양방향 최장일치법을 이용한 한국어 띄어쓰기 자동 교정 시스템

최 재 혁

부산여자대학교 컴퓨터교육과

Automatic Korean Spacing Words Correction System With Bidirectional Longest Match Strategy

Jae-Hyuk Choi

Dept. of Computer Education, Pusan Women's University

요 약

기존의 맞춤법 검사기의 단점인 오류 수정 작업과 처리 시간을 감소시키면서, 높은 오류 교정의 정확률을 보장하는 자동 오류 교정 시스템의 개발을 위한 첫 단계로써 한국어 오류의 80% 이상을 차지하는 띄어쓰기 오류에 대한 자동 교정 시스템을 개발하였다. 본 논문에서는 우리가 사용하는 일반 문서에서 띄어쓰기가 잘못된 단어에 대한 교정과 오류 단어에 대한 검색을 행하기 위하여, 띄어쓰기 교정 시스템의 개발 단계에서 현실적으로 고려해야 할 사항과 교정 정확률 및 처리 속도를 높이기 위한 본 시스템의 띄어쓰기 오류 루틴을 제시한다. 본 시스템의 처리 결과, 올바른 어절을 제외한 띄어쓰기가 잘못된 오류 단어(띄붙 오류와 붙띄 오류 포함)에 대해 약 98.7%의 띄어쓰기 교정 성공률을 보였다.

1. 서론

최근에 개발된 워드프로세서의 맞춤법 검사기는 오류 단어만을 탐색하고, 발견된 오류 단어에 대한 여러 오류 후보군을 제시하는 수준으로, 사용자가 모든 오류 단어에 대해 일일이 수정을 해야하는 번거로움을 초래할 뿐만 아니라, 이러한 수정 작업에 많은 시간을 낭비하게 된다. 따라서 빠르면서도 오류 단어를 자동으로 정확하게 교정해 주는 오류 교정 시스템의 개발이 절실히 요구되고 있다. 그러나 한국어 자체의 처리도 어렵거니와 글쓰는 사람의 의도에 따라 오류가 정답일 수도 있기 때문에 현재의 기술로는 모든 오류에 대한 100%의 교정은 거의 불가능하지만, 본 논문은 가능한 높은 수준의 교정률을 보장하고, 사용자의 불편을 최소화하는 오류 교정시스템을 개발하기 위하여 그 첫단계로써 한국어 오류의 80% 이상을 차지하는 띄어쓰기 오류 자동 교정 시스템을 구현하였다.

지금까지의 띄어쓰기 자동화에 대한 연구는 두 가지 접근 방법으로 분류할 수 있다[1].

(a) 어휘 지식과 휴리스틱을 이용한 접근 방법

(b) 통계 정보를 이용한 접근 방법

(a)방법은 주어진 입력 문장으로부터 띄어써야할 위치를 결정할 때 어휘 사전을 참조하게 되는데, 만일 띄어써야 할 가능성이 두군데 이상이면 휴리스틱을 적용하여 그 중 가장 적절하다고 판단되는 곳을 선택한다. (b)방법은 음절과 음절 혹은 어절과 어절 사이에 띄어써야

할 가능성이 어느 정도인지를 나타내는 통계정보를 바탕으로 띄어쓰기를 한다. 이러한 통계정보는 말뭉치와 같은 대량의 데이터로부터 자동 습득한다. 이러한 통계정보를 이용하는 방법은 필요한 정보를 구축하기는 매우 쉽지만 정확도를 향상시키는데는 한계를 가진다[2]. 본 시스템과 [2]는 (a)방법을, [3]은 (b)방법을 이용한다.

또한 기존의 한국어 띄어쓰기의 처리 시스템은 입력 문서의 형태에 따라 크게 2가지로 분류할 수 있다.

(a) 띄어쓰기가 전혀 이루어지지 않은 문서

(b) 띄어쓰기가 부분적으로 이루어지지 않은 문서

[1][2]는 (a)문서를, 본 시스템과 [3]은 (b)문서를 처리한다.

본 시스템 부분적으로 띄어쓰기가 이루어지지 않은 문장의 어절을 중심으로 띄붙오류와 붙띄오류를 처리하기 위하여 사전 참조 횟수가 가장 적은 형태소 분석 방법인 양방향 최장일치법을 이용하여 형태소 분석을 행한 후, 띄어쓰기 교정 루틴을 통해 교정을 행하게 된다.

