



RAN tag 결정 스코어 임계값 최적화 - 1



20151322 이한용

설명

- 태깅 모델은 새로 발견된 단어에 대해 낮은 스코어로 그 레이블을 예측한다.
- 랜섬웨어 이름은 새로운 단어인 경우가 많으므로 단어의 예측 스코어가 낮은 경우에 대해 랜섬웨어를 지칭하는 키워드가 근처에 있을 때 랜섬웨어로 태깅하는 방법을 사용할 것이다.
- 예측 스코어가 낮음을 판별하는 기준을 ‘스코어 임계값’ s_{θ} 으로 정의한다.
- 성능을 극대화 하는 s_{θ} 의 최적값을 탐색한다.

기호 정의

- s_θ : 스코어 임계 값
 - $\hat{L}(s_\theta)$: s_θ 에서의 레이블
 - L : 참값 레이블
 - n : 토큰의 개수
 - m : 레이블의 개수
-  $n \times m$ matrix
row : 레이블의 one-hot encoding vector

식 전개

1. $\operatorname{argmax}_{s_\theta} \operatorname{F1score}(s_\theta) \rightarrow \operatorname{argmin}_{s_\theta} \operatorname{MSE}(s_\theta)$
2. MSE (Mean Squared Error)를 다음과 같이 정의한다.

$$\operatorname{MSE}(s_\theta) := \frac{1}{n} \sum_{i=1}^n (L_i - \hat{L}(s_\theta)_i)^2 := \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^m (L_{ij} - \hat{L}(s_\theta)_{ij}) \right)^2$$

3. $L_{ij} \in \{0,1\}$

$$\hat{L}_{ij} \in \{0,1\}$$

4. 식 2에서 $\sum_{j=1}^m (L_{ij} - \hat{L}(s_\theta)_{ij})$ 부분을 $\Sigma(s_\theta)_i$ 으로 치환한다. 여기서 $\Sigma(s_\theta)_i$ 는 다음을 만족한다.

$$\Sigma(s_\theta)_i \in \{0,1\} \quad \because \sum_{j=1}^m (L_{ij} - \hat{L}(s_\theta)_{ij}) = \begin{cases} 0, & L_i = \hat{L}(s_\theta)_i \\ 1, & \text{otherwise} \end{cases}$$

식 전개

5. 식 2를 미분한다.

$$\frac{d}{ds_{\theta}} MSE(s_{\theta}) = \frac{d}{ds_{\theta}} \left(\frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^m (L_{ij} - \hat{L}(s_{\theta})_{ij}) \right)^2 \right) = \frac{2}{n} \sum_{i=1}^n \Sigma(s_{\theta})_i \frac{d\Sigma(s_{\theta})_i}{ds_{\theta}}$$

6. $\Sigma(s_{\theta})_i$ 는 미분 불가능 하므로 $\frac{d\Sigma(s_{\theta})_i}{ds_{\theta}}$ 가 성립하지 않는 오류가 있다.

계획 변경

- loss 함수의 미분값을 구할 수 없는 문제가 발생.
- s_θ 는 0~1 사이의 값을 가지므로 탐색 공간이 크지 않음.
- 따라서 s_θ 에 0부터 1까지 작은 값(e.g., 0.01)을 더해가며 F1-score를 매번 계산하여 최고의 F1-score를 탐색하는 방법을 사용함.