JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 27, No. 1: 41~46, Jan. 2023

빅데이터 분석을 위한 어텐션 기반의 단어 연관관계 분석 시스템

황치곤[1] · 윤창표[2] · 이수욱[3*]

Attention-based word correlation analysis system for big data analysis

Chi-Gon Hwang[1] · Chang-Pyo Yoon[2] · Soo-Wook Lee[3*]

[1]Visited Professor, Department of Computer Engineering, IIT, Kwangwoon University, Seoul, 01897 Korea
[2]Associate Professor,Department Of Computer & Mobile Convergence, GyeongGi University of Science and Technology, Siheung, 15073 Korea
[3*]Associate Professor, Glocal Education Center, Kwangwoon University, Seoul, 01897 Korea

요 약

최근, 빅데이터 분석은 기계학습의 발전에 따른 다양한 기법들을 이용할 수 있다. 현실에서 수집된 빅데이터는 단어 간의 관계성에 대한 의미적 분석을 바탕으로 같거나 유사한 용어에 대한 자동화된 정제기법이 부족하다. 빅데이터는 일반적인 문장으로 기술되어 있다. 이러한 문제를 해결하기 위해 문장의 형태소 분석과 의미를 이해해야 할 필요가 있다. 이에 자연어를 분석하기 위한 기법인 NLP는 단어의 관계성과 문장을 이해할 수 있다. 본 논문에서는 빅데이터에서 추출된 문장에서 단어를 추출하여 단어 간의 연관 관계를 생성하는 방법을 연구한다. 이에 트랜스포머 기술을 이용한다.

ABSTRACT

Recently, big data analysis can use various techniques according to the development of machine learning. Big data collected in reality lacks an automated refining technique for the same or similar terms based on semantic analysis of the relationship between words. Since most of the big data is described in general sentences, it is difficult to understand the meaning and terms of the sentences. To solve these problems, it is necessary to understand the morphological analysis and meaning of sentences. Accordingly, NLP, a technique for analyzing natural language, can understand the word's relationship and sentences. Among the NLP techniques, the transformer has been proposed as a way to solve the disadvantages of RNN by using self-attention composed of an encoder-decoder structure of seq2seq. In this paper, transformers are used as a way to form associations between words in order to understand the words and phrases of sentences extracted from big data.

키워드 : 빅데이터, 어텐션, 자연어처리, 트랜스포머

Keywords : Big Data, Attention, NLP(Natural Language Processing), Transformer

Received 4 December 2022, **Revised** 13 December 2022, **Accepted** 27 December 2022

*** Corresponding Author** Soo-Wook Lee(E-mail:wook@kw.ac.kr, Tel:+82-2-940-5649)
Associate Professor, Glocal Education Center, Kwangwoon University, Seoul, 01897 Korea

Open Access http://doi.org/10.6109/jkiice.2023.27.1.41 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

최근 정형, 비정형 데이터로부터 가치를 생성하여 의미를 부여하기 위한 기법으로 빅데이터 분석 기법을 이용하고 있다. 빅데이터는 기존의 데이터와 다르게 단순히 데이터를 수집하고 정제하여 가공하는 것만 아니라 수많은 데이터를 모아 한꺼번에 해석을 넘어서 새로운 사실 발견 및 예측을 가능하게 한다. 우리가 다양한 데이터들을 수집, 정제, 가공 그리고 종합한 자료들을 통해 결론을 추출하는 것과 다르게 빅데이터는 앞으로 발생할 미래를 예측할 수 있다[1].

