인지과학, 제19권 제3호
Korean Journal of Cognitive Science
2008, Vol. 19, No. 3, 283~295.

2단계 규칙을 이용한 해체된 한글 음절의 결합*

이 주 호 김 학 수†

강원대학교 컴퓨터정보통신공학전공

SMS나 메신저의 사용이 증가함에 따라 의도적으로 음절을 해체한 새로운 형태의 문장들이 젊은이들 사이에서 습관적으로 사용되고 있다. 이러한 상황에서 자연어 인터페이스 시스템을 개발하기 위해서는 해체된 한글 음절을 결합하여 올바른 문장을 만들어 주는 기술이 먼저 개발되어야 한다. 본 논문에서는 해체된 음절을 2단계 규칙을 이용하여 결합해주는 방법을 제안한다. 1단계에서는 수동으로 작성한 휴리스틱을 이용하여 단순하게 초성, 중성, 종성으로만 해체된 음절을 결합한다. 그리고 2단계에서는 매핑 테이블과 변환기반 학습을 이용하여 복자음까지 해체된 음절을 결합한다. 실험 결과, 제안한 방법은 단순 해체 음절의 결합과 복자음 해체 음절의 결합에서 각각 100%와 99.98%라는 매우 높은 정확률을 보였다.

주제어 : 한글 음절 해체, 한글 자소 결합, 변환기반 학습

---

* 본 연구는 삼성전자 정보통신트랙 과제의 지원을 받아 수행되었습니다.

† 교신저자: 김학수, 강원대학교 컴퓨터정보통신공학전공, 연구 세부 분야: 한국어정보처리
Email: nlpdrkim@kangwon.ac.kr

서 론

웹 기반의 전자 상거래 시스템들의 등장으로 사용자와 사업자 모두의 요구를 충족시킬 수 있는 효과적인 정보 검색 도구의 필요성이 제기되고 있으며, 이러한 요구가 무선 환경으로 빠르게 확산되고 있다. 그러나 현재 상용화되어 있는 PDA나 휴대폰 단말기의 사용자 인터페이스는 매우 지루하고 복잡한 메뉴들로 이루어져 있기 때문에 자연어 검색과 같은 보다 편리한 인터페이스의 개발이 필요하다. 하지만 이를 위해서는 인터넷 및 무선 환경에서 광범위하게 사용되고 있는 인위적인 철자오류를 보정하는 방법의 개발이 선행되어야 한다.

기존의 철자오류 보정 모델들은 크게 단어 기반 보정 모델과 자소 기반 보정 모델로 나누어진다. 단어 기반 보정 모델은 기구축된 보정사전을 n-gram이나 edit-distance를 이용하여 탐색함으로써 보정대상 후보 단어를 선택하고, 문맥 통계를 이용하여 사전에 등록되어 있는 새로운 단어로 교체하는 방법이다[4][5][6][8]. 단어 기반 보정 모델은 구어체 문장에서 자주 나타나는 신조어나 축약어 등을 모두 사전에 등록을 해야 하기 때문에 단어 사전의 크기가 매우 커진다는 단점이 있다. 이러한 문제를 해결하기 위해서 자소 기반 보정 모델들이 연구되고 있다. 자소 단위 보정 모델은 단어를 자소 단위로 해체한 후, 대용량의 학습데이터로부터 추출된 자소 변환 통계를 이용하여 자소를 삭제, 삽입하거나 새로운 자소로 교체하는 방법이다[9]. 그러나 지금까지 제안된 철자오류 보정 모델들은 '어케', '남친'과 같은 음절의 변형이나 '안녕하세여', '촙오'와 같은 자소의 변형을 복원하는데 초점을 맞추어왔으며, 통신체 문장에서 빈번히 발생하는 의도적 음절 해체 현상이나 오타에 의한 음절 해체 현상을 다루고 있지 않다. 그러므로 'ㅇㅏㅃㅏ', 'ㅅㅅㅏㄹㅣㄴㅏㅁㄱ'와 같이 해체된 음절을 기존의 단어 기반 보정 모델들의 입력으로 사용한다면 음절 정보를 알 수 없기 때문에 정확한 오류 보정이 불가능하다. 그리고 기존의 자소 기반 보정 모델을 사용한다고 하더라도 초성과 종성 정보를 구분할 수 없기 때문에 성능의 하락이 불가피하다. 이러한 문제를 해결하기 위해서 본 논문에서는 의도적으로 해체되거나 오타에 의해 해체된 음절들을 결합하는 철자오류 보정 전처리 방법을 제안한다. 음절 해체는 단순 음절 해체와 복자음 해체로 나누어진다. 단순 음절 해체는 'ㅇㅏㅃㅏ'와 같이 음절을 단순히 초성, 중성, 종성을 따

