석사학위논문

비정형 텍스트에서 BERT 기반 지식그래프 생성을 통한 BCG Matrix 시각화

2021년

한성대학교 지식서비스&컨설팅대학원

스마트융합컨설팅학과

스마트융합컨설팅전공

박 병 철

석 사 학 위 논 문
지도교수 정성훈

비정형 텍스트에서 BERT 기반 지식그래프 생성을 통한 BCG Matrix 시각화

BCG matrix visualization through BERT-based knowledge graph generation from unstructured text

2020년 12월 일

한성대학교 지식서비스&컨설팅대학원

스마트융합컨설팅학과

스마트융합컨설팅전공

박 병 철

석 사 학 위 논 문
지도교수 정성훈

비정형 텍스트에서 BERT 기반 지식그래프 생성을 통한 BCG Matrix 시각화

BCG matrix visualization through BERT-based knowledge graph generation from unstructured text

위 논문을 컨설팅학 석사학위 논문으로 제출함

2020년 12월 일

한성대학교 지식서비스&컨설팅대학원

스마트융합컨설팅학과

스마트융합컨설팅전공

박 병 철

박병철의 컨설팅학 석사학위 논문을 인준함

2020년 12월 일

심사위원장 _____(인)

심 사 위 원 _____(인)

심 사 위 원 _____(인)

국 문 초 록

비정형 텍스트에서 BERT 기반 지식그래프 생성을 통한 BCG Matrix 시각화

한성대학교 지식서비스&컨설팅대학원
스 마 트 융 합 컨 설 팅 학 과
스 마 트 융 합 컨 설 팅 전 공
박 병 철

자연어 처리란 전산 자원을 활용하여 사람들의 언어를 이해하고, 생성하거나 번역하는 등 다양한 활용 분야를 다루는 인공지능을 말한다. 최근 자연어 처리 모델은 딥러닝 기반으로 급격한 향상을 이루어내고 있으며 이들 언어 모델은 다양한 분야에서 인간의 지능에 근접한 능력을 보여주고 있다. 하지만 긴 문장에 대한 처리 일관성의 한계가 있으며 인간의 상식적인 수준의 물리적인 세계는 잘 이해하지 못하고 있다. 그리고 상호작용이 일어나는 복합적인 도메인에 대해서는 많은 문맥이 결여되어 있다. 도메인에 적합한 문장 수준의 임베딩 데이터를 학습시켜 활용하면 비즈니스 컨설팅 기반의 대용량 텍스트 데이터에서 차트, 그래프, 그림 등 원하는 보고서 생성을 위한 언어모델을 효율적으로 학습할 수 있다. 본 연구는 국내 증시 상장기업 전자공시 사이트에서 비정형 기업정보를 텍스트로 수집하여 BERT 기반의 딥러닝 자연어 처리 모델의 학습용 임베딩 데이터를 생성한다. 지식그래프는 멀티 도

메인 사이의 객체 및 관계에 대한 문제 해결 능력을 향상시키기 위한 Tool로써 활용하여 비즈니스 컨설팅 보고서를 생성하는 시각화에 관한 연구를 수행하였다. 또한 딥러닝 기반의 자연어 언어 모델의 학습에서 중요한 역할을 차지하는 임베딩 데이터의 정확성을 향상시키기 위한 Special Token인 컨설팅 카테고리 성격의 [CON] Token과 해당 리포팅 보고서의 관계(Relation) 값의 특성을 갖는 [REP]를 추가하는 연구를 수행하였다. 해당 데이터를 기반으로 BERT 사전학습모델을 이용하여 Special Token, 즉 [CON]토큰과 해당 [REP] 토큰을 추가하는 BPE (Byte Pair Encoding)을 통하여 임베딩 데이터를 생성한다. 그리고 컨설팅 도메인 영역과 이에 해당하는 로직 생성을 위한 BERT 기반의 사전학습(Pre-train)을 수행하여 어휘사전(Vocab)을 생성한다. 확장된 임베딩 데이터 기반으로 ETRI 엑소브레인의 API를 활용하여 개체 및 관계를 생성한다. 지식그래프의 개체명 분석 및 의존 구문 분석을 통하여 BCG Matrix의 요소인 개체 및 관계, [CON[과 [REP]의 강조를 통하여 유의미한 분류체계를 구현한다. 이를 기반으로 추가된 토큰과 지식그래프의 개체와 관계에 의해 맵핑된 임베딩 데이터에서 BCG Matrix 차트를 생성할 수 있는 요소(Element)를 추출하여 비즈니스 로직을 적용하여 시각화한다. 본 연구는 딥러닝 기반의 학습이 필요한 과정에서 자원과 기능상의 한계가 있었다. 그래서 비정형 텍스트에서 비즈니스 컨설팅 보고서 생성을 자동화하는 프로세스를 검증하는 방법으로 toy example 형태로 구현하였다. 또한 딥러닝 기반의 자연어 처리 모델의 학습에서 중요한 역할을 수행하는 임베딩 데이터의 품질을 검증하는 방법으로는 문장/단어 유사도를 사용하였다. 비록 학습데이터의 부족 및 학습 시간에 제약으로 보완이 필요하지만 향후 비즈니스 컨설팅 영역의 통합된 연구환경을 구성하면 자동화된 프로세스를 통하여 충분히 학습된 모델을 생성할 수 있다.

【주제어】 BERT, Text to Knowledge Graph, Word embedding, BCG Matrix

목 차

표 목 차

그 림 목 차

Ⅰ. 서론

1.1 연구의 배경

최근 자연어 처리 영역에서 문서 요약, 키워드 추출, Q&A 등 비정형 텍스트 데이터 기반의 자연어 처리에 대한 연구가 다수 진행되고 있다. 특히 GPT-3와 BERT 등 딥러닝 기반으로 급격한 향상을 이루어내고 있으며, 이들 언어 모델은 다양한 분야에서 인간의 지능에 근접한 능력을 보여주고 있다. 비즈니스 컨설팅에서의 비정형 데이터 처리의 중요성도 확대되고 있다. 특히 비즈니스 컨설팅 영역에서의 텍스트 데이터와 그림, 차트, 도표 등 통합적인 비정형 데이터 기반으로 자동화된 자연어 처리에 관한 연구가 필요할 것으로 보이나 현재까지 많은 연구가 이뤄지고 있지 않다. 이에 비정형 텍스트 데이터에서 딥러닝 기반 자연어 처리 모델인 BERT 기반으로 지식그래프를 활용하여 기업의 전략 포트폴리오 구성을 위한 BCG Matrix를 생성하는 연구를 진행하고자 했다. 이를 위해 기업공시 정보 및 비즈니스 영역의 대용량 텍스트 데이터의 처리 및 분석을 통하여 BCG Matrix 자동 생성에 대한 연구를 진행하게 되었다. 그러나 대부분의 자연어 처리 모델은 긴 문장에 대한 일관성의 한계를 가지고 있으며 인간의 상식적인 수준의 물리적인 세계를 잘 이해하지 못하고 있다. 그리고 상호작용이 일어나는 복합적인 도메인에 대해서는 많은 문맥이 결여되어 있다. 최신의 자연어 처리 모델은 텍스트 합성에서, 비록 전체적인 품질이 높지만, 충분히 긴 구절에 대해 일관성을 잃기 시작하고, 스스로 모순되며, 때로는 비순차적인 문장이나 단락을 포함하기도 한다. 또한 현실의 물리적인 세계를 잘 모른다. 모델 학습이 텍스트를 통해 이루어지면서 세상을 배웠기 때문이다. 특히 GPT-3가 약한 영역은 현실 세계의 일반적인 상식에서 취약함이 드러난다. 자연어 처리 모델은 실제 세계의 물리적 상호작용과 같은 경험의 다른 도메인에 기반을 두지 않으며, 따라서 복합적인 도메인에 대한 많은 문맥이 결여 되어 있다. 그래서 지도학습으로 예측을 확장하는 것은 한계에 도달할 가능성이 높으며 도메인의 제약을 극복하기 위하여 다른 접근 방식을 통한 증강 학습이 필요할 수 있다. 그리고 대

부분의 딥러닝 시스템에 공통적인 몇 가지 제약 사항이 있다. 도출된 결과는 쉽게 해석할 수 없으며, 인간보다 훨씬 더 성능에 대한 높은 편차 때문에 새로운 입력에 대한 예측에서 잘 보정되지 않고 학습된 데이터의 편향을 유지한다. 본 연구에서는 전문 용어 기반의 작은 빈도수로 인하여 비즈니스 컨설팅 영역 임베딩 데이터의 품질이 떨어지는 현상을 보완하기 위하여 Special Token을 추가하는 연구도 함께 진행하고자 한다. 이는 정제된 임베딩 데이터를 딥러닝 기반 자연어 처리 모델의 학습용 데이터로 활용하면 대용량 텍스트 데이터에서 차트, 그래프, 그림 등 원하는 보고서 생성을 위한 언어 모델을 통합하여 신뢰도 높게 학습시킬 수 있기 때문이다. 이를 위해 비정형 텍스트 데이터로부터 지식그래프의 개체 및 관계를 생성하고 BERT 기반의 정제된 임베딩 데이터 생성을 통하여 BCG Matrix의 요소들은 추출하고 이를 통해 자동으로 시각화하는 연구를 진행하고자 한다.

1.1.1 자연어 처리 최근 연구 동향

사람의 언어는 불연속적인 형태의 단어로 이루어진다. 각 단어가 갖는 의미는 서로 연관성이 있을 때도 있지만, 형태가 다른 경우에는 서로 얼마나 연관성이 있는지 겉 형태만으로 파악하기 어렵다. 그래서, 자연어 처리 분야에서 임베딩은 사람이 사용하는 자연어의 형태와 컴퓨터가 이해하는 벡터로 변환이 가능한 함수 또는 맵핑 테이블을 만들어 내는 과정이다. 이를 통해 자연어 처리에서는 단어나 문장, 문서를 벡터로 나타내는 것이 가능해졌다. 단어의 의미를 다루면서, 말뭉치(Corpus)로부터 단어의 특징을 추출하여 벡터로 만드는 과정을 거친다. 같은 데이터를 표현할 때 가능한 낮은 차원으로 표현할수록 쉽게 모델링하고 학습할 수 있으므로, 이러한 희소 벡터로 나타나는 것보다는 밀집(Dense) 벡터로 표현해주는 것이 훨씬 좋다. 예를 들면 '㈜카카오'의 '㈜카카오톡'이라는 서비스와 '네이버(주)'의 '라인'이라는 서비스 사이의 경쟁이라는 의미 차이가 밀집 벡터화된 임베딩 레이어에 함축되어 있으면 좋은 임베딩 데이터라고 말할 수 있다. 딥러닝 기반의 자연어 처리 모델인 BERT(Bi-directional Encoder Representations from Transformer)는 Transformer의 인코더로만 이루어진 언어 모델이다. 잘 만들어

진 BERT 언어 모델 위에 1개의 Classification layer만 부착하면 다양한 NLP Task를 수행할 수 있다. 이러한 자연어 처리 모델로 영어권에서는 11개의 NLP Task에서 최고의 성적을 거두었다.

BERT 모델은 BPE(Byte Pair Encoding)로 학습한 어휘 집합(vocab)을 쓴다. BPE 알고리즘을 통한 서브워드 단위 분절은 필수 전처리 방법이다. 한국어의 서브워드 분절(Tokenization) 방법에서는 기본적으로 한국어는 뜻을 가진 더 작은 서브 워드들의 조합으로 구성된다는 전제하에 적용되는 알고리즘이다. 서브 워드로 분절한 임베딩 데이터는 딥러닝 기반의자연어 처리 모델의 학습용 데이터로 활용할 수 있다. 해당 임베딩 데이터는 일부 후처리를 통해 언더바 (_) 문자를 ## 로 바꾸고 PAD Token, UNK Token, CLS Token, MASK Token, SEP Token 등 토큰을 더한다.

그림 1의 1번은 어휘사전(Vocab)을 만들기 위한 문서의 전처리 과정 기준이 되는 문서를 Character 단위로 분리한다. 2번은 Character 단위로 분리된 문서에서 가장 많이 등장하는 쌍(Pair)을 찾아 병합하는 과정이다. Iteration을 정해 놓고 주어진 횟수만큼 수행한다. 마지막으로 세 번째는 정의된 Iteration을 모두 수행 후 문서를 분절(Tokenize)해서 어휘사전을 생성한다.

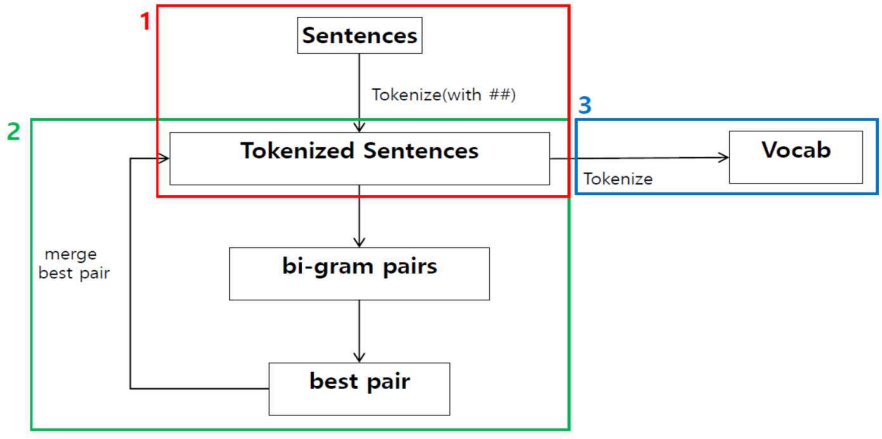


그림 1 BPE 순서도
(출처 : 『 Attention Is All You Need』 (Google Brain, Google Research, University of Toronto) 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.)

1.1.2 Text to Knowledge Graph

지식 그래프(Knowledge Graph)란 사람이 기억으로 생성하고 활용하는 지식 정보를 머신을 통해 좀 더 정확하게 많은 양의 지식 탐색을 위한 기법이다. 의미있는 데이터 개체의 연결을 통하여 새로운 아이디어를 발견할 수 있다. 대표적인 지식 그래프로는 구글 지식 그래프, 위키피디아, 위키데이터, 디비피디아 등이 있다. 구글의 지식 그래프는 검색엔진의 검색을 통한 발견에서부터 시작하여 좀 더 정확하고 좀 더 잘 요약되고, 좀 더 깊이 있는 정보로부터 지식을 발견하도록 제공하고 있다. Stanford CoreNLP (Manning et al., 2014)는 영국 Stanford NLP 그룹에서 배포한 자연어 처리 도구 소프트웨어이다. 대표적으로 개체명 인식(Named Entity Recognition), 품사 태깅 (Part-Of-Speech) 등 다양한 작업을 수행할 수 있다. 특히 개체 정의 인식기를 활용하여 개체를 인식하고 태그를 이용한 각 단어의 형태소 분석 결과를 활용하여 핵심단어 추출에 활용할 수 있다. 그림 2는 영어권 개체 인식 및 의존도 분석 사례이다.

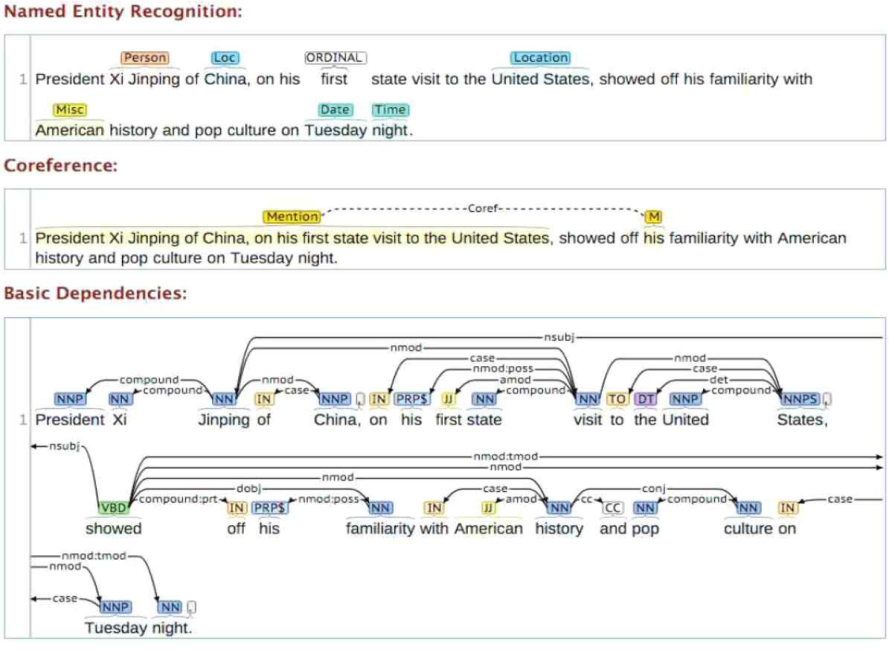


그림 2 Stanford CoreNLP 활용 예시
(출처: https://stanfordnlp.github.io/CoreNLP/)

언어분석에서 정보 추출(Information Extraction), Question & Answering 등의 응용 분야 기술은 개체명 인식 및 구문분석의 태그셋이 중요하다. 본 연구에서는 ETRI 엑소브레인 개체명 인식 기능, 개체명 태그셋은 다수의 대분류와 좀 더 세분화된 분류로 구성된 표준 개체명 태그셋을 사용하여 개체를 분류하였다. 여기에 비즈니스 컨설팅 영역의 개체명 인식 및 구문분석 태그를 추가하는 작업을 진행하여 좀 더 정확성을 높일 수 있다.

1.1.3 BCG Matrix

기업들의 미래사업 포트폴리오 전략을 수립하기 위하여 비즈니스 컨설팅에서는 매트릭스 분석(PPM : Product Portfolio Management)을 수행한다. 그 중에 BCG Matrix 분석은 Boston Consulting Group이 개발한 것으로 기업이 수행하고 있는 산업군 영역들을 상대적 시장점유율과 해당 사업이 속한 산업군 영역의 연평균 성장률로 핵심사업영역을 분석하고 그에 따른 미래사업 포트폴리오를 제시해 주는 PPM 분석 기법이며, 미래 사업 포트폴리오를 작성하는 데 광범위하게 활용되고 있다. 1사분면의 Problem Child는 성장성은 높으나 아직 시장점유율이 낮은 사업 부문으로 대부분 신규사업인 경우가 많아 새로운 투자가 필요한 영역이다. 따라서 집중적인 투자를 할 것인지 과감한 퇴출을 할 것인지에 대한 신중한 결정이 필요하다. STAR는 2사분면에 위치하며 고성장시장에서 시장점유율이 높은 사업이다. 선도기업의 지위를 유지하기 위한 자금의 투자가 필요하다. 3사분면인 Cash Cow는 성장성은 낮으나 시장점유율이 높아서 회사의 자금원 역할을 하는 사업이다. 해당 자금으로 Star, Problem Child의 성장을 지원한다. 마지막으로 4사분면인 Dog는 시장점유율과 성장성이 모두 낮은 사업으로 유지할 것인지, 축소 또는 철수할 것인지에 대한 의사결정이 필요하다. BCG Matrix를 생성하기 위하여 X축의 상대적 시장점유율은 국내 IT 플랫폼 및 콘텐츠 산업부문의 대표적인 두 기업인 네이버㈜와 ㈜카카오를 비교하였다. X축은 상대적 개념의 시장점유율인 RMS(Relative Market Share)을 사용하여 지수를 추출한다. Y축은 성장률로 cagr(Compound Annual Growth Rate)은 해당 IT플랫폼과 콘텐츠 산업부문

의 시장 성장률이며 2020년과 2016년 5년 단위 기준으로 매출액의 성장률을 산출한다. 마지막으로 원의 크기는 비교 대상 기준인 네이버(주) 사업의 매출 금액의 크기로써 가장 큰 사업을 기준으로 비교하여 원의 크기로 표현한다. 이렇게 생성된 BCG Matrix는 각 산업부문별 내부 성과 평가로도 응용하여 활용할 수 있다. 아래 그림 3은 BCG Matrix 구성 예시를 보여준다.

