방학 6주차

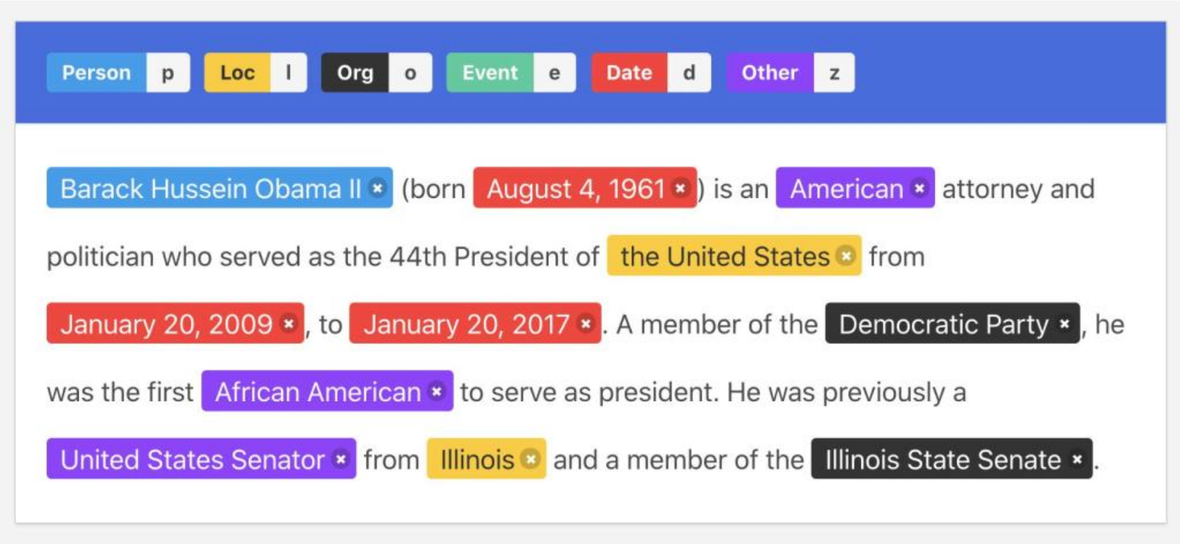
NER

NEKA

NER

NER이란 무엇인가

Named Entity Recognition의 준말로, 미리 정의해 둔 사람, 회사, 장소, 시간, 단위 등에 해당하는 단어(개체명)를 문서에서 인식하여 추출 분류하는 기법



NER

Named Entity

문자열 내에서의 기관명, 인물, 장소 뿐만 아니라 화폐, 시간, 퍼센테이지 표현까지 포괄하는 의미
ex) 철수 (인명), 서울역 (지명), 10시 (시간) 등

NER

NE의 종류 – Generic NEsG

일반적인 개체명에 해당하는 것들을 말함. 주로 NER 알고리즘을 적용함.

인물이나 장소 등의 명칭이 해당함.
ex) 삼성, 엑스포, 휴지 등

NER

NE의 종류 – Domain-specific NEs

특정 분야의 개체명에 해당하는 것들을 말함. 주로 TM(Translation Memory)을 적용함.

- TM : source-target으로 연결된 문장 pair로 구성된 DB

ex) Tetraphosphorus Decoxide, Diphosphorus Trisulfide 등

NER

NER이 필요한 이유

NER에 대한 고려 없이 TASK를 진행할 경우 원하지 않는 결과물이 나올 수 있음.

특히나 기계번역의 경우에는 번역의 품질과 사용자 경험에 적지 않은 영향을 줌.

ex) 주연이는 길을 걸었다. -> Main character walks.

