A Survey on Deep Learning for Named Entity Recognition

2021.10.03

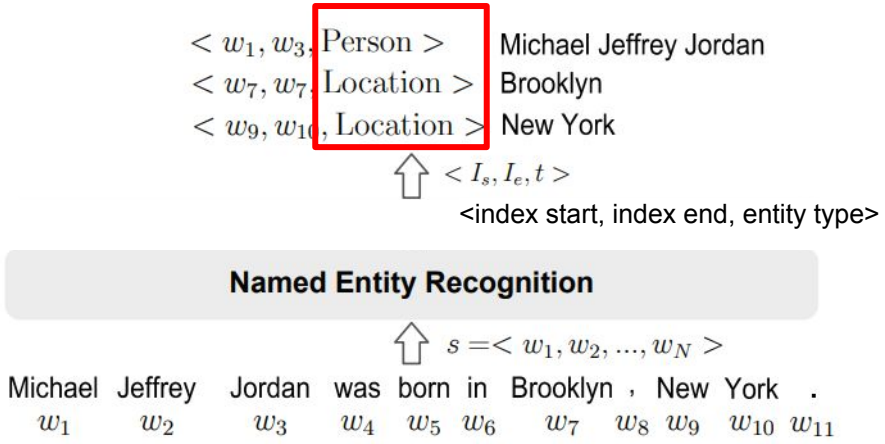
집현전 중급반 16조 류명현, 이수민, 이아름

목차

1. INTRODUCTION & BACKGROUND

1.1 NER이란

텍스트에서 사람, 위치, 조직 등과 같은 사전 정의된 엔티티를 인식하고 범주로 분류



1.2 NER 의 중요성

- information retrieval, question answering, machine translation과 같은 다운스트림 응용 프로그램에서 중요한 전처리 단계 역할을 함

- ex) semantic search 검색 엔진이 사용자의 질의 뒤에 있는 의미, 개념 및 의도를 이해할 수 있도록 하는 기술의 모음

  ➔ 검색 쿼리의 약 71%는 하나 이상의 Named Entity 포함
  ➔ 검색 쿼리에서 NE를 인식하면 사용자의 의도를 더 잘 이해하여 나은 검색 결과 제공할 수 있음

1.3 NER 리소스: 데이터 세트 및 도구

Person,
Location,
Organization,
Miscellaneous

TABLE 1
List of annotated datasets for English NER. "#Tags" refers to the number of entity types.

Corpus	Year	Text Source	#Tags	URL
MUC-6	1995	Wall Street Journal	7	https://catalog.ldc.upenn.edu/LDC2003T13
MUC-6 Plus	1995	Additional news to MUC-6	7	https://catalog.ldc.upenn.edu/LDC96T10
MUC-7	1997	New York Times news	7	https://catalog.ldc.upenn.edu/LDC2001T02
CoNLL03	2003	Reuters news	4	https://www.clips.uantwerpen.be/conll2003/ner/
ACE	2000 - 2008	Transcripts, news	7	https://www.ldc.upenn.edu/collaborations/past-projects/ace
OntoNotes	2007 - 2012	Magazine, news, web, etc.	18	https://catalog.ldc.upenn.edu/LDC2013T19
W-NUT	2015 - 2018	User-generated text	6/10	http://noisy-text.github.io
BBN	2005	Wall Street Journal	64	https://catalog.ldc.upenn.edu/LDC2005T33
WikiGold	2009	Wikipedia	4	https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500
WiNER	2012	Wikipedia	4	http://rali.iro.umontreal.ca/rali/en/winer-wikipedia-for-ner
WikiFiger	2012	Wikipedia	112	https://github.com/xiaoling/figer
HYENA	2012	Wikipedia	505	https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/hyena
N^3	2014	News	3	http://aksw.org/Projects/N3NERNEDNIF.html
Gillick	2016	Magazine, news, web, etc.	89	https://arxiv.org/e-print/1412.1820v2
FG-NER	2018	Various	200	https://fgner.alt.ai/
NNE	2019	Newswire	114	https://github.com/nickyringland/nested_named_entities
GENIA	2004	Biology and clinical text	36	http://www.geniaproject.org/home
GENETAG	2005	MEDLINE	2	https://sourceforge.net/projects/bioc/files/
FSU-PRGE	2010	PubMed and MEDLINE	5	https://julielab.de/Resources/FSU_PRGE.html
NCBI-Disease	2014	PubMed	1	https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/
BC5CDR	2015	PubMed	3	http://bioc.sourceforge.net/
DFKI	2018	Business news and social media	7	https://dfki-lt-re-group.bitbucket.io/product-corpus/

1.4 NER approaches

Traditional

- Rule-based approaches
- Unsupervised learning
- Feature-based supervised learning

NEW

- Deep-learning based approaches

1.4.1 Rule-based approaches

- 손으로 만든 규칙에 의존

  규칙: domain-specific 사전, 구문-어휘 패턴 등을 기반으로 설계

  ➡ **사전이 철저할 때만 잘 작동함**

  ➡ 도메인 고유 규칙과 불완전한 사전으로 높은 정밀도, 낮은 재현율 관찰,

  즉 다른 도메인으로 시스템을 이전 X

- ex)

