말뭉치를 이용한 형태소 분석 단계에서의 중의성 해결에 관한 연구

김경서, 김대철, 정강석, 송만석
연세 대학교 전산과학과

요 약

자연 언어 처리의 효율성은 대량의 정보를 담고 있는 사전을 잘 구성하는 데 있다. 사전을 잘 이용하기 위해서는 입력 어절에 대한 정확한 표제어(원형)를 효과적으로 찾아야한다. 입력 어절에 대한 표제어를 찾는 역할을 하는 형태소 분석기는 한 어절의 정보만 이용하기 때문에 입력 어절을 두 가지 이상의 표제어로 해석할 수 있다. 연세 대학교 사전편찬실이 갖고 있는 연세 말뭉치 I에 대해 10% 이상의 어절이 두가지 이상으로 분석되는 중의성을 가진다. 이렇게 중의성을 가지는 어절이 그대로 구문 구조 분석기에 전달되면 중의성을 해결하기 위해 구문 구조 분석기의 처리 과정이 복잡해진다. 본 논문은 표제어의 중의성을 보이는 어절을 구문 구조 분석기에게 전달하기 전에 형태소 분석기와 구문 구조 분석기 사이에서 정확한 표제어를 찾는 방법을 제안한다.

1. 서론

최근의 자연 언어 처리는 규칙(Rule)을 주로 사용해 처리하는 방식에서 어휘 정보(Lexicon)를 보강하는 방향으로 나간다. 이런 어휘 정보는 사전 형태로 저장되고, 처리기는 사전의 표제어를 참조함으로써 어휘 정보를 이용한다.

따라서 자연 언어 처리의 여러 단계 중에서 가장 먼저 위치하는 것은 입력 어절에 대해 처리기가 정보를 이용할 수 있도록 사전의 표제어를 정확하게 찾는 일이다. 이 처리과정을 형태소 분석이라 한다.

기존의 형태소 분석기는 자연 언어 처리에서 사전 정보의 중요성을 잘 인식하지 못하고, 주로 철자법 검사기 교정기, 띄어 쓰기 검사기 교정기에 응용되면서 입력되는 어절이 바른지 그른지 즉, 형태소 분석기가 분석할 수 있는지 분석할 수 없는지에 주로 관심이 있었다.

형태소 분석기가 하는 가장 중요한 일은 구문 구조 분석기 이후의 처리기가 한 어절에 대한 정보를 사전에서 참조할 수 있도록 사전의 표제어(원형)를 정확하게 찾는 것이다.

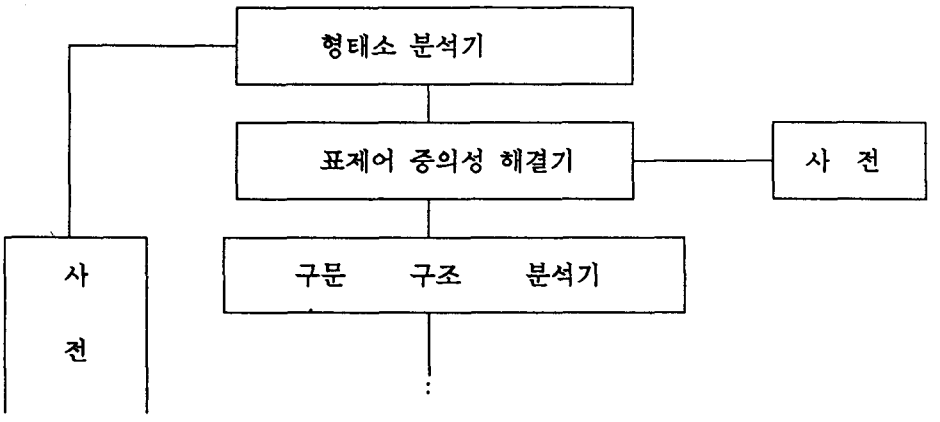
영어의 syntax analysis가 한국어의 구문 구조 분석과 동등할 때 한국어의 구문 구조 분석이 잘 안되는 것은 한국어의 구문 구조가 영어에 비해 복잡해서 구조 자체를 결정하기가 어려운 이유도 있지만 구문 구조 분석 이전에 있는 현재의 형태소 분석 단계가 syntax analysis 아래에 있는 lexical analysis 과 동등하지 않고 그 기능이 약

하기 때문이다.

