통계/의미 정보를 이용한 한국어 의존 파싱

장 명길*, 류 법모*, 박 재득*, 박 동인*, 맹 성현**
*시스템공학연구소 자연어정보처리연구부 **충남대학교 컴퓨터과학과

Korean Dependency Parsing Using Statistical/Semantic Information

Myung-Gil Jang*, Pum-Mo Ryu*, Jae-Deuk Park*, Dong-In Park*, Sung Hyun Myaeng**
***Dept. of Natural Language Information Processing, Systems Engineering Research Institute(SERI)**
****Dept. of Computer Science, Chungnam National University**

{mgjang,ryupm,jdpark,dipark}@seri.re.kr, shmyaeng@cs.chungnam.ac.kr

요 약

한국어 의존 파싱에서는 불필요한 의존관계의 과다한 생성과 이에 따른 다수의 구문분석 결과 생성에 대처하는 연구가 필요하다. 본 논문에서는 한국어 의존 파싱 과정에서 생기는 불필요한 의존관계에 따른 다수의 후보 의존 트리들에 대하여 통계/의미 정보를 활용하여 최적 트리를 결정하는 구문 분석 방법을 제안한다.

본 논문의 구문 분석에서 사용하는 통계/의미 정보는 구문구조부착 말뭉치(Tree Tagged Corpus)를 이용하여 구축한 술어 하위범주화 정보 사전에서 얻었으며, 이러한 정보를 활용한 구문 분석은 한국어 구문 분석의 모호성 해소에 적용되어 한국어 구문 분석의 정확도를 높인다.

1. 서 론

본 논문에서는 기본적으로 의존 문법을 이용한 의존 파싱 방법으로 한국어 구문분석을 시도하고 있다.

의존 문법에 의한 한국어 구문 분석 방법이 주목 받고 있는 이유는 한국어 어순의 자유성에 의한 문제점이 쉽게 해결되고 구성 요소의 불연속성이나 구성 요소의 생략 등과 같은 언어 현상에 큰 영향을 받지 않아 매우 견고성이 있는(robust) 파싱 방법의 구현이 가능하기 때문이다. 하지만, 이러한 잇점에도 불구하고 의존 문법에 의한 한국어 구문 분석은 불필요한 의존 관계의 생성을 억제하는 방법과 이에 따른 다수의 구문 분석 결과에 대처하는 방법에 대한 연구, 투사성의 원리에 맞지 않는 문장들과 지배소 유일의 원칙에 예외가 있는 문장들에 대한 처리 문제, 구조적인 정보의 이용에 대한 필요성, 모호성 문제 등을 해결해야 한다[나동 95].

한국어 의존 파싱에서 불필요한 의존 관계의 생성을 억제하는 방법에 대한 연구로는 [장명 96]과 [류법 96] 등이 있다. [장명 96]에서는 술어가 한국어 문장의 구성에서 뒤에 위치하면서 다른 문장 성분 요소들을 지배한다는 점을 중시하여, 술어 하위범주화 정보를 이용하여 의존 관계의 생성을 억제하는 의존 파싱 알고리즘을 제안한 바 있다. 하지만, 이러한 불필요한 의존 관계 생성을 억제하는 알고리즘의 적용에도 불구하고 여전히 의존 파싱에 의한

구문 분석에서는 많은 구문 분석 결과를 산출하여 실용적인 구문 분석기의 개발을 어렵게 한다.

본 논문에서는 의존 파싱 과정에서 생기는 불필요한 의존관계에 따른 다수의 후보 의존 트리들에 대하여 통계/의미 정보를 활용하여 최적 트리를 결정하는 구문 분석 방법을 제안한다. 이때 사용하는 통계/의미 정보는 구문구조부착 말뭉치(tree tagged corpus)를 이용하여 구축한 술어 하위범주화 정보 사전에 포함된 정보들이다.

다음 2장에서는 의존 파싱 관련 연구들에 대하여 살펴보고, 3장에서는 최적 트리 결정을 위한 통계/의미 정보를 활용한 구문 분석 방법을 제시한다. 다음 4장에서는 한국어 구문해석기 KOSA (Korean Syntactic Analyzer)의 구현과 평가 실험 결과를 보이며, 마지막으로 5장에서 결론을 맺는다.