이러한 이유로 기계를 통한 언어의 이해는 중요하다. 언어를 이해하기 위해 인공지능의 기법을 이용하여 연구하는 자연어 처리(NLP)는 인공지능의 기계학습 기술의 한 분야이다. 이는 형태소 분석, 품사 태깅, 구문 분석, 의미 추출 등 다양한 기술을 통해 자연어의 이해로 문장을 생성하고, 분석하는 기술을 다루는 커뮤니케이션 기술이다. 특히, 미리 학습된 언어 모델(PLM, Pre-trained Language Model)이 소개된 이후 감성 분석, 문장생성 등 자연어처리의 모든 분야에서 성능이 향상되었고, 이와 관련한 연구가 빅데이터 분야에서뿐만 아니라 언어학, 경영학 등 다양한 분야에 걸쳐 응용되어 진행되고 있다[2]. NLP는 컴퓨터가 사람처럼 언어를 이해하고 처리할 수 있도록 해주는 인공지능의 중요한 연구 분야이며 음성 인식, 정보 검색, 문서 자동 분류, 챗봇, 시스템 자동 번역 등 다양하게 응용되고 있다[3,4].

이러한 자연어처리를 위한 기술로 최근 트랜스포머[5]를 기반으로 하는 학습 모델들이 증가하고 있다. 이에 대해 본 논문에서는 빅데이터에서 추출한 문장들에서 중요 형태소들을 추출하여 트랜스포머를 기반으로 단어들의 연관성을 추출하기 위한 시스템을 제안하고자 한다. 이에 따라 본 논문에서는 2장에서 본 논문과 연관된 연구에 관해 기술하고, 3장에서 제안하는 시스템의 구성과 구성요소들에 관해 기술하고, 4장에서 실험 및 결과를 기술한다. 그리고 5장에서 결론을 기술한다.

Ⅱ. 관련 연구

2.1. NLP(Natural Language Processing)

NLP는 문장의 의미를 분석하여 컴퓨터가 처리할 수 있도록 하는 기법이다. 최근에는 기계학습 혹은 딥러닝을 이용하여 자연어 문서를 처리하는 기술이 발전함에 따라 기계에 언어를 학습시켜 이해를 시키고 있다[6]. 언어 모델은 기계학습의 딥러닝 이전에 주로 사용하던 통계 기반의 자연어처리 기법으로, 언어를 모델링하고 단어의 흐름(문장)에서 단어들의 확률을 할당하는 모델이다. 이러한 언어 모델은 나열된 순서가 중요한 언어 데이터를 처리하기 위한 기법으로, RNN, LSTM 기법이 있고, 이 기법의 단점을 해결하기 위한 트랜스포머[5], 트랜스포머를 기반으로 한 BERT[7], GPT[8]가 언어 모델에 기반을 두어 사전학습을 수행함으로써 언어의 구조와 문맥을 학습하는 방식으로 활용되고 있다.

2.2. Word2Vec

원-핫 인코딩 방식은 단어 집합을 단어를 다차원의 벡터로 표현하는 기법이다. 이는 단어 벡터 간 유사도를 계산할 수 없고 단어의 크기에 따라 벡터가 커진다는 단점이 있다. 이에 비해 Word2Vec은 인공신경망 기반의 워드 임베딩으로 원-핫 인코딩의 희소 행렬 방식이 아닌 밀집 행렬 방식으로 표현하여, 빠른 성능과 단어 간 유사도를 측정할 수 있고, 희소 행렬이 아닌 밀집 행렬 형식으로 표현하기 때문에, 단어 사이의 유사성을 가진다는 장점이 있다. 워드 임베딩은 문장을 구성하는 각 단어의 의미를 수치화할 수 있다. Word2Vec은 문장을 구성하는 단어들의 관계성을 신경망으로 학습시켜 단어의 의미를 내포하는 수치를 벡터로 표현한다[9].

원-핫 인코딩과 Word2Vec의 워드 임베딩 방식의 차이는 표 1과 같다.

Table. 1 The difference between one-hot encoding and word embedding

item	one-hot encoding	word embedding
demention(word size)	important	no matter
vertor type	sparse	dense
value	0 or 1	floating point

이처럼 Word2Vec은 단어의 의미와 문맥을 고려하여 단어를 벡터로 표현하기 때문에 의미상 유사한 단어들끼리 근접한 벡터 공간에 위치하게 된다. 같은 단어라도 문장 내에서 단어의 의미와 문맥에 따라 다른 벡터 공간에 학습될 수 있다. 이런 이유로 Word2Vec에서 사용되

는 워드 임베딩은 원-핫 인코딩과 같이 정수가 아닌 실수 값을 가진다[3].

