moHANA: 다차원 해석 사전을 기반으로 한 한국어 형태소 분석기

서승현* 강인호 김재동

주식회사 워드워즈
*shsuh@wordwords.co.kr
카네기멜론대학교/언어공학연구소
ihkang97@cs.cmu.edu, jdkim@cs.cmu.edu

moHANA: Morphological Hangul Analyzer using Multi-Dimensional Analysis Dictionary

SeungHyeon Seo*, In-Ho Kang, JaeDong Kim
* WordWords, Inc
Carnegie Mellon University/Language Technologies Institute

요 약

본 연구는 국어의 모든 언어적 특성을 기술하고 이를 실제 형태소 분석에 적용할 수 있도록 다차원 해석 사전을 이용하는 형태소 분석 시스템인 moHANA(**Mo**rphological **Han**gul **A**nalyzer)에 관한 연구이다. moHANA의 해석 사전은 태그정보 사전, 어휘 사전 그리고 문법 사전으로 구성된다. 태그정보 사전은 기존 형태소 해석기의 일차원적인 품사 정보와 달리 어류 태그정보, 형태적 정보, 통사적 정보, 의미적 정보 및 화용 정보의 5 차원 벡터 정보로 작성된다. 어휘 사전은 어휘와 그 어휘가 가질 수 있는 태그정보를 우선 순위에 기반하여 순서열로 가지며, 문법 사전은 특수 문법 연산자를 이용하여 태그정보 사전에 정의된 각각의 태그가 연결 가능한지 여부를 규정하는 문법이 구축되어 있다. 형태소가 가지는 태그정보를 다차원으로 정의하고 이에 따른 문법 규칙의 표현을 통해 보다 자세한 형태소 분석 및 새로운 형태소 태그의 삽입과 삭제의 용이함을 얻을 수 있다.

1. 서론

형태소 분석이라 함은, 주어진 입력 문자열을 대상으로 형태소 단위로 분리해 내고, 각각의 형태소의 원형을 복원하고 품사를 결정해 주는 과정을 의미한다[1]. 형태소 분석을 위해서 사용하는 해석 사전은 일반적으로 형태소 목록을 가지고 있는 어휘 사전 그리고 어휘가 가질 수 있는 태그정보 사전 그리고 태그 간의 연결 가능 여부를 표현하는 문법 사전으로 구성된다. 이러한 해석 사전을 기반으로 입력 문자열에서 발생 가능한 모든 형태소 열에 대해서 문법 사전에 기술된 연결 정보에 부합하는 형태소 열을 찾아낸다. 이러한 형태소 분석은 기계 번역, 정보 검색, 정보 추출 등의 자연언어처리 기술의 기반 기술로써 최소한의 의미를 가지는 형태소를 쉽게 파악할 수 있다.

모든 가능한 언어적 특성에 맞춰 다른 여러 자연언어처리기술에 응용되기 위해서는 국어의 어절 및 문장의 형태소를 제대로 표현할 수 있어야 한다. 그러나, 기존의 형태소 분석기에서는 단순하게 일차원적으로 형태소가 가지는 언어적 자질들을 기술함으로써, 여러 가지 서로 다른 언어적 자질들이 분리되지 못하고

혼용되어 있다[2][3]. 또한 음운, 통사, 화용적인 자질들과 같은 일반적인 언어적 자질 및 특성을 표현하기에는 부적절한 구조를 가지고 있거나, 이러한 자질들을 표현할 수 없는 경우들이 많다. 그래서 새로운 형태소 부류의 특성을 표현할 수 있게 하기 위해서는 형태소 분석기의 많은 부분을 수정해야 하거나 경우에 따라서는 형태소 분석기 전체 구조를 바꾸어야 하는 등의 문제점이 발생하는 경우가 있다.

이에 본 연구는 모든 자연언어 처리의 기본이라 할 수 있는 형태소 분석기가 일차적으로 형태소 분석을 정확하게 할 수 있게 할 뿐만 아니라, 형태소 분석이 된 형태소들을 국어의 특성에 맞게 활용할 수 있도록 하기 위한 다차원 해석 사전을 기반으로 한 형태소 분석기, moHANA(**Mo**rphological **Han**gul **A**nalyzer)를 소개한다.

moHANA의 해석 사전도 기존 해석기와 유사하게 태그정보 사전 어휘 사전 및 문법 사전을 기본으로 구성되며 부가적으로 분석기의 속도 향상과 사용의 편리를 위해서 기분석 사전과 사용자 사전을 포함한다. 형태소들의 자질 정보 혹은 부류를 나타내는 태그정보 사전은 어류 태그정보, 형태적 정보, 통사적 정보, 의미적 정보 및 화용 정보의 5차원의 벡터 정보가 부가되어 정의된다. 문법 사전은 이러한 5차원의 태그정보를 활용할 수 있도록 특수 문법 연산자를 이용하여 태그정보 사전에 정의된 각각의 품사가 연결가능한지 여부를 기술한다.

본 논문의 구성은 먼저 기존의 형태소 분석 애매성을 해결하기 위해서 사용해 온 자질 정보에 대해서 간략히 알아본 뒤, moHANA에서 사용하는 5차원 정보에 대한 정의 및 활용에 대해서 설명한다. 그리고 이러한 사전 구조를 가진 moHANA의 각각의 언어적 정보들에 따라 형태소 분석시의 정확률에 어느 정도의 영향을 미치는가를 실험하여, moHANA가 가지고 있는 다차원 해석 사전의 구조가 한국어 형태소 분석을 함에 정확률을 높이는 데에 효율적임을 보인다.

