Journal of The Korea Society of Computer and Information
Vol. 18, No. 10, October 2013
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2013.18.10.173

철자오류에 기인한 가의미 오류의 검출 및 교정 방법

김 동 주*

A Method for Detection and Correction of Pseudo-Semantic Errors Due to Typographical Errors

Dong-Joo Kim*

요 약

전자 문서의 초안 작성과정에서 추가되는 철자오류는 다른 유형의 오류보다 압도적으로 높은 비율을 차지한다. 입력 실수로 인한 이들 오류는 결과적으로 여전히 철자오류일 수도 있지만 상당수는 구문오류나 의미오류로 발전한다. 이러한 오류들 중 철자오류에서 발전된 가의미 오류는 순수 의미오류에 비해 문장 내에서 주변 단어의 의미에 대해 두드러진 상이성을 갖게 된다. 따라서 이러한 의미 오류는 그것이 가지는 두드러진 문맥 상이성으로 인해 간단한 동시발생 빈도에 기초한 알고리즘으로 검출 및 교정이 가능하다. 본 논문에서는 이러한 오류들을 검출하고 교정하기 위한 동시발생 빈도에 기초한 알고리즘을 제안한다. 제안하는 방법에서 동시발생 빈도는 의존 구조상에서 직접 의존관계에 놓인 단어만을 대상으로 계산하며, 가의미 오류 여부를 판단하기 위해서 코사인 유사도 측정 방법을 사용한다. 제시하는 실험으로부터 제안한 방법은 전체 맞춤법 검사기 검출율을 약 2~3% 수준까지 향상 시킬 수 있을 것으로 예측하였다.

▸ Keywords : 의미오류, 가의미 오류, 철자오류, 맞춤법검사

Abstract

Typographical mistakes made in the writing process of drafts of electronic documents are more common than any other type of errors. The majority of these errors caused by mistyping are regarded as consequently still typo-errors, but a considerable number of them are developed into the grammatical errors and the semantic errors. Pseudo semantic errors among these errors due to typographical errors have more noticeable peculiarities than pure semantic errors between senses of surrounding context words within a sentence. These semantic errors can be detected and corrected by simple algorithm based on the co-occurrence frequency because of their prominent

•제1저자 : 김동주
•투고일 : 2013. 10. 3 심사일 : 2013. 10. 16. 게재확정일 : 2013. 10. 21.
* 안양대학교 교양대학(College of Liberal Arts, Anyang University)

contextual discrepancy. I propose a method for detection and correction based on the co-occurrence frequency in order to detect semantic errors due to typo-errors. The co-occurrence frequency in proposed method is counted for only words with immediate dependency relation, and the cosine similarity measure is used in order to detect pseudo semantic errors. From the presented experimental results, the proposed method is expected to help improve the detecting rate of overall proofreading system by about 2~3%.

▸ Keywords : Semantic Error, Pseudo Semantic Error, Typographical Error, Proofreading System

I. 서 론

문서 편집기로 작성된 문서는 많은 오류들을 포함하고 있으며, 오류 원인의 대다수는 작성하려는 문서의 성격에 기인한다. 일반적으로 문서 편집 작업은 두 가지의 경우로 볼 수 있는데, 그 하나는 이미 문서화되어 있던 원문을 입력하는 경우가 있겠고, 또 하나는 새로운 내용의 문서를 작성하는 경우이다. 문서화된 원문을 옮겨 쓸 때 포함되는 오류에는 원문에 이미 포함되어 있던 원문오류 외에 옮겨 쓰는 과정에서 새롭게 추가되는 철자 오류가 포함된다. 원문오류는 원문이 출판되기 전에 교열(校閱) 능력을 지닌 전문가들 손에 의해 이미 여러 번의 교열 작업을 마친 후임에도 불구하고 포함되어 있는 오류들이다. 이러한 오류들은 전문가가 살펴보더라도 오류인지를 판단하기 힘든 경우가 많다. 따라서 맞춤법 비전문가인 일반 편집자가 원문을 옮겨 적을 때 원문 자체가 가지고 있는 오류는 새로 작성하는 문서 내에 거의 그대로 반영된다. 한편, 새로운 내용의 문서를 작성하는 경우에는 원문을 옮겨 적을 때보다 더 많은 오류들이 포함될 수가 있다. 교열 전문가가 작성하는 문서든 일반 편집자가 작성하는 문서든 작성 초기의 문서는 무수한 오류들을 포함하고 있으며, 이러한 내용들에 대하여 오류의 교열 작업을 반복적으로 수행해야 한다. 그런데 교열 작업은 문서의 양이 방대할 경우 여간 힘든 일이 아닐 수도 없으며, 또한 무수한 맞춤법 지식이나 표준어 규정을 정확히 알고 있기도 불가능한 실정이다. 이러한 일들을 손쉽게 하기 위해 교열 작업도 컴퓨터에 의존할 수가 있겠고, 현재 전문 문서 편집기뿐만 아니라 간단한 텍스트 편집기에서부터 다양한 용도의 많은 저작도구들에 맞춤법 검사 기능이 내장되어 있다.

맞춤법 검사기는 사용자의 실수 또는 잘못된 지식으로 인한 철자 및 구문 오류, 의미 오류, 문체 오류 등을 컴퓨터를 이용하여 자동으로 검사하고 교정(校訂)하는 시스템이다. 맞춤법 검사기는 오류를 검출해 내는 부분과 검출된 오류를 교정해 주는 부분으로 나눠진다. 검출과 교정은 사용되는 알고리즘이나 구축된 사전(Machine Readable Dictionary) 및 지식-베이스(Knowledge-Base)에 따라서 구분이 불분명한 경우도 있다.

맞춤법 오류의 유형을 좀 더 구체적으로 살펴보면 단순한 입력 실수로 발생하는 철자 오류, 입력 실수나 맞춤법 지식의 부족으로 발생하는 띄어쓰기 오류, 한국어 구문 지식의 부족으로 발생하는 구문 오류, 단어의 의미를 몰라서 틀리는 의미 오류, 철자법과 구문 및 의미적으로 오류가 없으나 적절하지 못한 낱말을 사용한다든가 표현 방법의 미숙, 정서에 맞지 않는 표현을 사용하는 문체 오류로 나눠진다[1]. 이외에도 비순화 용어를 사용한다든가 발음의 유사성으로 인해 틀린 단어나 사투리를 사용하는 맞춤법 규정상 오류인 맞춤법 오류도 있다[2]. 과거 연구에서 오류의 유형을 분류할 때, 발생 원인과는 관계없이 오류들의 검출을 위해 적용되는 언어 처리 기술 수준이 어디에 있느냐에 따라 조사하는 것이 일반적이었다[3~7]. 즉, 발생한 오류가 어절 내부의 구조를 분석하는 형태소 분석으로 검출 가능하다면 철자 오류로 분류했고, 어절 내의 구조 분석만으로는 오류인지를 판단할 수 없어 주변 어절을 같이 살펴야 하는 구문 분석이 요구된다면 구문 오류로 분류했다. 그리고 형태소 분석, 구문 분석 어느 것으로도 오류의 검출이 불가능하다면 의미 분석이 필요한 의미 오류로 간주하였다.