NER 시스템

문장을 토큰 단위로 나누고, 토큰을 각각 태깅해서 개체명인지 분간함

BIO 시스템

- 개체명 시작은 B, 개체명 중간~끝은 I, 개체명이 아니면 O

BIESO 시스템

- 개체명 시작은 B, 개체명 중간은 I, 개체명 끝은 E
- 개체명이 하나의 토큰이면 S, 개체명이 아니면 O

Sentence #1: change in mental status and increased respiratory distress
BIO format: change/B-problem in/I-problem mental/I-problem status/I-problem and/O increased/B-problem respiratory/I-problem distress/I-problem
BIESO format: change/B-problem in/I-problem mental/I-problem status/E-problem and/O increased/B-problem respiratory/I-problem distress/E-problem

Sentence #2: white blood cell count of 24.409 and platelets of 956,000 .
BIO format: white/B-test blood/I-test cell/I-test count/I-test of/O 24.406/O and/O platelets/B-test of/O 956,000/O ./O
BIESO format: white/B-test blood/I-test cell/I-test count/E-test of/O 24.409/O and/O platelets/S-test of/O 956,000/O ./O

NER의 성능평가

정밀도, 재현율, F1-Score 등을 토큰 단위로 진행함

TABLE 3
Summary of recent works on neural NER. LSTM: long short-term memory, CNN: convolutional neural network, GRU: gated recurrent unit, LM: language model, ID-CNN: iterated dilated convolutional neural network, BRNN: bidirectional recursive neural network, MLP: multi-layer perceptron, CRF: conditional random field, Semi-CRF: Semi-markov conditional random field, FOFE: fixed-size ordinally forgetting encoding.

Work	Input representation			Context encoder	Tag decoder	Performance (F-score)
	Character	Word	Hybrid			
[94]	-	Trained on PubMed	POS	CNN	CRF	GENIA: 71.01%
[89]	-	Trained on Gigaword	-	GRU	GRU	ACE 2005: 80.00%
[95]	-	Random	-	LSTM	Pointer Network	ATIS: 96.86%
[90]	-	Trained on NYT	-	LSTM	LSTM	NYT: 49.50%
[91]	-	SENNA	Word shape	ID-CNN	CRF	CoNLL03: 90.65%; OntoNotes5.0: 86.84%
[96]	-	Google word2vec	-	LSTM	LSTM	CoNLL04: 75.0%
[100]	LSTM	-	-	LSTM	CRF	CoNLL03: 84.52%
[97]	CNN	GloVe	-	LSTM	CRF	CoNLL03: 91.21%
[105]	LSTM	Google word2vec	-	LSTM	CRF	CoNLL03: 84.09%
[19]	LSTM	SENNA	-	LSTM	CRF	CoNLL03: 90.94%
[106]	GRU	SENNA	-	GRU	CRF	CoNLL03: 90.94%
[98]	CNN	GloVe	POS	BRNN	Softmax	OntoNotes5.0: 87.21%
[107]	LSTM-LM	-	-	LSTM	CRF	CoNLL03: 93.09%; OntoNotes5.0: 89.71%
[103]	CNN-LSTM-LM	-	-	LSTM	CRF	CoNLL03: 92.22%
[17]	-	Random	POS	CNN	CRF	CoNLL03: 89.86%
[18]	-	SENNA	Spelling, n-gram, gazetteer	LSTM	CRF	CoNLL03: 90.10%
[20]	CNN	SENNA	capitalization, lexicons	LSTM	CRF	CoNLL03: 91.62%; OntoNotes5.0: 86.34%
[116]	-	-	FOFE	MLP	CRF	CoNLL03: 91.17%
[101]	LSTM	GloVe	-	LSTM	CRF	CoNLL03: 91.07%
[113]	LSTM	GloVe	Syntactic	LSTM	CRF	W-NUT17: 40.42%
[102]	CNN	SENNA	-	LSTM	Reranker	CoNLL03: 91.62%
[114]	CNN	Twitter Word2vec	POS	LSTM	CRF	W-NUT17: 41.86%
[115]	LSTM	GloVe	POS, topics	LSTM	CRF	W-NUT17: 41.81%
[118]	LSTM	GloVe	Images	LSTM	CRF	SnapCaptions: 52.4%
[109]	LSTM	SSKIP	Lexical	LSTM	CRF	CoNLL03: 91.73%; OntoNotes5.0: 87.95%
[119]	-	WordPiece	Segment, position	Transformer	Softmax	CoNLL03: 92.8%
[121]	LSTM	SENNA	-	LSTM	Softmax	CoNLL03: 91.48%
[124]	LSTM	Google Word2vec	-	LSTM	CRF	CoNLL03: 86.26%
[21]	GRU	SENNA	LM	GRU	CRF	CoNLL03: 91.93%
[126]	LSTM	GloVe	-	LSTM	CRF	CoNLL03: 91.71%
[142]	-	SENNA	POS, gazetteers	CNN	Semi-CRF	CoNLL03: 90.87%
[143]	LSTM	GloVe	-	LSTM	Semi-CRF	CoNLL03: 91.38%
[88]	CNN	Trained on Gigaword	-	LSTM	LSTM	CoNLL03: 90.69%; OntoNotes5.0: 86.15%
[110]	-	GloVe	ELMo, dependency	LSTM	CRF	CoNLL03: 92.4%; OntoNotes5.0: 89.88%
[108]	CNN	GloVe	ELMo, gazetteers	LSTM	Semi-CRF	CoNLL03: 92.75%; OntoNotes5.0: 89.94%
[133]	LSTM	GloVe	ELMo, POS	LSTM	Softmax	CoNLL03: 92.28%
[137]	-	-	BERT	-	Softmax	CoNLL03: 93.04%; OntoNotes5.0: 91.11%
[138]	-	-	BERT	-	Softmax +Dice Loss	CoNLL03: 93.33%; **OntoNotes5.0: 92.07%**
[134]	LSTM	GloVe	BERT, document-level embeddings	LSTM	CRF	CoNLL03: 93.37%; OntoNotes5.0: 90.3%
[135]	CNN	GloVe	BERT, global embeddings	GRU	GRU	CoNLL03: 93.47%
[132]	CNN	-	Cloze-style LM embeddings	LSTM	CRF	**CoNLL03: 93.5%**
[136]	-	GloVe	Plooled contextual embeddings	RNN	CRF	CoNLL03: 93.47%