로 분리한 것이고, 복자음 해체는 'ㅇㅏㅂ ㅂㅏ'와 같이 초성 또는 종성이 복자음인 경우 다시 한 번 분리한 것을 말한다. 본 논문에서는 1단계로 간단한 휴리스틱을 이용하여 단순 음절 해체 문제를 해결한다. 그리고 2단계로 쌍자음 이외의 복자음 해체 문제는 간단한 매핑테이블(mapping table)을 이용하여 해결하고, 많은 애매성을 포함하는 쌍자음 해체 문제는 변환기반 학습(transformation-based learning)[2]에 의해 자동 추출된 규칙을 이용하여 해결한다.

본 논문의 구성은 다음과 같다. 2장에서 휴리스틱을 이용하여 단순 음절 해체 문제를 해결하는 방법을 설명하고, 3장에서 매핑 테이블과 변환기반 학습을 이용하여 복자음 해체 문제를 해결하는 방법을 설명한다. 4장에서 실험 및 오류 분석 결과를 보이고, 마지막으로 5장에서 결론을 맺는다.

단순 음절 해체 문장의 보정

단순 음절 해체 문장을 보정하기 위해서 본 논문에서는 한글의 초/중/종성 정보에 기초한 휴리스틱을 이용한다. 먼저, 복자음을 해체하지 않는 범위 내에서 입력된 문장에 포함된 모든 음절을 자소 단위로 해체한다. 즉, '닦ㅇㅏ서(닦아서)'와 같은 문장이 입력되었을 경우에 'ㄷㅏㄲㅇㅏㅅㅓ'와 같이 복자음을 해체하지 않는 범위 내에서 자소 단위로 분리한다. 그리고 연속된 2개 이상의 자음으로 구성된 자소열 중 복자음이 해체된 상태로 입력되었을 가능성이 있는 것(연속된 2개 이상의 자음이 쌍자음을 구성하는 경우와 3개 이상의 자음이 연속된 경우)은 단순 자소 결합 대상에서 제외한다. 즉, '느ㄱ김이 어ㅂㅅ다(느낌이 없다)'와 같은 문장이 입력되었을 경우에 'ㄴㅡㄱㄱㅣㅁㅇㅣ ㅇㅓㅂㅅㄷㅏ'로 자소 분리를 한 후, 2개 이상의 자음이 쌍자음을 구성하는 'ㄱㄱ'과 3개 이상의 자음이 연속된 'ㅂㅅㄷ'은 복자음 결합 대상으로 간주하고 단순 자소 결합 대상에서 제외한다. 물론 'ㄷㅏㄲㅇㅏㅅㅓ'와 같이 복자음이 해체되지 않은 상태로 입력된 자소열은 위의 두 경우에 해당하지 않기 때문에 단순 자소 결합 대상으로 삼는다. 단순 자소 결합 대상이 정해지면, 임의로 해체한 자소에는 그에 해당하는 올바른 초/중/종성 정보를 부착하고, 원래 해체되어 입력된 자음에는 무조건 종성 정보를 부착한다. 그 이유는 제

안한 방법이 임의로 해체한 자음은 초성/종성 정보를 알 수 있지만 사용자에 의해서 원래 해체되어 입력된 자음은 그것이 초성인지 종성인지 알 수 없기 때문이다. 표 1은 'ㅈㅏㅇㅕㄴㅇㅓ처리'라는 문장을 자소 단위로 해체한 후, 초/중/종성 정보를 부착한 예를 보여준다. 표 1에서 보는 것과 같이 제안한 방법은 원래 해체되어 입력된 'ㅈ', 'ㄴ', 'ㅇ'에는 무조건 종성 정보를 부착하고, 임의로 해체한 '여', '처', '리'에 포함되어 있는 'ㅇ', 'ㅊ', 'ㄹ'에는 초성 정보를 부착한다.

