論文94-31B-3-13

사전 정보와 차트 자료 구조를 이용한 효율적인 형태소 분석기 및 합성기(KoMAS)

(An Efficient Korean Morpheme Analyzer and Synthesizer using Dictionary Information and Chart Data Structure)

金 貞 海*, 李 相 祚**

(Jung Hae Kim and Sang Jo Lee)

要 約

본 논문은 한국어의 어절을 구성하는 형태소의 분석과 그 합성에 관해 기술한다. 형태소의 분석을 위해 사전내에 형태소 자질로 "morph"의 도입과 차트 자료 구조의 이용을 제안하고, 이를 이용하여 불필요한 형태소의 생성을 억제하며, 경험적 정보를 이용한 사전 검색을 최소화하면서 한 어절내에 모든 경우의 형태소 단위를 추출하였다. 또한 형태소의 합성을 위해 한 어절내에서 분리된 모든 경우의 형태소들은 미리 프로그램적으로 수행한 품사정보와 결합정보를 이용하여 분석된 모든 가능한 형태소간의 결합이 이루어지도록 구성하였다. 이러한 분석된 형태소의 합성은 자연언어처리의 다음 단계인 구문분석에 도움을 주기 위함이었다.

형태소의 분석과 합성을 위한 본 시스템은 분석된 형태소들의 자질들을 사전내의 구문·의미적인 자질요소들로 통합하여 하나의 어절을 생성하도록 하였으며, 이를 PC상의 C-언어로 구축하였다.

Abstract

This paper describes on the analysis of morphemes and it 's synthesis being constituted of Korean word phrases. To analyze morphemes, we propose the introduction of "morph" for morpheme features in lexicon and the usage of chart data structures. It controls over the generation of unnecessary morpheme, and extracts every possible morpheme unit in a word phrase which minimized lexicon investigation by using heuristic information. Moreover, to synthesize morphemes, it is composed of every possible analyzed morphemes in word phrases to take advantage of speech and union information which can be obtained for program. Therefore, the systhesis of analyzed morphemes were designed to aid a syntactic analysis next step of natural language processing.

This system for analyzing and systhesizing morpheme was to generate a word phrase by unifying syntactic and semantic features of analyzed morphemes in lexicon, and then established by C language of the personal computer.

* 正會員, 尙志專門大學 電算情報處理科
(Dept. of Computer Information Processing, Sangji Junior College)
** 正會員, 慶北大學校 컴퓨터工學科
(Dept. of Computer Eng., Kyungpook Nat'l Univ.)
接受日字 : 1993年 9月 21日

Ⅰ. 서론

컴퓨터에게 인간의 지적 능력을 이식하고자 하는 열망이 컴퓨터를 이용한 정보처리중에서 인공지능분야의 눈부신 발달을 야기시켰다. 특히 인공지능분야중에서 자연어처리는 인간의 언어로 컴퓨터 사용자에게 보다 편한 환경의 제공을 위해 자연어를 컴퓨터로 처리하고자 하는 것으로, 컴퓨터와 인간사이의 자연어 인터페이스, 기계번역시스템, 담화분석시스템 등에서 유용하게 이용되고 있다.[1,7,9]

자연어처리에 있어 형태소의 분석은 여러 처리 단계중에서 필수적으로 행해져야 하는 단계이다. 특히 한국어와 같이 형태소가 고정되어 있지 않는 언어에서는 형태소 분석 단계가 매우 중요한 부분을 차지한다. 형태소 분석 단계에서 얻어진 여러가지 정보는 구문 및 의미 해석 단계에서 사용되므로, 이 단계에서 처리되는 하나의 어휘에 대해 형태소에 관한 가능한 모든 품사 정보와 정보의 정확성이 형태소 이후의 단계에 커다란 영향을 미친다.

따라서 본 논문에서는 형태소의 분석 및 합성을 위해 필요한 어휘에 대한 품사분류, 형태론적인 결합정보를 제시하고, 자연어처리에서 중요한 위치를 차지하는 사전의 구성으로 형태소 분석을 위한 형태소 자질과 변칙·활용정보 자질을 도입하였다. 또한 차트 자료구조 및 한국어 어휘의 중요한 정보인 초성·중성·종성을 이용하여 어절내의 모든 가능한 형태소를 찾는데 있어, 사전을 검색하는 양을 최소화하며, 불필요한 형태소의 생성을 억제하여 어절내의 모든 경우의 형태소 단위를 추출하도록 하였다. 이렇게 추출된 형태소들은 구문 분석기에 최대한의 정보를 줄 수 있도록 각 형태소간의 결합을 통합이라는 메카니즘을 이용하여 하나의 가능한 어절들로 합성하여 모든 가능한 어절 단위의 정보를 추출하는 시스템을 구축하였다.

본 논문의 구성은 Ⅱ장에서는 형태소 분석 및 합성을 위한 한국어의 형태소적인 특징과 형태소의 분석 합성에 대한 정의를 살펴보고, Ⅲ장에서는 형태소 분석기 및 합성기의 설계를 위한 사전의 구성, 필요한 정보들 및 차트자료구조와 알고리즘에 대해 기술하고, Ⅳ장에서는 실험 결과와 시스템의 특성을 보여주며, Ⅴ장에서는 결론과 더불어 연구의 개선점 및 응용분야에 대해 기술하였다.

