기능동사 구문과 개념 유사도를 이용한 한국어 부사격의 의미역 결정

신명철° 이용훈 김미영 정유진 이종혁
포항공과대학교
{mcshin°, yhlee95, colorful, prizer, jhlee}@postech.ac.kr

Semantic Role Assignment for Korean Adverbial Case Using Support Verb Phrase and Concept Similarity

Myung-Chul Shin° Yong-Hun Lee Mi-Young Kim You-Jin Chung Jong-Hyeok Lee
Pohang University of Science and Technology

요 약

본 논문에서는 한국어에 있어 '에, 로'를 격표지로 하는 부사격에 대한 의미역 결정 모델에 대해 다루고 있다. 의미역 결정은 의미 분석의 핵심 과정 중 하나이고 자연언어처리에서 해결해야 할 중요한 문제이다. 본 논문은 기존 연구와 언어학 논저를 참고해서 의미역 결정에 유용한 자질들을 정리하였고 SVM을 이용하여 의미역 결정 모델을 구축하였다. 또한 기존 연구와 차별적으로 기능동사 구문의 처리와 지배소 개념의 유사도 보정 방법을 사용하여 보다 견고한 모델을 만들 수 있었다. 성능 평가 결과 개념(Concept)만을 사용한 기본 모델에 비해서 평균 9%의 정확률 향상을 보였다.

1. 서 론

의미 분석의 목적은 자연언어 문장의 의미적 구조를 분석하는 데 있으며 문장의 의미 분석 과정은 크게 단어 의미 중의성 해소(WSD : Word Sense Disambiguation) 단계와 논항의 의미역 결정(SRA : Semantic Role Assignment) 단계를 거쳐 이루어진다. 본 논문에서는 이러한 과정 가운데 논항의 의미역 결정에 관해 다루고자 한다.

의미역 결정이란 일반적으로 서술어-논항 관계에 적합한 의미 관계(Semantic Relation)를 정해주는 과정이라 할 수 있으며 이러한 의미 관계는 전통적으로 심층격(Deep Case), 격 관계(Case Role), 의미역(Thematic Role) 등으로 불려져 왔다[1].

자연언어처리에 있어서 의미역 결정은 기계 번역(MT), 정보추출(IE), 질의 응답 시스템(QA)의 성능 및 질 향상에 중요한 역할을 하기 때문에 최근 들어 정확하고 견고한 의미역 결정 방법론에 대한 필요성이 증가되고 있는 추세이다.

한국어의 경우 의미역 결정에 관한 다양한 연구[2,3,4,5,6]가 있었고 [2]은 부사격 조사에 대한 의미역 결정 문제를 중점적으로 다루었다. 한국어의 격조사는 구문 관계를 나타내는 격표지(Case Marker)라 할 수 있으며 하나의 격조사는 여러 가지 다양한 의미격을 표상할 수 있다. 특히 부사격 조사는 가장 많은 의미격을 나타낼 수 있어서 논항의 의미역 결정에 있어 심각한 문제를 드러내고 있다[2,5,6]. 따라서 본 연구는 한국어 부사격 조사 특히 애매성(Ambiguity)이 가장 큰 '에,로'를 격표지로 가진 부사격의 의미역을 결정하는 것을 목표로 한다.

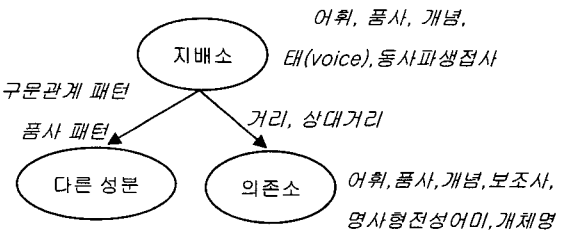
의미역 결정 모델의 구축을 위해 SVM(Support Vector Machine)과 다양한 자질을 이용하였으며 기능동사 구문 처리와 지배소의 개념 유사도 보정 기법을 사용하여 모델의 정확률 향상을 가져올 수 있었다.

