빅데이터를 활용한 영어학습 애플리케이션 설계 및 구현

이재훈, 김승범, 김창영, 양원석, 김도우
한국폴리텍대학 서울 강서캠퍼스 데이터분석과
alsltnpf1209@gmail.com, mar5924@naver.com, newshfkfk@gmail.com
yang991211@naver.com, dowoo2594@naver.com

English Learning Applications Using Big Data Development

Jae-hoon Lee, Seung-beom Kim, Chang-young Kim,
Won-seok Yang, Do-woo Kim
*Dept. of Data Analysis, Korea Polytechnic of Seoul Gangseo Campus

요 약 최근 교육분야에서는 IT 기술을 활용하여 교육을 혁신하는 것을 의미하는 에듀테크에 대한 관심이 높아지고 있다. 단순한 지식의 전달이 아닌 사용자의 수준에 맞춰진 학습을 하고 자신의 학습 내용을 스스로 모니터링할 수 있는 새로운 교육시스템이 필요하다.

이에 본 논문에서는 빅데이터를 활용한 영어학습 애플리케이션를 제안한다. 제안하는 애플리케이션은 영어뉴스 기사에서 추출한 빅데이터를 활용하여 사용자 수준에 맞춘 유용한 문장을 분석해 자동으로 문제를 생성하고 사용자의 음성데이터를 강세 분석 알고리즘으로 원어민 발음과 비교분석 하여 발음 및 강세를 교정할 수 있도록 설계 및 구현하였다.

1. 서론

최근 4차 산업혁명 시대에 맞춰 빅데이터의 수집 및 처리에 대한 활용의 중요성이 대두되고 있다. 이에 따라 교육 분야에서의 빅데이터 활용 사례 및 연구가 활발히 진행되고 있으며, 데이터와 소프트웨어를 융합한 에듀테크 산업에 관한 관심 또한, 높아지고 있다[1].

그러나 기존의 영어학습 애플리케이션들은 일상 영어회화 학습에 치중되어 있고 사용자의 영어 발음을 텍스트로 비교하여 분석하는 플랫폼이 대부분이다.

이에 본 논문은 영어뉴스 기사를 통해 시사와 영어를 동시에 학습하고 완전자동화를 목표로 한 영어 강세 교정 애플리케이션을 제안한다. 본 애플리케이션은 Java 기반의 Jsoup 웹크롤링 라이브러리를 활용하여 영어뉴스 기사의 데이터를 수집 및 자연어처리(Natural Language Processing, NLP)하여 학습 자료를 사용자에 제공한다. 또한, 음정의 높낮이, 소리의 세기를 음정 분석 알고리즘을 통해 원어민 발음과 비교하여 사용자에 제공한다.

논문의 구성은 2장에서 관련 연구로 기존의 영어 스마트러닝의 실태 및 요구분석과 제안하는 애플리케이션에서 사용될 웹 크롤러, 자연어처리, 웨이브넷 기반의 음성 합성기술에 관하여 기술한다. 3장에서는 제안하는 애플리케이션의 기술 구성 및 흐름도와 데이터베이스 설계, 빅데이터분석 및 처리, 음성분석 알고리즘에 대하여 설명한다. 4장에서는 프로젝트 구현을 설명한다. 마지막으로 5장에서는 본 논문에 대한 결론을 제시한다.

2. 관련 연구

2.1 기존 영어학습 플랫폼 실태 및 요구분석

스마트 영어학습 이용실태 보고서[2]에 따르면, 성인학습자의 77%는 시·공간에 구애받지 않고 자투리 시간을 활용하는 플랫폼을 요구하고 있는 것으로 나타났다. 또한, 기존의 영어학습 플랫폼은 개인적인 수준에 맞는 영어학습 부족, 학습 시 상호작용 부족을 단점으로 지적했다. 응답자는 영어학습 플랫폼을 하루 30분 이하의 자투리 시간과 출퇴근·통학 등의 이동시간에 학습하였으며 스마트러닝 영어학습 시 듣기, 말하기 등 직접 참여하는 것을 선호하는 것으로 나타났다. 이는 사용자의 요구를 충족시킬 새로운 영어학습 시스템의 필요성을 시사한다.

