워드 임베딩의 유사도 클러스터링을 통한 다중 문장 요약 생성 기법

이필원*, 송진수*, 신용태**
*숭실대학교 컴퓨터학과
**숭실대학교 컴퓨터학부
pwlee@soongsil.ac.kr

Multi Sentence Summarization Method using Similarity Clustering of Word Embedding

Pil-Won Lee*, Jin-su Song*, Yong-Tae Shin**
*Department of Computer Science, Soongsil Univ.
**Dept of Computer Science and Engineering, Soongsil Univ.

요 약

최근 인코더-디코더 구조의 자연어 처리모델이 활발하게 연구가 이루어지고 있다. 인코더-디코더기반의 언어모델은 특히 본문의 내용을 새로운 문장으로 요약하는 추상(Abstractive) 요약 분야에서 널리 사용된다. 그러나 기존의 언어모델은 단일 문서 및 문장을 전제로 설계되었기 때문에 기존의 언어모델에 다중 문장을 요약을 적용하기 어렵고 주제가 다양한 여러 문장을 요약하면 요약의 성능이 떨어지는 문제가 있다. 따라서 본 논문에서는 다중 문장으로 대표적이고 상품 리뷰를 워드 임베딩의 유사도를 기준으로 클러스터를 구성하여 관련성이 높은 문장 별로 인공 신경망 기반 언어모델을 통해 요약을 수행한다. 제안하는 모델의 성능을 평가하기 위해 전체 문장과 요약 문장의 유사도를 측정하여 요약문이 원문의 정보를 얼마나 포함하는지 실험한다. 실험 결과 기존의 RNN 기반의 요약 모델보다 뛰어난 성능의 요약을 수행했다.

