말뭉치의 통계정보를 이용한 한국어 글쓰기 도우미 시스템

이재승○, 유주현, 이현호, 이현아
금오공과대학교 컴퓨터소프트웨어공학과
grayhacker91@gmail.com○, wer053@naver.com, hyunho513@naver.com, halee@kumoh.ac.kr

Korean Writing Assistant System using Corpus Statistics

Jae-Seoung Lee○, Joo-Hyun Yu, Hyun-Ho Lee, Hyun Ah Lee
Dept. of Computer Sofrware Engineering, Kumoh National Institute of Technology

요 약

온라인을 통해 접하게 되는 잘못된 우리말 표현과 외국어 중심 교육 등으로 인하여 학생들의 한국어 능력, 특히 글쓰기 능력에 우려가 높아지고 있다. 본 논문에서는 잘 작성된 말뭉치에서 얻어진 데이터에 기반한 한국어 글쓰기 도우미 시스템을 제안한다. 시스템은 작성 중인 문맥에 맞은 단어를 추천하는 용언/체언 추천과 입력 문장의 주요 단어가 포함된 말뭉치의 문장을 제시하는 유사 문장 추천, 문서의 단어가 문서의 문맥 단어와 조화로운지를 확인하는 어휘 응집성 검사, 단어 중복도를 확인하기 위한 단어 빈도 검사 기능을 제공한다. 시스템에서는 사용자가 말뭉치를 추가하면 색인을 구축할 수 있어 원하는 분야에 맞는 추천과 검사 기능을 제공할 수 있다.

주제어: 글쓰기 도우미 시스템, 용언/체언 추천, 유사 문장 추천, 어휘 응집성

1. 서론

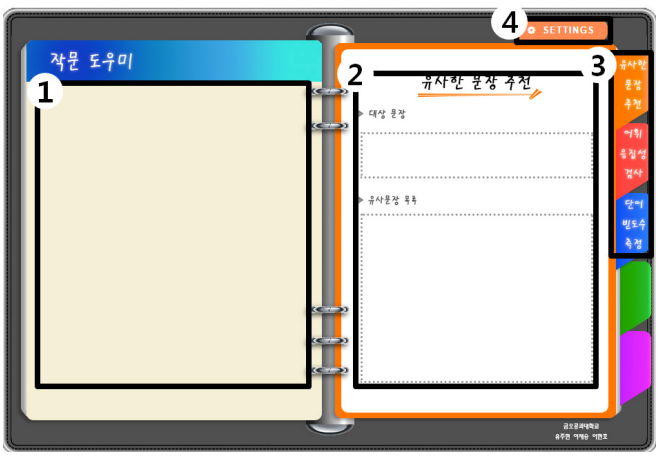
온라인을 통해 접하게 되는 잘못된 우리말 표현과 외국어 중심 교육 등으로 인하여 청년층의 한국어 능력에 대한 우려가 높아지고 있다. 2009년에 대학생 웹진 바이트에서 수도권 지역 대학생 300명을 대상으로 진행한 설문조사 '대학생의 한글 사용에 대한 어려움' 에 따르면 대학생의 80%가 한국어 사용에 어려움을 느끼고 있는 것으로 나타났다. 2013년 취업포털 사람인에서 실시한 기업 인사 담당자를 대상으로 한 설문조사에서는 62.2%가 신입사원을 뽑을 때 국어 실력을 평가할 필요를 느끼고 있으며, 신입사원에게 가장 부족한 능력으로 '기획안 및 보고서 작성능력' 을 뽑았다. 이와 같이 대학생이나 신입사원들은 보고서, 연구 논문, 취업 준비를 위한 자기소개서 등에서 높은 수준의 글쓰기 능력이 필요하지만, 이들 중 대다수는 글의 구성력이 미숙하여 문장의 호응관계가 맞지 않거나, 어휘 구사, 문장 표현 등에서 어려움을 호소하고 있다.