2.3. Transformer

NLP를 이용한 기술 현황은 최근 인공지능의 기계학습 기법의 발달에 따라 많은 분야로 확장되고 있다, 이는 언어모델, 문서분류, 문서생성, 문서 요약, 질의응답, 기계번역 등과 같은 분야로 확장되고 있다. 이를 위한 기법은 RNN, Seq2Seq 모형, 어텐션 기술, 트랜스포머, OpenAI의 GPT, 구글의 BERT 등과 같은 모델이 있다[10].

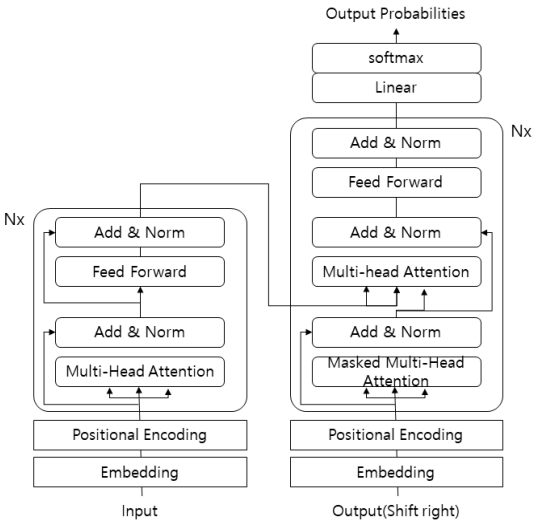

Fig. 1 Transformer Model Architecture[5]

트랜스포머는 2017년 구글이 발표한 논문 “Attention is all you need”라는 논문에서 나온 모델로써, Seq2Seq의 구조인 인코더-디코더를 이용하지만, RNN이나 LSTM이 아닌 어텐션으로 구성한 모델이다. 어텐션 메커니즘은 출력 예측하는 시점마다 입력 전체를 집중해서 참고하는 메커니즘이다[5].

이 모델은 그림 1과 같이 인코더나 디코더만 이용하는 방식의 모델에 관한 연구가 진행되었고, 이를 인코더와 디코더 구조로 구성된 부분에 대한 문제점을 해결해 발전된 대표적인 모델들로 BERT와 GPT가 있다. 이 모델들은 훈련데이터의 양과 파라미터의 수들이 이전의 AI보다 크게 증가한 거대 AI 기술로 자리 잡아 자연어 처리뿐만 아니라 이미지 처리에도 좋은 성능을 내고 있다. 트랜스포머는 RNN의 입출력이 종속적인 문제를 해결하기 위해 인코더-디코더 구조를 이용하여 성능 향상을 보였다.

Ⅲ. 제안 시스템

자연어처리에서 단어 벡터에 대한 유사도 추출을 통한 유사도를 학습에 이용함으로써 의미적 유사성을 측정하여 빅데이터 분석에 기여하는데 목적이 있다.

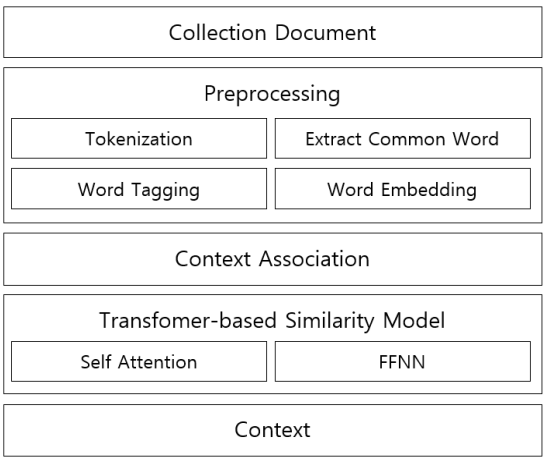

Fig. 2 Proposed System

수집된 문서를 바탕으로 전처리 과정과 컨텍스트 연관성을 이용하여 워드 유사도 산출을 위한 트랜스포머를 이용하고, 이에 관한 결과로 문장에서 단어의 연관성을 생성하는 시스템을 제안한다. 이에 제안하는 시스템의 구성은 그림 2와 같다. 시스템의 처리 과정은 전처리, 컨텍스트 연관성 축적, 트랜스포머 기반의 유사도 모델, 컨텍스트 생성의 순으로 진행된다. 각 구성요소는 다음과 같다.

