SMS 변형된 문자열의 자동 오류 교정 시스템
(Automatic Error Correction System for Erroneous SMS Strings)

강 승 식 [†] 장 두 성 [††]
(Seung-Shik Kang) (Du-Seong Chang)

요 약 휴대폰과 메신저 등 통신 환경에서 문자 메시지를 전송할 때 표준어가 아닌 왜곡된 어휘들을 사용하고 있으며, 이러한 변형된 어휘들은 음성 인식, 음성 합성, 문서 정보 추출 등 언어처리 및 관련 분야의 응용 시스템에서 많은 문제점을 유발시킨다. 본 논문에서는SMS 문장들의 변형 및 띄어쓰기 오류를 자동으로 교정하여 형태소 분석 및 품사 태깅의 성능 저하 문제를 방지하는 문자열 오류의 교정 방법을 제안하고 시스템을 구현하였다. 시스템의 성능에 가장 큰 영향을 미치는 변형된 문자열 사전을 구축하는 방법으로 (1) 통신 어휘집을 기반으로 수동으로 구축하는 방법, (2) 수작업으로 구축된 말뭉치로부터 자동으로 변형된 문자열을 추출하는 방법, (3) 자동으로 변형된 문자열을 추출할 때 좌우 문맥을 고려하는 방법에 대하여 시스템을 구현하고 실험을 통하여 비교-분석 및 성능 평가 결과를 제시하였다.

키워드 : 통신 언어, 오류 어휘, SMS 문장, 형태소 분석

Abstract Some spoken word errors that violate grammatical or writing rules occurs frequently in communication environments like mobile phone and messenger. These unexpected errors cause a problem in a language processing system for many applications like speech recognition, text-to-speech translation, and so on. In this paper, we proposed and implemented an automatic correction system of ill-formed words and word spacing errors in SMS sentences that has been the major errors of poor accuracy. We experimented three methods of constructing the word correction dictionary and evaluated the results of those methods. They are (1) manual construction of error words from the vocabulary list of ill-formed communication languages, (2) automatic construction of error dictionary from the manually constructed corpus, and (3) context-dependent method of automatic construction of error dictionary.

Key words : spoken word, error word, SMS sentence, morphological analysis

1. 서 론

단문 메시지 서비스(SMS)는 전세계적으로 널리 사용되는 무선 서비스로써 전자메일, 호출, 음성 메일 시스템들과 같은 외부 시스템과 모바일 가입자 사이의 문자 및 숫자 메시지 전송을 가능하게 해준다. 그러나 근래 한글 축약형, 비속어를 남발해 만들어낸 정체 불명의 신조어들은 물론, 한글과 이상한 문자들을 혼용한 속칭 '외계어'들이 급속히 퍼지면서 문어체 위주의 기존 형태소 분석 시스템으로는 이러한 SMS 영역 문장을 정확하게 분석할 수 없는 문제점이 발생하고 있다[1-3]. 이에 따라 통신상에서 의도적으로 혹은 통신환경의 특성에 따라 언어가 왜곡되는 현상에 대한 연구가 수행되어 왔다[4-6]. 또한, SMS 영역의 한국어의 문장 분석 정확도를 향상시키기 위하여 사용자가 직접 입력하는 문장에서 사용자들이 의도적으로 혹은 실수로 빈번하게 발생하는 오류를 자동으로 교정하고, SMS 영역의 문장들에 대한 형태소 분석 및 품사 태깅 오류를 최소화하는 연구가 필요하다[7].

