**연 구 용 역 보 고 서**

**페이크 뉴스 자동 검출 응용 서비스의 인공지능 표준 시스템 분석**

**수탁기관: 동의대학교 산학협력단**

한 국 전 자 통 신 연 구 원

**연 구 용 역 보 고 서**

**페이크 뉴스 자동검출 응용 서비스의 인공지능 표준 시스템 분석**

**수탁기관: 동의대학교 산학협력단**

한 국 전 자 통 신 연 구 원

**제 출 문**

**한국전자통신연구원장 귀하**

본 보고서를 “페이크 뉴스 자동검출 응용 서비스의 인공지능 표준 시스템 분석”의 최종 연구보고서로 제출합니다.

2020년 11월 07일

수 탁 기 관 : 동의대학교 산학협력단

수 탁 기 관 장 : 이 임건 (인)

연 구 책 임 자 : 김 근형

참 여 연 구 원 : 한 윤진

박 예진

이 승찬

**요 약 문**

1. **제 목**

페이크 뉴스 자동검출 응용 서비스의 인공지능 표준 시스템 분석

1. **연구의 목적 및 중요성**

**1) 연구의 목적**

**2) 연구의 필요성**

1. **연구의 내용 및 범위**
2. **연구 결과**
3. **활용에 대한 건의**



**목차**

**그림 목차**

**표 목차**

# 제 1 장 개요

## 제 1 절 인공지능 서비스 배경 및 현황

1. **서비스 활용 사례**
2. **시장/제품 분석**

## 제 2 절 신뢰성 기술

1. **정의**
2. **중요성, 필요성, 파급효과**
3. **국내외 현황**

# 제 2 장 인공지능 신뢰성 분석

## 제 1 절 인공지능 서비스 현황

(문제점, 해결 방안)

## 제 2 절 표준화 현황

1. **국외 표준화 현황**
2. **국내 표준화 현황**

# 제 3 장 인공지능 신뢰성 프레임워크

1. **서비스 현황 사례**

(참조구조 및 기능, 공통 인터페이스)

1. **인공지능 신뢰성 요구 사항**

(참조구조 및 기능, 공통 인터페이스)

# 제 4 장 설명 가능한 인공지능 기반 서비스

## 제 1 절 설명 가능한 리뷰의 감성 탐지 서비스

1. **임베딩**

일반적으로 컴퓨터는 자연어처리에서 단어의 의미와 상관없이 각각을 개별적인 기호로 취급한다. 컴퓨터가 자연어를 처리하도록 하기 위해선 자연어를 계산 가능한 형식으로 변형시켜야 한다. 문장에서 총 단어의 수만큼을 벡터의 차원으로 두고, 단어마다 해당하는 차원에 1로 표시하여 단어 벡터로 생성하는 것을 one-hot encoding이라고 한다. 각각의 단어 벡터들을 모두 더한 것을 해당 문장의 벡터라고 할 수 있는데, 단어의 의미를 담지 않은 단어 벡터들의 합이므로 문장의 의미 또한 표현하는데 어려움이 있다.

[http://doc.mindscale.kr/km/unstructured/11.html, 한국어 임베딩(책)]

이를 해결하기 위해 단어의 의미를 좀 더 세밀한 차원에 벡터로 표현하려 한 것이 단어 임베딩(word embedding)이다. 단어 임베딩은 one-hot encoding과 다르게 각각의 단어 벡터에 의미를 포함시키므로 비슷한 의미의 단어가 있는 경우 벡터 상에서 가까운 곳에 위치하게 된다. 그러므로 단어끼리 의미 비교를 위해 임베딩된 단어 벡터간에 유사도 계산이 가능하다. 유사도가 높으면 의미가 비슷하다는 것을 알 수 있다. 단어 임베딩의 종류에는 LSA, Word2Vec, GloVe, FastText 등이 있는데 이와 같은 단어 임베딩은 단어마다 벡터 값이 고정되어 있기 때문에 단어가 사용된 문맥은 고려하지 않는다. 즉 같은 단어이지만 다른 맥락에서 사용되었을 때도 같은 벡터 값으로 출력되기 때문에 차이를 알 수 없다.

