효율적인 한국어 형태소분석 방법

정일형, 양기주, 김영환
한국통신 소프트웨어연구소 인공지능연구실

An Efficient Method on Korean Morphological Analysis

I.H.Jung, G.J.Yang, Y.W.Kim
AI Section, Software Research Laboratories, Korea Telecom
Tel: (02)526-5914, Fax: (02)526-5909
E-mail: ihjung@aistar.kotel.co.kr

요 약

본 논문은 효율적인 한국어 형태소분석 방법을 제안한다. 기존의 형태소분석 방법에서는 분석속도와 분석정도가 상호보상 관계에 있으므로 형태소분석기가 이용되는 분야에 따라서 다른 분석방법이 사용되고 있다. 본 논문에서 제안한 형태소 분석 알고리즘은 하나의 어절을 이루는 형태소들 사이의 구성원리를 이용하여 각 어절 타입을 예측하고 각 타입에 적합한 분석을 함으로써 적은 회수의 형태소 분할로도 정확한 형태소분석이 가능하게 한다. 본 알고리즘은 많은 문장으로 형태소 분석 실험을 하였고 그 실험 결과는 기존의 방법 보다 우수하여 분석속도와 분석정도에 있어서 범용성이 입증되었다. 본 논문은 효율적인 형태소분석 방법을 제시하고 이를 반영한 형태소분석 시스템의 설계 및 구현에 관하여 기술한다.

I. 서론

소프트웨어 전쟁을 목하에 두고서 우리가 국내 소프트웨어 시장에서 일어설 수 있는 분야는 한글처리 부분이 아닌가 싶다. 그러므로 한글처리에서 가장 먼저 해결되어야 해결되어야 하고 한글 응용프로그램의 기반기술이 될 수 있는 범용 형태소 분석기의 구축이 시급하다 하겠다. 최근 들어서 모든 응용프로그램을 한글화 하고 있지만 아직은 자체 기술의 확보가 미흡하다 하겠다. 정확하고 빠른 형태소분석기의 구현은 철자 검색/교정기, 띄어쓰기 검사/교정기, 정보검색 시스템을 비롯한 응용프로그램의 기반기술일 뿐만 아니라 한글을 필요로 하는 모든 부분의 바탕이 된다. 본 논문은 정보검색과 관련하여 분석정도와 분석속도에 있어서 우수한 형태소 분석 방법을 고안하고 이의 구현에 관하여 설명한다.

II. 기존의 형태소분석 방법

기존의 형태소분석 방법이 형태소분석 결과의 완전성과 분석속도의 향상을 위해서 각기 개발되어 왔다. 그 결과 이 둘의 관계는 상보관계에 있어서 분석 결과의 질을 높이자면 처리속도가 떨어지고 빠른 처리속도를 얻자면 분석정보의 손실이 초래되었다. 이 둘의 문제가 형태소분석의 대표적 방법인 CYK와 최장일치에서 단점으로 지적되는 것이다. 본 형태소분석기는 하나의 어절을 이루는 형태소들 간의 구성원리(어절구조규칙)를 사용하여 어절 타입의 예측을 통한 효율적인 분석을 한다. 기존의 CYK 알고리즘은, 접속정보를 만들기 어렵고 많은 처리 시간이 필요하다는, 두가지 문제점을 안고 있다. 첫번째 문제는 포항공대의 잘 정리된 접속정보를 사용하여 해결하였고, 두번째 문제인 CYK 알고리즘 자체에서 비롯된 많은 처리시간은 본 논문에서 제안하는 효율적인 방법으로 해결하였다. 본 형태소 분석기는 분석정도와 분석속도에 있어서 이 둘을 모두 충족시키는 범용성을 지닌다.

