http://dx.doi.org/10.18841/2021.64.10

인공지능에 의한 일본어 텍스트 생성 및 일본어능력시험 난이도 분석 시스템 개발 연구

김 유 영*

< 目 次 >

Key Words : 일본어능력시험(JLPT), 일본어 텍스트 난이도(Readability of Japanese text), GPT-2, GPT2-JLPT-AOT, 파인튜닝(Fine Tuning), 인공지능(AI)

<要 旨>

본 연구에서는 인공신경망 자연어 처리 플랫폼 GPT-2를 기반으로 일본어능력시험 기출문제 데이터셋을 통해 파인튜닝을 수행하여, 인공지능을 통한 일본어능력시험 학습 및 교수를 위한 텍스트 생성의 가능성을 타진하고, 한발 더 나아가 생성된 텍스트에 대한 일본어능력시험 기준 난이도 분석을 수행할 수 있는 시스템의 개발을 수행했다.

이를 통해 인공지능을 통한 일본어 텍스트의 자동 생성에는 아직 한계가 있기는 하나, 머신러닝(파인튜닝)의 데이터셋의 양, 종류 그리고 학습 횟수 등 다양한 조건을 통제하여 적절한 실험을 수행한다면, 일본어 텍스트 자동 생성에 최적화된 파인튜닝의 초기 설정값을 찾아낼 수 있을 것이며, 이는 앞으로 인공지능이 더욱 정밀하게 사용자가 원하는 일본어 텍스트를 생성할 수 있게 하는 하나의 방안이 될 것이라는 점을 확인할 수 있었다.

또한 본고에서는 이와 같은 텍스트를 단순히 생성하는 것에서 그치지 않고 텍스트에 대한 어휘, 문법, 한자 차원의 JLPT 난이도 분석까지 실시할 수 있는 시스템을 개발할 수 있었는데, 앞으로 인공지능의 일본어학 및 일본어 교육 분야에 대한 응용 가능성이 크다는 점과 후속 연구의 필요성을 확인할 수 있었다.

* 동덕여자대학교 일본어과 부교수, 일본어학 전공, yuiyu1004@dongduk.ac.kr

Ⅰ. 머릿말

IT 기술의 발전에 따라 언어 교육에 IT 기술을 응용하고자 하는 노력은 끊임없이 지속되어 왔다. 그 결과 멀티미디어 활용, PC와 인터넷의 활용을 넘어 스마트 기기를 활용한 언어 교육에 이르기까지 이와 같은 다양한 시도는 종래의 전통적인 교수법 및 교재에 대한 혁신을 불러일으켰다. 특히 팬데믹 상황 속에서 언어 교육 분야에 있어서 스마트 기기를 활용한 원격 교육 및 평가 시스템은 그 효용을 여실히 증명했으며 이러한 가치는 다가올 포스트 팬데믹 시대에도 여전할 것으로 예측된다.

한편 이와 같은 하드웨어의 혁신과 함께 소프트웨어인 교육 콘텐츠에도 새로운 변화가 나타나기 시작했는데 바로 인공지능(AI) 기술을 교육에 활용하고자 하는 시도가 바로 그것이다. ICT 시장조사기관 IDC(2020)에 따르면, 다가오는 2025년이 되면 전 세계에서 생산되는 연간 디지털 정보의 양은 약 163조 기가바이트로, 하루에 약 4,466억 기가바이트의 데이터를 생산할 것이라고 한다. 참고로, 유사 이래 인류가 2,000년대 초반까지 생산한 정보의 총량이 고작 약 20엑사바이트(Exabyte, EB, 약 200억 기가바이트(Gigabyte, GB)에 불과하다는 점을 고려할 때 근래의 빅데이터의 확장 속도는 가히 폭발적임을 알 수 있다. 이와 같은 빅데이터의 탄생은 하드웨어 기술의 발달과 어우러져 머신러닝을 기반으로 한 인공지능 개발을 가능케 했다. 그리고 인공지능은 모든 분야를 넘나들고 있으며 최근 언어 교육 분야에도 응용되기 시작해, 실제로 몇몇 에듀테크 기업에 의해 언어 학습자들의 수험 데이터를 기반으로 특정 개인의 어학 실력을 유추한다거나, 인공지능이 어학 학습을 위한 문제를 만들거나 기존의 어학 문제의 난이도를 평가하는 등의 새로운 시도가 나타나고 있다.

이에 본 연구에서는 인공신경망 자연어 처리(Neuro-Linguistic Programming, NLP) 플랫폼 GPT-2[1]를 기반으로 커스터마이징된 일본어판 GPT-2(gpt2-japanese)[2]를 활용하여 일본어능력시험(이하, JLPT) 대비 텍스트를 자동으로 개

1) 2015년 12월 11일 일론 머스크와 샘 알트만이 공동 설립한 인공지능 회사인 openAI사가 개발하여 2020년 5월에 공개한 자기회귀 언어모델로, 정식 명칭은 Generative Pre-trained Transformer 2(GPT-2). 딥러닝을 이용해 마치 인간처럼 텍스트를 생성할 수 있는 인공지능 언어 모델.

발할 수 있는 시스템 개발에 대한 가능성을 엿보고자 한다.

Ⅱ. 선행연구 및 문제 제기

앞서 언급한 바와 같이 인공지능과 관련된 연구는 인공신경망을 기반으로 만든 알고리즘 중 하나인 딥러닝의 도입으로 근래에 큰 진전을 보이고 있는데, 예를 들어 인간과 유사한 통찰력으로 이미지를 인식 및 분류하는 기술, 인간의 언어를 다루는 자연어 처리, 그리고 자동차 자율 주행 등 지금까지 반드시 인간의 통찰력 개입이 필수 불가결했던 수많은 분야에서 비약적 발전을 이루어내고 있다. 이와 같은 흐름에 일본어학 분야에서도 먼저 인공지능이 응용될 수 있을 법한 분야인 교육 분야에 관심이 나타나기 시작했는데, 박강훈(2019), 신민철(2021) 등을 그 예로 찾아볼 수 있다. 이들 연구는 기존의 통계기반, 관용적 표현의 대응 관계에 기반한 어댑티드 기계번역이 아닌 인공신경망의 딥러닝을 기반으로 하는 인공지능 엔진에 의한 기계번역 프로그램을 교육에 활용하거나 그 효용성을 검토한 것으로, 구글 등 선도기업에 의해 기구축된 인공지능 엔진을 교육 분야에 시험적으로 적용한 연구라고 할 수 있다. 이어 정확히 인공지능 관련 선행연구라고 분류하기는 어려우나, 머신러닝을 가능케 한 또 다른 하나의 축인 빅데이터를 일본어 교육에 활용한 연구로서, 김유영(2014, 2015)와 같은 과거 20년간의 일본어능력시험 기출문제 빅데이터의 연구를 통해 텍스트의 가독성 및 난이도 분석을 수행한 연구, 그리고 빅데이터의 수집, 정제, 가공 등의 전처리뿐만 아니라 일본어 텍스트에 감정 분석의 기법을 도입한 김유영(2020)의 연구, 金蘭美(2016)와 같이 작문과제 등을 자동으로 채점하는 방안에 관한 연구 등을 관련 선행연구로 들 수 있겠다.

