온톨로지 정보를 이용한 범주 재편성: Roget 시소러스의 경우

양재군, 배재학
울산대학교 컴퓨터·정보통신공학부
e-mail:{jgyang, jhjbae}@ulsan.ac.kr

Category Reorganization with Ontology Information: Roget Thesaurus Case

Jae-Gun Yang, Jae-Hak J. Bae
School of Computer Engineering and Information Technology,
University of Ulsan

요약

본 논문에서는 Roget 시소러스의 범주를 재편성하여 문장추상화에 사용할 온톨로지를 구축하였다. Roget 시소러스의 표제정보와 참조정보를 이용해서 범주를 재편성한 각 결과를 토대로, OfN(Ontology for Narratives)을 구성하였다. 이렇게 하여 얻어진 OfN을 설화 문장추상화에 적용하여 이 온톨로지가 유의함을 확인하였다.

1. 서 론

인터넷의 대중화로 온라인화된 각종 문서의 양이 급격히 증가하고 있다. 이에 비례하여 문서 검색이나 요약의 필요성이 절실해지고 있다. 사람이 문서를 읽고 회상하고 요약하는 과정에서는, 문장을 구성하는 상세 정보보다 오히려 개념화되고 추상화된 문장이 처리 대상이 된다. 이러한 문장추상화[1]를 자동화하기 위해서 문장 안의 중요정보를 분별하는데 쓸 온톨로지(Ontology, 존재론)가 필요하다.

본 논문에서는 일곱 가지 범주로 구성된 OfN(Ontology for Narratives)[2]을 Roget 시소러스[3]를 재구성하여 얻었다. 이 온톨로지는 설화문장을 추상화시키는데 사용할 목적으로 설정되었는데, 다음과 같은 7가지 유형(Type)이 포함되어 있다: (1) 등장인물(Character) - 이야기에 등장하는 사람이거나 혹은 의인화 가능한 존재이다. (2) 심상(Affect State) - 등장인물의 감정 상태이다. (3) 사건(Event) - 이야기에서 어떤 중요한 일의 발생이다. (4) 상태(State) - 이야기의 등장인물이나 사물의 감정 상태를 제외한 나머지 상황이다. (5, 6) 시간과 공간의 변화(Delta-{Time, Space}). (7) 담화 표지(Discourse Markers) - 화자의 의도를 내포하는 단서구(Cue Phrase)이다. 이렇게 설정한 OfN을 구축하기 위해서 먼저 Roget 시소러스의 범주를 심상, 시간과 공간, 사건, 그리고 상태 등으로 재편성하였다.

2. Roget 시소러스의 구조

Roget 시소러스[3]는 의미 분류에 기초한 총 6개의 강(Class)으로 구성되었다. 각 강은 하부에 부(Division), 과(Section) 등의 계층구조로 세분화되었다. 각 계층은 저마다의 표제정보를 가지고 있으며 계층구조의 말단에는 총 1044개의 범주가 존재한다. 각 범주에는 품사별로 유의어 목록이 나열되어 있다. 한편, 유의어 목록에서 특정 어휘가 다른 범주를 참조하는 경우에는 "어휘 &c. (표제어) 표제번호"의 형식으로 표현한다.

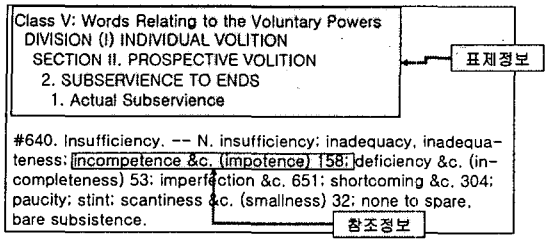

그림 1 Roget 범주 #640의 표제정보와 참조정보

3. 온톨로지 재분류

전처리된 Roget 시소러스[4]를 새로운 온톨로지로 재분류하기 위해서 서로 다른 두 가지 접근방식을 모색하였다. 첫 번째는 Roget 시소러스의 표제정보를 이용해서 분류하는 방법이고 두 번째는 참조기호를 탐색하여 얻은 정보를 기반으로 분류하는 방법이다.

