세종 형태분석 말뭉치의 오류 수정 도구 개발

최명길○, 남유림, 서형원, 전길호, 김재훈
한국해양대학교, 컴퓨터공학과
cmg5478@naver.com, zin1987@nate.com, wonn24@gmail.com, asone7705@nate.com, jhoon@hhu.ac.kr

Developing an Error Correction Tool for Sejong POS Tagged Corpus

Myung-Gil Choi○, Yoo-Rim Nam, Hyung-Won Seo, Kil-Ho Jeon, Jae-Hoon Kim
Korea Maritime University

요 약

한국어 정보처리에서 널리 사용되는 세종 형태분석 말뭉치는 품사정보와 문장정보 등 다양한 한국어 정보를 포함하고 있다. 이 말뭉치는 방대한 양의 정보들로 구축되었지만 많은 오류 또한 포함되어 있다. 예를 들면 철자 오류, 띄어쓰기 오류, 그리고 품사부착 오류 등이 있다. 하지만 세종말뭉치와 같이 대용량 말뭉치의 오류를 수정하는 것은 많은 인력과 시간이 필요하며 일관성 있게 오류를 수정하는 것은 쉽지 않다. 따라서 본 논문에서는 세종 형태분석 말뭉치에 포함된 오류를 빠르고 일관성 있게 수정하기 위한 오류 수정 도구를 구현하였다. 본 논문에서 수정 대상이 되는 오류는 어절과 형태소 분석 결과의 불일치에 관한 오류만 대상으로 한다. 이를 위해 세종 형태분석 말뭉치를 데이터베이스로 재구축하였으며, 본래의 어절과 품사가 부착된 형태소의 자모를 각각 분리하여 두 자모의 차이점을 분석하여 오류 후보를 선정한다. 오류 후보에서 동일한 오류 패턴을 갖는 모든 오류 후보에 대하여 동일한 방법으로 일관성 있고 빠르게 수정할 수 있다.

주제어: 세종 형태분석 말뭉치, 오류 수정 도구

1. 서론

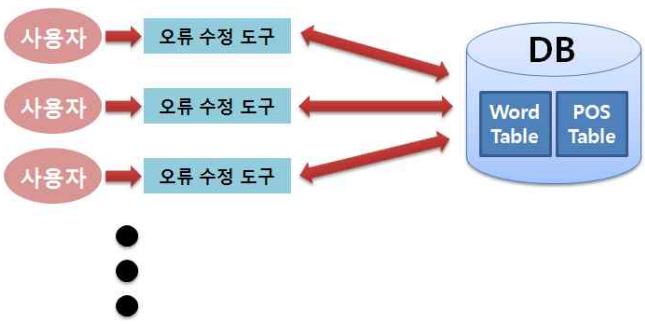
현재 한국어 정보처리 연구자가 쉽게 이용할 수 있는 말뭉치 중에는 세종 말뭉치[1]가 있다. 세종 말뭉치는 원시 말뭉치, 형태분석 말뭉치, 형태의미분석 말뭉치 그리고 구문분석 말뭉치를 포함하고 있으며 이런 다양한 말뭉치를 사용하면 기계번역, 질의응답 시스템, 정보검색, 문서분류 그리고 텍스트마이닝과 같이 다양한 시스템을 개발하는데 용이하다. 하지만 현재 구축된 세종말뭉치는 몇몇 문제점을 가지고 있다. 품사가 잘못 부착 되었거나 문장 내에서 단어가 잘못 분리된 경우, 그리고 불필요한 단어의 삽입 또는 삭제된 경우 등의 오류를 포함하고 있다. 이러한 오류가 포함된 말뭉치를 수정 없이 사용할 경우 앞서 언급한 시스템들의 좋은 성능을 기대하기 어렵다[2].

이러한 오류를 수정하는 방법은 쉽지 않다[3]. 다양한 패턴의 오류가 존재하고 오류의 개수 또한 매우 많기 때문이다. 또한 오류를 수정하기 위해서는 많은 인력과 시간이 필요하며, 결과적으로 많은 비용이 들게 된다. 많은 비용을 들여 수정한 말뭉치에는 여전히 오류가 존재할 수 있다. 여러 사람이 말뭉치를 수정하기 때문에 동일한 오류를 다른 방식으로 수정할 수 있기 때문이다. 그러므로 최대한 같은 정보에 대한 오류를 일관성 있고 효율적으로 수정하기 위해서 오류 수정 도구를 개발하였다.

본 논문은 세종말뭉치에서 약 1500만 어절로 구성된 형태 분석 말뭉치에 포함된 오류를 빠르고 일관성 있게 수정하기 위한 방법을 제시하고 구현한다.

이 논문은 2장에서 오류 수정 도구를 어떻게 구현하였는지 소개하고 3장에서 결론을 맺는다.

