문장분석에 활용할 종합적인 사용자 인터페이스

김명수, 김민찬, 배재학
울산대학교 컴퓨터 · 정보통신공학부
{zalhanda, tomatuli, jhjbae}@ulsan.ac.kr

An Integrated User Interface for Sentence Analysis

Myoung-Soo Kim, Min-Chan Kim, Jae-Hak J. Bae
School of Computer Engineering and Information Technology
University of Ulsan

요약

한 문장을 기계로 분석하기 위해서는 단편적인 정보가 아닌 그 문장에 대한 종합적인 정보를 필요로 한다. 이러한 문장분석 정보들의 관리나 처리작업에는 자동적 처리가 가능한 부분들도 있으나, 사람이 수작업으로 해결해야 할 부분들도 있다. 본 논문에서는 개별적인 문장분석 작업들을 통합하여 자동화하고, 그 분석의 체계성과 정확성을 높이고자 종합적인 사용자 인터페이스를 설계 및 구현하였다.

1. 서 론

문장분석에 활용할 도구와 자원들은 여러가지가 있다. 구문분석기나 유의어사전 등이 대표적인 예이다. 이러한 도구와 자원은 개별적으로 존재한다. 이는 문장분석에 필요한 정보들을 종합적으로 관리 및 처리하기에 어려움이 있음을 의미한다. 그러므로 문장분석의 효율을 높이기 위해서는 도구와 자원을 통합하고, 이들의 입·출력정보를 관리해줄 종합적인 인터페이스가 필요하다.

이에 본 논문에서는 문장을 좀더 체계적이고 편리하게 분석할 수 있도록 각 기능의 도구와 자원들을 통합, 관리할 종합적인 인터페이스를 설계 및 구현하였다. 구현한 종합적인 인터페이스에는 다음의 도구와 자원이 적용되었다: (1) 구문분석기가 밝혀낸 문장의 통사구조, (2) Roget 시소러스(Roget Thesaurus)[1]의 범주정보, (3) OfN(Ontology for Narratives) 정보, (4) 번역정보가 그것이다.

2. 종합적인 인터페이스에 적용된 도구와 자원

문장분석에 활용할 종합적인 사용자 인터페이스를 구현하기 위해서 구문분석기로 LGPI+를 사용하였고, 유의어 사전으로 Roget 시소러스를 사용하였다. 그리고 구문분석 후 분리된 어휘들을 Roget 시소러스에서 검색하기 위해 어형변화 처리과정을 추가하였다. 또한 어휘들에 대한 OfN 정보도 추가하였다.

2.1 인터페이스 실행예

다음 예문을 생각해보자: I do not know how to get past the pain. 다음의 [그림 1]은 종합적인 인터페이스에서 예시한 문장을 처리한 모습이다.

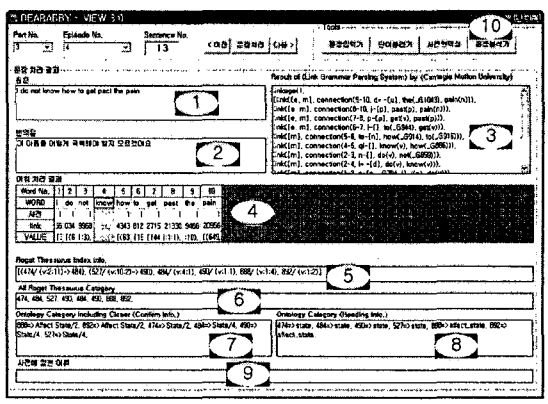

그림 1. 문장분석용의 종합적인 사용자 인터페이스

2.2 실험을 위한 설화 말뭉치 - Dear Abby

종합적인 인터페이스가 관리하는 말뭉치의 장르는 설화문이다. 이는 Dear Abby[5]라는 상담형식의 글이며, 그 중 40여편의 상담내용을 발췌하여 한 문장씩 처리하였다. [그림 1]의 ①은 말뭉치에서 선택한 문장을, ②는 미리 준비한 번역문을 보여준다.

2.3 LGPI+

본 논문에서는 문장의 통사구조를 밝히기 위한 도구로 LGPI+를 사용하였다. LGPI+는 SWI-Prolog[6] 패키지인 Link Grammar Parser Interface[4]를 확장시킨 것이다. LGPI+는 Link Grammar Parser[3]에 대한 SWI-Prolog API(Application Program Interface)를 제공한다. LGPI+는 6만 어형을 수록한 사전을 내장하고, 다양한 구문구조를 처리할 수 있다. 이 사전은 필요에 따라 확장이 가능하다. 입력문장에 대한 LGPI+ 구문분석 결과는, 표식고리(Labeled Link)의 집합으로 통사구조가 표현된다. 표식고리는 한 쌍의 단어를 연결함과 아울러 그것들의 문법적인 기능을 표시한다.

