(제 10회 한글 및 한국어 정보처리 학술대회)

구문 분석에서의 중의성 해소를 위한 일반화된 어휘정보의 자동 구축 및 적용

정후중, 황영숙, 곽용재, 박소영, 임해창

고려대학교 컴퓨터학과
서울시 성북구 안암동 5가 1 (우:136-701)
{hjchung, yshwang, yjkwak, ssoya, rim}@nlp.korea.ac.kr

Automatic Construction of Generalized Lexical Information for Syntactic Ambiguity Resolution

Hoojung Chung, Young-Sook Hwang, Yong-Jae Kwak, So-Young Park, Hae-Chang Rim
Department of Computer Science & Engineering
Korea University

요 약

구문 분석에서의 중의성을 해결하는 데 어휘정보가 유용하다는 것은 잘 알려져 있다. 그러나 기존의 어휘정보 구축 방법들은 많은 수작업을 요구하거나,자동으로 구축하는 경우에는 어휘 자체를 그대로 사용함에 따라 심각한 자료 희귀성 문제가 발생했다. 본 논문에서는 구문 분석에서의 중의성 해소를 위해 원시 코퍼스와 시소러스를 이용하여 개념 수준(conceptual-level)의 일반화된 술어-인자 어휘정보를 자동으로 구축하고, 이를 파서에 적용하는 방법을 제안하고자 한다. 제안한 방법으로 구축한 일반화된 어휘정보를 파서를 이용하여 명사구의 지배소 결정 실험에 적용하여 본 결과, 정확도가 85.9%에서 91.5%로 향상되었다. 또, 미지격 결정 실험에 대해서는 86.32 %의 격 결정 성공률을 보여주었다.

1. 서 론

구문 분석이란 주어진 문장의 구조를 주어진 구문 규칙에 따라 분석하는 작업을 말한다. 구문 분석 과정에서는 보통 하나 이상의 구문 구조가 구해지며, 이들 중 올바른 구문 구조를 선택하는 작업을 구조적 중의성의 해소 작업이라고 하는데, 일반적으로 PCFG나 PDG와 같은 확률 문법을 사용하여 이와 같은 문제를 해결하고 있다.

그런데, PCFG와 PDG의 구문 규칙은 일반적으로 구문 태그와 품사 태그로만 표현된다. 따라서 PCFG나 PDG로도 해결하지 못하는 구조적 중의성 문제가 여전히 존재하게 된다. 다음의 예문을 보자.

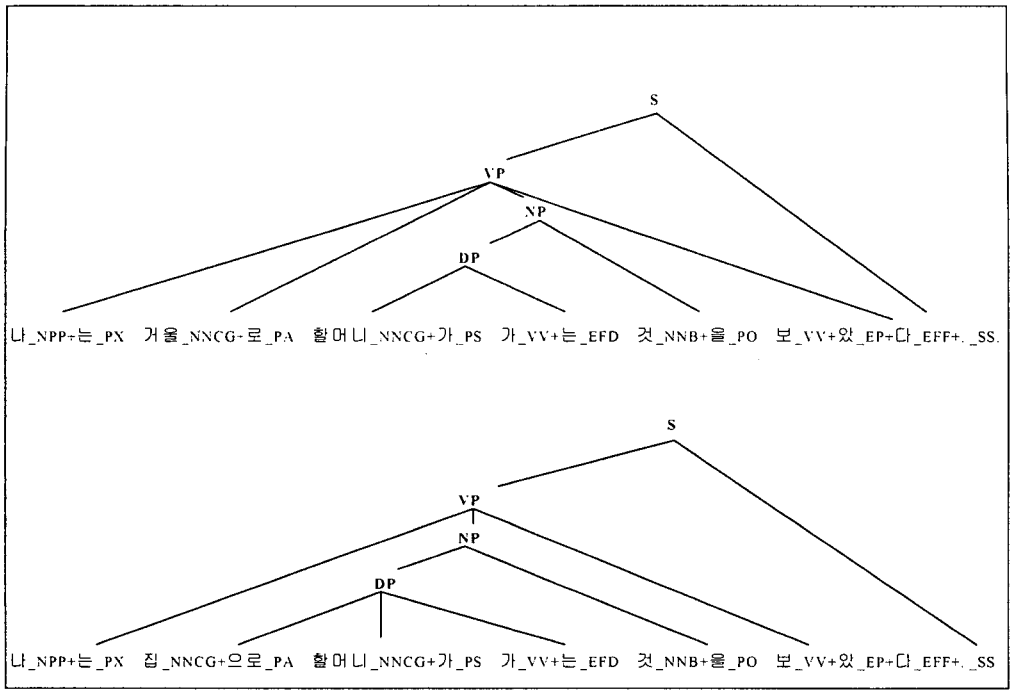
- 나는 거울로 할머니가 가는 것을 보았다.
- 나는 집으로 할머니가 가는 것을 보았다.

