가도까와(かどかわ) 시소러스를 이용한 구문관계에서 의미관계로의 사상(寫像) 규칙

박정혜° 강신재 이종혁

포항공대 정보통신대학원° 포항공대 컴퓨터공학과

{erica, sjkang, jhlee}@postech.ac.kr

Mapping Rules from Syntactic Relations to Thematic Relations by Using Kadokawa(かどかわ) Thesaurus

Jung-Hye Park° Sin-Jae Kang Jong-Hyeok Lee

Dept. of Graduate School for Information Technology, POSTECH°

Dept. of Computer Science & Engineering, POSTECH.

요약

본 논문에서는 의미분석을 위해서 구문관계와 의미관계를 자동으로 사상하는 규칙을 구축한다. 5 만개의 패턴을 수작업으로 사상해서 학습데이터로 만들고 이의 분석을 통해 규칙을 구축했다. 규칙에서는 의미역 결정을 위해서 가도까와 시소러스를 이용하는데, 본 논문에서는 한일 기계번역사전을 이용하여 추출한 구문 패턴을 대상으로 실험한 결과, 정확률 90%, 적용율 93.5%를 얻었다.

1. 서론

본 시스템은 의미분석을 위한 기초 작업으로서 구문관계에서 의미관계로의 사상 규칙을 구축한다. 의미분석은 기본적으로 의미역 이론(thematic theory)에 따른 선택제약(selectional restriction)과 논항구조(case frame)에 대한 정보에 의존하고 있다[3]. 의미역(thematic role, case relation)은 술어(predicate)가 논항(argument)에 부여하는 의미론적인 역할을 말한다. 따라서 의미역 결정을 위해서는 논항과 각각의 의미역에 관한 정의를 명확히 해야한다.

본 연구는 의미분석에서 핵심이 되는 작업으로서 선택제약과 논항구조에 대한 정보를 제공하여 단어의 의미 중의성 해소(word sense disambiguation)에 기여를 한다. 아울러 가도까와 시소러스를 이용하므로 시소러스의 개념들간의 의미관계 또한 파악할 수 있어서 온톨로지(Ontology)를 구축하는 기초 작업이 될 수 있다.

2 장에서 기존 연구와 문제점을 살펴보고 3 장에서는 논항과 의미역을 정의한다. 그리고 4 장에서는 알고리즘을, 5 장에서는 실험에 관해서 살펴보고, 6 장에서 결론을 내리고자 한다.

2. 기존 연구 및 문제점

의미관계의 결정에 대한 연구는 그리 많은 편이 아니다[3, 6]. Daniel(2000)[6]에서는 의미역이 태깅된 코퍼스(FrameNet) 학습에 기반한 통계적인 방법을 이용했다. 이 코퍼스는 의미역의 태그셋(frame elements)을 정의하고 이를 수작업으로 태깅한 49,013 문장을 포함하고 있다.

양단희(1998)[3]에서는 기계학습을 통해 격 원형성(prototypicalities)을 획득하는 방법을 이용했다. 원형성은 각 단어의 격 개념이 어느 정도 전형적인 예가 되는가에 대한 표현이다. 예를 들면, 명사 '수레' 가 조사 '로' 와 결합할 때 원형성은 {(목적지격으로 0.936), (자격으로 0.963), (도구격으로 2.321)}, 용언 '도주하다' 의 원형성은 {(목적지격으로 0.494), (자격으로 0.371), (도구격으로 0.509)} 이다. 이를 통해 '수레로 도주하다' 의 '수레로' 가 도구로 사용되었음을 알아낸다. 하지만 명사와 용언 이외의 요소가 의미역을 결정하는 '닻을 올리러 배로 갔다' 같은 문장에서는 정확한 의미역을 찾아낼 수 없다.

의미역은 논항에 부여하는 것이므로 의미역을 결정하기 위해서는 어떠한 요소를 논항으로 볼 것인가 그리고 논항에 어떠한 의미역을 할당할 것인가가 중요하다. 세종전자사전(1999)[1]에서는 논항을 서술어가 통사적으로 요구하는 필수적인 논항 뿐만 아니라 의미적인 필수항과 자주 쓰이는 부가항까지 포함해서 정의하고 있다. 그리고 대상, 행위주, 경험주, 동반주, 처소, 출발점, 도착점, 방향, 도구, 이유, 수령주, 자격, 기준치, 정도와 같은 14 개의 의미역을 정의했다. 세종전자사전에서는 구별가능한 의미역을 최대한 구분하여 기술하고 추후에 필요 없으면 구분했던 의미역을 하나로 통합하는 기준을 세우고 있다. 이는 의미역을 정의하는 것 또한 매우 힘든 작업이

라는 것을 말해준다.

