

# 양방향 인재매칭을 위한 BERT 기반의 전이학습 모델

오소진\* · 장문경\*\* · 송희석\*\*\*

## A BERT-based Transfer Learning Model for Bidirectional HR Matching

Sojin Oh\* · Moonkyoung Jang\*\* · Hee Seok Song\*\*\*

### Abstract

While youth unemployment has recorded the lowest level since the global COVID-19 pandemic, SMEs (small and medium sized enterprises) are still struggling to fill vacancies. It is difficult for SMEs to find good candidates as well as for job seekers to find appropriate job offers due to information mismatch. To overcome information mismatch, this study proposes the fine-tuning model for bidirectional HR matching based on a pre-learning language model called BERT (Bidirectional Encoder Representations from Transformers). The proposed model is capable to recommend job openings suitable for the applicant, or applicants appropriate for the job through sufficient pre-learning of terms including technical jargons. The results of the experiment demonstrate the superior performance of our model in terms of precision, recall, and f1-score compared to the existing content-based metric learning model. This study provides insights for developing practical models for job recommendations and offers suggestions for future research.

Keywords : HR Matching, Job Recommendation, BERT, Transfer Learning, Pre-trained Language Model, Fine-tuning Deep Learning Model

Received : 2021. 05. 31.    Revised : 2021. 07. 11.    Final Acceptance : 2021. 07. 11.

\* First Author, Doctoral Student, First Author, Department of Management Information Systems, Hannam university, e-mail: clicku123@gmail.com

\*\* Co-author, Assistant Professor, Department of Global IT Business, Hannam university, e-mail: mk.jang@hnu.kr

\*\*\* Corresponding Author, Professor, Department of Global IT Business, Hannam university, 70 Hannam-ro, Daedeok-gu, Deajeon, 34430, Korea, Tel: +82-42-629-8344, e-mail: hssong@hnu.kr

## 1. 서 론

최근 심각한 사회적 문제로 대두되고 있는 청년 취업난은 신종 코로나바이러스 감염증 여파로 인해 사상 최악 수준으로 치달고[Statistics Korea, 2020] 있으나, 이러한 상황에서도 기업은 여전히 구인난을 겪고 있다. 청년 취업난과 구인난의 가장 근본적인 원인 중 하나는 기업들이 현장에서 실제 원하는 직무역량(Job competency)과 청년들이 가지고 있는 직무역량 간에 차이가 발생하는 이른바 '일자리 미스매치(Job-Mismatch)' 현상 때문이라고 할 수 있다(Son, 2015). 2018년 KDI mismatch statistics에 따르면 2014~2017년 평균 실업률 3.62% 중 일자리 미스매치 실업률(93%)이 노동력 부족 실업률(7%)보다 절대적으로 높게 나타났다(Kim, 2018).

일자리 미스매치를 해소하기 위해 본 연구에서는 최근 자연어 처리 및 기계번역 분야에서 우수한 성능을 나타내고 있는 사전학습 모델인 BERT(Bidirectional Encoder Representations from Transformers) 모델을 인재매칭 분야에 적용한 모델을 제안하고 그 성능을 평가함으로써 제안한 인재매칭 모델의 유용성을 가늠하고자 한다. 본 연구에서 인재매칭이라 함은 특정 기업의 채용 공고에 대해 요구되는 역량과 스킬을 보유한 구직자를 추천하는 것과 구직자가 보유한 역량과 스킬에 적합한 채용 공고를 추천하는 등 양방향 일자리 및 인재 추천을 포함하는 개념으로 정의한다. 효과적인 양방향 인재매칭을 위해서는 구직자의 프로파일과 채용 공고의 요구사항을 추천에 활용하는 내용기반 추천이 바람직하다고 알려져 있다(Song, 2020). Song(2020)의 연구에서는 양방향 추천을 위해 내용기반 척도학습 모델(Content-Based Metric Learning Model)을 제안한 바 있는데 이 방식은 수집된 단어들의 출현 빈도(Frequency)에 집중하여 단어 사전(BoW: Bag of Words)을 만들고 학습을 진행하고 있다. BoW 구축은 자연어로 작성된 구직자의 프로파일과 채용 공고에 기술된 단어의 출현 빈도를 이용하여 유사성을 학습하기 때문에 문장에 나타난 단어와 단어 사이의 순서가 고려되지 않는 한계점이 있다. 또한 단어 사이에 내재되어 있는 숨은 의미와 관계를 파악하기 어렵고, 단어 사전에 구축되어 있지 않는 새로운 단어에 대해서는 유사성을 학습할 수가 없다는 단점이 있다.

이러한 한계를 극복하기 위해 본 연구는 자연어 처리 사전 학습 모델인 BERT를 이용한 양방향 인재매칭 모델을 제안하고, 기존의 내용기반 척도학습 모델(Content-based metric learning model)과 성능을 비교함으로써 제안한 모델의 유용성을 확인하고자 한다. 본 연구에서 제안한 인재매칭 모델을 상용화 한다면, 사전 학습된 BERT 모델을 적용하였기 때문에 추가로 단어 사전 구축이 필요하지 않기 때문에, 그와 관련한 비용 및 시간을 절약할 수 있을 것으로 기대한다. 제안한 모델은 구직자에게는 취업 정보 검색 비용을 낮추고, 기업에게는 채용을 위한 추가적인 비용을 줄일 수 있으며 나아가 구직자와 기업 모두에게 일자리 미스매치를 해소할 수 있는 플랫폼을 구현하는데 사용될 수 있을 것이다.

본 연구는 다음과 같이 구성된다. 먼저 제2장에서는 인재매칭을 위한 일자리 추천시스템과 자연어 처리 사전학습모델에 대한 기존 문헌을 검토한다. 제3장에서는 인재매칭을 위한 내용기반의 척도학습의 기존 모형을 검토하고, 이를 바탕으로 양방향 인재매칭을 위한 BERT 기반의 전이학습 모델 설계 대안을 제시한다. 제4장에서는 제안한 모델의 성능을 평가하기 위한 실험용 데이터 셋, 성능 평가지표 및 실험 결과를 요약한다. 마지막으로 제5장에서는 본 연구의 기여점과 후속 연구를 위한 제언을 제시한다.