사업명(계열이름)		A	B	C	D	E
M/S	상대적시장점유율(X축)	200 / 50 = 4.0	160 / 70 = 2.3	100 / 130 = 0.8	40 / 120 = 0.3	35 / 220 = 0.2
	(절대적시장점유율)	(58.8%)	(51.6%)	(29.0%)	(14.8%)	(9.9%)
성장률	공식	=RATE(경과년수,0,-첫해수치,마지막해수치)				
	cagr.(Y축)	7.2%	2.8%	6.7%	8.4%	0.9%
매출액	5년전 시장규모	240	270	250	180	340
	계(현재시장규모)	340	310	345	270	355
	당사(거품크기)	200	160	100	40	35
	경쟁사	50	70	130	120	220
	경쟁사	40	50	60	80	50
	경쟁사	30	20	40	20	30
	경쟁사	20	10	15	10	20

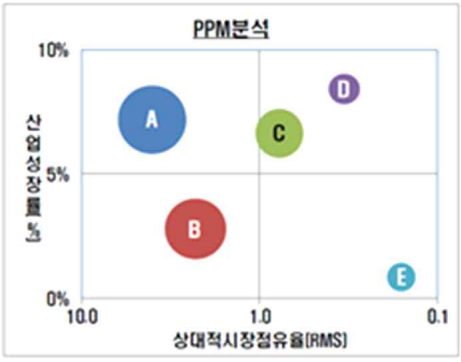


그림 3 BCG Matrix
(출처: 컨설팅 수행과 보고서 작성에 필요한 필수 Chart 작도 Skill
이봉철(KPC 경영기획/전략/재무부문 전문위원))

해당 데이터는 전자공시시스템(http://dart.fss.or.kr)의 상장회사 매출 및 영업 데이터에 대한 크롤링 및 OPEN API를 통해 수집할 수 있다. 그림 4는 해당 사이트에서 전자공시시스템에 있는 관련 데이터 예시를 보여준다.

그림 4 전자공시 기업정보 매출데이터 예시

네이버(주) 2020년

(2) 서비스별 영업현황

(누적) (단위 : 백만원,%)

구분	연결			
	제22기 반기		제21기 반기	
	금액	비중	금액	비중
영업수익	3,634,532	100.00%	3,141,137	100.00%
- 광고	318,686	8.8%	308,803	9.8%
- 비즈니스플랫폼	1,526,907	42.0%	1,385,227	44.1%
- IT플랫폼	328,377	9.0%	205,071	6.5%
- 콘텐츠서비스	134,982	3.7%	85,140	2.7%
- 라인 및 기타플랫폼	1,325,581	36.5%	1,156,896	36.8%

네이버(주) 2016년

(2) 영업부문별 매출현황

(누적) (단위 : 백만원,%)

구분	연결				별도			
	제18기 반기		제17기 반기		제18기 반기		제17기 반기	
	금액	비중	금액	비중	금액	비중	금액	비중
매출	1,924,562	100.0%	1,521,966	100.0%	1,201,285	100.0%	1,020,161	100.0%
- 광고	1,395,572	72.5%	1,088,464	71.5%	1,126,615	93.8%	945,427	92.7%
- 콘텐츠	473,131	24.6%	398,286	26.2%	52,577	4.4%	39,015	3.8%
- 기타	55,859	2.9%	35,216	2.3%	22,093	1.8%	35,719	3.5%

㈜카카오 2020년

(단위: 백만원)

구분	제26기 반기		제25기		제24기	
	매출액	비중	매출액	비중	매출액	비중
플랫폼 부문	934,516	51.3%	1,434,749	46.7%	1,039,315	43.0%
콘텐츠 부문	886,715	48.7%	1,635,363	53.3%	1,377,677	57.0%
합 계	1,821,230	100.0%	3,070,111	100.0%	2,416,992	100.0%

㈜카카오 2016년

(단위: 백만원)

사업부문	제22기 반기		제21기 반기		제21기	
	매출액	비중	매출액	비중	매출액	비중
광고 플랫폼	265,589	42.9%	300,425	65.1%	600,663	64.4%
콘텐츠 플랫폼	281,993	45.6%	137,169	29.8%	273,623	29.4%
기타	71,421	11.5%	23,280	5.1%	57,866	6.2%
합 계	619,003	100.0%	460,874	100.0%	932,152	100.0%

Ⅱ. 관련 연구

2.1 언어 모델에 관한 연구

최근에는 기계학습 방법론 중에서도 딥러닝을 이용한 개체명 인식이 가장 활발하다. 해당 연구에서는 대량의 말뭉치로부터 지도학습 기반의 음절 단위 임베딩 표현과 비지도 학습 기반의 어절 단위 임베딩 표현을 이용하여 시계열 데이터에 의미 있는 성능을 보이는 Bi-LSTM-CRF 모델로 개체명 인식을 수행하는 딥러닝 모델이 소개되었다. 단어의 의미가 숫자 벡터로 표현됨으로써 벡터 연산이 가능해지게 만드는 워드 임베딩의 개념도 도입되었다. 딥러닝을 이용한 개체명 인식 연구들은 대량의 말뭉치로부터 사전 학습된 언어 표현 모델을 이용한다. AllenNLP에서 공개한 ELMo, OpenAI에서 공개한 GPT, Google에서 공개한 BERT와 같은 모델들이 이러한 방식의 사전 학습된 자연어 처리 모델에 대한 대표적인 연구들이다. 이와 같은 언어 표현 모델 연구들의 특징은 모델의 디코더에서 단어를 예측하는 시점의 입력된 문장을 인코더에서 다시 한번 연산에 사용하는 어텐션 메커니즘을 사용하는 것이 두드러진 특징이다.

2.2.1 Transformer의 등장

RNN은 n번째에 대한 출력을 만들기 위해 n번째 입력과 n-1번째 Hidden Layer 상태를 이용했다. 해당 시계열 통계 방식으로 수행하는 예측 기반의 언어 모델에서는 문장의 시계열적인 특성이 유지될 수 있었다. 그러나, RNN의 경우 긴 문장에서 단어 간의 의미를 연결하여 유추하는 연관성(long-term dependency)에 취약함이 있었다. 이러한 문제점을 해결하기 위하여 단순히 RNN을 걷어내어 해결한 모델이 Google Research에서 발표한 "Attention Is All You Need"라는 논문에 나온다. Transformer는 인코더 및 디코더를 포함하는 복잡한 순환 신경망에 기초한 딥러닝 알고리즘이다. 자연어 처리

모델에서 입력과 출력의 예시를 들면 원본 문서를 입력으로 하여 텍스트 기반으로 문서를 요약하거나 이미지에서 캡션을 통해 해당 이미지 내용을 자동으로 생성하고 음성 인식 입력을 통해 자막을 자동으로 출력해 주기도 한다. Attention 모델은 인간이 정보처리를 할 때, 모든 시퀀스를 고려하면서 정보처리를 하는 것이 아니다. 인간의 정보처리와 마찬가지로, 중요한 특징(feature)을 더욱 중요하게 고려하는 것이 어텐션의 아이디어가 되었다. Attention 메커니즘을 통해 인코더와 디코더를 연결하여 구성했다. 트랜스포머 모델은 임베딩 벡터의 유사도에 따른 두 임베딩 데이터의 유사성을 분석하여 처리할 수 있는 형태로 표현하였다. 그림 5와 같이 다수의 인코딩 블록과 다수로 쌓은 디코딩 블록으로 적용되어 있다. RNN은 n번째 Hidden Layer 결과를 얻기 위해서 순차적으로 계산되기 때문에 병렬 처리가 불가능하고 이로 인해 연산 능력이 떨어진다. 트랜스포머에서 다수의 층으로 쌓인 블록의 출력은 다음 블록의 입력으로 사용되는 것은 RNN과 유사하다. 하지만 트랜스포머는 각 임베딩된 데이터의 유사도를 통하여 서로의 벡터를 연산할 때 병렬로 처리하는 멀티 헤드 어텐션(Multi head Attention)을 적용한다. Seq-to-seq의 문제였던 입력이 먼저 들어오는 노드가 정보를 주지 못했던 것을 Attention Mechanism으로 해결하였다. 또한 문장 단위의 임베딩 수준도 높아졌다. 트랜스포머의 핵심요소인 멀티 헤드 어텐션은 그림 6과 같다. 여러 개의 쿼리를 만들어 다양한 정보를 잘 얻어올 수 있도록 병렬 수행한 결과를 묶어서 출력한다. Attention 자체로도 정보의 인코딩과 디코딩이 가능함을 보여주었다.

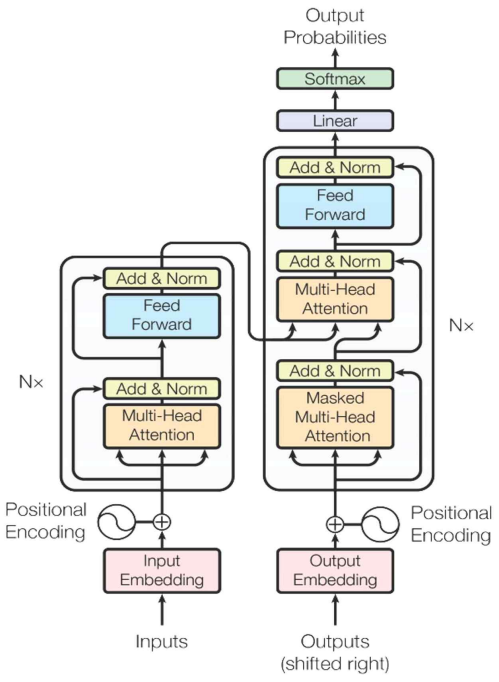


그림 5 The Transformer – model architecture

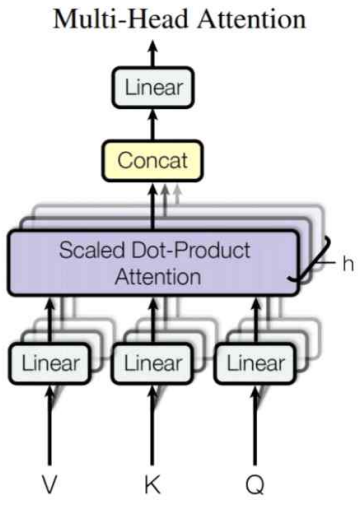


그림 6 Multi-head Attention 구조

(출처: Attention Is All You Need)

2.2.2 워드임베딩 개념

기계학습을 이용한 자연어 처리 분야 연구는 크게 임베딩 기법이 등장하기 전과 후로 나누어진다. 임베딩이란 자연어를 기계가 이해하고 효율적으로 처리할 수 있는 형태로 변환하기 위해서 벡터로 변환하여 벡터공간 안에 매핑하는 것을 말한다. 언어 모델에 대한 연구가 나오기 전까지는 대부분의 연구에서 이러한 임베딩 방식을 사용했다. 임베딩은 음절, 형태소, 단어, 어절, 문장, 문서 등 다양한 단위로 할 수 있다. 최근 자연어 처리 연구들은 방대한 양의 데이터세트와 연산량으로 비지도학습하여 범용 언어 표현 모델을 만드는 연구가 각광을 받고 있다. 언어 모델은 단어들의 시퀀스(Sequence)가 존재할 때, 특정 단어 다음에 어떤 단어가 나올 확률을 모델링하는 기법인데, 최근에는 이러한 방식의 언어 모델이 자연어 처리 응용프로그램에 많이 사용된다.

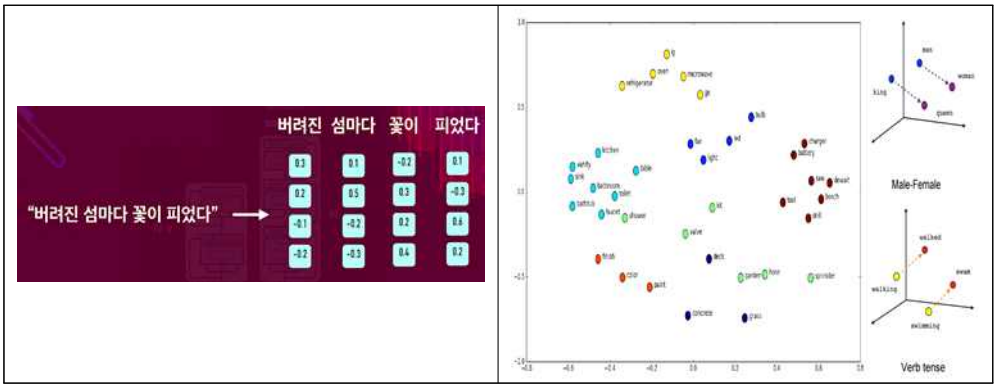


그림 7 워드임베딩 개념

임베딩에서 더욱 중요한 개념은 Self Attention이다. Self Attention이라는 개념을 통해 단어 간 관계를 강화시키기도 한다. 예를들면 “나는 공을 찬다”에서 “찬다”라는 단어는 누가 무엇을 찾는 지에 대한 단어 간 관계 강화를 Self Attention이라는 개념을 통해 제공한다. 그림 8을 보면 An attention function은 벡터 기반의 쿼리와 키-값 세트를 출력에 매핑하는 것으로 설명할 수 있으며 출력은 값의 가중치 합계로 계산된다. Query 벡터(Q)와 Key 벡터(K), 그리고 Value벡터(V)를 활용하여 문장 구조의 유사도를 추론하는 방법이다. Q와 K를 내적(MatMul) 한 뒤, V를 가중치로 최종 결과를 만들어 내는 방식을 사용한다.

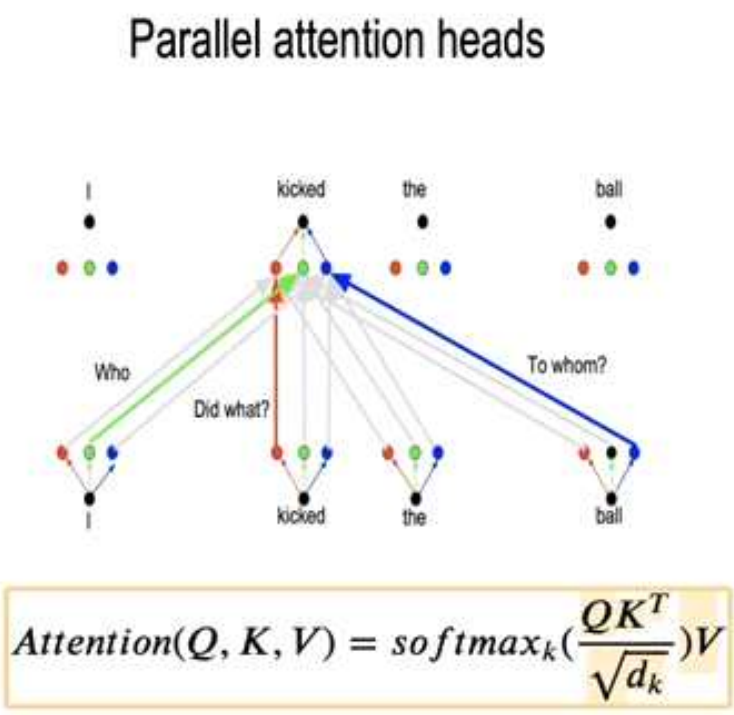


그림 8 An attention function

이는 딥러닝 기반의 자연어 처리의 혁명으로 인식되고 있다. 딥러닝 기반의 자연어 언어 모델인 BERT는 문맥 기반의 워드임베딩 결과를 제공해준다. 문서 단어 윈도우 크기에 따라 임베딩된 결과가 달라질 수 있다. 예를 들면, “그는 배를 탔다.” 와 “그는 배를 먹었다.”라는 문장이 있을 때, “배”라는 단어는 문맥 상 과일을 의미하는 “배”에 대한 임베딩을 출력하거나, 이동 수단 의미의 “배”로 다르게 임베딩된 값을 출력할 수 있다. 이러한 특징은 입력 문서에 발생하는 경우로 사전에서 빠진 단어를 문맥에 따라 정확하게 선택해야 하는 지도학습 모델에서 성능을 높이기 위한 중요한 요소이다. 본 연구에서 특징적으로 비즈니스 컨설팅 영역의 정제된 학습데이터 생성을 위한 핵심 연구에서 중요한 BERT 입력 표현에 대해 설명하고자 한다. BERT 임베딩 토큰은 Token embedding, Segment embedding, Position embedding으로 구성되어 있다. 주어진 토큰에 대해, 해당 토큰, 세그먼트 및 위치 임베딩을 합산하여 입력 표현을 구성한다. 토큰임베딩(Token embedding) 형태의 입력으로

문장의 시작은 [CLS], 문장의 종결 토큰 [SEP], 마스크 토큰[MASK], 배치데이터의 길이를 맞추기 위한 [PAD] 등의 4가지 입력 토큰을 사용하여 학습데이터를 구성하고 있다. 세그먼트 임베딩(문장의 순서를 구분하는 임베딩), 포지션 임베딩(문장내 절대적인 위치에 해당하는 임베딩)으로 구성되어 있다.

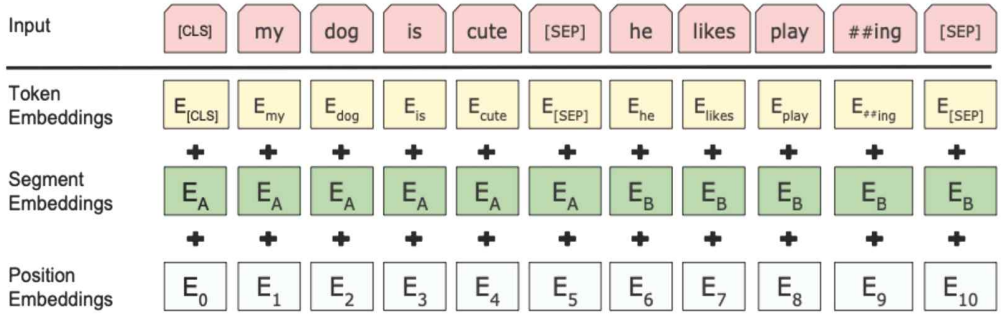


그림 9 BERT 입력

[출처] Google AI Language, 24 May 2019, BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT는 기존 언어 모델과 차이가 있는 것은 양방향 Transformer 구조로 되어 있으며 사전학습(Pre-training)을 통해 언어 전반에 대해 깊게 이해하는 단계와 미세조절(Fine-tuning)이라는 깊은 언어의 이해를 바탕으로 특정 문제에 맞춰 적응하는 단계를 거치는 언어 모델이다. 특정 문제에 맞춘 NLP 다운스트림 Task는 형태소 분석, 문장성분 분석, 의존관계 분석, 개체명 분석, 의미역 분석, 상호참조 해결, 의미관계 분석 등으로 이루어져 있다. Transformer와 BERT의 가장 큰 차이점은 마스크 언어 모델, NSP(Next Sentence Prediction) 과제를 수행하기 위한 마지막 예측 레이어의 존재 여부이다.

