문장 분석기 및 전자사전 구성에 대한 연구

○
윤 준태, 송 만석
연세 대학교, 전산과학과

Study on Sentence Analyzers and Electric Dictionary

Yoon Juntae, Song Manseok
Department of Computer Science, Yonsei University

요약

자연어를 분석하는데 있어 가장 중요한 것은 지식 베이스(Knowledge Base)가 얼마나 정확하고 많이 구축되어 있는가 하는 것이다. 일반적으로 이 지식 베이스는 사전으로 구성될 수 있는데 이를 전자 사전이라 한다. 또 지식 베이스의 정보들은 계속적으로 유지, 수정되는데 이는 말뭉치의 분석을 통해 얻어질 수 있다. 본 논문은 전자사전의 구성및 말뭉치의 분석과 관리를 구문 분석기를 통해서 알아 본다.

1. 서론

자연 언어를 분석하는데 있어서 사전 정보의 풍부함과 정확성은 가장 중요한 요소이다. 자연 언어 문장을 읽고 이해한다는 것은 그 문장이 표현하는 의미를 이해하는 것인데 자연어의 다양한 표현과 의미를 제대로 파악하기 위해서는 컴퓨터가 문장을 인식하기 위한 충분한 지식을 가지고 있어야 한다. 이 지식 기반이 사전의 정보로 들어 간다.

전자 사전이란 컴퓨터가 자연 언어를 다루기 위해서 그 언어에 대한 모든 정보를 가지고 있는 사전을 말한다. 외국의 경우, 이러한 연구는 이미 활발히 진행되어 왔다. 특히 기계 번역에 대한 연구를 중심으로 자신의 국어에 대한 연구는 물론이고 다른 나라의 언어도 전자화시키고 있다. 대표적으로 일본의 EDR(electronic dictionary research)이 그것이고, 유럽의 경우에도 각각 자기 나라의 언어현상을 전자 사전에 담는 연구를 계속적으로 진행시켜 왔다. 우리 나라의 예로는 현재 연세 대학교에서는 이와 같은 추세에 발맞춰 우리 나라의 언어현상을 그대로 반영할 수 있는 사전을 만들기 위해 한국어 사전 편찬실을 설립하고, 계속적으로 말뭉치를 구축하면서 이에 대한 연구가 계속되고 있다. 이와 병행하여, 한국어 처리를 위한 전자 사전의 구성을 연구 중이고 이들의 구성을 위한 여러 가지 도구들을 개발하고 있다.

본 논문에서는 전자 사전의 원형과 이를 이용하는 자연 언어 처리 도구의 하나로 파

서를 구성하였다. 한편 구성된 전자 사전은 계속적으로 갱신되어야 하는데, 사전을 평가하는데 가장 중요한 것이 말뭉치를 테스트하고 이를 평가하는 것이다. 본 논문은 현재 구성된 사전과 파서를 이용해서 작은 규모의 텍스트를 분석하고 분석된 말뭉치의 관리와 사전 정보의 갱신에 대해 연구하였다.

2. 사전의 구성

사전에 담길 정보는 형태소 분석을 위한 형태소 정보와 품사 정보와 같은 통사 정보와 개념 정보, 그리고 공기 정보 등이 있다. 여기서 형태소 분석기는 논외로 하고 파서의 구현을 위한 구문 정보와 개념 정보를 살펴 본다.

2.1 한국어의 품사 체계

문장 분석을 위해 품사는 각각 통사적인 특성에 따라 세분화된다. 이 중 명사와 동사의 분류 체계에 대해 살펴보면 다음과 같다.

(1) 명사

(2.1)

- 명사
  - 고유명사
  - 보통명사
    - 완형 보문 명사
    - 불구 보문 명사
    - 일반 보통 명사
  - 불완전명사
    - 단위성 불완전 명사
    - 상용 보문 불완전 명사
    - 완형 보문 명사
    - 불구 보문 명사
    - 부사성 불완전 명사

(2.1)은 명사의 앞에서 일어나는 통사적 특징을 통해 세분화된 명사의 갈래이다. 명사의 구분 기준에는 일반적인 갈래에 구문 구조적 특징인 관형절에서의 보문현상을 반영한다.

