문장추상화를 위한 Roget 시소러스 범주 재편성

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

Category Reorganization of Roget Thesaurus for Sentence Abstraction

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

요 약

본 논문에서는 Roget 시소러스의 범주를 재편성하여 문장추상화에 사용할 온톨로지를 구축하였다. Roget 시소러스의 표제정보의 범주 값과 참조정보의 범주 값을 산출한 후 가중 산술 평균을 구했다. 이 수치를 토대로 OfN(Ontology for Narratives)을 구성하였다. 최종적으로 Roget 시소러스와의 비교를 통하여 OfN을 확정하였다. 이렇게 하여 얻어진 OfN을 설화 문장추상화에 적용하여 이 온톨로지가 유의함을 확인하였다.

1. 서 론

인터넷의 대중화로 온라인화 된 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 설화문장을 추상화시키는데 사용할 목적으로 일곱 가지 범주로 구성된 OfN(Ontology for Narratives)[2]을 재구성하였다. 재구성에는 Roget 시소러스를 데이터로 이용하였다. 이 온톨로지의 범주는 다음과 같은 7가지 유형(Type)이 포함되어 있다: (1) 등장인물 - Character, (2) 심상 - Affect State, (3) 사건 - Event, (4) 상태 - State, (5) 공간 - Space, (6) 시간 - Time, (7) 담화 표지 - Discourse Markers.

표 1 OfN의 7가지 범주

범 주	의 미
Character (등장인물)	이야기에 등장하는 사람이거나 혹은 의인화 가능한 존재이다.
Affect State (심상)	등장인물의 감정 상태이다.
Event (사건)	이야기에서 어떤 중요한 일의 발생이다.
State (상태)	등장인물이나 사물의 감정 상태를 제외한 나머지 상황이다.
Space (공간)	공간
Time (시간)	시간
Discourse Markers (담화표지)	화자의 의도를 내포하는 단서구(Cue Phrase)이다.

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.

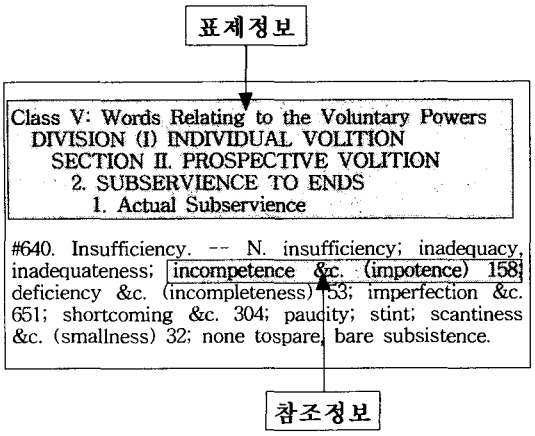

그림 1. 로젯 시소러스의 구조

3. 온톨로지 재구성

전처리된 Roget 시소러스[4]를 새로운 온톨로지로 재분류하기 위해서 Roget 시소러스 표제정보의 범주 값과 참조정보의 범주 값을 산출하였다. 산출된 각 범주 값의 가중치 평균을 근거로 OfN 범주를 결정하였다. 각 범주 값을 산출하는 구체적인 방법은 (1) 표제정보를 계층구조 순위의 특수성과 일반성에 따라 수치화 된 값으로 표현하였다. 표제정보내의 일반적인 분류정보에는 작은 값을 부여하였고 특수한 정보일수록 더 큰 값을 지정하였다. (2) 같은 개념을 참조정보에도 적용하여 반경 값이 작을수록 큰 값을 지정하였다.

각각의 범주 값을 병합하는 과정에서는 표제정보와 참조정보의 중요도를 조절하기 위해 가중치를 적용하였다. 가중치를 조절하는 실험을 통하여 적절한 가중치를 찾아내고 Roget 시소러스와의 비교를 통하여 선택한 값이 유의함을 확인할 것이다.

3.1 대표성과 선호도

문장 추상화를 위한 온톨로지인 OfN의 범주는 문장에서의 중요도에 따라 우선 순위를 아래와 같이 정할 수 있다.

