공학 석사 학위논문

시맨틱 텐서공간모델과 어텐션 메커니즘을 활용한 자동 문서분류

Automated Text Classification using Semantic Tensor Space Model and Attention Mechanism

2021년 2월

서울시립대학교 대학원

전자전기컴퓨터공학과

이 길 재

시맨틱 텐서공간모델과 어텐션 메커니즘을 활용한 자동 문서분류

Automated Text Classification using Semantic Tensor Space Model and Attention Mechanism

지도교수 김 한 준

이 논문을 공학 석사학위 논문으로 제출함

2020년 12월

서울시립대학교 대학원

전자전기컴퓨터공학과

이 길 재

이 길 재의 공학 석사학위 논문을 인준함.

심사위원장 김 용 철 인

심 사 위 원 김 한 준 인

심 사 위 원 이 영 민 인

2020년 12월

서울시립대학교 대학원

국문초록

문서분류는 자연어처리 분야의 대표적인 연구 중 하나로, 텍스트 임베딩과의 연관이 깊다. 이는 텍스트 임베딩에 사용한 모델이 단어나 문서를 구체적으로 표현하는 정도에 따라 문서분류 성능이 크게 달라지기 때문이다. 최근, 대용량의 텍스트 데이터로 사전학습 된 텍스트 임베딩 모델을 문서분류에 활용하여 분류 성능을 향상시키는 연구들이 수행되고 있다. 하지만, 대부분의 최근 임베딩 모델들은 단어의 의미를 예측할 때 주위 문맥만을 활용한다는 분명한 한계점을 갖는다. 이에 반해, 인간 지식기반을 활용한 텍스트 임베딩 모델들은 단어를 훨씬 구체적으로 표현할 수 있다는 연구 사례들이 존재한다. Yamada et al.은 위키피디아 문서집합으로 학습시킨 텍스트 임베딩 모델, TextEnt를 제안하고, 이를 활용한 문서분류 실험에서 'state-of-the-art'의 성능을 보인 바 있다. 이와 비슷하게, 본 논문 역시 인간 지식기반을 활용하여 문서를 단어와 컨셉으로 표현하는 텍스트 임베딩 모델, 텍스트 큐보이드(Text Cuboid)를 제안한다. 또한, 텍스트 큐보이드에 어텐션 네트워크를 차용한 텍스트 큐보이드-어텐션(Text Cuboid-Attention) 문서분류 모델을 제안하고, 제안 모델이 20Newsgroups와 R8 데이터 셋을 사용한 문서분류 실험에서 TextEnt의 결과를 넘어서는 점을 보인다.

주요어: 텍스트 마이닝, 문서분류, 텍스트 임베딩, 머신 러닝, 딥 러닝, 어텐션 메커니즘.

공학 석사 학위논문

시맨틱 텐서공간모델과 어텐션 메커니즘을 활용한 자동 문서분류

Automated Text Classification using Semantic Tensor Space Model and Attention Mechanism

2021년 2월

서울시립대학교 대학원

전자전기컴퓨터공학과

이 길 재

목차

그림 차례

표 차례

제1장 서론

제1절 연구 배경 및 내용

문서분류(Text Classification)는 자연어처리(Natural Language Processing) 분야의 대표적인 연구 중 하나로, 텍스트 임베딩(Text Embedding)과의 연관이 깊다. 이는 텍스트 임베딩에 사용한 모델이 단어나 문서를 구체적으로 표현하는 정도에 따라 문서분류 성능이 크게 달라지기 때문이다. 최근, 대용량의 텍스트 데이터로 사전학습(Pre-Trained)된 텍스트 임베딩 모델을 문서분류에 활용하여 분류 성능을 향상시키는 연구들이 수행되고 있다. Skip-Gram 방식으로 단어 표현(Word Representations)을 학습하는 Word2Vec을 필두로 [1], Bi-Directional한 언어모델(Language Model)을 활용하여 단어의 문맥적인 의미를 표현하는 ELMo [2], 트랜스포머(Transformer) 블록을 쌓아올린 구조의 GPT, BERT [3], [4]까지 Unlabeled된 대용량의 데이터로 사전 학습을 수행하는 모델들을 문서분류에 활용하는 연구 사례들이 증가하고 있다.