본 논문의 2장에서는 본 시스템의 구성도를, 3장에서는 띄어쓰기 교정시스템 개발 시의 고려사항을, 4장에서는 실험 및 평가 결과를 제시하고 5장에서는 결론 및 향후 연구 과제를 기술한다.

2. 띄어쓰기 교정 시스템

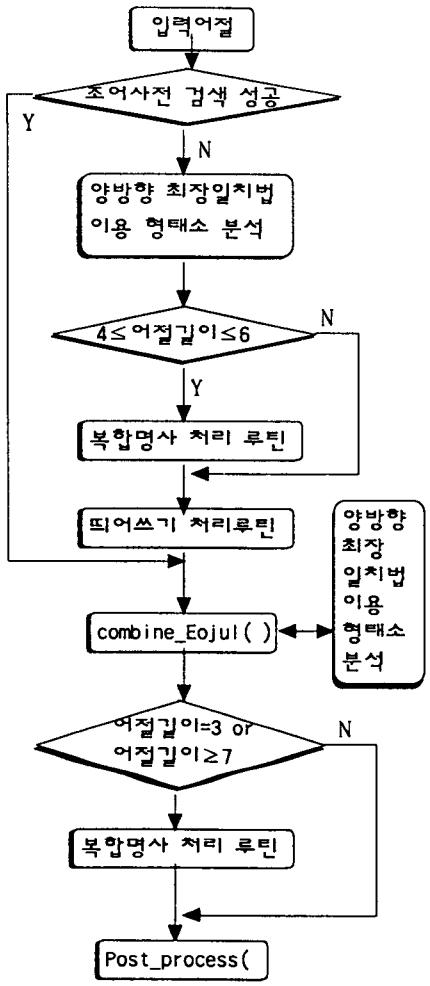
띄어쓰기 오류는 한국어 문서상에서 나타나는 오류의 약 80% 정도를 차지하는 가장 빈번히 나타나는 오류 유형이다. 특히 신문에서는 단순히 지면을 절약한다는 관점에서 띄어쓰기 규정을 무시하지만, 신문에 익숙해 있는 독자들의 문서에는 이를 그대로 사용하는 경우가 대부분이며, 띄어쓰기 규정을 정확히 알고 사용하는 사람은 극히 드물다. 이러한 띄어쓰기는 국어에 대한 문법 사항, 즉 품사에 대한 이해의 표현이므로 이러한 띄어쓰기 오류는 국어학의 관점에서 볼 때 심각한 문제로 받아들여야 할 것이다.

띄어쓰기 오류는 띄어써야 할 것을 붙여쓴 띄붙오류와 붙여써야 할 것을 띄어쓴 붙띄오류로 크게 나눌 수 있다. 띄붙오류와 붙띄오류에 해당하는 오류의 세부 유형을 표1에서 제시한다.

표1. 붙띄오류와 띄붙오류의 세부 유형

띄붙오류	1) 용언+어미+의존명사,명사 :말하는것은 2) 명사+열거단어+명사 : 선생및학생 3) 숫자+여 : 10여가지 4) 용언+조사+용언 : 만져도보고 5) 관형사,수사,부사+명사 : 새책 6) 명사+명사(복합명사) : 사회부조리 7) 용언+어미+용언 : 죽고싶다 8) 명사(성,이름 포함)+호칭,관직명 9) 외국어지명+지명접미사 :걸프만 10) 수사+단위명사 : 한가마 11) 몇+명사 : 몇사람 12) 숫자 : 12억3456만 13) 합성동사+보조용언 : 덤벼들어보아라 14) 부호+글자 : 눈,코,귀 15) 명사+조사+용언 : 힘이들다 16) 2음절 성+이름 : 황보영조 .....
붙띄오류	1) 명사+명사 : 내복 약 2) 보조용언 : 넘어 가다 3) 첩어 : 한들 한들 4) 명사+조사,부사 : 당신 보다 5) 명사+용언 : 힘 들다 6) 성+가/씨 : 최 씨 7) 고유명사+지명접미사 : 설악 산 8) 접두사+명사 : 풋 사랑 9) 조사/어미+조사/어미 : 학교에서 부터 10) 붙여써야 하는 의존명사 : 이 것 11) 몇+[수]+수사 : 몇 십년 12) 명사+하다,되다,지다,시키다 : 사랑 하다 .....