등장인물>심상>단서구>사건>상태>공간>시간

우선 순위에 따라 범주에 적용할 값을 서로 다르게 설정해야 할 것이다. 이 값을 선호도라고 한다. 선호도는 우선 순위가 높을수록 큰 값을, 우선 순위가 낮을수록 작은 값을 지정한다. OfN 범주 중에서도 특히 주목해야할 범주들이 있을 것이다. 이들을 주 범주로 분류하고 나머지를 부 범주로 분류하였다. 주 범주에 더 큰 비중을 두기 위해서 원래 설정된 값에 1을 더했다. 이와 같은

과정을 통해 설정된 초기 선호도는 등장인물:7, 심상:6, 사건:5, 상태:3, 공간:2, 시간:1 이다. 이 선호도는 OfN 범주 결정 실험을 통하여 유의한 값으로 변경될 것이다.

표 2. 초기 선호도

범 주	등장인물	심상	사건	상태	공간	시간
선호도	7	6	5	3	2	1

OfN 범주 결정에 변인으로 작용하는 또 다른 요인으로는 표제정보와 참조정보가 있다. 이들 정보에도 각각 값을 설정하며 이를 대표성이라 하였다.

표 3. 대표성과 선호도

분 류	주 범 주			부 범 주		
OfN범주	등장인물	심상	사건	상태	공간	시간
우선순위	1	2	3	4	5	6
선 호 도	7	6	5	3	2	1
표제정보 대 표 성	표제정보의 특수성과 일반성 (하위 계층일 수록 큰 값)					
참조정보 대 표 성	참조정보의 전체 반경(8)에 대한 보수 값 (반경이 작을수록 큰 값)					

3.2 표제정보의 범주 값

Roget 시소러스의 말단 범주 값을 얻기 위한 한가지 방법으로 해당 범주를 설명하는 정보중 하나인 표제정보를 수치화 하는 방법을 모색하였다. 이를 위해 모든 표제정보와 표제어를 취합하고 각 표제정보와 표제어를 Prolog의 술어형태로 치환한

후 OfN 테이블을 적용해서 OfN 형태의 표제정보를 얻었다.

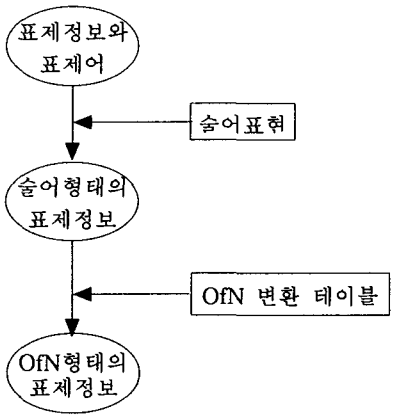

그림 2. 표제정보 변환 과정

표제정보의 범주 값은 표제정보 계층의 특수성과 일반성에 영향을 받는다. 표제정보내의 일반적인 분류정보에 대해서는 작은 값을 부여하였고 특수한 정보일수록 더 큰 값을 지정하였다. 이 값은 표제정보의 대표성이다. 또한, 표제의 각 계층 정보를 OfN 범주로 변환한 값은 표제정보의 선호도이다. 이러한 대표성과 선호도를 곱한 수치가 표제정보의 범주 값이다.

[표 4]는 Roget 범주 #411의 표제정보이다. 이 표제정보를 OfN 각 범주별 대표성으로 변환하면 [표 5]를 얻을 수 있다. 이때 각 항목 표제정보에 대한 선호도이다.

표 4. 표제정보의 대표성과 선호도

선호도		3	3	6	6	3	3
대표성	7	6	5	4	3	2	1
표제정보	cry	상태	상태	심상	심상	상태	상태

표 5. 표제정보 OfN의 대표성

OfN 범주	대 표 성						
	7	6	5	4	3	2	1
등장인물							
심 상				6	6		
사 건							
상 태	3	3	3			3	3
공 간							
시 간							

[표 5]를 행렬로 변환시키면 다음과 같은 행렬을 얻을 수 있다.

$$H=\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 3 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

