웹 검색을 이용한 자동 어학 문제 풀이 시스템

최현대, 윤형석, 이현아
금오공과대학교 컴퓨터공학부
e-mail : spankle@nate.com, truth4u@lycos.co.kr, halee@kumoh.ac.kr

Automatic Problem Solving System Using Web Information

HyunDae Choi, Hyung-Seok Yoon, Hyun Ah Lee
School of Computer & Software Engineering, Kumoh National Institute of Technology

요 약

현재 우리나라에서는 영어에 대한 중요성과 관심이 점점 커지고 있으며, 영어 능력을 평가하는 다양한 시험이 시행 중에 있다. 이런 시험들을 준비하기 위해 많은 문제들은 웹 상에서 손쉽게 구할 수 있는 반면에, 획득한 문제에 대한 정답을 원하는 순간에 구하는 것은 쉽지 않아 영어 문제를 푼 후에 정답을 확인할 수 없는 경우가 많다. 이런 불편함을 줄이기 위해 본 논문은 영어 문제의 정답을 추천해 주는 시스템에 대해서 논의한다. 단문 빈칸 채우기 형식의 문제에 대해서 해당 문제의 문장의 의미에 대한 이해 없이도 특정 어휘의 쓰임새나 빈칸 주변의 문맥 정보, 단어들 간의 공기빈도 정보를 이용하여 문제의 정답을 추천한다. 시스템에 필요한 정보를 위한 자료를 웹 상의 수많은 영어 문서들에 기술된 표현을 이용하여 수동 지식 구축과정 없이 문제를 해결한다.

1. 서론

세계화 시대에 따라 우리나라에서도 영어에 대한 수요가 급증하고 있고, 이런 추세를 반영하듯이 외국어 고등학교와 같은 특수목적 고등학교에서는 신입생 선발요건으로 일정 수준 이상의 영어 능력 검증 시험에서의 점수를 요구한다. 또한, 많은 대학에서도 신입생 선발 시 영어 능력 검증 시험에 대한 점수를 반영하고, 졸업 요건에도 일정 수준 이상의 영어 점수를 요구한다. 기업체에서도 물론 신입사원 선발이나, 승진, 해외파견에 있어서 영어 점수를 중요한 평가 요소로 사용한다.

이런 상황에서 자연스럽게 개인의 영어 능력을 검증하는 많은 시험들이 시행 중에 있다. 그 중 하나인 토익은 영어를 모국어로 사용하지 않는 사람들을 대상으로 일상생활과 비즈니스 현장에서 필요한 영어능력을 측정하는 실용영어 평가 시험이다. 이외에 서울대 어학 연구소에서 개발한 한국인을 위한 영어 시험인 TEPS, TOEFL, G-TELP, SEPT 등의 다양한 시험이 존재한다

우리나라에서 가장 많이 사용되고 있는 공인 어학능력 시험인 토익은 응시자의 영어 능력 검증을 위해서 크게 Listening Part 와 Reading Part 로 구성되어 있다. Listening Part 는 Part 1~Part 4 의 네 부분으로 구성되고, 전반적인 영어 듣기에 대한 평가를 한다. Reading Part 는 Part 5, 6, 7 의 세 부분으로 나뉘어져 있다. 그 중에서 Part 5 는 그 문장 가운데 한 부분이 비어져 있는 한 문장이 제시되고, 빈칸을 채우기 위해 사용되는 네 개의 보기가 제시된다. 여기에서는 응시자들의 어휘, 문법이 어느 정도 수준인지 평가한다. Part 6 는 편지, 광고, 알림 등의 지문이 제시되고, 그 문서 중 빈칸을 비워놓고 응시자에게 적절한 단어를 선택하게 한다. Part 5 와 같은 어휘, 문법에 대한 문제도 있지만 제시된 글의 앞, 뒤 문맥을 확인해서 풀어야 하는 문제도 출제된다. Part 7 는 하나 혹은 두 개의 잡지, 뉴스 기사, 편지, 광고와 같은 지문이 제시되고, 지문 당 두 개에서 다섯 개의 문제가 주어진다.