2. 관련 연구

형태소 분석기는 일반적으로 어절이나 문장과 같이 주어진 입력을 대상으로 해석 가능한 모든 형태소열을 찾아주는 것을 목적으로, 크게 두 가지 용도로 구분할 수 있다. 첫번째는 품사 태거의 입력을 제공하는 용도이다. 품사 태거는 형태소 분석기가 출력한 형태소열 중 가장 적합한 결과를 찾아내어 구문 해석이나 의미 해석을 위한 기본 정보 단위를 제공하는 용도이다[4][5]. 그리고 두 번째는 정보 검색을 위해 문서나 사용자의 질의에서 색인어를 추출하여 제공하는 용도이다[6][7]. 여기에서 색인어 추출기는 일종의 단어 단위 품사 태거를 이용한 것과 가장 적합한 결과를 최상위 결과로 제공하며 또한 입력 단어에서 추출 가능한 모든 색인어 후보를 제공한다. 구문 해석이나 의미 해석 용도에서는 일반적으로 문장 단위로 입력이 들어오는 반면, 색인어 추출에서는 사용자의 검색 질의어 분석을 목적으로 하기 때문에 단어 단위의 입력이 들어온다고 가정한다. 대부분의 경우 품사 태거의 입력 용도에서는 올바른 형태소 분석열을 포함하여 품사 태거의 결과물의 정확성을 향상시키는 것이 주된 관점이며 색인어 용도에서는 빠른 시간 안에 결과를 제공하는 것이 목적이다.

형태소 분석 과정에서의 애매성을 해결하기 위해서 품사간의 연결 여부를 사용하는 연구가 있어 왔다. 그 중 한 연구는 2차원 형태의 테이블을 작성하여 연속된 두 품사의 연결 가능여부로 애매성을 해소하였다[1]. 특히 품사 태거를 활용하는 경우 형태소 분석을 단순화하고 품사 태깅에서 세 단어 이상의 품사열이나 어휘 정보를 포함하는 품사열을 품사태깅 확률모델에 반영하여 애매성을 해결했다[4][5]. 또한 형태소가 가질 수 있는 품사를 좌접속정보, 우접속 정보로 명시하여 앞에 나타나는 형태소와 연결 여부를 조사할 때 사용하는 정보와 오른쪽에 나타나는 형태소와 조사할 때 사용하는 정보를 차별화하여 사용하는 연구가 있었다[10]. 품사 간의 연결 중 예외적인 결합 형태에 대해서는 품사를 더욱 세분화하거나 예외 리스트를 작성하여 추가 정보를 고려하는 예외적인 경우를 처리한다[2]. 또는 후처리나 기분석 사전을 활용하여 오류나 미등록어의 문제를 해결했다[8][9].

moHANA에서는 형태소 분석 애매성과 관련되는 정보들을 5차원의 벡터 정보로 표현하여 별도의 예외 리스트 작성이나 분석 프로그램의 수정 없이 해석 사전에 바로 반영하는 형태로 관련 작업을 최소화한다. 아울러 5차원의 벡터 정보를 형태소 분석기에 적용하고자 하는 최종 작업 시 필요한 품사 분류로 매핑하는 것이 가능하며, 형태소 분석기 사용 용도에 따라서 5차원의 정보 중 일부 정보를 첨가하거나 삭제함으로써 속도와 해석 결과의 정확성을 조절할 수 있다.

3. 다차원 해석 사전의 구성

그림 1은 moHANA의 해석 사전의 구성을 보여준다. 시스템에서 사용하는 형태소의 태그 및 언어적 자질 정보를 정의하는 태그정보 사전, 형태소와 그 형태소가 가질 수 있는 태그정보들을 가리키는 어휘 사전, 그리고 품사 간의 연결 가능 여부를 표현하는 문법 사전, 그리고 특정 영역의 사용자 표현이나 예외적인 형태를 수용하는 사용자 사전 그리고 자주 나타나는 어절에 대한 기분석 결과를 가지는 기분석 사전을 포함한다. 여기서 사용자 사전은 품사가 지정되지 않은 경우 기본 품사로 할당되며, 기분석 사전의 기분석

항목은 연결 문법에 맞지 않는 해석도 삽입이 가능하다.

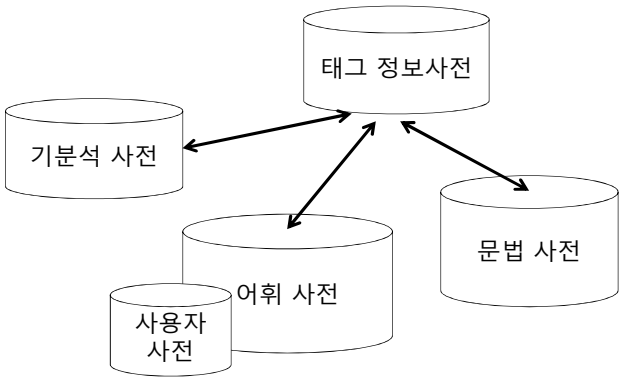

그림 1 moHANA 해석 사전의 구성

3.1 태그정보 사전

언어적인 특성에 따른 형태소의 부류를 나타내는 각각의 태그를 나타내기 위해, moHANA의 태그정보 사전은 어류 태그(word class tag) 정보, 형태적(morphological) 정보, 통사적(syntactical) 정보, 의미적(semantic) 정보 및 화용(pragmatic) 정보의 5가지 정보, 즉 5차원의 벡터로 정의되어 구성된다.

