1. 서론

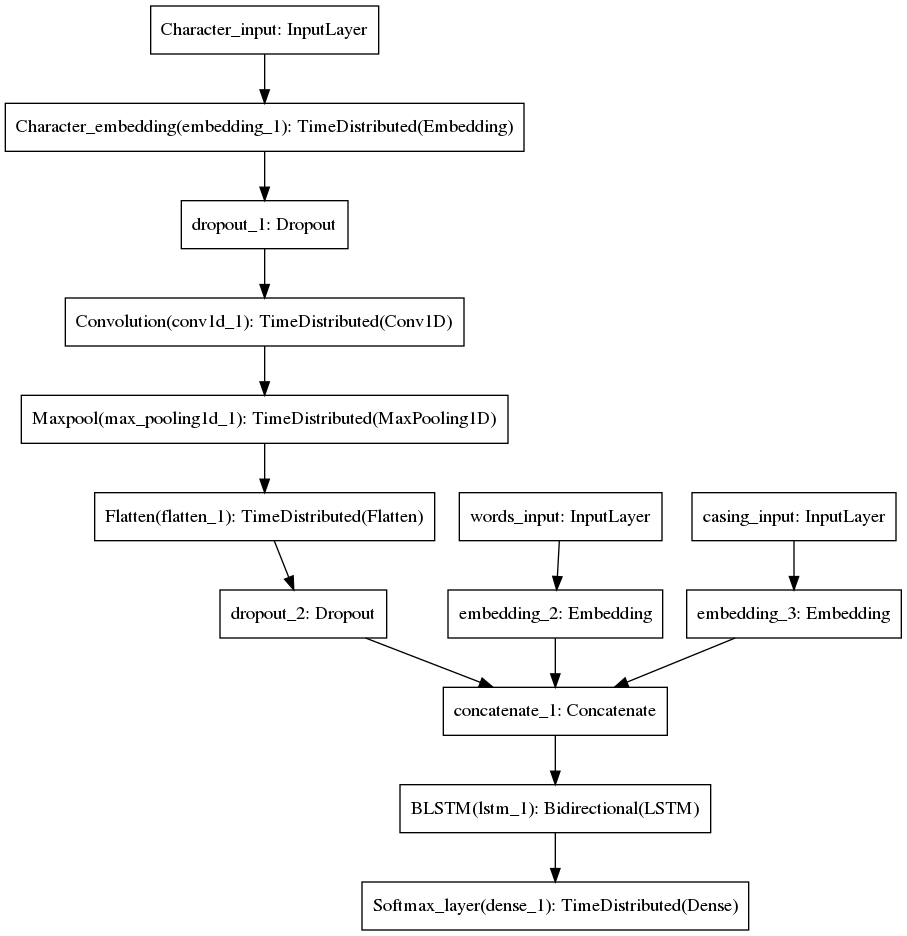
최근 Deep Neural Network Model은 다양한 자연어 처리 (NLP) 작업에서 최고의 성능 향상을 달성했지만, 여전히 해결해야 할 많은 과제가 남아 있습니다. 많은 기계학습 모델 중에서 DNN은 대량의 데이터를 필요로 하는 편이고, 모델이 복잡해질수록 학습에 필요한 장비와 시간이 증가하게 됩니다.

본 연구에서는 매우 적은 예제만 제공되는 환경에서도 일정 이상의 성능을 낼 수 있도록 구현된 모델을 기준으로 크기나 복잡도를 크게 높이지 않아 빠르게 학습되는 장점을 유지하지만, 실제 산업에서 활용되기 위해 필요한 수준의 높은 성능을 달성할 수 있도록 최적의 네트워크 구성에 관한 개선을 제안하고자 합니다.

실험 결과에 따르면 로 달성 할 수있는 69.3 %의 F1 점수는 78.87 %로 향상 될 수 있습니다.

2. 모델 설명 - 모델의 구조 등

가. Baseline 모델의 구조



Baseline 모델은 CoNLL-2003 텍스트 데이터 (J. Chiu & Nichols, 2016)에 대한 NER 아키텍처를 기반으로 합니다. 이 모델에서는 문자, 단어 및 케이싱의 세 가지 입력을 사용합니다. 각 입력은 텍스트의 다른 측면을 인코딩합니다 (그림 1 참조). 아키텍처는 이러한 세 개의 입력을 독립적으로 vector embedding한 다음 병합하여 Softmax 연산을 통해 Named Entity를 탐색합니다.  
 나. 각 레이어와 입력단의 특징

a. 문자 입력 layer (그림 1의 character\_input)는 숫자, 알파벳, 기호를 포함하는 97글자를 무작위로 초기화 된 30 차원 벡터로 임베딩합니다. 배치 당 입력 샘플 수 (그림 1의 b)와 샘플 당 단어 수 ( 'w')는 배치마다 다릅니다. 단어 당 최대 문자 수 ( 'c')는 52 자입니다.