하지만, 대부분의 텍스트 임베딩 모델들은 단어를 표현하는 데 분명한 한계점을 갖는다. 이는 모델이 단어의 의미를 예측할 때 주위 문맥만을 활용한다는 점인데, 문맥만을 통해서는 사람의 지식수준으로 단어의 의미를 표현할 수 없다. 또한, 대용량의 데이터로 사전학습을 수행하는 데 GPU, 메모리 등 많은 컴퓨팅 자원들(Computing Resources)이 요구되는 문제점도 갖는다. 이에 반해, 인간 지식기반(Human Knowledge Base)을 활용한 텍스트 임베딩 모델들은 더 적은 자원으로도 단어를 구체적으로

표현할 수 있다는 연구 사례들이 존재한다. Yamada et al.은 위키피디아(Wikipedia) 문서집합으로 학습시킨 텍스트 임베딩 모델, TextEnt를 제안하고, 이를 활용하여 20Newsgroups 및 R8 데이터 셋을 사용한 문서분류 실험에서 'state-of-the-art'의 성능을 보인 바 있다 [5]. 또한, TextEnt 모델에 어텐션 메커니즘(Attention Mechanism)을 적용시킨 문서분류 모델, NABoE를 제안하여 기존의 분류 성능을 능가하였다 [6].

이와 비슷하게, 본 논문 역시 인간 지식기반을 활용하여 문서를 단어(Term)와 컨셉(Concept)으로 표현하는 텍스트 임베딩 모델, 텍스트 큐보이드(Text Cuboid)를 제안한다 [7]. 또한, NABoE와 비슷하게 텍스트 큐보이드에 어텐션 네트워크(Attention Network)를 차용한 텍스트 큐보이드-어텐션(Text Cuboid-Attention) 문서분류 모델을 제안하고, 제안 모델이 20Newsgroups와 R8 데이터 셋을 사용한 문서분류 실험에서 TextEnt와 NABoE의 결과를 넘어서는 점을 보임으로써 그 성능을 증명한다.

제2절 논문 구성

본 논문의 구성은 다음과 같다.

제2장에서는 본 논문에서 수행한 연구 내용과 비슷하거나, 선행 연구가 되는 사례들을 소개한다.

제3장에서는 본 논문에서 제안하는 인간 지식기반의(Human Knowledge-Based) 텍스트 임베딩 모델, 텍스트 큐보이드(Text Cuboid)에 대해 설명한다.

제4장에서는 텍스트 큐보이드에 어텐션 네트워크(Attention Network)를 차용한 텍스트 큐보이드(Text Cuboid-Attention) 문서분류 모델을 제안한다.

제5장에서는 20Newsgroups 및 R8 데이터 셋을 활용한 문서분류 실험을 통해 제안 모델의 성능을 평가한다.

제6장에서는 본 논문의 전체적인 내용을 정리하고, 수행 연구의 의의에 대해 토론하며 논문을 끝마친다.

제2장 관련 연구

제1절 TextEnt

Document	Contextual Entities	Target Entity
Saturn is the sixth planet from the Sun and the second-largest in the Solar System, after Jupiter. It is a gas giant with an average radius about nine times that of Earth.	Planet, Sun, Solar System, Jupiter, Gas giant, Earth	Saturn