그림1은 본 시스템의 개략적인 흐름도이다. 그림1에서 형태소 분석에 실패한 어절에 한해 조사어미를 제외한 4음절에서 6음절 사이의 단어에 대해 복합명사 분리를 먼저 행한 후, 분리



<그림 1> 띄어쓰기 교정 시스템의 개략적인 흐름도

에 실패하면 띄어쓰기 처리 루틴을 호출한다. 이는 복합 명사 분리와 띄어쓰기 처리 루틴에서 동시에 분리 가능한 단어에 대해서는 복합명사의 분리가 정확한 분리일 가능성이 더 높은 것으로 실험 결과 나타났다. 그러나 형태소 분석에 실패한 7음절 이상의 단어에 대해서는 사전 참조 횟수를 감소시키기 위해 띄어쓰기 분리 루틴을 먼저 처리하고 실패하면 복합명사 분리 루틴을 호출하도록 하였다. 따라서 분리의 정확성과 처리 속도(디스크 참조 횟수)의 감소를 위해 복합명사 분리 루틴을 음절수에 따라 분리하여 처리하였다.

Combine_Eojul( ) 루틴은 어절과 현 어절을 합쳐 새로운 어절을 만든 후 형태소 분석을 행하여 분석에 성공하면 붙띄오류로 간주하고 두 어절을 붙여쓴다. 띄어쓰기 처리 루틴은 형태소 분석에 성공한 어절이더라도 붙띄오류 처리 루틴을 호출한 후, 분리에 실패한 어절에 대해서만 띄붙오류 처리 루틴을 호출한다.

띄어쓰기 교정을 위한 양방향 최장일치법은 일반 형태소 분석 과정에서 처리할 수 있는 띄어쓰기 오류는 바로 분리한다. 예를 들어, 접두사 처리시 '못하다', '못생기다', '못나다', '못되다'의 붙띄오류나, "라면되지"의 '라면 되지'로의 분리, "자연인들이"와 같이 접미사 '들'앞에 다시 '접미사'가 나타난 경우의 처리 등이다.

3. 띄어쓰기 교정 시스템 구현 시 고려 사항

지금까지 개발된 오류 교정 및 검사 시스템은 크게 형태소 분석을 중심으로 구현된 것[2][4]와 corpus를 중심으로 구현된 것[3]으로 나눌 수 있다. 형태소 분석을 이용한 교정 및 검사 시스템은 오류어에 대한 교정 후보의 범위를 최소화하기 위하여 형태소 분리가 요구되므로 형태소 분석기를 기반으로 하는 것이 효율적이며, 형태소 분석 결과를 철자 오류 교정, 문법 검사, 주제어 검색, 문서 요약 등 문서 편집기에 추가될 기능에서 공유될 수 있다는 장점이 있다. 또한 고수준의 교정 시스템을 구현하기 위해 필요한 구문 분석 결과를 추가하기 용이한 장점도 있다[4].

일반적인 교정 시스템에서 오류 어절에 대한 교정 결과 어절은 1개만 생성되어야 하며, 교정 결과가 100% 정확히 교정되었다는 보장을 할 수 없기 때문에, 교정 결과를 사용자에게 반드시 검증을 받아야 한다. 본 시스템에서는 사용

자의 편의를 위해 페이지 단위로 일괄적으로 오류 어절에 대한 교정 결과를 검증 받도록 하였다.

틀맞오류와 맞틀오류가 발생되지 않도록 고수준의 띄어쓰기 교정 시스템을 구현할 때 연구되어야 할 사항은 다음과 같다.

첫째, 어미/조사와 의존 명사와의 명확한 구분을 할 수 있는 방법을 연구하여야 한다.

예) 'ㄴ바', 'ㄴ지', 'ㄹ걸', '던데', 'ㄹ망정', 'ㄹ지', '은데', '은지', '을게', '을걸', '일지' 등의 어미에서 뒷 부분의 단어가 의존명사이면 띄어쓰기를 하여야 한다.

둘째, 접미사와 다른 품사(의존명사, 명사 등)와의 명확한 구분을 할 수 있는 방법을 연구하여야 한다.

