의생명 분야의 개체명 인식에서 순환형 신경망과 조건적 임의 필드의 성능 비교

조병철○, 김유섭
한림대학교 융합소프트웨어학과
max91128@naver.com, yskim01@hallym.ac.kr

Performance Comparison of Recurrent Neural Networks and Conditional Random Fields in Biomedical Named Entity Recognition

요 약

최근 연구에서 기계학습 중 지도학습 방법으로 개체명 인식을 하고 있다. 그러나 지도 학습 방법은 데이터를 만드는 비용과 시간이 많이 필요로 한다. 본 연구에서는 주석 된 말뭉치를 사용하여 지도 학습 방법을 사용 한다. 의생명 개체명 인식은 Protein, RNA, DNA, Cell type, Cell line 등을 포함한 텍스트 처리에 중요한 기초 작업입니다. 그리고 의생명 지식 검색에서 가장 기본과 핵심 작업 중 하나이다. 본 연구에서는 순환형 신경망과 워드 임베딩을 자질로 사용한 조건적 임의 필드에 대한 성능을 비교한다. 조건적 임의 필드에 N_Gram만을 자질로 사용한 것을 기준점으로 설정 하였고, 기준점의 결과는 70.09% F1 Score이다. RNN의 jordan type은 60.75% F1 Score, elman type은 58.80% F1 Score의 성능을 보여준다. 조건적 임의 필드에 CCA, GLOVE, WORD2VEC을 사용 한 결과는 각각 72.73% F1 Score, 72.74% F1 Score, 72.82% F1 Score의 성능을 얻을 수 있다.

주제어: 의생명 개체명 인식(Bio NER), 순환형 신경망(RNN), 조건적 임의 필드(CRFs), 워드 임베딩(word embedding)

1. 서론

의생명 분야의 개체명 인식 (Biomedical Named Entity Recognition: Bio NER) 이란 의생명 분야의 문서에서 단백질, RNA, DNA 와 같은 생물학적인 정보를 가지고 있는 용어들을 자동으로 추출하는 것을 말하는데, 전문 용어들이 다수 출현하는 바이오 문서의 특성상 이 처리는 매우 중요하다[1]. 텍스트로부터 유전자 이름을 찾는 이 작업은 신문에서 회사 이름과 사람 이름을 찾는 것과 비슷한 일이다. Bio NER 은 하나의 용어가 다양한 형태로 나타나고 분류 범주가 매우 유사하기 때문에, 일반적인 개체명 인식에 비하여 상당히 난이도가 높다[2].

본 연구에서는 Bio NER 을 위하여 순환형 신경망 (Recurrent Neural Networks: RNN) 과 조건적 임의 필드(Conditional Random Fields: CRFs)와 같은 기계학습 알고리즘을 사용하여 자동적으로 단어에 대한 개체명 인식을 하였다. RNN을 위해서는 Jordan type과 Elman type을 사용하였고, CRFs에서는 인공적으로 가공되지 않은 자질들을 사용하였는데, CCA[3], GLOVE[4], WORD2VEC[5] 에서 생성된 워드 임베딩만을 자질로 사용하였다. 본 논문에서는 BioNLP/NLPBA2004 [6] 말뭉치를 사용하여 실험을 하였다.

2. 방법론

Biomedical named entity에 대해 CRFs과 RNN인 기계학습 알고리즘을 사용하여 자동적으로 단어에 대한 개체명 인식(Named Entity Recognition: NER)을 하였다. 본 논문에서는 BioNLP/NLPBA2004 corpus를 사용하여 실험을 하였다. 이 corpus는 총 22,403 문장 중 학습 데이터로 18,546 문장을 사용하였고, 훈련 데이터로 3,857 문장을 사용하였다. 이 말뭉치에는 protein, DNA, RNA, cell line and cell type 이 주석이 달려있다.

3. 조건적 임의 필드(CRFs)

CRFs (Conditional Random Fields) 는 통계적 모델링 방법 중 하나로 패턴 인식과 기계 학습과 같은 구조적 예측에 사용된다. 일반적인 방법은 분류자 (Classifier) 가 이웃하는 표본을 고려하지 않고 단일 표본 라벨을 예측하는 반면 CRFs (Conditional Random Fields) 는 이웃하는 표본을 고려하여 예측한다. CRFs는 자연언어로 된 글 또는 생물학적 서열정보 일련의 데이터에 대한 라벨 예측, 분석에 사용 되기도 한다.[7]

본 논문에서는 crf suite[1]를 사용하였고, WORD2VEC, Global Vector(GLOVE), Canonical Correlation

1) http://www.chokkan.org/software/crfsuite

Analysis(CCA)에 대한 워드 임베딩을 각각 자질로 사용하여 의생명 개체명을 예측하였다.

4. 순환형 신경망(RNN)

Neural networks는 뉴런과 뉴런이 연결되어 일을 하는 것처럼 수학적 계산을 수행하는 방식을 기초로 하는 연산 타입을 나타낸다. Neural networks는 비선형 함수와 패턴 인식에 더 잘 맞춰져 있다. Neural networks는 많은 분야에서 응용되기 때문에 데이터 마이닝, 인공지능, 생물 정보 학 등 다양한 과정의 분야에서 관심이 많다. RNN은 다양한 자연어처리의 문제에 대해 뛰어난 성능을 보이고 있는 모델이다. 기본적인 인공 신경망 구조는 모든 입력과 출력이 각각 독립적이라고 가정했지만, RNN은 hidden state에서 과거의 입력 값에서 일어난 정보를 알 수 있고, 출력 결과는 이전의 계산 결과에 영향을 받는다. 본 논문에서는 RNN tutorial[2] 을 이용하여 RNN 알고리즘을 사용하였다. RNN에서는 Elman type network[8] 과 Jordan[9] type network 의 주요한 알고리즘이 있다.

