자연언어 처리에 있어서의 형태소 처리 및 어휘 사전

권 혁 철

언어는 인간의 고도한 지적 능력의 산물로, 기존 컴퓨터 혹은 인공지능 기술로는 인간 수준의 언어능력을 가진 기계를 개발하는 것은 불가능하다. 그러나 제한된 영역을 대상으로 하는 문장을 번역하거나, 이해하고 대화하는 것은 기존 기술로도 가능하다. 특히 사용자 지침서의 번역과 같이 처리 대상이 제한되고, 전문적 용어에 대한 지식이 요구되지만 사용되는 문장은 비교적 단순한 분야나, 항공 예약 시스템과 같이 인간이 종사하기에는 단순하고 비창조적인 분야를 기계에 맡기는 것은 상업적으로 성공할 수 있을 뿐만 아니라 인간에게도 도움이 된다.

자연언어 처리는 철자검사기와 같은 텍스트 처리, 자연언어 정보 검색 및 요약, 음성 합성/인식, 자연언어 사용자 인터페이스 및 기계번역 등 다양한 응용 분야를 가진다. 그런데 이 모든 자연언어 처리의 응용 분야가 공통적으로 가지는 것이 어휘 혹은 형태소 처리부다. 특히 교착어인 한국어는 어휘 처리와 이를 위한 사전의 역할이 중요하다.

어휘 사전이 가지는 정보는 응용 분야에 따라 형태소정보, 음성정보, 통사정보 및 의미정보 등 다양할 수 있다. 한편 어휘 사전의 기본단위, 즉 어휘 항목이 무엇인가는 어휘 사전이 가지는 정보와 함께 매우 중요하다. 어휘 항목은 형태소, 단어, 어절, 숙어, 발음 기호 등 응용 분야에 따라서 다르기 마련이다. 일반적으로 자연언어 처리 시스템은 형태소를 어휘 항목으로 설정한다. 그러나 자연언어 처리에서 사용되는 "형태소"라는 용어는 언어학에서 사용되는 "형태소"와 같이 엄격한 의미로 사용되지는 않고 있다. 즉 '짓누르다'의 '짓'은 틀림없이 형태소의 일부분이지만, 자연언어 처리 시스템은 '짓누르다' 자체를 한 개의 형태소로 설정한다. 그 이유는 형태소로부터 단어를 조성하는 과정이 기계로 처리하기에는 매우 어렵기 때문이다. 그러나 조사나 어미 등은 형태소로 분류한다. 즉, 자연언어 처리 시스템에서 형태소 분석을 위한 사전은 일률적으로 형태소나 단어만을 어휘 항목으로 취하지는 않는다.

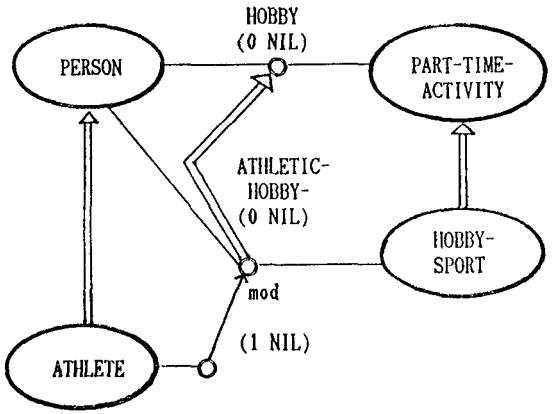
사전이란 특정 용어 혹은 항목에 대한 정보를 모아두고, 필요할 때 이용할 수 있게 해주는 정보의 집합체다. 그러나 사전은 단순한 정보의 모임이 아니고, 그 정보를 이용할 수 있는 효과적인 접근방법(access method)도 포함한다. 따라서 컴퓨터에 사전 정보를 효율적으로 표현하고, 내용을 손쉽게 이용할 수 있게 해주는 컴퓨터 내부의 표현구조와 보조기억장치에 사전을 기억하는 방법 및 사전 내용의 접근방법 등이 사전 연구의 중요한 문제다.

특히 단일화문법과 같이 사전 정보를 중심으로 문장 분석이 행해지는 경우에는 사전의 중요성이 더욱 크다. 단일화문법은 합성의 원리에 기초하여 사전의 정보로부터 단일화라는 연산에 의해 문장 분석을 행한다.