본 논문에서는 기존의 연구와는 방법을 달리하여 오류 유형을 분석하였으며, 전문가도 판별하기 힘든 오류의 검출은 배제하고 사용자 실수에 의한 오류에 역점을 두어 단어의 동시발생 가능성[8~10] 정보를 이용하여 일반적인 검사 범위를 뛰어넘어 원인별 발생 비율에서 상당 비중을 차지하는 의미 오류의 검사를 목표로 하고 있다.

II. 오류 유형에 따른 연구 동향

앞 장에서 기술한 바와 같이 오류들은 크게 원문서가 가지고 있는 원문 오류와 새로운 내용을 편집할 때 발생하는 편집상의 오류로 나뉜다. 원문 오류의 경우 가장 오류가 적은 것으로 알려진 교과서를 예를 들면 중고등학교 교과서 과목 평균 182개의 오류를 갖고 있다. 교과서의 경우는 이 정도이지만 소설이나 신문 잡지 등은 3배에서 10배에 가까운 오류를 포함하고 있다. 교과서의 오류에서 유형별 오류 형태를 정리한 것으로 표 1은 원문이 갖는 오류의 비율을 나타낸다[2]. 원문 오류의 특징은 편집자의 실수에 의한 오류는 거의 드물고 편집자 또는 교열자의 맞춤법 지식의 부족으로 인한 오류나 전문가가 봐도 오류로 판정하기 쉽지 않은 오류가 대부분을 차지한다. 즉, 표 1의 분류 중 맞춤법 오류, 띄어쓰기 오류, 외래어 표기 오류의 경우는 편집자나 교열자의 맞춤법 지식의 부족에 그 원인이 있겠고, 외국어투를 사용한다든지 정서상 부적절한 낱말의 사용한 경우와 구문 오류의 경우는 전문가들도 쉽게 오류인지를 구분 할 수 없는 경우이다. 전자의 경우에는 기존의 많은 맞춤법 검사 시스템에서 검출이 가능한 오류들이지만 후자의 경우에는 검출이 불가능한 실정이다.

표 1. 교과서에서 오류 유형별 비율
Table 1. Rates by error types in textbook

오류 유형	비율(%)
띄어쓰기 오류	29.5
어미와 조사의 오용	25.1
적절하지 못한 낱말	13.8
문맥에 호응하지 않는 문장	5.3
맞춤법 오류	4.8
기타	21.5

표 2는 경향신문사, 서울신문사를 통해 얻은 교열 작업을 하기 전의 신문 기사 4개월 치 분량과 통신상에 올려진 소설 등을 토대로 조사한 오류의 유형이다[2, 3]. 표 1에서 조사된 원문의 오류와 표 2에서의 편집 오류의 개수는 1:29의 비율로 편집 오류가 훨씬 더 많다. 원문에서 오류의 형태로 주로 쓰이는 것은 대부분이 편집시에도 오류의 형태로 쓰이게 되고, 또한 편집 오류의 절반 정도는 음소의 첨가, 탈락, 인접 키보드의 오타로 인해 발생하는 키보드 입력 실수로 인한 철자 오류가 대부분을 차지한다. 그 다음으로 비중이 높은 것은 띄어쓰기 오류가 있을 수 있겠는데, 띄어쓰기 오류의 경우는 잘못된 지식으로 인한 경우도 있겠고 맞춤법 자체의 모호성으로 발생하기도 한다.

표 2. 편집 오류의 유형별 비율
Table 2. Rates by error types in editing

오류 유형	비율(%)
철자 오류	47.5
띄어쓰기 오류	26.1
맞춤법 오류	10.7
구문 및 의미 오류	9.6
기타	6.1

초창기의 시스템은 철자 오류의 검사와 교정에 중점을 두었으며 입력 문장이 띄어쓰기 오류가 없는 것으로 가정하였고, 검사는 사전에 절대적으로 의존하였으며 교정은 사용자에게 많이 의존하여 실용성이 떨어지는 문제가 있었다. 일부 시스템은 철자 오류만을 검사하는 형태소 분석기를 통하여 띄어쓰기 오류까지 검사하기도 하였다. 초기 시스템들은 형태론적 관점에서 어절 내부의 오류의 검사와 교정을 하였을 뿐 주변 어절의 구조나 의미를 파악해야 하는 구문/의미 오류는 검사하지 못했으며 심지어는 실제 맞춤법 표준어 규정도 제대로 반영하지 못했었다. 80년대 말에 접어들면서부터 보다 정확한 철자 오류 검사 시스템들이 등장하기 시작하였고, 철자 오류 검사를 위한 다양한 방법이 시도되었으며, 이 때 사실상 형태소 분석기와 지식 베이스를 이용한 철자 오류 검사 시스템은 어느 정도의 한계점에 이르기 시작했다[4, 11].

철자 오류 검사의 한계와 더불어 90년대에 들어와서 문법 검사/교정에 관한 관심이 높아지면서 문법 검사/교정에 관한 시스템이 등장하기 시작하였다[4]. 가장 먼저 서울대학교에서 한국어 문법 검사/교정 방안을 제시하고 문제점들을 제기하였다. 이후 첨단요소기술 과제의 일환으로 한국과학기술원을 중심되어 실용적인 문법 검사 시스템을 개발을 시작하였다. 이렇게 문법 검사에 대한 연구가 시작되면서 철자 오류 검사는 형태소 분석기 성능에 대한 성과와 더불어 더욱 많은 연구가 진행되었으며, 띄어쓰기의 검사/교정은 철자 검사와는 별도로 연구되기도 하였다[6]. 다양한 방법의 철자 검사/교정에 관한 연구와 더불어 문법 및 의미 검사 시스템도 등장하였다[4]. 철자 오류 검사와는 방법론적인 측면에서 매우 상이하기 때문에 문법 오류 검사 시스템의 등장은 철자 오류 검사 시스템의 등장보다 한참 후의 일이었다. 그러나 현재 시도되고 있는 의미 오류의 검출은 실용적인 문법 검사 시스템과 방법론적인 측면에서 동일하기 때문에 문법 검사 시스템이 등장함과 거의 동시에 의미 오류 검출 시스템이 등장하게 되었다. 시도되는 주요 방법은 사람들이 틀리기 쉬운 의미적 오류의 유형들을 지식베이스를 구축함으로써 검출하는 것이다. 어떤

연구에서는 사람들이 틀리기 쉬운 의미 오류들을 수집하여 이를 지식베이스화 함으로써 실제 문서에서 나타나는 의미 오류의 80% 이상을 검출 가능하다고 주장하고 있다[4].

