# 의미 정보와 BERT를 결합한 개념 언어 모델[1]

이주상○, 옥철영

울산대학교, 한국어처리연구실
dosa510@naver.com, okcy@ulsan.ac.kr

# A Concept Language Model combining Word Sense Information and BERT

Ju-Sang Lee○, Cheol-Young Ock
University of Ulsan, Korean Language Processing Lab

**요 약**

자연어 표상은 자연어가 가진 정보를 컴퓨터에게 전달하기 위해 표현하는 방법이다. 현재 자연어 표상은 학습을 통해 고정된 벡터로 표현하는 것이 아닌 문맥적 정보에 의해 벡터가 변화한다. 그 중 BERT의 경우 Transformer 모델의 encoder를 사용하여 자연어를 표상하는 기술이다. 하지만 BERT의 경우 학습시간이 많이 걸리며, 대용량의 데이터를 필요로 한다. 본 논문에서는 빠른 자연어 표상 학습을 위해 의미 정보와 BERT를 결합한 개념 언어 모델을 제안한다. 의미 정보로 단어의 품사 정보와, 명사의 의미 계층 정보를 추상적으로 표현했다. 실험을 위해 ETRI에서 공개한 한국어 BERT 모델을 비교 대상으로 하며, 개체명 인식을 학습하여 비교했다. 두 모델의 개체명 인식 결과가 비슷하게 나타났다. 의미 정보가 자연어 표상을 하는데 중요한 정보가 될 수 있음을 확인했다.

주제어: Language model, BERT, word sense information, word Embedding

## 1. 서론

자연어 표상(Natural Language Representation)은 자연어를 컴퓨터가 이해할 수 있는 형태로 표현하는 방법이다. 인간은 텍스트로 표현된 자연어에 대해 사전에 습득한 지식을 통해 정보를 획득할 수 있다. 그러나 컴퓨터의 경우 단순 텍스트로 표현된 자연어에서 단순히 문자라는 정보만 획득이 가능하다. 인간과 비슷한 수준의 정보를 컴퓨터에게 제공하기 위해 자연어 표상을 사용한다. 최근 자연어 처리 분야에서 심층 학습(Deep Learning)이 주목받기 시작하면서 자연어 표상 방법의 중요도가 증가했다.

초기 자연어 표상은 학습을 통해 자연어를 고차원의 벡터로 표상했다. 자연어 표상을 통해 여러 자연어 처리 분야에서 좋은 성능을 보여주게 되었다. 하지만 벡터로 표상된 정보는 벡터 차원수에 의해 한정적이다. 자연어의 경우 복합적인 관계를 가지며 문법적 요소와 의미적 요소로 인해 단순히 고차원의 벡터로 표현하기에는 부족하다. 그리고 동형이의어에 대해 고려하지 않아 다른 의미로 사용된 동일한 형태의 자연어에 대해 분석 오류를 범할 수 있다. 예를 들어 '배'라는 단어가 "타는 배"로 학습하여 표상한 경우 "먹는 배"에 해당 정보를 사용하면 잘못된 분석이 되기 때문이다. 최근에는 현재 문맥을 분석하여 자연어의 쓰임에 따라 다르게 해석하여 벡터로 표상하는 모델이 등장했다. 대표적인 모델은 ELMo[1]와 BERT[2]가 있다. 두 모델은 문장에 따라 같은 형태의 자연어라 해도 다른 결과값을 도출한다. 예를 들어 '사과'라는 단어는 대표적으로 "과일의 한 종류"라는 의미와 "남에게 용서를 빎"이라는 두 가지로 사용이 가능하다. ELMo와 BERT의 경우 '사과'가 등장한 문장을 분석하여 사용된 의미에 맞는 '사과'의 결과값을 보여준다. 하지만 ELMo와 BERT는 많은 양의 학습 데이터를 사용하고 네트워크에서 많은 연산을 필요로 해 학습에 많은 시간이 필요하다. 학습 말뭉치를 적게 사용하면 편향적으로 학습될 가능성이 있다.