2. 관련 연구

의미역 결정의 방법론은 크게 격틀 사전에 기반한 방법(Case-frame-based)[7,8,9]과 말뭉치에 기반한 방법(Corpus-based)[2,3,4,1,10,11]으로 분류될 수 있다. 전자는 격틀 사전이라는 언어 지식이 없거나 부족한 경우 사용하기 힘든 방법이어서 본 논문에서는 고려하지 않겠다. 후자는 말뭉치의 각 문장마다 논항에 적합한 의미역을 부착하여 의미역 태깅된 말뭉치를 구축하고 통계적 혹은 기계학습 방법을 사용하여 의미역을 결정하는 방법론이라 할 수 있으며 본 연구에서도 같은 접근 방법을 사용하겠다.

통계적 혹은 기계학습의 관점에서 의미역 결정의 문제는 의미역이 부착된 말뭉치를 의미역 결정에 도움이 되는 자질로 표현된 학습 데이터로 변환한 후 모델을 학습하고 학습된 모델을 이용하여 새로운 데이터의 의미역을 결정하는 문제 즉 다중 분류 문제(Multi-class classification)로 변환될 수 있다.

이러한 문제의 해결을 위해서 기존 연구에서 사용되었던 자질들을 종합하여 한국어의 의존 트리 구조(Dependency tree structure)에 맞게 정리하면 [그림 1]과 같다.



[그림 1] 기존 연구에서 사용된 자질 집합

[1,10,11]에서 사용된 자질 중 영어에 의존적이거나 구문 트리 구조(Phrase tree structure)에 한정된 자질들은 제외시켰다.

3. 기계 학습을 이용한 의미역 결정

3.1 학습에 사용할 자질 선택

[3]는 부사격의 의미역 결정 유형을 세 가지로 분류하였다. [유형 I]은 의존소와 관계없이 지배소만으로 의미역을 결정할 수 있는 유형, [유형 II]는 의존소가 가질 수 있는 여러 의미격 중 지배소가 그 중 하나를 선택하여 결정할 수 있는 유형, [유형 III]은 의존소와 지배소만으로 의미역을 결정할 수 없는 유형을 말한다. [유형 III]은 일종의 문맥 정보를 이용해서 해결할 수 있는 유형이라고 볼 수 있다. [3]에서는 [유형 III]을 연구의 범위에서 제외시켰지만 [유형 III]과 다른 유형을 학습 데이터에서 분별하기도 어렵거니와 그렇게 제외시키는 것은 의미역 결정의 목적에 맞지 않기 때문에 본 연구에서는 다른 유형과 함께 구별없이 다루고자 한다. 따라서 이를 분별하기 위한 자질로서 주어과 목적어의 품사와 개념 코드를 추가적으로 고려하였다. 또한 [그림 1]에서 지배소와 의존소간의 거리와 상대 거리는 [2]에서 변별력이 없는 자질로 증명되었기 때문에 사용하지 않았고 의존소의 개체명(Named Entity)도 자질 집합에 추가하지 않았다. 결과적으로 학습을 위해 선택된 자질은 다음 [표 1]과 같다.

[표 1] 학습을 위해 선택된 자질 집합

예문, [그녀는 불길한 예감-에 몸을 벌떡 일으켰다]

자 질		약 어	예
의존소	어휘	LD	예감
	개념	CD	느낌-400
	품사	PD	CMCPA(서술성 명사)*
	명사형 전성어미	ED	-
	보조사	AD	-
지배소	어휘	LG	일으키
	개념	CG	일어남-353
	품사	PG	YBDO(일반동사)
	선어말어미	EG	-
	동사파생접사	SG	-
주어	품사	PS	CTP3(3인칭 대명사)
	개념	CS	타창-503
목적어	품사	PO	CMCN(일반 명사)
	개념	CO	육체-600
기타	조사 패턴	CP	는-에-을
	품사 패턴	PP	CTP3-CMCPA-CMCN-YBDO

