딥러닝 기반 기계번역 개념을 활용한 Text-to-Ontology 변환 사례

신유진*, 이지항*,+
*상명대학교 지능데이터융합학부 휴먼지능정보공학전공
201810780@sangmyung.ac.kr, jeehang@smu.ac.kr

A case study on Text-to-Ontology transformation on the basis of neural translation

Yu-Jin Shin*, Jee Hang Lee*,+
*Department of Human-Centered AI, Sangmyung University, Seoul, South Korea
+Corresponding author: Jee Hang Lee

요 약

온톨로지(Ontology)는 사람과 컴퓨터, 또는 컴퓨터 간의 개념 및 개념 표현을 공유하기 위한 개념화의 명시적 규약을 의미한다. 기존의 온톨로지 생성은 전문가에 의한 수작업에 의존되어 비용과 시간이 많이 드는 한계가 있다. 이에 본 논문에서는 딥러닝(Deep learning)기반의 기계번역 개념을 적용한 사례를 활용하여, 수작업의 의존성이 감소한 방법으로 텍스트로부터 온톨로지를 생성하는 방법을 구현하였다. 특히 기존 연구에서 제안한, 딥러닝을 이용해 텍스트로부터 지식 표현 시퀀스를 추출한 정보를 활용하여, 지식 표현 구조를 온톨로지로 변환하고 지식 베이스로 확장하는 과정을 통해 자동화 된 Text-to-Ontology 변환 방법론을 제안하고자 한다.

1. 서론

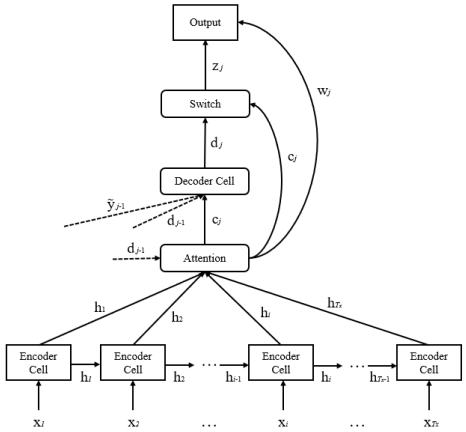
온톨로지(Ontology)는 사람과 컴퓨터, 또는 컴퓨터 간의 개념 및 개념 표현을 공유하기 위한 개념화의 명시적 규약을 의미한다[1]. 온톨로지는 특정 영역의 지식을 컴퓨터 프로그램에서 다룰 수 있는 정형화된 형식 (e.g. OWL[2], RDF[3], RDFS[4])으로 표현할 수 있어 검색, 추론, 지식표현 등 다양한 응용 시스템 개발에 활용된다. 그러나, 지식 정보를 체계화하기 위해서는 각 분야별 지식 정보에 따라 상당한 지식이 요구되며[5], 이는 결국 전문가의 수작업으로 이어져 비용과 시간이 많이 드는 한계가 있다[6].

이에 본 논문에서는 딥러닝 기반 기계번역의 어텐션 메커니즘을 적용한 사례[7]를 통해 지식 정보의 이해여부와 상관없이 문장의 문법적 구조에 기반하여 체계화된 지식 정보를 생성하는 Text to Logic-based Knowledge Representation 과정을 살펴보고자 한다. 이후 논리적 기호 사이의 계층 구조에 근거하여, 논리 기반 지식 표현 시퀀스를 지식 그래프(knowledge graph) 형태로 후가공하고, 지식 그래프 내 노드와 노드간 관계를 추출하여 head-relation-tail 형식의 온톨로지 형식으로 표현하고자 한다.

위 세 단계 과정, (Text) to (logic based knowledge representation) to (ontology)을 거쳐 자동화 된 Text-to-Ontology 변환 엔진은 딥러닝 기반 기술을 통해 수작업의 의존성이 감소한 방법으로 온톨로지를 생성할 수 있는 가능성을 제시할 것으로 기대한다.

2. 사례 연구

Petrucci *et al.*은 기계번역 맥락에서 text-to-knowledge sequence 추출을 구현하였는데[7], 여기서는 이 사례를 살펴보고자 한다.