기존의 시스템들은 결과론적인 관점에서 오류가 발생했을 때 그 오류를 검출하기 위해 도입되어야 할 자연언어 처리 기술에 관심을 두었다. 그런데 맞춤법 검사 시스템은 분석 시스템과 달리 견고한(robust) 측면보다는 정확성이 요구되고, 또한 한국어 자연언어 처리 기술 중 구문 분석과 의미 분석에 대한 성과가 아직 뚜렷하게 드러나고 있지 않기 때문에 맞춤법 검사에서 구문과 의미에 관한 자연언어 처리 기술을 쉽게 도입하지 못하는 문제가 발생하였다. 즉, 기존 맞춤법 검사 시스템들의 철자 검사 부분은 자연언어 처리 기술의 기반이 되는 형태소 분석 시스템을 도입하여 사용을 하였으나[3, 7] 구문 및 의미 오류의 검출에 자연언어 처리 기술인 구문 분석과 의미 분석 기술을 도입하기에는 자연언어 처리 기술 자체에 모호성과 과분석이 많아 오류 검출률이 매우 떨어지기 때문에 자연언어 처리 기술인 구문 분석기나 의미 분석기를 사용할 수 없게 되었다. 대안으로 사람들이 자주 틀리는 오류의 유형들을 지식베이스화하여 부분 구문 분석이나 변형된 의존 문법을 통하여 오류의 검출을 시도하였다. 이런 시도들의 단점 중 하나는 지식베이스의 구축에 많은 노력이 든다는 사실이다. 그리고 무엇보다 치명적인 것은 지식베이스에 등록되어 있지 않은 오류들은 검출이 불가능하다는 것이다.

기존 시스템들이 갖는 제한적인 검출율의 근본적인 문제점은 결과론적인 관점에서 오류들을 분류하고 해당 분류에 필요한 자연언어 처리 기술을 사용했기 때문이었다. 이러한 표면적 의미 오류의 원인은 많은 경우 사용자의 키보드 조작 실수로 발생하는 철자 오류에 있다. 또한 이렇게 철자 오류가 원인이 되는 경우는 표면적 철자 오류도 포함되고 의미 오류뿐만 아니라 높은 비율의 구문 오류와 띄어쓰기 오류까지도 포함된다. 표 2의 오류 유형을 표면적 결과와 무관하게 오류 발생 원인에 따라 다시 분류하면 표 3과 같다.

표 3. 원인별 편집 오류의 유형 비율
Table 3. Rates of error types by cause

오류 유형	비율(%)
철자 오류	62.6
띄어쓰기 오류	15.4
맞춤법 오류	10.4
구문 및 의미 오류	5.3
기타	5.3

표 2와 비교했을 때 맞춤법 오류를 제외하고는 대부분의 오류들에 대한 비율이 큰 폭으로 달라졌다. 즉, 맞춤법 오류는 10.7%에서 10.4%로 거의 변화가 없으나 띄어쓰기 오류는 26.1%에서 15.4%로, 구문 및 의미 오류는 9.6%에서 5.3%로 크게 감소한 반면, 철자 오류는 47.5%에서 62.6%로 크게 증가하였다. 표 2에서 띄어쓰기 오류와 구문 및 의미 오류로 분류되었던 오류들이 표 3에서는 철자오류로 분류된 것이다. 다시 말해 편집자의 입력 실수로 발생한 철자오류는 결과론적으로 판단하였을 때 여전히 철자오류로 분류되는 경우가 다수이기는 하지만 상당 비율이 띄어쓰기 오류와 구문 및 의미오류로 발전한다는 의미이다. 이와 같이 기존 연구에서 구문 오류, 의미 오류의 절반에 가까운 비율은 편집자의 입력 실수로 인해 발생했으며, 특히 의미 오류의 절반가량은 순수 의미 오류가 아닌 단순한 편집자의 입력 실수로 발생했다. 따라서 철자 오류의 특징을 분석하면 지식베이스의 구축과 같은 많은 수고를 들이지 않고도 기존 시스템에서 쉽게 해결할 수 없었던 구문 및 의미 오류의 검출이 가능할 것이다.

III. 의미 오류의 검출 및 교정

1. 의미 오류에 관한 기존 연구

전통적인 방법에서의 의미 오류의 검출은 먼저 문장에서 사용된 어휘의 의미가 무엇인지를 밝히는 의미 표지 부착이 필요하다[12, 13]. 의미 표지 부착 방법은 지식베이스에 기반한 전통적인 접근 방법이 일반적이고, 어휘의 표면적 정보를 이용하는 방법도 제시되었다[13, 14]. 그러나 지식베이스에 기반한 전통적인 접근 방법은 시간과 노력 면에서 매우 부담이 크기 때문에 쉽게 적용할 수 있는 방법이 될 수 없으며, 어휘의 표면적이 정보를 이용하는 방법은 짧은 시간에 많은 어휘에 대한 의미 표지 부착이 가능하지만 그 구축 범위가 매우 제한된다.

결과론적 관점에서의 의미오류인 표면적 의미오류에는 일반적으로 사용자가 어휘의 의미를 모르거나 의미를 혼동하여 발생하는 순수 의미오류와 철자 입력 실수가 원인이 되어 발생하는 가의미(假意味) 오류가 있다. 표면적 의미오류들을 원인론적 관점에서 분류를 하면 각각 철자 오류, 구문 오류, 의미 오류로 분류할 수 있다. 표 4는 기존의 시스템에서의 결과론적 관점에서 분류한 표면적 의미오류의 유형들 중 가의미 오류와 순수의미 오류에 대한 예들이다. 유형 A의 오류들은 실제로는 사용자의 타이핑 실수로 인하여 의미 오류로 발전한 가의미 오류이고, 유형 B는 사용자가 어휘에 대한 의미를 모르거나 혼동하여 발생한 순수 의미오류이다. 기존의 시스템에

서 표 4의 예를 모두 의미 오류로 분류한 까닭은 형태소 분석으로 오류를 검출할 수 없고, 구문 분석을 통해서도 오류를 검출할 수 없으며, 의미 분석을 해야만 오류를 검출할 수 있다고 판단했기 때문이다.