2. 오류 수정 도구 구현



(그림 1) 시스템 구성도

오류 수정 도구의 구성도는 (그림 1)과 같다. 데이터베이스는 Word table과 POS table로 구성되어 있다. 사용자는 오류 수정 도구를 통해 데이터베이스에 접속하여 데이터를 수정한다.

2.1 데이터베이스의 구축

오류 수정 도구를 구현하기 세종 형태분석 말뭉치를 여러 사람이 데이터를 수정하기 위하여 데이터베이스로 재구축하였다. 세종 형태분석 말뭉치에는 (그림 2)와 같은 형태로 표현되어 있다. 이 같은 형태의 데이터를 (그림

	①	②
BTAA0001-00000011	넓혀	넓히/VV + 어/EC
BTAA0001-00000012	프랑스의	프랑스/NNP + 의/JKG

(그림 2) 세종 형태분석 말뭉치 표기법

sid	wid	subwordid	word	history	etc	pattern
3	BTAA0001-00000011	1	넓혀	0	0	ㅓ<>ㅣㅇㅓ
3	BTAA0001-00000012	1	프랑스의	0	0	NULL

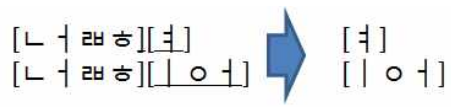
(그림 3) 데이터베이스의 Word table

sid	wid	subwordid	seq	morph	pos	history	etc
3	BTAA0001-00000011	1	1	넓히	VV	0	0
3	BTAA0001-00000011	1	2	어	EC	0	0
3	BTAA0001-00000012	1	1	프랑스	NNP	0	0
3	BTAA0001-00000012	1	2	의	JKG	0	0

(그림 4) 데이터베이스의 POS table

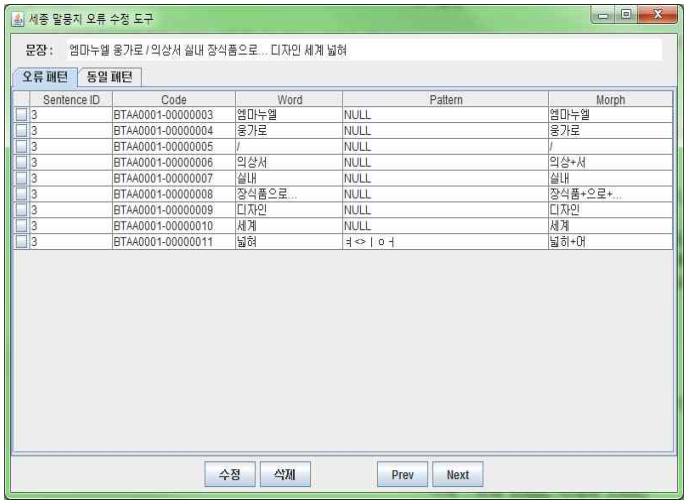
3)과 **(그림 4)**의 데이터베이스 테이블로 각각 변환하였다.

Word table과 POS table에는 sid, wid, subwordid, history 필드를 포함하고 있다. sid는 문장 식별번호(sentence id), wid는 각 어절에 대한 어절 식별번호(word id), subwordid는 띄어쓰기가 되지 않은 경우, 어절을 분리하는 과정에서 분리되었음을 표시하기 위한 필드이다. 그리고 history는 오류를 수정한 이력정보를 기록한다. 이력정보는 history필드의 기본 값에 −1을 해줌으로써 동일한 정보(sid, wid, subwordid)를 검색하였을 때 history의 값이 가장 작은 레코드가 최근에 수정된 레코드가 된다. 그리고 수정한 레코드를 삽입할 때에는 다시 history의 기본 값인 0으로 삽입하게 되는데 이때 값이 0이라는 것은 오류가 없는 어절임을 의미한다.

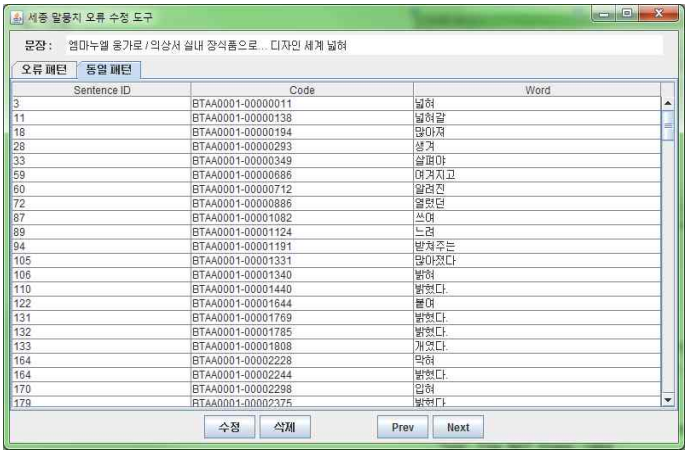


(그림 5) (그림 2)의 ①과 ②의 자모분리 결과를 비교

(그림 3)에서 Word table의 "pattern" 필드는 처음 형태분석 말뭉치를 구축하면서 발생할 수 있는 오류유형을 찾아내기 위하여 추가하였다. "pattern" 필드는 **(그림 2)**에서 본래의 어절 ①과 분석된 형태소 ②의 차이를 계산한다. 이는 ①과 ②의 자모분리를 통해서 **그림 5**와 같이 나타낼 수 있다. **(그림 2)**의 ②에서 부착된 품사를 제거하고 "넓히"와 "어"를 다시 결합한 후에 이것을 자모분리 할 경우 **(그림 5)**의 "[ㄴㅓㄹㅎㅣㅇㅓ]"와 같은 결과를 얻을 수 있다.



