JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 9, pp. 69-77, September 2024
https://doi.org/10.9708/jksci.2024.29.09.069

Evaluating the Impact of Training Conditions on the Performance of GPT-2-Small Based Korean-English Bilingual Models

Euhee Kim*, Keonwoo Koo**

*Professor, Dept. of Software Convergence, Shinhan University, Gyeonggi-do, Korea
**Lecturer, Dept. of English language and literature, Dongguk University, Seoul, Korea

[Abstract]

This study evaluates the performance of second language acquisition models learning Korean and English using the GPT-2-Small model, analyzing the impact of various training conditions on performance. Four training conditions were used: monolingual learning, sequential learning, sequential-interleaved learning, and sequential-EWC learning. The model was trained using datasets from the National Institute of Korean Language and English from BabyLM Challenge, with performance measured through PPL and BLiMP metrics. Results showed that monolingual learning had the best performance with a PPL of 16.2 and BLiMP accuracy of 73.7%. In contrast, sequential-EWC learning had the highest PPL of 41.9 and the lowest BLiMP accuracy of 66.3%($p < 0.05$). Monolingual learning proved most effective for optimizing model performance. The EWC regularization in sequential-EWC learning degraded performance by limiting weight updates, hindering new language learning. This research improves understanding of language modeling and contributes to cognitive similarity in AI language learning.

▸ **Key words**: GPT-2 model, sequential learning, continual learning, programmed plasticity, second langauge acquisition modeling

[요 약]

본 연구는 GPT-2-Small 버전 모델을 사용하여 한국어와 영어를 학습하는 이중 언어 모델의 성능을 평가하고, 다양한 학습 조건이 모델 성능에 미치는 영향을 분석하였다. 연구 방법으로 단일 언어 학습, 순차 학습, 순차-교차 학습, 순차-EWC 학습의 네 가지 조건을 설정하여 모델을 훈련하였다. 국립국어원 말뭉치와 영어 위키피디어 말뭉치를 사용하고, PPL과 BLiMP 지표를 통해 성능을 측정하였다. 연구 결과, 단일 언어 학습 조건에서 PPL 값은 16.2, BLiMP 정확도는 73.7%로 가장 우수한 성능을 보였다. 반면, 순차-EWC 학습 조건에서는 PPL 값이 41.9로 가장 높았고, BLiMP 정확도는 66.3%로 가장 낮았다($p < 0.05$). 단일 언어 학습이 이중 언어 모델 성능 최적화에 가장 효과적임을 확인하였다. 이는 결정적 시기 이론에 따라 모델이 단일 언어에 최적화될 때 더 나은 성능을 보인다는 것을 의미한다. 또한, 프로그래밍 가소성을 조절하는 EWC 정규화를 적용한 지속 학습 조건에서는 성능 저하가 두드러졌는데, 이는 정규화가 가중치 업데이트를 제한하여 새로운 언어 학습 능력을 저하시켰다는 것을 의미한다. 본 연구는 언어 모델링에 대한 이해를 높이고, AI 언어 학습에서 인지적 유사성을 개선하는 데 기여한다.

▸ **주제어**: 이중 언어 모델링, GPT-2 모델, 순차 학습, 지속 학습, 프로그래밍 가소성

• First Author: Euhee Kim, Corresponding Author: Keonwoo Koo
*Euhee Kim (euhkim@shinhan.ac.kr), Dept. of Software Convergence, Shinhan University
**Keonwoo Koo (rjsdnrn@gmail.com), Dept. of English language and literature, Dongguk University
• Received: 2024. 08. 05, Revised: 2024. 09. 12, Accepted: 2024. 09. 21.

Copyright © 2024 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

신경망 언어 모델의 발전은 자연어 처리 분야에서 큰 변화를 가져왔으며, 특히 GPT-2와 같은 거대 언어 모델(LLM)은 다양한 언어 처리 작업에서 뛰어난 성능을 보여주고 있다. 이러한 모델들은 인간의 언어 학습 과정을 모방하며, 자연 언어 생성, 번역, 요약 등 다양한 응용 분야에서 사용되고 있다. 그러나 인간의 언어 학습과 비교하여, 언어 모델이 언어를 학습하는 방식과 그 효과는 여전히 많은 연구가 필요한 부분이다[1].

결정적 시기(Critical Period, CP) 이론에 따르면, 청소년기 이후에는 제2언어 습득 능력이 감소하고 제1언어는 잘 잊히지 않는 경향이 있어, 이는 언어 학습 능력의 발달과 관련된 중요한 논쟁의 주제이다. 신경망 언어 모델의 발전은 이러한 논쟁에 활용될 수 있으며, 신경망 언어 모델이 인간과 유사한 문법 판단을 학습할 수 있음을 보여주고 있다[2].

기존 연구들은 주로 인간의 언어 학습을 관찰하는 데 초점을 맞추었으며, 이러한 연구들은 선천적 학습 구조를 가진 인간을 연구 대상으로 삼았다. Pallier et al.의 연구는 CP 동안의 언어 습득과 관련된 여러 현상을 관찰하였으나, 그 원인이 선천적인지 경험적인지에 대해서는 명확한 결론을 내리지 못했다. 최근에는 Warstadt et al.의 연구가 신경망 언어 모델을 활용하여 이러한 문제를 새로운 관점에서 접근하고 있으며, Zeng et al.의 연구는 사전 학습된 이중 언어 모델이 프로그래밍 가소성 조절과 정규화 기법을 통해 성능을 최적화하는 새로운 방법을 제시했다. 또한, Constantinescu의 연구는 GPT-2와 RoBERTa 신경망 언어 모델을 사용하여 서로 다른 언어들 간의 유사성이 제2언어 학습의 난이도에 미치는 영향을 연구하였으며, 제2언어에서의 성능이 제1언어와의 구문적 유사성과 높은 상관관계가 있는 언어를 순차적으로 학습할 때 CP 효과가 나타나지 않지만, 학습 도중 프로그래밍 가소성 감소를 통한 정규화 기법을 적용하면 CP 효과를 역설계할 수 있음을 발견했다[3-6].

