JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 18, No. 1 : 25~31 Jan. 2014

선택 제약 명사의 의미 범주 정보를 이용한 용언의 문맥 의존 오류 검사 및 교정

소길자[1*] · 권혁철[2]

The Detection and Correction of Context Dependent Errors of The Predicate using Noun Classes of Selectional Restrictions

Gil-ja So[1*] · Hyuk-chul Kwon[2]

[1]Department of Cyber Police and Science, Youngsan University, Kyungnam 626-790, Korea
[2]School of Computer Science & Engineering, Pusan National University, Busan 609-735 Korea

요 약

현재 실용화된 국내 문법 검사기는 경험적으로 구축된 오류 결정 규칙을 이용해 주위의 문맥을 보고 문법 오류를 판단하는 문맥 의존 오류를 처리하고 있다. 그러나 기존 문법 검사기의 오류 결정 규칙은 어휘 수준으로 구축되어 있어 검사기의 재현율이 낮다. 따라서 어휘대신 어휘 범주 정보를 사용하여 오류 결정 규칙을 일반화할 필요가 있다. 본 논문에서는 검사단어가 용언일 때 선택 제약 명사의 의미 범주를 국내에서 개발된 어휘의미망 KorLex에서 TCM과 MDL을 이용해 추출하고 추출된 의미 범주를 이용해 용언의 오류 결정 규칙을 일반화하는 방법을 제안한다.

ABSTRACT

Korean grammar checkers typically detect context-dependent errors by employing heuristic rules; these rules are formulated by language experts and consisted of lexical items. Such grammar checkers, unfortunately, show low recall which is detection ratio of errors in the document. In order to resolve this shortcoming, a new error-decision rule-generalization method that utilizes the existing KorLex thesaurus, the Korean version of Princeton WordNet, is proposed. The method extracts noun classes from KorLex and generalizes error-decision rules from them using the Tree Cut Model and information-theory-based MDL (minimum description length).

키워드 : 문법 검사기, 문맥 의존 오류, 선택 제약 명사 클래스, 트리컷 모델

Key word : grammar checker, context dependent error, noun class of selectional restrictions, TCM

접수일자 : 2013. 09. 24 심사완료일자 : 2013. 10. 17 게재확정일자 : 2013. 10. 30

*** Corresponding Author** Gil-Ja So(E-mail:kjso@ysu.ac.kr, Tel:+82-55-380-9530)
Department of Cyber Police & Science, Youngsan University, Kyungnam 626-790, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2014.18.1.25 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

맞춤법 검사에서 “문맥 의존 오류”란 좌우의 단어와의 관계를 살폈을 때 발견할 수 있는 오류를 의미 한다[1]. 영어권에서는 문맥 의존 오류를 찾는 방법으로 좌우 단어의 통계적 정보를 이용하거나[2-3] 구문 분석기를 이용하였으나 국내에서는 오류 패턴을 지식베이스화한 사전을 이용한 문법 검사기가 개발되어 실용화되었다(이후 PNU-Speller라고 한다)[4-5].

PNU-Speller에서는 한국어 문서에서 자주 발생하는 문맥 의존 오류 처리 규칙을 언어 전문가가 경험적으로 구축하는데, 어휘의 어근으로 표현된 ① 검사할 단어, ② 검사단어의 오류 여부를 확인시켜 줄 주변 단어의 정보들로 구성된 오류 결정 규칙, ③ 검사단어가 오류로 판정될 때 제시될 수 있는 후보 단어, ④ 도움말 등으로 구성한다[4].

오류 결정 규칙은 검사단어와 같이 사용될 수 없는 어휘, 품사, 범주화된 사전정보 등으로 구성되는데 문법 검사기의 재현율과 정확도에 영향을 주는 중요한 요인 중 하나다. 오류 결정 규칙에 명시된 정보가 어휘로 구성되면 정확도는 좋아지겠지만 오류 결정 규칙에 명시된 어휘와 검사단어가 문장에서 같이 사용되지 않으면 오류를 검출할 수 없으므로 재현율은 낮다. 따라서 어휘 대신 어휘의 범주 정보를 사용하여 오류 결정 규칙을 일반화할 필요가 있다.