2.2 웹 크롤러(Web Crawler)

웹 크롤러는 방대한 웹 문서를 제공하는 웹에서 특정 사이트의 웹 문서를 자동으로 수집하는 기술을 말한다. 특히, 빅데이터의 활용이 다양한 분야로 점

차 확산되고 웹 데이터가 매년 기하급수적으로 증가하면서 웹 크롤러의 중요성은 더욱 커지고 있다[3].

2.3 자연어처리

자연어처리는 자연어와 컴퓨터 간의 원활한 상호작용을 위해 기계학습(Machine Learning)을 통해 컴퓨터가 인간의 언어를 이해할 수 있도록 하는 작업이다. 효과적인 자연어처리를 위해서는 단어의 의미 및 문법적 속성, 문법 규칙과 같은 리소스와 어휘집(Lexicon), 온톨로지(Ontology) 등과 같은 다양한 지식 표현기술을 사용하게 된다[4].

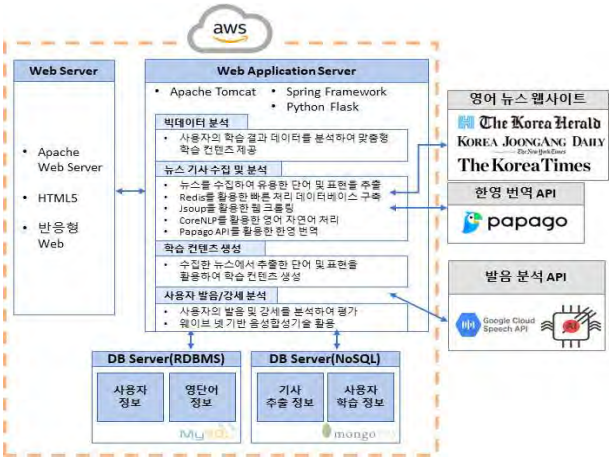
2.4 웨이브넷(WaveNet) 기반의 음성합성기술

웨이브넷은 원시 오디오를 생성하기 위한 심층 신경망이다. 이 기술은 실제 음성 녹음으로 훈련된 신경망 방법을 사용하여 파형을 직접 모델링 하여 상대적으로 사실적인 사람과 같은 음성을 생성할 수 있으며, 원시 파형을 생성하는 웨이브넷의 기능은 음악을 포함한 모든 종류의 오디오를 모델링 할 수 있음을 의미한다[5].

3. 빅데이터를 활용한 영어학습 애플리케이션

본 장에서는 빅데이터를 활용한 영어학습 애플리케이션에 대하여 설명한다.

3.1 시스템 구성도



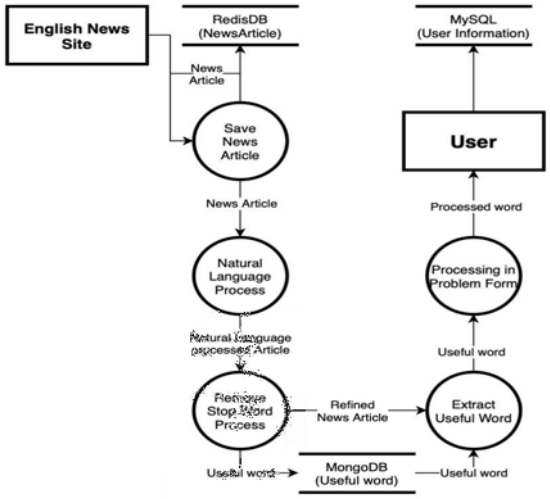
[그림 1] 빅데이터 영어학습 시스템 구성도

본 시스템의 구성은 크게 Web Server, Web Application Server와 DB Server 영역으로 구성되어 있다. Web Server는 사용자가 접근이 용이하도록 Apache Web Server로 사용하였고 PC, 스마트폰 등의 다양한 기기에서 이용할 수 있도록 HTML5 기반 반응형 웹으로 개발하였다.