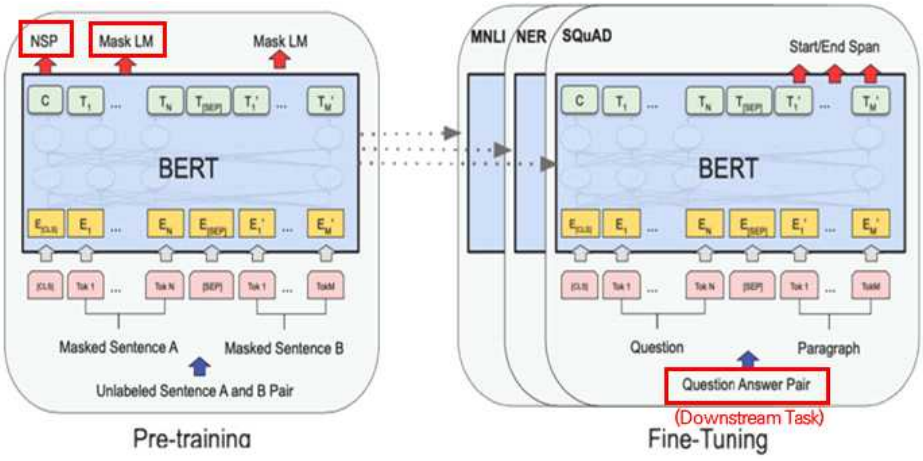


그림 10 BERT 구조 : 질문-답변 예시 BERT 실행 예

그림 11의 Mask 모델은 BERT의 성능에서 가장 중요한 역할을 하고 있다. BERT의 가장 중요한 특징은 임베딩 데이터의 15% 정도를 Masking 처리하여 학습시키는 방식으로 엄청난 성능 향상을 달성했다. 이렇게 학습된 BERT는 다음 단어를 예측하는 Next Sentence Prediction과 같은 다운스트림 타스크를 훌륭하게 수행한다.

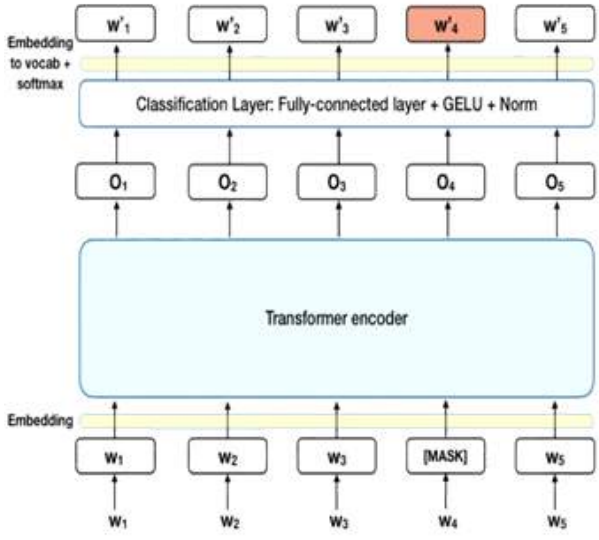


그림 11 Masked Language Model

최근에 화제가 되고 있는 GPT-3는 범용/마스터 Language Model을 추구하는 언어 모델로 많은 NLP 과제와 벤치마크에서 제로샷, 원샷, Few-shot 설정에서 강력한 성능을 보이는 1,750억 매개 변수 언어 모델을 제시했는데, 어떤 경우에는 최첨단 미세 조정 시스템의 성능과 거의 일치할 뿐만 아니라, 정의한 작업에서 고품질 샘플과 강력한 질적 성능을 고민 없이 빨리 창출할 수 있었다. 언어 모델을 확장하면 작업에 구애받지 않는 몇 번의 성능을 크게 향상시키고 때로는 이전의 최첨단 미세 조정 접근 방식으로 경쟁력에 도달한다는 것을 보여준다. GPT-3는 번역, 문제 해결, 클로즈(빈칸 채우기) 작업 등 많은 NLP 데이터셋에서 강력한 성능을 발휘하며, 문장의 참신한 단어를 사용하거나 3자리 산수를 수행하는 등 즉석 추론이나 도메인 적응이 필요한 여러 작업에서도 높은 성능을 발휘하고 있다.

2.2.3 Knowledge Graph & BERT

딥러닝 기반의 자연어 처리 기술이 활발하게 연구되고 활용되고 있는 지금, AI를 통해 인간에 근접하기 위한 지식을 학습시키기 위한 기술은 매우 중요한 기술이다. 인간의 지식이 투영된 단어의 관계를 지식 그래프로 만들고 이를 기반 데이터로 활용하게 되면 컴퓨터에게 쉽게 인간의 지식을 학습시킬 수 있다. 지식 그래프가 기존 지식 그래프와 달리 언어 종속성이 없고 새로운 용어에도 정확하게 매핑되고 확장 가능한 형태의 활용성이 있다는 것을 실험과 구현을 통해 검증했다. 최근 연구에서는 그림 12과 같이 지식 그래프의 확장에 대하여 실시간 비정형 데이터의 Knowledge Graph를 구성하였다. Social Crawler를 통하여 뉴스, 소셜미디어 등의 데이터를 수집하고 BERT 임베딩을 통하여 새로운 엔티티간의 Relation을 구성하였다. 이는 자연어 생성과 자연어 이해를 기반으로 Chatbot Layer를 구성하여 마이닝을 통하여 지식그래프를 확장하는 연구모델이다.

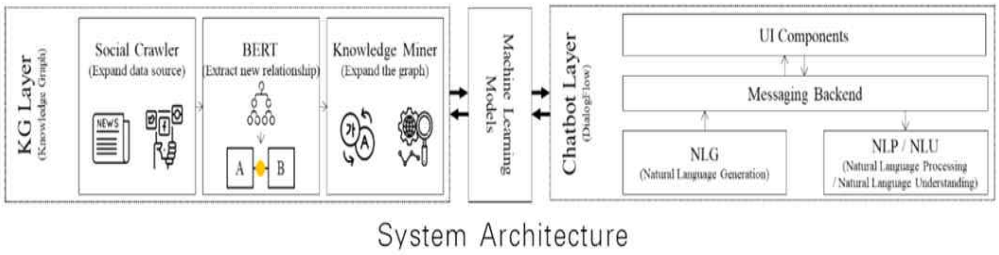


그림 12 BERT 기반 확장된 개체명 분석 및 구문 분석.
[출처] 유소엽(SoYeop Yoo), 정옥란(OkRan Jeong). (2019) .BERT 모델과 지식 그래프를 활용한 지능형 챗봇. 가천대학교

그림 13에서는 관계의 주체 및 대상에 대한 Token을 확장 구성하였다. [SUBJ]라는 관계의 주체가 되는 토큰과 [OBJ]라는 대상이 되는 토큰을 추가했다. 결과는 두 토큰 사이의 관계를 나타낸다.

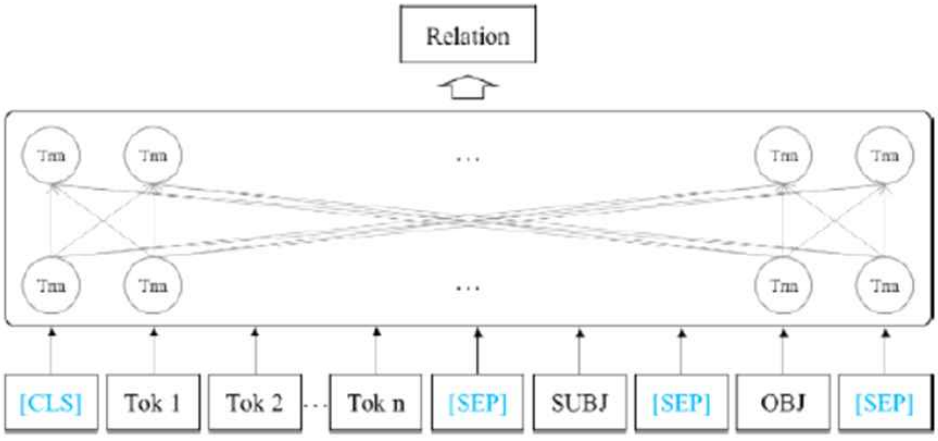


그림 13 Relation Extracting Using BERT.
[출처] 유소엽(SoYeop Yoo), 정옥란(OkRan Jeong). (2019) .BERT 모델과 지식 그래프를 활용한 지능형 챗봇. 가천대학교

BERT와 Knowledge Graph의 개념 및 관계 확장을 통하여 새로운 엔티티와 관계에 대한 예측을 수행하였다. 그림 14과 같이 BERT 모델의 임베딩 레이어를 추가하여 엔티티간 새롭게 한 쌍이 된 관계를 추론하고 해당 지식 그래프를 확장하는 연구이다. 해당 엔티티와 관계를 실시간 추출을 위해 TACRED 데이텃셋 기반으로 마이닝하여 미세조정된 BERT 모델에 임베딩된 데이터를 입력한다. 그리고 기존 지식 그래프에 관련성 깊은 노드를 찾아 연결한다.

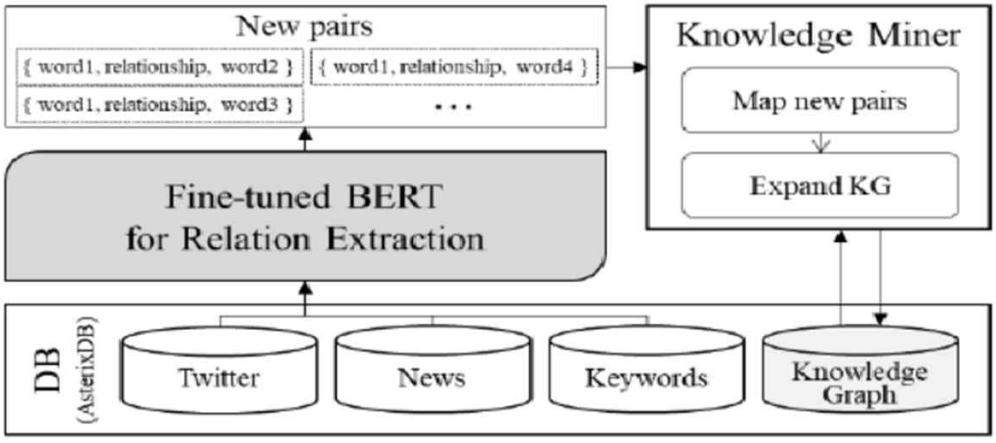


그림 14 확장된 지식그래프의 새로운 관계 추출 모델.
[출처] 유소엽(SoYeop Yoo), 정옥란(OkRan Jeong). (2019) .BERT 모델과 지식 그래프를 활용한 지능형 챗봇. 가천대학교

유사한 또 다른 연구로는 지식베이스 기반 관계 임베딩모델이 있다. BERT와 지식베이스 관계 임베딩의 융합을 통한 개방형 관계 임베딩 추출 방법으로 지식베이스 기반 모델과 BERT를 이용하여 자연어 문장으로부터 임의의 관계 임베딩을 생성하려는 연구이다. 지식베이스로부터 얻은 제한된 데이터셋을 통해 학습한 모델이 모든 임의의 관계를 추출할 수 있게 하기 위하여, 학습 목표를 임베딩으로 설정하고 일반화에 뛰어난 언어 모델인 BERT를 이용한 모델 구조를 제안하였다. TransE, ProjE와 같은 지식베이스 기반 관계 임베딩모델들은 지식베이스의 트리플들을 학습 데이터로 이용하여 트리플의 관계를 임베딩화 하는 심층 신경망(Deep neural network) 모델이다. 입력 문장의 각 단어는 BERT 언어 모델을 거쳐 임베딩 형태로 표현된다. 이는 다시 각각 개념 라벨 모델(그림15의 Entity)과 관계 임베딩 모델(그림15의 Relation)에 입력된다. 개념 라벨 모델은 입력으로부터 각 단어의 개념 점수(그림15의 s, 단어가 개념인지에 대한 점수)를 나타내는 개념 라벨(그림 15의 L)을 생성한다. 관계 임베딩 모델은 입력으로부터 문장의 관계를 200차원의 벡터로 표현한 관계 임베딩(그림15의 r)을 생성한다. 이 시점에서 개념 라벨로부터 h와 t로 표현되는 개념쌍(entity pair)을 얻을 수 있고, 관계의 표현형인 관계 임베딩을 얻을 수 있다. 즉, (h, r, t)으로 표현되는 트리플의 표현형

을 일차적으로 얻을 수 있다. 두 번째 관계 임베딩 모델은 BERT 단어 임베딩과 이전에 추출된 개념 라벨을 입력으로 받아, 개념과 관련된 관계의 임베딩(r')을 생성한다. 두 번째 개념 라벨 모델은 BERT 단어 임베딩과 이전에 추출된 관계 임베딩을 입력으로 받아, 관계와 관련된 개념 라벨(L')을 생성한다. 이 개념 라벨에서 개념 쌍(그림의 h',t')을 뽑으면, 최종 결과물인 (h', r', t'), 즉 트리플의 표현형을 얻을 수 있다. 개념 라벨과 관계 임베딩 모델들의 구조는 간단한 단일 레이어 신경망(single-layer Neural Network)이다. 이는 BERT 논문에서 미세 조절(파인 튜닝, Fine-tuning)을 위한 모델을 간단한 구조의 신경망으로 설정했던 것과 같이, BERT 언어 모델이 충분한 일반화 능력과 언어적 특성을 가진 임베딩을 제공하므로 간단한 구조의 신경망으로도 충분하기 때문이다.

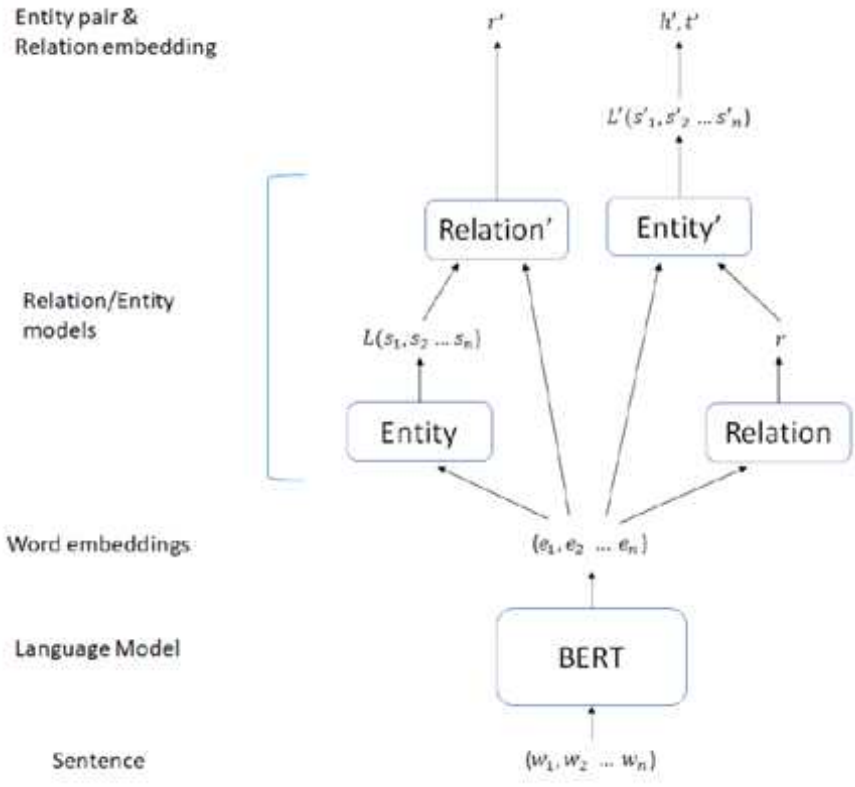


그림 15 지식베이스 기반 관계 임베딩모델

[출처] 홍기원, 맹성현. (2019) . BERT와 지식베이스 관계 임베딩의 융합을 통한 자연어 문장으로부터의 개방형 관계 임베딩 추출 방법. KAIST.

또 다른 언어 모델의 연구 영역은 기계 독해(Machine Reading Comprehension) 연구이다. 해당 연구에서는 Machine-Reading Model을 통하여 절차적 텍스트로부터 동적 지식 그래프를 구성하는 신경 기계 판독 모델을 제안했다. 모델이 구성하는 명시적이고 구조적이며 진화하는 지식 그래프 표현은 경험적으로 설명하는 것처럼 텍스트의 기계 이해력을 향상시키기 위해 다운 스트림 질문 응답 작업에 사용될 수 있다. 절차적 텍스트에는 뉴스 기사, 과학 매뉴얼 및 절차 텍스트 (예 : 레시피, 사용 방법 가이드 등)가 포함되어 있다. 이 데이터에서 KG를 구축하면 엔티티 간의 변화하는 관계를 연구하는 데 도움이 될 뿐만 아니라 암시적 정보를 보다 명확하게 만들 수 있다[21].

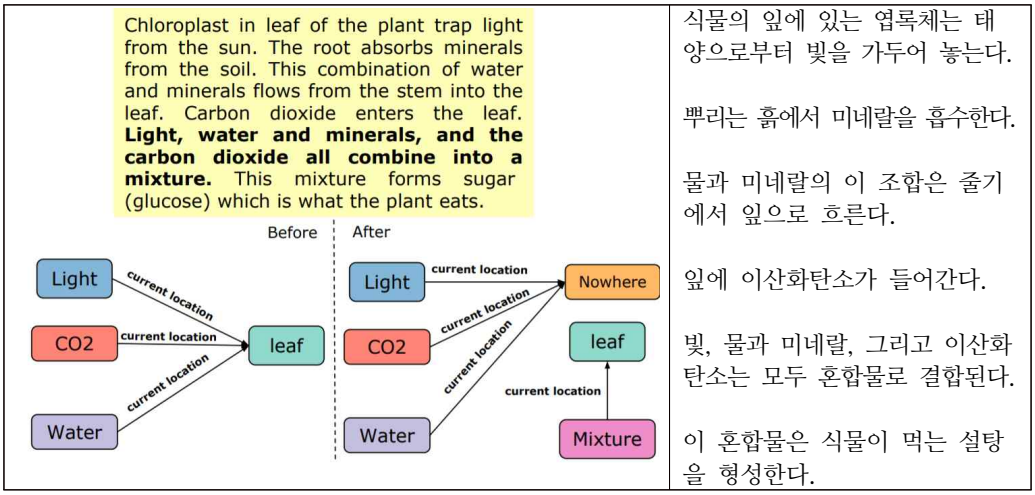


그림 16 지식 그래프의 스냅샷

[출처] Rajarshi Das Tsendsuren Munkhdalai Xingdi (Eric) Yuan Adam Trischler Andrew McCallum / ICLR 2019 | April 2019/College of Information and Computer Sciences University of Massachusetts, Amherst : Building Dynamic Knowledge Graphs from Text using Machine Reading Comprehension

2.2 연구 모형에 관한 연구

2.2.1 연구 모형

표 1에서 정제된 문장 수준의 임베딩 데이터 생성하기 위한 연구의 특장점을 보여주는 연구 개요를 설명하였다.

