어휘확장을 통한 문장분석 시스템의 개선

김민찬° 김곤 배재학
울산대학교 컴퓨터·정보통신공학부
{tomatuli°, gonkim, jhjbae}@ulsan.ac.kr

Improvement of a Sentence Analysis System through Lexical Expansion

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

요 약

본 논문에서는 미등록 어휘로 인한 구문분석의 실패를 해결하는 방법으로 WordNet의 유의어 정보를 이용하였다. 이 방법을 또한 설화용 온톨러지 OfN의 어휘확장에 적용하였다. 실험을 통하여 구문분석 과정에서 나타나는 미등록 어휘문제의 해결과 문장의 의미분석 과정이 순조롭게 진행될 수 있음을 확인하였다.

1. 서 론

구문분석은 문장의 구성성분들이 가지는 통사적인 관련성을 파악하는 작업이다[1]. 구문분석 중에 시스템에 등록되어 있지 않은 어휘가 출현하면 구문분석 오류가 발생한다.

본 논문에서는 문장분석 과정에서 시스템의 내장사전에 등록되어 있지 않은 어휘가 나타난다 할지라도, 다른 어휘자원인 WordNet[2]을 활용하여 이를 해결하고자 하였다[3]. 구체적으로는 구문분석기 LGPI+[4]의 통사 및 의미분석 과정에 이 방법을 활용하였다. 미등록 어휘문제를 해결함으로써 구문분석의 성공률을 높이고, 문장분석용 통합 사용자 시스템 ISAAC(An Interface for Sentence Analysis & Abstraction with Cogitation)[4]을 개선할 수 있었다.

2. 어휘확장

2.1 WordNet

어휘정보를 보완하기 위해 사용한 자원은 WordNet[]이다. WordNet은 인간의 어휘지식에 대한 심리언어학 연구성과를 토대로 1985년부터 프린스톤대학 인지과학 연구실이 구축해온 영어어휘 데이터베이스이다. 그림 1은 WordNet의 어휘 구성표이다[2].

POS	Unique Strings	Synsets
Noun	114648	79689
Verb	11306	13508
Adjective	21436	18563
Adverb	4669	3664
Totals	152059	115424

표 1 WordNet의 어휘

유의어 검색을 위해 동일한 의미정보를 가지는 유의어 집합(SYNSET)에서 찾은 어휘정보(그림 1)를 이용하였다. 그리고 유의어 집합들 사이의 관련성을 나타내는 형용사 관계정보(그림 2)도 포함하였다.

SN	SYNSET_ID	W_NUM	WORD	SS_	SENSE	TAG_
1	100001740	1	entity	n	1	11
2	100002056	1	thing	n	12	0
3	100002342	1	anything	n	1	0
4	100002452	1	something	n	1	0
5	100002560	1	nothing	n	2	0
6	100002560	2	nonentity	n	3	0
7	100002645	1	whole	n	2	0
8	100002645	2	whole_thing	n	1	0
9	100002645	3	unit	n	6	0

그림 1 단어의 의미 정보(Synset)를 포함: WN_S

SN	SYNSET_ID_1	SYNSET_ID_2
67	300010495	300011189
68	300010495	300011443
69	300010495	300011592
70	300010495	300011858
71	300010495	300011950
72	300010902	300010495