본 논문은 학습 시간을 개선하기 위해 기존의 BERT 모델과 문장 분석을 통한 단어가 가진 의미 정보를 융합한 모델을 제안한다. 문장에 대해 의미 분별을 실시한 후 한국어 어휘지도(UWordMap)[3]을 사용하여 의미 정보를 생성한다. 의미 정보에는 각 단어의 품사 정보와 명사의 계층 정보를 사용해 단어가 가지는 추상적 의미를 학습하여 적은 데이터에서도 효과적인 학습을 할 수 있도록 한다.

## 2. 관련 연구

초기 자연어 표상은 문장에서 주변 단어들을 활용해 학습한다. 대표적인 모델은 Word2Vec[4], GloVe[5] 가 있다. 두 모델은 단어를 단위로 하며, 학습을 통해 각 단어마다 고유의 벡터 값으로 표상한다. Word2Vec나 GloVe의 경우 단어를 기준으로 표상하기 때문에 신조어나 학습에 등장하지 않은 단어에 대해 표상이 불가능하다. 그리고 동형이의어에 대한 분별을 실시하지 않아 여러 의미를 가지는 하나의 표제어에 대해 정확한 학습이 어렵고, 원래 문장에서의 의도한 단어의 쓰임과 다른 의

[1] 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신/방송연구개발사업[2013-0-00179, 컨텍스트 인지형 Deep-Symbolic 하이브리드 지능 원천 기술 개발 및 언어 지식 자원 구축]을 받아 수행하였음

미로 분석하게 된다.

페이스북에서 개발한 fastText[6] 의 경우 신조어나, 미 학습 단어에 대한 처리를 가능하도록 설계된 모델이다. FastText는 각 단어를 n-gram으로 분리하여 부분 단어에 대해 학습한다. 학습된 부분 단어를 결합하여 원래 단어의 표상으로 사용하는 모델이다. 그러나 fastText도 부분 단어를 조합하여 사용하기 때문에 하나의 단어에 대해 하나의 벡터만을 도출하며, 동형이의어 문제를 해결하지 못한다.

최근 연구된 ELMo(Embeddings from Language model)와 BERT(Bidirectional Encoder Representations from Transformers)의 경우 여러 의미 정보를 내재한 표상 방법이다. 기존의 모델들은 벡터로 표상된 학습 결과만 사용했다. 그러나 ELMo와 BERT의 경우 학습된 모델 전체가 자연어 표상의 결과이며 최종 출력이 각 단어를 표현한 벡터이다. 두 모델은 문맥적 정보에 따라 단어의 표상이 달라지는 모델이다.

ELMo는 Bidirectional LSTM을 이용한 자연어 표상 모델이다. 문장의 양방향에서 생성된 정보를 조합하여 현재 문장에서 사용된 의미를 담고 있는 벡터로 표상하는 모델이다. 단어가 사용된 문장에 따라 같은 형태를 가진 단어라도 다른 결과를 얻게 된다. 그러나 ELMo는 Bidirectional LSTM으로 인해 이전 정보에 대한 의존성이 높아 학습 속도가 매우 느린 문제가 있으며 단어 간의 거리가 멀어지면 이전 정보에 대한 전파가 작아지는 문제를 가지고 있다.

BERT는 Transformer[7] 라는 encoder-decoder 모델의 encoder를 사용하여 자연어를 표상하는 모델이다. Transformer 모델은 self-attention을 사용하여 이전 단어의 정보에 대해 종속성을 가지지 않고 빠른 학습이 가능한 모델이다. BERT는 부분 단어의 형태로 단어를 분리한다. 단어의 분리를 위해 단어 조각 모델(Word Piece Model)을 통해 생성한 부분 단어를 사용한다. 단어 조각 모델은 글자 조합의 빈도수를 이용하여 부분 단어를 생성한다. 부분 단어의 조합과 문맥적정보를 활용하여 신조어나, 학습하지 않은 단어에 대한 접근도 가능한 모델이다.

## 3. 의미 정보와 BERT를 결합한 개념 언어 모델

### 3.1 문장 분석 정보와 의미 추상화 정보

본 논문에서는 기존 BERT 모델에 사용한 문장의 문맥 정보 이외에도 의미 정보를 추가로 학습한다. 본 논문에서 사용한 의미 정보는 형태소가 가진 품사 정보와 명사의 계층 구조를 추상화 한다.