Ⅱ. 형태소적인 특징 및 분석·합성의 정의

한국어의 문장에서 하나의 어절을 형성할 때 무한한 수의 형태소가 결합하는 것이 아니라 제한된 수의 형태소가 결합한다. 이는 한 문장내의 문장 성분들이 맨 뒤에 위치하는 서술어를 제외하고는 자유롭지만, 한 성분을 이루는 어절내의 형태소간의 순서는 자유롭지 않음을 의미한다. 이와 같이 한국어의 문장을 이루는 성분들간에는 많은 특성이 있겠지만, 이 장에서는 특히 형태소 분석 및 합성을 위한 한국어의 형태소적인 특징과 분석·합성에 대한 정의를 기술하였다.

1. 한국어의 형태소적인 특징

한국어는 어근을 중심으로 거기에 뜻을 더하거나 품사를 바꾸는 접사 또는 어미가 차례로 여럿이 덧붙여서 어절을 이루는 첨가적 성격을 띤 언어이다.[6] 따라서 한국어의 형태소적인 특징을 아래(1-6)으로 기술하고, 이러한 특징들은 한 어절내에서 형태소의 분리 및 하나의 어절로 합성시에 이용되는 결합정보표로 모두 구성하였다.

[특징 1] 체언은 조사 앞에 위치한다.
[특징 2] 본용언은 보조용언 앞에 위치한다.
[특징 3] 선어말어미의 순서는 일정하며 그 자리의 위치를 함부로 바꿀 수 없다.
[특징 4] 어미 가운데서도 용언중에 동사에는 사용되지만 형용사 및 서술격 조사에는 사용되지 못하는 것이 있다.
[특징 5] 조사가 중복해서 나타날 수 있다.
[특징 6] "체언+조용사"에 의해 용언화될 수 있다.

2. 형태소의 분석 및 합성의 정의

형태소 분석은 입력 문장에서 어절 단위로 분리하여 형태소 단위로 분석하는 과정으로, 한 어절 혹은 낱말의 형태적 중의성에 대해 다수의 다른 결과를 모두 분석할 수 있어야 한다. 따라서 본 논문에서 의미하는 형태소 분석을 아래와 같이 정의한다.

[형태소 분석] 한 개 이상의 형태소들의 결합으로 구성된 어절로부터 이들을 구성하는 부분 형태소들로 분해하여 그 최소의 의미단위인 형태소로 나누는 작업을 "형태소 분석"이라 한다.

예1) "집을 지었다" 라는 문장에서 "지었다"는 짓(동사어간), 었(과거시제선어 말어미), 다(평서형 어말어미)로 형태소 분석된다.

예2) "꽃이 아름답자"와 같은 문장에서 "아름답자"는 아름답(형용사의 원형), 자(명사, 청유형 어미, "자다"라는 동사의 어간)로 분석된다. 그러나 이 문장은 통사론적으로는 전혀 문제

가 없는 문장이나, 형태론적으로는 잘못된 문장임을 형태소의 합성과정에서 검사하여 한 어절로의 성립이 불가능함을 조기에 판단하여 통사 분석기의 부담을 줄이도록 하였다.

위의 예2)와 같이 한국어의 형태부가 고유의 자리를 차지하지만, 그것은 여전히 통사부와 밀접하게 연관되어 있으므로, 분석을 통하여 추출된 각각의 형태소들이 결합을 하여 하나의 올바른 단어나 어절을 이루게 하는 형태소의 합성이 필요하다. 따라서 형태소의 합성은 프로그램적으로 구축된 형태소의 결합정보를 이용하여 결합이 가능한 형태소들을 하나의 어절 단위로 뭉쳐주기 위한 것으로 아래와 같이 정의한다.

[형태소 합성] 한 어절내에서 모든 가능한 형태소들이 분석되었을 때 하나의 어절로 바르게 구성될 수 있도록, 추출된 형태소들은 통합(unification) [6, 15]이라는 기제를 사용하여 형태소간의 중요 성분(형태소, 구문, 의미정보)들을 결합하여 하나의 어절 단위로 생성하는 것을 "형태소 합성"이라 한다.

앞의 예1)에서 보인 "지었다"의 각각 분석된 형태소들은 각각 형태소들간의 결합정보를 조사하여 형태소 합성을 수행시에 한 어절 단위로 구성하기 위해 "지었다"로 합성하면 구문 의미 정보 - 과거시제, 평서형,종결어미 - 를 각각 통합하여 하나의 결합된 정보를 출력하나, 예2)의 문장은 비록 형태소의 분석은 이루어졌으나, 그들의 결합정보에 의해 "아름답(형용사원형)"과 "자(명사)", "아름답(형용사원형)"과 "자(청유형어미)", "아름답(형용사원형)"과 "자(동사어간)"가 있을 수 있으나, 추출된 "자"의 형태소 정보가 형용사와의 결합이 가능한 형태소가 아니므로 이들 경우 모두가 "아름답자"라는 하나의 어절로 형성할 수 없는 형태론적으로 잘못된 문장임을 출력한다.

