

BIO/BIOES tagging

- BIO tagging(또는 IOB tagging)은 beginning, inside, outside의 줄임 말로 전산언어학에서 chunking task 안의 토큰들을 태깅하는 공통적인 포맷이다.
- 의미상으로 같은 chunk에 속한 맨 앞 token에는 B 접두어를 붙이고, 같은 chunk에 속한 두번째 token 부터는 I 접두어를 붙인다. name entity가 아닌 token은 O로 태깅한다.
- “B-PER, I-LOC”과 같이 종류를 나타내는 태그 앞에 위치에 따라 접두어로 붙는다.
- BIOES는 BIO 태그를 확장한 포맷이다.
- chunk에 하나의 token만 있으면 S 접두어를 붙이고, 여러개의 token으로 이루어진 chunk에서는 시작 token에 B를, 시작과 마지막을 제외한 중간 token에는 I를, 마지막 token에는 E를 붙인다.

BIO/BIOES tagging 예시

```
Alex B-PER  
is O  
going O  
to O  
Los B-LOC  
Angeles I-LOC  
in O  
California B-LOC
```

BIO tagging

```
Alex S-PER  
is O  
going O  
with O  
Marty B-PER  
A. I-PER  
Rick E-PER  
to O  
Los B-LOC  
Angeles E-LOC
```

BIOES tagging

S2W 데이터셋 준비

ner tag 목록

- Flair 에서 제공하는 ner tagger의 태그 목록은 다음과 같다. (<https://huggingface.co/flair/ner-english>)
- 접두어로는 BIOES 규칙을 사용한다.

tag	meaning
PER	person name
LOC	location name
ORG	organization name
MISC	other name

추가한 tag 목록

- RAN : ransomware

tagging 예시

Simplelocker	S-REN	targets	O
,	O	Android	B-MISC
also	O	OS	E-MISC
known	O		
as	O		
AndrB-REN			
/	I-REN		
Slocker-A	E-REN		
,	O		
which	O		
is	O		
Tor-enabled	O		
mobile	O		
device	O		
ransomware	O		
,	O		

데이터셋 준비중 막히는 부분

- hand-annotate 한 corpus로 학습시켜도 새로 annotation한 token의 tagging confidence가 커지지 않음.
오히려 감소함
 - Embedding 을 변경하여 문제 해결 (glove만 사용 -> glove + forward FlairEmbedding + backward FlairEmbedding을 혼합한 StackedEmbeddings 사용)
- ner tagging 문제
 - “MISC” (miscellaneous) 으로 tagging하는 token의 범위가 애매함. (e.g., “P2P”, “AES”, “RaaS”, “Javascript”, “Bitcoins”, “SMBv1”, “C++” 등등 과 같은 기술 명은 named entity로 분류해야 하는가?)
 - “product” tag를 추가할지 여부에 대한 고민. (초벌 tagging에 사용한 ner tagger에는 LOC, PER, ORG, MISC의 4개 tag만 제공하기 때문에 “Android”, “Chrome”등의 product들은 모두 “MISC” tag로 분류됨. 제품명에 대한 분류가 필요하다면 해당 tag를 추가해야 함)

데이터셋 준비중 막히는 부분

- ner tagging 문제
 - “Mal/MSIL-BA” “Win32:Malware-gen”과 같은 경우는 이름 부분만 tagging해야 하는가? 아니면 전부 다 tagging해야 하는가?
- text editor로 일일이 tag를 수정하는 방법이 번거로웠다.
 - tag을 수정할 수 있는 gui 를 개발하여 사용하면 작업이 편하고 속도도 빨라질 것이다.
- exploit kit 의 이름도 많이 등장하는데, 이에 대한 tag도 따로 지정해야 하는지 여부 (e.g., Fallout1, Spelevo, EternalBlue, ...)
 - 현재는 “MISC” tag로 분류함

데이터셋 준비중 막히는 부분

- TrueCrypt 관련
 - TrueCrypt ransomware와 TrueCrypt open-source software에 대한 corpus가 혼재함