학습에 사용하는 품사 정보는 원시 문장을 형태소 분석기인 유태거(UTagger)를 통해 분석한다. 문장에서 생성한 품사 정보는 모델의 추가 입력 정보로 활용된다. 품사 정보는 해당 형태소가 가지는 기본적인 정보로 품사 간의 관계성과 문법적 특성을 자연어 표상 모델에 적용이 가능하다. 기존 BERT 모델의 경우 부분 단어로 나타낼 수 없는 자연어에 대해 주변 문맥 정보만 사용한다. 품사 데이터를 통해 부분 단어로 표현이 불가능한 단어에 대해서 추가 정보를 주변에 전달한다.

다른 의미 정보로 명사의 계층 구조를 추상화 한다. 명사들의 계층 정보는 울산대학교의 한국어 어휘지도(UWordMap)를 사용한다. 명사의 계층 구조는 상위 계층으로 갈수록 각 단어에 대한 추상적인 의미를 가진 단어로 표현된다. 명사 계층 정보를 활용한 의미 추상화는 학습 데이터를 확장하여 학습하는 효과를 가지게 된다. 명사 간의 관계성 및 유사한 개념을 가진 단어들의 군집할 수 있도록 도움을 준다. 예를 들어 '연필' 이라는 단어가 쓰인 문장은 상위어인 '필기도구' 로 추상화를 하면 '필기도구' 의 하위 개념들과 '연필' 이 관계성을 가질 수 있으며, '연필' 의 주변 개념들을 모델이 학습하는 효과를 가져온다.

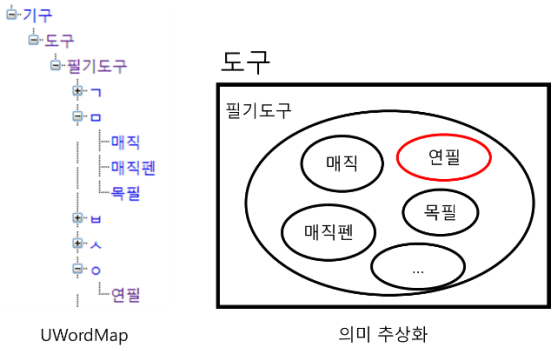


그림 1 '연필' 의 UWordMap의 상위어 구조 및 의미 추상화 표현

그림 1는 '연필' 이 실제 한국어 어휘지도에서 표현된 계층의 모습과 추상화를 표현한 그림이다. 그림에서 의미 추상화를 통해 '필기도구' 로 추상화된 '연필' 의 학습에 사용된 정보가 '필기도구' 의 하위 개념들인 '매직', '매직펜', '목필' 에 영향을 준다. 결국 주변 개념을 학습하지 않아도 '연필' 은 새로운 관계 정보를 추상화를 통해 획득하게 된다. 이렇게 의미 추상화를 통해 한정된 정보에 대해 확장이 가능하다.

### 3.2 형태소 단위의 단어 조각 모델과 부분 단어 구축 방법

본 논문에서 사용하는 단어 조각 모델(Word Piece Model)[8] 은 어절 단위가 아니라 형태소 단위의 부분 단어를 만들기 위해 사용한다. 예를 들어 '학교에' 라는 어절을 먼저 '학교' 와 '에' 로 분리하여 각각을 단어 조각 모델에 의해 분리하게 된다. 한국어는 교착어로 조사와 접사의 의미를 인식하기 위해 어절 단위보다 축소된 형태소 단위의 단어 조각 모델을 사용한다. 그리고 동일한 형태의 표제어가 여러 품사를 가지는 경우 구분을 위해 특수문자 "##" 을 사용한다. 예를 들어 '가' 라는 표제어는 명사, 접사, 조사로 사용이 가능하다. 명사로 사용된 '가' 와 접사나 조사로 사용된 '가' 는 쓰임새나, 의도가 다르다. 본 논문에서는 형태소의 품사

