https://doi.org/10.14400/JDC.2017.15.12.341

VAE를 이용한 의미적 연결 관계 기반 다중 문서 요약 기법

백수진
용인송담대학교 정보통신학과

Multi-Document Summarization Method Based on Semantic Relationship using VAE

Su-Jin Baek
Dept. of Information Communication, Yong-In Songdam College

요 약 많은 양의 문서 데이터가 증가됨에 따라 사용자는 해당 문서를 이해하기 위한 요약된 정보를 필요로 한다. 그러나, 기존 문서 요약 연구 방법들은 지나치게 단순한 통계에 의존함으로써 문장의 모호성 및 의미 있는 문장 생성을 위한 다중 문서 요약 연구가 미흡한 실정이다. 본 논문에서는 의미적 연결 관계에 대한 파악 및 불필요한 정보를 처리하기 위한 전처리 과정을 거치며, 어휘 의미 패턴 정보를 기반으로 VAE를 이용하여 문장 간의 의미적 연결성을 높인 다중 문서 요약 기법을 제안하였다. 문장을 이루고 있는 단어 벡터들을 이용하여, 잠재된 변수로 생성된 압축된 정보와 속성 판별기로부터 학습을 한 후 문장을 재구성함으로써 의미적 연결 처리가 자연스러운 요약문을 생성하였다. 제안된 방법과 다른 문서 요약 방법을 비교했을 시 미세하지만 더 향상된 성능을 나타냈으며, 이는 의미적 문장 생성 및 연결성을 높일 수 있음을 증명하였다. 앞으로, 다양한 속성 설정 값을 가지고 실험하여 의미적 연결 관계를 확장할 수 있는 방법을 연구하고자 한다.

주제어 : VAE, 다중 문서 요약, 자연어 처리, 딥러닝, 의미적 연결 관계

Abstract As the amount of document data increases, the user needs summarized information to understand the document. However, existing document summary research methods rely on overly simple statistics, so there is insufficient research on multiple document summaries for ambiguity of sentences and meaningful sentence generation. In this paper, we investigate semantic connection and preprocessing process to process unnecessary information. Based on the vocabulary semantic pattern information, we propose a multi-document summarization method that enhances semantic connectivity between sentences using VAE. Using sentence word vectors, we reconstruct sentences after learning from compressed information and attribute discriminators generated as latent variables, and semantic connection processing generates a natural summary sentence. Comparing the proposed method with other document summarization methods showed a fine but improved performance, which proved that semantic sentence generation and connectivity can be increased. In the future, we will study how to extend semantic connections by experimenting with various attribute settings.

Key Words : VAE, Multi-Document Summarization, Natural Language Processing, Deep Learning, Semantic Relationship

Received 2 November 2017, Revised 1 December 2017
Accepted 20 December 2017, Published 28 December 2017
Corresponding Author: Su-Jin Baek(Yong-In Songdam College)
Email: croso79@ysc.ac.kr

ISSN: 1738-1916

1. 서론

소셜 네트워크 서비스의 확산과 더불어 정보의 생성 및 공유가 활발히 이루어지고 있다[1]. 그러나 폭발적으로 늘어나고 있는 데이터의 양에 비해서 대부분의 데이터들은 잘 정제되어 있지 않으므로 간결하게 요약된 정보는 찾기 어려워졌다. 이에 따라 내용만을 자동적으로 파악하고 중복적인 정보를 제거하는 다중 문서 요약 기법의 필요성이 대두되고 있으며 중요성 또한 점점 커지고 있다. 다중 문서 요약이란 복수 개의 문헌으로부터 중심적인 내용만을 추려 하나의 요약문을 자동으로 요약하는 기법이다[2]. 이는 원문의 주제를 서술적으로 표현해야 하므로 복잡한 언어처리와 주제 분석 같은 고차원적인 문서 분석 기술을 필요로 한다.

