https://doi.org/10.5392/JKCA.2019.19.10.620

단어 빈도와 유사도 분석 기반의 회의록 요약 시스템 설계 및 구현

Design and Implementation of Minutes Summary System Based on Word Frequency and Similarity Analysis

허강호*, 양진우*, 김동현*, 복경수, 유재수*****

(주)바론*, 원광대학교 SW융합학과**, 충북대학교 정보통신공학부***

**Kanhgo Heo(kh@ibaron.co.kr)*, Jinwoo Yang(school@ibaron.co.kr)*,
Donghyun Kim(dh@ibaron.co.kr)*, Kyoungsoo Bok(ksbok@wku.ac.kr)**,
Jaesoo Yoo(yjs@chungbuk.ac.kr)*****

요약

의사 결정을 위한 토론이나 토의의 내용을 객관적 요약하고 분류하는 자동화된 회의록 요약 시스템이 요구되고 있다. 본 논문은 기존에 사용되었던 회의록 요약 시스템을 보완할 수 있도록 word2vec 모델을 이용한 회의록 요약 시스템을 설계하고 구현한다. 제안 시스템은 형태소 분석 과정에서 불용어를 제거하고 문서에서 공통적인 의견을 가진 대표 문장을 추출하기 위해 추가로 word2vec 모델로 학습을 수행한다. 제안 시스템은 회의 과정에서 수집되는 문서를 분석하여 자동으로 분류하고 다양한 의견들 중 안건을 대표하는 대표 문장을 추출한다. 회의 진행자는 제안 시스템을 통해 회의에서 다뤄지는 모든 안건을 보다 빠르게 확인하고 관리할 수 있다. 제안 시스템은 대규모 토론이나 토의의 여러 가지 안건을 분석하여 대표 의견이 될 수 있는 문장을 요약하여 빠른 정확한 의사 결정을 지원한다.

■ 중심어 : | 회의록 요약 | 단어 빈도 | 유사도 분석 | word2vec | 의사 결정 |

Abstract

An automated minutes summary system is required to objectively summarize and classify the contents of discussions or discussions for decision making. This paper designs and implements a minutes summary system using word2vec model to complement the existing minutes summary system. The proposed system is further implemented with word2vec model to remove index words during morpheme analysis and to extract representative sentences with common opinions from documents. The proposed system automatically classifies documents collected during the meeting process and extracts representative sentences representing the agenda among various opinions. The conference host can quickly identify and manage all the agendas discussed at the meeting through the proposal system. The proposed system analyzes various agendas of large-scale debates or discussions and summarizes sentences that can be representative opinions to support fast and accurate decision making.

■ keyword : | Minutes Summary | Word Frequency | Similarity Analysis | Word2vec | Decision Making |

* 이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-차세대정보·컴퓨팅기술개발사업의 지원(No. NRF-2017M3C4A7069432), 산업통상자원부와 한국산업기술진흥원의 "R&D재발견프로젝트"의 지원(과제번호: P0010202, 과제명: 소셜 빅데이터 기반의 개인맞춤형 취업 콘텐츠 추천(큐레이션) 및 디지털 증명서 발급 시스템), 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2019R1A2C2084257).

접수일자 : 2019년 07월 31일 심사완료일 : 2019년 09월 05일
수정일자 : 2019년 09월 05일 교신저자 : 유재수, e-mail : yjs@chngbuk.ac.kr

I. 서 론

현대 사회에서는 공동의 문제를 해결하기 위한 의사결정을 위해 다양한 형태의 토론이나 토의를 활용한다[1][2]. 다양한 주제로 진행되는 토론이나 토의에서 제시된 모든 의견들은 회의록을 만들어 문서화한다. 문서화된 의견들은 범주화하고 비슷한 의미를 갖는 회의 내용들을 통해 대표적인 의견을 도출한다. 이렇게 정리된 의견들은 수정 및 보완 단계를 거쳐 문제 해결을 위한 결론을 도출한다. 결과를 도출하는 일련의 과정을 진행하고 도와주는 퍼실리테이터(facilitatior)는 수작업으로 모든 과정을 진행한다. 일반적으로 퍼실리테이터 수는 토론이나 토의 참여자 수에 비례하며 이로 인해 대규모 토의나 토론일수록 인건비가 증가되고 최종 결론을 도출하기 위해 많은 시간이 요구된다.

토론이나 토의를 문서화한 회의록을 사람이 수작업으로 수행할 경우 잘못된 형식의 문장들을 포함할 수 있고 문서의 주제를 객관적으로 판별하지 못할 수 있다. 따라서 주제에 대해 정확한 표현으로 문서를 요약하고 분류하는 자동화된 회의록 요약 방법이 요구되고 있다[3][4]. 기존 회의록 요약 방법으로는 문서에서 불필요한 단어들을 제거한 압축된 문장들만을 이용하여 각 문장에 포함된 단어들의 통계적 정보를 이용한 방법이 있다. 이러한 방법들은 문장을 간결하게 만든 후 요약 정보를 생성하기 때문에 불필요한 단어들이 요약문에 포함되는 것을 방지할 수 있지만 유사 단어를 고려를 하지 못하는 문제점이 있다. 그로 인해 문장들 간의 유사도 역시 고려할 수 없어 여러 의견을 유사한 의견으로 분류하는데 어려움이 있었다.

일반적으로 문서를 자동 분류하는 방법은 이미 분류되어 있는 문서들로부터 문서 내에 나타나는 단어의 출현 횟수나 분포, 확률 등을 이용하는 통계적인 방법[5]과 자연어 처리를 통하여 문서 내에 있는 문장의 의미(semantic)나 구문(syntactic)을 분석하는 의미 분석 방법[6]이 있다. 보다 정확한 문장의 분류를 위해서는 자연어 처리를 통하여 문서의 내용을 파악하는 것이 바람직 하지만 자연어 자체의 모호성 때문에 문장의 의미 분석이 매우 어려워 의미 분석 방법은 한정된 영역에서 사용하기에 적합하다[7].