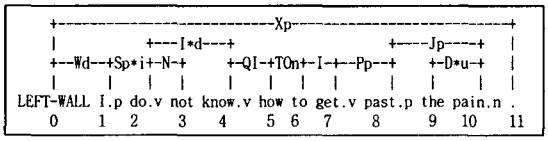

그림 2. Link Grammar Parser의 처리예

```
linkage(1,
[link([e, m], connection(9-10, d* -[u], the(_G1043),
      pain(n))),
 link([e, m], connection(8-10, j-[p], past(p), pain(n))),
 link([e, m], connection(7-8, p-[p], get(v), past(p))),
 link([e, m], connection(6-7, i-[], to(_G944), get(v))),
 link([m], connection(5-6, to-[n], how(_G914),
 to(_G916))),
 link([m], connection(4-5, qi-[], know(v), how(_G886))),
 link([m], connection(2-3, n-[], do(v), not(_G859))),
 link([m], connection(2-4, i* -[d], do(v), know(v))),
 link([m], connection(1-2, s-[p, _G794, i], i(p), do(v))),
 link([m], connection(0-1, w-[d], left-wall(_G764), i(p))),
 link([], connection(0-11, rw-[], left-wall(_G734),
      right-wall(_G736)))],
s(np(i/1), vp(do/2, not/3, vp(know/4,sbar(whadvp(how/5),
  s(vp(to/6, vp(get/7, pp(past/8, np(the/9, pain/10))))))))))
).
```

그림 3. 예시한 문장에 대한 LGPI+ 처리결과

LGPI+는 Link Grammar Parser의 [그림 2]와 같은 결과를 기계 가독형으로 바꾸어 [그림 3]과 같이 출력, 저장한다.

[그림 3]에서 4번째 줄은 past과 pain 간의 연결관계를 보여준다. 8-10은 문장에서 past가 나타나는 순서가 8번째이고, pain은 10번째에 나타남을 보여준다. j-[p]는 past과 pain 간의 연결유형을 나타낸다. 그리고 past(p), pain(n)에서 (p)와 (n)은 해당어휘가 문장에서 전치사와 명사로 각각 사용됨을 나타낸다. [그림 3]의 아래쪽에 표시된 정보는 문장에 나타나는 구성성분들을 구 단위로 묶은 것이다. LGPI+로 분석한 결과는 [그림 1]의 ③에 표시된다.

2.4 어형변화 처리

문장의 통사구조 분석을 통하여 문장을 구성하는 어휘들을 분리하였다. 그리고 어휘의 복수형이나 과거형, 불규칙 동사와 같은 변형어휘는 어형변화 처리를 통하여 원형어휘를 찾는다. 다시 이것은 Roget 시소러스의 색인정보를 검색하는데 사용된다.

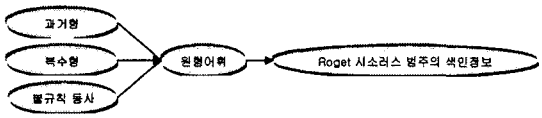

그림 3 어형변화 처리

LGPI+에서 분리된 어휘는 어형변화 처리를 거친 후 Roget 시소러스 색인에서 검색된다. 이 과정은 [그림 1]의 ④에 출력된다. 이때 원형어휘는 '사전' 항목에 1로 표시되며, 변형어휘는 '사전' 항목에 2로 표시된다. 만약 색인에 없는 어휘라면 ⑨에 그 어휘가 출력된다. 또한 ④에는 각 어휘에 대한 Roget 시소러스의 색인정보를 포함하고 있다. 즉, [그림 1]에서와 같이 get이라는 어휘를 선택하면 ⑤에 Roget 시소러스의 색인정보가 출력된다.

2.5 Roget 시소러스와 OfN

Roget 시소러스는 총 6개의 의미분류에 기초한 강(Class)으로 구성되며 각 부류는 하부에 부(Division), 과(Section) 등의 계층구조로 세분화 되어있다. 각 계층은 저마다의 표제정보를 가지고 있으며 계층구조의 말단에는 총 1044개의 범주가 존재한다. 각 범주에는 품사별 유의어 목록이 나열되어 있다.

설화 추상화용 온톨로지(Ontology, 존재론), OfN은 다음의 7가지 범주로 구성된다: 등장인물(Character), 심상(Affect State), 사건(Event), 상태(State), 시간과 공간의 변화(Delta-{Time, Space}), 담화표지(Discourse Marker). 설정된

OfN을 구축하기 위해서 먼저 Roget 시소러스의 범주를 심상, 시간과 공간, 사건, 그리고 상태 등으로 재편성된다. 이때, 등장인물 유형에 속하는 어휘들은 고유명사 자원[8]을 이용하여 선정하고, 담화표지의 경우는 수사구조의 연구결과[7]가 활용되었다. 이와는 달리 시공의 변화는 구문분석 후 문장 구성성분 간의 상호작용에 의하여 확인되는 유형인 바, 그 기본유형은 시간과 공간이다.