표 4. 표면적 의미오류의 원인
Table 4. Causes of superficial semantic errors

예	원인	유형
나의 의견이 옳다고 조장합니다. (주장) 1년 새 물자가 많이 올랐습니다. (물가) 체증이 많이 나갑니다. (체중)	철자	A
그 일로 매우 곤욕스러웠습니다. (곤혹스러웠습니다) 생각이 복잡하게 엉키었습니다. (얽혔습니다)	의미 혼동	B

즉, 유형 A의 오류를 검출하기 위해서 형태소 분석이나 구문 분석으로는 불가능하고 의미 분석이 필요하다는 것이다. 예를 들면 "나의 의견이 옳다고 조장합니다"라는 문장에서 '의견'이라는 단어는 의미적으로 인간의 생각이나 의지를 나타내는 추상명사로 '생각'이나 '의지'라는 의미 분류에 포함을 시킨다. 반면에 '조장하다'라는 동사는 '북돋는다'라는 의미를 지니고 '경향'이나 '분위기'의 의미 분류에 포함되는 목적어를 취하는 타동사이다. 기존의 방법에 의하면 각각의 단어가 이러한 의미 정보를 가지고 있고 문장 내에 나타난 단어의 쌍에 대해 각각 어휘의 의미 정보를 비교함으로써 오류를 판가름한다. 그러나 이러한 방법은 모든 어휘에 대한 의미 분류가 필요하여 구축하는데 많은 비용과 시간이 소모된다는 단점과 오류의 검출을 위해 별도의 수많은 규칙들이 필요하다는 단점을 지니고 있다. 현재 국내의 여러 연구 기관에서 이러한 방법으로 의미 사전의 구축을 시도하고 있지만 그것들을 적용하기에는 아직 많은 문제점들을 내포하고 있다.

어휘의 표면적 정보를 이용한 의미 분류에는 크게 어휘의 표면 패턴을 추출하여 의미를 파악하는 방안[9]과 어휘 자체의 형태로부터 의미 정보를 추출하는 방안[13]이 있을 수 있다. 표면 패턴을 추출하는 방법은 자동화가 불가능하므로 수동으로 직접 대량의 코퍼스로부터 의미와 관련된 패턴을 추출하여야 한다. 따라서 표면 패턴을 이용한 방법은 들이는 시간과 노력에 비해 성과가 뛰어 나지 못한 편이다. 그러나 어휘 자체적인 정보에 의한 방법은 간단한 몇 가지 수작업으로부터 많은 어휘에 대한 의미 정보를 파악할 수 있다.

앞서 설명한 두 가지 방법 중 간단한 수작업으로 많은 량의 어휘의 의미를 찾는 방법으로 대표적인 것이 접사의 의미를 이용하는 것이다[15]. 접사란 독립적으로는 단어가 될 수 없으며 체언, 용언의 앞이나 뒤에 붙어 의미를 더하는 품사를 말한다. 그런데 접사는 임의의 단어에 쉽게 결합할 수 있는 것이 아니라 각 접사마다 특정 품사나 의미를 갖는 단어에만 결합될 수 있다. 예를 들어 '-화(化)'라는 접미사는 상태성 명사의 뒤에 결합할 수 있고 결합된 단어는 상태의 변화를 나타낸다. 또한 '-님'이라는 접미사는 인명이나 대명사 뒤에만 붙을 수 있으며 '-님'이 붙은 단어는 존칭의 의미를 나타낸다. 따라서 코퍼스로부터 이러한 접미사와 함께 사용되는 단어를 추출함으로서 각 단어에 대한 의미 정보를 자동으로 부과할 수 있다.

이 방법은 미리 한국어에서 사용되는 접사에 대한 수집과 더불어 접사에 대한 분석이 필요하다는 것이 단점이며, 무엇보다 치명적인 단점은 접사와의 사용 예를 갖지 않는 단어에 대한 의미 정보를 파악할 수 없다는 것이다. 따라서 매우 제한된 범위에서 맞춤법 시스템에 응용될 수 있을 뿐이다.

2. 제안하는 방법

앞 장에서도 밝혔듯이 오류 원인별 발생 비율로 살폈을 때 사용자가 의미를 모르거나 혼동하여 발생하는 오류보다 다른 오류가 원인이 되어 발생하는 비율이 높은 편이다. 따라서 원인론적인 관점에서의 철자 오류의 특징을 분석하여 이를 모델링하면 결과론적 관점에서 의미 오류라고 분류했던 많은 오류를 검출할 수 있다. 철자 오류의 특징은 오류 단어의 발생성이 매우 희박하며 주변 단어와의 의미적 관계를 살펴보았을 때 명백히 다르다는 것을 알 수 있다. 표 4에서 보는 바와 같이 순수의미 오류는 오류어를 중심으로 주변 단어의 동시발생성이 현격히 차이나지 않는 반면 가의미 오류는 동시발생성이 현격히 차이난다. 즉, 수순 의미오류 단어인 '곤욕'은 동시발생하는 단어인 '일'이나 '매우'와 충분히 동시발생할 수 있는 반면, 가의미 오류 단어인 '조장'은 주변 단어인 '의견', '옳다'와 같은 단어와 동시발생 가능성이 현저히 낮다. 이와 같이 단순 입력 실수에 의해 발생한 가의미 오류는 주변 단어에 대한 동시발생성이 현격히 차이 나므로 단어에 대해 이웃하는 단어와의 동시발생성(co-occurrence)에 기반한 정보를 조사함으로써 오류 단어를 쉽게 검출할 수 있을 것이다[16, 17]. 즉, 두 단어가 특정 단어들과의 동시발생성 사용 사례가 유사하다면 두 단어는 의미적으로 사용례가 동일하다고 판단할 수 있고, 동일한 의미를 갖는 단어들의 사용은 동일한 패턴을 갖는다고 유추할 수 있을 것이다.

