Khann2 : 경험기반 고효율 한국어 품사태깅 도구

신창욱○ 박성재 차정원
창원대학교
papower2@gmail.com, tjdwo1289@gmail.com, jcha@changwon.ac.kr

Khann2 : The Korean Highly efficient ANNotation tool for POS corpora

Chang-Uk Shin○, Seong-Jae Park, Jeong-Won Cha
ChangWon National University

요 약

언어처리에서 사용되는 말뭉치를 제작하는 작업은 시간과 비용이 많이 들어간다. 본 논문에서는 품사부착 말뭉치 제작 시에 시간을 줄일 수 있는 도구를 제안한다. 태깅 효율을 높이기 위해 다음과 같은 기능을 구현하였다. 다중 사용자 태깅 환경 구축, 사용자 경험 기반 자동 오류 표시, 용례검색, 프로젝트 기반 관리, 집단 지식을 이용한 도움말, 사전 검색, 관리자 기능이다.

1. 서 론

인터넷 문서가 증가하면서 품사부착에 대한 요구는 증가되었다. 품사부착 말뭉치 작성은 이런 품사부착 시스템의 성능 향상을 위해 반드시 필요한 과정이다. 그러나 말뭉치 작성은 시간과 노력이 많이 든다. 따라서 효율적인 작성 도구의 개발이 요구된다. 현재 대부분의 품사부착은 전용 도구가 아닌 일반 에디터를 사용하여 사용자가 직접 품사를 수정하는 작업을 통해 이루어지고 있다. 일반 에디터를 이용하여 품사부착 말뭉치를 작성할 경우, 동시작업, 일관성 검사, 할당된 문서관리, 도움말 등에 어려움이 존재한다.

이처럼 수작업을 통한 말뭉치 작성의 어려움에 따라 최근 몇 년간 국내외 여러 분야에서 관련 분야에 대한 다양한 언어분석을 위한 기본 말뭉치 가공 도구들이 개발되었다.

현재 개발된 국외 도구에는 ‘MATE workbench’, 타이완의 ’ Opinion Annotation Tool(OAT)’, 미국의 ‘Annota -tion Graph AP’, 핀란드의 ’ DepAnn’ 등이 있다.

MATE workbench[1]는 음성 데이터나 문서 데이터를 가공하여 분석결과를 부착하고 표현하며 질의를 던질 수 있는 도구이다. 이 도구는 사람이 쉽게 가공된 문서를 생성할 수 있도록 설계되었으며 다른 사람들과 쉽게 공유하여 작업을 진행할 수 있고 다양한 부착 방법을 사용할 수 있다.

Opinion Annotation Tool(OAT)[2]는 의견을 부착할 수 있는 도구이다. 이 도구는 그래픽 사용자 인터페이스 형식으로 개발된 도구로서 각 레벨(단어, 부분문장, 전체문장)별, 주제별 의견 부착이 가능하다. 이 도구는 중국어, 영어, 일본어에 대한 문서 작성이 가능하다.

Annotation Graph AP[3]는 Time-Series Data에 대한 언어정보를 부착할 수 있는 도구이다. 이 도구는 사용자가 특별한 분야의 가공도구를 신속히 개발할 수 있도록 그래프를 이용하는 기반구조를 제공한다. 따라서 응용 프로그램 인터페이스, 입출력 라이브러리, 그래픽 인터페이스 등을 제공한다.

DepAnn[4]는 Dependency treebanks를 위한 언어정보 부착도구로 그래픽과 텍스트 기반 인터페이스를 제공한다. 사용자가 의견을 추가할 수 있도록 하며, 트리 구조를 손쉽게 이동 및 수정이 가능하다. 또한, 자동 일관성 검사를 통해 문장 구조와 주석, 부호화 검증을 수행한다.

솔트룩스의 ‘Semano’[6]는 자연어 처리와 텍스트 마이닝 기술에 기반, 사람이름・기업명・주소・기술명・제품명・프로젝트 등의 개체명과 사건・원인・결과・상황(상태) 등의 단위지식 정보를 추출하고, 온톨로지 인스턴스 생성 및 의미 테마데이터 생성 기능을 제공하는 시스템이다. 이 Semano에서는 각 도메인에 맞는 다양한 개체명 사전을 로딩하고 자체 개발한 XRE 문법과 규칙에 기반하여 의미 정보를 추출할 수 있다.

그러나 이러한 도구들은 표준화된 언어정보를 부착하는 것이 아닌 각자 기준에 근거하여 작성되고 있어 자원을 공유하는데 한계를 가진다.

