문서 내 전역 관계 추출: 생략된 개체의 고려

김규경○, 김경민, 조재춘, 임희석

고려대학교, NLP&AI 연구실

overmind22@korea.ac.kr, totoro4007@korea.ac.kr, jaechoon@korea.ac.kr, limhseok@korea.ac.kr

Global Relation Extraction for Documents: Regarding Omitted Entities

Kuekyeng Kim○, Gyeongmin Kim, Jaechoon Jo, Heuisoek Lim

Korea University, NLP&AI lab

요 약

최근 존재하는 대부분의 관계 추출 모델은 언급 수준의 관계 추출 모델이다. 이들은 성능은 높지만, 문서에 존재하는 다수의 문장을 처리할 때, 문서 내에 주요 개체 및 여러 문장에 걸쳐서 표현되는 개체간의 관계를 분류하지 못한다. 이는 높은 수준의 관계를 정의하지 못함으로써 올바르게 데이터를 정형화지 못하는 중대한 문제이다. 해당 논문에서는 이러한 문제를 타파하기 위하여 여러 문장에 걸쳐서 개체간의 상호작용 관계도 파악하는 전역 수준의 관계 추출 모델을 제안한다. 제안하는 모델은 전처리 단계에서 문서를 분석하여 사전 지식베이스, 개체 연결 그리고 각 개체의 언급횟수를 파악하고 문서 내의 주요 개체들을 파악한다. 이후 언급 수준의 관계 추출을 통하여 1차적으로 단편적인 관계 추출을 실행하고, 주요개체와 관련된 관계는 외부 메모리에 샘플로 저장한다. 이후 단편적 관계들과 외부메모리를 이용하여 여러 문장에 걸쳐 표현되는 개체 간 관계를 알아낸다. 해당 논문은 이러한 모델의 구조도와 실험방법의 설계에 대하여 설명하였고, 해당 실험의 기대효과 또한 작성하였다.

주제어: 관계 추출, 메모리 증강 신경망, 문서 분석, 자연어처리

1. 서론

문장은 개체를 설정하고, 해당 문장 내의 개체간의 상호작용을 나타냄으로써 의미를 얻는다. 그리고 이러한 상호작용들은 개체간의 시멘틱 관계 유형으로써 표현될 수 있다. 이것이 바로 관계 추출(Relation Extraction)이다.[1] 그러나 관계 추출은 문장 내의 개체 관계만 주목하는데, 여러 문장이 모여서 개체간의 상호작용이 문장 내에 국한되지 않고 여러 문장에 걸쳐 상호작용이 일어나는 경우에는 그러한 상호작용들을 관측하지 못하는 약점이 부각된다. 이러한 단점들은 주어나 목적어가 자주 생략되는 한국어에서 특히나 두드러진다. 이 논문에서는 이러한 단점을 극복하기 위하여 문서 내에 공통적으로 언급되어지는 주요 개체들과 타 개체들과의 관계에 주목한다. 제안된 모델은 다수의 문장을 처리할 때 해당 개체가 포함되지 않은 문장을 처리할 때에도 문장 내의 개체들과 포함되지 않는 주요 개체들 간의 관계 또한 분석하여 추출하려고 시도한다. 이후 해당 논문은 2.관련 연구에서 이 논문과 관련된 논문들 및 그에 대한 비교들을 설명한다. 3. 모델에서는 제안한 모델을 설명하고 4.평가방법 및 데이터에서는 이 모델에 사용된 데이터셋에 대한 설명과 제안된 모델의 성능 비교를 위한 평가방법을 설명할 것이다. 마지막으로 5. 결론에서는 이 논문에서 설계한 모델 완성 후 기대효과 및 향후 미래 연구에 대해 설명하며 이 논문을 마친다.