특히, 본 연구에 따라 새롭게 제안하는 '어류 태그(word class tag)'라고 하는 용어는 일반 언어학에서 정의내린 품사(part of speech)와는 다른 의미를 갖고 있다. 즉, 본 연구에서 '어류 태그'라는 것은 언어학적인 시각에서는 분류되지 못하는 단어 부류들이지만, 이하에서 더욱 상세히 설명하는 바와 같이, 전산상에서 형태소 분석을 용이하게 하기 위해 필요한 단어 부류 및 그 속성들을 정의해 놓은 것이다. 예를 들어 moHANA에서는 '한글자 명사'를 하나의 '어류 태그'로 정의하여 처리하고 있다. 그러나 일반언어학에서 '한글자 명사'들에 대해서는 하나의 품사로 규정하지 않으므로, 이에 따라 다른 일반 형태소 분석기의 품사 태그도 '한글자 명사'에 다른 명사와 구별하지 않고 처리하고 있다. 그러나, moHANA에서는 형태소 분석의 정확률을 높이기 위해서 '한글자 명사'를 위해 하나의 태그를 부여하여 처리하고 있다[1].

이에 본 연구에서는 형태소들의 분류를 일반 언어학의 품사와 구별하기 위해서 '어류 태그'라는 용어를 사용한다. 따라서 본 형태소 분석기의 어류 태그는 필요에 의해서 새로운 단어들의 부류를 자유롭게 확장하거나 축소하기 위해서 이용될 수 있다.

'어류 태그'를 제외한 4가지 정보인 형태적(음운론 포함), 통사적, 의미적, 화용적 정보 구성은 일반언어학의 연구 대상인 음운론, 형태론, 통사론, 의미론, 화용론의 정보를 모두 반영할 수 있도록 한 정보들의 구성이다[11][12]. 이중에 형태적 정보에는 형태음운적 정보(예를 들어, 마지막 음절의 종성 유무 {fc})도 함께 처리하였다. 이전의 방식들은 품사에 구문 및 의미 정보를 결합하여 일차원적으로 확장하였으나, 본 연구에 따르면, 품사에 해당하는 어류 태그를 최소한으로 유지하고 통사적인 정보와 의미 정보를 형태와 통계 정보처럼 5차원 벡터에서 별도의 차원으로 정의한다. 이는 총체적으로 품사(어류태그), 구문, 의미, 화용 정보들을 능률적으로 확장할 수 있는 이점을 제공한다.

이를 구체적인 예를 들어 설명하면 다음과 같다.

```
<명사 n>
  뉴딜 {ncn}{fc}{}{eco}{}
  히스라불로토프 {nq}{fc}{}{per:pol}{}
  객 {nc_one}{fc}{}{}{noidx}

<동사 pv>
  무서워하 {pv}{ir_yeo}{}{}{}
  열리 {pv}{rg}{}{}{}
```

즉, 위의 명사를 표현하는 예에 있어서, '뉴딜'과 '히스라불로토프', '객'이라고 하는 명사를 표현하기 위한 5가지 정보가 표현되어 있는데, 좌측부터 각각 어류 태그정보, 형태적 정보, 통사적 정보, 의미적 정보 및 화용 정보를 나타낸다. 이처럼, 단순히 기본 형태소 목록 또는 품사 간의 연결 가능 여부를 규정하고 있는 종래의 해석 사전과는 달리, 본 연구에 따르면, 소정의 형태소와 관련하여, 그 형태소를 어류 태그, 형태적, 통사적, 의미적 그리고 화용의 5가지의 정보 벡터로 구분하여 다차원 해석 사전의 데이터베이스를 구축한다.

좀 더 구체적으로 설명하면, 위의 기술 내용 중, <명사> 중에 '뉴딜'과 '히스라불로토프', '객'의 어류 태그는 각각 {ncn}, {nq}, 그리고 {nc_one}이며, {ncn}은 일반명사를, {nq}는 고유명사를, {nc_one}은 한글자 명사임을 표시한다. 여기서, {fc}는 final consonant의 약자로 명사의 종성이 다음의 형태소와 결합에 영향을 미치는 어류태그들에 대해서만 이 정보를 이용한다. 즉, 형태소의 마지막 음절에 따라서 'y'(유종성), 'n'(무종성), 'l'(ㄹ 종성),

[1] '명사' 중에 한글자인 명사들을 moHANA의 '형태적 정보'에서 처리할 수도 있으나, 문법을 기술함에, '한글자 명사'를 일반 명사와 같이 처리하게 되면, '한글자 명사'들이 일반 명사들과 달리 형태소간 연결 특성이 달라서 문법을 기술함에 더 많은 예외적인 규칙을 기술해야 하고 문법 기술이 복잡하기 되어, 하나의 어류태그로 분류해 내어 처리한 것이다.

‘c’(don’t care) 등의 값으로 시스템이 자동으로 구체화한다. 예제에서는 ‘뉴딜’은 ‘1’로 ‘히스라불로토프’는 ‘n’으로 구체화된다.

{eco}와 {per:pol}은 의미적 정보로서 {eco}는 ‘뉴딜’이 ‘경제’와 관련된 것이며, ‘히스라불로토프’는 사람 고유명사 중에 ‘정치’와 관련된 사람임을 정의해 놓은 것이다.

