SMS 영역에 대한 형태소 분석 사전의 구축

강승식*

일반적인 문서 영역을 기반으로 고안된 기존의 형태소 분석기를 SMS 영역의 문서에 적용하면 주로 SMS 영역에서만 자주 사용되는 미등록어에 대한 분석 오류가 빈번하게 발생한다. 본 연구에서는 SMS 영역에서 발생하는 형태소 분석 오류의 유형들을 분석하여 오류의 발생을 최소화하는 방안을 모색한다. 그 구체적인 방법으로 대량의 SMS 문장들에 대해 고빈도 어휘를 중심으로 미등록 어휘들을 품사별로 추출하고 이 어휘들에 대한 형태소 분석 사전을 보완할 수 있도록 함으로써 형태소 분석기의 성능을 개선하는 방법을 제시한다.

1. 서론

형태소 분석은 구문 분석이나 의미 분석의 전단계이면서 기계 번역이나 정보 검색 등 자연어 관련 연구 분야의 기본 과정이다. 형태소 분석에서 한국어와 같이 여러 개의 형태소가 하나의 단어를 이루는 언어는 단어를 구성하는 형태소들을 분리하는 방법이 매우 중요한 문제이다 (강승식 2002, Kwon et al. 1991). 단어의 의미가 어순보다 조사나 어미와 같은 형식 형태소에 의해 결정되는 한국어는 형

* 국민대학교 컴퓨터공학부 교수, sskang@kookmin.ac.kr. 본 논문의 작성 과정에 도움을 주신 장두성 박사님과 세 분의 심사위원께 감사를 표한다.

태소 분석이 구문 분석과 의미 분석에 미치는 영향이 매우 크다. 따라서 형태소 분석기의 성능 개선을 통해 구문 분석이나 의미 분석에서 어느 정도의 성능 개선 효과를 기대할 수도 있다. 효율적인 형태소 분석 방법론 및 분석 속도를 향상시키기 위한 방법으로 강승식(1994)은 한국어의 특성에 따라 음절 정보를 이용한 다단계 모델을 제안하였고, 이은철 (1992)은 접속 정보를 이용하였으며, 김재한(1994)은 어절 사전을 이용하는 방법을 제안하였다. 또한, 최재혁(1993)은 양방향 최장일치법을 통해 형태소 분석 효율을 제고하였으며, 임희석 (1995)은 배제 정보를 이용하여 분석 범위를 제한하는 방법으로 분석의 효율성을 높이고자 하였다.

일반 영역에 대한 범용 형태소 분석기는 SMS (Short Message Service) 영역의 문서에 적용할 때 SMS 영역에서 사용되는 미등록어로 인한 분석 오류가 빈번하게 발생한다. 따라서 형태소 분석기를 SMS 영역에 적용하려면 SMS 영역의 특성을 반영하여 미등록 어휘들을 형태소 사전에 추가하여야 한다.[1] 박봉래 (1998)는 용례 분석을 통해 미등록어를 추출하는 방법을 제시하였고, 김선호 (2002)는 동적으로 로컬 사전을 생성하여 미등록어를 분석하는 연구를 수행하였다. 미등록 어휘에 대한 형태소 분석 정확도를 향상시키기 위해 SMS 영역의 코퍼스를 분석하여 빈도가 높은 어휘들을 수집하고, 이를 형태소 분석 사전에 추가하여 형태소 사전을 증축함으로써 기존 형태소 분석기의 성능을 향상시킬 수 있다. 본 논문은 이와 같은 방법으로 일반 영역에 대한 범용 형태소 분석기의 성능을 SMS 영역

1) 본 연구에서는 최근에 주요 관심사가 되고 있는 SMS 영역에 대한 형태소 분석 오류를 분석 및 성능 개선을 추구하였다. 성능 개선 작업은 SMS 영역에 대해 수행되었으나 실제로 증축된 어휘들은 SMS 영역에서뿐만 아니라 일반 영역에서도 활용되기 때문에 증축된 사전을 domain 사전으로 제한하지 않고 일반 영역에서 사용하도록 하였다.

으로 확장하여 적용하는 방법을 기술한다.