이러한 형태소의 합성은 다음 단계인 구문 분석기의 부담을 감소시키기 위해 이들 각 형태소의 중요 자질 정보들을 통합이라는 메카니즘을 이용하여 결합시켰다.

Ⅲ. 한국어 형태소 분석기 및 합성기의 설계

본 형태소 분석 및 합성기는 한국어의 특성인 띄어쓰기를 하나의 어절 단위로 보고, 입력되는 문장에서 띄어 쓰기가 바르다고 가정하여 어절 단위로 처리하였다. 또한 입력되는 문자의 종류로는 한글과 영문자, 숫자를 허용하며, 한글은 2-바이트 조합형을 사용하여 PC상에서 C-언어로 구현하였으며, 전체 시스템의 구조도는 그림1과 같다.

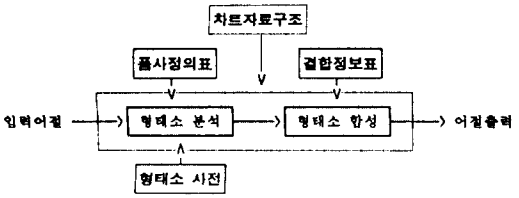


그림 1. 시스템 구조도
Fig. 1. System Diagram.

1. 사전의 구성

자연어처리에서 사전의 중요성은 여러 논문에서 발표되고 있으나, 아직까지 사전의 정규화된 형식에 대해서는 뚜렷한 방법이 제시되고 있지 않은 실정이며, 자연어처리의 범주에 따라 사전의 모양이나, 그 처리하는 방법을 각각 따로 정의하고 있는 실정이다. [2,4,7,8,9,11] 따라서 본 논문에서는 한국어 자연어처리 시스템의 각 단계에서도 이용가능하며, 특히 형태소의 효율적인 처리를 위해 사전의 구조를 그림2과 같이 "morph"자질을 정의하였다.

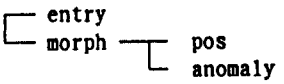


그림 2. 사전의 구조
Fig. 2. Structure of dictionary.

위 사전의 구조는 각 요소가 자질(feature) [3,6,7] 로서 표현되며, 기존의 어휘정보기반 문법이론 [4,7] 에서의 사전 구조에 한국어의 형태소 처리를 위한 형태소 자질인 morph를 도입하였다. 형태소 자질 morph는 특별히 형태소 분석 및 합성에 있어 필요한 정보 - pos자질은 표제어의 품사정보, anomaly자질은 용언의 변칙정보 - 를 가지고 있다. pos자질은 각 표제어에 대해 품사 분류표의 값을 하나 가지고, anomaly자질의 정보는 변칙이 일어나는 용언의 경우에는 표2의 변칙정보표의 값을 가지며, 나머지 변칙이 일어나지 않는 용언이외의 품사에 해당하는 체언이나 조사 및 등등의 경우에는 정칙의 경우로 보고 "0" 값을 주었다. 그림3은 명사 "집"과 불규칙 동사 "깨닫"에 대한 형태소 분석을 위한 형태소의 자질 정보만을 보인 사전 정보이다.

그림 3. 사전 표제어의 예
Fig. 3. Examples of dictionary entry.

2. 형태소 단위의 품사분류 및 결합정보

1) 형태소 단위의 품사분류

한국어의 어휘에 대한 품사 분류는 학자에 따라 그 분류가 각기 다르게 정의되며 많은 논란이 야기되는 분야이다. 이러한 논란의 대상이 되는 품사 분류를 어떤 특정의 분류에 한정시키지 않고 늘 유동적으로 수정이 가능하게 끔 처리하므로 시스템의 융통성(flexibility)을 높일 수 있게 구축하는 것이 바람직하다.

따라서 본 논문의 품사 분류는 컴퓨터를 이용한 자연어처리라는 점을 감안하여 형태소간의 결합 정보표 및 사전의 구성 그리고 시스템에서의 처리의 효율성을 위해 언어학의 품사론과 다른 각도의 품사를 분류할 필요성을 인식하여 품사의 분류를 계층화하였다. 품사의 계층화는 체언(명사,대명사,수사), 용언(동사,형용사,조용사), 조사, 접사(접두사,접미사), 어미(선어말〈존칭,시제1,시제2,양상1,…〉,어말〈종결어미,접속어미,전성어미〉), 수식언(관형사,부사), 감탄사로 나누고, 각 괄호안의 내용은 다시 세분화되어 계층화하였다. 이러한 계층화의 구성은 형태소간의 결합 정보표에 반영되어 한 어절내에서 좌우 형태소의 구별을 용이하게 해 준다.

계층화한 품사 분류표는 현재 이 시스템에서 구축하고 있는 형태소 시스템의 특징중의 하나로 형태소적인 품사 정보와 형태소간의 결합 정보표를 프로그램적으로 수행함으로 인해, 본 논문에서 정의한 품사 정보에다 더 세분화되거나 변화가 일어나는 경우에 기존의 시스템에서는 일일이 형태소간의 결합 정보표를 수정하였는데 반해, 본 시스템은 프로그램적으로 형태소간의 결합 정보의 값을 구함으로 인해 수정할 필요가 없어서 그 처리 효율이 매우 효과적이다.

