ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2019 June: 25(2): 25~38 http://dx.doi.org/10.13088/jiis.2019.25.2.025

뉴럴 텐서 네트워크 기반 주식 개별종목 지식개체명 추출 방법에 관한 연구

양윤석
연세대학교 투자정보공학과
(*midosto@naver.com*)

이현준
연세대학교 산업공학과
(*2wisedeep@yonsei.ac.kr*)

오경주
연세대학교 산업공학과
(*johanoh@yonsei.ac.kr*)

정보화 시대의 넘쳐나는 콘텐츠들 속에서 사용자의 관심과 요구에 맞는 양질의 정보를 선별해내는 과정은 세대를 거듭할수록 더욱 중요해지고 있다. 정보의 홍수 속에서 사용자의 정보 요구를 단순한 문자열로 인식하지 않고, 의미적으로 파악하여 검색결과에 사용자 의도를 더 정확하게 반영하고자 하는 노력이 이루어지고 있다. 구글이나 마이크로소프트와 같은 대형 IT 기업들도 시멘틱 기술을 기반으로 사용자에게 만족도와 편의성을 제공하는 검색엔진 및 지식기반기술의 개발에 집중하고 있다. 특히 금융 분야는 끊임없이 방대한 새로운 정보가 발생하며 초기의 정보일수록 큰 가치를 지녀 텍스트 데이터 분석과 관련된 연구의 효용성과 발전 가능성이 기대되는 분야 중 하나이다. 따라서, 본 연구는 주식 관련 정보검색의 시멘틱 성능을 향상시키기 위해 주식 개별종목을 대상으로 뉴럴 텐서 네트워크를 활용한 지식 개체명 추출과 이에 대한 성능평가를 시도하고자 한다. 뉴럴 텐서 네트워크 관련 기존 주요 연구들이 추론을 통해 지식 개체명들 사이의 관계 탐색을 주로 목표로 하였다면, 본 연구는 주식 개별종목과 관련이 있는 지식 개체명 자체의 추출을 주목적으로 한다. 기존 관련 연구의 문제점들을 해결하고 모형의 실효성과 현실성을 높이기 위한 다양한 데이터 처리 방법이 모형설계 과정에서 적용되며, 객관적인 성능 평가를 위한 실증 분석 결과와 분석 내용을 제시한다. 2017년 5월 30일부터 2018년 5월 21일 사이에 발생한 전문가 리포트를 대상으로 실증 분석을 진행한 결과, 제시된 모형을 통해 추출된 개체명들은 개별종목이 이름을 약 69% 정확도로 예측하였다. 이러한 결과는 본 연구에서 제시하는 모형의 활용 가능성을 보여주고 있으며, 후속 연구와 모형 개선을 통한 성과의 제고가 가능하다는 것을 의미한다. 마지막으로 종목명 예측 테스트를 통해 본 연구에서 제시한 학습 방법이 새로운 텍스트 정보를 의미적으로 접근하여 관련 주식 종목과 매칭시키는 목적으로 사용될 수 있는 가능성을 확인하였다.

주제어 : 뉴럴 텐서 네트워크, 검색엔진, 주식시장, 텍스트 마이닝, 금융 빅데이터 분석

논문접수일 : 2019년 2월 19일 논문수정일 : 2019년 5월 4일 게재확정일 : 2019년 5월 24일
원고유형 : 일반논문 교신저자 : 오경주

1. 개요

구글 검색엔진에 적용된 구글 지식저장소 (google knowledge vault) 프로젝트와 마이크로소프트의 Bing 검색엔진에 적용된 사토리 (Satori) 프로젝트 등 시멘틱 검색을 위한 지식기반 (knowledge-based) 애플리케이션의 상용화가 이루어지면서, 검색엔진이나 음성비서, 챗봇 같은 대화형 어시스턴트 서비스의 핵심인 지식기반기술이 주목받고 있다 (Dong et al., 2014). 지식기반기술은 웹상에서 텍스트를 단순한 스트링 정보로 처리하는 것이 아니라 의미적으로 접근하는 것을 말하며, 이는 텍스트가 가지는 지식을 기계가 이해할 수 있는 형태로 전환해야 한다는

문제를 필연적으로 수반한다. 이러한 기계가 이해할 수 있는 형태로의 지식표현과 관련된 다양한 선행연구가 시도되었고, 최근에는 RDF (Resource Description Framework) triple과 월드와이드웹 컨소시엄 (World Wide Web Consortium, W3C)에서 만든 웹 온톨로지 언어인 OWL (Web Ontology Language) 등의 마크업 언어 (Markup Language) 표현방식이 보편적으로 활용되고 있다 (Etzioni et al., 2011). 자연어로 이루어진 사람의 지식은 이런 마크업 언어로 표현된 온톨로지 구성을 통해 기계가 의미적으로 해석하고 처리할 수 있는 형태로 변환된다.