* KLE tag set

3.2 기계학습 모델의 선택

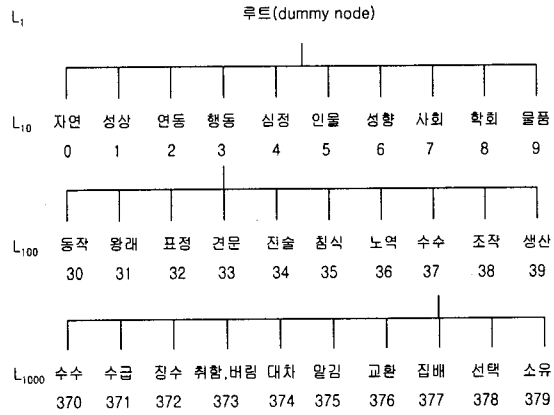
기계학습을 위한 모델로는 SVM(Support Vector Machine)을 사용하였다. [10]에서 밝힌 바와 같이 SVM은 이진 분류기(Binary classifier)이기 때문에 다중 분류 문제(Multi-class classification problem)를 위해서는 여러 개의 이진 분류 문제로 나누고 각 결과를 통합하는 방법을 사용해야 한다. 이 문제에 대한 일반적인 접근 방법은 PAIR-WISE 접근법과 ONE vs ALL 접근법이 있는데 본 연구에서는 PAIRWISE 접근법을 구현한 LIBSVM[12]을 사용하였다.

3.3 기능동사 구문 처리

술어명사(Predicative noun)[1] 란 명사적 특성과 술어적 특성을 동시에 가진 명사를 의미하며 기능동사(Support verb)[2] 란 어휘의미가 미약하거나(Light) 없으며 술어 명사의 현동화(Actualization)를 뒷받침해주는 동사를 말한다. 이런 성격의 기능 동사는 자신의 논항을 선택하지 못한다. 기능동사 구문이란 이처럼 술어명사가 기능 동사와 함께 실현되어 자신의 고유한 논항을 선택하는 기본 구문을 의미한다[13]. 의미역 결정에 있어서 이와 같은 기능동사 구문을 고려해야 한다. 예를 들어 '미국이 이라크-에 공격을 했다.'에서 의존소 '이라크'의 구문적 지배소는 '했다'이지만 의미적으로는 '공격'이 '이라크'를 논항으로 취하는 실제 지배소가 되어야 한다. 따라서 이와 같은 문장에서 '이라크-에'의 의미역 결정을 위한 자질 벡터는 '공격'을 지배소로 갖도록 재조정되어야 한다. 기능동사 구문의 식별의 위해 세종 전자 사전[15]에 기술된 기능동사 목록를 참고하여 만든 규칙을 사용하였다. 위의 예문과 같은 경우 '공격'의 품사가 서술성 명사이고 '했다'는 기능동사 목록에 들어 있으므로 '공격을 했다'를 기능 동사 구문으로 판별한다.

3.4 지배소의 개념 코드에 대한 의미 유사도 사용

본 연구에서 사용한 자질 중 개념 코드는 가도카와 시소러스(그림 2)를 통해서 얻어진다. 가도카와 시소러스는 총 1,110 개의 개념과 4 단계의 계층 구조를 가지고 있으며 L_1, L_{10}, L_{100} 레벨에 속해 있는 개념들은 각각 10개의 하위 개념들로 나뉜다.



[그림 2] 가도카와 시소러스의 계층 구조

지배소 어휘의 개념이 유사한 경우 의미역 결정에 있어 유사한 양상을 보이기 때문에 일반적으로 어휘의 개념을 자질로 사용한다. 이와 비슷한 방식으로 개념간 유사도를 자질로 사용할 수 있으며 특히 개념의 수가 많고 학습 데이터는 상대적으로 적은 경우는 유용하다. 본 연구에서는 [14]에서 개념 유사도를 자질로 사용한 방법을 그대로 이용하겠다. 두 개념간 유사도의 계산 공식은 다음과 같다.

$$sim(T,P_i) = \frac{2 \times level(MSCA(T,P_i))}{level(T) + level(P_i)} \times weight \quad (1)$$

