Encoder Layer를 이용한 의도 분류 성능 비교

안혁주○, 김혜영
하나금융 융합기술원
{nlpahj, rebeccakim}@hanafn.com

Comparing the Performances of Intent Classifications by Encoder Layer

Hyeok-Ju Ahn○, Hye-Young Kim
Hana Institute of Technology, Hana TI

요 약

본 논문에서는 분류 모델의 주류로 사용되고 있는 Encoder 기반 사전학습 모델(BERT, ALBERT, ELECTRA)의 내부 Encoder Layer가 하부 Layer에서는 Syntactic한 분석을 진행하고 상부 Layer로 갈수록 Semantic한 분석을 진행하는 점, Layer가 구성됨에 따라 Semantic 정보가 Syntactic 정보를 개선해 나간다 점에 기반한 기존 연구 결과를 바탕으로 Encoder Layer를 구성함에 따라 어떻게 성능이 변화하는지 측정한다. 그리고 의도 분류를 위한 학습 데이터 셋도 분류하고자 하는 성격에 따라 Syntactic한 구성과 Semantic한 구성을 보인다는 점에 착안하여 ALBERT 및 ELECTRA를 이용한 의도 분류 모델을 구축하고 각 데이터 셋에 맞는 최적의 Encoder Layer 구성을 가지는 모델을 비교한 결과, 두 데이터 셋 간에 다른 Layer 구성을 보이는 점과 기존 모델보다 성능이 향상됨을 확인하였다.

주제어: ALBERT, ELECTRA, Encoder Layer, Syntactic, Semantic, Intent Classification

1. 서론

의도 분류(Intent Classification)는 문장 또는 지문 전체에 구성된 키워드 구성 및 의미적 연관관계를 분석하여 사용자가 의도하고자 하는 바를 분류하는 기술을 말한다. 이전의 의도 분류 기술은 문장 내 주요한 키워드를 기반으로 한 규칙 기반 방법론, 핵심 자질들을 추출한 뒤 통계적으로 분류기준에 따른 자질들을 훈련하여 의도 분류하는 자질 기반 통계적 분류 방법론이 주류를 이루었지만, 최근 딥러닝(Deep Learning)을 이용한 인공신경망 기반 방법론이 기존 방법론들의 성능을 뛰어넘어 주류 기술로 사용되고 있다.

그 중 최근 각광받고 있는 사전학습 언어모델(Pre-trained Language Model)이 다른 딥러닝 기반 모델들의 성능을 앞지르면서 현재 다른 기술 뿐 아니라 의도 분류 기술의 주 된 모델로 사용되고 있으며, 사전학습 언어모델을 기반으로 한 BERT(Bidirectional Encoder Representations from Transformers)[1], ELECTRA (Efficiently Learning an Encoder that Classifies Token Replace-ments Accurately)[2], GPT-3(Generative Pre-trained Transformer 3)[3] 등의 모델이 주류 모델로 사용되고 있다.

한편, 학습을 위해 사용하는 모델 외에도 분류하고자 하는 목적에 따라 이를 학습하기 위한 데이터에 대한 분석 및 의도 분류 기준(intent)을 설계 및 구축하는 것이 의도분류 기술에 필요한 핵심 요소이다. 의도 분류를 위한 데이터는 크게 두 가지 방식으로 나뉘는데, 첫 번째로 학습하고자 하는 문장의 주요 키워드나 문법적 구성요소(Syntactic)로 인해 의도 분류 기준이 구성되는 경우, 두 번째로 학습하고자 하는 문장 내 키워드가 가지는 의미(Semantic)를 바탕으로 의도 분류 기준이 구성되는 경우이다. 표 1은 하나은행 챗봇에서 질의 의도를 분류하기 위해 구축한 데이터(이하 HIT_chatintent)와 문장이 의도하는 바를 분류하기 위한 데이터인 3i4k[4]의 구성 방식 차이를 보여준다. HIT_chatintent는 문장 내 주요 단어에 대한 의미론적 구성에 따라 분류 기준이 다르고, 3i4k는 문장 자체가 가지는 구문적 구성이나 담화 성분에 따라 분류가 달라지는 것을 확인할 수 있다.

