ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2019 June: 25(2): 141~166 http://dx.doi.org/10.13088/jiis.2019.25.2.141

주제 균형 지능형 텍스트 요약 기법

윤여일
국민대학교 경영대학
(*yunyi94@kookmin.ac.kr*)

고은정
국민대학교 비즈니스IT전문대학원
(*sbtm3459@kookmin.ac.kr*)

김남규
국민대학교 경영대학
(*ngkim@kookmin.ac.kr*)

최근 다양한 매체를 통해 생성되는 방대한 양의 텍스트 데이터를 효율적으로 관리 및 활용하기 위한 방안으로써 문서 요약에 대한 연구가 활발히 진행되고 있다. 특히 최근에는 기계 학습 및 인공 지능을 활용하여 객관적이고 효율적으로 요약문을 도출하기 위한 다양한 자동 요약 기법이(Automatic Summarization) 고안되고 있다. 하지만 현재까지 제안된 대부분의 텍스트 자동 요약 기법들은 원문에서 나타난 내용의 분포에 따라 요약문의 내용이 구성되는 방식을 따르며, 이와 같은 방식은 비중이 낮은 주제(Subject), 즉 원문 내에서 언급 빈도가 낮은 주제에 대한 내용이 요약문에 포함되기 어렵다는 한계를 갖고 있다. 본 논문에서는 이러한 한계를 극복하기 위해 저빈도 주제의 누락을 최소화하는 문서 자동 요약 기법을 제안한다. 구체적으로 본 연구에서는 (i) 원문에 포함된 다양한 주제를 식별하고 주제별 대표 용어를 선정한 뒤 워드 임베딩을 통해 주제별 용어 사전을 생성하고, (ii) 원문의 각 문장이 다양한 주제에 대응되는 정도를 파악하고, (iii) 문장을 주제별로 분할한 후 각 주제에 해당하는 문장들의 유사도를 계산한 뒤, (iv) 요약문 내 내용의 중복을 최소화하면서도 원문의 다양한 내용을 최대한 포함할 수 있는 자동적인 문서 요약 기법을 제시한다. 제안 방법론의 평가를 위해 TripAdvisor의 리뷰 50,000건으로부터 용어 사전을 구축하고, 리뷰 23,087건에 대한 요약 실험을 수행한 뒤 기존의 단순 빈도 기반의 요약문과 주제별 분포의 비교를 진행하였다. 실험 결과 제안 방법론에 따른 문서 자동 요약을 통해 원문 내 각 주제의 균형을 유지하는 요약문을 도출할 수 있음을 확인하였다.

주제어 : 문서 자동 요약, 워드 임베딩, 토픽 모델링, 텍스트 마이닝, 리뷰 요약

논문접수일 : 2019년 1월 3일 논문수정일 : 2019년 1월 3일 게재확정일 : 2019년 5월 6일
원고유형 : 학술대회(급행) 교신저자 : 김남규

1. 서론

정보통신기술의 발전과 스마트기기의 보급량 증가에 따라 비정형 데이터의 양이 꾸준히 증가하고 있으며, 특히 다양한 소셜 미디어를 통해 유통되는 텍스트의 양이 기하급수적으로 증가하고 있다. 텍스트는 데이터의 생성 및 수집이 용이하고 정보전달력이 우수하다는 특징을 갖고 있으며, 이로 인해 이러한 텍스트 데이터로부터 유의미한 정보를 추출하고 이를 효율적으로 관리하기 위한 많은 방법들이 고안되고 있다. 구체적으로는 텍스트를 특성에 따라 사전에 정의한 카테고리로 구분하는 문서 분류(Document Classification), 다량의 문서로부터 핵심 주제를 도출하는 토픽 분석(Topic Analysis), 텍스트를 통해 표출된 다양한 의견 및 감정에 대한 정보를 찾아서 이들의 양상을 파악하는 오피니언 마이닝(Opinion Mining), 그리고 방대한 내용을 축약하여 짧은 길이의 텍스트로 표현하는 문서 자동 요약(Automatic Text Summarization) 등에 대

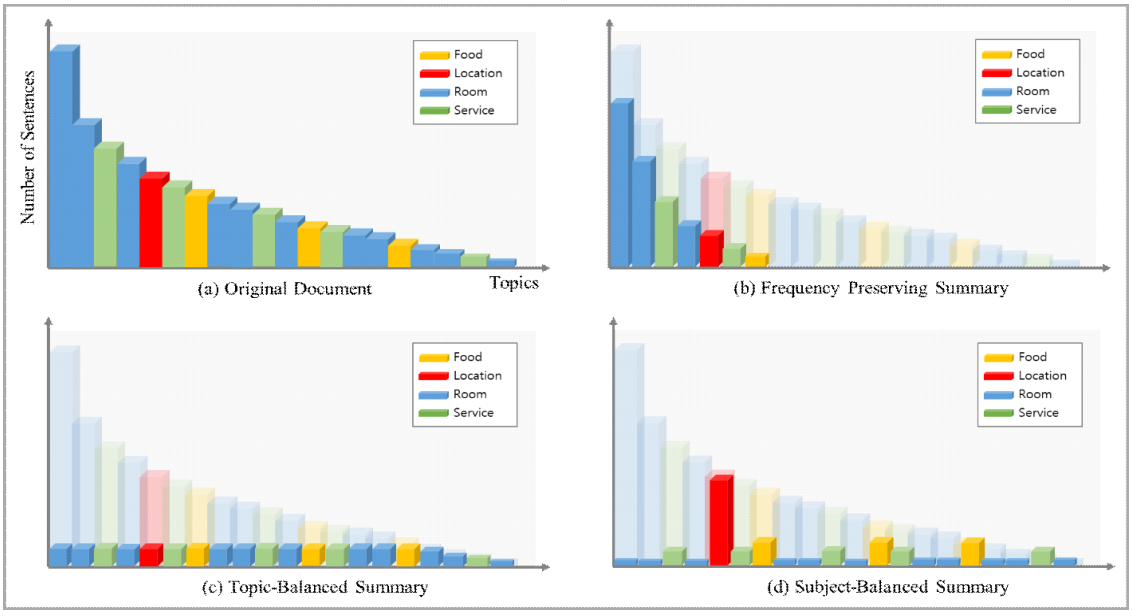
한 연구가 학계와 업계를 막론하고 활발하게 이루어지고 있다.