그러나 현재 문법 검사기에서 사용할 수준만큼 용언에 대한 선택 제약 명사의 의미 범주 정보와 그 의미 범주에 해당하는 구체적인 명사들에 대한 정보가 전자화된 자료가 없다. 다만, 최근에 명사, 동사, 형용사에 대해서 의미적 관계를 따져 계층적으로 어휘를 구축하고 이를 전자화한 어휘의미망이 개발 중이므로 어휘의미망의 서브 트리를 의미 범주로 사용할 수 있다. 현재 개발된 어휘의미망 중 KorLex는 프린스턴 대학에서 개발한 WordNet을 대역한 후 한국어 특성에 맞게 개념 및 의미를 다시 구조화한 어휘의미망이다. KorLex는 어휘의 개념을 나타내는 최소 단위로 신셋이라는 동일한 어휘 의미를 가지는 동의어 집합으로 정의하고, 어휘의 관계를 신셋의 계층적 구조로 나타내었다[6].

본 논문에서는 선택 제약 명사의 의미 범주 정보를 KorLex에서 추출하고, 추출된 의미 범주 정보를 이용해 용언의 오류 결정 규칙을 일반화한 후 용언의 문맥 의존 오류를 일반화된 오류 결정 규칙을 사용해 검사하고 교정하는 방법을 제안한다.

본 논문에서 제안한 문법 검사기의 문맥 의존 오류의 검사 및 교정 과정은 다음과 같다. ① 용언의 선택 제약 명사의 의미 범주 정보는 계층적 어휘의미망인 KorLex에서 추출한다. ② 추출된 선택 제약 명사의 의미 범주 정보를 이용해 혼동하기 쉬운 용언 쌍의 각 용언에 대해 오류 결정 규칙을 생성한다. ③ 문서에서 오류 가능성이 있는 용언이 발견될 때 의미 범주 정보로 일반화된 오류 결정 규칙을 이용해 용언의 문맥 의존 오류를 검사하고 교정한다.

Ⅱ. 선택 제약 명사의 의미 범주 정보를 이용한 문법검사기

문법 검사기는 문서에 있는 단어에 대해 철자 오류, 통사 오류, 의미 오류를 수행하는 시스템이다. 그림 1에서 보이듯이 문서에서 나타난 문장들은 어절로 분리되어 어절 버퍼 관리자에 저장되고, 형태소 분석기에서 각 단어의 형태소를 분석한다. 형태소 분석이 실패하면 한 어절 오류에 대한 교정을 시도하고, 형태소 분석에 오류가 없으면 다수의 어절을 이용해 오류를 검사하는 문맥 의존 오류의 검사 모듈이 구동된다. 문맥 의존 오류를 검사하는 모듈은 검사단어와 주위 단어와의 공기(collocation) 가능성을 오류 결정 규칙을 이용해 판단하는 모듈이다. 검사단어가 용언일 때 오류 결정 규칙은 검사단어인 용언과 같이 사용할 수 없는 의미 범주를 KorLex의 신셋 정보와 격정보로 표현된다.

2.1. KorLex를 이용한 선택 제약 명사의 의미 범주 정보 추출

선택 제약 명사의 의미 범주 정보를 추출하려면 먼저 부분 문장 분석기를 이용해 용언의 중요한 문장 성분인 주격, 목적격, 부사격 등으로 사용된 명사를 추출한다. 이 때 추출될 명사의 기준인 격 정보는 용언이 가질 수 있는 여러 개의 논항 중 혼동하기 쉬운 다른 용언, 즉, 오류라고 판단될 때 대치될 용언과의 의미를 더욱 잘 구별할 수 있는 논항이 가지는 격정보가 사용된다.