  음성 입력 : Brill의 품사 tagger을 기반으로 규칙을 자동 생성

  생의학 영역: 사전 처리된 동의어 사전을 활용하는 prominer 제안

1.4.2 Unsupervised learning approaches

- 일반적인 접근법은 클러스터링
  :컨텍스트 유사성을 기반으로 클러스터링된 그룹에서 NE를 추출
  (대규모 코퍼스에서 계산된 통계를 사용해 NE 언급 추론)

- ex)
  *Zhang, Elhadad: 생물의학 text에서 용어, 말뭉치 통계(TF-IDF, Context Vector), 얕은 구문 지식(명사구 등)에 의존하는 모델 제안

1.4.3 Feature-based supervised learning

- 지도학습을 적용하면 NER은 다중 클래스 분류나 시퀀스 레이블 지정 테스크가 됨
  :어노테이션된 데이터 샘플이 주어지면 feature들이 각 training 예시를 잘 나타내도록 설계됨

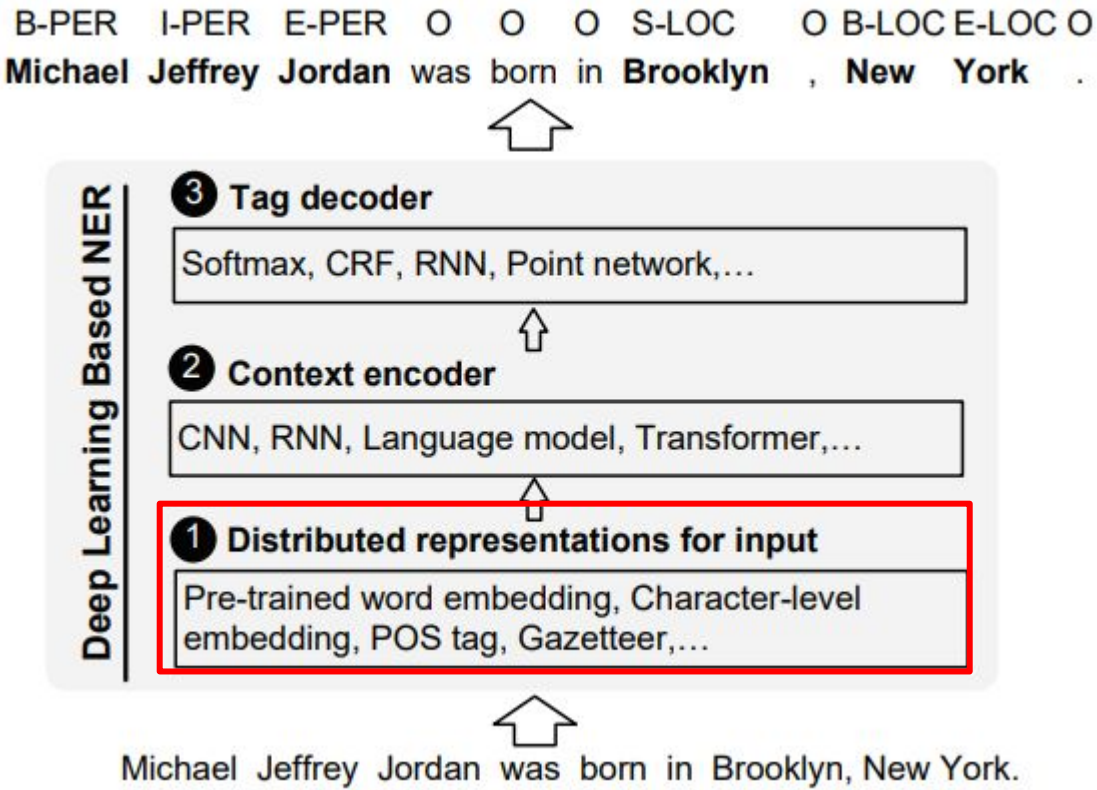
  ➡ 이후 기계학습 알고리즘을 사용해 데이터에서 유사한 패턴을 인식하는 모델을 학습함

- ex)
  - *HMM, 의사결정트리 등의 많은 기계학습 알고리즘이 supervised NER에 적용됨
  - *Szarvas et al: C4.5 의사결정 트리와 AdaBoostM1 알고리즘을 사용하여 다국어 NER 시스템 개발

2. DEEP LEARNING TECHNIQUES FOR NER

2.1 Why Deep Learning for NER?

1. 활성화함수(activation function)덕분에 비선형 변환이 쉽습니다.
2. NER 기능의 설계를 간단하게 해줍니다.
3. end-to-end 학습이 가능합니다.


B-PER I-PER E-PER O O O S-LOC O B-LOC E-LOC O
Michael Jeffrey Jordan was born in Brooklyn , New York .
Deep Learning Based NER
3 Tag decoder
Softmax, CRF, RNN, Point network,…
2 Context encoder
CNN, RNN, Language model, Transformer,…
1 Distributed representations for input
Pre-trained word embedding, Character-level embedding, POS tag, Gazetteer,…
Michael Jeffrey Jordan was born in Brooklyn, New York.