이처럼 다양한 텍스트 분석 응용 가운데 문서 자동 요약은 자연어(Natural Language) 형태의 입력에 대한 가공을 통해 자연어 형태의 결과물을 도출한다는 측면에서 다른 분석들과는 구별되는 측면이 있다. 일반적으로 문서의 요약은 문서의 수나 길이를 축약하여 나타내는 과정을 의미하며, 최소 분량의 텍스트로 원문의 내용을 최대한 다양하게 포함하는 것을 목적으로 한다. 수작업을 통해 텍스트 데이터의 원문을 확인하고 핵심을 파악하는 작업에는 막대한 시간과 비용이 소요될 뿐 아니라, 방대한 양의 텍스트를 모두 직접 확인하는 작업은 점점 더 불가능에 가까워지고 있다는 한계가 존재한다. 이러한 배경에서 사람의 주관이나 노력이 개입되지 않는 문서 자동 요약에 대한 관심이 더욱 집중되고 있다. 구체적으로 문서 자동 요약은 국내 인터넷 포털 Naver와 Daum의 자동 뉴스 요약 서비스, Agolo의 웹 요약 플랫폼, Mashape나 DeepAI가 제공하는 문서 요약 API 등 다양한 플랫폼과 서비스를 통해 제공되고 있으며, 학계에서도 이와 관련된 연구가 지속적으로 이루어지고 있다.

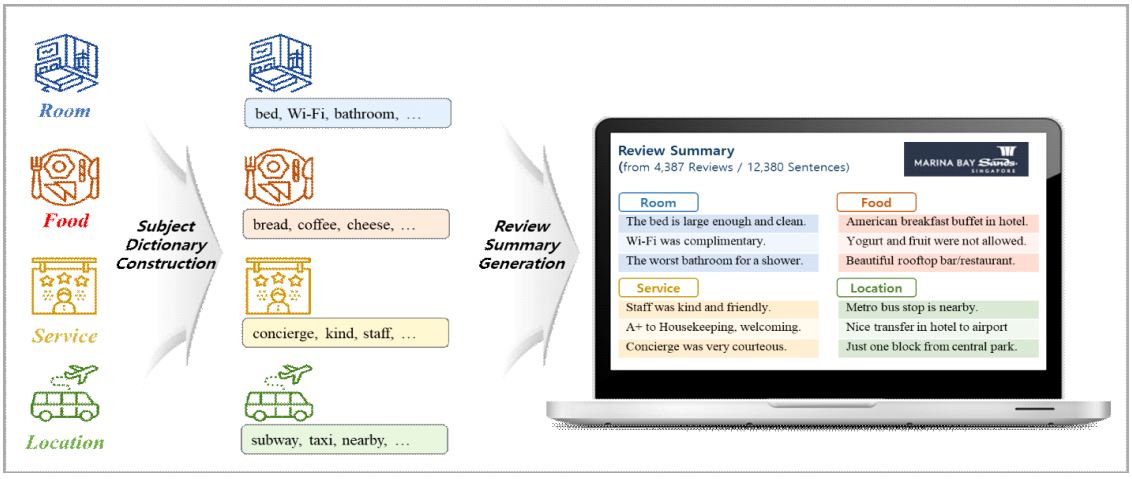
문서 자동 요약은 크게 문서의 주제를 파악하여 이를 새로운 형태의 문장으로 재구성하는 생성 요약(Abstraction-based Summarization)과 문서의 핵심을 포함하는 부분 문서를 탐색하고 발췌하여 이들을 요약문으로 나타내는 추출 요약(Extraction-based Summarization)으로 구분된다. 생성 요약은 자연어 처리에 기반을 두어 새로운 문장을 구성하므로, 요약문의 압축률과 정제 수준을 높일 수 있다는 장점을 갖는다. 하지만 생성 요약은 핵심 내용의 이해 외에 자연스러운 표현으로 문장을 재구성하기 위한 추가 노력이 필요하며, 이로 인해 대부분의 자동 요약 연구들은 추출 요약에 집중되어 있다. 이러한 기존 연구의 대부분은 출현 빈도에 기반을 두어 요약을 수행하기 때문에, 출현 빈도가 낮은 주제(Subject)에 대한 내용은 포함되기 어렵다는 한계를 갖는다. 즉 많은 수의 문장을 통해 언급되는 주제는 요약문에 다수의 문장으로 포함될 가능성이 높은 반면, 자주 언급되지 않은 주제는 요약문에서 아예 누락될 가능성이 높게 나타난다(Figure 1).

<Figure 1>은 호텔에 대한 리뷰를 세 가지 방식으로 요약한 가상 예를 보이고 있다. “Food”, “Location”, “Room”, “Service”의 네 개 주제에 대해 총 20개의 주제가 나타나 있으며, 원문의 경우 각 주제에 대응되는 토픽은 각각 3개, 1개, 11개, 그리고 5개이다. <Figure 1(b)>는 각 주제별의 빈도에 따라 요약을 수행한 결과이며, 이 경우 발생 빈도가 낮은 주제는 요약에 포함되지 않음을 알 수 있다. 한편 <Figure 1(c)>는 각 주제별로 동일한 수의 문장을 추출하여 요약문을 구성한 것이다. 토픽 분석에 기반을 둔 대부분의 기존 연구가 이러한 방식을 사용하고 있으며, 이 경우 요약문에서 주제별 균형은 유지되지만 “Room”과 관련된 문장에 비해 “Location”과 관련된 문장이 지나치게 적게 포함되는 주제별 불균형이 나타난다. 즉 극단적인 경우 요약문에서 “Location”에 대한 내용은 전혀 찾아볼 수 없는 경우도 발생할 수 있다. 한편 <Figure 1(d)>는 본 연구에서 제안하는 방식으로, 주제의 균형을 고려한 요약을 나타낸다. 예를 들어 “Location”의 경우 하나의 주제에서 11개의 문장을 추출하고 “Room”의 경우 각 11개의 주제에서 1개의 문장을 추출하여, 각 주제별 문장이 요약문에서 차지하는 비율을 균형있게 유지하는 방식이다.

본 연구에서는 문서가 포함하는 다양한 주제



〈Figure 1〉 Comparison of Various Summarization



〈Figure 2〉 Subject Dictionary Construction and Review Summary Generation

를 정의하고, 각 주제와 관련된 내용을 균형있게 요약하는 방안을 제시한다. 구체적으로 본 연구에서는 (i) 문서에 내포된 핵심 주제를 정의하고 각 주제에 해당하는 시드(Seed) 용어를 구성한 뒤, (ii) 워드 임베딩을 통해 주제별 용어 사전을 생성하고, (iii) 문장 간 유사도 및 각 문장의 주제별 대응도를 산출한 뒤, (iv) 요약문의 평가 지표인 완전성(Completeness)과 간결성(Succinctness)

의 개념을 활용하여(Ko and Kim, 2018), 요약문 내 내용의 중복을 최소화하면서도 원문의 다양한 내용을 최대한 포함할 수 있는 문서 자동 요약 기법을 제시한다. 이상의 전체 과정은 <Figure 2>와 같이 요약되며, 이러한 과정을 통해 최종적으로 다양한 주제의 내용을 균형있게 포함하는 요약문을 도출할 수 있다.