위의 내용에 따라 본 논문에서는 사전학습 언어모델 중 Transformer Encoder를 기반으로 한 모델인 ALBERT(A Lite Bidirectional Encoder Representations from Trans-formers)[5], ELECTRA를 이용하여 관련 연구를 참고한 의도 분류 모델을 구성하고, 성격이 다른 두 데이터에 Encoder Layer를 어떻게 구성하는 것이 가장 높은 성능을 도출할 수 있을지 확인하고자 한다.

표 1. 의도 분류 학습데이터의 차이

구분	문장	분류
Semantic	태국 바트 환전을 신청하고자 해	환전신청
	태국 바트 환전 시 환율이 궁금해	환율조회
	계좌에 있는 돈 얼마인지 확인할래	계좌조회
	계좌에 있는 돈 다른 계좌에 넣을래	이체신청
Syntactic	저 혼자서 그냥 갈게요	서술
	너 혼자서 갈 수 있는지 알려줘	질문
	제발 너 혼자 그냥 가라	명령
	너 혼자서 집에 가면 어떡해	수사의문

2. 관련 연구

최근 의도 분류 모델은 딥러닝 방법론이 주를 이루고

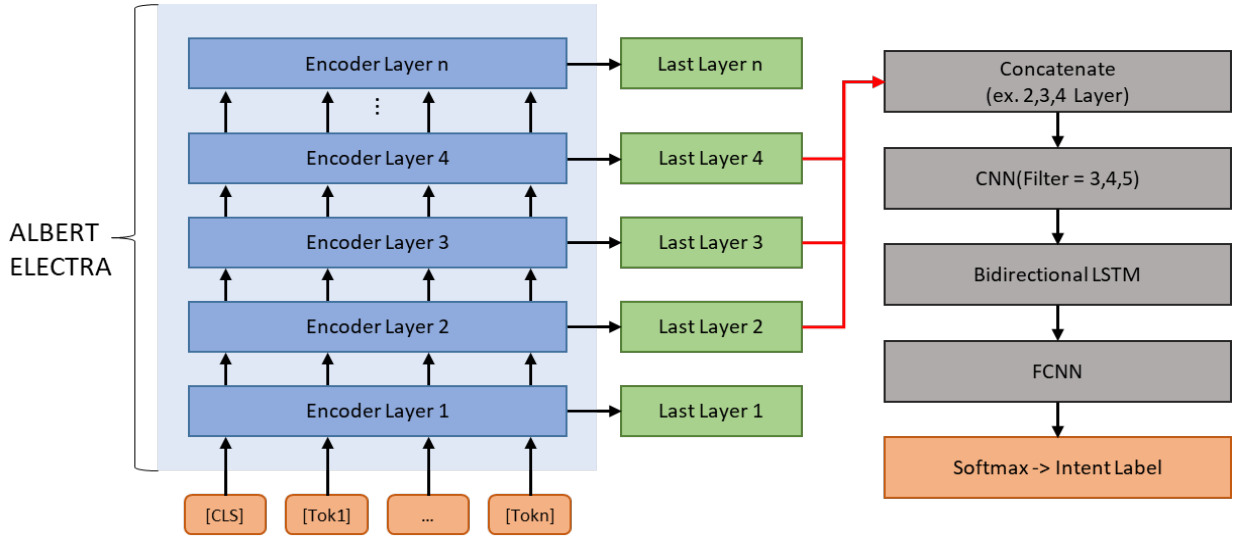

그림 1. Encoder Layer를 이용한 의도 분류 모델 구조도

있으며, 초창기에 많은 연구자들이 CNN(Convolutional Neural Networks), RNN(Recurrent Neural Networks) 등등의 방식을 도입했다. 몇 가지 예로 [6]은 CNN을 활용하여 기존 모델에 비해 문장 분류의 성능을 향상시켰고, [7]은 Bi-LSTM(Bidirectional Long Short-Term Memory)을 활용하여 의도 분류뿐만 아니라 개체명 인식 성능도 향상 시키는 연구 성과를 보였다. 이외에도 Bi-GRU를 이용한 기법, CNN과 LSTM을 결합한 기법 등의 연구들이 시도되어 왔다.