- 토큰 분리(Tokenization) : 입력된 문서 데이터를 바탕으로 처리를 위한 기본적인 토큰으로 분할하는 작업으로 기본적인 정제를 수행하는 과정이다.
- 공용 워드 추출(Extract Common Word) : 워드 임베딩을 위한 전처리로, 분리된 토큰을 이용하여 명사 추출과 불용어를 제거하여 입력된 문장에 대한 말뭉치(corpus)를 생성한다.
- 태깅(Word Tagging) : 추출된 단어의 유형을 결정하

여 개체명을 인식하거나 품사 태거를 만드는 단계이다. 이는 워드 임베딩을 통한 밀집 벡터를 만들기위한 각 단어에 대한 특징을 결정한다.

- 워드 임베딩(Word Embedding) : 토근 분리를 통하여 생성된 단어를 컴퓨터가 이해할 수 있도록 적절히 숫자로 바꾸는 작업으로, 원핫 인코딩과 같은 희소 행렬 형태로 표현하는 것은 단어의 개수가 늘어날 경우 벡터의 차원이 증가하는 이유로 밀집 벡터로 표현하기 위해 워드 임베딩으로 한다.
- 컨텍스트 연관성(Context Association) : 본 논문에서 제시하고자 하는 것은 빅데이터 분석을 위한 단어간의 연관성 추출을 목적으로 하므로 트랜스포머의 결과를 컨텍스트 연관성에 축적한다. 축적된 연관성을 통해 연관관계를 인식한다.
- 유사도 모델(Transfomer-based Similarity Model) : 어텐션은 주어진 질의에 대해 현재 시점이 아닌 전 시점을 참고하는데 유사도 가중치를 통해 연관된 부분을 집중적으로 참조하는 것이다. 트랜스포머는 이러한 어텐션 중 셀프 어텐션(self-attention)이 있다. 이 어텐션은 어텐션을 자신에게 수행함으로써 문장 내에 단어의 유사도를 추출할 수 있다.

트랜스포머의 입력은 임베딩 벡터의 조정값을 입력받아 셀프 어텐션과 포지션-와이즈 피드 포워드 신경망(FFNN)을 이용하여 트랜스포머 어텐션의 결과를 산출한다. 이 결과를 산출하기 위한 수식은 다음과 같이 표현된다[3]. 이를 통해 연관관계를 추출할 수 있다. 단어간의 관계성을 파악할 수 있다.

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

수식에서 이용되는 벡터 변수는 Q(쿼리), K(키), V(유사도가 반영된 값)이고, d_k는 K 벡터의 차원이다. 트랜스포머의 입력이 되는 임베딩 벡터의 값 조정으로 문장에서 단어의 연관성을 산출할 수 있다.

Ⅳ. 실험 및 결과

본 논문에서 실험은 문장에서 단어 간의 연관성 분석을 위해 빅데이터 분석을 위해 사용할 문장을 추출하여, 이 문장을 구성하는 단어 간의 연관성을 분석하였다.

4.1. 실험 환경

테스트에 사용할 문장을 빅데이터 분석 솔루션인 텍스텀을 이용하여 문장을 추출하였다. 실험은 위한 환경은 구글 코랩에서 텐서플로를 기반으로 Word2Vec, 트랜스포머를 이용하여 실험을 수행하였다. 실험에 이용된 데이터는 “웰니스 관광”으로 수집된 데이터를 이용하여 수행하였다. 추출된 데이터는 텍스텀의 부여 방식에 따라 일반명사(NNG), 고유명사(NNP), 의존명사(NNB), 단위명사(NNBC), 수사(NR), 대명사(NP)로 태깅하였다.