이 중 Reading Part 에서 단기간에 점수를 올리기 용이한 부분이 Part 5 이다. Part 6 와 Part 7 의 경우에는 주어진 지문의 전체적인 이해가 있어야 문제를 풀 수 있고, 이런 능력은 짧은 시간에 향상되지 않는다. 하지만 Part 5 는 기본적인 문법 지식만 알고 있으면 풀이가 가능하기 때문에 비교적 적은 시간으로 실력을 향상할 수 있다.

토익 Part 5 에서는 크게 문법 문제와 어휘 문제가 출제된다. 총 40 문제 중에서 어휘 문제는 15 문제가 출제가 되고, 25 문제가 문법 문제로 출제된다. 문법 문제의 경우에는 4 개의 보기가 모두 다른 품사인 경우가 50%이상 출제된다. 어휘 문제의 경우에는 숙어, 짝표현, 어울리는 표현에 대한 문제가 70%이상 출제된다.

사회적 필요성에 의해 다양한 어학 시험을 대비하기 위하여 토익과 관련한 인터넷 사이트들이 많이 생겨나고 있다[1,2,3]. 그리고 이런 사이트에서는 사용자들에게 토익 예상 문제를 제공한다. 온라인 영어교육 부분에서 1 위를 차지하고 있는 해커스 토익 사이트의 경우에도 토익 시험 이 치뤄지는 날짜를 기준으로 많은 토익 예상 문제들이 사이트에 업로드 된다. 사용자들은 그 문제들을 다운받아서 풀어 볼 수 있지만, 해당 문제에 대한 정답을 같이 포함하고

있는 경우는 드물다. 이런 상황에서 사용자들은 예상 문제를 풀이한다고 해도 정답을 바로 확인할 수 없다. 이런 문제들에 대한 정답이 문제가 업로드 된 뒤 1 주일 정도 후에 문제의 답이 따로 업데이트가 되기는 하지만, 그 동안 사용자들은 정답을 확인할 수 없어서 답답함을 느끼는 경우가 많다. 그리고 나중에 정답을 확인할 수 있다고 해도 확인하기 어려운 형태로 게시된다.

본 논문에서는 영어학습자들의 이런 불편함을 줄이기 위해 자동으로 영어 문제의 답안을 추천하는 시스템을 제안한다. 어학학습을 위한 자동화된 기존의 시스템이 언어학적인 지식에 기반한 영어 작문 위주로 구축되는 것[4,5,6]에 반하여, 본 논문에서 제안하는 시스템은 간단한 통계 정보에 기반하여 영어 능력 시험에 관한 풀이를 제공하는 것을 목적으로 한다.

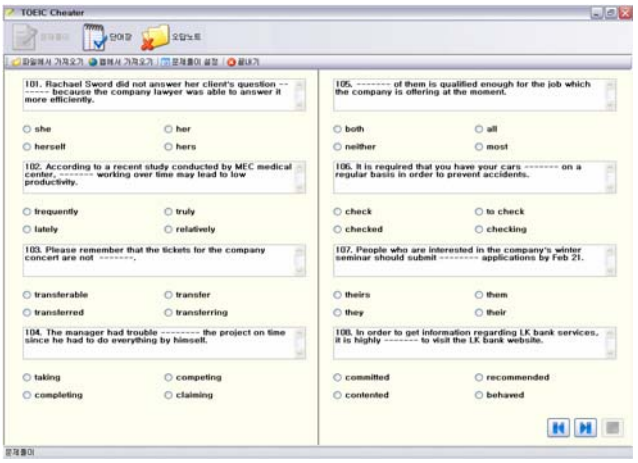
토익의 Part 5 의 경우 네 개의 보기가 제시되는 단문 빈칸 채우기 형태의 문제들이 출제된다. 이런 문제의 경우, 각 보기의 항목을 빈칸에 대입했을 때 문법에 가장 적절하고 문맥이 자연스러운 보기의 단어를 정답으로 볼 수 있다. 이런 단어를 찾기 위해 보기에 주어진 네 개의 항목을 각각 빈칸에 대입 시키고, 그 단어와 주변의 단어들의 어울림 정도를 확인하면 보기의 단어들 중 정답을 찾을 수 있다. 이를 위해서 사전에 발생 가능한 표현 정보들을 사전에 구축해 둘 수도 있다. 하지만, 수많은 표현을 위한 단어의 조합을 사전에 구축하는 것은 쉽지 않은 작업이다. 단어들 사이의 어울림 정도를 찾는 또 다른 방법으로는 그 표현이 얼마나 많이 사용되는지에 대한 문제로 바꾸어 생각할 수 있다. 이를 위해 본 시스템에서는 해당 표현이 얼마나 많은 문서에 포함되는지를 웹 검색을 통해 대량의 문서에서 획득하는 방식을 사용한다