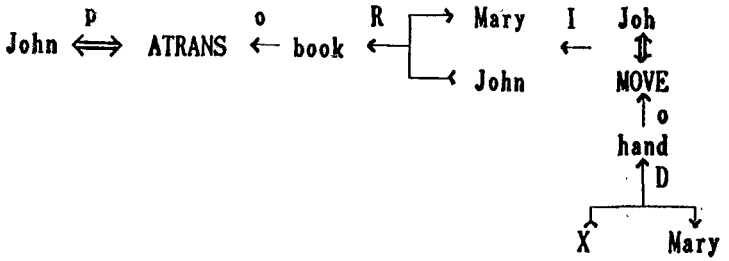
자연언어 처리에서 이용되는 사전의 내용은 어휘정보와 규칙정보로 나눌 수 있다. 어휘정보는 하위범주화정보, 자질정보, 불규칙정보, 음성정보 및 의미정보를 포함한다. 규칙정보는 형태소의 접속규칙, 불규칙 동사의 변형규칙, 동사의 피동 변화에 따른 하위범주화 변화규칙 등 어휘정보를 이용하여 추가의 어휘를 생성하거나 분석하기 위한 정보다.

어휘가 가지는 의미(semantics)를 표현하기 위해서는 자질구조 외에도 의미 공준이나 의미 해체를 이용하기도 한다. 의미 공준은 명사 의미의 표현에 일반적으로 많이 사용되며, 지식표현에서 사용되는 "is-a" 관계와 밀접하다. 이와 같은 의미 공준에 기반한 의미망(semantic network)인 KL-TWO가 BBN에서 자연언어 처리에 이용되었다. 의미 해체는 동사 의미의 표현에 많이 사용된다. Schank는 15가지의 의미소(semantic primitive)를 도입하여 동사의 의미를 표현하고자 시도했다.

그림 1)은 의미 공준과 의미 해체에 의한 자연언어 문장의 의미 표현 예를



An athlete is a person who has at least one athletic-hobby



John gave Mary a book by handing it to her

〈그림 1〉

보여준다.

형태소 분석과정에서는 세 가지의 어려운 문제가 발생한다. 그 첫번째가 중의성의 처리다. 예를 들면 우리말 어미인 "고"는 연결어미와 인용어미로 사용될 수 있다. 그러나 "고"의 중의성은 "고" 왼쪽에 나타나는 형태소에 의해 형태소 분석단계에서 해결될 수 있다. 한편 "겠"이 가지는 중의성은 통사분석 단계에서 자질구조를 이용함으로써 해결이 가능하다. 그러나 "감기는 병이다"에서 "병"이 가지는 중의성은 의미 분석 혹은 화용 분석을 요구한다.

두번째 문제는 미등록어 처리 문제이다. 만약 어휘 사전의 항목에 나타나지 않는 단어가 나오면 어떻게 할 것인가가 이 문제의 핵심이다. 인간은 문맥 등을 파악하여 그 어휘가 고유명사인지, 잘못 입력된 어휘인지를 쉽게 파악할 수 있다. 그러나 기계는 이와 같은 처리가 불가능하다. 그렇다고 모든 고유명사를 모두 사전에 넣을 수는 없다. 이 문제는 사용자가 사용목적에 따라 자신을 위한 독자적 사전을 이용하여 해결할 수도 있다. 그러나 기존 기술로는 사전에 없는 어휘는 모두 잘못 입력된 것으로 처리하는 것이 일반적이다.

세번째 문제는 형태소의 분리 및 재합성의 문제다. 형태소를 분리하는 과정에서 불규칙용언이나 줄임말은 처리과정을 복잡하게 한다. 특히 "가신다 신다"와 같이 두 어절이 합해져 이루어진 어절에서 형태소를 분리하는 것은 매우 어렵다.

그러면 컴퓨터에 의해 형태소가 분리되는 과정을 간략히 살펴보자.

