인공지능 자연어처리를 활용한  
맞춤형 영어 학습자료 생성

**1. 프로젝트 개요**

최근 영어학습을 위한 다양한 강의, 앱 등이 출시되고 있으며, 중학생, 고등학생 등 입시를 위한 영어뿐만 아니라 직장인을 위한 영어회화 등 영어학습의 수요가 다변화되는 추세입니다. 국내 교육 기관의 연구 보고서들을 살펴보면 디지털 기술을 기반으로 한 개인 맞춤형 학습이 미래 교육의 새로운 패러다임이 될 것이라 보고한 바가 있기도 합니다. 드라마나 영화 등 콘텐츠를 이해하고 즐기기 위해서 영어를 배우려는 사람들도 증가하고 있습니다. 그러나 현재 많은 영어학습 자료들은 실제 사람들이 즐기는 콘텐츠가 아닌, 다소 딱딱한 예시나 지루한 이야기가 중심이어서, 학습에 의욕을 느끼기 어려운 상황입니다.

저희는 영어를 비롯한 외국어 학습에서 사람들이 가장 어려움을 느끼는 부분이 의욕, 동기 부족이라는 점에서, 자신이 원하는 콘텐츠로 학습을 할 수 있다면 자연스럽게 흥미를 가지고 지속적으로 학습을 할 수 있다고 생각했습니다. 영상, 소설의 텍스트를 분석해서, 단어나 문장을 인공지능 자연어처리 기술로 추출한다면 단어 학습, 문장 학습, 받아쓰기 등을 진행할 수 있는 학습자료를 생성할 수 있다면, 누구나 원하는 콘텐츠로 학습을 하는 맞춤형 교육을 실현할 수 있다고 생각했습니다.

**2. 프로젝트 목적**

시장에서 개인 맞춤형 교육에 대한 수요가 높은 데 반해 디지털 기술을 기반으로 학습자 개인에게 맞는 자료를 생성하여 교육하는 서비스가 없었기 때문에, 맞춤형 영어 학습자료 생성 서비스를 개발하게 되었습니다. 온라인 학습 플랫폼을 통해, 단순히 교사의 지시를 따르는 것이 아니라 학생 스스로가 주체적인 의사 결정자가 되어 본인의 흥미와 필요에 맞는 교육 과정을 생성, 학습, 평가하는 새로운 형태의 교육 서비스를 개발하고자 했습니다.

**3. 프로젝트 내용**

3-1. 데이터의 정의

개발에 앞서 학습자료란 무엇인가에 대한 정의가 필요했습니다. 읽기, 쓰기, 듣기, 말하기라는 4개 유형의 학습을 위해 필요한 요소들이 무엇인지 분석한 결과, 다음과 같은 요소가 필수적이라는 것을 알아냈습니다.

- 단어: 단어의 철자, 뜻  
- 문장: 문장의 텍스트, 해석  
- 음성: 단어, 문장의 발음

따라서 학습자료의 생성을 위해서는, 임의의 텍스트에서 단어와 문장을 파악하는 것이 우선이고, 이후 각 단어와 문장의 뜻, 해석, 발음을 매칭시킬 필요가 있었습니다.

3-2. 데이터 전처리

우선 사용자가 입력하는 텍스트를 컴퓨터가 가공하기 좋도록 전처리가 필요합니다. 특히 본 프로그램은 영어 텍스트, 한국어 텍스트 등 다국어 텍스트가 입력될 수 있기 때문에 다국어 텍스트에 대해 일관성 있게 접근하고 동일하게 처리하기 위하여 입력되는 텍스트를 모두 UTF-8 인코딩 방식으로 변환했습니다. 또한 공백 문자가 중복되어 있을 경우에 이를 처리했습니다.

사람은 텍스트에서 무엇이 문장이고 무엇이 단어인지 파악할 수 있지만, 컴퓨터는 전부 똑 같은 문자열로 인식하기 때문에 문장 경계와 단어 경계를 알려주어야 합니다. 특히 문장 경계의 경우, 교육자가 실제 문장의 단위와는 달리 교육의 편의를 위해 문장을 나누어 놓은 경우도 있기 때문에 통상적인 문장 경계 규칙만을 적용해서는 안 되는 문제가 있었습니다. 이를 해결하기 위해 교육자가 텍스트를 입력할 때 나누어 놓은 문장 경계를 우선 순위로 하여 문장 단위를 나누어 저장한 후, 그 다음 문미 기호 등을 기준으로 문장 경계를 구분하였습니다.

3-2. 데이터 전처리

단어의 경우 언어마다 단어의 경계를 판별하는 것이 서로 다르기 때문에 학습 코퍼스 기반 기계학습 방법으로 적용하여 각 언어마다 단어 경계를 결정할 수 있도록 하였습니다. 이를 위해 각 언어마다 단어 및 형태소 단위로 구축된 학습 데이터를 사용하였습니다. 의존트리뱅크(Stanford Dependency Treebank)를 학습 데이터로 사용하여 모델의 분류 성능을 높이고 또한 학습 단계에 맞는 가중치 벡터를 학습시킴으로써 구문 분석 시 구조화된 분류가 가능했습니다.