따라서 본 연구의 목표는 GPT-2 모델 구조를 선택하여 구문적 유사성이 낮은 상관관계가 있는 한국어와 영어를 순차 학습하는 이중 신경망 언어 모델을 다양한 학습 조건에서 훈련하여 성능에 미치는 영향을 분석하고, 추가적인 엔지니어링을 적용하여 CP 효과를 유도함으로써 언어 모델의 인지적 유사성을 높이고, 인간의 언어 학습 과정을 더욱 정확하게 모방하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 연구 관련 주요 개념을 기술하고, 3장과 4장에서는 모델 제안과 방법론을 설명한다. 5장과 6장에서는 실험 및 실험 결과를 분석하며, 마지막으로 7장에서 결론을 맺는다.

II. Related Works

1. Deep Learning Language Models

최근 딥러닝의 발전으로 인해 순환 신경망(Recurrent Neural Networks, RNN)과 트랜스포머(Transformers)와 같은 복잡한 언어 모델들이 개발되었다. RNN은 입력 시퀀스를 처리하면서 내부 메모리를 유지하여 문맥 정보를 포착하는 구조로, LSTM(Long Short-Term Memory)와 같은 변형이 있다[7]. 그러나 RNN은 장기 종속성 문제와 기울기 소실 문제를 겪을 수 있으므로 이를 해결하기 위해 개발된 트랜스포머 모델은 어텐션(Attention) 구조를 사용하여 입력 시퀀스를 병렬로 처리하며, 긴 종속성을 더 효율적으로 포착할 수 있다[8].

트랜스포머는 BERT와 GPT와 같은 모델들로 발전되었으며, 특히 GPT-2는 단방향 트랜스포머 구조를 사용하여 자연스러운 텍스트 생성을 가능하게 해준다. GPT-2는 주어진 텍스트의 다음 단어를 예측하는 방식으로 작동하며, 이는 인간의 언어 학습 방식과 유사한 구조이다[9].

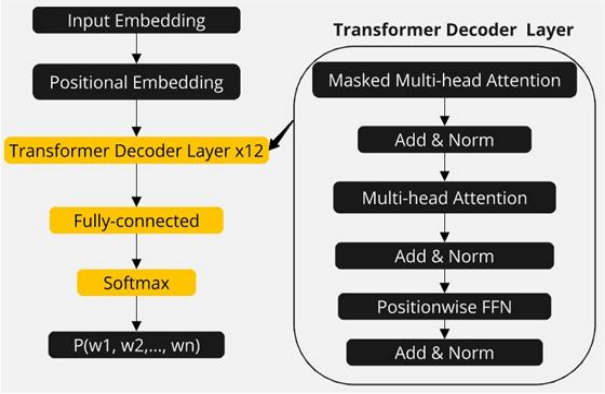

Fig. 1. The Architecture of the GPT-2-Small

Fig. 1은 GPT-2의 가장 작은 버전인 GPT-2-Small의 구조를 시각적으로 나타낸 것이다. GPT-2 모델은 트랜스포머 디코더 구조를 기반으로 하며, 입력 텍스트를 고정된 크기의 벡터로 변환하는 Embedding 계층과, 단어의 위치 정보를 추가하는 Positional Embedding 계층을 포함한다. Transformer Decoder Layer 계층은 입력 벡터를 처리하여 문맥 정보를 학습하며, 최종적으로 Softmax 계층에서 다음 단어를 예측한다.

Devlin et al.이 제안한 BERT(Bidirectional Encoder Representations from Transformers) 모델은 사전 학습된 트랜스포머 인코더를 사용하여 문맥을 양방향으로 이해할 수 있는 구조이다[10]. Pires et al.은 다중 언어 BERT(Multilingual BERT, m-BERT)가 마스크드(masked) 언어 모델링과 다음 문장 예측을 통해 사전 학습된 후, 특정 언어 과제에 맞게 미세조정(Fine-tuning)하여 여러 언어를 동시에 학습하여 언어 간에 공통된 표현을 전이 학습(Transfer Learning)함으로써 다양한 언어 간의 학습 성능을 향상시켜서 언어 간의 번역 및 이해 작업에서 우수한 성능을 보인다고 보고했다[11].

m-BERT 모델은 다양한 언어를 순차적으로 새로운 언어를 학습하는 대신 동시에 다양한 언어를 학습하며 양방향 문맥 이해 능력을 갖추고 있는 반면, GPT-2는 단방향 트랜스포머 구조로 순차적 학습 과정을 더 명확히 관찰할 수 있어, CP 효과 연구와 프로그래밍 가소성 감소 시뮬레이션에 적합하다.

따라서 본 연구에서는 GPT-2-Small 버전 모델 구조를 선택해서 이중 언어의 학습 단계를 명확히 구분하고, 순차 및 지속 학습 방식에 정규화 엔지니어링을 적용하여 언어 모델링하였다.

2. Catastrophic Forgetting, Critical Period & Programmed Plasticity

신경망 이중 언어 모델이 언어를 순차적으로 학습할 때 파국적 망각(Catastrophic Forgetting), 결정적 시기(Critical Period, CP) 효과, 프로그램 가소성(Programmed Plasticity) 문제가 발생하여 모델의 성능에 부정적인 영향을 미칠 수 있다.

