

<http://dx.doi.org/10.17703/JCCT.2021.7.2.427>

JCCT 2021-5-51

문학 텍스트를 활용한 머신러닝 언어모델 구현

Machine Learning Language Model Implementation Using Literary Texts

전현구*, 정기철**, 권경아***, 이인성****

Hyeongu Jeon*, Kichul Jung**, Kyoungah Kwon***, Insung Lee****

요약 본 연구의 목적은 문학 텍스트를 학습한 머신 러닝 언어 모델을 구현하는데 있다. 문학 텍스트는 일상 대화문처럼 질문에 대한 답변이 분명하게 구분되지 않을 때가 많고 대명사와 비유적 표현, 지문, 독백 등으로 다양하게 구성되어 있다는 특징이 있다. 이런 점들이 알고리즘의 학습을 용이하지 않게 하여 문학 텍스트를 활용하는 기계 학습의 필요성을 저해시킨다. 문학 텍스트를 학습한 알고리즘이 일반 문장을 학습한 알고리즘에 비해 좀 더 인간 친화적인 상호작용을 보일 가능성이 높다. 본 논문은 ‘문학 텍스트를 학습한 머신 러닝 언어 모델 구현’에 관한 연구로서, 대화형 기계 학습에 문학 텍스트를 활용하는 연구에서 필수적으로 선행되어야 할 세 가지 텍스트 보정 작업을 제안한다: 대명사 처리, 대화쌍 늘리기, 데이터 증폭 등에 대한 내용으로 기계 학습이 용이하고 그 효과도 높다고 판단됩니다. 인공지능을 위한 학습용 데이터는 그 의미가 명료해야 기계 학습이 용이하고 그 효과도 높게 나타난다. 문학과 같은 특수한 장르의 텍스트를 자연어 처리 연구에 도입하는 것은 새로운 언어 학습 방식의 제안과 함께 머신 러닝의 학습 영역도 확장시켜 줄 것이다.

주요어 : 자연어 처리, 머신 러닝, GPT-2, 시퀀스 투 시퀀스, 문학 텍스트

Abstract The purpose of this study is to implement a machine learning language model that learns literary texts. Literary texts have an important characteristic that pairs of question-and-answer are not frequently clearly distinguished. Also, literary texts consist of pronouns, figurative expressions, soliloquies, etc. They hinder the necessity of machine learning using literary texts by making it difficult to learn algorithms. Algorithms that learn literary texts can show more human-friendly interactions than algorithms that learn general sentences. For this goal, this paper proposes three text correction tasks that must be preceded in researches using literary texts for machine learning language model: pronoun processing, dialogue pair expansion, and data amplification. Learning data for artificial intelligence should have clear meanings to facilitate machine learning and to ensure high effectiveness. The introduction of special genres of texts such as literature into natural language processing research is expected not only to expand the learning area of machine learning, but to show a new language learning method.

Key words : NLP, Machine Learning, GPT-2, Sequence to Sequence, Literature

*준회원, 숭실대학교 글로벌미디어학부 (제1저자)

**정회원, 숭실대학교 글로벌미디어학부 (제2저자)

***정회원, 숭실대학교 글로벌미디어학부 (공동교신저자)

****정회원, 숭실대학교 영어영문학과 (공동교신저자)

접수일: 2021년 4월 17일, 수정완료일: 2021년 4월 30일

게재확정일: 2021년 5월 7일

.Received: April 17, 2021 / Revised: April 30, 2021

Accepted: May 7, 2021

*Corresponding Authors: kyounga.kwon@maat.kr

School of Global Media, Soongsil Univ.

peace@ssu.ac.kr

Dept. of English Language and Literature, Soongsil Univ.

1. 서론

자연어 처리를 위한 머신 러닝 언어 모델로는 기계 번역에서 주로 사용되는 시퀀스 투 시퀀스 (Sequence-to-Sequence, 이하 Seq2Seq) [1] 를 비롯해서 어텐션(Attention) 기능을 추가해 S2S의 단점을 보완한 트랜스포머(Transformer) [2], 대규모 파라미터가 저장된 번역 기반 언어 모델인 Generative Pre-Training(이하 GPT-1, GPT-2) [3] 등이 있다. 비교적 최근 OpenAI에서 발표한 GPT-3는 약 3천억 개의 토큰과 1,750억 개의 파라미터를 보유한 것으로 알려져 있다. 이처럼 방대한 양의 데이터를 사전에 학습한 GPT-3가 생성하는 텍스트는 인간의 답변과 견주어 볼 수 있을 정도로 높은 정확도를 보인다. 예를 들어, “파우스트는 누가 썼나요?”라는 질문을 입력했을 때 GPT-3는 “요한 볼프강 폰 괴테가 파우스트를 썼습니다.”라는 응답을 제시한다. 의학 지식과 관련된 질문을 했을 때도 유의미한 답변을 산출하며, 어떤 학습 모델은 스스로 기사를 작성하기도 한다 [4].

머신 러닝 언어 모델 구현을 위해 주로 사용되는 학습 자료는 ‘말뭉치 데이터’다. 말뭉치 데이터는 대화 학습을 용이하게 하기 위해 선별된 데이터로 짝지어진 대화 쌍들의 집합이다. 개발자가 직접 학습 자료를 수집하는 경우에는 ‘웹 크롤링(Web Crawling)’ 방식을 주로 사용한다. 웹 크롤링은 웹상에 존재하는 데이터를 수집하는 목적으로 개발된 자동화 기능이다. GPT도 이 기능을 이용해서 위키피디아(Wikipedia) 같은 웹상의 데이터를 수집해서 학습하는 것으로 알려져 있다. 말뭉치 데이터도 웹 크롤링 방식을 주로 이용해서 기본 자료를 수집하지만, 대화 학습을 위한 용도로 재가공한 후에 제공되기 때문에 웹 크롤링 방식만으로 수집된 데이터와는 차이가 있다. 이런 데이터를 이용해서 학습된 모델들은 일상 대화, 시사 상식 유형의 질문이나 응답 생성에 강점을 보이며, 간혹 인간이 말하는 것처럼 인식되는 답변을 산출하기도 한다.