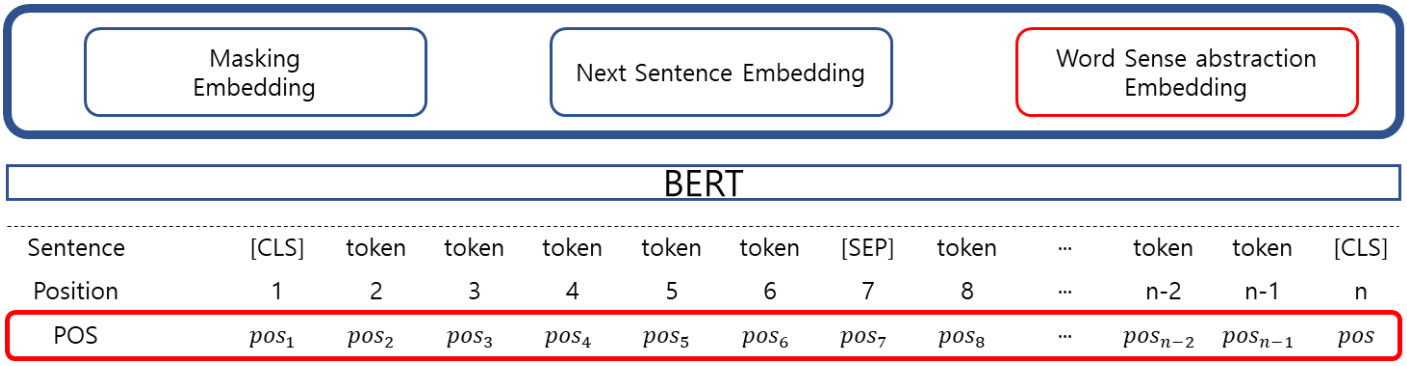


그림 2 의미 정보와 BERT를 결합한 개념 언어 모델

에 따라 특수문자인 "##" 을 추가한다. 다음의 표는 품사에 따른 특수문자를 추가하는 방식을 보여준다

.

표 1 품사에 따른 부분 단어 표현 규칙

| 품사 | 형태소 | 부분 단어 표현 |
|---|---|---|
| 명사 | 사과/NNG | 사과 |
| 용언 | 갈/VV | 갈## |
| 접두사 | 수/XPN | 수## |
| 접미사 | 가/XSN | ##가 |
| 어미 | 다/EP | ##다 |

표 1은 품사에 따른 부분 단어 표현 규칙을 보여준다. 용언과 접두사의 경우 해당 표제어의 뒤에 특수문자를 추가하며, 접미사와 어미의 경우 해당 표제어의 앞에 특수문자를 추가하여 표기한다.

본 논문에서 사용한 부분 단어 집합은 사전과 말뭉치를 이용해 구축했다. 먼저 사전에 존재하는 복합명사가 아닌 단어에 대해 형태소 단위로 모두 추출한다. 사전에서 추출한 단어가 말뭉치에 등장하면 해당 단어를 부분 단어 집합에 포함시킨다. 마지막으로 포함되지 않은 명사 단어에 대해 단어 조각 모델을 사용하여 최종적으로 부분 단어를 생성하게 된다.

### 3.3 의미 정보를 추가한 개념 언어 모델

본 논문에서 사용한 개념 언어 모델은 BERT와 동일하게 Transformer의 encoder를 사용하며, 의미 정보 데이터 생성을 위해 원시 데이터를 분석하는 과정을 먼저 실시한다. 다음의 그림은 개념 언어 모델에서 의미 정보 학습 데이터 생성 과정이다.

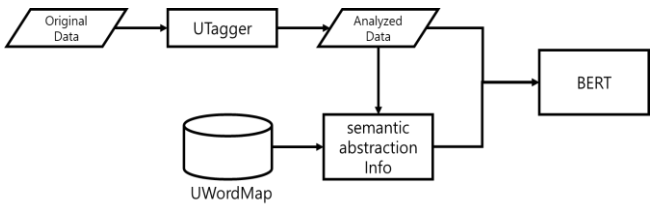


그림 3 개념 언어 모델 학습 데이터 생성 과정

그림 3에서 먼저 원시 말뭉치 분석을 실시한다. 분석된 말뭉치와 한국어 어휘지도에 있는 명사의 계층 정보를 활용해 최종 학습 데이터를 생성하게 된다. 생성한 정보로 개념 언어 모델을 학습한다. 다음은 실제 문장이 어떻게 분석되어 학습 데이터로 변화하는지 보여준다.