이처럼 언뜻 보기에는 일본어학 및 교육 분야에 인공지능을 도입하는 것에 대한 인식 및 분위기는 제반 기초연구를 통해 조성되어 가고 있는 듯 보이지만, 선행연구를 자세히 들여다보면 실상은 아직 상용 혹은 무료 인공지능 번역 서비스에 의존한 사용자 차원의 이용 사례의 연구 혹은 코퍼스 연구에서 파생된

2) GPT-2 기반 일본어판 인공지능 언어 모델 - https://github.com/tanreinama/gpt2-japanese

일본어 빅데이터 연구에 머무르고 있다는 것을 알 수 있다. 이에 본 연구에서는 자연어 처리 플랫폼의 사전 모델에 대한 전이학습(Transfer learning) 및 파인튜닝(Fine Tuning)을 통해 기존의 사전 모델에 대한 성능개선 및 일본어학 및 일본어 교육에 특화된 인공지능 시스템의 개발에 관한 연구를 수행하여 본격적으로 일본어학 분야에 인공지능 관련 연구를 도입하고자 한다. 구체적으로는 인공신경망 기술을 기반으로 개발된 자연어 처리 플랫폼 GPT-2를 기반으로 대규모 코퍼스를 사용하여 사전학습된 일본어 사전학습 모델 「GTP2-Japanese」를 지난 20년간의 일본어능력시험(이하, JLPT) 기출문제 데이터셋으로 파인튜닝을 수행 및 검증하여, JLPT에 특화된 인공지능 일본어 텍스트 생성 엔진을 개발하고자 한다. 그리고 동시에 인공지능에 의해 자동으로 생성된 일본어 텍스트를 JLPT의 어휘, 문법, 한자 난이도 기준에 따라 자동 분석 가능한 JLPT 텍스트 레벨 검증 시스템 개발에 관한 연구를 수행하여 일본어학 및 일본어 교육 분야에 대한 맞춤형 인공지능 텍스트 생성 및 레벨 검증 시스템의 도입 가능성을 엿보고자 한다.

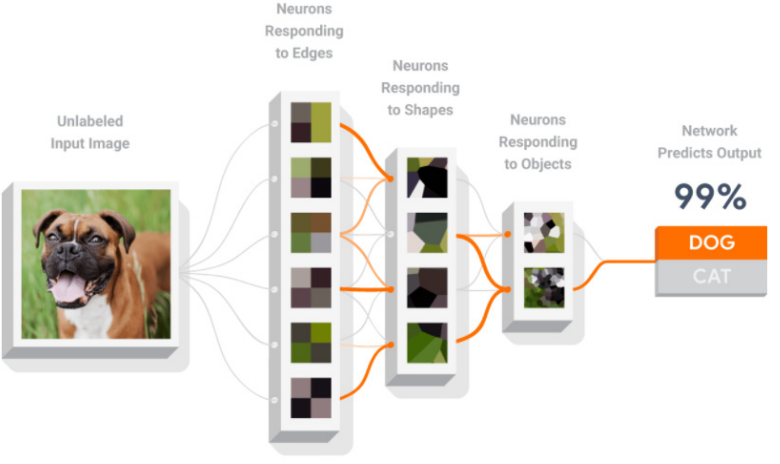
Ⅲ. 이론적 배경

1. 인공지능(인공신경망)과 머신러닝

인간의 신경계는 신체 내외부에서 일어나는 변화로 촉발되는 자극을 빠르게 뇌로 전달하여 반응을 생성하는 기관을 의미한다. 관련 기관으로 뇌, 감각기관, 뉴런 등을 들 수 있는데, 이 중 뉴런은 신경계를 이루는 가장 기본 단위를 가리킨다. 다층으로 복잡다단하게 연결된 수많은 뉴런은 정보를 받아 처리하여 전기 혹은 화학적 신호를 다음 단의 뉴런으로 전달하는데, 자주 보는 사람의 얼굴을 인식하게 된다거나, 필기체의 글자를 자주 접하면서 필기체를 읽을 수 있게 된다거나 하는 인간의 '학습'이라고 하는 능력은 반복된 자극 속에서 뇌를 포함한 신경계 속 뉴런의 연결 관계가 변화하면서 자리를 잡는 과정의 결과물이라고 할 수 있다.

본고에서 말하는 '인공지능의 학습' 혹은 '기계 학습', 다른 말로 '머신러닝'은 이와 같은 인간의 신경계 학습 프로세스를 본떠 만든 것으로, <그림 1>과 같이 수많은 인공 뉴런을 다층적으로 엮어서 예측을 수행하는 알고리즘(다층 퍼셉트

론 - Multi-Layer Perceptron, MLP)이라고 할 수 있다.



<그림 1> 신경망의 해부학적 구조 – TensorFlow 웹사이트

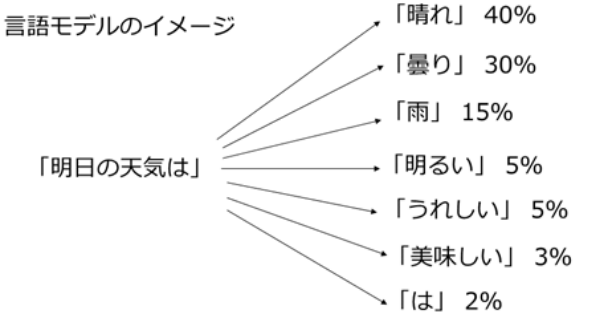
이와 같은 머신러닝에 대한 정의로서 다음의 두 설명이 가장 널러 인용되는데, 우선 아서 사무엘(Arthur Samuel, 1959)에 의하면 머신러닝이란 "컴퓨터에 명시적으로 프로그래밍하지 않고도 학습할 수 있는 능력을 부여하는 연구 분야"[3]라고 정의했다. 이어 톰 미첼(Tom Mitchell, 1997)은 "기계가 학습한다는 것은, 프로그램이 ***특정 작업***을 하는 데에 있어서 ***경험***을 통해 ***작업의 성능을 향상***시키는 것"[4]라고 정의했는데, 후자의 정의를 본 연구에 대입해 보자면 'JLPT 학습을 위한 일본어 텍스트 생성(문제출제 = 특정 작업)'을 '자동으로 신속하고 정확하게 수행(작업의 성능을 향상)'하고자 프로그램이 'JLPT 기출문제 빅데이터를 학습(경험)'하게 하는 것이라고 할 수 있겠다.

3) "the field of study that gives computers the ability to learn without being explicitly programmed." Samuel A.L. 1959.

4) "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." Tom Mitchell, 1997.

2. 인공신경망과 GPT-2

GPT는 Open AI가 2018년에 발표된 인공신경망 기반 언어모델로, 트랜스포머(Transformer[5])라는 딥러닝 체계를 기반으로 개발되었는데, 널리 알려진 구글의 BERT[6], MS의 Turing-NLG도 역시 트랜스포머를 기반으로 하는 언어모델이다. GPT-2는 <그림 2>와 같은 프로세스를 통해 인간의 글쓰기 능력을 모방할 수 있으며, GPT-3의 경우 언어 문제 풀이, 랜덤 텍스트 생성, 사칙연산, 번역, 간단한 웹 코딩 등이 가능하다. GPT의 장점은 무엇보다 매개변수(parameter)의 개수인데, GPT-1은 약 1억 1,700만 개, GPT-2는 최대 약 150억 개, GPT-3는 1,750억 개의 매개변수를 각각 갖고 있다. 한편 BERT는 Large를 기준으로 3.4억 개, Turing-NLG이 170억 개의 매개변수를 가지고 있어, GPT-3의 규모가 얼마나 큰 것인지는 쉽게 미루어 짐작할 수 있다.