표 1. 음절 해체 및 초/중/종성 정보 부착

입력 문장	ㅈㅏ여ㄴㅇㅓ처리										
음절 해체	ㅈ	ㅏ	ㅇ	ㅕ	ㄴ	ㅇ	ㅓ	ㅊ	ㅓ	ㄹ	ㅣ
	종성	중성	초성	중성	종성	종성	중성	초성	중성	초성	중성

음절 해체 및 초/중/종성 정보 부착이 끝나면, 1차로 '초성, 중성, 종성'이나 '초성, 중성'으로 이루어진 자소열을 한 음절로 결합한다. 표 2는 1차 자소 결합 방법에 따라 'ㅈㅏㅇㅕㄴㅇㅓ처리'라는 입력 문장이 처리된 결과를 보여준다.

표 2. 1차 자소 결합의 예

음절 해체	ㅈ	ㅏ	ㅇ	ㅕ	ㄴ	ㅇ	ㅓ	ㅊ	ㅓ	ㄹ	ㅣ
	종성	중성	초성	중성	종성	종성	중성	초성	중성	초성	중성
1차 자소 결합	ㅈㅏ연ㅇㅓ처리										

표 3. 2차 자소 결합의 예

음절 해체	ㅈ	ㅏ	ㅇ	ㅕ	ㄴ	ㅇ	ㅓ	ㅊ	ㅓ	ㄹ	ㅣ
	종성	중성	초성	중성	종성	종성	중성	초성	중성	초성	중성
종성 -> 초성	ㅈ	ㅏ	연			ㅇ	ㅓ	처		리	
	초성	중성				초성	중성				
2차 자소 결합	자연어처리										

1차 자소 결합이 끝나면, 2차로 중성 앞에 존재하는 종성을 모두 초성으로 변환한다. 그리고 '초성, 중성, 종성'이나 '초성, 중성'으로 이루어진 자소열을 한 음절로 결합한다. 표 3은 2차 자소 결합 방법에 따라 'ㅈㅏㅇㅕㄴㅇㅓ처리'라는 입력 문장이 최종적으로 처리된 결과를 보여준다.

복자음 해체 문장의 보정

복자음 해체 현상은 쌍자음 이외의 복자음이 해체된 경우와 쌍자음이 해체된 경우로 구분된다. 전자의 경우에 한글 초성 위치에 쌍자음 이외의 복자음이 올 수 없기 때문에 간단한 매핑 테이블('자소+자소 = 복자음' 형태의 테이블)을 이용하여 문제를 해결할 수 있다. 즉, 단순 자소 결합 대상에서 제외된 자소열 중에 'ㅇㅓㅂㅅㄷㅏ'의 'ㅂㅅㄷ' 과 같이 앞의 두 자소가 쌍자음을 구성하지 못하는 경우에 매핑 테이블을 참조하여 해당 종성 복자음으로 변경하고, 나머지들도 같은 방법으로 초성으로 변경함으로써 문제를 해결할 수 있다. 그러나 후자의 경우에 올바른 보정을 위해서는 다음과 같은 3가지 문제를 해결해야 한다. 첫째, 표 4에서 보는 것과 같이 쌍자음으로 결합 가능한 자소들이 연속해서 2회 이상 출현할 경우에 쌍자음이 해체된 경우인지 아닌지를 판단해야 한다.

표 4. 쌍자음으로 결합해야 하는 경우와 아닌 경우의 예

쌍자음으로 결합해야 하는 경우	쌍자음으로 결합하지 말아야 하는 경우
ㄴㅡㄱㄱㅣㅁ (느낌)	ㅎㅏㄱㄱㅛ (학교)
ㅇㅓㄷㄷㅓㅎㅏㄴ (어떠한)	ㄷㅡㄷㅏㄷㄱㅏ (듣다가)
ㅊㅐㅉㅈㅣㄱ (채찍)	ㅊㅏㅈㅈㅣ (찾지)
ㄷㅜㄹㄹㅓㅅㅆㅏㄷㅏ (둘러싸다)	ㅇㅓㅂㅅㅅㅡㅂㄴㅣㄷㅏ (없습니다)
ㅇㅖㅂㅂㅡㄱㅔ (예쁘게)	ㄷㅏㅂㅂㅕㄴ (답변)

둘째, 표 5에서 보는 것과 같이 쌍자음으로 결합 가능한 자소 'ㄱ', 'ㅅ'은 한글

에서 초성으로도 사용될 수 있고 종성으로도 사용될 수 있기 때문에 연속해서 3번 나타날 경우에 가운데 있는 자소가 종성으로 결합하는 것이 옳은지 초성으로 결합하는 것이 옳은지 판단해야 한다.