본 논문에서는 단어빈도 분석 뿐 아니라 단어 간의 유사도 분석을 통해 문장을 분류하여 더욱 효과적인 회의록 요약 시스템을 제안한다. 전체 문서에 나타나는 단어들의 출현 빈도에 대한 정보를 사용하여 문장을 분류한다. 그러나 단어의 출현 빈도만을 고려하는 것은 여러 문장에 동시에 출현하는 단어의 성질을 간과한 한계점을 가진다. 이에 출현 빈도를 고려한 특정 단어를 선정하고 특정단어를 포함하고 있는 문서를 정확하게 분류하기 위해 워드 임베딩을 활용한 접근법을 제안하며 워드 임베딩 모델 중 대표적인 word2vec 모델을 사용한다. word2vec 모델을 통해 문서의 각 단어의 벡터 공간에서 거리를 계산하고 문서의 특징을 포함하고 있는 각각의 자질들로 심층적 학습을 수행한다. 학습한 단어의 벡터 값을 이용해 유사단어를 출력하고 자동으로 문장을 분류한다.

II. 관련연구

1. 문서 분류

기존에 문서 분류를 자동화하기 위하여 단순히 문서에 나타나는 단어의 빈도를 이용하여 적합한 범주를 지정하는 통계적인 분류 방법을 이용하거나[8], 분류에 필요한 주요 단어들을 추출하고 추출된 단어들을 기반으로 K-NN의사결정 트리, 베이지언 네트워크, 인공 신경망 등의 데이터 마이닝 알고리즘을 이용한 연구가 진행되었다[9]. 최근에는 자연어 처리에 딥러닝(deep leaning) 알고리즘인 컨볼루션 신경망(CNN: Convolutional Neural Network)이 효과적이라는 것이 알려지면서, 단어를 벡터(vector)로 표현하는 방법인 word2vec 모델[9]과 CNN을 이용한 문장 분류 방법이 제안되었다[10].

[11]에서는 신경망 분석에 기반을 둔 비지도 학습 기법인 word2vec 모델을 제안하였다. 이 모델은 각 단어들이 학습 문헌 내에서 가지는 의미를 다차원의 벡터 값을 통해 수치적으로 표현하는 것을 목표로 한다. 이렇게 계산된 학습 결과를 다른 기계학습에서 학습 자질로도 사용하여 성능이 향상되었다.

2. 문서 표현

문서에 대한 수학적인 표현 모델로 가장 일반적인 것은 Bag-of-word 모델이다. 이 모델은 한 문서가 포함하고 있는 모든 단어와 각 단어가 문서 내에서 등장한 빈도수를 매칭시킨 짝(pair)들의 집합을 해당 문서에 대한 표현으로 정의하는 것이다. 이 방법을 통해 문서를 표현하게 되면 문서에 포함된 각각의 어휘는 해당 문서의 특징(feature)이 되고, 각 어휘에 대한 빈도수는 특징에 대한 값(feature value)이 된다. Bag-of-Word 모델에서는 문서 내에 단어가 등장하는 순서에 대한 정보는 보존 되지 않으며, 정확히 같은 종류의 어휘와 빈도수 분포를 가지고 있는 두 문서는 동일하게 표현되어 완전히 일치하는 것으로 간주되는 것이 특징이다[12].

단어의 순서와 의미를 내포하는 벡터의 형태로 단어를 표현하는 기법으로 word2vec 모델이 가장 대표적이다. word2vec 모델은 특정 embedding 공간상에서 같은 맥락을 갖는 단어들이 가까운 거리를 가진다는 전제에서 출발한다[13][14]. 단어 사이의 거리를 활용한 워드 임베딩 방식의 word2vec 모델은 주어진 문장에 대한 문법적 해석이 가능하며 단어의 거리를 통해 의미론적 추론도 가능하다. 즉, 주어진 문장을 구성하는 단어들의 전후 관계를 학습하여 단어의 의미를 내포하고 있는 벡터 값으로 문서를 구성하고 있는 자질들을 수치화한다. 이것은 기존의 통계적인 방식을 활용한 연구와는 다르게 별도의 유사도 계산이나 차원 축소 과정 없이 변별적인 특징을 내포하고 있는 벡터 값으로 단어를 수치화한다.

Ⅲ. 회의자료 분석 시스템

1. 전체 처리 과정

본 연구에서 제안하는 회의록 요약 시스템은 특정 주제 분야의 텍스트 집단 전체 문장과 회의록 문서 내부의 모든 텍스트를 분석하여 학습과정을 거친 후, 학습된 정보를 적용하여 대상 문서 내부에서 필요한 문장을 찾아낸다. 실험집단은 각 시도 교육청에서 이루어지는 토의를 기반으로 한 회의록을 대상으로 하였다.

[그림 1]은 제안 기법의 전체 처리 과정을 나타낸다. 제안 기법은 크게 학습 과정과 요약 과정으로 구분된다. 학습 과정은 여러 회의록 데이터를 단어로 쪼개는 자연어 처리, 불필요한 단어를 삭제하는 불용어 처리, 각 단어와의 연관성을 보여주는 word2vec 모델로 구성된다. 출력 과정은 문서에서 표현된 언급 순으로 단어의 빈도와 word2vec 모델 학습 결과로 출력된 유사 단어를 통해 문서에서 요약된 문장 중에서 대표 문장을 검색하여 선택하는 과정으로 구성되어 있다.