b. 문자 입력에는 Drop-out레이어 (그림 1의 char\_dropout1 및 char\_dropout2)가 2번 적용되어 지나친 과적합의 위험을 감소시킵니다. Drop-out은 0.5 혹은 0.68을 적용하였습니다.

c. 1-dimension Convolution 레이어(char\_conv)는 크기 3인 커널을 30 개 사용하여 문자 입력의 특징을 추출합니다. 첫 번째 Conv는 Embedding layer의 출력에, 두 번째 Conv는 Max Pooling된 feature를 Flatten 한 출력에 적용되었습니다.  
 d. Max-pooling 레이어는 windows size가 52이고, stride를 52로 하여 문자 크기를 크기 1로 축소합니다.

e. 초기화는 RandomUniform, glorot\_normal, glorot\_uniform, lecun\_uniform, he\_normal, lecun\_normal, he\_uniform 등을 테스트하였으며, lecun\_uniform의 성능이 가장 좋았습니다.

바이어스 항은 0으로 초기화됩니다. → 바꿔봐야겠다.

f. 워드 입력 레이어 (words\_input)는 단어를 50 차원 벡터로 임베딩합니다. GloVE에서 제공하는 [Wikipedia 2014 + Gigaword 5]의 50d 벡터를 사용하였습니다.  
 g. 케이싱 레이어 (케이싱 \_ 인풋)는 단어의 형태 정보를 벡터로 매핑합니다. 형태 정보는 모두 8가지를 사용하였습니다: 숫자, 전체 소문자, 전체 대문자, 주로 숫자(단어에 속한 문자의 50 % 이상이 숫자), 첫 글자가 대문자, 숫자가 포함된 단어, 패딩 및 기타 (카테고리가 없는 경우)

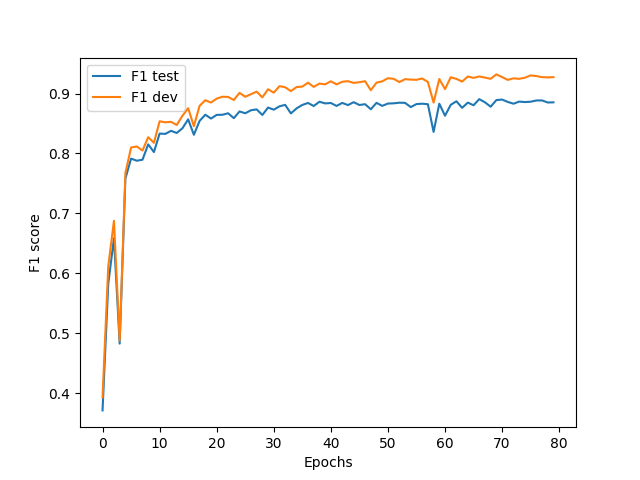
h. Concatenate 레이어(merge\_concatenate)는 벡터화된 세 개의 입력을 하나의 벡터로 연결합니다.

i. Bidirectional LSTM (Schuster & Paliwal, 1997) 레이어 (merge\_BLSTM)는 입력된 데이터를 200 단위의 두 벡터로 변환합니다. 커널은 Glorot uniform distribution에서 그리기로 초기화됩니다 Glorot & Bengio, 2010).

바이어스 항은 0으로 초기화됩니다. → 이것도 바꿔보자

다. Learning  
이 모델에서 모든 weight는 Nadam optimizer로 학습되고, Keras 및 TensorFlow 라이브러리를 사용하여 Python으로 구현되었습니다. 소스 코드는 [https://github.com/mxhofer/Named-Entity-Recognition-BidirectionalLSTM-CNN-CoNLL]의 코드를 사용하였습니다.

아무런 변경 없이 Epoch10회만으로 82.69의 F1을 기록하였습니다.



라. 모델의 성능

위 모델은 기본적으로 빠르고 훌륭한 성능을 보여주었습니다. 10 epoch를 시험삼아 학습시키는 경우 i5-4690 CPU 환경에서 7분 36초만에 학습을 종료했습니다.

Test set의 f1은 82.69, dev set의 f1은 84.71로 전반적으로 overfitting되며 test set 정확도가 떨어지는 모습은 관찰되지 않았습니다.

마. 모델의 개선

위 그림에서 보듯이 모델은 3가지 입력을 한꺼번에 NER task에 사용합니다. 캐릭터 레벨 입력은 Conv 레이어를 거치고, 워드 레벨 입력과 케이스 입력은 vector embedding만 적용 후 concatenated vector를 생성합니다. 생성된 context vector는 Bi-directional LSTM레이어를 거쳐 Softmax classifier에 입력됩니다.