본 논문의 이후 구성은 다음과 같다. 먼저 2장에서는 본 연구에서 사용하는 다양한 기술적 방법론 및 문서 요약에 대한 선행 연구의 성과를 요약하고, 3장에서는 본 연구에서 제안하는 방법론을 예시와 함께 소개한다. 4장에서는 제안 방법론을 실제 호텔 리뷰 데이터에 적용한 실험 결과를 제시하고, 마지막 5장에서는 본 연구의 기여 및 한계를 요약한다.

2. 관련 연구

2.1 텍스트 분석

텍스트 마이닝(Text Mining)은 기존의 데이터 마이닝의 다양한 기법을 텍스트 데이터에 접목한 분야로써, 텍스트 데이터로부터 알려지지 않은 새로운 사실을 발견하고 이를 통해 통찰을 도출하는 분야로(Tan, 1999) 정의된다. 텍스트 마이닝을 수행하기 위해선 텍스트 데이터의 수집과 분할, 정제와 같은 전처리 과정이 필요하며, 분석을 위해 자연어 형태의 텍스트를 정형화된 형태로 바꾸는 구조화의 단계가 선행되어야 한다. 이는 텍스트를 컴퓨터와 같은 정보 기기가 접근할 수 있는 구조로의 변환을 의미하며, 텍스트의 특성을 유지하면서 구조화를 진행하기 위한 다양한 기법들이 활용되고 있다. 이러한 기법들 중 가장 대표적인 구조화 방법으로써 텍스트를 구성하는 용어의 출현 빈도를 벡터 구조로 표현하는 Vector Space Model이(Salton et al. 1975) 사용되고 있으며, 이 중에서도 용어의 절대적 출현 빈도와 상대적 출현 빈도를 동시에 고려하는 TF-IDF(Term Frequency-Inverse Document Frequency)와 이를 변형한 구조화 기법들이 가장 대중적으로 활용되고 있다(Wen et al. 2011). 이처럼 용어의 출현 빈도를 이용하여 문서의 구조화를 진행하면 (문서) × (용어)의 행렬(Document-Term Matrix)을 도출하며 이를 통해 문서 별로 중요한 의미를 지닌 용어를 파악할 수 있는 장점을 갖는다. 하지만 이처럼 개별 문서가 전체 용어에 대한 정보를 포함할 경우, 저장 공간과 용량의 낭비를 초래하고 분석에 소요되는 시간을 증가시키는 한계를 갖는다. 이와 같은 한계를 극복하기 위해 차원 축소 기법을 활용한 연구(Bingham and Mannila, 2001), 용어의 의미적 중요도를 고려한 구조화 방법에 대한 연구(Erk and Pado, 2008) 등의 새로운 구조화 방법에 관한 연구가 이루어지고 있다.

2.2 문서 요약

문서 요약은 하나 이상의 문서로부터 중요한 정보를 찾아내고 해당 정보를 포함하는 짧은 문서로 축소하는 과정이라고 정의할 수 있으며(Eduard, 2015), 이렇게 도출된 요약문은 원문이 지닌 정보를 최대한 반영하여 정보의 손실을 최소화하면서도 객관적으로 전달해야 한다. 이에 따라 사람의 주관이 개입하지 않는 자동 요약에 대한 연구가 주목을 받아 활발히 이루어지고 있다(Nenkova, 2012). 자동 요약은 소프트웨어나 프로그래밍 언어가 제공하는 다양한 알고리즘을

사용하여 문서를 축약하며, 요약문에 포함될 핵심 문장 혹은 핵심 용어를 선정하거나 문서로부터 토픽을 도출하는 과정 등에서 데이터 마이닝이나 기계 학습의 다양한 기법이 사용된다(Joel et al. 2002). 문서 요약의 결과는 문서의 핵심을 나타내는 문장 혹은 용어 단위로 표현되며, 요약 방법에 따라 상이한 형태의 요약문을 도출한다.

문서 요약 방법에는 크게 생성 요약과 추출 요약 두 가지로 구분이 된다(Eduard and Lin, 1998). 이 중 생성 요약은 원문의 핵심을 찾아내고 이를 종합하여 간략하게 축약된 새로운 문서를 생성하는 방법이다(Zhang et al. 2018). 생성 요약은 요약문을 구성할 문장 또는 단어의 전후 문맥, 의미적 흐름 등을 고려한 하나의 문서를 생성하여 구조적으로 높은 완성도를 가진 요약문을 도출하며, 이를 위해 자연어 처리를 요약문 도출에 활용한 연구나(Li et al. 2018), 핵심 문장의 선택과 재구성을 통해 요약문을 생성하는 연구(Chen and Bansal, 2018) 등이 이루어지고 있다. 한편 추출 요약은 원문의 주요 문장 혹은 단어들을 선택하여 추출된 부분 문서들의 집합으로 요약문을 구성하는 방법으로써(Mittal et al. 2014), 요약문이 원문에서 가지고 있던 형태를 유지할 수 있어 의미적 왜곡이나 문서의 변형이 발생할 확률이 낮은 장점이 있다. 추출 요약에는 원문의 문단 별 대표 문장을 선택 후 이를 종합하는 방법이나(Gupta and Lehal, 2010), 전체 문장의 중요도를 파악하여 중요도가 높은 문장만을 추출하여 요약문으로 제시하는 방법(Sonawane et al. 2018) 등 문서를 대표할 수 있는 핵심 문장을 요약문으로 구성하는 방향으로 이루어지고 있다.

이를 바탕으로 생성 요약, 추출 요약 모두 원문을 대표할 수 있는 문장을 파악하는 것이 가장 중요한 과정이라는 사실을 알 수 있으며 (Goldstein et al. 1999), 이 중에서도 특히 추출 요약의 경우 결과로 제시된 요약문이 원문의 일부를 그대로 보여주기에 그 중요성이 더욱 강조된다. 추출 요약에 대한 연구에는 대표적으로 원문의 주요 토픽을 발견하고 토픽을 대표할 수 있는 문장을 선택하여 요약문으로 도출하는 연구나(Gong and Liu, 2001; Rachit and Ravindran, 2008), 문서의 부분 군집을 형성한 후 군집 별 대표 문장을 선택하여 요약문으로 종합하는 클러스터링에 기반한 연구(Wan and Yang, 2008) 등이 있다. 또한 원문을 구성하는 용어나 문장 간의 의미적 유사도를 계산하여 상대적 중요도를 파악하고 이를 바탕으로 중요도가 높은 문장을 뽑아내 해당 문장을 핵심 문장으로써 요약문에 포함하는 연구가 진행되었으며(Kim et al. 2000; Ramiz, 2009), 최근에는 인공신경망을 활용하여 문장이나 용어 단위의 학습을 통해 다양한 가중치를 도출하고 이를 요약문 도출에 활용하는 딥러닝(Deep Learning) 기반의 연구도 활발히 이루어지고 있다(Mohamed and Fuji, 2009; Chorpa et al. 2016).