표 1 확장된 문장 수준의 임베딩 데이터 생성 연구 개요

워드 임베딩 (Token Expansion)	임베딩 to KG Mapper	Business Logic Trigger
기업정보 대용량 텍스트 데이터를 BERT 기반으로 임베딩 및 사전학습을 통하여 [CON] :[REP] Pairs 데이터로 활용할 수 있도록 생성하여 딥러닝 모델의 정확성을 향상 시키는 연구.	생성된 임베딩 데이터 기반으로 지식그래프를 활용하여[CON]:[REP] Pairs와 객체[Entity]:관계[Relation] Pairs를 생성하고 이를 기반으로 전문 도메인 영역의 어휘사전과 모델을 생성.	생성된 차트 기반 데이터에서 비즈니스 로직 연산을 통해 BCG Matrix 생성을 위한 요소(Element)를 추출하고 차트 시각화를 수행.

지식 그래프의 개체(Entity)와 관계(Relation) 구조에 대해 BERT 기반의 임베딩 데이터 생성 및 사전학습(Pre-train)을 통해 어휘사전을 생성하고 사전학습된 BERT 모델을 생성한다. 해당 임베딩된 데이터를 ETRI의 엑소브레인에서 제공하는 개체명 인식 기능 및 구문 분석 기능을 활용하여 컨설팅 카테고리 성격의 Special Token [CON] 과 해당 카테고리 보고서의 결과 성격인 [REP]를 객체(Entity)와 관계(Relation) 데이터와 함께 학습시켜 확장된 임베딩 데이터를 생성한다. 생성된 임베딩 데이터에서 BCG Matrix를 생성하기 위한 요소들을 추출하고 Business Logic을 적용하여 시각화를 수행한다.

구체적인 연구 방법을 설명하면 비정형 텍스트인 IT플랫폼 및 검색서비스 기업인 네이버(주)와 ㈜카카오의 기업정보를 수집하고 정제한 후 해당 데이터를 파싱한다. BCG Matrix의 산업성장률은 주요 경쟁 기업간의 5년간 매출

비교를 통하여 산출하기 때문에 2020년과 2016년 기업정보를 수집한다. 데이터 수집을 위한 파이썬 팩키지인 BeautifulSoup를 활용하여 전자공시정보에서 기업정보 중 BCG Matrix 생성을 위한 요소를 추출하기 위하여 사업부문, 매출액, 시장규모 등 수집(Crawling)/정제(Cleansing) 및 HTML element 분석(태그, 속성, 값)을 통한 파싱(Parsing)을 수행한다.

두 번째로는 BERT 기반 임베딩을 통하여 Special Token인 [CON]과 [REP]를 추가하여 임베딩 레이어를 확장(Expansion)시킨다. 구글에서 공개한 언어 모델인 Sentencepiece 기반 BPE(Byte Pair Encoding)으로 임베딩 데이터를 생성한다. 빈도수에 기반하여 단어를 의미있는 패턴으로 구분하여 분절화(WordPiece Tokenization)를 수행한다. 해당 단계에서는 서브워드 단위 분절을 수행하여 희소성 감소 외에도 [UNK], 즉 Unknown 토큰에 대해 정제를 수행한다. 이렇게 생성된 임베딩 데이터를 어휘사전으로 활용한다. [CON] 토큰, 예를들면 사업부문, 매출, 시장규모와 [REP] 토큰에 추가된 사업 부문별 매출액, 산업 성장률 등을 파싱하여 해당 기반 데이터를 추출한다. 이후 BERT 사전학습을 통해 [CON], [REP]에 대한 개념(Entity) 및 관계(Relation)를 추출(Extracting)하고 Pari 형태로 맵핑(Mapping)이 수행된다. 이를 통해 BCG Matrix에 특화된 임베딩 데이터가 생성되는 것이다. 세 번째로는 Text to Knowledge Graph의 수행이다. 임베딩된 데이터를 ETRI 엑소브레인 API 기반으로 지식그래프 개체(Entity) 인식 및 의존 관계 구분분석을 수행한다. 이를 통해 형태소 분석 및 동음이의어, 개체명의 정보가 추출되며 구문분석을 통하여 개체과 개체간의 다중 지배소 형태의 구문 분석이 실행된다. 마지막으로 BCG Matrix생성에 필요한 개체와 관계를 추출하여 비즈니스 로직(Business Logic Trigger) 반영을 통한 결과 산출 및 이를 기반으로 BCG Matrix 차트를 생성한다.

본 연구에서는 그림 17과 같은 연구 모형을 제시한다. 우선 국내 대표적인 IT 플랫폼 및 검색서비스 기업인 네이버(주)와 ㈜카카오의 기업 정보를 전자공시 시스템을 통하여 사업보고서를 수집하고 전처리하여 본 연구의 기본 데이터셋을 만들었다. 본 연구에서는 수집 및 정제의 과도한 시간 소요와 GPU 환경 구성 등 제약사항으로 인하여 toy example 형태의 연구 모형을 구성하여 연구를 진행하였다. 아쉬운 부분이지만 해당 네이버(주)와 ㈜카카오(2020년과 2016년)의 전

자공시 사이트에서 수집된 데이터로도 유의미한 연구를 도출할 수 있었다. 간단히 연구모델을 설명하면 우선 BERT 언어 모델의 학습 데이터로 쓰이는 BPE(Byte Pair Encoding)로 임베딩 데이터를 만드는 과정을 거친다. 그리고 BERT 임베딩 Layer의 Special Token인 [CON]과 [REP]와 개념(Entity)와 관계(Relation)을 추가하여 BERT 기반의 사전학습을 통해 확장된 임베딩 데이터를 생성한다. 추가된 Special Token인 [CON]과 보고서 생성 기반 결과 값 성격인 [REP] Token으로 정의한다. 이렇게 Special Token이 추가된 임베딩 데이터를 BERT 기반으로 사전학습(Pre-train)된 어휘사전과 Pre-trained BERT Model을 생성한다. ETRI 엑소브레인 API를 활용하여 [CON]과 [REP]와 개체(Entity)와 관계(Relation)를 활용하여 지식그래프의 개체 및 의존관계를 분석하는 작업을 수행한다. 마지막으로 확장된 임베딩 데이터를 기반으로 Business Logic을 수행한다. 2가지 Token 정보로 학습하여 BCG Matrix의 X축, Y축에 사용될 상대적인 시장점유율과 산업성장률, 그리고 두 기업의 매출액을 기반으로 매핑하고 WinSim353테스트로 유의미한 성능이 검증된 Word embedding 데이터, 즉 [CON] : [REP]와 [Entity] : [Relation]에 대한 BCG Matrix를 시각화할 수 있는 비즈니스 로직을 반영한 결과를 산출한다.

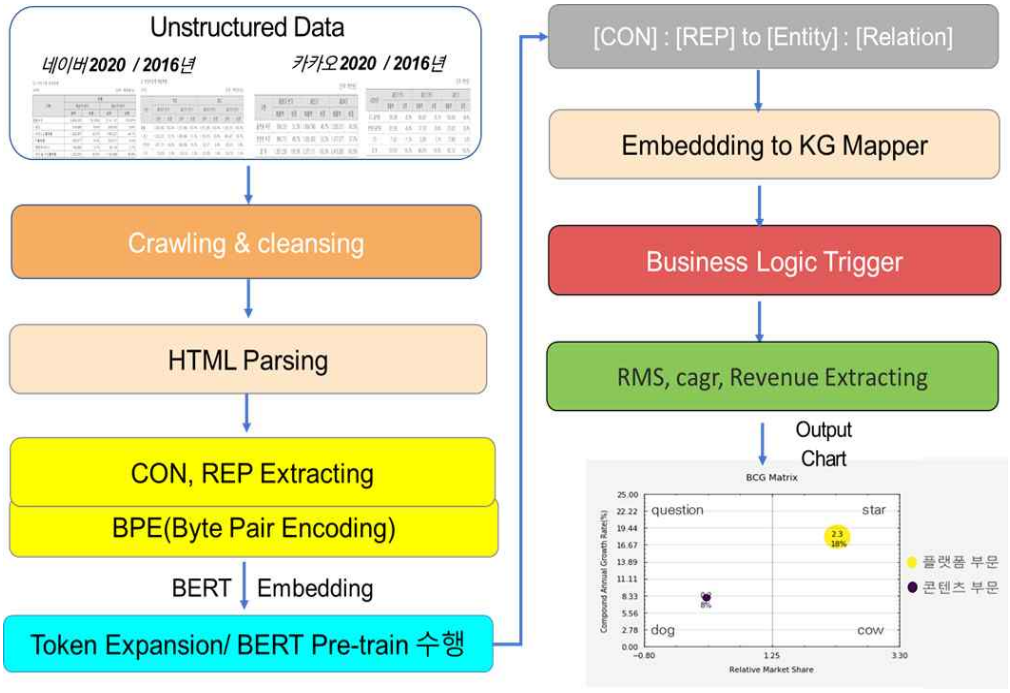


그림 17 전체 연구 모형 구조도

본 연구의 주요한 연구 분야에 대해 설명하고자 한다. 특히 비즈니스 컨설팅 도메인 영역의 Special Token을 추가하는 연구에 대해 설명하고자 한다. 우선 BERT 임베딩 과정 중에 Special Token 추가를 통한 비즈니스 컨설팅 도메인의 Self Attention을 강화한다. 이를 통하여 긴 문장에 대한 처리 일관성의 한계와 특정 도메인 지향적인 Downstream Task에 대한 제약을 보완할 수 있도록 확장된 임베딩 데이터를 제공하는 방안을 연구하였다. 원래 BERT 입력은 해당 Token, Segment 및 Positioning 임베딩을 합산하여 구성한다. BERT 입력 표현은 토큰임베딩(Token embedding), 문장의 시작은 [CLS], 문장의 종결 토큰 ([SEP]), 마스크 토큰[MASK], 배치 데이터의 길이를 맞추기 위한 [PAD] 등의 4가지 토큰을 사용하여 학습 데이터를 구성한다. 세그먼트 임베딩(문장의 순서를 구분하는 임베딩, 즉 첫 번째 문장인지, 두 번째 문장인지 구분), 포지션 임베딩(문장 내 절대적인 위치에 해당하는 임베딩)이 이에 해당한다. BCG Matrix와 같은 비즈니스 컨설팅 영역의 딥러닝 기반 언어 모델의 좀 더 정확한 학습을 위하여 그림 18과 같이 Special Token을 추가한다. 컨설팅 기반 확장 토큰 [CON], 리포팅 보고서 성격의 토큰 [REP] 개체를 추가하여 컨설팅 도메인에 특화된 임베딩 레이

어를 추가한다. "E CON A"와 "E REP A“는 문장이나 단어에 대한 어텐션이나 가중치를 부여하여 어휘사전에서 누락되지 않도록 확장된 임베딩 데이터를 생성한다. 이로인해 특정 비즈니스 컨설팅 영역의 도메인 지향적인 Downstream Task에 대한 확장된 학습용 임베딩 데이터를 제공할 수 있다.

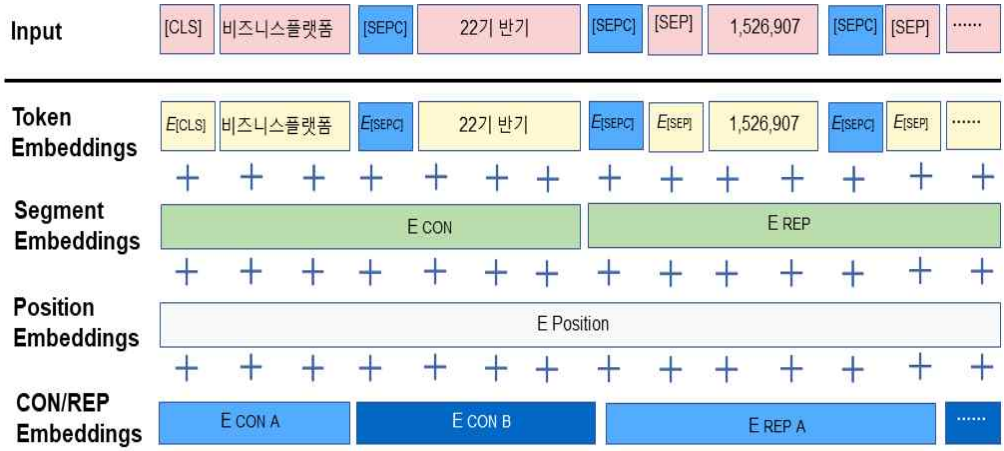


그림 18 BERT 임베딩 Special Token Expanding
[출처] Google AI Language, 24 May 2019, BERT

Ⅲ. 구현

3.1 설계 및 구현

3.1.1 모델 전체 구성도

Special Token을 추가한 BERT 기반 임베딩 데이터를 생성하고, 확장된 임베딩 데이터를 기반으로 Knowledge Graph의 개체명 분석 및 의존 구문분석을 수행한다. [CON], [REP]와 Entity, Relation에 대한 매핑을 통해 BCG Matrix 차트 생성에 필요한 데이터를 추출한다. 이를 기반 데이터로 활용하여 비즈니스 로직을 반영하고 시각화하는 연구를 그림 19와 같이 진행하였다.

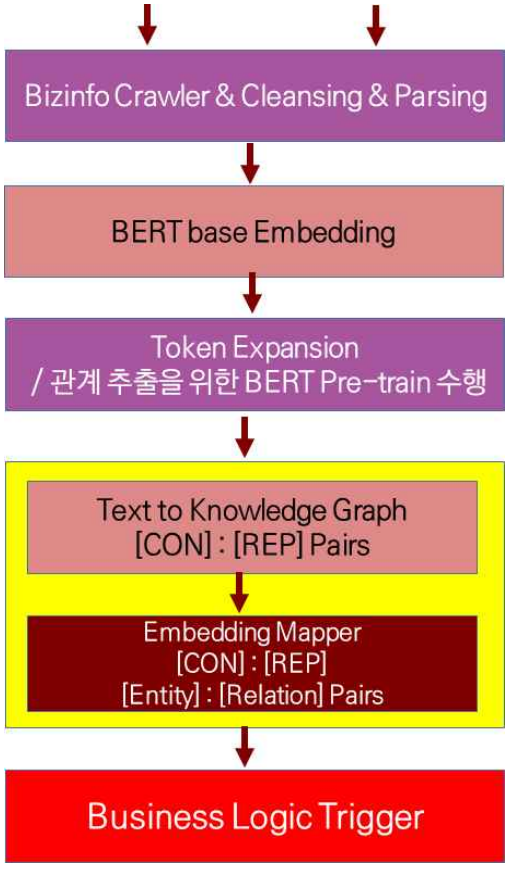


그림 19 Architecture

최근 자연어 처리 빅 모델의 대규모 자원 활용 제약으로 인하여 연구의 효율성을 기하기 위해 toy example 형태의 연구로 진행하였다. BCG Matrix 생성을 위한 기업 정보 추출을 위하여 전자공시 사이트에서 네이버(주)와 ㈜카카오의 해당 정보인 비정형 텍스트 형태의 데이터를 크롤링 및 정제하는 작업을 진행하였다. 우선 Python 팩키지인 BeautifulSoup으로 전자공시정보에서 기업 정보 중 사업 부문, 매출액, 시장규모 등의 데이터를 수집하였다. 파이썬에서 지원하는 데이터 형태인 데이터프레임 형태로 저장하고 전처리 및 정제를 수행하였다. Token Expansion을 위하여 BERT 임베딩 레이어의 Special Token인 컨설팅 카테고리의 [CON] 토큰과 해당 보고서 생성을 위한 데이터셋의 성격의 개체 정의를 위하여 [REP] 토큰을 추가하는 임베딩을 수행하였다. 좀 더 세분하여 기술하면 BERT 기반의 임베딩 데이터를 생성하기 위하여 Sentencepiece 기반 BPE(Byte Pair Encoding) 알고리즘을 적용하여 [CON], 예를들면 사업 부문, 매출, 시장규모 등과 [REP], 예를들면 부문별 사업의 매출액, 산업 성장률 등을 추출하여 BERT 기반 임베딩 입력 데이터를 생성한다. 임베딩된 데이터를 BERT 모델로 사전학습 시킨 후 확장된 임베딩 데이터를 생성한다. 추가적으로 임베딩(embedding)된 데이터를 지식 그래프(Knowledge Graph)의 개체와 관계 맵핑을 위하여 ETRI 엑소브레인 API를 사용한다. [CON]과 [REP] 토큰을 쌍(Pairs)으로 추출(Extractor)하고 지식그래프의 Entity와 Relation으로 맵핑할 수 있도록 지식그래프의 객체 인식 및 의미 부여할 수 있는 데이터를 생성한다. 마지막으로 컨설팅 보고서 생성을 위한 시각화로 Business Logic Trigger를 구현한다. BCG Matrix의 결과인 상대적 시장점유율(RMS), 산업별 성장률(cagr), 해당 경쟁기업 간 5년 단위 매출액(Revenue) 비교 산정을 통해 BCG Matrix 차트를 시각화한다.

3.2 데이터(코퍼스) 수집 및 전처리

본 연구에서는 BCG Matrix 생성을 위해 toy-example 형태로 파이썬 코드로 구현하여 데이터 수집을 진행하였다. 입력 데이터는 네이버(주)와 ㈜카카오 2020년, 2016년 데이터를 수집하여 생성하였다. 프로그래밍 툴은 파이참을 통해 구현하였으며 회사별, 연도별로 생성된 파일을 하나의 파일로 통합하여 초기 데이터

셋을 만들었다. 비정형 텍스트 크롤링 및 정제/가공(Bizinfo Crawler)을 위하여 전자공시시스템에서 해당 기업 정보를 크롤링한다. 도표 및 차트 등 다양한 형태로 데이터가 존재하지만 BERT기반의 임베딩을 위하여 텍스트 데이터 형태로 수집을 한다. Python BeautifulSoup 팩키지를 사용하여 전자공시정보에서 기업 정보 중 사업 부문, 매출액, 시장규모 등 수집 및 정제한다. 전처리 부분은 전각문자에 대한 반각 문자화 작업, 정규 표현식을 통한 빈칸 제거 등을 수행하였다. 기본적으로 한국어 문서에서의 숫자, 영자, 기호가 전각 문자일 때가 있다. 이러한 경우 일반적으로 사용되는 반각 문자로 변환해 주는 작업을 수행하였다[2]. 그리고 정규표현식(regular expression)을 활용하여 복잡한 규칙의 노이즈도 제거 치환하여 수집 데이터를 생성한다. 자연어 처리의 단계는 특수문자 제거, 공백 제거, 중복 표현 제거, 이메일, 링크 제거, 불용어 제거, 조사 제거, 띄어쓰기 등 수많은 노력과 시간이 들어간다. 해당 데이터에만 존재하는 전처리 및 가공 작업을 변환 및 수작업 등으로 진행하여야 한다. 그 외에 BCG Matrix 생성을 위한 고단한 전처리 작업을 인터랙티브하게 처리하는 방식으로 진행하였다. 정규 표현식, 해당 도메인의 특수한 경우 등 전처리는 향후 자동화하는 방안에 대한 연구가 추가적으로 필요할 것으로 보인다.

3.3 BERT 기반 임베딩

3.3.1 Sentencepiece 기반 BPE(Byte Pair Encoding) 생성

본 연구 수행의 핵심적인 역할을 하는 확장된 임베딩 데이터를 생성하기 위한 개발 환경을 구성하였다. BERT 모델은 GPU서버 환경에서 동작하게 되어 있다. 하지만 따로 고가의 개발환경을 구성하는데 어려움이 있다. 본 연구에서는 구글의 Colab을 사용하여 자연어 처리 모델의 기본 데이터가 되는 어휘사전(Vocab)을 BPE(Byte Pair Encoding)을 통해 구현하였다.

