구문구조부착 말뭉치를 이용한 술어의 하위범주화 정보 구축

류법모*, 장명길, 박수준, 박재득, 박동인
시스템공학연구소 자연어정보처리연구부
{ryupm,mgjang,soop,jdpark,dipark}@seri.re.kr

The Construction of Predicate Subcategorization Using Tree Tagged Corpus

Pum-Mo Ryu, Myung-Gil Jang, Soojun Park, Jae-Deuk Park, Doing-In Park
Detp. of Natural Language Information Processing, SERI

요약

한국어 문장에서 술어의 역할이 매우 중요하기 때문에 술어의 하위범주화 정보는 한국어 분석 및 생성에서 필수적이다. 그러나 기존의 한국어 술어의 하위범주화 사전은 전문가의 사전지식이나 직관에 의존하여 만들어졌기 때문에 주관적이고 오류의 가능성이 높으며 많은 수작업이 필요했다. 또 영역에 독립적인 하위범주화 정보를 구축하는 작업은 매우 어렵기 때문에 응용영역에 맞는 하위범주화 정보를 쉽게 구축하는 방법이 요구되었다. 본 논문에서는 구문구조부착 말뭉치를 이용하여 전문가의 제한된 개입만으로 통계정보와 명사의 의미정보를 포함하는 술어의 하위범주화 정보 구축 방법을 제안한다.

1. 서 론

구문분석을 위한 문법규칙은 구문 규칙이 인코드되는 위치에 따라 크게 규칙 기반 문법(rule based grammar)과 어휘 중심 문법(lexicalized grammar) 두 가지로 나눌 수 있다. 구문 태그 단위에서 별도의 규칙으로 구문 규칙을 표현하는 규칙 기반 문법은 어휘 중심 문법에 비해서 전자사전의 구조가 간단하고, 구문적 제약을 쉽게 규칙으로 표현할 수 있는 장점이 있다. 그러나 각각의 어휘를 하나의 구문 카테고리에 대응시키고, 일반적인 규칙을 찾아내는 작업이 매우 어렵고, 구문 카테고리가 변경되면 사전의 내용도 많이 바뀌어야 하는 단점이 있다. 더욱이 문장의 구문적 애매성은 구문 태그 단위에서는 해결이 불가능하기 때문에 각 어휘의 특성을 나타내기 위해서 별도의 정보를 표현하고 처리하는 방법이 필요하다. 따라서 최근의 구문분석을 위한 이론들은 구문 규칙을 별도의 문법 규칙으로 표현하기 보다는 사전에 각 어휘의 제약조건으로서 표현하는 어휘 중심 문법을 중심으로 발전하고 있다. 대표적인 어휘 중심 문법으로는 categorial grammar, linker grammar, lexicalized tree-adjoining grammar등이 있다. 그러나 이러한 어휘화 문법을 적용하게 되면 사전의 구조가 복잡해지고, 사전 구축에 많은 노력이 필요한 단점이 있다.

한국어에서는 술어의 역할이 매우 중요하기 때문에 규칙 기반 문법을 중심으로 하고, 술어의 어휘 특성 정보를 보조 정보로 이용하면 두 가지 문법 규칙의 장점을 함께 이용할 수 있다. 혼합된 방법을 사용하면 술어에 대해서만 사전에 자세한 정보를 표현하고, 나머지 카테고리에 대해서는 일반적인 규칙을 적용함으로써 사전 구축에 필요한 노력을 최소화할 수 있다. 그러나 기존의 한국어 술어의

[표 1] '보이다'에 대한 하위범주화 정보

보이다	
Sub-entry 1 Meaning : be watched or observed by others Probability : 0.427	
Case	Semantic Class
가 에게	{사람, 동물} {사람}
Sub-entry 2 Meaning : show something to others Probability : 0.573	
Case	Semantic Class
가 를 에게	{사람, 물품, 장소, 조직, 추상물, 추상사, 상태, 수량} {동물의 일부, 인공물, 추상물, 활동, 속성} {사람, 동물}

하위범주화 사전[김봉96][홍재97]은 전문가들의 직관에 의존하여 수작업으로 구축되었기 때문에 많은 오류를 포함하고 있다. 또 영역에 독립적인 하위범주화 사전을 구축하는 것은 매우 어려운 일이므로 상황에 따라서 각 응용영역에 맞는 하위범주화 사전을 쉽게 구축하는 방법이 필요하다. 따라서 본 연구에서는 구문구조부착 말뭉치를 이용해서 수작업을 최소화하여 반자동으로 통계, 의미정보가 포함된 술어의 하위범주화 사전을 구축하는 방법을 제안한다. 통계정보는 분석된 결과에 우선순위를 정할 때 사용되고, 의미정보는 일반화 과정을 통해서 선택제약 정보로 사용된다.

