구문분석기의 어휘확장

김민찬, 김곤, 배재학
울산대학교 컴퓨터・정보통신공학부
e-mail:{tomatuli, gonkim, jhjbae}@ulsan.ac.kr

Lexical Expansion of Sentence Parsers

Min-Chan Kim, Gon Kim, Jae-Hak J. Bae
School of Computer Engineering and Information Technology,
University of Ulsan

요 약

본 논문에서는 구문분석기의 어휘확장을 통해 구문분석의 성공률을 높이고자 하였다. 구문분석은 문장 내 구성성분들이 가지는 통사적인 관련성을 파악하는 작업이다. 구문분석 실패의 가장 빈번한 원인 중의 하나는 미등록 어휘의 출현이다. 결여된 어휘문제를 해결하는 것은 구문분석의 성공률을 높이고, 원문이해 시스템을 보다 더 견고하게 하는데 관건으로 작용한다. 이를 위하여, 본 논문에서는 구분분석기 LGPI+의 어휘 사전에 존재하지 않는 단어들을 또 다른 어휘자원인 WordNet을 이용하여 해결하고자 하였다. 구체적으로는, (1) 미등록 어휘를 WordNet에서 찾고, (2) 그 유의어 정보를 파악하여, (3) LGPI+ 어휘사전에 추가한다. 실험을 통하여 구문분석의 실패를 해결하고, 정확도와 성공률을 높일 수 있음을 확인하였다.

1. 서론

자연어처리 분야에서는 어휘자체에 대한 의미분석에서부터 구문분석, 의미해석으로 이어지는 단어, 문장, 글의 의미분석에 대한 연구가 계속되고 있다. 문장의 통사구조분석과 의미분석은 완전히 독립적인 단계에서 이루어 질 수 없다. 즉, 문장의 구성성분들이 가지는 문법적인 기능과 내포하는 의미는 서로 의존적인 관계에 놓여 있다. 따라서, 구문분석의 정확도와 성공률 향상은 의미분석의 중의성을 해소하고, 원문이해 시스템을 보다 더 견고하게 할 수 있다.

구문분석은 문장의 구성성분들이 가지는 통사적인 관련성을 파악하는 작업이다[1]. 이러한 구문분석의 오류를 야기하는 주된 원인으로는 알려지지 않은 어휘가 문장에서 나타나는 경우이다. 미등록 어휘문제를 해결하기 위한 연구로는, (1) LDOCE(Longman Dictionary of Contemporary English)를 활용하여, 계층적 구문 유형의 규칙들로부터 어휘의 의미정보를 파악하는 방법[2], (2) 특화된 전문용어의 선택적인 어휘 의미들 중에서 선별하는 방법[3] 등이 있다.

기존의 연구방법들은 미리 구축된 어휘사전에서 규칙 및 의미정보 분류를 활용한 방법이다. 본 논문에서는 이와는 달리, 문장에서 나타난 어휘를 찾을 수 없을 때, 또 다른 어휘자원을 활용하여 해결하는 방법을 모색해 보았다.

2. 어휘확장 알고리즘

구문분석기는 분석에 필요한 어휘 사전을 가지고 있다. 포함된 어휘의 수는 제한적이며, 정확한 구문분석을 위해서는 지속적인 어휘확장이 필요하다. 본

논문에서는 구문분석기의 어휘사전에 포함되지 않은 단어들을 다른 어휘자원을 활용하여 해결하고자 하였다.

활용한 구문분석기는 LGPI(Link Grammar Parser Interface[4])를 확장한 LGPI+이다. LGPI는 Link Grammar Parser[5]에 대한 SWI-Prolog API를 제공한다. LGP는 6만 어형을 수록한 사전을 내장하고 있으며 다양한 구문구조를 처리할 수 있다.

어휘 정보를 획득하기 위한 자원으로는 WordNet[6]을 이용하였다. WordNet은 인간의 어휘지식에 대한 심리언어학 연구의 성과를 토대로 1985년부터 프린스톤대학 인지과학연구실이 구축해온 영어어휘 데이터베이스이다. WordNet에서 단어의 의미는 synsets이라고 하는 유의어 집합으로 표현된다.