Web Application Server는 영어뉴스 웹사이트에서 뉴스를 수집하는 수집영역, Core NLP를 활용한 영어 자연어처리 영역, 사용자의 학습 결과 데이터를 분석하여 맞춤형 학습콘텐츠를 제공하는 빅데이터 분석영역, 맞춤단어를 분석하고 학습콘텐츠를 생성하는 문제생성 영역, 사용자의 발음 및 강세를 분석하는 음성분석 영역으로 나누어져 있다.

DB Server는 웹 크롤링한 뉴스의 원문을 저장하기 위해 비관계형 데이터베이스 관리시스템인 Redis로 접근속도와 데이터 과부화를 방지하였고, 자연어처리된 비정형 데이터를 관리하기 위해 도큐먼트지향 데이터베이스 시스템인 MongoDB를 이용하였으며, 빅데이터 분석에 활용할 학습데이터 및 사용자 정보는 관계형 데이터베이스 관리시스템(RDBMS)인 MySQL을 사용하여 시스템을 구성하였다.

3.2 시스템 흐름도



[그림 2] 데이터의 흐름 및 절차도

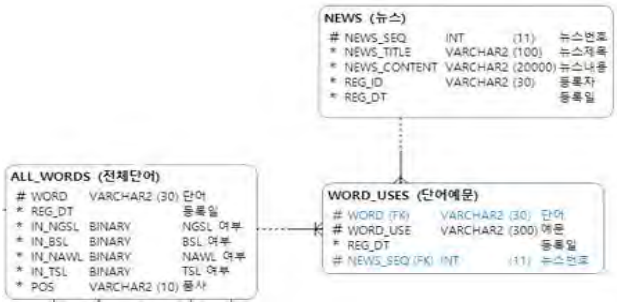
제안하는 애플리케이션의 시스템 흐름은 영어뉴스 사이트의 헤드라인기사 텍스트를 크롤링하여 “Key-Value” 구조의 비정형 데이터를 저장하고 관리하기 위한 오픈소스 기반의 Redis에 저장하여 사용자가 필요로 하는 뉴스기사의 원문을 제공한다.

MongoDB는 뉴스 원문을 자연어 처리하여 문장, 어휘의 원형을 저장한다. 문장의 유용성은 빅데이터 간 단어의 일치성을 비교 분석하여, 최다 누적된 단어가 포함된 문장을 식별하고 문제를 생성, 사용자에 제공한다.

발음 강세 및 교정 기능은 음성 녹음을 통해 사용자의 음성데이터를 전달받으면, 영어 강세 분석 알고리즘으로 분석하여 그래프의 형태로 재학습에 도움을 준다. 사용자가 학습한 데이터는 MySQL에 저장되며, 복습에 용이한 단어카드의 형태로 사용자에 제공한다.

3.3 데이터베이스 설계

제안하는 애플리케이션에서의 영어학습을 구현하기 위해서는 관심분야 DB와 뉴스DB는 필수적 요소이다. 관심분야DB는 일상(NAWL), 학술/논문(NGSL), 비즈니스(BSL), 토익(TSL) 총 4개의 테이블로 구성되어 있다. 뉴스DB는 매일 메인 뉴스로부터 추출되는 기사 제목, 내용이 저장되는 테이블 1개로 구성되어 있다.



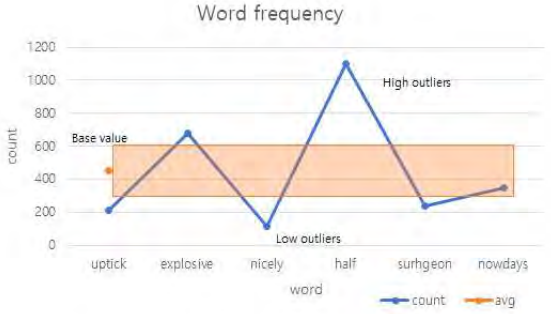
[그림 3] 엔티티 관계도

[그림 3]은 뉴스DB에 저장된 기사 내용 중 사용자가 선택한 관심분야 DB가 모인 ALL_WORDS(전체단어)테이블의 단어를 포함하고 있는 문장을 뉴스기사의 핵심 문장이 되면서 WORD_USES(단어예문)에 저장된다.