3.1 표제정보를 이용하여 구축한 OfN

온톨로지의 범주에 새로운 분류기준을 적용하면 쓰임새가 다른 새로운 온톨로지를 얻을 수 있을 것이다. 이점에 착안하여 Roget 시소러스의 표제정보를 이용해서 새로운 온톨로지로 재구성하였다.

모든 표제정보와 표제어를 취합한 후 각 표제정보와 표제어를 Prolog의 술어형태로 치환하였다. 이 술어 형태의 표제정보에 OfN 테이블을 적용해서 OfN 형태의 표제정보를 얻었다. OfN 형태의 표제정보에 나타나는 범주의 개수와 우선순위에 따라서 Roget 범주를 OfN범주에 할당하였다.

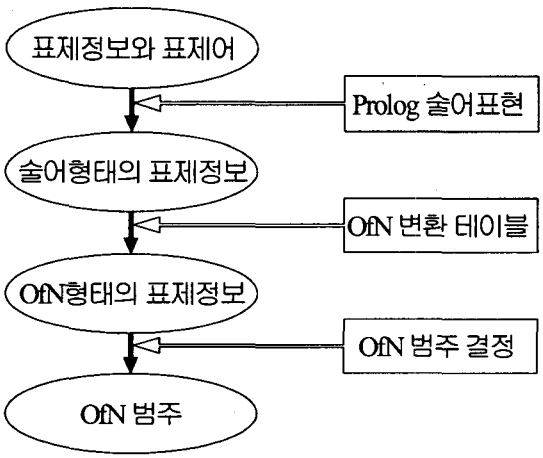

그림 2 표제정보를 이용한 OfN 추출 과정

'821'=>feeling:[[section, i]:[affections, in, general], [class, vi]:
[words, relating, to, the, sentient, and, moral, powers]].

↓ **Prolog 술어표현 적용**

'821'=>[feeling, affection(general),
word(sentient, power(moral))].

↓ **OfN 테이블 적용**

'821'=>[feeling, affect_state, affect_state].

↓ **OfN 범주 결정**

'821'=>affect_state

그림 3 Roget 범주 #821 "feeling"의 OfN 결정과정

표 1 OfN 변환 테이블 (일부)

표제정보	OfN 범주
act	event
being	state
dimension	space
future	time
affection(general)	affect_state
word(sentient, power(moral))	affect_state

이렇게 얻은 OfN은 모든 범주가 서로 중복되지 않는 장점이 있다. 또한 이 OfN은 Roget 시소러스의 범주 분류 취지를 그대로 계승한다. 한편 표제정보가 해당 범주의 유의어 집합을 모두 대변할 수는 없다. 또한 문장에 나타나는 어휘의 다의성도 고려되어야 할 것이다. 따라서 다른 각도에서 OfN 재구성 방법을 모색할 필요가 있다.

표 2 OfN: 표제정보를 이용해 구축한 경우

범주	Roget 시소러스의 범주 번호	개수
심상	11, 14, 33, . . . , 988, 989	318
사건	21, 35, 40a, . . . , 993, 994	188
상태	1, 2, 3, . . . , 819, 965	439
시간	106, 107, 108, . . . , 512, 513	43
공간	180, 180a, 181, . . . , 999, 1000	56
합계		1044

3.2 참조정보를 이용하여 구축한 OfN

다른 사전들처럼 Roget 시소러스도 어휘에 대한 부가적인 설명이나 참조가 필요한 경우, 어휘를 다른 표제어에 참조시킨다. 이러한 참조관계들의 연결인 참조 네트워크를 특정 기준으로 걸러 내거나, 참조 간에 관계를 맺은 후 재구성하면 새로운 온톨로지가 된다. 이 결과가 참조정보를 이용한 OfN이다.