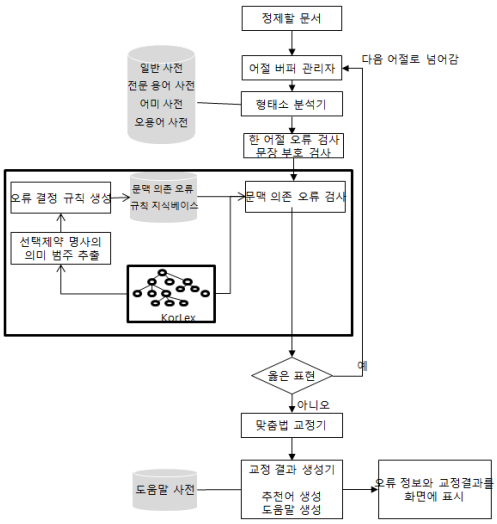

그림 1. 제안한 방법의 문법 검사기 구조도
Fig. 1 The structure of grammar checker

추출된 명사들의 의미 범주 정보는 계층적 명사 어휘 의미망 KorLex에서 TCM(tree cut model)과 MDL (minimum description length)을 이용해 추출한다. TCM과 MDL은 Abe와 Li가 워드넷에서 용언의 격틀 정보를 추출하기 위해 사용했던 방법이다[7]. TCM은 트리 안에서 단말 노드를 분할할 기준이 되는 노드의 집합인 여러 개의 tree cut과 실 데이터에서의 출현확률의 벡터로 모델을 표현한 방법이다. 예로 “먹다”의 선택 제약 명사가 “사과” 4번, “자두” 2번, “복숭아” 4번, “애호박” 2번, “오이” 4번, “양파” 4번 나타났고 이를 KorLex에서 해당하는 신셋을 찾았을 때 그림 2와 같은 그래프를 구성하면 5개의 컷 모델이 만들어 진다.

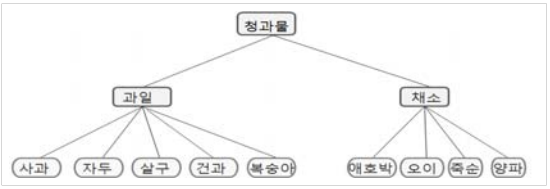

그림 2. [청과물]의 하위 노드
Fig. 2 The child noes of [chung-gwa-mul]

TM(1) = ([청과물], [1.0])
TM(2) = ([과일, 채소], [0.5, 0.5])
TM(3) = ([과일, 애호박, 오이, 죽순, 양파], [0.5, 0.1, 0.2, 0.0, 0.2])
TM(4) = ([사과, 자두, 살구, 건과, 복숭아, 채소], [0.2, 0.1, 0.0, 0.0, 0.2, 0.5])
TM(5) = ([사과, 자두, 살구, 건과, 복숭아, 애호박, 오이, 죽순, 양파], [0.2, 0.1, 0.0, 0.0, 0.2, 0.1, 0.2, 0.0, 0.2])

이 5개의 모델 중 가장 적합한 모델을 선택하는 방법은 MDL을 사용한다. MDL은 데이터 압축에 사용되는 이론으로 데이터를 일반화한 모델을 압축할 때 사용할 모델 정보량을 수식 1로 구하고 모델을 통해 실 데이터를 압축할 때 사용할 데이터 정보량을 수식 2로 구해서 두 정보량의 합인 모델 총정보량이 가장 작은 모델을 선택하는 방법이다.

$$\text{모델 정보량} = \frac{k}{2} log|S| \quad (1)$$

$$\text{데이터 정보량} = \left(-\sum_{n \in S} \log P(n|v,r) \right) \quad (2)$$

수식 1에서 k는 모델에 포함된 명사 클래스의 개수이다. S는 실험에 나타난 명사 리스트이다. $|S|$는 S의 크기, 즉 실험에 나타난 명사 출현 회수 합을 의미한다. 수식 2에서 $P(n|v,r)$은 MLE(maximum likelihood estimation) 방법으로 계산된 출현 확률이다. 이때 클래스 C에 속하는 명사 n의 확률 $P(n|v,r)$은 n이 속한 명사 클래스의 크기 $|C|$로 나눈다.