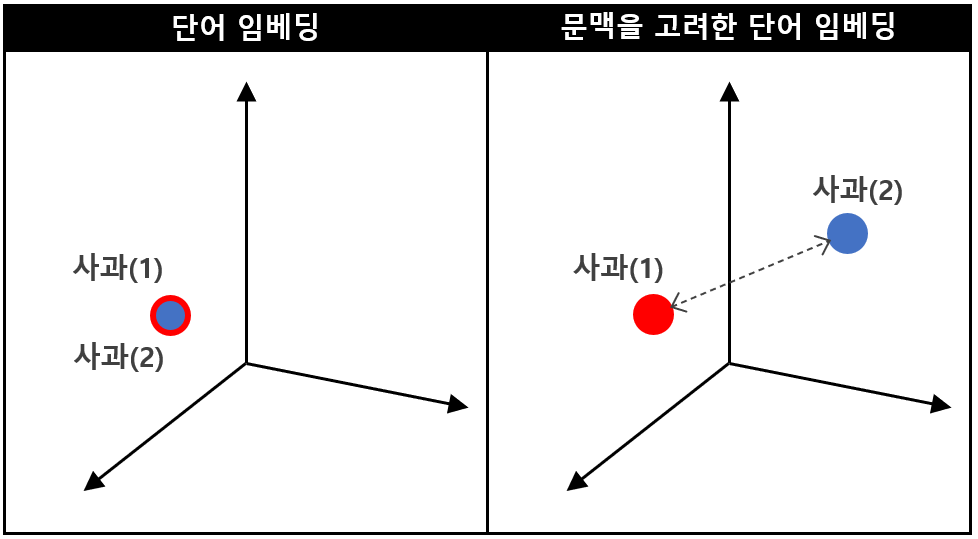


그림 1. 임베딩 방법 비교

단어가 사용된 문장의 맥락을 파악하기 위해 등장한 것이 문맥을 고려한 단어 임베딩(contextualized word embedding)이다. 문맥을 고려한 임베딩에선 개별 단어뿐만 아니라 문장 전체의 문맥적 의미를 함축하기 때문에 임베딩된 벡터를 인공지능을 통해 학습시켰을 때 기존의 단어 임베딩보다 학습 효과가 좋은 것으로 알려져있다. 그림 1은 단어 임베딩과 문맥을 반영한 단어 임베딩의 차이를 그림으로 나타낸 것이다. 단어 임베딩은 같은 문장에서 사용된 동일한 단어가 사용된 의미에 상관없이 같은 벡터를 출력하지만, 문맥을 고려한 임베딩의 경우 파악한 문맥에 맞게 각각의 벡터를 출력하게 된다. 문맥을 고려한 단어 임베딩 종류로는 BERT(Bidirectional Encoder Representtions from Transformer), ELMo(Embeddings from Language Models), GPT(Generative Pre-Training)등이 있다.

본 연구에서는 BERT가 사용되었다. BERT의 출력으로 얻을 수 있는 단어 벡터들은 문맥을 파악하여 생성된 벡터이기 때문에 단어의 의미뿐만 아니라 문장의 의미 또한 유지시킨다. 단어 벡터는 12개의 encoder layer를 거쳐 출력되는데, 마지막 레이어의 출력보다 마지막 4개 레이어 출력의 연결이 벡터의 표현을 더 잘 나타낼 수 있다. BERT는 출력 뒤에 원하는 분류 모델을 추가시켜 분류 학습을 하게 된다. 본 연구에선 설명가능한 인공지능을 위해 문장 벡터에 어떤 단어 벡터가 영향을 미쳤는지 해당 라벨로 분류되었는지 분석하였다.