4.2. 실험 결과

단어에 대한 연관성 추출을 위하여 빅데이터 분석 솔루션을 이용하여 데이터를 수집한 결과 7725의 단어를 추출하였고, 이에 따른 연관성 측정을 위하여 빈도수, 누적 빈도수와 TF-IDF를 산출한 결과를 표 2와 같이 구하였다. 키워드를 “웰니스 관광”으로 하여 추출된 키워드들은 대부분이 웰니스 관광이 가지는 의미인 웰빙(well-being), 행복, 건강이라는 뜻과 같이 그와 관련된 지역, 최근에 웰니스와 관련된 정책을 발표한 지역 등의 관련 키워드가 많이 나타났다. 단어의 빈도수는 단순한 출현 횟수를 말하기 때문에 단어의 중요성은 떨어진다고 볼 수 있다. 이에 특정 단어의 중요성을 확인하기 위해 TF-IDF 값을 통해 단어에서 불용어 또는 분석에 악영향을 미칠 수 있다. 이를 통해 표 2와 같이 나열된 단어에서 빈도수가 높지만, 의미가 중요하지 않은 단어들을 제거하였다. 이는 전처리 과정이 정확히 진행되었는지를 확인하기 위해 진행하였다.

Table. 2 Comparison of word frequencies, cumulative frequencies, and tf-idf

word	frequency	frequency rate	cumulative frequency	tf-idf
관광지	4397	2.30%	18.93%	5197.81
선정	2731	1.43%	20.36%	4101.91
제주	2709	1.42%	21.78%	5415.81
여행	2411	1.26%	23.04%	3732.58
치유	2081	1.09%	24.13%	3426.89
인천	1847	0.97%	25.09%	4899.65
힐링	1838	0.96%	26.06%	3111.14

word	frequency	frequency rate	cumulative frequency	tf-idf
연구	1728	0.90%	26.96%	4329.22
건강	1638	0.86%	27.82%	3041.41
코로나	1440	0.75%	28.57%	2846.75
의료	1282	0.67%	29.24%	3311.46
산업	1281	0.67%	29.91%	2812.69
육성	1202	0.63%	30.54%	2640.61
지역	1199	0.63%	31.17%	2807.92
한국	1183	0.62%	31.78%	2574.75
클러스터	1161	0.61%	32.39%	2578.77
경남	1142	0.60%	32.99%	3137.47
서울	1115	0.58%	33.57%	2553.71
관광공사	1090	0.57%	34.14%	2439.11
...	...	...	...	...

연관 관계 추출을 위한 유사도 측정은 앞의 실험을 통해 추출된 단어를 기반으로 수행되었으며, 빅데이터 분석을 위해 수집된 문장을 기반으로 문장 내 단어들 간의 유사도 측정을 수행하였다. 수행된 결과는 표 3과 같이 되었으며, 사용된 문장들에서 중요 단어만으로 수행하였다. 수행된 결과는 트랜스포머의 softmax 알고리즘에 따라 산정되었기 때문에 확률로 표현되었으며 각 단어와 같이 등장한 문장이 있는 경우에 따른 값을 추출한 것이다.

Table. 3 Measure similarity between words according to the proposed system

	관광지	여행	치유	힐링	의료
관광지	1	0.9613	0.7847	0.8135	0.6291
여행	0.9613	1	0.7895	0.9012	0.5412
치유	0.7847	0.7895	1	0.9731	0.6807
힐링	0.8135	0.9012	0.9731	1	0.8922
의료	0.6291	0.5412	0.6807	0.8922	1

이 값들은 문장의 구성이나 의미에 따라 달라질 수 있지만, 표1에서 추출된 키워드 중 제시된 키워드 “웰니스 관광”을 기반으로 추출된 문서들에서는 표 3과 같은 연관성 결과를 0~1사이의 값으로 산출되었고, 이는 추출된 단어의 연관성에 대한 확률이다. 이에 트랜스포머를 기반으로 문장에서의 단어 간의 연관성을 확인할 수 있었다.