문서의 요약 연구는 내용의 구성 방법에 따라 생성요약과 추출 요약으로 구분할 수 있다[3]. 생성 요약 기법은 원문으로부터 중요한 단어들을 선별한다. 그리고 자연어 처리 기법을 이용하여 새로운 문장을 구성을 하고 이를 요약문으로 제공하는 것이다. 기존의 생성 요약 기법들은 지나치게 단순한 통계에 의존함으로써 문서 주제 파악에 한계를 보이거나, 문장의 수사관계, 문장의 모호성 등을 보임으로써 해결해야 할 문제들을 많이 남아 있다. 추출 요약 기법의 경우 입력 텍스트로부터 의미 있는 문자들을 추출하여 요약문이 구성된다. 기존의 추출 요약 기법들은 대량의 데이터 처리 시 높은 정확도와 효율성이 입증된 딥 러닝 알고리즘을 활발히 사용하였으나 다중 문서 요약 연구에서는 이러한 경향이 미흡한 수준으로 반영되었다.

기존의 다중 문서 요약에 관한 연구들로는 TextRank 모델, 토픽 모델, 잠재 의미 분석 등을 통해 문서를 요약하고자 하는 방법들이 있다. TextRank 모델을 이용한 방법은 문서를 그래프로 표현하고, 그래프의 형태를 이용해 각 노드의 중요도를 결정하여, 중요한 문장을 선정하여 요약문을 생성한다[4]. 토픽 모델을 이용한 방법으로는 단어와 문장 매트릭스를 함께 활용하여 토픽 모델링을 수행하여 각 토픽에서 가장 확률 값이 높게 나타난 문장을 가지고 요약문을 작성한다[5]. 잠재 의미 분석((LAS : latent semantic analysis)은 통계적인 학습 방법으로 대상의 잠재된 의미를 파악하기 위한 기법으로 중요 문장을 추출한다[6]. 문서를 그래프로 표현하고 클러스터링 알고리즘을 통해 문장을 각 주제 클러스터로 할당한 후 각각의 주제 클러스터로부터 점수를 산출하여 이를 바탕으로 중요 문장과 주제어를 추출한다. 그 밖의 다중 문서 요약 연구로는 질의 기반 다중 문서 요약 모형에 딥 러닝 알고리즘을 접목한 방법[7]과 인공 신경망 분야 연구에서 순차적인 데이터를 학습하는데 범용적으로 사용되고 있는 RNN(recursive neural network)를 딥 러닝 알고리즘을 사용한 방법도 있다[8]. 그러나, 기존 연구로는 딥 러닝 알고리즘을 연계하여 문서 요약 모형을 개발한 연구는 많지 않으며, 현재 빅데이터 시대에 부합하는 다중 문서 요약 연구도 부재한 실정이다.

본 논문에서는 요약문의 더 긴밀한 관계 정보를 파악하고, 불필요한 정보 처리를 위한 전처리 과정을 한다. 이러한 정보를 바탕으로 의미 패턴 정보와 VAE의 잠재된 변수로 생성된 압축된 정보 및 속성 판별기로부터 학습한 정보를 기반으로 문장을 재구성함으로써 문장빈도수와 같은 단순한 통계에 의존하지 않고, 문장의 모호성을 해결하도록 문장 간의 의미적 연결성을 높인 다중 분석 요약 기법을 제안한다.

2. 관련연구

2.1 기존의 다중 문서 요약 연구

문서 요약은 요약문의 형태를 기준으로 생성 요약과 추출 요약으로 구분할 수 있다. 생성 요약은 추출한 문장들을 수정하거나 새로운 문장을 생성하여 요약문을 만들기 때문에 추출 요약에 비하여 요약문의 응집도나 가독성이 뛰어나다. 그러나, 문장을 새로 생성하는 자연어 처리 기술이 아직 완벽하게 구현되지 않기 때문에 대부분의 문서 요약 연구는 추출 요약을 기반으로 수행된다. 추출 요약은 입력된 텍스트로부터 의미 있는 문장들을 추출하여 요약하는 방법으로 현재 다양한 연구를 통해 문장이 가지는 단어의 빈도수 및 가중치를 통해 문장과 단어 간의 관계를 분석하여 중요 문장을 추출하는 방식으로 이루어지고 있다. 추출 요약은 구현하기 쉬우나 자동으로 분석된 문장의 가중치가 기존의 문서의 의미전달이 제대로 이루어지지 않는 경우 올바른 문장 요약이 이루어지지 않아 가독성이 떨어지는 단점이 있다[9].