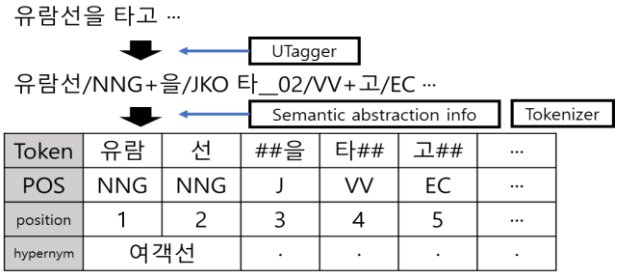


그림 4 예시 문장이 학습데이터를 생성하는 과정

그림 4은 "유람선을 타고 …" 라는 예시 문장에서 학습 데이터가 생성되는 과정을 보여준다. 원시 문장을 유태거를 사용해 분석하며, 토큰화를 이용해 문장을 부분 단어로 분리하며 각 단어의 품사 정보와 위치 정보를 부착한다. 그리고 명사 의미 계층 정보를 사용해 의미 추상화 정보를 생성한다.

그림 2는 본 논문에서 사용한 의미 정보를 개념 언어 모델에 적용하는 방법을 보여준다. 모델의 입력으로 토큰화를 통해 문장을 부분 단어로 구성한다. 구성된 부분 단어에 대해 품사 정보와 위치 정보, 문서의 다음 문장 여부를 사용한다. 만약 하나의 형태소 단어가 여러 개의 토큰으로 분리되면 원래 형태소가 가진 품사를 분리된 토큰에 동일하게 적용했다.

의미 추상화를 위한 명사 계층 정보 학습 방법은 기존 BERT의 Masked LM 학습 방법과 유사하다. 기존 BERT 모델은 일부 부분 단어에 대해 마스크를 씌워 해당 단어를 맞추도록 학습한다. 이와 유사하게 의미 추상화를 위해 입력으로 사용된 단어가 상위어를 가지게 되면 해당 상위어를 맞추도록 유도한다. 동형이의어의 경우 여러 상위어를 가질 수 있다. 해당 단어가 가진 상위어 토큰의 수가 10개 이하인 경우에만 상위어를 이용한 의미 추상화를 실시한다. 기존 단어와 상위어의 토큰 개수가 다른 경우가 많기 때문에 각 출력의 평균 값을 사용한다. 상위어의 경우 원래 단어와 동일한 형태가 아니므로 오류율의 90%만 적용하여 학습한다. 아래의 그림은 '유람선' 이라는 단어에 대해 의미 추상화를 하는 과정이다.

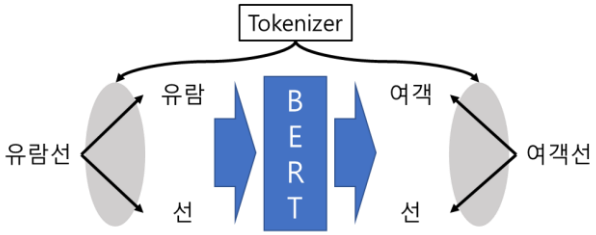


그림 5 '유람선'이 추상화되는 과정

그림 5은 '유람선'이 추상화되는 과정을 보여주고 있다. '유람선'의 상위어인 '여객선'을 '유람선'의 의미 추상화 학습에 목표 단어로 설정한다. 먼저 '유람선'의 토큰들을 BERT 모델을 통해 결과를 추출한다. 추출한 결과는 Feedforward NN을 통과하여 각 토큰에 대해 Softmax를 사용하여 결과를 생성한다. 생성한 결과에서 '여객선'의 토큰인 '여객'과 '선'의 위치에 해당하는 결과의 평균값을 구한다. 구한 평균값을 통해 '여객'과 '선' 토큰에 대한 오류를 구한다. 상위어 토큰들의 오류 값의 평균을 BERT 모델에 적용하여 학습한다.

## 4. 실험

본 논문의 모델의 실험을 위해 세종 형태 분석 말뭉치와 신문기사 등에서 수집한 500MB의 말뭉치를 사용했으며, 부분 단어의 수는 29,256개이다. 학습의 반복 횟수는 2000회로 설정했다. 학습에 사용한 컴퓨터의 CPU는 i7-5870k이며 메모리는 32GB이다. 총 학습에 10시간이 소요됐다.