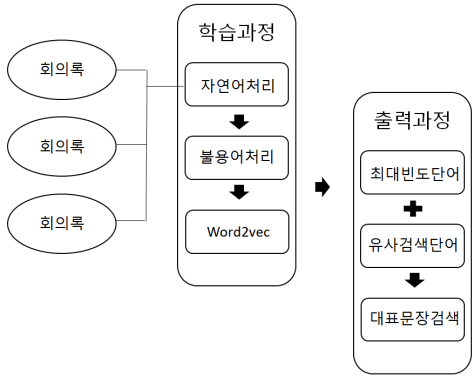

그림 1. 전체 처리 과정

2. 자연어 처리

자연 언어는 사람들이 생활 속에서 사용하는 언어를 의미한다. 회의록의 경우 국문, 영문 등 인간의 언어인 자연형태로 구성되어 있다. 자연어는 컴퓨터가 이해할 수 없는 언어이기 때문에 기계학습을 진행하기 위해 컴퓨터가 이해할 수 있는 형태로 표현해야 하는데 이에 따른 제반기술을 자연어 처리를 통해 수행한다.

형태소는 일정한 의미를 지닌 가장 작은 말의 단위로 문장 내에서 따로 떼어낼 수 있는 것을 나타낸다. 더 이상 분해하거나 분석하면 뜻이 없어지는 말의 가장 최소 단위로 추상적이며 다양한 형태로 나타내어질 수 있다. 자연어 처리 처리단계는 형태소 분석, 동사 분석, 의미 분석, 화용 분석으로 나눌 수 있다. 형태소 분석은 텍스트를 형태소 단위로 분석하여 출력하는 과정이다. 즉, 형태소를 비롯하여 어근, 접두사/접미사, 품사(part-of-speech, POS) 등 다양한 언어적 속성의 구조를 파악하는 것이다.

최근 전산 언어학에서 말뭉치(corpus)의 활용이 높

아지면서 형태소 분석과 품사 태깅 등 한국어 정보 처리를 위한 다양한 오픈 소스 라이브러리와 패키지가 공개되고 있다. 한국어 정보 처리를 위한 오픈 소스 라이브러리도 이미 1995년부터 KTS를 시작으로 다수 배포되고 있다. 그 중 대다수는 C/C++, Java 등으로 개발되어 있으며 형태소 분석기가 중심이다. 2014년에는 파이썬으로 한국어 정보처리를 할 수 있는 패키지인 KoNLPy(Korean NLP in Python)가 제안되었다. KoNLPy는 쉽게 습득하고 빠르게 사용할 수 있어야 한다는 설계 철학에 따라 파이썬을 이용해 한국어 정보처리를 지원한다. 또한, KoNLPy는 꼬꼬마, twitter 등의 여러 형태소 분석기를 사용할 수 있도록 지원한다[15]. 따라서 본 논문에서도 형태소 분석을 위해 KoNLPy 패키지를 이용하며 twitter 형태소 분석기를 이용하여 회의록의 형태소 분석을 수행한다. twitter 형태소 분석기를 이용하여 한글 문자열로부터 명사를 추출할 수 있다.

[그림 2]는 문자열로부터 명사를 추출하는 파이썬 프로그램의 명령어 이다. 입력데이터를 아래와 같이 설정하고, 출력데이터를 확인하였다.

① 명사 추출 입력 데이터
영희와 철수는 백구를 산책시키기 위해 한강에 갔다. 한강에 도착해 누렁이를 만났다.

② 명사 추출 수행결과
영희, 철수, 백구, 산책, 위해, 한강, 한강, 도착, 누렁이

```
# 명사 추출 소스코드

from konlpy.tag import Twitter
from konlpy.utils import pprint

twitter = Twitter()

sentence=u'영희와 철수는 백구를 산책시키기 위해 한강에 갔다. 한강에 도착하여 누렁이를 만났다.'

print('명사 : '+ str(twitter.nouns(sentence)))
```

그림 2. 명사 추출 실행 코드

어떠한 문서의 집합이 주어졌을 때 해당 집합 내의 어떤 문서에서든 보편적으로 자주 등장하는 단어들이 등장한다. 이러한 단어들 가운데 특히 문서 집합의 성격과 관계없이 해당 단어가 문장 내에서 가지는 기능으로 인해 자주 등장하는 단어들이 있는데 이러한 단어들을 불용어(stopword)라고 부른다. 이러한 불용어들은 대부분 그 자체로는 큰 의미가 없지만 문법적으로는 중요한 기능을 하는 단어들이다. 영어의 예를 들었을 때, 전치사(to, for, with, as 등), 접속사(and, but, so 등), 접속부사(however, therefore 등), 대명사(i, you, she, we, it 등), 관사(a, an, the), 관계대명사(which, who, that 등) 등이 이러한 경우에 속한다.

단순히 단어의 빈도수가 높은 단어를 문서에 대한 특정 값으로 삼는 모델의 경우 실제 의미를 반영한다고 보기 어려운 불용어에 대한 특정 값이 크게 부각될 수 있어 실제 문서가 가지는 중요한 특성을 희석시킬 수 있다는 문제가 있다. 이러한 문제를 회피하기 위해서 실제 문서에 대한 빈도수 기반의 모델을 구축할 때는 불용어를 문서에 대한 특징에서 제외시키는 것이 일반적이다.