BERT 모델은 BPE로 학습한 어휘 집합을 쓴다. 해당 결과를 BERT에 사용하려면 일부 후처리가 필요하다. 언더바(_) 문자를 ##으로 바꾸고 [PAD], [UNK], [CLS], [MASK], [SEP] 등 Token을 추가한다[1]. BERT 기반의 임베딩 데이터를

생성하기 위해서는 GPU 서버 분석 환경을 구성하고 여러 가지 제약사항을 극복하고 부족하지만 Gooogle Colab jupyter notebook을 활용하여 확장된 임베딩 데이터를 생성하였다.

그림 20 Colab 활용 GPU 환경 구성

```
!nvidia-smi
Wed Aug  7 12:10:17 2019
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 418.67       Driver Version: 410.79       CUDA Version: 10.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla T4            Off  | 00000000:00:04.0 Off |                    0 |
| N/A   70C    P8    18W /  70W |      0MiB / 15079MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

그림 21 딥러닝 학습을 위한 Tesla T4 GPU 서버 활용

그림 22는 구글에서 공개한 Sentecepiece 기반 언어모델에서 BPE (Byte Pair Encoding)을 통해 임베딩 데이터를 생성하고 어휘사전(Vocab)을 만들었다.

```
!python drive/My₩ Drive/Colab/src/make_vocab/wordpiece.py ₩
--corpus=drive/My₩ Drive/Colab/rsc/training_data/corpus_bcg_toyexam.refined_colab.txt ₩
--iter=20 ₩
--fname=drive/My₩ Drive/Colab/rsc/my_conf/my_vocab.txt

terminated vocabulary scanning
('매', '##출')
('##랫', '##폼')
('영', '##업')
('##0', '##0')
('1', '##,')
('##니', '##다')
('##니다', '##.')
('##.', '##0')
('##.0', '##%')
('##습', '##니다.')
('##텐', '##츠')
('제', '##2')
('비', '##중')
('1', '##00')
('플', '##랫폼')
('광', '##고')
('콘', '##텐츠')
('반', '##기')
('##5', '##,')
('부', '##문')
('100', '##.0%')
training bpe was done
```

그림 22 BERT 학습을 위한 Vocab 생성

3.3.2 BERT 학습데이터 구축

BERT 사전학습을 위한 Preprocessed data를 만들기 위하여 구글의 Colab을 사용하여 tensorflow==1.14 환경을 우선 구성한다. BERT 기반 Tokenization을 통하여 Special Token인 [CON], [REP] 추가한다. 그리고 BCG Matrix 임베딩 데이터로 학습된 관계 추출을 위하여 그림 23과 같이 BERT Pre-train 수행한다.

```
!python drive/My\ Drive/Colab/src/make_preprocessed_data/create_pretraining_data.p
--input_file=drive/My\ Drive/Colab/rsc/training_data/corpus_bcg_toyexam.refined_co
--vocab_file=drive/My\ Drive/Colab/rsc/conf/my_vocab_colab.txt \
--do_lower_case=False \
--max_seq_length=512 \
--output_file=drive/My\ Drive/Colab/rsc/my_preprocessed_training_data/corpus_bcg_t
```

그림 23 BERT 학습을 위한 Preprocessed data 생성

그림 24는 구글의 Colab에서 Preprocessed data 생성 코드를 실행 후 얻은 결과화면이다. corpus_bcg_toy example.refined_colab_512_tf.record이라는 텐서

플로우 기반의 임베딩 데이터가 생성되었다.

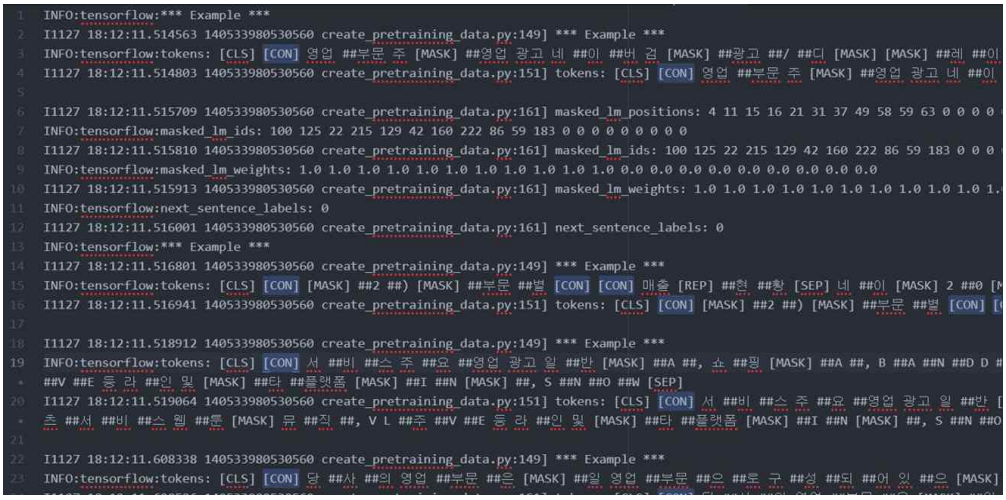


그림 24 BERT기반 Special Token Expanding 결과

그림 25는 BERT 사전학습 코드를 통하여 기존에 생성되던 [CLS], [SEP], [MASK]와 추가된 토큰인 [CON]과 [REP]이 포함된 관계 추출을 위한 Pre-trained BERT 모델을 생성한 코드화면이다. 미니배치 사이즈 및 학습률, 토큰 길이 등 하이퍼 파라미터를 설정하여 수행한다.

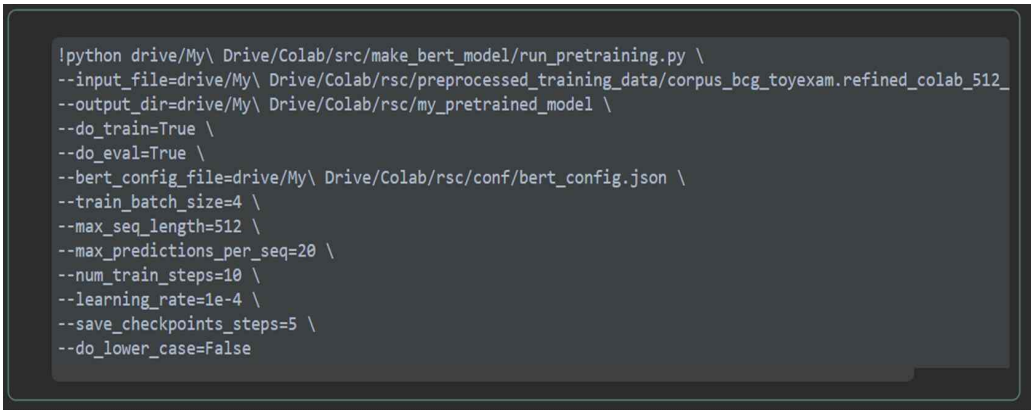


그림 25 관계 추출을 위한 Pre-trained BERT 모델 생성

그림 26은 위 작업을 통해 생성된 관계 추출을 위한 사전학습 (Pre-trained)된 BERT 모델을 보여준다.

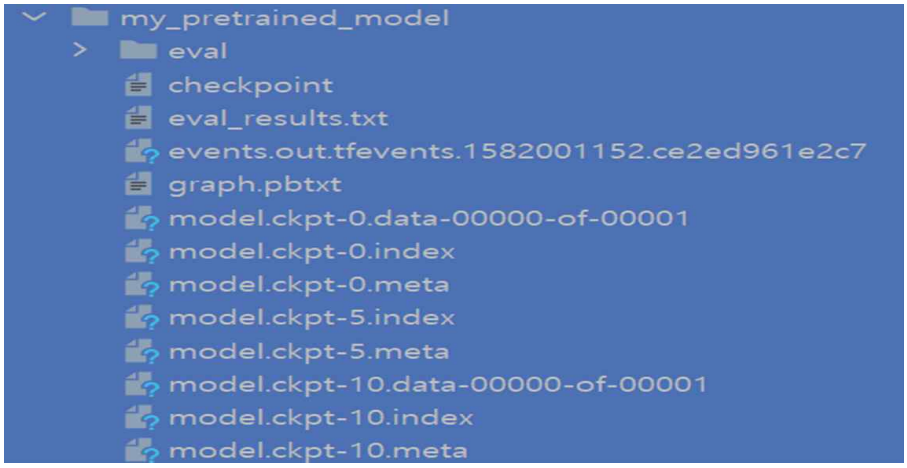
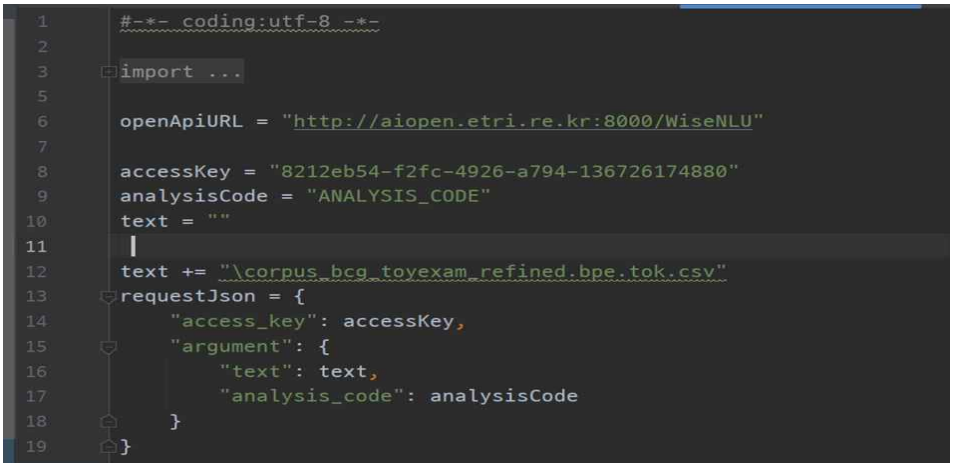


그림 26 BERT Pre-train 모델

3.4 Text to Knowledge Graph

3.3.1 지식 그래프 개체 및 관계 추출

개체명 인식은 자연어 처리 분야에서 각 단어가 개체명에 대한 분석을 진행하고 개체명이라면 어떤 개체명 카테고리에 속하는지 판단하는 전통적인 분류 문제(Classification Task)에 해당한다[7]. ETRI 오픈 API 서비스 포털(http://aiopen.etri.re.kr/guide_wiseNLU.php#group01) 개체명 태그셋은 15개 대분류 및 146개 좀 더 세분화된 분류체계로 구성된 표준 개체명 태그셋을 사용하였다. 그림 27는 ETRI API 발급을 통해 파이썬 스크립트에서 수행한 코드이다[부록4].



```
#-*- coding:utf-8 -*-

import ...

openApiURL = "http://aiopen.etri.re.kr:8000/WiseNLU"

accessKey = "8212eb54-f2fc-4926-a794-136726174880"
analysisCode = "ANALYSIS_CODE"
text = ""