이후 BERT를 위주로 한 사전학습 언어모델이 등장함에 따라 이를 이용한 다양한 의도 분류 방법론이 연구되었는데 그 중 [8]은 BERT를 이용해 의도 분류 및 슬롯 필링(Slot-Filling) 방법을 제안했다. 또한 [9]는 BERT와 Bi-LSTM을 병렬적으로 구성하여 의도 분류 성능을 향상시키는 방법을 제안했다.

한편 BERT, ALBERT, ELECTRA와 같은 Encoder를 이용한 모델에서 각 Layer 별로 어떤 차이를 보이는지에 대한 연구들이 진행되었는데 [10]은 BERT에서 하부 Layer는 Syntactic한 분석을, 상부 Layer로 갈수록 Semantic한 분석을 수행하며 이 정보를 계층적으로 저장한다는 결론을 도출했다. 그리고 Layer가 구성됨에 따라 Semantic 정보가 Syntactic 정보를 개선해 나간다는 점을 발견했다. [11]에서는 [10]의 연구결과를 바탕으로 특정 Layer를 앙상블 했을 때 기본 모델보다 성능이 향상함을 보였다.

3. 제안 모델

제안 모델의 전체 구조는 그림 1과 같다. 의도 분류 학습 데이터를 input으로 두어 모델을 학습시키면 Encode Layer들이 형성되는데 이 중 n개를 사용자가 임의로 지정하고, 지정한 Layer에서 Last Layer를 추출하여 연결(Concatenation)한다. 그리고 이 연결된 Layer를 CNN과 Bi-LSTM을 거쳐 특징 값 추출 및 양방향으로 연결한 결과를 완전 연결 신경망(Fully Connected Neutral Networks)을 적용하여 분류한다.

3.1 사전학습 모델

제안 모델은 직접 사전학습 모델(Pre-trained Model)을 자체적으로 구축하여 적용했다. 사전학습을 위한 데이터는 금융 관련 데이터(금융 관련 공개된 상품 설명서, 금융 관련 뉴스, 금융 관련 커뮤니티 게시글)과 일반 데이터(경제 관련 뉴스, 일반 커뮤니티 게시글, 한국어 위키백과)를 게시물 기준 2:8 비율로 수집하여 구축했다.

사전학습을 위한 전처리로 [12]의 방법론을 참고하여 SentencePiece[13]를 통해 서브워드(sub-word)를 분절하고, 조사 및 어미가 결합되어 출력되거나 복합 명사로 출력된 부분에 대해 Mecab-ko[14]를 이용하여 추출한 체언, 용언, 부사 사전을 바탕으로 추가 분절하는 과정을 거쳤다.

3.2 의도 분류 모델

제안 의도분류 모델은 앞서 설명한 그림 1과 같은 구조를 가지고 있으며 Encoder Layer를 선별하여 연결하는 과정은 실험을 통해 최종적으로 선정했다. 연결된 Encoder Last Layer는 CNN 층을 거치는데 CNN을 단순히 구성하지 않고 Convolution Layer에서 Filter를 직병렬로 연결하여 특징 값을 조금 더 추출할 수 있도록 구성했다. 또한 단순히 연결된 Layer를 바로 Bi-LSTM 층으로 거치게 구성하면 학습 및 추론속도가 상당히 느려지는 단점이 있기 때문에 Bi-LSTM을 거치기 전에 CNN 층을 구성했다.

CNN을 통해 거친 층을 플랫하게 구성한 뒤, 연결된 값들을 앞뒤로, 순차적으로 학습시키고 Vanishing Gradient Problem을 보완해주기 위해 Bi-LSTM 층을 구성

하여 거친다. 마지막으로 완전 연결 신경망과 Softmax를 통해 문장이 가진 의도를 분류한다.

4. 실험

4.1 학습 데이터 셋

본 논문의 실험 의도는 구조가 다른 두 학습 데이터에 대해 Encoder Layer를 어떻게 구성하는 것이 높은 성능을 이끌어 낼 수 있는지 실험하는 것이다. 따라서 본 논문에서는 상단의 표 1과 같이 Semantic한 구성을 이루는 데이터(HIT_chatintent), Syntactic한 구성을 이루는 데이터(3i4k)를 학습하여 실험하고자 한다. 데이터의 구성은 표 2와 같다.