3. 단어 벡터화

word2vec 모델은 단어를 수십~수백 차원의 벡터로 변환하여 단어의 의미를 효율적으로 추정하는 방법으로 자연어 처리 분야에서 비약적인 정밀도 향상을 가능하게 하였다. word2vec 모델은 인공 신경망을 기반으로 둔 방식으로 같은 맥락(context)에 있는 단어는 가까운 의미를 가진다는 전제에서 시작한다. word2vec 모델은 텍스트 문서를 통해 학습을 진행하며 문장 내에 한 단어와 같이 출현하는 다른 단어들을 관련 단어로써 인공 신경망에 학습시킨다. 연관된 단어들은 문서상에서 가까운 곳에 출현할 가능성이 높아지기 때문에 학습을 반복해 나가는 과정에서 주변 단어가 비슷한 두 단어는 가까운 벡터 공간에 놓이게 된다.

word2vec 모델은 단순하게 한 단어의 앞뒤로 서로 같은 정보가 있는지 없는지를 이용하여 학습하는 것이다. 따라서 아주 추상적인 동사나 형용사는 학습이 명사에 비해서 학습이 어려울 수 있다. 다만 수없이 많은 데이터를 보면 동사들이 어떤 목적어를 가지는지 규칙성을 파악함으로 어느 정도 동사들 간의 의미 관계도 학습이 가능하다고 볼 수 있다. 예를 들어, break, broken은 서로 비슷한 목적어를 가질 것이므로 두 동사는 비슷한 의미를 취할 것이라고 학습할 수 있을 것

이다. 또한, 충분히 많은 학습이 이루어지게 되면 break, broken의 벡터 공간에서의 거리가 have와 had의 벡터 공간에서 거리와 같아 질 수 있다. 이는 과거의 의미를 학습할 수 있다는 것이다. word2vec 모델은 심층 신경망(DNN: Deep Neural Network)이 아니다. 활성화 함수가 적용되지 않은 은닉층 1개와 softmax function이 적용된 출력층으로 구성된 인공 신경망이다. 그래서 일반적인 심층 신경망보다 학습속도가 굉장히 빨라서 매우 큰 데이터도 손쉽게 학습시킬 수 있다는 것이 큰 장점이다. word2vec 모델의 알고리즘은 내부적으로 [그림 3]과 같이 하나의 맥락으로 단어를 예측하는 CBOW(Continuous Bag Of Words)와 단어로 맥락을 예측하는 SG(Skip-Gram)라는 두 개의 신경망 모델을 이용해 문장을 학습하여 비슷한 의미의 단어들을 가까운 벡터 공간에 표현한다.

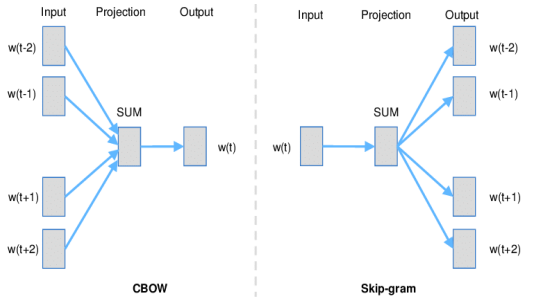

그림 3. word2vec 모델의 알고리즘

예를 들어, [그림 3]의 skip-gram은 w(t)가 입력 단어로 주어졌을 때, 입력 단어를 기준으로 지정된 윈도우 사이즈에 따라 앞, 뒤로 일정한 개수의 다른 단어에 대한 예측을 수행하는 것을 목표로 신경망을 훈련시킨다. OUTPUT은 입력 단어 w(t)를 기준으로 주변에 올 수 있는 단어 w(t-2), w(t-1), w(t+1), w(t+2)를 예측하는데 계산되는 가중치 값으로써, 가중치 값들이 w(t)를 나타내는 벡터 값이 된다. 실제로 연구자가 지정한 윈도우 사이즈로 OUTPUT 단어의 범위를 지정할 수 있다. word2vec 모델은 입력단어가 주어졌을 때, 출력단어의 조건부 확률인 softmax function을 사용하여 결과 값이 최대가 되도록 학습하는 것이다. 식 (1)은 word2vec 모델의 softmax function이다. $P(W_0|W_1)$은 입력단어 W_1가 주어졌을 때 출력 단어로 W_0가 나올 조건부 확률이다. 이에 따라 word2vec 모델에서 학습 문서 내 주위 단어의 분포가 가까운 단어일수록 산출되는 벡터 값이 유사해지며, 산출된 벡터 값이 비슷한 단어는 유사한 것으로 간주된다.

$$P(W_o|W_1) = \frac{\exp(W_1 \cdot W_o^T)}{\sum_{j=1}^{V} \exp(W_1 \cdot W_j^T)} \quad (1)$$

Ⅳ. 회의자료 분석 시스템 구현 및 평가

본 논문에서는 회의록 문서를 형태소로 분석하고 이를 토대로 word2vec 모델 skip-gram 알고리즘을 이용하여 학습한 데이터를 활용해 회의록 문서에서 대표의견을 찾아낸다. 실험을 위해 윈도우 10과 리눅스 환경에서 C 언어 기반으로 개발된 파이썬 언어를 사용하였으며, 파이썬 3.6.6 버전을 사용하였다. 원활한 실험을 위해 파이썬 라이브러리를 사용하였다. 형태소 분석을 위해서 KoNLPy패키지 twitter 형태소 분석기를 사용하였고, Gensim 패키지의 word2vec 모델을 사용하였다[16][17].

시스템에 사용된 데이터는 아이들이 행복한 지역교육환경을 위해 우리가 할 수 있는 일은(충북교육청), 소통과 협력을 위한 관행문화개선(경기교육청)의 회의록 데이터 외 14건을 사용했다. 처리 순서는 수집된 문서를 자연어 처리를 실시하여 형태소 형태로 표현한다. 이후 최대빈도단어 수집과 불용어 제거 작업까지 진행하도록 한다. 형태소 분석은 파이썬 환경에서 지원하는 형태소 분석기 패키지 'KoNLPy'를 사용했다. KoNLPy에서 지원하는 형태소 분석 및 품사 태깅 클래스는 Hannanum, Kkma, Komoran, Mecab, Twitter 형태소 분석기를 지원한다. 본 연구에서는 5가지 형태소 분석기 중에서 가장 좋은 성능을 보이던 twitter 형태소 분석기를 사용하였다. twitter 형태소 분석기를 이용해 데이터 처리 후 [그림 4]처럼 형태소를 확인 할 수 있다.