이런 장점 때문에 말뭉치 데이터 혹은 웹 크롤링 방식으로 수집된 데이터를 이용한 기계 학습 및 자연어 처리 연구가 다양하게 수행되고 있다. 하지만, 여기에는 그 한계도 분명히 존재한다. 문학 텍스트와 같이 특수한 문장에 대한 학습이 제대로 이뤄지지 않는다는 점이다. 그 이유는 첫째, 문학 텍스트는 일상 대화나 시사

상식처럼 불특정 다수를 대상으로 하지 않기 때문에 일반 문장에 비해서 사용 빈도가 낮고, 따라서 개발 필요성도 상대적으로 높지 않다. 둘째, 문학 텍스트는 “오늘 날씨 어때?” “응, 좋아”, 혹은 “비가 올 것 같은데?”라는 식의 일상 대화문처럼 질문에 대한 답변이 분명하게 구분되지 않는 경우들이 많다. 마지막으로, 문학 텍스트가 일반 문장과 가장 특징적으로 구별되는 점은 대부분의 문학 텍스트들이 ‘대명사’, ‘비유적 표현’, ‘지문’, ‘독백’ 등으로 구성되어 있다는 것이다. 영어의 문어체적 표현은 일반적으로 화자의 이름을 반복해서 사용하지 않고 처음 그 이름을 표기한 이후 문장부터는 해당 화자의 성별을 지시하는 대명사로 대체한다. 더 나아가, 대화문 앞뒤에 화자가 명시되지 않아서 텍스트의 앞뒤 문맥을 이해하지 않고서는 해당 대화문의 발화 주체를 파악하기 어려운 경우들도 종종 있다. 이런 경우에 알고리즘이 대화문의 각각의 화자들을 분명하게 인식할 수 없게 되어 등장인물별 특징에 따른 페르소나 학습 자체가 불가능해진다. 또한, 과일을 뜻하는 ‘사과’와 잘못을 인정하고 용서를 비는 의미의 ‘사과’가 동일한 단어를 사용하지만 각기 다른 의미를 내포하는 것처럼, 중의적이거나 비유적인 표현들도 본래의 의미를 잘못 해석할 가능성이 매우 높다.

본 연구의 목적은 문학 텍스트를 학습한 머신 러닝 언어 모델을 구현하는 것이다. 인공지능을 위한 학습 데이터는 그 의미가 명료해야 기계 학습이 용이하고 그 효과도 높게 나타난다. 문학 텍스트는 일상 대화문처럼 질문에 대한 답변이 분명하게 구분되지 않는 경우가 많고 대명사와 비유적 표현, 지문, 독백 등이 다양하게 구성되어 있는 특징이 있다. 이를 보완하기 위해, 본 연구는 대화형 기계 학습에 문학 텍스트를 활용하는 연구에서 필수적으로 선행되어야 할 세 가지 텍스트 보정 작업을 제안한다: 대명사 처리, 대화쌍 늘리기, 데이터 증폭이 그것이다.

대화형 언어 모델의 학습에 적절한 문학 텍스트는 전체 학습량을 채우면서도 개인화가 가능한 구조를 가져야 한다. 텍스트 안에 등장인물이 너무 많으면 스토리와 대화가 분산되어 알고리즘을 개인화할 수 있는 캐릭터를 특정할 수가 없다. 파스칼의 광세와 같이 자서전적인 독백 형식의 텍스트들은 대화문 자체가 구성되지 않는다. 따라서 챗봇형 언어 모델의 학습에 가장 적합한 텍스트는 소수의 인원을 중심으로 이야기가 전개

되는 방대한 서사 혹은 시리즈물이다. 소수의 인물을 중심으로 다양한 사건들이 전개되는 서사적 구조 혹은 시리즈물은 이와 같은 문제점을 보완하면서도 각 사건 별로 등장인물들의 다양한 변모를 확인할 수 있어 여러 방면의 페르소나가 학습된 알고리즘을 구현할 수 있다. 이러한 목적에 부합하는 텍스트로, 본 연구는 C. S. Lewis의 판타지 소설 시리즈인 『나니아 연대기』(The Chronicles of Narnia) (1950~1956) 1권부터 7권까지의 전체 텍스트를 알고리즘의 학습 데이터로 채택하였다 [5].

원문 텍스트 보정 및 학습 과정은 다음과 같다. (1) 원문 텍스트에서 대화문에 해당하는 데이터를 추출하여 학습에 사용하였다. (2) 대명사 및 인용부호 처리, 특정 등장인물의 대화량 부족 등의 문제로 모든 대화문을 한 인물로 간주하고 Seq2Seq에 학습을 수행하였다. (3) Seq2Seq에 기본적인 언어 학습이 선행되지 않아서 문법 오류가 많은 점을 보완하기 위해 영화 스크립트, 트위터 등의 데이터를 추가하여 학습을 진행하였다. (4) 원문 데이터가 너무 길고 복잡해서 학습이 제대로 수행되지 않는 복문 혹은 중문 일부를 단문으로 수정하고 긴 답변을 짧게 보정하였다. (5) 기 학습 데이터가 전혀 없는 Seq2Seq 모델의 한계를 해결하기 위해 언어 학습 모델을 GPT-2로 교체하여 학습을 수행하였다.