2) 형태소간의 결합정보

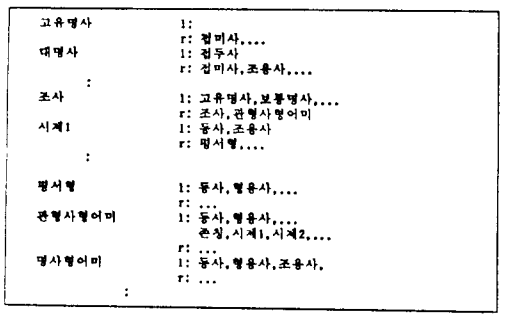
결합정보는 형태소 분석에서 추출된 형태소들을 하나의 어절단위로 구성하는 형태소 합성 단계에서 필요한 정보이다. 즉 형태소 합성을 행함에 있어 어절단위로 형태소들을 결합시키기 위해 한 형태소가 취하는 좌측과 우측에 결합될 수 있는 형태소 유형에 관한 정보를 파악하여 표로 구성한 것이 표1와 같은 결합 정보표이다.

기존 시스템[2]에서 좌우결합정보표를 따로 두고 사전내에 이들에 관한 정보를 일일이 기록함으로 인해 생기는 문제점을 개선시키고자 계층화한 형태소별 품사분류표와 표1에 나타난 형태소간의 결합정보표를 프로그램내부에서 구축하여, 프로그램의 실행과 더불어 한 어절의 처리시에 어절내에서 추출된 형태소들에 대해 좌우에 올 수 있는 형태소을 제한하여 처리함으로써 어절내에 올바른 형태소들의 결합이 이루어지게 하였다. 또한 이들의 처리를 프로그램내부에서 수행하기 때문에 형태소별 품사분류표와 형태소간의 결합정보표의 수정을 필요로 할 시 사전의 항목과는 관계없이 언제든지 쉽게 표 자체를 수정하면 되도록 동적으로 구성하였다.

이처럼 결합정보표는 형태소합성시 한 어절내에서 형태소간의 결합이 올바른가를 찾기위해 이용된다. 즉, 한 문장내에서 추출된 형태소에 대해 어절단위로 형태소 합성을 수행함으로 결합정보를 이용함은 불필요한 어절의 형성을 막아준다. 이는 구문분석의 입력이 되는 형태소분석·합성의 출력결과를 가장 적절한 결과만을 산출함으로 애매성을 최소화해 준다.

표 1. 형태소간의 결합 정보표

Table 1. Union information table of morphemes.



3. 용언의 활용

불규칙 현상의 처리는 사전 구조상에 형태소 자질인 morph내에 용언의 변칙이나 활용 현상을 처리하기 위한 anomaly 자질을 하나 도입하였다. 이 자질을 둠으로 지금껏 몇몇의 시스템에서 변칙이 야기되는 모든 어휘를 변칙된 어휘까지 사전내에 수록하여 사전의 표제어 수를 늘려서 운영하는 방식에서 탈피하였다.[10] 따라서 사전내에 변칙 어휘에 대해서는 원형이 되는 것을 표제어로 두어 처리함으로, 변칙처리 루틴에서 발견된 불규칙이 일어나는 어휘에 대해 원형의 모양으로 복구하여 사전을 검색한다.

용언의 변칙처리를 위해 불규칙 활용을 분류하여[10], anomaly자질내의 정보를 표2과 같이 표기하여 그 처리를 수행하였다. 이 anomaly 자질내에 "0"이 있으면 정칙인 용언을 나타내도록 하였다.

본 연구에서의 불규칙 처리는 표2의 정보를 이용 및 각 현상에 대해 표층어로 부터 현상을 인식하여 표제어를 찾을 수 있는 오토마타[2]를 구성하여

procedural하게 처리하였다.

표 2. 불규칙 및 활용 정보표

Table 2. Information table of irregularity and conjugation.

불규칙 활용	anomaly자질 정보
ㄷ	1
ㄹ	2
ㅂ	3
ㅅ	4
:	:

4. 차트 자료 구조

1) 정의

차트(chart) [13][14] 란 유한개의 노드(node)와 에지(edge)들로 이루어진 유향 그래프(directed graph)이다. 노드는 입력 문장내의 분리점의 역할을 하고, 에지는 부분 구조를 저장하는 역할을 가지며 시작 노드와 도착 노드, 그리고 연관된 정보 구조인 기호를 가지고 있는 자료 구조이다.

이러한 차트 구조는 자연어처리에 있어 주로 파싱(parsing)에서 많이 이용하고 있으나, 본 논문에서는 차트를 형태소 분석 및 합성과정에서 구성되는 음절 또는 음소로써 사전을 검색할 필요가 있는 즉, 하나의 형태소가 될 수 있는 어휘들을 보관하기 위한 bookkeeping기법으로 이용하여 처리하였다. 이렇게 형태소 분석에서 차트를 이용함으로 사전의 검색시에 각 음소의 초중종성 정보 및 차트의 에지 정보로 인해 사전의 검색을 단방향으로만 행함으로 검색 어휘의 수를 줄일 수 있었다.