표 5. 'ㄱ','ㅅ'이 연속 3번 나타나는 경우의 예

쌍자음해체 어절	결합 가능 후보
ㅅㅐㄱㄱㄱㅏㄹ	샊갈(오답),색깔(정답)
ㅁㅣㄴㅈㅗㄱㄱㄱㅣㄹㅣ	민죾기리(오답),민족끼리(정답)

셋째, 표 6에서 보는 것과 같이 자음을 이니셜로 사용한 경우에 쌍자음을 해체한 경우인지 아닌지를 판단하여야 한다.

표 6. 자음을 이니셜로 사용한 경우와 아닌 경우의 예

이니셜로 사용한 경우	이니셜이 아닌 경우
ㄱㄱㅛㅅㅜ (ㄱ교수)	ㄱㄱㅗㄱㅈㅣㅈㅓㅁ (꼭지점)
ㄷㄷㅏㅇㄱㅜㅈㅏㅇ (ㄷ당구장)	ㄷㄷㅏㅁㄷㄷㅣ (땀띠)
ㅂㅂㅕㄴㅎㅗㅅㅏ (ㅂ변호사)	ㅂㅂㅏㅇㅈㅣㅂ (빵집)
ㅅㅅㅓㄴㅅㅐㅇ (ㅅ선생)	ㅅㅅㅏㅇㅜㅁ (싸움)
ㅈㅈㅏㅇㅎㅏㄱㅅㅏ (ㅈ장학사)	ㅈㅈㅗㄱㅈㅣㅂㄱㅔ (쪽집게)

이러한 문제를 해결하기 위해서 본 논문에서는 변환기반 학습을 이용하여 대용량의 말뭉치로부터 쌍자음 결합 규칙을 자동으로 추출하는 방법을 이용한다. 변환기반 학습은 미리 정해놓은 규칙의 틀에 따라 학습 말뭉치로부터 규칙을 추출하고, 추출된 규칙을 반복적으로 학습 말뭉치에 적용해 가면서 최적의 규칙집합을 찾아주는 방법론이다[2]. 변환기반 학습은 1995년 Eric Brill에 의해 처음 소개된 이후에 형태소 분석[2], 전치사구 접속[3], 개체명 인식[1], 화행 분석[7] 등 많은 자연어처리 분야에서 폭넓게 활용되고 있다. 쌍자음 해체 문제를 해결하기 위해 본 논문에서 사용한 규칙 틀은 표 7과 같다.

표 7. 변환기반 학습을 위한 쌍자음 결합 규칙 틀

규칙 틀 1	ch-3 ch-2 ch-1 ch0 ch+1 ch+2 ch+3 => tch
규칙 틀 2	ch-2 ch-1 ch0 ch+1 ch+2 ch+3 => tch
규칙 틀 3	ch-3 ch-2 ch-1 ch0 ch+1 ch+2 => tch
규칙 틀 4	ch-3 ch-2 ch-1 ch0 ch+1 => tch
규칙 틀 5	ch-2 ch-1 ch0 ch+1 ch+2 => tch
규칙 틀 6	ch-1 ch0 ch+1 ch+2 ch+3 => tch
규칙 틀 7	ch-3 ch-2 ch-1 ch0 => tch
규칙 틀 8	ch-2 ch-1 ch0 ch+1 => tch
설명	ch0: 결합 후보 자소(해체된 쌍자음; ㄱㄱ,ㄷㄷ,ㅂㅂ,ㅅㅅ,ㅈㅈ) ch-1: ch0의 첫 번째 앞에 있는 자소 ch-2: ch0의 두 번째 앞에 있는 자소 ch-3: ch0의 세 번째 앞에 있는 자소 ch+1: ch0의 첫 번째 뒤에 있는 자소 ch+2: ch0의 두 번째 뒤에 있는 자소 ch+3: ch0의 세 번째 뒤에 있는 자소 tch: 규칙이 적용됨에 따라 ch0이 변환될 자소