2. 의존 파싱 관련 연구

한국어 의존 문법과 의존 파싱 관련 연구는 1980년대 말부터 현재까지 활발한 연구가 진행되고 있다. 최근에는 단순히 의존 문법에 의한 의존 파싱 알고리즘의 개발 수준을 넘어서 보다 실용적인 한국어 구문해석기의 개발에 필요한 여러 가지 의존 파싱 방법론에 대한 비교 연구가 주류를 이루고 있다. 특히 통계적 구문분석 방법의 연구에서 필요한 통계적 언어 정보의 획득과 이러한 정보를 활용한 효과적인 구문 분석 적용이 연구의 주요 관심 사항이 되고 있다.

[김형 95]에서는 한국어의 확률적 의존 문법(probabilistic dependency grammar)에 의한 의존 파싱 방법을 제안하였다. 한국어의 확률적 의존 문법의 구성은 수학적 모델에 바탕을 둔 효율적인 의존 파싱 알고리즘이지만 두 단어 사이의 의존 관계만을 고려하는 데서 연유한 구조적 분석의 한계를 가지면서 문장 수준에서의 구문 분석 정확도는 낮게 나타났다.

[Eisner 96]에서는 통계적 의존 파싱에서 두 어휘사이의 의존 관계의 확률을 고려하여 어휘 유사관계 모델(lexical affinity model), 의미 태깅 모델(sense tagging model), 생성 모델(generative model)의 세 가지 모델을 제시하고 이들 세 가지 모델에 대한 비교 실험을 실시하였다. 생성 모델이 다른 두 모델에 비하여 상당히 좋은 실험 결과가 나왔으나, 화자 주도(speaker-oriented)뿐만 아니라 청자 주도(hearer-oriented)의 확률적 구문 모델에 대한 연구가 더 필요하다.

의미정보를 이용한 구문분석 방법은 문장의 구문분석을 위하여 의미 표지(semantic marker)를 사용하는데, 각 단어마다 어휘 선호 지식으로 의미 표지를 부여하고 관련된 단어 사이의 의미적 관계를 고려하여 구문 분석을 한다. 그러나, 이 방법은 의미 표지 설정의 문제, 사람이 직접 의미 표지를 부여하는데 따른 문제로 실용적인 시스템에 사용되고 있지 않다.

3. 통계/의미 정보를 이용한 구문 분석 방법

본 논문의 구문분석 방법에서 채택한 기본적인 의존 파싱 모델은 [장명 96]에서 제안한 불필요한 의존관계의 생성을 억제하는 의존 파싱 알고리즘인 한국어 의존 문법을 가지고 지배가능경로(headable path)와 술어 하위범주화 정보를 이용한 의존 파싱 방법을 사용한다.

본 논문의 통계/의미 정보를 활용한 구문분석 방법은 앞서 설명한 [장명 96]의 기본 의존 파싱 모델에 최적 트리 결정을 위한 메커니즘을 추가하여 구성한다.

의존 파싱 결과로 생성된 다수의 후보 의존 트리들에 대하여 최적의 의존 트리를 결정하는 메커니즘은 구문구조부착 말뭉치와 명사 개념부여 작업으로부터 반자동으로 구축한 *「술어 하위범주화 정보 사전」*을 활용한다. 본 논문의 구문분석 방법에서 반자동으로(semi-automatically) 구축한 술어 하위범주화 정보 사전을 사용하는 이유는 국어 언어학 전문가의 수작업에 의해(manually) 구축되는 술어 하위범주화 사전이 구축에 드는 노력에 비하여 정보의 일

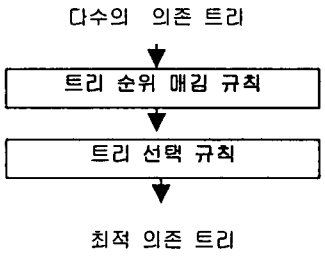
[표 1] 술어 '쓰다'의 하위범주화 정보 사전의 내용

쓰다	
Sub-entry 1 Meaning : use Probability : 61/128 = 0.47656	
Case	Semantic Class
가 을 에게	{사람, 인공물, 지역, 장소, 지적추상물} {인공물, 인공적 장소, 이성적 추상물, 활동, 수량} {인공물, 장소, 형식}
Sub-entry 2 Meaning : write Probability : 55/128 = 0.42968	
Case	Semantic Class
가 에게 에 를	{사람} {인간} {종이, 장소, 학문, 행위, 시간} {도구, 추상물}
Sub-entry 3 Meaning : make an effort Probability : 12/128 = 0.09375	
Case	Semantic Class
가 에 를	{사람} {지적 추상물, 활동} {물품, 사고내용, 이성적 활동}

관성이 결여되어 있고 술어의 여러 용법들의 빈도가 객관적이지 못하거나 나타나 있지 않은 단점이 있기 때문이다.