대표성의 개수를 l이라 할 때 행렬 H의 원소 a_{ij}의 첨자 i는 OfN 범주의 분류이고 j는 대표성 즉, 표제정보의 계층구조 레벨이다. 즉, 원소 a_{ij}의 값은 OfN 각 범주의 선호도이다. 이때, OfN 각 범주 값은 대표성과 선호도의 곱에 대한 합이라고 볼 수 있다.

$$H_i = \sum_{j=1}^{l} a_{ij} \qquad \text{(식 1)}$$

3.3 참조정보의 범주 값

다른 사전들처럼 Roget 시소러스도 어휘에 대한 부가적인 설명이나 참조가 필요한 경우, 해당 어휘를 다른 표제로 참조시킨다. 이러한 참조관계들을 탐색하면 탐색기준에 따라 Roget 시소러스를 재구성 할 수 있을 것이다. 이를 위해 OfN의 각 기저범주를 Roget 범주에서 선택하고 이 범주의 반경을 0으로 정한다. 다음 단계에서 이 기저범주가 참조하거나 기저범주를 참조하는 Roget 범주들을 취합했다. 이 범주들의 반경은 1이다. 같은 방법으로 모든 Roget 범주를 탐색하였다. 탐색 결과 반경 7 안에서 모든 Roget 범주를 탐색 할 수 있었다[5].

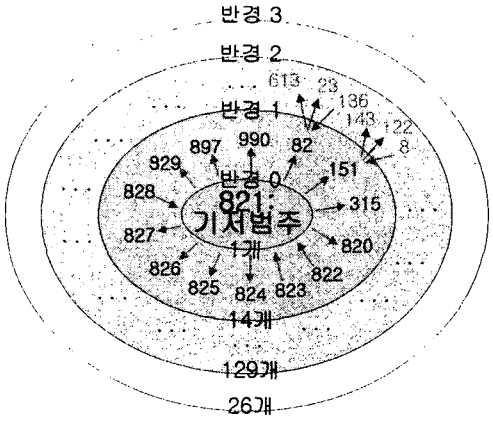

그림 3. OfN: 참조정보의 탐색

참조정보의 범주 값은 기저범주에 대한 반경에 영향을 받는다. 기저범주의 반경은 0이고 이 기저범주와 가까운 범주일수록 작은 반경 값을 갖는다. 그러므로 반경이 작을수록 범주 값에는 크게 적용되어야 할 것이다. 반경을 7까지 확장하면 모든 Roget 범주를 포함할 수 있으므로 반경 값에 대한 8의 보수가 참조정보의 대표성이다. 참조정보의 범주 값은 각 OfN 범주에 지정한 선호도와 대표성의 곱이다.

참조정보의 대표성을 r이라 하고 선호도를 p라 하면 각 OfN 범주 i에 대한 참조정보의 범주값은 다음 식으로 표현할 수 있다.

$$C_i = r_i \times p_i \qquad \text{(식 2)}$$

표 6. 로젯 범주 #411에 대한 참조정보

분 류	OfN범주	선호도	반경	대표성	비 고
Major	등장인물	7	5	3	참조정보의 선호도는 각 참조정보 반경의 최대반경에 대한 보수
	심 상	6	4	4	
	사 건	5	5	3	
Minor	상 태	3	5	3	
	공 간	2	6	2	
	시 간	1	5	3	

3.4 단위 대표성에 대한 선호도

표제정보의 범주 값과 참조정보의 범주 값을 하나의 수치로 표현하기 위해 단위 대표성에 대한 선호도를 산출한다. 대표성은 표제정보 계층구조의 레벨로 볼 수 있다. 이 대표성을 ℓ이라 하면 표제정보의 단위 대표성에 대한 선호도는 다음 식으로 표현 할 수 있다.

$$H_{ui} = \frac{H_i}{\text{대표성의 합}} = \frac{H_i}{\text{등차수열 } l\text{의합}} \qquad \text{(식 3)}$$

마찬가지로 참조정보의 단위 대표성에 대한 선호도는

$$C_{ui} = \frac{C_i}{\text{등차수열 } l\text{의합}} \qquad \text{(식 4)}$$

이다.