RDF나 OWL과 같은 지식표현은 표준화 방식에 차이가 있지만, 두 개의 노드 정보와 두 노드를 잇는 한 개의 관계 정보를 가지는 트리플 형식을 취한다는 공통점이 있다. 트리플 구조로 한 단위의 지식정보를 담아내고, 많은 트리플이 서로 연결되면서 지식 데이터베이스인 온톨로지가 만들어지게 된다. 온톨로지를 구성하는 방법은 사람이 직접 온톨로지를 만드는 방식과 알고리즘에 의해 자동화된 방식이 있는데, WordNet이나 Wikidata, 그리고 Freebase처럼 전문가나 집단지성을 이용하여 온톨로지를 구축하는 방법은 한정된 분야 내에서 높은 완성도의 온톨로지를 구축할 수 있다는 장점이 있으나, 그에 수반되는 비용과 시간 때문에 확장성이 낮다는 단점이 있다 (Bollacker et al., 2008). 특히 필요한 지식기반 데이터가 넓은 범위에 걸쳐있거나, 새로운 지식이 지속적이고 빠르게 발생하는 분야에서는 사람이 직접 지식기반을 구축하기에 더욱 난해하다. 따라서, 사람이 직접 온톨로지를 구축하지 않고 자동화된 지식 추출을 통해 온톨로지를 구축하는 방법이 꾸준히 연구되어왔다 (Lee and Sohn, 2013; Kim, 2017; Kim and Lee, 2019). 자동화된 온톨로지 구축 방법을 온톨로지학습 (Ontology Learning)이라고 하며, 사전적인 스키마 또는 외부로부터 주어지는 온톨로지가 없는 상태에서 기계가 텍스트로부터 구조화된 지식을 추출하는 것을 이른다 (Navigli and Velardi, 2004; Banko et al., 2007). 일반적으로 사전적인 스키마 없이 이루어지는 자동지식추출 모델은 비지도 학습의 특성을 갖게 된다 (Etzioni et al., 2011).

지식 추출 자동화는 필요한 지식의 범위와 목적이 광범위한 분야 또는 새로운 지식이 지속적으로 생성되는 공학 및 의학 분야 등에서 특히 활발하게 연구된다 (Nair, 2017). 금융 분야도 새로운 데이터 발생의 속도와 영향력을 고려하면 지식 추출 자동화의 필요성이 분명하다 (Kim et al., 2012). 특히 본 연구의 대상인 주식 투자 분야에서 자동화된 지식 추출이 가질 수 있는 효용을 고려하면, 개별 주식 종목에 대한 지식정보를 보유한 개체명과 그 트리플들을 사전적으로 파악함으로써 주식 정보 검색의 유효성을 높일 수 있다는 가정을 할 수 있다. 가령, 호텔신라 주식에 대해 학습한 결과 호텔신라에 대한 지식정보를 내포하고 있는 것으로 보이는 개체명들이 존재한다고 할 때, 이 개체명들과 교집합이 많은 뉴스 정보를 검색 결과에 상위 노출 시킬 수 있을 것이다. 특정 뉴스에서 호텔신라라는 단어가 포함되어 있지 않았을지라도 이 개체명들과 교집합이 많은 뉴스가 있으면 이를 호텔신라 검색 결과에 상위 노출시키는 것도 가능하다. 반대로 호텔신라 단어를 포함하고 있으나 호텔신라 주식에 대해 학습된 지식정보와 교집합이 적은 뉴스들은 상위 노출에서 제외할 수 있을 것이다. 이와 더불어 지속적인 학습 과정 업데이트를 통해 관련 개체명의 변화를 관찰한다면 호텔신라와 관련된 이슈의 변화를 더 쉽게 파악할 수도

있을 것이다.

본 연구에서는 모형의 성능 평가에 활용된 종목명 예측 테스트를 통해 의미적인 검색 목적으로 제시된 모형을 활용할 수 있는가를 판단하고자 한다. 그러나 앞선 방식과 같이 자동으로 지식 추출을 시도하는 경우에 아래와 같은 몇 가지 현실적 어려움에 직면한다.

- 자연어 텍스트로부터 트리플 정보 자동 추출
  : 뉴럴 텐서 네트워크 모형을 통해 지식정보를 학습하기 위해서는 우선 학습 대상이 되는 트리플 데이터를 자연어 텍스트로부터 추출하는 과정이 필요하다. 그러나 많은 경우 복잡한 언어학적인 전처리 과정이나 휴리스틱 접근 과정을 거쳐야 하는 경우가 많다. 주식 관련 지식정보는 다양한 분야에 걸쳐 다층적으로 분포하고 있고, 시간 경과에 따라 변화가 많다. 따라서 사람에 의한 전처리 과정을 지양하면서도 주식 관련 지식정보를 수용할 수 있는 형태의 자동화된 트리플 추출 방법이 필요하다.
- 개체명 추출 목적의 트리플 학습 방법론
  : 자연어 텍스트로부터 추출된 트리플 정보는 학습 과정을 거쳐 진위 확률값을 얻을 수 있다. 트리플은 두 개의 개체명과 이를 잇는 관계 정보로 구성되어있으며, 학습 모델은 이러한 트리플 구조 학습이 가능해야 한다. Socher 등과 Nickel 등이 연구한 뉴럴 텐서 네트워크 모델은 대표적인 트리플 데이터 학습에 특화된 모델로, 본 연구에서도 Socher의 방법론을 활용한다 (Socher et al., 2013; Nikel et al., 2015). 다만 기존의 뉴럴 텐서 네트워크 모델은 이미 존재하는 개체명 간의 관계 추론을 목표로 하고 있다. 즉, 사전에 데이터베이스로부터 주어진 개체명 정보를 학습 모델의 입력값으로 활용한다. 그러나 본 연구에서는 이와 반대로 개별 주식과 밀접하게 연관된 개체명 정보를 학습의 결과값으로 추출해야 한다.
- 지식정보 학습 방법론에 대한 평가
  : 본 연구에서 제시하는 모형은 각 트리플의 진위에 대한 사전적인 라벨이 없으므로 비지도 학습에 해당하고, 학습의 대상이 사용자의 관점과 목적에 따라 정의가 달라질 수 있는 지식정보라는 점에서 객관적인 평가가 어렵다. 따라서 현실적 효용과 목적을 반영한 구체적인 문제 정의가 필요하다.

