COSE474-2023F: Final Project
Leveraging Text-Augmentation Methods in Korean Text Classification

2019320137
Sangmin Hwang

1. Introduction

1.1. Motivation

데이터 증강(Data Augmentation)은 데이터가 부족한 환경에서 모델의 성능을 끌어올리는 방법 중 하나이다. 자연어 처리 분야에서는 텍스트의 일부분을 바꾸거나, 삭제하는 등의 방식이 이용되고 있다. EDA(Easy Data Augmentation), Backtranslation 등 다양한 방식의 텍스트 증강 기법이 제안되었으며, 각 기법들이 텍스트 분류를 포함한 여러 자연어처리 문제에서 모델의 성능을 향상시킬 수 있다는 것이 입증되었다.
그러나 이러한 기법들은 대부분이 영어 데이터셋을 대상으로 하였는데, 한국어 데이터셋을 대상으로도 이 기법들이 효과적인가에 대해서는 확인해 볼 필요성이 있다. 고립어인 영어와 교착어인 한국어의 언어적인 차이에서 비롯된 효과의 차이가 있을 수 있기 때문이다. 따라서 본 프로젝트에서는 기존에 제안된 텍스트 증강 기법을 한국어 데이터셋에 적용해보고, 유의미한 성능 향상 정도를 비교하며 한국어에 적절한 텍스트 증강 기법이 무엇인가에 대하여 탐구해보려고 한다.

1.2. Problem Definition

본 프로젝트에서는 데이터가 부족한 상황에서 한국어 텍스트 분류 문제의 성능을 가장 크게 향상시킬 수 있는 데이터 증강 기법에 대해 탐구해보는 것을 목표로 한다.
구체적으로는, 한국어 데이터로 Pretrain된 BERT 모델을 텍스트 분류 Task에 Fine-tuning하는 과정에서, 각 텍스트 증강 기법을 활용하였을 때의 성능을 비교해본다.
본 연구의 의의는 데이터가 부족한 상황을 최대한 현실적으로 가정하여 적절한 증강 기법을 제안하였다는 데 있다. 기존 연구들의 경우 Data augmentation을 위한 방법론을 구축하기 위해, 고비용의 API를 사용하거나 더 많은 데이터로 학습된 Language model을 활용하는 등의 문제점이 있었다(Dai et al., 2023). 그러나 Low resource 상황에서 그러한 비용을 감당하며 데이터를 증강할 수 있는 경우는 극히 제한적일 것이다. 그러나 본 연구에서는 Rule-base 기반 방법론, 무료로 제공되는 API, 기존 데이터만을 사용해 학습시킨 Language model 등을 데이터 증강에 활용하였으며, 현실과 유사한 문제 상황에 직접적으로 적용할 수 있는 방법론이라 생각된다.

2. Methods

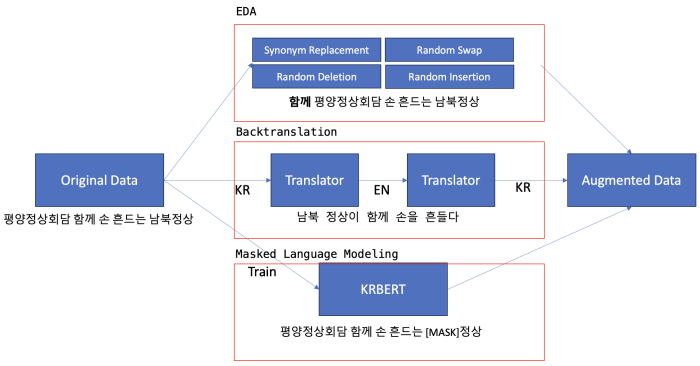

Figure 1. Overall Methodologies

다음 세 가지 기법을 데이터가 부족한 환경에서의 문장 분류 데이터셋을 증강하는 데 적용해본다. 각 방법론을 적용하여 원본 데이터 하나당 2개의 데이터를 추가로 증강하였으며, 최종적으로 증강된 데이터는 기존 데이터를 포함하여 총 3배 크기의 데이터가 된다.

2.1. EDA(Easy Data Agumentation)

EDA(Wei & Zou, 2019)는 간단한 텍스트 편집 기법을 이용하여 데이터를 증강하는 방식이다. 문장 내에서 임의의 단어를 유의어로 교체하고 (Synonym Replacement), 문장 내의 임의의 위치에 기존 단어들의 동의어를 삽입하고 (Random Insertion), 문장 내의의 임의의 두 단어의 순서를 바꾸고(Random Swap), 문장 내의 임의의 단어를 삭제하는(Random Deletion) 등의 4가지 방법을 사용하여 데이터를 증강한다.
원문에서는 Synonym Replacement와 Random Insertion 과정에서 영어 기반의 WordNet을 사용하였는데, 이 기법을 한국어에 적용하기 위해 한국어로 구축된 WordNet으로 대체하여 적용하였다. 원 문장의 10%에 해당하는 단어에 4가지 방법론 중 무작위로 하나를 선택하여 증강 데이터를 생성하였다.