SMS 영역에서 빈번히 발생하는 변형된 문자열로 인

· 본 논문의 SMS 관련 연구는 KT 미래기술연구소의 지원을 받아서 수행되었음. 본 논문에서 필요한 데이타 정리 및 프로그램 구현을 위하여 국민대학교 컴퓨터공학부 오대성 군이 수고하였음
· 이 논문은 2007 한국컴퓨터종합학술대회에서 'SMS 변형된 문자열의 자동 오류 교정 시스템'의 제목으로 발표된 논문을 확장한 것임

† 종신회원 : 국민대학교 컴퓨터공학부 교수
sskang@kookmin.ac.kr
†† 정 회 원 : KT 미래기술연구소 HCI연구담당 수석연구원
dschang@kt.com

논문접수 : 2007년 9월 27일
심사완료 : 2008년 4월 22일

정보과학회논문지: 소프트웨어 및 응용 제35권 제6호(2008.6)

한 어휘의 왜곡 현상은 음성 인식, 음성 합성, 문서의 유형 정보 추출 등 언어처리 분야의 응용 시스템에서 SMS 문장을 처리하는 것을 불가능하게 하고 있다. 이러한 문제점을 해결하기 위하여 변형된 문자열을 인식하고 정규화 및 띄어쓰기 등 전처리 단계를 통해 가능한 문법에 맞는 문자열로 변환하는 작업을 통해 SMS 문자열의 오류를 교정하는 연구가 수행되고 있다[8-10]. 본 논문에서는 아래의 각 단계별 오류 교정을 통해 SMS 문자열에 포함된 오류를 자동으로 교정하는 방법을 제시하고자 한다.

- 문자열 오류 교정
- 문법형태소 오류 교정
- 음소 단위 어미 오류 교정
- 어절 단위 오류 교정
- 형태소 분석 후 어절 단위 변환
- 불필요한 어휘 제거

각 단계의 오류 교정은 단계에 맞게 사전을 구축하고 해당 사전을 바탕으로 문자열을 변환하도록 되어 있다. 각 변환 단계는 서로 독립적이지만 순서에 따라 최종 결과에 영향을 줄 수 있다. 본 논문의 구성은 다음과 같다. 2장에서는 문자열 교정 단계를 자세히 설명하고, 3장에서는 실험 및 성능 평가 결과를 기술한다.

2. SMS 문자열의 오류 교정

2.1 변형된 문자열의 오류 교정

입력된 SMS 문자열의 띄어쓰기 제거 후 처음 적용되는 변환이 어휘 변환 단계이다. SMS 문자열 변환을 위한 변형 문자열 사전은 통신언어 어휘집에 수록된 통신언어를 사전을 기반으로 구축하여 문자열 변환에 적용이 가능하도록 한 것이다. 통신언어 어휘집에서 추출된 문자열 사전의 크기는 3,035개이다[11,12].

```
낼부터 내일부터
먹구 먹고
수욜 수요일
어떳누 어떻니
넘좋다 너무좋다
```

그림 1 문자열 변환사전 예

```
[입력 문장] 목은 좀 어떴누? 약 챙겨먹구. 9시간 연속이면 힘들겠다.
[어휘 변환]목은좀어떻니?약챙겨먹고.9시간연속이면힘들겠다.
[자동 띄어쓰기] 목은 좀 어떻니? 약 챙겨 먹고. 9시간 연속이면 힘들겠다.
[최종 결과] 목은 좀 어떻니? 약 챙겨 먹고. 9시간 연속이면 힘들겠다.

[입력 문장] 많이 바쁘세요? 오늘도 못오시는건가요? 수욜인데. 낼부터는 열심히하자구요!
[어휘 변환] 많이바쁘세요?오늘도못오시는건가요?수요일인데.
내일부터는열심히하자구요!
[조사어미변환] 많이바쁘세요?오늘도못오시는건가요?
수요일인데.내일부터는열심히하자고요!
[자동 띄어쓰기] 많이 바쁘세요? 오늘도 못 오시는 건가요?
수요일인데. 내일부터는 열심히 하자고요!
[최종 결과] 많이 바쁘세요? 오늘도 못 오시는 건가요? 수요일인데. 내일부터는 열심히 하자고요!
```