두 예문은 동일한 품사열로 구성되어 있으나, 구문 구조는 서로 다르다[그림 1]. 구문 태그와 품사태그만으로 표현되는 확률 문법들은 위와 같이 동일한 품사열을 가지지만 서로 다른 구문 구조를 가지는 문장들을 구별하지 못하고, 두 문장을 동일한 구조로 분석한다. 따라서 확률문법을 보완할 추가의 정보가 필요한데, 이 때 사용되는 정보에는 격틀 정보나 동사 하위 범주화 정보 같은 어휘정보가 있다.

그런데, 어휘정보를 구축한다는 것은 많은 수작업과 시간을 필요로 하기 때문에, 어휘정보를 좀 더 쉽게 구축하는 연구가 많이 진행되어 왔다. 구문 분석시 발생하는 중의성 해소에 사용되는 어휘정보 구축에 대한 연구들을 크게 두 부류로 나눈다면, 하나는 어휘정보를 반자동으로 구축하는 연구이고, 다른 하나는 어휘정보를 자동으로 구축하는 연구이다. 반자동으로 어휘정보를 작성하는 경우는 사람이 개입되므로 정확한 정보를 얻을 수 있고, 어휘 수준(lexical-level)이 아닌 개념 수준의 정보를 작성할 수 있다는 장점이 있으나 수작업을 많이 요구한다는 문제가 있다.

자동으로 어휘정보를 추출하여 사용하는 경우엔 코퍼스로부터 뽑은 어휘 공기 정보를 그대로 사용하기 때문에 구문 분석에 실제로 적용시, 자료 희귀성 문제가 발생된다는 문제점이 있다.

본 논문에서는 이런 문제점들을 해결하기 위한 방법으로 일반화 된 어휘정보를 자동으로 구축한 후 구문분석기에 적용하여 중의성을 해소하는 방법을 제안한다. 여기서 말하는 일반화 된 어휘정보란 술어와 그 인자로 쓰이는 개념들에 대한 정보를 말한다. 어휘 수준의 정보가 아니라, 개념



[그림 1] 동일한 품사열을 가지나 구문 구조가 틀린 문장

수준의 정보이기 때문에, 어휘 수준의 정보를 적용할 때 발생되는 자료 희귀성 문제가 어느 정도 해소된다.

앞에서 설명한 기존 방법의 문제점을 2장에서 다시 살펴본 후, 3장에서 이를 해소하기 위하여 원시 코퍼스와 시소러스를 이용하여 일반화된 어휘정보를 자동으로 추출하는 방법을 보여주고, 4장에서 이 정보를 명사구의 지배소 결정과 미지격 해결에 적용하는 방법을 제시한다. 5장에서는 실험을 통하여 본 논문에서 제안한 방법의 타당성을 보여주고, 5장에서 논문의 결론을 맺고자 한다.

2. 어휘정보 추출 관련 연구

술어와 인자 사이의 어휘정보를 반자동으로 작성하는 연구에는 [4]와 [5]가 있다. [4]에서는 원시코퍼스와 품사 태깅된 코퍼스로부터 관련이 있을 법한 동사-명사-격조사 어휘 리스트를 자동으로 추출한 후, 사람이 후처리를 통하여 잘못된 항목을 제거시키고 어휘 리스트 중 명사부분은 사람이 직접 시소러스에서의 클래스로 결정하여 어휘정보를 일반화시키는 방법을 사용하고 있다.

[5]에서는 반자동으로 술어의 하위 범주화 정보를 구축하는 방법을 제안하였다. 구문 구조가 부착된 코퍼스로부터 술어에 대한 인자(명사-조사)들을 추출하여 전문가가 술어가 사용된 용법에 따라 수작업으로 술어를 분류하고, 추출된 인자의 명사를 수작업으로 시소러스에 있는 의미 코드로 일반화하여 술어의 하위 범주화 사전을 생성한다.