2. SMS 영역의 말뭉치 규모

형태소 사전을 증축하기 위해 SMS 어휘를 추출하는데 사용된 말뭉치는 휴대폰에서 사용자들이 실제로 주고 받은 SMS 문자메시지 데이터를 축적하여 수집한 말뭉치이다.[2] 본 연구에서 사용된 SMS 말뭉치의 크기는 아래와 같다. 문자 메시지의 특성 때문에 한 문장의 길이가 여섯 어절 이하로 짧은 편이다.

- 문장 개수: 1,051,280 문장
- 어절 개수: 5,618,264 어절 (중복 제거시 295,631 어절)
- 텍스트 파일의 크기: 약 30M 바이트

SMS 데이터 말뭉치에서 어휘를 추출하는 방법으로 1차적으로 형태소 분석기의 형태소 분석 사전에 수록된 어휘와 수록되지 않은 미등록어들을 구분하여 어휘를 추출한다. 형태소 분석 사전에 수록된 것과 그렇지 않은 어휘들을 선별하기 위하여 범용 형태소 분석기를 사용한다. 그런데 형태소 분석기는 입력 어절에 대해 두 가지 이상의 분석 결과를 생성하는 경우가 있으며, 이 경우에 품사 태깅 결과로 선택된 형태소 분석 결과로부터 어휘형태소를 추출한다. 본 연구에서 사용한 성능 개선의 대상으로 삼은 형태소 분석기는 음성언어 처리용으로 사용되고 있는 형태소 분석기 및 품사 태거이다.

형태소 분석 및 품사 태깅 결과로부터 사전에 등록된 어휘와 미등

2) 말뭉치 수집 방법 및 출처를 자세히 밝혀야 하나 통신회사의 문자 메시지에는 개인정보가 포함되는 경우가 있어서 개략적으로 명시하였다.

록어를 선별하는 방법은 다음과 같다. 우선, 말뭉치에 대한 형태소 분석 결과 및 품사 태깅 결과를 출력한 후에 품사 태깅 결과에 해당하는 형태소 분석 결과를 역추적을 한다. 그 결과로 품사 태거가 선택한 형태소 분석 결과로부터 어휘 형태소를 추출한다. 품사 태깅 결과로부터 형태소 분석 결과를 역추적을 하는 이유는 연구에 사용된 형태소 분석 결과와 태깅 결과가 서로 다른 형식으로 저장되어 있어 직접 연결이 되어 있지 않은 특성 때문이다.

3. SMS 말뭉치의 분석 및 미등록어 추출

3.1 형태소 분석에 의한 어휘 형태소 추출

SMS 말뭉치로부터 형태소 분석 오류를 유발하는 형태소를 수집하기 위한 기초 작업으로 형태소 분석기와 품사 태거를 이용하여 말뭉치로부터 품사별로 어휘들을 추출하였다. 그 결과로 추출된 어휘들을 품사별로 구분한 결과는 표 1과 같다. 이 표에서 2개 이상의 명사들로 구성된 명사는 복합명사로 구분하였고, 형태소 분석 결과에서 어휘형태소가 형태소 분석 사전에 수록되지 않은 것은 '미등록어'로 간주하였다.

표 1. 말뭉치에서 추출된 어휘형태소

품사	추출된 어휘수
명사	20,033
수사	57
동사	2,365
형용사	1,123
부사	1,199
관형사	54
감탄사	59
복합명사	50,423
미등록어	44,483
합 계	119,798

그림 1, 그림 2, 그림 3은 형태소 분석기에 의해 분석된 어절들로부터 사전에 수록된 어휘형태소들에 대해 각 품사별로 <어휘, 품사정보, 출현 빈도>를 계산한 결과를 예시한 것이다.[3)]

명사 (20,033개)	수사 (57개)
가 Mvbgm 713	공 Syjy 11
가가 Mvbgm 26	구 Syjm 26
가감 Mvsty 1	넉 Sygy 4
가게 Mvbgm 401	네다섯 Sygy 1
가격 Mvbgy 602	네번째 Ssm 3
가격표 Mvbgm 15	넷 Sygy 33
가결 Mvstr 10	마흔 Sygy 9
가계 Mvbgm 31	만 Syjy 12299
가계부 Mvbgm 7	몇 Sgy 9
가계비 Mvbgm 1	몇몇 Ssy 10
가계약 Mvbgy 3	일곱 Sygy 66
히브리서 Mggm 2	일흔 Sygy 1
히스테리 Mvbgm 1	첫번째 Ssm 31
히터 Mvbgm 8	첫째 Ssm 54
히트 Mvbgm 6	칠 Syjr 139
힌트 Mvbgm 6	팔 Syjr 34
힐튼호텔 Mvbgr 1	하나 Ssm 1192
힘 Mvbgy 2601	한둘 Sygy 1