1. **BERT(Bidirectional Encoder Representtions from Transformer)**

BERT는 인공지능 기반 자연어처리 분야에 등장한 구글의 강력한 언어모델이다. 기본적으로 wiki나 book data와 같은 대용량 unlabeled data로 모델을 사전 학습(pre-training)하여 문맥을 파악할 수 있도록 구축된 모델이다. 사전학습된 모델 BERT에 분류하고자 하는 task에 맞게 labeled data로 추가 학습하여 분류 성능을 높인다. BERT는 사전학습에는 Masked Language Model(MLM)과 Next Sentence Prediction(NSP) 두 가지 방법이 사용된다. MLM은 랜덤한 위치의 토큰을 마스킹 시킨 뒤, 나머지 토큰들을 기반으로 마스킹된 토큰의 단어를 예측하고, NSP는 두 개의 문장을 입력하여 두 문장이 연속된 문장인지 예측하며 학습 성능을 높인다.

BERT Encoder는 MLM과 NSP를 위해 Encoder-Decoder 아키텍처로 이루어진 Transformer의 Transformer Encoder 구조를 기반으로 한다. Transformer는 RNN(Recurrent Neural Network)과 마찬가지로 시퀀스 데이터를 처리할 수 있도록 설계된 모델이다. 하지만 RNN과 다르게 Transformer는 시퀀스 데이터를 순차적으로 처리하지 않는다. 순차적으로 처리되는 RNN은 문장이 길어지는 경우 기울기 소실 문제(Vanishing gradient problem)로 인해 초기 토큰에 대한 정보가 사라질 수 있다는 단점이 있는 반면, Transformer는 병렬처리하여 초기 토큰의 정보를 보존할 수 있다. 또한 데이터 병렬처리 방법은 학습 시간을 단축시킨다.

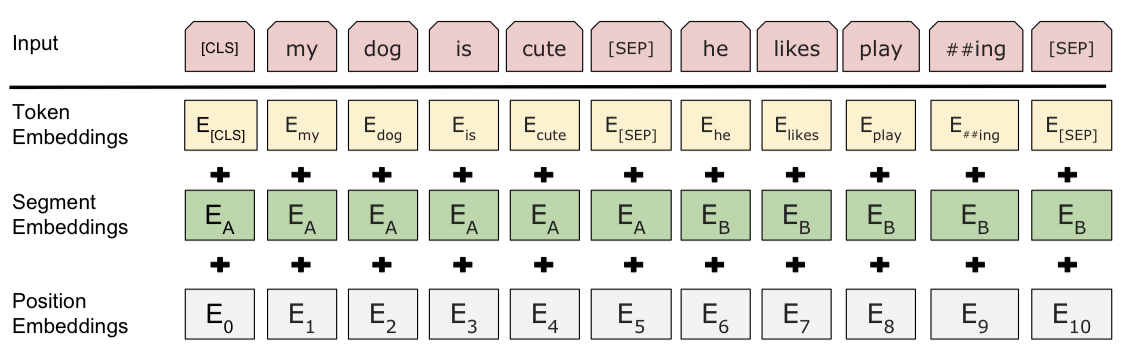


그림 2. BERT의 입력데이터 임베딩

BERT의 입력으로는 단어 토큰화된 토큰들이 사용된다. 그림 1의 Input에서 [CLS] 토큰은 BERT Encoder의 전체 layers를 거치고 난 후 token sequence의 결합된 의미를 가지게 되어 분류에 사용되는 토큰이며, [SEP] 토큰은 문장이 두 개 들어갈 때 문장을 구분하는 것에 사용되는 토큰이다. BERT Encoder에 입력하기 위해선 Embedding 과정을 거쳐야 하는데, Token embeddings과 position embeddings이 사용되는 Transformer Encoder과는 다르게, BERT Encoder에서는 토큰이 어느 문장에 속하는지 알 수 있도록 segment embeddings까지 더하여 입력 데이터를 생성한다.

