Term Project Proposal

Machine Learning과 Deep Learning을 활용한 개체명 인식

Sung Jae Hyuk
Department of Computer Science
Student ID: 2019320100
okaybody10@korea.ac.kr

Shin Seung Heon
Department of Computer Science
Student ID: 2016190329
heon4077@naver.com

Choe Min Seok
Department of Computer Science
Student ID: 2021320092
goro8pyo@korea.ac.kr

2023, Fall Semester

Course: Machine Learning(COSE362(02))

1 Objective

Named Entity Recognition(NER, 이하 개체명 인식)은 문장 내에서 의미를 가지는 개체들을 인식하는 것을 의미한다. 즉, 문장 속에서 각 단어가 어떤 태그에 속하게 되는지를 구분하는 문제이다. 한국어 기준 15종류의 명시된 개체명 가지고 있으며, 세분류로 150개의 개체명을 가지고 있다.[1] NER은 Word-Level에서 품사 태깅과 이루어지는 가장 기초적인 작업이며, 정보 추출(IE)을 위한 독립형 도구로 사용될 뿐만 아니라 텍스트 이해, 정보 검색, 자동 텍스트 요약, 질문 답변, 기계 번역, 지식 기반 구축 등과 같은 다양한 자연어 처리(NLP) 애플리케이션에서 필수적인 역할을 담당하고 있다.[2] 개체명 인식 작업을 수행하는 방법으로는 크게 Rule-Based, Feature-Based, Unsupervised-Based로 3가지가 존재하고, 이 이번 프로젝트에서는 개체명 분석 Task에서 문장의 문맥을 고려하는 것이 얼마나 많은 도움을 주는지, 유의미한 성능 발전이 있는지 알아보고자 한다. 또한 기존 모델을 단순히 활용하는 것과 추가로 Transfer Learning을 진행하였을 때의 정확률 차이에 대해서도 확인해 보고자 한다.

2 Datasets

현재 Public하게 제공하는 Korean Ner Dataset은 세 가지가 존재한다.

- **국립국어원 NER Dataset**
- 한국해양대학교 자연어처리 연구실 NER 데이터 셋
- Naver NLP Challenge 2018

이 중 첫번째 국립국어원에서 제공한 “개체명 분석 말뭉치 2022” Dataset을 활용하여 비교 예정이다.

3 Importance

개채명 인식 같은 경우 비정형화 되어있는 글에서 우리가 사용해야하는 몇몇 정형화 값들을 뽑아낼 수 있다는 점에서 아주 중요하다. 글 같은 경우 문맥에서 오는 정보들도 있지만 단어 자체로도 의미를 가지고 그것이 글이 말하고자하는 바에 직접적으로 기여하는 경우도 많다. 이러한 상황에서 개체명 인식을 통해 정보 추출, 기계 번역이나 Question Answering 과 같은 downstream NLP Task 외에도 특정 분야(*e.g.*, 의료, 금융, 법률) 등과 같이 단어 자체가 어렵고 의미를 많이

가지고 있는 분야에서는 아주 중요하게 여겨진다.
이런 단어들을 뽑아낼 때, 과연 문맥이 실제로 어느 정도의 영향을 끼치는지가 우리가 현재 진행할 프로젝트이다. 특히, 한국어 같은 경우 뜻이 정말 많거나 어쩔때는 기관으로 사용되는 것이 어떨 때는 다른 의미로 사용되기도 하는 경우가 빈번히 일어난다. 이는 일반적으로 문장 전체를 보고 판단하게 되는데, 딥러닝에서 이러한 작업을 수행하기 위해서는 더 많은 비용을 지불해야한다.
그렇기에 실제로 이걸 반영하였을 때 어느정도의 퍼포먼스가 차이가 나고, 더 많은 비용을 지불하면서까지 유의미한 결과를 얻을 수 있는지 분석하는 것이 프로젝트의 목표이다.

4 Approach

4.1 Margin Theory & Support Vector Machine(SVM)

분류 문제를 해결할 때 단순히 현재의 Dataset에 맞는 기준을 찾는 것 외에도 현재의 기준에서 다른 데이터가 들어왔을 때 실제로 정확한 라벨을 얼마나 잘 뱉는지도 중요하다. 이는 레이블들 간의 영역이 잘 나누어지고 그 영역이 크면 클수록 정확하게 분류할 확률이 높아진다. 먼저, Dataset $\mathcal{D} = \{(x_i, y_i)_{i=1}^n\}$가 완전히 분리된다고 가정하고 하나의 가설 h_θ에 대해 margin function $\text{margin}(x) := yh_\theta(x)$로 정의되고, dataset에 대해 최소 마진은 $\gamma_{min} := \min_{i\in\{1,\cdots,|\mathcal{D}|\}} y^{(i)}h_\theta(x^{(i)})$로 정의한다.[3] 즉, 분류기가 기준한 것들 중에서 데이터셋으로부터 가장 가깝게 붙어있는 데이터와 분류 기준까지의 거리를 의미한다.
γ_{min}을 최대화하는 Task를 수행을 하는 것이 Support Vector Machine이며, 현재 같은 경우 정확하게 fitting이 되므로 Hard Margin Problem으로 바뀌게 된다. Kernel Trick을 활용하지 않는 Linear한 모델 같은 경우 Duality를 이용하여 Convex Optimization Problem으로 바꾸고, KKT Condition으로부터 문제를 해결할 수 있다.
허나 Dataset이 완전히 쪼개지지 않거나 Margin이 너무 작아지게만 separable 되는 경우엔 예외를 조금 허용하는 것이 좋다. 이를 Soft Margin Problem이라 한다. 기존의 Objective function은 동일하게 유지하고, Constraint에 어느정도의 허용 범위를 추가해주고, Duality를 이용하여 동일하게 Convex Optimization Problem으로 바꿀 수 있고 KKT Condition 으로부터 문제를 해결할 수 있다.
해당 프로젝트에서는 ① Trick을 활용하여 비선형으로도 문제를 접근할 수 있고 ② 여러 개의 Class를 분류해야하는 문제이므로 SVM을 사용한다.

