부트스트래핑 알고리즘을 이용한 한국어 격조사의 의미역 결정

김병수° 이용훈 나승훈 김준기 이종혁
포항공대 정보통신대학원 정보처리학과, 포항공대 컴퓨터공학과, 첨단정보기술 연구센터+
{akirus82°, yhlee95, nsh1979, yangpa, jhlee}@postech.ac.kr

Bootstrapping for Semantic Role Assignment of Korean Case Marker

Byoung-Soo Kim°, Yong-Hun Lee, Seung-Hoon Na, Jun-Gi Kim, Jong-Hyeok Lee
Dept. of Graduate School for Information Technology, POSTECH, Dept. of Computer Science & Engineering, POSTECH, Advanced Information Technology Research Center(AITrc)

요 약

본 논문은 자연언어처리에서 문장의 서술어와 그 서술어가 가지는 명사 논항들 사이의 문법관계를 의미관계로 사상하는, 즉 논항이 서술어에 대해 가지는 역할을 정하는 문제를 다루고 있다. 의미역 결정은 단어의 의미 중의성 해소와 함께 자연언어의 의미 분석의 핵심 문제 중 하나이며 반드시 해결해야 하는 매우 중요한 문제 중 하나이다. 본 연구에서는 언어학적으로 유용한 자원인 세종전자사전을 이용하여 용언 격틀사전을 구축하고 격틀 선택 방법으로 의미역을 결정한 후, 결정된 의미역들에 대한 확률 정보를 확률 모델에 적용하여 반복적으로 학습하는 부트스트래핑(Bootstrapping) 알고리즘을 사용하였다. 실험 결과, 기본 모델에 대해 10% 정도의 성능 향상을 보였다.

1. 서론

일반적으로 의미 분석은 형태소 분석과 구문 분석의 과정을 거쳐 이루어지는 자연언어처리의 상위 단계로 크게 단어의 의미 중의성을 해소하는 문제와 서술어와 명사 논항 사이의 의미역을 결정하는 문제로 나눌 수 있다. 본 논문에서는 이러한 의미 분석 단계에서 의미역 결정에 대해 다루고자 한다.

의미역 결정이란, 문장의 서술어와 그 서술어가 취하는 명사 논항 사이에 적합한 의미 관계를 정하는 것이라고 할 수 있다. 즉, <그림1>과 같이 문장의 표층격(Surface Case)에 해당하는 문법관계를 심층격(Deep Case)에 해당하는 의미 관계로 사상하는 문제로 볼 수 있다.

한국어의 경우 격조사(Case Marker)에 의해 구문의 의미가 부여되고 하나의 격조사가 서술어의 특징에 따라 다양한 의미를 가지는 특징을 가진다. 특히 부사격 조사의 경우 쓰임이 다양하고 임의격으로 사용되는 경우가 많기 때문에 의미역을 결정하는데 있어 심각한 문제가 되고 있다. 한국어의 경우 의미역 결정에 관한 다양한 연구[1,6,8,11]가 있었고, 일부 부사격 조사에 대한 의미역 결정을 중점적으로 다룬 연구로는 [8,11]이 있었다.

의미역 결정은 자연언어처리에서 대규모의 의미역이 부착된 말뭉치를 필요로 하는 기계번역(MT), 질의응답(QA), 정보추출(IE) 시스템의 성능 향상에 중요한 역할을 한다. 이러한 중요성으로 인해, 최근 들어 자동으로 정확한 의미역을 결정하는 방법론[12]에 대한 연구들이 활발히 진행되고 있다.

2. 기존 연구

의미역 결정에 관한 기존 연구는 크게 격틀사전에 기반한 방법(Case-frame-based)[7]과 대규모 말뭉치에 기반하여 학습을 하는 방법(Corpus-based)[1,4,11]으로 나눌 수 있다.

격틀사전을 이용하는 방법은 격틀이라는 언어 지식을 이용하는 방법으로, 격틀사전에 기술된 격틀(frame)과 선택제약정보(selectional restriction) 등을 이용하여 입력 문장에 대하여 적합한 격을 할당하는 방법이다. 말뭉치를 이용하여 학습하는 방법은 대규모의 말뭉치에 의미역을 부착하고 기계적 혹은 통계적 학습 방법을 이용하여 의미역을 결정하는 방법이다. 최근에는 하나 이상의 방법을 통합하여 두 방법을 서로 보완하는 연구들[8,12]이 진행되고 있다. 즉, 정확률이 높은 격틀사전에 의한 방법과 적용률이 높은 말뭉치 학습을 하는 방법을 통합하거나 말뭉치를 이용한 학습 방법 간에 통합을 통하여 보다 견고하고 정확한 의미역 결정을 하고자 하는 방법이다. 이 방법은 의미역 결정 방법의 특징보다는 모델을 구축하는 측면의 비중이 큰 방법이라고 할 수 있다.