그러나 사람들이 사용하는 수십만 단어에 대한 동시발생성을 조사하는 것은 현실성이 떨어진다고 볼 수가 있다. 따라서 문장의 의미를 결정짓는 품사들의 쌍에 대해서만 동시 발생 빈도를 계산한다. 문장 내에서 의미를 결정짓는 단어는 내용어에 속한 단어들로 명사, 수사와 같은 체언류의 단어, 동사,

형용사와 같은 용언류의 단어와 부사와 관형사 일부이다. 대명사는 실질적으로 문장의 의미를 결정짓는 중요한 역할을 하며 대명사의 실제 대상이 무엇인지 결정되어야 하지만, 일부 대명사를 제외하고 대부분 형태의 대명사는 어떤 또 다른 품사의 단어들과도 동시에 사용될 수 있으므로 동시발생 패턴을 파악하기가 어려우므로 대명사는 동시 발생 빈도 추출에서 제외된다. 또한 대부분의 부사어나 관형사는 체언이나 용언의 뜻을 구체화시키거나 부가하는 역할을 하고 있어 문장의 의미를 결정하는 주체가 되지 못할 뿐만 아니라 많은 체언류의 단어와 용언류의 단어와 흔히 동시발생 한다. 그런데 부정 부사는 문장의 의미를 완전히 반대로 바꾸는 중요한 역할을 수행하므로 문장의 의미를 결정짓는 절대적인 품사라고 할 수가 있다. 그러나 몇몇 특수한 경우를 제외하고 부정 부사어는 어느 용언과도 쉽게 결합할 수 있으므로 동시발생 빈도만으로는 사용 예를 패턴화하기 쉽지 않다. 따라서 부사어와 관형사도 기본적으로 동시발생 빈도의 추출 대상에서 제외한다. 결과적으로 문장에서 동시발생 예로 추출 대상이 되는 품사는 명사, 수사, 동사, 형용사와 이들로부터 굴절되거나 파생된 일부 관형사와 부사로 한정한다.

(1) a. 아주머니가 사과를 팔았다.
b. 아저씨가 밤을 팔았다.
c. 내 동생이 사과를 먹었다.

간단한 예로 예문 (1)과 같은 세 개의 문장에 대한 동시발생 빈도는 다음과 같이 계산된다. '아주머니'라는 단어 t_1은 '사과'라는 단어 t_2와 '팔다'라는 단어 t_3와 한 문장 내에서 동시발생 되고, t_1과 t_2가 동시발생 하는 빈도 f_{12}와 t_1과 t_3가 동시발생 하는 빈도 f_{13}은 각각 1이다. 또한 t_3는 문장 (1-a)에서 t_1과 t_2뿐만 아니라 문장 (1-b)에서 t_4 '아저씨'와 t_5 '밤'이라는 단어와 동시 발생하며 각각의 빈도 f_{34}, f_{35}는 1이다.

표 5. 동시발생 빈도의 예
Table 5. An example for co-occurrence frequency

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
t_1	38	4	4	12	2	3	1	3
t_2	4	65	11	4	2	3	17	0
t_3	4	11	56	3	7	0	0	0
t_4	12	4	3	72	1	4	5	2
t_5	2	2	7	1	53	1	6	4
t_6	3	3	1	4	1	27	1	9
t_7	1	17	0	5	6	1	93	0
t_8	3	0	0	2	4	9	0	48

또한 '동생'이라는 단어를 t_6, '먹다'라는 단어를 t_7이라고 했을 때, 표 5는 KAIST 국어정보베이스[18]의 코퍼스 일부에서 추출한 동시발생 빈도를 보이고 있다. 표 5에서는 예문 (1-a)나 (1-b)에서 '팔다'라는 단어에서 첫 음절의 중성 'ㅏ'을 인접한 음소 'ㅜ'로 잘못 입력하여 '풀다'라는 가의미 오류가 발생했을 때 오류를 검출하는데 사용하기 위해 예문 (1)에는 존재하지 않지만 수집된 코퍼스로부터 '풀다'라는 단어 t_8에 대한 동시발생 빈도 또한 수집되어있다.

표 5의 동시발생 상대 빈도로부터 두 단어의 의미적 패턴에 관한 유사도의 계산은 유클리드 거리(Euclidean distance), 자카드 계수(Jaccard coefficient), 피어슨의 상관계수(Pearson's correlation coefficient), 코사인 유사도(cosine similarity) 등과 같은 다양한 방법을 사용할 수 있다. 물론 사용하는 유사도 척도에 따라 표 5에서의 절대빈도는 적절하지 않을 수도 있다. 즉, 표 5와 같은 절대 빈도는 두 단어의 빈도 경향이 유사하더라도 단어의 전체 빈도에 따라 유사한 경향이 파악되지 않을 수도 있다. 예를 들어, 유클리드 거리를 사용한다면 경향이 유사하더라도 단어의 전체 절대빈도가 낮은 단어와 높은 단어의 유사도는 매우 다른 것을 계산될 것이다.

본 논문에서는 절대빈도에 영향을 받지 않는 코사인 유사도를 사용한다. 단, 유사도 계산 대상이 되는 두 기준 단어의 전체 빈도와 두 기준 단어 간의 동시 발생 빈도는 제외한다. 따라서 코사인 유사도는 식 (1)과 같이 계산한다. 이 식에서 f_{ik}는 t_i와 t_k의 동시 발생 빈도를 의미한다.

$$SIM(t_i,t_j)=\frac{\sum_{k=1(k\neq i,j)}^{n} f_{ik}\times f_{jk}}{\sqrt{\sum_{k=1(k\neq i,j)}^{n} f_{ik}^2}\times\sqrt{\sum_{k=1(k\neq i,j)}^{n} f_{jk}^2}} \quad (1)$$

표 5와 같은 동시 발생 빈도를 사용하여 '아주머니가 사과를 풀었다'라는 문장으로부터 '아주머니', '사과', '풀다'라는 세 단어 t_1, t_2, t_8에 대해 서로 코사인 유사도를 계산하면 식 (2)와 같이 계산된다.

$$\begin{aligned} SIM(t_1,t_2)&=\frac{122}{283.44}=0.43\\ SIM(t_1,t_8)&=\frac{59}{138.53}=0.42\\ SIM(t_2,t_8)&=\frac{55}{223.72}=0.24 \end{aligned} \quad (2)$$

가의미 오류를 검출하기 위해서는, 먼저 임계치 이하의 유

사도를 갖는 단어쌍을 선택한다. 만약 임계치가 0.3이라고 한다면 식 (2)의 계산결과로부터 t_2와 t_8 단어쌍을 선택할 수 있을 것이다. t_2와 t_8의 유사도는 임계치 이하로 낮아, 이 두 단어의 의미적 사용 패턴이 다른 쌍들에 비해 가장 어울리지 않음을 알 수 있다. 그러나 t_2와 t_8 둘 중 어느 것을 가의미 오류로 간주할 것인가에 대한 의문이 남는다.

본 논문에서는 낮은 유사도를 갖는 쌍들의 중복된 개별 목록에서 다수인 단어를 선택(majority voting)한다. 즉, 한 문장 내에서 유사도 계산 대상이 되는 모든 단어 쌍에 대한 유사도를 계산하고, 유사도 순으로 정렬하였을 때 가장 낮은 유사도를 갖는 2개 이상의 쌍을 선택한 뒤 쌍들 간에 중복되어 있는 단어를 선택한다. 수식 (2)의 결과를 예를 들면, 가장 낮은 유사도를 갖는 단어쌍 (t_2,t_8)과 (t_1,t_8)이 선택되고, 이들 쌍에서 공통으로 포함하고 있는 단어 t_8을 가의미 오류로 선택하게 된다. 이때 낮은 유사도를 갖는 단어쌍을 몇 개를 선택할 것인가도 고민해야할 문제가 될 수도 있다. 본 논문에서는 가장 낮은 유사도를 갖는 단어쌍에서부터 선택을 시작하여 두 번 나타나는 단어가 존재할 때까지 높은 유사도를 갖는 단어쌍을 선택한다.