이를 위해 OfN의 각 기저범주를 Roget 범주에서 선택했다. 기저범주는 각 범주의 성격을 가장 잘 나타내는 범주이다. 이 범주의 반경은 0이다. 다음 단계에서 이 기저범주를 참조하는 Roget 범주들과 이 기저범주에서 참조하는 Roget 범주들을 취합했다. 이 범주들의 반경은 1이다. 같은 방법으로 참조정보를 탐색해서 범주의 반경을 한 단계씩 넓혔다. 탐색의 범위를 Roget 범주 누적 개수가 (1044 / 2)개를 넘지 않게 제한하였다. 탐색 결과, 이에 해당하는 반경을 3으로 정할 수 있었다. 즉, 참조정보 OfN은 각 기저범주에서 시작하는 참조 관계를 탐색해서 반경 0에서 반경 3까지 총 4단계의 범주로 구성하였다.

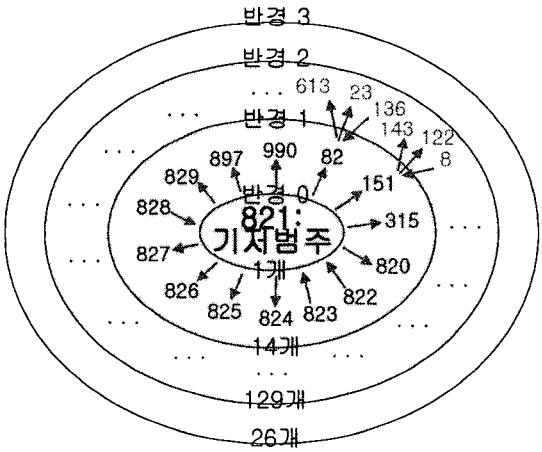

그림 4 OfN: 참조정보를 이용해 구축한 경우

이 방법으로 구성한 OfN은 Roget 시소러스의 내재적 구조를 밝힌 점에서 의미가 있다. 또한 범주의 반경을 정함으로서 해당 범주가 기저범주와 얼마나 밀접한지를 알 수 있다. 반경 정보는 다른 범주간의 우선순위 결정에도 이용할 수 있다. 한편 참조정보가 잘못된 경우 잘못된 정보가 파생시키는 오류의 범위가 크다. 또한 반경이 커질수록 범주들 사이의 교차 참조가 빈번하다. 이런 단점을 보완하기 위한 한 방법으로 표제정보 OfN과의 병합을 생각하였다.

3.3 표제 OfN과 참조 OfN의 병합

참조정보 OfN의 범주를 검토한 결과, 반경 3은 광범위해서 다른 범주와 많이 중복되었다. 그래서 이 범주 중에 다른 범주와 중복되지 않는 유일한 범주들을 추려내서 반경이 3인 범주로 택했다. 그리고 나머지 범주들을 모아 반경이 4인 범주로 정했다.

표제정보 OfN을 기반으로 참조정보 OfN을 비교한 결과, 참조정보 OfN 범주에서 누락된 Roget 범주들이 밝혀졌다. 누락된 범주들을 모아서 OfN에 반경이 5인 범주로 추가하였다. 마지막으로 표제정보 OfN을 기준으로 참조정보 OfN을 병합시켰다. 병합된 OfN은 모든 범주가 서로 중복되지 않으면서도 반경정보를 내포하는 특징이 있다.

표 3 OfN: Ontology for Narratives

OfN 범주	OfN 반경	Roget 범주	개수	
심상	0	821	1	318
	1	315, 820, 822, . . . , 829, 897	11	
	2	377, 378, 392, . . . , 981, 987	80	
	3	946, 947, 948, . . . , 975, 988	14	
	4	33, 61, 159, . . . , 944, 945	120	
	5	11, 14, 47, . . . , 986, 989	92	
사건	0	151	1	188
	1	152, 156	2	
	2	109, 142, 144, . . . , 789, 990	55	
	3	55, 145, 146, . . . , 790, 791	16	
	4	21, 63, 140, . . . , 992, 993	66	
	5	35, 40a, 41, . . . , 991, 994	48	
상태	0	7	1	439
	1	8, 240, 329	3	
	2	6, 81, 154, . . . , 550, 673	26	
	3	15, 247, 250, . . . , 596, 668	27	
	4	1, 4, 5, . . . , 747, 775	131	
	5	2, 3, 9, . . . , 819, 965	251	
시간	0	106	1	43
	1	108, 108a, 110, . . . , 120, 134	8	
	2	118, 119, 121, . . . , 138, 508	10	
	3		0	
	4	116, 117, 123, . . . , 507, 510	16	
	4	107, 114, 115, . . . , 512, 513	8	
공간	0	181	1	56
	1	189	1	
	2	182, 184, 193, . . . , 204, 1000	11	
	3	180a, 183, 200, 963	4	
	4	180, 185, 201, . . . , 967, 995	19	
	5	192, 216, 216a, . . . , 998, 999	20	
합 계				1044