2.2 Distributed Representations for Input

1. Word-level Representation

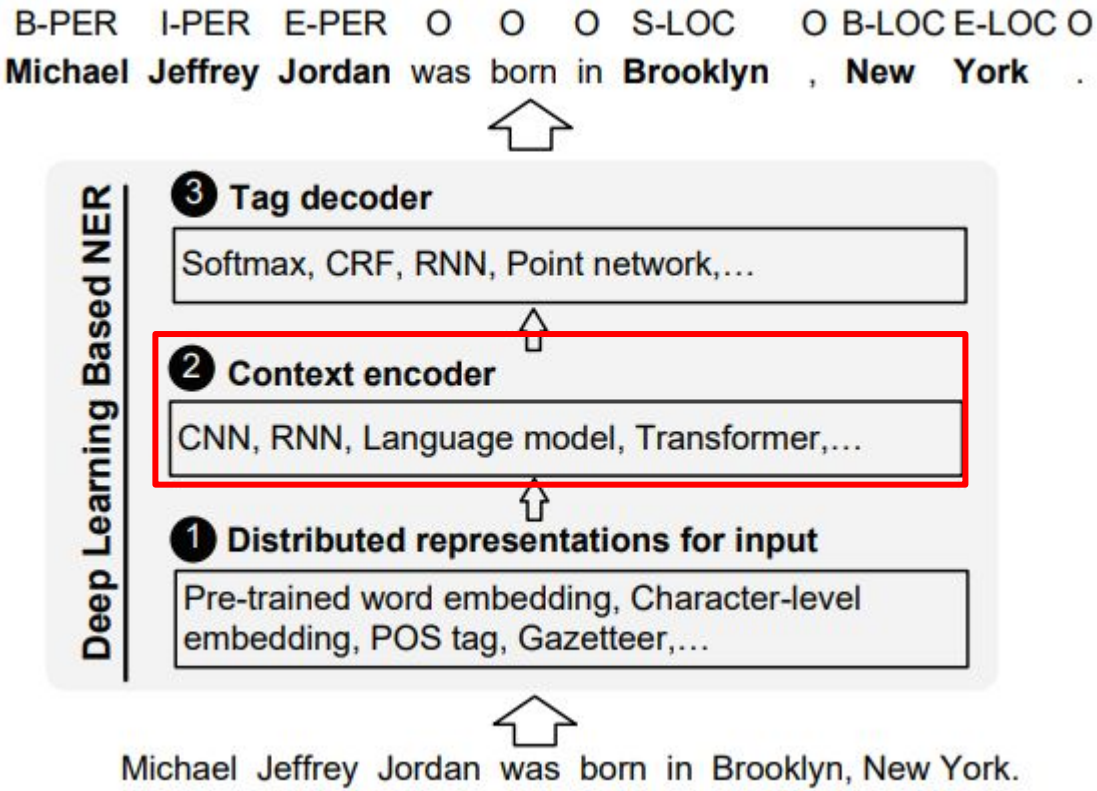
- pre-trained word embeddings : Google Word2Vec, Stanford GloVe, Facebook fastText , SENNA...
- training시에 pre-trained word embedding을 고정하거나 fine-tuning을 진행

2. Character-level Representation

- 접두사 및 접미사와 같은 명시적 하위 단어 수준 정보를 활용하는데 유용
- out-of-vocabulary를 자연스럽게 처리함 즉 보이지 않는 단어(unseen)를 표현하고 형태소 수준의 규칙성에 대한 정보를 공유
- 두가지 아키텍처 : CNN, RNN(기본적으로 LSTM or GRU)

3. Hybrid Representation

- 추가 정보(ex) 지명 목록, 어휘 유사성, 언어적 종속성및 시각적 특징)를 최종 representations of words에 포함
- 시스템의 일반성을 손상시키는 대가로 NER 성능이 향상


B-PER I-PER E-PER O O O S-LOC O B-LOC E-LOC O
Michael Jeffrey Jordan was born in Brooklyn , New York .
Deep Learning Based NER
3 Tag decoder
Softmax, CRF, RNN, Point network,…
2 Context encoder
CNN, RNN, Language model, Transformer,…
1 Distributed representations for input
Pre-trained word embedding, Character-level embedding, POS tag, Gazetteer,…
Michael Jeffrey Jordan was born in Brooklyn, New York.