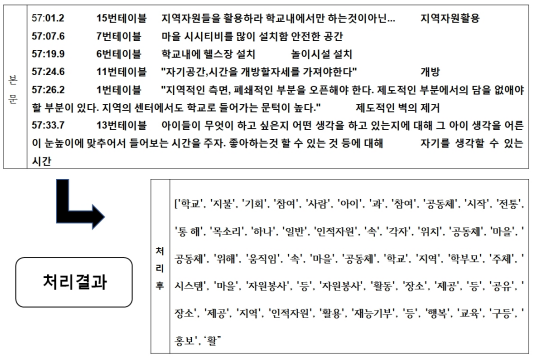

그림 4. 형태소 분석 데이터

회의록 문서 형태소 분석 이후 원활한 word2vec 모델 학습의 진행과 최종 결과 값의 신뢰도를 높이기 위해 불용어 사전을 기반으로 불필요한 단어를 제거한다. 불용어란 실제 단어는 아니지만 단어처럼 인식되는 요소들을 말한다. 예를 들어, '는', '을', '등', 등과 같은 조사들을 말한다. 조사가 단어로 인식되어도 크게 관계는 없지만 본 시스템에서는 빈도수가 높은 단어를 참조하여 활용하기 때문에 조사 불용어를 제거하도록 하고 처리된 데이터는 명사의 집합으로 구성하게 한다. 불용어를 제거 도구로는 파이썬 프로그램 NLTK 패키지의 stopword를 사용하였다. NLTK(Natural Language Toolkit) 패키지는 교육용으로 개발된 자연어 처리 및 문서 분석용 파이썬 패키지이다. 간단하게 인간의 언어인 자연어를 컴퓨터를 통해 처리하고 이용하려는 프로그램이라고 볼 수 있다. NLTK 패기지의 stopword를 이용해 불용어 처리작업 이후 [그림 5]과 같이 하나의 단어로 인식되었던 조사와 불용어가 삭제된 것을 확인할 수 있다.

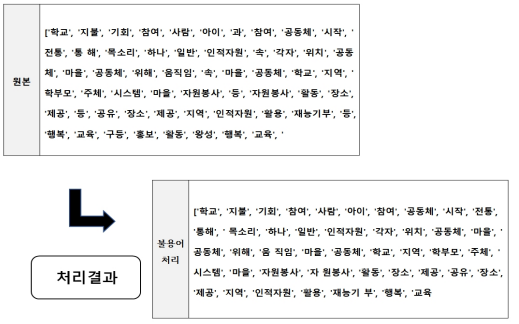

그림 5. NLTK패키지 stopword사용 불용어 처리

회의록 문서를 자연어 처리하는 과정으로 형태소 분석과 불용어를 제거한 후 word2vec 모델에서 학습을 진행하기 전에 출현 빈도가 높은 순으로 단어를 정렬하여 시각화한 결과이다. [그림 6]은 단어 출연 빈도를 나타낸다. 토론의 주제가 "아이들이 행복한 지역교육환경을 위해 우리가 할 수 있는 일은"이기 때문에 "아이",

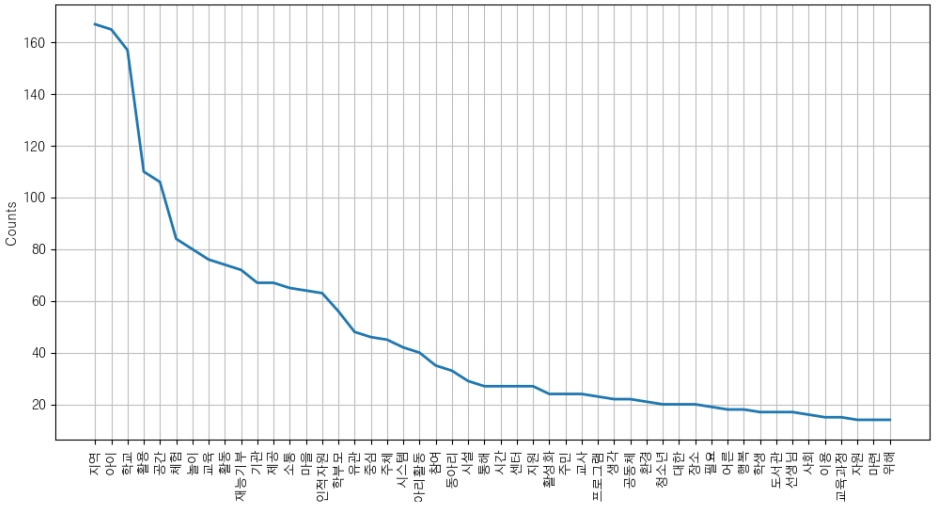

그림 6. 단어 출현 빈도 확인

"학교", "공간", "놀이", "활용" 등 주제와 연관성이 있다고 생각되는 단어들이 높은 순위에 올라 있는 것을 확인할 수 있다.

word2vec 모델의 학습 환경 구성을 위해 회의록에서 언급된 단어 상위 5개를 지정하여 확인했다. [표 1]은 형태소 분석을 통한 단어 빈도 및 상위 단어를 나타낸다. 주요 문장을 제안받는 방법으로는 상위 키워드만을 사용해도 가능하다. 그 이유는 형태소 분류를 통해 언급된 키워드는 문서에서 사용 빈도가 높은 순으로 정리되어 있고 다수의 문장들은 해당 단어를 포함한다고 볼 수 있기 때문이다. 본 연구에서는 대표 문장을 추출받기 위해 추가로 word2vec 모델로 학습하여 추출한 유사 단어 정보를 사용하기 때문에 다음으로 학습 작업을 진행한다.