기존의 다중 문서 요약 연구에는 TextRank를 사용하

여 문서를 그래프로 표현할 때 단어의 위치 정보나 더불어 소셜 폭소노미를 이용한 단어 간의 의미적 연관성을 고려하는 다중 문서 요약 기법이 있다. 이는 TextRank의 한계점을 보완하기 위한 것으로, 기존의 TextRank이 오직 문서상의 위치만을 고려하여 특정 노드의 중요도를 계산하기 때문에 노드 간의 의미적 유사성을 반영하지 못한다는 점에 착안하여 기법을 설계하였다[4]. 토픽 모델링을 이용한 기존 연구로는 베이지안 문장 기반 토픽 모델링을 통한 문서 요약이 있다[5]. 단어 문서 매트릭스와 가장 확률 값이 높게 나타난 문장을 가지고 요약문을 작성한 방법이다. 잠재적 의미 분석(LAS)를 활용한 다중 문서 요약 연구가 있다. 그 중 PLSA(Probavilistic Latent Semantic Analysis)알고리즘을 이용한 문서 요약 기법은 문서 내의 존재하는 단어들에 포함된 숨겨진 주제들 별로 클러스터들을 만들고 클러스터 된 단어와 문장과의 관계를 유사도 측정을 수행하여 문장에 점수를 부여하는 기법을 제안하였다[6]. 그리고 기존의 PLSA 알고리즘을 이용한 문서 요약 연구를 개선시키기 위해 클러스터링 알고리즘을 이용하여 문서를 요약한 기법도 있다. 이러한 기법들은 비교적 우수한 결과 값을 얻을 수 있지만 전처리 단계에서 사용되는 클러스터링 알고리즘과 분석단계에서 기계학습 알고리즘을 이용하여 복잡한 분석을 수행하기 때문에 높은 계산 비용과 분석 시간이 요구된다[10].

또 다른 방법으로는 문서의 중요 문장을 효율적으로 추출하기 위해 질의에 대한 확장과 분해 방법으로 문서를 요약한 연구가 있다[7]. 이러한 방법은 먼저 입력 문서들로부터 중요한 개념을 추출하고 요약문을 생성한 후 요약문을 다시 재수정하는 형태로 문서 요약을 수행하였다. 이때 딥 러닝 알고리즘은 입력 문서를 심층 구조의 그래프로 표현할 수 있게 하였으며, 내재된 계층에 분포한 정보까지 활용하여 요약문을 생성함으로서 정보량의 손실을 방지하는 역할을 하였다. RNN 모델을 이용한 방법은 입력 문서 데이터 셋을 대상으로 계층적 회귀 분석을 수행함으로써 문장들의 순위를 매기고 높은 순위를 차지한 문장들을 배열하여 요약문을 생성한다[8]. 그러나, 기존 연구로는 딥 러닝 알고리즘을 연계하여 문서 요약 모형을 개발한 연구는 많지 않으며, 현재 빅데이터 시대에 부합하는 다중 문서 요약 연구도 부재한 실정이다. 특히, 문서의 요약은 요약문 내의 문장들끼리 서로 연관성이 높아야하고 의미, 개념적 연결성이 높은 글이 되어야 함으로 추출 정보를 기반으로 하여 빈도수만을 고려하지 않고 문장 간의 의미적 연결성을 높인 다중 문서 요약 기법이 필요하다.

2.2 VAE

VAE(variational autoencoder) 는 단순히 입력 벡터를 출력에서 재구성하는 autoencoder 와는 다르게, 입력과 출력 사이에 변환 과정을 매개해주는 은닉된 랜덤 변수인 latent variable 이 존재한다는 전제를 가진다[11] 이는 잠재코드공간(latent code space)에서 실제 문장을 생성하면서 자연어의 풍부한 구조를 발견하는 연구로 사용되었다[12]. 이를 이용한 방법으로는 전체 문장의 분산잠재 표상을 통합한 RNN 기반의 VAE 생성 모델이 있다. 이 모델은 명시적인 전역 문장 표현에서 작동하며, 이 문장에 대한 사전확률(prior)로부터 뽑은 샘플은 다양하고 세련된 문장을 만들었다[13]. 지정된 의미를 가진 엉킴 없는(disentangled)잠재 표현을 학습함으로써 속성이 제어되는 문장 생성 기법을 제안한 방법도 있다[14]. 이 방법은 구조화된 변수 집합을 사용하여 VAE의 잠재코드를 만들었는데 각각은 중요하고 독립적인 문장의 피처를 목표로 한다. 모델은 VAE와 속성 판별기를 포함한다. 그럴듯한 텍스트를 생성하기 위해 생성자를 훈련시켜 실제 문장을 재구축하는 한편, 판별기는 생성자가 구조화된 코드와 일관된 속성을 생성하도록 한다. 이를 통해 주요 속성(시제, 감성)에 맞는 그럴듯한 문장들을 생성할 수 있음을 보여줬다. 따라서, 이들 방법을 이용하여 한국어 기반으로한 VAE 문장 생성 모델을 제안하며, 의미적 연결 관계를 기반으로 다중 문서 요약이 가능하도록 한다.

