칙 6에 의해 추출된 주제어간 관계 중 예외 예

1	A B	C B	상·하위 관계
	Automatic reasoning	Case-based reasoning	
2	A B	A′ B	동의 관계
	Text search	Full-text search	

추출 규칙 7 (관련어): 왼쪽 주제어의 마지막 단어와 오른쪽 주제어의 마지막 단어만 다르고 첫번째 단어는 동일, 구성 단어 개수는 2임 (예. [A B]와 [A C], 표 7 참조).

추출 규칙 4의 경우와 유사함.

표 7. 추출 규칙 7에 의해 추출된 주제어간 관계 중 예외 예

1	A B	A C	상·하위 관계
	Automatic reasoning	Automatic processing	
2	A B	A B′	동의 관계
	Automatic tagging	Automatic annotation	

추출 규칙 8 (관련어): 왼쪽 주제어의 첫번째와 마지막 단어가 오른쪽 주제어의 첫번째와 마지막 단어와 일치하고, 왼쪽 주제어의 단어 개수가 오른쪽 주제어의 단어 개수보다 하나 더 많음 (예. [A B C]와 [A C], 표 8 참조).

두 주제어 중 조어적 차집합에 해당하는 단어의 위치가 전방에 위치하는 경우에는 상·하위 관계를 주로 가지며, 후방에 위치하는 경우에는 추출 규칙 7의 경우와 유사함.

표 8. 추출 규칙 8에 의해 추출된 주제어간 관계 중 예외 예

1	A B C	A C	상·하위 관계
	Automatic program analysis	Automatic analysis	
2	A B C D	A C D	상·하위 관계
	Automatic parallel performance analysis	Automatic performance analysis	
3	A B C D	A B D	상·하위 관계
	Automatic test bench generation	Automatic test generation	
4	A B C D	A E D	동의 관계
	Automatic test bench generation	Automatic testcase generation	
5	A B C D	A B+C D	동의 관계
	Automatic test case generation	Automatic testcase generation	

추출 규칙 9 (관련어): 왼쪽 주제어의 첫번째와 마지막 단어가 오른쪽 주제어의 첫번째와 마지막 단어와 일치하고, 오른쪽 주제어의 단어 개수가 왼쪽 주제어의 단어 개수보다 하나 더 많음 (예. [A B]와 [A C B], 표 9 참조).

추출 규칙 8의 경우와 동일함.

표 9. 추출 규칙 9에 의해 추출된 주제어간 관계 중 예외 예

1	A B	A C B	상·하위 관계
	Automatic segmentation	Automatic text segmentation	
2	A B C	A D B C	동의 관계
	Automatic sense disambiguation	Automatic word sense disambiguation	
3	A B C	A D B′ C	상·하위 관계
	Automatic topic identification	Automatic web genre identification	
4	A B C	A B D C	상·하위 관계
	Automatic test generation	Automatic test pattern generation	

5. 결론

본 논문은 기존에 잘 알려지고 널리 사용되고 있는 어휘 의미망이나 시소러스를 활용하기 어려운 과학기술 분야, 특히 IT 분야에서 대용량 용어간 관계를 빠른 시간 내에 구축하기 위한 방법의 하나로써 조어적 기준에 의한 주제어간 관계 추출에 대해 기술하였다. 9개의 간단한 추출 규칙을 이용하여 추출된 관계는 2,660여 만 개에 달하는 데 일부 샘플을 분석하여 잘못된 유형으로 추출된 관계들을 살펴보았다. 이들 중 일부는 추출 규칙을 보완하여 해결할 수 있으며 (예, [A B C]와 [A D B C]의 경우에는 상·하위 관계 추출 규칙으로 재 기술), 일부는 추가적인 언어 자원을 도입하여 해결할 수 있을 것이다 (예. [A B] [A B′]의 경우에는 B와 B′의 개념이 같은 지의 여부를 검사하기 위해 WordNet의 Synset, Hypernym/Hyponym 등을 활용). 현재 추출 수준으로도 OntoFrame 서비스에의 적용이 충분히 가능하나, 향후 연구에서는 이러한 개선들을 통해 좀더 정교한 서비스의 제공을 추구할 예정이다.

참고문헌

[1] W. Sung, H. Jung, P. Kim, I. Kang, S. Lee, M. Lee, D. Park, and S. Hahn, "A Semantic Portal for Researchers Using OntoFrame", In Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, 2007.

[2] 류법모, 배선미, 최기선, "구성정보와 문맥정보를 이용한 용어의 전문성 측정 방법", 한국정보과학회 춘계학술대회, 2004.

[3] 구희관, 정한민, 이병희, 성원경, "수식어구를 이용한 한국어 용어의 전문성 측정", 한국컴퓨터종합학술대회, 2005.

[4] 강인수, 최석두, 김이겸, 한선화, 이상헌, 김도완, 박동인, 성원경, 정한민, "과학기술 분야 시소러스 구축 연구", 한국과학기술정보연구원, ISBN: 978-89-6211-025-8 93500, 2007.