표 1. 형태소 분석을 통한 단어빈도 및 상위단어

용어	지역	아이	학교	활용	공간
빈도수	167	165	157	110	106
순위	1	2	3	4	5
점유(%)	17.6	17.4	16.5	11.6	11.2

문서를 대표하는 문장을 추출하기 위해 word2vec 모델 패키지를 활용하여 중요 키워드 유사 단어 학습 과정을 진행한다. word2vec 모델의 알고리즘은 CBOW와 Skip-gram이라는 2가지 알고리즘으로 나눠진다. 두 알고리즘 모두 결과적으로 단어를 벡터화한다는 점은 동일하다. 하지만 예측하는 방법이 각 알고리즘별로 달라 경우에 따라 필요한 알고리즘을 사용하면 된다.

각 모델의 학습방법의 예를 비교하면 CBOW는 알고자 하는 주변 단어를 통해 해당 단어를 예측하여 단어를 벡터화하는 방식이지만 Skip-gram의 경우는 반대로 중심 단어의 주변에 어떤 단어가 올지 예측하는 방식이다. CBOW는 데이터의 양이 적을 때 적합하고 Skip-gram은 데이터의 양이 많을 때 적합하다고 알려져 있다. 하지만 말뭉치의 크기가 동일하더라도 한 번의 수행으로 학습을 진행하는 CBOW에 비해 여러 번 수행으로 주변에 올 단어를 예측하는 skip-gram의 데이터의 학습량이 보다 많아진다는 장점이 있다[8]. 본 연구에서는 빈도수가 높은 키워드를 확인하여 주요 문장을 찾아야하므로 skip-gram을 활용하여 데이터를 학습한다.

실험에 사용된 데이터는 충북교육청에서 "아이들이 행복한 지역 교육 환경을 위해 우리가 할 수 있는 일은"이란 주제로 진행된 회의록 데이터이다. 자연어 처리과정을 마친 문서데이터를 이용하여 word2vec 모델에 학습시키는 과정에서 사용된 parameter는 [그림 7]과 같다. word2vec 모델 임베딩 실행 환경을 300차원의 벡터 공간에서 실시한다. 또한, 문서 내에서 최소 10회 이상의 출현 횟수를 가진 단어만 추출하여 사용하며 데이터 학습 횟수는 2,000을 반복하도록 지정했다. 마지막으로 학습 방법은 skip-gram을 사용하도록 환경을 구성했다.

```
#word2vector 실행
embedding_model = word2vec.Word2Vec(tokenized_contents,
size=300, # 300차원 짜리 벡터 스페이스
 window = 5, # 분석할 단어 범위
 min_count=10, # 등장 횟수가 10미만인 단어는 무시
 workers=4, # CPU 사용조건
 batch_words= 1000, # 사전 구축 시 한번에 읽는 단어수
 iter=2000, # 학습 반복횟수
 sg=1) # 0이면 CBOW, 1이면 Skip-gram
```

그림 7. word2vec 모델 Parameter 설정화면

word2vec 모델은 중심 단어와 주변 단어 벡터의 내적이 코사인 유사도가 되도록 단어벡터를 벡터공간에 임베딩 한다. skip-gram을 이용한 word2vec 모델 학습 결과 시각화는 [그림 8]과 같다. 문서 내에서 많은 언급이 되고 중요 키워드끼리 뭉치가 형성되어 있다. 학습 데이터양이 많지 않지만 학습 반복 횟수를 늘려 의미가 비슷한 단어끼리 말뭉치가 형성된 것을 확인할 수 있다.

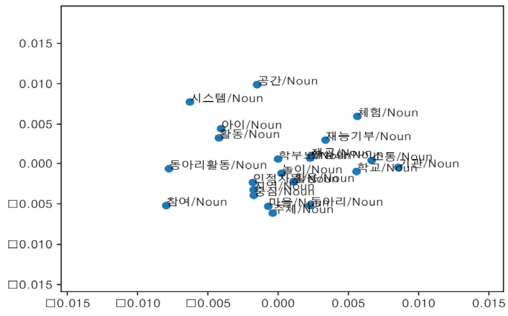

그림 8. word2vec 모델 학습결과 시각화

본 논문에서 제안하는 요약시스템은 [그림 9]와 같은 형태를 가진다. 회의록에서 의견을 대표하는 문장을 추출하는 과정은 형태소 분석 데이터에서 최대빈도 단어와 word2vec 모델 학습 데이터에서 제안하는 유사 단어를 조합하여 전체 회의록 문서에서 대표 문장을 검색하도록 한다.

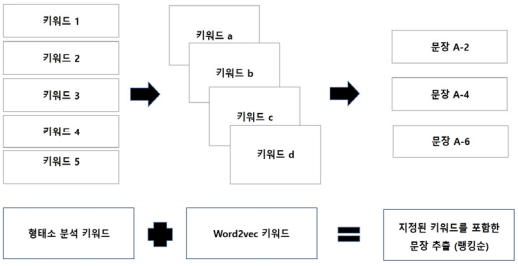

그림 9. 키워드 조합을 통한 대표 문장 제시

문서에서 공통적인 의견을 가진 대표 문장을 검색하기 위해 상위 키워드와 가장 유사한 단어를 word2vec 모델 학습 데이터를 이용한다. 결과 데이터는 [표 2]와 같다. 상위 키워드로 지정된 단어에 유사 키워드 벡터 값이 높을수록 두 단어간 의미가 비슷하며 의미가 비슷한 단어를 포함한 문장을 검색해 대표 문장으로 표현할 수 있다.