이와 같은 방법들은 어휘정보의 구축 시, 반드시 사람의 개입을 필요로 한다. 비록 반자동 방법이라고 하나, 정보를 작성해야 하는 어휘의 수가 엄청나기때문에 어휘정보 추출 과정과 일반화 과정 에서의 수작업 양은 여전히 많아지게 된다.

어휘정보를 자동으로 추출하여 사용한 연구에는 [7]과 [8]이 있다. [7]에서는 대량의 원시코퍼스로부터 술어-명사-격조사 공기 정보를 자동으로 추출하여 구문 분석기에서 중의성을 해결하는데 이용하고 있다. [8]에서는 구문구조가 부착된 코퍼스로부터 어휘 공기 정보를 추출하여 구조적 중의성 해결에 사용하였다. 두 연구 모두 일반화를 시도하지 않고 코퍼스로부터 추출한 어휘 수준의 공기 정보를 그대로 이용하였으므로 학습 코퍼스에서 등장하지 않은 어휘에 대해서는 자료 희귀성 문제가 발생한다. 실제로 [8]에서 사용한 어휘정보는 학습 데이터에 대해서는 괄목할만한 정확도 향상을 보이는 데 비해, 실험 데이터에 대해서는 약 1% 정도의 향상만을 보이고 있다고 한다. [7]의 경우에도 3000만 어절이라는 대용량 코퍼스에서 정보를 추출했으나, 구문 분석에 적용 시 자료 희귀성 문제가 발생하였다.

3. 일반화된 어휘정보의 자동 추출

앞서 설명한 어휘정보의 추출 방법들은 많은 수작업을 필요로 하거나, 자료 희귀성 문제를 유발하게 된다. 이런 문제를 해결하기 위하여 본 논

문에서는 코퍼스로부터 술어-인자에 대한 어휘 수준의 정보를 추출한 후, 자동으로 개념 수준의 정보로 일반화시키는 방법을 제안한다. 개념 수준의 정보로 일반화시킨다는 것은 어휘 수준의 정보 중 명사 부분을 개념으로 일반화시킨다는 것이다[그림 2]. 이렇게 해줌으로써 자료 희귀성 문제를 어느 정도 극복할 수 있다. 본 연구에서 제안한 방법은 크게 두 단계로 구성되어 있는데, 첫 단계에서는 원시코퍼스로부터 술어-격조사-명사 어휘 공기 정보를 추출하고, 두 번째 단계에서는 앞 단계에서 구한 공기 정보 중 명사 부분을 시소러스의 개념으로 일반화한다.

3.1. 술어-격조사-명사 정보의 추출

술어-격조사-명사간의 의존관계를 정확하게 알아내기 위해서는 구문태깅된 코퍼스를 사용하는 것이 가장 좋을 것이다. 그러나 원하는 어휘정보를 충분히 추출하기에는 구문태깅된 코퍼스의 양이 많지 않고, 구문태깅된 코퍼스를 사용하지 않는 한 술어와 의존관계를 갖는 모든 명사를 올바르게 추출하는 것은 불가능하다. 이에 본 논문에서는 어휘정보를 추출할 원시 코퍼스의 문장들을 자동품사태거로 태깅하여 용언의 연결어미나 종결어미를 기준으로 단문으로 분리한 후, 의존관계를 갖는 것이 확실한 술어-격조사-명사 리스트를 추출하는 방법을 사용한다. 즉, 각 단문의 마지막 술어와 그 앞 술어 사이에 있는 인자들만을 어휘정보 추출대상으로 삼음으로써 잘못된 술어-격조사-명사를 추출할 가능성을 배제시킨다. 이 단계에서 오류를 포함할 경우 다음 단계인 명사 일반화 단계에서 오류를 파생시킬 수 있기 때문에 추출되는 정보의 양이 적더라도 정확한 의존관계를 갖는 어휘정보들만을 추출하도록 한다.

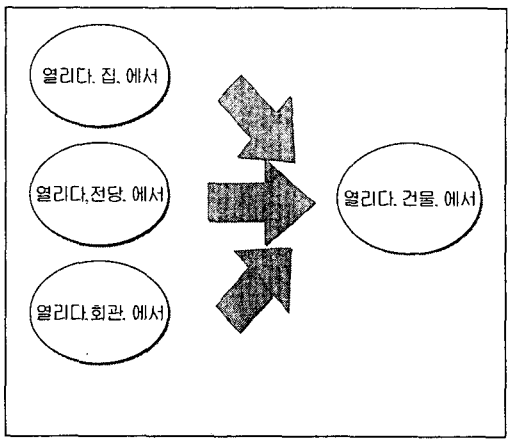
다음은 원시 코퍼스의 한 문장으로부터 술어-격조사-명사 정보를 추출하는 예이다.