2. 인터넷 검색을 이용한 자동 영어 문제 풀이

2.1. 시스템의 전체 구조

시스템은 크게 문장 분할, 문제 구간 인식, 답안 추천의 세 단계로 구성된다. 첫 번째 단계는 입력 받은 토익 문장에서 문제가 되는 부분을 찾기 위해 해당 문장을 어절 별로 분할한다. 이 단계에서는 쉼표나 줄표('-') 등의 특수 기호와 공백을 기준으로 문장을 분할한다. 두 번째 단계는 분리된 단어들의 집합에서 해당 토익 문제의 빈칸과, 찾아진 빈 칸의 앞의 n 개의 단어와 뒤의 m 개의 단어를 찾아낸다. 세 번째 단계에서는 위에서 찾은 빈칸에 토익 문제의 네 가지 답안을, 추출한 앞의 n 개 단어와 뒤의 m 개의 단어와 함께 웹 검색 엔진에 질의한다. 네 개의 보기 단어에 대해 검색한 결과의 문서의 수를 비교해서 그 사용빈도가 가장 높은 표현을 모범 답안으로 한다. 다음 절에서는 추천 답안 결정에 대해서 상세히 설명한다.

(그림 1)은 프로그램의 화면을 보인다. 프로그램에서는 학습자에게 편의를 제공하기 위해 포털 사이트 다음[7]에서 제공하는 토익 문제와, Microsoft Word 나 한글과 컴퓨터의 한글로 작성된 문제 파일을 열어서 문제를 풀고 자동으로 추천된 정답과 비교할 수 있는 환경을 제공한다.



(그림 1) 시스템의 유저인터페이스

2.2. 접근 방법

Part 5 와 같은 단문 빈칸 채우기의 문제를 풀기 위해서는 빈칸을 채울 단어의 문법적인 특성이나 그 주변 단어와의 문맥 상의 의미를 이해해야 한다. 하지만, 실제로 그들간의 의미를 자동으로 파악해서 문제를 풀거나 문법적 의미를 자동으로 파악하는 것은 어렵다. 본 논문에서는 이런 방법을 대체하기 위해서 단문 빈칸 채우기의 문제가 사지 선다형의 문제라는 것에 착안하여 문제를 해결한다. 문제에서 제시된 네 개의 보기 중 정답 단어나 어구는 다른 보기보다 올바른 문법으로 작성되거나 보다 자연스러운 문맥을 구성할 것이다. 따라서, 정답은 다른 보기보다 일반 영어 문서에서 자주 나타날 것으로 짐작할 수 있다. 본 논문에서는 이 점에 착안하여 보기에 나오는 단어 중 문제의 문맥 단어와 동시에 발생하는 빈도가 높은 단어를 정답으로 추천하는 방식을 제안한다.

단어나 어구의 사용빈도에 관한 정보들을 획득하기 위해서 영어로 된 많은 문서들이 필요하다. 이를 위해서 수 많은 문서들이 존재하는 웹을 이용하면, 전 세계의 문서들을 이용할 수 있다. 네 개의 보기에 주어진 단어들을 빈칸에 대입한 표현을 웹 검색 엔진에 질의하면, 해당 표현이 나타난 문서 개수를 얻을 수 있다. 이 개수는 문제에 주어진 문맥과 보기에 나타난 단어의 공기빈도로 볼 수 있으므로 이를 이용하여 정답의 여부를 결정한다.

검색에는 일반 검색과 연접 검색의 두 가지 방법을 쓸 수 있다. [are of the]의 세 단어를 검색 창에 나열하는 일반 검색의 경우에는 세 개의 단어가 인접하지 않고 각각의 단어가 문서에 포함되기만 해도 검색 결과로 인정한다. 연접 검색은 검색어가 문서에 연접하여 나타난 경우만을 검색 결과로 인정하는 방법이다. 구글의 경우에는 [“are of the”]와 같이 검색 창에 큰 따옴표 사이에 검색하기를 원하는 표현을 넣은 방식으로 연접 검색 서비스가 제공된다. 두 방법 중 단문 빈칸 채우기와 같은 문제를 자동으로 해결하기 위해서는 연접 검색을 이용한 검색 방법을 이용하는 것이 적합하다.