2) 차트 구성을 위한 경험적인 정보

형태소 분석 및 합성 단계에서의 한 어절에 대한 차트의 구성이 사전의 검색을 최소화할 수 있기 때문에, 본 시스템에서 이용한 차트 구성상의 경험적인 정보는 아래와 같다.

① 한글에서 초성으로 끝나는 형태소는 없다. 즉, 초성으로만 최소의 유의미인 형태소의 단위가 될 수 없다. 따라서 초성으로만으로 사전내의 검색을 할 필요 가 없다. 예를 들면 "집에"에서 초성 "ㅈ"이나 "ㅇ"만으로 형태소 정보가 될 수 없다.

② 중성으로만으로 시작되는 형태소도 없다.

③ 한 글자와 그 다음 글자의 초성과 결합된 형태로 사전의 검색는 할 필요가 없 다. 즉, 글자 + 초성은 하나의 형태소로 이루어지지 않는다. 예를 들면 "집에" 에서 "집+ㅇ"으로 사전의 검색은 무의미하다.

④ 한글의 종성중에서 "ㅁ", "ㄹ", "ㄴ"만은 한 어절의 구성에 있어 구문적인 정보를 가지므로 하나의 형태소 단위로 두고 나머지 종성 자음은 사전을 검색할 필 요가 없다. 여기서의 구문적인 정보라 함은 구문 처리시에 필요한 정보을 나타낸다. 즉, 종성 "ㅁ"은 용언의 어간과 결합하여 명사형 전성어미가 되고, 종성 "ㄹ"은 어간과 결합하여 관형사형 전성어미로 미래를 나타내며, 종성 "ㄴ"은 현재시제를 나타내는 선어말어미와 동사의 어간과 결합하여 과거시제, 형용사의 어간과 결합하여 현재의 사실을 나타내는 관형사형 전성어미를 지닌다. 따라서 이들에 관한 정보는 사전에 구성되어 있어야 하나의 어절로 합성시 그 정보를 이용할 수 있다.

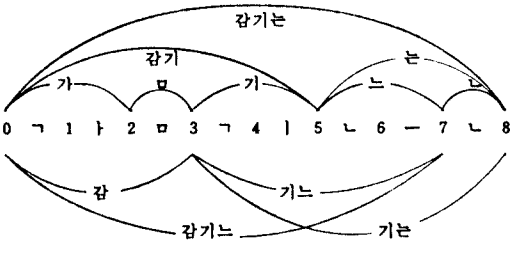
⑤ 추출된 형태소의 리스트에서 에지번호의 연속적인 연결과 접속정보를 이용하여 가능한 형태의 어절만을 구성한다.

이와 같은 경험적인 정보의 이용은 [4] 의 형태소 분석기에서 각 음소들의 차트구조에서 이들 정보를 이용하지 않음으로 인해 생기는 불필요한 사전 검색의 양을 줄였다. 즉, 훨씬 효과적으로 형태소를 찾기 위한 사전 검색의 양을 줄임과 동시에 한 어절내의 모든 형태소들을 추출하였고, 형태론적인 바른 어절을 구성할 수 있었다.

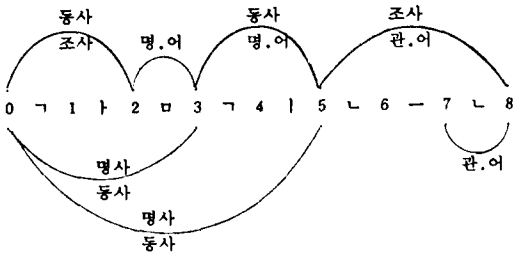
3) 차트 구성 예제

형태소 분석 및 합성 단계에서의 한 어절에 대한 차트 구조를 보면 그림 4와 같이 구성된다.

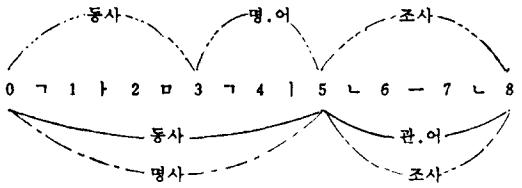
(a) 초기 음소들의 차트 구조



(b) 사전의 검색을 필요로 하는 음소 및 음절의 차트 구조



(c) 최종 사전의 검색에 대한 모든 형태소들의 차트 구조



(d) 결합 정보의 조사후 형성된 결합 가능한 형태소들로 합성된 차트 구조

그림 4. "감기는"에 대한 모든 형태소의 차트 구조 (a,b,c,d)

Fig. 4. Chart structure of all morphemes for "kamgineun(감기는)".

그림3의 (a)는 "감기는"의 어절을 음소들로 분해한 리스트이고, (b)는 사전의 검색을 필요로 하는 결합된 음소들 리스트이고, (c)는 사전내에서 발견된 모든 형태소들의 리스트이고, (d)는 (c)의 리스들로 형태소 단위의 결합을 통해 하나의 어절이 가능한 리스트들이다. 이때 (c)에서 한 어절내의 모든 형태소들에 대해 차트의 에지 정보를 이용하며 또한 형태소간의 결합 정보를 조사하여 형태소 합성을 행한다. 형태소 합성후의 위의 예제 "감기는"에 대한 가능한 형태소들의 결합 결과는 (d)의 그림처럼 "감(동사)" + "기(명사형어미)" + "는(조사)", "감기(명사)" + "는(조사)", "감기(동사)" + "는(관형형어미)"의 어절로 결합하여 출력한다.