그림 1. 명사, 수사의 추출 결과 및 빈도수

3) M은 명사, D는 동사, H는 형용사 등을 의미하며, 모든 품사 정의표를 추가하게 되면 불필요하게 많은 데이터가 논문에 삽입되어 생략하였다.

동사 (2,367개)	형용사(1,123개)
가 Dgma 19198	가깝 HbYabg 49
가 Dgma 5522	가깝 HbYebb 517
가공하 Dbmayg 1	가냘프 Hbmaeg 2
가까워지 Dgme 1	가느다랗 HbYahb 1
가꾸 Dgme 48	가늘 HbYerg 16
가누 Dgme 3	가득차 Hgma 22
가늘어지 Dgme 1	가득하 Hbmayg 1438
가늠하 Dbmayg 4	가렵 HbYebb 21
가다듬 Dgye 8	가렵 HbYebg 4
가도되 Dgme 2	가볍 HbYebb 82
가동하 Dbmayg 6	가볍 HbYebg 100
가두 Dgme 6	가뿐하 Hbmayg 13
가득채우 Dgme 1	가엾 Hgme 5

그림 2. 동사, 형용사의 추출결과 및 빈도수

부사 (1,199개)	관형사 (54개)
가까스로 Bgg 2	각 Geg 568
가까운 Bgg 44	갖가지 Ged 1
가까이 Bjgm 113	고Geg 106
가끔 Bjgy 317	그Geg 5936
가끔가다가 Bgg 1	그깟 Gg 8
가득 Bjgy 465	긴긴 Gg 3
가득가득 Bgg 29	까짓 Gg 6
가득히 Bjgm 16	네까짓 Gg 1
가뜩 Bjgy 1	다른 Geg 30
가뜩이나 Bgg 11	단돈 Ged 4

그림 3. 부사, 관형사의 추출 결과 및 빈도수

3.2 미등록어 추정 어휘의 분석

본 연구에서 성능 개선의 대상으로 한 형태소 분석기의 형태소 분석 결과는 미등록어를 추정할 때 표 2와 같이 품사별로 명사, 동사, 형용사, 부사, 감탄사로 품사 정보까지 추정한 분석 결과를 생성한다.[4] 형태소 분석기 사용자는 미등록어의 품사를 추정해 주기를 요

4) 미등록어가 포함된 어절의 분석 방식은 형태소 분석 알고리즘에 따라 다를 수 있다. “미등록어의 품사를 추정하는 방식”은 그 정확도가 낮을 경우 오히려 품사 추

구하고 있으며, 이러한 요구에 따라 미등록어의 품사를 추정한 것이다. 그러나 형태소 분석기가 미등록어로 추정한 결과는 많은 오류를 포함하고 있으며, 품사 추정이 잘못된 경우가 대부분이어서 품사별로 분류된 결과를 활용하기가 곤란하다. 따라서 미등록어는 추정된 품사 정보를 활용하는 것이 무의미하기 때문에 추정된 품사 정보를 무시한다.

표 2. 미등록어 추정 어휘의 품사별 어휘 수

품 사	추출된 어휘수
미등록 명사	25,765
미등록 동사	7,278
미등록 형용사	9,298
미등록 감탄사	169
합 계	44,483

형태소 분석 결과로부터 어휘를 추출할 때 제외된 어휘들은 영문자나 숫자를 포함하는 어절, 조사, 어말어미, 선어말어미, 관형사형 어미, 명사형 어미, 부사형 어미, 조용사, 기호 등이다. 전체 미등록어의 다수가 오류이므로 미등록어의 품사를 구분할 필요성이 없다고 판단하였다.