2.3 워드 임베딩

워드 임베딩은 임의의 용어를 실수 벡터로 표현하여 특정 차원의 벡터 공간으로 사상(Mapping)하는 구조화 방법이다. 용어를 벡터 형태로 표현하기 위한 다양한 기법들이 활용되어 왔지만, 이러한 기존의 워드 임베딩 모델은 구조화의 결과가 희소 행렬(Sparse Matrix)과 같은 비효율적인 저장 구조로 표현된다는 단점과 용어의 의미적 정보를 파악하기 어렵다는 한계를 가지고 있으며(Marco et al. 2014), 이런 한계를 극복하기 위한 다양한 워드 임베딩 모델에 대한 연

구가 진행되고 있다(Gao et al. 2017). 이 중 가장 대표적인 워드 임베딩 모델인 Word2Vec은 인공신경망을 활용한 용어 단위 학습 모델로써(Mikolov et al. 2013), CBOW(Continuous Bag of Words)와 Skip-Gram 두 가지 알고리즘을 활용하여 용어의 벡터화를 수행한다. CBOW는 대상 용어를 기준으로 주위에 등장하는 용어들에 대한 벡터를 도출하는 학습 모델이고, Skip-Gram 모델은 대상 용어와 인접하게 위치한 용어들을 바탕으로 대상 용어에 대한 벡터를 도출한다. Word2Vec은 기존의 임베딩 모델이 시도하지 않았던 개별 용어 관점의 구조화를 진행하였다는 점에 모델의 독창성을 인정받아 다양한 분야에서 활용되고 있다. 기존의 텍스트 마이닝 분야에서 이루어졌던 다양한 연구들도 Word2Vec이 접목되어 새로운 방향으로 제시되고 있으며, 문서 요약에도 마찬가지로 Word2Vec과 이를 활용한 알고리즘을 활용한 연구가 활발히 진행되고 있다. 그 예로써 Word2Vec의 Skip-Gram 알고리즘을 이용한 용어의 구조화와 이를 개량한 문장의 구조화를 통해 핵심 문장을 추출하는 연구(Kageback et al. 2014)와 Word2Vec을 통해 도출된 용어 벡터를 자연어 처리를 위해 구축된 다양한 신경망 모델에 대입하여 원문을 축소하고 재배열하는 연구(Nallapati et al. 2016) 등이 존재한다.

Word2Vec으로부터 도출된 용어 벡터를 활용하기 위한 다양한 지표들이 사용되고 있으며, 이 중에서도 코사인 유사도(Cosine Similarity)가 가장 대중적으로 사용이 되고 있다(Omer et al. 2015). 코사인 유사도는 비교 대상이 되는 두 벡터 간의 방향성을 나타내는 값으로써 벡터 간의 유사한 정도를 나타내는 지표로 활용이 된다. 유사도는 -1에서 1사이의 값으로 표현이 되며, 유사할 수록 1에 가깝게 유사하지 않을수록 -1에 가깝게 계산된다. 코사인 유사도는 어떤 차원의 벡터에도 적용이 가능하다는 장점이 있으며(Singhal, 2001), 특히 다차원 공간에 존재하는 벡터 간의 유사도 측정에 용이하다는 장점을 가지고 있어 많은 분야에서 활용되고 있다.

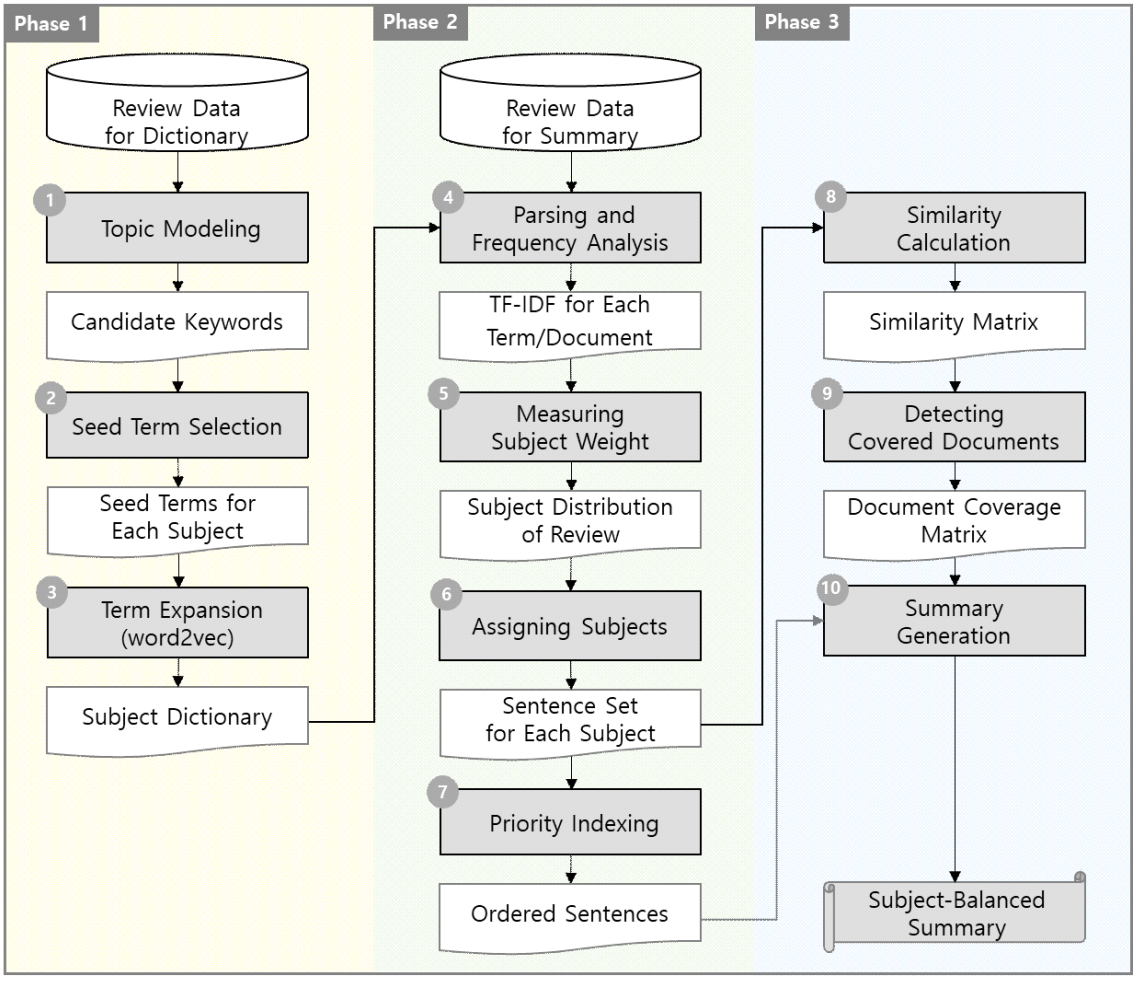
코사인 유사도를 문서 요약에 활용하는 경우 원문을 구성하는 문장 간의 유사도를 기반으로 요약을 진행하는 방향으로 주로 이루어지고 있다. 문장 간 유사도 가중치를 이용하여 네트워크를 구축하고 그 중심에 해당하는 문장을 핵심 문장으로 선정하여 요약문으로 구성하는 연구나(Yeh et al. 2008), 원문을 다수의 군집으로 분할한 뒤 군집 내 문장 간의 유사도를 도출하여 군집 별 대표 문장을 추출하는 연구(Zhang and Li, 2009) 등이 있다.