<그림 2> 언어모델의 이미지 -「自然言語処理モデル「GPT-3」の紹介」

5) 2017년 발표된 논문 “Attention is all you need”에서 공개된 모델로 어텐션(Attention)으로만 구현한 모델. 순환 신경망(RNN)의 순차처리기법의 한계를 극복하기 위해 개발되었으며, AI 딥러닝 학습에 있어서 언어 이해 능력을 높였다. - “we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs” (A Vaswani 외, 2017:2)

6) BERT : Bidirectional Encoder Representations from Transformers

그러나 거대한 매개변수는 곧 높은 컴퓨터 사양과 많은 시간을 요구하며, GPT-3의 경우 완전히 공개되어 있지 않다. 이에 본 연구에서는 컴퓨터 사양의 한계와 언어모델의 공개 여부의 조건을 고려하여, 공개 버전인 GPT-2를 기반으로 「コーパス2020[7]」를 사용한 사전학습을 통해 개발된 GPT2-Japanese에 파인튜닝을 실시하고자 한다. 참고로 GPT2-Japanese에는 그 규모에 따라 세 가지 사전학습 모델이 존재하며, 본고에서는 세 가지 모델을 모두 시험해 언어모델의 기반이 되는 코퍼스의 규모에 따른 효용성에 관하여 판단해 보고자 한다. 참고로 각 모델의 매개변수 크기 등 간단한 사양은 다음의 (1)과 같다.

(1) GPT2-Japanese – 모델 별 상세 사양

a. smallモデル : 매개변수 – 약 117M / 12heads, 12layers
b. mediumモデル : 매개변수 – 약 345M / 16heads, 24layers
c. largeモデル : 매개변수 – 약 774M / 20heads, 36layers

3. 파인튜닝(Fine Tuning)

파인튜닝은 이미 특정 도메인에 사전 학습된 모델을 다른 태스크에 활용하기 위해 학습시키는 과정이다(이치훈외, 2020). 이를 본 연구에 적용해본다면, 사전에 학습된 범용적인 일본어 텍스트 생성 언어모델을 기반으로 JLPT 기출문제와 유사한 텍스트를 생성하도록 이미 학습된 가중치와 매개변수를 학습시키는 것이라고 할 수 있다. 본고에서는 사전학습 모델의 파인튜닝을 위해 다음의 (2)와 (3)과 같이 구JLPT의 독해·문법 영역 기출문제의 독해 지문을 추출하여 'JLPT-DS-AOT' 데이터셋을 구축했다. 이는 JLPT 교육을 위한 텍스트 생성을 위함이며, 이 과정에서 (4)와 같이 텍스트의 일부를 수정 및 전처리 과정을 수행했다.

(2) [구JLPT 독해·문법 영역 기출문제의 독해 지문 선정 문항]

a. 구JLPT 1급

1) 1990~1991·1993·1998년 : I·II·III 1~5

7) コーパス2020 : 6,755,346개의 텍스트, 38.6억 token로 구성된 약 23.9G 용량의 텍스트 코퍼스 https://github.com/tanreinama/gpt2-japanese/blob/master/report/corpus.md

2) 1992·1994년 : I·II·III 1~7
3) 1995~1996·1999년 : I·II·III 1~6
4) 1997년 : I·III 1~5 (저작권 문제로 미공개 문항인 II 생략)
5) 2000~2001년 : I·II 1~4·III 1~3
6) 2002·2004·2008·2009년 1차 : I·II 1~4·III 1~4
7) 2003년 : I·II 1~3·III 1~4
8) 2005~2007·2009년 2차 : I·II 1~4·III 1~5

b. 구JLPT 2급

1) 1990년 : I·II·III 1~4
2) 1991·1999년 : I·II·III 1~7
3) 1992·1995년 : I·II·III 1~5
4) 1993~1994·1996~1998년 : I·II·III 1~6
5) 2000년 : I·II 1~3·III 1~6
6) 2001년 : I·II·III 1~4
7) 2002·2004~2005·2007·2009년 2차 : I·II 1~3·III 1~5
8) 2003·2008·2009년 1차 : I·II 1~3·III 1~4
9) 2006년 : I·II 1~3·III 1~3

c. 구JLPT 3급

1) 1990~1991년 : IV
2) 1992년 : IV 1~4
3) 1993~1996년 : V 1~4·VI
4) 1997~2009년 2차 : V·VI

d. 구JLPT 4급

1) 1990년 : V 1~6
2) 1991년 : V 1~5
3) 1992~1993년 : V 1~3
4) 1994~1995년 : VI 1~3
5) 1996~1999~2001년 : V·VI 1~3

김유영(2014, 216p)

(3) [구JLPT 2008년 기출문제 중 독해·문법 영역 III-(5)번 문항 예시]

III-(5) 下のグラフは、1971年度と2007年度に行われた調査での「会社を選ぶときどんなことを最も重視したか」という質問に対する新入社員の答えをまとめたものである。 김유영(2014, 216p)

(4) [구JLPT 기출문제 전처리 및 데이터셋 구축 예시]

a. 정확도 높은 머신러닝을 위한 텍스트 수정

e.g. A「日本人が昼ごはんによく食べるものは何だと思いますか。」
B「そうですね。日本人ならおすしでしょう。」
→「日本人が昼ごはんによく食べるものは何だと思いますか。」
「そうですね。日本人ならおすしでしょう。」

2009년 2회차 旧JLPT 3급, 문법 3-5 문제 중 일부

b. 태깅을 통한 전처리

e.g.「こんなにすわり心地のいいタクシーははじめてですよ。」

「そうですか。これは、自分で言うのもなんだけど、高級車なんですよ。」

→「こんなにすわり心地のいいタクシーははじめてですよ。」<|endoftext|>
「そうですか。これは、自分で言うのもなんだけど、高級車なんですよ。」
<|endoftext|>

1990년 旧JLPT 2급, 문법 2-2 문제 중 일부

Ⅳ. 개발 환경 및 빅데이터 DB

우선, 본고에서는 언어모델 튜닝 및 JLPT 텍스트 생성 및 분석 엔진 개발을 위해 다음의 (5)와 같은 하드웨어와 소프트웨어를 사용했다.

(5) 개발 및 배포용 하드웨어 및 소프트웨어

a. 엔진 개발용 컴퓨터
   1) CPU&RAM : Intel(R) Core(TM) i7-8700K CPU 3.70GHz RAM 32G
   2) GPU : NVIDIA Geforce GTX 1080 Ti
   3) 소프트웨어 및 버전
      3.1) OS : Windows10 pro 3.2) Python : 3.7.10
      3.3) tensorflow-gpu : 2.1.0
      3.4) Anaconda(Jupyter Notebook) : 2.0.3
      3.5) GPT-2

b. 파인 튜닝용 클라우드 서버
   1) Google Colab[8)]
   2) 소프트웨어 및 버전
      2.1) OS : Linux-5.4.104+-x86_64-with-Ubuntu-18.04-bionic
      2.2) Python : 3.7.10 2.3) tensorflow-gpu : 2.5.0
      2.4) GPT-2

c. JLPT 텍스트 생성 및 레벨 검증 인공지능 시스템 데이터
   1) GPT2-JLPT-AOT
      : http://www.japanese.or.kr/JapaneseNLP_main.aspx

8) Colab : Colaboratory는 Google 리서치팀에서 개발한 제품으로 브라우저를 통해 임의의 Python 코드를 작성하고 실행할 수 있는 클라우드 환경. Colab은 특히 머신러닝, 데이터 분석 및 교육에 적합하다. - https://research.google.com/colaboratory/faq.html

이어, 인공지능에 의해 생성된 텍스트를 JLPT 문자·어휘, 문법, 한자 난이도 기준에 따라 자동 분석하기 위해 상기 (2)와 마찬가지로 20년간의 구JLPT 기출문제로부터 다음의 <표 1>과 같은 난이도 기준 데이터를 추출하여 분석 시스템에 도입했다.