4.2 CRF(Conditional Random Field)

CRF(Conditional Random Field)는 패턴 인식을 할 때 자주 사용되는 Mrkov 기반의 모델이다. 그 주변의 데이터들을 반영하지 않고 단독적으로 결정하는 다른 classifier과 다르게 주변의 문맥을 고려하여 결과를 정하게 된다. 원래라면 개체의 주변을 인식하기 위해서 그래프 모델을 사용하여 Edge가 종속성을 가지고 있음을 두어 Graph Embedding을 수행하며, 현재와 같이 문장이 주어져있는 경우 나와 그 직전의 토큰들을 주변으로 인식하게 함으로써 확률을 계산할 수 있다. 이 기법을 Maximum Entroy Markov Model이라고 하며, Tag vector $\mathbf{T}$, word vector $\mathbf{V}$에 대해 다음과 같이 확률이 계산된다.[4]

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

다만, 주변만 보면 다소 편향적인 정보만을 활용하여 labeling을 할 수 있어서 정확률이 다소 떨어질 수 있는 Label Bias Problem이 일어날 수 있다. 이 문제를 해결하기 위해 순차적으로 이전의 라벨을 들고 오는 것이 아니라 양방향에서 데이터를 들고올 필요가 있다. 즉, Directed Graph를 활용했던 MEMM에서 이를 undirected graph로 바꾸어서 문제를

해결한다.[5] 이를 Conditional Random Field라 하며, 위와 동일한 상황에서 확률은 다음과 같다.[4]

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

해당 프로젝트에서는 문맥이 실제로 어떠한 영향을 미치는가에 대한 분석이기 때문에 Attnetion 기법을 활용하지 않는 선에서 주변 단어들간의 문맥을 가장 잘 활용하는 CRF 모델을 사용하기로 결정하였다.

5 Timeline

이후 예정된 일정은 다음과 같다.

- (~ 10/4) 국립국어원 데이터 신청 및 데이터 양식 확인
- (~ 10/9) 데이터 전처리
  - Naver측에서 제공하는 Dataset과 다르게 프로젝트에 사용할 Dataset에서는 BIO Tagging을 사용하고 있지 않음
  - 따라서 Tokenizer를 거쳤을 때 나온 결과물에 대한 Labeling을 BIO-Tagging으로, 세분류를 모두 제거하고 대분류로만 바꿔줄 필요가 있으므로 이에 대한 pre-processing을 진행해야함
- (~ 10/12) [First Compare] 단순 Support Vector Machine과 CRF 모델 훈련 및 Inference
- (~ 10/19) [Second Compare] BiLSTM Model 설계 및 SVM 및 CRF 모델 훈련 및 Inference
- (~ 10/23) [Third Compare] Pretrained된 Bert Model을 이용한 SVM / CRF Transfer Learning

6 Expected Result

양방향에서의 모든 문맥을 고려하는 BiLSTM과 Bert Model에서는 Support Vector Machine에서는 단어와의 관계에 주목하지 않고 오직 단어 하나만 주목하는 반면, CRF에서는 문장 내에서 단어들간의 관계를 모두 고려하기 때문에 CRF에 대한 정답률이 유의미한 수준으로 높을 것으로 예상된다.
다만, SVM과 CRF 단일 모델에서는 어느 모델이 더 좋다고 예상하기는 힘들며 만약 차이가 나더라도 크게 유의미한 차이는 나지 않을 것으로 판단된다. (단일 모델에서는 문장 내에서의 관계나 문맥을 파악하는데 많은 한계점이 존재할 것이고, 정확하게 캐치하기에는 너무 모델들이 단순하여 underfitting의 효과가 날 것으로 예상됨)
추가적으로 단순 모델보다는 BiLSTM이, BiLSTM보다는 Bert에서 전체적으로 성능이 더 좋을 것으로 기대된다.

7 References

[1] National Institue of Korean Language (2023). Nikl named entity corpus 2022 (v.1.0).

[2] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, 2020.

[3] Tengyu Ma. Lecture notes for machine learning theory (cs229m/stats214), June 2022.

[4] Weiwei Guo, Huiji Gao, Jun Shi, and Bo Long. Deep natural language processing for search systems (sigir 2019 tutorial). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 26–27. Association for Computing Machinery, 2019.

[5] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.