웹기반 TOEIC 문법 문제 자동 생성 시스템

정형구*, 김상철*, 채희락**, 이찬종**
*한국외국어대학교 컴퓨터공학과
**한국외국어대학교 언어인지과학과
arocer@naver.com

A Web-Based System for Automatic Generation of TOEIC Grammar Test

Hyungku Chung*, Sangchul Kim*, Heerak Chae**, Chan Jong Lee**
*Dept of Computer Sci. & Eng., HUFS
*Dept of Linguistics & Cognitive Science., HUFS

요 약

본 논문에서는 웹기반의 TOEIC 문법 문제 자동 생성 시스템을 제안한다. 본 시스템은 웹에서 문서를 가져온 후, 문서의 각 문장이 데이터베이스에 저장된 문제 패턴에 일치하는 지를 검사하여 문제를 생성한다. 본 시스템을 통해서 생성되는 문법 문제를 영어 전문가를 통해서 검증한 결과, 대부분의 문제가 TOEIC 실전 문제로 사용하기에 충분한 것이었다. 우리의 조사에 의하면, TOEIC 문법 문제의 자동 생성에 관한 기존 연구는 거의 발표되고 있지 않다

TOEIC, Grammer, Automatic Geneation a

1. 서론

영어 실력은 유학 및 취업 준비생이 필수적으로 갖추어야 하는 능력이다. 특히 국내 기업들은 사원들의 영어 능력 검증을 위해서 TOEIC 같은 공인 테스트를 이용하고 있다. 그런 이유에서 한해 200만명이 넘는 사람이 TOEIC 테스트에 응시하고 있는 실정이다. 그런 이유에서 시중에는 많은 종류의 TOEIC 테스트용 서적[1,2]들이 출판되어 있다.

최근 TOEIC 테스트는 크게 독해(Reading), 청취(Listening), 작문(Writing), 말하기 (Speaking)로 구성되어 있다. 독해 테스트의 중요한 부분 중 하나는 문법 (Syntax) 테스트이다. 영어 문법을 숙달하기 위해서는 이론 학습만으로는 부족하여 많은 유형의 문제들을 풀어보아야 한다. 그래서 시중의 TOEIC 문법 테스트용 책[2]에는 설전 문법 문제들이 많이 수록되어 있다. 하지만 책 한권에 수록된 문제는 지면의 한계로 인해서 수험생이 만족할 만한 정도라고 보기 어렵다.

본 논문은 웹 검색 기반의 영어 문법 문제 자동 생성 시스템을 제안한다. 본 시스템의 주된 기능은 웹상에서 문서를 가져오고, 그 문서에 수록된 문장들을 이용해서 TOEIC 문법 문제를 생성하는 것이다. 본 시스템은 TOEIC 테스트 준비생이 원하는 유형의 문법 문제를 무한정 풀어 볼 수 있는 기회를 제공할 뿐만이 아니라, TOEIC 문제지를 제작하는 업체[3]의 문제 출제 경비를 대폭 줄여 줄 것이다.

기존 연구에서 영어 단어 테스트나 문법 문제를 자동으로 생성하는 방법[4,5]은 발견할 수 있다. 우리의 조사에 의하면, TOEIC 문법 문제의 자동 생성에 관한 기존 연구는 거의 발표되고 있지 않다.

2. TOEIC 영어 문법문제 유형 분류

영어 문법문제를 자동으로 생성할 수 있는 소프트웨어를 개발하기 위해서는 먼저 문법문제의 유형을 분류하고 그 유형에 따른 데이터베이스를 구축하는 것이 필요하다. 여기서는 TOEIC 문법문제를 유형별로 분류하면서 간단한 설명을 덧붙이고 필요시 예문을 제시한다. TOEIC 문법 문제는 동사, 명사, 수

식어(형용사, 부사), 연결 장치 (접속사, 전치사) 관련 문제로 크게 나눌 수 있다.

2.1 동사 관련 문제

(1) 주어-동사의 수일치
(2) 수동태: 타동사 뒤에 목적어가 없으면 수동태 사용
(3) 시제의 선택: 사용하는 부사에 맞추어서 동사의 시제를 선택하는 문제이다.

ex) always, usually가 사용되는 경우에는 현재 시제를 사용하고, last, ago의 경우에는 과거 시제를 사용하는 문제

2.2. 명사 관련 문제

(1) 적절한 위치에서의 사용: 주어, 목적어, 전치사, 관사/소유격 뒤에 사용을 테스트한다. 관사/소유격 문제의 예는 다음과 같다:

ex) *Thank you for your interest in our new _____*
(a) product (b) products (c) of products (d) products'

(2) 대명사: 대명사의 수일치, 격일치, 부정대명사 선택. 예를 들면, 부정대명사의 선택은 둘일 때는 one과 the other를 사용하고, 셋일 때는 one, another 및 the other를 사용한다.