2.3 Context Encoder Architectures

1. Convolutional Neural Networks
2. Recurrent Neural Networks
3. Recursive Neural Networks
4. Neural Language Models
5. Deep Transformer

2.3.1 Convolutional Neural Networks

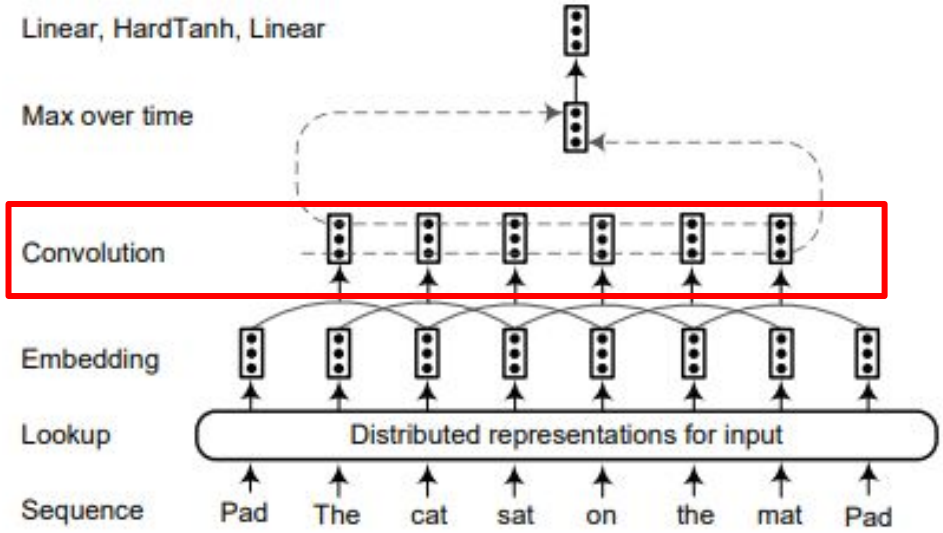
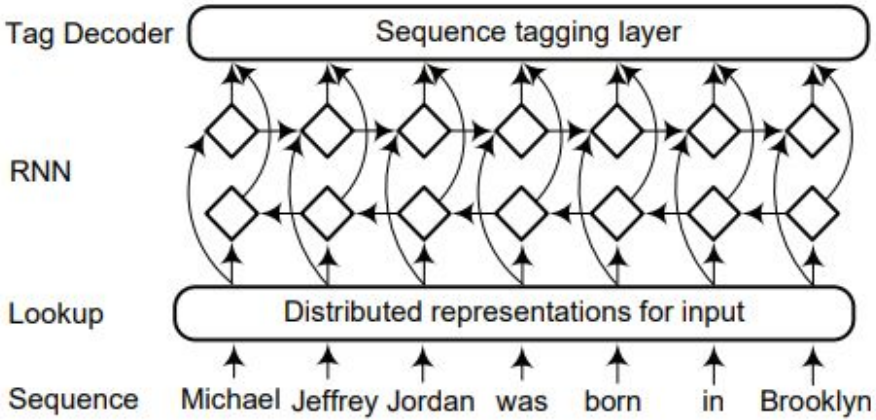


Fig. 5. Sentence approach network based on CNN [17]. The convolution layer extracts features from the whole sentence, treating it as a *sequence* with global structure.

각 단어 주위에 local features을 생성
(문장 단어 수에 따라 크기 결정)

→ Global feature vector는 Convolution layer에서 추출한 local features를 결합하여 구성됨
(문장 길이와 무관하게 고정)

2.3.2 Recurrent Neural Networks

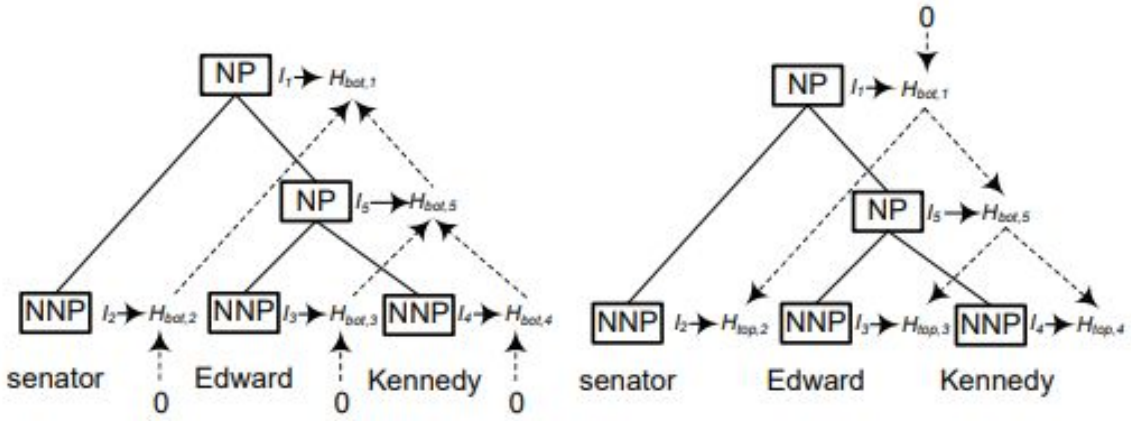


GRU, LSTM과 같은 변형과 함께, 순차적 데이터 모델링에서 좋은 성과

특히, Bi-RNN은 특정 시간 프레임에 대해 과거정보와 미래정보를 효율적으로 사용함.

