문서 데이터에서 추출한 지식기반 삼중 데이터 집합을 이용한 링크 예측 모델 학습

이명학, 엄정민, 이혜진, 강주희, 하정민, 한승진, 이재구*

국민대학교

jaekoo@kookmin.ac.kr

Learning the link prediction model using knowledge-based triplet dataset extracted from document data

Lee Myung Hak, Eom Jung Min, Lee Hye Jin, Kang Ju Hee, Ha Jung Min, Han Seung Jin, Lee Jae Koo*

College of Computer Science, Kookmin University.

요 약

본 논문은 객체(Entity) 간의 관계(Relation)를 예측하는 링크 예측(Link Prediction) 과업을 위한 모델을 학습하는 방법론에 관하여 서술하였다. 링크 예측 과업의 학습을 위해서는 객체-관계-객체(Entity-Relation-Entity)로 이루어진 지식기반 삼중 데이터 집합이 필요하다. 그러나 지식기반 삼중 데이터 집합의 수가 적으므로, 우리는 문서 형식의 데이터 집합으로부터 지식기반 삼중 데이터 집합을 추출해 새로운 데이터 집합을 만들어 이 문제를 해결하였다. 그리고 추출된 데이터 집합이 링크 예측 모델을 학습할 수 있는지 알아보기 위해 링크 예측 과업에서 사용하는 대표적 모델인 HAKE(Hierarchy-Aware Knowledge Graph Embedding)와 AcrE(Atrous Convolution and Residual Learning)의 학습을 진행해 보았고 그 결과 HAKE 모델과 AcrE 모델에서 각기 0.4328과 0.4812의 정확도를 보였다.

Ⅰ. 서 론

링크 예측 과업은 [그림 1]과 같이 객체-관계-객체(Entity-Relation-Entity)의 지식기반 삼중(Knowledge-based Triplets) 데이터 구조로 이루어진 그래프 형태의 데이터베이스에서 새로 들어온 객체와 기존 객체의 관계를 예측하는 과업이다. 영화 추천 과업을 예로 들면, 사람, 영화, 영화 종류가 객체가 되고 이들 사이의 선호도(Likes)와 장르(Genre)는 관계가 된다. 그리고 특정 사람이 과거에 봤던 영화의 선호도가 그래프 형태로 기록되어 있는 상황에서 새로운 영화가 그래프에 추가되었을 때 그 영화의 선호도를 예측하는 것이다.

최근에는 링크 예측 과업을 깊은 신경망을 이용하여 진행한 연구들이 많아지고 있다[1, 2]. 깊은 신경망을 훈련 시키기 위해서는 WN-18RR[3]이나 FB-15K-237[3]과 같은 지식기반 삼중 데이터 집합이 필요하나, 지식기반 삼중 데이터 집합의 수가 부족해 신경망을 학습하는데 제한이 있다. 따라서 우리는 이를 보완하기 위해 문서 데이터 집합으로부터 지식기반 삼중 데이터 구조를 추출하는 모델을 이용하여 삼중 구조 데이터 집합을 생성했다. 그리고 이 데이터 집합이 링크 예측 모델을 학습시킬 수 있는지 알아보았다.

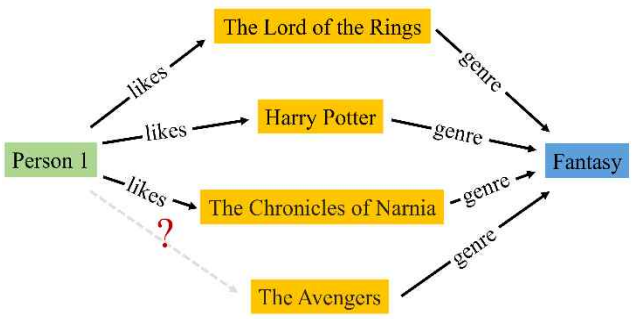

그림 1. 영화 추천 시스템에서 지식기반 삼중 데이터 집합을 이용한 그래프 데이터베이스의 예시

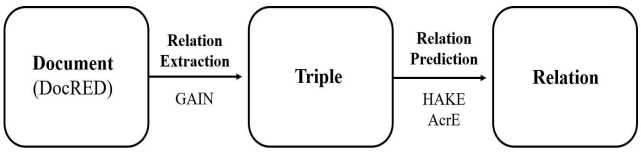

그림 2. 문서로부터 뽑은 지식기반 삼중 데이터 집합을 이용해 링크 예측 모델을 학습시키는 과정

Ⅱ. 본론

본 논문에서는 부족한 링크 예측 데이터 집합을 학습 시키는데 필요한 지식기반 삼중 데이터 집합을 보완하기 위해 문서 데이터 집합에서 삼중 데이터 집합을 추출하였다. 실험 과정은 [그림 2]와 같다. 우리는 실험을 크게 문서로부터 지식기반 삼중 데이터 집합을 뽑아내는 과정과 뽑아낸 데이터 집합으로부터 링크 예측 모델을 학습시키는 과정으로 나누었다.

첫 번째 단계에서는 문서 데이터 집합인 DocRED(Document-Level Relation Extraction Dataset)[4]를 이용해 문서로부터 지식기반 삼중 데이터 구조를 추출하는 모델인 GAIN(Graph Aggregation-and-Inference Network)[5]을 학습시켰다. GAIN은 BERT(Bidirectional Encoder Representations from Transformers)[6]나 GloVe(Global Vectors for Word Representation)[7]를 이용하여 단어 임베딩(Word Embedding)을 한 후 임베딩 된 벡터를 이기종 수준(Heterogeneous-level) 그래프를 이용하여 전체 문서에서 객체 간의 상호작용을 나타내는 모델이다. 우리가 학습한 GAIN의 평가지표는 정밀도(Precision), 재현율(Recall), F1-점수가 각각 0.5271, 0.6002, 0.5613으로 나왔다. 우리는 위와 같이 학습된 GAIN 모델을 이용해 DocRED의 테스트 데이터 집합에서 지식기반 삼중 데이터를 추출하였고 그 결과 638개의 객체와 66개의 관계로 이루어진 총 2,135개의 지식기반 삼중 데이터 집합을 얻을 수 있었다.