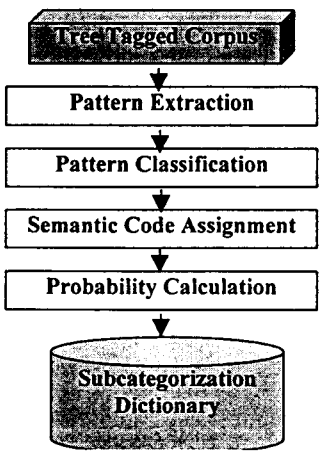
2. 술어의 하위범주화 정보 사전

술어의 하위범주화 정보는 문장의 구조를 분석할 때 사용될 수 있는 가장 중요한 정보의 하나이다. 하위범주화 정보 사전에는 술어의 인자(argument)의 종류 및 개수, 의미적 선택제약 정보, 하나의 술어가 여러 개의 용례으로 사용되었을 때 각각의 확률정보 등이 포함된다. [표1]은 술어 '보이'의 하위범주화 정보를 나타낸다.

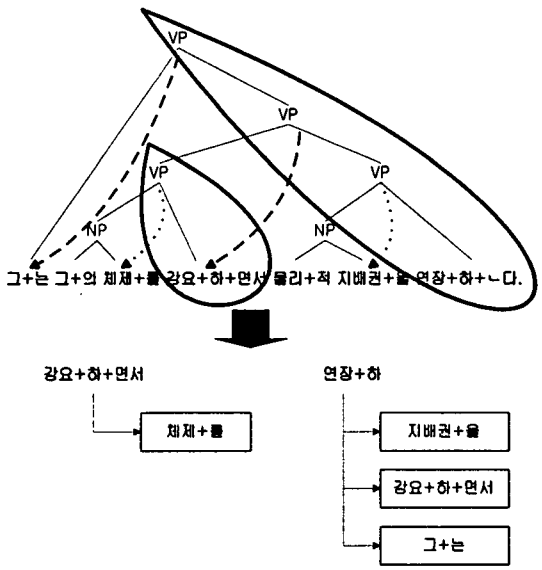
하위범주화 사전을 구축방법은 크게 전문가의 사전지식이나 직관에 의존하는 방법과 말뭉치 분석을 이용하는 방법 두 가지로 나눌 수 있다. 지금까지 한국어 술어의 하위범주화 사전은 대부분 첫 번째 방법에 의존해서 만들어졌다[김봉96][홍재97]. 그러나 이 방법은 작업자의 주관이 개입될 수 있고, 많은 수작업이 필요하며, 다른 영역으로 이식이 쉽지 않고, 확률정보를 얻을 수 없는 등 많은 단점이 있다. 한편 구문구조부착 말뭉치를 이용하면 객관적이며, 확률 정보를 얻을 수 있으며, 쉽게 다른 영역으로 이식할 수 있는 장점이 있다. 그러나 구문구조부착 말뭉치를 이용하더라도 몇몇 단계에서는 전문가의 수작업이 반드시 필요하다. 하나의 용언이 구문적으로 사용되는 방법이 달라서 여러 개의 하위범주화 정보를 가질 때, 그 용언이 실제 문장에서 어떤 하위범주화 정보로 사용되었는 지를 구분하는 작업은 사람의 판단에 의존해야 한다. 여러 연구에서 자동으로 각각의 용례를 구분해 주는 방법이 소개되고 있지만[briscoe97][Baek97] 아직 실험실 수준에 머물고 있다. 또 술어의 선택제약 정보를 입력하는 작업도 수작업에 의존해야 한다.

3. 반자동에 의한 하위범주화 정보 구축

본 연구에서 제안한 방법은 [그림 1]과 같이 4 단계로 구성된다.



