한국어 개념망 구축을 위한 지능형 워크벤치

허정°, 최미란, 장명길
한국전자통신연구원 지식마이닝연구팀
{jeonghur°, mirac, mgjang}@etri.re.kr

Intelligent Workbench for Korean Concept-Net Construction

Jeong Hur°, Mi-Ran Choi, Myung-Gil Jang
Knowledge Mining Research Team, ETRI

요 약

개념망은 상당히 도메인에 의존적인 언어자원에 해당한다. 따라서, 도메인이 다른 분야에 적용하고자 한다면, 많은 수정이 요구된다. 그러나 개념망의 편집은 언어 이해 능력이 뛰어난 언어학자들 조차도 상당히 많은 시간이 요구되는 작업이다. 대부분의 시간소요는 개념망의 전체적인 계층구조를 스캐닝하는 작업과 특정 노드를 검색하는 작업에 의한 것이다. 기 구축된 개념망을 분석하면 계층관계에 있는 어휘들간의 일관된 규칙을 발견할 수 있다. 이 논문에서는 어휘들의 뜻풀이와 상위어간의 관계성, 복합명사와 상위어간의 관계성을 통계적으로 분석하였다. 분석된 결과를 기반으로 확률모델을 이용하여 상위어 추천 기능을 구현하였다. 상위어 추천 기능의 시간 절감 효과를 실험하기 위해 실험자 2인을 대상으로 개념망 구축에 소요되는 시간을 측정하였다. 상위어 추천 기능이 있는 지능형 워크벤치를 이용할 경우 개념망 작업 시간은 약 65%정도로 단축되는 것을 확인할 수 있었다. 본 지능형 워크벤치는 다양한 도메인에서 요구되는 개념망 구축의 시간 비용 절감에 크게 기여할 것으로 기대된다.

1. 서론

개념망은 상당히 도메인에 의존적인 언어자원에 해당한다. 그러므로, 특정 분야를 목적으로 작성된 개념망을 다른 분야에 적용하고자 한다면 상당한 수정이 불가피하다. 개념망의 수정은 일반적으로 새로운 개념어휘의 추가 및 수정인데, 이 작업은 전체적인 개념망의 구조를 스캐닝해야 하고, 관련된 어휘들의 분포를 파악해야 하는 작업이다. 따라서, 많은 시간적 비용이 소요된다. 이와 같은 문제점을 최소화하기 위해서 개념망의 도메인 이식성을 고려하여 어휘들과 개념망 구조와의 관련성을 기반으로 한 지능적 개념망 구축 워크벤치가 요구된다. 지능형 워크벤치에서는 개념망에 새로운 어휘를 추가하고자 할 때, 개념망에 구축되어 있는 개념관계들과의 의미적 관련성을 고려하여 새로운 어휘가 위치해야 할 서브 노드를 추천하는 기능을 제공하고 있다.

지능형 워크벤치를 위해서는 기 구축된 개념망의 구조와 개념망에 포함되어 있는 어휘 정보들과의 의미적, 구조적 관계를 분석할 필요가 있다. 의미적인 관계를 위해서 어휘의 뜻풀이와 상위어와의 관련성을 분석하였고, 구조적인 관계를 위해서 복합명사의 구성 어휘와 상위어와의 구조적인 관련성을 분석하였다. 분석된 데이터를 기반으로 각각의 패턴에 대한 통계값을 계산하여, 패턴 지식으로 구축하였다. 구축된 패턴 지식을 기반으로 새로운 어휘에 대한 상위어 후보를 추천하는 지능형 워크벤치를 구현하였다. [그림 1]은 상위어 추천의 흐름도이다.

논문의 구성은 2장에서 통계 패턴 추출에 대해서 기술하고, 3장에서 상위어 추천에 대해서 기술한다. 4장에서는 지능형 워크벤치의 효율성에 대한 실험내용을 기술하고, 5장에서 결론 및 향후 연구에 대해서 기술한다.