본 연구는 자동으로 지식정보를 추출하는 경우 직면하게 되는 위와 같은 과제를 인지하고, 이를 극복하기 위해 뉴럴 텐서 네트워크 모형을 통한 실용적이고 자동화된 주식 개별종목 관련 지식 개체명 추출 모형을 제시하고자 한다.

본 연구가 가지는 의의는 다음과 같다. 첫 번째로, 본 연구는 주식 개별종목과 관련된 자연어 텍스트로부터 복잡한 언어학적 전처리나 휴리스틱 접근 없이 트리플 데이터를 추출하는 방법론을 제시한다. 주식 투자 분야처럼 복잡하고 광범위한 정보들을 수용하기 위해서는 오히려 유연하고 포괄적인 단순한 트리플 형태가 적합할 수 있다. 따라서 본 연구에서는 개체명 인식을 통해 추출된 개체명들 사이의 단순한 조합만으로 트리플을 생성하는 방법을 제시한다. 두 번째로, 주식 개별종목 관련 개체명 정보 추출을 목적으로 뉴럴 텐서 네트워크 모델의 활용 가능성을 확인한다. 기존 뉴럴 텐서 네트워크와 관련된 연구는 이미 존재하는 트리플 데이터베이스 내에서 관계를 추론하는 목적으로 이루어졌다. 그러나 본 연구에서는 사전적인 트리플 데이터베이스 없이 개체명을 추출하는 것을 목적으로 활용될 수 있는 가능성을 제시한다. 세 번째로, 정의가 쉽지 않고 모호한 개념의 지식정보를 추출하는

본 연구 모형에 대한 합리적 성능평가 방법을 제시한다.

이어지는 본 논문은 구성은 다음과 같다. 2절에서는 자동 지식 추출과 관련된 선행연구를 기술하였고, 3절에서는 주식 개별종목 관련 지식 트리플 추출을 위한 문제의 정의, 그리고 연구 방법론에 대한 정보를 제공한다. 4절과 5절에서는 각각 제시된 방법론을 활용한 실증 분석과 연구의 결론을 기술하고 있다.

2. 선행연구

지난 수십 년에 걸쳐 다양한 방법론을 활용한 자동 지식추출기 관련 연구가 이루어져 왔으며, 주로 지식표현 측면에서 왜곡되지 않은 지식을 내포한 트리플을 추출하는 방법을 제시하는 것에 초점이 맞춰져 왔다. 목표하는 특정 형태의 트리플 패턴 데이터를 구성하거나, 특정 형태의 패턴이 추출되도록 휴리스틱 규칙을 적용하기도 했다. Fader 등은 추출기 ReVerb에서, 그리고 Mesquita 등은 추출기 EXEMPLAR에서 수작업 추출패턴 (hand-crafted extraction rules)을 적용해 트리플을 추출하였다 (Fader et al., 2011; Mesquita et al., 2013). 한편 2013년 Del Corro와 Gemulla, 그리고 2014년 Schmidek와 Barbosa의 연구에서는 복잡한 구문을 트리플로 전환시키기 용이한 단순한 구문으로 전환시키는 과정을 통해 모형의 오류를 최소화하는 연구를 진행하였다 (Del Corro and Gemulla, 2013; Schmidek and Barbosa, 2014). 이와 유사하게 Mausam 등은 트리플 구조가 주로 문장구조에서 1, 2, 3형식에 적합하여 수식절 등을 포함한 복문구조에 취약한 점을 극복하고자 하였다 (Mausam, 2016). Banko 등과 Schmitz 등은 러닝 기법을 통해 알고리즘 내에서 스스로 라벨링이 이루어져 추출패턴이 결정되는 방법을 시도하였으며, 이를 통해 특정 도메인에 국한되지 않은 범용 패턴 추출을 시도하였다 (Banko et al., 2007; Schmitz et al., 2012).

상기 서술된 연구들이 주로 자연어 텍스트를 기계가 인식할 수 있는 형태의 트리플 데이터로 전환하는 지식표현에 초점을 맞추었다면, 다른 한편에서는 이미 존재하는 트리플 기반 데이터베이스 내에서 논리적·의미론적 추론을 통해 개체명 간의 새로운 관계, 즉 새로운 사실의 발견하는 연구가 진행되어왔다. 대표적인 두 연구가 2013년 Socher 등과 2015년 Nickel 등에 의해 이루어진 연구이다 (Socher et al., 2013; Nikel et al., 2015). 두 연구 모두 지식기반 데이터를 3차원 텐서 (tensor)로 정의한다는 공통점이 있다. 2개의 개체명이 하나의 관계 (relational edge)로 연결돼있는 트리플은 2개의 개체 차원과 하나의 관계 차원으로 매핑된 3차원 큐브 형태의 텐서 구조로 접근하기에 용이하다. 다만 두 연구에서 정의하는 텐서의 형태는 차이가 있는데, Nickel 등의 연구에서는 원천 트리플 데이터 전체가 하나의 거대한 텐서 속에 두 개의 개체명 축과 한 개의 관계 축에 매핑되며, 이는 결국 텐서 속에서 하나의 트리플이 하나의 점 형태가 되는 것을 의미한다. 이에 반해 Socher 등의 연구에서는 하나의 트리플에서 두 개의 개체명이 가지는 벡터값이 개체명 축에 매핑되며, 이는 하나의 트리플이 하나의 텐서값을 가지는 형태를 의미한다. 이런 정의에 따라 Nickel 등은 대규모 텐서 속에서 유의미한 지식 또는 사실을 추출하기 위해 텐서 팩터라이제이션 (factorization)이라는 잠재모형 (latent model) 접근법을 활용했다. 반면 Socher

등은 텐서에 대한 인공신경망 모델인 뉴럴 텐서 네트워크 모델을 활용하며, 이미 만들어져 있는 트리플 데이터베이스를 대상으로 새로운 노드 간의 관계를 발굴하고자 하였다.