## 2. 기존 연구

### 2.1 일자리 추천시스템

기존 일자리 추천 연구는 구직자의 프로파일과 채용 공고의 요구사항을 추천에 활용하는 내용기반 추천으로 주로 이뤄졌다. Pessemier et al.[2016] 연구에서는 사용자 이력서와 관심 채용 공고 클릭 여부 등의 지원 이력 및 관심도를 토대로 사용자가 선호하는 채용 공고 프로파일을 생성하고, 사용자의 프로파일-미지의 채용 공고와의 유사도 점수를 기반으로 추천을 수행하는 내용기반 추천(Content-based)과 K-NN 기반 추천을 사용자 지원 이력 및 관심도의 빈도로 가중 평균하는 하이브리드 일자리 추천방법을 제안하였다. Valverde-Rebaza et al.[2018] 연구에서는 구직자의 이력서 정보를 모두 사용하지 않고 일자리 추천에 영향을 미치는 구직자의 전문스킬 정보만을 추출하여

워드 임베딩을 얻는 방식을 사용하면 추천 성능에 영향을 줄 수 있다고 제안하였다. 전문 스킬에 대한 용어 추출 및 표현방식에 따라 벡터공간모델 기반의 워드임베딩 방식과 빈도기반의 TF-IDF(Term Frequency - Inverse Document Frequency) 방법을 사용하여 각각 성능을 비교하였다. IT분야의 스킬에만 국한하여 일자리 매칭 커뮤니티 링크드인(LinkedIn)에서 50명의 프로파일을 수집하고 Catho에서 3,877개의 일자리 프로파일을 수집하여 실험한 결과 CBOW(Continuous Bag Of Word)방식과 Skip-gram 방식 중 Skip-gram 임베딩을 활용한 추천이 가장 좋은 성능을 보였다고 보고하고 있다. Le and Mikolov(2014)는 BoW 및 TF-IDF와 같은 모형은 자연어 처리작업에 효과적이지만 단어의 출현 빈도에만 초점을 두기 때문에 맥락을 이해하는데 있어서는 부족하다고 지적하며, 맥락을 이해하도록 하기 위해서는 Word2vec 및 GloVe 등과 같은 임베딩 모형을 활용하여 벡터화 할 것을 권고하고 있다. Maheshwary and Misra(2018)은 구직자의 반구조화된(Semi-structured) 이력서를 채용 공고의 직무설명과 매칭하여 구직자에게 적절한 일자리를 추천하는 삼 신경망 기반의 일자리 추천 모형을 제안하였다. 이처럼 기존의 일자리 추천 연구에서는 구직자의 이력서 정보와 채용 공고를 기계학습 모형에 반영하는 전처리 과정에서 다양한 벡터화 기법에 초점을 맞춰 연구가 진행되고 있다.

Song(2020)의 연구에서는 상호작용 이력이 있는 채용 공고와 구직자 이력서 사이의 적합도를 학습하여 특정 구직자의 이력서에 적합한 일자리 또는 특정 채용 공고에 적합한 구직자를 매칭하는 양방향 척도학습모형을 제안하였다. 이 연구에서는 구직자ID와 일자리ID 대신 구직자 프로파일과 채용 공고문서의 텍스트를 BoW 벡터로 변환하여 동시에 입력한 후 심층 신경망을 거쳐 최종 출력층에서 적합도 여부를 학습하는 내용기반 척도학습모형을 제안하였다. 이 방법은 구직자 프로파일과 채용 공고 문서를 모형에 입력하기 위해 BoW 벡터로 변환하는 전처리 과정을 거친다. 이렇게 입력된 구직자 프로파일과 채용 공고문서 벡터는 각각 별도의 임베딩층을 거친 후 결합(Concatenation) 또는 내적(Dot Product)을 통해 통합되어 심층신경망을 통과하게 된다. BoW 입력층 위에 임베딩층을 둔 것은 동일한 구직자들이 지원한 일자리를 유사한 일자리 벡터

로 표현하거나 동일한 일자리에 지원한 구직자 프로파일을 유사한 구직자 프로파일 벡터로 투사함으로써 학습결과 입력벡터에 시맨틱을 부여하기 위함이다. 최종 출력층은 해당구직자 프로파일 벡터와 일자리 벡터 간 지원이력 등의 지원이력 내역을 나타내는 한 개의 노드로 구성되며 지원이력 내역이 있으면 1 없으면 0을 출력하도록 학습이 이루어진다. 내용기반 척도학습 모형은 협업필터링기반의 추천모형과 달리 신규구직자 또는 신규 채용 공고에 대해서도 추천이 가능한 장점이 있으나 구직자 프로파일과 채용 공고에서 수집한 데이터만을 활용하여 직접 훈련을 통해 임베딩 벡터를 획득하고 있기 때문에, 다양한 전문 용어에 대한 의미적 유사성을 모두 반영할 수 없는 단점이 있다. 이를 극복하기 위해 풍부한 사전 훈련된 언어모형을 활용한 전이학습이 주목받고 있다. 최근 딥러닝 기술이 활발하게 적용되고 있는 분야 중 하나가 자연어 처리(NLP; Natural Language Processing) 분야이다. 기존 자연어 처리 연구에서는 워드 임베딩 모형을 활용하여 벡터화 [Le and Mikolov, 2014] 하였으나, 자연어 처리 분야에서는 이미지 처리의 전이 학습처럼 사전 학습된 언어 모델을 이용하는 것이 성능향상에 도움이 된다고 보고하고 있다[Yoo and Jeong, 2019].