〈형태소 분석 방법〉

하나의 어절에 대한 형태소 분석 방법은 어절의 각 문자를 초, 중, 종성의 구성체로 보고 분석한다. 개략적으로 말하면, 각 초, 중, 종성의 자모를 하나씩 추가하면서 형성된 어절이 사전에 존재하는 형태소인지를 검사하여,

존재하면 하나의 형태소로 등록한다. 그 다음 계속하여 추가적인 어절 내의 형태소를 찾는다. 상세한 형태소 분석 방법을 예를 보면서 살펴보면 다음과 같다. 예로 "철수는"이란 어절에 대하여 각 단계별로 살펴보면 다음과 같다. ("철수는"을 "ㅊㅓㄹㅅㅜㄴㅡㄴ"으로 본다)

1. "ㅊ"에 대하여 사전 검사 : 형태소 형성 안됨.
2. "쳐"에 대하여 사전 검사 : 형태소 형성 안됨.
3. "철"에 대하여 사전 검사 : "철(iron)" 형태소 발견.

"수는"에 대한 형태소 분석.

3-1. "ㅅ"에 대하여 사전 검사 : 형태소 형성 안됨.
3-2. "수"에 대하여 사전 검사 : "수(number)" 형태소 발견.

"는"에 대한 형태소 분석.

3-2-1. "ㄴ"에 대한 사전 검사 : 형태소 발견 안됨.
3-2-2. "느"에 대한 사전 검사 : 형태소 발견 안됨.
3-2-3. "는"에 대한 사전 검사 : "는(조사)" 형태소 발견.

=> "철수는"에 대하여 "철(iron)"+"수(number)"+"는(어미)" 형태소 열 발견.

3-3. "수ㄴ"에 대하여 사전 검사 : 형태소 발견 안됨.
3-4. "수느"에 대하여 사전 검사 : 형태소 발견 안됨.
3-5. "수는"에 대하여 사전 검사 : 형태소 발견 안됨.

4. "철ㅅ"에 대하여 사전 검사 : 형태소 분석 안됨.
5. "철수"에 대하여 사전 검사 : "철수(고유 명사)" 형태소 발견.

"는"에 대한 형태소 분석.

4-1. "ㄴ"에 대하여 사전 검사 : 형태소 분석 안됨.
4-2. "느"에 대하여 사전 검사 : 형태소 분석 안됨.
4-3. "는"에 대하여 사전 검사 : "는(조사)" 형태소 발견.

=> "철수는"에 대하여 "철수(고유 명사)"+"는(조사)" 형태소 열 발견.

6. "철수ㄴ"에 대하여 사전 검사 : 형태소 발견 안됨.
7. "철수느"에 대하여 사전 검사 : 형태소 발견 안됨.
8. "철수는"에 대하여 사전 검사 : 형태소 발견 안됨.

==> 1) "철(iron)"+"수(number)"+"는(조사)"와
2) "철수(고유명사)"+"는(조사)"

중에서 적절한 형태소열 결정.

: 1) "철(iron)"+"수(number)"+"는(조사)"는 부적절.

따라서 어절 "철수는"의 형태소 분석 결과는 고유명사 "철수"와 조사 "는"이 결합된 형태임을 알 수 있다.

〈불규칙 변화를 하는 어절의 형태소 분석〉

불규칙 변화를 하는 어절의 형태소 분석에 대하여는, 먼저 사전에 불규칙 변화를 하는 형태소에 대한 정보가 있다고 가정한다. 예로 "아름다운"과 "아름답게"에 대하여 분석해 보자.

1. "ㅇ"에 대하여 사전 검사 : 형태소 발견 안됨.

...

7. "아름다"에 대한 사전 검사 : 사전에서 "ㅂ"불규칙하는 형태소의 표제어로 발견. (사전에는 "아름다"를 표제어로 두고 "ㅂ"불규칙하는 형태소라는 정보를 유지한다.)

7-1. 이때 문자가 "우"인지를 검사하여 "우"이면 "ㅂ"불규칙 변화한 "아름답(용언)" 형태소 발견 : "아름다운"인 경우.

"ㄴ"에 대한 형태소 분석.

7-1-1. "ㄴ"에 대하여 사전 검사 : "ㄴ(어미)" 형태소 발견.

=>"아름다운"에 대하여 "아름답(용언)"+"ㄴ(어미)" 형태소열 발견.