3.4 빅데이터 분석 및 처리

빅데이터 분석영역은 뉴스 빅데이터에서 중요어휘를 분석하여 사용자의 관심분야 별 문제생성에 이용한다. 저장된 ALL_WORDS(주요단어모음) 데이터는 각 단어의 속성에 일상, 학술/논문, 비즈니스, 토익을 배열로 저장할 수 있도록 하였다.

이를 활용하여 ALL_WORDS에 저장된 주요어휘가 당일의 뉴스문장에 포함되어있는지 확인하고, 포함이 되어있으면 문제로 만들어 학습 데이터베이스에 저장되어, 사용자는 자신의 관심분야 별로 맞춤형 문제를 제공 받을 수 있도록 한다.



[그림 4] 어휘 중요도 분석

[그림 4] 와 같이 ALL_WORDS에 저장된 각 단어들은 본 애플리케이션에 수집된 모든 뉴스 데이터와 비교하고, 사용빈도를 확인하여 중요도를 분석한다. 이는 사용자가 문제를 풀었을 때의 정답률과 오답률로 사용자의 레벨을 부여하여 word의 빈출 아웃라이어 (Highliers는 상, Basevalue는 중, Lowliers이면 하)에 따라 사용자에 맞게 제공할 수 있도록 데이터분석 알고리즘을 추가하였다.

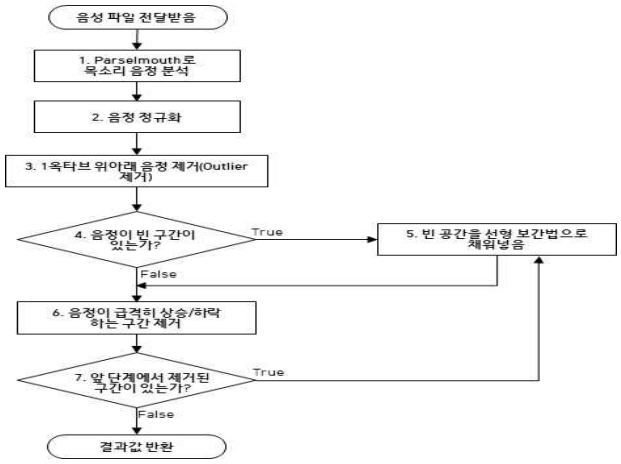
3.5 음성분석 알고리즘

음정 분석 및 가공 알고리즘은 사용자로부터 받은 음성데이터의 음정을 분석하고, 분석결과를 강세분석에 용이하게 가공하는 알고리즘이다. 분석을 위해 사용되는 Parselmouth 라이브러리는 모음(Vowel)에 대한 음높이만을 인식하여 분석하도록 설계되었지만, 간헐적으로 치찰음을 모음으로 잘못 인식하여 비정상적으로 높은 수치의 아웃라이어가 포함된 음높이 결과를 출력한다.

[그림 5] 아웃라이어 제거

아웃라이어가 포함된 분석결과는 추후 두 음성간의 비교분석 결과에 큰 오차를 발생시킬 수 있기에 이에 대한 처리가 필요하다.

이를 위해 [그림 5]와 같이 순간적으로 급상승 혹은 급 하락하는 음높이와, 음높이의 중앙값에 비해 2배 이상(1옥타브 이상), 1/2배 이하(1옥타브 이하)의 음높이는 아웃라이어로 판단하여 제거 한다



[그림 6] 음성분석 알고리즘 순서도

음정 분석 및 가공 알고리즘의 수행과정은 크게 3단계로 분석, 정규화, 보간 으로 이루어진다.

[그림 6]과 같이 목소리의 음정을 Parselmouth로

분석하고, 분석을 통해 얻은 음정을 중앙값으로 나누어, 중앙값을 기준으로 상대적인 음정(1옥타브 = 2의 제곱)이 되도록 정규화하고 1옥타브 위아래의 음정을 제거한다. (2초과, 0.5미만) 음정이 인식되지 않았거나 지난 처리로 인해 빈 구간이 있는지 확인하여 위에 서술한 조건이 맞는 경우 선형 보간법으로 빈 구간을 채워 넣는다.