➔ Bi-RNN은 텍스트의 깊은 컨텍스트 종속 표현을 구성하기 위한 표준이 됨

2.3.3 Recursive Neural Networks



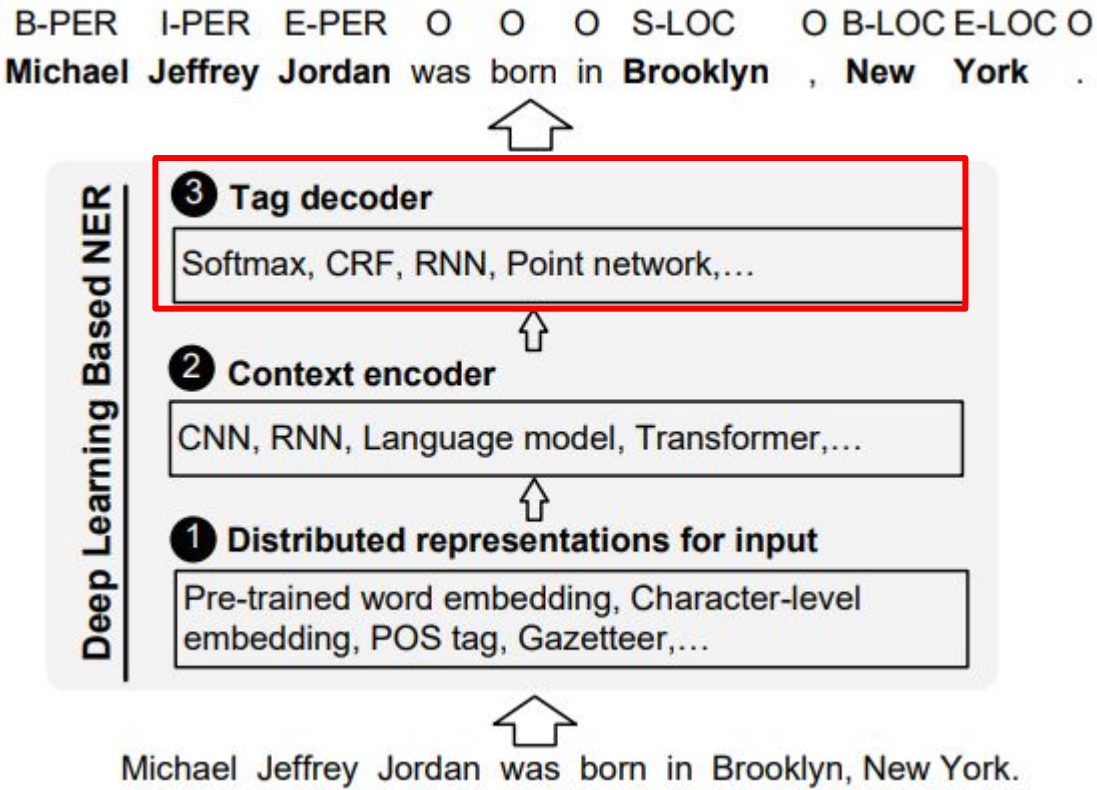
topological 순서로 주어진 구조를 순회함으로서 심층 구조 정보를 학습할 수 있는 비선형 적응 모델

2.3.4 Neural Language Models

- topological 순서로 주어진 구조를 순회함으로서 심층 구조 정보를 학습할 수 있는 비선형 적응 모델

2.3.5 Deep Transformer

- GPT, ELMO, BERT.....

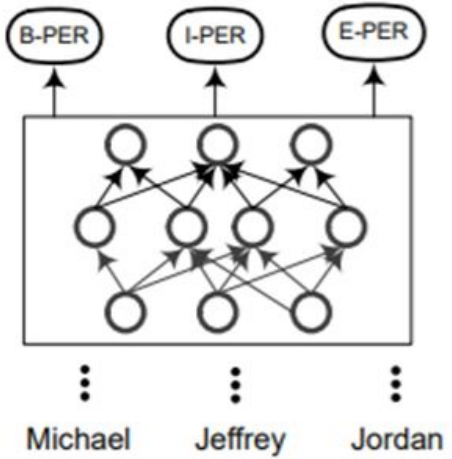

B-PER I-PER E-PER O O O S-LOC O B-LOC E-LOC O
Michael Jeffrey Jordan was born in Brooklyn , New York .
Deep Learning Based NER
3 Tag decoder
Softmax, CRF, RNN, Point network,…
2 Context encoder
CNN, RNN, Language model, Transformer,…
1 Distributed representations for input
Pre-trained word embedding, Character-level embedding, POS tag, Gazetteer,…
Michael Jeffrey Jordan was born in Brooklyn, New York.