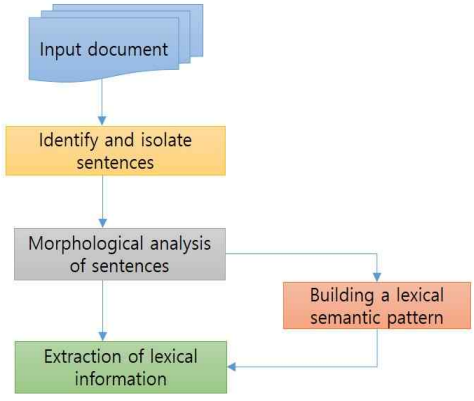
3. 본론

본 장에서는 의미적 연결 관계를 기반으로 문장을 추출하여 다중 문서를 요약할 수 있는 방법을 제안한다. 제안 방법은 크게 전처리 단계[15]와 VAE를 이용한 문장 추출 단계로 이루어진다.

3.1 전처리

전처리 단계에서는 주어진 한국어 문서를 각 구문이

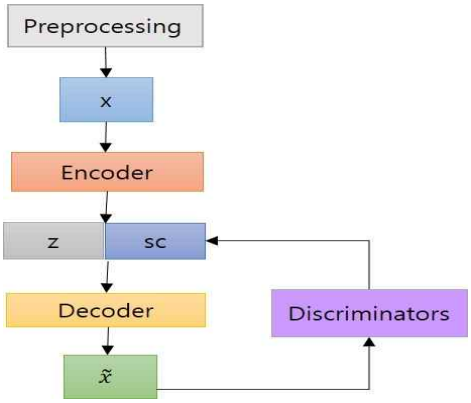
하나의 의미를 갖도록 각각의 문장으로 분리하고 그 후 각 문장의 품사를 판별하는 작업을 [Fig. 1]에서와 같이 처리한다. 이러한 형태소 분석 시 불용어 제거 및 활용할 어근을 추출하도록 한다. 주제와 관련된 단어의 의미 파악을 위해 사전을 기반으로 어휘 의미 패턴을 구축한다. 이것은 같은 의미를 가진 여러 문장 유형을 분석하기 위함이다. 이러한 어휘정보를 이용하여 단어를 숫자로 변환한다. 이때, Word2Vec를 이용하여 단어의 연결을 기반으로 한 형태소 벡터를 생성한다. 이 형태소 벡터를 기반으로 입력과 출력으로 VAE 모델에 학습시킨다. 이 논문에서는 VAE 모델을 이용한 방법에 초점을 두고 작성하였기 때문에 전처리 과정에 대해서 자세한 설명은 생략하였다. 어휘 의미 패턴 정보는 VAE 모델의 판별기 부분에서 문장 속성 판별 시 참조하도록 한다.