특히, 한글자 명사인 ‘객’의 경우, 화용적 정보에 {noidx}가 있는데, 이 정보는 실제 데이터에서 많이 쓰이느냐 혹은 색인어(index word)로 가치가 있느냐 하는 정보에 대한 것이다. ‘객’의 경우, 복합명사 분해시 ‘객’을 따로 분석해 내게 되면 오분석이 많으며, 실제 ‘객’으로 형태소 분석되어 쓰이는 일이 많지 않으므로, 이러한 한글자 명사들에 {noidx} 정보를 주었다. 반면에 한글자 명사인 ‘핵’과 같은 경우는 화용적 정보에 {idx}를 주어서 형태소 분석시 이 정보를 활용하게 된다. 이와 같이 “색인어인가 아닌가” 하는 문제는 순수한 언어적인 특성이 아닌, 말뭉치에서 한글자 명사들이 형태소 분석이 되어야 하느냐에 따라 자질들이 주어진 것인 실제 말뭉치에서 언어 사용자들의 직관에 의해 판단되는 언어적 현상에 대해 자질로서 분류하여 기술한 것이다. 이와 같이 차별된 화용 정보를 통해 분석 결과로 나타내게 하거나 형태소 분석기가 출력하는 결과 순서를 조정할 수 있다.

한국어 형태소 분석기는 복합명사 분해를 얼마나 잘 하는가도 형태소 분석기의 성능과 관련이 되어 있다. moHANA에서는 복합명사 분석의 정확률을 높이기 위해서 명사들의 복합명사 형성에 제약을 가지는 명사들을 분류해 내어 처리하고 있다. 즉, 한국어 명사의 형태적인 특성을 살펴보면, 모든 명사가 복합명사를 형성하는 것은 아니다. 오히려 복합명사 형성에 결합하지 않는 명사 부류들 때문에 복합명사 형태소를 분석할 때, 형태소 분석의 오분석을 만들어 내는 경우가 더 많다. 이와 같은 명사들에 대해 사전에서 이와 같은 특성을 나타내는 명사들의 형태적 정보에 복합명사 구성에 제약을 보이는 명사임을 표시해 줌으로써, 복합명사 분해의 정확률을 높이도록 하였다.

(1) 학교등에 학교_{ncn}+등_{nfix}+에_{j}
(2) 출판사의 출판사_{ncn}+의_{j}
출판_{ncp}+사_{nfix}+의_{j}

위의 분석은 moHANA에서 출력한 형태소 분석 결과로서[2], ‘등에’에 복합명사와 제약 정보를 줌으로써 ‘학교등에’에서 ‘등에’가 일반명사로 분석되는 ‘학교+등에(명사)’와 같이 분석되지 않는다. ‘출판사의’의 경우도, ‘사의’에 복합명사 형성 제약 정보를 준 것이므로, ‘출판+사의(명사)’로 분석되지 않는다.

<동사>의 경우, ‘무서워하’와 ‘열리’는 어류태그가 {pv}이며 이는 일반적인 동사를 나타내며, {ir_yeo}와 {rg}는 형태적 정보로 {ir_yeo}는 여불규칙 동사를, {rg}는 규칙 동사라는 정보를 기술한 것이다. {tra}와 {intra}는 용언의 통사 정보로서, {tra}는 타동사를 {intra}는 자동사라는 통사적 정보를 입력한 것이다. 이와 같이 각 어류 태그 및 각각의 형태, 통사, 의미, 화용 정보는 각 단어들 부류를 나눌 때의 기준과 속성에 따라 자의적으로 정의할 수 있다.

(3) 무서워하 {pv}{ir_yeo}{tra}{}{}
(4) 무서워하 {pv}{ir_yeo}{intra}{}{}

단, 어류 태그, 의미적 정보 등에 사용되는 각각의 약칭(예컨대, ncn, eco, pol, rg 등)은 사용자가 자신의 편의에 맞게 임의로 정하여 사용할 수 있는 것으로서, 본 연구는 이러한 개개의 약칭의 종류/형태에 제한되지 않는다는 점에 유의하여야 한다.

3.2 어휘 사전

어휘 사전은 형태소와 그 형태소가 가질 수 있는 품사열(태그열)의 리스트로 구성된다. 형태소는 일반적으로 의미를 가지는 최소한의 단위이다. 그러나 사용 영역에 따라서 의미를 가지는 최소한의 단위의 변화가 필요하다. 예컨대, 영화 정보를 검색할 수 있는 서비스에서는 영화명이 비록 여러 단어로 구성되어 있더라도 하나의 의미를 가지는 최소한의 단위로 파악되어야 정확한 검색 결과를 얻을 수 있다. 그러나, 종래의 형태소 분석에 따르면, 하나의 의미로 파악하는 것이 아니라 최소한의 단위에 대하여 형태소 분석을 수행하기 때문에, 사용자가 원하는 정확한 결과를 도출해낼 수가 없는 문제가 있다.

예를 들어 영화 제목이 포함된 문장 “나는 바람과 함께 사라지다가 좋아.”를 대상으로 기존 방식의 형태소 단위를 적용할 경우 ‘사라지다가’가 ‘사라+지다가’로 분석될 수 있어, 가능한 어떠한 형태소들의 결합도 “바람과 함께 사라지다”를 얻어낼 수 없다. 이에 따라, “바람과 함께 사라지다”를 영화명으로 가지고 있어도 검색할 수 없게 된다. moHANA에서는 기본적인 단위의 형태소를 이용하여 어휘 사전을 구축하는 것 외에, 여러 형태소로 이루어지는 단어들을 하나의 의미를 갖는 것으로 규정해 놓고 소정의 시스템이나 서비스에서 정의하는 의미 단위를 반영하는 형태로 입력 가능하게 한다. 즉 하나의 형태소가 시스템이나 서비스에서 정의하는 의미 단위를 반영할 수 있도록 포함하는

[2] 5 차원의 태그정보 중에서 1 차원의 태그 값만 보였다.