2. 관련 연구

관계 추출(Relation Extraction): 관계 추출은 구조화 도지 않은 데이터로부터 구조화된 정보를 추출하는데 핵심적인 역할을 수행한다. 이는 단순히 텍스트 데이터뿐만 아니라 개체와 개체간의 상호관계로 의미를 전달하는 모든 종류의 매체에 사용되어 질 수 있다.[2] 이러한 관계 추출은 크게 2가지 종류로 나눌 수 있는데 각각 전역 수준의 관계 추출(Global Level Relation Extraction)과 언급 수준의 관계 추출(Mention Level Relation Extraction)이 존재한다[3]. 현재 대부분의 높은 성능을 자랑하는 관계 추출 모델은 후자에 속하며, 문장 내의 관계 추출 성능은 우수하다. 그러나 이는 많은 양의 정보를 요약하고 그 주제를 파악하는데에는 한계가 있다. 이 논문에서는 전역 수준의 관계 추출을 한 번에 대량의 문서를 전부 분석하여 관계를 추출하기보다는 각 문장을 분석한 뒤, 문장 내에서 추출된 단편적인 관계들을 전체적으로 분석함으로써 전역 수준의 관계추출을 시도한다. 이러한 시도방법을 취함으로써 전역 수준의 관계 추출을 하되, 언급 수준의 관계 추출을 병행함으로써 정보의 누락을 최대한 방지한다.

메모리 증강 신경망(Memory Augmented Neural Network): 본 논문에서는 전역 수준 관계 추출과 언급 수준 관계 추출을 병행함으로써 많은 양의 개체 관계들을 처리하게 된다. 그러나 이러한 관계들 중 주요 개체가 포함된 개체들의 관계는 전역 수준 관계 추출에 커다란 영향을 끼친다. 그렇기 때문에 이들이 가지고 있는 관계들은 따로 샘플로 외부메모리에 저장하여 전역 수준 관계 추출 때 해당 관계 분류를 할 때 사용된다. 그리고 이렇게 추출된 주요 개체 간의 관계들은 다시 한 번 언급 수준 관계들을 재조정하는데 사용된다. 이는 메타 학습(Meta Learning)에서 사용되는 One-shot 학습 방법을 적용한

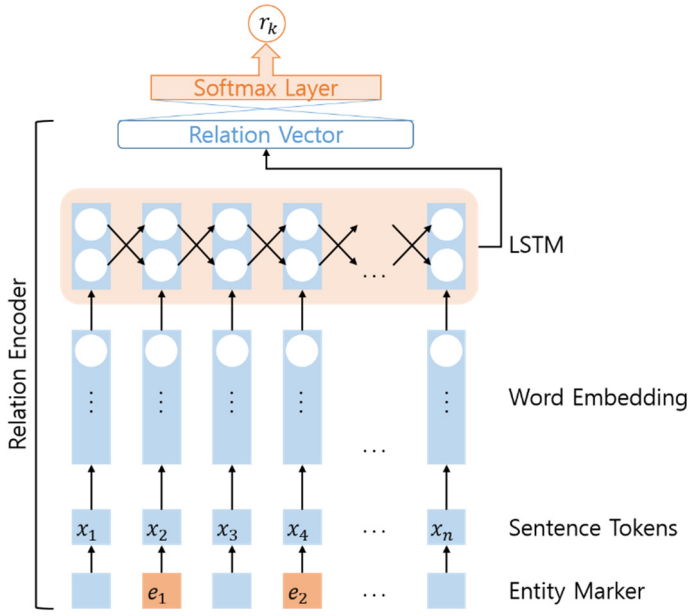

그림 1

것이다. 모델을 참조한 [4]에서는 첫 번째로 나온 분류 예측 결과를 외부 메모리에 샘플로 저장하고, 이를 차후 비슷한 유형의 샘플을 예측할 때 가져와서 사용하고, 이 결과에 따라 첫 번째 분류 예측 프로세스에 역전파 신호(Back Propagated Signal)를 보내어 세부 사항을 다시 한 번 재조정한다.

3. 모델

이 논문에서 제안하는 모델은 크게 3단계로 이루어져 있다. 각각 개체 분석, 단편적 관계 추출 단계 그리고 주요 개체 관계 추출 및 재검토의 단계의 순서로 진행된다.

해당 모델의 개체 분석은 입력된 문장의 전처리 후, 문서 내에 존재하는 모든 개체를 파악하는 것으로 시작된다. 이때 파악된 개체들이 문서 내에서 차지하는 중요성을 파악하기 위해 언급횟수 및 사전에 구축된 지식베이스(Knowledge Base)를 통한 개체 연결(Entity Linking)을 실행해 문서 내의 주요 개체들을 파악한다[5][6]. 이때 파악한 주요 개체들은 문서 내의 주제와 밀접한 개체들로써 문서 내에 존재하는 타 개체와 직·간접적으로 강한 연관성을 지니고 있다고 가정한다. 이리하여 문서 내의 주요 개체들을 기록함과 함께 개체 파악을 마친다.