자연언어처리에 기반을 둔 글쓰기를 돕거나 평가하기 위한 연구들이 이루어져왔다. 외국인을 위한 학습 도우미 시스템[1]은 짧은 문장에서의 조사/용언, 문장 추천을 통해 외국인의 한국어 글쓰기를 지원한다. 한국어 맞춤법/문법 검사기[2,3]는 다양한 한글 입력에 대한 맞춤법 교정을 제공하지만, 글쓰기보다는 오류를 수정하는 데에 초점이 되어 있으며, 우리말 배움터[4] 등과 같이 수동으로 구축한 지식에 의존적인 문제점을 가지고 있다. 근래 들어 글쓰기 평가에 대한 여러 연구가 진행되어 왔으나[5,6] 이 역시 수동으로 작성된 문법 체계나 어휘 사전을 중심으로 구성되어 있다.

본 논문에서는 대량의 문서에 나타난 통계정보를 활용한 글쓰기 지원 시스템을 소개한다. 시스템에서는 작문 중에 필요한 용언과 체언을 추천하는 용언/체언 추천 기능, 잘 작성된 말뭉치 문장과 유사한 문장을 제시하여 문장을 교정할 수 있게 지원하는 유사문장 추천 기능, 작성된 문장들에 사용된 단어들이 다른 문서에서도 공기하여 발생하는지를 점검하여 단어 사용의 적절성을 확인하는 응집성 검사 기능, 특정 단어의 지나친 반복 사용을 점검하기 위한 단어 빈도 측정 기능을 제공하여 한 단어, 글쓰기에서 발생하는 어휘 구사와 문장 표현의 어려움을 해결하고자 한다.

2. 한국어 작문 지원 시스템

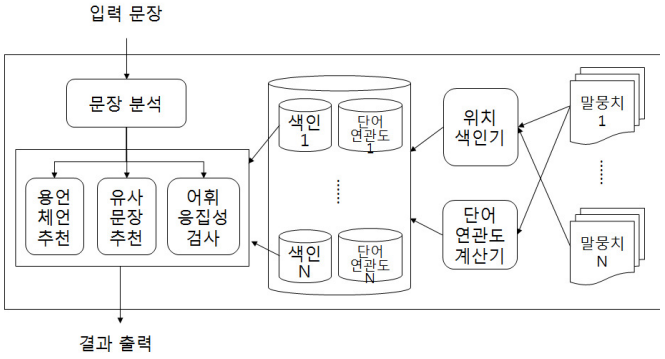
[그림 1]은 제안하는 작문 지원 시스템의 사용자인터페이스를 보인다. ①은 사용자의 작문 공간이며, ②에서



[그림 1] 사용자 인터페이스

는 ③의 각 기능에 대해 결과를 제시한다. ③은 탭으로 구성되어 시스템의 기능인 유사한 문장 추천, 어휘 응집성 검사, 빈도 검사를 선택할 수 있다. 유사문장 추천은 사용자가 입력을 일정시간 잠시 멈출 때마다 별도의 요청 없이 실시간으로 추천 결과를 제시하여 사용자 편의를 도모한다. 어휘 응집성 검사와 빈도 검사는 사용자가 요청하는 경우 작성된 문서에 대한 각 검사 결과를 제시한다. 시스템 기능 중 용언/체언추천 기능은 특정키를 입력하면 제공한다.

시스템에서는 각 기능의 빠른 처리를 위해 단어연관도 사전과 색인을 구축하여 사용한다. [그림 2]는 시스템 구조도를 보인다. 제안하는 시스템에서는 사용자가 작성하는 문서의 특성에 맞는 글쓰기 지원을 위해, 다양한 분야의 말뭉치를 추가 선택할 수 있게 구성한다. 시스템에서는 루씬 라이브러리[7]을 사용하여 사용자가 입력하는 다양한 말뭉치로부터 분야별 색인과 단어연관도 사전을 생성할 수 있게 기능을 제공하여, 각 분야에 맞는 추천이나 검사 결과를 제공하여 용도에 맞는 글쓰기를 지원한다. 아래에서는 각 부분에 대해 상세히 설명한다.