파국적 망각은 새로운 언어를 학습하면서 기존 언어 지식을 잃어버리는 현상으로, Kirkpatrick et al.는 Elastic Weight Consolidation (EWC) 정규화 기법을 사용하여 신경망에서 파국적 망각을 해결하는 방법을 제안하였다[12]. EWC 기법은 피셔 정보 행렬(Fisher Information Matrix)을 사용하여 각 파라미터의 중요도를 평가하고, 중요한 파라미터가 크게 변하지 않도록 정규화 항을 손실 함수에 추가하는 방식이다. 손실 함수는 수학적으로 다음과 같이 정의된다.

$$L(\theta) = L_{new}(\theta) + \frac{\lambda}{2}\sum_i F_i(\theta_i - \theta_i^*)^2 \quad (1)$$

여기서, $L(\theta)$는 전체 손실 함수, $L_{new}(\theta)$는 새로운 작업에 대한 손실 함수, λ는 정규화 강도를 조절하는 하이퍼파라미터, θ는 현재 모델의 파라미터, θ^*는 이전 작업에서 최적화된 파라미터, F_i는 파라미터 θ_i의 중요도를 나타내는 피셔 정보 행렬을 표시한다.

$$F_i = \mathrm{EXP}\left[\left(\frac{\partial L(\theta^*)}{\partial \theta_i}\right)^2\right] \quad (2)$$

여기서, EXP는 기댓값, $\frac{\partial L(\theta^*)}{\partial \theta_i}$는 손실 함수 L를 파라미터 θ_i로 편미분한 값이다. 이 값은 특정 파라미터가 손실 함수에 얼마나 중요한지를 나타내며, 중요한 파라미터일수록 큰 값을 가진다.

결정적 시기 효과는 초기 학습 단계에서 특정 언어에 최적화된 모델이 이후 다른 언어를 학습할 때 성능 저하가 발생할 수 있는 현상을 말한다. 이는 모델이 초기 언어 학습 시기에 형성된 구조에 의해 제약을 받을 수 있음을 의미한다. 이를 방지하기 위해 학습 전략의 신중한 설계를 요구한다.

프로그램 가소성은 모델이 초기 언어 학습 동안 형성된 신경망 구조를 기반으로 새로운 정보를 학습하면서도 기존 지식을 유지하고 조정할 수 있는 능력을 의미한다. 모델이 제1언어 학습 동안 형성된 특정 패턴이 제2언어 학습에 방해가 될 수 있으며, 이를 극복하기 위해 모델의 가소성을 조절하는 정규화 기법이 필요하다. Hernandez et al.의 연구는 이중 언어 학습에서 프로그램 가소성을 모델링하여 학습 도중 높은 가소성을 유지하다가 이후 의도적으로 감소시키는 개념을 적용했다[13].

본 연구에서는 제1언어 학습을 한 동일한 GPT-2 모델 구조를 이용하여 제2언어로 연속 학습시킬 때 EWC 정규화 기법을 적용하여 프로그램 가소성을 조절하고, 이를 통해 결정적 시기 효과를 검증하였다. 피셔 정보 행렬을 이용한 EWC 구현 알고리즘은 3장에서 소개한다.

III. Methodology

본 장에서는 인간의 언어 습득 과정을 모방한 신경망 이중 언어 모델링 관련 학습 전략을 제시한다. 먼저, 연구의 주요 주제를 제시하고, 이중 언어 모델의 학습 조건을 구체적으로 기술한다.

1. Research Design

본 연구는 이중 언어 모델링에서 결정적 시기(CP) 효과를 다루며, 두 가지 주요 실험 주제를 탐구한다. 첫째, 제2언어(L2) 학습에 대한 자연적인 CP 효과의 증거를 찾고,

둘째, 프로그래밍 가소성을 조절하여 인간의 언어 습득과 유사한 CP 효과를 유발하는지를 분석한다.

이를 위해 제1언어(L1)와 L2 노출 시점 및 프로그래밍 가소성 수준을 조절하여 단일 언어 학습, 순차 학습, 순차-교차 학습, 순차-EWC 학습의 네 가지 조건을 설정한다.

2. Bilingual Model Training Conditions

신경망 이중 언어 모델 학습 설정에는 먼저 한국어(L1)로 학습을 진행한 후, 주어진 학습 조건에 따라 영어(L2) 학습을 순차적으로 진행하는 방식이 포함된다. 이러한 설정을 통해 L1과 L2의 학습 시점과 방식이 모델 성능에 미치는 영향을 평가하고, 인간의 CP 효과와의 유사성을 분석한다.

또한, L2에 대한 단일 언어 학습 모델(En-Monolingual Model)은 L2 데이터만을 학습할 때 모델의 성능을 평가하는 기준을 제공하여, 이중 언어 학습 모델과 비교할 수 있도록 한다. 여기서, 단일 언어 학습 모델은 하나의 언어에만 노출되는 학습자를 모방한다. 이 모델은 GPT-2-small 구조를 사용하여 단일 언어 L2 데이터를 지정된 에포크(실험에서는 20 에포크)의 2배 동안 연속 학습한 모델이다.

Fig. 2는 이중 언어 모델 생성 과정에서 적용되는 순차 학습, 순차-교차 학습, 순차-EWC 학습의 조건을 시각적으로 설명한 다이어그램이다.

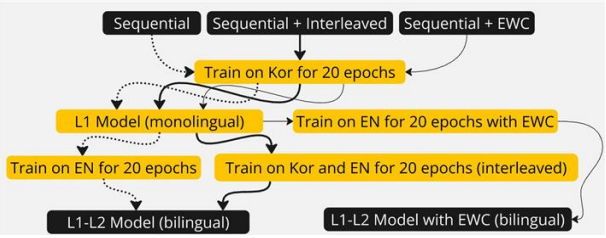

Fig. 2. A Visualization of the Training Conditions for L2

2.1. Sequential Learning

순차 학습(Sequential Learning)은 L1 학습 후 L2 학습을 추가로 진행하는 이중 언어 L2 학습자를 모방한다.

