The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 19, No. 6, pp.161-167, Dec. 31, 2019. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2019.19.6.161
JIIBC 2019-6-23

워드 임베딩을 이용한 질의 기반 한국어 문서 요약 분석 및 비교

Analysis and Comparison of Query focused Korean Document Summarization using Word Embedding

허지욱*

Jee-Uk Heu*

요 약 현재 ICT 기반의 웹 서비스 발달과 빠른 최신 기술의 보급으로 인하여 생성되는 정보의 양이 기하급수적으로 증가하고 있다. 이와 더불어 사용자들은 자신이 원하는 정보를 얻기 위해서는 많은 시간과 노력을 필요로 한다. 문서요약 기법은 사용자에게 주어진 문서의 문장과 핵심 단어들을 분석하여 효과적으로 요약문을 생성해주는 기술이다. 특히 한국어로 이루어진 문서는 언어의 특성상 기존 언어 분석 기법들을 적용하기 어렵다는 문제점이 있다. 따라서 한국어의 특성을 고려한 문서요약기법에 대한 연구가 필수적이다. 본 논문은 워드 임베딩 기법인 Word2Vec과 FastText를 활용하여 질의 기반의 한국어 문서요약 기법을 제안하고 그 결과를 비교 분석한다.

Abstract Recently, the amount of created information has been rising rapidly by dissemination of state of the art and developing of the various web service based on ICT. In additionally, the user has to need a lot of times and effort to find the necessary information which is the user want to know it in the mount of information. Document summarization is the technique that making and providing the summary of given document efficiently by analyzing and extracting the key sentences and words. However, it is hard to apply the previous of word embedding technique to the document which is composed by korean language for analyzing contents in the document due to the character of language. In this paper, we propose the new query-focused korean document summarization by exploiting word embedding technique such as Word2Vec and FastText, and then compare the both result of performance.

Key Words : FastText, Korean, Query-Based-Document-Summarization, Word2Vec

Ⅰ. 서 론

현재 제 4차 산업혁명이 도래된 가운데 인터넷 기술의 빠른 발달과 보급으로 인하여 ICT(Information & Communication Technology)기술의 발전과 스마트 기기의 대중화 그리고 이를 기반으로 페이스북(Facebook), 인스타그램(Instargram), 유튜브(Youtube) 등과 같은 다양한 사회 관계망 서비스(Social Network Service:

*정회원, 한양대학교 컴퓨터공학과
접수일자: 2019년 10월 12일, 수정완료: 2019년 11월 12일
게재확정일자: 2019년 12월 6일

Received: 12 October, 2019 / Revised: 12 November, 2019 / Accepted: 6 December, 2019
*Corresponding Author: jeeukheu@gmail.com
Dept. of Computer Engineering, Hanyang University, Korea

SNS)활성화로 인하여 생성되는 데이터의 양은 방대하며 이를 저장하고 처리하기 위한 기술 또한 발전하고 있다. IT 시장 조사 기관인 IDC의 조사에 따르면 2017년에 생성되는 데이터의 양은 33ZB(ZettaByte)이며, 2025년에 생성될 데이터는 약5배 정도인 175 ZB에 달할 것이라는 전망을 하였다[1]. 이에 따라 정보들을 단순히 생성 및 검색하는 것이 아니라 사용자에 따라 기존의 정보들을 가공하고 분석하여 유용하고 의미가 있는 정보들을 추출하는 기술들을 필요로 하게 되었다[2]. 특히 사용자가 원하는 정보를 획득하기 위해서는 다양한 검색 엔진을 통하여 직접 질의를 입력하고 결과를 확인해야 하는 번거로움이 발생된다. 문서 요약 기법은 이러한 사용자들의 어려움을 줄여주기 위하여 개발된 기법으로 주어진 문서의 문장들과 단어들을 분석함으로써 핵심 내용을 감지하여 사용자에게 단문으로 요약하여 제공해주는 기술이다. 문서요약을 위해서는 주어진 문서에 존재하는 단어들 중 문서의 핵심이 되는 단어의 추출이 필수적이며 이를 위해서 다양한 기법들이 연구되고 있다. 기존의 연구에서는 문서에 존재하는 단어의 빈도수나 TF-IDF등을 기반으로 한 통계적 방법[3-4], 또는 SVM(Support Vector Machine), Naive Bayesian, LDA(Latent Dirichlet Allocation)등과 같은 지도, 비지도 학습 기법들을 활용한 기계학습과 데이터 마이닝 기술을 이용하여 문서의 핵심적인 단어 또는 문장을 분석하는 방법들을 사용하여 문서를 요약하였다. 하지만 통계를 통한 기법은 단어나 문장이 가지고 있는 의미적인 내용이나 문맥 등을 고려하지 않으며 각 단어들 간의 의미적인 관계 분석 또한 쉽지가 않는다는 문제가 있으며 기계학습을 통한 분석은 학습을 위하여 많은 시간과 높은 비용이 필요하다는 문제점이 있다.