술어 하위범주화 정보 사전은 통계/의미 정보를 포함하고 있는데, 구문구조부착 말뭉치에서 나타나는 술어들의 상대적 출현 빈도가 확률값으로 통계적으로 구해져 포함되어 있고 술어 하위범주화 정보의 격(case) 정보와 개념 분류체계상의 분류에 기초한 명사 의미제약(semantic class) 정보는 말뭉치상에 없기 때문에 전문가에 의해 수작업으로 얻어져 등록되어 있다. [표 1]에서 보는 바와 같이 술어 '쓰다'의 하위범주화 정보 사전의 내용은 술어 '쓰다'의 세 가지 용법의 말뭉치에서의 확률값, 술어의 하위범주화 격 정보와 그 격의 명사 의미제약 정보를 포함하고 있다. 실제 입력 문장에서 술어 '쓰다'는 세 가지 경우로 분석될 수 있으므로 '정확한 술어 용법의 결정'(verb sense disambiguation)을 통한 최적의 구문분석 결과를 결정하여야 한다.

최적 트리 결정 과정은 [그림 1]과 같이 2 개의 규칙, 트리 순위 매김 규칙과 트리 선택 규칙이 차례대로 적용되어 입력 문장에 대한 정확한 술어 용법이 결정된다. 다음 절들에서 그 규칙들을 살펴본다.



[그림 1] 최적 트리 결정 과정

3.1 명사 의미제약 정보를 이용한 트리 순위 매김 규칙의 적용

트리 순위 매김 규칙은 입력 문장의 의존 파싱 결과인 다수의 의존 트리에 적용되어 이들 의존 트리들의 순위를 매기는 데 사용된다.

이 규칙은 입력 문장의 서로 다른 술어 용법들의 술어 의미제약 강도를 계산하여 비교한다. 즉, 입력 문장의 술어 p 의 의미제약 강도 $ST(p)$는 입력 문장의 술어 p 의 술어-인자에 사용된 명사들의 개념 유사도를 모두 곱하여 계산한다. 개념 유사도란 입력 문장의 술어-인자 a 번째 쌍에 나타난 명사 개념 카테고리와 술어 하위범주화 정보 사전의 해당 명사 개념 카테고리 사이의 유사도를 말한다. 술어 p 의 의미제약 강도 $ST(p)$를 구하는 식은 *(식 1)*과 같다.

(식 1)

$$ST(p) = \prod_{a \in subcat(p)} SIM(i,t)$$

단, $subcat(p)$: 술어 p 의 술어-인자 쌍
i : 입력 문장의 a 번째 술어-인자에 나타난 명사의 의미코드
t : 술어 하위범주화 사전의 a 번째 술어-인자에 사용된 명사 의미코드

*(식 1)*에서 사용한 두 개념사이의 유사도 계산식

*SIM(i,t)*은 [류법 97]에서 채택한 *유사도 측정법* [1]을 사용하고 있다. 입력 문장의 술어 의미제약 강도는 입력 문장의 술어-인자 쌍의 수가 적고 개념 유사도가 높은 경우에 큰 값을 가진다.

트리 순위 매김 규칙의 적용 예를 다음 입력 문장을 예로 들어 설명한다.

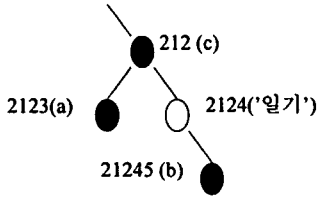
(예문) "*철수가 일기를 쓴다*"
1111(인간) 2124(지적 추상물)

