딥러닝 기반 다국어 데이터 정제 기법

이아영, 이동훈, 한민아, 김남규*
국민대학교

dkdud5068@kookmin.ac.kr, ldh1468@kookmin.ac.kr,

minahan0403@kookmin.ac.kr, ngkim@kookmin.ac.kr

Deep Learning-based Multilingual Data Purification Scheme

Lee Ahyoung, Lee Donghoon, Han Min Ah, Kim Namgyu
Kookmin Univ.

요 약

최근 딥 러닝 기술의 발전에 따라 방대한 양의 텍스트 데이터를 학습하고 그 결과를 활용하는 연구가 활발히 이루어지고 있다. 대부분의 텍스트 분석 응용에서 고품질의 학습 데이터 확보는 분석의 성패를 좌우하는 가장 중요한 요소로 인식되고 있으며, 특히 기계 번역 분야의 경우 두 가지 언어의 쌍으로 구성된 고품질 학습 데이터 확보의 중요성이 매우 강조되고 있다. 이러한 배경에서 저품질 데이터를 정제하여 고품질 데이터를 확보하려는 연구가 일부 시도되고 있지만, 각 언어로 작성된 두 내용의 의미가 서로 일치하지 않는 데이터를 사람의 주관적 개입 없이 식별하는 연구는 충분히 이루어지지 않았다. 이에 본 연구에서는 자동으로 저품질 데이터를 식별하기 위해, 사전학습 언어모델을 통한 문서 임베딩 및 벡터 얼라인먼트를 사용한 다국어 데이터를 정제 방법론을 제안한다.

Ⅰ. 서 론

최근 컴퓨팅 기술이 빠르게 발전함에 따라 대량의 데이터를 수집하는 것이 용이하게 되었으며, 이로 인해 이러한 데이터를 잘 학습할 수 있는 딥 러닝(Deep Learning)에 대한 관심이 높아지고 있다. 딥 러닝 모델이 좋은 성능을 내기 위해서는 모델 자체의 성능뿐만 아니라 학습에 사용되는 데이터의 품질 또한 중요하기 때문에, 좋은 품질의 데이터를 확보하는 것은 딥러닝 분석의 가장 중요한 요소로 인식되고 있다[5]. 특히 기계 번역 분야의 학습 데이터는 소스(Source) 언어의 데이터와 번역하고자 하는 타겟(Target) 언어의 데이터, 이렇게 두 가지 언어의 쌍으로 구성되며, 우수한 번역 모델의 개발을 위해 두 언어로 이루어진 고품질 학습 데이터를 확보하는 것은 가장 어려우면서도 중요한 과정으로 알려져 있다[2].

본 연구에서는 기계 번역에 필요한 고품질 데이터를 확보하기 위해, 저품질 데이터, 즉 소스 언어와 타겟 언어의 말뭉치 간 내용에 차이가 있는 데이터를 식별하여 제거할 수 있는 방안을 제안하고자 한다. 예를 들어 <표 1>은 AI-Hub 에서 제공되는 “한국어-영어 번역 말뭉치(기술과학)” 데이터의 일부로, 한국어 문장과 영어 문장이 쌍을 이루고 있다. 예시로 보인 세 문장 중 1 번과 2 번 문장은 비교적 한국어 문장과 영어 문장이 나타내는 의미가 유사한 반면, 3 번의 경우 영어로 작성된 문장이 한국어로 작성된 문장의 의미를 충분히 나타내고 있지 않음을 알 수 있다. 이에 본 연구에서는 사전학습 언어모델(Pre-trained Language Model)[6]을 통한 문서 임베딩 및 벡터 얼라인먼트(Alignment)를 활용하여[1] [3] [4], <표 1>의 3 번과 같은 저품질 다국어 데이터를 식별하는 방안을 제안하고자 한다.

