딥러닝 기법을 활용한 온톨로지 매칭 방법

이용주[1], 단홍조우[2]
[1]경북대학교 IT대학 컴퓨터학부
[2]경북대학교 IT대학 컴퓨터학부 박사과정
yongju@knu.ac.kr, caixiuming1984@163.com

Ontology Matching Method Using Deep Learning Technologies

Yongju Lee[1], Hongzhou Duan[2]
[1,2]Dept. of Computer Science and Engineering, Kyunpook National University

요 약

WordNet, Freebase, Wikidata와 같은 대규모 지식베이스에 대한 딥러닝 연구가 활발히 진행되고 있으나 이에 관한 온톨로지 매칭 연구는 거의 연구가 미비한 상태이다. 본 연구에서는 효율적인 온톨로지 매칭 방법을 개발하기 위해 데이터 수집 및 전처리, 텍스트 유사도 계산, 그리고 텍스트 유사도에 따른 결과 매칭의 세 가지 순차적 단계로 구성된 하나의 새로운 방법을 제안한다. 성능평가는 정보검색 분야에서 널리 활용되고 있는 재현율, 정밀도, F-측정값을 사용하였는데 제안한 방법이 기존의 모든 방법들보다 성능이 우수함을 보였다.

1. 서론

최근 WordNet, Freebase, Wikidata와 같은 대규모 지식베이스에 지식 그래프 및 합성곱 신경망을 활용한 연구가 활발히 진행되고 있으나, 지식베이스에 대한 온톨로지 매칭 연구는 상대적으로 거의 연구가 미비한 상태이다[1]. 대규모 지식베이스에 시멘틱이 어떻게 임베딩되어 딥러닝되고, 온톨로지 매칭에 딥러닝이 어떻게 활용되는지는 상대적으로 거의 연구가 되지 않고 있다.

온톨로지(ontology)란 지식의 어떤 특정 영역 내에 있는 실체와 그 실체가 의미하는 것 사이의 상호작용 데이터 모델을 말한다[2]. 이는 해당 분야의 지식을 개념화하고 이를 기계가 이해할 수 있게 명세화해야 한다. 이를 명세화할 수 있는 기법으로써 RDFS나 OWL 등의 언어를 이용해 표현한다. 온톨로지는 지식 표현의 한 방법으로 지식 베이스에서 해당 구조를 활용함으로써 많은 가치를 얻을 수 있다. 개념 간의 관계에 대한 지식을 통해 문제에 대한 솔루션을 식별하거나 어떤 제품이나 서비스가 다른 제품 세트와 함께 제공되는지에 대한 연관 지식을 얻을 수 있다.

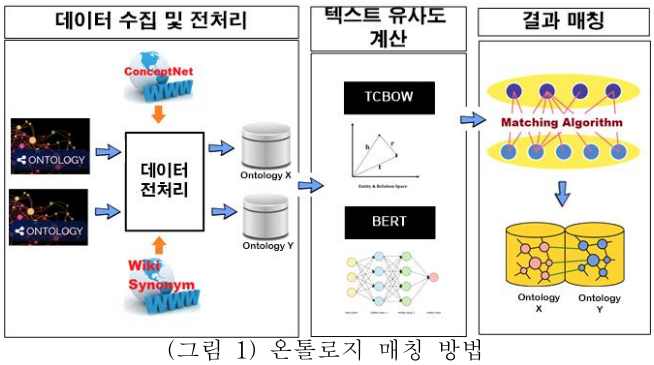
본 연구에서는 딥러닝 기법을 활용한 온톨로지 매칭 문제를 해결하기 위해 TCBOW(Twinned Continuous Bag Of Words), BERT(Bidirection Encoder Representations from Transformers), 그리고 SimRank 모델을 결합한 하나의 새로운 온톨로지 매칭 방법을 제안한다.

2. 관련 연구

지식 그래프 및 온톨로지는 지식 표현 방법으로써 RDF, RDF-S, OWL 등의 언어를 이용해 기술한다. 단어 임베딩은 유사한 단어들이 인접한 거리에 위치하도록 각 단어에 해당하는 벡터값을 찾는 것으로써 Word2Vec[3], GloVe[4], FastText[5] 등이 있다. 온톨로지에서 엔터티 간의 구조적 유사성을 계산하려면 온톨로지에 존재하는 Is-A와 SubclassOf 같은 관계를 고려하는 것이 필요한데, SimRank 알고리즘[6]은 두 온톨로지에서 추출된 그래프 구조 내 노드 간의 유사성을 발견한다.

3. 딥러닝 기법을 활용한 온톨로지 매칭 방법

(그림 1)은 온톨로지 매칭 방법에 대한 프로세스를 개략적으로 표현한 그림으로, 이 모델에서는 데이터 수집 및 전처리, 텍스트 유사도 계산, 텍스트 유사도에 따른 결과 매칭의 세 가지 순차적 단계로 구성된다. 이러한 단계는 궁극적으로 최종 매칭 결과 도출로 이어진다.