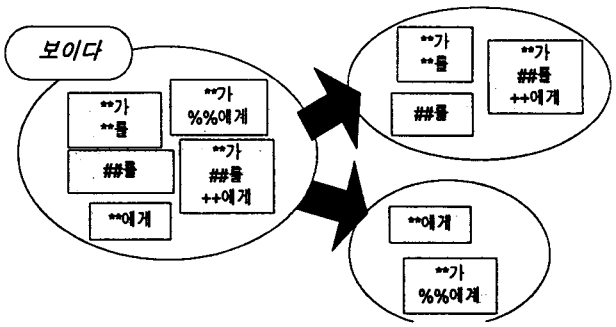
[그림 1] 구문구조부착 말뭉치를 이용한 하위범주화 사전 구축 과정



[그림 2] 구구조트리에서 의존트리로 변환

가. 구문구조 부착 말뭉치에서 술어-인자 쌍 추출

본 연구에서 사용된 구문구조부착 말뭉치[1]는 약 10,000 문장 100,000어절로 구성된다. 이 말뭉치에는 형태 및 구문태그가 같이 부착되어 있고, 문장의 구조는 괄호의 형태로 표현된다. 문장의 길이는 2-20어절 사이이고 평균 문장의 길이는 8.925어절이다. 구문구조부착 말뭉치는 구구조 문법을 기반으로 만들어졌기 때문에 술어-인자 쌍을 추출하기 위해서는 의존구조로 바꾸어주는 작업이 필요하다. 한국어 문장을 의존구조로 표현하였을 때 지배소가 항상 의존소의 뒤에 위치한다는 원칙에 따라서 어떤 구 안에서 가장 마지막 어절이 나머지 구의 head의 지배소로 작용한다고 볼 수 있다. [그림 2]는 구구조트리에서 의존트리로 변화시키는 과정을 보여준다. '강요하'를 포함한 서술구에서 마지막 어절 '강요하'는 자식 명사구의 head인 '체제를'의 지배소가 된다. 또 '연장하'를 포함하는 구는 세 개의



[그림 3] 술어의 용례에 따른 술어-인자 쌍의 분류

자식 구를 가지는데 각각의 구의 head인 '그는', '강요하면서', '지배권을' 을 의존소로 가진다.

나. 술어-인자 쌍의 분류

앞 단계에서 추출된 술어-인자 쌍들을 각 술어별로 분류한 다음, 각 술어에 포함된 술어-인자 쌍들을 사용된 패턴에 따라서 분류하는 작업이 필요하다. 구분하는 기준은 동일한 패턴의 격을 인자로 가지면 같은 그룹에 속하는 것으로 한다. [그림 3]에서 술어 '보이'가 구문적으로 두 가지로 사용될 수 있는데, 한 가지는 '가', '에게'를 인자로 가지고, 다른 한 가지는 '가', '를', '에게'를 인자로 가진다. 문장에서 용언이 어느 용법으로 사용되었는지를 구분하는 것은 사람이 그 용언이 사용된 문장을 직접 보면서 판단하는 수작업에 의존하여야 한다. 다음 단계에서는 이렇게 분류된 패턴을 이용해서 각 격과 어울릴 수 있는 명사를 일반화하는 작업이 필요하다.

다. 명사의 의미제약 할당 및 일반화

술어의 인자는 명사-조사 쌍으로 구성된다. 말뭉치를 이용해서 하위범주화 사전을 만들 때 조사는 종류가 많지 않기 때문에 각 술어의 용도에 따라 고정될 수 있지만, 명사는 종류가 매우 다양하기 때문에 자료의 희귀성 문제 (data sparseness problem)가 발생할 수 있다. 따라서 많은 연구에서 이 문제를 완화하기 위하여 클래스 즉 개념분류체계의 카테고리를 기반으로 하는 모델을 제안하고 있다. 이 모델을 이용하면 시스템이 학습 말뭉치에서 보지 못했던

[1] 이 구문구조부착 말뭉치는 한국과학기술원에서 과학기술처 STEP2000 과제의 세부과제로 구축함.

입력에 대해서도 유사도 계산을 통해 수용성을 결정할 수 있는 판단 근거를 제공한다.

이 연구에서 사용하는 명사의 개념분류체계는 [조정96]의 연구를 따른다. 이 개념분류체계는 트리 구조를 가지며, 3에서 5까지 서로 다른 깊이를 가지고, 단말노드는 60개 정도로 구성된다. 구문분석 단계의 애매성 해소를 위해서는 이 분류체계와 같이 비교적 큰 단위의 분류를 이용하고도 좋은 성능을 발휘한다. 또한 트리 구조는 상하위 관계를 쉽게 나타낼 수 있고, 간단한 연산을 통해서 개념 노드사이의 유사도를 쉽게 계산할 수 있다.

추출된 술어-인자 쌍과 원래의 문장을 보면서 각 인자의 명사가 어떤 의미 코드로 사용되었는지 결정한다. (a)와 (b)에서 명사 '병'은 서로 다른 의미로 사용되었다. 각각의 술어-인자 쌍과 원래 사용된 문장을 같이 보면서, 명사가 사용된 의미를 각각 2224(생리현상)와 1144(도구)로 분류한다.