## 2.2 어텐션 메커니즘(Attention mechanism)과 BERT

자연어 처리 분야와 기계 번역 분야에서 LSTM(Long Short-Term Memory)과 GRU(Gated Recurrent Units)와 같은 순환신경망(RNN; Recurrent Neural Network)은 대부분 SOTA(State-Of-The Art) 성능을 제공하고 있다 [Bahdanau et al., 2014; Cho et al., 2014; Sutskever et al., 2014]. 그러나 순환신경망 모델은 순차적으로 데이터를 처리해야 하는 특성 때문에 여러 작업을 동시에 수행하지 못하고, 병렬처리의 효율도 떨어지게 된다. 이러한 문제를 해결하기 위해 ByteNet, ConvS2S 등과 같이 합성곱 신경망(CNN; Convolutional Neural Network)을 사용하는 경우들도 있지만 합성곱 신경망을 사용하는 경우, 입력의 길이가 길어질수록 입력의 원소와 출력의 원소 사이의 의존성(Dependency)을 학습하는 것을 더 어렵게 만든다[Hochreiter et al., 2001]. 또한 입력 문장

이 긴 상황에서 자연어 처리나 번역의 품질이 떨어지는 현상이 나타났고 이러한 현상을 보정하기 위해 중요한 단어에 집중하는 어텐션 메커니즘(Attention mechanism)이 자연어 기계 번역을 위한 seq2seq(Sequence-to-sequence) 모델에 처음 도입되었다 [Sutskever et al., 2014; Cho et al., 2014]. seq2seq 모델은 하나의 인코더(Encoder)와 하나의 디코더(Decoder)로 이루어져 있으며, 인코더는 입력의 각 아이템을 처리하여 정보를 추출한 후 하나의 벡터인 컨텍스트(Context)로 만든다. 디코더는 컨텍스트를 받아 출력할 아이템을 하나씩 선택한다. 디코더는 하나의 고정된 벡터로 전체의 맥락을 표현하기 위한 방법으로, 긴 문장과 같이 처리하기 어려운 문제에 대한 해결 방법으로 제시된 것이 바로 어텐션이다 [Bahdanau et al., 2014; Luong et al., 2015]. 어텐션 메커니즘은 디코더에서 출력 단어를 예측하는 매 시점(Time-step)마다, 인코더에서의 전체 입력 문장을 다시 한 번 참고하도록 한다. 단, 전체 입력 문장을 전부 다 동일한 비율로 참고하는 것이 아니라 해당 시점에서 예측해야 할 단어와 연관이 있는 입력 단어 부분을 좀 더 집중적으로 보게 된다. 어텐션 메커니즘은 seq2seq 모델이 디코딩 과정에서 현재 스텝에서 가장 관련된 입력 파트에 집중할 수 있도록 해줌으로써 기계 번역의 품질을 크게 향상시켰다. 또한 다양한 작업에서 강력한 시퀀스 모델 및 변환 모델의 필수적인 부분이 되었으며, 입력과 출력의 원소들 사이의 거리와 무관하게 의존성을 학습할 수 있게 되었다 [Bahdanau et al., 2014; Kim et al., 2017]. 대부분의 경우 어텐션은 기존의 신경망과 함께 사용되어 왔으나 Vaswani et al. [2017] 연구에서는 기존 순환 신경망에서 발생하는 문제들을 해결하고 Self-attention을 이용하는 트랜스포머(Transformer)를 제안하였는데 이 연구에서는 반복(Recurrence)을 제거함으로써 오직 어텐션 메커니즘만을 이용해서 입력과 출력 사이의 전역 의존성(Global dependency)을 학습하게 된다. 반복 작업이 없어졌으니 당연히 더 손쉽게 병렬처리가 가능해진다. 또한, 합성곱 신경망을 사용했을 때 문제였던 원소들 사이의 거리가 매우 멀어지는 문제도 상수개의 연산으로 고정시킬 수 있는 매우 큰 장점이 있다. 트랜스포머의 가장 큰 특징 중 하나는 Self-attention을 수행한다는 것이다. Self-attention은 동일한 시퀀스(입력, 출력 등) 내의 원소들 사이에서 어텐션을 수행하는

것이다 [Vaswani et al., 2017].

BERT(Bidirectional Encoder Representations from Transformers)는 2018년에 구글이 공개한 사전 훈련된 모델이다. BERT는 트랜스포머를 이용하여 구현되었으며, 위키피디아(25억 단어)와 BooksCorpus(8억 단어)와 같은 레이블이 없는 텍스트 데이터로 사전 훈련된 언어 모델이다. 최근에는 104개 언어들을 지원하는 'BERT-Base, Multilingual Cased' 모델을 제공한다. 영어 뿐만 아니라 한글을 대상으로 BERT-Base, Multilingual Cased 모델에 한국어어를 더 학습시키는 연구가 진행되고 있다 [SKTBrain]. BERT는 모든 레이어에 양방향 문맥에서 공동으로 조절함으로써 라벨이 없는 텍스트로부터 깊은 양방향 표현들을 사전 훈련하도록 디자인되었다. 이에 대한 결과로, 사전 훈련된 BERT 모델은 작업별 구조 수정을 크게 할 필요 없이 질문 답변과 언어 추론과 같은 넓은 분야의 작업에서 하나의 출력 레이어만 추가하는 미세조정(Fine-tuning)을 통해 최고 성능(SOTA) 모델을 만들 수 있다 [Devlin et al., 2018]. BERT 모델 학습은 사전학습(Pre-training)과 미세조정단계로 나눌 수 있다. BERT 모델은 언어 표현을 위해 양방향(Bidirectional Representations)을 사전학습하게 디자인되어 있다. 위키피디아(25억 단어)와 BooksCorpus(8억 단어) 문장을 기반으로 언어 표현을 이해하기 때문에 단어 간에 드러난 의미와 관계뿐만 아니라 숨은 의미도 찾아낼 수 있다. 예를 들어 "J2EE 및 스프링프레임워크 개발자"를 찾는 채용 공고에 대해 "JAVA MVC" 역량이나 "자바서블릿" 역량을 보유한 구직자에 대한 적합도 예측이 가능하다. BERT의 이러한 특징을 활용하여 자연어 처리 분야에 BERT를 이용한 연구가 활발하게 이루어지고 있다. Hong et al. [2020]은 BERT를 이용한 알츠하이머병 치매와 조현병 진단모델을 제안하였고 Han et al. [2020] 연구에서는 BERT 언어 모델을 사용한 언어처리 기반 질의-상대 매칭 시스템을 제안하는 등 대화나 상담 글에서 직접적으로 나타나지 않는 숨은 의미와 연관성을 찾는 부분에서 BERT를 이용하는 연구가 증가하고 있다.

