SMS 변형된 문자열의 자동 오류 교정 시스템

강승식○ 장두성
국민대학교 컴퓨터공학부, KT 미래기술연구소
sskang@kookmin.ac.kr○, dschang@kt.co.kr

Automatic Error Correction System for Erroneous SMS Strings

Seung-Shik Kang○ Du-Seong Chang
School of Computer Science, Kookmin University; Advanced Technology Laboratory, KT

요 약

휴대폰과 메신저 등 통신 환경에서 사용되는 표준어가 아닌 SMS의 변형된 어휘 및 띄어쓰기 오류를 자동으로 교정하여 형태소 분석 및 품사 태깅의 성능 저하 문제를 방지하는 문자열 오류의 교정 방법을 제안하였다. 통신 어휘들의 문자열 사전 구축 방법으로 통신어 휘집을 기반으로 수동으로 구축하는 방법과 수작업으로 구축된 말뭉치로부터 자동으로 변형된 문자열을 추출하는 방법, 그리고 문맥을 고려하는 방법을 비교-분석하고 실험 및 성능 평가 결과를 제시하였다.

1. SMS 문자열의 오류 교정

SMS 영역에서 빈번히 발생하는 변형된 문자열 및 어휘의 왜곡 현상은 음성 인식, 음성 합성 등 언어처리 응용 시스템에서 SMS 문장을 처리하는 것을 어렵게 한다 [1,2]. 이러한 문제점을 해결하기 위하여 변형된 문자열을 인식하고 정규화 및 띄어쓰기 등 전처리 단계를 통해 문법에 맞는 문자열로 변환하는 작업을 통해 SMS 문자열의 오류를 교정하는 연구를 수행하였다. 본 논문에서는 어근 변형 오류, 문법형태소 변형 오류, 음소 단위 어미 오류, 어절 단위 오류, 형태소 분석에 의한 오류 교정 등 각 단계별 오류 교정을 통해 SMS 문자열의 오류를 자동으로 교정하는 방법을 제안하고자 한다.

문법형태소 및 음소 단위 오류 교정

SMS 문자열 변환을 위한 조사 및 어미 관련 문법형태소 관련 변형 문자열 사전은 통신언어 어휘집(문화관광부, 2001)에 수록된 문법형태소 오류어를 사전으로 구축하여 문자열 변환에 적용하였다. 구축된 사전의 크기는 조사 및 어미를 포함하여 720개이다. 문법형태소 변환은 어휘 변환 후, 자동 띄어쓰기 후 이렇게 두 차례에 걸쳐 이루어진다. SMS 문자열 변환을 위한 변형된 어말어미 중에서 'ㅆ/ㄴ/ㄹ/ㅁ/ㅂ' 으로 시작되는 것은 입력 문장에 대한 음소 단위 분할이 선행되어야 한다. 따라서 음소 단위의 변환이 필요한 문자열을 음소단위로 어말어미 176개를 구축하였다. 음소 단위 변환 사전을 적용할 때 적용되는 문자열의 길이가 짧아서 변환 오류가 발생하는 경우가 있다. 변환 문자열의 패턴이 과다 적용됨으로 인하여 발생하는 변환 오류를 방지하기 위하여 "변환 오류 방지 사전"을 도입하여 이 사전에 수록된 문자열은 변환이 되지 않도록 하였다.

어절 단위 오류 교정

SMS 문자열을 변환할 때 어떤 문자열들은 substring 단위로 적용할 경우 변환 오류가 발생하기도 한다. 이 경우에 어절 단위로 구분된 어절에 적용하는 것이 효율적인 경우에 자동 띄어쓰기를 적용한 후에 어절 단위로 오류를 교정한다. 어절 단위의 변환이 필요한 문자열을 사전으로 구축하였으며, 그 예는 '나두 → 나도', '낼 → 내일'과 같다. 어절 단위 변환을 적용할 때 변환 오류를 최소한으로 줄이기 위해 변환 대상이 되는 우측 문맥정보를 확인하여 변환 조건을 검사할 필요가 있다. 예를 들어, '낼'을 '내일'로 무조건 변환하는 것은 과다 변환 오류를 유발한다. 따라서 변환 조건으로 우측 어절이 보통명사인 경우에 한하여 변환하도록 제약 조건을 추가한다. 우측 어절의 품사태그를 고려한 변환을 하더라도 '~낼게', '~맘때쯤'과 같이 과다 변환되는 오류가 발생하는 경우가 있으며, 이러한 경우에는 "변환 오류 방지 사전(sma_err.dic)"에 변환 금지 문자열을 수록하여 변환이 적용되지 않도록 한다.

불필요한 어휘 제거

SMS 문장에는 불필요한 문자 및 기호들이 포함되기도 한다. 문장 내용과 무관하게 사용된 어휘들을 제거해

야 하는데 제거 대상이 되는 어휘는 제거 방법에 따라 스트링 단위, 어절 단위, 이모티콘 및 특수문자, 그리고 완성형 한글의 범위를 벗어나는 문자들로 구분된다.

2. 문자열 변환 모델

문자열 변환 시스템의 성능을 평가하기 위하여 데이터의 구축 방법에 따라 수작업으로 구축된 데이터를 적용한 모델-A, 문자열이 변형된 부분만 자동으로 추출한 데이터를 적용한 모델-B, 그리고 변형된 문자열의 앞뒤 음절 문맥을 고려하여 자동으로 추출한 모델-C로 구분하여 시스템을 구현하였다.