표 2. 상위빈도 키워드와 word2vec 모델 학습을 통한 유사 키워드

단어	키워드 1	키워드 2	키워드 3	키워드 4	키워드 5
지역	놀이	동아리활동	참여	활동	시스템
	0.114590942	0.113554440	0.064748942	0.053151577	0.049939170
아이	학교	시스템	활동	제공	재능기부
	0.126395180	0.116534307	0.091102123	0.050185315	0.047024309
학교	아이	제공	놀이	주체	인적자원
	0.126395180	0.104971289	0.092588156	0.081538178	0.071517258
활용	마을	동아리	중심	인적자원	체험
	0.082864031	0.061420358	0.061224631	0.053702723	0.052903652
공간	중심	주체	교육	마을	체험
	0.119040235	0.108881503	0.075241833	0.068808533	0.028565594

실험에서 제안하는 데이터의 형태소 분석을 통한 키워드와 word2vec 모델 학습 후 유사도 분석을 통한 단어를 조합하여 1개의 대표 키워드와 5개의 유사 단어를 이용하여 1개의 키워드당 3개의 문장을 추출하는 과정을 진행하였다. 추출된 데이터의 신뢰도를 확인하기 위해 기존 퍼실리테이터에 의해 수작업으로 분류된 결과와 본 시스템이 제안하는 결과를 비교 평가하였다. [표 3]은 제안 기법과 퍼실리테이션에 의한 제안 기법을 통해 비교 평가를 수행한 결과를 나타낸다. 퍼실리테이터에 의한 결과와 시스템이 제안하는 값이 완벽하게 일치 하진 않지만 유사한 점을 찾아볼 수 있다. 이와 같은 시스템의 장점은 빠른 시간에 회의록을 요약할 수 있다는 장점이 있고 시간을 절약함으로 비용을 절감할 수 있게 된다.

표 3. 시스템 제안결과 비교확인

순위	퍼실리테이션에 의한 제안	회의록 요약시스템 제안
1	체험중심의 동아리 활동	지역자원을 활용하고 인근학교와 연계한 동아리 활동을 추진하면 좋겠다.
2	아이들의 욕구를 표현할 수 있는 공감교육 필요	개인 사업장 활용에 대한 홍보차원의 기부나 바우처 지원을 통해 경제적 부담을 줄여 활용 할 수 있는 방안검토
3	지역시설(학교, 도서관) 개방	기관 내 놀이공간 확보
4	지역인적자원활용(재능기부등)	방과후 강사등에서 지역인적 자원 활용재료비등 지원하에 인적자원 활용
5	문제아동에게 관심 갖고 관계기관 연결해주기	지역인적자원을 활용한 과정을 교육활동으로만 보지 말고 놀이로써 볼 수 있도록 하자.
6	학부모 동아리 참여 기회 지원	놀이공간 확충
7	방과후 프로그램 지역사회(도서관)와 연계하여 운영	학교의 공적 기득권 내려놓고 지역과 함께 공감하고 소통하기.
8	유관기관과학교와의소통	도교육청과 지역청(학교)소통협의체 부채
9	마을 환경 체험하기	마을지도, 마을교과서 역사나 지리등 학부모와 협업해서 만들기 교재의 지역화 학교를 알리고 지역을 알리는 것이다.
10	아이들에게안전한놀이공간제공	지역주변에 있는 장소들을 활용을 해서 체험확대장거리가 아닌 지역체험을 활용
11	주민및 아이들에게 관심갖고 서로의 활동지원하기	지역 인적 자원을 하게 되면 재정적으로 지원이 가능한 기관에서 도움주기
12	아이들이 좋아하는 것을 할 수 있는 시간과 정보 제공하기	지역과 기관의 자율적 움직임중요, 재능기부나 가가자 개인의 체험활동비를 소액으로 책정하여 모든 학생들에게 부담없이, 또는 무상으로 프로그램에 참여하도록 한다.
13	등하교시 안전한 환경만들기	어린이를 위한 사회공헌사업
14	교육청의 컨트롤타워 역할	기관 내 놀이공간 확보
15	올바른 가정교육	경험을 통해 참여 할 수 있는 기회의 장열기

V. 결론

본 논문에서는 단어빈도 분석 뿐 아니라 단어 간의

유사도 분석을 통해 문장을 분류하여 더욱 효과적인 회의록 요약시스템을 설계하고 구현하였다. 이를 위해 교육청에서 진행한 토론에서 자유롭게 발언한 데이터를 기반으로 인공 신경망을 이용하여 문장을 제안 받는 방법에 대해 다루었으며 단어들을 수치화하여 벡터로 표현하는 word2vec 모델을 이용하여 주요 단어의 유사 키워드를 도출하였다. 제안 시스템을 토대로 회의록에서 대표 문장을 추출하는 서비스가 구현 가능하며 실제로 퍼실리테이터에 의해 수동으로 결정된 문장과 시스템에서 추출하는 문장의 내용이 약60%의 일치율을 보였다. 그러나 word2vec 모델로 학습을 진행하는데 있어 사용한 데이터의 양이 부족하다고 판단되었다. 보다 효과적인 결과도출을 위해서는 데이터를 추가 확보하여 분석 데이터를 늘려 학습 환경을 변경하는 작업이 요구된다. 짧은 문장 위주의 회의록을 분석하여 연구결과를 도출하였다. 향후 연구로 대용량의 문서와 긴 문장 회의록 데이터를 활용하여 다양한 실험 평가를 수행할 예정이다. 또한, 결과의 정확성을 보다 향상시키기 위한 연구를 진행할 예정이다.

참 고 문 헌

[1] H. Liu, X. Wang, Y. Wei, W. Shao, J. Liono, F. D. Salim, B. Deng, and J. Du, "ProMETheus: An Intelligent Mobile Voice Meeting Minutes System," Proc. International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, pp.392-401, 2018.

[2] H. Miura, Y. Takegawa, A. Terai, and K. Hirata, "Interactive Minutes Generation System Based on Hierarchical Discussion Structure," Proc. IEEE/WIC/ACM International Conference on Web Intelligence, pp.459-465, 2018.