(그림 1) Attention Mechanism 전체 구조.(© Petrucci et al., 2018, redrawn)

사례에서 사용한 딥러닝 기반 기계번역 모델은 두 개의 신경망 층으로 이루어진 Seq2Seq (Sequence to Sequence) 구조[8]로 구성되어 있다.

첫 번째 층은 인코더(encoder)로 연속된 데이터를 컴퓨터가 이해할 수 있는 고차원의 벡터로 변환하는 과정을 수행한다. 두 번째 층은 디코더(decoder)로 인코더에서 출력된 벡터를 번역하여 다시 데이터로 변환한다[9]. Recurrent encoder-decoder scheme[10]은 인코더 층에서 RNN(Recurrent Neural Network)구조를 기반으로 구성된다. RNN 구조는 인코더 층에서 입력 데이터에 대하여 마지막 타임 스텝의 결과 벡터(context vector)만을 이용하여 학습을 진행한다. 이때, 벡터는 X_1 부터 X_{Tx} 까지의 모든 입력을 고정된 길이로 표현되는데 긴 문장일 경우 정보의 손실이 발생하여 출력 결과의 품질이 떨어지는 한계가 있다.

Attention Mechanism ([11], 그림 1)은 기존 RNN 기반의 기계번역 모델에서 발생하는 정보 손실과 기울기 소실 문제를 해결하여 인코딩 이후 모든 타임 스텝의 결과 벡터를 이용하여 디코딩을 진행한다. 본 연구에서 차용한 모델의 경우, 디코딩 층 이후 온톨로지 단어 또는 관계를 출력할지 결정하는 별도의 층(switch)을 통해 디코더에서 출력된 값을 기반으로 단어 또는 논리적 기호를 최종 결과로 출력한다[7].

본 논문에서는 사례[7]에서 제안된 데이터셋과 실험 설정을 참고하여 실험을 진행하였다. 데이터셋은 사례에서 제시된 다양한 예시 중 2k 크기의 문장-타겟 쌍으로 구성된 닫힌 데이터셋(2k-closed)을 사용하였다. 데이터셋은 test 데이터셋(300M, 300 pairs)와 validation 데이터셋(1700M, 1700 pairs)으로 분할하여 어텐션 메커니즘 기반 기계번역 모델에 입력된다. 학습 이후 evaluation 데이터를 통해 온톨로지 변환 결과를 확인한다. evaluation 데이터셋은 사례[7]에서 제안된 30k(30000M, 30000 pairs)의 문장-타겟 쌍으로 구성된 데이터를 사용한다. 입력 문장은 동일한 문장 구조 내에서 닫힌 집합 내의 단어로 치환되는 과정을 통해 생성되어 문법적 구조는 올바르지만 사람이 해석하기 어색한 문장이 존재한다. 출력 결과는 임베딩(embedding)된 숫자 형태로 출력되어 이후 자체적인 별도의 매핑(mapping) 작업을 통해 결과값을 도출하였다. <표 1> 은 온톨로지 변환 결과에 대한 예시이다.

변환 결과를 통해 특정한 표현 체계에 의하여 지식 체계가 정형화되어 표현된 것을 확인할 수 있다. 예를 들어, 문장 "Every hand is also a hit of soap and a region pistol of rental"이 입력으로 주어졌을 때, hand 에 대한 설명인 hit of soap 와 region of rental 의 관계가 기호(^, :=)를 통해 표현되었다. 또한, 입력 문장 내에서 개념 간의 관계와, 변환 결과에서 논리적 기호의 의미를 비교하였을 때, 개념 사이의 관계를 나타내는 단어가 동일한 의미를 지닌 기호로 표현된 것이 확인된다. 이를 통해 어텐션 메커니즘을 통해 데이터셋 내의 text 가 logical representation 형태로 출력된다는 것을 확인할 수 있다.