이렇게 입력 실수로 발생한 가의미 오류를 선택할 수 있다. 남은 문제는 원거리 의존적인 단어쌍들을 동시발생 빈도의 계산에 포함시키면서도 문장의 길이가 길어졌을 때 서로 의미적 관계가 없는 단어쌍들에 대한 영향력을 제거하는 문제이다. 예문 (2)의 경우 표 5에서와 같은 방법으로 동시발생 빈도를 추출할 경우, '바람'이라는 단어에 대해 '날리다', '하얗다', '눈보라', '가로등', '떨리다', '깜빡이다'라는 동시발생 빈도 정보가 추출된다. 그러나 '바람'이라는 단어와 '하얗다', '가로등'이라는 단어들과는 명백히 의미적으로 직접적인 연관이 없다. 이러한 문제를 해결하기 위해 통계적 동시발생성에 대한 패턴 정보의 추출은 의존 문법[19, 20]을 이용한 부분 구문 분석 기법을 통해 문장의 의존 구조를 먼저 파악한 후 의존 관계에 놓인 단어들에 대해서만 조사한다. 예를 들어 예문 (2) 문장에 대한 의존 구조는 그림 1과 같다.

(2) 바람에 날리는 하얀 눈보라에 가로등마저 떨리는 듯 깜빡였다.

그림 1. 문장의 의존 구조
Fig. 1. Dependency tree of a sentence

따라서 '바람'이라는 단어에 대한 동시발생 빈도는 '날리다'라는 단어에 대해서만 계산된다. 물론 입력 문장에 대해 오류어를 검출할 때에도 모든 단어 쌍에 대한 의미 거리를 측정하는 것이 아니라 의존 관계에 놓인 단어에 대해서만 계산하게 된다. 그림 1에서 '떨다'*라는 단어는 '듯'과의 동시발생 빈도가 계산되어야 하지만 '듯'은 부사성 의존명사로 이 절의 첫 부분에서 설명한 대로 동시발생 빈도를 계산하는 대상이 되지 않는다. 따라서 '떨리는 → 듯 → 깜빡였다'라는 의존관계는 계산대상이 되지 않는 어절을 제거하고 '떨리는 → 깜빡였다'라는 관계로 축약되어 '떨다'라는 단어는 '깜빡이다'라는 단어에 대해서만 동시발생 빈도가 계산된다.

이상에서 설명한 동시발생 빈도정보에 기초한 가의미 오류 검출 알고리즘을 정리하면 크게 훈련 과정과 검사 과정으로 나눠진다. 훈련과정은 오류가 없는 코퍼스로부터 문장 단위로 다음 단계에 따라 수행된다.

가. 형태소 분석
나. 의존 구조 파악
다. 동시발생 빈도 계산

훈련 과정에서 계산된 동시발생 빈도는 편집 중인 문장에 대해 가의미 오류를 검출하기 위해 사용되며 검출과정은 다음과 같다.

가. 형태소 분석
나. 의존 구조 파악
다. 단어들 간의 유사도 계산
라. 임계치 이하의 유사도 단어쌍이 존재하는 문장 선택
마. 가장 낮은 유사도 단어쌍들으로부터 최다수 단어 선택

입력 실수에 의한 가의미 오류를 검출하는 과정에서 입력 문장에 훈련(training)되지 않은 단어가 포함 되었을 때에는 그 단어와 이웃하는 단어들에 대해 유사도를 계산하고 둘의 평균을 사용한다.

3. 교정 방법

가의미 오류에 대한 교정 방법은 비교적 간단하다. 가의미 오류의 원인이 인접한 키를 실수로 잘못 입력한 것이기 때문에 교정은 철자오류 교정방식과 유사한 방식을 취한다. 즉, 오류 어절을 구성하고 있는 음소들에 대해 키보드상에서 인접한 음소들 목록으로 하나씩 대치하거나 삽입, 삭제해가며 변경된 후의 어절에 대해 적법성 유무에 따라 판단하게 된다[21]. 철자 오류 교정과 다른 점은 단지 적법성 검사 방법에 있을 뿐이다. 철자 오류 교정에서 음소 변경된 후보의 적법성

* '떨리다'는 '떨다'라는 자동사 어근에 피동사로 변경시키는 피동접미사 '-리'가 붙어 파생된 것이므로 동시발생 빈도는 원형인 '떨다'에 대해 계산됨.

검사를 위해 형태소 분석기를 사용한다. 형태소 분석기의 분석 결과가 올바르거나 적법하다고 판단한다면 음소 변경된 어절을 하나의 교정 후보로 제시하게 된다. 반면에 가의미 오류의 적법성 검사에 있어서는 음소 변경된 어절에 대해 오류 검사에서 시행했던 것과 동일하게 주변 다른 어절들의 검사 대상 단어와 유사도를 계산하여 임계치보다 큰 유사도가 계산된다면 음소 변경된 어절에 대해 교정 후보로 제시한다.

철자오류 교정 방식에 비해 나은 부분은 적법성을 통과한 교정 후보들의 적합성을 보다 직관적으로 제시할 수 있다는 것이다. 음소를 변경해가며 적법한 어절들을 찾아내는 방식에서는 많은 교정 후보가 제시될 수 있는데, 철자오류 교정 방식에서 교정 후보들의 적합성은 적용된 음소 변형의 중요도나 자주 틀릴 가능성 등을 기준으로 임의로 할당된 점수를 통하여 자의적으로 계산한다. 그러나 가의미 오류 교정에서는 적법성 검사시 계산되는 주변 단어들과의 유사도를 이용하여 어울림 정도를 계산할 수 있으며, 이 방식이 보다 직관적일 뿐만 아니라 객관적이라 할 수 있다.