본 연구에서 제안하는 세 가지 텍스트 보정 작업, 즉 대명사 처리와 대화 쌍 늘리기, 그리고 데이터 증폭은 각각 (2)번째와 (3)번째, 그리고 (4)번째 단계에서 수행하였다.

II. 관련 연구

인공지능 시대의 글쓰기, 인문콘텐츠와 문화 기술 관해서는 융합 연구[6, 7, 8, 9]를 참조하였다. 머신러닝의 언어 학습과 캐릭터 유형화 관련해서는 [10, 11, 12]를 참조하였다. 본 연구에서 대화쌍 늘리기 작업은 자체적으로 수행하였고 이후 [13]를 통해 이를 확인하였다.

본 연구에서 실험에 사용한 머신러닝 언어 모델은 Seq2Seq와 GPT-2 두 종류이다. Seq2Seq는 기존의 딥러닝 네트워크(Dep Neural Networks)의 한계점을 보완한 모델로 2014년 구글에서 발표한 “Sequence to Sequence Learning with Neural Networks” 논문을 통

해 처음 소개되었다. Seq2Seq의 네트워크는 크게 인코더(encoder)와 디코더(decoder)로 구분된다. 인코더와 디코더는 순환신경망의 한 종류인 Long Short-Term Memory(LSTM) 여러 개가 연결된 형태다. 따라서 다양한 길이의 입출력 학습이 가능하고, 긴 시퀀스에서도 이전 단계의 정보를 다음 단계로 전달할 수 있어 초기 단계의 정보 손실이 최소화 된다. Seq2Seq는 주로 기계번역에 이용되지만, 질문과 답변으로 입출력 시퀀스를 구성하는 챗봇 유형의 언어 모델에서도 사용이 가능하다 [1]. GPT-2는 뉴스기사와 위키피디아 같은 웹 텍스트와 Conversational Question Answering Challenge (CoQA)의 12,7000+ 데이터 셋 등 방대한 양의 자료가 사전 학습되어 있어서 기 학습된 데이터를 바탕으로 뉴스 기사나 소설 등 다양한 형식의 글을 작성할 수 있다. GPT-2는 단어나 문장이 제시되면 새로운 문장을 만들어 내기도 하고 짧은 시간 안에 다량의 텍스트를 생성해 내는 것도 가능하다 [3].

[14]는 데이터 부족 문제를 해결하기 위해 NLP Cloud API를 이용하여 데이터를 증폭하는 방법을 제안한다. 본 연구에서는 구글 번역 API를 이용하여 기존 데이터의 왕복 번역을 통해 의미가 비슷한 데이터를 생성하는 방식을 시도했다.

[15]에서는 열린 담화에서의 높은 개인화가 사용자의 수용의도와 만족도에 긍정적인 영향을 준다는 것을 실험을 통해 제시한다. [16]에서도 사용자들의 선호도를 파악해서 개인을 그룹화 하여 서비스를 제공하는 방식이 추천의 정확도와 사용자의 만족도를 함께 높였다는 연구 결과를 제시한다. 본 연구는 문학 텍스트 안의 등장인물들의 대화를 학습한 개인화된 언어 모델 구현과 관련해서 위 연구들을 참조하였고, 본 논문의 향후 연구 과제인 챗봇의 개인화가 사용자의 만족도에 미칠 영향들을 사전 검토하였다.

III. 원문 텍스트 보정 작업

머신러닝 언어 모델의 문학 텍스트 학습을 위한 원문 텍스트 보정 작업은 기본 보정 작업과 학습 목표에 따른 추가 보정 작업으로 나눌 수 있다. 기본 보정 작업은 대소문자, 마침표 등 기본 문장을 교정하고, 중문이나 복문 일부를 단문으로 교체, 긴 답변을 짧게 보정하였으며, 마지막으로 일부 오타를 수정하고 구어체에서

불필요하게 사용되는 표현들을 삭제하였다. 학습 목표에 따른 추가 보정 작업은 세 가지 방식으로 진행하였다. 첫 번째 작업은 등장인물을 기준으로 질문-답변을 구성하는 방식이었으며, 두 번째 작업은 전체 텍스트를 기준으로 질문-답변을 구성하는 방식이었다. 마지막으로 특정 등장인물의 대화문만을 추출하는 방식으로 진행하였다.

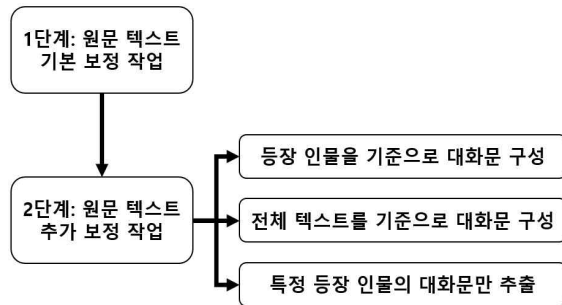


그림 1. 원문 텍스트 보정 작업 프로세스

Figure 1. Original text correction work process

1. 등장인물을 기준으로 질문-답변 구성

본 연구에서 채택한 학습 모델이 챗봇인 점을 감안해서 원본 텍스트에서 대사 형식으로 구성된 모든 문장을 추출하고 추출된 문장들을 각각 ‘질문-답변’의 형식으로 재구성하였다. 아래와 같이 하나의 질문에 여러 답변이 나오는 경우에는 각각의 답변에 동일한 질문을 재배치하였다.

<원문 텍스트 구성 예시>

"It's an owl," said Peter. "This is going to be a wonderful place. ..."

"Badgers!" said Lucy.

"Foxes!" said Edmund.