이어지는 관계 인코더는 그림 1과 같다. 이 단계에서는 문서를 각 문장 단위로 분석하기 시작한다. 이때 문장 x에 존재하는 각 토큰 $x = \{x_1, x_2 \dots x_n\}$들을 행렬 $W \in R^{|V| \times k}$을 이용하여 k차원의 벡터에 임베딩한다. 이때 $|V|$는 vocabulary의 크기를 나타낸다. 이 후 문장 내의 각 토큰들이 추출될 관계의 첫 번째 또는 두 번째 개체에 속하는지, 아니면 어느 개체와도 관계가 없는지 구분하여 마킹을 한다. 이는 구문분석을 통해 각 단어를 문법적으로 더 가까운 쪽으로 분류하도록 하며 이 거리가 일정 임계값을 넘으면 해당 토큰은 그 어느 쪽에도 속하지

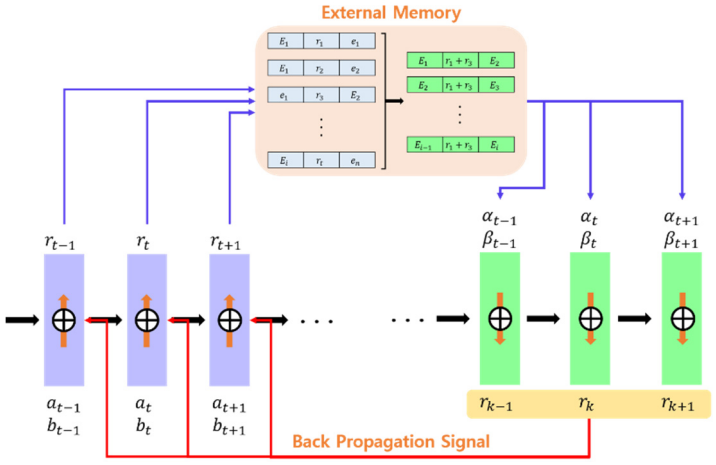

그림 2

않는 토큰으로 마킹된다. 이는 각 토큰의 포지션을 나타내는 부분으로써 사용된다. 이렇게 만들어진 토큰 임베딩들은 LSTM신경망(Long Short Term Memory Network)을 통하여 n개의 벡터들을 고정된 크기의 출력 벡터로 변환한다[7]. 이 출력 벡터는 문장 내 관계 인코딩을 시도하는 2 개체의 관계를 나타낸다.

이때 관계 인코딩을 시도하려는 2 개체 중 적어도 하나가 1단계에서 파악한 주요 개체들 중 하나라면 관계된 개체명과 출력벡터를 외부 메모리에 관계예제로써 저장한다. 마지막 단계인 관계 분류에서는 2단계에서 얻은 출력 벡터를 소프트맥스 레이어에 대입함으로써 해당 개체 쌍의 관계여부의 판별 및 관계 분류를 1차적으로 관계추출을 실행한다.

그러나 이렇게 하여 추출된 관계들은 오로지 하나의 문장 안에 존재하는 관계들을 추출한 것이 된다. 그러나 여러 문장에 걸쳐져있는 관계를 가진 개체들을 분석하기 위해서는 추가적인 조치가 요구된다. 가장 첫 번째로는 외부 메모리 내에 존재했던 각 주요 개체들의 관계를 통합하여 문서 내 주요 개체 간 관계를 판별하는 것이다. 이는 주요 개체들 간의 관계 사이에 위치하는 모든 종류의 출력벡터를 임베딩하여 새로운 예측 벡터를 생성한다. 이렇게 주요 개체간의 출력벡터를 얻어낸 후에는 문서내에 존재하는 타 개체들 간의 관계들을 정의해야 한다. 이를 위해서는 각각 개체와 주요개체간의 위치 및 거리를 분석하여 각각의 비주요 개체가 어느 주요 개체에 속하는지 마킹을 한다. 이후 첫 번째 개체가 속한 주요개체와 두 번째 개체가 속한 주요개체들 간의 관계를 나타내는 벡터, 각 비주요개체와 주요개체간의 관계를 나타내는 벡터를 합친다. 이후 얻어진 언급 수준의 관계들을 재분류하기 위하여 개체 간의 벡터들을 다시 한 번 판별하여 관계 여부의 판별 및 관계 분류를 최종적으로 수행한다.