표 1. WordNet의 데이터베이스 통계

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	114648	79689	141690
Verb	11306	13508	24632
Adjective	21436	18563	31015
Adverb	4669	3664	5808
Totals	152059	115424	203145

LGPI+를 이용한 구문분석 시, 사전에 포함되어지지 않는 어휘가 나타났을 경우, 또 다른 어휘정보를 이용한다. 이때, 사용되어진 알고리즘은 다음과 같다.

(1) LGPI+를 이용한 구문분석 시 사전에 포함되지 않은 어휘를 LGPI+의 품사정보를 이용하여 WordNet에서 검색한다.

(2) Synset_cnt로 그 어휘의 유의어 목록을 확인한다.

(3) WordNet에서 찾은 유의어가 LGPI+의 어휘사전에 있을 경우 동일한 위치에 어휘를 등록한다. .

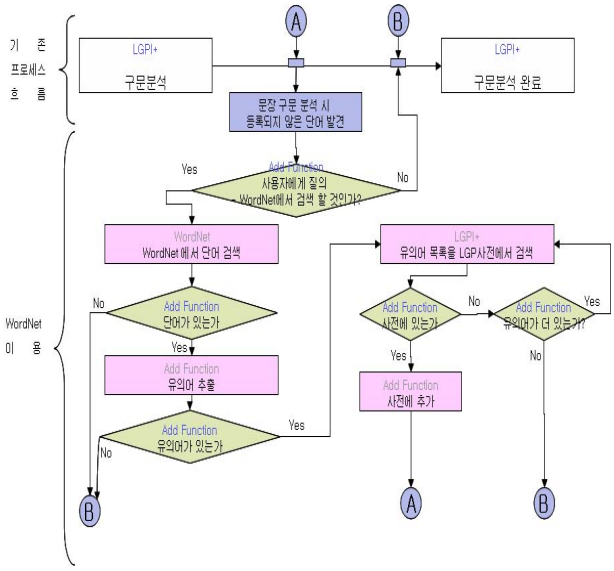

그림 1. 구문분석기의 어휘확장

3. 구문분석기의 어휘확장

실험을 위하여 LGPI+의 어휘사전에 포함되지 않은 단어를 이용하여 예문을 만들고, 구문분석 시 오류가 발생하였을 때, WordNet[6]을 이용하여 해결해 보았다.

실험에 사용한 문장은 다음과 같다. 문장에서 homemade는 LGPI+의 어휘 사전에는 포함되어 있지 않다.

"One lady often brings homemade cookies, cakes, candies and breads to the office."

그림 2는 어휘확장 알고리즘을 적용하지 않았을 때의 구문분석결과로 LGPI+의 출력 결과 중에서 첫 번째 Linkage정보를 출력한 것이다.

```
linkparser> ++++Time                                              0.56
seconds (0.56 total)
Found 28 linkages (28 had no P.P. violations)
  Linkage 1, cost vector = (UNUSED=0 DIS=0 AND=0 LEN=39)

    +-------------------------------------------------------Xp----------
    |                          +-----------------------------------MVp-----
    +-----Wd-----+------Ss-----+----------Op----------+
    |      +--Ds-+     +---E---+          +-----A-----+
    |      |     |     |       |          |           |
LEFT-WALL one lady.n often brings.v homemade[?].a cookies.n , cakes.n ,
-------------------------------------+
--------------------------+          |
                          +----Js---+  |
                          |  +--Ds--+  |
                          |  |      |  |
candies.n and breads[!].n to the office.n .
```

그림 2. 구문분석정보(어휘확장 이전)

그림 3은 그림 2에서 정확한 어휘정보를 찾지 못

한 homemade를 WordNet에서 검색한 것이다. 이때, LGPI+에서 찾은 동일한 품사정보로 검색한다.