표 7에서 보는 것과 같이 본 논문에서는 결합 후보 자소의 앞뒤 3자소를 최대로 하여 8개의 규칙 틀을 사용한다. '규칙 틀 1' 이외에 7개의 틀을 추가한 이유는 좌우 문맥 정보를 축소함으로써 데이터 부족 문제를 줄이기 위한 것이다. 표 8은 규칙 틀이 적용될 학습데이터의 일부를 보여준다.

표 7과 같은 규칙 틀을 이용하여 대용량의 학습말뭉치로부터 쌍자음 결합규칙들이 자동으로 추출되면, 문장이 입력되었을 경우에 추출 규칙들을 우선순위에 따라 반복적으로 적용함으로써 쌍자음이 해체된 자소들을 결합하게 된다. 그림 1은 본 논문에서 제안한 2단계 규칙을 이용하여 자소가 해체된 문장을 보정해 주는 방법을 도식화한 것이다.

표 8. 학습데이터의 예

입력	학습데이터 및 설명
ㅁㅣㄴㅈㅗㄲㄱㅣㄹㅣ (민족끼리)	ㄴ 0 0 // ch_{-3}: 결합 후보가 아님 ㅈ 0 0 // ch_{-2}: 결합 후보가 아님 ㅗ 0 0 // ch_{-1}: 결합 후보가 아님 ㄱㄱ ㄲ 0 // ch_0: 결합 후보(ㄱㄱ->ㄲ)지만 결합되지 않음 ㄱ 0 0 // ch_{+1}: 결합 후보가 아님 ㅣ 0 0 // ch_{+2}: 결합 후보가 아님 ㄹ 0 0 // ch_{+3}: 결합 후보가 아님
ㅁㅣㄴㅈㅗㄱㄱㅣㄹㅣ (민족끼리)	ㅈ 0 0 // ch_{-3}: 결합 후보가 아님 ㅗ 0 0 // ch_{-2}: 결합 후보가 아님 ㄱ 0 0 // ch_{-1}: 결합 후보가 아님 ㄱㄱ ㄲ ㄲ // ch_0: 결합 후보(ㄱㄱ->ㄲ)이고 'ㄲ'으로 결합됨 ㅣ 0 0 // ch_{+1}: 결합 후보가 아님 ㄹ 0 0 // ch_{+3}: 결합 후보가 아님

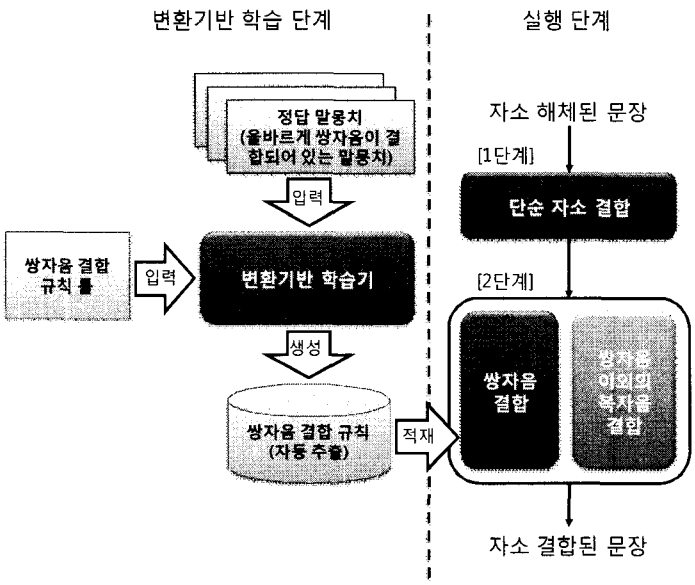

그림 1. 한글 자소 결합 모델

그림 1에서 보는 것과 같이 단순 자소 결합이나 쌍자음 이외의 복자음 결합은 별도의 학습 과정없이 수동으로 만들어진 규칙에 의해서 이루어진다. 그러나 쌍자음 결합은 학습 단계와 실행 단계로 구성된다. 학습 단계에서는 변환기반 학습 방법론을 이용하여 규칙 집합을 자동으로 생성한다. 그리고 실행 단계에서는 생성된 규칙을 적재하여 입력 문장에 적용함으로써 자소 결합을 수행한다.