표 1은 수식 1과 수식 2를 이용해 5개의 컷모델의 정보량을 계산한 결과다. [청과물]의 총정보량이 5개의 컷 모델 중에서 가장 낮다. 이는 하위 노드들의 특성을 대표할 수 있는 의미 범주 정보로 상위 노드인 [청과물]을 사용할 수 있음을 의미한다. 이는 부분적인 모델에서도 같은 결과를 보인다. 예로 [사과,자두,살구,건과,복숭아]의 모델 총정보량이 33.8631이고 상위 노드로 구성된 모델 [과일]의 모델 총정보량은 33.2192가 되어 [과일]이 [사과, 자두, 살구, 복숭아]를 대표할 수 있는 노드로 선택된다. 이 사실을 이용하여 선택 제약 명사의 의미 범주 정보의 추출 과정은 KorLex의 모든 서브 트리에 대해 모델을 만들지 않고, 하위 노드들로 구성된 모델과 이 하위 노드들의 상위 노드로 구성된 모델에 대해서만 정보량을 계산한 후 계층적 구조를 따라 하위 노드들로 구성된 모델의 정보량이 상위노드로 구성된 모

델의 정보량보다 작을 동안 반복적으로 수행한다.

표 1. 5개 컷모델의 정보량 계산 결과
Table. 1 The amounts of information of 5 cut models

모델	모델 정보량	데이터 정보량	총 정보량
TM(1)	0.0000	63.3985	63.3985
TM(2)	2.1609	63.2193	65.3802
TM(3)	8.6439	58.4386	67.0825
TM(4)	10.8048	55.2193	66.0241
TM(5)	17.2878	50.4386	67.7264

2.2. 선택 제약 명사의 의미 범주 정보를 이용한 문맥 의존 오류 처리 시스템

문맥 의존 오류 처리 시스템은 추출된 명사의 의미 범주 정보를 오류 결정 규칙으로 만드는 오류 결정 규칙 일반화 모듈과 이 모듈에서 일반화된 오류 결정 규칙으로 구축된 지식베이스를 이용해 문장에서 오류를 검사하고 교정하는 문맥 의존 오류 검사 모듈로 구성된다.

그림 3은 혼동하기 쉬운 두 용언 쌍의 오류 결정 규칙을 일반화하는 과정이다. 명사의 의미 범주 정보를 추출하는 모듈은 앞 절에서 설명한 대로 두 용언의 선택 제약 명사의 의미 범주를 이용해 KorLex에서 추출한다. 각 용언의 오류 결정 규칙은 검사단어인 용언의 선택 제약으로 사용할 수는 없으나 교정 단계에서 대치단어로 제시될 용언의 선택 제약은 될 수 있는 의미 범주가 격 정보와 함께 표기된다.

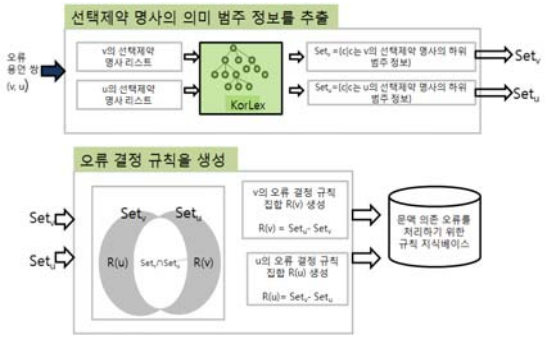

그림 3. 혼동하기 쉬운 용언 쌍의 오류 결정 규칙을 일반화하는 과정
Fig. 3 The process of generalization of error decision rules for usually confused predicate sets

예로 동사 "맡다"와"맞다"선택 제약 명사의 의미 범주 중 두 용언과 같이 쓰일 수 있는 의미 범주를 제외하고 "맡다"의 오류 결정 규칙은 "맞다"의 선택 제약 명사의 의미 범주로 만들어진다.