IV. 실험 및 평가

제안하는 방법의 검증을 위해서는 교열 이전의 자연스러운 오류가 포함된 대량의 문서를 필요로 한다. 본 논문에서는 서울신문사와 경향신문사의 3개월 분량의 기사 원고 초안을 사용하였다. 본 논문에서의 제안하는 방법은 맞춤법 검사 시스템에서 단독으로 사용될 수 없으며 일반적인 문서편집기가 가지고 있는 오류 유형별 검사기와 함께 사용되는 것이 바람직하다. 즉, 철자 오류의 검사, 문법 오류의 검사, 그리고 제한적이기는 하지만 지식베이스로 구축된 자주 틀리는 의미오류의 검사 기능을 통과하고 여전히 남아 있는 오류를 검출 대상으로 한다. 이런 까닭에 서울신문사 및 경향신문자의 기사 원고 초안 데이터를 그대로 사용하기에는 적합하지 않다. 이 원고에는 표 2에서 제시한 바와 같이 맞춤법 오류, 철자오류, 띄어쓰기 오류, 구문 오류, 순수 의미오류를 모두 포함하고 있기 때문이다.

따라서 이 데이터는 기존의 다른 워드프로세서와 맞춤법 전문가의 손을 빌려 가의미 오류를 제외한 나머지 오류는 미리 수정하였다. 그렇게 수정된 실험 데이터 전체 3,212개 어절 중 87개 어절 86개 문장에 가의미 오류가 남아 있는 상태이다. 단 한 개의 문장을 제외하고 가의미 오류가 발생한 문장은 단 한 개의 가의미 오류만 존재한다. 동시발생 빈도 정보는 모든 오류가 수정된 어절 3,212개를 대상으로 추출하였지만, 정확도 검증을 위해서는 가의미 오류를 포함하고 있는 86개 문장 중 두 개의 가의미 오류가 존재하는 한 개 문장을 제외하고 85개 문장만을 사용하였다. 그 이유는 제안하는 알고리즘은 문장 단위로 수행되므로 특정 가의미 오류를 검출하는 데에는 그 가의미 오류를 포함하고 있는 문장 이외의 다른 문장은 사용되지 않기 때문이다.

그림 2는 임계치를 0.1에서부터 0.9까지 0.1 단위로 변경해가며 검출 정확률을 측정한 결과이다. 그림 2에서 보는 바와 같이 신문데이터의 경우 임계치 0.3을 기준으로 정점을 기록하였으며 임계치가 0.3보다 작거나 0.3보다 클 경우 검출 정확율이 하락함을 알 수 있다.

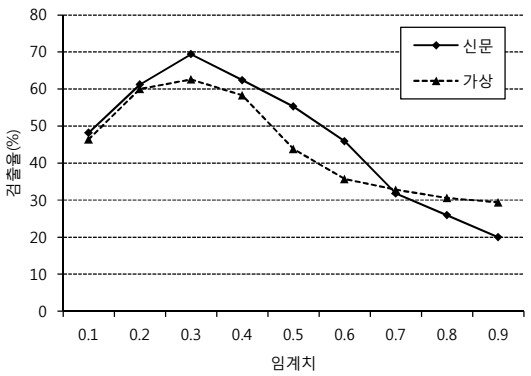

그림 2. 검출율
Fig. 2. Detection rates

신문 데이터는 문서 작성자가 입력 실수한 의미오류로 발전한 철자오류를 포함하고 있어 실제 실수할 가능성이 있는 오류들을 포함하고 있다. 그러나 평가 데이터 량이 많지 않아 평가의 신뢰성을 높이기 위해 별도의 평가 데이터를 마련하였다. 별도의 평가 데이터를 위해서 국어정보베이스1과 ETRI 코퍼스에서 트리 태깅된 약 650여 문장을 사용하였다. 의존구조 분석은 구구조 트리 태깅된 정보로부터 몇 가지 규칙을 통해 자동으로 의존구조로 변화하였으며, 오류가 없는 문장으로부터 동시 발생 빈도를 조사하였다. 가의미 오류는 모든 문장에서 임의로 선택된 한 단어에 대해 어근이나 어간 위치에 철자의 교체, 삽입, 삭제 등을 통하여 자동으로 생성하였다. 이때 철자의 교체는 50%, 삽입은 25%, 삭제는 25% 비율로 균질하게 적용하였으며, 교체와 삽입 대상이 되는 음소의 위치(초/중/종성)와 교체 및 삽입되는 음소는 모두 키보드 상에서 원래 음소의 키에서 이웃하는 키의 음소로 임의로 선택하였다. 이는 편집자가 실수하는 입력 오류는 원래 자소의 키에 이웃하는 키를 잘못 입력하는 경향을 반영하기 위한 것이다. 이렇게 생성된 가상의 철자 오류들 중 가의미 오류만을 남기

기 위해 생성된 철자 오류를 철자 검사기용 형태소 분석기의 분석을 거쳐 분석 성공한 어절만을 오류로 확정하였다. 물론 이들 어절에는 가의미 오류뿐만 아니라 결과론적 구문 오류도 일부 포함될 것이며, 또한 매우 제한적이기는 하나 맞춤법 오류도 포함될 것이다. 이들이 정확한 가의미 오류 검출율을 측정하는데 방해가 될 것임에 분명하나 이 데이터를 사용한 실험은 개략적인 수준에서의 검출율의 영향력을 파악하기 위한 것이므로 결론에 이르는데 그 효과는 큰 영향을 받지 않을 것이다.

이렇게 생성한 가의미 오류는 235개였으며 문장당 한 개의 가의미 오류를 포함하고 있다. 이 데이터로부터 검출 정확률을 측정한 결과 그림 2에서 보이고 있다. 그림 2의 가상으로 생성한 데이터에 대한 검출율은 신문상의 오류 데이터보다 전체적으로 낮아졌으며, 제일 높은 검출율을 갖는 임계치 0.3보다 다른 임계치가 현저히 낮아진 것을 알 수가 있다.

표 3으로부터 가의미 오류의 비율을 전체 오류에서 대략 4% 이내로 유추할 수 있다. 그리고 그림 2에서 보는 바와 같이 임계치가 0.3일 때 거의 70%에 가까운 검출율을 보이고 있으므로 제안하는 방법이 실제 맞춤법 검사기에 적용되었을 때 전체 약 2~3% 수준의 검출율을 개선할 수 있을 것으로 예상된다.