3. 제안 방법론

3.1 연구 모형

본 장에서는 문서가 내포한 다양한 주제를 식별하고, 각 주제와 관련된 내용을 균형있게 요약하는 요약문 도출 방법론을 간단한 예시와 함께 소개한다. 제안 방법론의 전체 개요는 <Figure 3>과 같다.

우선 Phase 1은 주제별 용어 사전을 구축하는 과정이다. 이를 위해 사전 구축용 문서에 대한 토픽 모델링을 수행하여 주요 용어를 확인하고(1), 전체 주제를 아우를 수 있는 주요 주제를 선정한 후 각 주제별 시드 용어를 선정한다(2). 주제별 용어를 모두 수작업으로 선정하는 것은 현실적으로 매우 어렵기 때문에, 각 주제별 시드



〈Figure 3〉 Research Overview

용어와 유사한 의미를 갖는 용어를 Word2Vec 알고리즘을 활용하여 발굴한다(3).

다음으로 Phase 2는 주제별 용어 사전을 활용하여 각 리뷰를 주제별로 재구성하는 과정이다. 우선 구축된 용어 사전을 기준으로 요약문을 추출할 대상 문서에 대해 빈도 분석을 수행하고(4), 주제별 용어에 대한 TF값과 IDF값을 도출하여 각 문장이 다양한 주제에 대해 대응되는 가중치를 산출한다(5). 이러한 가중치에 기반을 두어 각 문장을 주제별 문서 집합으로 재구성하고(6), 각 주제별 문장 집합 내에서 해당 주제의 가중치 값의 크기에 따라 주제별로 문장을 정렬한다(7).

마지막으로 Phase 3은 주제별 문장 집합으로부터 주제별 요약문을 도출하고 이를 통합하는 과정이다. 우선 각 주제 내에서 문장 간 유사도를 산출하고(8), 특정 임계값을 기준으로 유사도

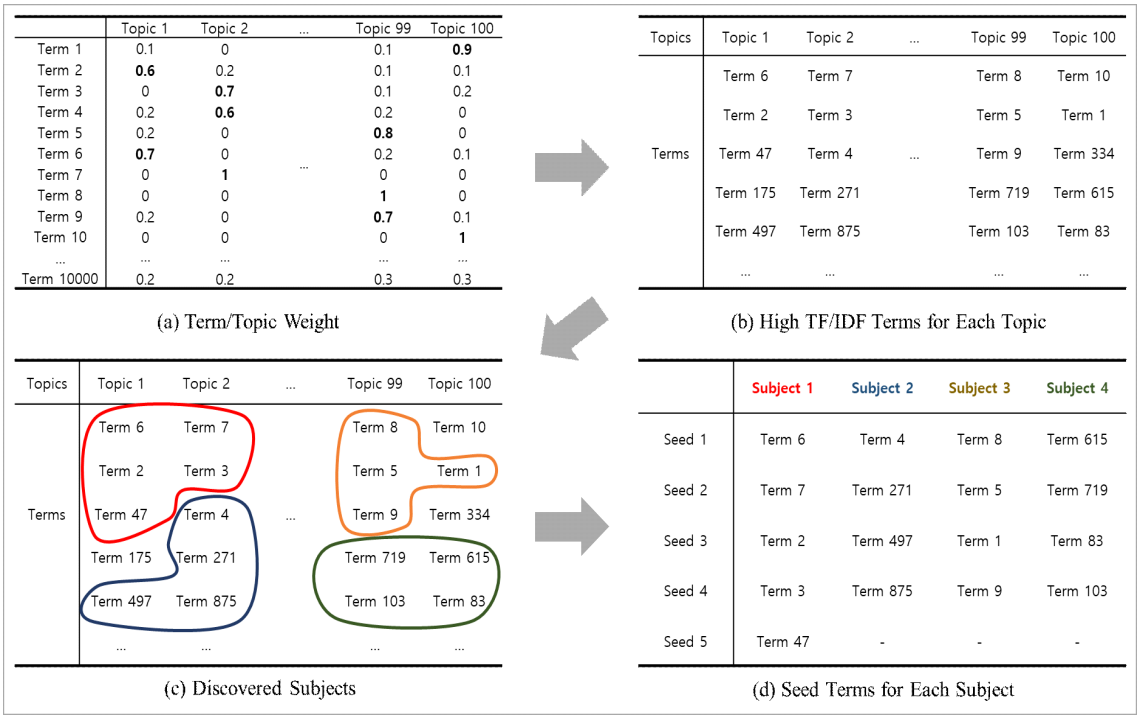
가중치를 0과 1의 이진 행렬로 변환한다(9). 다음으로 주제별 문장 집합에 대해 완전성과 간결성을 고려한 주제별 요약을 수행하고, 최종적으로 각 주제별 요약문을 종합하여 전체 요약문을 구성한다(10).

본 장의 이후 부분에서는 제안 방법론의 동작 과정을 간단한 가상 예를 통해 소개하며, 실제 데이터에 제안 방법론을 적용한 실험 결과는 4장에서 제시한다.