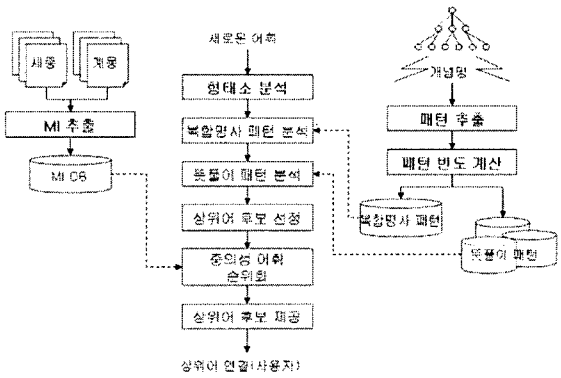

그림 1. 상위어 추천 흐름도

2. 기 구축된 개념망의 어휘들간의 관계 분석

기 구축된 개념망[2,3]은 일반적으로 어휘들 간의 관계와 어휘들에 대한 사전 정보들로 구성된다. 상위어 추천을 위한 기 구축된 개념망 분석을 위해서 다음 두 가지의 유형에 대해서 분석하였다. 첫째, 개념망을 구성하는 어휘들 간의 관계에 대한 분석으로, 복합명사와 그 상위어간의 패턴 구조 분석이 있고, 둘째, 어휘의 뜻풀이와 상위

어 간의 패턴 구조 분석이 있다.

2.1. 복합명사와 상위어간의 관계 분석

개념망을 구성하는 어휘들 중 복합명사의 비율은 약 20%정도 된다. 복합명사를 구성하는 개별 어휘들의 의미와 복합명사의 의미는 상호간에 밀접한 연관성이 있다. 그러므로, 복합명사와 복합명사를 구성하는 개별 어휘들의 의미적 관계를 고려하면 개념망에 복합명사를 추가할 때 상위어를 추천할 수 있다.

복합명사에 대한 분석 결과, 개념망에 포함된 어휘 63,075개 중 복합명사의 수는 12,828개였다. 복합명사의 개별 어휘들 중에 복합명사의 상위어가 있는 경우가 5,916개였고, 없는 경우가 6,912개로 약 46%정도는 복합명사를 구성하는 개별 어휘들 중에 복합명사의 상위어가 있는 것으로 분석되었다. 또한, 복합명사를 구성하는 명사들 중 제일 마지막에 위치한 명사가 상위어인 형태가 대부분이다. 이와 같은 특징을 반영하기 위해 복합 명사의 품사열을 기반으로 패턴을 통계적으로 추출하였다.

표 1. 복합명사와 상위어간의 관계 품사열 패턴

패턴	확률
NN+[NN]	0.079460
PF+[NN]	0.008307
NN+[NN+SN]	0.002504
NN+NN+[NN]	0.002425

표 2. 복합명사와 상위어간의 관계 예

복합명사	상위어
공상과학소설	소설
무형문화재	문화재
호두엿	엿

2.2. 뜻풀이와 상위어간의 관계 분석

어휘의 뜻풀이에는 어휘와 관련된 다양한 의미적 정보가 포함되어 있다. 개념망의 어휘관계도 뜻풀이를 통해 추론할 수 있다 이를 위해, 어휘들의 개념와 뜻풀이 간의 관계 분석은 개념망의 상위어 추천에서 중요한 지식으로 활용될 수 있다.

[1]은 명사 계층 구조를 뜻풀이를 기반으로 하여 반자동으로 구축하는 방법에 대해서 제시한 논문이다. 이 논문에서는 뜻풀이 유형을 11가지로 분류하고 있다.

본 논문에서 기 구축된 개념망의 뜻풀이와 상위어의 관계를 분석한 결과는 크게 5가지로 분류할 수 있다.

(1) 명사구 형태의 뜻풀이로 마지막에 상위어가 있는 형태
(2) 명사구 형태의 뜻풀이로 마지막에 상위어가 접속조사나 접속부사로 연결된 형태
(3) " ~를 이르는 말" 의 형태
(4) 동사로 파생된 명사가 상위어인 형태
(5) 상위어에 관형격 조사가 붙은 형태

뜻풀이의 분석결과, 실험대상이 된 어휘는 총 82.867이고, 이 중 뜻풀이에 상위어가 있는 어휘의 개수는 49.701개이다. 즉 59.98%가 뜻풀이에 상위어를 포함하고 있는 것으로 분석되었다.