부트스트래핑 알고리즘을 의미역 결정에 적용한 연구는 [9,10]이 있다. [9,10]에서는 단순한 격틀 매칭을 통해서 초기 확률 정보를 구축하고 일부 제한된 동사들에 대해서만 평가가 이루어졌다. 그러나 한국어의 경우 어순이 자유롭고 격조사가 영어의 전치사에 비해 의미가 다양하기 때문에 [9,10]의 방법을 한국어에 직접 적용하기에는 어려움이 있다.

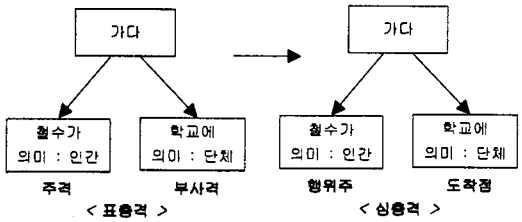


그림 1. 의미역 결정의 예

3. 연구 범위

본 논문은 문장의 구문 분석을 통해 얻은 서술어와 명사 논항 사이의 의존트리(Dependency Tree)를 이용하여 문장에 나타나는 주격, 목적격, 일부 부사격(에, 로, 에서, 에게)에 대한 의미역을 자동으로 결정하는 시스템을 구축하는 것을 목표로 한다.

한국어의 경우 영어권의 FrameNet이나 PropBank와 같이 의미역이 부착된 대규모의 말뭉치가 없기 때문에 말뭉치를 이용한 학습 방법을 하기 위해서는 수작업으로 말뭉치에 의미역을 부착해야 하는 문제점이 있다. 의미역을 부착하는 작업은 의미역의 수나 기준에 따라 언어학자들도 적용하는데 어려움이 있

+본연구는 첨단정보기술연구센터를 통하여 과학재단의 지원을 받았음.

는 작업이므로 본 연구에서는 고려하지 않겠다. 따라서 대규모의 말뭉치 대신에 격틀사전을 이용하여 비교적 정확한 학습 데이터를 구축하고 이를 확률 모델에 적용하여 반복적으로 학습하도록 부트스트래핑 알고리즘을 사용하여 시스템을 구축하였다.

4. 격틀사전을 이용한 의미역 결정

세종전자사전의 용언사전은 서술어가 요구하는 통사적인 논항인 격틀정보뿐만 아니라 의미적인 논항인 선택제약정보에 대해서 의미역을 정의하여 기술되었다. 세종전자사전에서는 행위주, 경험주, 심리경험주, 동반주, 대상, 장소, 방향, 도착점, 결과상태, 출발점, 도구, 영향주, 기준치, 목적, 내용 등 총 15개를 정의하였다. 현재 2005년까지 세종전자사전의 용언사전은 18,618개의 표제어에 대해 기술되어 있다.

4.1 용언사전으로부터 격틀사전 구축

용언사전에는 의미역 결정에 필요한 정보들이 XML 형태로 기술되어 있기 때문에 이를 활용하기 위해서는 전산적 처리가 쉽도록 재구축해야 할 필요가 있다. 따라서 프레임, 격조사, 선택제약정보 등 의미역 결정에 필요한 정보들을 선별하여 격틀사전을 구축하였다. 용언사전은 2007년에 완성을 목표로 현재 기술이 미흡한 부분에 대한 보완이 이루어지고 있다. 따라서 이 부분은 격틀사전의 표제어에서 제거하였다. 그 결과 표제어당 평균 1.97개의 격틀이 있음을 확인하였다.

4.2 세종 명사의미부류 정보 이용

격틀사전의 선택제약정보는 세종전자사전의 명사의미부류를 기준으로 기술되어있다. 세종 명사의미부류는 트리구조로 최상의 의미부류 구체물, 집단, 장소, 추상적대상, 사태 등에 대해 총 582개의 의미부류로 분류된다.

