

S2W 데이터셋 학습



20151322 이한용

학습 세팅

- 모델
 - Flair 에서 제공하는 ner tagger model을 fine-tuning 함 (<https://huggingface.co/flair/ner-english>)
 - Flair embeddings 와 LSTM-CRF을 사용한 모델
- dataset split
 - corpus에 중복된 문장이 몇 개 포함되어 있었음 (51쌍). 해당 문장의 중복을 제거
 - corpus는 문장들을 Python의 random.shuffle() 으로 섞은 다음에 train, dev, test을 약 4:1:1의 비율로 나눔
 - 따라서 4908 train + 1232 dev + 1227 test sentences 으로 구성되었음. (121172 train + 30852 dev + 30903 test tokens)

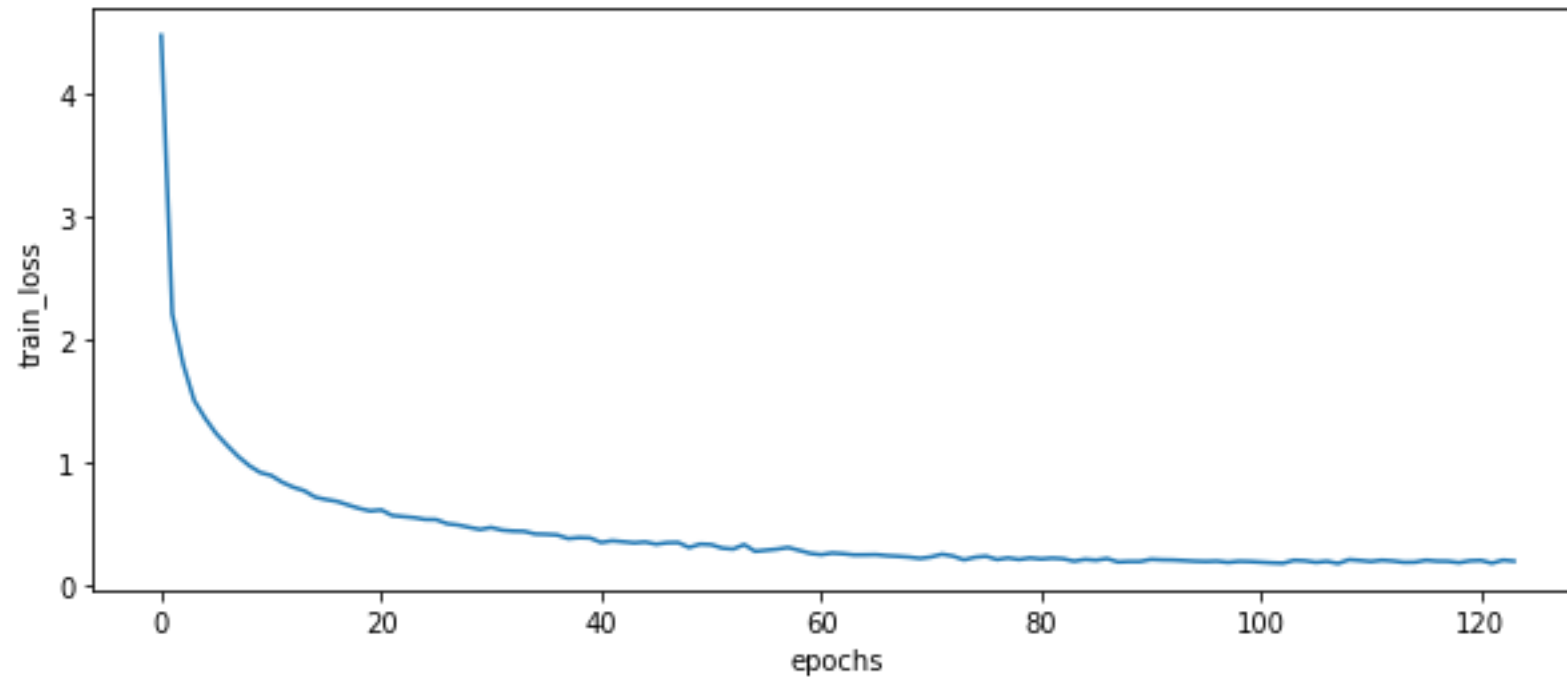
학습 세팅

- 학습 세팅
 - Flair 의 ModelTrainer 를 이용
 - Parameters:
 - learning_rate: "0.1"
 - mini_batch_size: "32"
 - patience: "3"
 - anneal_factor: "0.5"
 - max_epochs: "150"
 - shuffle: "True"
 - train_with_dev: "True"
 - batch_growth_annealing: "False"