[그림 2] 시스템 구조도

2.1 용언/체언 추천

글쓰기에서 발생하는 어려움 중의 하나는 작성하는 문장의 문맥에 맞는 단어를 선택하는 문제이며, 이러한 문제는 내용어인 명사와 동사, 형용사에서 더욱 크게 나타난다. 시스템의 용언/체언 추천에서는 말뭉치에서 나타나는 단어간 공기 빈도를 활용하여, 현재 작성 중인 문서에 적합한 용언이나 체언을 추천한다.

용언과 체언 추출에서는 앞 형태소의 품사를 활용한다. 예를 들어 아래 예문에서 '제안하는'과 '부모님의' 뒤에 나타나는 적합한 단어는 체언이고, '기술을'과 '적용하여' 뒤에 적합한 단어는 용언이다. 시스템에서는 앞에 나타나는 형태소가 '용언+관형사형 어미'인 경우에는 해당 용언에 적합한 체언을, 앞 형태소가 '체언+관형사격 조사 또는 접속격 조사'인 경우에도 해당 체언에 적합한 체언을, 앞 형태소가 '용언+연결어미'나 '체언+이외 조사'인 경우에는 체언이나 용언과 쌍을 이를 수 있는 용언을 추천한다. 예문 (다)의 경우 '사랑은' 뒤에는 '끝이'가 발생하지만 시스템에서는 '사랑'의 지배소에 해당하는 '없다'를 추천하는 것을 목표로 한다. 만일 앞 형태소가 어미나 조사를 가지지 않는 경우에는 용언과 체언을 구분하지 않고 추천한다.

(가) 제안하는 방식에 적합한 기술을 선택하였다.
(나) 신기술을 적용하여 완성하였습니다.
(다) 부모님의 사랑은 끝이 없다는 것을 깨달았다.

용언과 체언 추천에서는 단어 연관성 점수를 사용한다. 만일 두 단어가 주어진 분야의 문서에서 자주 연속되어 자주 발생한다면 연관성이 높다고 볼 수 있다. 하지만, 저자의 작문 습관에 의해 한 문서에서 중복되어 나타나는 표현은 각기 다른 문서에서의 발생하는 표현과 동일하게 간주되어서는 안 된다. 아래 식 (1)은 문서 d에서의 두 단어 w_i와 w_j의 가중치가 반영된 공기빈도를 계산한다. 수식에서 $freq_d(w_iw_j)$는 문서 d에서 단어 w_i와 w_j가 연속하여 발생하는 빈도를 나타낸다. 단어 w_i와 w_j는 체언과 용언만을 대상으로 하며, 문장에서 발생하는 기능어를 제거한 뒤 연속하여 발생하는 빈도만을 추출한다. 예를 들어 앞의 예문 (가)에서는 '제안하다 방식 적합하다 기술 선택하다'를 추출한다. 얻어진 빈도에 [0,1]의 범위를 가지는 가중치 α를 적용하여 가중치 반영 빈도를 구한다. 단일 문서에서 중복되어 발생하는 경우의 중요도를 낮게 볼수록 α에 작은 값을 부여한다.

$$wfreq_d(w_iw_j) = \alpha \cdot freq_d(w_iw_j) + 1 - \alpha \qquad (1)$$

두 단어의 연관도(WRS, Word Relatedness Score)는 대상 분야 말뭉치 C에 포함되는 모든 문서 d_k에서의 가중치 반영 빈도를 합산하여 식 (2)으로 구한다. 시스템에서는 사용자가 입력한 마지막 어절의 용언이나 체언 w에 대하여 $WRS_C(ww_i)$의 내림차순으로 추천 단어를 제시하여 다음 어절로 적합한 용언이나 체언을 추천한다.