선택제약정보는 서술어의 논항과 유사도 계산을 하는 과정을 통해서 격틀을 선택하는데 이용된다. 따라서 서술어의 논항은 유사도 계산을 위해서 하나의 세종 명사의미부류로 사상이 되어야 한다. 본 연구에서는 이를 위해 명사 논항이 가질 수 있는 의미부류에 대해서 선택제약정보와 최대 유사도를 가지는 의미부류를 선택하는 방법을 이용하였다.

4.3 격틀 선택

격틀사전을 이용하여 입력 문장에 대해서 적합한 격틀을 선택하는 과정은 의존 트리의 논항과 격틀에 기술된 선택제약정보 사이의 유사도를 계산하여 가장 높은 점수를 갖는 격틀을 선택하는 과정으로 생각할 수 있다. 각 논항에 대해 <수식 1>[5]을 이용하여 유사도를 계산하고, 이들의 합이 최대가 되는 격틀을 최종적으로 선택하도록 한다.

$$\mathrm{sim}(c_i,c_j)=\frac{2*\min\limits_{\mathrm{pths}(\mathrm{mscs}(c_i,c_j),rt)} \mathrm{len}_e p}{\min\limits_{\mathrm{pths}(c_i,c_j)} \mathrm{len}_e p + 2*\min\limits_{\mathrm{pths}(\mathrm{mscs}(c_i,c_j),rt)} \mathrm{len}_e p}$$

* pths(x) : set of paths between the concepts x and y
* $\mathrm{len}_e(x)$: length in number of edges of the path x
* mscs(x,y) : the most specific common subsumer of x and y

수식 1. Wu & Palmer's measure

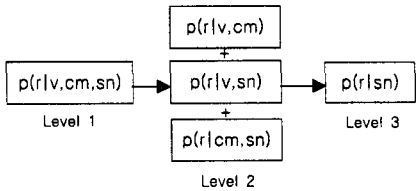
5. 확률 모델을 이용한 의미역 결정

5.1 초기 확률 정보 구축

격틀사전을 이용한 의미역 결정 과정을 거치게 되면 일부 서술어와 명사 논항에 대해서 의미역이 결정되게 된다. 격틀에 의한 방법은 정확률이 높은 방법으로 이 결과를 초기 정보로 이용하여 확률 정보를 구축하였다. 그러나 격틀사전에 의한 방법은 적용률이 낮기 때문에 충분한 학습 데이터를 위해서는 적절한 수준에서 격틀을 선택하여 정확하고 충분한 학습 데이터를 얻는 것이 중요하다고 하겠다.

5.2 확률 모델

확률 모델은 [9]에서 제시한 Backoff 모델을 세종전자사전에서 얻을 수 있는 정보에 맞게 수정을 하였다. 확률 모델은 확률 정보를 이용하여 <그림2>의 각 레벨을 거치면서 격틀 방법에서 의미역을 결정하지 못한 논항에 대해 의미역을 결정한다. 의미역 결정은 서술어, 격조사, 선택제약정보가 주어질 때 논항에 적합한 의미역을 결정하는 레벨 1의 확률 수식으로 생각 할 수 있다. 그러나 레벨 1은 조건이 모두 만족해야 확률값을 가지므로 자료 부족(Data Sparseness) 문제가 발생할 수 있다. 이를 해결하기 위해 각 레벨을 좀 더 일반성을 가지는 확률 모델로 분해를 하였다. 각 레벨은 정해진 임계값에 도달하지 못하면 다음 레벨로 넘어가 확률값을 계산하여 의미역을 결정한다.



* r(의미역), v(서술어), cm(격조사), sn(세종 명사의미부류)

그림 2. 확률 모델

5.3 부트스트래핑 알고리즘을 이용한 모델 학습

확률 모델로 의미역을 결정한 후, 임계값에 도달하지 못한 논항들의 의미역을 결정하기 위해서 부트스트래핑 알고리즘을 사용하였다. 부트스트래핑 알고리즘은 초기 정보를 이용하여 점진적으로 정보를 확장하여 전체 문제를 해결하는 방법으로 본 연구에서는 새로 의미역이 결정될 때 마다 확률 정보를 갱신하고 이를 반복적으로 확률 모델에 적용하여 의미역이 결정되지 않은 모든 논항들의 의미역이 결정되도록 하였다. 이때 알고리즘의 반복횟수를 결정하는 임계값의 변화폭에 따라 확률 정보가 갱신되는 범위가 결정되므로 확률 정보의 정확성에 중요한 영향을 미치게 된다. 확률 모델의 경우, 확률값이 0에서 1사이의 값을 가지므로 이 사이에서 임계값을 줄여나가면서 반복 학습하였다.