<표 1> JLPT 난이도 기준 통계

수치 / JLPT섹션	N1초과	N1	N2	N3	N4	N5	합계
문자·어휘	32	4,459	3,855	2,483	2,503	882	14,214
문법	0	684	559	719	380	33	2,375
한자	0	1,172	399	337	184	103	2,195

참고 1) 문자·어휘의 경우 기출문제를 우선, 차선으로 네이버 사전을 기준으로 설정
2) 한자의 경우 「或」「伊」와 같은 상용한자 외의 한자가 포함되어 2,136자를 초과 함

<표 2> JLPT 난이도 기준 DB 예시(일부 발췌)

1) 문자・어휘

```
改訂,,1,名,21,かいてい,1,명사,개정,,,,↓
街道,,1,名,22,かいどう,1,명사,가도,교통상 중요한 도로,,,,↓
海抜,,1,名,23,かいばつ,1,명사,해발,표고,,,,↓
斡旋,,1,名,24,あっせん,1,명사,알선,주선,,,,↓
生き甲斐,いきがい,1,名,25,いきがい,NONE,NONE,보람,,,,↓
遺跡,,1,名,26,いせき,1,명사,유적,유지,,,,↓
隠居,,1,名,27,いんきょ,1,명사,은거,살림의 책임을 물려 주고,,,,↓
衛星,,1,名,28,えいせい,2,명사,위성,,,,↓
閲覧,,1,名,29,えつらん,1,명사,열람,,,,↓
```

2) 한자

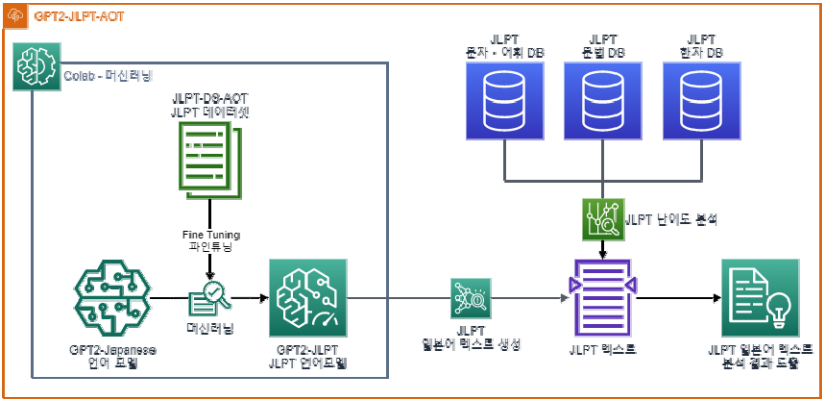
```
2167,惧,,,,1,Jouyo,,,,,,,,,,,,,,,,,,↓
2168,慄,,,,1,Jouyo,,,,,,,,,,,,,,,,,,↓
2169,憚,,,,1,Jouyo,,,,,,,,,,,,,,,,,,↓
2170,拉,,,,1,Jouyo,,,,,,,,,,,,,,,,,,↓
2171,摯,,,,1,Jouyo,,,,,,,,,,,,,,,,,,↓
2172,曖,,,,1,Jouyo,,,,,,,,,,,,,,,,,,↓
2173,楷,,,,1,Jouyo,,,,,,,,,,,,,,,,,,↓
2174,鬱,,,,1,Jouyo,,,,,,,,,,,,,,,,,,↓
2175,炒,,,,1,NotJouyo,,,,,,,,,,,,,,,,,,↓
```

3) 문법

```
,を鑑みて,,1,,명사+,~를 감안하여 / ~를 반영하여↓
,をかんがみて,,1,,명사+,~를 감안하여 / ~를 반영하여↓
,にしたって,,1,,동사 보통형+/ い형용사 보통형+/ な형용사 어간(+である)+/ 명사(+である)+,~라도 / ~인 경우에도↓
,にしたところで,,1,,동사 보통형+/ い형용사 보통형+/ な형용사 어간(+である)+/ 명사(+である)+,~라고 한들 / ~이라도↓
,に忍びない,,1,,동사 사전형+,(차마) ~할 수 없다↓
,ないではすまない,,1,ないではすまない、ずにはすまない,동사 ない형+,~하지 않으면 끝나지 않는다 / 반드시 ~해야 한다
```

Ⅴ. 일본어 텍스트 생성 및 JLPT 난이도 분석 인공지능 시스템

JLPT를 포함하여 일본어 교육에 필요한 텍스트를 생산하는 것은 다년간의 경험을 가진 일본어 교육 전문가에게도 쉬운 일은 아니며, 이는 경험이 부족한 교수자에게는 두말할 필요가 없다. 물론 신문 기사, 웹 문서 등 기존의 텍스트를 응용하는 방법도 유효하지만, 저작권 문제는 물론이거니와 문체, 테마, 톤 등 텍스트의 질적 문제도 함께 따라오게 된다. 또한 생성한 텍스트의 어휘와 문법 그리고 한자가 어느 정도의 난이도 즉, 어느 정도의 JLPT 레벨인지 판별하는 것 역시 전문가조차도 어려움을 겪을 수밖에 없다. 이와 같은 어려움을 해결하기 위해 본고에서는 크게 다음의 두 가지 인공지능 시스템의 개발을 시도했다. 첫 번째가 JLPT에 특화된 일본어 텍스트를 생성하는 시스템이며 두 번째가 전 단계에서 생성한 일본어 텍스트를 JLPT 기출문제를 기반으로 난이도를 검증하여 학습용 컨텐츠로 가공하는 시스템이다. 이와 같은 흐름을 도식화하면 다음의 <그림 3>과 같다.



<그림 3> GPT2-JLPT-AOT

1. 파인튜닝을 통한 JLPT 언어모델 구축

우선 아래 <표 3>과 같이 사전학습 완료된 GPT2-Japanese의 세 언어모델

(small, medium, large)에 대해, 3.3절에서 언급한 바와 같이 구JLPT의 독해·문법 영역 기출문제의 지문을 추출하여 정제한 후, 전처리 과정을 거쳐 생성한 JLPT-DS-AOT 데이터셋으로 머신러닝(파인튜닝)을 수행했다.