- 학교에서 열린 학회에 참석하지 않고, 우리는 밥을 먹었다.
  → 학회에 참석하지 않고 +
  우리는 밥을 먹었다.
  → <참석하다-학회-에>
  <먹었다-밥-을>

예문의 "학교에서"라는 명사+격조사 어절은 지배 술어가 "열린"인지 "참석하지"인지 알 수가 없으므로 <열리다, 학교, 에서>라는 어휘 쌍은 오류의 가능성이 있다고 여겨져 추출되지 않는다. 또 "우리는"이라는 어절은 술어 "먹었다"와 의존 관계를 갖는 것이 분명하나, 격조사가 아닌 보조사가 사용되어 격관계를 알 수 없으므로, <먹었다,우리,는> 어휘 쌍 또한 추출되지 않는다.

3.2. 명사의 일반화

추출된 술어-격조사-명사 정보를 어휘정보 그 자체로 사용한다면 학습 코퍼스에서 나타나지 않았거나 낮은 빈도를 갖는 명사들로 인하여 심각한 자료부족 문제가 발생할 수 있다. 이를 해결하기 위해 추출한 술어-명사-격조사 정보에서 명사 부분을 시소러스의 개념 항목으로 대체하여 명사에 대해 일반화를 수행한다. 명사를 시소러스의 개념항목으로 대체하기 위해서는 명사가 사용된 의미(word sense)를 알아야 한다. 예를 들어 [9]에서 생성한 시소러스에서 명사 '방문'은 '문'이라는 개념의 하위 항목인 의미와, '글'이라는 개념의 하위 항목인 의미, 두 가지를 갖는데, 어떤 의미로 사용되었는지를 모른다면 적절한 개념으로 일반화해 줄 수가 없다.



[그림 2] 동사 "열리다"의 부사격조사 "에서"와 함께 사용되는 인자의 일반화

명사의 의미를 알아내기 위해 본 논문에서는 다음과 같은 가정을 사용하였다.

- 가정 : 같은 격조사 및, 같은 술어와 함께 사용되는 명사들은 유사한 의미 자질을 갖는다.

두 단어가 유사한 의미 자질을 갖는다는 것은 시소러스에서 두 단어 사이에 MSCA(Most Specific Common Abstraction)가 존재하며, 그 단어들 사이의 거리가 가깝다는 것을 의미한다. 예를 들어, 술어-격조사 쌍인 <'열리다'-'에서'>와 함께 사용된 명사들 "집", "전당"은 시소러스에서 두 단어 사이에 MSCA가 존재하고, 두 단어 사이의 거리가 가깝다는 것이다.

그런데 시소러스에서 두 명사의 MSCA 깊이가 깊을 수록 두 명사의 의미가 더욱 유사하다는 것을 고려해야 하므로, 시소러스에서 두 명사 n과 m 사이의 상대적 거리 $d(n,m)$ 를 다음과 같이 정의한다.

$$d(n,m) = \frac{dist\_in\_thesaurus(n,m)}{depth(MSCA(n,m))}$$

위 식에서 $dist_in_thesaurus(n,m)$은 시소러스에서 두 단어 사이의 실제 거리를 의미하고,

열리다_VV	
조사	클래스(의미코드) : 어휘들
가	문(0) : 문, 대문, 창문, 방문 모임(0) : 총회, 공청회, 연주회, 토론회, 전시회, 음악회, 대회, 회의 기관(0) : 국회, 위원회 열매(0) : 과일, 열매
로	앞(1) : 처음, 앞
서	건물(0) : 집, 전당, 회관 지역(0) : 각지, 지역

[표 1] 동사 "열리다"에 대한 어휘정보

*depth(w)*는 시소러스에서의 주어진 단어 w의 깊이를 의미한다.