최근 단어의 중요도를 분석하기 위하여 주어진 문서에 존재하는 단어들을 n-차원의 실수 벡터로 변환 하고 컴퓨터가 자연어를 인식할 수 있는 워드 임베딩(Word Embedding) 기술이 각광을 받고 있다. 워드 임베딩은 단어를 특정 벡터로 투사하기 때문에 각 단어들 간의 관계 또한 분석할 수 있으며 이를 기반으로 자연어 분석, 인공지능 처리, 정보 검색 등 다양한 분야에서 활용된다. 기본적으로 워드 임베딩을 위해서는 주어진 문장의 공백이나 띄어쓰기를 기준으로 하여 구분된 어절들이 입력값이 되며, 영어와 같이 고립어로 구성된 언어들은 단어의 문법적인 정보를 헤치지 않고 워드 임베딩을 적용하는데 어려움이 없다. 반면, 한국어는 교착어의 특성을 가지고 있어서 문장에서 독립된 단어가 아닌 조사나 다양한 어미가 단어에 같이 사용 되는 형태가 많다. 때문에 띄어쓰기로 구분된 어절을 입력하는 기존의 기법을 동일하게 적용할 경우 단어의 형태가 훼손된 상태에서 워드 임베딩이 생성되며 이 경우 정상적인 단어의 분석 및 성능을 발휘 하지 못한다는 문제점이 있다.

이를 해결하기 위해서는 Konlpy[5]와 같은 다양한 한국어 형태소 분석기를 활용하여 다양한 어간과 조사와 같은 불필요한 문자들을 제거해야 하는 전처리 과정을 거쳐야하기 때문에 형태소 분석기의 성능이 한국어 임베딩의 결과에 크게 영향을 주게 된다. 따라서 한국어 문서를 워드 임베딩 할 경우 한국어의 특성을 고려하는 것이 중요한 요소 중 하나이다.

본 논문에서는 주어진 한국어 질의를 통하여 검색된 문서 결과를 사용자에게 효과적으로 요약하여 제공해 주기 위하여 두개의 워드 임베딩 기법인 Word2Vec과 FastText 활용하고 그 결과를 평가 및 비교한다. 또한 각 워드 임베딩이 한국어 문서 요약에 높은 효과를 보이는 학습 파라미터 값을 알아보기 위한 실험 및 그 결과를 비교한다.

본 논문의 구성은 다음과 같다. 2장에서는 문서요약과 워드 임베딩에 대한 기존 연구들을 소개 하고, 3장에서는 Wrod2Vec과 FastText를 활용한 질의 기반 한국어 문서요약 기법을 제안한다. 4장에서 실험을 통하여 각 워드 임베딩을 활용한 문서 요약 결과에 대한 비교 분석을 하고 5장에서는 결론 및 향후 연구에 대해 다룬다.

II. 관련 연구

1. 문서요약기법

문서요약기법은 해외의 DUC (Document Understanding Conference) 또는 TAC(Text Analysis Conference) 등과 같은 단체들을 선두로 연구가 진행되고 있으며, 문서요약 평가를 위한 다양한 테스트 데이터 셋을 제공해 주고 있다. 문서요약기법은 요약 방식에 따라 크게 문서의 추출(Extractive)요약과 생성(Abstractive)요약으로 나누어진다. 추출요약은 주어진 문서의 각 문장을 분석하여 중요도 또는 점수를 부여하고 이를 중심으로 요약된 문장을 제공해주는 방식이며, 생성요약은 주어진 문서의 전체적인 내용을 파악하고 중요한 정보들을 중심으로 자연어 생성 기법을 이용하여 문장을 새롭게 생성하여 요약문을 제공해주는 방식이다. 일반적으로 자연어 처리를