표 2-1. TextEnt에서 Entity 추출 예시

Yamada et al.은 인간 지식기반(Human Knowledge Base)을 활용한 텍스트 임베딩 모델, TextEnt를 제안하고, 이를 활용하여 20Newsgroups 및 R8 데이터 셋을 사용한 문서분류 실험에서 'state-of-the-art'의 성능을 보였다. Yamada et al.은 위키피디아(Wikipedia) 문서집합을 활용하여 각 문서에서 중요한 키워드가 되는 문맥 개체들(Contextual Entities)을 추출하고, 추출한 문맥 개체들과 문서에 출현하는 단어들을 바탕으로 문서의 제목이 되는 타겟 개체(Target Entity)를 예측하도록 TextEnt를 학습시켰다. **표 2-1**에서 Yamada et al.이 제시한 개체 추출의 예시를 확인할 수 있다. TextEnt는 Word2Vec과 비슷하게 학습을 통해 각 개체 표현(Entity Representations)을 배우고, 최종 학습된 개체 표현과 출현 단어들을 활용하여 문서를 표현한다.

제2절 어텐션 메커니즘

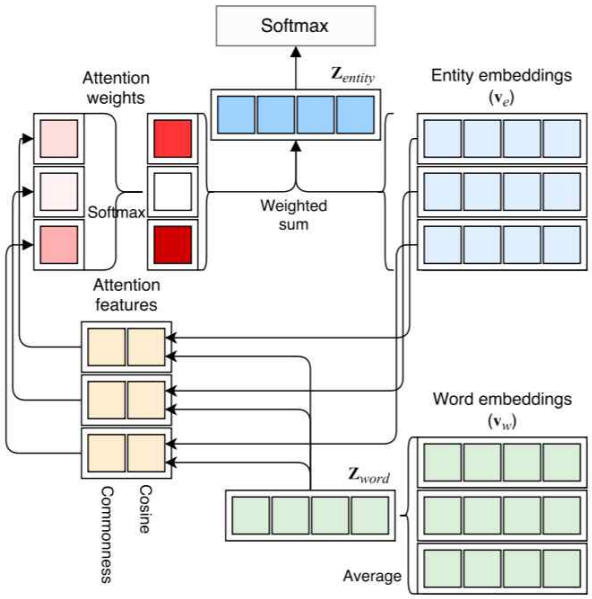

그림 2-1. NABoE에서의 어텐션 메커니즘

어텐션 메커니즘(Attention Mechanism)은 최근 자연어처리 모델들의 기반이 되는 알고리즘으로 신경망 기계번역(Neural Machine Translation) 분야를 중심으로 다양한 연구에 활용되고 있다. Bahdanau et al.은 순환 신경망 기반(RNN-based) 인코더-디코더(Encoder-Decoder) 구조를 갖는 신경망 기계번역 모델에 어텐션 메커니즘을 도입하여, 번역 이전과 이후 문장 간 상관성이 높은 부분에 가중치를 부여함으로써 번역 성능을 높인 바 있다 [8]. Vaswani et al.은 어텐션 메커니즘의 변형 알고리즘인 Self-Attetion을 바탕으로 하는 트랜스포머(Transformer) 모델을 제안하였고 [9], 트랜스포머 블록을 쌓아올린 구조를 갖는 GPT, BERT같은 모델들은 다양한 자연어처리 분야들에서 좋은 성능을 보이고 있다.

Yamada et al.은 TextEnt 모델에 어텐션 메커니즘을 도입한 NABoE 문서분류 모델을 제안하여 기존의 분류 성능을 넘어선 바 있다(**그림 2-1** 참조). NABoE는 어텐션 메커니즘을 통해 문서를 표현하는 개체들과 출현 단어들 간의 유사도를 계산하고, 이를 바탕으로 개체들에 가중치를 부여하여 기존 TextEnt에서의 개체 표현을 개선시켰다.