<표 1> 온톨로지 변환 결과

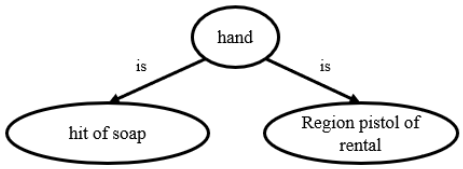
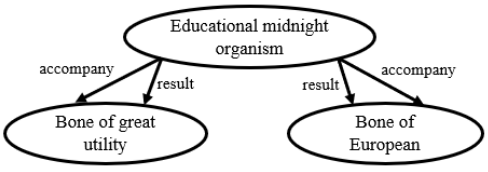
유형	값
Sentence	every hand is also a hit of soap and a region pistol of rental .
Predicted	hand := hit of soap ^ region pistol of rental
Sentence	a educational midnight organism result or accompany also bone of great utility and of European .
Predicted	educational midnight organism := E (result U accompany) . (bone of great utility ^ bone of European)

3. 고찰

사례 [7]를 통해 어텐션 메커니즘을 적용하여 문장을 논리적 기호로 표현하는 과정을 살펴보았다. 실험 결과를 통해 출력 값이 문장의 문법적 구조에 기반하여 logical representation 의 형태로 출력되는 것을 확인하였다.

여기서, 변환되어 출력된 문장의 논리적 관계는 논리 기호의 성격에 따라 개념 사이의 계층 구조로 표현될 수 있으며, 이를 기반으로 개념 간의 관계를 정점과 간선을 활용하여 지식 그래프(Knowledge Graph)[12]의 형태로 표현할 수 있다. <표 2>는 <표 1>의 예시를 기반으로 지식 그래프 형태로 가공한 결과이다.

<표 2> 지식 그래프 생성 결과

Logic	hand := hit of soap ^ region pistol of rental
Knowledge Graph	
Logic	educational midnight organism := E (result U accompany) . (bone of great utility ^ bone of European)
Knowledge Graph	

첫 번째 결과를 보자. 논리적 기호로 표현된 knowledge sequence 를 활용, 각 개념(hand, hit of soap, region pistol of rental)사이의 관계(is)를 노드-관계-노드로 표현된 tree 구조로 표현된 (중첩된) 지식 그래프

로 표현할 수 있다. 이때 지식 그래프에서 노드-관계-노드 연결은 계층 구조에 기반하여 head-relation-tail 형식의 온톨로지 형식으로 표현할 수 있다.

즉, 서술된 일련의 과정에 따라 문장을 logical representation 단계를 거쳐 온톨로지 형식으로 변환할 수 있다. 문장 세트가 입력되었을 때, (i) 기계 번역의 관점에서 문장을 논리 기반의 지식 시퀀스로 변환하고, (ii) 지식 표현 간 논리적 구조를 지식-관계-지식 형태의 계층적 지식 그래프 구조로 치환하여 (iii) 노드-관계-노드 연결을 head-relation-tail 형태의 온톨로지로 최종 변환이 가능하다.

4. 결론

본 논문에서는 딥러닝 기반 기계번역 개념을 활용한 온톨로지 생성 사례를 통해 지식 정보의 이해 여부에 상관없이 문장을 logical representation 의 형식으로 변환하고, 이를 간단한 후처리를 통해 지식 그래프 형태로 가공하는 3 단계 과정을 수행하여, 온톨로지를 획득하는 실험을 진행하였다. 데이터셋은 2k 크기의 문장-타겟 쌍으로 구성되었으며, 모델은 Seq2Seq 구조를 기반으로 한 어텐션 메커니즘을 적용하여 RNN 기반의 기계번역 모델에서 발생하는 정보 손실과 기울기 소실 문제를 최소화하여 학습을 진행하였다. 3k 크기의 데이터셋을 통해 학습 결과를 확인하였으며, 문장의 문법적 구조에 따라 텍스트를 logical representation 을 반영한 결과로 출력하였다. 논리적 관계를 기반으로 지식 표현 시퀀스를 계층적 지식 그래프의 형태로 표현하였으며, 각 노드 간의 계층적 구조에 기반하여 온톨로지 형식으로 표현할 수 있다는 것을 확인하였다.