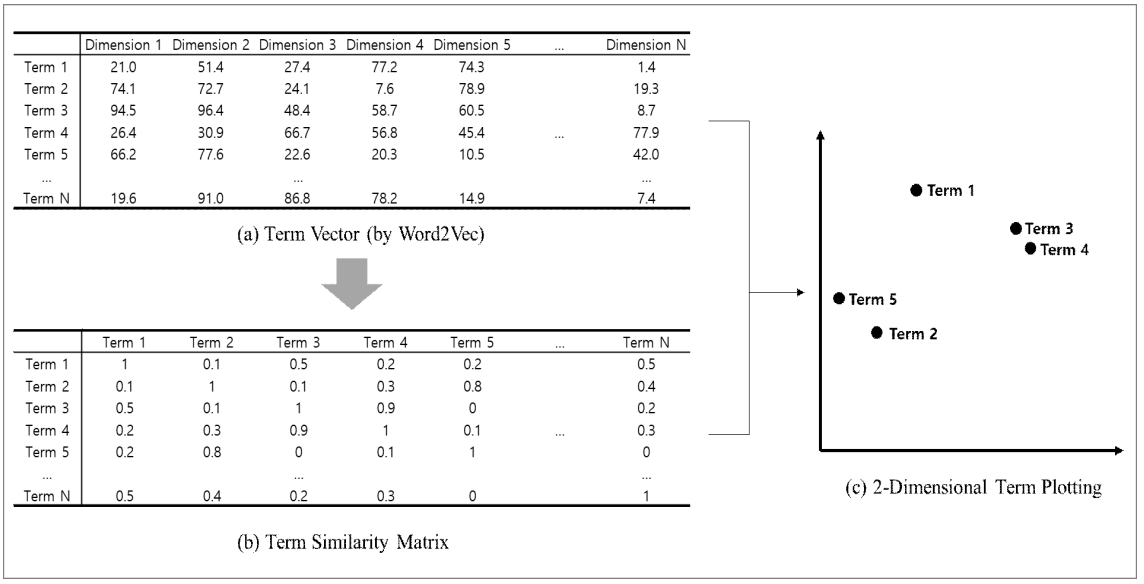
3.2 주제별 용어 사전 구축

본 절은 <Figure 3>의 Phase 1에 해당하는 과정, 즉 주제별 용어 사전을 구축하는 과정을 소개한다. 우선 주제 정의 및 시드 용어 선별을 위해 사전 구축용 문서에 대한 토픽 모델링을 수행한다. <Figure 4(a)>는 토픽 모델링 결과에 대한 가상 예를 나타내며, 각 용어가 100개의 토픽 각각에 대응되는 가중치를 보이고 있다. 이들 용어 중 각 토픽에 대해 가중치가 높은 주요 용어만을 선정하여 각 토픽의 가중치 순으로 정렬한 결과가 <Figure 4(b)>에 나타나 있다. <Figure 4(b)>에 수록된 주요 용어들로부터 전체 문서에서 중요하게 다루고 있는 주제를 <Figure 4(c)>와 같이 파악할 수 있으며, 이들 주제를 대표할 수 있는 핵심 용어를 <Figure 4(d)>와 같이 시드 용어로 선정할 수 있다.

하지만 이와 같은 방식으로 각 주제별 용어를 모두 선정하는 것은 매우 비효율적이므로, 이들 시드 용어를 기본으로 하고 시드 용어와 유사한



〈Figure 4〉 Discovering Subjects and Seed Terms



	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	...	Dimension N
Term 1	21.0	51.4	27.4	77.2	74.3		1.4
Term 2	74.1	72.7	24.1	7.6	78.9		19.3
Term 3	94.5	96.4	48.4	58.7	60.5		8.7
Term 4	26.4	30.9	66.7	56.8	45.4	...	77.9
Term 5	66.2	77.6	22.6	20.3	10.5		42.0
...			...				...
Term N	19.6	91.0	86.8	78.2	14.9		7.4

(a) Term Vector (by Word2Vec)

	Term 1	Term 2	Term 3	Term 4	Term 5	...	Term N
Term 1	1	0.1	0.5	0.2	0.2		0.5
Term 2	0.1	1	0.1	0.3	0.8		0.4
Term 3	0.5	0.1	1	0.9	0		0.2
Term 4	0.2	0.3	0.9	1	0.1	...	0.3
Term 5	0.2	0.8	0	0.1	1		0
...			...				...
Term N	0.5	0.4	0.2	0.3	0		1

(b) Term Similarity Matrix

(c) 2-Dimensional Term Plotting

〈Figure 5〉 Word Embedding with Word2Vec

의미와 맥락을 지닌 용어를 자동으로 추출하여 주제별 용어 사전을 확장한다. 이를 위해 대상 문서에 포함된 용어를 벡터로 변환한 후 벡터 간 유사도에 따라 유의어를 식별한다. 본 연구에서는 용어의 벡터 변환에는 Word2Vec 알고리즘을 활용하며, 벡터 간 유사도는 코사인 유사도로 측정한다(Figure 5).

<Figure 5(a)>는 Word2Vec을 이용하여 전체 용어를 100차원 공간에 사상한 것이며, <Figure 5(b)>는 용어 벡터 간 코사인 유사도를 산출한 결과이다. 또한 직관적인 이해를 돕기 위해 100차원 용어 벡터를 2차원으로 압축하여 도식화한 결과가 <Figure 5(c)>에 나타나 있다. <Figure 5(b)>에서 유사도가 높게 나타난 Term3/Term4, 그리고 Term2/Term5는 <Figure 5(c)>에서 근접하게 나타남을 알 수 있다.

다음으로 각 시드 용어의 유의어로 추출된 용어에 대한 검증을 통해 각 주제별 용어 사전을 구축한다. 이 작업은 각 시드 용어와 높은 유사도를 갖는 확장 용어 후보를 유사도 순으로 정렬한 후, 해당 주제에 적합하지 않은 용어를 사전에서 제거하는 방식으로 수행된다. 예를 들어 <Figure 6>은 각 주제별로 5개의 시드 용어를 갖는 4개의 주제 “Room”, “Food”, “Service”, 그리고 “Location”에 대해 유의어를 확장하는 가상 예를 보인다. <Figure 6(a)>는 각 시드 용어에 대해 유사도 순으로 5개씩의 용어를 확장 용어 후보로 추출하였으며, 해당 주제와 관련이 있는 것으로 파악된 용어를 굵은 글씨로, 관련이 없는 것으로 파악된 용어를 흐린 글씨로 표시하였다. 이렇게 확장 용어 후보에 대한 검증을 통해 주제별 용어 사전을 구축한 결과의 예는 <Figure 6(b)>와 같다.