(2) 동사

한국어의 동사 유형을 볼 때, 크게 다음과 같이 큰 갈래를 둘 수 있다.

(2.2)

1) S + V

(예) 그 아이는 혼자 잘 논다.
2) S + C + V
(예) 그는 결국 의사가 되었다.
3) S + S + V
(예) 그녀는 눈이 예쁘다.
4) S + O + V
(예) 우리는 함께 아침밥을 먹었다.
5) S + I.O + D.O + V
(예) 나는 그에게 책을 주었다.
6) S + O + O.C + V
(예) 그는 아들을 의사로 만들었다.
7) S + O + O + V
(예) 그는 그 아이를 사람을 만들려고 노력했다.

위의 분류 중 이중 주어 유형이나 이중 목적어 구문의 경우 양상이 매우 복잡하고 실제로 2 개 이상의 주어가 올 수도 있으므로 좀 더 연구가 필요한 부분이나 일단은 위와 같이 분류한다. 또 각 경우에 있어서도 좀 더 세분화된 동사 유형이 있다. 예를 들어 주어+보어+동사 구문의 경우에 보어는 보격조사 혹은 동사의 부사형 등 여러 가지 형태로 실현된다.

2.2 머리 개념

개념(concept)이란 어떤 사물 혹은 객체에 대해 누구나 공통적으로 가지고 있는 생각의 대상이다.[10] 머리 개념(head concept)이란 어떤 명사를 대표할 수 있는 개념을 표시하는 기호이다. '사람'은 '영희'의 머리개념이 될 수 있다. 본 구문분석기에서 의미적인 요소는 제외되나, 파서가 갖추어야 할 기본적인 기능을 위해서 다루어져야 할 기본 개념을 둔다.

(예문) ① 철수는 영희와 공원에 갔다.
② 영수는 빵과 우유를 먹었다.

예문의 ①에서 조사 '와'는 동반을 의미하는 부사격 조사로 쓰였고 ②의 '과'는 접속 조사로서 쓰였다. 이러한 관계는 두 명사의 개념 비교를 통해 검사할 수 있다. 예문 ①에서 '영희'와 '공원'은 각각 인간과 장소의 개념으로 병립할 수 없다. 이 과정을 통해서 '과'가 갖는 중의성을 해결함으로써 파싱 과정에서 효율성을 높여 줄 수 있다. 또 이 머리개념은 후에 의미 네트워크와 인터페이스를 제공하는 정보가 된다.

2.3 공기 정보

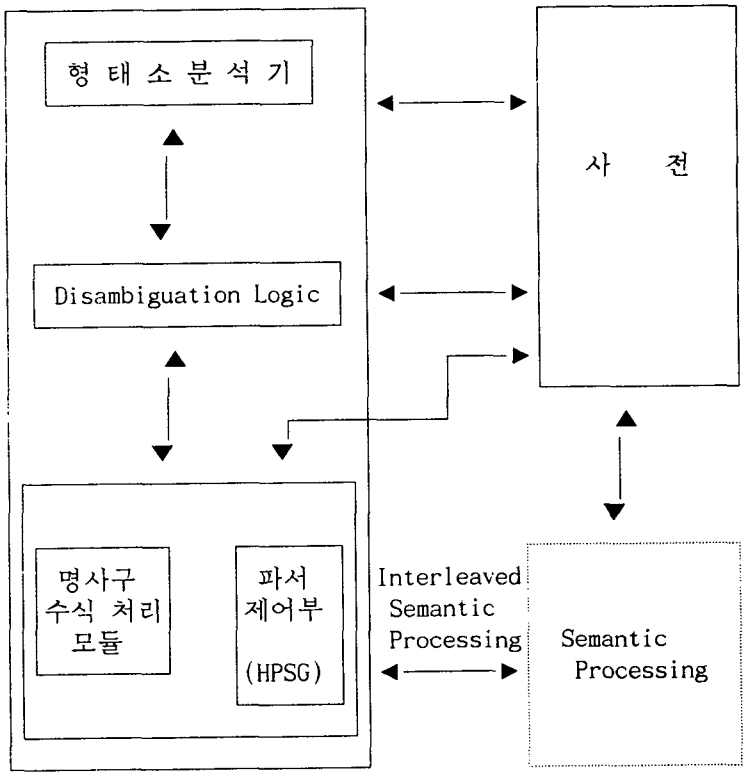
여기서 구문적 공기(syntactic collocation)라 함은 문장내에서 특별한 단어들이 서로 연접해 나타나는 현상을 말한다. 예를 들어, 다음은 구문적 공기이다.