필요로 하는 생성요약은 추출 요약 기법보다 효과적으로 내용이 압축된 요약문을 생성하는 장점이 있지만 문법의 오류나 부자연스러운 문장 생성 등의 문제점 때문에 상대적으로 추출요약 기법 위주로 연구가 진행되었다. 그러나 최근에는 Sequence-to-Sequence 모델을 기반으로 한 RNN(Recurrent Neural Network)등과 같은 심층학습(Deep learning)활용한 새로운 기법들의 발달로 이를 활용한 생성요약 기법에 대한 연구가 활발하게 연구가 진행되고 있다[6-7].

문서요약기법은 요약 결과에 따라 주어진 문서의 전반적인 내용을 요약해주는 기법(Generic)뿐만 아니라 사용자의 질의에 따라 검색된 결과를 요약해주는 질의기반(Query-focused)문서요약기법을 위주로 하는 연구도 진행되고 있다. 질의 기반 문서요약기법은 사용자가 입력한 질의어들 간의 의미적인 조합 문맥적 구문을 사용하여 주어진 문서에서 의미 있는 문장을 추출하는 기법이다. 사용자가 질의를 할 때 검색하고자 하는 대상이 명확할 경우 검색결과의 정확도가 높지만 그렇지 않을 경우에는 입력 받은 질의어를 통하여 사용자의 의도를 자동으로 파악하여 사용자가 원하는 정보를 제공해주어야 한다. 따라서 질의어를 분석하여 사용자의 의도를 파악하는 것이 질의기반 요약 기법의 핵심이라고 할 수 있다.

대부분의 문서 요약기법들은 영문 문서를 대상으로 문서를 분석하기 때문에 교착어의 특성을 가지는 한국어를 동일한 기법으로 동일하게 적용하였을 경우 정상적인 결과가 나타나지 않는 다는 문제점이 있다. 최근에는 한국어의 특성을 고려한 문서요약기법에 대한 연구가 진행중이다[8-9].

2. 워드 임베딩

워드 임베딩(Word Embedding)은 자연어 처리에서 주어진 단어를 분석 및 표현하기 위하여 N-차원의 벡터의 공간을 생성하고 실수의 형태로 수치화 하여 표현하는 방법이다. 최근 심화 학습(Deep Learning)과 신경망(Neural Network)기술이 발달이 되면서 워드 임베딩에 대한 다양한 연구가 진행되고 있다. Word2Vec[10] 는 구글에서 개발한 가장 널리 알려준 워드 임베딩 기법이며, 2016년에 스탠포드 대학에서 GloVe[11] 기법을 발표했다. 2017년에는 페이스북에서 FastText[12] 를 발표 하였으며, 그리고 최근에는 Bert[13]와 등과 같은 다양한 워드 임베딩 발표되었다. 최근 개발되는 워드 임베딩 기법들의 가장 큰 특징은 임베딩 시 가능한 단어의 차원을 줄여 계산 복잡도를 줄이고 의미적으로 유사한 단어들은 벡터 공간에서 근접한 벡터로 맵핑하여 생성한다는 점이다. 따라서 임베딩 된 수치로 표현된 단어의 벡터간의 거리 계산을 통하여 주어진 단어들의 의미적인 유사도를 구할 수 있으며 이를 활용하여 인공지능, 자연어 처리, 번역기 등과 같은 정보검색과 같은 다양한 분야에서 적용하여 성능을 향상 시키는데 많은 기여를 하고 있다. 하지만 워드 임베딩을 위하여 학습 시 주어진 입력 값에 등장하지 않은 단어(Out Of Vocabulary: OOV)를 고려하지 않기 때문에 해당 단어에 대한 분석이 어렵기 때문에 그 정확도가 떨어지게 된다. 뿐만 아니라 교착어의 특성을 갖고 있는 한국어의 형태 때문에 학습이전에 형태소 분석기를 이용한 전처리 과정이 필요할 경우가 존재하며 이 과정에서 조사나 다양한 어미와 같은 의미적으로 중요한 정보들이 손실된다는 문제점이 있다.

Ⅲ. 워드 임베딩을 이용한 질의 기반 한국어 문서요약 기법

본 논문에서는 2개의 워드 임베딩 기술인 Word2Vec과 FastText를 활용하여 주어진 문서를 요약하고 그 성능을 비교해 본다.