"Rabbits!" said Susan.

<보정 결과 예시>

"This is going to be a wonderful place. ..." said Peter

"Badgers!" said Lucy.

"This is going to be a wonderful place. ..." said Peter

"Foxes!" said Edmund.

"This is going to be a wonderful place. ..." said Peter

"Rabbits!" said Susan.

다음으로, 학습 모델이 각 대화문의 화자가 누구인지를 분명하게 인식할 수 있도록 ‘별명’이나 ‘일반명사,’ 또는 Mr.와 같이 성명 앞에 붙여 사용하는 ‘의존명사’로 표기된 명칭들을 모두 대표 이름으로 통일하고 ‘대명사’로 표기된 경우에는 문맥에 따라 해당 화자의 이름으로 교체하였다.

<별명 수정 예시>

1. Lu = Lucy

2. Su = Susan

3. Ed = Edmund

<일반명사, 의존명사 수정 예시>

1, the Faun, Mr. Tumnus = Tumnus

2. Lion = Aslan

3. Mr. Rumblebuffin = Giant Rumblebuffin

기타 보정 작업은 다음과 같다. (1) 텍스트의 형식이 대화문처럼 인용부호(quotation mark) 안에 포함되더라도 내용이 대화가 아닌 경우에는 제외하였다. (e.g., 3권에서 ‘유스터스(Eustace)’의 일기). (2) 텍스트가 연대기 방식으로 구성되어 있어서 동일 인물일지라도 이야기 전개 과정에서 특정한 이유로 다른 페르소나로 표현된 경우에는 각각의 명칭을 유지하는 방식을 채택하였다. (e.g., 4권에서 ‘릴리안(Rillian)’과 릴리안이 저주를 받아 빙의된 상태를 지칭하는 ‘블랙 나이트(Black Knight)’). (3) 본래의 이름이 텍스트 후반에 가서야 밝혀지는 경우, 텍스트 설정 상으로는 한 인물이지만 각기 다른 인물로 구분하고 각각의 이름을 그대로 표기하였다. (e.g., 5권에서 ‘사스타(Shasta)’와 사스타가 태어날 때 부여 받은 이름인 ‘코어(Cor)’). (4) 한 인물의 대화가 지문을 사이에 두고 연속해서 이어지는 경우에는 해당 지문을 삭제하고 하나의 대화문으로 통일하였다.

2. 텍스트를 기준으로 질문-답변 구성

서론에서 언급한 바와 같이, 본 연구는 알고리즘 학습량 및 개인화 가능성을 고려하여 소수의 캐릭터들을 중심으로 사건이 전개되는 시리즈물을 학습 텍스트로 채택하였다. 하지만, 선택된 텍스트가 전체 7권으로 구성되어 있고 이야기 또한 소수의 캐릭터들을 중심으로 전개되었음에도 불구하고, 캐릭터별 학습 데이터로는 알고리즘의 개인화 가능성을 도출하기에 한계가 있었

다. 그래서 고안된 방법이 전체 대화문을 하나의 인물로 간주하고 학습하는 것이었다. 이전 텍스트 보정 작업에서는 각 인물별로 질문-답변을 구성하고 이에 따라 인물을 지칭하는 명칭들을 통일시키는 것에 주안점을 두었다면, 이번 단계에서는 등장인물들을 구분하지 않고 텍스트 전체를 하나의 인물로 간주하고 보정하였다. 먼저, 대화문의 구성이 챗봇의 학습에 용이한지를 판별하기 위해 모든 대화문을 세 가지 기준, 즉 질문-대답만으로 의미가 성립되는 경우와 질문-대답만으로는 의미가 성립된다고 보기 어렵지만 텍스트 전체 맥락에 비출 때 유의미하다고 판단되는 경우, 마지막으로 질문-대답 자체가 성립되지 않는 경우로 분류하였다.

<보정 예시 1>

(원문) "We've fallen on our feet and no mistake," Peter said, "This is going to be perfectly splendid. That old chap will let us do anything we like."

-> (질문) A: "We've fallen on our feet and no mistake," Peter said,

-> (대답) B: "This is going to be perfectly splendid. That old chap will let us do anything we like."

이와 같은 경우 질문-대답이 성립되므로 1번으로 분류.

<보정 예시 2>

A: "But it wasn't a hoax at all," said Lucy, "really and truly. It was all different a moment ago. Honestly it was. I promise."

B: "Come, Lucy," said Peter

이와 같은 경우에는 질문-대답 자체의 유의미성은 완전하지 않지만 텍스트의 맥락을 어느 정도는 반영하므로 2번으로 분류.

<보정 예시 3>

A: "Foxes!" said Edmund

B: "Rabbits!" said Susan

이와 같은 대화 쌍은 질문-대답 자체가 성립되지 않기 때문에 3번으로 분류.

문학 텍스트는 일반적으로 한 인물의 '말'에 대한 다른 인물의 '말'로 구성되지만, 특정 장면에서 이와 같이 진행되지 않을 경우에는 한 인물의 '말'을 잘라서 강제로 질문-대답 형식으로 변환하고 질문-대답 성립 여부를 판별하였다. 또, 한 쌍의 인용부호 안에 포함되는 대화문의 길이가 너무 긴 경우는 학습이 용이하지 않을 것으로 판단하고 학습 데이터에서 제외하였다.

1단계 작업은 텍스트의 맥락을 파악하지 못해도 선별 작업이 가능하지만, 2단계 작업은 텍스트의 전체 내용을 정확하게 파악하지 못하면 작업 자체가 불가능하다. 또, 선별 작업자가 텍스트의 맥락을 어느 정도 이해하고 있더라도 이해도의 수준에 따라 동일한 대화 쌍을 2번으로 분류할 수도 있고 혹은 3번으로 분류할 수도 있다. 2단계 작업은 1단계에 비해서 고난이도 분석 작업에 해당한다.