(예문) ① 그는 살기 위해서 그 곳을 도망쳤다.
② 우리는 그 일을 해낼 수 있다.

이러한 관점에서 보면 보조용언도 특수한 공기어로 볼 수 있다. 공기 정보는 특히 문장내에서 조사의 생략을 다루거나 중의성 처리에 중요한 정보가 된다.

3. 구문 분석기의 구현

3.1 시스템의 구성



<그림 1> 시스템의 구성도

시스템의 개략적인 구성도는 <그림 1>과 같다.

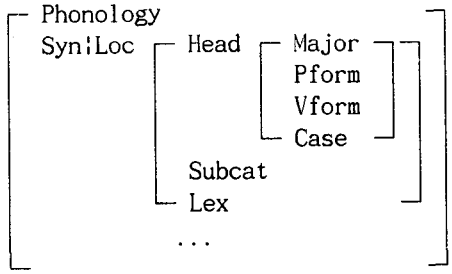
3.2 파서 제어부

파서 제어부는 전체 파싱 시스템의 규칙 적용, 파싱 메모리의 관리를 하는 파서의 핵심적인 부분이다. 자연어 문장을 파싱하기 위해서는 먼저 언어를 모델링하고 형식화하는 작업이 필요하다. 본 논문에서 구현한 파서는 언어 모델로 HPSG를

선택했다. HPSG는 GPSG의 변형으로 아주 간단한 몇 개의 규칙과 원리(Principle) 그리고 복잡한 어휘 정보로 구성 요소들 간의 결합 가능성을 검사한다. 모든 어휘 정보는 자질 구조에 일관되게 구성되고, 단일화를 연산자로 사용한다.

HPSG의 어휘부는 (3.1)과 같다.

(3.1)



한편, 각 부호들을 결합시켜서 보다 복잡한 형태를 만들기 위해 문법 규칙과 원리들을 설정한다. HPSG는 어휘내 속성들이 단일화를 위한 많은 정보를 포함하고 있으므로 문법 규칙이 매우 간단하다. 문법 규칙은 소수의 원리를 바탕으로 이루어지는데 본 시스템은 HPSG의 다음 원리를 이용하고 있다.

(1) Head Feature Principle

머리어의 HEAD 자질값은 모 노드(Mother Node)의 HEAD 값과 동일하다. 이것은 HPSG의 기본 원리로 머리어의 어휘 정보가 문장의 구성에 이용됨을 뜻한다.

(2) Subcategorization Principle

모노드, 머리어, 비머리어 딸노드의 SUBCAT 자질값은 다음 관계의 하나를 만족해야 한다.

① M -> C H
② M -> A H
③ M -> H1 H2

이것은 머리어의 SUBCAT으로부터 범주 하나를 뽑아 보어와 단일화시켜 모 노드를 생성하는 것이다. 특히 ③은 머리어-머리어 결합으로 한국어의 등위 접속문의 구성에 필요한 규칙이다. ①은 동사에 필수적으로 요구되는 논항을 찾는 규칙에 해당하며, ②는 수식어 피수식어 관계를 밝혀 주는 규칙이다.

또 SUBCAT은 한국어가 부분 자유 어순이기 때문에 SUBCAT 자질은 stack 대신에 set으로 구성한다.

4.3 Disambiguation Logic과 명사구 처리

일반적으로 형태소 분석 후에 나오는 출력은 많은 중의성을 포함한다. 이 중의성은 시스템의 성능을 급격히 떨어뜨리는 요인이 된다. 그러므로 가능한 한 중의성을 없애주면 파서가 효율적이 될 뿐 아니라 정확한 분석을 할 수 있게 된다.