본 연구에서는 채용 공고나 구직자 프로파일과 같은 문서 내 단어의 숨은 의미를 이해할 수 있는 사전 학습 모델을 적용하면 기존 일자리 추천에서 사용되는 단순 워드 임베딩을 활용한 학습 모델보다 높은 성능을 나타

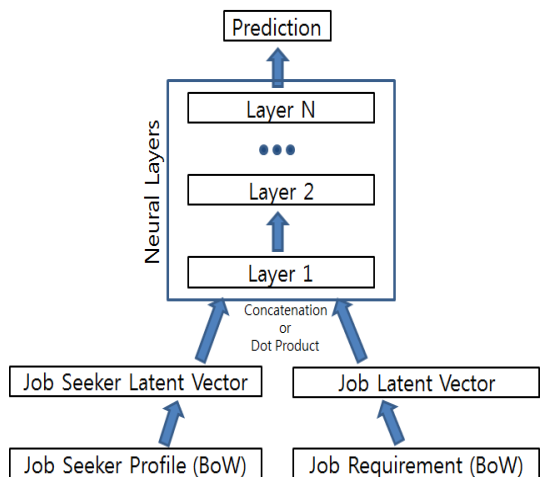
널 것이라는 아이디어에 착안하여 BERT 기반의 전이 학습 모델을 제안하고 성능을 평가하기로 한다.

### 3. BERT 기반의 전이학습 모델 설계

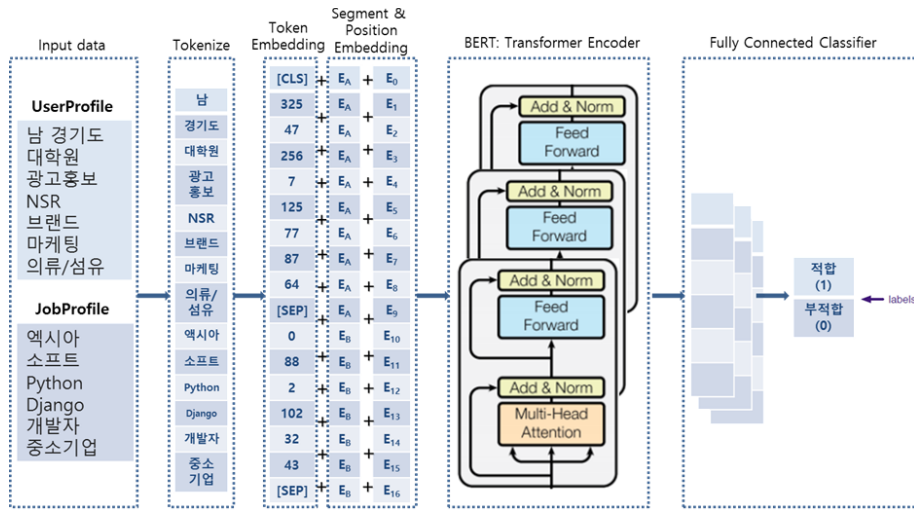
본 연구는 기업에게는 적합한 구직자를 추천할 수 있어야 하며, 구직자에게는 보유한 역량과 관심에 적합한 일자리를 추천하는 양방향 인재매칭 모델을 제안하는 것을 목표로 한다. 여기서는 구직자의 프로파일(이력서)과 채용 공고 정보를 BertTokenizer를 이용하여 BERT 사전 학습 모델에 입력하며, 최종 출력층에는 구직자 정보와 채용 공고의 적합도가 출력되도록 하는 학습 구조를 제안한다. 본 연구에서는 104개 언어들을 지원하는 'BERT-Base, Multilingual Cased' 모델을 적용하여, 구직자 정보와 채용 공고의 적합도가 출력되는 미세조정 모델(Fine-tuning model)을 구성한다(Devlin et al., 2018).

본 연구에서는 구직자 프로파일과 채용 공고문서의 텍스트를 BoW 벡터로 변환하여 동시에 입력한 후 심층 신경망을 거쳐 최종 출력층에서 적합도 여부를 학습하는 내용기반 척도학습모형(Song, 2020)을 양방향 인재추천 모형의 벤치마킹 대상모형으로 선정하여 성능평가를 실시하기로 한다. 내용기반 척도학습모형(Content-based metric learning model for HR matching)은 <Figure 1>과 같이 구직자 프로파일과 채용 공고 문서를 모형에 입력하기 위해 BoW 벡터로 변환하는 전처리 과정을 거친다. 이렇게 입력된 구직자 프로파일과 채용 공고문서 벡터는 각각 별도의 임베딩층을 거친 후 결합(Concatenation) 또는 내적(Dot Product)을 통해 통합되어 심층신경망을 통과하게 된다. BoW 입력층 위에 임베딩층을 둔 것은 동일한 구직자들이 지원한 일자리들을 유사한 일자리 벡터로 표현하거나 동일한 일자리에 지원한 구직자 프로파일을 유사한 구직자 프로파일 벡터로 투사함으로써 학습결과 입력벡터에 시맨틱을 부여하기 위함이다. 최종 출력층은 해당구직자 프로파일 벡터와 일자리 벡터 간 지원이력 등의 상호작용 내역을 나타내는 한 개의 노드로 구성되며 상호작용 내역이 있으면 1 없으면 0을 출력하도록 학습이 이루어진다. 내용기반 척도학습모형은 텍스트를 BoW로 변환할 때 BoW의 희박도(Sparsity)가 높아 학습이 어려울 뿐 아니라 단어의 순서를 고려하

지 않기 때문에 문맥적인 의미가 무시되고, 사전에 구축되지 않은 단어(OOV: Out Of Vocabulary)에 대해서는 의미적 관계를 파악하기 어렵다는 단점이 있다. 따라서 본 연구에서는 이미 훈련된 임베딩 벡터를 활용하는 미세 조정 모델을 <Figure 2>와 같이 제안한다. BERT 모델을 이용한 양방향 인재매칭 모델은 <Figure 2>과 같이 구직자의 프로파일(이력서)과 채용 공고를 BERT 사전 학습 모델에 반영 할 수 있도록 Tokenize, Token Embeddings, Segment Embeddings & Position Embeddings과 같은 전처리 과정을 거친다. 구직자 프로파일과 채용 공고는 각각 구직자 프로파일(UserProfile)과 채용 공고(JobProfile)를 컬럼으로 변환 후 하나의 데이터로 통합한다. 이렇게 통합된 데이터를 BertTokenizer를 이용하여, 텍스트를 더 이상 나눌 수 없는 단어 형태(Token)로 분리한다. 분리한 단어들은 Token Embeddings을 통해 단어의 고유 값(ID)을 부여하고 Segment Embeddings & Position Embeddings 통해 문장의 순서와 위치 값을 설정한다. 이러한 전처리 과정을 거친 입력 데이터를 BERT 모델의 입력(Input)으로 한다. 최종 출력층을 추가하는 미세조정 단계에서는 BERT 모델의 입력(구직자 프로파일-채용 공고)에 대한 지원이력 등의 상호작용 내역을 반영한 한 개의 노드로 구성하여 매칭이 부적합하면 0 적합하면 1을 출력하도록 학습이 이루어진다.