$$WRS_C(w_i, w_j) = \sum_{d_k \in C} wfreq_{d_k}(w_iw_j) \qquad (2)$$

2.2 유사 문장 추천

유사 문장 추천은 입력된 텍스트와 유사한 말뭉치의 문장을 찾아 유사도 내림차순으로 사용자에게 제시한다. 시스템에서는 입력된 문장에서 키워드를 추출하여 키워드 열(keyword sequence)를 얻고, 입력 문장의 키워드 열과 말뭉치 문장의 키워드 열의 유사도를 편집거리를 구하는 Levenshtein Distance[8] 알고리즘으로 계산한다. 얻어지는 값은 유사할수록 작은 값을 가지므로, 시스템에서는 거리값이 3이하인 문장을 거리의 오름차순으로 사용자에게 제시한다.

예를 들어 사용자가 입력한 미완성된 문장 "최고가 되기 위해 노력"에서는 키워드 열로 (최고, 노력)을 추출할 수 있다. 키워드 열 추출에서는 형태소 분석 결과로 얻어지는 체언과 용언을 사용한다. 만일 말뭉치의 문장 중에 "어떤 일에 최고가 되기 위해서는 더 많은 노력이 필요하다"가 존재하면 다음과 같은 키워드 열(어떤, 일, 최고, 노력, 필요)을 얻을 수 있다 입력된 문장의 키워드 열은 (최고, 노력)이고 편집거리를 구하면 3이므로 사용자에게 문장을 추천 한다.

만약 편집거리로 말뭉치에서 유사한 문장이 하나도 추천 되지 않는 경우에는 입력된 텍스트의 키워드 열과 말

뭉치 문장의 Term Frequency Vector의 유사도를 계산하여[7], 가장 유사한 문장들을 찾아 추천한다.

2.3 어휘 응집성 검사

글쓰기에서 발생하는 오류 중 하나는 작성하는 문서의 특성에 맞지 않은 용어를 사용하거나, 글쓰기의 문맥이 맞지 않은 내용으로 전이하는 것이다. 어휘 응집성이란 문서 내의 어휘들을 통하여 문서가 동일한 대상이나 주제에 대해 이야기를 하고 있는지에 대한 척도이다[6]. 제안하는 시스템에서는 어휘 응집성을 검사하여 작성하고 있는 문서의 단어와 주제의 일관성을 검사하고자 한다.

어휘 응집성은 작성된 문서의 체언과 용언으로 구성된 키워드 집합을 이용하여서 측정한다. 아래의 식 (3)은 대상 말뭉치 C에서 단어 w_i의 적합도(CSS, Category Suitability Score)를 계산한다. 예를 들어 w_1, w_2, w_3, w_4의 네 단어가 작성 중인 문서에 발생하는 경우, 대상 말뭉치에 w_1은 w_2~w_4와 같은 문서에서 자주 발생되는 빈도가 높지만, w_4는 w_1~w_3과 같은 문서에서 발생되는 빈도가 낮다면, w_4는 응집성이 떨어지는 용어로 판단할 수 있다. 식 (3)에서 $df_C(w_i)$는 대상 말뭉치 C에서의 단어 w_i의 문서빈도(document frequence, df)를 의미한다. 두 단어가 같이 발생한 문서빈도를 각 단어의 문서빈도의 곱으로 나누어, 같이 발생하는 단어가 많을수록 높은 점수를 가지되, 단어 자체의 발생빈도가 높아 우연히 높은 점수를 가지지 않도록 조정한다.

$$CSS(w_i, C) = \sum_{w_j \in C} \frac{df_C(w_i \cap w_j)}{df_C(w_i) * df_C(w_j)} \quad (3)$$

시스템에서는 얻어진 적합도 값이 임계치 이하인 단어를 [그림 1]의 오른쪽 화면에 목록으로 보이고, 목록에 포함된 단어를 왼쪽 화면에 밑줄로 표시하여, 응집성이 떨어지는 단어를 눈에 보기 쉽게 제시한다.