(예제 1)의 한 토익 문제를 살펴보자. 이 문제는 단문 빈

> Advertising agencies and public relations firms are _____ the largest contractors for the services of professional printers.
> (a) of (b) some
> (c) many (d) among

(예제 1) 토익 예제

칸 채우기의 형태를 가진 토익 Part 5 의 한 문제이다. 본 논문에서는 전체 문장을 대상으로 하지 않고, 빈 칸을 중심으로 앞, 뒤 세 단어를 기준으로 빈 칸에 적합한 정답을 찾고자 한다. 이 때 빈 칸의 위치 t_i 라 했을 때 앞뒤 단어들은 아래와 같이 표현할 수 있다.

relations	firms	are	____	the	largest	contractors
t_{i-3}	t_{i-2}	t_{i-1}	t_i	t_{i+1}	t_{i+2}	t_{i+3}

여기서 앞쪽 문맥의 크기 m 과 뒤쪽 문맥의 크기 n 은 3 이 된다. t_i 의 위치에 들어갈 각 j 번째 보기 예제 단어를 t_{ij} 라 하자. 아래의 (표 1)은 t_{i-1}, t_{ij}, t_{i+1} 의 연접검색 결과의 개수를 보인다. 표에서 볼 수 있듯이 (d)가 다른 제시된 단어들보다 검색된 문서가 많으므로 이를 정답으로 볼 수 있다.

(단위 : 1,000)

보기 문항(j)	(a)	(b)	(c)	(d)
연접 검색 수	2,040	88	215	13,700

(표 1) 연접 검색 $f_q(t_{i-1}, t_{ij}, t_{i+1})$의 결과 개수

본 논문에서는 웹 검색의 결과 개수를 이용한 정답 추천의 세 가지 모델을 사용한다. 검색에서 t_i 의 위치에 들어갈 각 j 번째 보기 예제의 검색 결과를 $f(t_{ij})$라 하자. 검색 대상이 되는 앞, 뒤 n, m 크기의 문맥 $t_n, \ldots, t_{i-1}, t_i, t_{i+1}, \ldots, t_m$ 에 대한 일반 검색 결과의 개수를 $f_s(t_n, \ldots, t_{i-1}, t_{ij}, t_{i+1}, \ldots, t_m)$라 하고, 동일 문맥에 대한 연접 검색의 결과 개수를 $f_q(t_n, \ldots, t_{i-1}, t_{ij}, t_{i+1}, \ldots, t_m)$라 하자. **모델 1** 에서는 연접검색 결과의 개수 $f_q(t_n, \ldots, t_{i-1}, t_{ij}, t_{i+1}, \ldots, t_m)$의 값으로 가장 큰 값을 가지는 보기 단어를 정답 단어로 결정한다. **모델 2** 에서는 연접 검색 결과의 개수를 보기 단어의 검색 결과 개수로 나눈 $f_q(t_n, \ldots, t_{i-1}, t_{ij}, t_{i+1}, \ldots, t_m) / f(t_{ij})$를 기준으로 정답을 추천한다. **모델 3** 에서는 일반검색의 결과 개수를 단어의 검색 결과 개수로 나눈 $f_s(t_n, \ldots, t_{i-1}, t_{ij}, t_{i+1}, \ldots, t_m) / f(t_{ij})$를 기준으로 정답을 추천한다.