NER의 접근 방식

NER에는 크게 3가지 접근 방식이 있음

(1) 규칙 기반 접근 (Rule-based Approaches)

(2) 비지도 학습 접근 (Unsupervised Learning Approaches)

(3) 변수 기반 지도 학습 접근 (Feature-based Supervised Learning Approaches)

NER 접근 방식 1 - 규칙 기반 접근

사전(Gazetteer)이나 패턴을 적용해서 접근함

정확도는 높으나 재현율이 낮고, 다른 도메인에 잘 맞지 않음

- Gazetteer : Geographical Dictionary로, 지명이 담긴 사전처럼 인명에 대한 dictionary라고 할 수 있음

NER 접근 방식 2 - 비지도 학습 접근

문맥적 유사도에 기반한 클러스터링을 통해 접근함

사전을 제작할 때 비지도 학습을 이용하는 방식으로 통계적 정보나 얇은 수준의 통사적 지식에 의존함

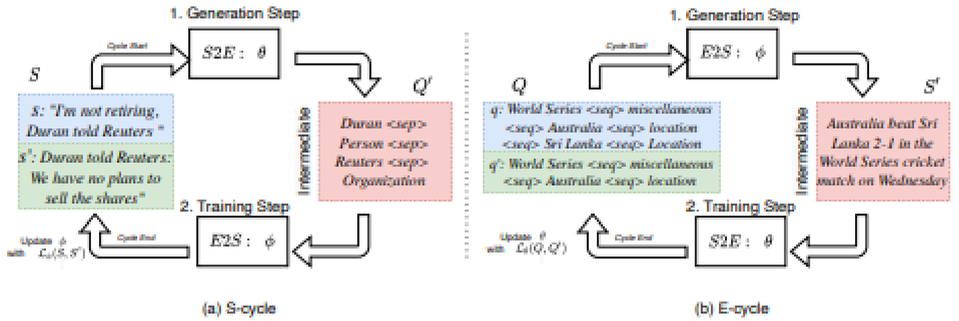


Figure 3: Training cycles and steps in CycleNER.