III. 본 시스템의 형태소분석 방법

3.1 사전

조사와 어미를 별개의 사전 (이후에는 조어사전이라 부른다)에 등록시키고 이들의 복합형태도 등록시켜 가능한 조사와 어미의 분석이 (디스크에 저장된) 사전을 검색하지 않고 주기억장치의 조어사전 검색만으로 분석가능하도록 한다. 복합조사(또는 복합어미)는 원칙적으로 모두 사전에 등록한다. 사전에 등록하지 않고 이들 간에 결합정보를 부여하여 처리할 경우는 이들에 대한 접속정보계층을 구성하기 어렵다. 또한 이들은 여러개가 복합적으로 접속이 가능하다. 따라서 복합형태를 따로 분리하여 처리하지 않기로 한다 [1]. 본 시스템은 조어사전의 정보를 통하여 어절 타입을 예측하고 이에 따라서 어절, 음절 또는 자소 단위의 사전 검색을 한다. 그러므로 잘 정리된 조어사전은 본 시스템의 성능에 중요한 역할을 한다.

3.2 어절구조규칙의 이용

본 논문은 한국어 형태소 분석에 어절구조규칙을 이용한다. 본 논문에서 제안한 형태소 분석 알고리즘은 어절안에서 형태소 구성의 분석을 형태소간의 결합형태에 대한 예측을 통하여 적은 회수의 형태소 분할로도 알맞게 형태소를 분리할 수 있는 방법을 제시한다.

한국어 어절 생성전이도에서 알 수 있듯이, 모든 어절은 <명사상당어구+조사>, <용언상당어구+어미>, 또는 <단독 어절> 형태를 가진다 [2]. 그리고 하나의 문장을 이루는 어절의 타입 중에서 <명사상당어구+조사>가 <용언상당어구+어미>나 <단독어절> 보다 많은 것을 쉽게 알 수 있다. 서술어가 하나 이상의 <명사상당어구+조사>를 가지기 때문이다. 이는 실험에서 입증되었다.

3.2.1 어절타입의 예측

하나의 어절에서 우좌분석을 하여 조사 또는 어미를 조어사전에서 검색후, 어절 구조규칙에 따라서, 그것이 조사라면 그 앞을 명사상당어구로, 어미라면 용언상당어구로 예측하여 각각에 알맞은 방법으로 분석한다. 어절타입의 예측은 3.2.2.의 미등록어처리나 IV의 개선된 CYK 알고리즘이 가능케 한다. 어절구조규칙을 예측할 수 있기 위해서는 3.1에서 설명한 것처럼 잘 정리된 사전이 준비되어야 한다.

3.2.2 미등록어처리

미등록어인 경우에, 기존의 시스템에서처럼 <미등록어>라는 결과를 내지 않고 조어사전의 정보를 이용하여 미등록어의 품사를 예측하여 결과를 낸다. 예를 들면 "킬라베스체육관에서"를 분석하면 기존의 시스템은 미등록어처리를 하는 반면에 본 시스템에서는 [명사? + 조사]의 결과를 낸다. 그리고 이 예측을 통하여 정보검색의 고유명사 처리에 상당히 좋은 결과를 얻었다.

3.2.3 어절타입의 평가

어절구조규칙을 사용하여 어절타입에 맞지 않는 분석결과는 제거한다. 예를 들어 "이"의 분석에 있어서 기존의 알고리즘에서는 "이"가 조용보조어간이나 조사로 분석되기도 하지만 어절구조 규칙에 따르면 조용보조어간이나 조사는 단독으로 어절의 시작 위치에는 올 수 없기 때문에, 본 알고리즘에서는 이를 틀린 분석으로 처리한다.

IV. 개선된 CYK 알고리즘

CYK 알고리즘의 단점은 분석시간이 많이 걸린다는 것이다. 그 근본 원인은 CYK 알고리즘이 한 어절의 분석 가능한 모든 경우를 다 분석한다는 데 있다. 그런데 이 이유가 오히려 CYK의 장점으로 여겨지므로 이 장점을 충분히 살리면서도 속도를 향상시키는 것이 본 논문이 제안하는 CYK의 개선된 점이다. 그러면 아래에서 CYK 알고리즘의 단점을 분석해 보고 이를 개선시키기 위한 방법을 살펴본다.