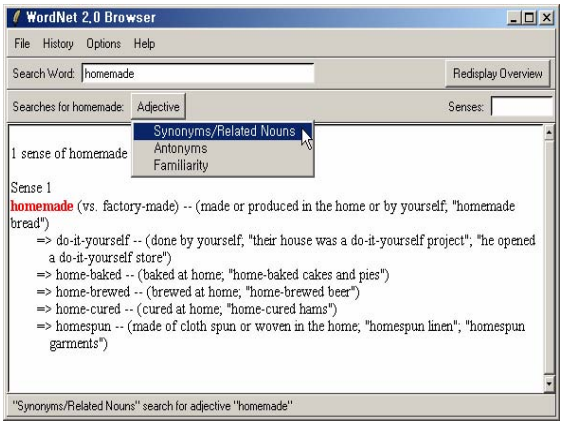

그림 3. WordNet 검색결과: homemade

```
$$1
$$1=>homemade (vs. factory-made)
        do-it-yourself
        home-baked
        home-brewed
        home-cured
        homespun
```

그림 4. 추출한 어휘 목록

추출한 어휘 목록 중에서 LGPI+의 어휘 사전에 등록되어 있는 단어를 선별한다. 그림 5는 추출한 어휘 목록들이 LGPI+ 어휘사전에 등록되어 있는지 확인한 결과이다. 이때, 그 어휘가 포함된 사전의 파일명과 위치정보를 보여준다.

```
linkparser> !!do-it-yourself
    "do-it-yourself" matches nothing in the dictionary.
linkparser> !!home-baked
    "home-baked" matches nothing in the dictionary.
linkparser> !!home-brewed
    "home-brewed" matches nothing in the dictionary.
linkparser> !!home-cured
    "home-cured" matches nothing in the dictionary.
linkparser> !!homespun
Matches:
        homespun.a          165  <words/words.adj.1>
linkparser>
```

단어의 위치 *사전파일명*

그림 5. LGPI+에서 어휘 검색 결과

그림 5에서 확인한 바, WordNet에서 추출한 어휘 목록 중에서 LGPI+ 어휘사전에 등록되어 있는 단어는 homespun이며 이를 포함하고 있는 파일명은 “words/word.adj.1”임을 알 수 있다. 따라서 homemade는 homespun과 같은 위치에 놓으면 된다.

그림 6은 homemade를 LGPI+ 어휘사전에 추가한 결과이다.

```
histrionic.a hoarse.a hoary.a hobnailed.a hollow.a holy.a homeless.a
homely.a homeopathic.a homesick.a homemade.a homespun.a homeward.a homey.a
homicidal.a homogeneous.a homophobic.a homosexual.a honest.a
honorable.a honorary.a honourable.a hooked.a hopeless.a horizontal.a
horned.a horny.a horrible.a horrid.a horrific.a horticultural.a
hospitable.a hostile.a hot.a hourly.a housebound.a houseproud.a
```

그림 6. LGPI+ 어휘사전에 추가

그림 7은 homemade가 LGPI+ 어휘사전에 추가된 후, 예문에 대한 구문분석 결과이다.

```
linkparser> ++++Time                                          0.42 seconds
(0.42 total)
Found 19 linkages (19 had no P.P. violations)
  Linkage 1, cost vector = (UNUSED=0 DIS=0 AND=0 LEN=39)

   +-------------------------------------------------------Xp----------
   |                          +-------------------------------------MVp---
   +-----Wd-----+------Ss-----+---------Op--------+
   |      +--Ds-+     +---E---+         +----A----+
   |      |     |     |       |         |         |
LEFT-WALL one lady.n often brings.v homemade.a cookies.n , cakes.n ,
-----------------------------------------+
--------------------------+              |
                          +----Js---+    |
                          |  +--Ds--+    |
                          |  |      |    |
candies.n and breads[!].n to the office.n .
```

그림 7. 구문분석 결과: 어휘확장 이후

본 논문에서 기술한 어휘확장 알고리즘을 문장 분석용 통합 사용자 인터페이스 ISAAC[7]에 적용하였다.