즉, lexical analysis에 비해 형태소 분석기는 사전 표제어에 대한 모호성을 많이 보이기 때문에 구문 구조 분석기가 syntax analyzer가 하지 않는 표제어에 대한 중의성을 제거해야하는 부담이 생기게 된다.

그러나, 현재의 형태소 분석기는 한 어절만 보고 간단한 규칙과 한 어절이 가질 수 있는 간단한 정보만을 이용하기 때문에 어떤 어절에 대해서는 두가지 이상으로 분석되는 중의성을 피할 수 없다.

구문 구조 분석기가 syntax analysis 와 비슷한 출발점을 가지게 하기위하여 형태소 분석기가 가능한 모든 가능성의 표제어(원형)을 출력할 때 이를 구문 구조 분석기에 전달 하기 전에 표제어의 중의성을 해결하는 단계를 만든다.



2. 문제 정의

한국어 사전 편찬실에 있는 연세 말뭉치 I (300만 어절, 36만 어절 종류)을 신기철, 신용철 사전의 표제어와 품사를 담고 있는 사전을 이용해 가능한 모든 원형을 찾는 형태소 분석기로 분석한 결과 중의성을 보이는 어절 종류가 약 5만 단어가 나왔다.

이 결과에 따르면 보통의 형태소 분석기는 일반적인 문장의 어절을 분석할 때 13% 정도의 어절에 대해서 원형이 두개 이상으로 분석되는 사전 표제어 중의성을 띤다.

사전 표제어 중의성을 분류하면 다음과 같다.

1). 용언류 + 용언류 : 물어 1. 묻다 2. 물다
2). 용언류 + 체언류 : 한 1. 하다 2. 한
3). 체언류 + 체언류 : 강의 1. 강의 2. 강

체언류 : 홀로 독립적으로 쓰일 수 있는 것(체언, 관형사, 부사, 감탄사)
용언류 : 형용사, 동사

위의 경우이외에도 (체언류 + 체언류 + 용언류, 체언류 + 용언류 + 용언류,)등 여러 형태가 있으나 이들은 크게 위의 세가지 경우의 하나에 속한다고 본다.

1)의 경우 동사나 형용사는 문장내에서 그 통사적인 역할이 비슷하기 때문에 문장의 구조가 많이 결정 된 다음 해결이 가능하다.

3)의 경우 대부분이 (명사 + 명사)의 꼴로서 동사, 형용사, 명사의 의미 분류가 되어야 정확한 해결이 가능하다.

2)의 경우 체언류과 용언류는 문장 안에서 특히 그 문법적 역할이 다르므로 의미 정보를 많이 이용하지 않고, 문장안에서 나타나는 통사적 정보와 통계적인 정보로써 구별 가능하다.

본 논문에서는 위의 세 가지 경우중에서 2) 경우로 분류되는 어절의 사전 표제어 중의성을 처리하도록 한다.

3. 품사 중의성 문제

문장 안에서 한 어절이 여러 개의 품사로 분석될 때 이를 하나의 품사로 결정하는 것이 품사 중의성 문제이다.

품사 중의성 문제는 한 어절이 가질 수 있는 여러 가능한 것(품사)중 하나를 선택한다는 점에서 표제어 중의성 문제와 비슷하다. 그러나 자연 언어 처리 단계에서 그 위치는 확실하게 다르다.

즉, 품사 중의성은 구문 구조 분석기나 그 이후의 처리기가 사전을 참조하여 그 사전 안에 있는 정보를 이용하여 품사를 결정하고 표제어 중의성은 구문 구조 분석기가 사전을 이용하기 전에 다른 처리기가 존재하여 이 처리기만의 정보를 이용해서 사전의 표제어를 결정한다.

이 두 문제는 그 위치는 다르나 이용하는 정보와 처리하는 비슷하기 때문에 표제어 중의성 해결 방식은 품사 중의성 해결 방식으로 확장될 수 있다.