2.4 Tag Decoder Architectures

1. MLP + Softmax
2. CRF(Conditional Random Fields)
3. RNN

2.4.1 MLP + Softmax

- NER은 sequence labeling problem.
- MLP + Softmax 계층을 태그 디코더 계층으로 사용하면 sequence labeling problem이 다중 클래스 분류 문제로 변환
- 각 단어에 대한 태그를 독립적으로 예측



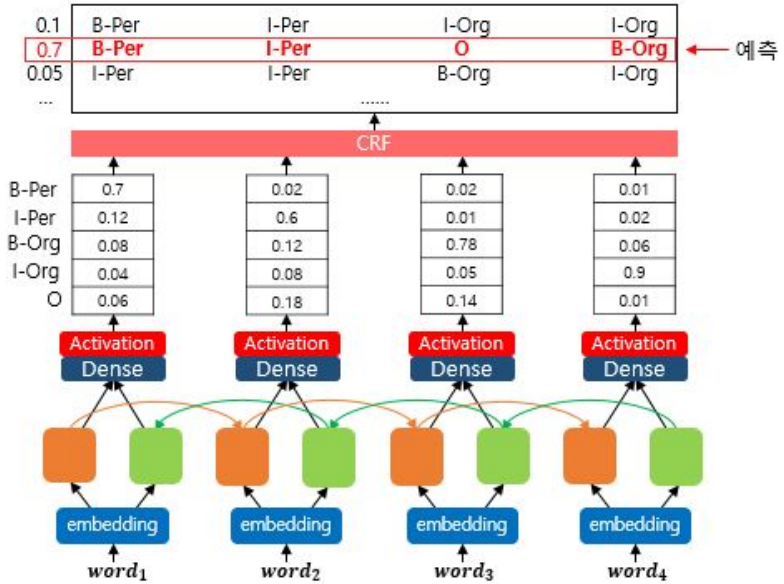
(a) MLP+Softmax

2.4.2 CRF(Conditional Random Fields)

- 레이블의 인접성에 대한 정보를 바탕으로 레이블을 추측하는 기법

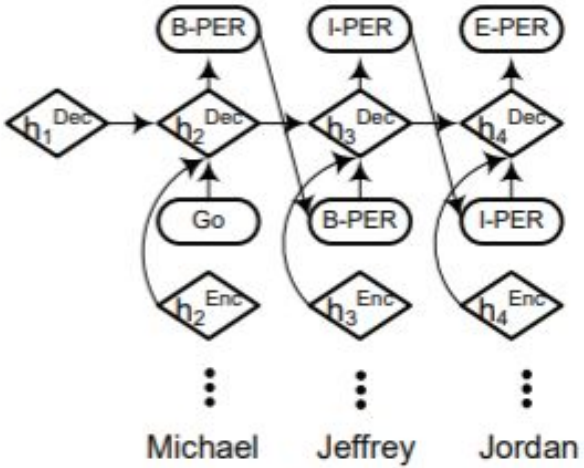
- 기존 모델들은 활성화 함수를 지난 시점에서 개체명을 결정했지만, CRF층을 사용하므로써 전체 문맥을 고려

- 즉 가장 높은 점수를 갖는 **시퀀스**를 예측



2.4.3 Recursive Neural Networks

- RNN 태그 디코더가 CRF보다 성능이 뛰어나고 entity 유형의 수가 많을 때 학습 속도가 더 빠름
- 최초 [GO] 토큰으로 시작이 되고, 각각의 시간 단계 i에서 이전단계 태그와 hidden state 그리고 현재 단계의 hidden state를 계산
- 최종적으로 softmax를 이용해 디코딩 후 출력 태그 예측



(c) RNN

2.5 Summary of DL-based NER

- NER 시스템의 성공은 input representation에 크게 의존함

- 사전학습모델에 fine-tuning을 진행하는 것이 NER의 새로운 패러다임(2018년 기준)으로 성능이 크게 향상됨

2.5 Summary of DL-based NER

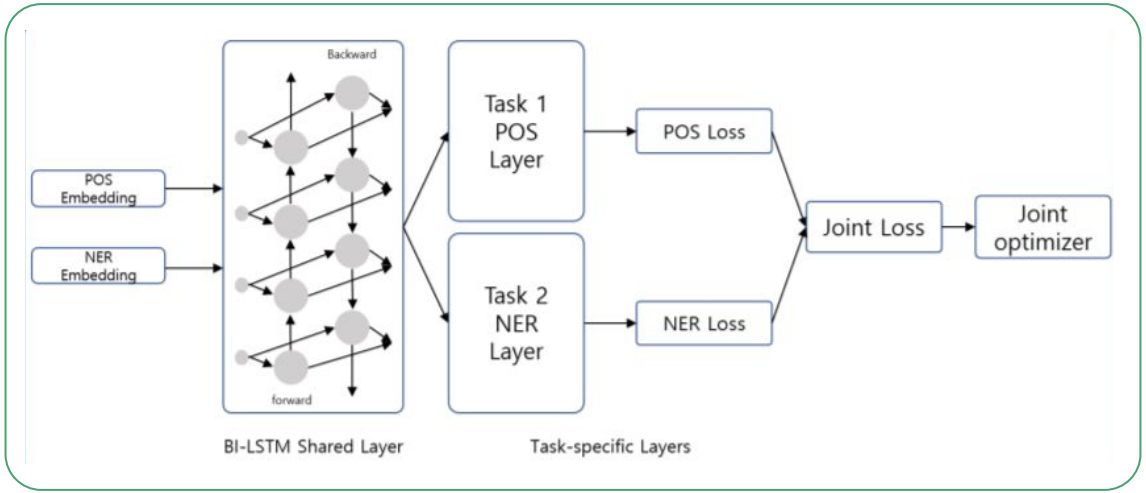
- 하지만 아직 새로운 문장 또는 단어가 DL기반 NER모델에 어떻게 추가할지에 대한 방법이 없음
  ➡ NER 성능이 외부 지식으로 향상될 수 있음을 보여주지만, 비용이 많이들고, 학습에 부정적인 영향