[Fig.1] Preprocessing

3.2 VAE를 이용한 문장 추출 방법

의미적 연결 관계를 기반으로 자연스러운 문장 추출하기 위해 가변 자동 인코딩 방법(VAE)을 사용하며, 이는 인코더와 디코더 네트워크로 구성된다[16]. 인코더는 입력된 벡터(x)로부터 잠재된 변수(z)로 생성하며, 가우시안 분포를 가진 랜덤 변수라할 때 압축된 정보의 평균과 분산을 생성할 수 있다. 디코더는 압축된 잠재 변수 정보를 이용하여 입력 벡터를 재구성하는 네트워크를 나타낸다. 여기서 VAE 모델의 입력층은 현재 형태소의 형태소임베딩 벡터이고, 출력층은 다른 형태소의 형태소임베딩 벡터가 된다. 본 논문에서는 위키피디아의 한국어 데이터를 사용하였으며, KoNLPy의 트위터 형태소 분석기를 이용하였다[17]. 따라서, [Fig. 2]와 같이 입력된 벡터(x)는 전처리에서 생성되는 형태소임베딩 벡터이며, 출력 벡터($\tilde{x}$)들은 잠재된 변수(z)로 생성된 압축된 정보로부터 문장을 재구성한다. sc는 어휘패턴 정보를 참고로 각 문장의 속성 코드가 존재하게 되며, 제어할 문장 속성을 대상으로한 구조화된 코드이다. 속성 판별기(Discriminators)는 생성된 샘플들을 설명할 수 있을 뿐 아니라 지정된 의미를 수반할 수 있도록 하였다. 여기서의 인코더를 $q_\theta(z|x)$로 표현하며, 디코더는 $p_\phi(x|z)$로 표현한다. p는 평균 0과 분산 1을 갖는 표준정규분포로 $p(z) = Normal(0,1)$이다.



[Fig 2] VAE model with discriminator added for semantic connection

VAE 입력 벡터 x에 대한 학습 시, 목표함수는 다음과 같이 정의된다.

$$\mathcal{L}_{VAE}(\theta,\phi;x) = -KL(q_\theta(z|x) \parallel p(z)) + E_{q_\theta(z|x)q_D(sc|x)}[\log p_\phi(x|z,sc)] \quad (1)$$

여기서의 ϕ, θ 는 각각 디코더와 인코더의 매개변수를 나타내며, VAE는 관찰된 실제 문장의 재구성 오차를 최소화하도록 최적화되며 동시에 이전의 p(z)에 근접하도록 인코더를 정규화합니다. $KL(q_\theta(z|x) \parallel p(z))$ 는 인코더로부터 생성된 잠재 변수의 확률 분포와 잠재변수의 사전 확률 분포 간의 KL 분산(Kullback-Leibler divergence)을 의미하며, 인코더에서 생성된 잠재변수의 확률 분포가 가우시안 분포와 유사할 수 있도록 제약해주는 역할을 한다. $E_{q_\theta(z|x)q_D(sc|x)}[\log p_\phi(x|z,sc)]$은 VAE의 입력과 출력 사이의 교차엔트로피(cross-entropy)를 나타내며, 재구성한 입출력 벡터 간의 차이를 줄여주기 위한 것이다.

$q_D(sc|x)$는 sc의 각 구조화된 변수에 대한 판별기 D로 정의된 조건부 분포이다. 표기 단순화를 위해 하나의 구조화된 변수와 하나의 판별기로 가정하지만 여러 속성으로 적용할 수 있다. 현실적인 문장을 생성하도록 디코더를 동작하며, 판별기는 디코드를 제어하여 구조화된 sc와 일치하는 일관된 속성을 생성하도록 추가 학습 신호를 제공한다. 개별 샘플들을 통한 판별기로부터 변화도를 전파할 수 없으므로 결정론적인 연속 근사치를 사용한다. 근사치는 각 단계에서 샘플된 토큰들을 softmax 함수를 사용하여 확률 벡터로 나타내며, 현재 단계의 출력과 의사결정 순서에 따른 다음 단계의 입력으로 사용된다. 확률 벡터로 생성된 문장들의 결과는 목표 속성에 적합성 측정을 위해 판별기로 보내지는데 이때 디코더를 향상시키며 손실이 일어난다. 따라서 이러한 손실을 학습을 통해 최소화시키도록 하며, 판별기 D에서는 잡음이 많은 데이터들을 완화하고, 모델 최적화의 견고성을 보장하도록 식(2)와 같이 최소 엔트로피 정규화를 한다.

$$\mathcal{L}_D(\theta_D) = E_{P_\phi(\hat{x}|z,sc)p(z)p(sc)}[\log q_D(sc|\hat{x}) + \beta H(q_D(\acute{sc}|\hat{x}))] \quad (2)$$

위의 식(2)에서 θ_D는 판별기의 매개변수를 나타내며, β는 균형 매개변수를 나타낸다. $\beta H(q_D(\acute{sc}|\hat{x}))$는 디코더에 의해 생성된 문장 $\hat{x}$에서 평가된 경험적 섀넌 엔트로피(Shannon entropy)분포 q_D이다. 따라서, 판별기는 레이블이 지정된 샘플들과 잡음이 많은 문장-속성 쌍으로 합성된 샘플들을 사용하여 최소 엔트로피 정규화 모델이 레이블 예측에 대한 높은 확신을 줄 수 있도록 학습한다.