<Figure 1> Content-based Metric Learning Model for HR Matching [Adapted from Song, 2020]



〈Figure 2〉 BERT-based Fine-Tuning Model for HR Matching

## 4. 실험

### 4.1 데이터 셋

실험에 사용한 데이터는 국내 헤드헌터 업체에 축적된 데이터로 6,028명의 구직자 프로파일과 352개의 채용 공고 정보, 그리고 지난 수년간 이들 구직자가 특정 채용 공고에 지원한 이력 데이터 60,362건으로 구성되어 있다. 이 데이터는 긍정사례 17,995건(29.8%)과 부정사례 42,367건(70.2%)으로 구성되어 있다. 구직자 프로파일은 성별, 소재지, 대학교, 최종학력, 전공, 경력연수, 근무회사, 스킬, 희망 연봉, 전문경력, 경력상세정보와 장문의 이력서 정보를 텍스트 형태로 포함하고 있다. 채용 공고는 헤드헌터가 기업으로부터 의뢰받은 정보로 채용직무, 기업형태, 회사소개, 직무소개, 자격요건, 우대사항 등의 텍스트 형태이다. 마지막으로 지원 이력정보는 구직자 ID, 채용 공고 ID, 지원 여부(1: 지원, 0: 지원 안함)로 구성된 데이터이다. 훈련용 데이터는 전체 데이터의 80%, 시험용 데이터는 전체 데이터의 20%로 임의 추출하여 실험을 진행하였다.

### 4.2 성능평가 방법

여기서는 인재매칭의 방식으로는 성능 평가지표를 제시하여 성능평가를 수행한다. 추천문제를 정의하는 방법은 특정 구직자가 특정 채용 공고에 적합한지 여부

를 예측하는 방법으로, 추천문제를 0 또는 1의 이진 분류문제로 정의한다. 0은 채용 공고에 적합 또는 구직자의 적합하지 않을 경우 부정을 나타내며, 1은 채용 공고에 적합 또는 구직자의 적합한 것으로 긍정을 나타낸다. 성능평가를 위한 지표로 정밀도와 재현율, f1-score, AUC(Area Under Curve)를 사용하였다. 본 연구에서 정밀도는 “추천모형에서 긍정(1)이라고 예측된 구직자-채용 공고 쌍 중 실제 지원한 적이 있던 구직자-채용 공고 쌍의 비율”로 계산한다. 재현율은 “실제 지원한 적이 있던 구직자-채용 공고 쌍 중 추천모형에서도 긍정(1)으로 예측된 구직자-채용 공고 쌍의 비율”로 계산한다. f1-score는 정밀도와 재현율을 이용하여 조화평균(Harmonic mean)을 이용한 값으로, 주로 분류 클래스 간 데이터가 심각한 불균형(Imbalanced data)을 이루는 경우에 사용한다. 각 성능지표의 계산 공식은 다음과 같다.

$$\text{정밀도(Precision)} = \frac{\|P \cap A\|}{\|P\|}$$

$\|P \cap A\|$  : 긍정(1)으로 예측된 구직자-채용 공고 쌍 중 실제 지원이력이 있는 구직자-채용 공고 쌍의 수

$\|P\|$  : 긍정(1)으로 예측된 구직자-채용 공고 쌍의 수

$$\text{재현율(Recall)} = \frac{\|A \cap P\|}{\|A\|}$$

$\|A \cap P\|$  : 실제 지원이력이 있는 구직자-채용 공고 쌍 중 추천모형에서도 긍정('1')으로 예측된 구직자-채용 공고 쌍의 수

$\|A\|$  : 실제 지원이력이 있는 구직자-채용 공고 쌍의 수

$$f1-score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

AUC값은 ROC(Receiver Operating Characteristic) 그래프의 면적으로 계산하며 최대값은 1이 된다. ROC 그래프는 모든 임계값에서 분류 모델의 성능을 보여주는 값이다. 일반적으로 이진분류 문제의 경우 분류 임계값(Classification threshold)에 따라 예측 정확도가 달라질 수 있다. 분류 임계값을 낮게 설정하면 더 많은 사례가 양성으로 분류되기 때문에 참 양성율과 거짓 양성율이 모두 높아지게 된다. AUC는 클래스를 구별하는 모델의 성능을 해석할 때 사용되며, 특정모형에 대해 임의의 긍정사례를 임의의 부정사례보다 더 우수한 성능으로 평가할 확률로 해석할 수 있다.

이외에도 실제 추천 시스템에 적용하기 위해서는 특정 사용자에게 가장 적합할 것으로 예상되는 N개의 일 자리를 추천하는 Top-N 추천이 중요하다. 본 연구에서는 Top-N 추천의 성능을 평가하기 위해 추천하고자 하는 Top-N 추천 재현율(REC@N: Recall and Top N)과 Top-N 추천 정확도(AAS@N: Average Accuracy Score and Top N)의 두 가지 성능평가 지표를 사용하였다[Song, 2020]. REC@N은 과거 사용자가 지원 이력이 있는 채용 공고가 Top-N 추천 리스트에도 포함되어 있는지를 계산하는 것으로, 사용자별로 과거에 지원한 채용 공고가 Top-N 추천 리스트에 포함되어 있으면 1, 없으면 0으로 계산하여 전체 사용자에게 대해 평균한 값으로 계산된다.

$$REC@N = \frac{1}{m} \sum_{i=1}^m \Pi_i^{j_i}, \Pi_i^{j_i} = \begin{cases} 1, & \text{if } j_i \in \text{Top } N \text{ list for } i \\ 0, & \text{else} \end{cases}$$

where  $i$ : user,  $j_i$ : Job post  $j$  to which user  $i$  had been applied

AAS@N은 보다 정밀하게 Top-N 추천의 정확도를 평가하기 위한 지표로, 과거 구직자가 지원이력이 있는 채용 공고가 Top-N 추천 리스트에 몇 번째로 나타났는지를 고려한 지표이다. 특정 구직자가 지원한 바 있던

채용 공고가 Top-N 추천 리스트의 맨 처음에 나타나면 1을, k번째 나타나면  $(N-(k-1))/N$ 을, 추천 리스트에 없으면 0을 점수로 부여하는 방법으로 다음과 같은 식으로 계산한다.