1) 모델-A. 통신 어휘집 기반으로 수동 데이터 구축

모델-A는 문화관광부 사업으로 구축된 통신언어 어휘집에 수록된 1,878개의 통신어휘를 기반으로 하여 변환 데이터를 구축하였다[3]. 또한, 8만 문장 규모의 단문 메시지 말뭉치에서 발견되는 변형된 문자열들을 수작업으로 추출하여 변형된 문자열 변환 사전을 구축한 것으로 최종 데이터 개수는 3,035개이다.

2) 모델-B. 변형된 문자열의 자동 구축

모델-B는 교체(substitution) 현상에 대한 변형된 문자열 쌍을 자동으로 추출하여 변형된 문자열 사전을 구축하였으며, 추출된 데이터 개수는 9,851개이다. 이 때 삽입(insertion) 및 삭제(deletion) 오류는 데이터 구축에서 제외하였는데, 그 이유는 삽입과 삭제된 문자열 데이터는 범용으로 사용하기에 부적합하여 시스템의 성능 개선에 도움이 되지 않았기 때문이다. 모델 B를 약간 변형한 모델-B'은 자동으로 추출된 데이터를 수작업을 통해 검토하여 변환 오류를 유발하는 문자열을 제거하는 작업이 반영된 것이다.

3) 모델-C. 문맥을 고려한 변형된 문자열의 자동 구축

모델-B는 문자열이 변형된 문장에서 앞뒤 문맥을 전혀 고려하지 않았다. 이에 비해, 모델-C에서는 앞뒤 각 두개의 음절을 변형을 위한 필요조건으로 하여 데이터를 구축한 것이다. 즉, 변환 문자열을 추출할 때 변형된 문자열 부분을 중심으로 선행 2음절, 후행 2음절을 추출하여 선행 및 후행 음절들이 일치하는 문맥에서만 문자열 변형 데이터가 적용되도록 하였다. 이 모델에서 사용한 말뭉치는 실험 및 성능 평가에 사용한 것과 동일한 말뭉치에서 추출된 것을 대상으로 하였다. 모델-C를 약간 수정하여 모델-C'와 모델-C''을 구성하였는데, 모델-C'은 앞뒤 문맥 음절의 길이를 2에서 1로 축소한 것이고, 모델-C''은 앞뒤 문맥 음절을 각각 선행 1음절과 후행 2음절, 그리고 선행 2음절과 후행 1음절로 수정하여 데이터를 구축한 것이다.

3. 실험 및 성능 평가

SMS 문자열 변환 성능 실험을 위하여 실제 사용자들의 문자 메시지로부터 구축한 SMS 단문 데이터 파일을 사용하였다. SMS 단문 말뭉치는 시스템 구축용과 성능 실험을 위한 실험 데이터로 구분하였다. 성능 평가를 위해 수집된 단문 메시지 15,241 문장에서 99%는 시스템 구축용 데이터로 사용하고, 나머지 152 문장(1,335 어절)을 실험 데이터로 사용하였다. 세 가지 유형의 문자열 변환 모델 및 혼합 모델에 대한 성능을 측정 실험을 수행한 결과는 표 2와 같다.

표 1. SMS 문자열 변환 성능 실험 결과

실험 모델	미변환 오류	변환 오류1	변환 오류2	변환 성공 (재현율 %)	정확률 (%)
A	52	5	2	71 (55.5%)	91.0%
A+B	50	7	2	71 (55.5%)	88.8%
A+B'	50	5	2	73 (57.0%)	91.3%
A+C	52	5	3	71 (55.5%)	91.0%
B	81	5	2	42 (32.8%)	85.7%
C	83	4	3	41 (32.0%)	85.4%
B+C	81	4	3	43 (33.6%)	86.0%
C'	73	5	5	50 (39.1%)	83.3%
C''	78	6	4	44 (34.4%)	81.5%

"미변환 오류"는 오류 문자열이 변환되지 않은 경우다. 이 오류는 전체 오류의 가장 많은 비중을 차지하는데 해당 어휘들을 사전에 추가함음로써 성능 개선을 기대할 수 있다. "변환 오류1"은 변환이 일어나기는 했지만 잘못 변환되어 올바른 교정이 일어나지 않는 경우다. 예를 들어 "눈이 온다. 것두 무지 많이…"의 입력에 대해서 "것두"를 단순히 "것도"로 변환했을 때 "그것도"로 변환하지 못해 올바른 변환이 일어난 것은 아니다. "변환 오류2"는 옳은 문자열을 변환하여 오답을 만든 경우를 의미한다. 중의성 어휘를 일괄 변환하는 경우 출현 빈도에 따라 자주 출현하는 어휘로 변환하지만 가끔 출현하는 어휘가 이처럼 잘못 변환되는 경우도 있을 수 있다. 또한,

참고문헌

[1] 차인태, "PC 통신 언어 분석", 음성과학, 8권 3호, pp.75-91, 2001.

[2] 임동희, 강승식, 장두성, "음성 인식 후처리를 위한 띄어쓰기 오류의 교정", 한국 컴퓨터 종합 학술대회 (KCC 2006) 논문집, Vol.33, pp.25-27, 2006.

[3] 조오현, 김경용, 박동근, "통신언어의 실태와 개선 방안", 통신언어 어휘집, 문화관광부, 2001.