3.5 가중 산술 평균

표제정보 범주 값과 참조정보 범주 값을 하나의 수치로 묶는 과정에서 필요에 따라 어느 한 쪽에 더 큰 비중을 줄 수 있을 것이다. 그러기 위해서 각 단위 대표성에 대한 선호도에 표제정보와 참조정보간의 중요도를 고려해서 가중 산술 평균을 구한다. 이 값이 최종 범주 값이다.

표제정보에 가중치 2를 설정하고 참조정보에 가중치 1을 설정한다면 표제정보와 참조정보의 가중 산술 평균은 다음과 같다.

$$S_i = \frac{H_{ui} \times 2 + C_{ui} \times 1}{3} \qquad \text{(식 5)}$$

각 i 에 대한 S_i의 값을 구한 후 가장 큰 값을 가지는 범주를 해당 Roget 범주의 OfN 범주로 결정한다.

3.6 OfN 범주 결정 예

Roget 범주 #411의 OfN 범주 결정을 전술한 과정을 따라 예시하면 다음과 같다.

· 표제정보의 범주 값 (상태 범주)

식 1에 따라서

$H_{상태} = (6+5+2+1) \times 3 = 42$

식 3에 따라서

$H_{u상태} = \frac{42}{28} = 1.5$

· 참조정보의 범주 값 (상태 범주)

식 2에 따라서

$C_{상태} = (8-5) \times 3 = 9$

식 4에 따라서

$$C_{u_{상태}} = \frac{9}{36} = 0.25$$

· 가중 산술 평균

식 5에 따라서

$$S_{상태} = \frac{H_{u_{상태}} \times 2 + C_{u_{상태}} \times 1}{3} = 1.0833$$

다른 OfN 범주에 대해서도 같은 계산을 수행한 후 각 OfN 범주의 가중치 평균을 비교하여 가장 큰 값을 가지는 범주를 OfN 범주로 결정한다. 심상 범주 값이 1.2222이고 상태 범주 값이 1.0833이므로 Roget 범주 #411은 OfN 범주 중에 심상 범주로 결정한다.

표 7. Roget 범주 #411의 OfN 범주 판단 예

OfN범주	가중산술평균	결 과
등장인물	0.1944	심상범주의 가중 산술평균값이 가장 크므로 Roget 범주 #411은 OfN 범주의 "심상" 범주로 결정한다.
심 상	1.2222	
사 건	0.1389	
상 태	1.0833	
공 간	0.0370	
시 간	0.0278	

3.6 OfN 범주 결정 시스템

OfN 범주 결정의 효율을 높이고 자동화하기 위하여 OfN 범주 결정 시스템을 구현하였다. 이 시스템은 표제정보와 참조정보의 범주 값을 산출하고 각 결과에 대한 가중 산출 평균을 구해내고, 그 값에 따라 OfN 범주를 결정하는 전 과정을 처리한다. 또한, 범주 결정에 변인으로 작용하는 OfN 범주의 선호도와 참조정보의 최대 반경 값, 표제정보와 참조정보의 중요도를 조절하는 가중치를 변경할 수 있게 설계하였다. 마지막으로 계산과정을 검토할 수 있도록 OfN 범주 결정 과정의 상세 정보를 출력하도록 하였다.