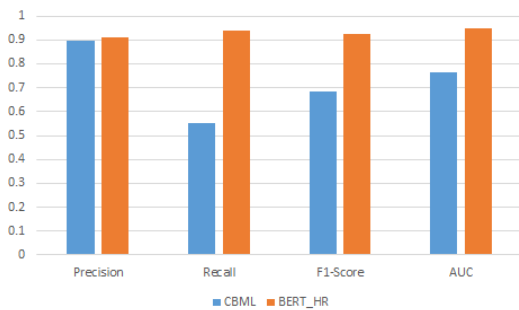
$$AAS@N = \frac{1}{m} \sum_{i=1}^m X_i^{j_i}, X_i^{j_i} = \frac{N - (\text{rank}_{j_i} - 1)}{N}$$

where  $\text{rank}_{j_i}$ : rank of  $j_i$  in Top N list for  $i$

### 4.3 실험 결과

실험에 사용된 첫 번째 모델은 채용 공고 벡터와 구직자 프로파일 벡터를 결합(Concatenation)방식에 의해 통합하는 내용기반 척도학습 모형(CBML)이다. 두 번째 모형은 BERT 모델을 이용한 양방향 인제매칭 모델(BERT\_HR)이다. 내용기반 척도학습 모형(CBML)은 채용 공고 안에서 단어 사전을 만드는 방식인 BoW를 사용하여 상위 2,000개의 어휘를 사용하여 학습에 사용하였고, BERT 모델을 이용한 양방향 인제매칭 모델(BERT\_HR)은 구직자 프로파일 6,028건, 채용 공고 352건, 지원한 이력 데이터 60,362건의 데이터를 BERT 사전학습 모델에 임베딩하기 위해 UserProfile(구직자 프로파일), JobProfile(채용 공고), Label(지원이력)을 하나의 파일로 통합하는 전처리 과정을 통하여 59,694건의 통합 데이터를 구성하였다. 훈련용 데이터는 전처리를 통한 통합 데이터의 80%, 시험용 데이터는 전체 데이터의 20%로 임의 추출하여 훈련용 데이터 47,755건, 시험용 데이터 11,939건을 실험에 사용하였다. 모델 학습에는 한국어 포함 104개 언어들을 지원하는 'BERT-Base, Multilingual Cased' 사전 학습된 모델을 사용하였고, 임의로 추출된 훈련용 47,755건의 데이터를 BertTokenizer를 이용하여 BERT 사전 학습 모델에 임베딩 할 수 있는 입력층을 통해 추가 학습에 사용하였다. 성능 평가에서는 학습에 사용되지 않은 시험용 데이터 11,939건을 사용하여 정밀도, 재현율, f1-score, AUC를 측정하였다. 실험 모델의 하이퍼 파라미터(Hyper parameter) 설정은 다음과 같다. 먼저 옵티마이저(Optimizer)는 RectifiedAdam을, 학습율은 1.0e-5, 손실함수(Loss)는 BinaryCrossentropy를 사용하고, 과대적합(Overfitting) 문제를 해결하기 위해서 Weight decay를 0.0025 설정하였다. 모델 훈련은 총 4번(epochs = 4)을 진행하였다.

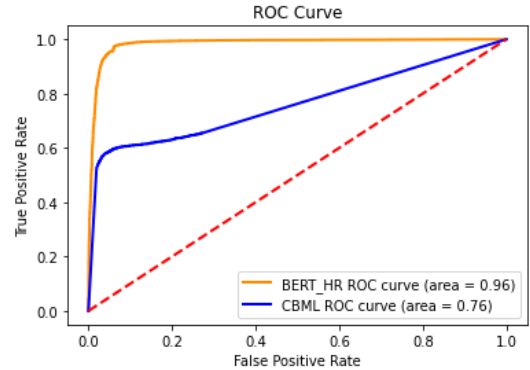
(Figure 3)는 각 모형별 정밀도와 재현율, f1-score, AUC를 보여준다. 정밀도와 재현율, f1-score, AUC 측면에서 CBML 모형이 0.8961, 0.5514, 0.6827, 0.7628을 기록해 정밀도는 높지만 재현율은 낮게 나타났다. BERT\_HR 모형은 0.9104, 0.9371, 0.9236, 0.9487을 기록해 CBML 모형보다 모든 측면에서 높게 나타났다. 구직자에게 적합한 모든 채용 공고가 추천되었는지를 확인 하는 지표인 재현율은 BERT\_HR 모형이 CBML 모형에 비해 아주 높은 성능을 나타냈다. 이는 CBML 모형의 경우 BoW를 사용하기 때문에 학습되지 않은 실험 데이터의 사전에 구축되지 않은 단어 문제를 해결하기 어렵지만, BERT\_HR 모형은 이미 사전 학습된 모델을 적용했기 때문에 새로운 데이터에 대해서 구직자 프로파일과 채용 공고 사이에 다양한 전문용어 간의 의미적 관계를 파악 할 수 있었기 때문으로 해석된다. 실제 인재매칭 추천 시스템에서는 채용 공고가 구직자에게 실제로 적합한 채용 공고 인지를 확인하는 지표인 정밀도를 높게 유지하는 것이 중요하다. 채용 공고와 구직자 추천문제에서 재현율보다 정밀도가 보다 중요한 지표라는 사실을 감안하더라도 두 가지 모델 중 BERT\_HR 모형이 모든 성능지표에서 가장 높은 성능을 나타냈으며, BERT\_HR 모형이 인재매칭 시스템 실제 적용이 가능할 것으로 판단된다.