학습 결과 – F1-score

- Results:
 - F1-score (micro) 0.9692
 - F1-score (macro) 0.9591
- By class:
 - LOC tp: 48 - fp: 0 - fn: 0 - precision: 1.0000 - recall: 1.0000 - f1-score: 1.0000
 - MAL tp: 495 - fp: 10 - fn: 13 - precision: 0.9802 - recall: 0.9744 - f1-score: 0.9773
 - MISC tp: 484 - fp: 27 - fn: 30 - precision: 0.9472 - recall: 0.9416 - f1-score: 0.9444
 - ORG tp: 181 - fp: 6 - fn: 17 - precision: 0.9679 - recall: 0.9141 - f1-score: 0.9403
 - PER tp: 20 - fp: 3 - fn: 1 - precision: 0.8696 - recall: 0.9524 - f1-score: 0.9091
 - RAN tp: 942 - fp: 24 - fn: 7 - precision: 0.9752 - recall: 0.9926 - f1-score: 0.9838
- loss:
 - 0.6280

학습 결과 – train loss curve



학습 결과 – error sample

dev error

index	token	truth	predicted
27	TrueCrypt	B-MISC	B-RAN
149	police	B-MAL	O
150	virus	I-MAL	O
152	Reveton,which	O	B-RAN
328	GOST	O	B-MISC
397	Sandboxie	B-MISC	B-RAN
1001	Android	B-MISC	O
1039	Cowboys	O	B-MISC
1168	Opera	B-MISC	B-ORG
1170	Yandex	B-MISC	B-ORG
1178	Comodo	B-MISC	B-ORG
1180	Chromium	B-MISC	B-ORG
1182	DynDNS	B-MISC	B-ORG
1184	Filezilla	B-MISC	B-ORG
1186	FlashFXP	B-MISC	B-ORG
1188	Outlook	B-MISC	B-ORG
1190	Netscape	B-MISC	B-ORG
1215	DOCX	O	B-MISC
1250	Themida	B-MISC	B-RAN
1419	Windows	O	B-MISC
1601	Horse	O	B-MAL
1616	Troy	B-PER	B-RAN
1769	EITest	B-MISC	B-MAL
2039	Backdoor.TeamViewer	B-MAL	B-MISC

test error

index	token	truth	predicted
53	bleepingcomputer	B-ORG	O
567	EternalBlue	B-MISC	B-RAN
580	DoublePulsar	B-MISC	B-RAN
801	SetFileAttributes	O	B-MISC
1157	System	O	I-MISC
1840	Sodinokibi	B-RAN	B-PER
1842	Unknown	O	B-ORG
2004	BitcoinBlackmailer.exe	O	B-MISC
2659	tuyuljahat	B-ORG	B-RAN
3567	Malwarebytes	B-MISC	B-PER
4109	Internet	O	B-MISC
4110	of	O	I-MISC
4111	Things	O	I-MISC
4377	LockerPin.A	B-MISC	B-RAN
4436	FileCrypt	B-MISC	B-RAN
4746	Safe	O	B-MISC
4747	Mode	O	I-MISC
4976	Ransomware	O	B-RAN
5255	EmoCheck	B-MISC	B-MAL
5652	SQLite	B-MISC	O
5661	Java	B-MISC	I-MISC
6283	IAALIXR	B-ORG	O
6829	XAgentOSX	B-MAL	B-MISC
7388	Malwarebytes	B-MISC	B-ORG

학습 결과 - 분석

- PER tag의 F1-score가 낮게 나왔는데, 이것은 PER으로 tagging 된 token의 수가 적기 때문으로 판단된다.
- 회사와 제품 이름이 같은 경우에는 회사는 ORG, 제품 이름은 MISC 로 tagging했다. 따라서 같은 token이라도 다르게 tagging된 경우가 좀 있었는데, 이 때문에 모델의 학습에 혼동이 있었을 것으로 생각된다.