이 문장은 본 연구실에서 개발 중인 한국어 구문해석기 KOSA의 중간 의존 파싱 결과로 의존 트리가 출력된다.

```
(a)
<PRED> 쓰 {TYPE(pv)SUBCAT(jcs,jco,jca1)SUBCAT_SEM({1111,114,
          121,123,2124},{1111,114,123,2123,221,2223,2345},{114,123,
          2334}) PROB(0.47656) MOOD(ef2)}
       <COMN> 철수 {CASE(jcs)TYPE(nq)SEM(1111)}
       <COMN> 일기 {CASE(jco)TYPE(nc)SEM(2124)}
(b)
<PRED> 쓰 {TYPE(pv)SUBCAT(jcs,jco,jca4)SUBCAT_SEM({1111},
          {1144, 21245},{1111})PROB(0.42968)MOOD(ef2)}
       <COMN> 철수 {CASE(jcs)TYPE(nq)SEM(1111)}
       <COMN> 일기 {CASE(jco)TYPE(nc)SEM(2124)}
(c)
<PRED> 쓰 {TYPE(pv)SUBCAT(jcs,jco,jca3)SUBCAT_SEM({1111},
          {1146,212,2213},{2124,221})PROB(0.09375)MOOD(ef2)}
       <COMN> 철수 {CASE(jcs)TYPE(nq)SEM(1111)}
       <COMN> 일기 {CASE(jco)TYPE(nc)SEM(2124)}
```

위 분석 결과는 [표 1]의 술어 하위범주화 정보 사전에 구축된 술어 '쓰다'의 세 가지 용법으로 분석된 것이다. 각 술어 용법의 목적격 명사의 의미제약 정보는 (a) 전기, 비누, 참기름 따위의 명사들을 명사로 취하는 경우와 (b) 소설, 일기, 책, 글 따위의 명사들을 명사로 취하는 경우, 그리고 (c) 힘, 권력 따위를 명사로 취하는 경우의 서로 다른 종류의 술어 용법으로 분석된다.

이들 세 가지 경우에서 술어 의미제약 강도 계산에 의한 트리 순위 매김을 살펴본다. 먼저, 입력 문장의 중간 분석 결과 ***(a), (b), (c)***에 나타난 술어 ***Pa, Pb, Pc***의 술어 의미제약 강도 ***ST(Pa), ST(Pb), ST(Pc)***를 구하기 위해 입력 문장의 술어-인자 쌍에 나타난 명사 *'철수'(1111)*와 *'일기'(2124)*의 개념 유사도를 각각 계산한다.



[그림 2] 명사 "일기"의 개념 유사도 측정

[그림 2]는 명사 '일기'의 개념 유사도 측정에 관련된 그림이다. 이때 개념 유사도는 공통 조상 중 가장 깊이가 깊은 개념, 형제 개념, 조상 개념의 순으로 높게 나타난다. 입력 문장의 술어 용법에 대한 술어 의미제약 강도는 다음과 같이 계산되어 진다.

$$ST(Pa) = SIM(\{1111\},(\{1111,114,121,123,2124\}) * SIM(\{2124\},\{1111,114,123,2123,221,2223,2345\}) = 1 * 0.444$$

$$ST(Pb) = SIM(\{1111\},\{1111\}) * SIM(\{2124\},\{1144,21245\}) = 1 * 0.909$$

$$ST(Pc) = SIM(\{1111\},(\{1111\}) = * SIM(\{2124\},\{1146,212,2213\}) = 1 * 0.4$$

따라서, 의존 트리의 순위 매김은 술어 의미제약 강도에 따라 ***(b) 0.909, (a) 0.444, (c) 0.4***의 순서로 매겨진다. ***(b)***의 경우에 입력 문장의 목적격 명사 '일기'의 의미코드와 술어 하위범주화 정보 사전의 목적격 명사 의미제약 정보의 개념 유사도가 높아 술어 의미제약 강도가 가장 높게 나타난 것이다.

---

[1] 본 유사도 측정법은 두 개념사이의 유사도 계산을 '공통조상' MSCA(Most Specific Common Abstraction)와 '자손 가중치' Is-a penalty 의 두 요소에 의한 다음 (식)으로 계산된다.

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

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

3.2 확률값에 의한 트리 선택 규칙

트리 선택 규칙에 의한 최적 트리 결정은 [류법 97]에서 구문구조부착 말뭉치로부터 얻은 술어 용법들의 상대적 출현 빈도에 의한 확률값을 활용한다.

이 확률값은 말뭉치에 있는 전체 술어 집합 n에서 술어 용법 $pred_i$가 그 술어의 하위범주화 정보 $Sub_{i,k}$ $(1 \le k \le m)$로 사용될 확률 $prob(Sub_{i,k}/pred_i)$로 (식 2)에 의해 구해진다.