또한, 음정이 급격히 상승하거나 하락하는 구간을 제거하고 제거된 구간이 있을 경우 선형 보간법으로 다시 빈 구간을 채워 넣는다. 해당하는 구간이 없는 경우 결과 값을 반환하여 두 음성 간의 비교분석 결과가 가장 정확하게 나올 수 있도록 음성데이터를 정제한다.

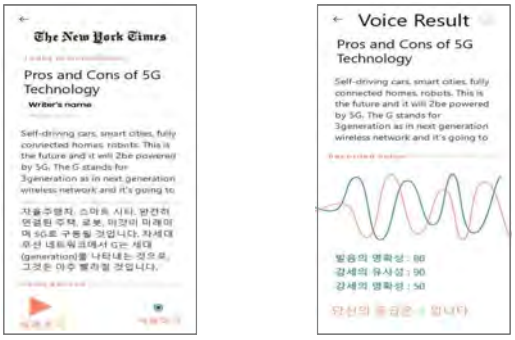
4. 프로젝트 구현

(a) 메인화면 (b) 오늘의 문장 (c) 문제풀이

[그림 7] 제안하는 시스템의 주요화면

[그림 7-a]는 웹 크롤러를 활용하여 뉴스 기사를 가져온 후 사용자가 읽기 편하도록 자연어 처리하여 문장으로 분절한다. [그림 7-b]는 분절한 문장을 '오늘의 문장'이라는 이름으로 사용자에게 보여준다.

[그림 7-c]는 '오늘의 문장'에서 중요한 단어를 빈칸 넣기 형태로 만들어 사용자에게 제공하여 단순히 기사를 읽는 것뿐 아니라 직접 입력하게 하여 한 번 더 기억할 수 있게 학습 방향을 설정하였다.



(a) 녹음화면 (b) 녹음분석

[그림 8] 제안하는 시스템의 주요화면

[그림 8-a]는 '오늘의 문장'의 발음을 들어보고 직접 녹음한다. [그림 8-b]는 사용자의 음성데이터를 음성분석 알고리즘을 활용하여 정제한 후 원어민의 발음과 강세를 비교하여 더욱 정확한 영어 발음을 구사하고 교정할 수 있도록 하였다.

5. 결론

본 논문에서는 빅데이터를 활용한 영어학습 애플리케이션을 제안하였다. 영어학습 콘텐츠의 다양성을 위해 시사영어 학습데이터를 제공하는 플랫폼을 제시하였고 기존의 콘텐츠 제작비용과 관리비용을 줄인 자연어처리를 이용한 완전 자동학습 콘텐츠 생성을 구현하였다. 영어 발음뿐만 아니라 기존에 없는 영어 강세에 대한 교정 방향을 제시함으로써 기존의 영어학습 애플리케이션과의 차별 점을 두어 사용자의 선택 폭을 넓혔다.

지속적으로 축적되는 학습데이터는 향후 사용자에게 빅데이터를 활용한 맞춤형 콘텐츠 생성이 가능하다. 급변하는 스마트시대에서 현재와 미래를 예측하는 대규모 데이터를 이용해 사용자의 영어 능력을 향상시킬 수 있는 양질의 학습효과를 기대한다.

사사

'본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트의 결과물입니다.'

REFERENCES

[1] Youngseok Lee, Jungwon Cho Study on Educational Utilization Methods of Big Data. Journal of the Korea Academia-Industrial cooperation Society Vol. 17, No. 12 pp. 716-722, 2016

[2] Soeun Lee, "Current States and Needs of Smart Learning: Focused on Adult L2 Learner" Master's Thesis, Hanyang University, 2018

[3] Dongmin Seo, Hanmin Jung, "Intelligent Web Crawler for Supporting Big Data Analysis Services" Journal of Academic Contents Society, Vol. 13 No. 12, 2013

[4] Ji Hee Lee, "Using unstructured text data within the construction industry Global Research Trends Based on Natural Language Processing (NLP)" Construction engineering and management v.20 no.2 2019

[5] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi "Speaker-dependent WaveNet vocoder" 2017