<표 3> JLPT-DS-AOT 데이터셋 예시

マンションの十階の部屋を出てエレベーターを待っていたら、扉が開くやいなや、ニャーといって猫が降りてきた。人はだれも乗っていない。驚いて猫を見ていると、猫は別段あわてる様子もなく、非常階段のほうへすたすた歩いていってしまった。屋上のほうを目指しているようだった。一体どうやって乗ってきたのだろう。人が乗っていれば追い出すだろうから、無人で扉が開いたときに乗ってしまったのだろうか。そうだとすれば、ボタンが押されるまで、ずっと中でおとなしく待っていたに違いない。<|endoftext|>
一時間後、用をすませてマンションに戻り、十階に着いてエレベーターの扉が開いた瞬間、ニャーといってさっきの猫が中に入ってきた。屋上を見物したあと階下に降りるべく、エレベーターの前で待っていたらしい。化かされているような気がしたが、一階まで送ってやると、ものたりなそうな顔をしながら、外のやみに消えていった。<|endoftext|>
ちょっとありえないような不思議な話だ。でも、そういえば、ひとりで電車に乗ってる猫もいたっけ。出張の旅先でローカル線に乗っていたら、車両の中に猫がいたのだ。ちょこなんと座っているので、だれかが連れてきたのだろうと思っていたら、ドアが開くなり、さっとホームに飛び降り、改札口から外へ出て、そのままどこかへ行ってしまった。土地の人にその話をしたところ、「ここではわりと知られている猫なんですよ。」とのことだった。猫は猫なりにどこかに用事でもあったのだろうか。<|endoftext|>
…… 후략 ……

단, <표 4>와 같이 파인튜닝을 수행하는 데에 있어 컴퓨터 사양의 문제와 과적합(Overfitting[9]))을 방지하기 위해 각각 300회의 학습으로 한정했으며, 그 결과 생성한 JLPT 언어모델 – 「GPT2-JLPT」는 다음의 (6)과 같으며, (6)d의 웹페이지에도 이를 공개해 두었다.

9) 언어 모델에 파인튜닝을 수행할 때, 학습용 데이터셋을 과도하게 암기하여 데이터셋에 들어 있는 노이즈까지 학습한 상태를 가리킴.

<표 4> 파인튜닝 – 'GPT2-JLPT_small'의 경우

```
2021-07-14 03:31:28.778299: I tensorflow/stream_executor/platform/default/dso_loader.cc:53]
Successfully opened dynamic library libcudart.so.11.0
2021-07-14 03:31:30.520511: I tensorflow/core/platform/cpu_feature_guard.cc:142] This
TensorFlow binary is optimized with oneAPI Deep Neural Network Library (oneDNN) to use
the following CPU instructions in performance-critical operations:  AVX2 FMA
To enable them in other operations, rebuild TensorFlow with the appropriate compiler flags.
2021-07-14 03:31:30.523929: I tensorflow/stream_executor/platform/default/dso_loader.cc:53]
Successfully opened dynamic library libcuda.so.1
2021-07-14 03:31:30.533878: E tensorflow/stream_executor/cuda/cuda_driver.cc:328] failed call
to cuInit: CUDA_ERROR_NO_DEVICE: no CUDA-capable device is detected
2021-07-14 03:31:30.533935: I tensorflow/stream_executor/cuda/cuda_diagnostics.cc:156] kernel
driver does not appear to be running on this host (ab5560f78c04): /proc/driver/nvidia/version
does not exist
2021-07-14 03:31:32.588418: I tensorflow/core/platform/profile_utils/cpu_utils.cc:114] CPU
Frequency: 2299995000 Hz
Loading checkpoint gpt2ja-small/model-10412700
Loading dataset...
Training...
[1 | 7.26] loss=3.40 avg=3.40
[2 | 10.02] loss=2.97 avg=3.18
[3 | 12.64] loss=3.27 avg=3.21
                                        …… 중략 ……
[298 | 762.14] loss=1.51 avg=2.08
[299 | 764.74] loss=1.65 avg=2.08
[300 | 767.31] loss=1.59 avg=2.07
Saving checkpoint/gpr2ja-small-finetune-210714/model-301
```

(6) GPT2-JLPT 언어모델

a. GPT2-JLPT_small (gpr2ja-small-finetune-210714)
b. GPT2-JLPT_medium (gpr2ja-medium-finetune-210714)
c. GPT2-JLPT_large (gpr2ja-large-finetune-210714)
d. 언어모델 배포 -「4.2 GPT2-JLPT 언어모델」
http://www.japanese.or.kr/JapaneseNLP_main.aspx

2. GPT2-JLPT에 의한 자동 일본어 텍스트 생성

본고에서는 인공신경망 자연어처리 언어모델을 사용하여 일본어 텍스트를

자동으로 생성할 수 있는 프로그램[10]을 파이썬을 기반으로 작성하고, 상기 (1)과 같은 「GPT2-Japanese」의 세 언어모델(small, medium, large)을 사용하여 500자 이내의 일본어 텍스트를 각각 100건씩 생성케 했다. 마찬가지로 파인튜닝을 수행하여 구축한 상기 (6)과 같은 「GPT2-JLPT」의 세 언어모델을 사용하여 500자 이내의 일본어 텍스트를 각각 100건씩 생성케 했다. 이렇게 생성한 총 600건의 텍스트는 웹페이지[11]에서 확인할 수 있으며, 인공지능에 의해 생성된 텍스트에는 본고의 분량 관계상 무의미한 개행을 삭제하는 것을 제외하고는 어떤 가공도 하지 않았다. 각각의 언어모델 별 첫 번째 생성 텍스트는 아래 <표 5·6>과 같다.

<표 5> GPT2-Japanese 기반 자동 생성 일본어 텍스트

1. GPT2-Japanese – gpt2ja-small : GPT2JAP_small.txt – 24,492자 / 1,279단어

1.6kgの缶缶チューブが入っています。
この缶チューブに水が入っているため、水が入っているのに水が引っかかり、入らないようです。
缶チューブが入っていない缶チューブは、水分が出てしまうといいです。
他の方も仰っていますが、缶チューブだけの水なんて出せますか?
また、缶チューブの水分が出てしまっているのでしょうか?
ちなみに、缶チューブは水と一緒に出てくるので気になりました。
ただ、缶チューブと水だけのものは缶チューブを使用しています。
ご回答よろしくお願いします

2. GPT2-Japanese - gpt2ja-medium : GPT2JAP_medium.txt – 29,395자 / 1,814단어

「それで、これからどうするんだよ」
「うーん……」
彼は悩ましげに唸ると、その左手で顎の下の髪を耳にかけて掻き回す。「……う、うむ……これで、よい」
彼は少しためらいもなくそんなことを言う。
彼女は思った。『そのままでいいなんて言ってはいけない』と。
そして再び彼女は考え込む。「まあ、いい。これでこそオレ様が正しい」
「やっぱり、私もそうなってもいいですか?」
「ん……ああ、お願いする」
そうして、彼女は彼と二人、この城で一人の男女の生活をしていく。
ある日、二人を見つけた商人の商人が、「ようこそ『神の館』へ」と二人を出迎える。

10) 「gpt2-jptxtgen_aot_auto.ipynb」
11) GPT2-JLPT TextData - http://www.japanese.or.kr/JapaneseNLP-JText.aspx

「ようこそ『神の館』へ」という男は、そう言った男のことを知っているらしい。
そして男は言う。『お前がここにいるということは、神は何らかの罰を受けたようだな』と。
その罰は、『呪いにかかった』という、簡単なものだった。
「──いや、お前も呪いをかけられている？ ……いや、そんなはずはないだろう。このオレ様が裁きなど受けずに、誰かの呪いをかけたとでもいうのか？」
そのような男の心の声が聞こえて、男は言う。『神を裁くというのは、神ではなくて、その人を裁くということだ