제3장 Text Cuboid

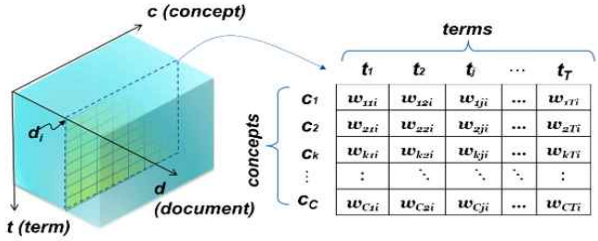

그림 3-1. Text Cuboid의 구조

텍스트 큐보이드(Text Cuboid)는 문서를 단어(Term)와 컨셉(Concept) 축으로 구성된 2차원의 행렬로 표현하는 인간 지식기반의(Human Knowledge-Based) 텍스트 임베딩 모델이다(**그림 3-1** 참조). 컨셉이란 인간이 단어를 생각할 때 떠올리는 추상적인 개념들을 표현하기 위해 새롭게 정의한 특성(Feature)으로, 사전에 위키피디아 페이지로 색인하여(Indexing) 놓은 검색엔진을 활용하여 얻을 수 있다. 텍스트 큐보이드는 문서집합에서 자주 출현하는 단어집합을 추출하고, 이를 사전에 색인된 엘라스틱서치(Elasticsearch) 검색엔진에 질의하여 컨셉집합을 얻는다(**그림 3-2** 참조). 예를 들어, 문서집합에 '사과', '오렌지', '포도'와 같은 단어들이 자주 출현하여 이를 엘라스틱서치에 질의하면 '과일', '주스' 등의 위키피디아 페이지를 얻을 수 있는데, 이 경우 '과일'과 '주스'가 컨셉집합에 포함된다. 위와 같은 과정을 통해 얻은 단어 및 컨셉들은 각각 단어와 컨셉 축을 구성하여 문서를 표현한다. 문서집합 전체를 임베딩 했을 때 생성되는 3차원의 텐서공간으로 인해 제안 모델을 '텍스트 큐보이드'라고 명칭하게 되었다.

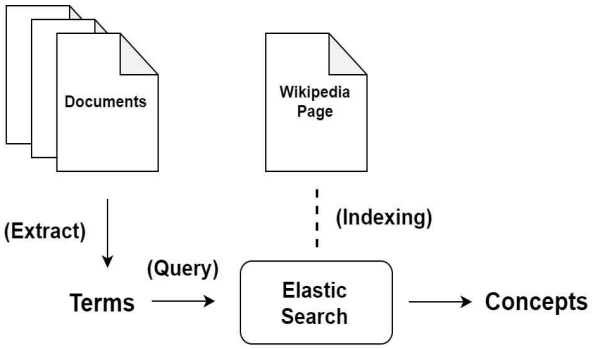

그림 3-2. 위키피디아 페이지를 활용한 컨셉 추출 과정

Word2Vec, ELMo 그리고 TextEnt와 같은 텍스트 임베딩 모델들은 단어를 각 차원이 특별한 의미를 갖지 않는 n차원의 벡터로 표현한다. 이에 반해, 텍스트 큐보이드는 단어를 각 차원이 컨셉이란 특정한 의미를 갖는 시맨틱 텐서공간(Semantic Tensor Space)에 매핑(Mapping)한다. 위와 같은 텍스트 큐보이드의 특성은 단어 표현을 더 의미 있게 만든다. 그리고 이러한 장점을 활용하기 위하여 다음 장에서 중요한 컨셉차원을 강조하는 문서분류 모델을 제안한다.