단어의 개수에 제약이 없다. 따라서, 본 연구에 따른 다차원 해석 사전의 어휘 사전에는 “바람과 함께 사라지다”가 하나의 형태소로서 등록된다. 아울러 태그정보에 필요한 값을 삽입하는 형태로 영화명의 주위에 자주 나타나는 정보를 형태소 해석 과정에서 사용할 수가 있게 된다.

한편, 각각의 형태소는 하나 이상의 관련 품사 정보를 가진다. 예를 들어 형태소 ‘가’는 명사, 조사 그리고 접미사와 관련된 5차원 값의 품사 정보를 가진다. 여기서의 품사열은 기술 순서에 따라서 분석 과정에 사용되는 순서를 뜻한다. 즉 ‘가’를 해석하기 위해 명사일 경우를 먼저 가정해서 살펴 보고 분석이 실패할 경우 조사 그리고 접미사의 순으로 해석을 시도한다. 본 연구에서는 이러한 관련 품사의 순서를 두 가지 레벨의 규칙을 이용하여 정의한다. 즉 품사들간의 우선 순위, 그리고 각 어휘별 품사 우선 순위 규칙을 이용한다. 한 형태소의 품사열 순위는 특정 어휘인 경우 사전 작성된 품사 선호 순위에 따라서 결정되며, 그 외의 어휘인 경우는 일반적인 품사들간의 우선 순위에 따라서 결정된다.

3.3 문법 사전

moHANA에서는 5차원의 벡터로 구분하여 구축되어 있는 기본 해석 사전을 활용하기 위하여, 그에 상응하는 형태로 형태소 분석을 위한 연결 문법을 구축한다. 즉, 태그 항목 간의 연결 여부 및 강도를 기술한다. 형태소들간의 결합 강도 여부를 계층적으로 표현할 수 있어서, 띄어쓰기 오류(예, 안먹는, 못가는 등등)처럼 국어문법에서는 비문법적인 표현들이지만, 실제적 사용에서는 허용되는 부분을 처리한다. 연결 강도가 0인 규칙은 연결되지 않음을 나타내고 연결 강도가 1인 규칙이 제일 우선시 되고 그 다음 강도를 가지는 규칙을 차례대로 적용한다. 각각의 형태소 간의 연결 규칙 작성에 대해 개략적으로 기술하면 다음과 같다.

■ 연결 가능 문법 규칙

moHANA에서는 특별한 연결 규칙이 작성되어 있지 않은 경우 연결이 불가능함을 뜻한다. 각 형태소 간의 연결이 가능한 문법 기술은 다음과 같이 한다. 예를 들어 동사는 어미와 연결이 가능하다는 품사에 따른 연결 여부 표현하는 문법 규칙은 다음과 같다.

```
{pv}{*}{*}{*}{*} <-> {ef}{*}{*}{*}{*} 1
```

여기에서 {pv}는 동사 어류 태그 그리고 ef는 어미 어류 태그를 뜻한다.

또한, 다음은 종성의 여부에 따라서 연결 가능한 조사의 형태가 달라짐을 뜻하는 형태에 따른 연결 여부를 표현하는 문법 규칙이다.

```
{*}{n}{*}{*}{*} <-> {j}{n}{*}{*}{*} 1
{*}{y}{*}{*}{*} <-> {j}{y}{*}{*}{*} 1
```

여기에서 {n}은 종성이 없는 형태소를 {y}는 종성이 이쓴 형태소 임을 나타내는 형태 정보이다.

■ 연결 불가능문법 규칙

각각의 형태소들 간에 연결이 불가능한 경우에 대해서는 연결 강도를 0으로 설정한다. 예를 들면, 다음은 조사와 조사는 연결 가능하지 못함을 표현하는 문법 규칙이다.

```
{j}{*}{*}{*}{*} <-> {j}{*}{*}{*}{*} 0
({j}는 조사에 대한 어류태그)
```

■ 대원칙과 세부 문법 규칙

문법 사전에 기술되는 문법 항목은 작성 순서에 따라서 적용된다. 즉 제일 처음 항목이 우선적으로 적용되며 그 아래에 나타나는 항목들이 차례로 적용되어서 전체 문법 규칙을 제한하고 변경한다. 따라서 대원칙을 우선적으로 기술하고, 각각의 문법의 대원칙에 예외인 세부 문법 기술을 하는 것을 원칙으로 한다.

대원칙과 예외 규칙에 대해 간단히 설명하면, 다음과 같다. 예를 들어, 연결 문법 규칙에는 “명사” + “조사”가 서로 연결될 수 있다는 규칙은 국어학에서 변하지 않는 국어의 기본 원칙이므로 이 규칙을 먼저 기술한다.

대원칙

```
{n*}{*}{*}{*}{*} <-> {j}{*}{*}{*}{*}  1
```

그러나, 명사나 조사의 정보 자질들에 따라 명사와 조사의 종류에 따라 연결될 수 없는 경우가 있다. 다음과 같이 인성의 의미적 특성을 지닌 명사는 장소(처소격 등)의 의미를 가진 조사와는 결합하지 않는다[3].

---

[3] “인성”을 지닌 경우에 장소를 나타내는 조사 “에서”가 결합하기도 한다. 이 경우의 조사 “에서”의 의미는 “장소”가 아닌 “출발점”의 의미를 가지고 있으므로, 태거나 구문 분석, 자연어 처리시에 각각의 문장에서 쓰여진 “에서”의 의미 및 기능을 기술하기에 편리하다.

예 1) *철수에서 논다.
예 2) 그 이야기는 철수에서 시작해서 영희로 끝난다.