3. 연구모형

서론에서 언급하였듯, 자동지식추출과 관련된 연구는 다양한 문제와 한계점에 부딪힌다. 이러한 어려움 속에 텍스트로부터 지식을 추출하는 것에 관한 문제를 명확하고 실용적으로 정의하기는 쉽지 않다. 텍스트로부터 추출된 트리플 데이터가 과연 지식을 포함하고 있는가에 대한 근원적 의문이 제기될 수 있으며, 이런 의문이 존재하는 이유는 트리플 데이터가 지식을 내포하고 있다고 볼 수 있는, 즉 참이라고 볼 수 있는 확증 데이터가 없기 때문이다. Freebase나 WordNet처럼 이미 라벨링된 데이터베이스를 활용하거나, 전문가 집단 또는 Amazon Mechanical Turk 등을 통한 크라우드 라벨링을 활용하는 경우를 제외하면 지식 트리플 추출은 기본적으로 비지도 학습의 성격을 가진다. 따라서 자동지식추출은 비지도 학습의 특성상 그 성과의 평가가 어렵고, 자칫 맹목적인 트리플 텍스트 데이터 추출의 반복에 지나지 않을 수 있다.

본 연구에서는 사전적인 말뭉치나 추출된 지식의 참·거짓 여부에 대한 라벨 없이 주식 개별종목에 대한 지식을 추출하고 추출 결과에 대한 평가를 진행한다. 현업에 종사하는 전문가인 주식 애널리스트는 본인이 분석하는 종목과 관련된 핵심적인 개체명 또는 개체명 조합을 알고 있을 것이며, 이를 이용하여 혼재된 다양한 정보 속에서 본인이 분석하는 종목과 연관된 정보와 콘텐츠를 구분하여 찾아낼 수 있을 것이다. 따라서 만약 자동지식추출기로 추출된 정보를 이용하여 혼재된 정보 속에서 특정 종목과 관련이 높은 콘텐츠를 분류해낼 수 있다면, 이는 정보 분류 작업에 관하여 전문가처럼 지식을 추출했다고 간주할 수 있을 것이다. 지식은 정의와 활용 방법에 따라 다양한 범주가 있겠으나, 본 연구에서는 주식 투자 관점에서 개별종목과 관련이 높은 콘텐츠를 색인화하여 분류해내는 것을 목표로 한다.

이러한 접근은 앨런 튜링의 튜링 테스트와 유사한 방식이다 (Epstein et al., 2009). 튜링은 지능의 정의조차 합의가 쉽지 않은 모호한 개념임에도 불구하고 기계가 지능을 가졌다고 판단할 수 있는 테스트 방식을 제시하였다. 즉, “지능을 가진 실험 참가자가 기계의 응답과 사람의 응답을 구분하지 못한다면 기계가 지능을 가졌다고 간주할 수 있다”라는 논리를 통해 기계의 지능 보유 여부에 대한 테스트 방법을 제시한 것이다. 이 방법은 기계의 지능 보유 여부에 관한 문제를 우회하여 해결해주었다.

일반적인 지식 추출에서의 트리플은 두 개의 개체명과 두 개체명 사이의 관계로 정의된다. 그러나 본 연구에서는 트리플의 관계 부분을 종목명으로 치환하여 사실상 세 개의 개체명으로 구성된 트리플을 정의한다. 이는 트리플 안에 존재하는 두 개체명들이 일반적으로 서로 관련이 있다기보다는, 특정 주식 종목의 관점에서 봤을 때 두 개체명이 서로 관련이 있다고 보는 것이다. 예를 들면, 면세점과 사드 (THAAD)라는 두 개체명은 일반적으로는 직접적으로 관련이 높다고 보기 어려울 수 있으나, 호텔신라라는 종목 관점에서는 이벤트로서 깊이 연관되어 있다. 따라서 트리플에서 개별 종목명을 관계로

정의하는 것은 해당 종목에 대한 투자 관점에서 보았을 때 두 개체명이 깊은 연관성을 가진다는 의미가 된다.

본 연구는 증권사의 주식 개별종목 리포트에서 추출된 개체명들을 트리플 형태로 전환하고 뉴럴 텐서 네트워크 모델을 학습시킨다. 학습을 마친 뉴럴 텐서 네트워크 스코어 함수는 해당 종목과 깊이 연관된 트리플일수록 높은 값을 가짐으로써 핵심 개체명을 보유한 트리플을 파악할 수 있게 해준다. 따라서 본 연구에서 뉴럴 텐서 네트워크의 사용 목적은 핵심 개체명 또는 개체명 조합을 파악하는 것이라고 할 수 있다.

반면 Socher 등과 Nickel 등의 기존 연구는 주로 이미 존재하는 노드에 대해 새로운 관계를 찾아내는 것이 목적이었다. 본 연구에서는 주식 개별종목이 주어졌을 때, 해당 종목에 대한 지식을 내포하고 있는 개체명을 추출하고, 그 단어들의 유의미한 연결 조합이라고 할 수 있는 지식 트리플 (개체명1-종목-개체명2)를 만드는 것을 목적으로 한다. 일반적으로 지식을 추출하고자 하는 대상 분야에 대해 충분한 말뭉치가 만들어져 있는 경우는 드물며, 따라서 본 연구에서는 개체명 추출을 위해 대표적인 개체명인식기 (NER, Named Entity Recognition)인 꼬꼬마 (KKMA)를 활용한다 (Lee et al., 2010).