이 방식에서는 먼저 L1 데이터를 사용하여 지정된 에포크 동안 L1 모델을 훈련한 후, L1 모델 체크포인트에서 L2 데이터를 또 다른 에포크 동안 추가로 학습한다. 이 과정을 통해 최종적으로 이중 언어 모델(Bilingual L1-L2 Model)을 생성하게 된다(Fig. 2의 점선). 주목할 점은 L1에서 L2로의 전환이 점진적이지 않고, L1 학습을 완전히 중단한 후 L2로 갑작스럽게 전환된다는 것이다.

2.2 Sequential-Interleaved Learning

순차-교차학습(Sequential-Interleaved Learning)은 L1 학습 후 동시에 두 언어에 노출되는 이중 언어 L2 학습자를 모방한다.

이 방식에서는 L1 데이터를 지정된 에포크 동안 먼저 학습한 후, L2 데이터를 또 다른 에포크 동안 추가로 학습하면서 L1 데이터를 같이 학습한다. 즉, L1 모델 체크포인트에서 L2 교차 훈련을 시작하여, L1과 L2 데이터를 번갈아 가며 훈련한다. 이 과정을 통해 최종적으로 이중 언어 모델(Bilingual L1-L2 Model)을 생성한다(Fig. 2의 굵은 실선).

2.3 Sequential Learning with EWC

순차-EWC 학습(Sequential Learning with EWC)은 새로운 언어를 학습하면서도 기존 언어 지식을 유지할 수 있는 이중 언어 L2 학습자를 모방한다.

이 방식은 먼저 L1을 지정된 에포크 동안 학습한 후, L1 모델 체크포인트에서 L2 데이터를 또 다른 에포크 동안 추가 학습하되, 손실 함수에 EWC 정규화 기법을 적용한다. 이를 통해 L1에서 L2로의 전환 시 가중치 업데이트를 제한하여, 최종적으로 이중 언어 모델(L1-L2 Model with EWC)을 생성한다(Fig. 2의 실선). 이 학습 과정을 통해 성인의 제2언어 습득에서 나타나는 CP 효과를 반영하며, EWC 정규화 기법으로 프로그래밍 가소성을 조절하여 모델이 새로운 언어를 학습하면서도 기존 언어 지식을 손실하지 않도록 파국적 망각을 해결한다.

IV. Proposed GPT-2-Small Based Bilingual Language Modeling

본 장에서는 인간의 언어 습득 과정을 모방한 결정적 시기 효과를 다루기 위해, GPT-2-Small 버전 모델 구조를 기반으로 이중 언어 모델을 구축하고 평가하는 방법을 제안한다. 이를 위해 제1언어와 제2언어의 노출 시점 및 프로그래밍 가소성 수준을 조절하여 3장에서 설명한 단일 언어 학습, 순차 학습, 순차-교차 학습, 순차-EWC 학습의 네 가지 조건에 따라 4개의 모델을 훈련하고 비교 분석하고자 한다.

Fig. 3은 이중 언어 모델을 처음부터 훈련하고 평가하는 과정을 시각적으로 설명한다. 훈련 데이터는 BabyLM 첼린지와 모두의 말뭉치에서 가져오며, 데이터 전처리 과정을 통해 맞춤형 이중 언어 Byte Pair Encoding(BPE) 토

크나이저를 생성한다. 또한, 순차-교차 학습 모델에 사용할 이중 언어 쌍을 위한 교차 데이터 셋도 생성한다. 이후, GPT-2-Small을 기반으로 3장에서 설명한 학습 조건에 따라 훈련을 진행하며, 특히 Sequential-EWC 조건에서는 EWC 알고리즘을 적용하여 훈련을 진행한다. 훈련 설정 단계에서는 하이퍼파라미터 튜닝과 EWC 알고리즘이 포함되며, 훈련된 모델은 PPL, BLiMP-accuracy 등의 지표를 사용하여 평가된다.

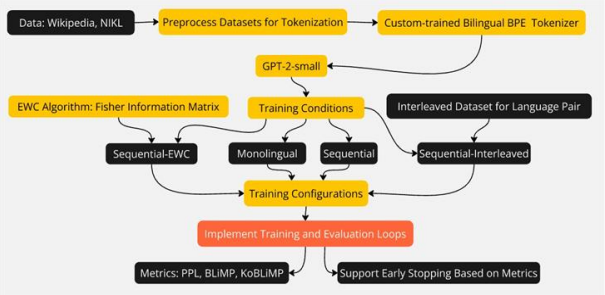

Fig. 3. Proposed GPT-2-Small Based Bilingual Modeling

V. Experiments

본 장에서는 Fig. 3의 이중 언어 모델링 구현과 관련된 실험 환경 구성, 데이터 수집, 전처리, 다양한 훈련 조건, EWC 알고리즘, 하이퍼파라미터 설정 및 평가 방법 등을 기술한다. 또한, 3장에서 제안한 다양한 학습 조건을 이용하여 언어 모델링을 위한 기본 모델을 훈련하며, 이 과정에서는 미세 조정을 하지 않는다.

1. SW Library & HW Environment

이중 언어 모델링 작업의 효율성을 높이고, 전처리 및 평가 과정을 체계적으로 수행하기 위해 여러 가지 라이브러리를 사용하였다. 첫째, 언어 모델링 구현에 사용한 Constantinescu의 제안한 최적화 기법과 PyTorch와 Hugging Face Transformers 라이브러리를 사용하여 이중 언어 모델 훈련, BabyLM 챌린지에서 제공된 평가 파이프라인을 사용하여 성능 평가, BPE Tokenizers 라이브러리를 사용하여 특정 데이터셋에 맞춘 맞춤형 BPE 토크나이저를 수행하였다[14-18].