문장에서 사용된 명사 n의 의미를 결정할 때는 다음과 같은 식을 이용하여 두 명사 사이의 상대적 거리 d가 최소가 되는 명사의 의미 i를 선택한다.

$$\text{명사 } n\text{의 의미} = argmin_i(d(n_i, M_j))$$

식에서 M은 n과 같은 <술어-격조사>와 사용되는 명사들의 집합이다. n_i란 명사 n의 i번째 의미에 해당하는 시소러스에서의 개념을 말한다. 가정에 의해서 위 수식의 결과인 i가 명사 n의 의미가 된다. 이를 이용하면 앞 단계에서 추출한 술어-명사-격조사 리스트를 가지고 리스트 내 명사들의 의미 결정을 할 수 있게 된다.

명사의 의미를 결정했다면, 비슷한 개념을 가지는 명사들을 묶어 일반화하는 작업이 필요한데. 이는 시소러스에서 찾은 명사들의 MSCA를 이용한다. 동일한 <술어-격조사>를 가지는 모든 명사들을 쌍으로 묶어 클러스터로 만든 후, 두 클러스터의 MSCA 사이에 조상-자손 관계가 존재하는 클러스터들을 하나의 클러스터로 묶으면 된다.

모든 동사와 격조사와 함께 추출된 명사에 대해서 이 작업을 수행하면 일반화된 어휘정보를 얻을 수 있다. [표 1]은 위와 같은 과정을 거쳐 얻은 동사 "열리다"의 일반화된 어휘정보이다.[1)]

4. 추출한 어휘정보의 적용

4.1. 어휘 연관도

추출한 어휘정보를 파서에서 사용할 때에는 어휘 연관 정도를 나타내는 *Assoc()* 함수를 이용한다. *Assoc()* 함수는 추출한 어휘정보에 포함되어 있는 통계 정보를 이용하여 최적의 어휘 연관 관계를 구한다. 연관도 함수 *Assoc()*는 다음과 같이 표현한다. [4] [7]

$$Assoc(v, n, j) = \alpha \cdot Req_{NJ}(v, n, j) + (1-\alpha) \cdot Req_J(v, j) \qquad (0 \le \alpha \le 1)$$

연관도 함수 *Assoc(v, n, j)*는 동사-명사-격조사가 통계적으로 어느 정도 관련이 있는지를 표현하는 함수이다. 연관도 값은 "명사-격조사 요구도"인 Req_{NJ}와 "격조사 요구도"인 Req_J의 합으로 이루어지며, 변수 α는 "명사-격조사 요구도"와, "격조사 요구도" 사이의 비율을 결정한다. Req_{NJ}와 Req_J는 다음과 같이 계산한다.

$$Req\_NJ\ (v, n, j) = \max(P(n, j|v), \frac{P(class(n), j|v)}{N}) \cong \frac{\max(f(v, n, j), \frac{f(v, class(n)j)}{N})}{f(v)} \qquad (when\ f(v) \neq 0)$$

$$Req\_J(v, j) = P(j|v) \cong \frac{f(j, v)}{f(v)} \qquad (when\ f(v) \neq 0)$$

*Req_NJ*는 술어가 주어졌을 때, 명사와 격조사를 요구하는 정도인데, 개념으로 일반화된 정보와 일반화되어 있지 않은 어휘 수준의 정보 중, 빈도가 높은 정보를 이용하여 구한다. *class(n)*이란 명사 n이 포함된 개념을 의미하고, N은 개념 *class(n)*에 속한 명사들의 갯수를 말한다.

*Req_J*는 술어가 주어졌을 때 격조사가 요구될 확률이다. 연관도 값, *Assoc()* 를 구하는 데 *Req_NJ*뿐만 아니라 *Req_J*도 함께 사용한 이유는 자료 희귀성 문제를 보완하기 위해서이다. 만약 연관도를 구하려는 명사가 함께 주어진 술어-격조사와 한번도 함께 발생하지 않았고, 명사가 속하는 어떠한 개념에서도 어휘정보를 찾을 수 없다면, 술어의 격조사 요구 정도만을 가지고 연관도 값을 구한다.

4.2. 어휘 연관도의 적용

본 논문에서 어휘정보를 적용하여 해결하려는 문제는 의존 파싱에서 발생하는 명사구의 지배소 선택 문제와 미지격 결정 문제이다. 이 두 부분은 어휘정보를 사용하지 않고는 해결하기가 어려운 부분이다. 명사구의 지배소 선택 문제는 명사구가 가지는 여러 지배소 후보들 중에서 올바른 지배소를 선택하는 문제이다. 또 미지격 결정 문제란 격이 생략된 술어와 명사 사이의 올바른 격관계를 결정하는 문제를 말한다.

어휘 연관도를 이용하여 명사구의 지배소를 선택하는 예를 살펴보자.