미등록어로 추정된 어휘 중 종성 '씨'이 포함된 것 (6,976개)과 "느" (1,049개), "는" (495개), "니" (928개), "다" (610개)로 끝나는 용언 미등록어는 명사를 추출하기 위한 미등록어가 아니므로 제외한다.[5] 그 밖에 '며/까/잖/없'으로 끝나는 것들도 미등록어가 아닌 것으로 판

정을 하지 않는 것이 나을 수 있다.

5) "씨/느/는/니/다"로 끝나는 것들은 형태소 분석기가 미등록어로 잘못 추정한 것이다. 즉, 어말어미를 제거한 후에 미등록어를 추정해야 하는데 어말어미까지 포함시켜 추정한 오류이다. 이 오류는 사전 증축이 아니라 미등록어를 추정하는 알고리즘을 수정하여 개선되어야 할 사항이다.

단하여 제외한다. 이렇게 하여 미등록어로 최종 결정된 어휘의 개수는 34,425개이다. 형태소 분석에 의해 추출된 복합명사 50,423개는 그림 4에 그 일부를 예시하였다,

```
4089 연락_주_세 : 연락주세요
2932 진료_예약 : 진료예약
2613 납입_지연 : 납입지연
2367 확인_요망 : 확인요망
2329 자동_이_체 : 자동이체
1883 영대_병원 : 영대병원
1877 전_화주_세 : 전화주세요
1725 수신_문서 : 수신문서가
1639 미_입금 : 미입금
1423 결혼_기념일 : 결혼기념일
1175 회_비 : 회비
1144 반_납일 : 반납일
1092 중앙_도서관 : 중앙도서관
970 대봉_도서관 : 대봉도서관
958 건강_보험료 : 건강보험료
```

그림 4. 복합명사 추출 결과: 출현 빈도순

미등록어 추정오류가 분명한 “ㅆ/느는/니/다”로 끝나는 어휘를 제외한 미등록어 34,425개 어휘들의 예는 그림 5와 같다.

```
2742 파이팅 (M*(미등록-명사)) 파이팅
1647 되시기 (D*(미등록-동사)) 되시길
1624 없 (H*(미등록-형용사)) 없거나
1480 하루되세 (M*(미등록-명사)) 하루되세요
1304 샬로 (H*(미등록-명사)) 샬롬
1292 핸드포 (H*(미등록-형용사)) 핸드폰
1165 월례 (M*(미등록-명사)) 월례
1066 되기 (D*(미등록-동사)) 되길
1064 아무트 (H*미등록-명사)) 아무튼
1048 뭐 (M*(미등록-명사)) 뭐데
1014 입금 (M*(미등록-명사)) 입금되야
```

그림 5. 미등록어 추출 결과: 출현 빈도순

4. 미등록어와 복합명사의 인식 오류

4.1 미등록어의 인식 오류

형태소 분석기의 분석 오류의 유형을 분석해 보면 '미등록어 인식 오류'와 '복합어 분석 오류'로 구분된다. 미등록어 인식 오류는 그 오류의 발생 원인에 따라 '어휘형태소 인식 오류'와 '문법형태소 분리 오류', 그리고 '형태소간 결합 조건 검사 오류' 로 세분화된다.

- 미등록어 인식 오류: 어휘형태소 인식 오류, 문법형태소 분리 오류, 형태소간 결합 조건 검사 오류
- 복합어 분석 오류: 복합명사 분해 오류, 복합용언 분해 오류, 접두사/접미사 분리 오류

문법형태소 분리가 성공적인 어절에서 어휘형태소 인식 오류는 해당 어휘형태소를 사전에 등록함으로써 분석 오류를 방지할 수 있다. 그러나 문법형태소 분리 오류가 발생하는 경우는 어휘형태소의 사전 등록 여부와 무관하게 분석 오류가 발생한다. 문법형태소를 분리하지 못하는 오류는 적용범위가 넓고 사전 보완 작업으로 해결할 수 없는 문제이다. 즉, 문법형태소 사전을 보완하고, 결합제약 조건 관련 모듈을 수정해야 하며, 알고리즘의 관련 부분을 수정하는 작업이 요구된다.

미등록어로 추정된 형태소 분석 오류는 위 미등록어 인식 오류 유형에서 발생 원인을 알 수가 없으므로 단순히 어휘형태소 사전에 추가하는 방법만으로는 근본적인 문제가 해결되지 않는다. 따라서

미등록어로 추정된 입력 어절들의 오류 유형을 분석하여 각 유형별로 문제점을 파악하여 그 문제점을 해결해야 한다.