※ MSCA : most specific common ancestor
※ sim(T, P_i) = 0 if MSCA(T, P_i) is level 1
※ weight = 1 (direct descendent), 0.5 otherwise

4. 실험 및 평가

4.1 학습 데이터

한국어 형태소 분석기 및 품사태거 평가 워크숍을 위해 제공받은 말뭉치(MATEC' 99)에서 임의로 추출한 3,400 문장과 세종 전자 사전에 기술되어 있는 예문 4,353 문장을 학습을 위한 기본 코퍼스로 사용하였다. 코퍼스로부터 자질을 추출하기 위해 포항공대 지식 및 언어공학 연구실에서 자체 개발한 형태소 분석기(KoMA), 구문 분석기(KoPA)를 사용하였으며 단어 의미를

1. 서술 명사 또는 동사적 명사(Verbal noun)이라 부르기도 한다.
2. 'Light verb'라고도 한다.

얻기 위해 한일 기계번역기(COBALT-KJ)의 단어 의미 중의성 모듈을 이용하였다[6]. 형태소 및 구분 분석 그리고 단어 의미 중의성 해소 결과는 오류를 포함하고 있기 때문에 수작업을 통해 이러한 오류를 수정하였다.

4.2 실험 결과

[2]는 분산 클러스터링 알고리즘을 통해 지배소와 의존소의 클래스를 얻고 이를 학습을 위한 자질로 사용하였다. 클래스는 일종의 개념(Concept)라고 볼 수 있기 때문에 본 연구에서는 개념만을 사용한 경우를 기본 모델(Baseline)로 설정하였다. 다음 [표 3]은 기본 모델(Baseline)에 각각의 자질들을 추가하였을 때 나타나는 효과을 보여준다.

[표 3] 기본 모델에 추가된 자질 별 효과(10-fold CV)

자 질	정확률(%)	
	에	로
CD+CG (baseline)	68.22	72.07
+ PD + PG	68.56	72.55
+ ED	68.25	72.19
+ AD	68.15	72.25
+ EG + SG	68.45	72.43
+ PS + PO	68.73	72.43
+ CS + CO	69.07	71.24
+ CP	69.47	73.62
+ PP	68.62	71.77
+ LD + LG	77.57	77.86
+ LG	78.22	77.38
+ LG + CP	78.87	78.52
+ LG + CP + SPP	78.94	78.88
+ LG + CP + SIM	79.07	78.16
+ LG + CP + SPP + SIM	79.41	78.22
MFC	43.86	33.95

* SPP : 기능 동사 구문 처리 * SIM : 지배소 개념 유사도 보정
* MFC : Most Frequent Class
* SVM Kernel function : RBF, cost = 40

아래 [표 4]는 위의 결과 중 가장 높은 정확률을 보인 자질 조합에 대해서 의미역 별 정확률 및 학습 데이터에서의 빈도수를 나타낸다.

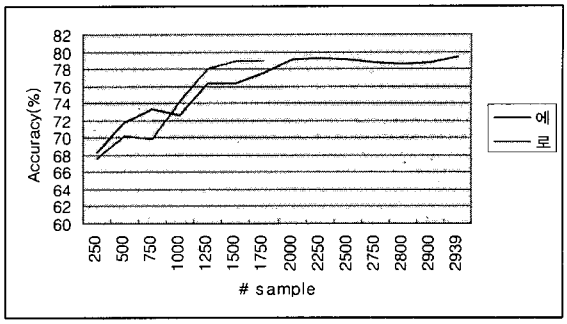
[표 4] 학습 데이터에 나타난 의미역 별 정확률 및 빈도수

의미역	에		로	
	정확률(%)	빈도수	정확률(%)	빈도수
행위주	78.18	62	100	8
대상	77.61	157	75	12
장소	80.45	1289	53.84	30
도착점	78.33	963	82.28	569
결과상태	0	4	73.14	420
원인주	82.49	276	78.51	138
도구	62.5	63	79.70	482
기준치	71.05	106	75	11
동반주	0	3	0	0
경험주	60	9	0	0
출발점	100	7	100	6
계	79.41	2939	78.88	1676
엔트로피		1.44		1.46

[그림 3]은 학습 데이터의 증가에 따른 모델의 성능 향상 효과를 보여준다.