4. 평가

본 연구에서는 의미적 연결 관계를 고려한 요약문을 생성하기 위해 위키피디아의 한국어 데이터를 사용하였으며, KoNLPy의 트위터 형태소 분석기를 이용하였다. 문장 수준의 제어를 위해 단어 수준의 레이블을 사용하며, 이는 어휘 의미 패턴 정보를 통해 자주 사용되는 단어들을 파악하며 이에 따른 속성 판별을 위해 단어 뒤에 붙는 조사, 동사의 시제를 파악하고 분류하여 생성된 문장에 레이블을 지정하였다. 기존의 문서 요약에 대한 평가를 위해서 딥 러닝 알고리즘을 연계한 RNN방법과 RNN+LSTM 방법, 그리고 제안한 VAE를 이용한 방법을 비교하였다. 이를 위해 ROUGE의 시스템 중 ROUGE-N을 이용하여 논문의 실험을 평가하였다[18]. n-gram을 이용하여 전문가가 직접 요약한 문서와 자동으로 요약된 문서를 비교평가한다. <Table 1>은 unigram을 사용한 ROUGE-1의 정확률(P), 재현률(R), F1-score(F)와 기본(all words), 어간(word stem) , 불용어 제거(elimination of stop words) 를 한 내용으로 각각을 비교하여 나타내었다. 다른 문서 요약 방법을 통해 생성된 각각의 요약문들과 전문가가 직접 요약한 문서를 비교했을 때 제안한 방법(VAEMS)이 미세하지만 조금씩 성능이 다 향상됨을 볼 수 있다.

<Table 1> ROUGE-1 Score in terms of Precision, Recall, F1-score and all words(Basic), stemmed words, elimination of stop words.

Categories		RNN	RNN+LSTM	VAEMS
Basic	P	0.413	0.455	0.468
	R	0.394	0.441	0.465
	F	0.418	0.448	0.466
Stem	P	0.442	0.487	0.498
	R	0.411	0.479	0.488
	F	0.439	0.481	0.499
Stem + elimination of stop words	P	0.364	0.443	0.457
	R	0.387	0.416	0.442
	F	0.379	0.438	0.455

5. 결론

많은 양의 데이터를 중심적인 내용만을 추려 하나의 요약문을 만들어내기 위한 다중 문서 요약 기법의 필요성과 중요성이 대두되고 있다. 그러나, 기존 문서 요약 연구 방법들은 지나치게 단순한 통계에 의존함으로써 문서 주제 파악에 한계를 보이거나, 문장의 모호성 등을 보임으로써 해결해야 할 문제들이 많이 남았다. 또한, 대량의 데이터 처리 시 높은 정확도와 효율성이 입증된 딥 러닝 알고리즘을 연계한 문서 요약 모형을 개발한 연구 많지 않으며, 다중 문서 요약 연구도 부재한 실정이다.

본 논문에서는 의미적 연결 관계를 파악 및 불필요한 정보를 처리하기 위한 전처리 과정을 거치며, 어휘 의미 패턴 정보를 기반으로 VAE 를 이용하여 문장 간의 의미

적 연결성을 높인 다중 문서 요약 기법을 제안하였다. 전처리 정보를 이용하여 문장을 이루고 있는 단어 벡터들을 입력 벡터로 사용하며, 잠재된 변수로 생성된 압축된 정보와 속성 판별기로부터 학습하여 문장을 재구성함으로써 의미적 연결 처리가 자연스러운 요약문을 생성하였다. 제안된 방법과 다른 문서 요약 방법을 통해 생성된 각각의 요약문들과 전문가가 직접 요약한 문서를 비교했을 때 재현률, 정확률, F1-score 측면에서 미세하지만 조금씩 더 향상된 성능을 나타낸다는 것을 알 수 있었다. 이는 해석 가능한 잠재 표현을 학습하고 어휘 의미 패턴을 이용한 속성을 가진 의미 있는 문장 생성 가능하다는 것을 보여주며, 빅데이터 대상의 문서 요약 수행 시 문장 간의 의미적 연결성을 높일 수 있다는 것을 알려준다.

