형태소간의 의존 관계에 따른 오류 유형 추정 함수를 이용한 한국어 철자 오류 교정*

심철민, 권혁철
부산대학교 전자계산학과

Korean Spell Correction Using Collocation of Morphemes

Chul-Min Sim, Hyuk-Chul Kwon
Department of Computer Science, Pusan National University

요 약

기존 철자 검사/교정기들은 한 어절을 구성하는 형태소들의 품사 정도만을 이용하고 있다. 때문에 철자 검사나 교정의 정확도 면에서 한계를 가진다. 본 논문에서는 한국어의 구문적 연관 관계 및 구문 내에 존재하는 단어들 간의 의미적 연관 관계 등을 바탕으로 오류 유형을 추정하는 오류 유형 추정 함수를 제안하고, 이를 이용한 철자 교정기를 구현하였다. 본 논문에서 구현한 오류 유형 추정 함수를 이용한 철자 검사/교정기는 한 어절에 국한되었던 철자 검사/교정의 범위를 여러 어절로 확장하고자 하는 시도의 시발이라 할 수 있다. 따라서 구문 검사 및 의미 검사를 수행하는 문체 검사기의 원형으로서 그 의의를 가진다.

I.서론

한국어에는 한 어절을 구성하는 형태소간의 품사적 연관성 외에도 한 어절 내의 단어 결합상의 의미적 연관성, 문장 구성상의 구문적 연관성 및 문장을 구성하는 단어들 간의 의미적 연관성 등이 있다. 그러나 기존의 대부분 철자 검사/교정기들은 한 문장을 구성하는 어절들 간의 구문적, 단어 의미적 연관성을 이용하지 않고 있다[1,2].

한편 한 어절을 구성하는 형태소들 간의 연관성은 품사 정도만을 이용하고 있다. 때문에 철자 검사나 교정의 정확도 면에서 한계가 있다. 즉 문맥상으로 확실히 틀린 어절임에도 불구하고 철자 검사 과정에서는 올바른 어절로 간주되는 경우가 발생한다. 뿐만 아니라 철자 교정 과정에서는 앞뒤 어절간의 연관 관계를 이용하지 않기 때문에 구문상, 의미상으로 틀린 후보 어절을 제시하기도 한다.

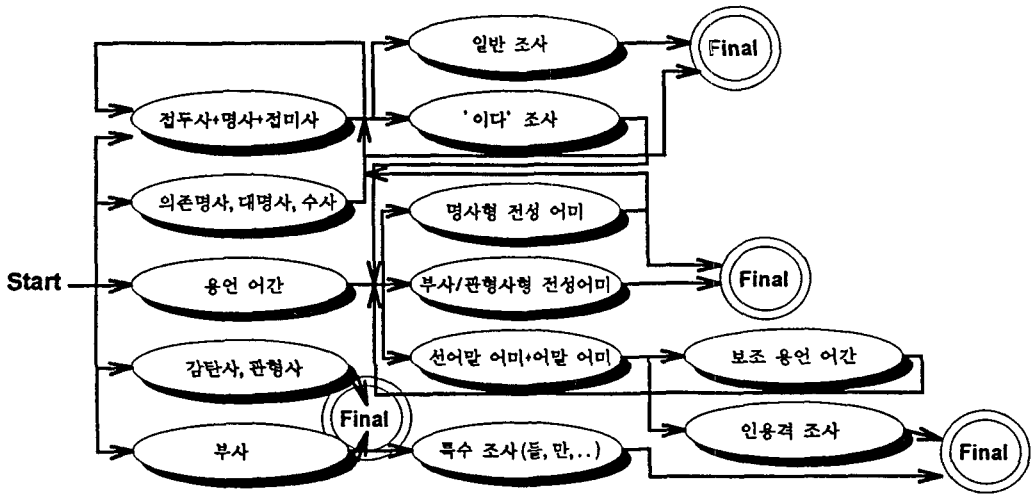
본 논문에서는 한국어의 구문적 연관 관계 및 구문 내에 존재하는 단어들 간의 의미적 연관 관계 등을 바탕으로 오류 유형을 추정하는 오류 유형 추정 함수를 제안하고, 이를 이용하여 철자 교정기를 구현하였다. 오류 유형 추정 함수는 지식 베이스를 참조하여 오류 유형을 추정하며, 철자 교정기는 오류 유형 추정 함수에서 추정된 결과를 바탕으로 교정 계획을 수립한다. 수립된 교정 계획에 따라 교정 과정에서 중점적으로 검사할 형태소 영역과 교정 기법 적용 순서가 동적으로 재조정된다.