Subject	Seed Terms	Similar Term 1	Similar Term 2	Similar Term 3	...	Similar Term N
Subject A	breakfast	sandwiches	buffet-style	pizzas		meals
	drinks	tea	coffee	espresso		herbal
	bread	pancake	waffle	meats	...	cheeses
	buffet	buffet-style	pizzas	seafood		cafe
	coffee	cappuccino	espresso	tea		teabags
Subject B	bed	beds	mattress	pillows		linen
	small	tiny	big	king-size		robe
	dirty	clean	dirt	messy	...	sofa
	cover	pillow	cotton	linen		curtain
	clean	bathroom	cleanness	livingroom		shower
Subject C	traffic	transportation	blocks	tucked		station
	subway	metro	stops	train		tunnel
	avenue	street	ave	square	...	park
	terminal	port	railway	bus		route
	nearby	close-by	cloesest	seconds		hub
Subject D	bell	bellman	clerks	doorman		professional
	desk	frontdesk	concierge	reception		rude
	doorman	doormen	clerks	helpful	...	greeting
	staff	staffs	employee	bellmen		team
	kindly	polite	gracious	professional		upbeat

(a) Similar Terms of Each Seed Term

	Subject A	Subject B	Subject C	Subject D
Seed Terms	breakfast	bed	traffic	bell
	drinks	small	subway	desk
	bread	dirty	avenue	doorman
	buffet	cover	terminal	staff
	Coffee	clean	nearby	kindly
Expanded Terms	buffet-style	mattress	metro	frontdesk
	cookies	pillow	station	clerks
	bistro	curtain	bus	polite
	fries	linen	park	greeting
	cappuccino	bathrobe	mall	professional
	meats	shower	square	concierge
	non-alcoholic	hanger	shops	bellman
	espresso	Wi-fi	nearest	24hours
	chocolate	soap	riverside	cleaning
	cheese	dryer	downtown	manner

(b) Subject Dictionary

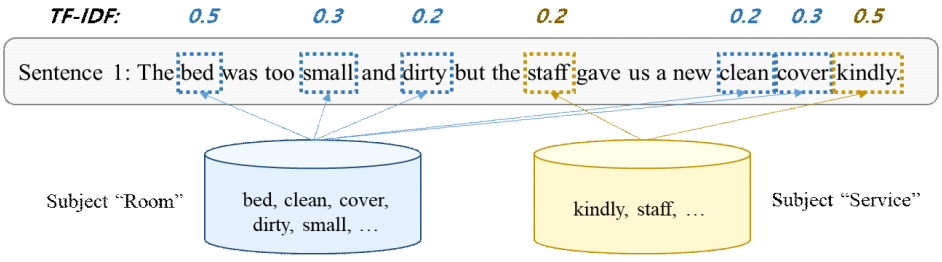
〈Figure 6〉 Similar Term Expansion and Subject Dictionary Construction

3.3 리뷰의 주제별 재구성

본 절에서는 <Figure 3>의 Phase 2에 해당하는 과정, 즉 각 리뷰를 주제에 따라 재구성하는 과정을 설명한다. 우선 각 문장에 대한 파싱을 통해 각 용어가 각 문장에서 갖는 TF-IDF를 도출한다. 용어의 TF 값은 각 문서에서 해당 용어가 출현한 빈도수를 나타내며 IDF 값은 해당 용어를 포함하는 문서가 전체 문서에서 차지하는 비율에 로그를 취한 값으로 계산된다. TF-IDF 값은 TF 값과 IDF 값을 곱하여 도출하며, <Figure 7(a)>는 “Sentence 1”의 주요 용어에 대한 TF-IDF 값의 예를 보이고 있다.

<Figure 7(a)>에서 “bed”, “clean”, “cover”, “dirty”, 그리고 “small”은 모두 주제 “Room”과 관련된 용어이며, 이들의 TF-IDF 값의 합은 1.5이다. 따라서 본 연구에서는 “Sentence 1”이 주제 “Room”에 대해 갖는 TF-IDF 문장 가중치를 1.5로 산출한다(Figure 7(b)). 마찬가지 원리로 “Sentence 1”이 주제 “Service”에 대해 갖는 TF-IDF 문장 가중치는 0.5가 됨을 알 수 있다. 이러한 방식으로 도출된 TF-IDF 문장 가중치는 문장의 길이가 길어질수록 증가하는 경향을 갖게 되므로, 전체 문장의 상대적 비교를 위해서는 절대적 TF-IDF 값을 상대적인 비율로 변환하는 과정이 필요하다. 따라서 각 문장의 주제별 TF-IDF 값을 전체 주제에 대한 TF-IDF 값의 총합으로 나누어 정규화한 값을 이후 분석에 사용한다. <Figure 7(c)>에서 “Sentence 1”은 “Room”에 대한 내용 75%와 “Service”에 대한 내용 25%로 구성되어 있다.

이렇게 산출된 가중치, 즉 각 문장의 주제별 정규화된 TF-IDF 값을 활용하여 각 주제별로 리뷰를 재구성하며, 이 과정은 <Figure 8>을 통해 확인할 수 있다. <Figure 8(a)>는 8개 문장에 대해 각 주제별로 정규화된 TF-IDF 값을 산출한



(a) Calculating Sentence TF/IDF from Term TF/IDF

	Room	Food	Service	Location
Sentence 1	1.5	0	0.5	0
…		…		

(b) Sentence TF/IDF

	Room	Food	Service	Location
Sentence 1	0.75	0	0.25	0
…		…		

(c) Normalized Sentence TF/IDF

〈Figure 7〉 TF-IDF Value of Each Subject

Sentence	Weight (Room)	Weight (Service)	…
Sentence 1	0.75	0.25	
Sentence 2	0	0.8	
Sentence 3	0.58	0.39	
Sentence 4	0.47	0.44	
Sentence 5	0.30	0.12	…
Sentence 6	0.21	0.20	
Sentence 7	0.10	0.65	
Sentence 8	0.64	0	

(a) Normalized Sentence TF/IDF

Sentence	Weight	Priority
Sentence 8	0.64	1
Sentence 3	0.58	2
Sentence 1	0.50	3
Sentence 4	0.47	4

(b) Sentences in Subject "Room"

Sentence	Weight	Priority
Sentence 2	0.8	1
Sentence 7	0.65	2
Sentence 4	0.44	3

(c) Sentences in Subject "Service"

〈Figure 8〉 Normalized Sentence TF-IDF for Subject "Room" and "Service"

가상 결과를 보이고 있다. 이 때 특정 임계값 이상의 가중치를 갖는 문장만을 선별하여 각 주제별 문장 집합을 구성할 수 있다. 예를 들어 가중치를 0.4로 설정한 경우, Sentences 1, 3, 4, 8은 "Room"과 관련있는 문서로 파악되며(Figure 8(b)), Sentences 2, 4, 7은 "Service"와 관련있는 문서로

파악된다(Figure 8(c)). 그림에서 Sentence 4는 “Room”과 “Service”의 두 가지 주제 모두에 대응되며, Sentence 5는 어떠한 주제에도 대응되지 않음을 알 수 있다.