(a) 오류 후보를 포함하는 문장



(b) (a)와 동일한 오류 유형의 레코드 목록

(그림 6) 세종 형태분석 말뭉치의 오류 수정 도구

"넓혀"의 자모분리 결과와 비교할 경우 "[ㄴㅓㄹㅎ]"가 일치하며 두 어절의 차이점이 "[ㅕ]"와 "[ㅣㅇㅓ]"임을 알 수 있다. 이러한 과정을 통하여 오류 후보를 결정하며, **(그림 6)의 (b)**와 같이 오류 후보를 확인할 수 있다. 앞에서 언급한 방법으로 데이터를 수정할 경우 우리말의 특성인 불규칙 활용에 대한 패턴을 알 수 있고, 사용자가 불규칙 활용에 대한 패턴임을 인지할 경우 동일한 패턴을 수정하지 않고 데이터를 보존함으로써 올바른 불규칙 활용을 잘못 수정하는 상황을 줄일 수 있다[4]. **(그림 4)**에서 POS table은 **(그림 2)**에서 ②을 품사단위로 분리하여 데이터베이스를 구현한다. POS table에는 seq필드를 따로 포함하고 있다. seq필드는 같은 어절이지만 품사로 인해 분리될 경우 순서를 기록해 놓은 필드이다. 예를 들어 **(그림 2)**에서는 "넓히"와 "어"로 분석되었는데, 이때 "넓히"의 seq값은 1이고 "어"의 seq값은 2이다.

2.2 오류 수정

(그림 6)의 (a)는 기본적으로 오류후보가 포함되어 있는 문장을 데이터베이스로부터 불러들이고 오류라고 판단되는 레코드를 직접적으로 선택하여 수정할 수 있다. 그리고 동일한 패턴을 가지는 레코드를 **(그림 6)의 (b)**와 같이 볼 수 있다. 이때 사용자는 어절과 패턴을 확인함으로써 불규칙 활용으로 인한 패턴인지 오타로 인한 패턴인지 확인할 수 있다.

레코드를 수정하기 위해서 사용자가 레코드를 선택하고 수정버튼을 누르고 편집 상태로 넘어가며 이를 저장함으로써 데이터베이스에 업데이트가 가능하다. 데이터가 수정된 이후에 다음 수정되지 않은 오류패턴을 다시 데이터베이스에서 불러온다.

이런 작업을 반복적으로 함으로써 말뭉치의 오류를 줄이고 수정작업을 통한 2차적인 오류발생을 줄일 수 있다.

3. 결론 및 향후 연구

본 논문에서는 세종 형태분석말뭉치에서 발생하는 오타를 빠르고 일관성 있게 수정하기 위한 오류 수정 도구를 구현하였다. 오류 수정 도구를 개발하기위해 기존에 텍스트로 구축된 말뭉치를 데이터베이스로 재구축하였으며, 본래의 어절과 품사가 부착된 어절을 자모를 각각 분리하여 두 자모의 차이점을 분석하여 오류 후보를 결정한다. 오류 후보에서 동일한 오류 패턴을 갖는 모든 오류 후보에 동일한 방법으로 일관성 있고 빠르게 수정할 수 있다.

오류 수정 도구를 사용해서 말뭉치를 수정하는 시간은 여전히 짧지 않다. 그러므로 사용자가 좀 더 빠르고 편리하게 데이터를 수정할 수 있도록 모든 작업을 단축키로 수행 가능하도록 사용자 인터페이스를 개선해야 한다.

참고문헌

[1] 21세기 세종계획, http://www.sejong.or.kr

[2] 김형준, 임동희, 강승식, 은지현, 장두성, “세종 계획 말뭉치를 이용한 품사 태거의 성능 개선”, 한국정보과학회 2007 한국컴퓨터종합학술대회 논문집, 제34권, 제1호, pp. 177-180 2007.

[3] 김은혜, 최기선, “품사 부착 코퍼스 수정 방안에 대하여”, 제12회 한글 및 한국어 정보처리 학술대회, pp. 361-367, 2000.

[4] 김재훈, 서형원, 전길호, 최명길, “세종말뭉치의 오류 수정 방법”, 한국마린엔지니어링학회 공동학술대회 논문집, pp. 435-436, 2010.