text += "\corpus_bcg_toyexam_refined.bpe.tok.csv"
requestJson = {
    "access_key": accessKey,
    "argument": {
        "text": text,
        "analysis_code": analysisCode
    }
}
```

그림27 ETRI 개체명 분석 API

BCG Matrix에서 X축의 상대적 시장점유율 해당 산업 군에서 가장 큰 시장점유율을 가지는 경쟁사들과 비교하여 시장점유율 및 산업 성장률을 산출한다. 그래서 IT 플랫폼 및 검색서비스 사업의 가장 큰 두 회사인 네이버(주)와 ㈜카카오를 비교하여 분석하게 되었다. ETRI API를 통해 정제된 임베딩 데이터셋을 기반으로 네이버(주) 2020년 / 2016년 반기보고서에 대한 개체명 분석 및 구문 분석 결과를 도출하였다. 일반적인 자연어 처리 분석 과정은 형태소에 대한 분석, 구문에 대한 분석, 의미역 분석의 순서를 따라 진행된다. 학습 데이터 양이 충분하지 않아 일부 개체명 인식 및 구문 분석 결과가 잘 분류되지 않은 경우가 있었지만 대체로 TTA 표준 개체명 태그셋에 정의된 분류 체계에 대부분 속하여 형태소 및 동음이의어, 개체명이 분류된 것을 확인할 수 있었다. 구문 분석 결과도 네이버(주)의 사업부문과 해당 사업부문별 가중치가 높게 나옴을 유추할 수 있었다. 몇 가지 분석 사례를 보면 그림 28의 네이버(주) 2020년 반기보고서 개체명 분석 및 구문 분석 결과의 개체명 분석에서도 네이버(주)의 경우 "경제 관련 기관/단체, 기업"이라는 분류기준에 부합하게 분류되고 있었다. 사업분야도 "IT플랫폼"의 경우 "과학관련 이론/법칙/방식/양식/체계/학설"이라는 분류체계를 잘 따르고 있음이 보여준다. 의존 관계 구문 분석의 경우 "분기 보고서"는 "네이버(주)&2020&영업수익"이란 다중 지배소 형태의 관계를 형성하는 것을 보여주고 있다. 또한 "IT 플랫폼"은 사업분야에서 "매출액 - 1,385,227"과 "비지니스 플랫폼&콘텐츠 서비스" 라는 의존 관계를 보여주고 있다. 이처럼 BERT 학습으로 확장 생성된

임베딩 데이터를 기반으로 ETRI 엑소브레인을 활용하여 유의미한 결과를 도출하였다. 좀 더 자세히 설명하면 그림28의 개체명 분석 결과에서는 네이버(주) 및 사업부문이 ETRI 개체명 표준을 준수하여 분류되었음을 확인할 수 있다. 구문분석에서 각 어절은 자신 어절과 지배소 어절 사이의 관계를 표현하는 의존 관계 태그를 가진다. TTA 표준 개체명 태그셋에서는 지배소 후위 원칙에 따라 각 어절의 지배소는 자신보다 뒤에 위치하도록 분석한다. 318,686이라는 광고 부문 매출액은 네이버(주)의 2020년 플랫폼 부문 매출에 다중 지배소 형태의 의존관계를 형성하고 있다. 이는 해당 부문에 카테고리에 매칭되어 BCG Matrix 부문별 매출을 통합할 때 분류 체계에 적용된다.

그림 28 네이버(주) 2020년 반기보고서 지식그래프 분석 결과 예시(부록2)

형태소/개체명/동음이의어 결과 | 구문분석 결과 | 의미역 인식 결과

No	Word	형태소			동음이의어			개체명		
		단어	태그	설명	단어	형태소 태그	의미번호	단어	태그	설명
0	네이버㈜	네이버 ㈜	NNP SW	고유명사 기타 기호 (논리 수학기호, 기호 등)	네이버 ㈜	NNP SW	00 00	네이버㈜	OGG_ECONOMY	경제 관련 기관/단체, 기업
1	2020년	2020 년	SN NNB	숫자 의존명사	2020 년	SN NNB	00 02	2020년	DT_YEAR	년
2	플랫폼	플랫폼	NNG	일반명사	플랫폼	NNG	00			
3	부문	부문	NNG	일반명사	부문	NNG	06			
4	광고	광고	NNG	일반명사	광고	NNG	02			
7	BAND	BAND	SL	외국어	BAND	SL	00			
8	DA,	DA ,	SL SP	외국어 쉼표, 가운뎃점, 콜론, 빗금	DA ,	SL SP	00 00	BAND DA	OGG_ECONOMY	경제 관련 기관/단체, 기업
9	네이버TV	네이버 TV	NNP SL	고유명사 외국어	네이버 TV	NNP SL	00 00	네이버 TV	OGG_ECONOMY TMI_HW	경제 관련 기관/단체, 기업 IT 하드웨어 용어

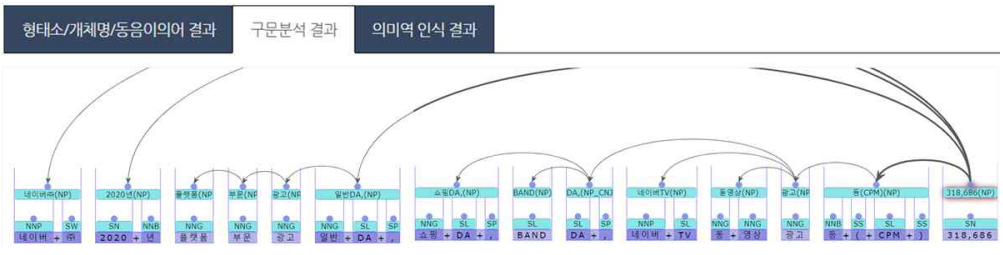


그림 29는 BCG Matrix에서 시장 성장률을 계산하기 위하여 현재와 5년 전 시장규모에 대한 원천 데이터를 비교하기 때문에 네이버(주) 2020년과 2016년의

5년 전 데이터를 같이 분석하였다. 해당 개체명 분석도 2020년도와 동일하게 분류되었다. 그런데 ETRI 의존 구문 분석 어휘사전에서도 빈도수에 의하여 ”광고“라는 키워드가 다중 지배소 역할을 하는 것으로 나타났으며 네이버(주) 및 2016년, 플랫폼 부문이라는 개체명을 통해 의존관계를 형성하고 있다.

그림 29 네이버(주) 2016년 반기보고서 지식그래프 분석 결과 (부록2)

형태소/개체명/동음이의어 결과 | 구문분석 결과 | 의미역 인식 결과

No	Word	형태소			동음이의어			개체명		
		단어	태그	설명	단어	형태소 태그	의미번호	단어	태그	설명
0	네이버㈜	네이버 ㈜	NNP SW	고유명사 기타 기호 (논리 수학기호, 기호 등)	네이버 ㈜	NNP SW	00 00	네이버㈜	OGG_ECONOMY	경제 관련 기관/단체, 기업
1	2016년	2016 년	SN NNB	숫자 의존명사	2016 년	SN NNB	00 02	2016년	DT_YEAR	년
2	플랫폼	플랫폼	NNG	일반명사	플랫폼	NNG	00			
3	부문	부문	NNG	일반명사	부문	NNG	06			
4	네이버	네이버	NNP	고유명사	네이버	NNP	00	네이버	OGG_ECONOMY	경제 관련 기관/단체, 기업
5	검색광고/	검색 광고 /	NNG NNG SP	일반명사 일반명사 쉼표, 가운뎃점, 콜론, 빗금	검색 광고 /	NNG NNG SP	00 02 00	검색광고	TMI_SERVICE	IT 서비스 용어
6	디스플레이광고/	광고 /	NNG SP	일반명사 쉼표, 가운뎃점, 콜론, 빗금	광고 /	NNG SP	02 00			
7	지식쇼핑	지식 쇼핑	NNG NNG	일반명사 일반명사	지식 쇼핑	NNG NNG	02 00			
8	수수료,	수수 료 ,	NNG XSN SP	일반명사 명사파생접미사 쉼표, 가운뎃점, 콜론, 빗금	수수료 ,	NNG SP	00 00			
9	라인	라인	NNG	일반명사	라인	NNG	01			
10	광고	광고	NNG	일반명사	광고	NNG	02			
11	등	등	NNB	의존명사	등	NNB	05			
12	1,395,572	1,395,572	SN	숫자	1,395,572	SN	00	1,395,572	QT_PRICE	금액

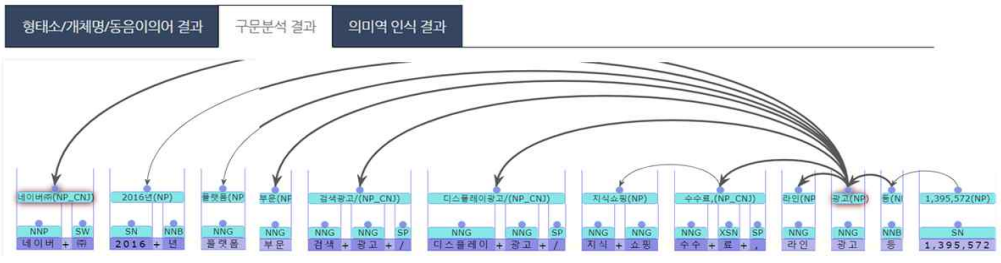


그림 30에서는 ‘카카오’라는 개체명 인식에서 분류 매핑이 잘못되는 오류가 발생하였다. 이는 과일으로 잘못 인식하는 오류였으며 이는 추가적인 기업 명칭

‘㈜카카오’라는 어휘사전 추가를 통한 보완을 하였다. 특정 도메인에 특화된 어휘사전 생성이 현재로서 부족한 부분이 많이 있어 기업공시정보 및 유관 정보에 대한 대용량의 비정형 텍스트 데이터의 수집 및 본 연구의 프로세스를 반영한 Special Token을 추가한 BERT 기반의 임베딩 데이터 생성을 통해 보완하면 전문용어나 도메인 특화 개체의 인식률을 높일 수 있다.

그림 30 ㈜카카오 2020년 반기보고서 지식그래프 분석 결과 (부록2)

형태소/개체명/동음이의어 결과 | 구문분석 결과 | 의미역 인식 결과

No	Word	형태소			동음이의어			개체명		
		단어	태그	설명	단어	형태소 태그	의미번호	단어	태그	설명
0	㈜카카오	㈜ 카카오	SW NNP	기타 기호 (논리 수학기호, 기호 등) 고유명사	㈜ 카카오	SW NNP	00 00	㈜카카오	OGG_ECONOMY	경제 관련 기관/단체, 기업
1	2020년	2020 년	SN NNB	숫자 의존명사	2020 년	SN NNB	00 02	2020년	DT_YEAR	년
2	플랫폼	플랫폼	NNG	일반명사	플랫폼	NNG	00			
3	부문	부문	NNG	일반명사	부문	NNG	06			
4	톡비즈,	톡비즈 ,	NNP SP	고유명사 쉼표, 가운뎃점, 콜론, 빗금	톡비즈 ,	NNP SP	00 00			
5	포털비즈,	포털 비즈 ,	NNG NNG SP	일반명사 일반명사 쉼표, 가운뎃점, 콜론, 빗금	포털 비즈 ,	NNG NNG SP	00 00 00			
6	신사업	신 사업	XPN NNG	체언접두사 일반명사	신사업	NNG	00			
7	934,516	934,516	SN	숫자	934,516	SN	00	934,516	QT_OTHERS	기타 수량 표현

형태소/개체명/동음이의어 결과 | 구문분석 결과 | 의미역 인식 결과

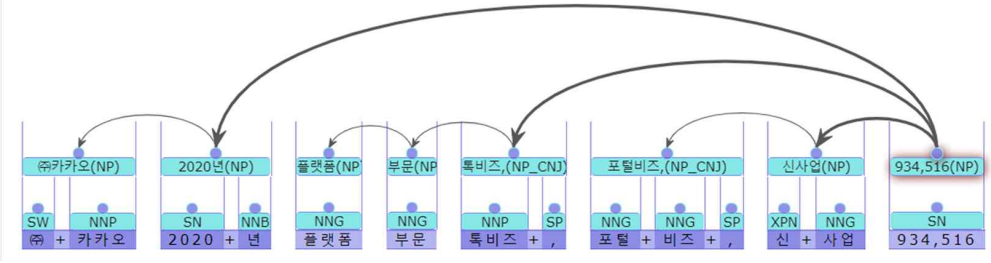
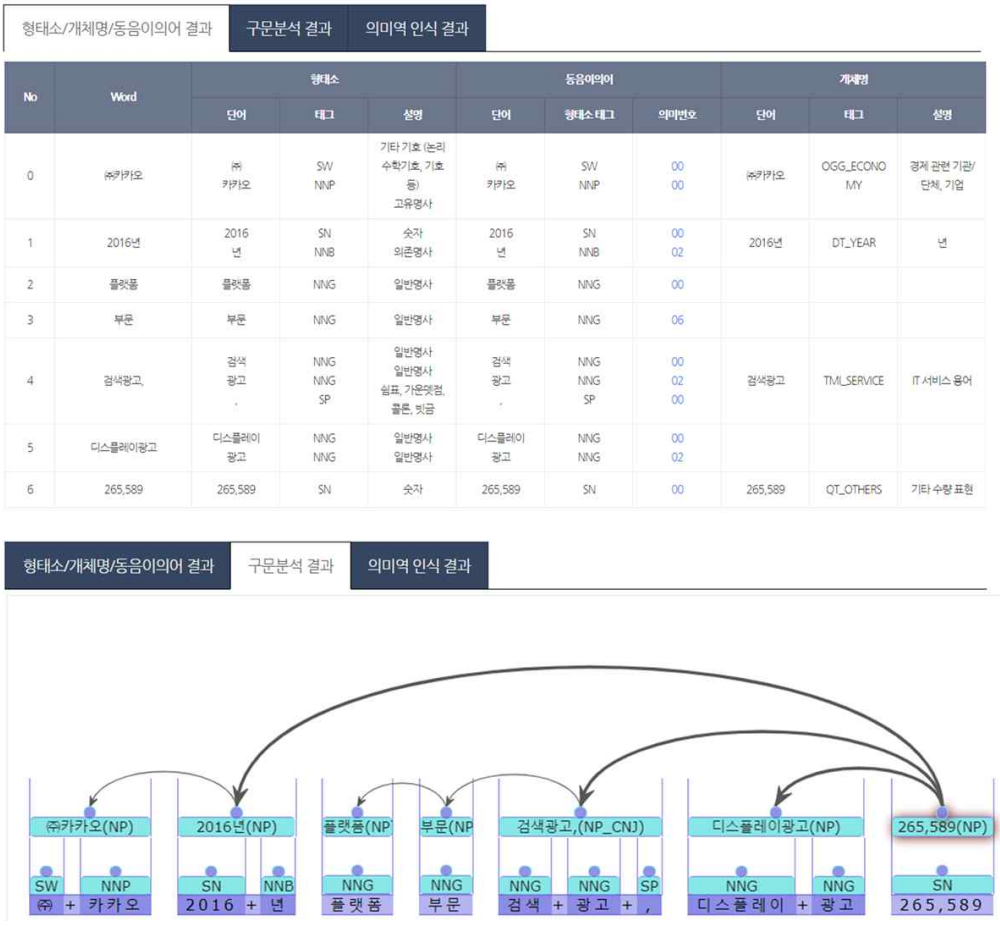


그림 31의 경우에도 ㈜카카오의 개체명 분류가 ETRI 분류기준에 맞게 생성되었으며, TTA 표준 개체명 태그셋에서는 지배소 후위 원칙에 따라 가장 후위에 위치한 플랫폼 부문의 매출액 ‘265,516’이 다중 지배소의 역할을 하고 있다. 이는 BCG Matrix를 생성하기 위한 요소 추출할 때 Business Logic 적용을 정확하게 수행하기 위한 기반이 된다.

그림 31 ㈜카카오 2016년 반기보고서 지식그래프 분석 결과 (부록2)



형태소/개체명/동음이의어 결과 | 구문분석 결과 | 의미역 인식 결과

No	Word	형태소			동음이의어			개체명		
		단어	태그	설명	단어	형태소 태그	의미번호	단어	태그	설명
0	㈜카카오	㈜ 카카오	SW NNP	기타 기호 (논리 수학기호, 기호 등) 고유명사	㈜ 카카오	SW NNP	00 00	㈜카카오	OGG_ECONOMY	경제 관련 기관/단체, 기업
1	2016년	2016 년	SN NNB	숫자 의존명사	2016 년	SN NNB	00 02	2016년	DT_YEAR	년
2	플랫폼	플랫폼	NNG	일반명사	플랫폼	NNG	00			
3	부문	부문	NNG	일반명사	부문	NNG	06			
4	검색광고,	검색 광고 ,	NNG NNG SP	일반명사 일반명사 쉼표, 가운뎃점, 콜론, 빗금	검색 광고 ,	NNG NNG SP	00 02 00	검색광고	TMI_SERVICE	IT 서비스 용어
5	디스플레이광고	디스플레이 광고	NNG NNG	일반명사 일반명사	디스플레이 광고	NNG NNG	00 02			
6	265,589	265,589	SN	숫자	265,589	SN	00	265,589	QT_OTHERS	기타 수량 표현

다만 아쉬운 부분은 학습 데이터 전처리 및 가공에 대한 제약이 없었다면 자동화된 대용량 데이터셋 수집, 전처리 및 가공을 통해 사전학습된 어휘사전의 지식그래프 개체 및 관계에 의한 가중치가 좀 더 명확해질 것으로 생각한다. 또한 ETRI 엑소브레인 태깅 사전이 비즈니스 컨설팅 도메인에 특화된 분류체계를 세부적으로 제공하지 않아 분석을 확대하기 어렵고 정량적으로 비교하기 힘든 부분은 아쉬운 부분이었다. 더불어 2020년과 2016년 산업군 변화에 따른 사업 부문의 변동을 반영하지 못한 부분은 보완이 필요하다. 시계열적인 변화를 반영하기 위하여 다변화된 산업군을 포괄하는 추가적인 분류 체계도 필요할 것으로 보인다. 구문 분석 결과에서 분절(Subword)이 너무 세세하게 나뉘어 사업 부문과 매출액

이 정확하게 매칭되지 못한 사례도 발생하였다. BERT 기반의 Pre-train 데이터 셋이 너무 적어 학습효과가 나타나지 못한 것으로 해석할 수 있다. 이 부분은 향후 수집 및 정제, 가공의 자동화 방안을 수립하고 GPU 분석 환경 등이 충분히 확보되면 추가 연구를 통하여 보완할 필요가 있다.

Ⅳ. 실험 결과 및 분석

4.1 문장/단어 유사도 분석

일반적으로 임베딩 데이터로 만들어 딥러닝 기반의 자연어 처리 모델에 활용하는 이유는 문장/단어 간의 유사도 측정이 용이할 뿐만 아니라 관계 파악에도 용이하기 때문이다. 벡터 연산을 통한 추론도 가능하다. 본 연구 결과에 대한 성능 검증은 여러 가지 제약사항으로 다운스트림 Task 수행을 통한 리더보드 검증 등 벤치마크를 통해 성능향상에 대한 결론을 내리기 어려운 부분이 있어 임베딩된 문장에 대해 문장/단어 유사도 분석을 통해 임베딩 품질에 대한 결론을 유도해 보고자 한다. 해외 사례로 Word embedding 성능 검증은 주로 WordSim353으로 실행한다. WordSim353은 13-16명의 사람이 annotate한 두 단어의 유사도 정답을 제시한다. 이를 기반으로 두 문장/단어 벡터의 코사인 유사도(Cosine similarity)를 구한 후 정답과 Spearman`s Rank-order correlation값을 산출하여 경험상 0.7 이상의 값이 나오면 임베딩이 잘 수행되었다고 평가하는 방식이다. 국내에서는 문장/단어 유사도 평가를 위한 국내 특정 도메인에 대한 테스트 데이터셋이 별도로 공개되지 않아, ETRI 엑소브레인 문장 단위의 유사도 측정 API를 통해 검증을 수행하였다.

4.1.1 문장 단위 유사도 분석

임베딩 데이터의 품질을 검증하는 방법으로 우선 문장 유사도를 적용하여 검증하였다. 통상 코사인 유사도(Cosine Similarity)가 7 이상이면 임베딩 품질이 괜찮다고 평가한다. 본 연구의 결과에서 해당 임베딩 데이터에 대한 문장 유사도 분석을 통해 7.7 정도로 유의미한 유사도를 나타 내었다. 그림 34에서와 같이 문

장 단위의 유사도는 별도의 한글화된 테스트 셋이 없어도 임베딩이 잘 수행되었음을 보여주고 있다. 네이버(주) 2020년 vs ㈜카카오 2020년 문장 유사도 분석이 7.76으로 임베딩이 적정하게 분류되었음을 유추할 수 있다.

그림 32 문장 단위 유사도 분석

4.1.2 단어 유사도 분석

문장 유사도와 더불어 임베딩 데이터의 품질 검증을 위하여 테스트한 단어 유사도는 7 보다 낮게 나타났다. 이는 학습데이터 양과 BERT 사전 학습 시간에 대한 자원 제약으로 인한 것으로 보인다. 이로 인한 제약으로 본 연구에서는 단어 유사도로 BCG Matrix 연산을 위하여 양의 관계를 가진 개체(Entity)에 대해 가중치를 보정하여 기준을 낮춰 적용하였으며 일부 유의미한 결과를 도출하였다. 그림 35의 단어 유사도 분석은 해외 유사도 분석 방법인 WordSim353 형태의 국내 해당 도메인 영역의 테스트 셋 구축이 되어야 개선될 것으로 보인다. 본 연구에서 여러 제약사항이 있었지만 향후 특정 도메인의 대용량 데이터 수집 환경을 만들어 보완하는 연구가 필요할 것으로 보인다.

Entity A	Entity B	similarity
영업현황	매출현황	84%
영업부문	주요영업	75%
광고	디스플레이광고	69%
광고 플랫폼	콘텐츠 플랫폼	65%
IT플랫폼	IT서비스	64%
광고	네이버 검색광고	60%
2020	2016	55%
분기보고서	반기	26%
콘텐츠	웹툰	25%
합계	1,821,230	22%
금액	1,526,907	22%
비즈니스플랫폼	쇼핑검색	20%
콘텐츠	뮤직	15%
광고	네이버TV	14%
광고	수수료	14%
광고	동영상	13%
매출액	265,589	13%
광고	쇼핑DA	12%
광고	지식쇼핑	12%
영업수익	3,634,532	12%
IT플랫폼	네이버페이	10%
합 계	619,003	8%
광고	318,686	6%
매출	1,924,562	5%
비즈니스플랫폼	1,526,907	5%

그림 33 단어 유사도 분석

4.2 BERT 기반 임베딩 데이터 생성

Toy example의 한계는 있었지만 대용량 텍스트 데이터에서 [CON] : [REP] 추가 및 임베딩을 통하여 최신 자연어 처리 모델의 비지도 학습 데이터를 생성하여 긴 문맥에 대한 한계점과 도메인 제약, 편향성 등을 개선할 수 있는 BERT 기반의 임베딩 데이터를 생성해 보았다. BERT 성능에 영향을 미치는 말뭉치의 크기, 특히 전문 용어와 같이 저빈도 핵심 어휘는 임베딩 사전에서 누락되는 문제가 발생한다. 그리고 말뭉치의 도메인 영역의 사전, 말뭉치 분절(Tokenizing)과 어휘사전(Vocab) 크기의 제약으로 인해 결론 도출에 제약이 있었다. 하지만 이러한 보완점은 향후 대용량 말뭉치 수집 및 비정형 비즈니스 컨설팅 영역의 논문 자료 등 대용량 환경하에서 개별 프로세스를 통합하여 자동화시킬 수 있는 기반을 마련한 연구였음을 확인할 수 있었던 것에 의의가 있다.

Ⅴ. 결론

5.1 BCG Matrix 학습용 데이터 생성 및 시각화

데이터 전처리 및 분절(Tokenization)과 형태소 분석, 개체명 인식 및 구문 분석의 과정을 거쳐 전반적인 실험을 진행하였다. 단계별로 실험 환경 및 기반이 상이한 부분이 많아 전체적인 통합 구현을 통한 완전한 자동화 부분은 부족한 부분이 있으나 전반적인 단계별 실험은 진행을 하였다. 주요 부분을 설명하면 아래와 같다. 임베딩 데이터 생성 부분은 Special Token 추가를 통하여 [CON], [REP]와 Entity, Relation에 대한 비즈니스 로직을 반영한 결과를 산출하고 산출된 결과에 대한 BCG Matrix 시각화 / 차트를 생성하였다. BCG Matrix 생성을 위하여 파이썬 데이터프레임 형태로 Pandas 패키지를 사용하였다. 그림 32과 같이 주피터 노트북에서 데이터프레임 형태로 데이터 분석 팩키지인 Pandas를 사용하여 BCG Matrix 생성에 필요한 데이터를 추출하여 해당 표를 생성하였다.

그림 32 BCG Matrix에 산출된 수식을 설명하면 상대적 시장 점유율(RMS : Relative Market Share) 는 네이버(주)와 ㈜카카오의 2020년 매출액의 비교하여 해당 회사 중 가장 큰 매출액을 비교 기준인 네이버(주)의 매출액과 백분율로 나눈 비율을 나타낸다. 관련 로직을 적용하여 플랫폼 부문은 2.3이 산출되었으며 콘텐츠 부문은 0.2가 산출되었다. 두 번째 요소인 시장 성장률 cagr. (Compound Annual Growth Rate) 은 5년 기간인 2020년/2016년 전체 기업 합계 매출액을 재무함수인 RATE(기간, 0, 첫 번째 수치, 마지막 수치)를 활용하여 산출한다. 플랫폼 부문 cagr은 재무함수 =RATE(5, 0, 3108486, 1021697)을 적용하여 18.1%가 산출되었다. 콘텐츠 부문은 재무함수 =RATE(5, 0, 1354053, 680279) 을 적용하여 8.5%가 산출되었다. 마지막으로 비교 기준 회사인 네이버(주)의 매출액의 크기인 상대적인 원의 크기를 시각화하기 위한 플랫폼 부문 매출액인 2,173,970과 콘텐츠 부문 매출액 134,982을 산출하였다.

그림 34 BCG Matrix Extracting

```
12 import pandas as pd

   train_data = pd.read_csv('./corpus_bcg_toyexam_refined.csv')
   train_data.head()
```

12

	Unnamed: 0	RMS	cagr	Before volumn	Present volumn	Naver Revenue	kakao Revenue
0	Platform	2.3	18%	1,354,053	3,108,486	2,173,970	934516
1	Contents	0.2	8%	680,279	1,021,697	134,982	886715