BERT Encoder는 Multi-Head Attention이 사용되는데 이는 임베딩된 입력 데이터의 차원을 Head 수만큼 나누어 각각 Attention을 계산하고 추후에 합치는 과정이다. Attention 은 모든 토큰 간의 상관관계에 따라 가중치를 계산하는 메커니즘이다. 주어진 토큰 벡터 Query에 대해서 다른 토큰 Key와의 관계를 계산하여 가중치를 구하게 되는데 이를 Value에 적용시켜 출력값을 생성한다. 모든 토큰에 대한 attention 계산은 다음과 같이 최적화된 행렬 계산으로 표현될 수 있다.

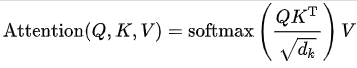


그림 3. Attention 계산 수식

설정한 layers 수만큼 BERT Encoder가 반복된 후 [CLS] 토큰은 token sequence의 결합된 의미를 가지게 되는데, 여기에 분류 모델을 추가되면 문장 분류 학습이 가능하게 된다. BERT는 다양한 task에 사용될 수 있으며, 입력 데이터와 분류 모델을 어떻게 Fine-tuning 하느냐에 따라 다른 유형으로 변형 가능하다.

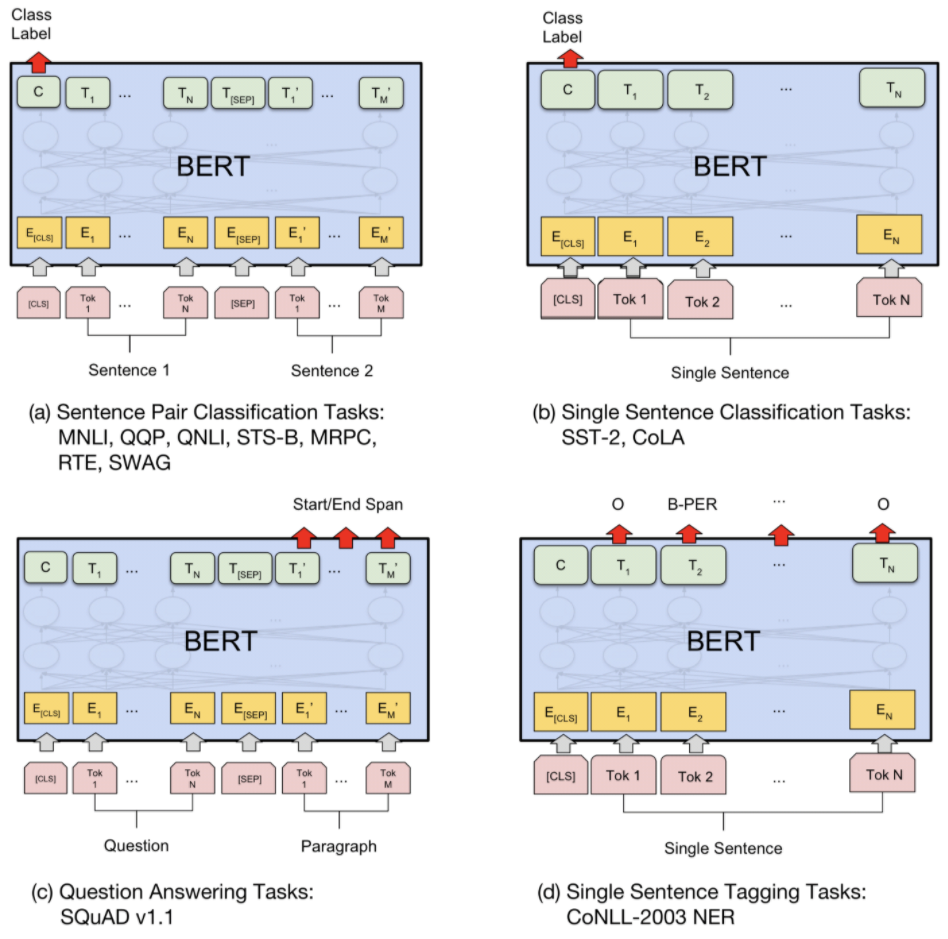


그림 4. BERT의 사용 유형

그림 3의 (a)와 (b)와 같은 경우는 분류하고 싶은 개수에 따라 분류 모델을 조정하고, 문장의 의미를 가지는 [CLS] 토큰의 벡터를 사용하여 class를 분류하는 유형이다. (c)와 같은 Question Answering(QA) task 경우엔 Question와 Paragraph를 [SEP] 토큰으로 분류하고, Question에 정답을 Paragraph에서 찾아내는 유형이다. (d)와 같은 경우는 Named Entity Recognition(NER)이나 형태소 분석처럼 각 토큰이 어떤 class를 가지는지 각각을 분류하는 유형이다.