실험 설계 - 데이터셋 설명, hyperparameter, 비교 대상 모델

실험 결과 및 분석 – 기존 타 연구 와의 성능 비교 포함

- 성능 평가시 F1 Evaluation Metric 준수,

결론

1. Introduction

전자 건강 기록 (EHR)은 일반의 (GP) 및 병원에서 환자의 병력을 구축하고 저장하는 데 사용되는 데이터베이스입니다 (O. A. Johnson, Fraser, Wyatt, & Walley, 2014).  
이러한 기록에는 약물 투여 이유, 과거 질환 및 과거 치료 결과 등의 정보가 포함되어 있으며 이들은 의학 연구에서 경험적 자료의 가장 큰 출처이며 암과 같은 매우 관련성이 높은 질환에서 중요한 과학적 발견을 가능하게합니다 알츠하이머 병 (Perera, Khondoker, Broadbent, Breen, & Stewart, 2014).  
그러나 이러한 EHR에 보관 된 대부분의 정보는 자연 언어의 형태로되어 통계 분석에 거의 접근 할 수 없게됩니다 (Murdoch & Detsky, 2013).  
이 정보의 잠금을 해제하면 생물 의학 연구에 상당한 발전을 가져올 수 있습니다.

규칙 기반 시스템은 추출 된 정보가 규칙적인 음성 패턴 (예 : ICD9와 같은 규제 의료 코드의 사례) (Karystianis et al., 2018)을 따르는 경우와 같이 간단한 상황에서는 의료 정보를 정확하고 정확하게 추출 할 수 있습니다.  
그러나 이러한 시스템은 복잡한 패턴 (예 : 증상 설명), 텍스트 패턴의 변형 (예 : 미국 영어와 영국식 영어) 또는 심하게 구조화 된 텍스트 (예 : 표준화 된 약어가 EHR에서 매우 일반적 임)에 잘 맞지 않습니다. EHR에서 발견되는 상황과 유사합니다.  
또한 규칙 기반 시스템을 설계하는 데는 시간이 많이 걸리고 전문적인 현장 지식이 필요합니다.  
이러한보다 복잡한 상황에서 기계 학습 (ML) 기반 방법은 일반 알고리즘을 기존 데이터로 조정하여 규칙 기반 방법을 능가합니다.  
가장 최근에, 신경망은 전통적인 규칙 기반 및 다른 ML 기반 방법이 실패한 복잡한 NLP 작업 (Young et al., 2017)에서 특히 성공적이었습니다 (Cambria & White, 2014).  
예를 들어, 긴 단기 메모리 (LSTM) 유형의 반복적 인 신경망 (RNN)과 컨볼 루션 신경망 (CNN)의 조합이 NER 작업 기반의 새로운 최첨단 성능을 설정하는 데 성공적으로 적용되었습니다. CoNLL-2003 및 OntoNotes 5.0 데이터 (J. Chiu & Nichols, 2016).

신경 네트워크의 주된 한계는 많은 양의 주석이 달린 텍스트의 필요성에 있으며, 이는 특히 EHR에 적용하기에는 까다로운 내용입니다.  
즉, 의료용 NLP에 대해 공개적으로 사용 가능한 데이터 세트는 거의 없으며 NLP 작업 (예 : 문서 분류 또는 슬롯 채우기)에 대한 주석이 더 적습니다.  
이 데이터 세트 중에서 가장 잘 알려진 것은 MIMIC-III (A. E. W. Johnson et al., 2016)와 i2b2 (i2b2, 2018, p.2)입니다.  
따라서 매우 적은 주석 예제가있을 때 신경망의 성능을 향상시키는 것은 생물 의학 연구에서 최우선 순위로 남아있다.

이 논문에서는 주석이 달린 예제가 거의 없을 때 신경망의 학습 능력에 대한 5 가지 순차적 개선 효과를 보여줍니다.  
우선 i2b2 2009의 NER 작업에서 성능을 최적화하는 목표를 설정하고 무작위로 선택된 주석 처리 된 방전 요약을 10 개만 사용합니다.  
최첨단 NER 아키텍처를 기준으로 (J. Chiu & Nichols, 2016), 우리는이 목표에 대한 성능을 크게 향상시키는 여러 가지 개선 사항을 순차적으로 설계하고 적용합니다.  
이는베이스 라인에서 69.3 %에서 78.87 %로 성능을 향상시킵니다.