본 논문에서 사용한 교정 함수는 교정 속도는 감소시키면서 교정률은 증가시킴으로 전체적으로 철자 교정기의 교정 성능이 향상되었다.

II.오류 유형 추정과 형태소 분석

한국어 문서 내에서의 맞춤법 및 문구상의 오류는 크게 어절 내의 오류와 어절간의 오류로 나눌 수 있다. 어절 내의 오류는 음소 변형 오류, 오용어, 띄붙 오류 등을 들 수 있으며 어절간의 오류는 붙띄 오류와 일부 수사 오류, 문장 부호 오류, 구문 오류, 의미 오류 등을 들 수 있다[3].

*본 논문은 STEP2000('94 핵심S/W 기술개발 사업) 중 '국어정보처리 기술 개발'사업의 과제인 한국과학기술연구원 지능형 처리기 개발 과제 연구비에 의해 연구되었음.



<그림 1> 한국어 한 어절의 형태소 오토마타

이러한 오류들을 교정할 때 원래의 오류 유형에서 어긋나는 교정을 시도하는 경우 즉, 띄어 쓰기 오류인데 음소 변형 교정을 시도하는 경우와 같이 불필요한 교정을 시도할 때가 있다. 기존 철자 교정기에서는 교정 과정에서 최대 130여 회에서 평균 50여 회까지 형태소 분석을 시도하고 있다. 그러나 오류 유형을 정확히 추정할 수 있을 경우에는 추정된 오류에 집중된 교정만을 시도하므로 평균 32회 정도로 형태소 분석 시도 횟수가 감소된다.

2.1 철자 오류의 형태소 오토마타적 표현

한국어에서 한 어절은 <그림 1>과 같은 한 어절 형태소 오토마타로써 표현될 수 있다. 본 논문에서는 이러한 형태소 오토마타의 측면에서 다음과 같이 철자 검사 및 교정 과정을 정의한다.

정의 1. 철자 검사는 한국어 어절 오토마타 상에서 주어진 어절 내의 음절들을 소비하며 상태를 이동하면서 수렴 여부를 검사하는 과정이다.

정의 2. 철자 교정은 한국어 어절 오토마타에 의해 인식되지 않는 어절을 가공하여 한국어 오토마타에서 인식될 수 있는 어절로 만드는 과정이다.

이와 같은 한국어 오토마타 상에서의 어절 내 오류는 다음과 같이 정의된다.

오류 1. 띄붙 오류는 한국어 어절 오토마타 상에서 종료 상태(final state)를 만났음에도 불구하고 어절이 끝나지 않았을 경우이다.

오류 2. 붙띄 오류는 2가지 유형으로 구분된다. 하나는 오토마타 상에서 현재 상태가 종료 상태가 아니고 음절도 모두 소모되지 않았는 데도 다음 상태로 이동하기 위한 제약 조건을 만족하는 상태가 하나도 없는 경우이다. 또 다른 하나는 모든 경우의 형태소 분석에 실패한 경우이다.

오류 3. 음소 변형 오류는 한국어 어절 오토마타 상에서의 띄붙 오류와 붙띄 오류의 경우를 모두 포함한다.

어절 간 오류는 오류 형태에 따라 2가지로 분류된다. 하나는 각각의 어절들은 모두 형태소 오토마타에서 수렴되므로 올바른 어절로 분석되지만 문구상 오류로 판단할 수 있는 경우이다. 또 다른 하나는 오류 어절을 발견하였지만 정확한 교정을 위해서는 앞 뒤 여러 어절을 참조해야만 하는 경우이다. <표 1>은 이 2가지 어절 간 오류의 예를 보여 준다.

구문 오류		여러 어절 오류	
오류	대치어	오류	대치어
만의 하나 마주 치다 먹기 마련	만에하나 마주치다 먹게마련	아브라함 링칸 4.19 의가가 MBC 방송굳	아브라함 링컨 4.19 의거가 MBC 방송국

<표 1> 어절 간 오류 예

2.2 오류 유형 추정을 위한 형태소 분석의 특징

오류 유형을 추정하기 위한 형태소 분석은 가능한 분석 결과를 모두 생성하는 일반적인 형태소 분석과는 다른 점이 있다. 다음은 오류 유형을 추정하기 위한 형태소 분석 정보를 생성하는 원칙들이다.