그림 4는 본 논문에서 제안하는 문맥 의존 오류 검사 모듈의 구성도이다. 규칙 검색부에서는 용언의 오류 검사에 적용할 규칙을 문맥 의존 오류 규칙 지식베이스에서 검색한다. 검색된 규칙에는 용언의 오류를 결정할 명사 논항이 가지는 조사 제약 조건이 명시되어 있는데, 이 조건에 맞는 명사 논항을 부분 문장 분석을 통해 입력된 문장에서 검출한다. 오류 결정 규칙 처리부는 검출된 논항 명사가 속하는 의미 범주를 KorLex에서 추출하고 이 명사의 의미 범주가 검사 중인 용언의 오류 결정 규칙에 있는지 검색한다. 만약 오류 결정 규칙에 있으면 맞는 대치어를 제시한다.

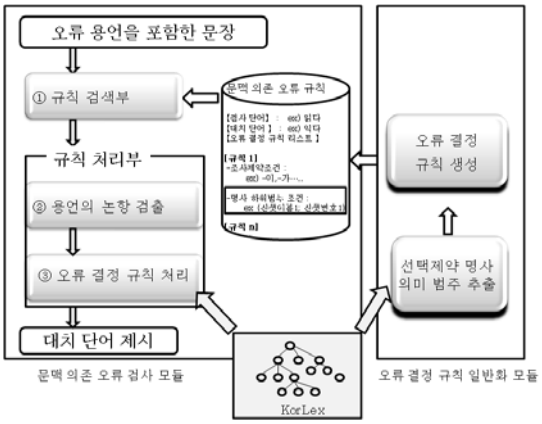

그림 4. 선택 제약 명사의 의미 범주 정보를 이용한 문맥 의존 오류 처리 모듈
Fig. 4 The context dependent error checkers using selectional restrictions

Ⅲ. 실 험

실험은 용언의 선택 제약 명사의 의미 범주 정보를 이용해 오류 결정 규칙을 일반화하고, 일반화된 오류 결정 규칙을 이용해 수행된 용언의 문맥 의존 오류의 검사 성능을 평가한 후, 같은 실험 환경에서 수행된 기존 문법 검사기의 성능과 비교한다. 표 2는 각 오류 유형에 따라 실험에 사용될 용언 쌍이다. 혼동하기 쉬운

용언 쌍은 문맥 의존 오류 규칙에서 각각 검사단어거나 대치단어가 되는데 표 14에서 검사 및 교정 방향을 나타내는 화살표의 시작은 검사단어, 끝은 대치단어를 의미한다. 용언_1과 용언_2는 검사 방향에 따라 때로는 검사단어, 때로는 대치단어가 될 수 있다.

예로 “부치다”가 검사단어일 때 “붙이다”는 대치단어이고, “붙이다”가 검사단어일 때 “부치다”는 대치단어이다.

표 2. 혼동하기 쉬운 용언 쌍
Table. 2 Confusion sets of predicate

오류 유형	용언_1	검사 및 교정 방향	용언_2
의미 오류	부치다	⇔	붙이다
	쌓이다	⇔	싸이다
	들어내다	⇔	드러내다
	마치다	⇔	맞히다
	맡다	⇔	맞다
	늘이다	⇔	늘리다

각 용언의 선택 제약 명사는 세종 말뭉치에서 추출되었다. 추출된 명사들의 의미 범주 정보가 될 신셋 정보를 KorLex에서 추출하고, 이를 이용해 오류 결정 규칙을 생성하였다.