2. Methods

2.1. Data

여섯 가지 의료용 NER 데이터 세트가 사용되었습니다.  
하나는 목표 작업을 정의하므로 감독 교육 및 시험에 사용됩니다 (i2b2 2009).  
(i2b2 2010 및 i2b2 2012)의 가중치에 대한 사전 감독을위한 두 가지  
맞춤 단어 임베딩 (BioNLP-2016, MIMIC-III 및 UK CRIS)에 대한 감독되지 않은 훈련을 위해 3 개.  
또한 비 의학적 데이터 세트 (CoNLL-2003)는 감독 된 사전 체중 훈련에 사용되었습니다.  
감독 방식으로 사용 된 각 데이터 세트 (즉, i2b2 세트 및 CoNLL-2003)는 라벨로 적용되는 많은 수의 타겟 NER 카테고리를 제공했지만 (표 1 참조), 감독되지 않은 패션 원 주석으로 사용 된 데이터 세트에서는  
무시되었다 (표 2 참조).  
우리는 i2b2b 2009 데이터를 전체 교육 데이터 세트에서 추출한 무작위 샘플 10 개로 제한했습니다.  
통합 생물학 및 베드 사이드 (i2b2) 비영리 재단은 매년 NLP의 일련의 과제를 수행하여 주석이 달린 손으로 쓴 모든 디지털 사본을 출판합니다.  
신원 미상의 임상 기록.  
2009 년 i2b2 도전 과제는 Partners Healthcare (O. Uzuner, Solti, & Cadag, 2010)의 확인되지 않은 퇴원 요약에서 약물 정보를 추출하는 데 중점을 둡니다.  
2010 i2b2 / VA Relations Challenge는 Partners Healthcare, Beth Israel Deaconess Medical Center 및 University of Pittsburgh Medical Center (2011 년 South, Uzuner, South, Shen 및 DuVall)의 퇴원 요약 및 진행 보고서를 사용했습니다.  
2012 년 i2b2 과제에는 파트너 헬스 케어 (Partners Healthcare) 및 베스 이스라엘 디코 누스 메디컬 센터 (Beth Israel Deaconess Medical Center)가 제공 한 문제, 테스트, 치료, 임상 부서, 증거 자료 (즉, 정보 출처를 나타내는 이벤트) 및 사건 (입원, 이동 등) (Sun, Rumshisky, & Uzuner, 2013).  
자연 언어 학습 회의 (CoNLL)는 SIGNLL (ACL의 특별 관심사  
자연어 학습 그룹).  
CoNLL-2003은 이전의 세 그룹 (Sang & De Meulder, 2003)에 속하지 않은 사람, 지사, 조직 및 기타 주체의 주석이 첨부 된 영어 뉴스 기사를 포함합니다.

BioNLP-2016은 PubMed Central Open Access 서브 세트 (PMC)와 PubMed (B. Chiu, Crichton, 그림 1의 char\_dropout2)의 텍스트 데이터를 기반으로 단어 벡터를 제공하며, 드롭 율 0.5 Korhonen & Pyysalo, 2016)

MIMIC III v1.4 데이터베이스는 매사추세츠 주 보스톤에있는 Beth Israel Deaconess Medical Center의 58,000 개의 중요한 진료 병원 기록으로 구성되어 있습니다 (A. E. W. Johnson 등, 2016).

영국 CRIS는 영국 전역의 정신 건강 병원의 EHR에 대한 통제 된 접근을 허용하며, 그 중 약 50 만 건의 환자 기록 (Callard 외, 2014; Stewart et al., 2009)을 포함하는 Warneford Hospital의 Oxford 정신 건강 클리닉에 접근했습니다.  
이 데이터 세트는 영국 국립 보건 서비스 (NHS)에서 관리하며, 윤리적 승인 및 감사 프로세스 (https://crisnetwork.co/에서 자세한 정보 참조)를 통해서만 액세스 할 수 있습니다.

2.2 Baseline model

기본 모델은 CoNLL-2003 및 OntoNotes 5.0 텍스트 데이터 (J. Chiu & Nichols, 2016)에 대한 최첨단 NER 아키텍처를 기반으로합니다.  
모델에는 문자 수준, 단어 수준 및 케이싱 입력의 세 가지 입력이 있습니다. 각 입력은 텍스트의 다른 측면을 인코딩합니다 (그림 1 참조).  
아키텍처는 이러한 세 개의 입력을 독립적으로 처리하기 시작한 다음 병합하여 추가 처리합니다.  
이 아키텍처는 수많은 원자 연산 또는 레이어를 통해 수행되며, 다음과 같이 설명됩니다.

• 문자 삽입 layer (그림 1의 char\_input)는 가능한 97 자의 어휘를 U (-0.5,0.5)에서 무작위로 초기화 된 30 차원 임베딩으로 매핑합니다.  
배치 당 입력 샘플 수 (그림 1의 b)와 샘플 당 단어 수 ( 'w')는 배치마다 다릅니다.  
단어 당 최대 문자 수 ( 'c')는 52 자입니다.

• 드롭 율이 0.5 인 드롭 아웃 레이어 (그림 1의 char\_dropout1 및 char\_dropout2)는 문자 수준 입력에 적용되어 지나친 조정의 위험을 완화합니다.

• 1D 길쌈 레이어 (char\_conv)는 폭 3의 30 개 커널로 1 차원 문자 입력을 처리합니다.  
이 계층 다음에는 창 크기 52와 스트라이드 52의 1d maxpool 조작 (char\_maxpool)이 있으며 문자 크기를 효과적으로 크기 1로 컴파일합니다.  
커널은 Glorot uniform distribution (Glorot & Bengio, 2010)에서 그리기로 초기화됩니다.  
바이어스 항은 0으로 초기화됩니다.