원칙 1. 철자 검사에서는 기본적으로 한 어절 오토마타에서 수렴되면 형태소 분석을 종료한다.
원칙 2. 추가의 분석 정보를 필요로 하는 특별한 오류의 경우 형태소 분석을 계속한다.
원칙 3. 형태소 분석 과정에서 오류 유형 추정 정보를 남긴다.
원칙 4. 특별한 경우에 한하여 부분적인 형태소 분석도 가능하다.

철자 검사 단계는 형태소 오토마타를 따라가며 어절을 소모하는 과정이다. 이 과정에서는 사전 정보에 따라 한국어 한 어절 오토마타에서 수렴 여부를 시도해 보고 수렴되는 경로가 있으면 그 경로를 저장하고 종료한다. 이 방법은 기존 철자 검사/교정기들이 사용하는 기본적인 방법이다.

기본적인 방법으로는 처리할 수 없는 모호성을 가진 특별한 오류의 경우는 형태소 오토마타에서 수렴되었음에도 불구하고 추가의 형태소 분석을 해야만 한다. 이 경우는 필요한 형태소 분석이 끝난 뒤 오류 유형 추정 함수에서 최종적으로 오류 여부를 판단한다.

오류 유형 추정을 위해서는 가능한 한 정확한 정보를 필요로 하므로 철자 검사 과정에서 중요한 분석 정보는 따로 저장한다. 예를 들면 띄어 쓰기 위치나 최대한 옳다고 판단되는 위치 등의 정보로써 띄어 쓰기 오류를 대부분 추정하고 교정할 수 있다.

오류 유형이 추정된 후 철자 교정기는 추정된 오류의 종류별로 부분적인 한국어 형태소 오토마타를 수행시키는 것이 효율적이다. 이를 위해서는 철자 검사에서 품사별로 경로를 검사하는 기능이 제공되어야 한다.

III. 오류 유형 추정 함수를 이용한 철자 교정 기법

3.1 오류 유형 추정 함수

본 논문에서 구현한 오류 유형 추정 함수는 다음과 같이 정의된다.

정의. I가 오류 어절의 정보를 포함한 입력 파라미터의 tuple이고 O가 오류 유형 추정 결과 tuple일 때 오류 유형 추정 함수 F()는 다음과 같이 정의된다.

$$F : I \rightarrow O$$

$$F(I) = \max\left(\sum_{i=1}^{k} G_i(I) * W_i(I)\right)$$

$G_i()$: 오류 유형별 추정 함수

$W_i()$: 추정 함수별 오류 유형 가중치

$I = \langle M, P, V \rangle$

$O = \langle S, D, W \rangle$

M : 형태소 분석 정보

P : 띄어 쓰기 후보 위치

V : 최대한 올바른 위치

S : 오류 기법 적용 순서

D : 교정 기법 적용 형태소 영역

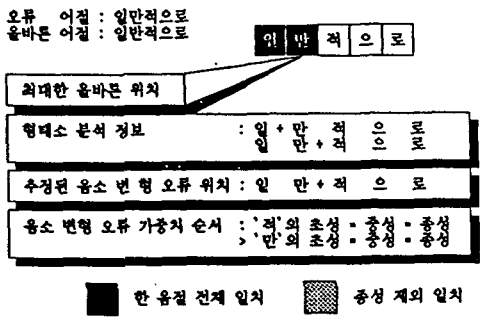
W : 가중치

철자 교정기는 이렇게 동적으로 재조정된 교정 계획에 따라 가중치가 높은 교정 기법부터 우선으로 시도한다. 후보 어절을 생성하는 과정에서는 철자 검사를 통하여 그 후보 어절을 검증하는데 이 때 오류 유형 추정 함수에서 구한 형태소 영역만을 검사한다. 이 방법은 불필요한 형태소 영역의 검사에 의해 교정 효율이 저하되는 경우를 줄인다.