선택 제약 명사의 의미 범주 정보로 구축된 오류 결정 규칙의 검사 및 교정 성능을 평가하기 위해 한겨레신문 2천만 어절을 대상으로 용언의 문맥 의존 오류 검사를 수행하고, 같은 신문에서 기존 문법 검사기로 검사한 결과와 비교하여 제안한 방법의 성능이 우수함을 보인다.

생성된 오류 결정 규칙을 이용해 문법 검사기 성능을 평가하는 실험은 규칙을 생성할 때 사용한 문서와는 다른 문서에서 정확도와 재현율을 측정한다. 그리고 같은 실험 환경에서 기존 문법 검사기의 정확도와 재현율을 측정하고, 제안한 방법과 비교하여 성능이 향상되었음을 보인다.

많은 오류를 찾으면 그중에는 맞는 단어를 틀렸다고 보는 오류(False Alarm)가 많이 포함될 수 있다. 그러므로 문법 검사기를 정확도와 재현율 두 가지 기준을 모두 고려하여 평가하려고 F-measure를 사용한다.

$$F-measure = \frac{2 \times 정확도 \times 재현율}{정확도 + 재현율}$$

3.1. 실험 대상 문서

현재 실험에 사용할 수 있는 대용량 말뭉치들은 신문 기사나 출판된 서적, 교과서 내용을 포함하고 있다. 이런 문서들은 전문가에 의해 작성되거나 교정된 상태이므로 오류 문장이 매우 적어 제안한 방법을 테스트하기에 부적합하다. 이런 이유로 다수의 연구에서 인위적으로 오류 문장을 생성하고 이를 이용해 제안한 방법을 평가하는 실험을 수행하였다[8-10]. 따라서 본 논문에서도 문장에 나타난 용언을 그 용언과 혼동하기 쉬운 다른 용언으로 교체한 문장을 생성하여 제안한 방법의 검사기 성능을 평가한다. 검사 대상은 한겨레신문 2천만 어절을 대상으로 수행하였다. 검사 방법은 검사단어인 용언별로 오류 문장으로 구성된 검사 적합 문서를, 오류가 없는 문장들로 구성된 비 적합 문서를 각각 만든다.

먼저 한겨레신문에서 검사단어가 용언으로 사용되고 선택 제약 조건에 맞는 논항을 포함한 문장을 추출한다. 이는 오류로 검출해서는 안 되는 문장이므로 비 적합 문서로 정한다. 검사단어와 쌍을 이루는 후보 용언을 포함한 문장에서 후보 용언을 모두 검사단어로 바꾸어 실험용 적합 문서를 생성한다. 그리고 실험은 이 적합 문서와 비 적합 문서를 섞어 만들어진 테스트 문서를 대상으로 정확도와 재현율을 측정하였다.

3.2. 기존 문법 검사기와 성능 비교

본 논문에서 제안한 방법의 효율성을 평가하기 위해 기존 문법 검사기를 같은 실험환경에서 정확도와 재현율을 측정하여 비교하였다. 기존 문법 검사기는 오류 결정 규칙이 언어전문가에 의해 수작업으로 만들어지므로 정확도는 100%이다. 그러나 규칙이 어휘로 만들어져 있어 같은 단어가 나타나지 않으면 오류 검출이 되지 않으므로 재현율이 낮게 나타난다. 또한, 검사단어별 재현율 차이가 커서 일관성 있는 성능을 기대하기 어렵다. 이에 비해 선택 제약 명사의 의미 범주 정보를 오류 결정 규칙으로 사용한 시스템은 범주 안에 의미가 유사한 단어들이 포함되므로 재현율이 높게 나타난다. 그러나 의미 범주 정보를 사용하면 맞는 문장을 틀렸다고 보는 잘못된 검출이 늘어나 정확도가 떨어질 수 있다.

표 3은 용언 5쌍에 대해서 기존 문법 검사기와 본 논문에서 제안한 방법의 문법 검사기로 수행한 검사 및 교정 결과이다. 표 3에서 P는 정확도, R은 재현율, F는 F-measure값이다.