4.1 기존 CYK의 문제분석 및 해결안

예를 들어 "건설업체를"이라는 어절을 분석하면 (<건설+업체+를>과 <건설업체+를>)을 결과로 얻는다. 이 경우 가장 진보된 CYK 알고리즘을 사용하여 [3] 분석하였을 경우에도 총 22번의 디스크를 검색한다. 그 중에서 총 13번은 불필요한 디스크의 검색이다. 이를 분석해 보다.

첫째, (결과적으로) <명사상당어구+조사>의 어절 분석에, 어미인 'ㄹ'과 그 앞 자소들의 결합형태를 전체어절의 분석 끝까지 사전검색 함으로써 총 8번의 불필요

한 검색이 이루어진다. 그 예는 아래와 같다.

: ㄹ, 체르, ㅂ체르, 업체르, ㄹ업체르, 설업체르, ㄴ설업체르, 건설업체르

둘째, 종성과 그 뒤 음소들의 결합형태를 사전검색 함으로써 총 5번의 불필요한 검색이 이루어진다. 그 예는 아래와 같다.

: ㅂ체, ㄹ, ㄹ업체, ㄴ설, ㄴ설업체

첫째와 같은 결과가 초래된 것은, 어미인 'ㄹ'로 <명사상당어구+조사>의 어절 형태인 어절을 분석하려고 했기 때문이다. 이는 어절형태의 예측이 전혀 없었기 때문이다. 본 알고리즘에서는 조어사전의 정보로 이를 예측할 수 있으므로 어절형태를 예측 못한 결과인 불필요한 검색을 피할 수 있다.

둘째와 같은 결과가 초래된 것은, 명사상당어구는 음절 단위의 구성을 이룸에도 불구하고 자소 단위로 분석했기 때문이다. 본 알고리즘에서는 <명사상당어구+조사>로 예측되는 어절의 분석에 있어서는 자소 단위가 아닌 음절 단위의 CYK 분석을 함으로써 불필요한 검색을 피할 수 있다.

그러므로 기존의 총 22번의 디스크 검색에 반하여, 기존의 알고리즘은 9번의 필요한 검색만 하고 이는 2번의 주기억장치에 적재된 조어사전의 검색과 7번의 디스크의 사전 검색이다. 2번의 조어사전 검색에서는 'ㄹ'과 '를'을 검색하고 나머지 7번 중에서 전체어절인 "건설업체를"을 제외하면, 명사합성어를 분석하기 위해 필요한 사전검색은 6번에 불가하다. 그 예는 아래와 같다.

: 체, 업체, 설, 설업체, 건설, 건설업체

4.2 본 시스템에서 사용한 CYK

본 논문에서는 CYK 알고리즘의 적용에 있어서 한국어 어절구조규칙을 기반으로 어절타입을 예측하고 그 타입에 맞게 어절 , 음절, 자소 단위의 분석방법을 제안한다. 기존의 형태소 분석이 모두 자소 단위로 처리된 이유는 한국어 용언의 불규칙이 음절 단위로 발생하는 것이 아니라 자음과 모음에 대해 발생하므로 음절 단위로 처리할 경우, 사전검색 회수가 용언의 불규칙처리로 인하여 오히려 더 늘어나기 때문이다 [4]. 그렇다면 <용언상당어구+어미>의 경우를 제외한 형태의 어절은 자소 단위로 분석할 필요가 없다. 오히려 이 경우 자소 분석은 더 많은 사전검색을 초래한다. 본 시스템에서는 어절 형태의 추측을 통하여, <단독 어절>은 어절, <명사상당어구+조사>는 음절, <용언상당어구+어미>는 자소 단위로 형태소 분석을 한다.