추출된 개체명들을 대상으로 트리플을 생성하고, 생성된 트리플을 뉴럴 텐서 네트워크에 적용하여 신경망 모형을 구축하였다. 총 N개의 개별종목에 대하여, 종목별로 M_N개의 각 증권사 리포트를 활용하였다. 각 리포트 내의 동일한 문장에 속해있는 개체명 $e_1, e_2 \in R^N$을 연결하여 Table 1의 $T(t,Doc,t)$와 같은 트리플을 생성시킨다. 서로 다른 문장에 속해있는 의미적 연결성이 떨어지는 개체명 쌍을 제거하고, 원활한 모형 학습을 위해 같은 문장에 속해있는 개체명들로만 트리플을 생성시켰다. 각 개체명은 N개의 종목에 대해 추출된 총 개체명 수에 따라 원-핫 인코딩 (One-hot Encoding)을 통해 벡터화시키며, 문장이 속한 리포트에 관련된 종목을 관계 $R \in N$으로 간주하여 (e_1, R, e_2) 와 같은 형태의 트리플을 생성시킨다 (Zhang et al., 2015; Turian et al., 2010).

이렇게 생성된 트리플에 대해서, 개별종목별로 뉴럴 텐서 네트워크 모형을 구축한다 (Liu et al., 2015). 뉴럴 텐서 네트워크 모형은 두 개의 개체명이 특정한 관계 (종목) 하에서 얼마나 긴밀하게 연관되어 있는지를 반영하며, 학습이 완료되면 전체 N개의 네트워크 모형이 구축된다 (Figure 1). 개별 트리플에 대한 이러한 평가와 반

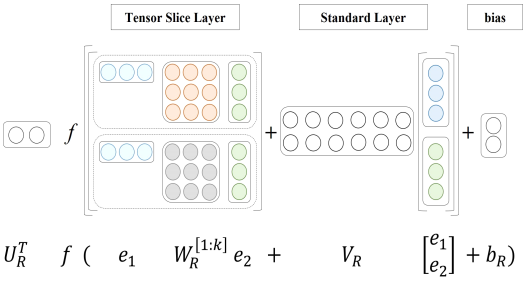
〈Table 1〉 Illustration for Graph Generation

$Documents_1$				$Documents_m$			
$Sentence_{1,1}$		$Sentence_{1,n_1}$		$Sentence_{m,1}$		$Sentence_{m,n_m}$	
$t_{1,1,1} \cdots t_{1,1,l_1}$	…	$t_{1,n_1,1} \cdots t_{1,n_1,l_{n_1}}$	…	$t_{m,1,1} \cdots t_{m,1,k_1}$	…	$t_{m,n_m,1} \cdots t_{m,n_m,k_{n_m}}$	…
$T(t_{1,1,1}, Doc_1, t_{1,1,2})$		$T(t_{1,n_1,1}, Doc_1, t_{1,n_1,2})$		$T(t_{m,1,1}, Doc_1, t_{m,1,2})$		$T(t_{m,n_m,1}, Doc_1, t_{m,n_m,2})$	
…		…		…		…	
$T(t_{1,1,l_1-1}, Doc_1, t_{1,1,l_1})$		$T(t_{1,n_1,l_1-1}, Doc_1, t_{1,n_1,l_1})$		$T(t_{m,n_m,k_{n_m}-1}, Doc_1, t_{m,n_m,k_{n_m}})$		$T(t_{m,1,k_1-1}, Doc_1, t_{m,1,k_1})$	

영은 아래의 스코어 함수 $g(e_1, R, e_2)$를 활용하여 이루어진다. 스코어 함수는 특정 트리플이 입력되면 스코어 값을 산출하는 역할을 하며, 그 값이 클수록 개체명이 해당 종목(R)과 관련 있다고 해석할 수 있다. 함수 f는 쌍곡탄젠트(hyperbolic tangent)이다. 본 논문에서는 텐서 $W_R^{[1:k]} \in R^{N^* N^* k}$에서 $k = 2$인 쌍선형 (bilinear) 텐서 곱을 적용한다.

$$g(e_1, R, e_2) = u_R^T f(e_1 W_R^{[1:k]} e_2 + V_R \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} + b_R)$$

제시된 모형의 학습은 Socher 등이 선행연구와 같이 contrastive max-margin objective 함수를 통해 이루어지며, 최적화 과정에서도 Socher 등이 선행연구에서 사용했던 L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) 알고리즘을 활용하였다.



〈Figure 1〉 Neural Tensor Network

4. 실증 분석

제시된 모형의 실증 분석에는 2017년 5월 30일부터 2018년 5월 21일까지 1년간 국내 증권사에서 발행된 26,667개의 종목 분석 리포트 중 발행 빈도 기준으로 상위 30개 종목과 관련된 5,600개의 리포트를 활용하였다. 총 5,600개의 리포트 중 실험 기간의 절반에 해당하는 2017년 5월 30일부터 2017년 11월 24일 사이에 발행된 3,074개의 리포트를 학습데이터로 사용하고, 나머지 2,526개의 리포트를 모델 검증에 사용했다.