표 2. 학습데이터 구성

구분	Train	Test	Intent
HIT_chat intent	38,822	3,317	348
3i4k	55,134	6,121	7

HIT_chatintent의 학습 데이터는 상담 질의, 지식 구축 등으로 구성된 데이터로 질의의 핵심 단어, 구문 구성, 질의 출현 빈도를 분석하여 각 분류 기준당 약 50~300개 정도로 구성한 38,822개의 데이터이다. 테스트 데이터는 사용자 실제 질의 로그를 분석하여 분류기준 빈도 별로 각 분류 기준 당 5~20개 정도로 구성한 3,317개의 데이터로 구성되어 있다.

3i4k의 학습데이터는 7개의 분류 기준(조각구, 서술, 질문, 요구, 수사의문문, 수사명령문, 억양의존성 발화)으로 구성된 문장을 바탕으로 불균형 없이 구성한 55,134개의 데이터이다. 테스트 데이터 또한 7개의 분류 기준을 바탕으로 구성되었다. 실험에 앞서 사전학습 모델 및 전처리가 금융 도메인에 맞춰 구성되어 있으므로 3i4k를 학습하기에 최적의 조건이 아님을 미리 밝힌다.

표 3. HIT_chatintent 학습 모델 테스트(ALBERT)

구분	Layer	Accuracy(%)
ALBERT Base	8	86.70
ALBERT + CNN + Bi-LSTM	8	87.52
	2~8	88.11
	5,6,7,8	88.03
	2,3,4,5	88.60
	3,4,5,6	88.53
	3,4,5	**88.65**
	2,3,4	88.50
	3,4	88.45
	4,5	88.48

4.2 실험 방식

본 논문의 목표는 의도분류 모델의 데이터 구성에 따른 최적의 Encoder Layer 구성 방식과 그 차이를 보여주는 것이다. 따라서 학습이 여러 번 진행되어야 하며, 전체에서 일부 Layer를 소거하는 방식으로 실험을 진행했다. 본 실험에 구축된 모델은 총 8개 의 Encoder Layer로 구성되어 있으며, 최하단 Layer를 제외하고 구성하여 연결한 모델을 Tesla-V100 3대를 이용하여 3번 학습시킨 뒤 테스트 결과로 도출된 Accuracy의 평균을 표기했다. 표 3은 HIT_chatintent를 학습한 모델을 소거해 나가며 측정한 결과로, 상단 Layer보다 중하단 Layer의 조합에서 비교적 높은 성능을 보였고, 3,4,5 Layer를 연결한 모델이 가장 성능이 높음을 확인할 수 있다.

4.3 실험 결과

본 논문에서 수행한 모든 실험은 의도 분류 학습 데이터와 Encoder Layer를 선정하는 부분을 제외하고 모두 같은 조건에서 실험했다. ALBERT 및 ELECTRA 모델은 사전학습 당시 최적화를 위해 Base 모델에서 몇몇 파라메터를 조정했다. 앞서 설명한 바와 같이 본 실험은 8개의 Encoder Layer를 가진 모델을 기준으로 실험을 진행했다. 표 4는 HIT_chatintent와 3i4k에 대하여 4.2의 실험방식을 적용하여 실험한 결과이다.

표 4. 두 데이터에 대한 비교 테스트

구분	Accuracy(%) [Layers]	
	HIT_chatintent	3i4k
ALBERT Base	86.70 [8]	89.56 [8]
ALBERT + CNN + Bi-LSTM	87.52 [8]	89.74 [8]
	88.65 [3,4,5]	89.89 [7,8]
ELECTRA Base	88.12 [8]	90.82 [8]
ELECTRA + CNN + Bi-LSTM	88.35 [8]	91.03 [8]
	88.68[2,3,4,5]	91.25 [6,7,8]

실험 결과, 단순 Base 모델을 의도 분류 모델로 구축하는 방법론, Base 모델에 추가적인 신경망을 구성하여 구축하는 방법론보다 특정 Encoder Layer의 Last Layer를 연결하는 방식이 더 높은 성능을 보임을 확인할 수 있다. 그리고 데이터의 구성에 따라 Encoder Layer 선정 기준이 다름을 알 수 있다. 다만 Syntactic한 구성을 보이는 3i4k가 더 상위 Layer에서 높은 성능을 보이는 이유는 [10]에 의거, Finetuning이 진행되면서 Encoder Layer가 구성되어 나가는 과정에서 상부 Layer로 진행할수록 Syntactic한 정보를 개선해 나가는 것으로 판단된다. 그리고 HIT_chatintent 데이터의 구조 상 한정된 금융, 은행 관련 키워드로 이루어져 있어 하부 Layer에서 Semantic 정보가 어느정도 구축된 것으로 추측해 볼 수

있다. 결론적으로 데이터의 구성에 따라 성능 향상에 기여하는 Encoder Layer가 다름을 실험 결과로 도출할 수 있다.