제4장 Text Cuboid-Attention

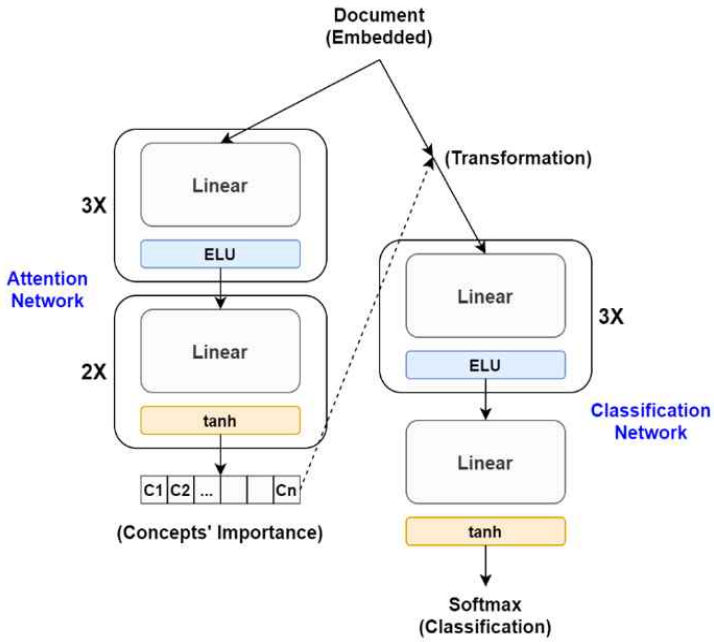

그림 4-1. Text Cuboid-Attention의 구조

본 논문에서 제안하는 텍스트 큐보이드-어텐션(Text Cuboid-Attention)은 어텐션 네트워크(Attention Network)을 차용한 심층신경망 기반의(DNN-based) 문서분류 모델이다. **그림 4-1**에서 보는 바와 같이 제안 모델은 어텐션 네트워크와 분류 네트워크(Classification Network)로 구성되어 있으며, 모델의 입력 문서들은 모두 텍스트 큐보이드를 통해 임베딩한다. 제안 모델의 핵심 원리는 입력 문서에서 중요한 의미를 갖는 컨셉을 찾아내어 해당 차원에 추가 가중치를 부여하는 점이다. 어텐션 메커니즘의 역할을 수행하는 어텐션 네트워크는 컨셉들의 중요도를 학습하고, 분류 네트워크는 문서들의 실질적인 분류를 수행한다.

$$W_{weighted} = \sum_{i=1}^{n(concepts)} \sum_{j=1}^{n(terms)} W_{i,j} * (1 + C_i - 0.3), (-1 < C_i < 1)$$

수식 1. Attention Network를 통해 컨셉차원에 부여되는 가중치

어텐션 네트워크는 입력 문서를 구성하는 컨셉들의 중요도를 학습하고, 중요도에 따라 해당 컨셉차원에 가중치를 부여한다. 컨셉 중요도(Concept Importance)는 분류 과정에서 컨셉이 미치는 영향력을 나타내는 수치이다. 어텐션 네트워크는 다중 선형 레이어(Linear Layer 혹은 Fully-Connected Layer) 구조를 가지며, 어텐션 네트워크의 최종 출력 텐서는 입력 문서의 컨셉 수와 동일한 차원을 갖는다. 출력 텐서의 n번째 차원 값은 입력 문서의 n번째 컨셉의 중요도를 나타내며, 마지막 선형 레이어의 활성화 함수인 tanh함수로 인해 -1~1 사이의 값을 갖는다. 출력 텐서를 통해 컨셉차원에 부여되는 가중치는 **수식 1**과 같다. **수식 1**에서 W는 입력 문서의 가중치, C는 출력 텐서를 일컫는다. **수식 1**을 살펴보면, 중요한 의미를 갖는 컨셉의 경우 C의 값이 1에 가까운 값을 가짐으로써 기존 가중치의 2배에 수렴한 값으로 조정된다. 반면, 중요하지 않은 컨셉의 경우 C의 값이 -1에 가까워짐으로써 0혹은 음의 값으로 가중치가 조정된다. **수식 1**에서 괄호 안에 0.3의 값을 빼준 이유는 어텐션 네트워크의 앞선 3개의 활성화 함수인 ELU함수가 양의 방향으로 치우쳐 있기 때문에, 입력 문서의 가중치가 전체적으로 커질 수 있기 때문이다. 위 값의 크기는 경험적으로 결정되며, 실험에 사용하는 데이터에 따라 달라진다.

