moHANA: 해석 정보 정밀화를 통한 형태소 해석 정확률 향상 방안

moHANA: Improving Korean Morph ological Analysis using Information Specificatio n

서승현*, 김재동*, 강인호**
* 주식회사 워드워즈 (shsuh@wordwords.co.kr, ihkang@wordwords.co.kr)
** 카네기멜론대학교/언어공학연구소 (jdkim@cs.cmu.edu)

요 약

본 연구는 해석 정보 정밀화를 통해 복합명사, 파생명사와 같은 복합어와 해석 실패 어절을 중심으로 형태소 분석기 moHANA의 분석 정확률을 높이기 위한 방안을 제시한다. 이와 같은 방안은 한국어의 형태적인 언어적 특성을 기반으로 한 것으로, 어절들을 형태소 분석함에 각각의 형태소의 언어적인 정보와 언어정보간의 문법을 정밀화함으로써 형태소 분석의 정확률을 높인다.

keywords : 단어조어, 복합명사, 파생명사, 다단계 해석 문법, 부분 기분석 사전

1. 서 론

한국어는 형태적인(morphological) 언어적 특성이 강한 언어이다. 다른 언어와는 달리 형태적인 특성이 강하여, 언어의 어휘적인 의미와 문법적인 의미 및 기능을 나타내는 다양한 의미와 기능의 형태소가 발달되었다. 특히 형태적 특성이 강한 언어는 새로운 단어들을 만들어내는 단어조어법이 발달하여, 한국어는 복합어 및 파생법이 발달하였다. 그래서 한국어를 형태소 해석하기 위해서는 이러한 한국어의 단어조어적인 특성을 효율적으로 기술하는 것이 형태소 분석의 정확률 및 기능을 높이는 방법이다. 특히 요즘 인터넷 상에서 자신의 글을 자유자재로 올리게 되면서 띄어쓰기가 잘못된 어절이나 오탈자 및 신조어의 빈번한 활용으로 형태소 분석의 어려움이 가중되어 분석 실패하거나 잘못 분석되는 경우가 많다. 잘못된 형태소 분석은 정보 검색기가 올바른 결과를 추출하지 못하게 하며 문서 상에 나타난 정보를 추출하는 데에도 어려움을 유발한다.

이에 본 논문에서는 다차원 해석 사전을 이용하는 형태소 분석기 moHANA(morphological Hangul Analyzer)[1]에서 한국어의 형태적인 특성을 반영한 복합어 및 파생어의 처리방식과 이를 이용한 띄어쓰기 어절 처리 방식에 대해 소개한다.

2. 복합어 분석

한국어 형태소 분석기의 기능을 판가름할 수 있는 것은 한국어의 복합어(파생어 포함)와 분석실패 어절 처리를 얼마나 효과적으로 처리할 수 있느냐이다.

2.1 복합명사 처리

한국어 어절을 형태소 분석할 경우 어려움 중의 하나가 복합어의 처리이다. moHANA에서는 한국어의 복합어를 처리하기 위한 한 방법으로 각각의 형태소의 언어적 특성을 다차원 해석 사전에서 기술한다. moHANA의 다차원해석 사전에서는 일반적인 언어적 특성인 음운적(phonological), 형태적(morphological), 통사적(syntactical), 의미적(semantic), 화용적(pragmatic) 특성을 모두 기술할 수 있다[2][3]. 이러한 특성들을 이용하여 한국어의 복합어를 사전에 하나의 올림말로 올리지 않고 처리할 수 있도록 하였으며, 과도한 분석을 막아 한국어의 복합어의 형태소 분석을 효율적으로 하도록 하였다.

한국어는 명사와 명사가 결합하여 새로운 단어를 만들거나 표기상 명사들을 붙여 쓰므로, 이와 같은 복합명사들을 한국어 형태소 분석에서 정확하게 분석해내는 것이 그 성능의 하나의 척도가 될 수 있다.

예제(1), (2)의 경우처럼, 한국어 어절을 분석하다 보면, 용언간의 결합이 명사의 결합으로 분석이 되거나 오분석을 하는 경우가 허다하다.