NER 접근 방식 3 - 변수 기반 지도 학습 접근

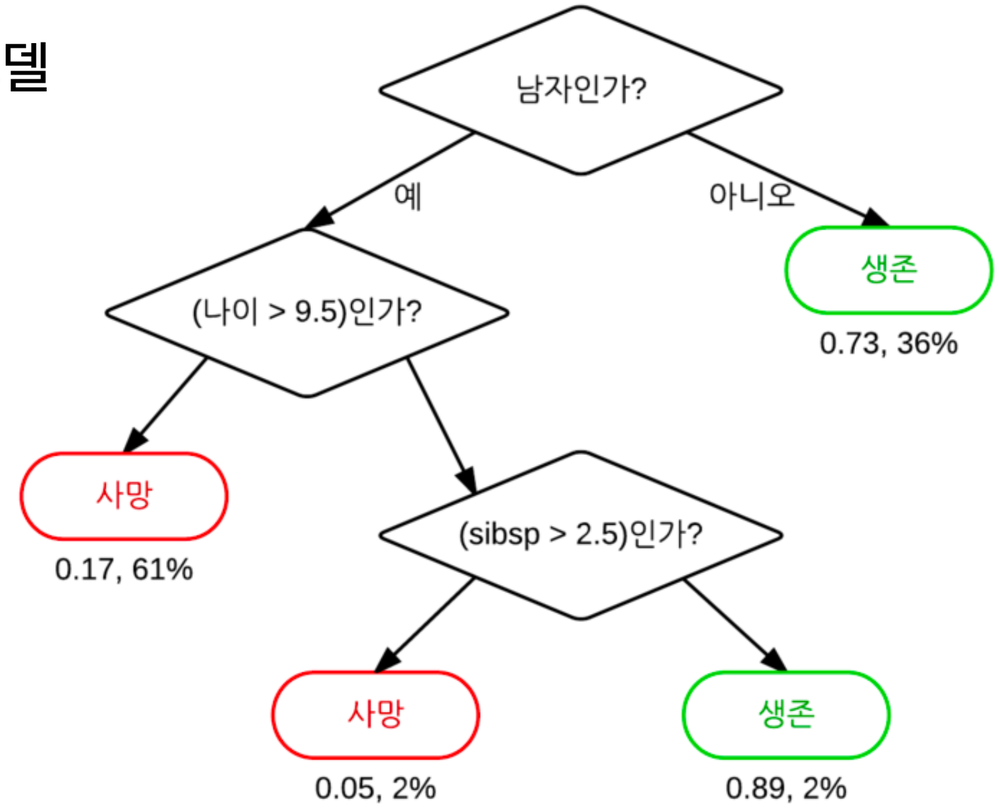
다중 클래스 분류나 시퀀스 레이블링 작업을 통해 접근함

HMM, Decision Tree, Maximum Entropy Model, SVM, CRF 등 다양한 모델을 사용함.