분류 네트워크는 어텐션 네트워크로부터 조정된 가중치를 갖는 입력 문서들을 실질적으로 분류한다. 분류 네트워크 역시 다중 선형 레이어 구조

를 가지며, 빠른 학습 수렴을 위해 활성화 함수로 ELU함수를 사용한다. 마지막 활성화 함수로 사용되는 tanh함수는 Softmax 레이어 이전에 입력 문서들의 가중치를 정규화(Normalize)한다. 분류 네트워크를 통한 분류 결과는 교차 엔트로피 손실 함수(Cross-Entropy Loss)로 계산되어, 어텐션 및 분류 네트워크를 학습시키는 데 사용된다. 어텐션과 분류 네트워크는 동시에 같은 손실 함수로 학습된다. 어텐션 네트워크는 중요한 컨셉차원에 더 큰 가중치를 줄 수 있도록 학습되고, 분류 네트워크는 실제 분류 성능이 향상될 수 있도록 학습된다.

본 논문에서 제안한 어텐션 네트워크는 기존의 어텐션 메커니즘 혹은 NABoE에서 사용된 어텐션 메커니즘과 같이 단순히 단어와 단어 혹은 단어와 개체 간의 유사도를 계산하는 것이 아닌, 실제 분류 과정으로부터 컨셉의 중요도를 구하기 때문에 분류 작업에 직접적으로 도움이 되는 방향으로 가중치를 부여하게 된다. 개념적으로 생각하면, 어텐션 네트워크가 기존의 어텐션 메커니즘을 완벽히 따른다고 보기는 힘들지만, 분류 작업에 있어서 더 특화된 알고리즘이라고 생각할 수 있다.

제5장 실험 및 결과

제1절 실험 데이터 및 방법

본 논문에서 제안한 텍스트 큐보이드-어텐션의 성능을 평가하기 위해 20Newsgroups 및 R8 데이터 셋에 대한 다중 클래스 문서분류 (Multi-Class Text Classification) 실험을 수행하였다. 'by date' 버전의 20Newsgroups 데이터 셋은 20개의 클래스에서 11,314개의 학습 데이터와 7,532개의 테스트 데이터를 포함한다. Reuters-21578의 하위 집합인 R8 데이터 셋은 5,485개의 학습 데이터와 2,189개의 테스트 데이터를 포함한다. R8 데이터 셋의 경우 8개의 클래스로 구성되어 있으며, 20Newsgroups 데이터 셋과 달리 클래스 별 데이터의 수가 불균형하다.

본 논문에서는 텍스트 큐보이드-어텐션의 객관적 평가를 위해 4개의 비교 모델들을 선정하여 실험을 수행했다. BoW-SVM 모델은 과거의 텍스트 임베딩인 Bag-of-Words을 활용하여 문서를 표현하고, Support Vector Machine을 통해 분류를 수행하는 모델이다. TextEnt-full, NABoE-full 모델은 Yamada et al.이 제안한 인간 지식기반을 활용한 모델들로, 제안 모델과 유사한 특징을 갖는다. NABoE-full 모델은 어텐션 메커니즘을 차용했다는 점에서 제안 모델과 공통점을 갖는다. Text Cuboid-DNN은 제안 모델에서 분류 네트워크만을 포함하는 모델로, 어텐션 네트워크의 효과를 증명하기 위해 선정되었다.

제2절 실험 결과

		BoW-SVM	TextEnt-full	NABoE-full	TC-DNN	TC-Attention
20News groups	Accuracy	.790	.845	.868	.866	**.871**
	F1-Score	.783	.839	.862	.872	**.875**
R8	Accuracy	.947	.967	.971	.983	**.986**
	F1-Score	.851	.910	.917	.958	**.966**

표 5-1. 실험 결과(TC는 Text Cuboid의 준말)