예외규칙
{*}{*}{*}{per*}{*} <-> {j}{*}{*}{loc}{*} 0

이러한 문법 기술 방식은, 형태소 간의 연결 문법 규칙의 수를 줄일 수가 있으며 태그의 종류를 확장하여도 문법을 따로 기술할 필요가 없게 된다. 예를 들어, 정보검색에서 '경제'에 관련된 명사부류에 대해, 새로운 어류 태그를 부여하여 그 명사부류만을 따로 처리하고 싶다면, 해석 사전에 명사로서 가지는 각각의 값을 기술 한 뒤, 의미 정보에 '경제'를 뜻하는 의미 자질(eco)을 기술해 주면 된다. 만약 이 명사 부류가 형태소 간의 연결에 특별한 특성이 없다면 따로이 이 명사 부류에 대해 연결 문법 규칙을 첨가할 필요가 없다.

■ 허용 문법 규칙

한국어로 된 말뭉치를 형태소 분석을 하다 보면, 한국어의 맞춤법, 띄어쓰기 등 현재 한국어 문법에 어긋난 표현인 어절들이 많으며 이러한 어절들에 대해서 분석을 할 경우, 이 어절을 이루는 각 형태소들의 연결이 불가능하다. 그러나, moHANA에서는 이와 같이 비록 문법적으로는 어긋난 표현이지만, 많은 말뭉치에 나타나는 빈도수가 많은 경우의 어절들을 처리하기 위해서, 형태소 간 연결 문법 규칙을 유연하게 하여 이와 같은 어절들을 정확하게 분석할 수 있도록 하는 허용 문법 규칙을 두어, 이들 어절들을 처리할 수 있도록 하였다.

예를 들어서, 말뭉치를 분석하다 보면 많은 데이터에서 "안먹는다, 잘간다, 못놀겠다…"과 같이 띄어쓰기 오류인 어절들을 많이 찾아 볼 수 있다. 이와 같은 어절들은 한국어 문법에 의하면 "안 먹는다, 잘 간다, 못 놀겠다.."와 같이 써야 한다. 그래서 아래의 예)1처럼 형태소 분석에 실패하게 된다.

(5) 안먹는다 안먹는다_{unk}
잘간다 잘간다_{unk}
못놀겠다 못놀겠다_{unk}

여기에서 {unk}는 미등록어를 나타내는 어류 태그이다.

이러한 어절을 해결하기 위해서 아래와 같은 허용 문법 규칙을 적용할 수 있다.

{ad}{fc_adpred}{*}{*}{*} <-> {pv}{*}{*}{*}{*} 2
{ad}{fc_adpred}{*}{*}{*} <-> {pa}{*}{*}{*}{*} 2

여기서 {ad}는 부사를 나타내는 어류 태그이며, {fc_adpred}는 종성 여부를 나타내는 'fc'와 용언과 잘 쓰이는 부사를 나타내는 'adpred'가 결합하여 기술된 형태 정보이다.

위의 문법을 적용한 후에, 예 (6)처럼 "안먹는다, 잘간다, 못놀겠다…" 와 같이 띄어쓰기 오류인 어절들에 대해 분석이 가능하게 된다.

(6) 안먹는다 안_{ad} + 먹_{pv} + 는다_{ef}
잘간다 잘_{ad} + 가_{pv} + ㄴ다_{ef}
못놀겠다 못_{ad} + 놀_{pv} + 겠다_{ef}

■ 연결의 제한 문법 규칙

상기 문법 규칙 이외에, 연결의 제한/특수성을 표현하기 위하여 양방향 화살표가 아닌 단방향 화살표를 허용한다. 예를 들어, 오른쪽 화살표(->)의 경우, 화살표 왼쪽의 복합 품사는 화살표 오른쪽의 복합 품사와만 연결이 가능함을 나타낸다. 왼쪽 화살표(<-)는 반대 의미를 나타내며, 이는 연결 문법을 넓은 품사 영역에서 좁은 품사 영역으로 효과적으로 제한하며 기술하는 것을 가능하게 한다. 보다 구체적인 예를 들어 설명하면,

{nq}{*}{*}{per*}{*} <-> {j}{*}{*}{*}{*} 1
{ncn}{*}{*}{*}{*} <-> {j}{*}{*}{*}{*} 1
{nfix}{*}{*}{*}{*} <-> {j}{*}{*}{*}{*} 1

여기서 {nq}는 고유명사, {ncn}는 일반명사, {nfix}는 명사화접미사에 대한 어류태그를 나타낸다. 위의 표현은 {nq}, {ncn}, {nfix}의 어류 태그를 가진 단어들은 {j} 어류 태그를 가진 형태소와 결합이 가능함을 의미한다. 즉, 의미적 정보에 {per}를 가지고 있는 어류 태그는 명사, 접미사 등 여러 어류 태그가 있다. 이들 어류 태그는 통사적 정보에 {accu}를 가지고 있는 조사{j} 이외의 다른 조사와 자유롭게 결합할 수 있다. 반면에, 통사적 정보가 {accu}를 가진 조사{j}는 의미적 정보에 {per}를 가진 어류 태그와 결합이 가능하다. 이를 문법에서 표현하면 다음과 같다.