표 6. 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	

OfN 구축을 위한 Roget 시소러스의 범주 재편성 정보는 [그림 1]의 ⑦과 ⑧에 나타난다. 그 과정은 ⑤의 한 어휘(예시: get)에 대한 Roget 시소러스 색인정보에서 Roget 범주들을 추출하여 ⑥에 나열한다. 이때, Roget 표제정보에서 찾은 범주는 ⑧에 표시하고, OfN에서 찾은 범주는 범주 우선순위와 반경이 작은 순서에 따라 ⑦에 출력한다. ⑦의 출력형식을 설명하면 다음과 같다. 144=>Event/2에서 144는 선택한 어휘가 참조하는 Roget 범주이고, Event는 OfN의 사건(Event)범주이며, 마지막에 표시된 2는 OfN 사건범주의 기저범주에 대한 반경이다. 범주반경은 기저범주를 0으로 하고, 해당범주와 기저범주 간의 참조단계에 따라 나누어 놓은 값이다. 즉, 반경이 짧을수록 범주의 성격을 잘 나타낸다.

3. 처리과정

위에서 설명한 도구와 자원들이 통합된 종합적인 인터페이스의 처리과정을 간략하게 살펴보면 다음과 같다. 저장되어 있는 말뭉치에서 한 문장을 선택하면 원문과 번역문이 표시된다. 그리고 원문을 LGPI+로 처리하여 저장된 결과를 표시한다. 다음으로 분리된 어휘들의 어형을 판별, 처리하여 Roget 범주의 색인정보를 찾는다. 이 색인정보에서 추출한 Roget 범주에 대한 OfN범주를 출력한다. 이를 도식화하여 나타내면 [그림 4]와 같다.

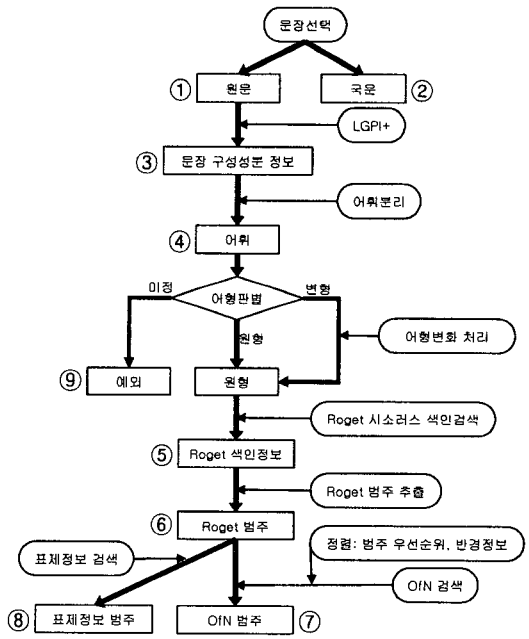

그림 4. 종합적인 인터페이스에서 문장 처리과정

새로운 문장을 입력하여 분석하고자 하는 경우에는 문장분석 모듈을 호출하는 버튼(⑩)을 선택한다.

4. 결론

문장분석에 활용할 종합적인 사용자 인터페이스의 설계 및 구현을 통하여 다음과 같은 결과를 얻었다:

(1) 개별적인 문장분석 도구와 자원에서 분석정보를 얻고, 이를 수작업으로 조합, 처리하였던 과정을 하나의 도구 안에서 해결할 수 있도록 하였다.

(2) 구문분석기 LGPI+의 문장분석 결과와 Roget 시소러스 및 OfN의 관련 어휘정보를 효과적으로 파악할 수 있게 하였다.

(3) 문장추상화 작업에 종합적인 문장분석 결과를 활용할 수 있도록 하였다.

참고문헌

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

[2] 양재군, 배재학. "온톨로지 정보를 이용한 범주 재편성: Roget 시소러스의 경우." (준비중)

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

[4] SWI-Prolog -- Link Grammer Parser Interface. http://gollem.swi.psy.uva.nl/twiki/pl/bin/view/Library/LinkGrammer/.

[5] DearAbby. http://www.dearabby.com/.

[6] What is SWI-Prolog. http://www.swi-prolog.org/.

[7] 배재학. "언어학적인 방법론을 취하는 자동 문서요약에 대한 연구." 공학 연구논문집, 제 29권 2호, pp.351-363, 울산대학교, 1998.

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