〈Figure 3〉 Precision, Recall, f1-score, and AUC by Models

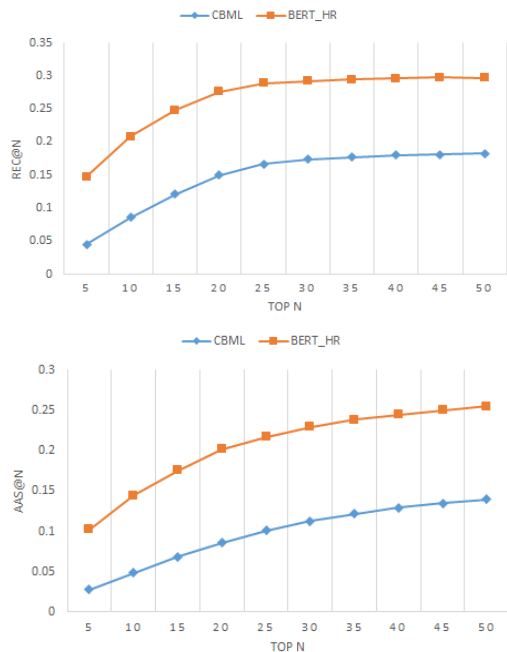
(Figure 4)는 두 모형의 ROC 그래프를 나타내고 있다. ROC 그래프의 곡선(Curve)은 모든 임계값에서 분류 모형의 성능을 보여주는 것으로, 이진 분류 모형에서 두 클래스(0: 부적합; 1: 적합)를 더 잘 구별할 수 있을 때 그래프의 좌측 상단에 더 가까워지게 된다. BERT\_HR 모형이 CBML 모형보다 ROC 곡선이 그래프의 좌측 상단에 더 가깝게 그려지고 있기 때문에

더 좋은 분류 모델이라고 해석할 수 있다.



〈Figure 4〉 Receiver Operating Characteristic Curve by Models

다음으로, 모형별로 Top-N 추천 개수를 최소 5개에서 50개까지 변화시키면서 추천 성능을 비교하였다. 〈Figure 5)는 BERT\_HR, CBML 모형의 REC@N과 AAS@N를 각각 나타내고 있다. 실험결과 REC@N과 AAS@N 모두 BERT\_HR 모형이 CBML 모형 보다 높은 성능을 나타내고 있다. 그리고 Top-N 개수에 따라 성능차이가 나타나지만 Top 25개 이후 비교적 일



〈Figure 5〉 REC@N and AAS@N by Models



정한 성능을 보인다. 따라서 실제 인제매칭 추천시스템에서의 Top N 기본값을 25로 설정하는 것이 효과적일 것으로 예상된다.

## 5. 결론

본 연구에서는 특정 기업의 채용 공고에 대해 요구되는 역량과 스킬을 보유한 구직자를 추천하는 것과 구직자가 보유한 역량과 관심에 적합한 채용 공고를 추천하는 양방향 인제매칭을 위해 자연어 처리 사전 학습 모델인 BERT 모델을 이용한 양방향 인제매칭 모델을 제안하였다. 제안 모델은 기존 내용기반 척도학습모형보다 정밀도와 재현율, f1-score, AUC, Top-N 추천인 성능평가 REC@N과 AAS@N 에서 높은 성과를 나타내었다. 특히, 재현율을 볼 때 학습되지 않는 새로운 데이터인 신규 이력서 및 채용 공고의 의미를 분석하여 추천하는 부분에서 사전 학습 모델이 높은 성능을 낼 수 있다는 것을 보여주고 있다. 또한 Top-N 추천 성능평가에서도 BERT 기반 모델이 높은 성능을 나타내었다.

본 연구는 다양한 측면에서 의의를 가진다. BERT 모델을 이용한 양방향 인제매칭 모델을 제안하였으며, 제안 모델은 양방향 추천이 가능할 뿐 아니라, 사전 학습 모델을 사용하여 문맥적인 의미가 무시되었던 문제점을 해결한다. 기존 연구에서 활용되었던 BoW, TF-IDF 등과 같은 자체적인 사전 구축을 통한 임베딩에서 문체시 되고 있는 희소 행렬로 학습 시 계산의 복잡성이 증가하고 단어의 순서를 고려하지 않기 때문에 문맥적인 의미를 파악하기 어려웠다. 사전에 구축되지 않은 단어(예. 줄임말, 의미는 같으나 다른 단어 등)에 대한 의미적 관계 파악하기 어려웠던 문제점 또한 제안 모델에 사전 학습 모델인 BERT를 적용하여 해결할 수 있었다. 구직자의 프로파일 텍스트와 채용 공고 등과 같은 긴 장문의 텍스트 데이터 전처리 작업을 BERT에서 제공하는 임베딩을 통해 별다른 전처리 작업을 거치지 않고 토큰, 인덱스, 세그먼트, 포지션으로 쉽게 데이터를 학습시킬 수 있었다. 그리고 추천 시스템에서 중요하게 해결하고자 하는 문제인 콜드 스타트 문제(Cold-start problem)도 BERT 모델을 통해 해결할 수 있었다. BERT 모델은 문장을 기반으로 언어 표현을 이해하기 때문에 신규 구직자, 신규 채용 공고에 대해서 재학습 없이 인제매칭이 가능하기 때문이

다. 제안 모델은 자체적인 용어 사전 구축을 위한 많은 학습 데이터의 획득을 위해 비용이 많이 들어가는 인제매칭 분야에서 사전 학습 모델을 적용하여 데이터 획득 비용을 줄이고 매칭 성능을 향상시킬 수 있는 효과적인 방법으로 평가되었다.

향후 연구로는 우선 본 연구에서 제안한 모델의 성능 개선을 위해 하이퍼 파라미터 튜닝이 가능할 것이다. 제안 모델은 4회의 훈련을 통해 높은 성능을 보여주었으나 여전히 개선 여지가 남아 있고 매칭 오류를 개선하기 위해서는 훈련 횟수(Epoch)와 배치사이즈(Batch size), 손실함수 등 변경 등의 경험적 실험을 통해 최적의 설정을 위한 하이퍼파라미터 튜닝이 가능할 것이다. 그리고 한국어 인제매칭에 더욱 특화된 모델 개발을 위해 BERT 기본적으로 제공하는 'BERT-Base, Multi-lingual Cased' 모델 외에 대량의 한국어 코퍼스를 통해 기존 BERT 다국어 모델을 재학습시킨 SKTBrain의 KoBERT 모델을 적용하여 성능을 비교하고, 지속적인 성능 개선을 진행할 수 있을 것이다. 또한 채용 공고나 인제 매칭 시 조건을 활용한 필터링 기능 등 정량적 구간 조건 비교 기능이 추천시스템과 결합된다면 정교한 추천 및 매칭 정확도를 높일 수 있을 것이다.