(3) 재귀 대명사의 사용
(4) 소유 대명사의 사용
(5) 가주어와 가목적어 사용
(6) 단수 명사 that와 복수 명사 those의 사용

ex.) Sales in this quarter are similar to ____ of previous quarter.
(a) that (b) those (c) these (d) ones

2.3 수식어 (형용사, 부사) 관련 문제

(1) 형용사 사용: 명사 수식 형용사, 보어로서 형용사
(2) 부사 사용
(3) 수량 형용사 사용

ex) many/ a number of, several/ a few, few, other, these/those 뒤에 복수명사가 오는 지를 묻는 문제.

(4) 비교급, 최상급 사용

2.4. 연결 장치 (접속사, 전치사) 관련 문제

(1) 접속사와 전치사의 구별:

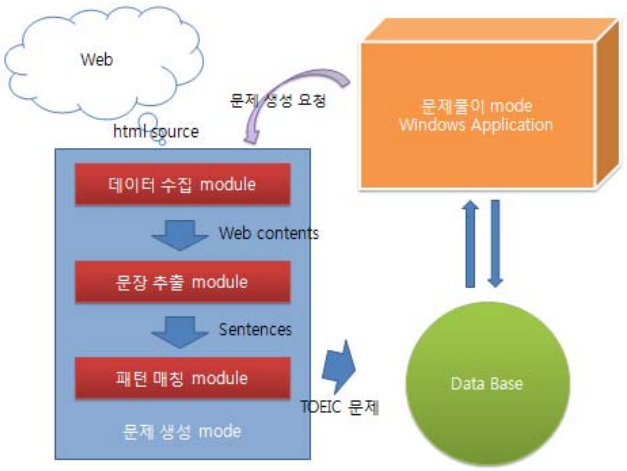
ex) [접속사 + 주어 + 동사] 구조와 [전치사 + 명사] 구조에 대한 올바른 사용을 묻는 문제

(2) 등위 접속사와 상관 접속사 사용
(3) 관계 대명사 사용
(4) 명사절의 that/what 사용
(5) 가정법 if 사용
(6) 도치 구조

3. 웹 기반 문법문제 자동 생성 엔진

3.1 전반적인 동작

본 시스템은 TOEIC에 사용할 수 있는 문제를 생성하여 데이터베이스(DB)에 저장하는 문제 생성 모드)와 DB에 축적된 문제들을 풀이할 수 있는 문제 풀이 모드로 크게 나뉘어 구성되어 있다.



위 그림에서와 같이 문제 생성 모드에서 사용자는 문제 추출하기에 적합한 웹페이지를 자신이 취약한 문제 유형으로 설정한 후 그 주소를 입력한다. 이때 데이터 수집 모듈이 입력된 웹페이지 주소에 해당하는 html 소스를 가져와 html 태그를 제거한다. html 태그가 제거가 되어도 사용자가 웹에서 본 텍스트 문장이 아니므로 문장 추출 모듈로 완전한 텍스트 데이터만을 추출한다. 추출된 텍스트 데이터를 패턴 매칭 모듈에서 사용자가 설정한 토익 문제 패턴과 유사여부를 파악하여 이에 합당한 텍스트를 문제 유형으로 변환하여 데이터베이스에 저장한다.

문제풀이 모드에서 사용자는 취약한 문제 유형을 데이터베이스에서 불러와 이를 풀어봄으로써 자신의 취약한 유형을 집중적으로 공부하여 토익 실력을 향상 시킬 수 있다.

3.2. 데이터 수집 모듈

TOEIC 문제로 사용할 재료를 선정하기 위하여 고려해야 할 사항들이 있다. 우선 방대한 웹을 대상으로 소프트웨어가 지능적으로 정보를 가져오는 것은 어려운 일이다. 따라서 우리는 위키피디아[6] 백과 사전을 이용한다. 사전에 등재하기 위한 문체들로 구어체나 인터넷 문체가 아닌 TOEIC 문제에 적절한 형태의 문장이 풍부하다. 수백만의 콘텐츠는 중복 검색을 쉽게 피할 수 있고 그 분야 또한 무궁무진하다. 사용자는 백과사전에 등재되어 있을 법한 단어만 입력하면 데이터 수집 모듈이 자동으로 위키피디아의 해당 페이지를 직접 접근한다.