3.4 OfN을 문장추상화에 활용한 예

OfN 범주 중에서, 등장인물 유형에 속하는 어휘들은 고유명사 자원[5]을 이용하여 선정하였다. 담화표지의 경우는 수사구조의 연구결과[6]를 활용하였다. 이와는 달리 시공의 변화는, 구문분석 후 문장의 구성성분간의 상호작용에 의하여 확인되는 유형인 바, 그 기본유형은 시간과 공간이다.

구문 분석기[7]로 문장 구조를 분석한 후 추상화 도구[1]에 OfN을 활용하여 보았다. 그 결과, [표 4]에 예시한 것처럼 OfN을 문장추상화에 적용하는 것이 가능함을 알 수 있었다. OfN을 문장추상화에 활용할 경우, 어휘가 복수의 범주에 해당되면 다음의 우선순위를 적용한다:

등장인물 > 심상 > 단서구 > 사건 > 상태 > 공간 > 시간

표 4 문장추상화의 예

문장	He suggested to paul that he get away for a weekend
결과	affect_state:***suggested***/ (paul<-mike) delta(space):***away***/ (paul<-mike) delta(time):***weekend***/ (paul<-mike) state:***get***/ (paul<-mike)
문장	But paul said he wasn't interested
결과	cue_phrase:***but***/ (paul<-paul) affect_state:***interested***/ (paul<-paul) state:***wasn't***/ (paul<-paul)

4. 결 론

본 논문에서는 Roget 시소러스의 온톨로지 정보를 재구성해서 OfN을 얻었다. 재구성 과정에는 두 가지 방법을 사용하였는데, 하나는 표제정보를 이용한 방법이고, 다른 하나는 참조정보를 탐색한 방법이다. 이 결과를 정리하면 다음과 같다: (1) 표제정보 OfN과 참조정보 OfN을 병합하여 얻어진 OfN은 범주간에 서로 중복되지 않으면서, 또한 각 범주는 반경 정보를 내포하는 특징을 가지고 있다. (2) 표제정보 OfN을 구성할 때 사용한 방법은, 원 온톨로지의 분류원칙을 유지하면서 새로운 온톨로지로 재편성하는데 활용할 수 있다. 또한 참조정보 OfN 구축에 사용한 방법은, 주어진 온톨로지의 내재적 구조를 밝혀내는데 활용할 수 있을 것이다. 그리고 (3) 설화문장 추상화에 OfN을 적용하는 것이 가능함을 확인하였다.

참고 문헌

[1] 김곤, 배재학. "문서요약을 위한 문장추상화." (준비중)

[2] Bae J.-H. J. and Lee J.-H. "Topic Sentence Selection with Mid-Depth Understanding." Proc. of ICCPOL, pp. 199-204, 2001.

[3] Roget's Thesaurus. http://promo.net/cgi-promo/pg/t9.cgi?entry=22&full=yes&ftpsite=ftp://ibiblio.org/pub/docs/books/gutenberg/.

[4] 양재군. "시소러스의 기계 가용화에 대한 연구." 울산대학교 석사학위논문, 2000.

[5] Proper Names Wordlist. http://clr.nmsu.edu/cgi-bin/Tools/CLR/clrcat#I4.

[6] .Knott, A. A Data Driven Methodology for Motivating a Set of Coherence Relations. Ph.D. thesis, University of Edinburgh, 1996.

[7] Link Grammar. http://www.link.cs.cmu.edu/link/.