표 3. 기존 문법 검사기와의 성능 비교 결과
Table. 3 Comparing the previous and proposed grammar checker

용언 쌍	기존 문법 검사기			제안한 방법		
	P	R	F	P	R	F
부치다/붙이다	1.00	0.47	0.66	0.81	0.55	0.65
들어내다/드러내다	1.00	0.04	0.08	0.67	0.31	0.42
마치다/맞히다	1.00	0.08	0.16	0.68	0.47	0.56
맡다/맞다	1.00	0.11	0.19	0.57	0.40	0.47
늘이다/늘리다	1.00	0.49	0.66	0.91	0.49	0.64

정해진 단어에 대해서만 검사를 수행하는 기존 문법 검사기에 비해 본 논문에서 제안한 방법의 검사기의 정확도는 낮게 나왔다. 정확도가 낮은 이유는 의미 범주를 추출하는 과정에서 의미 중의성 때문에 용언과 관련 없는 의미 범주가 추출될 수 있고, 이를 오류 결정 규칙으로 사용하여 오류를 검출하면 맞는 문장을 잘못된 문장으로 판단하기 때문이다. 그러나 기존 문법 검사기의 재현율은 제안한 방법이 전체적으로 높거나 같게 나타난다. 따라서 정확도와 재현율 모두를 고려한 F-measure로 문법 검사기의 성능을 평가하면 표 3에서 보이듯이 제안한 방법 모두 기존 문법 검사기보다 2배 이상 성능이 높게 나타났다.

Ⅳ. 결 론

문맥 의존 오류 처리 규칙에서 검사단어의 오류 여부를 확인시켜 줄 주변 단어의 정보인 오류 결정 규칙은 검사기의 성능에 영향을 주는 중요한 요인이다. 기존 문법 검사기인 PNU-Speller의 오류 결정 규칙은 품사나 10여 가지 분류 정보를 사용하고 있으나 대다수가 어휘로 구축되어 있어 정확도는 높지만 재현율이 낮은 문제점이 있다.

본 논문에서는 이를 해결하고자 국내에서 구축된 어휘의미망 Korlex를 이용해 선택 제약 명사의 의미 범주 정보를 추출하고, 추출된 의미 범주 정보를 이용해 용언의 오류 결정 규칙을 일반화한 후 일반화된 오류 결정 규칙으로 용언의 문맥 의존 오류의 검사와 교정을 하는 방법을 제안하였다.

제안된 방법에서는 첫째, 용언의 선택 제약 명사의 의미 범주 정보를 계층적 어휘의미망인 KorLex에서 추출하였다. 둘째, 추출된 선택 제약 명사의 의미 범주 정보를 이용해 혼동하기 쉬운 용언 쌍의 각 용언에 대해 오류 결정 규칙을 생성하였다. 셋째, 문서에서 오류 가능성이 있는 용언이 발견될 때 의미 범주 정보로 일반화된 오류 결정 규칙을 이용해 용언의 문맥 의존 오류를 검사하고 교정하였다.

제안한 선택 제약 명사의 의미 범주 정보를 오류 결정 규칙으로 한 문법 검사기의 성능을 평가하기 위하여 기존 문법 검사기와 성능 비교실험을 수행하였다. 먼저, 5개의 용언 쌍에 대해 선택 제약 명사의 의미 범주 정보를 추출한 후, 제안한 방법으로 생성된 오류 결정 규칙을 사용하는 문법 검사기와 기존 문법 검사기의 정확도, 재현율, 그리고 F-measure 값을 비교 분석 하였다.