2.3. 문맥 크기

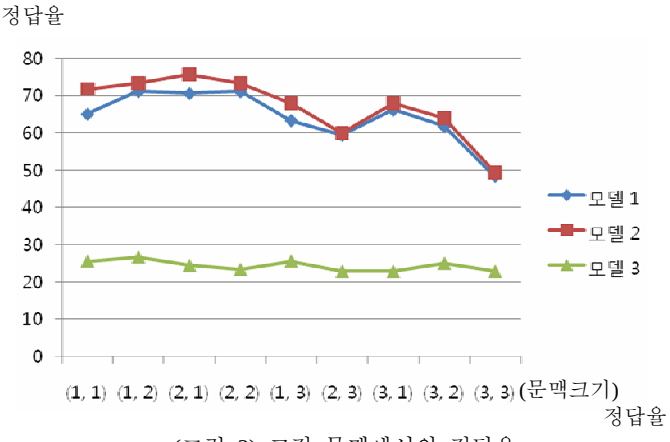
자동으로 정답을 확인하는 시스템에 있어서 정확도를 높이기 위한 또 하나의 고려사항은 문맥의 크기로 빈칸을 중심으로 얼마까지 범위를 잡아서 검색할 것인가에 따라 정답율이 달라질 수 있다. 빈칸을 중심으로 문맥의 크기가 크게 하는 경우, 검색되는 문서의수는 줄어들 것이다. 하지만, 이 경우 잘못된 표현의 경우에는 검색된 문서의 수가 없을 가능성이 높기 때문에, 추천된 정답의 정확율은 높아질 것이다. 반대로, 문맥의 크기가 작아지면, 각 표현에 대해서 검색의 결과는 많이 나올 것이다. 그러나, 이 경우에는 정답이 아닌 보기의 단어를 이용한 검색 결과에서도 그 수가 많이 검색될 것이고, 추천된 정답의 정확도는 떨어질 것이다.

본 논문에서는 이 두 측면을 고려해서 문맥의 크기를 변경시키면서 검색을 하는 유동 문맥의 방법을 사용한다. 검색을 시작할 때는 일정한 크기의 문맥으로부터 시작한다. 최초 크기의 문맥으로 검색한 결과가 모든 보기의 단어를 대입했을 때에도 검색이 되지 않으면, 앞쪽의 단어 하나를 빼서 문맥을 하나 줄이고 검색한다. 여기서도 모든 보기의 단어에 대해 검색 결과가 나오지 않는다면 뒤쪽의 단어를 빼고, 문맥의 크기를 또 하나 줄인다. 이렇게 앞, 뒤의 단어를 하나씩 줄여나가면서 정답이 추천될 때까지 검색을 이어나간다.

이런 방식의 유동 문맥을 사용해서 문맥의 크기가 클 때의 문제점과 문맥의 크기가 작을 때의 문제점을 해결함으로써 사용자에게 보다 높은 정확율을 가진 정답을 추천할 수 있다.

3. 실험

본 논문에서 제안한 방식을 실험하기 위해서 포털 사이트 다음에 게재된 토익 Part 5 문제를 이용했다[7]. 게재된 웹 문서를 분석해서 문제 부분만 추출해서 이용하였으며, 무작위로 추출한 180 개의 문제를 대상으로 실험을 수행하였다. 검색 엔진은 구글[8]을 사용했다.

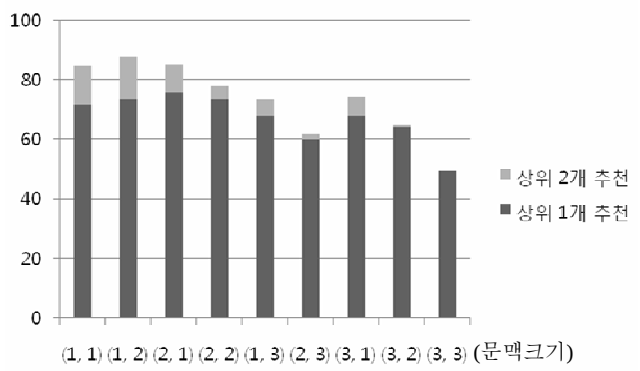


(그림 2) 고정 문맥에서의 정답율

(그림 2)에서는 본 논문에서 소개한 세 가지 방법의 모델을 이용하고, 각 모델에서 고정 문맥의 크기를 바꾸어 가며 실험을 한 결과를 보여준다. 이 그림에서 연접 검색 결과를 단어의 검색 결과로 나누는 **모델 2** 사용하면서, 빈칸을 중심으로 앞 두 단어, 뒤 두 단어로 실험을 한 결과 75.6%의 정답율로 가장 높았다.