표 3. 뜻풀이와 상위어간의 유형 예

개념어	뜻풀이	상위어	유형
각도기	각의 크기를 재는 기구	기구	(1)
모종삽	모종할 때 쓰는 작은 삽	삽	(1)
분별	세상 물정에 대한 바른 생각이나 판단	판단	(2)
진도	일의 진행 속도 또는 정도.	정도	(2)
가난뱅이	가난한 사람을 낮추어 이르는 말.	사람	(3)
가스	기체 물질을 통틀어 이르는 말	기체	(3)
가불	(봉급 따위를) 기일 전에 지불함	지불	(4)
간난	몹시 힘들고 고생이 됨	고생	(4)
가시밭	고난과 애로가 덮친 환경의 비유.	환경	(5)
가야금	우리나라 고유 현악기의 하나	현악기	(5)

5가지 유형으로 분류된 뜻풀이들의 패턴 통계자료를 추출하기 위해 2종류의 패턴형식을 정의하고 이를 기반으로 통계정보를 추출하였다. 유형(1), (2)를 반영하기 위해서 단순 품사열을 기반으로 한 통계패턴과 유형(3), (4), (5)를 위한 품사열과 어휘에 기반한 통계패턴이 있다. 품사열과 어휘를 기반으로 한 통계패턴은 2가지 형태로 구분하여 구축하였다. [표 5]는 품사열과 어휘에 기반한 통계패턴의 예인데, 동사 어휘에 기반한 유형과 그렇지 않은 유형으로 구분할 수 있다.

표 4. 품사열 통계 패턴

패턴	빈도	확률
[NN]_SY	28193	0.397157
[NN+NN]_SY	1549	0.021821
[NN+SN]_SY	860	0.012115
[NN]_SV+EM+SY	583	0.008213

표 5. 품사와 어휘에 기반한 통계 패턴

구분	패턴	빈도	확률
Type1	[NN]_하나/NU+./SY	397	0.005593
	[NN+NN]_하나/NU+./SY	118	0.001662
	[NN+SN]_한/DN+분야/NN+이/CP	70	0.000986
Type2	[NN]_이르/VV	1044	0.014707
	[NN]_하/VV	62	0.000873
	[NN+NN]_이르/VV	59	0.000831

3. 상위어 추천

워크벤치는 윈도우 기반에 MFC를 이용하여 구축되었다.

[그림 2]는 상위어 추천 모듈의 실행 화면이다. 추천된 어휘들의 정보를 기반으로 사용자는 원 클릭으로 개념망에 어휘를 추가할 수 있다.

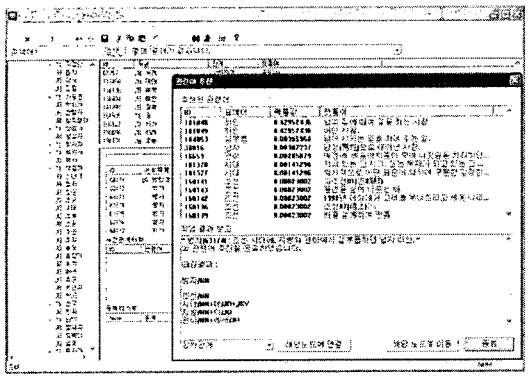

그림 2. 워크벤치 화면

3.1. 상위어 후보 순위화

2장에서 기술된 통계패턴을 기반으로 개념망에 추가될 어휘와 뜻풀이를 형태소 분석하여 후보군에 속하는 명사어휘를 추출하고, 해당 명사가 상위어일 확률값을 계산한다.

$$R_{W_i} = P^1_{W_i} + P^2_{W_i} + P^3_{W_i}$$

2장에서 언급된 3개의 통계패턴의 합으로 해당 후보어휘의 상위어 추천값을 계산하였다.

3.2. 중의성 어휘

상위어 후보에 속한 어휘들 중에는 의미적 중의성을 가진 어휘가 있다. 따라서 중의성 어휘들에 대한 어휘 중의성 해소 모듈이 요구된다. 그러나, 중의성 해소 모듈의 정확률에 한계가 있으므로, 중의성 해소 모듈의 확률값을 기반으로 순위화하여 제시한다.