NER 접근 방식 3 - 변수 기반 지도 학습 접근

Decision Tree

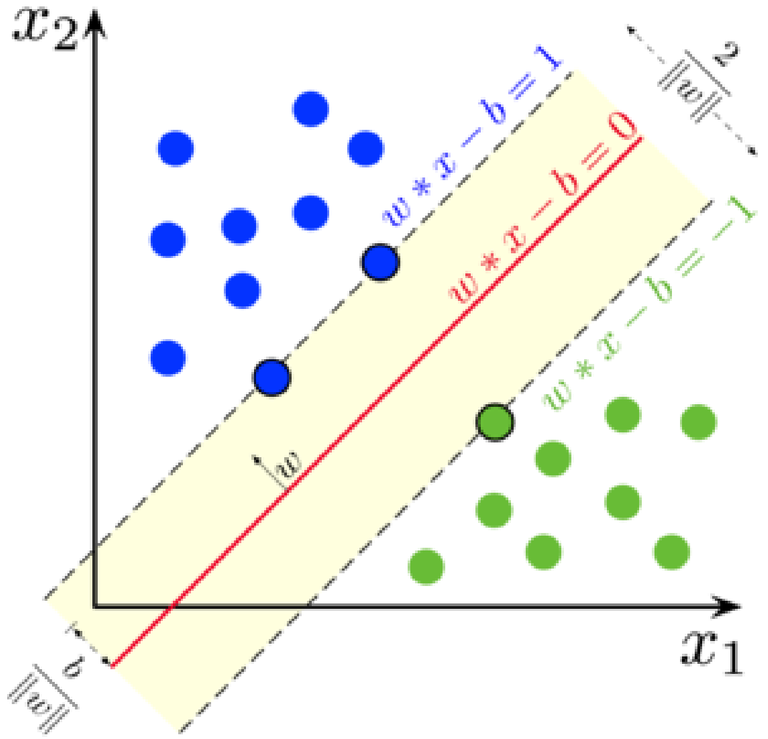
일련의 분류 규칙을 통해 데이터를 분류하는 모델



NER 접근 방식 3 - 변수 기반 지도 학습 접근

Soft Vector Machine

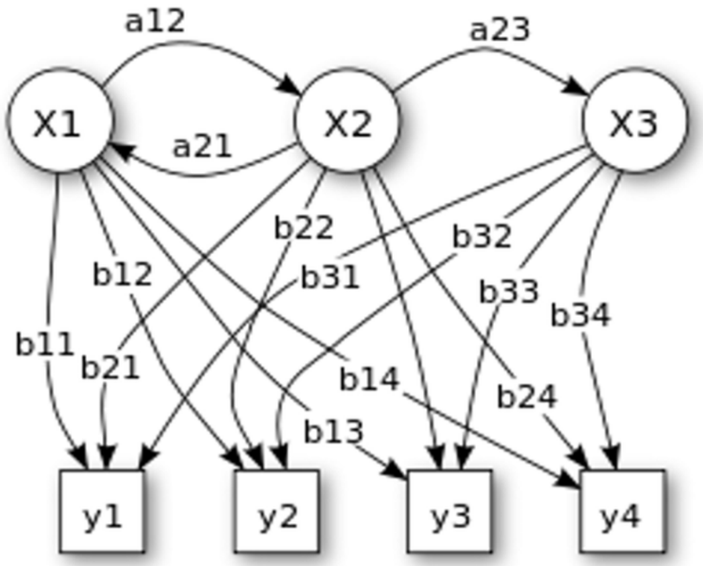
이진 선형분류 모델을 만들고 그 중에서 가장 폭이 긴 경계를 만드는 모델을 찾기



NER 접근 방식 3 - 변수 기반 지도 학습 접근

HMM

바로 직전에 변화된 상태를 기준으로 다음 상태를 추론하기

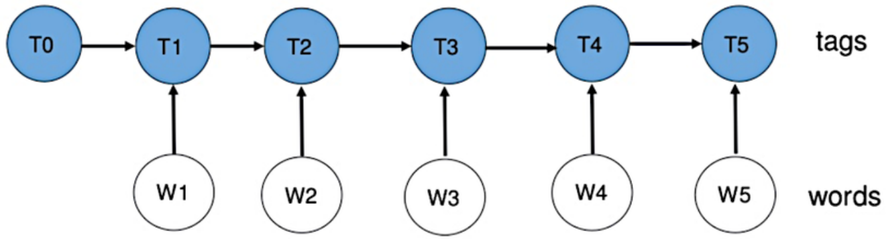


NER 접근 방식 3 - 변수 기반 지도 학습 접근

Maximum Entropy Model

다항 로지스틱 회귀라고도 함.