(식 2)

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

단, $1 \le i \le n$

트리 선택 규칙에서 이 확률값의 적용은 앞서 설명한 트리 순위 매김 규칙에서 술어 의미제약 강도가 서로 같은 경우에 한해서 최적 트리 선택 결정에 이용되도록 한다. 이것은 술어 의미제약 강도가 서로 다른 경우에는 서로 다른 술어 용법의 적용으로 인한 트리 순위 매김에 의해 최적 트리의 결정이 가능하기 때문에 상대적으로 술어 용법의 빈도에 의한 통계적 확률값을 무리하게 적용하지 않아야 하기 때문이다.

실제로 이 규칙이 적용되는 경우는 많지 않을 것으로 예측된다. 왜냐하면, 본 연구에서 사용하는 술어 하위범주화 정보 사전의 개념분류체계상의 명사 의미코드는 일반화(generalization) 과정을 거쳐 만들어 졌기 때문에 반대로 입력 문장의 술어 하위범주화 정보 사전의 명사 의미코드로의 직접적인 매핑이 이루어지지 않아 유사도 값은 대부분 1보다 적은 서로 다른 값들을 가질 가능성이 높기 때문이다.

앞의 예문에서 술어 '쓰다'의 세 가지 용법의 말뭉치내에서의 확률값이 ***(a) 0.47656, (b) 0.429680, (c) 0.09375*** 로 나타났으나 트리 순위 매김 규칙에 의한 술어 의미제약 강도가 모두 같지 않아 이미 트리 순위가 매겨졌기 때문에 이 경우 확률값에 의한 트리 선택 규칙은 적용되지 않는다.

4. 한국어 구문해석기 구현 및 평가 실험

4.1 시스템 개요

한국어 구문해석기 KOSA는 크게 세 부분으로 구성되는데, 1) 구문구조부착 말뭉치로부터 통계/의미 정보를 구축하는 부분(통계정보 베이스 및 술어 하위범주화 정보 사전 구축부)과 2) 형태소 해석기의 수행 결과를 입력으로 받아 전처리 과정을 거친 후, 의존 파싱을 수행하는 부분(의존 파싱부), 다음으로 3) 불일치 의존 관계의 제거를 통한 의존 제약을 적용하여 최적의 구문 분석 결과를 산출하는 부분(의존 제약부)이다[장명 96]. 의존 제약부는 의존 파싱의 결과인 후보 의존 트리들 중에서 의존 제약 적용 루틴과 최적 트리 결정 루틴으로 구성되는데, 본 논문은 최적 트리 결정 루틴의 구현 부분으로 트리 순위 매김 규칙과 트리 선택 규칙을 차례로 적용하여 최적의 구문 분석 결과를 생성한다.

4.2 실험

4.2.1 구문해석기 평가용 문장

한국어 구문해석기 KOSA의 성능 평가를 위한 실험은 ***[SERI Test Suites '97]***[성원 97]에 기술된 한국어 구문해석기 평가용 문장을 사용한다. SERI Test Suites '97은 술어 구문 평가문 180문과 문법 현상 평가문 292문의 총 472문장으로 구성되는데, 본 실험에서는 ***술어 구문 문형 평가문 180 문장*** 만을 사용하여 실험한다. 이것은 본 연구의 주된 연구 초점인 구문 분석에서 술어 하위범주화 정보 중심의 구문해석기 성능 평가를 위하여 구문구조부착 말뭉치에서 상위 고빈도로 나타난 20개 술어로 술어 구문 문형 평가문을 구성하였기 때문이다.

4.2.2 구문분석 실험

구문분석 실험은 SERI Test Suites '97의 술어 구문 문형

을 가지고 음 가지

첫번째는 본 논문의 통계/의미 정보를 이용하지 않는 [장명 96]의 의존 파싱 방법에 의한 '한국어 구문해석기 KOSA 1.0'의 평가 실험 ***[실험 1]***이고 두번째는 본 논문의 최적 트리 결정 루틴을 이용한 의존 파싱 방법에 의한 '한국어 구문해석기 KOSA 2.0'의 평가 실험 ***[실험 2]***이다.

술어 구문 문형 평가문에 대한 한국어 구문해석기 KOSA 1.0 과 KOSA 2.0 의 평가 실험은 SERI Test Suites '97 의 「평가표 작성 및 평가 요령」에 따른다. 즉, 평가문의 구문분석 수행 결과를 분석한 후 그 구문 트리가 '올바른 구문 트리'(wft; well-formed tree)와 '잘못된 구문 트리'(ift; ill-formed tree)인 경우를 계산하고, 평가 기준에 따라 1 에서 5 까지의 점수를 매겨 구문 분석의 정확도 평가를 실시하는 것이다.