3.2 형태소 연관 관계를 이용한 추정 정보

● 어절 내 오류 추정

어절 내 오류 중 오류 유형 추정의 대상이 되는 오류는 띄어 쓰기 오류와 음소 변형 오류이다. 띄어 쓰기 오류의 경우는 철자 검사 과정에서 제약 조건의 불일치 등에 의해 발생한 경우이다. 띄어 쓰기 오류를 추정하기 위해 철자 검사 과정에서 제약 조건이 어긋나는 위치가 종결 상태일 경우 그 위치 정보들을 띄어 쓰기 후보 위치로서 저장한다. 이 후보 위치들에 대한 띄어 쓰기 검사를 가중치 순서대로 시도함으로써 띄어 쓰기 오류를 교정할 수 있다.



<그림 2> 음소 변형 오류 추정 예

음소 변형 오류를 추정하기 위해서는 최대한 옳다고 추정되는 위치와 띄어 쓰기 오류를 추정하기 위해 구해진 후보 위치들을 참조한다. 띄어 쓰기 오류는 사전 검색에는 성공했지만 제약 조건이 일치되지 않을 때의 위치에 가중치를 더 주는데 비해 음소 변형 오류는 사전 검색 과정 및 활용형 처리에서의 실패 위치에 가중치를 더 준다. <그림 2>는 음소 변형 오류를 추정하는 예를 보여 준다.

음소 변형 오류 추정은 한 음절 전체 일치와 종성 제외 일치가 겹칠 경우에는 오류 위치의 음소 변형 오류 교정만을 시도하며, 음절 오류만이 존재할 경우에는 앞 음절의 음소 변형 오류로 인해 형태소 분석이 잘못되었을 가능성이 있으므로 오류 음절과 바로 앞의 음절까지 음소 대치를 시도한다.

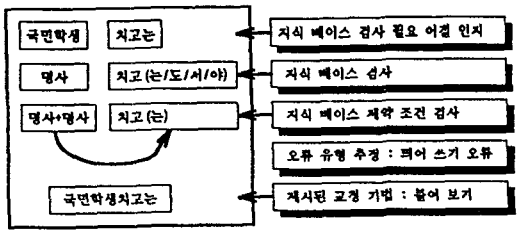
● 어절 간 오류 추정

어절 간 오류는 의미 오류를 제외하고 앞 뒤 n어절 정도를 고려하여 처리할 수 있는 오류를 처리 대상으로 한다. 이 때 n은 오류 종류에 따라 다르며, 대부분의 경우 앞 뒤 1~2어절 정도를 참조한다. 그러나 외국인의 이름이나 지명 등 여러 어절에 걸친 명사들은 3~4어절까지 고려해야 할 경우가 있다. 어절 간 오류는 어절 간 오류 지식 베이스로 구축되어 있으며 오류 유형 추정 함수에서는 지식 베이스의 어절 거리 정보에 따라 간단한 구단위 파싱을 행한다.

다음은 어절 간 오류 지식 베이스의 tuple이다.

```
유형 1. 앞 뒤 1어절 고려
   < 앞 단어열, 앞 단어 형태소 정보,
     뒤 단어열, 뒤 단어 형태소 정보,
     앞 단어 대치어, 뒤 단어 대치어,
     가중치, 도움말 번호 >
유형 2. 여러 어절 명사
   < 단어열, 연속 정보, 대치어,
     가중치, 도움말 번호 >
```

어절 간 오류 지식 베이스는 오류 추정 뿐만 아니라 철자 검사 과정에서도 참조되어 검사의 질을 높인다. <그림 3>은 구단위 파싱을 통해 어절 간 오류를 교정하는 예를 보여 준다.



<그림 3> 어절 간 오류 교정 예

3.3 heuristics를 이용한 추정 정보

형태소 분석 정보의 이용으로 어느 정도 정형화된 오류 유형은 추정할 수 있다. 그러나 실제 한국어 문서에는 상당한 공통점을 가지는 오류들이 존재한다[1,4]. 심지어 많은 사람들이 서로 다른 종류의 문서에서 똑같은 오류를 범하기도 한다. 이러한 일정한 오류들은 heuristics로써 처리함으로써 교정 시간과 교정률을 개선할 수 있다. 일반적으로 자주 범해지는 오류 유형은 많은 자료를 분석함으로써 구해질 수 있다.

구해진 오류들은 heuristics 지식 베이스에 오류 유형과 적절한 교정 방법이 함께 저장된다. heuristics 지식 베이스는 띄어 쓰기, 붙여 쓰기, 음절 대치의 3가지 tuple로 나뉜다. 다음은 heuristics 지식 베이스의 tuple들이다.