3. GPT2-Japanese – gpt2ja-large : GPT2JAP_large.txt – 26,478자 / 1,504단어

ご訪問ありがとうございます
老け顔にさよならサロン~
ゆったり~のメニューとお申込み
ビフォーアフターご感想　人気記事一覧

<표 6> GPT2-JLPT 기반 자동 생성 일본어 텍스트

1. GPT2-JLPT_small : GPT2JLPT_small.txt – 12,646자 / 326단어

いいこともいっぱいありますよ。たとえば、料理はおいしいし、おいしい。

2. GPT2-JLPT_medium : GPT2JLPT_Medium.txt – 10,474자 / 418단어

「もう、そんな時間か。そろそろ帰ろう」

3. GPT2-JLPT_large : GPT2JLPT_large.txt – 16,735자 / 254단어

「私も子供の頃はそういう気持ちが強かったよ。だから、子どもの自分を大切にしないこと。自分に自信を持って、彼に勝とうとかう努力をして、自分が活躍したいと思うことをやろうと思っていた。だけど、彼は違う。この誇りを持って、自分が活躍できる人になろうと思ったのだ。それだけの楽しみをあえて持っていた。ただそれは、自分を救うことで、誰かを救うことにもあるから、だから彼は努力しない。また、自らの損な精神も、ある程度は人より心に影響を与えるものであった。「あれなあに?」

위 <표 5>와 <표 6>에서, 우선 GPT2-JLPT 기반 텍스트가 GPT2-Japanese와 비교해 볼 때 상대적으로 문장의 길이가 짧으며, 글의 주제가 협소하다는 점을 확인할 수 있었다. 그중에서 우선 문장 길이의 경우, GPT2-JLPT가 학습한 데이터셋이 JLPT의 문법 문제의 짧은 지문, 그리고 독해의 짧은 답안을 다수 포함하고 있기 때문이라 판단된다. 한편 <표 5・6>의 각각 생성된 텍스트의 언어량을 관찰하자면, 생성 텍스트의 길이가 언어모델의 크기와 비례하는 것은 아니었다.

이어 주제의 다양성의 차이는 학습 텍스트의 장르에서 기인하는 것으로 판단된다. 구체적으로 오직 JLPT의 다소 경직된 문장을 학습한 GPT2-JLPT와 달리 GPT2-Japanese는 다양한 장르의 글을 포함하고 있는 코퍼스[12]를 학습했기 때문이라고 할 수 있다.

위 두 차이점에 대한 관찰을 통해 머신러닝(파인튜닝)에 따라 인공지능이 어떻게 변화하지는 지를 명확히 확인할 수 있었으며, 이는 곧 머신러닝의 효과와 필요성을 증명하고 있다고 할 수 있겠다. 앞으로 학습 데이터의 양, 종류 그리고 학습 횟수 등 다양한 조건을 통제하여 지속적으로 실험을 수행한다면, 일본어 텍스트 자동 생성에 최적화된 파인튜닝의 초기 설정값을 찾아낼 수 있을 것이며, 이는 인공지능이 더욱 정밀하게 원하는 일본어 텍스트를 생성할 수 있게 하는 하나의 방안이 될 것으로 생각된다.

한편 텍스트가 길어지면 길어질수록 일관되지 못한 텍스트를 생성하는 경우가 많았다. 이는 주제 혹은 키워드가 부여되지 않았기 때문이 아닐까 가정하여, 앞선 실험에서는 인공지능으로 하여금 글자 수의 제한 이외에 어떠한 제약도 두지 않고 자유롭게 텍스트를 생성시켰으나 이번에는 인공지능에게 「学校」라는 키워드를 주고 각각 1개씩의 텍스트를 생성시켜 보았으며 그 결과는 다음과 같다.

<표 7> GPT2-Japanese 기반 자동 생성 일본어 텍스트 – 키워드 学校

1. GPT2-Japanese - gpt2ja-small

学校の友達と行ったんだけど、その子がいなくて。その子とは、仲良くしようとして、家を探していて。でも、どうしたらいいか分からなくて。でも、誰も見つけられないと思って、一人で探していた矢先に声を掛けてきて、声に反応できるようになってきたんだ。だから、何を聞いても、どうしていいか、分からなくて」

「それは……、いいの……？」

「何も知らなかった時は本当に、ひどくうろたえた、って思っていたから」

「……」

私の話、聞いていたのか。私に話してくれた人、かなり気まずい空気の中を、ここまで話し始めて、その雰囲気に飲まれていたのに。

「けど。その後、ずっとかみしんだような、寂しさを感じてた。そして、気がつくとその子と仲良くしていたらしくて。僕が見つけてくれた子は、その子の好きな人だったんだ。……僕も、初めて女の子を見たけど、本当の可愛いところは全然見えなくて。ただ、

12) 개인·기술 블로그, 질문·기사 모음·채점 사이트, 웹 사전, 뉴스, 웹소설이 포함된 「コーパス2020」

その子に自分の見える範囲の可愛さは全くなかった。多分、可愛い姿を見せるために女の子を集めてたのだと思うけど、その姿がどうしても好きになれずに、ずっと一人で彷徨っていたんです、けど。やっぱり、そんなのダメだと思って、諦めようとした瞬間でし

2. GPT2-Japanese - gpt2ja-medium

学校、小・中学校に通う場合、各学校の各学年に1日乗車券(ICO)が配布されています。
※ICOはICTAの規定に基づき発行される乗車券で、発売所によってはすでに完売しているものも存在します。
※ICOは、<DATE>現在のものです。
ICO回数を「1,500」から「2,000」に増加させるシステムです。
(1)　初回購入時
・ICO回数が「1,500」から「2,000」に推移している時、「乗車券」を「2,000」のいずれかの回数に、乗車することができます。また、「乗車券」の有効期間中に利用できるといったプラスの使い道も存在します。(詳細はUAISのページをご確認ください)
・有効期間の最終有効利用日の有効期間中に「ICO」を利用した場合、その回数と有効期間の合計が「1,500」となります。(詳細はUAISのページをご確認ください)
また、有効期間の最終利用日の利用条件も「2,000」となるので、「乗車券」利用後に「ICO」利用する場合は「乗車券」の有効期間終了後に「ICO」回数の有効期限が切れます。

3. GPT2-Japanese - gpt2ja-large

学校の教育費の一部を児童買会に寄付する寄付から総額を調べあげる寄付への繰越利益を寄付した業者より収益を得られる所得益の一部を家庭に支払う税金に、NHKの老人放送局が消費者の支払い義務を守っている。なお、家庭内の社会問題の分かる範囲で所得を減じさせるという社会的差別化がある。「他人に注意するのもいい。注意しないなら自分の利益のために相手の利益のために自己を金で買うべきだ」

<표 8> GPT2-JLPT 기반 자동 생성 일본어 텍스트 – 키워드 学校

1. GPT2-JLPT_small

学校へ行きなさい。きっと、あなたには分かっているはずだ。

2. GPT2-JLPT_medium

学校の食堂も、いつか自分の家のようになるだろうと思っていた。だが、そんな当てが外れたことで、少し意外なところにいいものが入ったのだと、残念に思っていた。

3. GPT2-JLPT_large

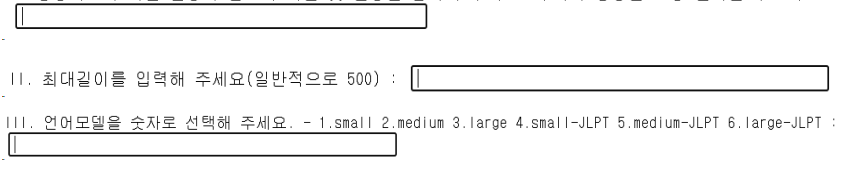
学校の教育費の一部を児童買会に寄付する寄付から総額を調べあげる寄付への繰越利益を寄付した業者より収益を得られる所得益の一部を家庭に支払う税金に、NHKの老人放送局が消費者の支払い義務を守っている。なお、家庭内の社会問題の分かる範囲で所得を減じさせるという社会的差別化がある。「他人に注意するのもいい。注意しないなら自分の利益のために相手の利益のために自己を金で買うべきだ」

상기 <표 7・8>의 텍스트를 통해, 인공지능에 키워드를 주고 키워드와 연관된 텍스트를 생성하게 하는 것이 아무런 힌트 없이 텍스트를 생성하는 것 보다 주제의 통일성, 그리고 문장의 완성도 측면에서 바람직하다는 것 또한 확인할 수 있었다.