조일영(1998)[5]에서는 'NP'로에 관한 의미역만을 논하는데도 무려 15 개의 의미역을 정의하고 있다. 동사에 의한 의미역과 명사에 의한 의미역으로 나누어 두단계에 걸쳐 의미역을 결정한다. 동사에 의해서 방법, 결과, 처소, 원인 중 하나가 결정되며 동사에 의한 의미역의 하위 집합격인 명사에 의한 의미역으로 도구, 수단, 재료(방법), 경로, 방향, 지향점(처소) 등이 결정된다. 그러나 동사에 의한 의미역은 결정적이지 못하다. "버스로 학교에 가다'의 '버스로'는 동사에 의해 방법이, 명사에 의해 수단이 할당되며, "산길로 학교에 가다"의 '산길로'는 동사에 의해 처소가, 명사에 의해 경로가 할당된다. '가다'가 두 가지 추상적인 의미역을 가지는 이 예는 조일영(1998)에 따르면 모순이라 할 수 있다.

고영근, 남기심(1993)[2]에서는 보어를 '되다/아니다'와 느낌, 감정형용사 앞에 오는 주어가 아닌 '이/가'와 결합하는 요소로 정의하고 있다. 반면 이홍식(1996)[4]은 보어를 술어가 요구하는 필수논항으로 엄격하게 정의하고 조사에 따라 '이/가', '와', '에', '로'로 나누고 있다. 그리고 주어에는 행위주, 대상, 경험주가 할당될 수 있고 목적어와 보어 ' 이/가' 에는 대상이, 보어 '와'에는 동반주가, '로'에는 도달점, 재료, 결과가, '에'에는 처소, 원인, 대상 등이 각각 할당될 수 있다고 했다.

3. 논항과 의미역

기존연구에서 살펴본 것처럼 논항의 정의는 어렵다. 본 논문에서는 논항을 주어, 목적어, 보어, 부사어로 규정한다. 보어는 '되다/아니다' 앞에 오는 주어가 아닌 성분과 '무섭다'와 같은 심리형용사, '싫다'와 같은 느낌, 감정형용사 앞에 오는 주어가 아닌 성분으로 정의한다. 그리고 부사어는 주어, 목적어, 보어가 아닌 '체언 + 부사격 조사'로서 술어의 필수적인 요소이어야 한다.

의미분석을 위해서는 논항 뿐만 아니라 부가항에 대해서도 지배소와의 의미관계를 밝혀야 한다. 그러므로 구별하기 힘든 필수적인 요소와 부가적인 요소를 구별하여 보어와 부사어로 규정하는 것이 의미분석에서는 큰 의미가 없기 때문에 본 논문에서는 보어를 위와 같이 정의하였다.

그리고 의미역은 16 개로 정의하였다. 그 정의는 아래와 같다.

대상(Theme) : 동작의 대상, 동작이나 과정에 영향을 입는 요소

행위주(Agent) : 행위를 야기시키는 인물, 행위의 주체

경험주(Experiencer) : 사건의 심리적인 주체, 경험의 주체

동반주(Companion) : 행위주나 대상과 동등한 지위를 가지는 요소

처소(Location) : 일정한 면적을 가진 이동의 의미가 없는 요소

도착점(Goal) : 이동 후에 이르게 되는 곳, 변화, 구분, 판단의 결과

출발점(Source) : 동작이 이루어지거나 시작하는 시점, 어떤 행위의 유래

방향(Direction) : 행동이 진행되는 방향

도구(Instrument) : 행동이 가능하도록 하는 직접적인 장치나 수단

재료(Material) : 결과물의 요소

경로(Path) : 경유지

자격(Apparaisee) : 적합한 능력을 갖춘 인물이나 인물의 성향을 가진 명사

이유(Reason) : 사건의 원인, 가치판단을 내재하지 않은 표준

수령주(Recipient) : 혜택을 받는 사람, 단체, 사물

기준치(Criterion) : 비교시 기준이 되는 것

정도(Degree) : 구체적인 수량이나 가격의 차이

세종전자사전(1999)에서 구분 가능성을 제시한 재료와 경로만을 추가하였다. 재료와 경로의 정의는 조일영(1998)을 따랐다.

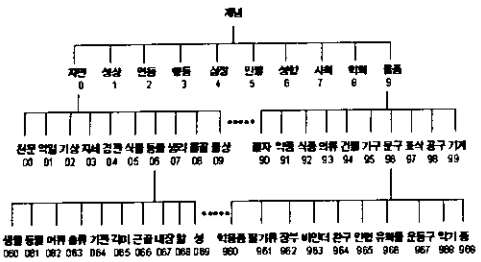
그리고 각각의 문법관계에 따른 의미역은 아래와 같다.