1) 각 어휘 및 격조사가 갖는 확률은 표시되어 있지 않다.

• 바람이 열린 창문으로 들어왔다.

위의 예문에서 어절 "창문으로"의 지배소는 어절 "들어왔다"라는 것이 분명하다. 그런데 어절 "바람이"는 "열린"과 "들어왔다"라는 두 지배소 후보를 갖기 때문에 올바른 지배소를 결정해야 하는 문제가 발생한다. 이런 경우 앞서 설명한 연관도 값을 사용하는데, *Assoc* (열리다, 바람, 이)와 *Assoc*(들어오다, 바람, 이)를 다음과 같이 비교하여 연관도 값이 더 높은 술어 "들어오다"를 "바람이"의 지배 어절로 결정한다. (α = 0.999)

$$\begin{aligned} &Assoc\ (\text{열리다}, \text{바람}, \text{이}) \\ &\quad = 0.999 \times 0 + 0.0001 \times 0.560 \\ &\quad = 0.0005 \end{aligned}$$

$$\begin{aligned} &Assoc\ (\text{들어오다}, \text{바람}, \text{이}) \\ &\quad = 0.999 \times 0.003 + 0.0001 \times 0.389 \\ &\quad = 0.0037 \end{aligned}$$

이번에는 미지격을 결정하는 예를 살펴보자. 위의 예문에서 관형형인 "열린"과 수식을 받는 "창문"의 사이에는 주격 관계가 성립한다. 이 미지격 관계를 결정할 때에는 다음과 같이 모든 격조사에 대해서 *Assoc(열리다, 창문, 모든 격조사)* 값을 계산하여 최대 어휘 연관도 값을 가지는 격조사를 선택해주면 된다.

$$\begin{aligned} &Assoc\ (\text{열리다}, \text{창문}, \text{이}) \\ &\quad = 0.999 \times 0.0255 + 0.0001 \times 0.5603 \\ &\quad = 0.0261 \end{aligned}$$

$$\begin{aligned} &Assoc\ (\text{열리다}, \text{창문}, \text{을}) \\ &\quad = 0.999 \times 0 + 0.0001 \times 0.0165 \\ &\quad = 0.00001 \end{aligned}$$

$$\begin{aligned} &Assoc\ (\text{열리다}, \text{창문}, \text{에}) \\ &\quad = 0.999 \times 0 + 0.0001 \times 0.0884 \\ &\quad = 0.00008 \end{aligned}$$

$$\begin{aligned} &Assoc\ (\text{열리다}, \text{창문}, \text{로}) \\ &\quad = 0.999 \times 0 + 0.001 \times 0.101 \\ &\quad = 0.0001 \end{aligned}$$

$$\begin{aligned} &Assoc\ (\text{열리다}, \text{창문}, \text{서}) \\ &\quad = 0.999 \times 0 + 0.0001 \times 0.232 \\ &\quad = 0.0002 \end{aligned}$$

주격인 경우의 연관도 값이 가장 높으므로 "열린"과 "창으로"의 격관계를 주격으로 결정한다.

5. 실험 및 평가

어휘정보의 일반화가 구문 분석에서 발생하는 중의성 해소에 어느 정도 도움을 줄 수 있는지를 평가하고 검증하기 위해서 명사구의 지배소를 선택하는 실험과 미지격을 결정하는 실험을 해보았다.

실험을 위하여 어휘정보를 추출한 코퍼스는 신문 기사, 잡지 기사, 소설, 설명문 등으로 구성된 약 800만 어절의 원시 코퍼스이며, 이 원시 코퍼스를 자동 품사 태거로 품사를 붙여준 후, 앞에서 설명한 방법으로 술어-명사-격조사 쌍들을 생성하였다. 격조사는 주격, 목적격, 보격과, 대표 조사가 '에', '로', '서'인 부사격 조사[2] 만을 고려하였다. 이렇게 얻은 술어-명사-격조사 쌍이 약 624,200 개의 분량이다.

일반화에 사용한 시소러스는 [9]에서 bottom-up 방식으로 생성한 시소러스로서, 12,833 개의 명사 항목으로 구성되어 있으며 최대 깊이는 17이다. 추출된 어휘정보의 명사를 일반화시킬 때는 빈도가 3 이상인 명사를 대상으로 했으며, 두 명사 사이의 상대적 거리가 0.85 이하일 때만 두 명사가 유사한 의미를 가진다고 고려하여 일반화를 시켰다. 결과 5000개 정도의 술어-개념-격조사 쌍이 추출되었다.