- RNN의 경우, 현재단계의 입력이 이전 단계의 출력을 필요로 하기 때문에 속도의 개선이 필요

- 최종적으로 사용자가 선택할 architecture 는 데이터 및 도메인 작업에 따라 다름
  - 데이터가 많을 경우 : 처음부터 RNN을 사용하여 모델을 훈련하고, 상황에 맞는 언어 모델을 fine-tuning
  - 데이터가 적을 경우 : 미리 사전 훈련된 모델이 많기 때문에 특정 영역(의료, 소셜 미디어)의 경우 상황에 맞는 언어 모델을 fine-tuning

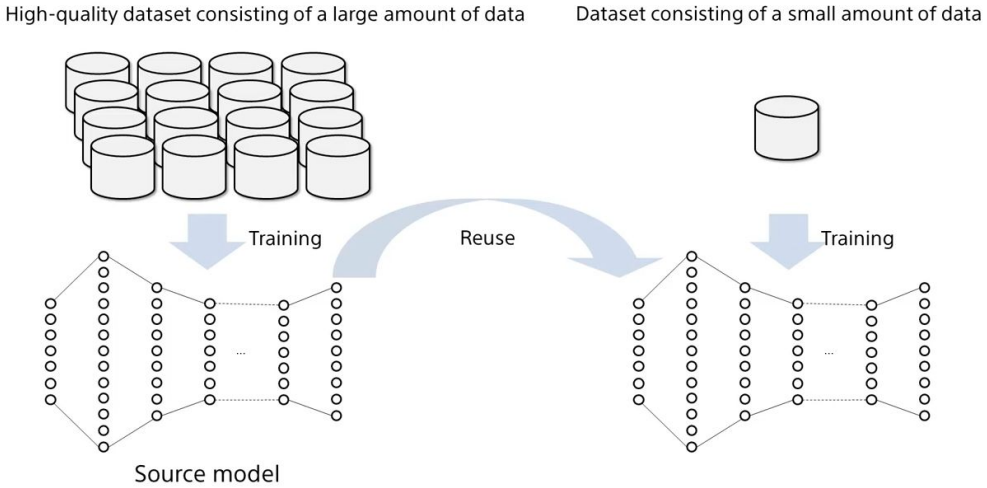
3. APPLIED DEEP LEARNING FOR NER

3.1 Deep Multi-task Learning for NER



서로 연관 있는 과제들을 동시에 학습함으로써
모든 과제 수행의 성능을 전반적으로 향상시키려는 학습 패러다임

3.2 Deep Transfer Learning for NER

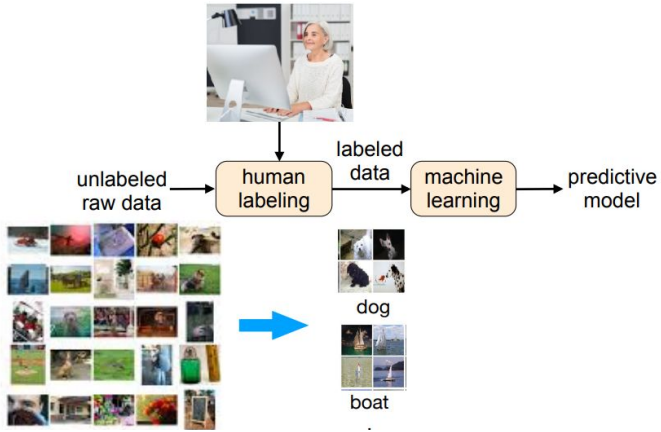


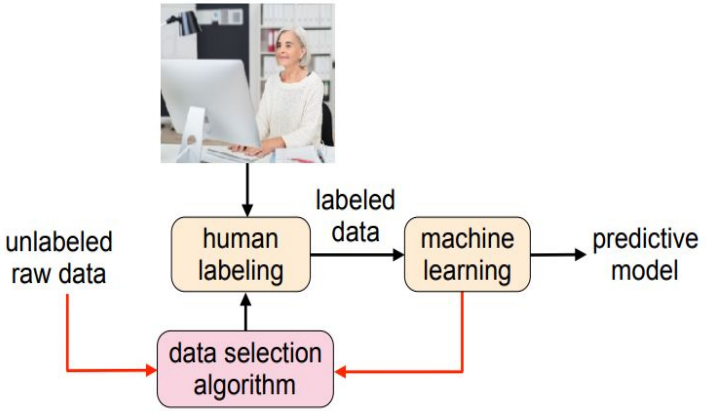
이미 학습된 정보를 이용해 새 데이터셋에 대해 보다 잘 학습하는 방법

ex) pretrained model -> fine tuning

https://www.youtube.com/watch?v=J53A4pCvCkc

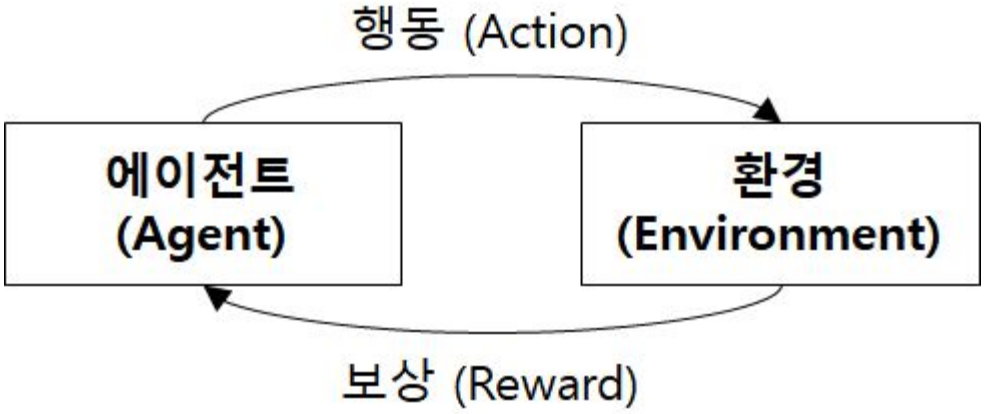
3.3 Deep Active Learning for NER





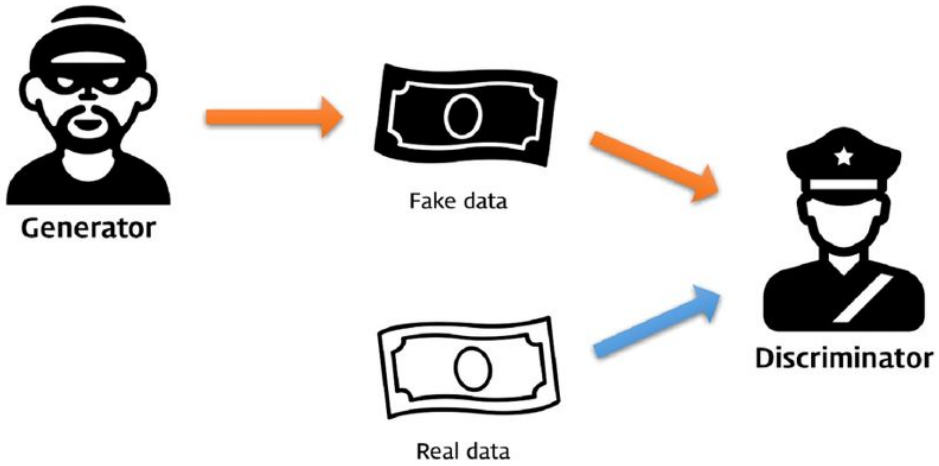
어떤 데이터가 필요한지 기계가 판단하여 사람에게 라벨링을 부탁하는 방식

3.4 Deep Reinforcement Learning for NER



어떤 행동에 대하여 학습의 누적 보상을 최대화하는 방향으로 학습하는 방법

3.5 Deep Adversarial Learning for NER



Generator는 real data와 거의 유사한 fack data를 만드는 방향

Discriminator는 real data와 fack data를 구분하기 노력하는 방향으로 학습하는 방법

https://sites.google.com/site/aidysft/generativeadversialnetwork