(1) 고아지지 *고아_{ncn[1]} + 지지_{ncp}[2]
고_{pv[3]} + 아_{ef} + 지_{aux} + 지_{ef}
(2) 기독교의 기독교_{ncn}+의_{j}
*기독_{ncn} + 교의_{ncn}

그러나, moHANA에서는 위와 같이 복합명사로 오인하여 분석하는 것을 막기 위해서, '지지', '교의'와 같은 명사들에 ¡복합명사 형성 제약¡정보를 두어 처리하였다. 즉, 복합명사 형성시, 이들 명사들은 그 복합명사의 마지막에 나타날 수 없다는 위치 제약을 두어 처리하면, 다음과 같이 분석이 된다.

---

[1] {ncn}은 일반명사에 대한 태그이다.
[2] * 표시는 비문법적임을 표시한다
[3] {pv}는 동사에 대한 태그이다.

(3) 고아지지 고_{pv} + 아_{ef} + 지_{aux} + 지_{ef}
(4) 기독교의 기독교_{ncn} + 의_{j}

이러한 특성들을 가지는 명사들은 일반 명사뿐만 아니라, 지명, 인명, 회사명 등의 고유명사에도 있으므로, 이러한 복합 명사 형성 제약적인 특성을 지닌 명사들의 사전에 이와 같은 특성을 입력하면, 기술을 효율적으로 할 수 있다. 뿐만 아니라 분석된 결과들 중에 잘못된 결과를 줄이고 분석이 정확한 결과만을 내놓으므로, 형태소 분석의 정확률을 높인다.

2.2 파생명사 처리

한국어의 단어조어방식 중에 발달한 것이 ¡접미사, 접두사¡와 같은 파생접사를 이용한 파생법이다. 이러한 접사들은 새로운 단어의 형성을 생산적으로 하므로, 이 모든 파생된 단어들을 사전에 올리는 것은 비효율적일 뿐만 아니라, 파생된 단어들을 사전에 모두 올리는 것도 불가능한 일이다. 특히, 단어조어에 있어서 매우 생산적이라 함은 그 방법이 규칙적이라는 것을 의미하므로, 생산적인 접사들의 언어적 특성을 찾아내어 규칙화할 수 있다. 따라서 생산적인 접사들의 특성을 사전과 문법에 기술하여 처리하는 것이 효율적이다.

단어조어에 생산적인 접사들은 결합하는 형태소에 의미적, 형태적 언어특성에 따라 단어조어에 제약을 준다.

예를 들어. '현', '주'와 같은 접미사는 지명을 나타내는 고유명사와 특히 외국지명과는 결합하여 단어를 파생하지만, 이들 명사 부류 이외의 명사 부류와는 결합하지 않는다.

(5) 나카타현 나가타_{nq_forloc[4]} + 현_{nfix[5]}
미네소타주 미네소타_{nq_forloc} + 주_{nfix}

만약 '현', '주'와 같은 접미사를 모든 명사와 결합할 수 있도록 하면, 예제(6), (7)과 같이 잘못된 분석이 된다.

(6) 성주현 성주현_{nq_per[6]}
*성주_{ncn} + 현_{nfix}
*성주_{nq_loc[7]} + 현_{nfix}

(7) 이성주 이성주_{nq_per}
*이성_{ncn} + 주_{nfix}

이와 같이 한국어의 접미사는 결합하는 명사들의 의미적 특성에 따른 결합 제약이 있다.

따라서 정확한 한국어 형태소 분석을 위해서는 형태소의 음운, 형태, 의미 등과 같은 언어적인 특성을 기술하고 이들 언어적 정보들을 이용하여야 한다.

[4] {nq_forloc}는 외국지명 고유명사에 대한 태그이다.
[5] {nfix}는 명사파생접미사에 대한 태그이다.
[6] {nq_per}는 인명 고유명사에 대한 태그이다.
[7] {nq_loc}는 한국지명 고유명사에 대한 태그이다.

3. 분석실패 어절 분석

형태소 분석에 어려움이 생겨 분석에 실패한 어절을 분석하기 위해서, moHANA에서는 형태소 간의 결합할 수 있는 문법의 강도를 조절하거나, 띄어쓰기 오류의 패턴을 찾아 사전에 기술하여 형태소 분석에 반영할 수 있도록 하는 두 가지 장치를 마련하여 해결한다.