4. 처리 정보

표제어 중의성 해결기가 사용하는 정보는 다른 자연 언어 처리기가 이용하는 사전과는 독립적으로 존재하며 사전에 담겨 있는 정보(음운, 품사, 의미등)와는 전혀 다르다. 사전은 단어를 표제어로 하고 표제어 아래에 일반적인 언어 정보와 처리기에 유용한 정보를 담고 있으나 표제어 중의성 사전은 중의성을 보이는 어절을 표제어 하고 표제어 아래에 어절이 가지고 있는 특수한 정보를 가진다.

각 어절이 서로 쓰이는 환경이 다르고, 가지고 있는 정보가 다르기 때문에 처리 방식을 일괄적으로 분류를 하기 어렵다. 그러나 다음과 같이 어절이 가지고 있는 정보를 대략 나누어서 처리하는 것이 효과적이다.

4.1. 사전 재구성

형태소 분석기를 비롯한 모든 처리기가 효율적이려면 사전의 표제어가 방대해야한다. 그러나 이런 경우 보통의 형태소 분석기는 표제어가 많아서 중의성을 낼 수 있다.

큰 사전에는 대표적인 단어 이외에도 고어, 이두, 은어등을 수록하고 있어 형태소 분석기가 이를 이용할 때 어떤 어절에 대해서 중의성을 보일 수 있다.

예) 넌 : 넌(체), 녀다(동) + ㄴ(어미)
"녀다" 는 고어로써 "가다"라는 뜻

이 경우는 표제어 중의성 해결기가 해결하기 보다는 사전에서 "녀다" 라는 동사를 제거해 형태소 분석기에서 중의성을 보이지 않게 한다.

4.2. 용례적 제한

용례는 말뭉치에서 실제로 나타나는 언어 현상을 말한다. 언어 현상을 잘 정리한 것이 문법이다. 문법에는 항상 예외적인 경우가 있으며, 모든 언어 현상을 나타내지는 못한다. 문법에서 볼 때 말뭉치에서 자주 나타날 것 같은 것이 자주 나타나지 않거나 특별하게 문법적으로 정의 되지 않은 것이 말뭉치에서 자주 나타나는 경우도 있다.

문법적으로 잘 정의되지는 않았지만 말뭉치에서 잘 나오는 현상을 정보로 이용하고, 비록 문법적이라고 할지라도 말뭉치에서 나타나지 않는 것은 처리에서 제외한다.

어떤 어절은 용례에서 반드시 앞에 특별한 조사 뒤에서만 사용되거나, 뒤에 오는 어절이 항상 같거나 하는 문법적으로 꼭 그렇다고 할 수는 없으나 실제 언어 생활에서 마치 문법같이 사용되는 것이 용례적 정보다.

◆ 형태적 용례 제한

형태적 용례 제한는 어절이 특정한 조사 뒤에서만 사용되거나 그 어절 다음에는 대개 어떤 단어가 쓰인다든지 따위의 정보이다.

예) 대한 : 대한(체), 대하다 (용) + 다(어미)
그런 예비 지식이 없는 고교생으로서는 이 사실에 대한 정확한 이해가
안 되지 않겠는가
국정 운영의 유연성과 어떤 변화에 대한 기대를 낳게하는 대목이다.
*) 헌법 제 1조에 대한 민국은 민주 공화국이라고 되어있다.

"대한" 어절이 용언류(어미 활용을 하는 종류)로 사용되면 앞의 어절의 조사가 모두 "에"로 끝난다. 이때 "대한"이 문법적으로 조사 "에"를 제한하는 것은 아니지만 거의 모든 경우가 이 꼴로 나타나 "대한"이라는 어절을 만났을 때 앞에 조사 "에"가 있으면 "대한(체언류,부사)"로 분석할 수 있다.

그러나 *)의 경우에서 고려할 때 "대한"이 문법적으로 앞의 조사 "에"를 제한 하지 않기 때문에 예외적인 경우가 생길 수 있다. "대한" 이 용언류로 잘못 분석 될 수 있다. 이런 예는 거의 나타나지 않으며 정보를 추가해서 해결할 수 있다.

◆ 문법적 용례 제한

문법적인 용례 제한은 형태적인 용례 제한이 어절이 앞뒤에 특정한 형태의 단어나

기능어를 제한하는 것과 달리 어절이 특정한 품사나 구분할 수 있는 부류의 기능어(예, 관형격 어미)와 함께 잘 사용되는 것이다.