미등록어로 추정된 어휘 34,425개 중에는 대부분 특정 유형의 형태소 분석 오류로 인한 어휘 분석 오류가 많다. 빈번하게 발생하는 형태소 분석 오류를 발견하기 위하여 미등록어로 추정된 어휘들의 끝음절 빈도를 조사하였으며, 그 결과는 그림 6과 같다.[6] 그림 6의 좌측은 품사 정보를 고려한 끝음절 빈도이고, 우측은 품사 정보를 무시하고 계산한 끝음절 빈도이다. 이 유형들에 대한 형태소 분석 오류의 원인을 분석하여 오류의 원인을 제공함으로써 형태소 분석 성능을 개선하여야 한다. 즉, 오류 유형을 분석함으로써 빈도가 높은 유형에 대한 성능 개선 작업이 쉽게 이루어지도록 하기 위한 것이다.

품사 고려	품사 무시
649 며 (H*)	1169 기
645 기 (D*)	805 며
551 라 (M*)	769 까
474 잖 (M*)	737 하
448 고 (M*)	681 지
401 래 (M*)	681 라
394 지 (M*)	587 시
344 스 (M*)	522 래
341 려 (M*)	521 되
338 기 (H*)	516 잖
337 마 (H*)	514 이
331 서 (M*)	490 으
321 까 (H*)	470 거
318 하 (M*)	467 고
314 에 (M*)	458 마
313 네 (M*)	451 려
291 아 (M*)	447 스
281 리 (M*)	441 드
278 까 (M*)	431 리
271 이 (M*)	427 서
264 시 (M*)	389 가

6) 이 유형의 오류어들은 성능 개선을 위한 수작업 대상이 아니므로 사전 증축 및 보완을 하기 위한 미등록어 추출하는 범위에서 제외된다. 즉, 이 유형들을 일괄적으로 제외함으로써 사전 증축 및 보완 작업을 효율적으로 수행하였다.

260 가 (M*)	388 구
257 세 (M*)	374 네
251 군 (M*)	366 에
250 봐 (M*)	362 더
243 거 (H*)	357 봐
241 하 (D*)	342 아
239 으 (H*)	312 야
239 나 (M*)	307 나
234 건 (M*)	282 러
232 더 (M*)	276 어
216 구 (M*)	270 세

그림 6. 미등록어 끝음절: 출현 빈도순

그림 7은 미등록어 분석 오류로부터 수작업을 통해 사전에 수록되어야 할 미등록 어휘형태소와 복합 용언, 그리고 조사/어미를 수집한 것이다.[7)]

미등록어	복합용언	조사/어미
까르푸	가까워지	으로는요
뉴케이에스	가까이하	는요
능사	가득하	는가
다빈치	가보	로는요
대장금	간지럽히	에게는
동창회	감사받	보다도
디시	거슬리	보단
디카	거침없	에다가
띵동	걱정없	는구려
락카	걸리	ㄴ구려
레포트	걸맞	는데
로딩	걸어다니	ㄴ다는데
로또	걸어오	ㄴ가
모닝코피	격려바라	는다니
모닝플러스	공연하	는지
문자메시지	구워먹	는지요
미주랜드	궁금해지	ㄹ는지

7) 본 연구의 사전을 증축하기 위한 미등록어 추출 방법론은 형태소 분석 오류들의 유형 및 오류 개선 방법을 포괄적으로 제시하여 형태소 분석기 개발자가 성능을 개선할 때, 그리고 형태소 분석기 사용자들이 어떠한 형태소 분석 오류가 발생하는지를 예측할 수 있도록 함으로써 이 오류들에 대해 대처하는 방안을 모색하는데 도움이 될 수 있다.