• 워드 임베딩 레이어 (words\_input)는 'V w'단어의 어휘를 50 차원 임베딩에 매핑합니다.  
달리 명시하지 않는 한, 우리는 6B 토큰 (Pennington, Socher, & Manning, 2014)과 함께 GloVE Wikipedia 2014 및 Gigaword 5 임베딩을 사용합니다.  
케이싱 임베딩 레이어 (케이싱 \_ 인풋)는 'V ca'케이싱 타입의 어휘를 V ca 차원 임베딩에 매핑합니다.  
기본적으로 숫자, allLower,  
allUpper, 주로 영숫자 (단어의 문자의 50 % 이상이 숫자), initialUpper, contains\_digit, padding 및 기타 (카테고리가 적용 가능하지 않은 경우)  
연결 계층 (merge\_concatenate)은 처리 된 문자 수준 (샘플 입력 당 30 개의 차원 벡터), 단어 수준 (50 차원) 및 대소 문자 (V ca 차원)  
데이터를 80 + V ca 차원의 벡터로 변환합니다.  
양방향 LSTM (Schuster & Paliwal, 1997) 레이어 (merge\_BLSTM)는 이전에 연결된 데이터를 200 단위의 두 벡터로 변환합니다. 하나는 입력에 앞으로 및 다른 뒤로 순환을 적용합니다. 커널은 Glorot uniform distribution에서 그리기로 초기화됩니다 Glorot & Bengio, 2010).  
바이어스 항은 0으로 초기화됩니다.  
이 기본 모델에서 모든 매개 변수는 기본 매개 변수 (Keras 버전 2.2.0에 정의 된대로)를 사용하여 Nadam 최적화 알고리즘으로 교육되어 데이터를 64 개 배치로 나눕니다.  
기본 모델은 Keras 및 TensorFlow 라이브러리를 사용하여 Python으로 구현되며 https://github.com/mxhofer/Named-Entity-Recognition-BidirectionalLSTM-CNN-CoNLL에서 사용할 수 있습니다.

2.3. Single pre-training

주석 처리 된 예제가 거의없는 컴퓨터 비전 작업에서 일반적으로 성공하는 접근 방법은 샘플을 사용할 수있는 다른 관련 작업에 대한 사전 교육으로 구성됩니다.  
예를 들어, 하나의 연구는 정지 이미지의 대형 분류 된 교육 코퍼스에 대한 사전 교육을 실시하고 비디오 인식을 위해 덜 깔끔하게 분류 된 이미지 코퍼스로 학습을 성공적으로 옮겼습니다 (Su, Chiu, Yeh, Huang, & Hsu, 2014).  
이 사전 교육 접근 방식은 기본 아키텍처 성능을 향상시키기 위해 적용한 첫 번째 방법입니다.  
네트워크 매개 변수는 목표 작업 (i2b2 2010 및 i2b2 2012, 의학 텍스트)과 동일한 도메인에 속한 두 개의 서로 다른 NER 작업에 각각 별도로 사전 교육을 받았고 다른 도메인 (CoNLL-2003, 비 의학적 텍스트).  
세 가지 다른 초기화 전략을 비교했습니다. 사전 훈련 된 가중치로 모든 레이어 초기화. 오직 레이어 merge\_BLSTM 초기화 (다른 레이어들은 무작위로 초기화 됨); 과  
merge\_BLSTM 이외의 모든 것을 초기화합니다 (merge\_BLSTM은 무작위로 초기화됩니다).  
모든 경우에, words\_input의 삽입은 훈련되지 않고 GloVE의 값으로 동결됩니다.

2.4. Hyperparameter tuning

우리의 두 번째 향상은 3.3 절에서 설명 된 것 이외에, 그리드 검색을 통해 하이퍼 매개 변수 값을 조정하는 것으로 구성되었습니다.  
이것은 WikiQA 및 SemEval-2016 (Min, Seo, & Hajishirzi, 2017)과 같은 여러 NLP 작업에 성공적으로 적용된 일반적인 방법입니다.  
우리가 미세 조정 한 하이퍼 패러미터는 최적화 기 (확률 적 경사 구배 (SGD) 또는 나담 중 하나 선택); 예비 교육 데이터 세트 (i2b2 2010 또는 i2b2 2012); SGD 학습 률 (0.04 또는 0.08); 배치 정규화 (유무에 관계없이); 훈련 가능한 단어 임베딩 (객관적인 작업에 훈련되거나 글로브 값에 고정 된 계층 'words\_input'의 가중치) 및 학습 속도 감소 (일정 또는 일정 시간).

2.5. Combined pre-training

개별 데이터 세트의 사전 교육 외에도 (2.3 절 참조), 결합 된 데이터 세트의 사전 교육은 대상 작업의 성능을 더욱 향상시킬 수 있습니다.  
예를 들어, 건강 정보학 연구는 CT (폐 CT) 스캔을 위해 여섯 개의 공개 된 데이터 세트를 사용하여 컨볼 루션 신경 네트워크를 사용하여 폐 패턴을 식별함으로써 목표 도메인에서 2 %의 성능 향상을 가져 왔습니다 (Christodoulidis, Anthimopoulos, Ebner, Christe, & Mougiakakou, 2017).  
우리의 목표 작업이 사전 학습을 결합하여 이루어 졌는지 여부를 테스트하기 위해 이제 i2b2 2010과 2012의 학습을 결합했습니다.  
이는 두 가지 가능한 방향 중 하나에서 각 데이터 세트를 순차적으로 학습함으로써 달성됩니다. i2b2 2010에서 무작위로 초기화 된 모델을 먼저 교육 한 다음 i2b2 2012에서 계속 교육을 수행합니다. 또는 2010 년, 그리고 2012 년에 첫 번째 훈련.  
두 경우 모두 객관적인 과제를 훈련 할 때 두 번째 훈련에서 얻은 최종 가중치를 초기 값으로 사용했습니다.