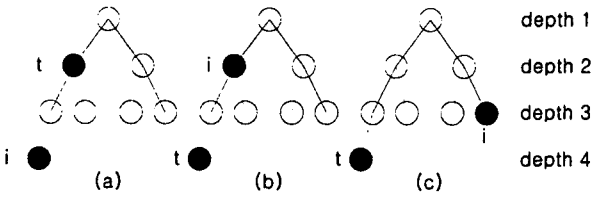
(a)

병이 들어서 몸이 아픈 경우도 있다.
병/ncn 이/jcs 들/pvg
2224(생리현상)

(b)

그 사람은 병을 들고 손님들을 위협했다.
병/ncn 을/jcs 들/pvg
1144(도구)

일반화는 개념분류체계 상의 서로 다른 위치에 있는 두 개의 개념 사이에 유사도를 계산하는 방법이다. 지금까지 일반화에 관한 연구는 많이 있었지만[Framis94][Li95][Kim94], 실용적인 수준에서 만족스러운 결과는 내지 못하고 있다. [Framis94][Li95]에서 제안한 방법은 엄청나게 많은 계산량과 복잡한 연산이 필요하기 때문에 실제 적용하기에는 많은 문제점이 있다. 본 연구에서는 [Kim94]에서 사용한 유사도 측정법을 사용한다. 이 방법은 계산이 간단하고, 트리 모양의 개념분류체계의 상하위어 관계를 잘 이용할 수 있다. 두 개념 카테고리 사이의 유사도는



[그림 4] Most Specific Common Abstraction 계산

MSCA(Most Specific Common Abstraction), Is-a penalty 두 가지 요소에 의해 결정된다. MSCA는 개념분류체계에서 두 개념 카테고리의 공통 조상 중 가장 깊이가 깊은 개념 즉 두 개념에 가장 가까운 공통 조상을 말한다. [그림 4]의 (a), (b)에서 개념 t와 개념 i의 MSCA의 깊이는 2이고, (c)의 개념 t와 개념 i사이의 MSCA의 깊이는 1이다. 실제 사용되는 명사의 개념 카테고리는 단말노드의 개념일 수도 있고, 일반화된 중간노드의 개념일 수도 있다. 입력문장에서 사용된 의미 i가 하위범주화정보의 의미 t의 자손 노드일 때는 다른 경우보다 두 개념이 더 유사하다고 볼 수 있기 때문에 Is-a penalty를 사용하여 차별성을 둔다. 의미 i가 의미 t의 자손 노드일 때 Is-a penalty는 1이고 그렇지 않은 경우는 0.5이다. [그림 4]에서 (a)의 경우는 Is-a penalty가 1이고 (b),(c)의 경우는 0.5이다.

두 개념카테고리 사이의 유사도는 아래와 같이 계산된다. SIM(i,t)의 첫 번째 항은 두 개념 사이의 거리가 가까울수록 MSCA의 깊이가 깊어진다는 근거에서 만들어 졌다.

$$SIM(i,t) = \begin{cases} \dfrac{2*level(MSCA(i,t))}{level(i)+level(t)} * Is\text{-}a\ penalty(i,t) \\ 1 \qquad if\ two\ categories\ are\ matched \end{cases}$$

$$Is\text{-}a\ penalty(i,t) = \begin{cases} 1 & if\ t\ is\ the\ ancestor\ of\ i \\ 0.5 & otherwise \end{cases}$$

i: 입력 문장에서 사용된 명사의 의미코드

t: 하위범주화 사전의 인자에서 사용된 명사 의미코드

라. 확률값 계산

하나의 술어가 여러 개의 하위범주화 정보를 가질 때 각각의 경우에 빈도수를 기반으로 확률값을 줌으로써 분석결과에 우선순위를 매길 수 있다.

전체 술어의 집합 predicate를 아래와 같이 표시한다.

$$\mathbf{predicate} = \{ \mathbf{pred_1}, \mathbf{pred_2}, \ldots, \mathbf{pred_n} \}$$

이 때 각각의 술어 $pred_i$는 여러 개의 하위범주화 정보 $Subcat(pred_i)$를 가질 수 있다.

$$\mathbf{Subcat(pred_i)} = \{ \mathbf{Sub_{i,1}}, \mathbf{Sub_{i,2}}, \ldots, \mathbf{Sub_{i,m}} \}$$

술어 $pred_i$의 각 하위범주화 $Sub_{i,k}$가 사용된 빈도수를 아래와 같이 나타낼 수 있다.