문법관계		의미역
주어		행위주, 대상, 경험주, 수령주
목적어		대상
부사어	어	수령주, 행위주, 기준치, 출발지, 대상, 이유, 처소, 도착점
	로	재료, 경로, 방향, 정도, 이유, 자격, 도착점, 도구
	와	동반주

[표 1] 문법관계에 따른 의미역

4. 알고리즘

[표 2]는 패턴 변환의 한 예이다. 지배소와 의존소에 세자리의 숫자를 볼 수 있는데 이는 가도까와 시소러스의 의미코드이다[7]. 가도까와 시소러스는 의미분류체계에 따른 의미코드를 포함하고 있다[그림 1]. 4 단계의 계층구조를 이루고 있는데, L1 - L100 까지의 각 계층의 개념은 10 개의 하위개념으로 분류되고 개념은 숫자로 표현한다. 용언 '내리다'를 예로 살펴보면 '비가 내리다'의 '내리다'는 기상의 의미을 담고 있는 이슬의 개념인 025 의 의미코드를 가진다. 반면 '영희가 내리다'의 '내리다'는 승강의 개념인 217, 315 를 가진다. 이 의미코드를 이용해서 선행하는 예의 주어는 대상이며 후행하는 예의 주어는 행위주임을 알 수 있다. 예에서 살펴본 것처럼 지배소와 의존소의 의미코드를 이용해서 문법관계에 상응하는 의미관계를 결정할 수 있다.

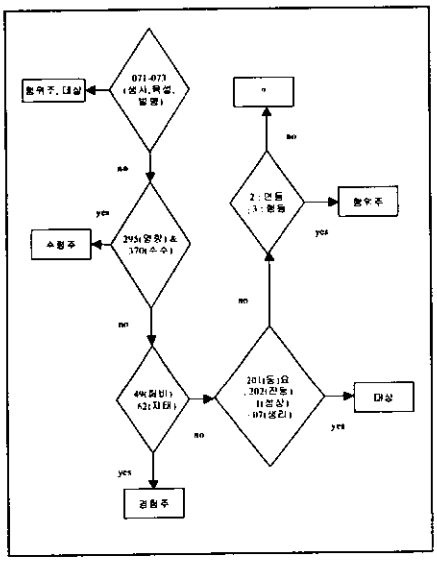


[그림 1] 가도까와 시소러스의 계층구조

		지배소	문법관계	의존소
Input	:	071(죽)	주어	572(학생)
		071(사살하)	목적어	561(종업원)
		189(어울리)	와|과	543(장관)
Output	:	071(죽)	대상	572(학생)
		071(사살하)	대상	561(종업원)
		189(어울리)	동반주	543(장관)

[표 2] 패턴 변환의 예

규칙은 [표 2]과 같은 패턴 변환을 수작업으로 하여 데이터 셋을 만들어 학습시킨 결과이다. 여기서는 간단히 주어를 결정하는 알고리즘만 보이도록 하겠다[그림 2].



[그림 2] 주어의 의미역을 결정하는 알고리즘

5. 실험 및 결과

본 논문에서는 한일 전자사전에서 추출한 54,013 개의 패턴을 대상으로 실험을 했으며 90.7%의 정확률과 93.5%의 적용율을 나타냈다. 규칙 구축에 실패한 예를 들면 '죽다' 와 '죽이다' 는 가도까와 시소러스에서 의미코드 071(생사)을 가진다. 두 용언 모두 의존소의 의미코드도 동일하게 사람(5)을 취하므로 의미코드와 문법관계만으로는 이 둘의 의미를 구별할 수 없다.

문법관계		패턴 수	정확률(%)	적용율(%)
전체		1,077	90.7	93.5
주어		784	90.6	100
목적어		152	100	100
부사어	에	85	90.5	49.4
	로	44	17.6	38.6
	와	12	100	100

[표 3] 실험 결과

목적어와 부사어 '와' 는 당연한 결과이며 주어는 90% 이상의 상당한 정확률을 보이지만 부사어 '에' 와 '로' 는 매우 적용율이 매우 낮다. 이는 학습 데이터가 다양한 사상관계를 반영하지 못했기 때문이다.

6. 결론 및 향후 연구

본 논문에서는 의미분석을 위한 문법관계에 따라 의미관계를 결정하는 규칙을 가도까와 시소러스를 이용해서 구축했다. 이 연구는 의미분석을 가능하게 할 뿐만 아니라 단어의 중의성을 해소하고 정보 추출과 같은 자연언어처리의 여러 분야에도 적용될 수 있다.

앞으로 패턴을 확장하고 통계적인 방법을 적용한다면 더 나은 성능을 기대할 수 있을 것이다.

7. 참고 문헌

[1] "21 세기 세종계획 - 전자사전 개발- " 연구보고서, 문화관광부, 1999, pp198-221

[2] 고영근, 남기심, "표준국어문법론", 탑출판사, 1993

[3] 양단희, "기계학습에 의한 단어의 격 원형성 자동획득", 한국정보과학회논문지(B) Vol.25 No. 7 1998 pp1116-1127

[4] 이홍식, "국어문장의 주성분 연구", 서울대학교 박사학위논문, 1996

[5] 조일영, "' NP 로' 의 의미역", 제 16 차 한국어학회 전국 학술 대회, 1998, pp56-65

[6] Daniel Gildea, Daniel Jurafsky, "Automatic Labeling of Semantic Roles", 38th Annual Meeting of the Association for Computational Linguistics, Proceedings of the ACL Conference, 2000, pp512-520

[7] 大野晋, "類語新辭典", 角川書店, 1981