4. 평가 방법 및 데이터

제안된 모델의 성능을 평가하기 위해서는 크게 2가지 부분을 중점적으로 확인한다. 첫 번째는 타 한국어 관계

추출 모델과 동일한 데이터셋을 통한 성능비교이다[8]. 그러나 해당 비교평가만으로는 여러 문장에 걸쳐 상호작용하는 개체들의 관계를 얼마나 정확하게 추출하는지 성능평가를 통해 파악하기 힘들기 때문에 두 번째 실험을 추가한다. 동일한 데이터셋을 기반으로 만들어진 온톨로지와 해당 모델의 결과물과 유사도를 비교하여 주요개체들과 비주요 개체들 간의 관계추출을 얼마나 정확하게 하였는지 평가하는 것이다[9].

이 논문에서 훈련 및 평가에 쓰이는 데이터셋은 위키 형식의 한국어 문서로 기술한 데이터셋을 사용된다. 두 번째 실험에 사용되어지는 온톨로지는 동일한 데이터셋을 수동으로 온톨로지로 바꾸며, 이때 사용되는 관계분류기준은 온톨로지와 제안된 모델의 분류를 통일한다.

현재 해당 실험을 위하여 위의 정의와 겹치는 데이터셋을 제작하는 중에 있으며, 이러한 평가를 통해 향후 추가적인 성능 개선을 할 계획이다.

5. 결론

이 논문에서는 문서 관계 추출을 시도할 시 문서 내의 여러 문장을 통해 관계가 형성되는 개체들의 관계도 추출하려고 하는 모델을 제안하였다. 자연어로 이루어진 문서에서는 단 하나의 문장 내에서 성립되는 관계보다는 여러 문장에 걸쳐 표현되어지는 개체들 간의 관계가 더 해당 문서내의 주제를 더 잘 반영한다. 그렇기 때문에 이러한 여러 문장에 걸쳐서 표현되어지는 관계를 추출하는 것이 해당 문서를 더욱 정확하게 요약하며, 단순히 단편적인 관계들을 사용하는 것보다 훨씬 정형화되고 사용하기 편한 지식베이스를 구축할 있도록 해준다[8].

향후 미래 연구로는 현재 평가를 위한 데이터셋을 제작하는 것을 마친 후 성능평가를 수행함과 함께 그에 대한 결과에 따라 제안된 모델의 성능을 더 높일 계획이다. 이후 해당 연구가 끝나면 이를 문서 단위로 관계가 정의되어지는 개체들의 관계추출이 가능한지 알아보고, 이를 통하여 자동적인 지식베이스 구축을 시도해볼 계획이다.

사사문구

본 연구는 문화체육관광부 및 한국콘텐츠 진흥원의 2017년도 문화기술 연구개발 지원 사업으로 수행되었음 [R2017030045].

참고문헌

[1] Hyun, Kim et al. "디지털 인문학 - 아카이브와 인문학 연구의 통섭" Proceedings at 디지털 인문학. 2017.
[2] Nguyen, Thien Huu, and Ralph Grishman. "Relation extraction: Perspective from convolutional neural networks." Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015.
[3] Pawar, Sachin, Girish K. Palshikar, and Pushpak Bhattacharyya. "Relation Extraction: A Survey." arXiv preprint arXiv:1712.05191 (2017).
[4] Santoro, Adam, et al. "Meta-learning with memory-augmented neural networks." International conference on machine learning. 2016.
[5] Shen, Wei, Jianyong Wang, and Jiawei Han. "Entity linking with a knowledge base: Issues, techniques, and solutions." IEEE Transactions on Knowledge and Data Engineering 27.2 (2015): 443-460.
[6] Pappu, Aasish, et al. "Lightweight multilingual entity extraction and linking." Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017.
[7] Sorokin, Daniil, and Iryna Gurevych. "Context-Aware Representations for Knowledge Base Relation Extraction." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.
[8] Gábor, Kata, et al. "Semeval-2018 Task 7: Semantic relation extraction and classification in scientific papers." Proceedings of The 12th International Workshop on Semantic Evaluation. 2018.
[9] Paulheim, Heiko. "Knowledge graph refinement: A survey of approaches and evaluation methods." Semantic web 8.3 (2017): 489-508.