자동 문제 추천에 의한 부정확한 정답 추천으로 인한 사용자의 불편을 감소시키기 위해 결과의 상위 두 개의 정답을 추천할 수 있다. 상위 한 개의 정답을 추천할 때의 정답율과 상위 두 개의 정답을 추천할 때의 정답율을 비교한

결과가 (그림 3)에 있다.



(그림 3) 연접검색의 추천 수에 따른 정답율

상위 한 개의 정답을 추천했을 때는 연접 단어 검색 결과를 단어의 검색 결과로 나눈 **모델 2** 일 때 정답율이 가장 높았다. 그러나 상위 2 개의 정답을 추천하는 경우에는 인접 단어의 검색 결과만 사용한 **모델 1** 일 때의 정답율이 가장 높았다. (그림 3)에는 **모델 1** 을 사용한 경우, 정답의 추천 수와 문맥의 크기에 따른 정답율을 보여준다. 그 결과 상위 한 단어를 추천한 경우에는 71.1%의 정답율을 보였으나, 두 단어를 추천한 경우, 문맥의 크기를 앞쪽으로 한 단어, 뒤쪽으로 2 단어를 이용했을 때 정답율이 87.8%까지 증가하는 것을 볼 수 있다.

(표 2)는 **모델 2** 를 사용했을 때 고정 문맥과 유동 문맥을 사용하고 상위 한 개의 정답을 추천한 경우에 정답율을 비교하고 있다. 유동 문맥의 경우 문맥의 크기는 앞, 뒤로 두 개의 단어씩을 사용하여 최초 검색을 실시했다. 고정 문맥의 경우에는 가장 높은 결과를 보인 앞, 뒤 두 단어씩을 사용한 결과이다. 이 표에 의하면, 유동 문맥을 사용한 경우가 고정 문맥을 사용해서 추천한 것보다 높은 정답율을 보인다. 그리고 (그림 3)에서 보인 **모델 1** 을 사용하여 두 개의 정답을 추천한 결과보다 유동 문맥으로 하나의 정답만 추천한 경우가 더 높은 정답율을 나타낸다는 것을 알 수 있다.

문맥 별	고정 문맥	유동 문맥
정답율	75.6%	89.9%

(표 2) 정답율 비교

4. 결론 및 향후 연구

본 논문에서는 단문 빈칸 채우기 형식의 영어 문제에 대한 정답을 웹에서의 검색 결과를 이용해서 자동으로 추천하는 방법을 제안하였다. 이를 위해 적절한 검색방법과 검색에 사용되는 단어의 수 등에 대해서 알아보았다. 문맥의 크기를 조절하면서 검색하는 방식을 사용하면 일정 수준 이상의 정답율을 나타낼 수 있다는 것을 보았다. 유동 문맥을 사용하여 하나의 정답을 추천하는 방식이 정답율이 가장 높았으며, 이런 방식에서 정답을 상위 두 개에 대해서 추천한다면 정답율은 더 높아질 것으로 보인다.

이런 방식으로 다른 영어 시험에 대해서 비슷한 유형의 문제에 대해서도 문제를 추천해 줄 수 있다. 또한 간단한 수준의 의미분석을 적용하여 지문의 내용을 자동으로 파악할 수 있다면 Part 6 나 Part 7 의 정답을 추천해 줄 수 있을 것으로 보인다. 또한 입력된 문장에 대해 특정 처리를 통해서 그 문장에 대한 문법적 구성을 파악하여, 단문 빈칸 채우기 형식의 문제를 자동으로 생성하는 연구도 진행 중이다.

참고문헌

[1] http://www.hackers.co.kr
[2] http://www.et-house.com/html/index.asp
[3] http://tomato.et-house.com
[4] 김지은, 이공주, “중학생 영작문 실력 향상을 위한 자동 문법 채점 시스템 구축,” 한국콘텐츠학회논문지 Vol. 7 No.5 2007.
[5] M.Bowden and R.A.Fox, “Diagnostic Approach to the Detection of Syntactic Errors in English for Non Ntive Speakers”, Technical Report 2002.
[6] J.Burstein and D.Higgins, “Advanced Capabilities for Evaluation Student Writing: Detection Off-Topic Essays Without Topic-Specific Trainin,” Proceedings of the International Conference on artificial Intelligence in Education, July 2005
[7] http://engdic.daum.net/dicen/toeic_question_part.do?q=018
[8] http://www.google.co.kr