예) 할 : 할 (체), 하다(용) + 다(어미)
그는 3 년 연속 3 할 이상의 타율를 치고 있다.
총선에서 겨우 3 할 남짓 득표율을 올리고 전체 의석의 6 할 가량을 차지하였다.

"할"이 체언으로 사용될 경우 모든 경우가 수사(숫자)가 앞에 온다. 위의 형태적인 용례에서와 같이 특정 단어나 기능어를 제한 하는 것이 아니라 특별히 구문할 수 있는 (여기에서는 숫자)가 어절과 항상 함께 사용된다. 그래서 숫자가 오면 "할"은 체언으로 보고 아니면 "하다"에서 활용된 것으로 본다.

◆ 의미적 용례 제한

품사는 다르지만 문장에서 같은 기능을 해 같은 문장 성분으로 사용되면 통사적으로 쉽게 구분이 되지 않지만 어절의 의미상 자주 같이 쓰이는 어절이 적게 있을 때 이런 어절을 참조하여 중의성을 해결하는 것을 의미적 용례 제한이라 한다.

예) 온 : 온 (체, 관형사), 오다(용, 동사) + ㄴ (관형격 어미)
우리는 위대한 민주 장정에 온 국민의 동참과 성원을 바란다.
국방 태세가 완벽하다는 것을 온 세계에 알리려 하고 있다.
*) 조용한 아침의 나라에서 온 식구들에게 아주 만족하고 있다.

온(체) 는 "오다"가 활용한 "온" 문장 안에서 같이 뒤에 체언을 꾸민다. 온(용)은 "오다"에서 활용했기 때문에 '에서 '오다 라는 것으로 자주 사용되지만 결정을 할 수 있을 만큼 확실한 정보가 되지 못하다. 온(체)의 '모두, 전부' 라는 의미가 있기 때문에 "온"과 자주 사용되는 단어(국민, 민족, 세계등)가 있다 이런 정보를 지원하기 위해서는 명사가 그 의미에 따라 분류가 되어 있어야한다.
*)에서 알 수 있듯이 '식구'는 "온"과 같이 쓰이어 자연스럽지만 '에서'라는 조사가 앞에 있어 "온"이 '오다', '온' 두 경우로 다 사용될 수 있어 의미, 개념 분석이 있는 후에라야 구별이 가능하다.

위에서 같이 용례적인 정보를 이용해 중의성을 해결하려 할 때 예외적인 경우(주어진 정보가 오류를 범하는 경우)가 생길 수 있다. 이 때 이런 어절은 이 단계에서 해결을 하지 않고 구문 구조 분석기가 두 표제어를 모두 참조하게 하지만 어떤 표제어를 먼저 참조 할 것인가에 대한 정보를 줄 수 있다. 또는 예외적인 경우에 대한 특별한 처리를 하여 어느 정도의 오류를 범할 가능성이 있어도 하나의 표제어를 선택하게 한다.

4.3. 구문 정보

구문 정보는 어절과 어절사이의 수식 관계를 말한다. 이 정보를 찾는 것이 구문 구조 분석기의 주요 목표이며, 구문 구조 분석이 안 된 경우 일반적인 구분 정보를 알기

는 힘들다. 그러나 어떤 어절들의 수식 관계, 구문 정보는 전체 문장을 보지 않고 앞뒤의 몇개의 어절만으로 판단할 수 있는 경우가 있다. 이렇게 중의성을 갖는 어절이 간단하게 앞뒤 어절과 수식 관계가 명확하게 나타나면 이를 중의성 해결의 정보로 사용할 수 있는데 이를 구문 정보라 한다.

예) 번 : 번(체, 의존 명사) , 벌다(용, 동사) + ㄴ(관형격 어미)
성급한 대복이 없는 지 다시 한 번 찬찬히 살펴 보자
문제점이 무엇인지 지난 번 사태를 통해 확실히 증명되었다.