BERT는 위의 내용처럼 downstream task 학습을 통해 사전학습된 parameter들을 변경하며 진행되는 Fine-tuning 접근법뿐만 아니라 Feature-based 접근법으로도 사용 가능하다. Feature-based 접근법은 BERT를 통해 진행된 Feature extraction을 특정 분류를 위한 추가 모델에 입력하여 학습시키는 형태이다. Feature-based 접근법은 다른 모델인 ELMo(Embeddings from Language Model)에서도 사용되는 접근법으로, 사용하는 이유는 Fine-tuning 접근법을 사용했을 때 사용되는 pooler는 분류 모델에 들어가기 전에 [CLS] 토큰을 처리하는 과정인데, pooler의 출력 값은 분류에는 도움을 주지만 문장의 의미를 표현하기엔 좋은 값이 아니기 때문이다.

* (This output is usually not a good summary of the semantic content of the input, you’re often better with averaging or pooling the sequence of hidden-states for the whole input sequence.) [출처]

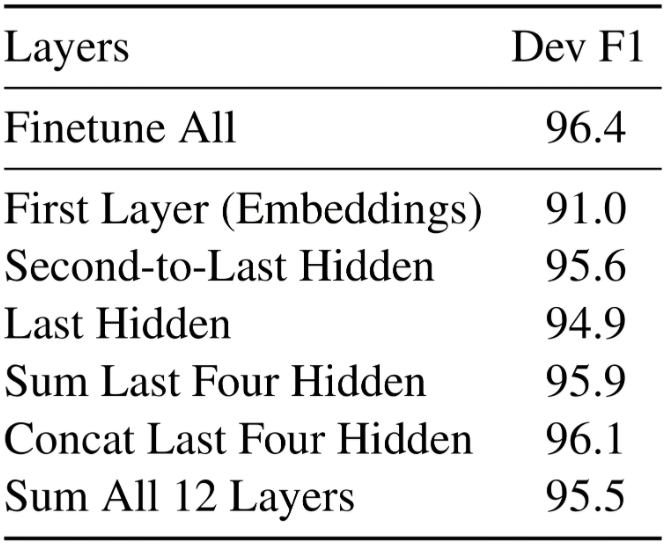


표 1. BERT의 Feature-based 접근법

그림 4는 BERT 논문에서 Feature-based 접근법들을 사용하여 실험한 결과를 정리한 표다. 그림 4에서 볼 수 있듯이, Concat Last Four Hidden 값을 사용하면 Fine-tuning 접근법의 성능과의 차이가 크게 나지 않는다. 이를 보아, BERT는 Feature-based에서도 효과적이라고 볼 수 있다.