```
13 train_data[['RMS', 'cagr','Naver Revenue']] #['RMS', 'cagr','Naver Revenue']] # 필요한 항목
```

13

	RMS	cagr	Naver Revenue
0	2.3	18%	2,173,970
1	0.2	8%	134,982

5.1.1 BCG Matrix의 데이터 기반 시각화

BCG Matrix 생성을 위한 Business Logic을 적용하여 해당 데이터에서 Python matplotlib의 Scatter 시각화 팩키지를 사용하여 Visualization을 수행하였다. X축인 상대적 시장점유율은 네이버(주)와 ㈜카카오를 비교하는 해당 도메인 영역에서 주요 경쟁사와 비교하는 당사와 최대경쟁자와의 매출 기반 상대적 시장점유율(RMS:Relative Market Share)는 플랫폼 영역에서 지수가 2.3이 생성되었으며 콘텐츠 부문에서는 0.2가 생성되었다. 성장률인 Y축은 산업군성장률(Compound Annual Growth Rate)로 해당 IT플랫폼과 콘텐츠 산업의 시장 성장률 (해당 도메인 영역의 마켓 성장률이며 그 기준이 분석 대상 기업이 속한 국가의 국내 총생산과 비교를 통하여 플랫폼 영역에서 18%가 생성되었으며 콘텐츠 부문에서는 8%가 생성되었다. 마지막으로 매출액인 원의 크기는 네이버(주) 사업의 매출액의 크기로써 비중이 가장 큰 사업을 기준으로 상대적인 크기로 표현되고 있다. 이는 분석 대상 기업인 네이버(주)

는 Star사업부문인 플랫폼 부문에 집중하고 dog 사업부문인 콘텐츠 부문은 투자를 줄이는 포트폴리오를 구성하는 전략을 보여준다. 본 연구에서는 특정 산업군과 특정 기업에 국한된 toy-example 형태의 연구였지만 BERT 임베딩 데이터 생성을 위한 비즈니스 컨설팅 영역의 충분한 어휘사전 생성 등 자원 확보와 학습 시간을 추가하면 다양한 컨설팅 보고서 생성을 기대할 수 있다.

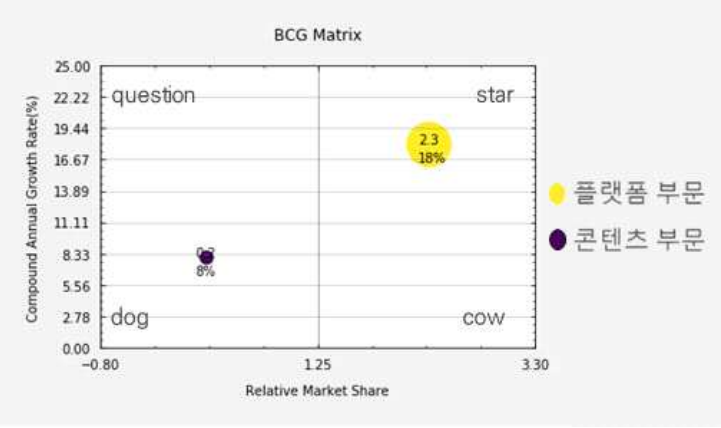


그림 35 BCG Matrix 시각화

5.1.2 BCG Matrix의 한계

자동으로 생성된 비즈니스 컨설팅 보고서의 경우에도 아직까지 상용화하기에는 부족한 부분이 많은 것이다. BCG Matrix의 경우 논문이나 자료에 나타난 형태의 산출이 되더라도 몇 가지 한계점이 존재한다. 시장점유율과 수익성은 정비례 하지 않을 수도 있다. 경쟁이 치열한 시장에서 출혈경쟁이나 일어나면 시장점유율이 상승할수록 수익은 감소할 수 있다. 또한 시장 성장률과 시장 점유율의 높고 낮음을 판단하는 기준이 상대적이라 기준에 따라 잘못 적용될 수도 있다. 과거와 달리 시장 성장률만이 시장 기회를 결정하는 것이 아니라 기술이나 생산성도 영향을 미친다. 이러한 한계들이 여전히 전문 컨설턴트의 역량과 보완적인 수단으로 딥러닝 기반의 자연어 처리 모델이 함께 고민하여 연구를 고도화해야 할 것이다.

Ⅵ. 추가연구 계획

6.1 conBERT : 자동 컨설팅 보고서 생성

6.1.1 복합적인 도메인 간의 학습 데이터 생성

최신 언어모델을 활용, 텍스트 기반의 지식 그래프 자동 생성을 통해 학습 데이터 생성의 한계를 극복하고 여러 도메인 간의 관계 추출을 통해 복합적인 도메인 간의 새로운 관계를 찾아내 문제 해결 방안을 제시할 수 있도록 하는 연구를 추가로 진행하고자 한다.

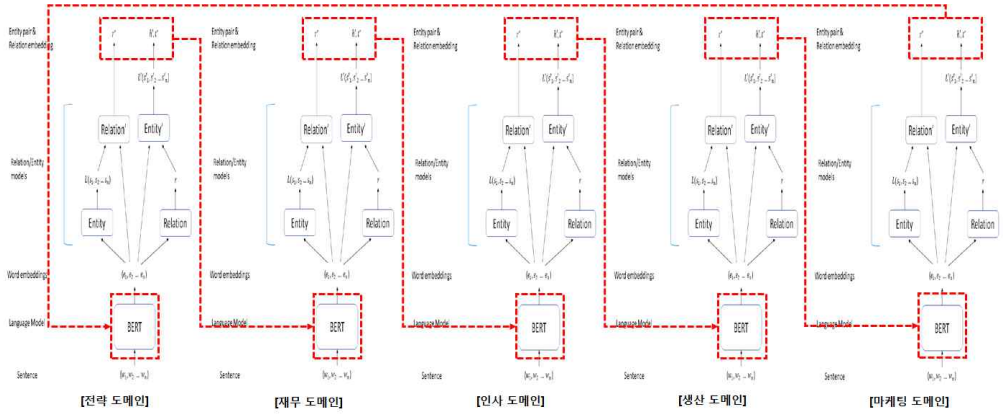


그림 36 복합적인 도메인간의 학습데이터 생성
[참조 : 홍기원외 (2019). BERT와 지식베이스 관계 임베딩 추출 방법. 한국정보과학]

6.2 자동 콘텐츠 생성 연구 진행

향후 추가적으로 텍스트 기반에서 비정형화된 데이터(차트, 그래프, 그림 등)를 통합하여 자동 콘텐츠 생성 연구를 지속하고자 한다. 기존의 정형화된 논문이나 학술 논문을 통한 전문 지식 컨텐츠에서 텍스트 외에 차트, 그래프, 그림 등 비정형화된 데이터를 융합하여 자동 컨텐츠의 신뢰도나 정확성을 향상시키는 모델을 구현하고자 한다. BERT 등 최신 자연어 처리 모델을 이용

하여 컨설팅 관련 자료를 Q/A 방식으로 학습하고 이를 이용하여 비즈니스, IT, 빅데이터 등 다양한 컨설팅 분야에 적용하는 연구를 진행하고자 한다. 향후 핵심적인 개발 분야로는 정형화된 데이터와 비정형화된 데이터(그림, 그래프, 차트 등)를 통합해서 학습할 수 있도록 하고 잘못된 결과에 대한 체계적인 분석을 통하여 해결할 수 있는 방안을 도출한다. 최신 자연어 처리 모델이 Q/A 시스템에서 상당히 높은 수준의 성능을 보여주고 있고, 정형, 비정형화된 데이터를 통합하고 학습하여 결과의 신뢰도를 높인다는 것을 향후 고도화 방향으로 지속적인 연구를 하고자 한다. 정형, 비정형화된 데이터 통합 연구가 확산 초기 단계여서 향후 자동화된 컨텐츠 생성 및 Q/A 시스템의 핵심 주제가 될 것으로 예상되기 때문이다.

참 고 문 헌

1. 국내문헌

김기현.(2019). 자연어처리 딥러닝, 한빛미디어, pp. 111-120, P133-138

조휘열. (2017). “딥러닝 기반 텍스트 질의응답을 위한 지식 추출 데이터 증강 기법” 서울대학교 대학원 석사학위 논문. pp 9

양성민.(2019).“개체명 인식을 위한 BERT 언어 모델 기반 강화학습 적용에 대한 연구”. 가천대학교 대학원 석사학위 논문. pp 4-7

유소엽(SoYeop Yoo), 정옥란(OkRan Jeong). (2019) .BERT 모델과 지식 그래프를 활용한 지능형 챗봇. 가천대학교.

이기창. (2019).『한국어임베딩』. 에이콘, pp. 28-30, pp.104, pp.237.

이동준, 임유빈, 권태경, 서울대학교 컴퓨터공학부.(2015).『Journal of KIISE』, Vol. 45, No. 5 : 형태소 기반 효율적인 한국어 단어 임베딩 , p444-450.

이태석. (2019). “맥 기반의 워드임베딩과 Out-of-Vocabulary 마스킹을 통한 문서 요약”. 국민대학교 일반대학원 박사학위 논문. pp 29-33

임준호, 배용진, 한국전자통신연구원 지식마이닝연구실, 울산대학교 국어국문학과1, 충남대학교 컴퓨터공학과. (2015). 제27회 한글 및 한국어 정보처리 학술대회 논문집 : 의존 구문분석을 위한 한국어 의존관계 가이드라인 및 엑소브레인 언어분석 말뭉치

한국정보통신기술협회.『개체명 태그 세트 및 태깅 말뭉치』.(TTAK.KO-10.0852).

홍기원, 맹성현. (2019) . BERT와 지식베이스 관계 임베딩의 융합을 통한 자연어 문장으로부터의 개방형 관계 임베딩 추출 방법. KAIST.

Dong Bok Lee1, Seanie Lee1, Woo Tae Jeong. KAIST1, AITRICS2, 42Maru Inc. Generating Diverse and Consistent QA pairs from Contexts with Information-Maximizing Hierarchical Conditional VAEs, 211

Jinhyuk Lee, Wonjin Yoon, Korea University, Seoul 02841, Korea : 『

BioBERT: a pre-trained biomedical language representation model for biomedical text mining』

2. 국외문헌

Ashish Vaswani, Noam Shazeer : *Attention Is All You Need,* (Google Brain, Google Research, University of Toronto) 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Denis Lukovnikov, University of Bonn. (2017). *Neural Network-based Question Answering over Knowledge Graphs* on Word and Character Level.

Jacob Devlin, Ming-Wei Chang. *BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding,* (Google AI Language, 24 May 2019)

Martin Reeves, Sandy Moose, and Thijs Venema. (2014). *BCG Classics Revisited: Perspectives, The Growth share matrix,* The Boston Consulting Group, Inc.

Rajarshi Das Tsendsuren Munkhdalai Xingdi (Eric) Yuan Adam Trischler Andrew McCallum / ICLR 2019 | April 2019/College of Information and Computer Sciences University of Massachusetts, *Amherst : Building Dynamic Knowledge Graphs from Text using Machine Reading Comprehension*

Rico Sennrich, Barry Haddow, Alexandra Birch School of Informatics, University of Edinburgh : *Neural Machine Translation of Rare Words with Subword Units.* pp. 3

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah. (2020). *Language Models are Few-Shot Learners,* - OpenAI, arXiv:2005.14165v4 [cs.CL] 22 Jul 2020, 33-34

Zhuosheng Zhang, Department of Computer Science and Engineering,

Shanghai Jiao Tong University : *Semantics-aware BERT for Language Understanding,* (finearXiv: 1909.02209v2 [cs.CL] 20 Nov 2019)

부 록

1. 수집 Crawling 소스코드

```
import requests
from bs4 import BeautifulSoup
import re

# 네이버(주) 매출정보 크롤링 #
rq_naver = requests.get(

"http://dart.fss.or.kr/report/viewer.do?rcpNo=20200814001974&dcmNo=7446945&eleId=10&offset=177077&length=157288&dtd=dart3.xsd")
rqctnt_naver = rq_naver.content
soup = BeautifulSoup(rqctnt_naver, "html.parser")
naver = str(soup.find_all("tr"))
naver = re.sub('<.+?>', '', naver, 0).strip()
fd = open("naver.csv","w",encoding='utf-8')
fd.write(naver)
print(naver)


# 네이버(주) 2016년도 매출정보 크롤링 #
rq_naver_2016 = requests.get(

"http://dart.fss.or.kr/report/viewer.do?rcpNo=20160816002079&dcmNo=5262442&eleId=10&offset=129708&length=132448&dtd=dart3.xsd")
rqctnt_naver_2016 = rq_naver_2016.content
soup = BeautifulSoup(rqctnt_naver_2016, "html.parser")
naver_2016 = str(soup.find_all("tr"))
naver_2016 = re.sub('<.+?>', '', naver_2016, 0).strip()
fd = open("naver_2016.csv", "w", encoding='utf-8')
fd.write(naver_2016)
print(naver_2016)

# ㈜카카오 매출정보 크롤링 #
rq_kakao = requests.get(

"http://dart.fss.or.kr/report/viewer.do?rcpNo=20200814002188&dcmNo=7447711&eleId=10&offset=287158&length=197448&dtd=dart3.xsd")
rqctnt_kakao = rq_kakao.content
soup = BeautifulSoup(rqctnt_kakao, "html.parser")
```

```
kakao = str(soup.find_all("tr"))
kakao = re.sub('<.+?>', '', kakao, 0).strip()
fd = open("kakao.csv", "w", encoding='utf-8')
fd.write(kakao)
print(kakao)

# ㈜카카오 2016년도 매출정보 크롤링 #
rq_kakao_2016 = requests.get(

"http://dart.fss.or.kr/report/viewer.do?rcpNo=20160816002028&dcmNo=5262283&eleId=10&offset=271961&length=146730&dtd=dart3.xsd")
rqctnt_kakao_2016 = rq_kakao_2016.content
soup = BeautifulSoup(rqctnt_kakao_2016, "html.parser")
kakao_2016 = str(soup.find_all("tr"))
kakao_2016 = re.sub('<.+?>', '', kakao_2016, 0).strip()
fd = open("kakao_2016.csv", "w", encoding='utf-8')
fd.write(kakao_2016)
print(kakao_2016)
```

2. 지식그래프 개체 및 관계 추출

(ETRI 개체명 분석 API 활용 : http://aiopen.etri.re.kr/guide_wiseNLU.php#group01) 개체명 태그셋은 15개 대분류 및 146개 세분류로 구성된 TTA 표준 개체명 태그셋(TTAK.KO-10.0852)과 의존 구문 분석 말뭉치 구축을 위한 의존 관계 태그 세트 및 의존 관계 설정 방법 정보통신단체표준(국문표준 - TTAK.KO-10.0853)을 적용)

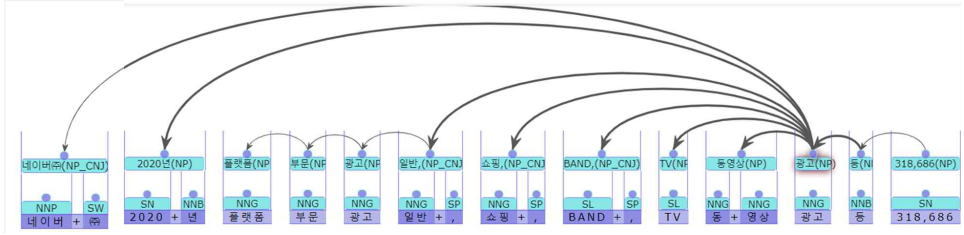
네이버(주) 2020년 반기보고서 개체명 분석 결과

형태소/개체명/동음이의어 결과 | 구문분석 결과 | 의미역 인식 결과

No	Word	형태소			동음이의어			개체명		
		단어	태그	설명	단어	형태소 태그	의미번호	단어	태그	설명
0	네이버㈜	네이버 ㈜	NNP SW	고유명사 기타 기호 (논리 수학기호, 기호 등)	네이버 ㈜	NNP SW	00 00	네이버㈜	OGG_ECONOMY	경제 관련 기관/단체, 기업
1	2020년	2020 년	SN NNB	숫자 의존명사	2020 년	SN NNB	00 02	2020년	DT_YEAR	년
2	플랫폼	플랫폼	NNG	일반명사	플랫폼	NNG	00			
3	부문	부문	NNG	일반명사	부문	NNG	06			
4	광고	광고	NNG	일반명사	광고	NNG	02			
5	일반,	일반 ,	NNG SP	일반명사 쉼표, 가운뎃점, 콜론, 빗금	일반 ,	NNG SP	02 00			
6	쇼핑,	쇼핑 ,	NNG SP	일반명사 쉼표, 가운뎃점, 콜론, 빗금	쇼핑 ,	NNG SP	00 00			
7	BAND,	BAND ,	SL SP	외국어 쉼표, 가운뎃점, 콜론, 빗금	BAND ,	SL SP	00 00	BAND	TR_SCIENCE	과학 관련 이론/법칙/방식/양식/체계/학설
8	TV	TV	SL	외국어	TV	SL	00	TV	TMI_HW	IT 하드웨어 용어
9	동영상	동 영상	NNG NNG	일반명사 일반명사	동영상	NNG	00			
10	광고	광고	NNG	일반명사	광고	NNG	02			
11	등	등	NNB	의존명사	등	NNB	05			
12	318,686	318,686	SN	숫자	318,686	SN	00	318,686	QT_COUNT	개수/빈도

네이버(주) 2020년 반기보고서 구문분석 결과

형태소/개체명/동음이의어 결과 | 구문분석 결과 | 의미역 인식 결과



네이버(주) 2016년 반기보고서 개체명 분석

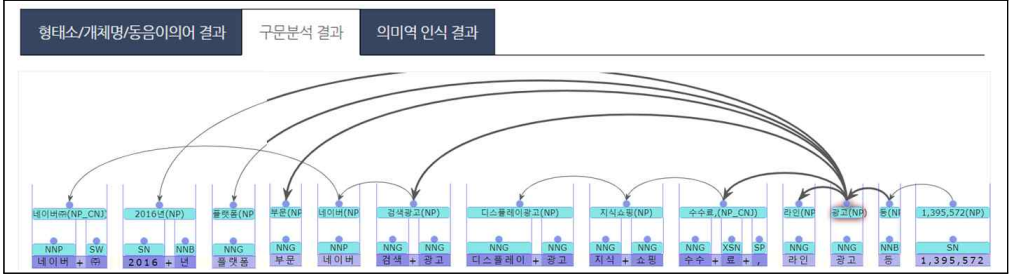
분석 문장 선택

- (●) 네이버㈜ 2016년 플랫폼 부문 네이버 검색광고 디스플레이광고 지식쇼핑 수수료, 라인 광고 등 1,395,572
- ( ) 네이버㈜ 2016년 콘텐츠 부문 라인 게임 스티커, 네이버 게임 뮤직 웹툰 등 473,131

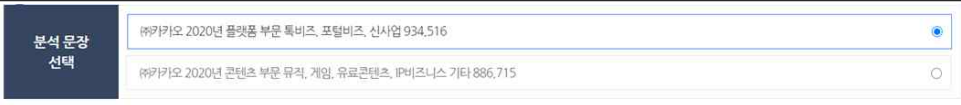
형태소/개체명/동음이의어 결과 | 구문분석 결과 | 의미역 인식 결과

No	Word	형태소			동음이의어			개체명		
		단어	태그	설명	단어	형태소 태그	의미번호	단어	태그	설명
0	네이버㈜	네이버 ㈜	NNP SW	고유명사 기타 기호 (논리 수학기호, 기호 등)	네이버 ㈜	NNP SW	00 00	네이버㈜	OGG_ECONOMY	경제 관련 기관/단체, 기업
1	2016년	2016 년	SN NNB	숫자 의존명사	2016 년	SN NNB	00 02	2016년	DT_YEAR	년
2	플랫폼	플랫폼	NNG	일반명사	플랫폼	NNG	00			
3	부문	부문	NNG	일반명사	부문	NNG	06			
4	네이버	네이버	NNP	고유명사	네이버	NNP	00	네이버	OGG_ECONOMY	경제 관련 기관/단체, 기업
5	검색광고	검색 광고	NNG NNG	일반명사 일반명사	검색 광고	NNG NNG	00 02	검색광고	TMI_SERVICE	IT 서비스 용어
6	디스플레이광고	디스플레이 광고	NNG NNG	일반명사 일반명사	디스플레이 광고	NNG NNG	00 02			
7	지식쇼핑	지식 쇼핑	NNG NNG	일반명사 일반명사	지식 쇼핑	NNG NNG	02 00			
8	수수료,	수수 료 ,	NNG XSN SP	일반명사 명사파생접미사 쉼표, 가운뎃점, 콜론, 빗금	수수료 ,	NNG SP	00 00			
9	라인	라인	NNG	일반명사	라인	NNG	01			
10	광고	광고	NNG	일반명사	광고	NNG	02			
11	등	등	NNB	의존명사	등	NNB	05			
12	1,395,572	1,395,572	SN	숫자	1,395,572	SN	00	1,395,572	QT_PRICE	금액

네이버(주) 2016년 반기보고서 구문분석 결과



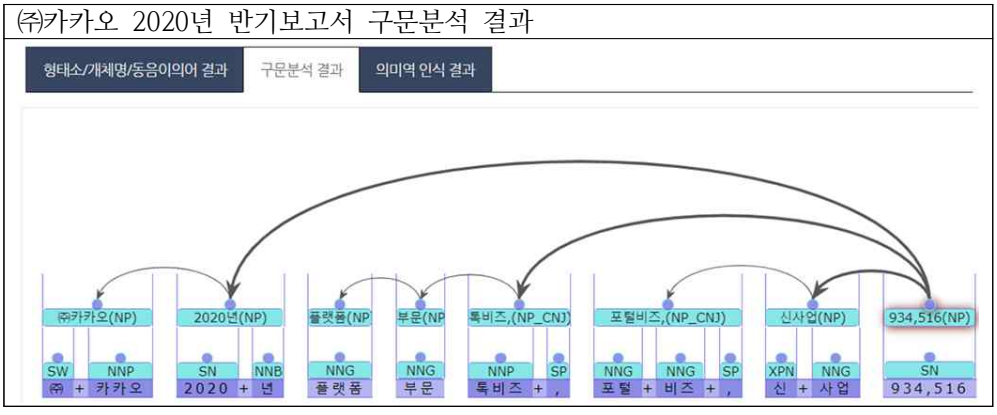
㈜카카오 2020년 반기보고서 개체명 분석



형태소/개체명/동음이의어 결과 | 구문분석 결과 | 의미역 인식 결과

No	Word	형태소			동음이의어			개체명		
		단어	태그	설명	단어	형태소 태그	의미번호	단어	태그	설명
0	㈜카카오	㈜ 카카오	SW NNP	기타 기호 (논리 수학기호, 기호 등) 고유명사	㈜ 카카오	SW NNP	00 00	㈜카카오	OGG_ECONOMY	경제 관련 기관/단체, 기업
1	2020년	2020 년	SN NNB	숫자 의존명사	2020 년	SN NNB	00 02	2020년	DT_YEAR	년
2	플랫폼	플랫폼	NNG	일반명사	플랫폼	NNG	00			
3	부문	부문	NNG	일반명사	부문	NNG	06			
4	톡비즈,	톡비즈 ,	NNP SP	고유명사 쉼표, 가운뎃점, 콜론, 빗금	톡비즈 ,	NNP SP	00 00			
5	포털비즈,	포털 비즈 ,	NNG NNG SP	일반명사 일반명사 쉼표, 가운뎃점, 콜론, 빗금	포털 비즈 ,	NNG NNG SP	00 00 00			
6	신사업	신 사업	XPN NNG	체언접두사 일반명사	신사업	NNG	00			
7	934,516	934,516	SN	숫자	934,516	SN	00	934,516	QT_OTHERS	기타 수량 표현

㈜카카오 2020년 반기보고서 구문분석 결과



㈜카카오 2016년 반기보고서 개체명 분석 결과

분석 문장 선택

㈜카카오 2016년 플랫폼 부문 검색광고, 디스플레이광고 265,589

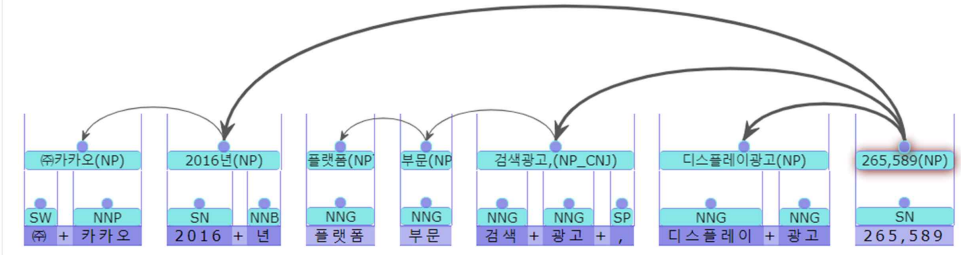
㈜카카오 2016년 콘텐츠 부문 게임, 음악 281,993

형태소/개체명/동음이의어 결과 | 구문분석 결과 | 의미역 인식 결과

No	Word	형태소			동음이의어			개체명		
		단어	태그	설명	단어	형태소 태그	의미번호	단어	태그	설명
0	㈜카카오	㈜ 카카오	SW NNP	기타 기호 (논리 수학기호, 기호 등) 고유명사	㈜ 카카오	SW NNP	00 00	㈜카카오	OGG_ECONOMY	경제 관련 기관/단체, 기업
1	2016년	2016 년	SN NNB	숫자 의존명사	2016 년	SN NNB	00 02	2016년	DT_YEAR	년
2	플랫폼	플랫폼	NNG	일반명사	플랫폼	NNG	00			
3	부문	부문	NNG	일반명사	부문	NNG	06			
4	검색광고,	검색 광고 ,	NNG NNG SP	일반명사 일반명사 쉼표, 가운뎃점, 콜론, 빗금	검색 광고 ,	NNG NNG SP	00 02 00	검색광고	TMI_SERVICE	IT 서비스 용어
5	디스플레이광고	디스플레이 광고	NNG NNG	일반명사 일반명사	디스플레이 광고	NNG NNG	00 02			
6	265,589	265,589	SN	숫자	265,589	SN	00	265,589	QT_OTHERS	기타 수량 표현

㈜카카오 2016년 반기보고서 구문분석 결과

형태소/개체명/동음이의어 결과 | 구문분석 결과 | 의미역 인식 결과



3. Word embedding-FastText 단어 유사도 분석 소스 코드