현재 구축되어 있는 Disambiguation Logic은 보조 용언 처리 루틴과 공기 관계 처리 루틴 그리고 heuristic을 이용한 Disambiguation 루틴이다. 공기 처리 루틴은 형태소 분석이 끝나면 형태소 분석이 된 결과로부터 공기 관계가 존재하는지를 살핀다. 공기 관계가 발견되면 각 노드에 공기 관계 태그를 할당하고 그 노드의 다른 중의성을 가진 노드를 제거한다.

또 한국어의 명사구는 어느 정도 순서를 부여할 수 있다. 즉, 이것은 finite automata 형식으로 구현할 수 있음을 의미한다. 수식어-피수식어 관계가 밝혀지면 그 관계를 설정해 놓음으로써 바로 A,H 단일화의 규칙 활성화 루틴으로 모노드를 생성하게 된다.

또 명사구 처리에서도 Disambiguation Logic이 있다. 예를 들어

(예) 사과 한 개

의 경우 사과는 중의성이 없는 단어인데 다음의 '한'은 관형사나 동사+관형형어미로 분석될 수 있다. 또 '개'의 경우도 일반명사나 단위성 불완전명사로 분석되는데 이는 <명사+수.관+단.불.명>의 구조로 파악할 수 있고 이로써 중의성이 해결된다.

4. CORPUS의 분석

현재 구성된 사전과 문장 분석기에는 검증이 필요하다. 검증에 가장 중요한 자료는 말뭉치의 분석이다. 많은 텍스트를 분석해 봄으로써 현재 기록된 사전의 내용과 문장 분석기의 오류를 수정하고 올바름을 검증할 수 있다. 사전과 처리 도구의 개발 과정은 <그림 2>와 같다.

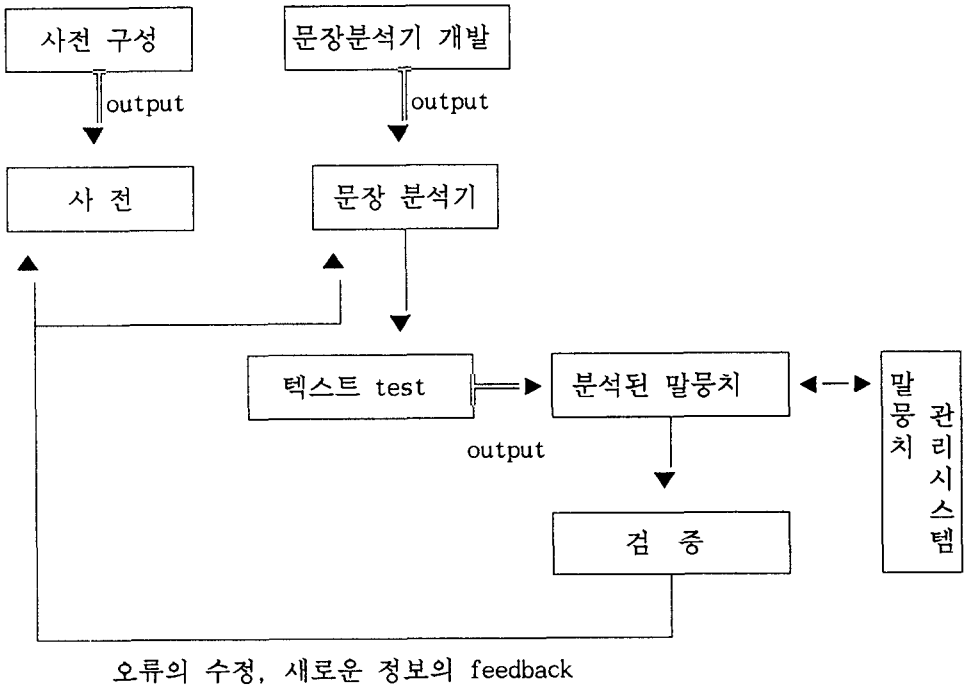
한편, 분석된 말뭉치는 다음과 같은 구조를 갖는다.