2.6. Customized word embeddings

이전에 (2.2 절, 2.3 절, 2.4 절 및 2.5 절) 우리는 입력 단어 (그림 1의 word\_input)에 대해 Glove 삽입을 사용했지만 이러한 벡터 표현은 의료 용어에 대해 부정확 할 것으로 예상됩니다. 글로브.  
이 문제를 해결하기 위해 우리의 네 번째 개선점은 CRIS, MIMIC III 또는 BioNLP-2016에서 훈련 된 자체 단어 임베딩 개발이었습니다.  
CRIS와 MIMIC III 임베딩은 Facebook의 FastText (Bojanowski, Grave, Joulin, & Mikolov, 2016, Joulin, Grave, Bojanowski, & Mikolov, 2016) 알고리즘, 최소 단어 수 5, 초기 학습 속도 0.05 및 문맥 창 크기는 5입니다.  
BioNLP-2016은 공식 저장소 (Cambridge Language Technology Lab, 2018)에서 다운로드하여 추가 사전 처리없이 사용했습니다.

2.7. Optimizing OOV words

데이터 조사에 따르면 대상 데이터 세트 (i2b2 2009)에는 어휘 밖 (OOV) 단어가 많이 포함되어있었습니다.  
이것은 퍼포먼스에 매우 해로운 영향을 미칠 수있는 퍼짐의 어휘에 포함되지 않은 단어입니다.  
우리의 경우, OOV 용어의 상당 부분은 "복용량 증가 : 일주일에 +20 mgs"와 같이 후행 주인공을 포함합니다.  
OOV를 줄이기 위해 마지막 개선 사항은 텍스트 전처리에 다음 두 단계를 추가했습니다.  
• 후행 ":", ";", "."및 "-"를 제거하십시오.  
• 인용문 삭제  
• 선행하는 "+"

3. Results

소개에서 설명한 바와 같이, 먼저 2009 년 i2b2 도전 과제의 공식 목표로 구성되었지만 10 개의 주석이 달린 무작위로 추출 된 배출 요약에 대해서만 학습을 허용하는 객관적 과제를 배우는 몇 번의 촬영을 정의했습니다.  
그 후, 우리는 최첨단 NER 아키텍처 중 하나를 구현하고이 기본 모델에 비해 5 가지 순차적 개선 효과를 평가했습니다.

3.1. Baseline model

베이스 라인 모델은 CoNLL-2003 및 OntoNotes 5.0 코퍼러리 (J. Chiu & Nichols, 2016)에 대해 처음 제안 된 최첨단 NER 아키텍처를 기반으로합니다.  
방법들 (그림 1 참조)에서 설명한 바와 같이, 그것은 케이싱, 단어 및 문자 레벨 입력이있는 BLSTM과 하나의 컨볼 루션 및 드롭 아웃이있는 BLSTM으로 구성됩니다.  
레이어 'words\_input'은 Wikipedia 2014 및 Gigaword (6B 토큰)에서 학습 된 GloVE 삽입으로 초기화됩니다.  
이러한 embedding은 초기화 후에 동결되며 backpropagation 중에는 더 이상 수정되지 않습니다.  
'char\_input'과 'casing\_input'의 삽입은 균일 분포 U (-0.5, 0.5)로 무작위로 초기화됩니다.  
다른 모든 매개 변수는 Keras (버전 2.2.0) 기본값에 따라 무작위로 초기화됩니다.  
그런 다음 데이터는 64 개의 배치로 나뉘며 모든 고정되지 않은 포함, 가중치 및 편향이 기본 Keras 매개 변수로 Nadam에 의해 교육됩니다.  
이 초기화 및 교육 방법을 사용한이 기본 아키텍처는 객관적인 작업에서 69.30 %의 F1 점수를 획득했습니다 (그림 2 참조).