예) "출장중이다"에서 '중'은 접미사로 붙여쓰고, "연필중에서"에서 '중'은 의존명사로 띄어써야 한다. 또한 "언어권에서"의 '권'은 접미사로 붙여쓰고, "한권"에서의 '권'은 단위명사로 띄어써야 한다.

셋째, 붙띄오류를 처리하기 위해서는 앞 어절 혹은 뒷 어절의 단어와 현 어절을 합쳐서 한 단어로 만들어 분석을 해야 하는데, 이는 처리 속도에 엄청난 영향을 끼치고 모호성을 유발할 수 있다. 따라서 결합 제약 조건 및 rule에 대한 연구가 이루어져 꼭 필요하지 않은 단어들간의 결합을 행하지 않도록 하여야 한다.

예) "살 수"를 결합하여 "살수"로, "그대로 서 있어"는 "그대로서 있어"와 같은 원치 않는 분석 결과가 발생될 수 있다.

넷째, 분리의 모호성이 존재할 경우 처리하는 기준을 마련하여야 한다.

예) "각시대마다"에서 "각 시대마다"와 "각시 대마다"로 분리 가능하다.

다섯째, 미등록어를 포함한 어절이나 복합명사를 포함한 어절에서 문법형태소를 분리한 후, 이후에 나타나는 문법형태소를 제외한 어휘형태소가 같은 어절에 대해서는 오류 어절로 출력하지 않고 어휘형태소의 교정 결과를 그대로 적용하도록 함으로써 출력해야 하는 오류 단어의 수를 감소시키도록 하여야 한다.

여섯째, 오류 교정의 높은 정확률을 보장하면서 처리속도를 빠르게 할 수 있는 algorithm 개발에 중점을 두어야 한다. 교정시 형태소 분석을 행하는 시스템은 사전 참조가 필수적이며, 가급적 사전 참조를 적게 하도록 하여 처리 속도를 높이는 방안을 연구하여야 한다.

본 시스템에서는 위에서 제시한 6가지에 대해 다음과 같이 처리하였다.

첫 번째와 두 번째에서 제시한 조사/어미, 접미사와 의존명사/명사와의 구분은 조사/어미, 접미사 뒤에 붙은 조사에 의해 의존명사로의 구분이 가능한 것은 띄붙오류로 처리하여 교정을 행하고, 구분이 불가능한 것은 설명문과 더불어 예를 제시하여 사용자가 쉽게 판단할 수 있도록 하였다. 세 번째, 붙띄오류 처리를 위한 두 어절 결합 제약 조건은 '살 수', '해 지', '한 자리', '나 이를' 등의 결합해서는 안되는 두 단어를 조사하여 이를 <결합 불가능 Table>에 저장하여 처리하였으며, 또한 결합한 어절의 길이가 7음절 이상이면 결합하여 처리하지 않도록 하였다. 이는 한국어에서 7음절 이상의 단어의 빈도수가 복합명사를 포함하여 0.5%가 되지 않고[4], 이들 한 단어가 두 단어로 잘못 분리하여 사용되는 경우가 거의 발생되지 않음으로 처리의 overhead를 줄이기 위함이다.

네 번째, 분리 모호성이 발생하는 어절에 대한 처리를 위해, 본 시스템은 복합명사 분리 루틴과 표2에서 제시한 띄어쓰기 오류 루틴의 처리 순서에 의해 1개의 확률 높은 처리 결과만

을 출력하도록 하였다. 궁긍적으로 오류 어절이 복합명사나 띄어쓰기 오류 사전에 저장되어 있다면 복합명사 분리나 띄어쓰기 오류 처리 루틴을 호출하지 않고 단 한번의 사전 참조로 교정이 가능하므로, 이러한 오류 사전이 말뭉치를 얼마나 많이 수록하고 있는지가 처리 속도 및 교정 정확률의 향상에 가장 큰 영향을 미친다. 그러나 모든 오류를 사전에 수록한다는 것은 불가능하므로 표2의 띄어쓰기 처리루틴의 성능이 교정 시스템의 성능을 사전 다음으로 좌우하게 된다.

다섯 번째, 같은 어휘형태소를 가지는 어절에 대해서는 한번만 교정 결과가 나오도록 처리하였다. 여섯 번째, 본 시스템은 형태소 분석시 사전참조를 가장 적게 하는 양방향 최장일치법의 이용과 띄어쓰기 오류 루틴에서 참조를 적게 해도 분리 가능한 오류 어절을 먼저 처리하도록 함으로써 해결하였다.