추출된 어휘정보는 의존 파서를 통하여 구문 분석 과정에 적용되었다. 이 의존 파서는 일반적으로 사용되는 지배소 후위 제약, 투영의 제약, 지배소 유일 제약, 일문 일표층격 제약 등이 가해졌으며, 최적의 분석 결과 하나 만을 출력한다. 최적의 의존 구조는 앞서 설명한 *Assoc()* 함수를 사용하여 구한다.

5.1. 실험 1 : 명사구 지배소 선택 실험

첫 번째 실험인 명사구의 지배소를 선택하는 실험은 어휘를 개념으로 일반화시켜 사용하는 것이 어휘 자체를 사용하는 것보다 유용하다는 것을 보여주기 위한 실험이다.

실험 대상 100 문장은 어휘정보를 추출한 학습 코퍼스에서 뽑은 문장 50개, 학습 코퍼스 이외의 신문 기사, 초등학교 교과서, 소설 코퍼스에서 뽑은 문장 50개이다. 본 논문에서 제안한 방법은 명사+격조사 어절과 술어의 결합 부분에만 영향을 미치므로 이 부분의 정확도만을 측정했다. 또한, 명사+격조사 어절이 결합할 수 있는 지배소 후보가 단 하나인 경우에는 중의성이 존재하지 않으므로, 이 경우도 제외했다. 이렇게 했을 때 학습 코퍼스와 실험 코퍼스에 각각 71개의 중의성을 갖는 명사+격조사 어절이 남았다. 의존 파서를 통해 구조적 중의성 해소 실험을 수행한 결과, 정확도는 [표 2]와 같았다.

표에서 보듯이 자동 추출한 어휘정보를 학습 코퍼스에 적용한 경우, 어휘정보를 일반화시킨 경우나 추출한 어휘정보를 그대로 사용한 경우나 중의성 해소율에 차이가 없었으나, 실험 코퍼스에 본 논문에서 제안한 방법을 적용했을 때, 약 6%의 정확도가 향상되었다.

일반화된 어휘정보로도 해결이 안된 경우는 명사구의 지배소 후보인 술어들이 비슷한 개념의

2) "에", "에는", "에도", 등 : 대표조사 "에"
"로", "으로", "으로는", 등 : 대표조사 "로"
"에서", "서", "에서는", 등 : 대표조사 "서"

	평균 중의성 정도	어휘정보만 사용한 경우 정확도	어휘정보와 개념정보를 함께 사용한 경우의 정확도
학습 코퍼스	2.52 개	90.1 %	90.1 %
실험 코퍼스	2.49 개	85.9 %	91.5 %

[표 2] 명사구의 지배소 선택 실험 결과 (평균 중의성 정도는 각 명사구가 가지는 평균 지배소 후보 갯수)

인자들을 요구하여 올바른 지배소를 선택하지 못한 경우였다. 본 실험에서 사용한 시소러스의 문제점때문에 오류가 발생하기도 했다. 예를 들어 '사람'이라는 개념이 '물건'의 하위 개념으로 분류되어 있기때문에 일반화 과정에서 오류를 유발했다.

5.2. 실험 2 : 미지격 결정 실험

반자동으로 어휘정보를 구축한 [4]에서는 미지격의 격결정 실험을 통하여 반자동으로 작성한 어휘정보의 유효성을 검증하였다. 미지격 결정 실험이란 동사의 관형형과 그것이 수식하는 명사 사이의 격관계, 격조사가 생략된 명사구와 지배동사의 격관계, 그리고 보조사가 포함된 명사구와 지배 동사의 격관계를 결정하는 실험을 말한다.

본 논문에서는 [4]에서 사용한 방법을 통하여 본 논문에서 제시한 방법으로 획득한 어휘정보의 타당성을 검증하고, 수작업으로 작성한 어휘정보와의 비교를 수행하여 보았다. [4]의 미지격 결정 실험에서 사용한 동사 10개[3]를 사용하여 학습 및 실험 코퍼스에서 그 동사가 포함된 50문장씩을 골라 미지격 결정 실험을 수행하였다. 총 500개의 문장에서 나온 미지격을 포함한 의존 관계는 534개였다. [표 3]은 실험 결과이다.