5. 결론 및 향후 계획

실험 결과 의미역 결정에 결정적 기여를 하는 자질은 지배소의 어휘(LG)인 것으로 나타났으며 이를 단독으로 사용하는 것보다 개념(CG)과 함께 사용하는 것이 성능 향상에 도움이 되었다. 본 연구에서 제시한 기능동사 구문 처리와 지배소의 개념 유사도 보정 방법이 성능 향상에 기여를 한다는 것은 분명한 것으로 보이며(McNemar test, p = 0.1573) 조사 패턴(CP)을 함께 사용하는 경우 가장 높은 정확률을 보였다. 본 연구와 기존 연구의 직접적인 비교는 어렵지만 단순히 지배소와 의존소의 개념만을 사용한 기본 모델보다 약 9%의 성능 향상을 보였다. 차후 의미역 결정에 결정적인 자질에 관한 연구와 부사격이 아닌 다른 논항의 의미역을 먼저 결정한 이후 이를 자질로 사용하는 방법에 대한 연구를 수행할 예정이다.



[그림 3] 데이터의 수에 따른 학습 커브

6. 참고 문헌

[1]Daniel Gildea and D. Jurafsky. Automatic Labeling of Semantic Roles, Computational Linguistics, 28(3):245-288, 2002

[2]S.B. Park. Decision Tree Based Disambiguiation of Semantic Roles for Korean Adverbial Postpositions, IEICE Transaction Information and System, Vol.E86-D, No. 8, 2003

[3]양단희, 송만석. 기계학습에 의한 단어의 격 원형성 자동 획득, 정보과학회지, 제 25권, 제 7 호, pp. 1116-1127, 1998

[4]Kang, W.S., et al. A Neural Network Method for the Semantic Analysis of Prepositional Phrases in English-to-Korean Machine Translation, An International Journal of the Chinese Language Computer Society, vol.8, no.2, pp. 143-162, 1994

[5]Jung-Hye Park. Determination of Thematic Roles according to Syntactic Relations Using Rules and Statistical Models, MS Thesis, Pohang University of Science and Technology, 2002

[6]강신재, 박정혜. 대규모 말뭉치와 전산 언어 사전을 이용한 의미역 결정 규칙의 구축, 정보처리학회 논문지 B 제 10-B권 제 2호, 2003

[7]Hui-Feng Li. Conceptual Graph Generation from Syntactic Dependency Stuructures for an Interlingua-Based MT System, Phd Thesis, Pohang University of Science and Technology, 1998

[8]Beale, S., S. Nirenburg and K. Mahesh. Semantic Analysis in The Mikrokosmos Machine Translation, In Proc. Of Symposium on NLP, Kaset Sart Univerity, Bangkok, Thailand, 1995

[9] Kurohashi, S., and Nagao, M. A Method of Case Structure Analysis for Japanese Based on Examples in Case Frame Dictionary. IEICE Transactions on Information and System, vol.E77-D, no.2, pp. 227-239. 1994

[10]Kadri Hacioglu, et al. Shallow Semantic Parsing Using Support Vector Machines. CSLR Tech. Report, CSLR-TR-2003-1, 2003

[11]Kadri Hacioglu, et al. Semantic Role Labeling Using Dependency Trees, In COLING 2004, 2004

[12]C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machine, Software available at http://www.csie.ntu.edu-.tw/~cjlin/libsvm , 2001

[13]이성헌. 전자사전에서의 기능동사구문 처리 문제-세종 체언사전의 경우-, 한국사전학회 제5회 학술대회 발표자료집, 2004

[14]You-Jin Chung, et al. Word Sense Disambiguation Using Neural Networks with Concept Co-occurrence Information, NLPRS 2001. pp. 715 - 722, 2001

[15]홍재성 외. 21세기 세종계획 전자사전개발 연구보고서, 문화관광부, pp. 110 ~ 116, 2004