2.2. Backtranslation

Backtranslation(Edunov et al., 2018)은 원문을 다른 언어로 번역한 후, 번역된 문장을 다시 원문의 언어로 번역함으로써 데이터를 증강하는 방식이다. 번역 과정에서 문장의 순서나 어휘 등의 조금씩 바뀜으로써 원문의 표현을 다르게 만들 수 있다는 장점이 있다.
원문에서는 딥러닝 기반의 언어 모델을 사용하여 Backtranslation을 진행하였으나, 한국어를 출발/도착 언어로 사용할 수 있는 언어 모델이 부족하여 웹 상에서 공개되어 있는 한국어 번역 API[1]를 사용하여 번역을 진행하였다. 번역에 사용한 중간 언어는 영어, 일본어, 중국어 간체, 스페인어 총 4가지이며, 이 중 중복된 결과를 제외하고 증강 데이터로 활용하였다.

2.3. Masked Language Modeling

Masked Language Modeling 기반 Data Augmentation은 (Kumar et al., 2021)에서 제안된 방법론으로, 원 문장의 일부 토큰을 마스킹한 뒤 마스킹한 토큰을 예측함으로써 데이터를 증강하는 방법이다. 이러한 방법론 구현을 위해 MLM을 위한 모델을 하나 더 두었으며, 이는 원 Task의 Classifier와 동일한 모델을 사용하였다. 또한 학습 데이터로 원 task(Topic classification)에 사용한 것과 동일한 데이터를 사용하여 학습하였다.

3. Experiments

3.1. Computing Resource

실험 환경: Google Colaboratory Free plan
GPU: Tesla T4 GPU (15GB VRAM)
OS: Ubuntu 22.04.3 LTS
개발 환경: Python 3.10.12, Huggingface Transformers 4.35.2

3.2. Dataset

KLUE-ynat 데이터셋은 한국어 언어 모델의 주제 분류 (Topic Classification) 성능을 측정하기 위해 구축된 데이터셋[2]이다. 2016년부터 202년까지의 연합뉴스 기사의 헤드라인을 모아 7가지의 카테고리(정치, 경제, 사회, 생활문화, 세계, IT과학, 스포츠)로 분류되어 있다.

3.3. Model

Baseline 및 Masked Language Modeling 방법론에 사용한 모델은 KRBERT-Medium[3][4]이다. 약 100M 개의 파라미터를 가지고 있으며, 한국어 위키피디아, 신문 기사 등으로 Pretrain되어 있다.

Split	Number of data
Train	280
Validation	70
Test	7000

Table 1. Low resource dataset setting. Each Category labels are equally distributed.

Methodology	**Accuracy**
Baseline	0.707
EDA	0.799
Backtranslation	0.792
MLM	0.739

Table 2. Topic classification accuracy

3.4. Experimental Setup

3.4.1. Low Resource Setting

본 프로젝트에서는 데이터가 부족한 상황을 가정하여, Table 1과 같이 각 카테고리별로 50개의 데이터만을 샘플링하여 학습 데이터로 구축하였다. 이중 40개의 데이터는 실제 학습에 사용하고, 10개의 데이터는 검증을 위한 Validation Set으로 활용하였다. 최종 성능을 측정하는 테스트 데이터는 카테고리별로 1000개의 데이터로 구성하였다. 또한, 한국어 데이터에서의 온전한 성능 측정을 위해, 한문 등의 한국어 문서와 동떨어진 텍스트는 제외하였다. Masked Language Modeling 방법론에 사용한 데이터셋도 이와 동일하다.

3.4.2. Evaluation

Train data를 이용하여 모델을 학습시킨 뒤 Validation Accuracy가 가장 높은 모델을 사용하여 평가를 진행하였다. 각 증강 방법론별로 학습 Epoch 수는 모두 동일하게 설정하였으며, 증강을 적용하지 않은 Baseline의 경우, 동일한 조건에서 실행하기 위해 증강된 데이터 수만큼 Epoch 수를 늘려 증강된 경우와 Iteration 수를 동일하게 맞춰주었다. Batch size, Optimizer 등의 하이퍼파라미터는 모든 방법론이 동일하다.

4. Results

4.1. Quantitative Results

실험 결과는 Table 2와 같다. 세 방법론 모두 증강을 적용하지 않은 경우보다 높은 Accuracy를 기록하는 모습을 보이며 데이터셋이 부족한 상황에서 데이터 증강이 모델 성능을 향상시키는 데 도움이 된다는 것을 알 수 있었다.

[1] https://papago.naver.com
[2] 기존 Proposal에서 선정하였던 Toxic-comment 데이터셋은 실험 결과 Task의 난이도로 인해 Low-resource setting에서의 성능 향상 폭이 미미하여 변경하였다.
[3] Proposal에서 선정한 KOBERT 모델은 Tokenizer 문제가 발생하여 인해 변경하였다.
[4] https://huggingface.co/snunlp/KR-Medium

EDA를 적용한 경우가 기존보다 13% 가량 성능이 향상되며 가장 좋은 결과를 보였으며, Backtranslation을 적용한 경우도 이와 비슷한 성능을 보였다. Masked language modeling의 경우에는 4.5% 향상되는 데 그치며 두 방법론보다는 성능 향상 폭이 덜했다.

