웹의 협업 환경을 이용한 확장 형태소 사전 관리*

황인범○ 이동주 연종흠 이상구

서울대학교 컴퓨터공학부

{inbeom, therocks, jonghm, sglee}@europa.snu.ac.kr

Managing Extended Morpheme Dictionary through User Collaboration on the Web

Inbeom Hwang○ Dongjoo Lee Jongheum Yeon Sang-goo Lee

Department of Computer Science and Engineering, Seoul National University

형태소 분석은 한국어 자연어 처리에 반드시 필요하다. 워드 프로세서에서 맞춤법을 검사할 때나, 문서에서 검색 엔진이 사용할 색인어를 추출할 때 등 여러 방면에서 형태소 분석을 널리 사용하여 왔으며, 최근 오피니언 마이닝(Opinion Mining) 등 정보 추출에 대한 요구가 생기면서 활용 범위가 점차 넓어지는 추세이다. 형태소 분석은 자연어 문장을 구조화하는 과정의 첫 단계인데, 자연어가 담고 있는 정보를 최대한 보존하고 활용하기 위해서는 정확도 높은 형태소 분석 방법을 이용하여 문장 구성 요소를 분석해내야 한다. 형태소 분석의 활용 범위가 넓어지는 것과 동시에 형태소 분석을 적용해야 할 데이터 또한 많아지고 있다. 웹을 통해 생산되는 자연어 데이터 양이 빠른 속도로 증가하고 있기 때문이다. 많은 데이터를 자동화 처리하기 위해서, 형태소 분석기는 정확한 결과를 보여야 할뿐만 아니라 빠른 속도로 분석을 수행할 수 있어야만 한다. 한국어를 다른 언어와 비교할 때 형태소 분석의 중요성이 더욱 두드러진다. 영어 등 고립어로 분류할 수 있는 언어들과 달리, 교착어로 볼 수 있는 한국어는 어근이 접두사 및 접미사와 결합하여 다양한 형태로 변화하기 때문에 복잡한 형태소 분석 알고리즘이 필요하며, 한국어의 이러한 특성 때문에 뛰어난 형태소 분석기를 구현하기 어려웠고, 현재까지 많은 연구자가 이 문제 해결을 위해 노력하여 왔으며 우수한 연구 결과들이 배출되어 왔고 널리 이용되어 왔다.

특정 형태소 분석 방법의 우수성을 평가할 때 단순히 정확도와 분석 속도를 측정하는 것과 함께, 문장에 담긴 언어적 오류에 대응하는 능력도 중요한 요소로 고려해야 한다. 자연어 데이터에는 많은 오류가 있게 마련이다. 언어 사용 형태는 꾸준히 변화하고 이러한 변화는 일상 생활에서도 흔히 볼 수 있다. 새로운 어휘나 용법이 등장할 때, 이들을 기존 형태소 분석기를 이용해 분석하면 오류가 발생할 가능성이 크다. 또한 문장 작성자가 부주의로 인해 틀린 문장을 쓰거나 오타가 발생했을 때도 올바른 분석이 불가능하다. 형태소 분석기는 이러한 오류에 대응할 수 있어야 하며 대응을 통해 작성자가 가진 의도를 충분히 반영하는 분석 결과를 내어놓을 수 있어야 한다. 문장에 담긴 오류는 웹에서 얻을 수 있는 자연어 데이터에서 빈번하게 찾아볼 수 있다. 빠른 정보 생산 주기로 인해 작성자들이 문장을 주의 깊게 검토하지 않는 경우가 많고, 이로 인하여 단순한 오류나 용법에 맞지 않는 문장이 다수 존재한다. 그러나 기존 형태소 분석기들은 바른 문장을 분석한다는 것을 전제해왔기 때문에, 기존 분석기들은 웹과 같이 오류가 많은 환경에서 사용하기 어렵다는 문제점을 안고 있다. 지금까지 여러 형태소 분석기들이 정확한 문장을 잘 분석해내는 것에 초점을 맞추어 왔다면, 앞으로는 용법에 맞지 않는 문장을 분석하여 문장에 담긴 정보를 보존할 수 있는 방안 또한 고려 대상이 되어야 한다.

형태소 분석기는 단순한 패키지 프로그램과 같이 인식되고 개발되어 왔다. 그러나 개발 주기가 긴 패키지 프로그램으로는 변화하는 언어 사용 양상에 빠르게 대응하여 형태소 분석기의 분석 품질을 높게 유지하는 것이 불가능했다. 이 문제를 해결하기 위해서는 유지보수가 용이하여 여러 오류에 빠르게 대응할 수 있는 형태소 분석기 개발 체계를 갖추어야 한다.

보통 형태소 분석기는 분석 알고리즘과 이 알고리즘이 참조하는 사전으로 이루어지는데, 구현이

* 본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 육성・지원사업 (NIPA-2010-C1090-1031-0002)의 연구결과로 수행되었음.