그림 2 문자열 변환 예

2.2 문법형태소 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. 조사-어미 변환 사전의 크기는 약 720개의 데이타로 구성되어 있다. 문법형태소 오류의 교정은 어휘 변환 과정의 다음 단계에서 적용되고 있으며, 또한 자동 띄어쓰기를 적용한 후에도 적용하게 되는데 이는 적용률을 높이기 위한 것이다.

```
데루 대로
데염 데요
데영 데요
셔용 셔요
시쥬 시죠
힘내구 힘내고
```

그림 3 문법형태소 변환 사전 예

```
[입력 문장] 어제술마이드셨져? 괜찮으셔용? 오늘도수고하셔용 아자!
[어휘 변환] 어제술많이드셨져?괜찮으셔용?오늘도수고하셔용아자!
[조사어미 변환] 어제술많이드셨져?괜찮으셔요?오늘도수고하셔용아자!
[어미 변환] 어제술많이드셨죠?괜찮으셔요?오늘도수고하셔용아자!
[자동 띄어쓰기] 어제 술 많이 드셨죠? 괜찮으셔요? 오늘도 수고하셔용 아자!
[조사어미 변환] 어제 술 많이 드셨죠? 괜찮으셔요? 오늘도 수고하셔요 아자!
[최종 결과] 어제 술 많이 드셨죠? 괜찮으셔요? 오늘도 수고하셔요 아자!
```

그림 4 문법형태소 변환 예

2.3 음소 단위 어미 오류 교정

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 별도의 사전으로 구축하였으며, 그 예는 그림 5와 같다. 음소단위 어미 변환 사전의 크기는 음소 단위의 어말어미 176개가 구축되어 있다. 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(eomi_err.dic)"을 도입하여 이 사전에 수록된 문자열은

ㄹ꺼 ㄹ거
ㄹ께 ㄹ게
씨남 씨어

그림 5 음소 단위 어미 변환 사전 예

[입력 문장] 이사는 잘했남? 왜 연락안했엉. 도와주러 갈려구 했는뎅 집들이행. 선물해줄께
[어휘 변환] 이사는잘했남?왜연락안했어.도와주러가려구했는뎅집들이해.선물해줄께
[어미 변환] 이사는잘*했어*?왜연락안했어.도와주러가려구했는데집들이해.선물해*줄게*
[자동 띄어쓰기] 이사는 잘 했어? 왜 연락 안 했어. 도와주러 가려구했는데 집들이 해. 선물 해 줄게
[어휘 변환] 이사는 잘 했어? 왜 연락 안 했어. 도와주러 가려고했는데 집들이 해. 선물 해 줄게
[최종 결과] 이사는 잘 했어? 왜 연락 안 했어. 도와주러 가려고 했는데 집들이 해. 선물 해 줄게

그림 6 음소 단위 어미 변환 예

변환이 되지 않도록 하였다.

2.4 어절 단위 오류 교정

SMS 문자열 변환을 할 때 어떤 문자열들은 부문자열(substring) 단위로 적용할 경우 변환 오류가 발생하기도 한다. 이 경우에 어절 단위로 구분된 어절에 적용하는 것이 효율적인 경우에는 자동 띄어쓰기 모듈을 적용한 후에 어절 단위로 오류를 교정한다. 어절 단위의 변환이 필요한 문자열을 사전으로 구축하였다. 현재까지 구축된 어절 단위의 변환 문자열은 156개이다.