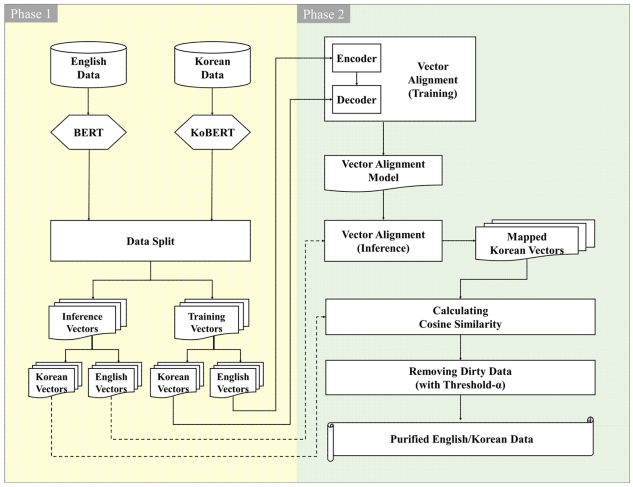
<표 1> 기계 번역 학습을 위한 한·영 데이터 쌍의 예

번호	한국어	영어
1	로봇에 대한 수요 대부분은 본인을 위한 수요로 파악되나 특수 목적의 로봇은 조금 달랐다.	Most of demand for robots is understood as demand for themselves, but special-purpose robots are a little different.
2	산소를 제공함으로써 미토콘드리아의 기능 저하를 막는데 도움을 줄 수 있었을 것이다.	By providing oxygen, it may have helped prevent the deterioration of mitochondrial function.
3	폐가스 가습조 내부의 용수에 용존하는 암모니아 질소 및 NO - 이온의 시간에 따른 농도 추이는 Figs. 13 및 14 와 같다.	Figs.Figs.Figs. Same as 13 and 14.

제안 방법론은 (1) 한국어와 영어 문서를 각각의 사전학습 언어모델을 통해 임베딩하고, (2) 영어 문서의 벡터를 한국어 문서 벡터 공간으로 매핑하는 벡터 얼라인먼트를 수행한 뒤, (3) 영어에서 한국어로 매핑된 벡터와 원래의 한국어 벡터와의 유사도를 비교하여 유사도가 낮은 데이터를 저품질 데이터로 판정한다. 본 논문에서는 제안 방법론의 전체 구조를 소개하고, 제안 방법론의 성능을 평가하기 위해 인위적으로 저품질 데이터를 삽입하여 이를 식별해 내는 실험을 수행한 결과를 소개한다.

Ⅱ. 본론

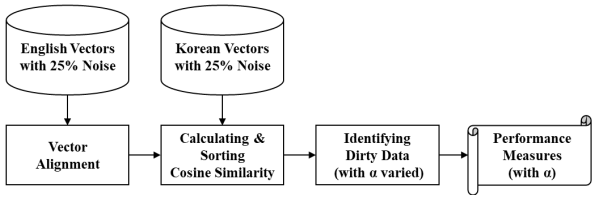
본 연구의 제안 방법론은 다음 <그림 1>과 같다. <그림 1>에서 Phase 1 은 각 언어별로 입력 데이터의 임베딩을 수행한 후, 이를 학습 데이터와 추론 데이터로 분할하는 과정이다. 이때 영어 문서는 BERT 를 통해, 한국어 문서는 KoBERT 를 통해 임베딩을 진행한다.



<그림 1> 제안 방법론 전체 개요

다음으로 Phase 2 에서는 앞에서 분할한 학습 데이터를 사용하여 영어 문서 벡터의 한글 벡터 공간 매핑을 위한 딥러닝 기반 벡터 얼라인먼트 모델을 학습한다. 본 연구에서는 이를 위해 단순한 층만 쌓아서 구성한 모델보다 일반적으로 성능이 더 우수한 것으로 알려진 DNN(Deep Neural Network) 기반의 Encoder-Decoder 구조를 사용하여 벡터 얼라인먼트를 수행하였다. 이렇게 생성한 벡터 얼라인먼트 모델에 추론 데이터의 영어 벡터를 입력으로 제공하면, 얼라인먼트를 통해 한국어 공간으로 매핑된 벡터가 도출된다. 도출된 한국어 벡터와 실제 정답인 한국어 벡터 간의 코사인 유사도를 산출하고, 미리 정한 임계값 이하의 유사도를 갖는 데이터를 저품질 데이터로 간주하여 제거하는 방식으로 데이터 정제가 이루어진다.