3.1 다단계 해석 문법

형태소 분석에 실패한 어절들을 분석해 보면, 언중들이 자주 틀리는 패턴이 있고, 그 패턴들을 분석해 보면 그 형태소들 간에 규칙성을 찾을 수 있다. 이와 같은 경우는 형태소 결합 문법의 강도를 조절하여 처리함으로써 분석 실패율을 줄인다. 즉, 문법에 맞는 입력이라고 가정하여 해석을 수행하였으나 해석 실패할 경우 다음 단계의 문법을 활용하여 입력에서의 오류를 추정하여 해석을 시도한다.

(8) 학생및선생 학생_{ncn} + 및선생_{unk}
안적는 안적는_{unk}
엄마와아이 엄마_{ncn} + 와아이_{unk}

이러한 어절을 해결하기 위해서 부사와 용언 (동사, 형용사)의 연결을 2단계에서는 허용하는 문법 규칙[8]을 작성한다.

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

이와 같은 허용 규칙을 통해 띄어쓰기 오류가 있는 어절을 다음과 같이 형태소 분석을 한다.

(9) 학생및선생 학생_{ncn} + 및_{ad} + 선생_{ncn}
안적는 안_{ad} + 적_{pv} + 는_{ef}
엄마와아이 엄마_{ncn} + 와_{j} + 아이_{ncn}

예외적인 경우와 이에 따른 연결 여부를 쉽게 기술하기 위해서, moHANA는 다차원의 형태소 해석 정보를 이용하여 결합 문법 강도 조절에 해당하는 형태소들을 쉽게 정의한다.

3.2 부분 기분석 사전

형태소 분석 실패한 어절을 분석하여 형태소 간의 결합패턴을 찾을 수 없지만, 언어 사용자들이 자주 틀리는 띄어쓰기 오류 중에서, 자주 쓰이는 형태적 패턴을 찾아 기분석 사전에 등록하여 형태소 분석시 이 사전의 분석결과를 이용할 수 있도록 하였다. 이러한 예외 처리는 오류 패턴의 발생 위치, 좌우에 나타나는 형태소의 유형과 이에 따른 해석 결과물을 쌍으로 가진다. 예를 들어 '에대한', '에가는', '하기위한', '으로인한'과 같은 띄어쓰기 오류 패턴에 대해서 마치 공백이 들어간

[8] {unk}는 미등록어, {ad}는 부사, {pv}는 동사, {pa}는 형용사에 대한 태그이다.

것처럼 내부적으로 해석을 시도하는 과정을 추가하여 아래 예제 (10)과 같은 띄어쓰기 오류 어절을 올바르게 분석한다.

(10) 편지에대한 편지_{ncp} + 에_{j} + 대하_{pv} + ㄴ_{ef}

청주에가는방법 청주_{ncn} + 에_{j} + 가_{pv}는_{ef} + 방법_{ncn}

생존하기위한방법 생존_{ncp} + 하_{vfix} + 기_{ef} + 위하_{pv} + ㄴ_{ef} + 방법_{ncn}

태풍으로인한피해 태풍_{ncn} + 으로_{j} + 인하_{pv} + ㄴ_{ef} + 피해_{ncp}

moHANA에서의 띄어쓰기 오류 부분에 대한 처리는 형태소 분석시 띄어쓰기 교정 프로그램이 없이도 띄어쓰기 오류 어절을 처리할 수 있도록 한 것이다.

또한, 형태소 간의 문법 결합 강도를 조절하고, 부분 기분석 사전을 사용함으로써 다음과 같은 어절을 처리할 수 있다.

(11) 지가공시및토지등의평가에대한법률
지가공시_{ncp} + 및_{ad} + 토지_{ncn} + 등_{nfix} + 의_{j} + 평가_{ncp} + 에_{j} + 대하_{pv} + ㄴ_{ef} + 법률_{ncn}
성매매알선등행위의처벌에대한법률
성매매_{ncn} + 알선_{ncp} + 등_{nfix} + 행위_{ncn} + 의_{j} + 처벌_{ncp} + 에_{j} + 대하_{pv} + ㄴ_{ef} + 법률_{ncn}

이와 같이 형태소 분석 시 각 형태소의 구체적이고 세분화된 정보를 이용하면, 형태소 분석의 정확률을 높이고 분석을 효율적으로 할 수 있다.