나두 나도
낼 내일

그림 7 어절 단위 변환 사전 예

[입력 문장] 그래. 나두영화보구싶따. 영화머보나? 영화본지가언젠지. 낼쉬니깐놀러와.
[어휘 변환] 그래.나두영화보고싶따.영화뭐 보나?영화본지가언젠지.낼쉬니깐놀러와.
[자동 띄어쓰기] 그래. 나두 영화 보고 싶따. 영화 뭐 보냐? 영화 본지가 언젠지. 낼 쉬니깐 놀러 와.
[어휘 변환] 그래. 나두 영화 보고 싶다. 영화 뭐 보냐? 영화 본지가 언젠지. 낼 쉬니깐 놀러 와.
[어절 변환] 그래. *나도* 영화 보고 싶다. 영화 뭐 보냐? 영화 본지가 언젠지. *내일* 쉬니깐 놀러 와.
[최종 결과] 그래. 나도 영화 보고 싶다. 영화 뭐 보냐? 영화 본지가 언젠지. 내일 쉬니깐 놀러 와.

그림 8 어절 단위 변환 예

2.5 형태소 분석 후 어절 단위 변환

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 제한할 수 있다. 형태소 분석에 의한 우측문맥 조건검사가 필요한 어절 단위 변환 문자열을 사전으로 구축하였다. 우측 어절의 품사 태그를 고려한 변환을 하더라도 '~낼게', '~맘때쯤'과 같이 과다 변환되는 오류가 발생하는 경우가 있으며, 이러한 경우에는 "변환 오류 방지 사전"에 변환 금지 문자열을 수록하여 변환이 적용되지 않도록 한다.

낼 내일 Mvb
넘 너무 H,D,B,Mvs,Mvb,T
마니 많이 H,D
맘 마음 D,Mvb,Me,H,B
머 뭐 H,D
멀 뭘 D
모 뭐 H,D
몰 뭘 H,D
함 한번 D

그림 9 형태소 분석 후 어절 단위 변환 사전

[입력 문장] 주말잘보내셨어요? 데이콤시외전화신청관련 함알아보셨는지요?
[자동 띄어쓰기] 주말 잘 보내셨어요? 데이콤 시외 전화 신청 관련 함 알아보셨는지요?
알아보/D+ 시/s+ 었/s+ 는지요/ez
[형태소 분석 후 변환] 주말 잘 보내셨어요? 데이콤 시외 전화 신청 관련 *한번* 알아보셨는지요?
[최종 결과] 주말 잘 보내셨어요? 데이콤 시외 전화 신청 관련 한 번 알아보셨는지요?

그림 10 형태소 분석 후 어절 단위 변환 예

2.6 불필요한 어휘 제거

SMS 문장에는 불필요한 문자 및 기호들이 포함되기도 한다. 따라서 문장 내용과 무관하게 사용된 어휘들을 제거해야 하는데 그 예는 그림 11, 12와 같다. 불필요한 어휘를 제거하는 방식은 띄어쓰기 단위인 어절 단위 제거와 부문자열 제거로 구분된다. 불필요한 문자의 많은 경우가 이모티콘이며, 이모티콘은 문자 메시지에 감정을 전달하는 효과가 있다. 그러나 문장 중심의 텍스트 처리

개쉑	TT
개코여	Zz
꾸벅	바보
뜨바야	쌍
바오앙	에고
씨바	에구
아흠	와와
젠장	우씨
쩝	으앙
캬캬	으유
흐흐흐	쩝
흑흑	쳇
히히힛	캬캬

그림 11 불필요한 어휘 제거 사전

```
뭐라고? (-_-) → 뭐라고?
황당한걸~ (@_@)→ 황당한걸
야호! (>_<) → 야호!
나 졸려~(~.-) → 나 졸려
```

그림 12 이모티콘 어휘 제거 예

에서는 이모티콘을 처리하기가 곤란하므로 제거한다. 그 방법으로 문장 끝의 4가지 문장부호('./,/?/!')를 제외한 모든 특수문자를 제거한다. 또 다른 유형의 특수문자에는 'ㅋㅋ' 등과 같이 완성형 한글의 범위에서 벗어나는 한글 문자들이 있다. 이와 같이 음절 구성이 안되는 문자들도 불필요한 문자로 간주하여 제거하였다.