본 연구에서 제안한 BERT 모델을 이용한 양방향 인제매칭 모델이 인제매칭 시스템으로 상용화된다면 기업과 구직자 모두에게 개인화된 일자리와 인재를 양방향으로 추천하는 서비스를 제공할 수 있을 것이다. 특히 유래 없는 구인난을 겪고 있는 중소기업과 '체감 실업률'이 사상 최악 수준을 경험하고 있는 청년들의 취업 미스매치를 해소하고 국가의 고질적인 취업난을 해결하는데 기여할 수 있을 것으로 기대한다.

## References

- [1] Bahdanau, D., Cho, K., and Bengio, Y., "Neural machine translation by jointly learning to align and translate", CoRR, abs/1409.0473, 2014.
- [2] Cho, K., van Merriënboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y., "Learning phrase representations using rnn encoder-decoder for statistical machine translation", CoRR,

- abs/1406.1078, 2014.
- [3] Han, J., Kim, J., Jeon, W., Kim, H., and Hong, Y., "A Counseling Matching System Using BERT Language Model", The Korean Institute of Information Scientists and Engineers, pp. 1566-1568, 2020.
  - [4] Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J., "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies", 2001.
  - [5] Hong, S., Na, S., Kim, K. W., Shin, B., and Chung, Y., "BERT for Alzheimer's Disease and Schizophrenia diagnosis", The Korean Institute of Information Scientists and Engineers, pp. 419-421, 2020.
  - [6] Jacob, D., Chang, M., Lee, K., and Kristina, T., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv preprint at arXiv:1810.04805v1, 2018.
  - [7] Kim, J., "Analysis on the Increase in Unemployment Rates Since 2014", KDI (Korea Development Institute) Feature Article (2018.11. 06) Eng, 2018.
  - [8] Kim, Y., Denton, C., Hoang, L., and Rush, A. M., "Structured attention networks", In International Conference on Learning Representations, 2017.
  - [9] Le, Q. V. and Mikolov, T., "Distributed Representations of Sentences and Documents", Proceeding of International Conference on Machine Learning(ICML), 2014.
  - [10] Luong, M., Hieu, P., and Christopher, D. M., "Effective approaches to attentionbased neural machine translation", arXiv preprint arXiv:1508.04025, 2015.
  - [11] Maheshwary, S. and Misra, H., "Matching Resumes to Jobs via Deep Siamese Network", Proceedings of the The Web Conference 2018 (WWW '18), 2018, pp. 87-88.
  - [12] Pessemier, T. D., Vanhecke, K., and Martens, L., "A scalable, high-performance Algorithm for hybrid job recommendations", Proceedings of the Recommender Systems Challenge (RecSys Challenge '16), 2016.
  - [13] SKTBrain, "Korean BERT pre-trained cased (KoBERT)", <https://github.com/SKTBrain/KoBERT>.
  - [14] Son, B., "NCS Utilization in Recruitment and Selection : The case of HRDKorea.", Korean Management Consulting Review, Vol. 15, No. 4, 2015, pp. 217-228.
  - [15] Song, H. S., "A design of content-based metric learning model for HR matching", Journal of Information Technology Applications & Management, Vol. 27, No. 6, 2020, pp. 141-151.
  - [16] Statistics Korea, "December 2020 and Annual Employment Trends", 2020,12.
  - [17] Sutskever, I., Vinyals, O., and Le, Q. V., "Sequence to sequence learning with neural networks", In Advances in Neural Information Processing Systems, 2014, pp. 3104-3112.
  - [18] Valverde-Rebaza, J., Puma, R., Bustios, P., and Silva, N. C., "Job Recommendation based on Job Seeker Skills: An Empirical Study", Proceedings of the First Workshop on Narrative Extraction From Text (Text2Story 2018), 2018, pp. 47-51.
  - [19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., and Polosukhin, I., "Attention is all you need", arXiv preprint arXiv:1706.03762, 2017.
  - [20] Yoo, S. and Jeong, O., "An Intelligent Chatbot Utilizing BERT Model and Knowledge Graph", The Journal of Society for e-Business Studies, Vol. 24, No. 3, 2019, pp. 87-98.

■ 저자소개



오 소 진

한남대학교 경영정보학과 학사, 한남대학교 경영정보학과 석·박사 통합과정을 수료하였다. 현재 딥러닝 기술을 활용한 정보 시스템 개발 회사인 포에스비 주식회사 대표로 재직 중이다. 관심분야는 전이학습 알고리즘, 추천알고리즘, 딥러닝, 머신러닝, IT인력 이다.



장 문 경

경희대학교 경영학부 학사 및 석사, 서울대학교 경영학과 박사학위를 취득하였으며, 롯데정보통신, Tyco International에서 ERP 프로그래머로, 싱가포르 난양공과대학교와 고려대학교에서 연구원으로 근무하였으며, 현재 한남대학교 글로벌IT경영학과 조교수로 재직 중이다. 관심 분야는 디지털 콘텐츠, 온라인구진, 미디어 산업 등으로, 주요 연구 결과를 Computers in Human Behavior, Electronic Commerce Research and Applications, Telecommunications Policy 등의 학술지에 발표했다.

송 희 석



고려대학교 경영학과 학사, 한국과학기술원 경영과학과 석사 및 경영공학과 박사학위를 취득하였으며 대우전자, 대우정보시스템 기술연구소 근무를 거쳐 현재는 한남대학교 글로벌IT경영학과 교수로 재직 중이다. 관심분야는 CRM과 Data

Mining, 소셜네트워크, 유비쿼터스 비즈니스, 비즈니스모델 등이며 주요연구 결과는 Knowledge-based systems, Expert systems with applications, Artificial Intelligence Review 등의 학술지에 발표하였다.