2.4 단어 빈도 측정

주제에 대한 글쓰기에서 동일한 단어를 지나치게 자주 반복하여 사용하는 것은 좋은 글쓰기 방법이 아니다. 제안하는 시스템에서는 작성된 문서에 포함된 단어의 빈도를 내림차순으로 정렬하여 제시하는 기능을 제공하여 세련된 글쓰기를 지원한다.

3. 평가 및 결론

시스템의 평가를 위해 우수 자소서, 사전 예문, 뉴스 사설, 세종말뭉치의 문장들에 대한 색인을 수행하고 각 기능을 평가하였다. 유사 문장 검색에서는 “성실함을 무기로서”로 검색한 결과 “입사가 허락이 된다면 저의 무기인 성실함과 책임감, 여기에 열정을 쏟아 부어서 귀사에 도움이 되는 인재 역량 있는 인재가 될 것입니다”와 “제가 가진 가장 큰 무기는 "성실함과 끈기, 그리고 타인과의 친화력"입니다” 등의 문장이 추천되어, 글쓰기에 도움이 되는 것으로 분석되었다. 하지만, 유사 문장 추천 기능의 경우 사용자가 잘못 활용하는 경우 표절의 우려가 있어 이에 대한 대처 방안이 필요한 것으로 분석되었다. 용언/체언 추천 기능에서는 “아이디어를”을 입력하는 경우 “낼 수 있는”, “통해”, “생각해”, “이끌어” 등의 다양한 용언이 추천되어, 글쓰기에서 적합한 단어가 떠오르지 않을 때에 도움이 될 수 있는 것으로 분석되었다. 아래 예제는 어휘 응집성 검사에서의 예제를 보인다. 예문 (라)의 결과에서는 ‘동아리’, ‘가입’, ‘활동’, ‘선배’, ‘역할’, ‘조직’, ‘책임감’, ‘리더십’ 등은 말뭉치에서 서로 자주 발생하는 단어이지만, ‘운동’과 ‘테니스’는 함께 잘 사용되지 않는다고 판단하여 밑줄로 표시된다. (마)에서는 영업 관리에 대해서 ‘선봉장’이라는 표현도 잘 쓰이지 않아 밑줄로 표시된 결과를 보인다. 시스템 사용자는 응집성 결과를 참고하여 글쓰기 결과를 수정할지 자신이 작성한 문장의 독창성을 살려 그대로 둔 것인지 여부를 결정할 수 있다.

(라) 운동을 좋아하기 때문에 테니스 동아리에 가입을 하여 활동하였습니다. 또 선배로서의 역할을 통해 한 조직을 이끌어 감으로써 책임감과 리더십을 배웠습니다.
(마) 고객과의 접점에 있는 영업 관리 직무는 선봉장이라고 할 수 있습니다.

향후 연구로는 제안한 시스템에 대한 정성적 평가와 정량적 평가를 통한 성능 향상과 추가 기능 개발을 계획하고 있다. 추가 기능으로는 문법 검사와 맞춤법 검사, 의미론적인 수준에서의 글쓰기 지원 방안을 예정하고 있다.

참고문헌

[1] 박기태, 이태훈, 황소현, 김병만, 이현아, 신윤식, "자동 추출된 지식에 기반한 한국어 학습 지원 시스템", 정보처리학회논문지 소프트웨어 및 데이터 공학 제1권 제2호, pp. 91-102, 2012.
[2] http://speller.cs. pusan.ac.kr/
[3] http://www.naver.com
[4] http://urimal.cs.pusan.ac.kr/urimal_new/learn/writing/main.asp
[5] 김지은, 이공주, "중학생 영작문 실력 향상을 위한 자동 문법 채점 시스템 구축", 한국콘텐츠학회 논문지 제7권 제5호, pp. 36-46, 2007.
[6] 김동성, 김상철, 채희락, "문법성과 어휘 응집성 기반의 영어 작문 평가 시스템", 인지과학 제19권 제3호, pp. 223~255, 2008.
[7] http://lucene.apache.org/
[8] https://en.wikipedia.org/wiki/Edit_distance