실험 및 평가

단순 음절 해체와 복자음 해체 문장의 자소 결합에 대한 실험을 위하여 '21세기 세종 계획 말뭉치(858,035문장)'로부터 2종류의 실험데이터를 구축하였다. 하나는 단순 음절 해체 문장에 대한 자소 결합을 실험하기 위한 것으로 세종 계획 말뭉치 내의 모든 음절을 초성, 중성, 종성으로 단순 분리한 말뭉치이다. 다른 하나는 복자음 해체 문장에 대한 자소 결합을 실험하기 위한 것으로 단순 분리한 말뭉치에서 복자음으로 이루어진 초성이나 종성을 다시 분리한 말뭉치이다. 실험은 학습데이터와 평가데이터를 9:1의 비율로 나눈 후, 10배 교차 검증(10-fold cross validation)을 수행하였다. 표 9는 변환기반 학습의 결과로 추출된 쌍자음 결합 규칙들 중 상위 5개를 보여준다.

단순 해체 음절 결합과 쌍자음 이외의 복자음 결합에 대한 실험 결과는 100%의 정확률을 보였다. 그리고 쌍자음 해체 음절 결합에 대한 실험 결과는 표 10과 같

표 9. 변환기반 학습 결과로 추출된 쌍자음 결합 규칙

순위	규칙 및 설명
1	ㄱㄱ ㅏ => 0 //결합하지 않음
2	ㄱㄱ {ㅑ or ㅒ or ㅕ or ㅖ or ㅛ or ㅠ or ㅢ} => 0 //결합하지 않음
3	{ㅝ or ㅜ or ㅟ} ㄱㄱ => 0 // 결합하지 않음
4	ㄱㄱ o {space} =>0 // 결합하지 않음
5	{ㅗ or ㅚ} ㄱㄱ => 0 // 결합하지 않음

표 10. 쌍자음 해체 음절 결합에 대한 실험 결과

	정확률	잘못 결합된 쌍자음 수/실험대상 쌍자음 수	학습된 규칙 수
Fold-01	99.97%	274/1,077,001	761
Fold-02	99.98%	237/1,076,067	757
Fold-03	99.98%	246/1,075,561	766
Fold-04	99.98%	254/1,076,184	759
Fold-05	99.98%	257/1,076,841	755
Fold-06	99.98%	240/1,076,700	775
Fold-07	99.98%	233/1,070,834	748
Fold-08	99.98%	263/1,071,291	757
Fold-09	99.98%	247/1,072,674	765
Fold-10	99.98%	245/1,075,563	756
평균	99.98%	249.60/1,074,872.00	759.90

표 11. 데이터 희소 문제로 발생하는 오류의 예

평가데이터에 출현한 형태	잘못 결합된 형태
조끄만	족그만
도끼를	독기를
뜨바띠도	뜨받디도
걷든가	거뜬가
돋더니	도떠니
부뜨베르그교수	붇드베르그교수
들어뿌렀어	들업부렀어
아레자도오스빠에게	아레자도오습바에게
비숍박물관	비쇼빡물관
밥버러지	바뻐러지
잡부덜도	자뿌덜도
굽발치	구빨치

표 12. 규칙 우선순위 문제로 발생하는 오류의 예

평가데이터에 출현한 형태	잘못 결합된 형태
백가지	배까지
악기	아끼
음악가	음아까
성악가	성아까

이 평균 99.98%의 정확률을 보였다. 이것은 본 논문에서 제안한 방법이 음절 해체 문제를 해결하는데 매우 효과적임을 보여준다.

쌍자음 결합을 잘못한 경우를 살펴본 결과, 다음과 같이 2가지 종류의 에러 유형이 있음을 알았다. 첫째는 데이터 희소 문제로 인해 발생하는 오류이다. 표 11은 학습 말뭉치에서 출현 빈도가 매우 낮아 발생하는 오류들의 예를 보여준다.