(그림 1) 온톨로지 매칭 방법

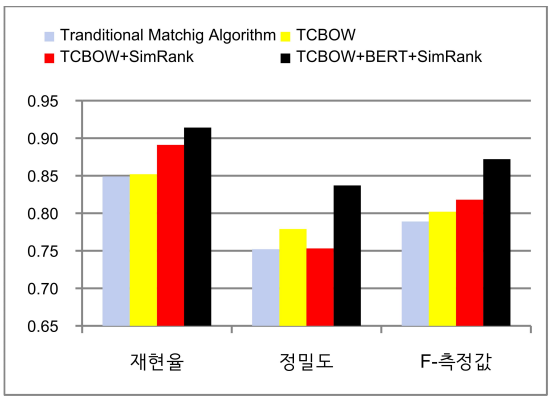
텍스트 유사도 계산에서는 고품질의 문장 임베딩을 효율적으로 산출하기 위하여 TCBOW 신경망 모델을 사용한다. 문장에 대한 임베딩을 계산하는데 하나의 효율적이고 성공적인 방법은 문장을 구성하는 단어의 임베딩을 평균화하는 것이다. 즉, 한 쌍의 문장(x_i, x_j)에 대해 훈련 데이터에서 문장이 서로 인접할 가능성을 반영하는 확률 $p(x_i, x_j)$를 정의한다. 또한, 이 단계에서 온톨로지의 텍스트 요소에 의미론적 표현을 할당하기 위해 BERT 방법을 활용한다.

결과 매칭 단계에서는 SimRank 알고리즘을 활용하여 개체 간의 구조적 유사성을 계산한다. 이전 단계에서 얻은 텍스트 유사성 정보를 사용하여 두 그래프 구조 간의 연결을 설정하고 이를 하나의 그래프로 결합한 후, SimRank 알고리즘을 두 온톨로지의 엔터티 간 구조적 유사성을 측정하는 데 적용한다.

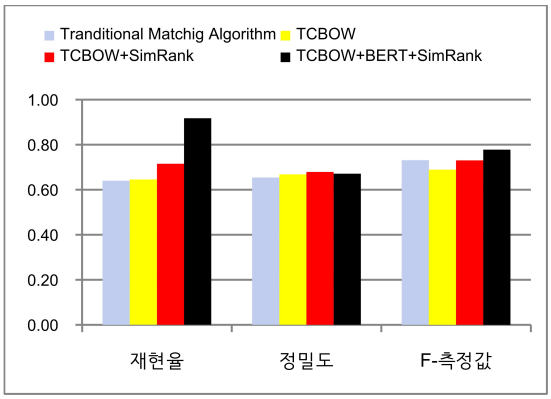
4. 성능 평가

성능 평가 방법은 정보검색 분야에서 가장 널리 활용되고 있는 재현율(recall)과 정밀도(precision), 그리고 F-측정값(F-measure)을 이용한다. 실험 데이터로는 FMA, SNOMED, 그리고 NCIT 데이터셋을 사용한다. FMA는 해부학적 정보를 위해 분산 프레임워크에 통합된 정보 리소스이고, SNOMED는 SNOMED 국제 조직에서 관리하는 컴퓨터 처리 가능한 의학 용어 모음이다. 그리고 NCIT는 미국 국립 암 연구소에서 개발한 공개적으로 사용 가능한 용어집이다.

우리는 FMA-NCIT와 FMA-SNOMED 데이터셋을 사용하여 기존 전통적인 매칭 알고리즘은 물론이고 여러 다양한 알고리즘에 대한 전반적인 성능을 측정하였는데, (그림 2)와 (그림 3)은 이러한 알고리즘들에 대한 성능 측정 결과를 보이고 있다. 이 그래프로부터 TCBOW, BERT, SimRank를 결합한 우리의 접근 방식은 FMA- SNOMED의 정밀도 성능만 제외하고 다른 모든 방법보다 우수한 것을 알 수 있다.



(그림 2) FMA-NCIT 데이터셋 결과



(그림 3) FMA-SNOMED 데이터셋 결과

4. 결론

본 논문에서는 새로운 온톨로지 매칭 방법을 제안하기 위해 TCBOW, BERT, 그리고 SimRank 모델을 결합한 딥러닝 기법을 활용한 매칭 방법을 제안하였다. 제안된 온톨로지 매칭 방법에서는 데이터 수집 및 전처리, 텍스트 유사도 계산, 텍스트 유사도에 따른 결과 매칭의 세 가지 순차적 단계로 구성되고 이들 결과를 통합하여 최종 매칭 결과를 산출한다. 성능평가 방법은 정보검색 분야에서 보편적으로 사용되고 있는 재현율, 정밀도, F-측정값을 사용하였는데 본 논문에서 제안한 방법이 기존의 모든 방법들보다 성능이 우수함을 보였다.

감사의 글

본 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2016R1D1A1B02008553). 본 논문은 교육부 및 한국연구재단의 4단계 BK21 사업(경북대학교 컴퓨터

학부 지능융합 소프트웨어 교육연구단)으로 지원된 연구임 (4120240214871).

참고문헌

[1] 이용주, 순위샹, “그래프 합성곱 신경망과 임베딩 기법을 적용한 지식 그래프 엔티티 매칭 방법,” 한국정보기술학회논문지, 제21권, 제6호, pp. 9-19, 2023.

[2] 조명대, 링크드 데이터, 서울시 마포구, 커뮤니케이션북스, 2017.

[3] J. Pennington, R. Socher, and C. Manning, “GloVe: Global Vectors for Word Representation,” Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1532-1543.

[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with subword Information,” Transactions of the Association for Computational Linguistics, Vol. 5, pp. 235-146, 2016.

[5] T. Kenter, A. Borisov, and M. Rijke, "Siamese CBOW: Optimizing Word Embeddings for Sentence Representations,” 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 941-951.

[6] G. Jeh and J. Widom, "SimRank: A Measure of Structural-Context Similarity,” 8th ACM KDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 538-543.