1. 시스템 구성도

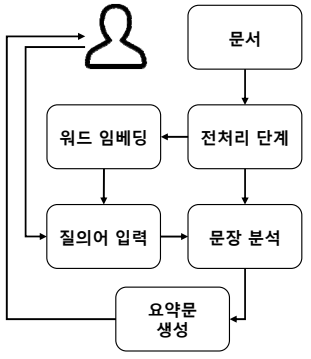

그림 1. 시스템 구성도
Fig. 1. System Architecture

그림1은 본 논문에서 제안하고 있는 질의 기반 문서요약 시스템을 보여주고 있다. 문서를 요약하기에 앞서, 주어진 문서들은 워드 임베딩에 적용하기에 부적절한 기호와 특수 문자들을 전처리 단계에서 제거하고 문서에 존재하는 단어들에 대한 워드 임베딩 기술을 적용한다. 그 후 사용자로 부터 질의어를 입력 받으면 워드 임베딩

기술을 통하여 질의어 분석을 수행하고 질의어와 유사한 단어와 함께 주어진 문장들 중 관련성이 높은 문장에 점수를 부여한다. 마지막으로 문장의 점수들을 기반으로 요약문을 생성하여 사용자에게 제공한다.

2. 워드 임베딩 단계

최근 연구되는 워드 임베딩 기법들은 유사한 의미를 가지는 단어들은 동일한 문맥에서 발생된다는 분포 가설(Distributional Hypothesis)을 기반으로 모델을 생성되고 있다[14]. 그중 Mikolov가 제안한 Word2Vec 기법은 워드 임베딩 시 주어진 문서 존재하는 단어들의 빈도와 위치들을 파악하고 단어의 확률적인 출연 까지 고려하여 학습하는 모델이다. Word2Vec 은 2개의 신경망을 구축하여 학습하며 이를 기반으로 주어진 단어들에 대한 벡터들을 생성하며 학습 시 CBOW(Continuous Bag of Words)모델 또는 Skip-gram모델이 사용된다.

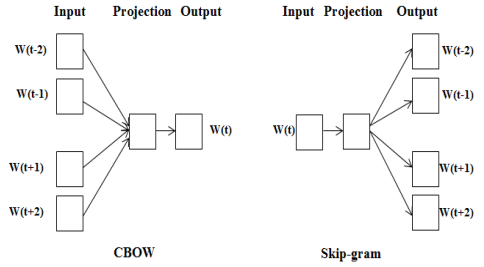

그림 2. Word2Vec의 CBOW 모델과 Skip-gram 모델
Fig. 2. CBOW and Skip-gram model of Word2Vec Architecture

그림 2.는 Word2Vec에서 학습되는 모델인 CBOW와 Skip-gram의 학습 신경망을 보여주고 있다. CBOW 모델은 입력 층에서 각 주변의 단어들을 $t-2$, $t-1$, $t+1$, $t+2$의 입력 값으로 하며 이들에 가중치 행렬을 적용하여 출력 층에서 N차원의 벡터가 형성된다. 이를 기반으로 형성된 모델은 실제 예측하려고 하는 대상 단어인 t 를 비교 하며 학습하게 된다. Skip-gram모델은 CBOW모델과 반대로 대상이 되는 단어 t를 입력 층의 입력 값으로 사용하고 가중치 행렬을 적용하며 N차원의 벡터를 형성하여 출력 층으로 내보낸다. 그 후 만들어진 벡터를 기반으로 예측하려고 하는 주위 단어들을 $t-2$, $t-1$, $t+1$, $t+2$를 비교 학습을 하게 된다. Word2Vec을 사용하여 워드 임베딩 시 일반적으로 학습하는 양이 Skip-gram모델이 CBOW모델보다 많기 때문에 Skip-gram모델을 주로 많이 사용한다.