"번"이 체언류(의존 명사)로 사용 될 때는 홀로 사용되지 않고 항상 앞에 "이, 그, 저"등의 관형사와 "첫, 한"등의 수사 "지난" 등의 관형격 어미를 동반한다. 이러 관형사, 관형격어미, 수사 다음에는 "벌다"에서 활용한 "번"은 올 수 없다.
위의 예에서는 중의성을 갖는 어절이 특수한 품사(의존 명사)나 특별한 범주에 있어서 앞에 오는 어절에 따라 중의성을 해결한다.

예) 안 : 안(체, 부사) 알(용, 동사) + ㄴ(관형형 어미)
글쎄 별 문제가 안 될 걸 가지고 그러는 것 같다.
더 잘 알면서 일체 대답을 안 했다.

만약 "안"이 부사로 사용될 때는 뒤에 몇개의 부사가 사이에 오더라도 반드시 동사가 와야한다. 그러나 "안"이 알다(용언)에서 관형형 어미가 활용한 것이면 뒤에는 반드시 체언이 와야한다.

대개의 어절은 위에서 열거한 용례적 정보 구문적 정보중 어느 하나만을 가지는 것은 아니다. 용례적 정보를 이용해서 상당 부분을 해결하고 예외적인 부분은 구문적인 정보를 사용해서 결정한다. 구문적인 정보로도 어느 정도 해결할 수 있지만 특히 용례에서 많이 타나나는 용례들을 추가적으로 사용할 수 있다.

5. 처리 정보 구성

체언류와 용언류로 두가지 이상의 원형의 표제어로 분석될 수 있는 어절이 5500개나 되며, 어절에 대한 중의성을 해결하는 정보는 각 어절마다 다르다. 각 어절마다 서로 다른 정보를 그대로 이용하면 한 어절의 중의성을 해결하기 위해 하나의 처리 함수가 필요하다. 이렇때 표제어 중의성 해결기는 5만개 이상의 함수가 필요하다. 이런 중의성 해결기는 효율이 떨어질 뿐 아니라 새로운 정보를 추가하거나 기존의 정보를 갱신 하기가 어렵다.
중의성 해결기가 간단하고 효율적이며 확장성을 가지기 위해서는 해결기가 이용하는 정보가 사전에 효과적으로 있어야한다.
각 어절의 정보는 기본적인 몇개의 정보를 하나 또는 그이상을 이용해 표현 가능하다. 정보를 더이상 조개면 해결기에 필요한 정보로소 의미가 없어지게 되는 단위(atomic)로 나누면 모든 정보는 이들 단위 정보들의 Boolean Expression으로 나타날 수 있다. 이런 단위 정보들을 어느 정도 발견한다면 처리하고자 하는 어절이 늘어난다

해도 단위 정보를 늘릴 필요없이 기존의 정보를 조합해서 나타낼 수 있다.

단위 정보를 찾는 함수를 단위 함수라 하며 이것이 사전에 수록된다.

헤결기가 이용하는 사전은 아래와 같이 중의성을 갖는 어절을 표제어로 하고 표제어 아래 각 필드(Field)는 분석 가능한 원형과 이 원형이 분석되기 위해 만족해야하는 기본 함수들의 Boolean Expression이 있다.

```
표제어 : 원형 1 :(fre1): B_expr1(fn1_1(param1_1,,), fn2_2(param1_2,,),,)
(어절)   원형 2 :(fre2): B_expr2(fn2_1(param2_1,,), fn2_2(param2_2,,),,)
           ........
           ........
         원형 N :(freN): B_exprN(fnN_1(paramN_1,,), fnN_2(paramN_2,,),,)
```

위에서 fre는 나타나는 말뭉치에서 실제로 나타나는 빈도를 나타낸다. Fn은 단위 함수를 나타내며 이 함수에 매개 변수를 전달할 수 있다. 현재 이용하고 있는 대표적인 단위함수는 다음과 같으며 함수 이름은 번호로 나타나는데 이는 간단하게 사전에 수록하기 위한 것이다.

10(string) : 앞에 있는 체언의 조사가 매개 변수와 같으면 참.
100(number) : 이 원형은 통계적 정보로 해결하며 확률은 number * 100
88(category) : 뒤에 오는 어절이 품사 분류가 매개변수와 같으면 참.