4. 실험 및 고찰

본 논문에서 제안하는 성능 향상 방안의 유용성을 살펴보기 위해 실제 상용 검색 엔진 (daum[9])에 사용되었던 질의 로그 2만개를 대상으로 형태소 분석기의 성능 변화를 보인다. 사용자 질의는 신문 기사나 교과서의 문장과 달리 띄어쓰기 오류와 오탈자가 빈번하여 실제 사용되는 문장에 대한 해석 성능을 살펴볼 수 있다. 형태소 분석기의 분석 성능은 입력 어절당 출력하는 해석 결과 수 (avg # cand), 출력한 해석 결과 중 정답을 포함하는 제대로 해석된 어절의 비 (prec), 그리고 분석 실패한 어절의 비 (fail rate)를 이용하여 보인다. 즉 적은 해석 결과수를 보이면서 낮은 실패율과 높은 정확률을 보이는 형태소 분석기가 좋은 분석기를 뜻한다. 표1은 사용하는 정보 세분화에 따른 각 성능 단위의 변화를 보인다.

표 1 정보 정밀화를 통한 형태소 분석 성능의 비교

	avg # cand	prec	fail rate
Base	2.976	0.859	0.068
+ 접미사 세분	1.676	0.862	0.111
+ 부분 기분석 사전	1.671	0.866	0.104
+ 다단계 해석 문법	1.654	0.914	0.047

표 1에서 'Base'는 기존 형태소 분석기[4]와 같이 1차원적인 품사 위주의 품사 정보를 이용한 기본 모델을 뜻하며, '+ 접미사 세분'은 의미 정보를 활용한 형태소 품사 정보의 세분화를 통해 얻어낸 모델을 뜻한다. 그리고 '+ 부분 기분석 사전' 과 '+ 다단계 해석 문법'은 바로 윗줄의 모델에 부분 기분석 사전과 다단계 해석을 통한 모델의 확장을 의미한다. 접미사 세분을 통해서 평균 약 3개 정도의 후보를 제공하던 형태소 분석기가 평균 약 1.7개의 결과를 출력하면서도 유사한 정확률을 유지하는 것을 알 수 있다. 반면 세분화를 통해 형태소 연결 가능 여부 조건이 강화되어 해석 실패율이 증가한다. 그러나 빈번한 오류를 해결하기 위한 부분 기분석 사전과 다단계 해석 문법의 적용을 통해서 정확률을 향상시키며 해석 실패율을 57% 가까이 감소시키는 것을 알 수 있다.

5. 결 론

한국어 어절을 정확하게 형태소 분석을 하기 위해서는 한국어의 언어적 특성을 고려하려야 한다. 이에 본 논문에서는 각각의 형태소의 정보를 세밀화하고 각각의 정보들간의 언어적인 특성을 분석하고 분류해내어, 이를 형태소 분석에 이용하여 분석의 정확률을 높였으며 분석 실패한 어절의 수를 줄일 수 있었다.

이와 같이 정확한 형태소의 분석은 정보검색의 정보 제공에서의 정보의 정합성을 높이고, 이어서 각각의 정밀화된 형태소의 언어적 정보를 이용한다면, 자연언어처리에 효율적인 방법론을 제시할 수 있을 것이다.

참고문헌

[1] 서승현, 강인호, 김재동, ¡moHANA: 다차원 해석 사전을 기반으로 한 한국어 형태소 분석기¡, *한글 및 한국어 정보처리 학술 대회*, pp. 99-106, 2007

[2] 허웅, "국어학", *샘문화사*, 1984

[3] 김진우, "언어학", *탑출판사*, 1996

[4] 강인호, 김재훈, 김길창, "최대 엔트로피 모델을 이용한 한국어 품사 태깅", *한글 및 한국어 정보처리 학술 대회*, 1998

[9] http://www.daum.net