FastText는 페이스북에서 제안한 기법으로 기본적으로 Word2Vec의 학습방법과 동일하나 단어 기반으로 학습되는 Word2Vec과는 달리 n-gram의 문자(character) 단위로 학습한다는 차이점이 있다. 즉 입력 층에서 단어를 n-gram으로 문자로 나눈 각 단어의 subword들로 학습이 진행되며, 각 subword들로 표현된 벡터들의 합이 한 단어의 벡터로서 출력 층에 표현되게 된다. 따라서 학습된 FastText 모델은 오타나 잘못 입력된 형태의 단어가 입력되었다 하더라도 비슷한 단어가 예측이 가능하게 된다. FastText 학습 시 계산되는 워드 임베딩시 계산되는 각 단어의 $Score$ 값은 다음과 같이 표현된다.

$$s(w,c) = \sum_{g \in G_w} z_g^{\top} v_c \qquad (1)$$

위 식(1)의 v_c는 문서의 문맥에 포함된 단어의 벡터를 나타내며, $z_g^{\top}$는 n-gram에 의해 생성된 단어의 subword의 벡터를 의미한다. 최종적으로 각 subword의 벡터 총합이 하나의 단어 벡터로서 표현이 된다.

3. 문서요약 단계

문서를 요약하기 위해서는 주어진 문서에 존재하는 문장들을 분석하고 이를 기반으로 각 문장에 점수를 부여하는 과정이 필요하다. 본 논문에서는 질의 기반의 문서 요약을 위하여 주어진 질의어 그리고 이와 의미적으로 유사한 단어들은 워드 임베딩 모델을 활용하여 찾아낸다.

우선 요약을 위해 주어진 문서 중 가장 의미 있는 문장을 추출하기 위해서 사용자 질의어와 가장 유사한 단어가 포함된 문서들을 검색을 하여 점수를 부여한다. 각 문장에 점수를 부여하기 위하여 아래의 식(2)를 적용하여 계산한다.

$$SenScore_i = FindSen(q, SimWord(q)) * tf \bullet idf(q) \qquad (2)$$

$FindSen$ 함수는 사용자 질의어와 $SimWord$ 함수를 통하여 획득된 가장 유사도 높은 단어가 포함된 문장i를 찾고, 만약 문서에 문장i 가 존재하면 단어들 간의 유사도 값을 반환한다. 질의어와의 유사도 값은 $SimWord$ 함수를 이용하며 $SimWord$ 함수는 워드 임베딩에서 제공해주는 함수로서, 입력된 단어와 가장 유사한 단어 벡터단어와 유사도를 반환해준다. 그 후 질의어의 $tf \bullet idf$값을 추가적으로 적용하여 해당 문장i의 점수 $SenScore_i$값을 부여하게 된다. 계산된 문장들은 높은

점수가 주어진 문장을 기존으로 요약문을 생성하게 된다.

Ⅳ. 실험 및 결과

본 논문에서는 실험을 위하여 2019년 1월부터 2019년 3월 까지 약 3개월 간 생성된 네이버 인터넷 뉴스 3만 건을 수집하였다. 수집된 뉴스기사는 “사회”, “정치”, “경제”, “IT과학”, “오피니언”, “생활문화”등으로 다양한 주제들을 다루고 있다. 정확한 실험을 위하여 수집된 뉴스 데이터 중 내용이 2~3문장의 짧은 단문으로 이루어지거나 내용은 없고 제목만 있는 기사들은 실험에 사용하지 않았다. 질의 기반 문서 요약의 평가를 위한 질의어는 네이버 데이터 랩(https://datalab.naver.com/)에서 제공되는 서비스인 ‘급상승검색어’를 참조하여 2019년 1~3월에 해당되는 기간 중 앞서 수집된 뉴스기사들과 관련성이 높은 질의어들만으로 추려서 실험을 진행하였다.

문서요약 평가 방법으로는 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)[15]를 사용하였다. ROUGE는 정답 데이터와 생성된 요약문의 문장 간에 존재하는 단어와 단어가 출연하는 순서의 일치 하는 정도를 정확율과 재현율 그리고 F-Measure를 측정하는 방법으로 요약문의 성능을 측정하고 우수성을 판단하는 척도로 사용된다. ROUGE-N의 평가 방법은 아래의 수식 (3)과 같다.

$$ROUGE-N=\frac{\sum_{S\in S_H}\sum_{g_n\in S} C_m(g_n)}{\sum_{S\in S_H}\sum_{g_n\in S} C(g_n)} \qquad (3)$$

S_H는 정답 데이터 S 의 집합을 의미하며 $C_m(g_n)$은 생성된 요약문과 정답 요약문과의 n-gram 만큼의 얼마나 일치하는지를 나타낸다. 본 논문에서는 ROUGE 측정 시 수집된 뉴스기사의 제목을 정답 요약문으로 설정을 하였으며, 생성된 각 뉴스의 요약문들과 비교하여 실험을 진행 하였다.