그러나 본 논문에서 제시한 검사 방법은 몇 가지 문제점에 노출되고 있다. 첫 번째 문제점은 문장내에서 지역적으로 발생하는 단어들의 동시발생성은 의미적으로 유사하지 않고 단지 의미적으로 연관돼 있을 뿐이라는 것이다. 따라서 여기에서 사용된 동시발생 빈도 기반 의미 정보는 특정 응용 분야에서만 적용이 가능하다. 두 번째, 훈련되지 않은 단어에 매우 취약한 면을 나타낸다. 특히 문장이나 구가 매우 짧을 경우 문제는 더욱 심각해진다. 예를 들어 "나는 집에 간다"라는 입력 문장이 있고 '집'이라는 단어가 훈련되지 않았다고 가정한다면 '집'이라는 단어의 의미를 결정하기 위해 주변 단어 '나'와 '간다'라는 단어의 동시발생 빈도의 평균을 사용하게 된다. 이 때 '집'이라는 단어와 '간다'라는 단어의 의미적 거리의 계산시에 이미 '집'이라는 단어에 '간다'라는 단어에 대한 의미적 정보가 포함되기 때문에 오류 검출률이 저하될 수 있을 것이다. 세 번째, 단어마다 어휘의 빈도가 불균일하다는 것이다. 불균일한 빈도의 어휘를 동일한 공간상으로 유사도를 계산하면 낮은 빈도의 어휘에 대해서는 검출에 대한 편차가 심해질 수 있다.

V. 결 론

본 논문에서는 기존의 맞춤법 오류 검출을 위한 방법을 살펴보았고, 기존의 시스템들이 의미 오류라고 분류한 오류들을 오류 원인별로 구분을 시도하였으며, 그들 중 입력 실수로 인하여 발생한 가의미 오류를 동시발생 빈도 정보와 이 빈도 정보를 기반으로 한 유사도 계산 알고리즘으로 검출을 시도하였다. 가의미 오류는 순수의미 오류에 비해 의미적으로 오류 단어가 주변 단어의 문맥에서 발생 가능성이 현저히 떨어지는 현상을 보인다. 이러한 특징을 반영하는 동시발생 빈도에 기반한 방법을 제시하였다. 동시발생 빈도에 기반한 방법에서 불필요한 의미적 관계를 제거하기 위해 의존구조 정보를 활용하였다. 제안하는 방법은 성능 측면에서 포화상태에 있는 맞춤법 검사기 성능을 개선하는데 도움이 될 뿐만 아니라, 나아가 제한적이기는 하지만 문법 오류 검출에도 도움을 줄 수 있을 것으로 기대한다.

보다 뛰어난 성능을 위한 향후 연구 과제로는 제시한 세 가지의 문제점을 해결해야 할 것이며, 특히 두 번째 문제점은 오류 검출률을 저하시키는 가장 큰 요인으로 시급히 해결되어야 할 것이다.

참고문헌

[1] Byung-hoon Lee, Korean Spelling Corrector Based on Corpus Analysis, MS Thesis, Yonsei University, 1993.

[2] Dong-joo Kim, "Detecting Spelling Errors by Comparison of Words within a Document," Journal of the Korea Society of Computer and Information, Vol. 16, No. 12, pp. 83-92, 2011.

[3] Dong-joo Kim, et al., "Design and Implementation of Morphological Analyser for Korean Spell Checker," Proceedings of IEEK Summer Conference, IEEK, Vol. 20, No. 1, pp. 255-258, 1997.

[4] Hyuk-chul Kwon, "Korean Spelling and Grammar Checker", Journal of the Korea Society of Computer and Information, Vol. 15, No. 10, pp. 24-34, 1997.

[5] Hall, Patrick A. V., et al., "Approximate string matching," ACM Computing Surveys, vol. 12, No. 4, pp. 381-402, December, 1980.

[6] Jae-Hyuk Choi, "Automatic Korean Spacing Words Correction System With Bidirectional Longest Match Strategy," Proceedings of the 9th

Conference on Hangul and Korean Information Processing, pp. 304-315, 1997.
〔7〕 Seung-Shik Kang, et al., "Morphological Analysis and Spelling Check Function of Korean Morphological Analyzer HAM," Proceedings of the 8th Conference on Hangul and Korean Information Processing, pp. 246-252, 1996.
〔8〕 Hang Li, et al., "Word Clustering and Disambiguation Based on Co-occurrence Data," The On-Line Proceedings of the ACL, 1998.
〔9〕 Ellen Riloff, "Atomatically Generating Extraction Patterns from Untagged Text," Proceedings of the AAAI-96, pp. 1044-1049, 1996.
〔10〕 Kong-joo Lee, et al., "Automatic Word Classification and Wordtags in Korean," Proceedings of the 23th KISS Spring Conference, Vol. 23, No. 1, pp. 961-964, 1996.
〔11〕 Young-sin Lee, et al., "Automatic Spelling Correction using an Error-tolerant Morphological Analyzer and Co-occurrence Information," Proceedings of the 24th KISS Spring Conference, Vol. 24, No. 1, pp. 411-413, 1998.
〔12〕 Ted Pedersen, et al., "A New Supervised Learning Algorithm for Word Sense Disambiguation," Proceedings of the AAAI-97, pp. 604-609, 1997.
〔13〕 Marc Light, "Morphological Cues for Lexical Semantics," The On-Line Proceedings of the ACL, 1996.
〔14〕 Kamal Nigram, et al., "Learning to Classify Text from Labeled and Unlabeled Documents," Proceedings of the AAAI-98, pp. 792-799, 1998.
〔15〕 Yun-jin Nam, et al., "Constructing Dictionary Information for the Processing of Derivational Suffixes of Nouns based on Corpus Analysis," Journal of the Korea Society of Computer and Information, Vol. 23, No. 4, pp. 389-401, 1996.
〔16〕 Fernando Pereira, et al., "Distributional Clustering of English Words," ACL On-line proceeding, 1994.
〔17〕 Lillian Jane Lee, Similarity-Based Approaches to Natural Language Processing, Ph. D. Thesis, Harvard University, 1997.
〔18〕 Young-soog Chae, et al., "Introduction of KIBS (Korean Information Base System) Project," International Conference on Language Resources and Evalution (LREC2000), Serial. 2, Athens, Greece, pp. 1731-1735, 2000.
〔19〕 Dae-seon Choi, et al., "A Two-Phase Dependency Parser of Korean," Proceedings of the natural language pacific rim symposium, 1995.
〔20〕 Jong-hyeok Lee, et al., "Structural Disambiguation Using Constraint-Satisfaction Algorithm for Dependency Parsing," Proceesings of the International Conference on Computer Processing of Oriental Language, pp. 213-216, 1995.
〔21〕 Hyung-jong Noh, et al., "A Joint Statistical Model for Word Spacing and Spelling Error Correction Simultaneously," Journal of the Korea Information Science Society: Software and Applications, Vol. 34, No. 2, pp. 131-139, 07.

저 자 소 개

김 동 주
1996: 한양대학교 전자계산학과 공학사.
1998: 한양대학교 전자계산학과 공학석사.
2007: 한양대학교 컴퓨터공학과 공학박사
현 재: 안양대학교 교양대학 교수
관심분야: 맞춤법검사, 기계번역, 한국어정보처리, 의견검색, 감정인식
E-mail: djkim@anyang.ac.kr