[표 2] 한국어 구문해석기 KOSA 1.0 과 2.0 의 SERI Test Suites '97 의 술어 구문 문형에 대한 구문분석 실험 평가 결과

평가 등급 / 구문분석 결과의 평가 기준		각 항에 속하는 문장의 수	
		[실험1]	*[실험2]*
A	모든 wft만 포함	42	97
B	모든 wft뿐만 아니라 ift 도 포함	88	35
C	일부의 wft만 포함	13	18
D	일부의 wft와 ift를 포함	5	9
E	모두가 ift인 경우	6	6
정확도 δ = (A*5+B*3+C*3+D*3+E*0)/(165*5)		*0.701*	*0.813*

[실험 1]과 ***[실험 2]***의 평가 실험 결과는 [표 2]에 나타나 있다. [표 2]의 ***[실험 2]*** 평가 실험 결과를 살펴보면, 평가 기준 A 항에 속하는 문장의 수가 97 문이고, B 항, C 항, D 항에 속하는 문장의 수는 각각 35 문, 18 문, 9 문이며, E 항에 속하는 문장의 수는 6 문장으로 나타났다. 따라서, 술어 구문 문형 평가문 총 180 문에 대한 한국어 구문해석기 KOSA 의 정확도 δ 는 (5*97 + 3*35 + 3*18 + 3*9 + E*6) / (165*5) = 0.813 로 계산되었다. 물론 여기에는 전자사전에 등록된 형태, 구문 정보의 미비로 인한 형태소 분석 결과 오류의 구문분석 실패 문장 15 개는 SERI Test Suites '97 의 평가 기준에 따라 구문해석기의 성능 평가 실험에서 제외되었다.

4.2.3 실험 결과 분석

SERI Test Suites '97 의 술어 구문 문형 평가문에 대한 한국어 구문 해석기 KOSA 1.0 ***[실험 1]*** 과 KOSA 2.0 ***[실험 2]*** 의 구문 분석 정확도 δ 는 ***0.701*** 과 ***0.813*** 으로 나타났다. 이 평가 결과는 술어 하위범주화 구문 관계 정보만을 이용하여 구문분석한 경우 - 불필요한 의존 관계의 생성에 따른 다수의 wft, ift 를 생성함 - 보다는 통계/의미 정보를 이용한 구문분석의 경우가 11.2 % 정도 향상된 구문분석 정확도를 얻은 것이다. 이 수치 11.2 %는 ***[실험 1]*** 의 평가 기준 B 의 많은 문장들이 ***[실험 2]*** 에서 평가 기준 A 로의 분석 결과를 얻었기 때문이고, 그 밖에 ***[실험 1]*** 에서 평가 기준 A 와 B 를 받은 일부 문장들이 ***[실험 2]*** 에서는 C, D, E 의 평가 기준의 문장으로 잘못 분석 결과가 나오기도 하였으나, ***[실험 1]*** 의 평가 기준 D 항의 일부 문장이 ***[실험 2]*** 의 C 항으로 좋게 분석되었도 하였다.

본 실험에서 사용한 SERI Test Suites '97 의 평가 방법에서 평가 기준과 정확도 계산 방법은 이번 실험 결과에서 보듯이 실제 구문해석기의 비교 평가에 큰 변별력을 제공하였다고 생각된다. 즉, ***[실험 1]*** 과 ***[실험 2]*** 의 평가 결과에서 보면, 모든 wft 가 포함된 경우를 평가 기준으로 하는 평가 방법의 경우는 - 여기서는 평가 기준 A 와 B 를 하나의 평가 기준으로 하는 경우 - 두 구문해석기의 구문분석 정확도가 비슷하게 나와 이 경우 평가 방법으로 문제가 있으나, 본 평가 방법에서는 평가 기준 A 와 B 가 분리, 설정되어 있기 때문에 변별력 있는 구문해석기의 평가 기준으로 작용될 수 있음을 확인할 수 있다.