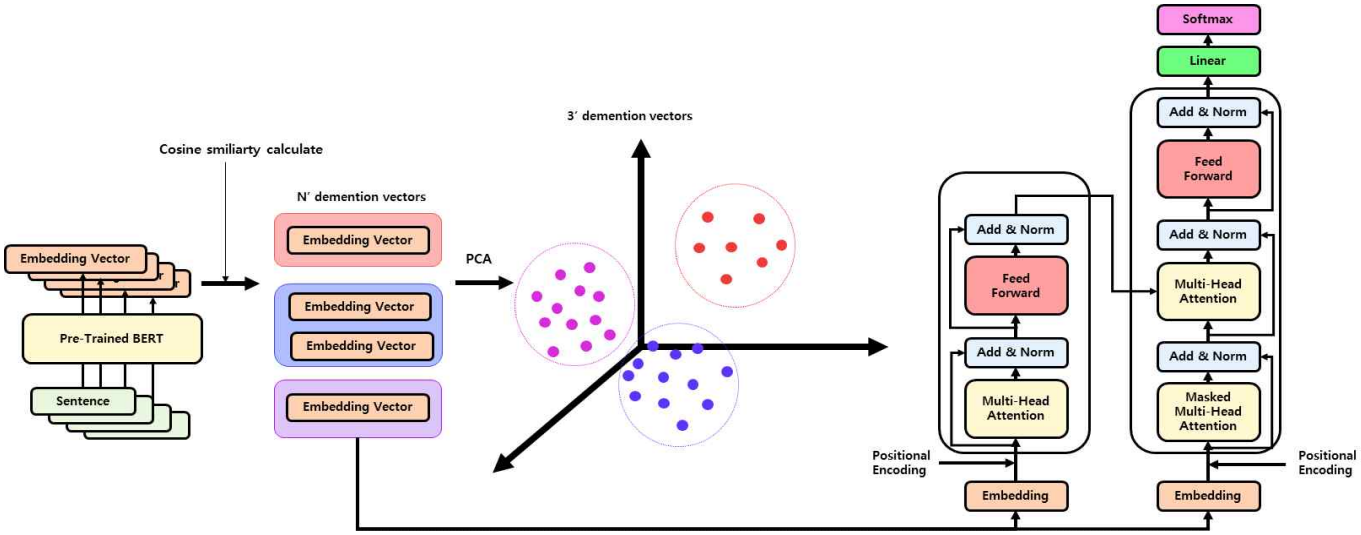
1. 서론

문서 요약은 일반적으로 추출(Extractive)[1]요약과 추상(Abstractive)[2]요약으로 나누어진다. 추출 요약은 문서 내에서 중요한 문장을 평가하여 해당 문장을 추출하는 방법으로 자연스러운 요약과는 거리가 멀다. 추상요약은 문서의 가장 핵심적인 정보를 추출하고 전체적인 글의 내용을 바탕으로 새로운 문장을 생성하여 요약하는 방법으로 최근 활발히 연구가 이루어지는 분야다. 단일 문서의 추상요약은 순환신경망(RNN)의 종류 중 하나인 LSTM(Long Short-Term Memory)을 적용한 인코더-디코더 언어 모델 Sequence-to-Sequence가 뛰어난 성능을 보였다[3]. 이후 RNN을 활용하지 않고 어텐션(Attention)만을 활용하는 트랜스포머(Transformer) 모델이 RNN기반 언어모델 보다 뛰어난 성능을 보였다[4]. 그러나 기존 언어모델의 성능은 지속적으로 발전하고 있지만 여전히 단일 문서를 기준으로 설계되었기 때문에 하나의 핵심적인 정보가 아닌 다수의 핵심 정보를 가지고 있는 다중 문서 요약에 적합하지 않다. 만약 다중 문서로 요약을 수행하여도 요약의 성능이 떨어지는 문제가 발생한다. 따라서 본 논문에서는 사전 학습이 되어있는 워드 임베딩을 통하여 다중 문서 각각의 문장의 임베딩을 추출하고 문장 간의 유사도를 산출하여 클러스터를 구성한다. 또한 클러스터별로 추상 요약을 수행하여 다중 문서가 내포하는 정보를 문장으로 생성한다. 본 논문의 구성은 다음과 같다. 2장에서는 기존 인공 신경망 기반 언어모델의 연구 현황과 언어모델의 구조에 대해 서술하고 기존의 문장의 유사도를 평가하는 방법에 대해서도 서술한다. 3장에서는 본 논문에서 제안하는 워드 임베딩의 유사도 클러스터링을 통한 다중 문서 요약 생성 기법에 대해 설명한다. 4장에서는 제안하는 기법의 성능을 분석하고, 마지막 5장에서는 결론을 제시한다.

2. 관련 연구

2-1.인공 신경망 기반 문서 요약

인공 신경망 기반 문서 요약 모델은 인코더-디코더 구조를 바탕으로 구성된다. 인코더에 입력문장을 입력하면 문장의 단어가 순차적으로 학습된 인공 신경망에 입력되어 결과적으로 하나의 벡터 값 Context Vector가 도출된다. Context Vector는 디코더에 입력되어 단어를 순차적으로 하나씩 단어를 도출하여 문



(그림 1) 제안하는 요약기법 시스템 구성도

장을 완성한다. 요약 모델 중 가장 대표적인 Sequence-to-Sequence 모델은 LSTM 또는 GRU로 구성된 인코더-디코더 구조 모델이다. 이후 RNN 모델의 단점인 기울기 소실문제로 입력문장이 길어지면 요약 성능이 떨어지는 것을 보완하기 위해 주위 집중 매커니즘인 Attrention[5]이 모델에 결합되어 성능을 높였다. 최근에는 RNN을 활용하지 않고 Attention만 활용하여 인코더-디코더를 구성하는 트랜스포머(Transformer) 방식이 연구되고 있으며 성능이 Sequence-to-Sequence 보다 우수한 것으로 나타났다[4]. 따라서 본 논문에서는 문서 요약 모델은 트랜스포머 모델을 기반으로 구축한다.

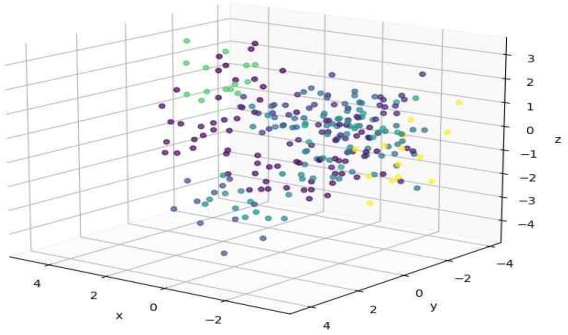
2-2.문장 유사도 평가

추출 요약에서 대표적으로 활용되는 TextRank 모델은 문서에서 중요한 문장에 점수를 평가하여 순위를 나타낼 수 있다[6]. TextRank는 앞서 서술한 추상요약 모델과 비교적으로 오랜 기간 연구가 진행된 모델이다. 그러나 본 논문에서 분석하려는 상품 리뷰와 같은 다중 문서에서는 성능을 기대하기 어렵다. 이에 따라 다중 문서 요약을 위해 문장을 벡터로 표현할 수 있는 워드 임베딩(Word Embedding)을 활용한다. 워드 임베딩은 전통적으로 Word2Vec가 오랜 기간 활용되었지만 최근 트랜스포머 모델의 등장으로 워드 임베딩 역시 영향을 받아 2018년 Google에서 BERT라는 트랜스포머 모델기반 워드 임베딩이 현재 단어 간의 관계를 표현하는데 가장 높은 성능을 가진다. 따라서 본 논문에서는 BERT 기반 워드 임베딩을 활용하여 다중 문서의 각각의 문장을 벡터로 표현하고 코사인 유사도(Cosine Similarity)를 연산하여 유사도가 높은 문장 별로 클러스터를 구성하여 구별하며 수식으로 나타내면 다음과 같다.

$$Similarity = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

3. 제안하는 모델

제안하는 모델은 (그림 1)과 같이 사전 학습된 BERT를 활용한다. 한국어 리뷰를 분석하기 위해 한국어로 학습이 이루어진 SKTBrain이 제공하는 KoBert를 활용한다. KoBert는 한국어 위키를 학습 세트로 활용하여 범용성이 높은 BERT 모델로 상품 리뷰 요약 활용에 적절하다. BERT를 활용하여 각각의 문장의 임베딩 벡터를 얻고 코사인 유사도 연산을 통해 유사도가 95% 이상인 문장 별로 클러스터를 생성한다. 이 때 클러스터를 생성하기 위해 K-means와 같은 비지도 학습은 활용하지 않고 요사도를 기준으로 클러스터를 인덱싱한다.