3. SMS 오류 교정 모델

문자열 변환 시스템의 성능을 평가하기 위하여 프로그램을 실험할 때 적용된 데이타의 구축 방법에 따라 수작업으로 구축된 데이타를 적용한 모델-A, 문자열이 변형된 부분만 자동으로 추출한 데이타를 적용한 모델-B, 그리고 변형된 문자열의 앞뒤 음절 문맥을 고려하여 자동으로 추출한 모델-C로 구분하여 시스템을 구축하였다.

1) 모델-A. 통신 어휘집을 기반으로 수동 데이타 구축

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

2) 모델-B. 변형된 문자열의 자동 구축

모델-B는 SMS-1에서 교체(substitution) 현상에 대한 변형된 문자열 쌍을 자동으로 추출하여 변형된 문자열 사전을 구축하였으며, 추출된 데이타 개수는 9,851개이다. 이 때 삽입(insertion) 및 삭제(deletion) 오류는 데이타 구축에서 제외하였는데, 그 이유는 삽입과 삭제된 문자열 데이타는 범용으로 사용하기에 부적합하여 시스템의 성능 개선에 도움이 되지 않았기 때문이다. 이 모델을 약간 변형한 모델-B'은 자동으로 추출된 데이타를 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 데이타를 구축한 것이다. 즉, 변환 문자열을 추출할 때 변형된 문자열 부분을 중심으로 선행 2음절, 후행 2음절을 추출하여 선행 및 후행 음절들이 일치하는 문맥에서만 문자열 변형 데이타가 적용되도록 하였다. 이 모델에서 사용한 말뭉치는 SMS-2로서 실험 및 성능 평가에 사용한 것과 동일한 말뭉치에서 추출된 것을 대상으로 하였다.

2음절 + 변형된 문자열 + 2음절

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 데이타를 구축한 것이다.

4. 실험 및 성능 평가

SMS 문자열 변환 성능 실험을 수행하기 위하여 실제 사용자들의 문자 메시지로부터 구축한 2개의 SMS 단문 데이타 파일을 사용하였다. 각 말뭉치는 원시 상태로 수집된 말뭉치에 대해 통신 환경에서 사용되는 문법이 파괴된 어휘들을 수작업으로 교정한 것이다. 따라서 원문과 교정된 문장을 비교해 보면 삽입/삭제/교체 오류를 추출할 수 있다. SMS 단문 말뭉치는 SMS 문자열 오류를 교정하는 시스템을 구축하는데 사용한 2개의 데이타 파일과 구축된 시스템의 성능을 실험하기 위해 사용된 실험 데이타로 구분하였다. 성능 평가를 위해 사용한 시스템 구축 및 실험 데이타는 표 1과 같다.

표 1 시스템 구축 및 실험에 사용된 데이타

데이타 유형	데이타 명칭	데이타의 크기
시스템 구축용 데이타	통신언어 어휘집	1,878 단어
	SMS-1	80,820 문장
	SMS-2	15,089 문장

성능 평가를 위한 실험 데이타는 시스템 구축에 사용된 SMS-2과 동일한 유형의 자료로 구축하였다. 즉, 말뭉치로 수집된 총 단문 메시지 15,241 문장에서 99%는 시스템 구축용 데이타 SMS-2로 사용하고, 나머지 1%는 성능 평가용으로 사용하기 위한 목적으로 15,241 문장의 각 100번째 문장들을 추출하였다. 성능 평가에 사용된 실험 데이타는 1,335 어절로 구성되었으며 변환되어야 할 총 문자열 개수 128 개이다. 세 가지 유형의 문자열 변환 모델 및 각 모델들을 조합하여 아래와 같이 동일한 실험 데이타에 대한 성능을 측정하는 실험을 수행하였는데 그 결과는 표 2와 같다.