5. 형태소 분석 및 합성 알고리즘

형태소의 분석 및 합성기의 구축에서 사용된 사전의 구조는 그림2와 같으며, 아래에 한국어 형태소 분석 및 합성을 위한 알고리즘을 제시하였다.

```
algorithm KoMAS( )
{
  1. 결합정보표 set /* · 품사 분류표와 결합 정보표를 읽어들여 형태소간의 결합 가능성을 나타내는 표(table)를 만듬
                     · 입력되는 품사분류표와 결합정보표의 내용이 바뀔 때마다 표의 값이 자동적으로 구축됨 */

  while( 한 문장 != EOF ) {
    2. 하나의 어절을 가져와 음소별 리스트를 만듬
       /* · 하나의 어절을 음소단위로 분해함
          · 각 음소에 대해 초중종성 및 영어, 숫자 정보를 넣음
          · 정의되지 않은 문자는 그냥 돌려줌 */

    3. 음소 리스트로 불규칙 현상을 처리하여 차트에 적재함

    4. 한 어절에 대한 음소 리스트에서 형태소를 구성하면서 분석
       /* · 음소 리스트가 한글, 영어, 숫자 인지를 검사하여
          · 음소 리스트에서 초중종성 정보를 검사하면서 사전을 검색할 필 요가 있는 음소와 음소들을 연결시켜 하나의 음절로 형성하여 word_string을 만듬
          · 음절로 구성된 word_string으로 사전을 검색
          · 사전에 있는 음절들만을 차트에 적재시킴 */

    5. 가능한 형태소의 리스트에서 결합 정보에 의해 결합하여 어절 단위로 출력
       /* · 한 어절내에서 추출된 형태소들 중에서 결합 가능한 형태소들을 결합하여 모든 경우의 어절 단위로 묶어 출력함
          · 연속적인 에지번호의 연결이 되도록 형태소간의 결합정보를 조사하여 합성시킴
          · 이때 각 형태소 사전의 중요 자질 정보를 통합시켜 하나의 어절로 구성 */
  }
}
```

Ⅳ. 실험 결과 및 시스템의 특성

이 장에서는 몇 어절에 대해 시스템 [4] 와 비교한 실험 결과를 제시하고, 실험 자료로 "어린이 보호 헌장"의 처리 결과와 본 시스템의 특성에 대해 기술하였다.

1. 실험 결과

1) 실험 비교

시스템 [4] 는 모든 가능한 형태소의 조합으로 분리해 내고, 분리된 각 형태소에 대하여 다품사를 포함한 자질 정보 및 다중에지정보를 구문 분석단계로 넘겨준다. 따라서 구문 분석단계에서 하나의 어절로 구성하여 구문처리를 수행한다. 또한 이 시스템은 용언의 불규칙 처리를 하지 못하므로 사전에 불규칙 어휘의 모든 경우를 등재해야 하고, 가능한 모든 조사(으로부터, 에서는,…)의 경우도 역시 사전에 등재하며, "명사+조용사"(사랑+하)의 경우도 사전에 각각 등재해야 함으로 인해 많은 양의 어휘를 사전에 등록해야 한다.

반면 본 시스템은 형태소 합성 루틴에서 결합정보를 이용하여 하나의 어절로 묶어 구문분석기의 입력이 되게끔 하였다. 또한 불규칙 처리, 조사와 조사의 결합 처리 및 품사정의표에 조용사의 도입으로 사전에 등재되는 어휘의 양을 줄였다.

표 3. 문장 "나는 새를 보다"와 어절 "감기는"에 대해 비교 분석표

Table 3. Table of Comparison Analysis for "kamgineun" and "naneun saereul boda".

어절 / 분석내용 / 시스템	나는		새를		보다		감기는	
	A	B	A	B	A	B	A	B
사전의 검색을 필요로 하는 차트수	15	6	15	6	10	3	36	12
사전 검색후의 차트수	8	4	8	5	5	5	16	13
합성후의 어절수	2	2	1	1	1	1	3	3

(여기서 [4] 를 시스템 A라 하고, 본 시스템을 B라 한다.)

표3의 비교 분석표에서와 같이 차트자료구조를 이용한 두 시스템에서, 한국어의 특성을 반영하여 경험적인 정보를 이용한 본 시스템이 훨씬 효율적임을 볼 수 있다. 즉, 본 시스템의 처리에서는 사전 검색에 필요한 차트에 유지되는 음절수, 사전 검색후의 형태소들 수, 가능한 모든 올바른 어절의 합성을 모두를 시스템 [4] 보다 효과적으로 얻을 수 있었다.

2) 실험 자료

"어린이 보호 헌장"의 각 어휘에 대해 사전을 구성[12] 하고, 그 처리를 수행한 결과는 아래와 같다.