미리 정의된 제한조건을 만족하고 그 외의 값은 동일하게 함.



Discriminative model, model conditional probability Pr(T | W) directly.

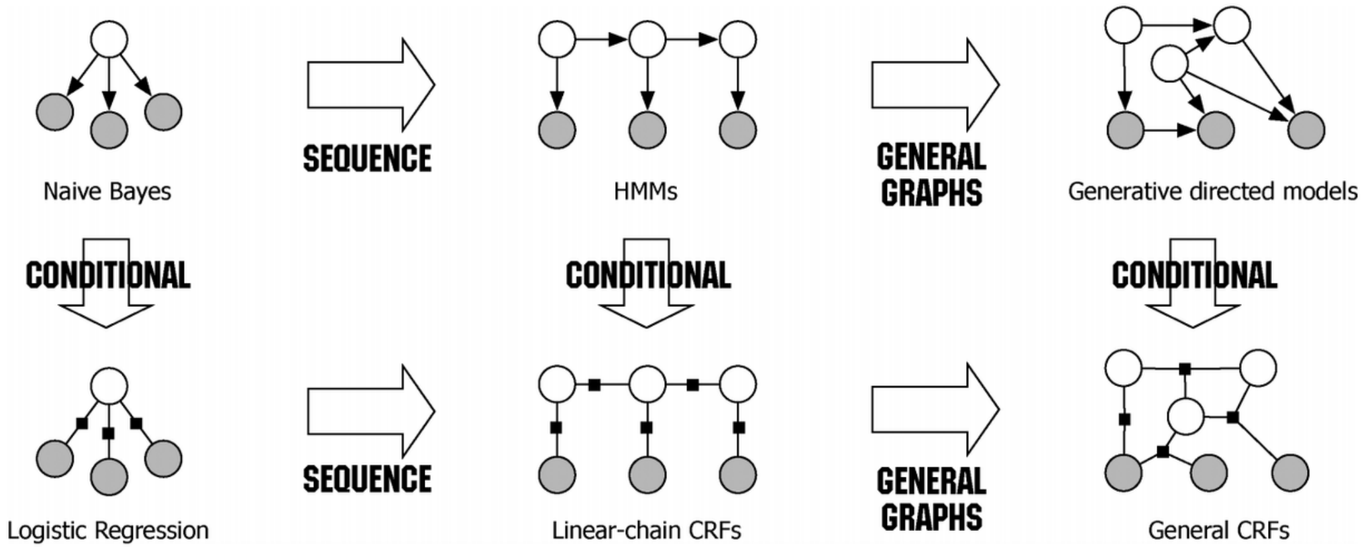
$$\Pr(\boldsymbol{T}|\boldsymbol{W}) = \prod_{i=1}^{L} \Pr(t_i|t_{i-1}, w_i) = \prod_{i=1}^{L} \frac{\exp(\sum_j \beta_j f_j(t_{i-1}, w_i))}{Z(t_{i-1}, w_i)}$$

t_0 is a dummy start state.

NER 접근 방식 3 - 변수 기반 지도 학습 접근

Conditional Random Fields

가능성 있는 후보를 몇개 선택하고 그 중에서 가장 적합한 하나를 고르기



NER 데이터셋

NER 데이터셋은 데아터를 얼마나 자세하게 분류하냐에 따라 크게 3가지가 있음

(1) 4가지 분류 방식 : 기관명, 인명, 제품명, 저작물명

(2) TTA 대분류 : 인공물, 짐승, 문화, 기간, 사건, 분야, 지역, 원소, 기관, ...

(3) TTA 소분류 : 사람 이름, 캐릭터 이름, 동물 이름, 과학 분야, 사회과학 분야, ...

과제

오늘 배운 개념을 바탕으로
간단한 NER 모델 구축하기

NEKA

THANK YOU