또한 이중 언어 모델링 구현 실험에 사용한 하드웨어는 각 실험 실행의 자원 할당 및 총 자원 사용량을 Table 1에 정리하였다.

Table 1. Hardware for Experiment Execution

Resource Type	Description	
GPU	NVIDIA GeForce RTX 5000 X 4	
CPU	i9-10980XE	
RAM	32 GB	
Training Run	16 CPUs with 4 GPUs	
Evaluation Run	4 CPUs with 16GiB RAM	
Training Time (Epoch/Hours)	En-Monolingual 40/33	
	Ko-En Sequential Model	20/32
	Ko-En Sequential-Interleaved	20/38
	Ko-En Sequential+EWC	20/36

2. Language Pairs Selection

본 실험의 주요 목적은 한국어와 영어의 구문적 상이함이 이중 언어 학습에서 결정적 시기(CP) 효과에 미치는 영향을 분석하는 것이다. 이를 위해 한국어-영어 언어 쌍을 선택한 이유는 우선, 영어는 잘 연구된 리소스를 통해 계산 비용을 줄일 수 있다. 한국어는 영어와 전혀 다른 언어 계통에 속하여 언어 모델의 수용성을 평가하는 데 유리하다. 또한, 영어 BLiMP 데이터셋을 쉽게 구할 수 있어 수용성 평가가 가능하다. 이로 인해 다양한 학습 조건에서의 L2 모델 평가가 효율적으로 이루어질 수 있다[19-20].

3. Datasets

한국어-영어를 사용하여 언어 모델을 훈련하기 위해, 영어 데이터는 BabyLM 첼린지에서, 한국어 데이터는 국립국어원의 모두의 말뭉치에서 수집하였다. BabyLM 첼린지란 아이들이 언어를 학습하는데 주로 사용되는 데이터를 제공한다. 또한, BabyLM에서 제공하는 전처리 도구를 사용해 XML 태그를 제거하고 텍스트 파일로 변환하였으며, 국립국어원 말뭉치는 제공된 텍스트 데이터를 사용하였다[19-22].

첫째, BabyLM 데이터셋의 크기는 0.7GB, 총 60백만 개의 문장, BNC, CHILDES, Gutenberg, Opensubtitles, 그리고 Simple English Wikipedia로 구성되며 총 1억 개의 단어로 이루어져 있다. 둘째, 모두의 말뭉치 데이터셋의 크기는 0.8GB, 총 5백만 개의 문장, 그리고 1억 개의 단어를 포함하고 있다. 이러한 데이터셋은 언어 모델링을 위한 훈련 데이터로 사용되었으며, 영어와 한국어 각각의 대규모 텍스트 데이터를 활용하였다.

4. Preprocessing

Table 2는 영어 BabyLM과 모두의 말뭉치를 사용하여 훈련 데이터를 구성하는 데이터셋 크기를 설명한다. 데이터 정제 과정에서는 먼저 불필요한 공백 및 특수 문자를 제거하고, 정제된 데이터를 무작위로 섞어 83%, 8.5%, 8.5% 비율로 훈련, 검증, 테스트 파일로 나누었다.

Fig. 3의 순차-교차 학습 과정에서 필요한 한국어와 영어 쌍에 대해 교차된 데이터 세트를 생성하였다. 이전 단계에서 샘플링된 동일한 텍스트 블록을 순서를 유지하며 번갈아 배치하여, 순차 훈련과 동일한 순서로 모델에 데이터를 제공하였다.

Table 2. Statistics of the Training Datasets

Train Dataset	Size (GB)	Words (M)	Token (M)
English	0.74	100	163
Korean	1.5	100	224

5. Custom Bilingual BPE Tokenization

Byte-Level BPE 토크나이저를 사용하여 맞춤형 이중 언어 토크나이저를 훈련하였다. HuggingFace Tokenizer 패키지를 이용하여 어휘 크기를 32,000으로 설정하여 최소 빈도 2로 BPE를 훈련시켰다. 이 토크나이저는 한국어와 영어 데이터 모두에 동일하게 적용되며, 훈련 데이터는 512 토큰 크기의 블록으로 나누어진다[16].

6. EWC Algorithm

Constantinescu가 제안한 EWC 알고리즘을 적용하여 이중 언어 모델을 구현하였다[18]. Fig. 4은 이중 언어 순차 학습에 EWC 알고리즘의 적용 과정을 설명한다. 먼저, 첫 번째 언어(Task L1)로 모델을 훈련한 후, Fisher Information Matrix를 계산하여 첫 번째 작업에서 중요한 파라미터를 저장한다. 두 번째 언어(Task L2)로 모델을 훈련할 때, 정규화 항을 추가하여 첫 번째 작업의 성능을 유지하면서 모델을 업데이트한다. 정규화 항은 Fisher Information Matrix를 이용해 중요한 파라미터가 크게 변경되지 않도록 한다. 이렇게 하면 모델이 두 언어 모두를 잘 처리할 수 있게 된다. 두 번째 언어 학습이 완료되면 EWC 알고리즘을 종료한다. 이 과정은 모델이 새로운 언어를 학습하면서도 이전 언어의 성능을 유지할 수 있게 되어, 다양한 작업을 순차적으로 학습하는 상황에서 유용한 프로그래밍 가소성을 구현할 수 있다.

Fisher Information Matrix를 계산하기 위해 훈련 데이터에서 무작위로 선택한 10개의 샘플로 미니배치를 구성하고, 손실을 계산한 후 각 파라미터에 대한 그라디언트를 구한다. 이 과정을 여러 번 반복하여 다양한 샘플에서 얻은 값을 평균내거나 누적하여 최종적으로 사용할 Fisher Information Matrix를 결정한다.