필요한 알고리즘보다는 사전 정보가 유지보수하기에 용이하다는 점은 쉽게 짐작 가능하다. 본 연구에서는 사전 관리를 용이하게 하여 실용적이고, 언어 사용 양상이 변화하더라도 높은 분석 품질을 꾸준히 제공할 수 있는 형태소 분석기 '꼬꼬마(KKMA, http://kkma.snu.ac.kr/)'를 구현하였다. 꼬꼬마는 인접 조건 검사에 의한 고속 형태소 분석 방법[1]에 기반하였는데, 이 방법은 분석 정확도가 높을 뿐 아니라 다른 분석 방법들에 비해 빠른 분석 속도를 보여서 양이 많은 데이터를 처리하는 데 적합하다. [2]를 비롯한 기존 형태소 분석 방법은 자소 단위로 결합이 이루어지는 형태소를 찾기 위해 접두사 및 접미사가 결합하는 부분의 음절을 자소 단위로 나눈 후 사전을 탐색하여 분석을 시도한다. 이에 반해, [1]은 음절 단위에서만 연산이 이루어지는데 이로 인해 사전 탐색 횟수가 적고 빠른 분석 속도를 보인다. 확장 형태소 사전[3](기분석 사전)을 이용해 음절 단위 분석이 가능하다. 기존 형태소 분석기들이 이용하는 형태소 사전과 대조적으로 확장 형태소 사전은 어간의 원형뿐만 아니라 접두사, 접미사 결합으로 인해 변형된 어간들이 갖는 형태도 모두 포함하며, 형태소 분석기는 간단한 조건 검사를 통해 이 항목들을 적절히 결합한 분석 결과를 내어놓는다. 때문에 사전 내용이 언어학적 사전과 다르고 포함하는 항목 수가 많다. 또한 각 항목과 결합되는 항목의 조건을 명시하기 위해 품사 결합 조건, 음운 결합 조건 등 언어적 규칙까지 사전에 기록해야만 한다. 이러한 변형과 언어 규칙은 대부분 자동화하여 기록할 수 있으나, 실용적으로 사용하기 위해서는 사전 관리에 많은 노력을 들여야만 한다. 그러나 분석 알고리즘이 단순한 동적 프로그래밍 형태로 구현할 수 있을 정도로 간단하고 형태소 분석에 필요한 언어적 정보를 모두 사전에 담을 수 있어 형태소 분석기의 기능과 언어적 정보를 뚜렷하게 분리하여 관리할 수 있다. 형태소 분석기 개발자에게 유지보수가 까다로운 알고리즘 대신, 사전 정보만을 관리하여 형태소 분석기 동작을 제어할 수 있기 때문에 알고리즘에 대한 의존도가 큰 다른 방법보다 형태소 분석기 유지보수에 적합한 구조라고 볼 수 있다.

사전 정보를 적절히 유지하는 것은 좋은 형태소 분석 결과를 얻기 위해 반드시 필요한 작업이다. 꼬꼬마는 API를 공개하여(Open API) 웹에서 협업을 통해 확장 형태소 사전을 관리할 수 있도록 한다. 사용자들은 공개 API를 사용해 사전 항목을 검색, 추가, 수정, 삭제하는 것이 가능하다. 사용자들은 분석 오류를 발견했을 때 그에 대응하는 사전 항목을 추가하거나 수정하여 올바르지 않은 분석 결과를 바로잡을 수 있다. 또한 사용자들이 사전 항목을 직접 수정하는 것이 불가능한 상황이나 논의가 필요한 오류에 대해서는 오류 보고 API를 사용하여 사전 수정을 요청할 수 있다. 이렇게 공개된 사전 관리 방법을 이용해 형태소 분석기 개발자만에 의한 품질 유지를 수행하는 것보다 효율적으로 사전 품질을 유지할 수 있다. 형태소 분석기를 적용하는 영역이 크게 다르고 각 영역에서 사용하는 어휘가 뚜렷한 차이를 보이는 경우에는 이러한 접근 방식이 정확도를 떨어뜨리는 요인이 될 수도 있으나, 같은 언어를 사용하는 환경에서 그런 경우는 찾아보기 힘들다. 또한 그런 경우에는 특정 응용의 형태소 분석기만이 사용하는 사전을 오프라인에서 추가로 관리하는 것도 가능하다. 형태소 분석기 개발자나 응용 개발자가 사전을 관리하는 경우, 처리해야 할 데이터가 많을 때 분석 품질을 유지하는 작업은 더욱 어려워진다. 한글 형태소 분석이 필요한 응용 영역이 다르더라도 데이터가 많아질수록 비슷한 오류에 대응하는 빈도가 커질 것이고, 이러한 관점에서 볼 때 사전을 관리하는 사용자가 많아질수록 품질 개선의 가능성이 높아질 것이라는 점은 명백하다. 여러 사용자가 변경한 사전 내용은 바로 확장 형태소 사전에 반영되어 틀린 형태소 분석 결과를 보정한다. 이는 사전 관리에 집단 지성을 발현하고자 한 첫 시도로 평가할 수 있다.

수정이 쉽고 발전 가능성이 높은 자유 소프트웨어인 꼬꼬마는 기존 형태소 분석기 구조에 확률 모델을 도입하고 잘못된 띄어쓰기를 보정[4]하기도 하는데, 장차 이러한 알고리즘을 개선하여 다양한 오류 상황에 대응할 수 있을 것으로 기대된다. 또한 라이브러리 형태로 어떤 응용에서나 편리하게 사용할 수 있기 때문에 앞으로 폭 넓은 연구 및 응용에서 활용할 수 있을 것이다.

참 고 문 헌

[1] 심광섭, 양재형, 인접 조건 검사에 의한 초고속 한글 형태소 분석기, 정보과학회논문지: 소프트웨어 및 응용 제31권 제1호 pp. 89-99, 2004. 1

[2] 강승식, 음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 컴퓨터공학과 박사학위 논문, 1993. 2

[3] 양승현, 김영섬, "부분 어절의 기분석에 기반한 고속 한국어 형태소 분석 방법", 정보과학회논문지: 소프트웨어 및 응용 제27권 제3호, 2000. 3, pp.290-301

[4] 강미영, 정성원, 권혁철, 어절 내의 형태소 범주 패턴에 기반한 통계적 자동 띄어쓰기 시스템, 정보과학회논문지: 소프트웨어 및 응용 제33권 제11호, 2006. 11