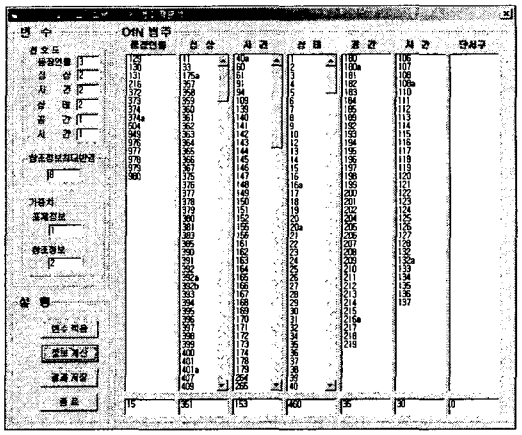

그림 4. OfN 범주 결정 시스템

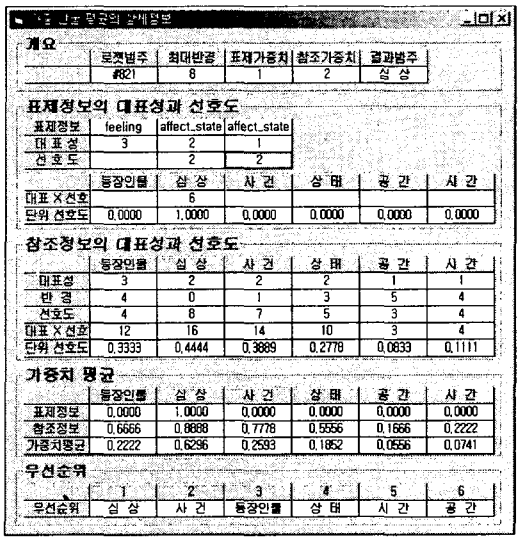

그림 5. OfN 범주 결정 상세정보

OfN 범주 결정 시스템을 이용해서 각종 변인을 조절하였다. 이 실험 결과 온톨로지의 표본을 Roget 시소러스와 비교한 결과, 표제정보 범주 값과 참조정보 범주 값의 가중치는 1:2로 설정하는 것이 가장 바람직하였다. 또한, 범주 분포에 있어서 가장 유의한 선호도는 다음 표와 같았다.

표 8. 최종 선호도

범 주	등장인물	심상	사건	상태	공간	시간
선호도	3	2	2	2	1	1

표 9는 확정한 변인을 적용해서 Roget 시소러

스을 재분류한 OfN이다.

표 9. OfN : Ontology for Narratives

범 주	Roget 범주	개 수
등장인물	129, 130, . . . , 979, 980	15
심 상	11, 33, . . . , 999, 1000	351
사 건	40a, 60, . . . , 994	153
상 태	1, 2, . . . , 819, 965	460
공 간	180, 180a, . . . , 218, 219	35
시 간	106, 107, . . . , 136, 137	30
합 계		1044

3.7 OfN을 문장추상화에 활용한 예

OfN의 유용성을 검토하기 위해서 구문 분석기 [6]로 문장 구조를 분석한 후 추상화 도구에 온톨로지를 적용하여 보았다. OfN 범주 중에서 담화표지의 경우는 수사구조의 연구결과[7]를 활용하였다. 그 결과, [표 10]에 예시한 것처럼 OfN을 문장 추상화에 적용하는 것이 가능함을 알 수 있었다.

표 10. 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 Roget 시소러스 표제정보의 범주 값과 참조정보의 범주 값을 산출하였다. 산출된 각 범주 값의 가중치 평균을 근거로 OfN 범주를 결정하였다. 또한, OfN 범주 결정의 효율을 높이고 자동화하기 위하여 OfN 범주 결정 시스템을 구현하였다. 이 시스템은 범주 결정의 자동화뿐만 아니라 범주 결정에 변인으로 작용하는 요인들을 조절할 수 있게 함으로써 보다 유의한 OfN을 얻을 수 있었다. 재구성한 OfN을 문장 추상화에 적용한 실험에서는 추상화가 가능함을 알 수 있었다.

본 논문에서 이용한 온톨로지 재구성 방법은 서로 다른 분류 기준을 하나의 분류 기준으로 묶을 수 있었다. 또한 범주 결정 과정에서 변인으로 작용하는 요인들을 조절 할 수 있게 하고, 그 결과 표본을 원 온톨로지와 비교해 봄으로써 보다 유의한 온톨로지 재구성이 가능하였다. 즉, 필요한 온톨로지를 얻기 위한 방법으로는, 다른 온톨로지의 범주 정보를 여러 가지 분류 기준에 따라 추출한 후 그 분류 기준 사이의 조율을 거친다면 새로운 온톨로지 재구성이 가능할 것이다.

참고 문헌

[1] 김곤, 배재학. "문서요약을 위한 문장추상화." 한국정보처리학회, 제 9권, 제 1호, pp.531-534, 2002.

[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] 양재군, 배재학. "온톨로지 정보를 이용한 범주 재편성: Roget 시소러스의 경우." 한국정보처리학회, 제 9권, 제 1호, pp.515-518, 2002.

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

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