표 1은 워드 임베딩 시 FastText의 임베딩 벡터 차원 수에 따른 요약 결과를 보여주고 있다. 임베딩 차원의 수가 높을수록 좋은 성능을 보여주고 있으나, 벡터의 차원을 높일수록 소요되는 학습과 수행시간이 기하급수적으로 늘어나는 것을 확인 할 수 있었다. 따라서 성능의 효율성을 고려 할 경우 임베딩 벡터 차원의 수가 300일 경우가 가장 효과적이라고 할 수 있다.

표 1. FastText 차원의 수에 따른 문서요약 성능
Table 1. Document summarization performance for number of dimensions on Fasttext word embedding

Embedding Dimension	ROUGE-1	ROUGE-2	ROUGE-L
100	0.234	0.179	0.230
200	0.252	0.178	0.259
300	0.276	0.180	0.263
400	0.271	0.184	0.262
500	0.273	0.183	0.265

표 2. FastText와 Word2Vec 윈도우크기에 따른 문서요약 성능
Table 2. Document summarization performance for number of window size on FastText and Word2Vec word embedding

Window size	ROUGE-N	Word2Vec	FastText
3	ROUGE-1	0.224	0.254
	ROUGE-2	0.155	0.179
	ROUGE-L	0.196	0.226
4	ROUGE-1	0.231	0.252
	ROUGE-2	0.159	0.178
	ROUGE-L	0.189	0.219
5	ROUGE-1	0.253	0.276
	ROUGE-2	0.167	0.180
	ROUGE-L	0.223	0.263

표 2는 FastText와 Word2Vec 워드 임베딩 시 차원의 수를 300으로 고정하고 윈도우 크기(window size)에 변화 주고 생성한 요약문의 평가를 비교한 ROUGE 결과를 보여주고 있다. Word2Vec은 한국어의 특성을 고려하지 않기 때문에 전처리 단계 없이 워드 임베딩 기술을 적용하면 단어의 형태가 훼손된 상태로 임베딩이 생성되므로 MeCab 형태소 분석기를 사용하였다. 그 후 주어진 문서에서 명사(NN)만을 추출하여 워드 임베딩을 생성하였다. 표2에서 확인할 수 있듯이 질의어를 입력할 때 요약 시 Word2Vec 보다 FastText의 워드 임베딩 기술을 적용하는 것이 우수한 성능을 보이며 약 11.9% ~ 18.1%의 향상된 결과를 보여주고 있다. 실험 시 주어진 질의어가 부정확한 표기가 있을 경우 Word2Vec기반의 워드 임베딩은 질의어의 유사도 분석이 정상적으로 수행되지 않았으며 그로 인하여 요약문의 결과 또한 부정확하다는 것을 확인할 수 있었다. 반면에 FastText는 n-gram 단위로 학습하였기 때문에 질의어가 부정확한 표기로 입력 받는다 하더라도 이와 유사한 단어에 대한 추측이 가능하기 때문에 이를 기반으로 정상적인 요약문 생성이 가능하였다.

V. 결 론

본 논문에서는 사용자로부터 주어진 한국어 질의를 통하여 검색된 문서 결과를 사용자에게 효과적으로 요약하고 제공해 주기 위하여 두개의 워드 임베딩 기법인 Word2Vec과 FastText 활용하여 요약문을 생성하고 그 결과를 평가 및 비교하였다. Word2Vec은 한국어의 특성을 고려하지 않기 때문에 단어의 형태가 훼손된 상태로 워드 임베딩을 생성하므로 형태소 분석기를 통한 전처리 단계를 거친 다음에 워드 임베딩을 생성하였다. 반면 FastText는 n-gram의 문자 단위로 단어를 학습하여 워드 임베딩을 생성하기 때문에 전처리 단계에서 형태소 분석기를 사용하지 않고 적용하였다. 실험 결과 Word2Vec 보다 FastText를 워드 임베딩하여 활용한 생성된 요약문 더 우수한 성능을 보였다.