(그림 2) 유사도 기준의 클러스터링

(그림 2)는 임베딩 벡터를 주성분분석(PCA)를 통해 3차원으로 차원 축소한 클러스터링 결과다. 클러스터링이 완료되면 각각의 문장에 클러스터 인덱싱이 완료된 데이터 세트가 완성된다. 클러스터 별로 문장 요약을 수행하기 위해 트랜스포머 모델을 구현한다. 트랜스포머 모델은 입력을 처리하는 인코더와 출력을 처리하는 디코더로 구성되어 있다. 인코더는 N개의 레이어로 구성되며 각 레이어의 입력은 이전 레이어의 출력으로 연결되고 각 레이어의 출력은 다음 에이어 입력으로 연결되어 있다. 각 레이어는 어텐션 관계를 Q(quert), K(key), V(value) 매트리스로 표시하는 scaled dot-product 과정과 멀티헤드 어텐션으로 구성되며 수식으로 나타내면 다음과 같다.

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (2)$$

최종적으로 트랜스포머 모델을 통해 요약된 문장의 개수는 클러스터의 개수와 같다.

4. 성능평가

논문에서 제안하는 기법의 성능을 평가하기 위해 실제 네이버의 쇼핑 리뷰 데이터 약 4,000건을 수집하여 테스트를 수행한다. 문서 요약 모델을 평가할 때 활용되는 대표적인 지표는 실제 요약문과 예측 요약문을 비교하는 ROUGE 스코어가 활용된다. 그러나 본 논문에서는 다중 문장 요약을 수행하여 요약된 문장이 원문의 의미를 내포하고 있는지 여부를 살펴봐야하기 때문에 ROUGE 스코어는 적절하지 않다. 따라서 전체 문장과 요약 문장의 코사인 유사도를 통해 요약문장이 원문의 정보를 어느 정도 내포하고 있는지 평가한다. 비교 기준 모델은 Wang et al.[7]이 제안한 Sequence-to-Sequence와 성능을 비교한다.

<표 1> 요약문과 전체 문장의 유사도 실험 결과

Summarization Model	Similarity
Sequence to Sequence	76.61
Proposed Method	**88.77**

실험 결과 제안하는 모델은 기존의 LSTM 기반 인코더-디코더 모델과 비교해 다중 문서 요약에서 더 많은 정보를 나타낼 수 있다는 결과를 도출했다.

5. 결론

기존의 RNN 기반의 인코더-디코더 모델의 요약 기법은 최근 발전에 따라 주의 집중 메커니즘 기반의 트랜스포머 모델까지 발전하여 높은 성과를 나타내고 있다. 그러나 요약 모델은 단일 문서를 전제로 설계되었기 때문에 다중 문서를 요약할 때는 성능이 낮아지는 문제점이 있었다. 본 논문에서는 BERT를 활용한 워드 임베딩의 유사도 클러스터링을 통해 유사한 문장을 분류하여 문장 요약을 수행하여 다중 문서가 내포하고 있는 다양한 측면의 정보를 요약할 수 있다. 향후 트랜스포머 요약 모델의 튜닝을 통해 완성도가 높은 한국어 문장을 생성하기 위한 연구가 필요하다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음" (IITP-2020-2020-0-01602)

참고문헌

[1] J. Kupiec, J. Pedersen and F. Chen, "A trainable document summarizer," Advances in Automatic Summarization, 1999, pp. 55-60, 1999.

[2] P. E. Genest and G. Lapalme, "Fully abstractive approach to guided summarization," Proc. of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 354-358, 2012.

[3] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to sequence learning with neural networks," Advances in neural information processing systems, pp. 3104-3112, 2014.

[4] Jacob Devlin, et al., "BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding," In proceedings of NAACL, pp.171-4186, 2019.

[5] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," ICLR, 2016.

[6] Y. J. Bae, H. T. Jang, T. W. Hong, and H. Y. Lee, "Automatic Meeting Summary System using Enhanced TextRank Algorithm", Korea Information Electron Communication Technology, vol. 11, No. 5, pp. 467-474, October 2018.

[7] L. Wang, J. Jiang, H. Chieu, C. Ong, D. Song, and L. Liao, "Can Syntax Help? Improving and LSTMbased Sentence Compression Model for New Domains," Proc. of the ACL, pp. 1385-1393, 2017.