5. 결론

본 논문에서는 Encoder 기반 사전학습 모델이 Encoder Layer의 구성에 따라 성능이 향상되는지, 그리고 데이터의 구성에 따라 Encoder Layer도 다른 구성을 보이는지에 대한 연구를 진행하였다. 그 결과 단일 Encoder Layer를 구성하여 신경망을 구성하는 방법보다 특정 Encoder Layer들의 Last Layer 조합을 연결하는 방법이 보다 높은 성능을 보임을 확인할 수 있었고, 데이터의 구성에 따라 높은 성능을 보이는 Encoder Layer 조합이 달라짐을 확인할 수 있었다.

하지만 본 논문에서 진행한 연구는 데이터 구성이 다른 두 종류의 학습 데이터 셋으로만 진행했기 때문에 조금 더 확정적인 결론 도출을 위해서는 데이터 셋을 추가로 확보한 뒤 이에 대한 실험이 필요할 것이다. 그리고 실험 환경의 한계로 인해 보다 많은 Encoder Layer를 두고 실험을 진행하지 못하였다는 한계가 있다.

향후 연구로 추가적인 실험 환경이 제공될 때 Encoder Layer 층을 더 두어 보다 더 세밀한 Layer 계층간의 차이를 비교 연구하는 실험, 의도 분류 데이터 셋 외에도 분류기준이 다양한 양질의 데이터 셋을 확보하여 추가적인 검증 실험을 진행할 예정이다.

참고문헌

[1] Jacob Devlin et al, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805, 2018.

[2] Kevin Clark, et al, ELECTRA: Pre-training text encoders as discriminators rather than generators, ICLR, 2020. [Online]. Available: https://openreview.net/pdf?id= r1xMH1BtvB.

[3] Tom B. Brown et al, Language Models are Few-Shot Learners, arXiv preprint arXiv:2005.14165, 2020.

[4] Won Ik Cho et al, Speech Intention Understanding in a Head-final Language: A Disambiguation Utilizing Intonation-dependency, arXiv preprint arXiv:1811.04231, 2019.

[5] Zhenzhong Lan et al, ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, arXiv preprint arXiv:1909.11942, 2019.

[6] Yoon Kim, Convolutional neural networks for sentence classification. Proc of the 2014 Conference on Empirical Methods in Natural Language Processing, 1746-1751. 2014.

[7] Akson Sam Varghese et al, Bidirectional LSTM Joint Model for Intent Classification and Named Entity Recognition in Natural Language Understanding, 2018 Intelligent Systems Design and Applications(ISDA), 58-68. 2018.

[8] Qian Chen et al, BERT for Joint Intent Classification and Slot Filling, arXiv preprint arXiv:1902.10909, 2019.

[9] Aysu Ezen-Can, A Comparison of LSTM and BERT for Small Corpus, arXiv preprint arXiv:2009.05451, 2020.

[10] Ian Tenney, et al, BERT rediscovers the classical NLP pipeline, the 57th Annual Meeting of the Association for Computational Linguistics, 4593-4601, 2019. In Proceedings of the 2020 Conference of the International Conference on Learning Representations, 2020.

[11] Wei-Tsung Kao et al, BERT's output layer recognizes all hidden layers? Some Intriguing Phenomena and a simple way to boost BERT, arXiv preprint arXiv:2001.09309, 2021

[12] Chang Woong, BERT with MECAB for Korean text, https://github.com/changwng/BERT-MECAB-Korean-Model

[13] Taku Kudo et al, SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, arXiv preprint arXiv:1808.06226, 2018.

[14] 은전한닢, Mecab-ko, https://bitbucket.org/eunjeon/mecab-ko/src/master/