본 논문에서는 모델 성능 실험을 위해 한국어 개체명 인식을 사용했다. 2016년에 국어 정보 경진 대회용으로 배포한 개체명 말뭉치를 사용하며, 5개의 개체명에 대한 분류 성능으로 실험했다. 학습에 3,556 문장을 사용했으며 실험에 502 문장을 사용했다.

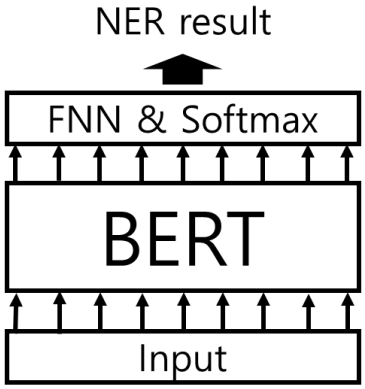


그림 6 실험을 위해 사용한 개체명 학습 모델

그림 6은 실험에 사용한 개체명 모델을 그림으로 표현한 것이다. BERT 모델의 비교를 위해 1개의 은닉층을 사용하는 Fully connect Neural Network와 Softmax를 사용했다. 개체명 정답의 비교는 형태소 단위로 실시했다. 부분 단어로 분리된 단어의 경우 가장 마지막 조각이 가지는 개체명 태그로 해당 단어의 개체명을 결정했다. 모델의 성능 비교를 위해 ETRI에서 배포한 형태소 기반의 BERT를 사용한다. ETRI에서 배포한 형태소 기반의 한국어 BERT는 23GB의 대용량 말뭉치를 사용하여 학습한 모델이다. 다음의 표는 각 개체명에 대해 성능을 비교한 표이다.

표 2 모델별 개체명 실험 결과

| 개체명 | 테스트 개수 | 개념 언어 모델 | ETRI BERT 모델 |
|---|---|---|---|
| 인명 | 725 | 458(63.1%) | 478(65.9%) |
| 지명 | 253 | 199(78.6%) | 188(74.3%) |
| 조직명 | 667 | 406(60.8%) | 421(63.1%) |
| 날짜 | 800 | 744(93%) | 681(85.1%) |
| 시간 | 111 | 100(90%) | 80(72%) |
| 일반 단어 | 15,839 | 15,066(95.1%) | 15,090(95.2%) |

표 2는 본 논문에서 제안한 모델과 ETRI에서 배포한 형태소 기반의 한국어 BERT 학습 결과를 비교한 결과이다. 전체 개체명에 대해서 유사한 성능을 나타내고 있다. 그러나 학습에 사용된 데이터의 양과 학습 시간을 비교하면 큰 차이를 보이게 된다. 그렇기 때문에 의미 정보가 자연어 표상에 도움을 줄 수 있다고 볼 수 있다.

## 5. 결론

본 논문에서는 기존의 BERT 학습이 대용량 말뭉치와 대형 네트워크 구조를 사용하여 학습이 느린 문제점을 해결하기 위해 의미 정보와 BERT를 결합한 개념 언어 모델을 제안했다. 의미 정보로 품사 정보와 명사의 계층 정보를 사용하여 학습했다. 품사 정보를 통해 문법적 특성을 자연어 표상에 적용하며, 명사의 계층 정보를 활용하여 의미 추상화를 실시했다. 개체명 데이터를 학습하는 실험에서 ETRI에서 공개한 형태소 기반의 한국어 BERT 모델과 유사한 성능을 보였다. 비교를 위해 사용한 개체명 데이터가 작고 다양한 분야에서의 모델의 비교를 하지 못했다.

향후에는 의미 정보를 사용한 자연어 표상 방법과 기존의 말뭉치 기반을 결합한 새로운 자연어 표상 방법에 대한 연구를 진행할 계획이다.

## 참고문헌

[1] Peters, Matthew E., et al. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365 ,2018.

[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 2018.

[3] 배영준, 옥철영, "한국어 어휘지도(UWordMap)와 API 소개", 제 26회 한글 및 한국어 정보처리 학술 대회

논문집, 27-31, 2014.
[4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, “Distributed Representations of Words and Phrases and their Compositionality”, Advances in neural information processing systems, 2013
[5] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
[6] Bojanowski, Piotr, et al. "Enriching word vectors with subword information." Transactions of the Association for Computational Linguistics 5 (2017): 135-146.
[7] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.
[8] Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." arXiv preprint arXiv:1609.08144 (2016).