실험은 한겨레신문 2천만 어절을 대상으로 수행하였다. 실험 결과 기존 문법 검사기는 정확도는 높지만 규칙이 어휘로 만들어져 있어 같은 단어가 나타나지 않으면 오류 검출이 되지 않으므로 재현율이 낮게 나타났다. 이에 비해 제안된 방법은 한 범주 안에 의미가 유사한 단어들이 포함되므로 재현율이 높게 나타났다. 그러나 의미 범주 정보를 사용하면 맞는 문장을 틀렸다고 보는 잘못된 검출이 늘어나 정확도가 떨어진다. 그러나 정확도와 재현율 모두를 고려하면 본 논문에서 제안한 선택 제약 명사의 의미 범주 정보를 사용한 문법 검사기가 기존 문법 검사기보다 2배 이상 성능이 높게 나타났다.

제안된 방법에서 의미 중의성이 있는 단어 또는 부문 문장 분석에서 추출된 선택 제약 명사에 오류가 있을 경우 오류 결정 규칙으로 용언과 관련 없는 의미 범주 정보가 추출이 되어 문법검사기의 정확도가 낮아진다. 향후 정확도 향상을 위하여 의미 중의성 제거 시스템을 전처리 단계에 활용하거나 선택 제약 명사를 검증하는 추가의 과정에 대한 연구가 필요하다.

감사의 글

이 논문은 2013년도 영산대학교 교내 연구비 지원에 의하여 연구되었음.

REFERENCES

[1] K. Kukich, “Techniques for automatically correcting words in text,” *ACM Computing Surveys*, vol. 24, no. 4, pp. 377-439, 1992.

[2] A. R. Golding, and D. Roth, “A winnow-based approach to context-sensitive spelling correction,” *Machine Learning*, vol. 34, no. 1-3, pp. 107-130, 1999.

[3] A. Carlson, and I. Fette, “Memory-based context-sensitive spelling correction at web scale,” *in Proceeding of The 6th International Conference on Machine Learning and Applications*, pp. 166-171, 2007.

[4] M. Y. Kang, A. S. Yoon, H. C. Kwon, "Improving Partial Parsing Based on Error-Pattern Analysis for Korean Grammar-Checker", *TALIP ACM*, vol. 2, no. 4,pp. 301-323, 2003.

[5] J. L. Kong, S. Y. Hwang, “A Korean Grammar Checker based on the Trees Resulted from a Full Parser,” *Journal of KIISE : Software and Applications*, vol. 30, no. 10, pp. 992-999, 2003.

[6] A. S. Yoon, S. H. Hwang, E. R. Lee, H. C. Kwon, “Construction of Korean Wordnet 「KorLex 1.5」,” *Journal of KIISE : Software and Applications*, vol. 36, no. 1, pp. 92-108, 2009.

[7] H. Li, and N. Abe, “Generalizing case frames using a thesaurus and the MDL principle,” *Computational Linguistics*, vol. 24 no. 2, pp. 217-244, 1998.

[8] G. Hirst, and D. S. Onge, "Lexical chains as representations of context for the detection and correction of malapropisms,” *WordNet*, The MIT Press, pp. 305-332, 1995.

[9] G. Hirst, and A. Budanitsky, “Correcting real-word spelling errors by restoring lexical cohesion,” *Natural Language Engineering*, vol. 11, no. 1, pp. 87-111, 2005.

[10] A. Islam, and D. Inkpen, “Real-word spelling correction using Google web lT 3-grams,” *in Proceeding of The 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1241-1249, 2009.

소길자(Gil-Ja So)

2012년 부산대학교 컴퓨터공학과 박사
2002~현재 영산대학교 사이버경찰학과 조교수
※관심분야 : 자연언어처리, 인공지능, 정보보안

권혁철(Hyuk-Chul Kwon)

1982년 서울대학교 컴퓨터공학과 학사
1984년 서울대학교 컴퓨터공학과 박사
1987년 서울대학교 컴퓨터공학과 박사
1987년~현재 부산대학교 정보컴퓨터공학부, 인지과학협동과정 교수
※관심분야 : 인간언어공학, 정보검색, 인공지능