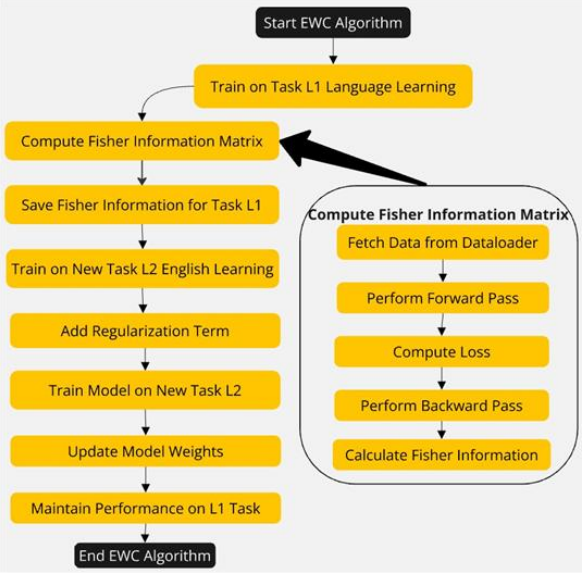

Fig. 4. The Working Principle of EWC with the Fisher Information Matrix

7. Evaluation Metrics

이중 언어 모델의 언어 능력을 평가하는 데 사용된 지표는 Perplexity(PPL)과 BLiMP-accuracy을 이용하여 수용성 평가를 수행하였다.

PPL은 언어 모델의 성능을 평가하는 기본적인 지표 중 하나이다. PPL은 모델이 주어진 텍스트를 얼마나 잘 예측하는지를 수치로 나타낸다. 낮은 PPL 값은 모델이 텍스트를 잘 예측함을 의미한다.

BLiMP 평가는 언어 모델의 텍스트에 대한 영어 문법 규칙 준수, 즉 수용성을 평가하기 위해 BLiMP 데이터 세트를 사용한다. 이 데이터 세트는 구문, 형태론, 의미론의 특정 현상을 평가하는 67개의 하위 데이터 세트로 구성된 67,000 문장의 테스트 세트이다. 이 데이터 세트는 문법적으로 맞거나 틀린 최소 쌍의 문장을 포함하며, 모델이 문법적으로 맞는 문장에 더 높은 점수를 부여하는지를 평가한다. 본 연구에서는 BabyLM 평가 파이프라인을 사용해 모델의 학습 에포크 전반에 걸쳐 정확도에 대해 BLiMP task 평가를 수행하였다[21-22].

VI. Results

본 장에서는 5장의 실험 설계를 바탕으로 PPL과 BLiMP 지표를 사용하여 이중 언어 모델의 성능을 평가하고, EWC 정규화 기법을 적용하여 프로그램 가소성을 조절한 후, 모델 학습 과정에서의 파라미터 변화를 관찰하였다. 이를 통해 L1과 L2에서 모델이 어떻게 작동하는지를 분석했다.

PyTorch와 Transformers 라이브러리를 통해 다양한 조건에서 Hugging Face Transformers 라이브러리의 GPT-2 모델을 처음부터 훈련했으며, 세 가지 순차적 학습 조건(순차, 순차-교차, 순차-EWC) 모두에서 L2 훈련은 동일한 L1 모델 체크포인트에서 새로운 옵티마이저로 시작하며, 동일한 설정을 유지했다. 모델이 학습하지 못하면 학습률을 50% 증가시켰으며, 훈련 중 발산할 경우 학습률을 50% 감소시켰다[16-18].

Fig. 3의 GPT-2-Small 버전의 순차 학습 신경망 언어 모델 학습과 일반화 기법에 사용된 하이퍼파라미터를 Table 3에 정리하였다.

Table 3. Overview of GPT-2 Hyperparameters

Hyperparameter	Value
n_layer, n_positions, n_embed	12, 1,024, 768
activation_function	gelu_new
optimizer	adamw_hf
lr_scheduler	linear
device_train_batch_size	32
adam_beta1, adam_beta2	0.9, 0.999
adam_epsilon	1e-8
max_grad_norm	1
layer_norm_episilon	1e-5
weight_decay, dropout	0, 0.1

이 신경망 이중 언어 모델에서는 12개의 레이어와 1,024개의 포지션, 768차원의 임베딩 벡터를 사용하여 모델의 깊이와 입력 데이터 처리를 한다. 활성화 함수로는 gelu_new를 사용하고, 옵티마이저로는 adamw_hf를 채택하여 학습 효율을 높였다. 학습률 스케줄러는 4로 설정되어 있으며, 한 번에 처리하는 배치 크기는 32로 하였다. Adam 옵티마이저의 베타 값은 0.9와 0.999로 설정되었고, 작은 상수인 1e-8이 추가되어 수치적 안정성을 보장하였다. 그래디언트 클리핑의 최대 노름은 1로 설정되어 있으며, 레이어 정규화를 위한 작은 상수는 1e-5로 하였다. 가중치 감쇠는 0으로 설정되었고, 드롭아웃 확률은 0.1로 설정되어 과적합을 방지하였다.

3장에서 설계한 실험 결과는 이중 언어 모델의 L2 성능 평가를 포함하여 Fig. 5부터 Fig. 8까지의 그래프에 정리되었으며, Mon(단일언어), seq(순차), seq-int(순차-교차), seq-ewc(순차-EWC)로 줄여서 표시하였다.