3.2. Single pre-training

단일 사전 훈련은 기본 모델을 통해 구현 된 첫 번째 개선 사항입니다.  
베이스 라인 모델이 모든 파라미터를 무작위로 초기화하는 동안 (레이어 'words\_input'의 임베딩 제외), 단일 사전 학습은 다른 데이터 세트에서 사전 학습을 통해 부분 또는 전체 신경 네트워크를 초기화했습니다.  
CoNLL-2003 (+ 0.52 %)에서 사전 훈련 된 가중치보다 무작위로 초기화 된 가중치보다 i2b2 2010 (기본 F1 대비 평균 3.06 % 증가) 또는 2012 년 i2b2 (+ 1.58 %)에 대한 사전 교육을 통해 레이어를 초기화 할 수 있습니다. (69.3 %).  
또한 모든 레이어 초기화는 BLSTM (71.21 %) 만 초기화하거나 BLSTM 레이어 (70.65 %)를 초기화하는 것보다 낫습니다.  
모든 조합의 결과를 표 2에 나타내었다.  
모든 레이어를 사전 교육하고 i2b2 2010에서 기준선을 기준으로 + 4.52 %의 F1 점수로 최고의 성능을 달성 한 조합이었습니다.  
따라서 우리는이 개선을 후속 실험을위한 모델에 통합했습니다.  
Embedding 흥미롭게도, 의료 영역의 데이터 세트에 대한 사전 교육은 다른 방법보다 빠른 학습을 가능하게했습니다.  
2010 년과 2012 년의 i2b2 평균 평균 점수는 58.98 % 였지만 CoNLL-2003의 경우 50.2 % 였고 무작위 초기화의 경우 28.47 %였습니다 (그림 2 참조).

3.3. Hyperparameters

하이퍼 파라미터 튜닝은 테스트 된 두 번째 개선이며, 단일 사전 교육을 통해 최고의 성능을 발휘하는 모델을 통해 구현되었습니다.  
평가 된 모든 하이퍼 매개 변수 중 가장 큰 영향을주는 것이 최적화 프로그램이었으며 Nadam은 평균 70.41 %, SGD는 50.56 %를 달성했습니다.  
두 번째로 중요한 하이퍼 매개 변수는 신경망 사전 교육에 사용되는 데이터로, 2010 년 i2b2 2010은 최대 + 2.34 %, i2b2 2012 최대 + 1.58 %를 달성했습니다.  
그림 3에서 볼 수 있듯이 다른 하이퍼 매개 변수 (일괄 정규화, 학습 가능 삽입, 학습 속도 및 학습 속도 감소)의 영향은 결정적이지 않았습니다.  
따라서 우리는 최적화 프로그램을 Nadam에 고정 시켰고 사전 교육을 위해 i2b2 2010을 사용하면서 다른 hyperparameter 변경을 거부했습니다.

3.4. Combined pre-training

여러 데이터 세트와 동시에 사전 교육을 거친 세 번째 개선 사항은 미세 조정 된 하이퍼 매개 변수의 최상의 모델을 통해 구현되었습니다.  
그 결과, i2b2 2010과 별도의 사전 교육은 어느 방향 으로든 예비 교육을 결합하여 수행 한 결과보다 우수합니다.  
2010 년과 2012 년의 프리 트레이닝은 F1이 -1.85 %, 반대 방향이 -1.66 %로 나타났습니다.  
따라서 우리는 결합 된 사전 교육을 거부하고 후속 실험을 위해 i2b2 2010 만 계속 사용했습니다.

3.5. Customizing word embeddings

사전 훈련을 통해 레이어 'words\_input'의 삽입을 사용자 정의하는 것은 테스트 된 네 번째 개선이며 미세 조정 된 매개 변수의 최상의 모델을 통해 구현되었습니다 (결합 된 사전 학습은 거부되었습니다).  
모든 이전 모델은 공개적으로 사용할 수있는 값 (https://nlp.stanford.edu/projects/glove/)으로 고정 된 50 개의 차원으로 된 Glove 단어 임베딩을 사용했습니다.  
이들을 MIMIC III (50 차원 또는 200 차원의 단어 벡터)에서 훈련 된 삽입물로 대체하면 3.3 절의 최상의 결과와 관련하여 F1 점수 (200 차원)가 3.78 % 향상되었습니다. BioNLP-2016 (200 차원, 창 크기, 2 또는 30의 w) -7.4 %; 영국 CRIS (50 또는 200 차원) -2.16 %.  
완료를 위해, 세 가지 가능성 모두가 i2b2 2010 또는 i2b2 2012에서 나머지 네트워크를 사전 교육하는 동안 평가되었습니다 (각 조합의 결과는 표 3 참조). 가장 좋은 조합은 78.07 % (+ 3.78 %)의 F1 점수를 가진 i2b2 2010의 MIMIC III였습니다.  
흥미롭게도, MIMIC III는 객관적인 과제와 관련하여 OOV 용어가 가장 적은 데이터 집합이기도합니다 (표 4 참조).

3.6. Optimizing OOV words

'words\_input'레이어 삽입시 OOV 단어의 수를 줄이는 것이 테스트 된 다섯 번째이자 마지막 개선이었으며 가장 맞춤형 퍼가기를 사용하여 모델을 통해 구현되었습니다.  
OOV 단어의 수는 단어 embedding을 생성하는 데 사용 된 데이터 세트에 따라 다릅니다 (표 4 참조).  
섹션 3.7에 설명 된 추가 전처리 단계는 MIMO III에서 3 %, CRIS에서 7 %, Glove에서 11 %, BioNLP-2016에서 11 % 감소 된 OOV 단어를 감소시킵니다.  
결과는이 단계가 모든 경우에 F1 점수를 약간 향상시킬 수 있지만 결코 성능을 저하시키지 않는다는 것을 보여줍니다.  
사전 트레이닝 가중치에 i2b2 2010을 사용하면 MIMIC III 임베딩 및 추가 사전 처리 단계를 통해 F1 점수가 78.07 %에서 78.87 %로 향상됩니다.