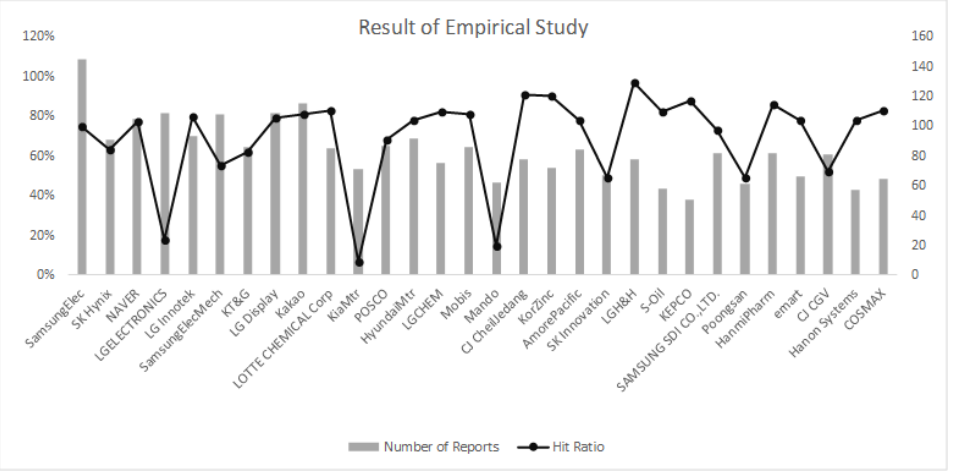
학습 대상이 되는 리포트 3,074개를 종목별로 취합한 뒤, 각 리포트를 문장 단위로 개체명인식기에 적용하여 개체명을 추출하였다. 그러나 추출된 개체명 중에는 지식이나 의미를 내포하고 있다고 보기 어려운 단어들도 많으며, 종목별로 추출된 개체명 개수 차이 때문에 벡터화 과정에서 차원이 달라지는 문제가 있다. 따라서 종목별 발생빈도가 높은 상위 100개 개체명으로 제한하여 차원을 일치시키는 제약조건을 설정하였다. 결과적으로 30개 종목에서 추출된 총 3000개의 개체명을 원-핫 인코딩을 통해 벡터화시켰으며, 앞선 Table 1과 같은 방법으로 트리플을 생성하여 학습을 위한 입력 데이터가 완성된다.

구축된 모형의 성능 검증을 위해, 리포트 상의 종목명이나 종목코드가 삭제된 상태에서 모형이 스스로 개체명들로부터 종목을 예측하도록 하였다. 테스트 데이터로부터 추출된 트리플을 종목별로 학습된 30개의 스코어 함수에 적용하여 스코어 값을 산출하며, 결국 개체명은 산출된 스코어 값이 가장 큰 함수의 종목과 관련된 콘텐츠라고 해석할 수 있다. 예측력 검증 결과, 총 30개 종목의 2,526개 리포트에 대한 예측 적중률 (Hit Ratio)은 평균 약 69.3%를 기록하였다. 이는 모형을 구축하는 과정에서 설정된 현실적인 제약조건들에도 불구하고 유의미하게 높다고 할 수 있다 (Table 2).

〈Table 2〉 Result of Empirical Study

	Training	Test
Period	2017.05.30. ~ 2017.11.24	2017.11.25. ~ 2018.05.21
Number of Report	3,074	2,526
Average Hit Ratio	82.2% (2,469/3,074)	69.3% (1,734/2,526)

30개 종목별 예측 성과를 살펴보면, LG전자, 기아자동차, 그리고 만도 3개 종목의 예측률만 현저하게 낮음을 그래프로 확인할 수 있다 (Figure 2). 이러한 결과는 유사 종목과의 간섭 효과와 새로운 지식의 발생 등을 원인으로 작용했을 수 있다.



〈Figure 2〉 Result of Empirical Study by Stocks

5. 결론

넘쳐나는 정보 속에서 관심 있는 주제에 부합하는 정보를 효과적으로 찾는 일의 중요성은 확대되고 있다. 구글, 빙 등의 검색엔진들은 시멘틱 트리플 데이터 구축과 적용으로 사용자의 쿼리 (query)에 대해 더욱 의미적인 접근을 시도한다. 이러한 접근 방법은 기존 키워드 중심 검색 방식보다 사용자의 의도에 더 부합하는 방식으로, 사용자의 정보 접근성을 한 차원 높이려는 대표적인 시도이다. 본 연구 역시 주식 투자 분야에 대해 기존의 키워드 중심의 검색을 넘어 사용자의 의도에 더 부합하는 정보 탐색이 이루어질 수 있는 개선된 방법론을 제시하고자 하였다. 또한, 주식 투자 분야의 경우에는 새로운 정보가 끊임없이 생성되고, 투자자 관점에서는 일반적으로 새롭게 생성된 정보일수록 더 중요하다는 점을 고려하여 자동화된 지식 추출 방법론을 제시하고자 하였다.