3. 특정 캐릭터의 대화문만 추출한 경우

본 단계는 본 연구의 최종 목적인 문학 텍스트 내의 가상 캐릭터 구현 가능성을 탐색하기 위한 기초 작업으로 수행되었다. 학습을 위해 선택된 등장인물 루시(Lucy)는 텍스트에서 이야기를 이끌어 가는 중요한 캐릭터 중의 하나다. 본 연구에서는 GPT-2를 이용해 새로운 가상 캐릭터를 생성하고 기계 학습을 통해 창조된 캐릭터에 기초해서 대화문을 확장하기 위한 목적으로 시도되었다.

먼저, 순수한 대화문장만을 학습시키기 위해 인용부호(“)와 “said Lucy”와 같이 화자가 누군지를 표현하는 문구를 모두 삭제하고 추출된 문장과 그 문장이 포함하는 단어의 개수를 함께 표기하여 표로 작성하였다. 문장에 포함된 단어의 개수를 산출한 이유는 GPT-2가 주어진 단어 개수에 맞춰 응답을 산출하도록 개발되어 있기 때문이다. GPT-2의 응답 결과를 보기 위해서는 응답에 사용할 수 있는 단어의 개수를 기본적으로 지정해주어야 한다.

앞서 수행한 두 단계의 보정 작업과 마찬가지로 이번 단계에서도 유효한 대화쌍이 성립되는 경우와 그렇지 않은 경우가 있었다. 이에 따라 이어지는 대화문을 하나의 문장으로 통일할 지 아니면 각각의 대화문으로 분리할 지를 결정하는 작업이 수행되었다. 가령, 원문의 구성이 "Thank you very much, Mr Tumnus," said Lucy. "But I was wondering whether I ought to be

getting back.“ 와 같이 이어지는 경우, 이전 작업에서는 인물을 중심으로 텍스트를 재구성하였기 때문에 한 인물의 대화로 간주하여 한 문장으로 통일했지만, 이번 작업에서는 비록 동일 인물의 대화일지라도 각각의 문장이 독립적이라 판단하고 두 개의 문장으로 분리해서 재구성하였다. 아래 <보정 예시 4>에서처럼 이어지는 두 문장이 서로 종속된다고 여겨질 경우에는 하나의 문장으로 합치하였다.

<보정 예시 4>

“Why, sir?”, “I think I don't know but I could be brave enough.”

IV. 학습 과정

1. Seq2Seq 모델 학습

Seq2Seq 학습은 총 5단계에 걸쳐 진행되었다. 원문 텍스트에서 대화 내용만 추출하는 전처리 및 초기 설정 작업이 우선 수행되었고, 기본적인 대화문 구조를 생성한 후에는 대화쌍 늘리기, 데이터 증폭, 말뭉치 데이터 활용 같이 학습 데이터를 늘리기 위한 작업들이 차례로 시도되었다.

1) 텍스트 전처리 과정 및 초기 설정

소설 속의 데이터를 텍스트 파일로 모두 변환한 뒤 등장인물들의 대화 내용만 추출하기 위하여 인용부호(“”)가 포함된 문장을 파이썬을 통해 추출하였다. 추출된 데이터는 공통적으로 “대화 + said + 등장인물“ 로 이루어져 있었다. 가령, “I think he's an old dear,” said Susan. 과 같은 방식이다. 대부분은 이 방식을 따르고 있었지만 간혹 said 대신에 asked, exclaimed 으로 표현된 경우도 있었고, 영어 텍스트 특성상 등장인물의 이름이 대부분 대명사로 바뀌어 있었다.

학습 초기에는 알고리즘이 소설 속의 주요 등장인물들의 대화를 학습하여 그 인물에 상응하는 가상 캐릭터를 생성하는 것을 목표로 하였지만, 대명사에 해당하는 등장인물을 교체하는 과정 없이는 학습 자체가 불가능했다. 주요 등장인물인 루시(Lucy)의 경우, 루시가 등장하는 총 5권에서 대명사를 제외하면 대화쌍이 대략 400쌍 정도에 불과하다.

2) 대화문 구조 생성

대화문 구조는 원문 텍스트의 모든 대화문을 하나의 언어 모델로 간주한 후 [그림 2]와 같이 질문-대답 쌍을 구성하였다. 추출된 대화문은 총 4,864문장이었으며 학습에 사용된 대화쌍은 2,432쌍이다.

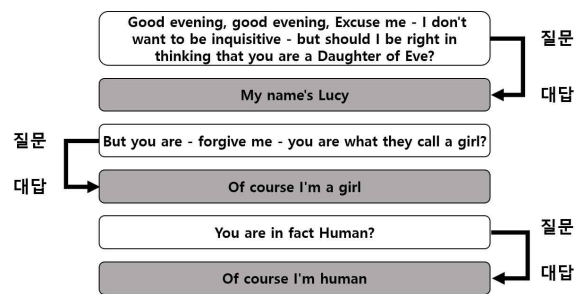


그림 2. 질문-대답 쌍 구축

Figure 2. Building a question-answer pair

3) 대화쌍 늘리기

대화쌍 늘리기는 [그림 3]과 같이 이전 대화 쌍의 대답을 질문으로, 다음 대화 쌍의 질문을 대답으로 교체하는 방식으로 기존 2,432쌍을 3,648쌍으로 늘려서 학습을 시도했다. 소설 속의 대화는 질문이 대답이 될 수도, 대답이 질문이 될 수 있다는 점에서 착안했다.