표 5-1은 20Newsgroups 및 R8 데이터 셋에 대한 텍스트 큐보이드-어텐션(Text Cuboid-Attention 모델)의 분류 결과를 보여준다. 제안 모델은 두 데이터 셋에서 모두, 비교 모델들에 비해 높은 분류 정확도와 F1-Score를 보였다. 20Newsgroups 데이터 셋을 사용한 실험에서 Text Cuboid-DNN 모델은 86.8%를 기록한 NABoE-full 모델에 비해 약간 낮은 86.6%의 분류 정확도를 보였지만, 어텐션 네트워크를 적용함으로써 NABoE-full의 결과를 넘어선 87.1%로 분류 정확도를 향상시킬 수 있었다. R8 데이터 셋을 사용한 실험에서는 텍스트 큐보이드만을 활용하여 NABoE-full 모델에 비해 1.2% 높은 98.3%의 분류 정확도를 기록했으며, 어텐션 네트워크를 차용하여 0.3%의 추가 성능 향상을 이루어냈다.

추가적으로, 본 논문에서는 Yamada et al.이 모델의 성능 평가를 위해 수행했던 것과 비슷하게 클래스 레벨에서의 F1-Score 결과를 분석했다. **표 5-2**와 **표 5-3**는 각각 20Newsgroups와 R8 데이터 셋에서의 F1-Score 결과를 보여준다. 20Newsgroups 데이터 셋을 사용한 실험에서, 모든 클래스에 대하여 압도적인 성능을 보이는 모델은 없었다. 예를 들어,

‘alt.atheism’, ‘talk.politics.misc’ 등의 클래스에서는 텍스트-큐보이드가 가장 좋은 성능을 보였고, ‘comp.graphics’, ‘sci.electronics’과 같은 클래스에서는 NABoE-full 모델이 다른 모델에 비해 좋은 분류 결과를 보였다.

위와 같이 클래스 별로 모델 간의 성능 차이가 존재하는 것은 모델이 문서를 표현하는 기본 방식이 다르기 때문이다. Bow-SVM, TextEnt-full, NABoE-full 모델은 기본적으로 문서에 출현하는 단어들을 바탕으로 문서를 표현한다. 그렇기에 3개의 모델들은 좋은 분류 결과를 보이는 클래스와 그렇지 않은 클래스가 전반적으로 비슷하다. 반면, TC-DNN, TC-Attention 모델은 문서를 표현하는 데 컨셉이라는 새로운 특성을 활용한다. 전체 문서집합을 활용한 실험에서 증명하였듯 컨셉은 전체적으로 분류를 하는 데 도움이 되는 정보를 가지지만, 일부 클래스들에는 혼란을 야기할 수 있다. TC-DNN, TC-Attention 모델에 비해 NABoE-full 모델이 좋은 분류 결과를 보이는 클래스들은 새롭게 정의된 컨셉들이 갖는 정보가 불분명한 경우이다. 비록, 클래스 간 성능 차이는 있지만 컨셉은 통상적으로 분류과정에 도움이 되는 정보를 갖는다. 다만, 컨셉들을 설정하는 데 있어 일정 수준의 Tuning을 필요로 하는 점이 존재한다.

	BoW-SVM	TextEnt-full	NABoE-full	TC-DNN	TC-Attention
alt.atheism	.699	.783	.820	.916	**.920**
comp.graphics	.702	.773	**.818**	.754	.776
comp.os.ms-windows.misc	.714	.742	.802	**.867**	.864
comp.sys.ibm.pc.hardware	.673	.721	.754	.857	**.862**
comp.sys.mac.hardware	.778	.840	.865	**.877**	.872
comp.windows.x	.779	.846	**.867**	.841	.849
misc.forsale	.846	.829	.834	**.907**	.905
rec.autos	.817	.909	**.929**	.848	.843
rec.motorcycles	.900	.943	**.968**	.820	.821
rec.sport.baseball	.895	.941	**.969**	.879	.885
rec.sport.hockey	.935	.960	**.981**	.941	.954
sci.crypt	.890	.934	.940	**.961**	.960
sci.electronics	.721	.757	**.806**	.651	.662
sci.med	.803	.891	**.900**	.831	.825
sci.space	.892	.900	.923	**.949**	.948
soc.religion.christian	.823	.904	.906	.945	**.950**
talk.politics.guns	.781	.810	.828	.908	**.913**
talk.politics.mideast	.837	.944	.940	.952	**.956**
talk.politics.misc	.699	.678	.694	.866	**.879**
talk.religion.misc	.590	.672	.702	**.867**	.862