1. **BERT 세부 기능 분석**

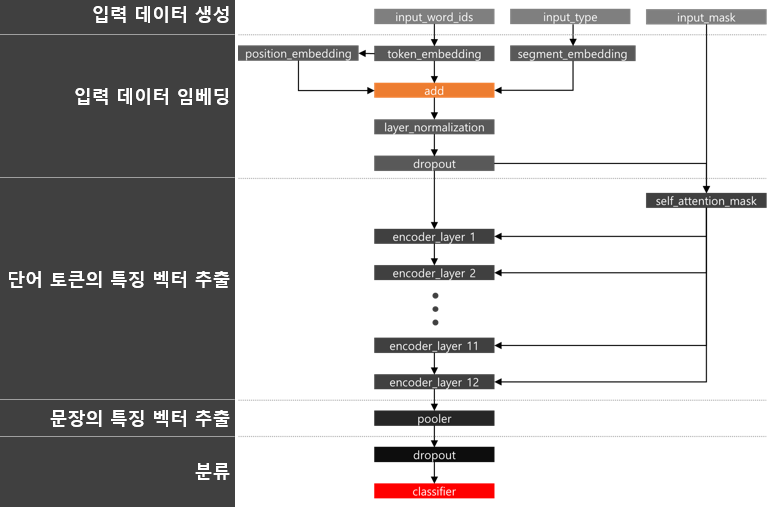


그림 4. BERT의 블록 다이어그램

BERT의 입력으로는 세 가지 데이터가 준비되어야 한다. 첫 번째로 입력 문장을 토큰화 한 뒤, 토큰을 index 값으로 변환시킨 input\_word\_ids 데이터가 필요하다. 두 번째로는 여러 문장이 입력으로 들어올 때 토큰이 어느 문장에 소속되어 있는지 알 수 있도록 input\_type 데이터가 있어야 하며, 마지막으로 토큰과 토큰이 아닌 패딩 위치를 알 수 있도록 해주는 input\_mask 데이터가 필요하다.

입력 받은 데이터들은 임베딩 과정을 거치게 되는데 input\_word\_ids 데이터로 token\_embedding 벡터와 토큰의 순서를 알 수 있는 position\_embedding 벡터를 얻게 된다. input\_type 데이터로 소속 문장을 알 수 있도록 해주는 segment\_embedding 벡터를 얻게 되며, input\_mask는 다음 단계인 encoder 단계에서 토큰과 패딩을 구분할 수 있도록 활용된다. 임베딩 과정을 통해 얻은 세 가지 벡터를 합하여 임베딩된 입력 데이터를 얻을 수 있다.

임베딩된 입력 데이터는 토큰의 특징 벡터를 추출할 수 있는 encoder로 입력된다. 이 과정에서 데이터는 Multi-Head Attention을 통해 각 단어 사이의 관계성에 맞게 가중치가 계산된다. 설정된 여러 layer를 거치게 되면서 입력된 문장의 문맥을 파악하게 되고 그에 맞는 특징 벡터를 출력한다.

encoder 단계에서 출력된 토큰의 특징 벡터들 중에서 첫 번째 토큰인 [CLS] 토큰은 다음 단계인 pooler의 입력으로 사용된다. pooler에서는 분류 모델이 분류를 더 효율적으로 진행할 수 있도록 문장 벡터인 [CLS] 토큰을 처리하는 과정이다. 분류 모델은 pooler의 결과인 문장 벡터를 입력으로 받아 결과에 맞는 label을 학습하게 된다.

1. **실험**

|  |  |
| --- | --- |
| 모델 | BertForSequenceClassification |
| 토크나이저 | BertTokenizer |
| 데이터셋 | Naver sentiment movie corpus |
| 데이터 수 | 학습 데이터: 150,000 |
| 테스트 데이터: 50,000 |
| 라벨 | 긍정: 100,000 |
| 부정: 100,000 |
| 테스트 정확도 | 87% |

실험을 위해 네이버 감정분석 영화 리뷰 데이터[https://github.com/e9t/nsmc]를 사용하였다. BERT 모델에 긍정 및 부정으로 이루어진 영화 리뷰 데이터를 학습시켰으며, 각 라벨로 분류될 때 어떤 단어의 영향이 컸는지 테스트했다. 임베딩된 벡터는 사칙 연산을 통해 단어와 문장 사이의 의미적 관계를 도출할 수 있다. 이를 기반으로 우리는 문장 벡터와 단어 벡터들 사이의 유사도를 측정하였을 때, 유사도가 높은 벡터가 문장에 영향을 크게 미쳤다고 판단했다. 표 2는 해당 테스트를 위한 실험 환경이다.

## 제 2 절 설명 가능한 가짜 뉴스 탐지 서비스

# 제 4장 결론

**참고문헌**

[1]