```
from gensim.models.word2vec import FastText
import gensim

path = 'my_vocab_colab'
sentences = gensim.models.word2vec.Text8Corpus(path)

model = FastText( sentences, min_count=5, size=100, window=5)
model.save('ft_model')

saved_model = Word2Vec.load('ft_model')

word_vector = saved_model['네이버(주)']

saved_model.similarity('네이버(주)', '㈜카카오')

saved_model.similarity_by_word('네이버(주)')
```

4. 개체명 태그 세트 및 태깅 말뭉치(TTAK.KO-10.0852)

태그 세트 분류 체계

(개체명 태그셋은 15개 대분류 및 146개 세분류로 구성된 TTA 표준 개체명 태그셋 (TTAK.KO-10.0852))

대분류	세분류	정의
PERSON (PS)	PS_NAME	- 사람 이름
LOCATION (LC)	LC_OTHERS	- LC계열의 세부 유형이 아닌 기타 장소
	LCP_COUNTRY	- 국가명
	LCP_PROVINCE	- 도, 주 지역명
	LCP_COUNTY	- 군, 면, 읍, 리, 동 등과 같은 세부 행정구역명
	LCP_CITY	- 도시명
	LCP_CAPITALCITY	- 수도명

대분류	세분류	정의
	LCG_RIVER	- 강, 호수, 연못
	LCG_OCEAN	- 해양, 바다
	LCG_BAY	- 반도, 만
	LCG_MOUNTAIN	- 산, 산맥, 능선, 고개/재, 봉
	LCG_ISLAND	- 섬, 제도
	LCG_CONTINENT	- 대륙
	LC_TOUR	- 관광명소
	LC_SPACE	- 천체 명칭
ORGANIZATION (OG)	OG_OTHERS	- 기타 기관/단체
	OGG_ECONOMY	- 경제 관련 기관/단체, 기업
	OGG_EDUCATION	- 교육 기관/단체, 교육관련 기관
	OGG_MILITARY	- 군사 기관/단체 및 유형, 국방기관
	OGG_MEDIA	- 미디어 기관/단체, 방송관련 기관/기업
	OGG_SPORTS	- 스포츠 기관/단체
	OGG_ART	- 예술 기관/단체
	OGG_MEDICINE	- 의학/의료 기관/단체
	OGG_RELIGION	- 종교 기관/단체, 종파 포함
	OGG_SCIENCE	- 과학 기관/단체
	OGG_LIBRARY	- 도서관 및 도서관 관련 기관/단체
	OGG_LAW	- 법률 기관/단체
	OGG_POLITICS	- 정부/행정 기관, 공공기관, 정치기관
	OGG_FOOD	- 음식 관련 업체/회사
	OGG_HOTEL	- 숙박 관련 업체
ARTIFACTS (AF)	AF_CULTURAL_ASSET	- 건물/서적/작품 등 문화재
	AF_BUILDING	- 건축물/토목건설물, 운동장이름, 아파트, 다리, 등대, 분수
	AF_MUSICAL_INSTRUMENT	- 악기 명칭
	AF_ROAD	- 도로/철로 명칭
	AF_WEAPON	- 무기 명칭
	AF_TRANSPORT	- 교통수단/자동차/선박 모델 및 유형, 운송수단, 놀이기구
	AF_WORKS	- AFW의 세부 작품명에 해당하지 않는 기타 작품명
	AFW_DOCUMENT	- 도서/서적 작품명 - 지리서, 지도 - 의학 관련 서적 - 종교 서적, 각 종교의 경전 - 철학 관련 서적 - 시/소설/희곡 등의 문학 작품명

대분류	세분류	정의
		- 역사 서적 - 기타 서적
	AFW_PERFORMANCE	- 춤/무용 작품명 및 춤 종류 - 연극/공연명/가극
	AFW_VIDEO	- 영화 작품명 - TV 프로그램 이름
	AFW_ART_CRAFT	- 미술 작품명(그림/미술품)
	AFW_MUSIC	- 음악 작품명
	AF_WARES	- 상품/제품 이름
DATE (DT)	DT_OTHERS	- DT 계열의 세부 유형이 아닌 기타 날짜
	DT_DURATION	- 기간
	DT_DAY	- 날짜/절기
	DT_MONTH	- 달
	DT_YEAR	- 년
	DT_SEASON	- 계절
	DT_GEOAGE	- 지질시대
	DT_DYNASTY	- 왕조시대
TIME (TI)	TI_OTHERS	- 기타 시간
	TI_DURATION	- 기간
	TI_HOUR	- 시각
	TI_MINUTE	- 분
	TI_SECOND	- 초
CIVILIZATION (CV)	CV_NAME	- 문명/문화 명칭
	CV_TRIBE	- 민족/종족 명칭, 국가를 구성하는 국민
	CV_SPORTS	- 스포츠/레포츠/레저 명칭
	CV_SPORTS_INST	- 스포츠 용품/도구
	CV_POLICY	- 제도/정책 명칭
	CV_TAX	- 조세 명칭
	CV_FUNDS	- 연금, 기금, 자금, 펀드 명칭
	CV_LANGUAGE	- 언어 명칭
	CV_BUILDING_TYPE	- 건축양식 명칭
	CV_FOOD	- 음식 명칭, 음식재료
	CV_DRINK	- 음료수, 술 명칭
	CV_CLOTHING	- 의복/섬유 명칭
	CV_POSITION	- 직위/직책 명칭, 스포츠 포지션
	CV_RELATION	- 인간 관계 명칭
	CV_OCCUPATION	- 직업 명칭
	CV_CURRENCY	- 통화 명칭
	CV_PRIZE	- 상과 훈장

대분류	세분류	정의
	CV_LAW	- 법/법률 명칭
	CV_FOOD_STYLE	- 음식 종류
ANIMAL (AM)	AM_OTHERS	- 기타 동물 명칭
	AM_INSECT	- 곤충
	AM_BIRD	- 조류
	AM_FISH	- 어류
	AM_MAMMALIA	- 포유류
	AM_AMPHIBIA	- 양서류
	AM_REPTILIA	- 파충류
	AM_TYPE	- 동물 분류 명칭
	AM_PART	- 동물 몸의 한 부분(신체 부위) 명칭
PLANT (PT)	PT_OTHERS	- 기타 식물 명칭
	PT_FRUIT	- 과일 이름 (식용 가능한 식물의 열매)
	PT_FLOWER	- 꽃 이름
	PT_TREE	- 나무 이름
	PT_GRASS	- 풀 이름
	PT_TYPE	- 식물 유형 명칭
	PT_PART	- 식물의 한 부분에 대한 명칭
QUANTITY (QT)	QT_OTHERS	- 기타 수량 표현
	QT_AGE	- 나이
	QT_SIZE	- 크기/넓이
	QT_LENGTH	- 길이/거리/높이
	QT_COUNT	- 개수/빈도
	QT_MAN_COUNT	- 인원수
	QT_WEIGHT	- 무게
	QT_PERCENTAGE	- 백분율, 비율, 농도
	QT_SPEED	- 속도
	QT_TEMPERATURE	- 온도
	QT_VOLUME	- 부피
	QT_ORDER	- 순서, 순차적 표현
	QT_PRICE	- 금액
	QT_PHONE	- 전화번호
	QT_SPORTS	- 스포츠 관련 수량 표현
	QT_CHANNEL	- TV/라디오 채널 번호
	QT_ALBUM	- 앨범 관련 수량 표현
	QT_ZIPCODE	- 우편번호
STUDY_FIELD (FD)	FD_OTHERS	- 학문 분야 및 학파
	FD_SCIENCE	- 과학 학문 분야
	FD_SOCIAL_SCIENCE	- 사회과학 학문 분야 및 학파

대분류	세분류	정의
		- 정치/경제/사회와 관련된 분야
	FD_MEDICINE	- 의학 관련 학문 분야 및 분과
	FD_ART	- 예술관련 학문 분야 및 학파
	FD_PHILOSOPHY	- 철학 관련 학문 분야 및 학파
THEORY (TR)	TR_OTHERS	- 기타 이론/법칙/원리
	TR_SCIENCE	- 과학 관련 이론/법칙/방식/양식/체계/학설
	TR_SOCIAL_SCIENCE	- 사회과학 이론/법칙/방법/원리/사상, 정치사상
	TR_ART	- 예술관련 이론/법칙/방식/양식, 예술사조
	TR_PHILOSOPHY	- 철학 이론/사상
	TR_MEDICINE	- 의학 요법/처방, 의학 진단법
EVENT (EV)	EV_OTHERS	- 기타 사건/사고 명칭, ~사태
	EV_ACTIVITY	- 사회운동 및 선언
	EV_WAR_REVOLUTION	- 전쟁/혁명
	EV_SPORTS	- 스포츠/레저 관련 행사
	EV_FESTIVAL	- 축제 명칭
MATERIAL (MT)	MT_ELEMENT	- 원소명
	MT_METAL	- 금속물
	MT_ROCK	- 암석 명칭
	MT_CHEMICAL	- 모든 화학물질을 나타냄
TERM (TM)	TM_COLOR	- 색
	TM_DIRECTION	- 방향
	TM_CLIMATE	- 기후지역 명칭
	TM_SHAPE	- 모양/형태 명칭
	TM_CELL_TISSUE	- 세포/조직/기관 명칭 (외부에서 안 보이는 조직/기관)
	TMM_DISEASE	- 증상/증세/질병
	TMM_DRUG	- 약/약품명
	TMI_HW	- IT 하드웨어 용어
	TMI_SW	- IT 소프트웨어 용어
	TMI_SITE	- URL 주소
	TMI_EMAIL	- 이메일주소
	TMI_MODEL	- 각종 제품의 세부 모델명, 부품류
	TMI_SERVICE	- IT 서비스 용어
	TMI_PROJECT	- 프로젝트 명칭
	TMIG_GENRE	- 게임 장르
	TM_SPORTS	- 스포츠/레저 용어 (기술/규칙 명칭)

5. 태그 세트 분류 체계

(의존 구문 분석 말뭉치 구축을 위한 의존 관계 태그 세트 및 의존 관계 설정 방법 정보통신 단체표준(국문표준 - TTAK.KO-10.0853))

○ 구문 태그 세트

구문 태그	의미
NP	체언 (명사, 대명사, 수사)
VP	용언 (동사, 형용사, 보조용언)
AP	부사구
VNP	긍정 지정사구 (명사+이다)
DP	관형사구
IP	감탄사구 (호칭 및 대답 등의 표현)
X	의사 구 (pseudo phrase, 조사 단독 어절 또는 기호 등)
L	부호 (왼쪽 괄호 및 따옴표)
R	부호 (오른쪽 괄호 및 따옴표)

○ 기능 태그 세트

기능 태그	의미
SBJ	주어
OBJ	목적어
MOD	관형어 (체언 수식어)
AJT	부사어 (용언 수식어)
CMP	보어
CNJ	접속어 (~와)

ABSTRACT

BCG matrix visualization through BERT-based knowledge graph generation from unstructured text

Park, Byoung-Cheol

Major in Smart Convergence Technology Consulting

Dept. of Smart Convergence Consulting

Graduate School of Knowledge Service & Consulting

Hansung University

Recently, natural language processing models are making rapid improvements based on deep learning, and these language models show capabilities close to human intelligence in various fields. However, there is a limit of processing consistency for long sentences, and the physical world at the level of human commonsense is not well understood. And there is a lack of many contexts for multi-domains in which interactions occur. If you learn and use sentence-level embedding data suitable for a domain, you can efficiently learn a language model for generating desired reports such as charts, graphs, and pictures from large-capacity text data based on business consulting. This paper creates embedding data for training of a deep learning natural language processing model based on BERT by collecting unstructured company information as text from an electronic public announcement site of a domestic stock market. The

knowledge graph was utilized as a tool to improve problem-solving ability for objects and relationships between multiple domains, and a study was conducted on visualization that generates business consulting reports. In addition, [REP], which has the characteristic of the relationship value between [CON] Token in the nature of consulting category, which is a special token, to improve the accuracy of embedding data, which plays an important role in learning of natural language models based on deep learning. Based on the data, BPE (Byte Pair Encoding) that adds Special Token, that is, [CON], which is a consulting category classification token, and [REP] token, for generating the corresponding report, using BERT pre-learning model It generates embedding data through. In addition, a vocabulary dictionary (Vocab) is created by performing BERT-based preliminary learning for creating a consulting domain domain and corresponding logic. Based on the expanded embedding data, entities and relationships are created using the API of ETRI Exobrain. Through the entity name analysis and dependency syntax analysis of the knowledge graph, a meaningful classification system is implemented through emphasis on entities and relationships, [CON[and [REP]], which are elements of the BCG Matrix. Based on this, an element that can create a BCG Matrix chart is extracted from the embedded data that is mapped to the entity and relationship of the added token and the knowledge graph and visualized by applying business logic. As a limitation of this paper, the process that requires deep learning-based learning was conducted in the form of a toy example because of limitations in resources and functions. It is implemented in the form of a toy example as a method to verify the process of automating the generation of business consulting reports from unstructured text. Through this, it was found that the study of creating a BCG matrix in the form of a toy example is also meaningful. In

addition, sentence/word similarity was used as a method to verify the quality of embedded data, which plays an important role in learning a natural language processing model based on deep learning. Although it is necessary to supplement due to the lack of unstructured data and constraints on the learning time, if an integrated research environment in the business consulting area is configured in the future, a sufficiently learned model can be created through an automated process.

【Keywords】 BERT, Text to Knowledge Graph, Word embedding, BCG Matrix