또한, 본 실험에서도 역시 평가문에 기술된 평가문의 평가 기준과 실제 문장을 분석하는 시스템에 구축된 언어

지식의 불일치로 인한 낮은 평가 결과를 얻는 경험을 하였다. 즉 실제 술어 하위범주화 정보 사전에 구축된 술어 하위범주화 정보와 술어 구문 평가문의 주석에 평가 기준으로 기술된 술어 하위범주화 정보의 상이함으로 인한 구문분석 정확도의 하향 원인은 구문해석기의 평가 방법에 있어서의 보다 근본적인 문제로 판단된다.

마지막으로 술어 하위범주화 정보 사전의 명사 의미제약 정보와 입력 문장에 나타난 명사 의미코드의 매칭의 부정확성으로 인한 잘못 구문 분석된 경우가 있었는데, 개념분류체계에 의한 명사 의미 코드 부여와 의미코드 일반화(혹은 그룹화) 문제와 관련하여 모든 개별 어휘에 대하여 의미코드를 직접 부여하는 것이 더 바람직한 지에 대한 보다 심도있는 연구[Kurohashi 94]가 더 필요하다고 생각된다.

5. 결 론

본 연구에서는 구문구조부착 말뭉치를 이용하여 구축한 술어 하위범주화 정보 사전의 통계/의미 정보를 활용한 구문 분석 방법을 제안하였다. 이러한 구문 분석 방법은 의존 파싱에서 발생하는 다수의 의존 트리들에 대한 정확한 트리 선택 결정을 통하여 구문해석기의 정확도를 높일 수 있었다.

앞으로 대규모의 구문구조부착 말뭉치의 확충과 이를 이용한 대량의 술어 하위범주화 정보 사전의 정교한 구축 등이 필요하다. 이를 통하여 신문이나 web 에서 사용되는 술어를 포함한 다양한 한국어의 실제 문장들에 대한 구문분석 실험을 통해서 구문해석기의 안정성과 성능을 높이는 작업을 계속할 계획이다.

참고 문헌

[Caroll 92] G. Caroll, E. Charniak, "Two Experiments on Learning Probabilistic Dependency Grammars from Corpora", In Workshop Notes, Statistically-Based NLP Techniques, AAAI, p1-13, 1992.

[Charniak 93] Eugene Charniak, Statistical Language Learning", Cambridge, MA., MIT Press, 1993.

[Covington 88] T. A. Covington, "Parsing Variable Word Order Languages with Unification-Based Dependency Grammar", ACMC Research Report 01-0022, Univ. Of Georgia, 1988.

[Eisner 96] Jason Eisner, "Three New Probabilistic Models for Dependency Parsing: An Exploration", In the Proceedings of COLING-96, p340-345, Copenhagen, 1996.

[Magerman 91] D. Magerman and M. Marcus, "Pearl: A Probabilistic Chart Parser", In the Proceedings of European ACL, Berlin, 1991.

[Kurohashi 94] Sadao Kurohashi and Makoto Nagao, "A Method of Case Structure Analysis for Japanese Sentences Based on Examples in Case Frame Dictionary", IEICE Trans. INF. & SYST., Vol. E77-D, No. 2, p227-239, February 1994.

[김형 95] 김형근, "확률적 의존문법과 한국어 구문분석", 한국과학기술원 전산학과 석사학위논문, 1995.

[나동 95] 나동렬, "한국어 파싱에 대한 고찰", 정보과학회지, 12 권 8 호, p33-46, 1995.

[류법 96] 류법모 외 2 인, "한국어 파서에서의 지역 의존 관계의 이용", 제 8 회 한글 및 한국어 정보처리 학술대회 발표 논문집, p464-468, 1996.

[류법 97] 류법모 외 4 인, "구문구조부착 말뭉치를 이용한 술어의 하위범주화 정보 구축", 제 9 회 한글 및 한국어 정보처리 학술대회 발표 논문집, 1997.

[서정 96] 서정연 외 1 인, "통계적 방법을 이용한 구문분석", 정보과학회지 14 권 7 호, p58-70, 1996 정보처리 학술대회 발표 논문지, 1997.10.

[성원 97] 성원경 외 5 인, "SERI Test Suite '97: 한국어 구문해석기 성능 평가용 문장 모음", 제 9 회 한글 및 한국어 정보처리 학술대회 발표 논문집, 1997.

[장명 96] 장명길 외 4 인, "술어 하위범주화 정보를 이용한 한국어 의존 파서", 제 8 회 한글 및 한국어 정보처리 학술대회 발표 논문집, p452-463, 1996.