본 연구에서는 개별 주식 분석 보고서로부터 자동화된 지식정보를 추출하기 위해 개체명 인식과 문장 단위의 트리플 생성 알고리즘을 활용

하였으며, 추출된 트리플 데이터에 대하여 뉴럴 텐서 네트워크 모델을 구축하여 실증 분석을 진행하였다. 30개 종목에 대해 종목명 예측 결과, 테스트 기간에 속한 총 2,526개 리포트에 대한 종목명 예측 적중률은 평균 약 69%를 나타냈다. 결과적으로 제시된 방법론은 개별 주식의 주요 개체명 데이터를 통해 새롭게 주어진 텍스트가 어느 종목에 관련된 내용인지 파악 가능한 수준이며, 더욱 효과적인 정보 검색 방법론으로써 활용 가능성이 있다고 판단된다. 다만 몇 개의 종목에 대한 예측력이 현저하게 낮다는 점 등에서 제시된 방법론은 보완할 여지를 확인할 수 있으며, 이러한 오차는 영어와 합성어 등으로 구성된 개체명을 처리하는 과정에서의 문제점, 유사 종목과의 간섭 효과, 그리고 새로운 지식의 추가적인 발생 등의 원인이 작용했을 것으로 예상된다. 상기된 요인들은 뉴럴 텐서 네트워크 모델의 문제가 아니라 입력 데이터의 질과 관련된 문제로, 개체명인식기 선택이나 개체명사전 구축 등을 통한 추가 연구를 통해 극복하기 위한 후속 연구가 필요할 것이다. 이러한 실증 연구 결과들을 통해 우리는 종목명 예측 테스트를 통해 본 연구에서 제시한 학습 방법이 새로운 텍스트 정보를 의미적으로 접근하여 관련 주식 종목과 매칭시키는 목적으로 사용될 수 있는 가능성을 확인하였다.

본 연구는 복잡한 언어학적인 분석이나 휴리스틱 접근을 지양하고 개별 주식에 관하여 실용적인 트리플 데이터 추출 방법을 제시하였다는 의의가 있다. 개체명의 인식과 조합 알고리즘만을 거쳐 자연어 텍스트로부터 자동화된 트리플 데이터를 추출하는 것은 금융 투자 분야처럼 정보의 흐름이 빠르고 방대한 분야에서 지속적인 정보 업데이트에 필요한 가장 현실적인 방법이라고 할 수 있다. 특히 문장 단위로 트리플 데이터를 생성함으로써 문서 내에서 서로 가깝게 존재하는 유효한 지식 정보가 추출될 수 있는 가능성을 높였다. 또한 개체명 추출 목적으로 뉴럴 텐서 네트워크 모델의 활용 가능성을 제시하였는데, 이는 이미 존재하는 데이터 속에서 관계를 추론하는 목적으로 사용된 선행 연구들과 차별점이 있다. 마지막으로 비지도 학습의 특성을 가진 지식정보 추출 문제에 대하여 적절한 성능 평가 방법을 제시하였다. 본 연구에서는 종목명 예측 문제를 정의하여 모형의 성능을 평가할 수 있도록 하였는데, 이는 추출된 트리플의 진위 여부를 평가를 전문가가 직접 검토하는 것이 아닌 우회적인 방법으로 수행할 수 있다는 가능성을 제시한다.

참고문헌(References)

Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, “Open information extraction from the web,” *IJCAI*, Vol.7, (2007), 2670~2676.

Bollacker, K., C. Evans, P. Paritosh, T. Sturge, and J. Taylor, “Freebase: a collaboratively created graph database for structuring human knowledge,” *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, (2008), 1247~1250.

Del Corro, L., and R. Gemulla, “Clausie: clause-based open information extraction,” *Proceedings of the 22nd international conference on World Wide Web*, (2013), 355~366.

Dong, X., E. Gabrilovich, G. Heitz, W. Horn, N.

Lao, L. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, (2014), 601~610.

Epstein, R., G. Roberts, and G. Beber, *Parsing the Turing test*, Springer, Dordrecht, 2009.

Etzioni, O., A. Fader, J. Christensen, S. Soderland, and M. Mausam, "Open information extraction: The second generation," *IJCAI*, Vol.11, (2011), 3~10.

Fader, A., S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," *Proceedings of the conference on empirical methods in natural language processing*, (2011), 1535~1545.

Kim, H., *Knowledge Graph*, Communication Books, 2017.

Kim, J. H., and M. Lee, "Knowledge Extraction Methodology and Framework from Wikipedia Articles for Construction of Knowledge-Base," *Journal of Intelligence and Information Systems*, Vol.25, No.1(2019), 43~61.

Kim, Y., N. Kim, and S. R. Jeong, "Stock-Index Invest Model Using New Big Data Opinion Mining," *Journal of Intelligence and Information Systems*, Vol.18, No.2(2012), 143~156.

Lee, D., J. Yeon, I. Hwang, and S. Lee, "KKMA : A Tool for Utilizing Sejong Corpus based on Relational Database," *Journal of KIISE : Computing Practices and Letters*, Vol.16, No.11(2010), 1046~1050.

Lee, H. J., and M, Sohn, "Dynamic Virtual Ontology using Tags with Semantic Relationship on Social-web to Support Effective Search," *Journal of Intelligence and Information Systems*, Vol.19, No.1(2013), 19~33.

Liu, P., X. Qiu, and X. Huang, "Learning Context-Sensitive Word Embeddings with Neural Tensor Skip-Gram Model," *IJCAI*, (2015), 1284~1290.

Mausam, M, "Open information extraction systems and downstream applications," *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, (2016), 4074~4077.

Mesquita, F., J. Schmidek, and D. Barbosa, "Effectiveness and efficiency of open relation extraction," *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (2013), 447~457.

Navigli, R., and P. Velardi, "Learning domain ontologies from document warehouses and dedicated web sites," *Computational Linguistics*, Vol.30, No.2(2004), 151~179.

Nair, S., "A Biomedical Information Extraction Primer for NLP Researchers," *arXiv preprint arXiv:1705.05437*, (2017).

Nickel, M., K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proceedings of the IEEE*, Vol.104, No.1(2016), 11~33.

Schmidek, J., and D. Barbosa, "Improving Open Relation Extraction via Sentence Re-Structuring," *LREC*, (2014), 3720~3723.

Schmitz, M., R. Bart, S. Soderland, and O. Etzioni, "Open language learning for information extraction," *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (2012), 523~534.

Socher, R., D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," *Advances in neural information processing systems*, (2013), 926~934.