<명사상당어구+조사>의 어절 분석에 있어서, 음절 단위의 CYK를 사용하면 모든 가능한 결과를 얻을 수 있고, 최장일치를 사용하면 최장의 명사복합어의 결과를 얻을 수 있어서 모든 가능한 결과는 아니지만 그만큼 더 분석속도를 향상시킬 수 있다. 위의 "건설업체"의 경우 최장일치를 사용하면 "건설업체"가 사전에 있으면 한 번 (건설업체) "건설"과 "업체"가 사전에 있으면 4번 (건설업체, 건설업, 건설, 업체)의 검색만이 필요하다. 본 시스템에서는 범용성을 위하여, 음절 단위의 CYK 분석을 통하여 가능한 명사들의 결합을 구하고, 후처리에서 정보검색을 위하여 그 중에서 최장의 결과를 사용한다.

4.3 본 알고리즘의 우수성

본 알고리즘의 우수성의 뒷받침은 다음과 같이 정리할 수 있다.

첫째) 잘 정리된 조사, (선어말)어미사전을 가진다.
둘째) 이를 사용하여 어절의 타입을 추측할 수 있다.
세째) 전체 어절 타입 중에서 <명사+조사>의 형태가 상당한 비중을 차지한다.
네째) <명사+조사> 중에서 복합명사가 상당한 비중을 차지한다.
다섯째) 각각의 어절타입에 알맞은 방법으로 분석한다.
여섯째) 어절구조규칙을 이용하여 잘못된 분석결과를 제거한다.
일곱째) 어절타입의 추측을 통하여 미등록어 [illegible]사를 추정한다.

V. 실험 결과

본 형태소분석기의 성능검사를 위하여, 200개의 신문기사에서 총 15243개의 어절을 SUN4 (CYPRESS CY7C601 SPARCRISC Microprocessor 33.33 MHz Solbourne Series 700)에서 테스트하였다. 어절 타입을 살펴보면, 이 중에서 <명사상당어구+조사>가 9617 (63%)개, <용언상당어구+어미>가 3546 (23%)개, <단독 어절>이 2080 (14%)개 이었다. 그리고 이 데이타를 분석하는데, 기존의 CYK 알고리즘으로는 약 95분 (0.37초/어절)이 걸렸고 본 시스템 (KSmarT)으로는 약 25분 (0.09초/어절)이 걸렸다. 결과적으로 본 시스템의 분석속도는 기존의 시스템 보다 약 4배의 속도 향상을 가져왔다. 또한 분석결과에 있어서도, 어절구조규칙의 이용과 미등록어의 처리로 인하여 기존의 시스템에 비하여 좋았다.

VI. 결론

본 논문에서 제안하고 구현한 알고리즘은 첫째, 알고리즘 측면에서 CYK 알고리즘을 개선하여 분석속도를 향상시켰고 둘째, 어절구조규칙을 이용하여 형태소 분석 정도를 향상시켰고 세째, 형태소 분석정도와 분석속도에 있어서 범용성과 우수성이, 많은 어절의 형태소분석 실험을 통하여 입증되었다. 본 알고리즘은 형태소 분석에 있어서 상당한 효율성을 가져왔다. 앞으로 더 나은 형태소분석 속도를 위해서 명사합성어의 분석에 휴리스틱을 적용하여 개선할 것이다. 또 형태소분석에 구문정보를 이용함으로서 그 분석정도의 질을 높이는 것이 앞으로 남은 과제이다.

참고문헌

[1] 이은철, "CYK법에 기반한 한국어 형태소 분석에서의 개선기법," 석사학위논문, 포항공과대학, 1993.

[2] 임희석, 이호, 임혜창, "형태소 분석 단계에서 발생하는 어절의 중의성 분석 방안," 제20권 1호, pp.769-772, 1993.

[3] 김은자, 이종혁, "일-한 기계번역 시스템의 구현 : 휴리스틱을 이용한 일본어 형태소 해석 기법," 한국정보과학회 학술발표논문집, 제20권 1호, pp.797-800, 1993.

[4] 최재혁, 이상조, "양방향 최장일치법을 이용한 한국어 형태소 분석기," 한국정보과학회 학술발표논문집, 제20권 1호, pp.769-772, 1993.