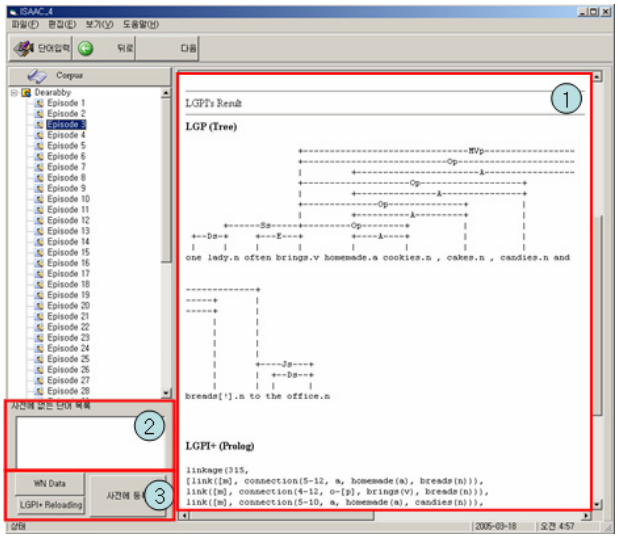

그림 8. ISAAC의 구문분석 화면

그림 8의 ①은 구문분석결과 화면이다. ②는 구문

분석 시 사전에 등록되어 있지 않은 단어의 목록을 나타낸다. ③은 ②에서 나타난 어휘들에 대하여, WordNet 유의어를 검색하고, 찾은 유의어들을 LGPI+ 어휘사전에서 확인한 후, 사전에 등록하는 기능이다.

4. 결론

LGPI+의 어휘사전은 각 어휘별 품사 및 특성에 따라 분류된 파일로 구성되어 있다. 따라서, 어휘확장 시 정확한 파일과 위치에 기록되어야 구문분석의 성공률과 정확도가 높아진다.

그림 2의 homemade가 사전에 없을 경우에는 28개의 linkages를 보이고, 그림 7의 사전에 추가한 경우에는 19개의 linkages를 보여준다. 이를 통하여, 어휘정보를 추가하였을 경우 linkages의 수가 감소함을 알 수 있다.

본 논문에서는 구문분석 시 미등록된 어휘를 확장하는 방법으로 WordNet의 유의어 정보를 이용하였다. 실험을 통해 구문분석기의 성공률과 정확도를 높일 수 있음을 알 수 있었다. 정확한 구문분석은 언어번역, 정보검색, 문법분석, 음성처리, 원문처리 및 원문이해 등 자연언어처리 분야의 전반적인 연구분야뿐만 아니라 전문용어 및 다양한 전문가 시스템이나 학제 간 연구의 중요한 기초자료가 될 수 있을 것으로 기대된다.

<Acknowledgements>

본 연구는 한국과학재단 목적기초연구 R05-2004-000-12362-0 지원으로 수행되었음. 또한 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 육성·지원사업의 연구결과로 수행되었음.

[참고문헌]

[1] 김곤, 배재학. “문서요약을 위한 문장추상화.” 한국정보처리학회 춘계 학술대회 논문집, 제 9권, 제 1호, pp.531-534, 2002.

[2] Hiyan Alshawi. "Processing Dictionary Definitions with Phrasal Pattern Hierarchies." Computational Linguistics, Volume 13, Numbers 3-4, 1987.

[3] Walker, D. and Amsler, R. "The use of Machine Readable Dictionaries in Sublanguage Analysis." In Analyzing Language in Restricted Domains, Lawrence Erlbaum Associates, 1986.

[4] SWI-Prolog. http://www.swi-prolog.org/.

[5] Sleator, D. and Temperley, D. "Parsing English with a Link Grammar." Third International Workshop on Parsing Technologies, August 1993.

[6] WordNet. http://wordnet.princeton.edu/.

[7] 김곤, 김민찬, 배재학. 이종혁, "ISAAC: 문장분석용 통합시스템 및 사용자 인터페이스", 한국정보처리학회 논문지B , 제11-B권, 제 1호, pp. 107-116, 2004.