표 2에서 실험 과정에서 발생한 변환 오류를 유형별로 미변환 오류, 변환 오류1, 변환 오류2로 구분하여 오

표 2 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%

류 유형을 정리하였다. "미변환 오류"는 사전에 데이타가 없어서 오류 문자열이 변환되지 않은 경우다. 이 오류는 전체 오류의 가장 많은 비중을 차지하는 부분으로 해당 어휘들을 사전에 추가하는 작업을 통해서 성능 개선을 기대할 수 있다.

"변환 오류1"은 변환이 필요한 부분에서 변환이 일어나기는 했지만 잘못 변환되어 올바른 교정이 일어나지 않는 경우다. 예를 들어 "눈이 온다. 것두 무지 많이…"의 입력에 대해서 시스템은 "눈이 온다. 것도 무지 많이…"로 변환함으로써 "것두"를 단순히 "것도"로 변환하여 변환이 발생하기는 했지만 "그것도"로 변환하지 못해 올바른 변환이 일어난 것은 아니다. "변환 오류2"는 원래 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.

각 모델에 대한 실험 결과로 A+B' 이 재현율과 정확률 모두 가장 좋은 성능을 보여주었다. 즉, 재현율과 정확률은 통신어휘집을 이용하여 구축된 데이타를 기반으로 하고, 여기에 자동으로 추출된 변형된 문자열을 수작업을 통해 부작용(side effect)를 유발하는 것을 제외하여 정련을 함으로써 정확도를 향상시킨 방법을 적용했을 때 가장 좋은 성능을 얻을 수 있음을 확인할 수 있다. 모델 C는 변형된 문자열의 좌우문맥을 고려하기 때문에 문맥을 고려하지 않은 모델 B보다 정확률이 더 높다. 모델 B의 데이타를 가공하여 SMS 변환에 도움이 되는 것만 선별한 경우와 정확률이 비슷하다. 그러나 모델 C는 좌우문맥으로 인하여 적용율이 낮아지므로 모델 B에 비해 재현율이 매우 낮은 단점이 있다. 따라서 모델 A와 모델 B'을 결합한 형태가 가장 성능이 좋은 방법이라고 할 수 있다.

자동 띄어쓰기 모듈을 적용하는 과정에서 발생한 에러가 변환에 영향을 주는 경우도 있을 수 있다. 실험 결과를 분석할 때 이러한 부분도 감안할 필요가 있다. 전체 실험 데이타 파일의 어절 수 1,385중 자동 띄어쓰기 오류는 모델 A의 경우 50개이다. 각 실험 모델에 따라 구축된 문자열 변환 데이타의 크기는 표 3과 같다.

표 3 모델별 문자열 변환 데이타 개수

실험 모델	변환 데이타 개수	파일 크기(Kb)
A	3,035	44
A+B	12,885	280
A+B′	10,562	226
A+C	6,792	134
B	9,851	228
C	3,758	87
B+C	13,491	326
C′	4,338	67
C′′	8,400	162

< 이보세요전화을못받는거요*아님*안봐서못받은거요
> 이 보세요 전화를 못 받는 거요 *아님* 안 봐서 못 받은 거요

< 왜문자씹냐 범진오빤*모래*?
> 왜 문자 씹냐범진 오빠는 *모래*?

< *인제* 기상 이다 어제 조앗지.
> *인제* 기상이다 어제 좋았지.

< 오빠뭐해? 어떻게*사러* 그때지원했던놈은잘다녀?
> 오빠 뭐해? 어떻게 *사러* 그때 지원했던 놈은 잘 다녀?

그림 13 SMS 문자열 변환의 문제점

SMS 문자열을 변환할 때 좌우 문맥정보를 이용하여 변환하기 곤란한 경우가 발생하며 그 예는 그림 13과 같다.