수집된 한국어로 이루어진 뉴스들 중 제목이 추상적이거나 지나치게 자극적일 경우 문서에 대한 정상적인 요약문이 생성이 되어도 해당 요약문에 대한 평가를 할 수가 없다는 한계점이 존재하였다. 이를 보완하기 위하여 실험 전 수집된 문서들에 대한 충분한 검토가 필수적이며 각 뉴스 문서에 대한 정답 요약문 구축을 할 예정이다. 뿐만 아니라 특정 도메인이 아닌 다양한 주제를 가지고 있는 뉴스들 또한 요약문 생성이 가능한 문서요약 시스템을 구축에 대한 연구를 진행 할 예정이다. 뿐만 아니라 최근에 공개되어 각광받고 있는 워드 임베딩 기술인 GloVe와 Bert를 한국어의 특성을 고려하여 다양한 정보검색 분야 서비스를 제공할 수 있도록 지속적으로 연구를 할 예정이다.

References

[1] D. Reinsel, J. Gantz, J. Rydning, “The Evolution of Data to Life-Critical”, Data Age 2025, 2017.

[2] D. J. Shiin, J. H. Park, J. H. Kim, K. J. Kwak, J. M. Park, J. J. Kim, “Dig Data-based Processing and Analysis for IoT Environment”, The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 13, No. 5, pp. 37-47, Oct 2013.
DOI: https://doi.org/10.7236/JIIBC.2019.19.1.117

[3] N. G. Kim. S. J. Kang, “Relevant Image Retrieval of Korean Documents based on Sentence and Word Importance,” Journal of the Korea Academia-Industrial cooperation Society(JKAIS), Vol. 20, No. 3, pp. 43-48, 2019.
DOI: https://dx.doi.org/10.5762/KAIS.2019.20.3.43

[4] D. S. Park, and H. J. Kim, “A Proposal of Join Vector for Semantic Factor Reflection in TF-IDF Based Keyword Extraction”, The Journal of KIIT, Vol. 16, No. 2, pp. 1-16, Feb 2018
DOI: https://dx.doi.org/10.14801/jkiit.2018.16.2.1

[5] E. J. Park, and S. Z. Cho, “KoNLPy: Korean natural language processing in Python”, Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology, pp. 133-136, Oct 2014.

[6] Baumel, Tal, Matan Eyal, and Michael Elhadad. "Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models." arXiv preprint arXiv:1801.07704, 2018.

[7] Chopra, S., Auli, M., & Rush, A. M. “ Abstractive sentence summarization with attentive recurrent 0neural networks”, In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 93-98, Jun 2016. DOI: https://doi.org/10.18653/v1/n16-1012

[8] J. S. Seol, S. G. Lee, “lexrankr: LexRank based Korean multi-document summarization”, The Korean Institute of Information Scientists and Engineers, pp. 458-460, Dec 2016.

[9] K. H. Choi, C. Lee, “End-to-end Korean Document Summarization using Copy Mechanism and Input-feeding”, The Korean Institute of Information Scientists and Engineers, Vol. 44, No. 5, pp. 503-509, 2017.

[10] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J, “Distributed representations of words and phrases and their compositionality”, In Advances in neural information processing systems, pp. 3111-3119, 2013.

[11] Pennington, J., Socher, R., & Manning, C., “Glove: Global vectors for word representation”, In Proceedings of the 2014 conference on empirical methods in natural language processing, pp. 1532-1543, Oct 2014.

[12] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T., “Enriching word vectors with subword information”. Transactions of the Association for Computational Linguistics, pp. 135-146, 2017.

[13] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., “Bert: Pre-training of deep bidirectional transformers for language understanding”. arXiv preprint arXiv:1810.04805. 2018.

[14] Sahlgren, M., “The distributional hypothesis”, Italian Journal of Disability Studies, Vol. 20, pp. 33-53, 2008

[15] Lin, C. Y., “Rouge: A package for automatic evaluation of summaries”. In Text summarization branches out, pp. 74-81, 2004.

저 자 소 개

허 지 욱(정회원)

- 2007년 한림대학교 컴퓨터공학과 학사 졸업,
- 2009년 한양대학교 컴퓨터공학과 석사 졸업,
- 2016년 한양대학교 컴퓨터공학과 박사 졸업,
- 현재 한양대학교 박사 후 과정,

- 관심분야 : 멀티미디어 정보검색, SNS 분석, 빅데이터 분석, 워드 임베딩, 다중문서요약 등