Head 문장	문장의 구문구조

또 구문 분석기를 통해서 출력이 된 문장에 대한 구문 구조의 예는 다음과 같고 이는 그대로 분석된 말뭉치로 관리된다.

(예) 삼촌과 나는 그늘에 앉아서 포도를 먹었습니다 .

((SUB (삼촌과 ((삼촌 0)(과 15)) 사람)(나는 ((나 1)(는 15)) 사람))
(((그늘에 ((그늘 0)(에 15)) 장소)(앉아서 ((앉 6)(아서 17))))
((OBJ 포도를 ((포도 0)(를 15)) 음식)
(먹었습니다 ((먹 6)(었 14)(습니다 17))))))



<그림 2> 사전 및 문장 분석기의 개발 과정

(1) 말뭉치의 테스트

테스트는 검증을 위해 가능한 많은 문장을 임의로 선택한다. 하지만 이들 텍스트들은 또한 언어학적으로 의미가 있는 문장일 필요가 있다. 분석된 결과는 후에 사전의 구축에 필수적으로 작용하는 요소들이므로 텍스트의 선정은 매우 중요하다.

(2) 분석된 말뭉치

분석되 결과를 담은 말뭉치들로부터 실제 응용을 하기 위해서는 색인이 되어 있어야 한다. '먹다'라는 단어의 쓰임에 대해 정보를 수집하고 사전의 검증을 위해서는 분석된 말뭉치에 색인을 해 놓음으로써 보다 쉽게 처리할 수 있다. 결국 분석된 말뭉치는 말뭉치 관리기를 통해 유지된다. 현재 말뭉치 관리기는 기본적인

형태로 구성되어 있으며 이는 분석된 말뭉치의 관리를 위해 확대될 예정이다.

5. 결론

본 연구에서는 자연 언어 처리를 위한 전자 사전 중에서 특히 문장 분석과 관련된 사전 정보와 그 사전을 이용한 구문 분석기를 구현하였다. 또 이들의 관리와 유지, 수정을 위해서 사전과 분석 도구를 이용하여 문장을 분석하고 그 출력을 유지하면서 정보를 갱신할 수 있도록 했다. 이는 후에 말뭉치 관리 도구를 통해서 색인이 된 상태로 관리되어야 한다.

현재는 수집한 어휘의 양이 많지 않아 완전히 개발이 되지는 않은 상태로 본 논문은 그 원형을 제시하였다. 현재 분석된 문장과 어휘는 국민학교 6학년과 중학교 국어 교과서의 약 800 문장으로부터 발췌되었다. 실제 사전은 한국어의 언어현상을 충분히 반영해야 하므로 이보다 훨씬 많고 다양한 말뭉치의 선정을 통해 충분한 정보를 담도록 구성해야 한다.

참고 문헌

[1] 한국어 분류표, 한국어 정보 처리 연구실 Memo

[2] 남기심, 고영근, 표준 국어 문법, 탑출판사, 1990

[3] 최재희, 한국어의 접속문 구성 연구, 탑출판사, 1992

[4] 김승렬, 국어 어순 연구, 한신 문화사, 1988

[5] 남기심, 국어 완형 보문법 연구, 국어학회, 1989

[6] 조성원, 사전에 기반한 한국어 문장 해석 시스템 원형의 연구, 연세 대학교 석사학위 논문, 1992

[7] 배순일, HPSG를 기반으로 한 한국어 문장 분석기의 설계, 연세 대학교 석사학위 논문, 1992

[8] 양재형, HPSG에 기반한 한국어 분석기의 연구, 서울 대학교 석사학위 논문, 1992

[9] 우승균, 구문관계를 이용한 한국어 구문 분석, 한국 과학 기술원 석사학위 논문, 1992

[10] EDR, An Overview of the EDR Electronic Dictionaries, Japan EDR Institute, LTD., 1990

[11] EDR, Japanese Word Dictionary, Japan EDR Institute, LTD., 1990

[12] Pollard, Lectures on HPSG, 형식 문법 이론 연구 자료 제 2 집, 1985