유형 1. 띄어 쓰기
< 단어열, 오류 위치, 앞 어절 형태소 정보,
뒤 어절 형태소 정보, 가중치 도움말 번호 >
유형 2. 붙여 쓰기
< 단어열, 앞 어절 형태소 정보,
현재 어절 형태소 정보, 가중치,
도움말 번호 >
유형 3. 음절 대치
< 음절, 오류 위치, 형태소 정보, 가중치,
도움말 번호 >

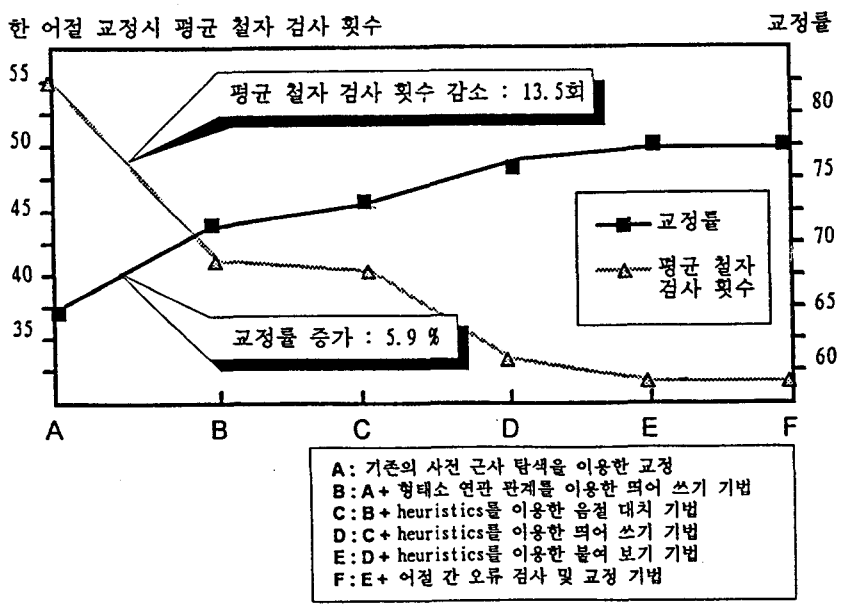
음소 대치는 일반적인 음소 변형 오류를 교정하는 기법이다. 그러나 "건너-"와 "건네-"와 같이 빈번히 발생하는 음절 단위의 변형 오류들은 음소 대치보다 한 단계 높은 음절 단위 대치 기법을 사용함으로써 교정 속도와 교정률을 개선할 수 있다.

VI. 실험

오류 유형 추정 함수의 타당성을 검토하기 위해 본 논문에서는 신문 기사 및 소설, 하이텔 통신 자료를 실험 자료로 사용한다. 그리고 교정률의 비교를 위해서 오류 유형 추정 함수를 이용한 철자 교정기와 오류 유형 추정 함수를 사용하지 않고 사전 근사 탐색에 기반한 기존의 철자 교정기를 실험 대상으로 한다[3].

철자 교정기들은 각자 독특한 교정 기법을 사용하고 있으므로 절대적인 교정 시간을 측정하는 것이 어렵다. 그러므로 본 논문에서는 상대적인 교정 시간을 측정한다. 즉, 철자 교정기의 교정 시간을 오류 어절에 변형을 가하여 철자 검사기를 통해 검증하는 과정에서의 소요 시간이라 판단하고 오류 어절 당 평균 철자 검사 횟수를 측정 단위로 한다.

한편 철자 교정률은 올바른 후보 어절이 후보 어절 중에 존재하는 지의 여부 뿐만 아니라 불필요한 후보 어절의 제시에 의한 모호성 증가가 고려되어야 한다. 그러므로 후보 어절의 개수에 반비례하는 교정률 측정 함수를 이용한다[3].



<그림 4> 교정 기법별 교정률 및 교정 시간 변화

미등록어 및 사전 정보의 잘못, 여러 번 틀린 어절에 대한 교정의 시도는 무의미하다. 이러한 오류를 교정하고자 하는 과정에서 불필요하게 대치어를 생성한 경우 및 대치어를 전혀 생성하지 못한 경우 또한 교정률 산출시에 고려되어야 한다. 그러나 이러한 경우에 대한 적당한 교정률 산출 함수를 구하는 것이 어려우므로 본 논문에서는 교정률 산출시 고려 대상에서 제외하였다.