4.2. Qualitative Results

Figure 2는 실제 증강된 데이터셋의 예시이다. EDA의 경우에는 전반적으로 문장의 주제를 유지하면서 문장을 증강하는 모습을 보였다. 간혹 Random Insertion과 Synonym Replacement 과정에서 문맥과 관련없는 단어가 들어가거나, Random Swap으로 인해 온전한 문장의 형태라 보기 힘든 문장이 있었다.
Backtranslation의 경우에는 문장의 의미나 문장의 형태가 잘 보존되었다. 그러나 사용된 단어의 다양성이 조금 떨어지는 모습을 보였고, 번역 과정에서 원본에는 없었던 종결 어미가 붙는 등의 현상이 발생하였다.
MLM의 경우, Mask Infilling 과정에서 다양한 단어를 생성하여 표현이 다양하다는 장점이 있었으나, 문장의 의미와 상관없는 이모지, 한자 등이 문장에 노이즈가 많이 삽입되는 모습이 있었다.

4.3. Figures

Methodology	Sentence (Label: 정치)
Original	선택 4.13 고려인 동포들도 국민으로서 한 표
EDA	선택 4.13 고려인 동포들도 국민 한 표 선택 들 고려인 동포4.13도 국민으로서 한 표
Backtranslation	4.13 고려인 동포들도 국민으로 한 표를 행사했다 선택4.13 고려인 동포들도 국민으로서 한 표를 얻었어요
MLM	선택 4. 13 고려인 동포들도 국민으로서🐸해야 제주 4. 13 고려인 동포 대한민국 국민으로서 한마음

Figure 2. Augmentation results

4.4. Dicussions

증강된 데이터로 미루어보았을 때, Masked Language Modeling 방법론의 경우 Language Model 학습 과정에서 사용된 데이터가 부족하여 증강된 문장에 노이즈가 많이 삽입된 것으로 보인다. 성능 향상 폭이 가장 적었던 것도 이러한 이유인 것으로 생각된다. Low resource setting에서는 데이터 증강 모델을 충분히 학습시키기에는 어려움이 있을 것으로 생각된다. 그러나, 결과로 나온 단어의 다양성이 가장 높은 모습을 보여주었기에, 공개된 PLM의 Pre-trained knowledge를 활용한다면 다른 방법론과 다른 성능 향상을 기대할 수 있을 것이다.
EDA의 경우 증강된 데이터로 미루어 보았을 때 실제 문장이라고 보기 어려운, 단어의 나열에 가까운 문장들이 다수 있었으나 오히려 성능 향상 폭이 가장 컸다. 원문에서 언급하듯 문장 형태의 붕괴가 일종의 노이즈로 작용한 것으로 판단되며, Task가 Topic classification이었기에 문장의 온전한 형태를 유지하는 것이 최종 성능에 크게 영향을 미치지 않았을 것으로 생각된다. 따라서 Classification task에는 적절한 방법론이라고 생각되나, Semantic Textual Similarity (STS), Natural language Generation (NLI) 등의 다른 Task에 적용하기에는 무리가 있을 것으로 판단된다.
Backtranslation은 문장의 의미를 잘 보존하는 경향이 있으나 번역 과정에서 문장의 어체가 바뀌는 문제가 있었다. Topic classification, NER 등의 Classification 기반 Task나 STS 등의 Task에는 적용하기 좋은 방법론이라고 생각되나, 어체 변화로 인해 Natural Langauge Generation과 같은 분야에 사용하게 되면 문제가 발생할 여지가 있을 것이다.

5. Future Works

본 프로젝트에서는 Topic classification Task를 대상으로만 데이터 증강 기법의 효용성을 검증하였다. 그러나, 더욱 정밀한 성능 검증이 되려면 다양한 태스크에서의 테스트가 필요할 것이다. 태스크별로 적절한 증강 기법이 각각 다를 것으로 생각되며, 이러한 과정이 후행된다면 태스크별로 Low resource 상황에서의 문제점을 해결할 수 있는 방법론이 적절히 제시될 수 있다.
또한, 증강 과정에서 Low resource setting에서의 데이터 수(Label별 50개)와, 증강한 데이터 규모(Data당 2개)를 임의로 설정하여 실험을 진행하였는데, 데이터 크기에 따른 증강 기법별 성능도 차이가 있을 것으로 생각된다. 다양한 데이터 크기에 대해 증강 성능을 테스트해보아야, 데이터가 부족한 정도에 따라 적절한 증강 기법을 적용할 수 있을 것이다.

References

Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., Li, S., Zhu, D., Cai, H., Sun, L., Li, Q., Shen, D., Liu, T., and Li, X. Auggpt: Leveraging chatgpt for text data augmentation, 2023.

Edunov, S., Ott, M., Auli, M., and Grangier, D. Understanding back-translation at scale, 2018.

Kumar, V., Choudhary, A., and Cho, E. Data augmentation using pre-trained transformer models, 2021.

Wei, J. and Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019.