7-2. "우"가 아니면 "ㅂ"인지 검사하여 "ㅂ"이면 "ㅂ"불규칙 하지 않은 "아름답(용언)" 형태소로 결정 : "아름답게"인 경우.

"게"에 대한 형태소 분석.

7-2-1. "ㄱ"에 대한 사전 검사 : 형태소 발견 안됨.

...

7-2-3. "게"에 대한 사전 검사 : "게(어미)" 형태소 발견.

=> "아름답게"에 대하여 "아름답(용언)"+"게(어미)" 형태소 열발견.

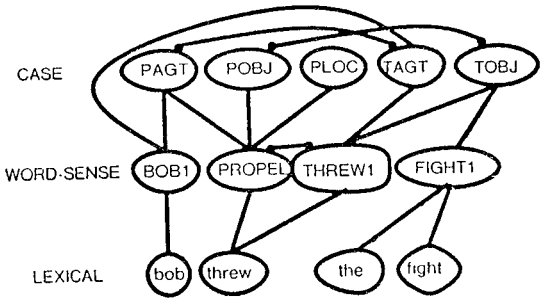
==>

"아름다운"과 "아름답게"에 대하여 각각 하나의 형태소열이 만들어지므로 적절한 형태소열의 선택이 필요하지 않음. 따라서 "아름다운"은 "아름답(용언)"+"ㄴ(어미)"로, "아름답게"는 "아름답(용언)"+"게(어미)"로 형태소 분석된다.

실제 자연언어 처리를 위한 사전을 구성하는 것은 매우 어렵고 힘든 일이다. 따라서 사전을 만들기 위한 여러가지 보조 프로그램의 도움 하에서 사전

을 구현한다. 이와 같은 방법의 하나가 인간이 사용하는 사전으로부터 기본적인 정보를 가져오는 방법이다. 현재 Oxford 사전과 Longman 사전으로부터 자연언어 처리에 필요한 정보를 입수하는 방법이 다양하게 연구되고 있다.

최근에는 단어의 의미를 고정하기 위해 연결주의적 접근(connectionist approach), 다른 말로는 신경망적 접근(neural network)이 시도되고 있다. 다음 그림 2)는 연결주의적 접근에 의해 "Bob threw the flight"를 단어의 의미를 고정하고, 격을 결정하는 예를 보여준다



Subset of the network for "Bob threw the fight"

〈그림 2〉

어휘 사전은 자연언어 처리에 있어서 매우 중요한 역할을 한다. 그러나 어휘 사전을 만드는 것은 매우 방대하며, 시일이 소모되는 작업이므로, 국가적 차원에서 공동의 노력이 요구된다. 특히 전산학, 언어학 및 인지심리학자들이 서로 도와서 자료를 모으고 분석하여, 자연언어 처리를 위한 핵심 사전을 만드는 것은 자연언어 처리 시스템의 빠른 실용화를 위해 매우 절실하다.

참 고 문 헌

Bolt Beranek and Newman Inc.(C. Sidner et al.) (1984) Research in Knowledge Representation for Natural Language Understanding, National Technical Information Service, Report No. 5694.

Cheng-ming Guo (1989) Constructing A Machine Tractable Dictionary From Longman Dictionary of Contemporary English, MCCS-89-156, National Techni-University of New Mexico State.

Dan Fass (1988) Collative Semantics: A Semantics for Natural Language Processing Ph.D Dissertation, MCCS-88-118, Computing Research

Laboratory, University of New Mexico State.
Garrison W. Cottrell (1989) A Connectionist Approach to Word Sense Disambiguation, Pitman Publishing.
Roger C. Schank with Peter G. Childers (1984) The Cognitive Computer: On Language, Learning, and Artificial Intelligence, Addison-Wesley Publishing Compony Inc.

ABSTRACT

The Role of Morphological Analysis and Lexicon in NLP

Hyuk-Chul Kwon

This paper shows the role of morphological analysis and lexicon in natural language processing. The information in lexicon varies by the application fields. One of the difficulties in constructing lexicon is the represention of semantics. Meaning postulate and semantic decomposition are generally used for representing meaning. The morphological analysis method of Korean is given as an example.

607-735
부산시 금정구 장전동
부산대학교 자연대학
전자계산학과