바이오	기약없	ㄹ는지요
바코드	기어다니	는지요
베니건스	기운나	려는데

그림 7. 미등록어, 복합용언, 조사-어미 사전

4.2 복합명사 분해 오류의 개선 방안

3.2절에 예시한 바와 같이 복합명사로 분석된 50,423개의 어절로부터 복합명사 분해 오류어를 선별하여야 한다. 이 오류에는 복합명사 분해 오류 이외에도 문법형태소 분리 오류로 인한 복합명사 인식 오류가 다수 포함되어 있으며, 형태소 분석 성능 개선에서 복합명사 분해 모듈은 독립적인 기능을 수행하므로 복합명사 분해보다는 복합명사를 인식하는, 즉 문법형태소와 분리하는 과정에서 발생한 오류를 개선하여야 한다. 그 해결 방안으로 형태소 분석기가 복합명사로 분석한 결과 중에서 복합명사가 아닌 경우만을 선별하였다. 이 어절들에 대한 형태소 분석 오류를 수정함으로써 복합명사 관련 형태소 분석 성능을 개선할 수 있다. 그림 8은 이러한 오류 발생 어절들을 고빈도어 순으로 정렬하여 예시한 것이다.

4089 연락주세요	229 보관중	116 마시길
1877 전화주세요	204 희망찬	113 속상해
890 인해	195 지속적	111 전화가
833 입금해	193 연락주시면	110 입금치
780 전화주고	189 전화해도	106 귀하신
607 활기찬	188 지연중	105 부터
569 중이오니	185 보내시길	105 꽃피는
549 가득하시길	177 사용시	100 전화피하고
543 참석바람	177 불참시	99 제게
536 하시길	174 힘내시고	99 오시길
525 미납시	174 전화상	98 평안하시고
519 힘내세	167 연락주시기	93 평안한
463 입금하시고	166 행복하시길	88 정도면
454 한시적	161 활기차게	87 충만하시길
440 입금하시어	159 부득이하게	87 미납중

433 법조치	156 미결제시	86 주길
427 궁금해	145 연체시	83 하지마
417 주시길	143 이상해	83 깃들기
405 힘내고	142 입금하시면	82 피하지
386 내원해	142 금일중	81 자라길
362 일해	140 시청하시나요	81 별세하셔서
349 예약도	137 전화로	78 생기길
348 연체중	134 전화해	78 미납부시
327 입금일	129 전화주시기	77 전해주고
325 입금하시기	127 혹시나	77 보고파
306 건강하시길	125 발신자정보표시	74 참석하시길
266 경기도당	121 내원하시고	74 전해주세요
247 우리당	117 경매지	73 문자보시면

그림 8. 형태소 분석 오류 어절: 출현 빈도순

5. 실험 및 성능 평가

실험 및 성능 평가는 SMS 데이터에 대해 사전 증축 전후의 형태소 분석 성능 개선 효과에 대한 실험 대신에 음성 언어용 형태소 분석기와 KLT version 2.0[8]의 성능을 비교하였다. 그 이유는 실험 데이터 및 성능 개선 효과의 객관성 확보가 어려웠기 때문이다. 실험 데이터는 미등록어가 많이 포함되어 있는 분야로 신문기사를 선택하였으며 어절 개수는 2,317개이다.[9] 형태소 분석 결과에 대한 분석 정확도는 '분석 성공한 경우', '옳은 분석 결과가 생성되지 않은 미분석 오류', '틀린 분석 결과가 포함된 과분석 오류'의 세 가지 유형으로 구분하여 정확도를 측정하였다.

8) http://nlp.kookmin.ac.kr/

9) 성능 개선 대상의 형태소 분석기는 음성언어를 대상으로 개발된 것이고, 비교 대상으로 사용한 것은 국내에서 보편적으로 많이 사용되고 있는 형태소 분석기이다. SMS에 대한 실험 결과를 제시하는 것이 바람직하나 실제 SMS 문장에는 철자 오류 및 맞춤법에 어긋나는 오류어와 통신용 은어가 다수 포함되어 있어 실험용으로 사용하기가 곤란하여 신문기사를 실험 대상으로 하였다.

표 3. 형태소 분석기 성능 실험

	분석 성공(%)	미분석 오류(%)	과분석 오류(%)
음성언어용 형태소 분석기	2,163 어절	30 어절	124 어절
	93.35%	1.30%	5.35%
KLT version 2.0	2,283 어절	6 어절	28 어절
	98.53%	0.26%	1.21%

음성 언어용 형태소 분석기의 경우에 과분석 결과가 2개 이상인 경우가 다수 포함되어 있는데, '과분석 오류' 어절에는 과분석 결과가 2개 이상이 포함되어 있을 때 과분석 개수로 계산하지 않고 해당 오류가 포함된 어절 개수로 계산하였다.