3. JLPT 난이도 분석

본고에서는 앞선 절에서의 실험을 통해 사전 언어모델을 적절한 통제조건 하에서 파인튜닝을 실시하여 재구성한 언어모델로 일본어 텍스트를 생성하되, 키워드를 부여하여 일관성 있고 주제에 부합하는 텍스트를 생성할 필요가 있다는 결론에 도달했다. 또한 생성된 텍스트는 앞선 예시에서 확인할 수 있는 바와 같이 아직 인간처럼 완벽한 문장을 구사할 수 없으므로 생성된 텍스트 일부분을 발췌하는 방식으로 적합한 텍스트를 얻어내고 이를 JLPT 난이도 분석에 사용한다면 JLPT의 학습 및 교육에 필요한 텍스트를 추출할 수 있을 것이라 기대할 수 있겠다.

이에 본고에서는 우선 <그림 4>와 같이 키워드, 문장 길이, 언어모델을 선택하여 인공지능에 의한 테스트 생성을 수행할 수 있도록 검증 프로그램을 설계했으며, <그림 5>와 같이 「私は」와 같은 키워드 혹은 첫 단어, 500자 이내, 언어모델은 GPT2-JLPT-large를 선택하여 다음과 같은 일본어 텍스트를 자동 생성할 수 있었다.



<그림 4> 텍스트 자동 생성을 위한 조건 설정 화면 1

I. 생성하고자 하는 문장의 힌트가 되는 첫 문장을 입력해 주세요. 무작위 생성은 그냥 엔터를 누르세요 : 私は
II. 최대길이를 입력해 주세요(일반적으로 500) : 500
III. 언어모델을 숫자로 선택해 주세요. - 1.small 2.medium 3.large 4.small-JLPT 5.medium-JLPT 6.large-JLPT : 6
*선택하신 언어모델은 checkpoint/gpr2ja-large-finetune-210714입니다

<그림 5> 텍스트 자동 생성을 위한 조건 설정 화면 2

<표 9> 키워드, 글자 수, 언어모델 설정을 기반, 자동 텍스트 생성 예시

私は何を言いたいのだろうか。頭の処理が追いつかず、言葉が記憶されなくなるくらい複雑になる。これは、言葉が頭に残っているからだろうか。言葉を失った後である。だから、心のどなげを失う形ができるのかもしれない。考えてみると、これは自分のことを考えから変えたつもりだったが、その決心は鈍にせよ残りの五、六年のうちに打つことができた。

이어, <표 9>의 텍스트 중 일부인,「私は何を言いたいのだろうか。頭の処理が追いつかず、言葉が記憶されなくなるくらい複雑になる。これは、言葉が頭に残っているからだろうか。」부분을 발췌하여 <그림 6>과 같이 JLPT 난이도 분석 시스템을 통해 분석을 시행했으며, 그 결과는 다음의 <그림 7・8・9>와 같다.

◎ AI 딥러닝 Tensor-AOT에 의해 자동 생성된 일본어 텍스트는 다음과 같습니다.

1. 私は何を言いたいのだろうか。頭の処理が追いつかず、言葉が記憶されなくなるくらい複雑になる。これは、言葉が頭に残っているからだろうか。言葉を失った後である。だから、心のどなげを失う形ができるのかもしれない。考えてみると、これは自分のことを考えから変えたつもりだったが、その決心は鈍にせよ残りの五、六年のうちに打つことができた。

2. 私は何を言いたいのだろうか。頭の処理が追いつかず、言葉が記憶されなくなるくらい複雑になる。これは、言葉が頭に残っているからだろうか。言葉を失った後である。だから、心のどなげを失う形ができるのかもしれない。考えてみると、これは自分のことを考えから変えたつもりだったが、その決心は鈍にせよ残りの五、六年のうちに打つことができた。

==========

IV. 사용하고자 하는 텍스트 번호를 입력해 주세요. 직접 입력하고자 한다면 3번을 입력하고, 직접 텍스트를 잘라붙여 주세요 : 3
IV-2. 직접 텍스트를 잘라붙여 주세요 : 私は何を言いたいのだろうか。頭の処理が追いつかず、言葉が記憶されなくなるくらい複雑になる。これは、言葉が頭に残っているからだろうか。

==========

<그림 6> JLPT 난이도 분석 시스템을 통한 텍스트 분석

<그림 7> JLPT 난이도 분석 시스템을 통한 텍스트 분석 결과 - 어휘

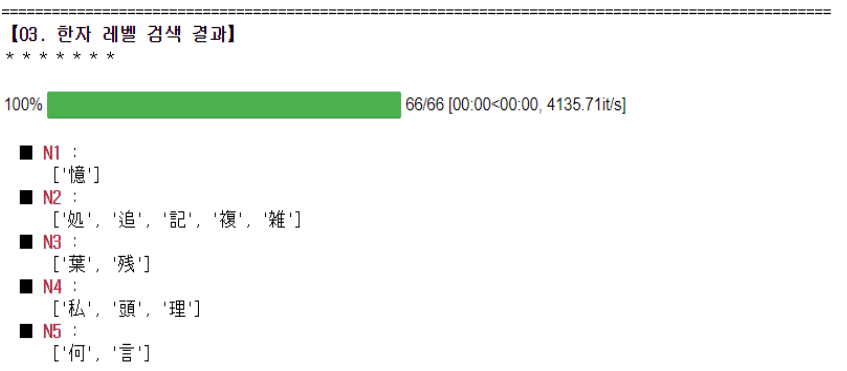
어휘 분석은, <그림 7>과 같이 「출현(사전형), 기출문제 기준 JLPT 레벨(네이버 사전 기준 레벨), 독음, 의미」 순으로 분석 결과를 제시하여, 학습자 혹은 교수자가 원하는 JLPT 레벨에 맞는 어휘와 이를 상회하는 어휘를 구분할 수 있게 되며, 이를 취사선택한다면 특정 레벨의 학습에 적합한 문장으로 정제할 수 있게 된다.

이어 <그림 8>과 같이 문법의 경우, 「해당 문법이 포함된 문장, 유사 표현, 접속 형태, 해석」 순으로 분석 결과를 제시하여 마찬가지로 학습자와 교수자 모두에게 JLPT 학습 혹은 학습자료 작성에 도움을 줄 수 있게 되며, 특정 레벨의

학습에 적합한 문장의 정제도 가능하게 했다.