5. 결 론

본 논문에서는 SMS 영역 문장의 오류나 의도하지 않은 변형을 교정하는 방법으로 복수개의 사전을 단계적으로 적용하는 방법을 제안하고 구현하였다. 구현된 시스템은 통신 어휘집 및 자체 구축한 SMS 영역 데이타 파일로부터 자동으로 추출한 사전을 사용하였을 때 비교적 만족스러운 성능을 보이면서 사전 구축 자동화

가능성을 확인하였다. 그러나 좌우 문맥 정보를 최대한 이용함으로써 가능한 많은 변환을 이끌어 내고 있음에도 불구하고 여전히 문제점을 가지고 있다. 그 대부분이 사전 어휘의 부재로 변형이 발생하지 않는 경우이다. 급속도로 변화하는 통신 어휘인 만큼 그 변화에 맞춰 데이타 및 사전을 갱신하는 것 또한 안정적인 성능을 보장하는 필수 요소라 하겠다. 중의성을 가진 어휘에 대한 사전의 적용 여부도 성능에 많은 영향을 준다. 향후 연구에서는 중의적인 어휘에 대해 연구하고 그 결과를 시스템에 반영할 필요가 있다. 어휘의 중의성을 판별하고 그 의미에 따라서 선택적으로 사전을 적용한다면 더 좋은 성능을 기대할 수 있다. 지속적으로 통신 어휘 데이타가 갱신되는 환경에서 어휘의 중의성을 감안한 문장 오류 교정 시스템을 본 논문에서 제안한 것과 같이 구현한다면 최적의 시스템이 될 수 있을 것이다.

참 고 문 헌

[1] 권연진, "컴퓨터 통신어의 언어학적 연구", 언어과학, 5권, 2호, pp. 58-62, 1998.

[2] 조찬식, "인터넷상에서의 언어 사용에 관한 연구", 한국문헌정보학회지, 35권 4호, pp. 177-196, 2001.

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

[4] 김보영, 강승식, "자모 빈도에 의한 통신 언어의 특성 연구", 제19회 한국 정보처리학회 춘계 학술발표 논문집, 10권 1호, pp. 501-504, 2003.

[5] 이정복, "컴퓨터 통신 분야의 외래어 사용", 새국어생활, 8권 2호, 국립국어연구원, 1998.

[6] 이정복, "통신 언어 문장 종결법의 특성", 우리말글, 22집, pp. 123-151, 2001.

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

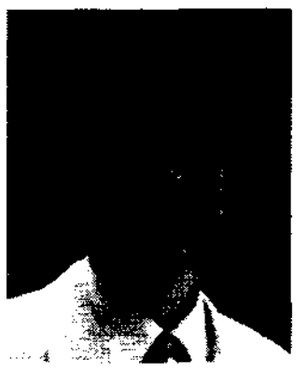
[8] 이재성, "영한 병렬 코퍼스로부터 외래어 표기 사전의 자동 구축", 컴퓨터교육학회논문지, 한국컴퓨터교육학회, 6권, 2호, pp. 9-21, 2003.

[9] Christian Jacquemin, Spotting and Discovering Terms Through Natural Language Processing, MIT press, 2001.

[10] Seung-Shik Kang, Kyu-Baek Hwang, "A Language Independent n-gram Model for Word Segmentation", AI'2006, pp. 557-565, 2006(LNAI 4304).

[11] 김용경, 조오현, 박동근, 컴퓨터 통신 언어 사전, 역락사, 2002.

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



강 승 식

1986년 서울대학교 컴퓨터공학과(학사)
1988년 서울대학교 컴퓨터공학과(석사)
1993년 서울대학교 컴퓨터공학과(박사)
1994년~2001년 한성대학교 정보전산학부 부교수. 2001년~현재 국민대학교 컴퓨터학부 부교수. 관심분야는 한국어 정보처리, 정보검색, 텍스트마이닝 등

장 두 성

1990년 전남대학교 전산학과(학사). 1993년 KAIST 전산학과(석사). 2005년 KAIST 전산학과(박사). 1993년~현재 KT 미래기술연구소 수석연구원. 관심분야는 한국어 정보처리, 음성언어처리, 대화시스템 등