5. 향후 연구 계획

현재 구현된 온톨로지 생성 파이프라인은 기존 연구에세 제공한 데이터를 중심으로 사례를 확인한 바, 이후 새로운 학습 데이터 및 테스트 데이터셋을 기반으로 성능을 확인할 필요가 있다. 사전 연구가 기계번역의 개념으로 semantic 과 syntactic 을 고려하여 텍스트 내용의 논리적 구조를 추출한 바, 이를 사전학습된 모델이라 가정하고 학습에 사용되지 않은 새로운 텍스트를 입력하여 온톨로지 생성 결과를 확인할 수 있다. 생성 품질에 따라, 새로운 학습 데이터를 이용하여 제안한 온톨로지 생성 파이프라인을 학습시킨 후 상기 일련의 과정을 수행할 필요도 있다.

생성 품질에 대한 비교 분석 또한 요구된다. 생성한 온톨로지 결과를 최신 온톨로지 생성 연구 결과와 비교하여, 우리가 제안한 방법의 성능을 확인할 필요가 있다. 특히, 다양한 분야에서 수집한 데이터셋을 기반으로 개방형 언어추출모델(open IE, Open Information Extraction)[12] 및 기타 온톨로지 생성 모델[13-14]를 통해 온톨로지를 생성하고, 이들 연구에서 사용한 평가 지표를 통해 성능 비교를 진행한다. 대표적으로 Petrucci *et al.*에서 제안한 자동 분석 방법[7]과 더불어 Petrova *et al.* 이 제안한 사용자 평가 방법 및 지표[15]를 적용하여 본 제안의 우수성을 확인하고자 한다.

Acknowledgement

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1G1A1102683). 본 연구는 삼성미래기술육성센터의 지원을 받아 수행하였음 (No. SRFC-TC1603-52). 본 결과물은 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 사회맞춤형 산학협력 선도대학 (LINC+) 육성사업의 연구결과임.

참고문헌

[1] Choe, Gi-Seon, and Beop-Mo Ryu. "온톨로지의 구축과 학습: 상하위 관계" *Communications of the Korean Institute of Information Scientists and Engineers* 24.4 (2006): 24-30.

[2] Völker, Johanna, Pascal Hitzler, and Philipp Cimiano. "Acquisition of OWL DL axioms from lexical resources" *European Semantic Web Conference*, Springer, Berlin, Heidelberg, 2007.

[3] Manola, Frank, Eric Miller, and Brian McBride. "RDF primer" *W3C recommendation* 10.1-107 (2004): 6.

[4] Munoz, Sergio, Jorge Pérez, and Claudio Gutierrez. "Simple and efficient minimal RDFS" *Journal of web semantics* 7.3 (2009): 220-234.

[5] Choe, Ho-Seop, et al. "온톨로지 구축 방법과 사례 " *Communications of the Korean Institute of Information Scientists and Engineers* 24.4 (2006): 31-44.

[6] Olson, Judith Reitman, and Henry H. Rueter. "Extracting expertise from experts: Methods for knowledge acquisition" *Expert systems* 4.3 (1987): 152-168.

[7] Petrucci, Giulio, Marco Rospocher, and Chiara Ghidini. "Expressive ontology learning as neural machine translation" *Journal of Web Semantics* 52 (2018): 66-82.

[8] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks" *Advances in neural information processing systems*. 2014.

[9] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation" *arXiv preprint arXiv:1406.1078* (2014).

[10] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing*

systems. 2017.
[11] Singhal, Amit. "Introducing the knowledge graph: things, not strings" google blog, 2012.
[12] Stanovsky, Gabriel, et al. "Supervised open information extraction" *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018.
[13] Cimiano, Philipp, and Johanna Völker. "text2onto" *International conference on application of natural language to information systems*. Springer, Berlin, Heidelberg, 2005.
[14] Fortuna, Blaz, Marko Grobelnik, and Dunja Mladenic. "Ontogen: Semi-automatic ontology editor" *Symposium on Human Interface and the Management of Information*. Springer, Berlin, Heidelberg, 2007.
[15] Petrova, Alina, et al. "Formalizing biomedical concepts from textual definitions" *Journal of biomedical semantics* 6.1 (2015): 1-17.