<그림 8> JLPT 난이도 분석 시스템을 통한 텍스트 분석 결과 - 문법

마지막 한자 레벨 검색 결과는 <그림 9>와 같이, 급수별로 구분하여 제시했는데, 그 기준은 기본 JLPT가 제시한 한자 기준에 <표 1>과 같은 기출문제의 한자 출현 이력을 통해 본고에서 독자적으로 부여한 기준을 기반으로 설정했다. 이를 통해 급수별 학습자료 등 일본어 교육용 텍스트의 적절한 한자의 사용 여부를 한눈에 확인할 수 있게 된다. 참고로 본 연구를 통해 개발된 「JLPT 난이도 분석 시스템」은 (5)c를 통해 연구자 한정으로 추후 공개할 예정이다.



<그림 9> JLPT 난이도 분석 시스템을 통한 텍스트 분석 결과 - 한자

Ⅵ. 맺음말

본 연구에서는 인공신경망 자연어 처리 플랫폼 GPT-2를 기반으로 일본어능력시험 기출문제 데이터셋을 통해 파인튜닝을 수행하여, 인공지능을 통한 일본어능력시험 학습 및 교수를 위한 텍스트 생성의 가능성을 타진하고, 한발 더 나아가 생성된 텍스트에 대한 일본어능력시험 기준 난이도 분석을 수행할 수 있는 시스템의 개발을 수행했다. 이를 통해 인공지능을 통한 일본어 텍스트의 자동 생성에는 아직 한계가 많이 있다는 점을 확인함과 동시에, 머신러닝(파인튜닝) 및 조건 통제를 통해 수준 높은 일본어 텍스트를 생성할 수 있는 언어모델 개발의 가능성을 확인할 수 있었다. 또한 이와 같은 텍스트를 단순히 생성하는 것에서 그치지 않고 텍스트에 대한 어휘, 문법, 한자 차원의 JLPT 난이도 분석까지 수행할 수 있는 시스템을 개발할 수 있었고, 이를 통해 앞으로 인공지능을 통한 일본어학 및 일본어 교육 분야에의 응용 가능성 또한 확인할 수 있었다. 금후 본 연구를 통해 알게 된 사실을 기반으로 인공신경망 일본어 생성 시스템 개선을 수행하여 본 시스템의 일반 공개를 목표로 지속적 연구를 수행하고자 한다. 또한 금번 연구에서는 텐서플로를 기반으로 실험 및 시스템 개발을 수행했으나, 또 다른 인공지능 플랫폼인 파이토치 등 새로운 시스템의 도입도 적극적으

로 고려하여 상호 비교・대조 분석을 통해 인공지능에 의한 일본어 텍스트 생성 뿐만 아니라 성능 향상 또한 꾀하고자 한다.

【参考文献】

金蘭美「メール文の自動採点のための評価基準とタスクータスクの達成の可否と必須語に焦点を当てて」,『일본어학연구일본어학연구』제48집, 한국일본어학회, 2016, pp.3-19

김유영「일본어 텍스트 가독성 분석 ―구JLPT 기출문제 기반 일본어텍스트의한자 및 한자 어휘 가독성 판단 프로그램의 개발―」,『言語文化』27, 韓国日本言語文化学会, 2014, pp.211-236

金暎泳「日本語のテキストの可読性レベル分析 ― 旧日本語能力試験の既出問題データにおける統計的検証を中心に」、『日本學報』Vol.103、韓国日本学会、2015、pp.21-40

김유영「텍스트마이닝 기법을 활용한 일본 미디어의 한국 뉴스의 감정 추이에 대한 분석-파이썬을 활용한 단어감정극성대응표 분석기법의 수행을 통해-」,『日本語學研究』第65輯, 2020, pp.25-43

박강훈「인공신경망 번역 엔진을 활용한 멀티링구얼 문법 교육―韓・日・英 삼중언어를 중심으로―」,『일본문화학보』80, 한국일본문화학회, 2019, pp.23-43

신민철「인공지능(AI) 번역기를 활용한 한일 한자어 비교의 가능성 모색」,『일본근대학연구』71권, 한국일본근대학회, 2021, pp.97-112

이상민「인공신경망 기계번역의 오류분석 및 퀄리티 평가」,『외국어교육연구』34권4호, 한국외국어대학교 외국어교육연구소, 2020, pp.129-153

이치훈・이연지・이동희「사전 학습된 한국어 BERT의 전이학습을 통한 한국어 기계독해 성능개선에 관한 연구」,『韓國IT서비스學會誌』제19권 제5호, 2020, pp.83-91

A Vaswani, N Shazeer, N Parmar, J Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Attention Is All You Need, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA., 1-11

Samuel A.L. “Some studies in machine learning using the game of checkers.” IBM J Res Dev. 1959; 3: 210-229

T. Mitchell, “Machine learning”, McGraw Hill, 1997

コラム「自然言語処理モデル「GPT-3」の紹介」人工知能(AI)、エヌ・ティ・ティ・データ先端技術株式会社
https://www.intellilink.co.jp/column/ai/2021/031700.aspx(검색일: 2021.07.12)

コーパス2020 https://github.com/tanreinama/gpt2-japanese/blob/master/report/corpus.md (검색일: 2021.07.12)

BERT – google-research/bert https://github.com/google-research/bert(검색일: 2021.07.12)
Colab – Colaboratory https://research.google.com/colaboratory/faq.html(검색일: 2021.07.12)
gpt2-japanese - Japanese GPT2 Generation Model
https://github.com/tanreinama/gpt2-japanese(검색일: 2021.07.12)
openAI https://openai.com(검색일: 2021.07.12)
TensorFlow「TensorFlow를 사용해야 하는 이유」
https://www.tensorflow.org/about?hl=ko(검색일: 2021.07.12)

▣ 접 수 일 : 2021년 07월 15일
심사개시 : 2021년 07월 28일
심사완료 : 2021년 08월 15일
게재결정 : 2021년 08월 20일

〈要旨〉

人工知能による日本語のテキストの生成及び日本語能力試験の難易度分析システムの開発に関する研究

本研究では人工神経網の自然語処理フラットフォームであるGPT-2を基盤として，人工知能を利用して日本語能力試験の学習と教育のためのテキストの生成の可能性を確認し，そこからもう一歩進んで生成されたテキストに対する日本語能力試験の難易度分析を遂行できるシステムの開発を遂行した。

このような研究を通じで，人工知能を利用した日本語のテキストの自動生成にはまだ限界があるものの，マシンラーニング(ファインチューニング)及びその条件の統制を通じてレベルの高い日本語のテキストを生成できる言語モデルの開発の可能性を確認できた。また，このようなテキストを単純に生成することに留まらず，テキストに対する語彙･文法･漢字のJLPTにおける難易度の分析まで実施できるシステムを開発できた点で，人工知能の日本語学及び日本語の教育分野に対する応用の可能性は計り知れないという点を確認できた。

Research on the development of artificial intelligence systems for the generation of Japanese texts and the difficulty analysis of the Japanese Language Proficiency Test

In this study, we confirmed the possibility of generating textbooks for learning and teaching the Japanese Language Proficiency Test using artificial intelligence, based on GPT-2, which is a natural language processing flat form of an artificial neural network. In addition, we went one step further and developed a system that can perform difficulty analysis of the Japanese Language Proficiency Test for the generated text.

Through such research, although there is still a limit to the automatic generation of Japanese text using artificial intelligence, it is possible to generate high-level Japanese text through machine learning (fine tuning) and control of its conditions. We were able to confirm the possibility of developing a “language model”. In addition to simply generating such texts, we were able to develop a system that can analyze the difficulty of vocabulary, grammar, and kanji in JLPT for texts. It was confirmed that the possibility of application to the educational field is immeasurable.