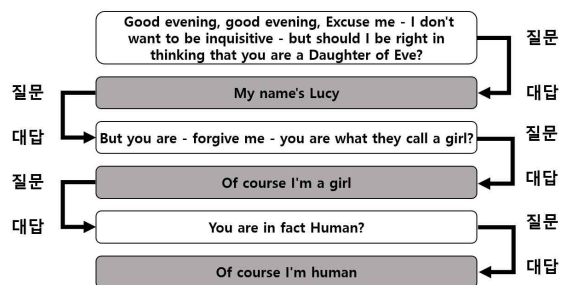


그림 3. 대화쌍 늘리기

Figure 3. Increasing conversation pairs

4) 데이터 증폭

데이터 증폭은 구글 번역기와 googletans 라는 API를 이용했다. 기 데이터를 한 문장씩 불러와 영어에서 특정 언어로 번역을 한 뒤 번역 결과를 또 다른 언어로 재차 번역하는 과정을 반복한 후, 마지막에 다시 영어로 재번역하여 마무리하는 방식으로 작업을 진행하였

다. (그림 4) 이렇게 하면 의미는 동일하지만 단어, 어순 등이 바뀐 새로운 텍스트를 확보할 수 있다. 번역 횟수가 늘어날수록 문장의 변형도 함께 늘어나며, 사용된 언어에 따라서 결과도 달라진다. 이런 방식으로 증폭된 데이터량은 원본 데이터를 기준으로 약 10배 정도 된다.

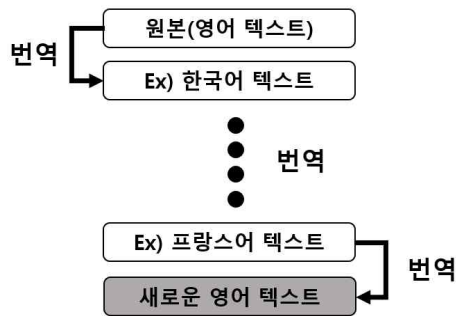


그림 4. 번역 API를 통한 데이터 증폭
Figure 4. Data amplification through translation API

5) 말뭉치 데이터 보충 학습 및 원문 데이터와의 학습 비율 조정

앞서 언급한 바와 같이, Seq2Seq는 GPT-2처럼 학습된 데이터가 전무하기 때문에 기본적으로 문법에 맞는 문장을 산출하는 능력이 부족하다. 이를 보충하기 위해 트위터와 영화 데이터를 추가하여 학습을 시도하였다. 하지만, 원문 텍스트와 직접적으로 연관되지 않는 말뭉치 데이터를 너무 많이 학습할 경우 문학 텍스트 학습을 통해 특정한 성격을 가진 언어 모델을 구현한다는 본래의 목적에 부합되지 않는 결과가 나올 수도 있는 단점이 있다. 그래서 Seq2Seq에 적절한 부차적 학습 데이터양을 가늠하기 위해 원문 데이터와 외부 말뭉치 데이터의 이용 비율을 각각 1:1, 1:2, 1:100으로 조정하면서 작업을 시도하였다. (1 = 8,000 문장). 외부 말뭉치 데이터 비율이 높아질수록 원문 텍스트에는 없는 단어를 출력하는 답변이 늘어나는 것을 확인할 수 있었다.

2. GPT-2 모델 학습

GPT-2에 『나니아 연대기』 1권부터 7권까지의 모든 원문 텍스트 파인 튜닝(fine tuning)을 시도했다. 파인 튜닝은 GPT-2에서 기본적으로 제공하는 기능으로 기존 모델의 파라미터로부터 학습을 업데이트하여 아키텍처를 새로운 목적에 맞게 변형할 수 있는 방식이

다. 파인튜닝을 거친 GPT-2에 원문 텍스트 “Oh, don’t you know? Aslan is the great Lion who comes from over the sea.”를 입력하고 출력 문장의 길이를 20으로 지정했을 때, 학습 알고리즘이 “That’s right. But no one ever said so before in Narnia.”라는 문장을 생성했다. 이처럼 문학 텍스트의 등장인물이 대답하는 것처럼 보이는 응답문도 있지만, “I don’t care what they say, we wouldn’t do it.”처럼 원문 텍스트의 내용과는 무관한 답변들도 있었다. 사전 학습 모델이라는 GPT-2의 특성상 문법적, 문맥적으로는 높은 문장 완성도를 보였지만, 추가로 학습된 문학 텍스트의 영향력은 기 학습 데이터의 영향력보다 적은 것으로 나타났다.

V. 결과 및 평가

1. Seq2Seq 학습 결과

원문 데이터의 질문-대답 2,432쌍을 학습한 Seq2Seq에 질문을 던졌을 때 95% 이상의 답변이 문법적 오류를 나타냈다. 소설 속에 존재하는 질문인 “You are in fact human?”이라는 문장을 입력하면 간혹 “of course but i’m not a kid.”와 같이 문맥이 연결되는 문장도 보였다. 위 문장에 대한 대화 쌍은 본래 “of course I’m human.”이었다. 또한, 이 문장을 질문으로 넣었을 때 출력되는 문장 중 약 50%는 “of course”로 시작을 하였고, 그 뒤에는 kid, girls, he와 같이 human에 대응하는 단어들이 나오는 것을 확인할 수 있었다. 이런 점들에 비추어 볼 때, 신경망이 human에 해당하는 단어들을 kid, girls, he와 같은 부류로 학습했다는 것을 유추해 볼 수 있다.