제안 방법론의 우수성을 평가하기 위해, 저품질 데이터를 제안 방법론이 얼마나 정확하게 식별하는지를 확인하는 실험을 수행하였다. 이러한 실험을 수행하기 위해서는 한국어/영어 데이터 쌍에 대해 고품질 혹은 저품질로 레이블이 부여된 데이터가 필요하므로, 본 연구에서는 인위적으로 저품질 데이터를 생성하여 실험을 수행하였다. 실험에 사용한 데이터는 AI hub 의 기술과학 분야 영·한 병렬 말뭉치 데이터로, 총 10 만 건의 영어 문장과 한국어 문장의 쌍으로 구성되어 있다. 10 만 건의 데이터 중 6 만 건은 학습용, 4 만 건은 추론용으로 사용하였으며, 추론 데이터 4 만 건 중 3 만 건은 그대로 사용하고, 1 만 건은 한국어와 영어 문서가 서로 엉뚱하게 연결되도록 처리하였다. 즉 75%의 고품질 데이터와 25%의 저품질 데이터를 사용하여 성능 평가를 수행하였으며, 실험의 개요는 <그림 2>와 같다.



<그림 2> 실험 과정 전체 개요

제안 방법론에 따라 먼저 실험에 사용한 데이터를 6:4 의 비율로 학습 데이터와 추론 데이터로 분할한 후, 학습 데이터로 벡터 얼라인먼트 모델을 학습하였다. 이후 추론 데이터 4 만 건에 대해 유사도의 임계값을 변화시켜가며 정확도를 산출하였다. <표 2>의 실험 결과 임계값의 증가에 따라 정확도와 F1-Score 는 점차 증가하다가 감소하는 현상을 보였으며, 가장 높은 정확도와 F1-Score 는 각각 0.805 와 0.794 로 나타남을 확인하였다.

<표 2> 저품질 데이터 식별 성능 (Noise: 25%)

정제할 하위 데이터 개수 (4만 중 하위 α %)	Accuracy	Precision	Recall	F1-score
2000 (5%)	0.779	0.794	0.159	0.719
4000 (10%)	0.798	0.74	0.296	0.764
6000 (15%)	0.805	0.683	0.41	0.787
8000 (20%)	0.802	0.63	0.504	0.794
10000 (25%)	0.794	0.587	0.587	0.794
12000 (30%)	0.776	0.544	0.652	0.783
14000(35%)	0.755	0.507	0.709	0.766
16000 (40%)	0.729	0.473	0.757	0.745

Ⅲ. 결론

본 연구에서는 영·한 병렬 말뭉치 데이터 중 저품질 데이터를 자동으로 식별하기 위한 방안을 제안하였다. 또한 실제 데이터 10 만 건을 분석한 실험을 통해 제안 방법론의 성능을 평가한 결과, 제안 방법론이 인위적으로 삽입된 저품질 데이터를 우수한 성능으로 식별해 냄을 확인하였다. 제안 방법론을 통해 저품질 데이터를 제거하여 영·한 병렬 말뭉치 데이터의 품질을 향상시킴으로써, 병렬 말뭉치의 학습을 요구하는 다양한 딥 러닝 응용의 성능도 향상될 것으로 기대한다.

참 고 문 헌

[1] 김준우, 윤병호, and 김남규. “전문어의 범용 공간 매핑을 위한 비선형 벡터 정렬 방법론,” 지능정보연구, 28(2), Jun, 2022.

[2] 박찬준 and 임희석. “공공 한영 병렬 말뭉치를 이용한 기계번역 성능 향상 연구,” 디지털융복합연구, 18(6). Jun, 2022.

[3] M. Artetxe, G. Labaka, and E. Agirre, “Learning Principled Bilingual Mappings of Word Embeddings while Preserving Monolingual Invariance,” In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016.

[4] M. Biesialska, and Marta R. Costa-jussà, “Refinement of Unsupervised Cross-lingual Word Embeddings,” arXiv:2002.09213, Feb, 2020.

[5] S. U. Park, “Analysis of the Status of Natural Language Processing Technology Based on Deep Learning,” Korean Journal of BigData, Vol. 6, Aug, 2021.

[6] J. Devlin, W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv:1810.04805, Oct, 2018.