사전 정보의 실제 예는 다음과 같다.

```
일 :     일:(150): 87(수사) V 86(관형격)
         일다:(2):11(주격)
         이다:(1):11(목적격)
   87() : 앞에 있는 어절의 품사 검사
   86() : 앞에 있는 용언의 어미 분류 검사
   11() : 앞에 있는 체언의 조사 분류 검사
```

위의 예에서 처리기가 "일"이라는 어절을 입력으로 받으면, 사용 빈도수에 따라 순서화 되어있는 필드를 순서대로 검사한다. 즉, 앞의 어절의 품사가 수사이거나 용언의 관형격어미가 활용된 것이면 '일(체)'로 판단하고 그렇지 않으면 다음 필드를 검사한다.

6. 결론

사전이 자연 언어 처리에서 중요한 위치를 차지하게되면서, 표제어를 효과적으로 찾는 것도 중요한 위치를 갖게 된었다. 표제어 중의성이라는 문제를 제기하게 됨으로써 형태소 분석기의 정확한 기능(사전의 표제어를 정확하게 찾는일)을 알게 되었으며, 한국어 구문 구조 분석기가 나아가야 할 방향이 사전 중의성 해결이 아니라 구문 구조 그 자체를 결정이라는 것을 확실하게 했다.

아직 한국어 구문 구조 분석기가 대량의 말뭉치에서 실험되지 않았고 표제어 중의성을 해결하지 않고 구문 구조를 분석할 때 얼마나 어려움이 있는지 알수 없고, 표제

어 해결기도 한 어절에 대한 정보를 찾는 데 많은 시간아 걸려 실제 말뭉치에서 실험할 정도가 되지 않아 표제어 해결기에 대한 평가는 직관적일 수 밖에 없다.

그러나 표제어 해결기가 확실한 위치를 가지려면 다음과 같은 문제점들에 대한 충분한 고려가 있어야 한다.

우선, 과연 표제어 중의성 해결이라는 단계가 특별히 필요한가? 중의성을 가진 어절을 그대로 입력으로 받아서 구문 구조 분석기가 구조를 정하면서 중의성을 해결하는 것이 좋지않은가? 이 것은 어떠한 구문 분석 전략을 사용할 것인가와도 연결되어 교려되어야한다.

다음으로 구문 구조 단계 이전에 표제어 중의성을 해결하는 것이 가능한가? 완전한 구문 구조 분석이 없이 중의성을 해결할 수 있는 어절이 매우 적다면 이 단계의 처리는 의미가 없을 것이다.

끝으로 문장의 완전한 구조를 알지 않고서 중의성을 해결한다면 얼마 만큼 신용할 수 있할 수 있나? 모든 정보를 말뭉치에서 찾아내 때 말뭉치에서 나타나는언어 현상과 잘 나타나지 않는 어떤 문법사이에서 말뭉치에서 자주 등장하는 것을 택할 때 어떤 근거를 가질 수 있나? 용례적 제한등 본 연구에서 찾은 말뭉치적인 정보들이 언어 처리 영역이 확대될 때에도 어느 정도 일반성을 갖고 확장될 수 있을까?

참고 문헌

[1] Small, Cotrell, Tanenhaus, "Lexical Ambiguity Resolution", Morrgan Kaufman Publishers, Inc, 1988

[2] James Allen, "Natural Languagee Understanding", Bejamin/Cummings Publishing Company, Inc., 1977

[3] 최현배, "우리 말본", 정음 문화사, 1989

[4] 정찬섭외 공저, "새 한국어 사전 편찬을 위한 사전 편찬학 연구", 연세 대학교 한국어 사전 편찬실, 1990

[5] 형태소 분석기, 용례 분석기, 연세 대학교 한글 정보 처리 연구실

[6] 조성원 송만석,"사전에 기반한 한국어 문장 해석 시스템 원형의 연구", 1991 가을 학술 발표 논문집 정보 과학회 예정

[7] 배순일 송만석, "HPSG를 이용한 한국어 구문 구조 분석기에 관한 연구", 1991 가을 학술 발표 논문집 정보 과학회 예정

[8] 남기심, "표준 국어 문법론", 탑출판사, 1990

[9] 신기철, 신용철, "새우리말 큰사전", 삼성 출판사, 1978