4. Discussion

제한된 주석 샘플을 갖는 문제는 원샷 (Fei-Fei, Fergus, & Perona, 2006, Vinyals, Blundell, Lillicrap, Kavukcuoglu, & Wierstra, 2016) 및 제로 샷 학습 (Xian , Lampert, Schiele, & Akata, 2017).  
그러나 우리가 아는 한,이 문제는 문제가 특히 과세하는 의학 텍스트에서는 다루어지지 않았습니다.  
즉, EHR의 의학 텍스트는 다른 코로나와 다르므로 임상 적으로 훈련받지 않은 독자는 기술 용어, 비표준 약자 및 비공식적 인 속기의 광범위한 사용으로 인해 EHR 텍스트를 거의 이해할 수 없습니다.  
이러한 이유 때문에 다른 더 잘 주석이 있지만 비 의료용 (예 : 표준 장갑 삽입 또는 Wikipedia에서 훈련 된 언어 모델 사용)의 표준 이전 학습 방법은 좋지 않은 수행이 예상됩니다.  
따라서 우리는이 논문과 같은 공헌이 특히 중요하다고 생각합니다.

최종 모델 외에, 우리의 실험은이 도전에 적용 할 수있는 많은 관찰을합니다.  
기준 모델에 비해 첫 번째 개선 (3.2 절)에서 얻은 결과는 대상 도메인과 유사한 텍스트를 기반으로 사전 훈련 된 가중치가 성능을 크게 향상시키는 것을 의미합니다 (표 2 및 그림 2 참조).  
도메인 내에서보다 도메인 내에서 더 나은 전송 학습은 다른 곳에서도보고되었습니다. (Tan, Zhang, Pan, & Yang, 2017).  
우리가 관찰 한 향상은 학습의 첫 번째 단계에서 특히 과감했으며 오랫동안 학습 한 후에 점근선의 양수 상수로 감소하는 경향이 있습니다 (그림 2 참조).  
두 번째 모델 개선 (3.3 절)은 BN (batch normalization) 및 TE (trainable word embedding)를 사용하는 것보다 옵티 마이저의 선택이 F1 점수에 더 큰 영향을 미친다는 것을 나타냅니다.  
BN과 TE는 결론적이지 않은 결과를 보여 주지만 Nadam 최적화 도구는 테스트 한 모든 상황에서 SGD를 능가합니다 (그림 3 참조).  
이는 NER 작업에 대한 다양한 하이퍼 매개 변수의 영향을 평가 한 최근 연구와 일치합니다.  
그들은 Nadam 최적화 도구가 가장 빠른 속도로 수렴하는 동안 가장 잘 수행함을 발견했습니다 (Reimers & Gurevych, 2017).  
사전 교육 데이터와 세 번째 개선, 성능 저하 (3.4 절)를 결합했습니다.  
 이에 대한 가능한 설명은 두 가지입니다.  
첫째, 3.3 절의 개별 데이터 세트에서 하이퍼 매개 변수를 최적화했습니다.  
둘째, 모델이 첫 번째 데이터 집합에 대해 학습하고 두 번째 데이터 집합의 가중치를 조정하면 초기 값이 부적합하다는 것을 알게되고 나중에 가중치를 임의로 초기화하는 것보다 훨씬 더 많은 조정을해야합니다. 별도의 사전 훈련 된 가중치를 생성합니다.  
마지막으로 네 번째 개선 (3.5 절)의 결과는 객관적인 작업이 특히 성능을 향상시키는 동일한 도메인의 텍스트에 대한 사용자 지정 단어 포함을 제안합니다 (표 3 참조).  
이것은 부분적으로 도메인 특정 코사인에 대한 임베딩 교육으로 인한 OOV 단어가 적기 때문일 수 있습니다 (표 4 참조).  
최종 의견으로, 우리의 연구는 여전히 몇 가지 한계를 가지고 있으며, 앞으로의 연구는 이러한 문제를 해결해야합니다.  
첫째, 개선 단계가 순차적으로 적용되었으므로 다른 순서는 객관적인 작업의 성능을 더욱 향상시킬 수 있습니다.  
이 검색 공간 제한을 감안할 때이 연구에서 발견 된 결과는 보수적 인 것으로 간주 될 수 있습니다.  
마지막으로, 78.87 %의 F1 점수는 대형 주석 처리 된 코로나로 얻을 수있는 성능에 비해 여전히 상당히 낮습니다.

Acknowledgements

UK CRIS는 Dementias Platform UK 및 Oxford Health National Health Service (NHS) 재단의 NIHR Biomedical Research Centre를 포함하여 국립 보건 연구원 (NIHR)과 의학 연구위원회 (MRC)에서 지속적으로 자금을 지원 받고 있습니다. 트러스트와 옥스퍼드 대학교.