중의성 해소 모델은 어휘의 상관계수들 중 MI를 이용하여 수행하였다. 뜻풀이에 출현하는 중의성을 가진 상위어 후보는 뜻풀이에서 공기한 어휘들와 중의성 어휘의 뜻풀이에 나온 어휘들간의 평균 MI값으로 순위화 할 수 있다.

$$SE = \{SW_1, SW_2, ..., SW_n\} .....(1)$$

$$AW = \{TE_1, TE_2, ..., TE_n\} .....(2)$$

$$TE_i = \{TW^i_1, TW^i_2, ...TW^i_m\} .....(3)$$

$$TER_i = \frac{\sum_{x=1}^{n}\sum_{y=1}^{m} MI(SW_x, TW^i_y)}{n \times m} ....(4)$$

SE는 개념망에 추가되는 어휘의 뜻풀이를 의미하고, SW는 뜻풀이를 구성하는 어휘들이다. AW는 상위어 후보 중 중의성을 가진 어휘를 의미하고, TE는 AW의 다양한 의미에 대한 뜻풀이를 의미한다. TW는 TE의 뜻풀이를 구성하는 어휘들이다. 중의성 어휘의 확률값은 추가될 어휘의 뜻풀이와 중의성을 가진 후보의 의미별 뜻풀이와의 평균 MI값으로 계산된다. 수식(4)는 이를 반영하고 있다. MI추출을 위해서는 ETRI 개념망 사전, 계몽사전, 세종코퍼스를 활용하였고, 추출된 어휘쌍은 약 2,040만쌍이다.

4. 실험 및 분석

실험은 워크벤치를 이용하였을 때의 작업 효율에 관한 것으로 개념망 구축 소요 시간에 대한 실험이다. 100개의 어휘에 대해서 실험자 2명을 이용하여 상위어 추천기를 사용했을 때와 사용하지 않았을 때의 작업 시간을 분석하였다.

표 6. 상위어 추천 모듈의 작업 속도 효율성 실험 결과

	사용(A)	사용 하지 않음(B)	A/B
실험자 1	68 분	102 분	0.67
실험자 2	52 분	85 분	0.61

[표 6]에서 알 수 있듯이, 상위어 추천 모듈이 없을 때의 작업에 비해 상위어 추천 모듈이 있을 경우, 시간적으로 약 0.64((0.67+0.61)/2)의 시간 투자로 동일한 작업을 할 수 있다. 즉, 작업의 속도 측면에서 상당한 효율이 있다는 것을 알 수 있었다. 특히, 실험자 1과 실험자 2는 서로 작업의 숙련도에서 다소 차이가 있는데, 숙련도가 높은 실험자 2가 상위어 추천 모듈이 없을 때보다 있을 때의 작업 속도 비율이 작다는 것은 고무적인 일이다. 즉, 숙련도가 높은 작업자에게 상위어 추천 모듈이 작업 효율향상에 더욱 영향을 미친다는 것이다.

5. 결론 및 향후 연구방향

개념망은 의미적인 관계를 설정하는 것으로 사람의 개입이 원천적으로 요구되는 자원이다. 이는 많은 시간이 소요되는 작업으로 이를 최소화하기 위해 기 구축된 개념망을 분석하여 패턴 통계를 구축하고 이를 기반으로 상위어를 추천하는 지능형 워크벤치를 개발하였다.

상위어 추천 모듈의 효율성에 대한 실험에서 작업 속도의 효율은 상위어 추천 모듈을 이용할 때 약 0.64 의 노력으로 이용하지 않을 때와 동일한 작업을 할 수 있다는 것을 알 수 있었다. 이처럼 상위어 추천 모듈이 내장된 지능형 워크벤치는 개념망 구축의 시간 비용 절감에 상당한 긍정적 역할을 할 것으로 기대된다.

앞으로 진행되어야 할 작업은 상위어 추천의 과분석에 의한 문제를 해결하기 위해 상위어 후보 필터링 기술과 중의성 어휘들에 대한 처리 방안을 연구하여야 한다. 이를 위해서는 정확률이 향상된 어휘 의미 분석 모듈이 내장되어야 한다.

참고문헌

[1] 조평옥 외 3명, 사전 뜻풀이에서 구축한 한국어 명사 의미계층구조, 인지과학회 논문지 제10권 제4호, 1999년.

[2] 최미란 외 2명, Constructing Korean Lexical Concept Network for Encyclopedia Question-Answering System, IECON 04, 2004년

[3] 왕지현 외 1명, 정보검색을 위한 한국어 명사 개념망 구축에 관한 연구, 제1회 한국시소러스연구회 국제학술포럼, 2003년.