기존의 부트스트래핑 알고리즘은 수작업으로 초기 정보를 구축하였으나 본 연구에서는 세종전자사전이라는 신뢰할 수 있는 언어학적 자원을 이용하여 일종의 규칙으로 자동으로 초기 정보를 구축하였다는 차이점이 있다. 의미역 결정의 경우 수작업으로 초기 정보를 구축하기에 어려움이 있기 때문에 이는 큰 장점이라 할 수 있다.

6. 실험 및 평가

6.1 학습 데이터

세종전자사전에 기술되어 있는 예문들을 추출하여, 의미역 결정에 적합하지 않은 문장, 문법적 오류가 있는 문장들을 선별하여 57,238개의 문장은 학습 데이터로 1,000개의 문장은 평가 데이터로 구축하였다. 격틀사전을 이용한 방법에 필요한 정보를 추출하기 위해 포항공과대학 지식 및 언어공학 연구실에서 개발한 형태소 분석기(KoMA), 구문 분석기(KoPA)를 사용하였다. 좀 더 정확하고 충분한 확률 정보를 구축하기 위해 형태소 분석, 구문 분석 결과에서 발생하는 오류를 수정할 수 있으나 본 연구에서는 자동으로 의미역을 결정하는 시스템 구축에 목표를 두었으므로 그 결과를 그대로 이용하였다.

6.2 실험 결과

<표1>은 의미역 결정 모델에 따른 결과를 보여준다. 본 연구에서는 격조사에 대해 결정된 의미역 중 최다 빈도를 가지는 의미역으로 할당하는 모델을 기본 모델로 설정하였다. 격틀모

델은 격틀에 일치하는 일부 입력 문장에 대해 의미역이 결정되므로 정확률(presicion)과 재현률(recall)로 평가하였다. 제안한 확률 모델은 기본 모델에 비해 제안한 확률 모델은 평균 10% 정도의 성능 향상을 보였다.

모델	정확률					
	이/가	을/를	에	로	에서	에게
기본모델	73.48	98.33	50.13	29.11	55.84	62.50
격틀모델	82.68	96.58	77.03	78.75	94.12	92.59
	58.86	39.38	41.28	39.87	21.05	52.63
확률모델	79.84	97.77	70.54	62.66	61.04	88.42

(단위 : %)

*정확률 : 의미역 결정이 맞은 개수 / 의미역이 결정된 개수

표 1. 모델에 따른 의미역 결정 결과

<표2>는 부사격 조사의 의미역 별 정확률 및 빈도수에 대한 결과를 보여준다. 부사격 조사에 대한 성능이 전체적으로 낮은 이유는 부사격 조사 자체가 서술어의 특징에 따라 다양한 의미를 가지기 때문이라고 할 수 있다. 세종전자사전의 의미역 기술 지침[3]에 따르면 이전 연구들과 다르게 장소와 방향점, 장소와 출발점에 대한 구분 등이 격조사에 의해 일반적으로 결정되는 것이 아니라 서술어의 특징에 따라 다양하게 구별하기 때문에 부사격 조사에 대한 결과가 낮게 나왔다고 할 수 있다. 또 <표1>에서 알 수 있듯이 주격이나 목적격 조사에 비해 격틀사전을 이용한 의미역 결정 후 확률 모델에 이용되는 학습 데이터가 정확하지 못하거나 그 수가 부족한 것을 알 수 있다. 그 이유는 부사격 조사는 필수격(Obligatory Case)보다 임의격(Optional Case)으로 사용되는 경우가 많기 때문에 격틀사전에 나타나지 않아 이후 확률 모델을 통해 학습 과정을 반복하게 돼도 이 부분에 대해서는 학습이 이루어지지 않기 때문이다. 따라서 격틀사전에 추가적으로 임의격도 기술함으로써 성능을 향상시킬 수 있다.