{nq}{*}{*}{per*}{*} <- {j}{*}{accu}{*}{*} 1
{nfix}{*}{*}{per*}{*} <- {j}{*}{accu}{*}{*} 1

여기에서 {per*}은 사람의 의미적 정보, {accu}는 여격에 대한 통사적 정보를 나타낸다. 위의 표현의 경우, 통사적 정보에 {accu}를 가지고 있는 조사는 어류 태그 중에 의미적 정보가 {per*}인 명사류와만 결합될 수 있음을 의미한다. 예를 들어, 조사{j}의 부류에 속하는 '에게'는 통사적 정보가 {accu}인 조사이다. 이 경우, '에게'는 '책상에게', '의자에게', '텔레비젼에게', '창문에게', '군대에게' 처럼 의미적 정보에 {per*}가 아닌 경우는 '에게'와 결합할 수 없으나, '선생님에게',

‘어머니에게’, ‘의사에게’처럼 의미적 정보가 {per*}인 명사군들과는 결합할 수 있다. {accu}의 통사적 정보를 가지는 조사{j}는 일반적인 다른 조사와는 달리 선택적으로 결합할 수 없음을 위와 같이 표현할 수 있다. 이와 같이 특별히 moHAHA에서는 형태소를 분석함에 의미정보를 활용할 수 있는 언어적 정보 필드를 마련하였다.

현재 moHANA는 1,000여개의 문법 규칙으로 일반적인 언어적 특성 및 실제 문서에서 사용되는 형태와 관련한 내용을 기술하고 있다.

3.4 사용자 사전 및 기분석 사전

일반적인 형태소 분석 시스템에서 어휘 사전은 언어학 전문가에 의해 구축되는데, 이에 따라, 한번 구축되면 계속 발생하는 신조어나 형태소 분석기가 적용되는 사이트만의 요구를 자주 반영하기 쉽지 않다. 그러므로, 본 연구에 따르면, 형태소 분석기에 대한 전문 지식이 없는 사용자도 쉽게 신조어를 등록할 수 있도록 사용자 사전이 제공된다. 본 연구의 사용자 사전 역시 동일한 기능을 가지며, 사용자 사전에 등록된 형태소들은 어휘 사전에 존재하는 형태소들보다 우선적으로 적용되게 구성된다. 사용자가 품사를 특별히 지정하지 않는 한, 가장 많이 나타나는 품사인 ‘명사’를 품사값으로 갖도록 구성한다. 빈번히 나타나는 입력이나 애매성 해소에 많은 시간이 걸리는 입력에 대해서 미리 해석한 결과를 기분석 사전에 기록하여 해석 시간을 단축한다. 또한 예외적인 형태에 대한 해석 내용도 기록하여 일반적인 문법 규칙에 어긋나는 현상에 대해서도 기록하여 사용한다.

4. 실험 및 결과

5차원 벡터값으로 표시된 moHANA에서 사용하는 각각의 언어적 정보들의 유용성을 살펴보기 위해서 KAIST 말뭉치를 사용하여 실험했다. KAIST 말뭉치는 신문, 사설, 교과서, 소설, 홍보물 등의 다양한 내용의 글들로 175,524 어절로 구성되어 있다[1].

moHANA 시스템에서 출력하는 복수개의 결과 중 제1결과만을 사용하여 형태소 단위의 정확률과 재현율 그리고 어절 단위의 정확률을 측정하였다. 실험에 사용한 측정 단위는 다음과 같다.

$$Precision = \frac{\#correct\ \ morphemes}{\#result\ morphemes}$$

$$Recall = \frac{\#correct\ \ morphemes}{\#answer\ \ morphemes}$$

$$Eojeol\ Acc. = \frac{\#correct\ \ eojeols}{\#eojeols}$$

표 1은 KAIST 말뭉치를 대상으로 태그정보에서 어류 태그 값을 제외한 나머지 네개의 값을 각각 삭제하였을 때의 moHANA의 분석 성능을 보여준다. Avg #Res.는 한 어절당 출력되는 해석 결과 수를 보여준다.

표 1 정보 삭제에 따른 분석 성능의 변화

삭제정보	*Prec.*	*Rec.*	*Eoj Acc.*	*Avg #Res*
-	0.875	0.822	0.791	2.13
형태정보	0.876	0.821	0.791	2.17
통사정보	0.872	0.811	0.786	2.15
의미정보	0.860	0.799	0.770	2.67
화용정보	0.871	0.818	0.787	2.46

moHANA의 정확률은 79%정도인데 이는 KAIST 말뭉치에서 사용하는 형태소의 단위나 활용된 형태소의 원형에 대한 정의가 moHANA에서 사용하는 내용과 차이가 있기 때문이다[4]. 그러나, 이와 같은 moHANA의 형태소 분석 정확률의 결과는 KAIST와 분석 기준이 다르고, n-gram과 같은 확률 정보 없이 분석된 것이므로, 동일한 환경에서 moHANA의 형태소 분석 결과 정확률은 기존의 형태소 분석기의 분석 결과와 유사하거나 더 높을 것이라 기대된다.

표에 따르면, moHANA에서 사용된 각각의 언어적 정보 중에서, 의미 정보가 분석 정확률에 가장 큰 영향을 미치며 화용 정보, 통사 정보 그리고 형태 정보의 순으로 분석에 영향을 미치는 것을 알 수 있다. 이와 같은 실험 결과는 언어를 분석함에 형태적, 통사적 정보뿐만 아니라, 의미적 정보가 매우 중요한 기능을 함을 증명한 것이다.

형태 정보를 삭제한 경우, 성능에 큰 차이를 보이지는 않았지만 종성 여부에 따른 조사 결합 제한이 제대로 이뤄지지 않아 평균 해석 결과수가 증가함을 알 수 있다.