<그림 4>는 각 교정 기법을 추가 적용해 나감에 따른 교정률과 교정 시간의 변화를 그래프로 표현한 것이다. 교정률 향상폭은 형태소 연관 관계를 이용한 띄어 쓰기 처리에서 크게 나타난다. 교정률이 향상됨에 따라 음소 대치 전단계에서 교정되는 경우가 많아지므로 철자 검사 횟수의 감소폭도 이 구간에서 가장 크게 나타난다.

V.결론

본 논문에서는 형태소 간의 연관 관계와 heuristics를 이용하여 오류 유형을 교정 전단계에서 추정하는 오류 유형 추정 함수에 의한 교정 기법을 제시하였다. 기존의 한 어절에 국한된 철자 검사 및 교정 영역을 여러 어절로 확장하였으며 오류 유형 추정 함수를 이용한 동적 교정 기법을 사용하였다.

실험 결과 오류 유형 추정 함수에 의한 교정 기법을 사용함으로써 기존의 교정 기법에 비해 평균 철자 검사 횟수는 54.7회에서 31.7회로 감소되었으며, 교정률은 약 10% 가량 향상되었다. heuristics 지식 베이스 및 여러 교정 지식 베이스의 보강으로 교정률은 더욱 향상될 수 있을 것이다. 특히 본 논문에서 구현한 오류 유형 추정 함수를 이용한 철자 검사/교정기는 한 어절 처리에 국한되었던 검사/교정의 범위를 확장하려는 시발이다. 따라서 구문 검사 및 의미 검사 시스템의 원형으로서 그 의의를 가진다.

참고 문헌

[1] Chul-Min Sim, Min-Jung Kim, Hyuk-Chul Kwon, "Automatic Revision of Korean Texts by Collocation Words", Proc. of the '94 International Conference on Computer Processing of Oriental Languages, pp.280-284, 1994

[2] Hyck-Chul Kwon, Aesun Yoon, "Unification-Based Dependency Parsing of Governor-Final Languages", Proc. of Second International Workshop on Parsing Techonology, Cancun, Maxcio, pp. 182-192, 1991

[3] 이영식, "사전 근사탐색과 Heuristics를 이용한 한국어 철자 오류 교정 시스템 구현", 부산대학교 전자계산학과 석사학위 논문, 1994

[4] 이병훈, 윤준태, 송만석, "말뭉치를 기반으로 한 한국어 철자 교정기의 구현", 한글 및 한국어 정보처리 학술발표논문집, pp.285-293, 1993

[5] 김병희, 임권묵, 송만석, "형태소 접속 특성과 인접 말마디 정보를 이용한 형태소 분석기", 한글 및 한국어 정보 처리 학술발표논문집, pp.395-404, 1993

[6] 정한민, 이근배, 이종혁, "자판 특성을 이용한 Neuro-Fuzzy 한국어 철자 교정기의 구현", 한글 및 한국어 정보 처리 학술발표논문집, pp.317-328, 1993

[7] 이종현, 오상헌, "N-GRAM 한글 사전을 이용한 오인식 단어의 교정 알고리즘", 한글 및 한국어 정보처리 학술발표논문집, pp.271-283, 1993

[8] 강승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울대학교 컴퓨터공학과 박사학위 논문, 1993

[9] 박종만, "철자 검색기에서 틀린 어절의 처리", 국어정보학회 우리말 정보화 잔치 '91 논문집, pp.179-186, 1991

[10] 채영숙, 김재원, 김민정, 권혁철, "한국어 철자 검색을 위한 형태소 분석 기법", '91 우리말 정보화 잔치, 국어정보학회, pp.179-186, 1991

[11] 강재우, "접속 정보를 이용한 한국어 철자 띄어쓰기 검사기의 설계 및 구현", 한국 과학 기술원 전산학과 석사학위 논문, 1990

[12] 강승식, 이호석, 문유진, 김영택, "한국어 문법 검사/교정 시스템의 설계", '90 춘계 논문집, 17권 1호, 한국정보과학회, 1990

[13] 부산일보, "부산일보사 style book", 부산일보사

[14] 중앙일보, "중앙일보사 style book", 중앙일보사