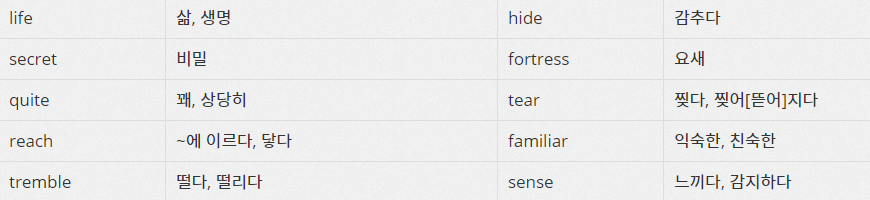
영어 교육 현장에서 문법 교육이 중요하게 다뤄지기 때문에, 단어와 해당 단어가 문장 내에서 어떠한 품사로 사용되는지도 함께 제시할 필요가 있었습니다. 단어와 해당 단어가 입력되어 있는 문장이 서로 연결되어 있는 모델을 구축하여, 이러한 모델을 통해 특정 단어가 자신이 속한 문장 내부에서 어떠한 품사 역할을 하는지 결정할 수 있습니다.

단어와 문장에 대해 해석과 발음 데이터를 매칭시키기 위해서 Google Cloud Platform의 TTS 및 STT 서비스를 활용했습니다.



[그림1] 프로젝트 처리 흐름

3-4. 주요성과



[그림2] 실제 유튜브 영상의 내용에서 단어를 추출한 예시



[그림3] 실제 유튜브 영상의 내용에서 문장을 추출하고, 학습을 위해 빈 칸을 뚫은 예시

유튜브 영상을 비롯한 임의의 콘텐츠에 포함된 텍스트에서 단어, 문장 등을 추출하고, 이를 학습에 활용할 수 있는 자료의 형태로 생성하는 것을 위의 이미지와 같이 수행할 수 있었습니다.

3분 길이의 영상에서 단어, 문장을 추출해 위와 같은 형태의 자료로 만드는 것이 기존에 인력을 활용한 방식으로는 3시간 이상이 소요된 반면, 이번 프로젝트로 개발한 모델을 사용한 결과 순수한 모델 작동에는 5초 이내, 이외 준비 작업 및 정리 등을 포함해 7분이 소요됐습니다.

3-5결론 (발전방향)

현재 다양한 자연어처리 기술과 서비스가 등장하고 있으며, 모든 요소를 직접 구현하는 것뿐만 아니라 이미 존재하는 최신의 서비스를 Google Cloud Platform 등에서 가져와 통합하는 것으로 우수한 성능의 모델을 빠른 시간에 구현할 수 있었습니다.

앞으로도 이러한 방식으로 지금까지 인공지능이 적용되지 않은 분야에 기술이 적용되고 새로운 방식의 활용사례가 늘어날 것으로 예상됩니다. 이번 프로젝트에서는 교육 분야에 자연어처리 기술을 적용했으며, 이후에도 각종 자연어처리 및 인공지능 서비스를 추가하여 외국어 학습을 수월하게 만들 수 있을 것이라 기대됩니다.

**4. 주요 기술**

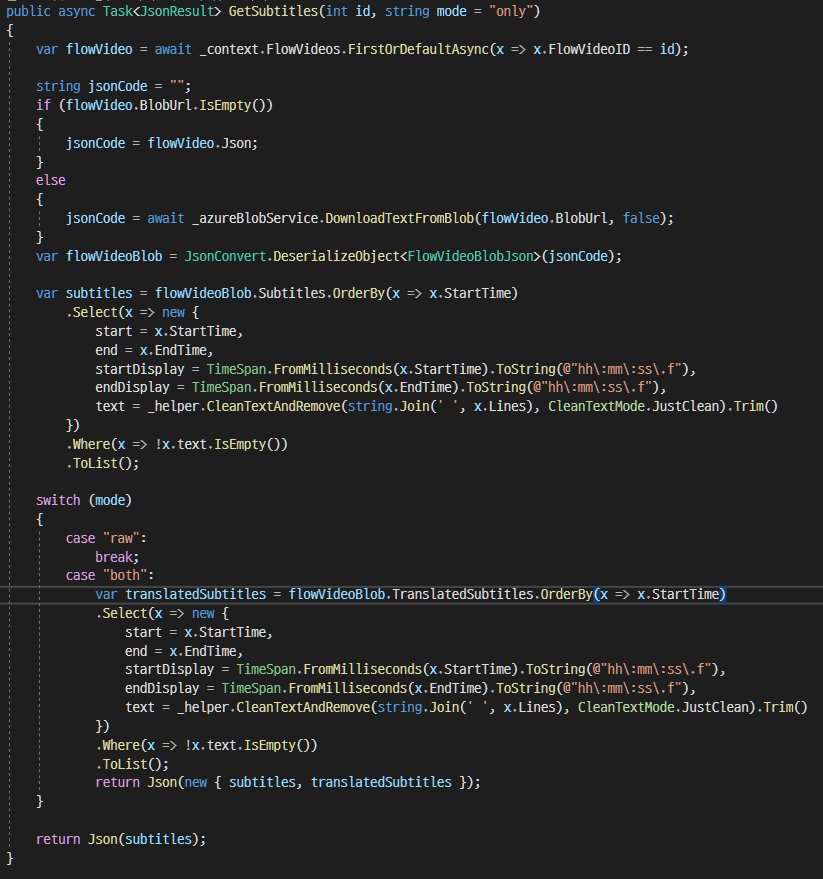
- 언어 : C# (AI 로직 등 주요 부분), Javascript (웹페이지, 학습 관련 동작)

- 모델 : NLP 관련 복수 모델 사용 (TTS, STT, 형태소분석 외)

- 프레임워크 : ML.NET

**5. 주요 소스코드**

- 유튜브 콘텐츠 자막 획득 및 처리 관련 코드



- 텍스트의 문장 성분 추출 관련 코드