3.6 Neural Attention for NER

- 주어진 단어에 모든 정보를 다 활용하는 것이 아닌, 어떤 단어의 정보를 얼마나 가져올 지 결정하는 방법.
- ex)

  Rei et al. [105] applied an attention mechanism to dynamically decide how much information to use from a character- or word-level component in an end to-end NER mode

4. CHALLENGES AND FUTURE DIRECTIONS

4.1 Challenges

- **Data annotation :**
  - 많은 데이터를 위한 시간과 비용의 문제
  - 주석 작업을 수행하기 위해서는 전문가가 필요함, 아직 자원이 부족한 많은 언어와 특정 도메인이 많은 문제
  - 언어의 모호성으로 인한 주석의 품질과 일관성의 문제
    - ex) "Baltimore de feated the Yankees"라는 문장의 "Baltimore"는 MUC-7에서는 Location으로, CoNLL03에서는 Organization으로 레이블이 지정되어있음

      ➜ 하나의 명명된 엔터티에 여러 유형이 할당될 수 있는 중첩 엔터티와 세분화된 엔터티 모두에 적용할 수 있는 공통 주석 체계를 개발이 필요함

- **Information Text and Unseen Entities :**
  - 신조어, 비공식 텍스트(예: 트윗, 댓글, 커뮤니티)에 대한 낮은 정확도

    ➜ 이전에 볼 수 없었던 비형식적 텍스트를 식별하는 능력 필요

4.2 Future Directions

- Fine-grained NER and Boundary Detection
- Joint NER and Entity Linking
- DL-based NER on Informal Text with Auxiliary Resource.
- Scalability of DL-based NER.
- Deep Transfer Learning for NER
- An Easy-to-use Toolkit for DL-based NER

THANK YOU :)