Turian, J., L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," *Proceedings of the 48th annual meeting of the association for computational linguistics*, (2010), 384~394.

Zhang, X., J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, (2015), 649~657.

Abstract

A Study on Knowledge Entity Extraction Method for Individual Stocks Based on Neural Tensor Network

Yunseok Yang* · Hyun Jun Lee** · Kyong Joo Oh***

Selecting high-quality information that meets the interests and needs of users among the overflowing contents is becoming more important as the generation continues. In the flood of information, efforts to reflect the intention of the user in the search result better are being tried, rather than recognizing the information request as a simple string. Also, large IT companies such as Google and Microsoft focus on developing knowledge-based technologies including search engines which provide users with satisfaction and convenience. Especially, the finance is one of the fields expected to have the usefulness and potential of text data analysis because it's constantly generating new information, and the earlier the information is, the more valuable it is. Automatic knowledge extraction can be effective in areas where information flow is vast, such as financial sector, and new information continues to emerge. However, there are several practical difficulties faced by automatic knowledge extraction. First, there are difficulties in making corpus from different fields with same algorithm, and it is difficult to extract good quality triple. Second, it becomes more difficult to produce labeled text data by people if the extent and scope of knowledge increases and patterns are constantly updated. Third, performance evaluation is difficult due to the characteristics of unsupervised learning. Finally, problem definition for automatic knowledge extraction is not easy because of ambiguous conceptual characteristics of knowledge.

So, in order to overcome limits described above and improve the semantic performance of stock-related information searching, this study attempts to extract the knowledge entity by using neural tensor network and evaluate the performance of them. Different from other references, the purpose of this study is to extract knowledge entity which is related to individual stock items. Various but relatively simple data processing methods are applied in the presented model to solve the problems of previous researches and to enhance the effectiveness of the model. From these processes, this study has the following three

* Department of Investment Information Engineering, Yonsei University
** Department of Industrial Engineering, Yonsei University
*** Corresponding Author: Kyong Joo Oh
Department of Industrial Engineering, Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea
Tel: +82-2-2123-5720, Fax: +82-2-364-7807, E-mail: johanoh@yonsei.ac.kr

significances. First, A practical and simple automatic knowledge extraction method that can be applied. Second, the possibility of performance evaluation is presented through simple problem definition. Finally, the expressiveness of the knowledge increased by generating input data on a sentence basis without complex morphological analysis. The results of the empirical analysis and objective performance evaluation method are also presented.

The empirical study to confirm the usefulness of the presented model, experts' reports about individual 30 stocks which are top 30 items based on frequency of publication from May 30, 2017 to May 21, 2018 are used. the total number of reports are 5,600, and 3,074 reports, which accounts about 55% of the total, is designated as a training set, and other 45% of reports are designated as a testing set. Before constructing the model, all reports of a training set are classified by stocks, and their entities are extracted using named entity recognition tool which is the KKMA. for each stocks, top 100 entities based on appearance frequency are selected, and become vectorized using one-hot encoding. After that, by using neural tensor network, the same number of score functions as stocks are trained. Thus, if a new entity from a testing set appears, we can try to calculate the score by putting it into every single score function, and the stock of the function with the highest score is predicted as the related item with the entity. To evaluate presented models, we confirm prediction power and determining whether the score functions are well constructed by calculating hit ratio for all reports of testing set.

As a result of the empirical study, the presented model shows 69.3% hit accuracy for testing set which consists of 2,526 reports. this hit ratio is meaningfully high despite of some constraints for conducting research. Looking at the prediction performance of the model for each stocks, only 3 stocks, which are LG ELECTRONICS, KiaMtr, and Mando, show extremely low performance than average. this result maybe due to the interference effect with other similar items and generation of new knowledge.

In this paper, we propose a methodology to find out key entities or their combinations which are necessary to search related information in accordance with the user's investment intention. Graph data is generated by using only the named entity recognition tool and applied to the neural tensor network without learning corpus or word vectors for the field. From the empirical test, we confirm the effectiveness of the presented model as described above. However, there also exist some limits and things to complement. Representatively, the phenomenon that the model performance is especially bad for only some stocks shows the need for further researches. Finally, through the empirical study, we confirmed that the learning method presented in this study can be used for the purpose of matching the new text information semantically with the related stocks.

Key Words : Natural Language Processing, Neural Tensor Network, knowledge Entity, Stock, Artificial Intelligence

Received : February 19, 2019 Revised : May 4, 2019 Accepted : May 24, 2019
Publication Type : Regular Paper Corresponding Author : Kyong Joo Oh

저 자 소 개

양 윤 석
서울대학교에서 기계항공공학 학사, 한국과학기술원(KAIST)에서 경영공학 석사를 취득하였다. 현재 연세대학교 투자정보공학과 박사과정에 재학 중이며, 삼성자산운용에서 OCIO운용 팀장을 역임 중이다. 주요 관심분야는 자연어처리, 그래프, 시멘틱 검색 등이다.

이 현 준
연세대학교에서 물리학 학사를 취득하고, 현재 동 대학원 산업공학 박사과정에 재학 중이다. 주요 관심분야는 금융, 머신러닝, 빅데이터 분석, 시계열 분석, 데이터 마이닝 등이다.

오 경 주
한국과학기술원(KAIST)에서 경영정보공학 박사학위를 취득하였다. 금강기획 마케팅전략연구소와 현대증권 리서치 센터에서 근무하였으며, 현재 연세대학교 산업공학과 교수로 재직 중이다. 금융공학연구실에서 시스템트레이딩, 핀테크전략과 스마트금융기술, 로보어드바이저를 연구하며 학생들을 지도하고 있다.