본 논문에서는 21세기말뭉치 계획 프로젝트[7] 품사태그 기준에 따라 말뭉치를 작성하고, 다중 사용자 환경 및 자동 학습 기능을 가진 품사부착 도구(Khann)[1]를 제안한다. 본 논문에서 제안한 도구에서는 사용자가 각각의 어절에 대한 후보를 입력하지 않고, 미리 생성된 후보를 단순 선택하는 방식을 택함으로써 오류를 줄이고 시간을 단축할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 구현된 도구에 대한 구조 및 주요 기능을 설명한다. 3장에서 결론을 내린다.

---

1) https://air.changwon.ac.kr/khann/에서 다운로드가 가능하다.

2. 주요 기능 설명

말뭉치 제작 도구들이 작업에 편리한 인터페이스를 제공한다면 약간의 훈련만으로도 충분히 도구를 다룰 수 있어 말뭉치 작성에 효율적일 것이다. 그러나 도구 사용에 능숙한 사용자라도 관련 분야의 지식이 부족하다면 작업에 많은 제약이 따르게 된다.[8)]

특히, 품사 부착의 경우 정확도 높은 작업을 위해 다양한 작업이 동시에 이루어져야 하므로 작업시간의 증가 및 혼란과 번거로움이 발생할 수 있다.

이러한 문제점을 최소화하기 위해 본 논문에서는 사용자 경험을 학습할 수 있는 품사 부착 도구를 개발하였다. Khann에서는 품사 부착 시 필요한 다양한 작업을 한 작업 공간 내에서 모두 이루어질 수 있도록 구성하고, 완료된 문서에 대해 빠른 확인 및 수정이 가능하다. 또한, 다른 사람과 작업 결과를 공유하고 의견을 달거나 다른 사람의 의견을 확인할 수 있다.

2.1 다중 사용자 환경 구축

일반적으로 품사부착 작업은 다수의 인원이 참여하는 작업이다. 따라서 이들이 동시에 작업하면서 정보를 공유하는 환경을 제공해야 한다. Khann은 다중 사용자가 동시에 작업을 하면서 서로의 결과를 참조할 수 있고 정보를 공유할 수 있는 환경을 제공한다.

2.2 제시된 후보로부터의 선택

품사부착 작업에서 발생하는 오류의 많은 부분은 장시간 작업으로 인한 작업자의 오류에서 비롯된다. Khann은 어절에 대해 생성된 후보를 제시하고 선택만 하면 작업이 완료되도록 함으로써 이러한 오류를 최소화하도록 설계했다. 만약 알맞은 후보가 존재하지 않을 경우, 직접 입력할 수도 있도록 하였다.

2.3 사용자 경험기반 자동 오류/후보 표시 기능

Khann은 제시된 어절이 사용자들에 의해서 빈번히 수정되었을 때 추가 작업 시 사용자에게 알려주어 품사 부착 작업의 효과를 증가시킨다. 오류가 빈번히 발생하는 부분에 대해서는 사용자가 더 관심을 가지고 작업을 할 수 있을 뿐 아니라, 추천 후보에 대한 반복적 오류 수정을 통해 작업의 질을 향상시킬 수 있다.

Khann은 또한 사용자가 작업한 내용이 반복적으로 나타날 경우, 작업을 기억하여 반복적 선택을 최소화할 수 있도록 설계했다.

또한 제시된 후보와 용례 데이터베이스를 비교하여 후보에는 없고 용례에는 있는 후보가 있을 경우 최빈 용례 데이터베이스 후보를 제시하여 추가입력을 최소화 하였다.

2.4 용례 검색 기능

용례는 품사부착 작업 시에 많은 도움을 주는 정보이다. Khann은 부착하고자 하는 어절에 해당하는 용례를 자동으로 검색하여 옆에 표시해준다. 이렇게 함으로써 작업자의 추가적인 작업이 필요없게 구성하였다.

2.5 집단 지식을 이용한 도움말 확장

집단 지식을 이용한 도움말의 확장은 용례검색의 확장이라고 말할 수 있다. 태깅 중 어절을 선택할 때 마다 자동으로 출력되는 코멘트 형식의 집단 지식은 품사부착 작업의 효율을 높일 수 있다. 또, 어절이 아닌 형태소마다 코멘트를 작성함으로써 효율을 높이고자 하였다.

2.6 사전 및 품사 표기법 검색

단어마다 다양한 의미가 존재할 수 있다. 품사 부착 작업을 하면서 용례검색을 통한 도움말을 이용할 수 있지만, 사전적인 의미가 필요한 경우 국어사전을 검색해야 할 필요가 있다. 이를 위해 국어사전 검색 기능을 추가하여 어절의 문맥적 의미와 품사를 참조할 수 있어 작업이 편리하도록 한다. 작업과 동시에 검색이 가능하며, 단축키를 이용하여 팝업창으로 추가 페이지를 호출할 수도 있다.