Ⅴ. 결 론

트랜스포머는 RNN이나 CNN 같은 신경망의 문제를 해결하기 위해 셀프 어텐션을 통해 집중적 연관성을 제공할 수 있어 앞으로 신경망을 이용할 분야에 중요한 부분이 될 수 있다. 이에 본 논문에서 제안하는 시스템은 빅데이터 분석을 위한 단어 간의 연관성을 제공하기 위한 시스템으로 트랜스포머를 이용하였다. 단어는 문장의 문맥에 따라 유사성이나 동음이의어 같은 이질성을 가진다. 이러한 문제를 미리 정의함으로써 빅데이터 분석에 도움을 줄 수 있다. 현재 타 연구에서 언어의 인식이나 이미지의 인식 부분에 트랜스포머를 적용한 연구가 많이 진행되고 있다. 이에 향후 이미지 인식이나 생성을 위해 적용해야 할 필요가 있다.

ACKNOWLEDGEMENT

This paper was researched by Kwangwoon University's intramural academic research fund support in 2022.

REFERENCES

[1] J. M. Jo, “Effectiveness of Normalization Pre-Processing of Big Data to the Machine Learning Performance,” *The Journal of the Korea institute of electronic communication sciences*, vol. 14, no. 3, pp. 547-552, Jun. 2019. DOI: 10.13067/JKIECS.2019.14.3.547.

[2] J. M. Park, “A Study on the Performance of Document Summarization Using Transformer-Based Korean Pre-Trained Language Model,” M. S. thesis, Ewha Womans University, Korea, 2022.

[3] S. M. Kim, I. S. Na, and J. H. Shin, “A Method on Associated Document Recommendation with Word Correlation Weights,” *Journal of Korea Multimedia Society*, vol. 22, no. 2, pp. 250-259, Feb. 2019. DOI: 10.9717/kmms.2019.22.2.250.

[4] S. Y. Yoo and O. R. Jeong, "Korean Contextual Information Extraction System using BERT and Knowledge Graph," *Journal of Internet Computing and Services(JICS)*, vol. 21, no. 3, pp. 123-131, Jun. 2020. DOI: 10.7472/jksii.2020.21.3.123.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in *Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach: CA, USA, 2017.

[6] S. U. Park, "Analysis of the Status of Natural Language Processing Technology Based on Deep Learning," *The Journal of Big Data*, vol. 6, no. 1, pp. 63-81, Aug. 2021. DOI: 10.36498/kbigdt.2021.6.1.63.

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of NAACL-HLT*, Minneapolis: MN, USA, pp. 4171-4186, 2019.

[8] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," [Internet]. Available: https://s3-us-west-2.amazonaws.com/openaiassets/research-covers/languageunsupervised/language understanding paper. pdf, 2018.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint, arXiv:1301.3781*, 2013. DOI: 10.48550/arXiv.1301.3781.

[10] H. S. Yun and J. J. Jung, “Automated Fact Checking Model Using Efficient Transfomer,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 25, no. 9, pp. 1275-1278, Sep. 2021. DOI: 10.6109/jkiice.2021.25.9.1275.

황치곤(Chi-Gon Hwang)

2012년 광운대학교 컴퓨터과학과 (공학박사)
2006년~2015년:(주)인찬 연구원
2016년~2018년: 경민대학교 인터넷정보과 교수
2019년~현 재: 광운대학교 정보과학교육원 컴퓨터공학과 교수
※관심분야 : 모바일 클라우드, 멀티미디어 온톨로지, 기계학습, NLP

윤창표(Chang-Pyo Yoon)

2012년 : 광운대학교 컴퓨터과학과 (공학박사)
2012년~현 재: 경기과학기술대학교 컴퓨터모바일융합과 교수
※관심분야 : 기계학습, 모바일 시스템, 네트워크 보안, 무선 네트워크, 온톨로지

이수욱(Soo-Wook Lee)

2002년 : 광운대학교 경영학과(경영학박사)
2007년~2013년: 광운대학교 정보과학교육원교수
2013년~현 재: 광운대학교 글로컬교육센터교수
※관심분야 : 재무회계, 빅데이터, 경영정보시스템