5. 실험 및 결과

본 논문에서는 BioNLP/NLPBA 2004 shared corpus를 사용하여 실험을 하였다. RNN을 Jordan type network, Elman type network의 성능과 CRFs에 CCA, Glove, Word2Vec를 통하여 생성된 워드 임베딩을 각각 자질로 사용한 성능을 비교한다. RNN을 Jordan 유형으로 사용하였을 때는 60.75%의 F1-score가, Elman 유형을 사용하였을 때는 58.80%의 F1-score가 나왔다. CRFs에서는 오직 n-gram (unigram, bigram, trigram) 만을 자질로 사용한 것을 기준점으로 결정하였고, 그 결과 71.09%의 F1-score 성능을 보였다. 여기에 추가적으로 CCA, GloVe, Word2Vec를 통하여 생성된 워드 임베딩을 각각 자질로 사용하였고, 차원수 (10,30,50,80,100), window size (3,5,7,9,11), 최소 빈도 (3) 등을 다양하게 변화시키며 실험하였다. CRFs의 실험 결과는 다음과 같다 :

	3	5	7	9	11
10차원	72.05	72.27	72.33	72.47	72.41
30차원	72.5	72.66	72.27	72.35	72.58
50차원	72.77	72.27	72.58	72.66	72.41
80차원	72.61	72.82	71.99	72.38	72.17
100차원	72.34	72.2	72.35	72.38	72.15

표 2 파라미터 값에 따른 WORD2VEC 성능 결과

	3	5	7	9	11
10차원	72.21	72.22	72.41	72.35	72.32
30차원	72.14	72.47	72.55	72.72	72.4
50차원	72.28	72.6	72.28	72.63	72.74
80차원	72.01	72.33	72.57	71.77	72.02
100차원	72.23	72.34	72.13	72.42	72.28

표 3 파라미터 값에 따른Global Vector 성능 결과

	3	5	7	9	11
10차원	72.15	72.73	72.73	72.73	72.73
30차원	72.15	72.4	72.4	72.4	72.4
50차원	72.26	72.46	72.46	72.46	72.46
80차원	72.4	72.1	72.1	72.1	72.1
100차원	72.48	72.12	72.12	72.12	72.12

표 4 파라미터 값에 따른 Canonical Correlation Analysis 성능 결과

각 파라미터를 조절한 결과, 각 임베딩 방법에서 각각 72.73% (CCA), 72.74% (GloVe), 72.82% (Word2Vec) 의 성능을 보였다.

6. 결론

Bio-NER은 다른 분야의 NER보다 더 어려운 일이다. NER은 일반적으로 사전, 규칙, 기계학습에 기반하여 개체명 인식을 한다. 사전에 기반하는 방법은 사전에 등록하지 않은 고유명사나, 등록 되었더라도 중의성 문제에 따라 문맥에 따라 개체명 인식이 다르게 될 수 있다. 규칙에 기반하는 방법은 규칙 및 패턴을 정의하여 개체명 인식을 한다. 본 논문에서는 이러한 문제를 더 효과적으로 처리하기 위해 RNN과 CRFs을 이용한 기계학습 이용하여 연구했다. CRFs는 일반적으로 자질들을 사용하는데, 본 논문에서는 인공적으로 가공 되지 않은 자질들을 사용하였고, RNN과 성능 비교를 하였다. CRFs 모델에 자질을 각기 다른 워드 임베딩인 WORD2VEC, GLOVE, CCA를 사용하였을 때, WORD2VEC이 72.82% 성능을 보였다. RNN모델은 Jordan type, Elman type network를 사용하였을 때 Jordan type network가 60.75%가 가장 좋은 성능을 보였다.

참고문헌

[1] Leaman, Robert, and Graciela Gonzalez. "BANNER: an executable survey of advances in biomedical named entity recognition." Pacific Symposium on Biocomputing. Vol. 13. 2008.
[2] Wilbur, John, Lawrence Smith, and Lorraine

2) http://deeplearning.net/tutorial/rnnslu.html

Tanabe. "Biocreative 2. Gene mention task." Proceedings of Second BioCreative Challenge Evaluation Workshop. 2007.
[3] Hotelling, H. (1936). Relations between two sets of variates. Biometrika,28(3/4), 321-377.
[4] J. Pennington, R. Socher and C.D. Manning, “Glove: Global vectors for word representation”, Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014), 12, (2014), pp. 1532-1543
[5] T. Mikolov, K. Chen, G. Corrado et al. “Efficient estimation of word representations in vector space”, arXiv preprint arXiv: 1301.3781, (2013)
[6] Kim, J. D., Ohta, et.al. (2004, August). Introduction to the bio-entity recognition task at JNLPBA. In Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (pp. 70-75). Association for Computational Linguistics.
[7] A. Mahmoud, A. Pattar and A. Hamdulla, ―Uyghur Stemming Using Conditional Random Fields‖, International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 8, no. 8, (2015), pp. 43-50.
[8] Elman,J.L. “Finding Structure in Time.” Cognitive science, vol. 14, 1990, pp. 179-211
[9] Jordan, Michael I. "Serial order: A parallel distributed processing approach."Advances in psychology 121 (1997): 471-495.