$\mathbf{count(Sub_{i,k}/pred_i)}$: 술어 $\mathbf{pred_i}$가 $\mathbf{Sub_{i,k}}$으로 사용된 경우의 수$(\mathbf{1 \le k \le m})$

이 때 술어 $pred_i$가 각 하위범주화 정보 $Sub_{i,k}$로 사용될 확률은 다음과 같다.

$$prob(Sub_{i,k}/pred_i) = \frac{count(Sub_{i,k}/pred_i)}{\sum_{j=1}^{m} count(Sub_{i,j}/pred_i)} \quad (1 \le i \le n)$$

4. 실험

본 연구에서는 빈도수가 높고, 구문적으로 여러 개의 용법을 가지는 술어 20개를 선택하여 위의 방법으로 술어의 하위범주화 정보를 구축하였다. 구문구조부착 말뭉치의 규모가 작아서 모든 술어에 대해서 충분한 수의 술어-인자 쌍을 얻지 못하기 때문에 빈도수가 높은 술어 20개를 선택했다. 대상 술어는 최대 빈도수 399회에서 최소 빈도수 75회로 구성된다. 이 방법으로 구축한 사전의 구문 분석의 애매성 해소에 대한 평가를 위해서 구문적으로 여러 개의 용법을 가지는 술어를 선택했다. 전체 구축과정에서 하나의 술어에 대해서 두 명의 전문가들이 맡아서 용례별 분류 작업을 했고, 의미 코드 할당 작업도 한 번 작업 후 검증 단계를 통해 작업자의 주관이 개입되는 것을 최소화 했다.

구축된 의미, 통계정보가 포함된 하위범주화 정보를 본 연구실에서 개발 중인 구문분석기 KOSA(Korean Syntactic Analyzer)에 적용시킨 결과 기존의 단순한 술어 하위범주화 정보만을 이용하였을 경우와 비교해서 구문분석 결과의 정확도를 높일 수 있었다. 정확도 평가에 대한 자세한 내용은 [장명97]에 설명되어 있다.

5. 결론 및 향후 연구

본 연구에서는 구문구조부착 말뭉치를 이용해서 술어의 하위범주화 정보를 손쉽게 구축하는 방법을 제안하였다. 구문구조부착 말뭉치에서 술어-인자 패턴을 추출하고, 각 패턴을 용례에 따라 나눈 다음, 명사에 의미코드를 넣고, 각 용례별로 확률값을 계산하여 하위범주화 사전을 완성하였다. 또 이 방법이 실제 구축과정 및 응용시스템의 적용을 통해서 유망한 방법임을 보였다.

그러나 대량의 정보 베이스 구축을 위해서는 대규모의 구문구조부착 말뭉치가 필요하고, 수작업으로 진행하였던 술어의 용례별 구분 작업과 명사의 의미코드 할당 작업을 자동화할 수 있는 연구가 더 필요하다. 그리고 구축된 하위범주화 정보 사전을 객관적으로 평가할 수 있는 평가 및 검증 방법에 대한 연구가 더 필요하다.

참 고 문 헌

[Baek97] Dae-Ho Baek, Ho Lee, Hae-Chang Rim, "Conceptual Clustering of Korean Concordances Using Similarity between Morphemes," Proceeding of ICCPOL '97, Vol 1, pp. 19-24, April 1997.

[Briscoe97] T. Briscoe, J. Carroll, "Automatic Extraction of Subcategorization from Corpora," Proceedings of 5th Conference on Applied Natural Language Proceedings, pp.356-363, 1997.

[Framis94] F. R. Rramis, "An Experiment On Learning Appropriate Selectional Restrictions From A Parsed Corpus," Proceedings of COLLING ,1994.

[Kim 94] E. J. Kim, J. H. Lee, G. B. Lee, "A Lexical Transfer Model Using Extended Collocational Patterns in COBALT J/K," Proceedings of International Conference On Chinese and Oriental Language, 1994.

[Li 95] H. Li, N. Abe, "Generalizing Case Frames Using a Thesaurus and the MDL Principle," Proceedings of International Conference on Recent Advances in Natural Language Processing, pp. 239-248, 1995.

[김봉 96] 김봉모, "한국어 문장 분석을 위한 용언의 하위 범주화," 국어정보처리기술개발과제 최종보고서, 부산대학교 국어국문학과, 1996.

[장명 97] 장명길 외 3 명, "통계/의미 정보를 이용한 한국어 의존파싱," 제 9 회 한글 및 한국어 정보처리 학술대회 발표 논문지, 1997.10.

[조정 96] 조정미, 김길창, "한국어 의미 해석시 중의성 해소에 대한 연구," 정보과학회지 제 14 권 제 7 호, pp. 71-83, 1996.7.

[홍재 97] 홍재성 외 9 명, "현대 한국어 동사 구문 사전," 두산동아, 1997.