표 5-2. 클래스 레벨에서의 20Newsgroups 분류결과(F1-Score)

	BoW-SVM	TextEnt-full	NABoE-full	TC-DNN	TC-Attention
grain	.824	.889	.889	**.947**	**.947**
ship	.781	.829	.817	.912	**.928**
interest	.745	.873	.885	.939	**.957**
money-fx	.687	.876	.894	.901	**.933**
trade	.897	.918	.924	**.993**	**.993**
crude	.929	.929	.958	**.996**	.992
acq	.956	.977	.980	.979	**.981**
earn	.986	.988	.990	**.996**	**.996**

표 5-3. 클래스 레벨에서의 R8 분류결과(F1-Score)

제6장 결론

본 논문에서는 위키피디아 페이지를 활용하여 문서를 단어와 컨셉으로 표현하는 인간 지식기반의 텍스트 임베딩 모델인 텍스트 큐보이드와 어텐션 네트워크를 차용한 텍스트 큐보이드-어텐션 문서분류 모델을 제안했다. 20Newsgroups와 R8 데이터 셋을 사용한 다중 클래스 문서분류 실험을 통해 제안 모델이 인간 지식기반의 다른 모델들에 비해 좋은 성능을 보이는 것을 증명할 수 있었고, 어텐션 네트워크가 문서분류 작업에서도 효과적임을 보였다. 본 연구가 신경망 기계번역 분야에서 중심적으로 사용되는 어텐션 메커니즘이 더 많은 모델에 효과적으로 적용될 수 있는 계기가 될 것으로 기대하고 있으며, 추후 연구로서 어텐션 네트워크가 학습한 컨셉 중요도를 바탕으로 한 텍스트 데이터의 증식을 수행할 계획이다.

참고문헌

[1] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013): 3111-3119.

[2] Peters, Matthew E., et al. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365 (2018).

[3] Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018): 12.

[4] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[5] Yamada, Ikuya, Hiroyuki Shindo, and Yoshiyasu Takefuji. "Representation learning of entities and documents from knowledge base descriptions." arXiv preprint arXiv:1806.02960 (2018).

[6] Yamada, Ikuya, and Hiroyuki Shindo. "Neural attentive bag-of-entities model for text classification." arXiv preprint arXiv:1909.01259 (2019).

[7] Kim, Han-joon, et al. "Towards perfect text classification with Wikipedia-based semantic Naïve Bayes learning." Neurocomputing 315 (2018): 128-134.

[8] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

[9] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017): 5998-6008.

Abstract

Automated Text Classification using Semantic Tensor Space Model and Attention Mechanism

Gil-jae Lee

Department of Electrical and Computer Engineering

University of Seoul

Supervised by Prof. Han-joon Kim, Ph. D.

Text classification is deeply related to text embeddings. This is because text classification performance depends on how specifically text embeddings represent words and documents. Recently, text embedding models pre-trained with a large text corpus have outperformed conventional text embeddings. However, most text embedding models have a clear limitation that the models predict the meaning of words just by the context. Some studies showed that text embedding models trained with a human knowledge base could represent words more specifically. Yamada et al. introduced TextEnt, the model trained with a Wikipedia document set, and showed state-of-the-art text classification performance using TextEnt. Likewise, we introduce Text Cuboid that is a human knowledge-based text embedding model representing a document by terms and concepts. Also, we propose Text Cuboid-Attention, a text classification model that applied an attention network to Text Cuboid, outperforming the models

using TextEnt on both 20Newsgroups and R8 datasets.

Keyword: Text Mining, Text Classification, Text Embedding, Machine Learning, Deep Learning, Attention Mechanism.