표2. 띄어쓰기 오류 루틴의 내용

1. 붙띄 오류 처리 루틴 - 하다/되다/지다 의 붙띄오류/띄붙오류 처리 루틴 - 수사의 붙띄오류 처리 루틴 - 접미사와 명사 구별 처리 - 어미/조사 와 의존명사 구별 처리 루틴
2. 띄붙 오류 처리 루틴 - '있다/없다/아니다/아니하다/않다/싶다' 포함 어절 처리 루틴 - '몇' 포함 어절 처리 루틴 - '는/은/을/를'포함 어절 처리 루틴 - '게/고/야/지'포함 어절 처리 루틴 - 조사 포함 어절 처리 루틴 - 조사 포함 어절 처리 루틴 - ㄴ/ㄹ 로 끝나는 용언 포함 어절 처리 루틴 - 관형사/부사 포함 어절 처리 루틴 - 어미 포함 어절 처리 루틴 - '전/후/앞/뒤/...'포함 어절 처리 루틴 - 호칭/관직 포함 어절 처리 루틴 - 외국어+지명접미사 처리 루틴 - '씨'포함 어절 처리 루틴 - 분리 가능 특수 음절 포함 어절 처리 루틴 - 수사 처리 루틴 - '타/신/구/...' 등의 접두사 포함 어절 처리 루틴 - 성+이름+'씨/가' 포함 어절 처리 루틴 . . .

표2에서 제시한 처리 루틴의 순서는 사전 참조 많이 하지 않으면서 분리 가능한 것을 먼저 처리하되, 분리 모호성이 발생하는 처리 루틴들은 정확한 분리의 확률이 높은 것을 우선적으로 처리하도록 하였다.

4. 실험 및 평가

본 시스템에서 구성한 한국어 어휘사전은 약 24만 단어, 영어 어휘 사전은 약 11만 단어를 수록하고 있으며, IBM PC상에서 Boland C++로 구현하였다. 일반적인 오류 교정 시스템은 형태소 분석을 위한 사전 참조에 교정을 위한 추가적인 사전 참조가 이루어지므로 수행 시간과 정확률이 가장 큰 문제점으로 대두된다. 본 시스템은 형태소 분석시 기존의 형태소 분석 방법론 중 가장 사전 참조 횟수가 적은 양방향 최장 일치법을 띄어쓰기 교정에 적합하게 변형하였으며, 띄어쓰기 교정 시에도 교정 정확률을 높이면서 가능한 사전 참조 횟수를 적게 하도록 하는 수행 순서를 제시하였다.

실험 대상 문서 신문 사설, 논문, 일반 서류 문서 등에서 추출한 띄어쓰기 오류 어절 780개를 대상으로 처리한 결과 접미사/조사/어미 와 명사(의존명사 포함)와의 모호성 어절이 4개, "교육과정의특성"(교육 과정의 특성, 교육과 정의 특성) 등과 같은 분리 모호성 어절이 3개, "주장하던지주위를더욱놀라게했다는"과 같이 일반 문서에서 거의 나타나지 않는 5개이상의 어절이 띄어쓰기가 되지 않은 어절 1개가 처리되지 않은 것으로 나타나 본 시스템은 98.7%의 교정 성공률을 보였다.

그러나 이러한 교정 성공률은 처리 대상 문서의 종류에 따라 달라질 수 있어 큰 의미를 부여할 수 없지만, 본 시스템은 해결하기 힘든

종류의 띄어쓰기 오류 어절을 제외한 대부분의 오류 어절은 교정이 가능하였고, 실험 결과 나머지 모호성 어절에 대해서도 오류의 설명문을 추가하여 사용자가 큰 불편함과 어려움 없이 정확한 교정을 할 수 있도록 하였다.

다만 본 시스템은 현재 완전히 개발이 끝나지 않은 관계로 사전의 오류 및 미등록어와 오류 어절 사전에 대한 말뭉치 단어 보완이 추가되고, 앞으로 테스트 및 튜닝 작업을 거쳐 좀 더 보완한다면 조만간 상용화가 가능할 것으로 예상된다.