의미역	에		로		에서		에게	
	정확률	빈도수	정확률	빈도수	정확률	빈도수	정확률	빈도수
행위주	100	1	0	0	33.33	3	90.91	11
경험주	100	3	0	0	0	0	70	10
심리경험주	0	0	0	0	0	0	50	2
동반주	0	0	100	1	0	0	0	0
대상	65.71	35	100	2	0	1	100	2
장소	83.59	195	25	4	41.86	43	75	4
방향	0	0	80	5	0	0	0	0
도착점	59.05	105	75.56	45	0	0	96.67	60
결과상태	50	2	90	40	0	0	0	1
출발점	100	1	0	0	93.33	30	60	5
도구	80	5	38.10	42	0	0	0	0
영향주	29.63	27	30.77	13	0	0	0	0
기준치	50	12	16.67	6	0	0	0	0
목적	100	1	0	0	0	0	0	0
내용	0	0	0	0	0	0	0	0

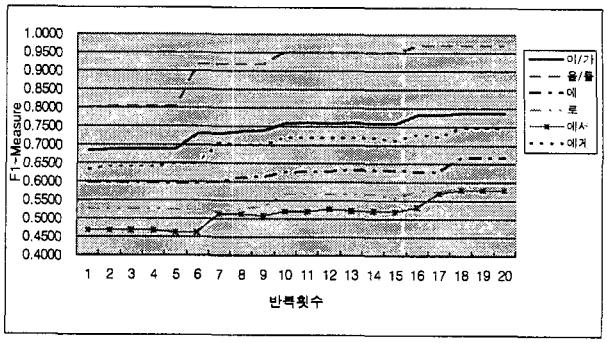
(단위 : %)

표 2. 부사격 조사 별 정확률 및 빈도수

<그림3>은 특정 임계값에서 반복 학습을 하였을 때 확률 모델의 학습 곡선을 나타낸다. 확률 모델이 반복 학습할수록 서서히 성능이 증가하며 올바른 방향으로 학습이 되는 것을 알 수 있다.

7. 결론 및 향후 연구

본 논문에서는 대량의 의미역 부착된 말뭉치에 의존하지 않고 자동으로 한국어의 의미역 결정을 하는 방법을 제시하였다. 세종전자사전을 이용하여 용언격틀사전을 구축하여 의미역 결정을 하고, 비교적 간단한 확률 모델을 이용하여 반복적으로 학습함으로써 격틀사전을 이용한 방법에서 의미역 결정을 하지



*F1-Measure : 2 * 정확률 * 재현률 / (정확률 + 재현률)

그림 3. 반복 학습에 따른 학습 곡선

못한 문장들에 대해서 의미역 결정을 하였다.

실험 결과, 확률 모델은 초기에 결정된 확률 정보에 의존적인 경향을 나타내는 문제가 있음을 알 수 있었다. 따라서 좀 더 견고한 모델 구축을 위해서 확률 정보 외에 구문 분석 결과에서 얻을 수 있는 정보를 이용하여 기계학습을 하는 방법을 생각할 수 있다. 또는 유사한 의미를 가지는 서술어들 사이에는 비슷한 격틀 정보를 갖는 점에서 동사의 클래스 정보를 확률 모델에 추가하는 방법을 이용할 수도 있다. 향후 격틀사전의 보완 및 모델의 보완을 통해 추가적인 부사격 조사에 대한 적용을 수행할 계획이다.

8. 참고 문헌

[1] 양단희, 송만석, *기계학습에 의한 단어의 격 원형성 자동 획득*, 정보과학회지, 25권 제7호, pp.1116-1127, 1998

[2] 이희자, 이종희, *한국어 학습용 어미·조사사전*, 2001

[3] 홍재성 외, *21세기 세종계획 전자사전 개발 연구보고서*, 국립국어원, pp.62-66, 2005

[4] Daniel Gildea and Daniel Jurafsky, *Automatic Labeling of Semantic Roles*, Computational Linguistics, Vol.28, No.3, pp.245-288, 2002

[5] Emmanuel Blanchard, et al. *A typology of ontology -based semantic measures*, EMOI - INTEROP, 2005

[6] 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

[7] Kurohashi, S, and Nagao, M. *A Method of Case Structure Analysis for Japanese Based on Examples in Case Frame Dictionary*, IEICE Transaction Information and System, Vol.E77-D, No.2, pp.227-239, 1994

[8] Myung-Chul Shin, *Integration of Case-Frame Dictionary into Machine Learning Techniques for Semantic Role Assignment of Korean Adverbial Cases*, MS Thesis, Pohang University of Science and Technology, 2006

[9] Robert S. Swier and Suzanne Stevenson, *Unsupervised Semantic Role Labelling*, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp.95-102, 2004

[10] Robert S. Swier and Suzanne Stevenson, *Exploiting a Verb Lexicon in Automantic Semantic Role Labelling*, HLT/EMNLP, 2005

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

[12] Xavier Carreras et al, *Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling*, In Proceeding of CoNLL-2005