“Who is she?”라는 질문에 대해서는 “the king of narnia”라고 답변한 결과가 있었다. 학습 데이터의 원문 대답은 “Why, it is she that has got all Narnia under her thumb.”이었다. 이런 사례는 알고리즘이 원문과는 다르지만 명확한 답변을 내놓는 것을 보여준다.

[그림 3]의 방식으로 질문-대답 쌍을 3,648쌍으로 늘리고 파라미터는 [표 1]과 동일하게 하여 학습을 진행하였다. 결과는 2,432쌍을 학습한 이전 결과와 유사하게 나타났다.

표 1. Seq2Seq 학습 파라미터

Table 1. Learning Parameters of Seq2Seq

Hyperparameter		값
voca_size	단어사전 갯수	10000
incoeder_seq_length	인코더 길이	20
decoder_seq_length	디코더 길이	20
emb_dim	임베딩 차원	1024
batch_size	배치 사이즈	32
num_epochs	에폭	100

2. GPT-2 학습 결과

GPT-2 학습 결과를 확인하기 위한 입력문으로 원문 텍스트에서 총 10개 문장을 선택하였다.

1. Hadn't we all better go to bed?
2. This must be a simply enormous wardrobe.
3. I am very pleased to meet you, Mr. Tumnus.
4. I don't think you're a bad Faun at all.
5. But what'll they do to him, Mr. Beaver?
6. What's wrong? Can't you tell us?
7. They're doing something worse to Him..
8. What an extraordinary place!
9. What Aslan did for him? Does he know what the arrangement with the Witch really was?
10. It will not go out of my mind that if we pass this post and lantern either we shall find strange adventures or else some great change of our fortunes.

위 문장들은 각각 문장의 난이도, 문장의 길이, 문장의 개수, 평서문, 의문문, 감탄문 등 다양하다. 이것은 학습 알고리즘이 문장의 차이를 인식하고 변별할 수 있는 지를 확인하기 위함이다. 결과적으로, 문장 유형에 따른 변별력은 크지 않은 것으로 분석되었다.

그러나 문법적으로는 완성도가 높은 문장을 산출하는 것으로 평가되었다. 문맥이 대부분 납득할 만한 수준이었고 문장의 완성도도 높았다. 과인 튜닝을 통한 학습이 긍정적인 결과를 생성하는 것을 확인할 수 있었다.

GPT-2의 출력 단어 제한 개수는 각각 10단어, 20단어, 30단어, 40단어, 50단어, 100단어, 200단어를 기준으로 시도되었다. GPT-2가 이미 방대한 데이터 학습이 이뤄진 상태임에도 불구하고 10단어 이하의 경우에 문

장 자체가 성립되지 않는 경우가 많았다. 100단어가 넘어가면 원문 텍스트의 맥락에서 크게 벗어나 새로운 이야기가 전개되는 경우가 대부분이었다. 원문 텍스트와의 연관성에 비추어볼 때, 가장 유의미한 응답이 가능한 단어의 출력 범위는 20단어에서 50단어 사이인 것으로 평가되었다.

본 연구는 문학 텍스트에서 대화문을 인용하는 방식의 차이가 GPT-2의 응답에 미치는 영향을 확인하기 위해 두 가지 방식의 학습을 별도로 진행하였다. 하나는 인용부호와 인용부호 바로 뒤에서 화자가 누구인지를 말해주는 표현(예, "said, Peter.")들을 모두 포함해서 학습시킨 경우이고, 다른 하나는 인용부호와 화자를 설명하는 문구를 모두 삭제하고 대화문 자체만을 학습시킨 경우다. 학습 파라미터는 [표 2]와 같이 설정하였다. 이 경우, 전자는 GPT-2는 인용부호나 화자와 관련한 어떠한 단어도 산출하지 않았다.

<결과 예시 1>

인용부호와 화자와 관련한 문구 없이 모든 원문 텍스트를 학습한 GPT-2에게 "Hadn't we all better go to bed?"라는 문장을 입력하고 20자로 응답할 것을 지시한 경우

Hadn't we all better go to bed? We've all been there.

No, we haven't,

Oh well, you've been

=====

Hadn't we all better go to bed?

Well, we must change places. I suppose it's only a few hours' walk from here

=====

Hadn't we all better go to bed?

I'm sorry, I forgot to mention it to you.

Yes, It's a great

=====

Hadn't we all better go to bed? Oh, you're so tired. And you're so sleepy. And you're so sleepy.

=====

Hadn't we all better go to bed?

I don't know that I want to go out of this way, Aslan, not for

<결과 예시 2>

인용부호와 화자와 관련한 문구를 포함한 모든 원문 텍스트를 학습한 GPT-2에게 전자와 동일하게 “Hadn’t we all better go to bed?”라는 문장을 입력하고 20자로 응답할 것을 지시한 경우

Hadn’t we all better go to bed? I’m tired from the march.”

“No, I haven’t,” said Digory.

=====

Hadn’t we all better go to bed? I should think so, I should think.”

“I don’t know what you mean,” said

=====

Hadn’t we all better go to bed? I mean, why should you be King of us?”

“I say, Lucy,” said Peter

=====

Hadn’t we all better go to bed? ”

“But we’ve got to go back and have fun again,” said Jill.

<결과 예시 1>에서는 보이지 않던 인용부호와 화자를 설명하는 “said”라는 문구가 <결과 예시 2>에서 출력되는 것을 확인할 수 있다. 이것은 알고리즘이 문학 텍스트의 대화문 형식으로 응답문을 도출하도록 학습하려면, 인용부호나 화자를 설명하는 문구를 삭제하지 않는 것이 더 유의미한 결과를 산출할 수 있음을 보여 준다. 이 결과는 향후 가상캐릭터 연구에 매우 중요한 방향을 제시한다.