① 사전내에 등재된 품사별 음절당 수

품사 \ 음절	1	2	3	4	총수
체언	25	56	7	4	92
용언	15	23	10	1	49
조사	20	6	x	x	26
접두사	12	x	x	x	12
접미사	11	x	x	x	11
어미	17	5	x	x	22
관형사	3	1	x	x	4
부사	2	5	x	x	7
감탄사	1	x	x	x	1
총 수	106	96	17	5	224

어린이 보호 헌장의 어절수가 모두 155개이지만 중복적인 어절을 제외한 어절의 수는 127개이다. 여기서 사전에 등재된 수가 224개 인것은 한 어휘가 여러 개의 품사를 지닐 수 있음을 의미한다.

② 사전내에 등재된 표제어에서 음절당 다품사를 지니는 수
(품사적인 중의성을 지니는 형태소 수)

음절수 \ 품사수	1	2	3	4	5
1음절 형태소	20	13	7	6	3
2음절 형태소	83	5	1	0	0
3음절 형태소	15	1	0	0	0
4음절 형태소	5	0	0	0	0

③ 음절당 어절수와 각 어절당 형태소 분석에서 사전검색에 필요한 수와 사전내에서 검색된 형태소수를 음절별로 비교

음절	어절수 (총: 127개)	사전검색이 필요한 음절수(A)		사전내에서 검색된 형태소 음절 수(B)	
		합계	평균	합계	평균
1	3	5	1.67	11	3.66
2	21	113	5.38	104	4.95
3	63	653	10.37	406	6.44
4	30	439	14.63	229	7.63
5	7	157	22.43	64	9.14
6	2	57	28.50	22	11.00
7	1	35	35.00	5	5.00

여기서 A와 B의 차이는 각 음절에서 다품사를 가지는 형태소가 있는 경우와 하나의 형태소를 찾기 위해, 음소들의 연결로 음절을 만들어 사전을 검색하기 위한 차트을 구성할 때 그만큼 많은 양의 차트의 형성이 있음을 나타낸다.

④ 형태소 분석 및 합성후의 출력된 어절의 구성수

어절의 구성 \ 음절	1	2	3	4	5	6	7
단독 명사	3	3	1	1	x	x	x
관형사	2	1	x	x	x	x	x
수사	1	x	x	x	x	x	x
접두사	3	x	x	x	x	x	x
접미사	2	x	x	x	x	x	x
단독 동사	x	x	1	x	x	x	x
체언 + 조사	x	3	35	13	4	2	1
체언+접미사	x	1	x	1	x	x	x
체언+접미사+조사	x	x	x	2	x	x	x
체언+접미사+어미	x	x	x	1	x	x	x
체언+조용사+어미	x	1	1	8	2	x	x
용언+어미	1	13	25	7	1	x	x
용언+어미+조사	x	1	1	x	x	x	x
접두사+용언+어미	x	x	2	x	x	x	x
부사	x	3	x	x	x	x	x
총 수	11	26	66	33	7	2	1

출력된 어절의 구성수를 보면 146개인데 이는 19개의 어절에서 중의성이 있는 것으로, 각 음절에서 어절의 합성후 출력된 어절의 구성수가 2개 이상인 경우는 중의성이 있는 어절로 형태소 처리시에 해결되는 것이 아니고, 구문 의미단계에서 처리되어야 하므로 본 시스템의 형태소 분석 합성단계에서는 모든 가능한 어절형성을 출력하였다.

2. 시스템의 특징

형태소 분석기의 성능을 판단하는 가장 중요한 기준은 모든 주어진 어절에 대하여 가능한 모든 경우의 형태소 결과를 생성하는가(completeness)하는 점과 분석 오류가 발생하는가(soundness)하는 점, 그리고 알고리즘의 효율성이다. 자연어처리 시스템에서 알고리즘의 완전성 여부는 사전 표제어의 수와 사전 정보의 완전성에 의존하는데, 본 시스템에서는 하나의 통합된 사전을 구성하고 사전의 표제어는 형태소 단위로 수록하였으며, 사전의 구조에 형태소 분석을 위한 형태론적인 자질을 추가하여 보다 효율적인 처리를 하였다.

또한 구문분석의 단위가 한국어의 특성상 어절이라는 점을 감안하여 추출된 형태소들을 하나의 어절로 합성하여 한 어절내의 모든 구문·의미 정보를 결합하여 생성시킴으로 구문 분석기에서 처리해야 할 일의 양을 많이 줄어들 수 있도록 하였다. 이러한 처리에 있어 본 시스템의 특성은 아래와 같다.

1) 입력 문장에서 모든 형태소를 추출하기 위해 차트자료구조를 이용하여 상향식(bottom-up)방법으로 구현하였다.
2) 품사정의표와 결합정보표는 표의 내용이 바뀔 때 프로그램의 실행과 더불어 자동적으로 구축된다.
3) 사전에 형태소 처리기에 필요한 새로운 "morph"라는 자질을 도입하였다.
4) 기억 장치의 크기를 고려하여 최소 유의미인 형태소 단위로 처리한 후 형태소간의 결합 정보표를 이용하여 한 어절단위로 처리한다. 이렇게 첨용, 활용 관계에 대한 처리를 한 결과를 통사 분석의 입력으로 줌으로 통사 분석기의 부담을 줄이도록 하였다.
5) 처리에 있어 효율적인 사전의 검색을 위한 특성은 3-4절의 차트 구성을 위한 경험적인 정보와 같다.