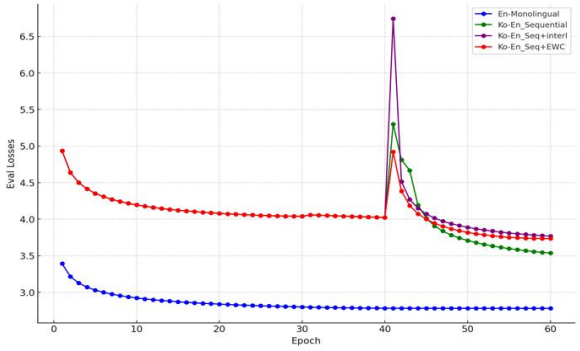

Fig. 5. CE on the L2 validation set

Fig 5의 그래프에서 L2의 성능이 Mon 조건에서 초기와 최종 손실(Cross-Entropy, CE)이 낮은 값을 보인다. 손실이 가장 급격한 감소 패턴을 갖고 있는 seq 조건이 초기 손실에서 가장 급격히 감소하여 최종 손실이 가장 낮다(약 3.5). 반면, 가장 큰 손실을 가진 seq-ewc 조건이 초기엔 가장 낮은 손실을 보여주지만 최종 손실에선 가장 큰 값을 보인다 (초기: 약 4.9, 최종: 약 3.7).

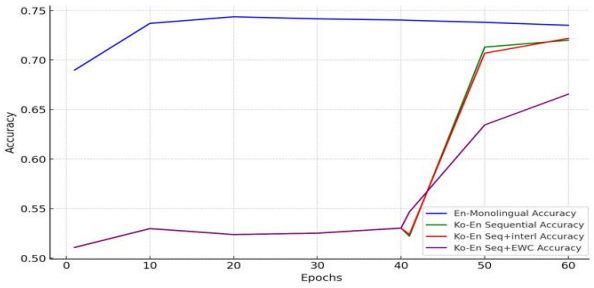

Fig. 6. BLiMP performance over epochs for L2

Fig. 6의 그래프는 BLiMP task의 정답률을 보여준다. Mon 조건이 초기와 최종 BLiMP 정답률 모두에서 가장 높은 값을 보인다. seq+ewc 조건은 초기엔 다른 조건들 보단 높은 값으로 시작되지만, 최종 정확도 모두에서 가장 낮은 값을 보인다(초기: 약 54, 최종: 약 66).

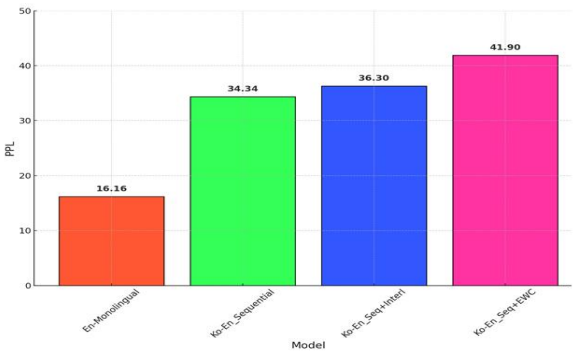

Fig. 7. PPL on the L2 validation set after L2 training

Fig. 7은 가장 낮은 PPL이 mon 조건 (16.16)으로, 이는 모델의 학습 성능이 가장 우수함을 나타낸다. 가장 높은 PPL은 seq-ewc 조건 (41.9)으로, 이는 모델의 학습 성능이 가장 떨어짐을 나타낸다.

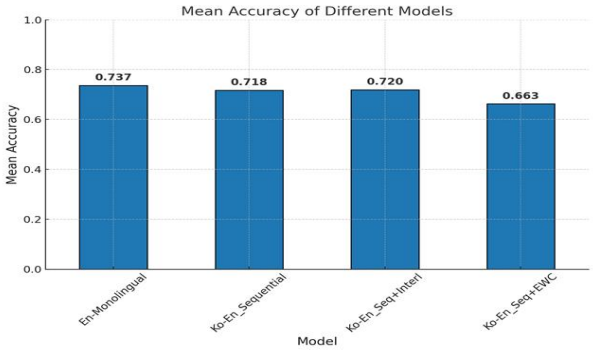

Fig. 8. Final accuracy on BLiMP (average across all tasks) after L2 training

Fig. 8은 가장 높은 최종 L1-L2 언어 모델의 BLiMP 정확도는 Mon 조건 (73.7)으로, 이는 모델의 성능이 가장 우수함을 나타낸다. 가장 낮은 최종 BLiMP 정확도는 seq-ewc 조건 (66.3)으로, 이는 모델의 성능이 가장 떨어짐을 나타낸다.

실험 결과, 제2언어 학습 성능은 단일 언어 학습 조건이 가장 우수한 성능을 보였으며, 이는 모델이 단일 언어에 최적화되었기 때문이다. 이 조건에서 모델은 가장 낮은 PPL과 가장 높은 BLiMP 정확도를 기록했다. 순차적-교차 학습 조건은 L1 노출이 L2 학습을 방해하여 성능이 저하되었다. 반면, 순차적-EWC 학습 조건은 가장 높은 PPL와 가장 낮은 BLiMP 정확도를 기록하며, 성능이 가장 저조했다. 이는 정규화로 인한 가소성 감소가 L2 학습에 부정적 영향을 미친다는 것을 시사한다. 전반적으로, 단일 언어 학습이 가장 효과적이며, EWC가 적용되지 않은 이중 언어 학습도 성능에 부정적인 영향을 미칠 수 있음을 확인했다.

Fig. 9에서는 순차-EWC 조건은 단일 언어 조건과 통계적으로 유의미한 차이($P < 0.05$)를 보여 그 부정적 영향을 강화했지만, 교차 조건 간의 유사성은 발견되지 않았다.

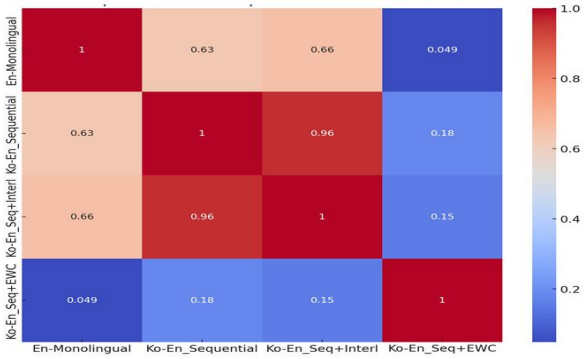

Fig. 9. Permutation test p-values for L2 performance

VII. Conclusions

본 연구에서는 이중 언어 모델의 제2언어 성능을 평가하기 위해 PPL과 BLiMP-accruacy 지표를 사용하고, 다양한 학습 조건에 따른 모델 성능 변화를 분석하였다.