2.7 프로젝트 단위 관리

일정 수의 사람이 함께 말뭉치를 제작하는 프로젝트에 참여하게 된다. 이 모임은 모두 전문가일 수도 있지만 초보자가 포함될 수도 있다. 따라서 프로젝트에서 일관성 문제가 발생할 수 있다. 본 도구에서는 작업을 프로젝트 단위로 생성할 수 있게 설계하였다. 생성된 프로젝트에는 작업자들 등록할 수 있고 문서를 배분할 수 있으면 일관성 검사를 할 수 있다. 프로젝트에 참여하는 숙련자들은 초보자들의 결과를 수정해 줄 수 있고 이러한 과정을 통해 작업을 능률을 향상시킬 수 있다.

2.8 관리자 기능

관리자는 품사부착 작업을 위해 사용자에게 파일에 대한 권한을 할당하거나 취소하고, 요청한 파일에 대한 처리 및 회원 관리가 가능하다.

회원으로 등록된 사용자에게 작업할 파일에 대한 권한 처리가 가능하다. 또 원활한 로그아웃 처리가 되지 못했을 경우 관리자가 그 사용자의 접속을 끊거나 불필요한 계정의 탈퇴처리가 가능하다. 기존 회원의 경우, 회원의 정보 수정 및 등급 조정이 가능하다.

도움말기능(용례검색)에 문제가 발생하거나 존재하지 않는(이미 작업 완료 후 제거된) 파일의 정보가 데이터베이스에 들어가 있어 사용자들에게 잘못된 도움말(용례)를

제공할 경우, 관리자는 데이터베이스에 접속하여 용례 수정도 가능하다.

3 결론

본 논문에서는 사용자의 경험을 기반으로 한국어 품사 태깅의 효율성을 향상시킬 수 있는 도구에 대해서 기술하였다. 본 논문의 기여점은 다음과 같다.

1. 사용자 경험기반 고빈도 오류어절 표시
2. 집단 지식을 이용한 용례사전 및 의견 활용
3. 분석 후보 중 선택을 통해 최종 분석어절 결정
4. 웹기반 공동 작업 환경 구축
5. 자동 일관성 검사를 통한 효율성 증가

국내의 경우 언어처리를 위한 말뭉치 가공 도구가 일반적이지 않다. 하지만 많은 양의 데이터를 활용하고자 하는 요구는 증가하며, 품사 부착 작업의 처리 효율성을 높이기 위해 편리한 도구의 필요성 또한 높아지고 있다. 지금까지 개발된 도구들은 동시에 많은 사용자들을 관리하기 어려우며, 느리고, 사용자들 간의 정보 공유가 어렵다는 것 등의 단점을 가지고 있다.

본 논문에서 제안한 도구는 설계 시 이러한 단점들을 극복하기 위한 기능들의 구현에 중점을 두었다. 또한 실험을 통해서 본 도구의 효과와 필요성을 보였다.

본 논문에서 제안한 툴에서 제공되는 일관성 유지 문제의 경우 문맥에 대한 일관성 자동 검사 기능만을 제공하고 있다. 그러나 작업할 동일 파일을 다른 사용자에게 할당한 후 그 결과를 서로 비교한다면, 일관성 유지 문제에 있어 더욱 높은 성능을 기대할 수 있을 것이다. 또한, 품사 부착 결과의 평가 방법으로 전문가가 작업한 품사 부착 결과를 바탕으로 각 사용자의 결과와 비교하여 추가적인 평가가 이루어진다면 사용자 평가와 동시에 각 문서의 정답 문장(또는 어절) 확률로서의 문서 신뢰도를 평가할 수 있을 것이다. 이와 같은 방법들을 이용하여 평가방법을 보안한다면 품사 부착 작업에서 효율은 더욱 증대될 것이다.

4 참고 문헌

1. Multilevel Annotation Tools Engineering, http://mate.nis.sdu.dk/, 2000.
2. Lun-Wei Ku, Yu-Ting Liang and Hsin-Hsi Chen, “Opiion Extraction, Summarization and Tracking in News and Blog Corpora“, AAAI, p. 100-107, 2006.
3. sourceforge, AGTK : Annotation Graph Toolkit, http://agtk.sourceforge.net/(2003).
4. Tuomo Kakkonen, “DepAnn - An Annotation Tool for Dependency Treebanks“, ESSLLI, Vol.11, p. 214-225, 2006.
5. 차정원, 일반화된 한국어 미등록어 추정, 석사학위 논문, 1999.
6. Saltlux, http://in2.saltlux.com/, 2006.
7. 국립국어원, 21세기 세종계획, http://www.sejong.or.kr, 2007.
8. 이인근, 황도삼, 권순학, 온톨로지 구축 시스템 (An Ontology Construction System), 한글 및 한국어 정보처리 학술대회, Vol.18, p. 220-227, 2006.