표 1. 링크 예측 작업을 위한 데이터 집합인 FB15K-237, WN18RR과 DocRED로부터 추출한 지식기반 삼중 데이터 집합을 이용해 링크 예측 모델인 HAKE와 ACRE의 평가지표.

Dataset	Model	MR	MRR	Hits@1	Hits@3	Hits@10
WN18RR[3]	HAKE[1]	3360.82	0.4973	0.4531	0.5144	0.5817
	AcrE[2]	4773.8	0.4324	0.3998	0.4454	0.4983
FB15K-237[3]	HAKE[1]	183.84	0.3449	0.2477	0.3824	0.5412
	AcrE[2]	188.5	0.3501	0.2589	0.3838	0.5323
DocRED[4] (Our Data)	**HAKE[1]**	**140.5054**	**0.4785**	**0.4328**	**0.5000**	**0.5699**
	AcrE[2]	**83.7390**	**0.5301**	**0.4812**	**0.5591**	**0.6075**

두 번째 단계에서는 링크 예측 과업을 위한 대표적 데이터 집합인 WN18RR과 FB-15K-237을 우리가 앞서 뽑은 지식기반 삼중 데이터 집합과 비교하여 추출된 데이터 집합의 합리성을 확인해 보았다. 본 논문에서는 비교를 위해 각 데이터 집합에 대하여 링크 예측 모델인 HAKE(Hierarchy-Aware Knowledge Graph Embedding)[1]와 AcrE(Atrous Convolution and Residual Learning)[2]를 학습시키고 그 평가지표를 구해보는 방식을 사용하였다. HAKE는 그래프상에서 객체와 관계를 모두 노드로 본 후 관계를 나타내는 노드는 서로 멀어지고 같은 관계 노드에 묶인 객체 노드끼리는 서로 가까워지는 방식으로 군집을 만들어 링크 예측을 하는 모델이다. 그리고 AcrE는 표준 합성곱(Standard Convolution), 확장된 합성곱(Dilated Convolution), 그리고 잔차 연결(Residual Connection)을 사용한 KGE(Knowledge Graph Embedding)모델이다.

앞서 구한 결과를 우리는 [표 1]에 나타내었다. [표 1]에서 Hits@k는 데이터 리스트에서 모델이 예측한 상위 k번째 안에 정답 데이터가 존재할 확률을 나타내는 지표이고, MR(Mean Rank)은 전체 데이터 집합의 각 리스트에서 모델이 예측한 정답 데이터의 순위에 대한 평균이며 MRR(Mean Reciprocal Rank)은 모델이 예측한 정답 데이터의 순위에 대한 역수의 평균이다. 그리고 우리가 첫 번째 단계에서 DocRED로부터 추출한 지식기반 삼중 데이터 집합은 굵은 글씨로 표시하였다.

그 결과, 우리가 DocRED로부터 지식기반 삼중 데이터를 추출하여 만든 데이터 집합은 다른 링크 예측 작업을 위해 만든 데이터 집합과 비교해도 성능 면에서 크게 차이 나지 않았다. 즉 우리는 관계 추출(Relation Extraction) 모델을 이용해 만든 지식기반 삼중 데이터 집합을 가지고 링크 예측 신경망을 학습할 수 있음을 증명하였다.

Ⅲ. 결론

본 논문에서는 문서 기반 데이터 집합으로부터 지식기반 삼중 데이터 집합을 뽑은 후 이를 이용해 링크 예측 모델을 학습시켰다. 그 결과, 준수한 성능을 확인할 수 있었다. 향후 우리는 조금 더 안정적인 모델 학습을 위하여 DocRED 안의 같은 뜻을 가진 객체를 묶는 방법을 시도해볼 것이다. 그리고 더 나아가 보안 등의 특정 세부 분야에 집중된 신경망에서도 본 논문에서 진행하였던 실험과 유사하게 준수한 성능을 갖는지 실험할 것이다.

ACKNOWLEDGMENT

" 본 연구는 2016년 과학기술정통신부 및 정보통신기획평가원의SW중심대학사업의 연구결과로 수행되었음"(2016-0-00021)

" 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00994)

참 고 문 헌

[1] Z. Zhang, J. Cai, Y. Zhang, and J. Wang, "Learning hierarchy-aware knowledge graph embeddings for link prediction," in Proc. AAAI, 2020, pp. 3065?3072.

[2] Feiliang Ren, Juchen Li, Huihui Zhang, Shilei Liu, Bochao Li, Ruicheng Ming, Yujia Bai, Knowledge Graph Embedding with Atrous Convolution and Residual Learning, arXiv:2010.12121v2

[3] T. Dettmers, P. Minervini, and P. Stenetorp, ''Convolutional 2D knowledge graph embeddings,'' in Proc. 32nd AAAI Conf. Artif. Intell., 2018

[4] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 764-777.

[5] Shuang Zeng , Runxin Xu1, Baobao Chang, and Lei Li, Double Graph Based Reasoning for Document-level Relation Extraction, arXiv:2009.13752v1

[6]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint

arXiv:1810.04805, 2018.

[7] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532 - 1543.