5. 결 론

본 시스템은 기존의 맞춤법 검사기의 단점인 오류 수정 작업과 처리 시간을 감소시키면서, 높은 오류 교정의 정확률을 보장하는 자동 오류 교정 시스템의 개발을 위한 첫 단계로써 한국어 오류의 80% 이상을 차지하는 띄어쓰기 오류에 대한 자동 교정 시스템이다. 본 시스템은 오류 어절에 대해 한 개의 교정 결과만을 제시하여 한 페이지 단위로 교정 결과를 일괄적으로 보여주면, 사용자가 교정 결과를 확인하도록 하였으며, 확인된 결과는 그 다음부터는 출력되지 않고 그 파일이 끝날 때까지 자동 교정되도록 하였다. 본 논문에서는 우리가 사용하는 일반 문서에서 띄어쓰기가 잘못된 단어에 대한 교정과 오류 단어에 대한 검색을 행하기 위하여, 띄어쓰기 교정 시스템의 개발 시 고려해야 할 사항과 교정 정확률 및 처리 속도를 높이기 위해 본 시스템에서 구축한 띄어쓰기 오류 루틴을 제시하였다. 본 시스템의 처리 결과, 띄어쓰기가 잘못된 780개의 오류 어절(띄붙 오류와 붙띄 오류 포함) 중 분리 모호성을 가진 10개의 오류 어절을 제외한 나머지 770개의 오류 어절에 대해 정확한 교정을 행하여 약 98.7%의 띄어쓰기 교정 성공률을 보였다.

앞으로 오류 교정 시스템을 개발하기 위하여 일반 사용자가 가장 많은 오류를 범하는 '로써'와 '로서', '음으로'와 '으므로', 본 논문에서 제시한 조사/어미/접미사와 의존명사/명사 등의 구분 방안을 연구하고, 오류 어절 사전의 보완 등이 이루어지면 상용화도 가능할 것이다.

참 고 문 헌

[1]. 심광섭, "음절간 상호 정보를 이용한 한국어 자동 띄어쓰기", 한국정보과학회 논문집 23권 9호, pp.991-1000, 1996

[2]. 김계성, 이현주, 이상조, "음절 정보를 이용한 한국어 띄어쓰기의 구현", 정보과학회 추계 학술발표 논문집, 제24권 2호, 1997

[3]. 박봉래, 임해창, "코퍼스용 철자 및 띄어쓰기 오류 교정 시스템", 한국어 정보 처리 소식 제 3권 제1,2호, pp.15-27, 1995

[4]. 강승식, "한국어 형태소 분석기 HAM의 형태소 분석 및 철자 검사 기능", 제8회 한글 및 한국어 정보처리학술대회 발표논문집 pp. 246-252, 1996

[5]. 최재혁, "양방향 최장일치법에 의한 한국어 형태소 분석기의 구현", 경북대학교 공학박사 학위논문, 1994

[6]. 최재혁, "형태소 분석을 통한 한·영 자동 색인어 추출 시스템", 한국정보과학회 논문집 23권 12호 pp.1279-1288, 1996

[7]. 최재혁, "음절수에 따른 한국어 복합명사 분리 방안", 제8회 한글 및 한국어정보처리 학술대회 발표논문집 pp.262-267, 1996

[8]. 강승식, "음절정보와 복수어 단위 정보를

이용한 한국어 형태소 분석", 서울대학교 공학 박사 학위논문, 1993

[9]. 김성용, 최기선, 김길창, "Tabular Parsing 방법과 접속 정보를 이용한 한국어 형태소 분석기", 한국정보과학회 춘계 인공지능 발표논문집 pp. 133-147, 1987

[10]. 심철민, 김현진, 김영진, 권혁철, "연어정보를 이용한 한국어 철자 검사기의 기능 개선", 제7회 한글 및 한국어 정보처리 학술대회 발표논문집 pp.86-90, 1995

[11]. 이병훈, 윤준태, 송만석, "말뭉치를 기반으로 한 한국어 철자교정기의 구현", 제5회 한글 및 한국어 정보처리 학술대회 발표논문집 pp.285-293, 1993

[12]. 정한민, 이근배, 이종혁, "자판 특성을 이용한 Neuro-Fuzzy 한국어 철자 교정기의 구현", 제5회 한글 및 한국어 정보처리 학술대회 발표논문집 pp.317-328, 1993

[13]. 미승우, "새 맞춤법과 교정의 실제", 어문각, 1990

[14]. 손세모들, "국어 보조용언 연구", 한국문화사, 1996