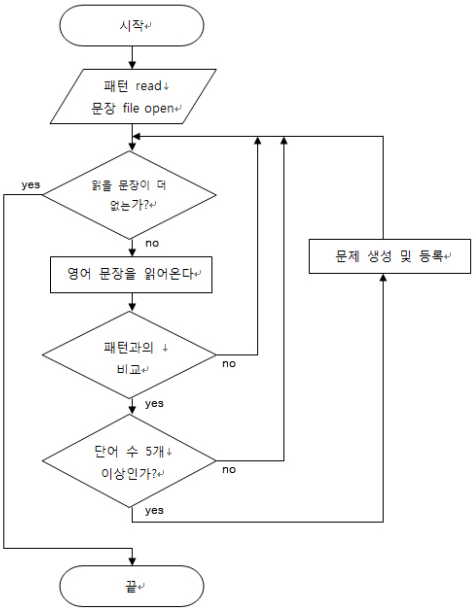
3.3. 문장 추출 모듈

데이터 수집 모듈에서 얻은 데이터는 바로 쓸 수 없는 순수 html 형태의 원시 데이터이기 때문에 문장 추출 모듈에서는 먼저 html 태그를 제거하고, 다음으로 얻어진 데이터를 가비지(garbage)인지 단순 단어인지 혹은 TOEIC 문제로 활용할 수 있는 영어 문장인지 선별하는 두 가지의 작업을 거친다.

html 태그를 완벽하게 제거하기 위해서는 파싱 작업이 불가피하고, Java-script, 공백 문자 기호, 이중 태그 등 태그의 규칙성 예외 등은 정규식이나 간단한 알고리즘으로는 처리가 불가능한 것으로 알려져 있다. 이러한 작업을 모두 처리하는 것은 상용화 소프트웨어에서도 불가능하다. 따라서 가비지가 존재하고 손실되는 문장이 발생하더라도 어느 정도의 온전한 문장을 보전하는 알고리즘과 코드를 구현하여 개발하였다. TOEIC에서 주로 사용되는 단어만으로 구성된 문장을 최종 선택한다.

3.4. 패턴 매칭을 통한 TOEIC 문제화 모듈

패턴 매칭 모듈은 위의 과정을 통하여 얻어진 문장들을 일반적인 TOEIC 문제의 유형으로 바꾸는 역할을 한다. TOEIC 문장을 생성하기 위해서는 얻어진 문장을 영어 구문분석기를 통하여 분석하고 이 정보를 토대로 실제 사람의 손으로 TOEIC 문제를 생성하는 작업을 모델링하여 알고리즘화 해야 한다. 이렇게 얻어진 패턴을 위의 과정을 통해 나온 문장들과 비교해서 문제화 시키는 작업을 담당한다. DB에는 미리 준비된 문제의 구체적인 유형이 존재하고 보기 항목과 정답도 DB에 미리 입력된 상태에서 이에 부합하는 문장을 검색하여 문제화 하고 이를 문제 테이블에 삽입한다.



4. 결론

본 논문에서는 웹기반의 TOEIC 문법 문제 자동생성 시스템을 제안하였다. 본 시스템은 TOEIC 문법 문제 유형의 패턴들을 데이터베이스에 저장하고 있다. 문제 생성 과정은 웹에서 문서를 가져온 후, 문서의 각 문장이 데이터베이스에 저장된 패턴에 일치하는 지를 검사하면 된다.

프로그래밍 도구로서 VC++를 사용 하고, MS Access를 사용하여 개발하였다. 본 시스템을 통해서 생성되는 문법 문제를 영어 전문가를 통해서 검증한 결과, 대부분의 문제가 TOEIC 실전 문제로 사용하기에 충분한 것이었다.

참고문헌

[1] 서경주, 매운 TOEIC, BCM 미디어, 2007.

[2] David Cho, Hackers Grammar Start, 해커스어학연구소, 2004.

[3] 내부보고서, (주) 에듀소프트, 2007

[4] C.-Y. Chen, et al, “FAST: an automatic generation system for grammar tests,” Proceedings of the COLING/ACL on Interactive presentation sessions, 2006

[5] J. C. Brown, et al, "Automatic question generation for vocabulary assessment," Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing HLT '05

[6] www.wikipedia.com