[3] Z. Zhao, H. Pan, C. Fan, Y. Liu, L. Li, and M. Yang, "Abstractive Meeting Summarization via Hierarchical Adaptive Segmental Network Learning," Proc. The World Wide Web Conference, pp.3455-3461, 2019.

[4] T. Huang, C. Hsieh, and H. Wang, "Automatic meeting summarization and topic detection system," Data Technologies and Applications, Vol.52, No.3, pp.351-365, 2018.

[5] 이병수, *어휘의 동시 발생 빈도와 분포를 이용한 다중 주제 회의록 요약,* 성균관대학교, 석사학위논문, 2015.

[6] 김선공, *word2vec모델과 RNN을 이용한 영화 리뷰의 감성분석,* 동국대학교, 석사학위논문, 2016.

[7] 노현아, *단어 빈도 가중치를 이용한 자동 문서 분류,* 전남대학교, 석사학위논문, 2013.

[8] 백민지, *word2vec모델 학습을 통한 의미 기반 해외 유사 특허 검색 방안,* 국민대학교, 석사학위논문, 2017.

[9] 김정미, 이주홍, "word2vec모델을 활용한 RNN기반의 문서 분류에 관한 연구," 한국지능시스템학회 논문지, 제27권, 제6호, pp.560-565, 2017.

[10] Y. kim, "Convolutional Neural Network for Sentence Classification," Proc. Conference on Empirical Method in National Language Processing, pp.1746-1751, 2014.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," Proc. Annual Conference on Neural Information Processing Systems, pp.3111-3119, 2013.

[12] 김성민, *단어 벡터화를 통한 특징 단어 기반 문서 관련성 분석 방법 영화 스크립트 비교를 중심으로,* 건국대학교, 석사학위논문, 2016.

[13] G. Yoav and O. Levy, "word2vec Explained: deriving Mikolov etal's negative-sampling word-embedding method," CoRR abs/1402.3722, 2014.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Proc. International Conference on Learning Representations Workshop, 2013.

[15] 김도우, 구명완, "Doc2Vec과 Word2Vec을 활용한 Convolutional Neural Network 기반 한국어 신문기사 분류," 정보과학회논문지, 제44권, 제7호, pp.742-747, 2017.

[16] Python 형태소 분석기 Konlpy, http://konlpy-ko.readthedocs.io/ko/v0.43/api/konlpy.tag/

[17] word2vec 모델, https://radimrehurek.com/gensim/models/word2vec.html

저 자 소 개

허 강 호(Kangho Heo) 정회원

- 2009년 2월 : 충북대학교 구조시스템학과(이학사)
- 2019년 2월 : 충북대학교 전자정보공학과(공학석사)
- 2010년 2월 ~ 2015 12월 : 엘지유플러스 과장
- 2016년 1월 ~ 현재 : ㈜바론 부장

〈관심분야〉 : 데이터베이스 시스템, 빅데이터 처리 등

양 진 우(Jinwoo Yang) 정회원

- 2015년 2월 : 서원대학교 정보통신공학과(공학사)
- 2015년 7월 : 와이즈멘토리더십센터 강사
- 2016년 2월 ~ 현재 : 주식회사 바론 대리
- 2018년 3월 ~ 현재 : 충북대학교 전자정보공학과 공학석사 과정

〈관심분야〉 : 데이터베이스 시스템, 소셜 네트워크 서비스, 빅데이터 처리 등

김 동 현(Donghyun Kim) 정회원

- 2002년 2월 : 충북대학교 구조시스템공학과(공학사)
- 2006년 2월 : 충북대학교 구조시스템공학과(공학석사)
- 2009년 8월 : 충북대학교 구조시스템공학과(공학박사수료)
- 2017년 2월 ~ 현재 : 충북대학교 빅데이터학과 박사과정
- 2003년 ~ 2005년 : 서울대학교 선진화 연구단 실무담당
- 2006년 ~ 2007년 : 교차로, 내일신문 교육 칼럼리스트
- 2006년 ~ 2011년 : ㈜바른교육 창업콘텐츠 연구소장
- 2016년 ~ 현재 : ㈜바론 대표이사

〈관심분야〉 : 데이터베이스 시스템, 이동 P2P 네트워크, 소셜 네트워크 서비스, 빅데이터 처리 등

복 경 수(Kyoungsoo Bok) 종신회원

- 1998년 2월 : 충북대학교 수학과(이학사)
- 2000년 2월 : 충북대학교 정보통신공학과(공학석사)
- 2005년 8월 : 충북대학교 정보통신공학과(공학박사)
- 2005년 3월 ~ 2008년 2월 : 한국과학기술원 정보전자연구소 Postdoc
- 2008년 3월 ~ 2011년 2월 : 가인정보기술 연구소 차장
- 2011년 3월 ~ 2019년 8월 : 충북대학교 정보통신공학부 초빙교수
- 2019년 9월 ~ 현재 : 원광대학교 SW융합학과 조교수

〈관심분야〉 : 데이터베이스 시스템, 이동 객체 데이터베이스, 이동 P2P 네트워크, 소셜 네트워크 서비스, 소셜 IoT, 빅데이터 처리 등

유 재 수(Jaesoo Yoo) 종신회원

- 1989년 2월 : 전북대학교 컴퓨터공학과(공학사)
- 1991년 2월 : 한국과학기술원 전산학과(공학석사)
- 1995년 2월 : 한국과학기술원 전산학과(공학박사)
- 1995년 2월 ~ 1996년 8월 : 목포대학교 전산통계학과 전임강사
- 1996년 8월 ~ 현재 : 충북대학교 정보통신공학부 정교수

〈관심분야〉 : 데이터베이스 시스템, 멀티미디어 데이터베이스, 센서 네트워크, 바이오 인포메틱스, 소셜 네트워크, 사물인터넷, 빅데이터 등