통사 정보를 제거했을 경우에는 종결 어미와 연결 어미간의 구분이 모호해져 보조동사와의 해석에 오류가 발생하는 경우와, 품사 전성이 나타나는 경우에 제대로 된 결과를 내지 못했다. 예를 들어 ‘노력해주기’ 입력에 대해서 ‘노력_{ncp} + 하_{vfix} + 여_{ef} + 주_{aux} + 기_{ef}’와 같이 용언과 보조 동사의 해석 대신 ‘노력_{ncp} + 해_{nc_one} + 주기_{ncp}’와 같은 복합명사 구성인 어절로 분석하였다.

의미 정보를 제거한 경우에는 종결 어미나 조사에 대해서 1음절짜리 명사나 고유 명사로 해석하는 오류가

많았다. 특히 어류태그가 사람 고유명사인 형태소 중에 의미적 정보로 '이름의 성'을 가진 형태소들에 대한 연결 제한 규칙이 제대로 적용되지 못하는 분석의 오류를 보였다. 예를 들어, 어절'사태임에'가 '사태_{ncn} + 이_{jp} + ㅁ_{ef}+ 에_{j}'에 대신에 '사태_{ncn} + 임_{nq} + 에_{j}'로 '임_{nq}'을 사람의 이름으로 잘못된 해석 결과를 보였다. 고유명사인 '임_{nq}'이 이름의 성씨를 나타내는 정보가 사라져 연결을 제한하는 문법의 적용이 제대로 이뤄지지 않았다.

화용 정보를 제거한 경우에는 색인용 어휘와 비색인용 어휘의 구분이 모호해져 출력되는 평균 해석수가 늘어나서 잘못된 결과를 1순위로 출력하는 오류를 보였다. 예를 들어, 어절 '최고위원'이 '최고_{ncn} + 위원_{ncn}' 대신에 '최고위_{ncn} + 원_{nc_one}'로 잘못된 해석 결과를 상위 결과로 보였다. 비색인용 1음절짜리 명사인 '원'의 정보가 삭제됨에 따라 '최고위_{ncn}'와의 연결을 제한하는 문법의 적용이 제대로 이뤄지지 않았다.

5. 결론

보다 정확한 형태소 분석을 위해서는 형태소 분석기에 모든 가능한 언어적 특성을 반영할 수 있어야 한다. moHANA는 어류 정보, 형태적 정보, 통사적 정보, 의미적 정보 및 화용 정보의 5차원의 벡터 값을 가지게 하여 언어적 특성에 따른 형태소의 분류를 쉽게 지정할 수 있게 한다. 이러한 다차원 해석 사전에서의 태그 간의 연결 여부를 표현하기 위해 양방향 화살표와 단방향 화살표 및 연결 강도를 나타낼 수 있는 문법 연산자를 개발하여 문법 사전을 기술한다. 기존 일차원으로 정의된 품사 사전과 달리 다차원으로 태그가 기술되어 있어서 새롭게 발견한 형태소의 부류나 예외의 경우로 처리하고 싶은 형태소 부류에 대해서 손쉽게 수정을 가능하게 한다. 이는 형태소 분석기를 사용하려는 작업에 맞춰서 형태소의 단위 및 부류를 쉽게 정의할 수 있게 한다. 또한 문법 규칙 항목에 연결 강도를 지정할 수 있어 해석 대상 도메인의 특성에 따라서 띄어쓰기나 문법 오류를 포함한 입력을 처리할 수 있다.

본 논문에서 소개한 5차원의 정보는 형태소 분석을 중심으로 한 내용이다. 그러나 5차원 이상의 고차원으로 확장하여 실제 적용하는 작업에 맞춰서 작업에 사용하는 온톨로지의 정보를 자동으로 반영하여 형태소 분석 과정에서 관련 정보를 사용하는 방법의 연구가 앞으로 필요하다.

참고 문헌

[1] 김재훈 1996. 오류-보정 기법을 이용한 어휘 모호성 해소. *박사학위논문, 한국과학기술원 전산학과*

[2] 임희석, 윤보현, 임해창, 1995. 배제 정보를 이용한 효율적인 한국어 형태소 분석기, *한국정보과학회 논문지, 22권 6호, pp. 957-964*

[3] 강승식, 1993, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, *서울대학교 공학박사 학위 논문*

[4] 정성영 1996. 마코프 랜덤 필드를 이용한 영어 품사 태깅 시스템. *석사학위논문, 한국과학기술원 전산학과*

[5] 강인호, 김재훈, 김길창 1998, 최대엔트로피 모델을 이용한 한국어 품사 태깅, *한글 및 한국어 정보처리 학술대회*

[6] 심광성, 양재형 2004. 인접 조건 검사에 의한 초고속 한국어 형태소 분석, *한국정보과학회 논문지: 소프트웨어 및 응용, 31권 1호 pp.89-99*

[7] Kwangseob Shim and Jaehyung Yang 2002. "MACH: A Supersonic Korean Analyzer," *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002), pp. 939-945*

[8] 양승현, 김영섬, 2000 "부분 어절의 기분석에 기반한 고속 한국어 형태소 분석 방법" *정보과학회 논문지: 소프트웨어 및 응용, 27권, 3호, pp. 290-301*

[9] Gary Geunbae Lee, Jong-Hyeok Lee, and Jeongwon Cha, 2002. Syllable-pattern-based unknown-morpheme segmentation and estimation for hybrid part-of-speech tagging of Korean, *Computational Linguistics, vol. 28, pp. 53-70*

[10]CJK Moran, 2007. MoranSoft 주식회사 *http://www.moransoft.com*

[11]허웅, 1984, 국어학, *샘문화사*

[12]김진우, 1996, 언어학, *탑출판사*