향후 연구로는 성능 평가부분에서 제한된 데이터를 사용하였으므로 다른 데이터 집단을 가지고 좀 더 객관적인 성능 평가를 하여 내용을 보완하고자 한다. 또한 어휘 의미 패턴 정보를 통한 속성 또한 실험 결과에 영향을 주었을 수도 있으므로 다양한 속성 설정 값을 가지고 실험하여 확장할 수 있는 방법을 연구하고자 한다.

REFERENCES

[1] Jinsu Kim, “Emotion Prediction of Document using Paragraph Analysis”, Journal of Digital Convergence, Vol. 12, No. 12, pp.249-255, 2014.

[2] J. Goldstein, V. Mittal, J. Carbonell, & M. Kantrowitz, “Multi-document summarization by sentence extraction”, In NAACL-ANLP 2000 Workshop on Automatic summarization, pp. 40-48, 2000.

[3] O. Sornil, K. Gree-ut, “An Automatic Text Summarization Approach using Content-Based and Graph-Based Characteristics”, In Proceedings of IEEE Conference on Cybernetics and Intelligent Systems, pp. 1-6, 2006.

[4] Mihalcea, Rada, and P. Tarau, “TextRank : Bringing order into texts,” Association for Computational Linguistics, 2004.

[5] D. Wang, S. Zhu, T. Li, & Y. Gong, “Multi-document summarization using sentence-based topic models”. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers: pp.297-300, 2009.

[6] I. Mani, “Automatic Summarization”, John Benjamins Publishing Company, pp.114-125, 2001.

[7] X. Wan, J. Yang, “Multi-Document Summarization Using Cluster-based Link Analysis”, Proceeding of the International Conference(SIGIR’08), 2008.

[8] Z. Cao, F. Wei, L. Dong, S. Li, & M. Zhou, “Ranking with recursive neural networks and its application to multi-document summarization”. In AAAI: pp.2153-2159, 2015.

[9] Won-Chul, Kim “Scalable Multi-document Summarization Using Deep Learning-based Topic Modeling”, M.S. thesis, Yonsei University, 2016.

[10] Henning, Leonhard, “Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis”, Proceeding of the International Conference RANLP’09, 2009.

[11] J. Y. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengio, “ A recurrent latent variable model for sequential data”, In Neural Information Processing Systems (NIPS), 2015.

[12] D. P. Kingma, M. Welling, “Auto-encoding variational Bayes” In Proceedings of the International Conference on Learning Representations (ICLR), 2014.

[13] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, “Generating sentences from a continuous space“, arXiv preprint arXiv:1511.06349, 2015.

[14] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, “ Controllable text generation”, arXiv preprint arXiv:1703.00955v2, 2017.

[15] Kyung-Ae Kim, Jin-Hee Ku, “A Study on the Change of the View of Love using Text Mining and Sentiment Analysis”, Journal of Digital Convergence, Vol. 15, No. 2 , pp.285-294, 2017.

[16] A. Graves, “Generating sequences with recurrent neural networks”, In Arxiv preprint arXiv:1308.0850, 2013.

[17] Cheol-Jung Yoo, Yong Kim, Bo-Hyun Yun, “A Study on Utilization of Wikipedia Contents for

Automatic Construction of Linguistic Resources", Journal of Digital Convergence, Vol. 13, No. 5 , pp.187-194, 2015.

[18] P. McNamee, J. Mayfield, "Character N-Gram Tokenization for European Language Text Retrieval", Information Retrieval, Vol 7, No. 1-2, pp.73-97, 2004.

백 수 진(Baek, Su Jin)

・2002년 2월 : 경희대학교 컴퓨터공학과(공학사)
・2004년 2월 : 경희대학교 컴퓨터공학과(공학석사)
・2012년 8월 : 경희대학교 컴퓨터공학과(공학박사)
・2015년 5월 ~ 현재 : 용인송담대학교 정보통신학과 겸임 교수
・관심분야 : 자연어처리, 딥러닝, 적응형 소프트웨어
・E-Mail : croso79@ysc.ac.kr