그림 2 유사한 의미를 가지는 형용사 관계: WN_SIM

2-2 LGPI+(구문분석기)의 어휘 확장

구문분석기로 LGPI(Link Grammar Parser Interface[5])를 확장한 LGPI+를 이용하였다. LGPI+는 6만 어형을 수록한 사전을 내장하고 있으며 다양한 구문구조를 처리할 수 있다. 이 사전은 필요에 따라 확장이 가능하다. LGPI+를 이용한 구문분석 시, 내장사전에 포함되어지지 않는 어휘가 나타났을 경우, WordNet의 정보를 이용한다. 사용한 알고리즘은 다음과 같다.

```
Tx: 어휘사전에 포함되지 않은 단어
Txp: Tx의 품사정보
Txs: Tx의 SYNSET_ID
WNs: 단어에 대한 의미정보(SYNSET 정보)의 집합
WNsim: 유사한 SYNSET에 속하는 형용사의 관계집합
Ls:WNs에서 검색한 어휘 리스트
Ssim:WNsim에서 검색한 SYNSET 정보
Lr:WNsim에서 검색한 어휘 리스트
Ts: 유의어 목록(Ls)에서 선택되어진 단어
Ld: LGPI+의 어휘사전

입력 :
LGPI+ 구문분석시의 Tx와 Txp
출력 :
문장분석가에 의해 선택되어진 Ts
Tx를 Ts의 위치에 Ld에 추가
Main :
TLs ⇐ null
Search_Wns(Tx, Txp)
If Txp = 'adjective' Then
        Search_Wnsim(Txs)
        Search_Wns(Lr)
Ls에서 Ts 선택
Ld에 Tx 등록(Ts의 위치)

WordNet 검색 :
function: Search_Wns(X)
      Loop(WNs)
      Ls ⇐ search(X)
      End

function: Search_Wnsim(SYNSET_ID)
      Loop(WNsim)
      LRws ⇐ search(SYNSET_ID)
```

그림 3 LGPI+의 어휘확장

2-3. 설화용 온톨로지(OfN) 어휘확장

OfN은 문장에서 중요정보를 분별하고 이야기를 이해하기 위한 온톨러지로 Roget 시소러스를 심중사전(Lexicon)으로 삼아 이를 재구성하여 얻은 것이다. 이 온톨로지는 설화문장을 추상화시키는데 사용할 목적으로 구축되어졌다[6].

표제정보	OfN 범주
act	event
being	state
dimension	space
future	time
affection(general)	affect_state
word(sentient, power(moral))	affect_state

표 2 OfN의 일부

OfN확장은 LGPI+의 어휘사전 확장과 동일한 방법으로 이루어 진다. 따라서 변경되는 사항만을 기술하면 다음과 같다.

```
Tx: OfN에 포함되지 않은 단어
Tofn: Tx의 OfN 정보(색인정보, 범주정보)
Ld: OfN의 범주정보 DB

출력 :
문장분석가에 의해 선택되어진 Ts
Search_OfN(Ts)
Ld에 Tx와 Tofn 정보 추가

OfN 검색 :
function: Search_OfN(Ts)
      Loop(Ld)
      Tofn ⇐ search(Ts)
      End
```

그림 4 OfN의 어휘확장

3. ISAAC의 개선

본 논문에서 제시하고 있는 구문분석기와 OfN의 어휘확장방법을 문장분석용 통합 사용자 인터페이스 ISAAC[4]에 적용하였다. 아래 예문에서 'homemade'는 LGPI+와 OfN에 등록되어있지 않은 어휘이다.

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

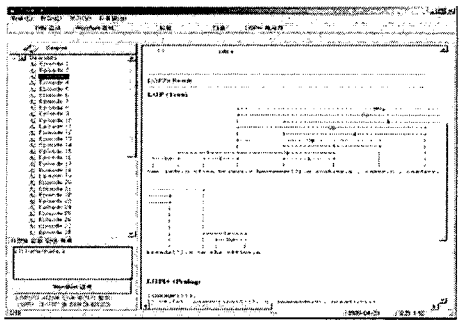

그림 5 ISAAC의 문장분석(LGPI+)

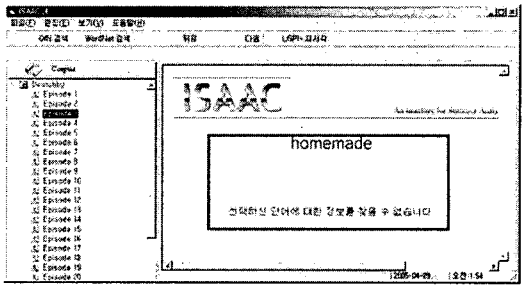

그림 6 ISAAC의 문장분석(OfN)

그림 5는 입력된 문장의 LGPI+의 구문분석정보를 보여준다. 그림 6은 선택된 단어의 OfN정보를 보여준다. 그림 5와 6에서 표시된 부분은 어휘사전에 등록되지 않은 단어를 발견한 화면이다.