연구 결과, 단일 언어 학습 조건에서 가장 우수한 성능을 보였으며, 이는 단일 언어에 최적화된 모델이 안정적이고 일관된 성능을 보장하기 때문이다. 반면, 순차 학습 및 순차-EWC 조건에서는 성능 저하가 두드러졌으며, 특히 순차-EWC 조건에서 성능이 가장 낮았다. 이는 EWC 정규화 기법이 모델의 가중치 업데이트를 제한하여 새로운 언어 학습 능력을 저하시켰기 때문임을 시사한다.

결정적 시기(Critical Period, CP) 이론을 이중 언어 모델의 학습 과정에 적용해보면, 단일 언어 학습이 최적의 성능을 보이는 이유를 이해할 수 있다. 모델이 단일 언어에 최적화될 때, 이중 언어 학습에서는 언어 간 상호작용이 부정적인 영향을 미치며, 특히 EWC 조건에서는 다른 조건에서보다 성능 저하가 더 발생한다. 이러한 결과는 CP 이론과 일치하며, 프로그래밍 가소성의 관점에서, 다중 언어 모델이 새로운 언어를 학습할 때 가중치 업데이트의 유연성을 유지하는 것이 중요함을 보여준다. 또한, GPT-2 신경망 언어 모델을 사용하여 제2언어(영어)에서의 성능이 제1언어(한국어)와의 구문적 유사성이 매우 낮은 상관관계가 있는 언어를 순차적으로 학습할 때도 CP 효과가 나타났음을 확인했다. 이는 모델이 초기 언어 학습 시기에 형성된 신경망 구조에 의해 제약을 받을 수 있음을 시사한다.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Education of the Republic of Korean and the National Research Foundation of Korea (NRF-2020S1A5A2A01044957)

REFERENCES

[1] A. Radford et al., “Language models are unsupervised multitask learners,” OpenAI blog, 1(8), 9, June 2019.
[2] J. Johnson et al., “Critical Period Effects in Second Language Learning: The Influence of Maturational State on the Acquisition of English as a Second Language,” Cognitive Psychology, 21(1), 60-99, 1989.
[3] C. Pallier et al., “Brain Imaging of Language Plasticity in Adopted Adults: Can a Second Language Replace the First?,” Cerebral Cortex, 13(2), 155-161, 2023.
[4] A. Warstadt and S. R. Bowman, “What artificial neural networks can tell us about human language acquisition,” Algebraic Structures in Natural Language, pp. 17−60. CRC Press, 2022.
[5] A. Zeng et al., “GLM-130B: An Open Bilingual Pre-trained Model,” Oct. 2023, DOI:arXiv.2210.02414.
[6] Constantinescu, Ionut-Laurentiu, “Exploring the Language Acquisition Critical Period Effect in Language Models,” pp. 80, May 2024, DOI: doi.org/10.3929/ethz-b-000674036.
[7] S. Hochreiter and S. Jurgen, “Long short-term memory,” Neural Computation, Vol. 9(8), pp. 1735-1780, Nov. 1997.
[8] A. Vaswani et al., “Attention Is All You Need,” Advances in neural information processing systems, 30, Dec. 2017, DOI: arXiv.1706.03762.
[9] A. Radford et al., “Language Models are Unsupervised Multitask Learners,” OpenAI. June 2019.
[10] J. Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv preprint arXiv:1810.04805, June 2019.
[11] T. Pires et al., “How multilingual is Multilingual BERT?,” arXiv preprint arXiv:1906.01502, 2019.
[12] J. Kirkpatrick et al., “Overcoming catastrophic forgetting in neural networks,” Proceedings of the National Academy of Sciences, 114(13), pp. 3521-3526, 2017.
[13] A. E. Hernandez et al., “The emergence of competing modules in bilingualism. Trends in Cognitive Sciences, 9(5),” pp. 220-225, Oct. 2005.
[14] T. Wolf et al., “Transformers: State-of-the-Art Natural Language Processing,” 2020. arXiv preprint arXiv:1910.03771.
[15] A. Warstadt et al., “BabyLM: A Benchmark for Language Models on Child-Directed Speech,” 2023 arXiv preprint arXiv:2301.11796.
[16] Gpt2, https://huggingface.co/docs/transformers/model_doc/gpt2
[17] Babylm, https://github.com/babylm/evaluation-pipeline-2023
[18] https://github.com/iconstantinescu/lm-critical-period
[19] BabyLM Challenge, https://babylm.github.io
[20] NIKL, https://kli.korean.go.kr/corpus.main/requestMain.do
[21] BLiMP, https://github.com/alexwarstadt/blimp

Authors

Euhee Kim received the M.S. degrees in Computer Engineering from Dongguk University, Korea, in 2002 and Ph.D. degrees in Mathematics from The University of Connecticut, U.S.A in 1995.

Euhee Kim is currently a Professor in the Department of Software Convergence at Shinhan University. She is interested in AI, NLP and Big Data computing.

Keonwoo Koo received the B.S., M.S. degrees in English language and education, currently Ph.D. Candidate in Computational linguistics from Dongguk University, Korea, in 2019, 2021 and 2021-, respectively.

Keonwoo Koo is currently a Ph.D candidate in the Department of English language and literature, Dongguk University. He is interested in computational linguistics, computational psycholinguistics, and psycholinguistics.