Ⅴ. 결론

형태소 처리는 하나의 어절을 형성하는 모든 경우의 형태소들을 분석하고, 분석된 형태소들을 다시 결합하여 하나의 어절을 만드는 과정을 말한다. 따라서 본 논문에서는 형태소의 효과적인 분석을 위해 사전의 구조에 morph라는 형태소자질을 도입하였다. 이러한 사전의 구조와 차트자료구조 및 경험적인 정보를 이용하여 효과적으로 모든 가능한 형태소들을 분석하였고, 어절내의 모든 형태소의 추출에 있어 사전 검색의 양을 최소화할 수 있었다. 또한 형태소의 분석 과정에서 불규칙 현상을 처리하며, 추출된 형태소

들을 결합시키기 위한 품사분류와 형태소간의 결합 정보를 제시하여 모든 가능한 어절을 생성하였다. 어절의 생성에 있어 추출된 모든 경우의 형태소들은 그들내의 중요 자질 성분끼리는 통합이라는 기제를 통해 한 어절내의 모든 형태소들의 구문·의미 정보 자질들을 모두 통합시켰다. 이와 같은 처리는 구문 분석기에서 처리할 수 없는 형태론적인 모호성을 줄임과 동시에 구문분석기의 처리 부담을 줄일 수 있다.

그러나 본 시스템이 음소 단위로 결합하여 음절을 형성하여 처리함으로 인해, 형태소를 찾기 위한 사전의 검색에 있어 여전히 많은 양의 불필요한 음절들로 차트를 구성한다. 따라서 각 형태소를 이루는 음절들의 정보를 이용한다면 사전의 검색을 위한 차트에 적재되는 음절의 양 및 처리의 효율성등 보다 나은 처리가 될 것으로 보고 현재 연구수행중에 있다.

본 시스템은 자연어의 여러 응용 시스템에 이용될 수 있으며, 특히 한글 철자교정기나 기계번역시스템에 직접 응용될 수 있을 것이다.

參 考 文 獻

[1] 권 혁철, "자연언어 처리 동향", 한국정보과학회, 인공지능연구회, 1991.

[2] 김 성용, TABULAR PARSING 방법과 접속 정보를 이용한 한국어 형태소 분석기, 한국과학기술원 전산학과 석사학위논문, 1986.

[3] 김 혜리, 한국어 어휘부-기반 문법 정립을 위한 기초연구, 서울대 언어학과 석사학위논문, 1988.

[4] 김 호영, 개선된 차트 파싱 방법을 이용한 한국어 해석기의 구현, 경북대 컴 퓨터공학과 석사학위논문, 1990.

[5] 남 기심, 고 영근, 표준 국어 문법론, 탑출판사, 1985.

[6] 박 원철, 형태론적 연산으로서의 통합, 서울대 언어학과 석사학위논문, 1989.

[7] 서 영훈, 의미 정보를 이용하는 중심어 주도의 한국어 파싱, 서울대 컴퓨터공 학과 박사학위논문, 1991.

[8] 손 우형, 한국어 기계 번역을 위한 형태소 분석에 관한 연구, 한국과학기술원 전산학과 석사학위논문, 1986.

[9] 이 상조, 한국어 자연어 인터페이스를 위한 사전 구성에 관한 연구, 한국전자 통신연구소 위탁과제 최종연구보고서, 경북대 전자기술연구소, 1991.

[10] 조 규빈, 하이라이트 국문법, 지학사, 1984.

[11] 최 형석, 이 주근, "자연어 어절 처리 알고리즘", 한국정보과학회, '84 가을 학술발표논문집, 1984.

[12] 민중서림 편집국, 엣센스 국어 사전, 민중서림, 1992.

[13] James Allen, *Natural Language Understanding*. The Benjamin / Cummings Publishing Company Inc., 1987.

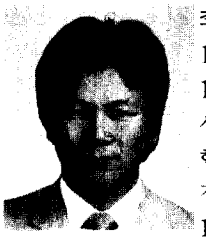
[14] Winograd, *Language as a Cognitive Process*. vol.1: Syntax, 1883.

[15] S. M. Shieber, et al, "A Compilation of papers on Unification-based Grammar Formalisms", *CLSI-86-48*. 1986.

著 者 紹 介

金 貞 海(正會員)

1988年 계명대학교 전자계산학과(학사). 1990年 경북대학교 전자계산기공학과(석사). 1992年 경북대학교 컴퓨터공학과(박사과정 수료). 1992年 ~ 현재 상지전문대학 전산정보처리과 전임강사로 재직중. 주관심 분야는 기계번역, 자연어처리 등임.



李 相 祚(正會員)

1974年 경북대학교 수학과(학사). 1976年 한국과학원 전산학과(석사). 1993年 서울대학교 컴퓨터공학과(공학박사). 1991年 1993年 경북대학교 전자계산소 소장 역임. 1976年 ~ 현재 경북대학교 공과대학 컴퓨터공학과 교수로 재직중. 주관심분야는 프로그래밍언어, 기계번역, 자연어처리 등임.