머신러닝

최신기술 특론

term project 제안서

M2017517 조영훈

20153204 이가영

목차

1. 프로젝트 개요

자연어 처리 분야의 Part-Of-Speech tagging (이하 pos tagging)은 Rule-Base 혹은 Statistical 방식을 이용한다. 보통 새로운 규칙에 대해 적용이 안되는 Rule-Base 방식보다 기존의 데이터를 통해서 규칙을 찾아내서 tagging을 하는 Statistical 방식을 많이 사용한다. 이 중에서도 Hidden Markov Model(이하 HMM)을 많이 사용한다.

과거에 비해 최근에 PC의 성능이 좋아져서 머신러닝 중 딥러닝의 성능이 많이 상향된 것을 기존의 많은 실험을 통해서 알 수 있다. 그래서 기존의 pos tagging에 사용되는 통계학적 방식인 HMM을 최근에 통계학적 모델로 많이 사용이 되는 딥러닝을 이용한다면 더 좋은 성능이 나올 수 있다고 생각을 하였다.

2. 프로젝트 내용

2.1 학습 Model

pos tagging을 더 효율적으로 학습하기 위해서 딥러닝의 많은 모델 중에서 적합한 model을 찾던 중에 sequence data에 대해서 성능이 좋은 Recurrent Neural Networks(RNN)의 한 종류인 Long Short-Term Memory(이하 lstm)을 사용하기로 했다. 그 이유는 pos tagging은 한 단어만 중요한 것이 아니라 앞 뒤의 단어가 중요하다. 예를 들자면 "나는" 이라는 단어를 봤을 때, 후보 군은 여러가지가 나올 수 있다.

- 자신을 지칭하는 인칭 대명사.
- 날고있다는 동사.

하지만 "나는" 이라는 단어의 앞 뒤의 내용이 있으면 구분이 쉽다.

- 나는 김철수이다 -> 인칭 대명사
- 하늘을 나는 비행기 -> 동사

이와 같이 pos tagging을 할 때 한 단어만 사용되는 것이 아니라 앞 뒤의 순차적인 데이터를 필요로 한다. 그러므로 딥러닝 모델 중 lstm이 적합하다고 생각한다. Sequence data를 앞에서 뒤로, 뒤에서 앞으로 두 방향으로 학습을 하는 Bidirectional Long Short-Term Memory(이하 blstm)을 이용을 한다.

2.2 Data 형식

딥러닝에 사용 될 data는 tagged corpus를 사용할 것이다. 여러 문장이 특수 기호로서 나뉘

어져 있다. 문장안의 각 단어는 품사가 tag가 되어 있다. 이 tagged corpus(Figure 1)는 세종말뭉치(출처: 국립국어원)에서 수집한다. 이 세종말뭉치의 tag의 수는 총 45개이다.

```
남미풍의    남미/NNP + 풍/XSN + 의/JKG
강렬한  강렬/XR + 하/XSA + ㄴ/ETM
원색끼리의  원색/NNG + 끼리/XSN + 의/JKG
조화,   조화/NNG + ,/SP
수채화  수채화/NNG
같이    같이/MAG
안온한  안온/NNG + 하/XSA + ㄴ/ETM
배색    배색/NNG
등   등/NNB
색의    색/NNG + 의/JKG
분위기를    분위기/NNG + 를/JKO
강조하는    강조/NNG + 하/XSV + 는/ETM
기하학적    기하학/NNG + 적/XSN
무늬,   무늬/NNG + ,/SP
꽃무늬  꽃무늬/NNG
디자인이    디자인/NNG + 이/JKS
주류를  주류/NNG + 를/JKO
이루고  이루/VV + 고/EC
있다.   있/VX + 다/EF + ./SF
```

Figure 1 tagged corpus

분류	세종 품사 태그	
	태그	설명
체언	NNG	일반 명사
	NNP	고유 명사
	NNB	의존 명사
	NR	수사
	NP	대명사
용언	VV	동사
	VA	형용사
	VX	보조 용언
	VCP	긍정 지정사
	VCN	부정 지정사

Figure 2 세종말뭉치의 부분 tag set

2.3 Output

최종 결과물은 일반적인 자연어 문장이 입력으로 들어오면 각 단어가 알맞은 tag로 tagged 된 문자열이다.

3. 멤버 구성

학번	이름	역할
M2017517	조영훈	학습 모델 구축, 테스트
20153204	이가영	데이터 수집, 데이터 가공

4. 일정표

	1주차	2주차	3주차	4주차	5주차
데이터 수집					
데이터 가공					
모델 구축					
정확도 향상					
테스트					

5. GitHub 주소

https://github.com/YoungHunCho/pos-tagger-using-tensorflow