둘째는 규칙의 우선순위 문제로 인해 발생하는 오류이다. 예를 들어, '아까'와 '음악가'라는 두 단어가 학습 말뭉치에서 100:1의 비율로 출현했다면, 'ㅇㅏㄱㄱㅏ'를 '아까'로 결합하는 규칙이 'ㅇㅡㅁㅇㅏㄱㄱㅏ'를 '음악가'로 결합하는 규칙보다 우선순위가 높게 된다. 그 결과 '음아까'라는 잘못된 결합이 발생한다. 표 12는 규칙의 우선순위 문제로 발생하는 오류들의 예를 보여준다.

결 론

본 논문에서는 통신체에서 자주 나타나는 음절해체 문제를 해결하기 위하여 2단계 한글 자소 결합 방법을 제안하였다. 제안한 방법은 1단계로 휴리스틱을 이용하여 단순하게 초성, 중성, 종성으로만 해체된 음절을 결합한다. 그리고 2단계에서 변환기반 학습 방법을 이용하여 쌍자음까지 해체된 음절을 결합한다. 단순 해체 음절 결합과 쌍자음 해체 음절 결합에 대한 실험에서 제안한 방법은 각각 100%와 99.98%라는 매우 높은 정확률을 보였다. 실험 결과, 제안한 방법이 음절 해체 문제를 해결하는데 단순하지만 매우 효과적임을 알 수 있었다. 마지막으로 제안한 방

법을 기존의 철자오류 보정 시스템들의 전처리기로 활용한다면 정확률 향상에 기여할 수 있을 것으로 생각된다.

참고문헌

[1] Black, W. J. and Vasilakopoulos, A. (2002), Language-independent named entity classification by modified transformation-based learning and by decision tree reduction, Proceedings of CoNLL'2002.

[2] Brill, E. (1995), Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging, Computational Linguistics, 21-4, 543-565.

[3] Brill, E. and Resnik, P. (1994), A rule-based approach to prepositional phrase attachment disambiguation, Proceedings of COLING'94.

[4] Kashyap, R. L. and Oommen, B. J. (1984), Spelling correction using probabilistic methods, Pattern Recognition Letters.

[5] Li, M., Zhu, M., Zhang, Y., and Zhou, M. (2006), Exploring distributional similarity based models for query spelling correction, Proceedings of ACL 2006.

[6] Mays, E., Damerau, F. J., and Mercer, R. L. (1991), Context based spelling correction, Information Processing and Management.

[7] Samuel, K., Carberry, S., and Vijay-Shanker, K. (1998), Dialogue act tagging with transformation-based learning, Proceedings of COLING/ACL'98, 1150-1156.

[8] 강승식 (2001), 음절 bigram를 이용한 띄어쓰기 오류의 자동교정, **음성과학회논문지**, 8-2.

[9] 노형종, 차정원, 이근배 (2006), 띄어쓰기 및 철자 오류 동시교정을 위한 통계적 모델, **제 18회 한글 및 한국어 정보처리 학술대회 논문집**.

1 차원고접수 : 2007. 9. 6
2 차원고접수 : 2007. 10. 24
최종게재승인 : 2008. 4. 30

(*Abstract*)

Assembling Disjoint Korean Syllables Using Two-Step Rules

Jooho Lee Harksoo Kim

Kangwon National University

With increasing usages of a messenger and a SMS, many young people are habitually using a new-style of sentences with intentionally disjoint Korean syllables. To develop a natural language interface system in these environments, we should first develop a technique that converts a sequence of disjoint Korean syllables to a correct sentence. Therefore, we propose a method to assemble a sequence of disjoint Korean syllables into a correct sentence by using two-step rules. In the first step, the proposed method assembles CVC (consonant-vowel-consonant) forms of simple-disjoint Korean syllables by using manual heuristic rules. In the second step, the proposed method assembles CCVCC forms of double-disjoint Korean syllables by using a mapping table and a transformation-based learning technique. In the experiment, the proposed method showed the perfect precision of 100% in assembling simple-disjoint Korean syllables and the high precision of 99.98% in assembling double-disjoint Korean syllables.

Keywords : Disjointing Korean syllables, assembling Korean syllables, transformation-based learning