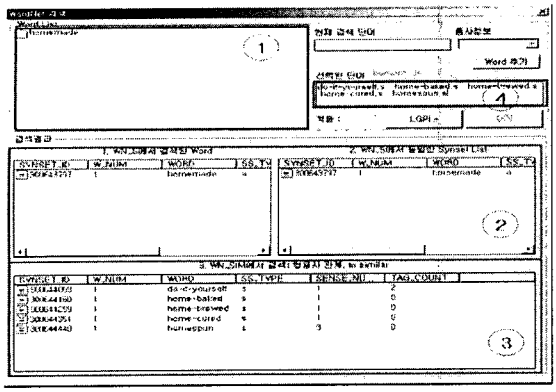

그림 7 WordNet 검색

그림 7은 미 등록단어의 WordNet 검색화면이다. ①은 미 등록단어의 목록이다. ②는 해당어휘가 포함된 WordNet의 유의어집합을 검색한 결과이다. ③은 비슷한 의미관계에 있는 SYNSET의 검색결과이다. ④는 ②③에서 선택한 단어의 목록이다.

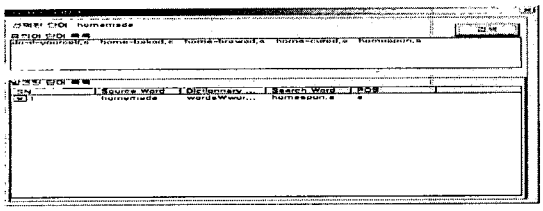

그림 8 어휘사전 검색(LGPI+, OfN)

그림 8은 그림 7의 ④에서 선택한 단어가 어휘사전에 등록되어 있는지 확인하는 화면이다. 단어목록에서 'homespun'이 어휘사전에 등록되어 있다. 따라서 LGPI+의 어휘사전에 'homespun'과 동일한 위치에, 그리고 OfN의 어휘사전에는 'homespun'의 OfN정보와 동일하게 'homemade'정보가 등록된다.

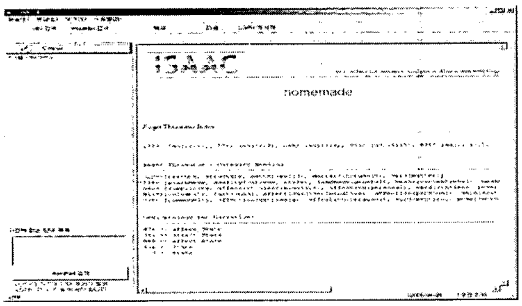

그림 9 'homemade'의 OfN정보

4. 결 론

본 논문에서는 문장분석용 통합 사용자 인터페이스 ISAAC의 문장분석 성공률을 높이고자 WordNet의 어휘정보를 이용하였다. 구체적으로는, 구문분석기 LGPI+가 문장에서 나타난 어휘를 찾지못해 실패한 경우, WordNet에서 유의어를 찾고, 이를 활용하는 어휘사전에 등록한 후 문장을 재분석하게 된다. 또한, 문장의 주요 구성성분들의 온톨러지 OfN 정보를 찾을 수 없을 때, WordNet의 유의어 목록에서 선택한 적절한 어휘의 OfN을 등록해서 의미정보를 보완한다.

ISAAC은 주어진 문장의 구문분석과 의미분석이 동시에 가능한 시스템으로, 문장의 구문분석 정보와 OfN정보를 문장분석가에게 제공한다. 구문분석기 LGPI+와 OfN의 어휘사전이 보유하고 있는 정보는 문장분석에 필요한 어휘보다 제한적이다. 따라서 본 논문에서는 문장분석의 성공률을 높이고, 보다 더 견고한 문장분석 및 원문이해 시스템을 위해 WordNet의 어휘정보를 이용하여 문장분석 시 미등록 어휘문제를 해결하였다.

<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] WordNet. http://wordnet.princeton.edu/.

[3] 김민찬, 김곤, 배재학, "구문분석기의 어휘확장", 한국정보처리학회 춘계 학술대회 게재예정. 2005.

[4] 김곤, 김민찬, 배재학, 유해영, 이종혁, "문장분석용 통합사용자 인터페이스 ISAAC의 개선", 한국정보처리학회 춘계 학술발표대회 논문집, 제10권 제1호, pp325-328. 2003.

[5] Link Grammar. http://www.link.cs.cmu.edu/link/.

[6] 양재군, 배재학, "온톨로지 정보를 이용한 범주 재편성: Roget 시소러스의 경우". 한국정보처리학회, 춘계학술발표논문집 9권 1호, pp.515-518, 2002.4