표 2. GPT-2 파인 튜닝 파라미터

Table 2. Fine Tuning Parameters of GPT-2

Parameter		값
model_name	모델명	124M
step	학습 횟수(=에폭)	20
batch_size	배치 사이즈	10
OUTPUT_length	출력 길이	20

VI. 결론 및 향후 연구

본 연구는 처음에 목표했던 만큼 알고리즘으로부터 만족스러운 답변을 얻지 못했다. 하지만, 본 연구의 커다란 의의는 일상적인 문장 혹은 1:1 매칭 번역을 중심

으로 진행되어 온 언어 기반 기계 학습에 문학 장르라는 특수한 텍스트를 도입함으로써 새로운 언어 학습 방식을 제시하고 기계 학습 영역을 확장했다는 점이다. 특히, 문학 텍스트 학습 과정에서 요구된 3가지 작업 유형—대명사 처리, 대화 쌍 늘리기, 데이터 증폭—은 향후 이와 유사한 학습을 시도하는 머신 러닝 연구에서는 필수적인 선행 작업으로 고려되어야 할 요소들이다.

GPT-2는 방대한 사전 학습 데이터가 내장되어 있어서 텍스트 전처리 과정을 거치지 않아도 원문 텍스트의 맥락과 연관되는 답변을 산출하였지만, Seq2Seq는 GPT-2처럼 기 입력된 데이터가 전혀 없기 때문에 다양한 텍스트 보정 작업을 거치면서 학습을 수행하였음에도 불구하고 문법에 맞는 문장을 구성하는 데 한계를 보였다. 이런 측면에서 보면, Seq2Seq는 문학데이터 학습에 적절하지 않아 보인다. 하지만, Seq2Seq에 지속적인 학습이 이루어져 어느 정도 유의미한 답변이 가능해 진다면, 문학 텍스트와 무관한 방대한 데이터를 기 학습한 GPT-2보다는 문학 텍스트에 기반을 둔 가상 캐릭터 생성에 더 높은 효과를 낼 수 있다고 판단된다.

특정 문학 작품을 머신 러닝 언어 모델에 활용하는 연구는 단지 문학 데이터를 학습하는 것만을 의미하지 않는다. 문학 작품의 시대적 배경이나 작가 혹은 각 등장인물 별 페르소나 분석 등은 여러 방향에서 인공 지능의 해석 및 활용 능력을 높일 수 있는 연구의 가능성을 제시한다. 향후 연구는 위와 같은 맥락에서 진행하고자 한다.

References

- [1] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. arXiv preprint arXiv: 1409.3215.
- [2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv: 1706.03762.
- [3] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
- [4] Artificial Intelligence Times. (2020, August 14). GPT-3, the most outstanding ability in human history, "language artificial intelligence". Retrieved from <https://www.aitimes.kr/news/article>

- View.html?idxno=17370
- [5] C. S. Lewis. (2001). *The Chronicles of Narnia*. New York: HarperCollins Publishers.
- [6] S. Kim. (2019). In the Age of Artificial Intelligence, Is Human Writing still Possible? – Critical Prediction of Changes in Writing(text). *The Korean Journal of Literacy Research*, 10(3), 283–315. doi: 10.37736/kjlr.2019.06.10.3.283
- [7] D. Han, H. Choi, J. Kim, I. Doo, S. Kim (2012). Media and Culture Technology, Humanities Contents. *Humanities Contents*, (27), 53–59.
- [8] H. Chae. (2013). A Study on Humanities and Engineering based on Human-centered Design: Providing Human Interface for Media Contents in User Equipment. *Lingua Humanitatis*, 15(1), 153–167.
- [9] N. Oh, B. Rhee. (2019). A Study on the Expression of Poetry Converged with Augmented Reality(AR). *The Korean Society of Science & Art*, 37(4), 191–204. doi: 10.17548/ksaf.2019.09.30.191
- [10] S. Choi, S. Park, K. Kim. (2017). Korean Abbreviation Generation using Sequence to Sequence Learning. *KIISE Transactions on Computing Practices*, 23(3), 183–187. doi: https://doi.org/10.5626/KTCP.2017.23.3.183
- [11] W. Kang, B. Kim. (2020). A Study on the Types of Classic Fiction Using Deep Learning Techniques – Focusing on Hero Novels and Romantic Novels -. *Korean Language and Literature in International Context*, 84, 9–35. doi: http://dx.doi.org/10.31147/IALL.84.1
- [12] H. Park. (2020). A Study on the of Historical Character Design by Non-verbal Communication Theory. *The Treatise on The Plastic Media*, 23(1), 177–184. doi: https://doi.org/10.35280/KOTPM.2020.23.1.19
- [13] S. Ha. (2017). "Dialogue response generation using RNN-based sequence-to-sequence learning (Doctoral dissertation, Graduate School of Seoul National University).
- [14] Coulombe, C. (2018). Text data augmentation made simple by leveraging nlp cloud apis. arXiv preprint arXiv: 1812.04718.
- [15] S. Park, Y. Jung, H. Kang, "Effects of Personalization and Types of Interface in Task-oriented Chatbot," *The Journal of the Convergence on Culture Technology (JCCT)* 7.1 (2021): 595–607. doi: http://dx.doi.org/10.17703/JCCT.2021.7.1.595
- [16] Kim, H., Jung, S., & Ryu, G. (2020). A Study on

the Restaurant Recommendation Service App Based on AI Chatbot Using Personalization Information. *International Journal of Advanced Culture Technology*, 8(4), 263–270. doi: https://doi.org/10.17703/IJACT.2020.8.4.263

※ 이 연구는 2019년도 송실대학교 교내연구비 지원(융합연구)에 의한 연구임.