표에서 보듯이 동사에 따라서 [4]의 실험 결과와 어느 정도의 차이가 있지만, 전체 평균 정확도로 봤을 때, 본 논문의 방법으로 자동 생성한 어휘정보를 사용한 경우나, 사람이 개입하여 작성한 [4]의 어휘정보나, 미지격의 결정을 하는 데 큰 차이가 없음을 알 수 있다. 실험 대상 문장이 [4]의 실험에서 사용한 문장과 동일하지 않기때문에 단순 정확도 비교만을 할 수는 없지만, 어휘정보를 작성하는 데 들어가는 엄청난 수작업의 양을 고려할 때, 본 실험은 본 논문에서 제안하는 자동화된 방법의 유효성을 보여준다.

동사	[4]의 정확도	본논문에서 제안한 방법의 정확도
내리	81.81 %	83.01 %
만들	85.71 %	81.48 %
먹	85.14 %	90.38 %
받	86.43 %	85.96 %
보내	90.90 %	86.79 %
쓰	91.37 %	89.65 %
앉	88.24 %	78.84 %
열/열리	77.27 %	90.00 %
짓	77.78 %	96.22 %
타	88.89 %	79.24 %
평균	86.26 %	86.16 %

[표 3] 미지격 결정 실험 결과

6. 결론 및 향후 연구

본 논문에서는 원시 코퍼스와 시소러스를 이용하여 자동으로 일반화된 어휘정보를 구축하여 구문 분석에 적용하는 방법을 소개하였다. 원시 코퍼스를 자동 태거로 품사 태깅하여 단문으로 분리, 정확한 술어-명사-격조사 패턴을 추출하고, 자동으로 명사의 의미를 결정하고 일반화시켜 개념-술어-격조사로 이루어진 어휘정보를 구축하였다. 또, 이 정보를 구문 분석기에 적용하여 구조적 중의성 해소 및 격 중의성 해소 실험을 수행하였다. 한국어에서 술어의 역할이 매우 중요하기 때문에 술어-인자에 대한 어휘정보는 구문 분석 과정에서 유용하게 사용될 수 있을 것이다.

구문 분석 단계에서 어휘정보가 좀 더 유용하게 작용하기 위해서는 술어가 갖는 여러 인자들을 동시에 고려할 수 있어야 한다. 이러한 다항 인자 정보와 다항 인자 정보의 적용에 대한 연구가 더 필요하다.

3) [4]의 미지격 결정 실험에서 사용한 동사의 종류는 모두 11개이나, 동사 "흐르다"의 경우 실험한 예문이 한 문장에 불과하므로 본 실험에서는 제외했다.

참 고 문 헌

[1] Donald Hindle, Mats Rooth, "Structural Ambiguity and Lexical Relations", Computational Linguistics, Vol. 19, No 1. 1993

[2] Hui-Feng Li, Jong-Hyeok Lee, Geunbae Lee, "Identifying Syntactic Role of Antecedent in Korean Relative Clause Using Corpus and Thesaurus Information"

COLING-ACL, 1998

[3] Philip Stuart Resnik, "Selection and Information: A Class-Based Approach to Lexical Relationships", Ph.D. Thesis, Univ. of Pennsylvania , 1993

[4] 김선호, "통계 정보를 기반으로 한 어휘 관계 예측", 연세대학교 대학원 석사 학위 논문, 1996

[5] 류법모, 장명길, 박수준, 박재득, 박동인, "구문구조부착 말뭉치를 이용한 술어의 하위범주화 정보 구축", 제9회 한글 및 한국어 정보처리 학술대회 pp.116-121, 1997

[6] 엄미현, 신대규, 임병준, 나동렬, "다중파스 여과에 기반한 한국어의 구조적 중의성 해소", 제 8회 한글 및 한국어 정보처리 학술대회, pp.443-451, 1996

[7] 윤준태, "공기 관계 기반 어휘 연관도를 이용한 한국어 구문 분석", 연세대학교 대학원 박사 학위 논문, 1997

[8] 이공주, 김재훈, 김길창, "중심어간의 공기 정보와 구문 규칙을 기반으로 한 확률적 한국어 구문 분석", 제9회 한글 및 한국어 정보처리 학술대회 pp.332-338, 1997

[9] 조평옥, 옥철형, "한국어 명사 의미 계층 구조 구축", 제9회 한글 및 한국어 정보처리 학술대회 pp.129-135, 1997

[10] 홍재성 외 9명, "현대 한국어 동사 구문 사전", 두산동아, 1997