6. 결론

일반적인 영역에 대한 기존의 형태소 분석기를 SMS 영역에 대해 적용할 때 미등록 어휘로 인한 오류를 개선하기 위해 SMS 영역의 말뭉치에서 미등록어를 추출하고 분석 사전을 보완하여 증축함으로써 SMS영역에 대한 형태소 분석 성능을 개선하였다. 실험 결과에 의하면 사전 증축으로 인한 형태소 분석 성능 개선이 있었으나 타 형태소 분석기와 비교했을 때 사전의 증축 작업만으로는 성능 개선 효과에 한계가 있음을 알 수 있다. 또한, SMS영역 말뭉치의 가공 및 미등록어 추출 과정에 수작업의 비중이 커서 변화 및 활용이 심한 SMS 영역 코퍼스에 대해 반복 적용하기가 어려운 점이 있다. 그러므로 SMS 영역 코퍼스에 대한 형태소 분석 성능을 개선하고자 한다면 알고리즘을 개선하는 것이 더 효과적일 것으로 추정된다. 다만,

형태소 분석 알고리즘이 규칙 기반 알고리즘이 아니라 사전을 기반으로 하는 통계적인 방법을 사용하는 경우라면 이러한 결론은 재고되어야 할 것이다.

참고문헌

강승식. 2002. 「한국어 형태소 분석과 정보검색」, 홍릉과학출판사.

강승식. 1994. "다층 형태론과 한국어 형태소 분석 모델." 제6회 한글 및 한국어정보처리 학술발표논문집. 140-145.

김재한, 옥철영. 1994. "어절 사전을 이용한 한국어 형태소 분석." 한국정보과학회 봄학술발표 논문집. 813-816.

이은철, 이종혁. 1992. "계층적 기호 접속 정보를 이용한 한국어 형태소 분석기의 구현." 제4회 한글 및 한국어 정보처리 학술발표 논문집. 95-104.

임희석, 윤보현, 임해창. 1995. "배제 정보를 이용한 효율적인 한국어 형태소 분석기." 정보과학회 논문지, 22권 6호. 957-964.

최재혁, 이상조. 1993. "양방향 최장일치법에 의한 한국어 형태소
분석기에서의 사전 검색 횟수 감소 방안." 정보과학회 논문지, 20권 10호. 1497-1507.

김선호, 윤준태, 송만석. 2002. “한국어 문서 처리를 위한 동적 생성 로컬 사전 기반 미등록어 분석.”정보과학회논문지: 소프트웨어 및 응용. 29권 5-6호. 407-416.

박봉래, 황영숙, 임해창. 1998. “용례 분석에 기반한 미등록어의 인식.”정보과학회논문지(B). 25권 2호. 397-407.

H.C. Kwon, Y.S. Chae, and G.O. Jeong. 1991. "A Dictionary-based Morphological Analysis." In Proceedings of NLPRS'91. 87-91.

K. Shim and J. Yang. 2002. "MACH: A Supersonic Korean Morphological Analyzer." In *Proceedings of the 19th International Conference on Computational Linguistics*. 939-945.

K. Koskenniemi. 1983. "Two-level Model for Morphological Analysis." In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*. 683-685.

L.J. Cahill. 1990. "Syllable-based Morphology." In *Proceedings of the 13th International Conference on Computational Linguistics*, Vol.3. 48-53.

Language Information 9, 2008, pp. 5-21

Constructing Morphological Analysis Dictionary for SMS Domain

Seung-Shik Kang

Keywords Morphological analysis, dictionary construction, unregistered word, SMS domain

Abstract

When we apply a general-purpose morphological analyzer to SMS (Short Message Service) domain, some errors are included in the morphological analysis results because there are many unregistered words in SMS sentences. Our goal of this research is to investigate a method of minimizing the errors through the analysis of morphological error types in SMS sentences. As a practical method, high-frequency lexical morphemes are extracted from the large volume of SMS sentences that cause morphological analysis errors. Unregistered morphemes are manually reviewed and they are added to the lexical dictionary to improve the accuracy of the morphological analyzer.

논문투고일 : 2008년 1월 16일
심사완료일 : 2008년 2월 29일
게재확정일 : 2008년 3월 7일