3.4 주제 균형 요약문 생성

본 절은 <Figure 2>의 Phase 3의 과정을 소개한다. 각 주제별 문장 집합으로부터 문장 간 유사도를 기준으로 각 주제를 대표할 수 있는 문장을 추출하고, 이를 종합하여 주제별 문서의 요약문을 생성한다. 문장 간 유사도 도출을 위해 우선 각 문장을 벡터로 변환한다. 문장의 벡터 변환은 용어 임베딩에 사용된 Word2Vec을 문서 임베딩에 확장한 Sen2Vec을 활용하여 수행한다. Sen2Vec은 문장을 구성하는 모든 용어를 Word2Vec을 통해 용어 벡터로 변환한 후, 차원별 벡터의 평균을 구하여 그 값을 해당 차원의 값으로 갖는 하나의 문장 벡터를 도출한다. 이를 통해 도출된 벡터는 용어 벡터와 동일한 차원의 벡터 공간에 표현이 되고, 형태 역시 용어 벡터와 동일하게 구성된다. 용어 벡터를 문장 벡터로 변환하는 가상 예가 <Figure 9>에 나타나 있다. <Figure 9>는 “Sentence 1”이 6개의 용어로 구성되어 있는 경우를 가정하며, 각 용어는 100 차원의 벡터로 표현되어 있다. 6개 용어에 대해 각 차원별 값의 평균을 구함으로써 “Sentence 1”을 하나의 100차원 벡터로 표현할 수 있다.

Sen2Vec을 통해 벡터로 사상된 각 문장에 대해, 문장 벡터 간 코사인 유사도를 산출하여 문장 간 유사도 행렬을 도출할 수 있다. 문장 벡터 간 코사인 유사도 도출 역시 용어 벡터 간 유사도 도출과 동일한 과정을 통해 진행된다. 문장 간 유사도 행렬을 도출한 후, 해당 행렬을 특정 유사도 임계값을 기준으로 유사도가 임계값 이상일 경우 1로, 미만인 경우에는 0으로 변환하여 이진 행렬을 새롭게 구축한다. <Figure 10>은 주제 “Room”의 문장 간 유사도 행렬을 임계값 0.4를 기준으로 이진 행렬로 변환한 결과를 나타낸다.

	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	...	Dimension N
Term 1	8	3	2	7	1		8
Term 2	7	6	2	1	3		2
Term 3	1	8	7	1	2		2
Term 4	1	9	9	8	3	...	2
Term 5	0	1	6	0	3		8
Term 6	4	2	6	8	1		4

	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	...	Dimension N
Sentence 1	3.6	4.9	5.4	4.1	2.1	...	4.4

〈Figure 9〉 Calculating Sentence Vectors from Word Vectors

	Snt. 8	Snt. 3	Snt. 1	Snt. 4	Snt. 11	Snt. 12	Snt. 13
Snt. 8	1	0.1	0	**0.5**	-0.2	-0.3	0.1
Snt. 3	0.1	1	-0.5	0.2	**0.4**	**0.6**	0
Snt. 1	0	-0.5	1	0.2	-0.8	-0.5	0.3
Snt. 4	**0.5**	0.2	0.2	1	-0.1	0	**0.4**
Snt. 11	-0.2	**0.4**	-0.8	-0.1	1	**0.4**	-0.3
Snt. 12	-0.3	**0.6**	-0.5	0	**0.4**	1	0.3
Snt. 13	0.2	0	0.3	**0.4**	-0.3	0.3	1

	Snt. 8	Snt. 3	Snt. 1	Snt. 4	Snt. 11	Snt. 12	Snt. 13
Snt. 8	1	0	0	1	0	0	0
Snt. 3	0	1	0	0	1	1	0
Snt. 1	0	0	1	0	0	0	0
Snt. 4	1	0	0	1	0	0	1
Snt. 11	0	1	0	0	1	1	0
Snt. 12	0	1	0	0	1	1	0
Snt. 13	0	0	0	1	0	0	1

〈Figure 10〉 Document Similarity Matrix

원문에 포함된 내용 중 최대한 다양한 주제를 포함하면서 요약문 내 문장 간의 내용 중복을 최소화하기 위해, 본 연구에서는 요약문 평가의 지표로 최근 발표된 개념인 완전성(Completeness)와 간결성(Succinctness)에 기반을 두어 문서 요약을 수행한다(Ko and Kim, 2018). 완전성은 요약문이 원문의 내용을 최대한 누락없이 포함해야 한다는 성질이고, 간결성은 요약문 내용의 중복이 최소가 되야 한다는 성질이다. 완전성과 간결성은 서로 상충 관계(Trade off)에 있는 개념으로, 두 가지 지표는 문장 간 유사도에 기반하여 측정된다. 본 연구에서는 두 문장 간 유사성 여부를 이전 단계에서 도출한 이진 행렬에 기반을 두어 판단한다. 즉 이진 행렬에서 임의의 문장과 유사도가 1로 표현되는 문장은 해당 문장과 유사한 문장인 것으로 간주한다. 따라서 “Sentence 1”이 요약문에 포함될 경우 이 문장과 유사한 다른 문장은 요약문에 포함시키지 않음으로써 요약문의 내용 중복을 최소화할 수 있다.

본 연구에서는 이와 같이 유사한 내용을 갖는 문장 집합 중 요약문에 포함되는 문장을 “Selected Sentence”, 그리고 이 문장과 유사한 것으로 인정되어 굳이 요약문에 포함될 필요가 없는 문장을 “Covered Sentence”로 명명한다. 구체적으로 제안 방법론은 각 주제별 문장 집합 중 우선순위가 가장 높은 문장을 “Selected Sentence”로 선정하여 요약문에 추가하고, 이 문장을 포함하여 이 문장과 유사한 것으로 인정되는 문장들을 “Covered Sentence”로 선정하여 이진 행렬에서 제거한다. 다음으로 이진 행렬에 남아있는 문장들 중 우선순위가 가장 높은 문장을 다시 “Selected Sentence”로 선정하고, 이진 행렬에 있는 모든 문장이 제거될 때까지 위의 과정을 반복 수행한다. 이처럼 주제별 우선순위 및 문장 간 유사도에 기반하여 주제별 요약문을 생성하는 가상 예가 <Figure 11>에 나타나 있다.

<Figure 11>의 좌측 행렬은 <Figure 10>의 우측에 나타난 이진 행렬과 동일하다. 또한 행렬 내의 모든 문장은 해당 주제 내 각 문장의 우선순위에 따라 정렬되어 있다. 제안 방법론에 따라 우선순위가 가장 높은 “Snt. 8”이 가장 처음 요약문에 포함되며, 이와 동시에 “Snt. 8”이 Cover하는 “Snt 4, 8”은 행렬의 행과 열에서 제거된다. 다음으로 “Snt. 3”이 요약문에 포함되며, “Snt 3, 11, 12”가 행렬에서 제거된다. 다음으로 “Snt. 1”이 요약문에 포함되면서 행렬에서 제거되고, 마지막으로 “Snt. 13”이 요약문에 포함되면서 행렬에서 제거된다. 결과적으로 총 7개 문장 중 “Snt.

	Snt. 8	Snt. 3	Snt. 1	Snt. 4	Snt. 11	Snt. 12	Snt. 13
Snt. 8	1	0	0	1	0	0	0
Snt. 3	0	1	0	0	1	1	0
Snt. 1	0	0	1	0	0	0	0
Snt. 4	1	0	0	1	0	0	1
Snt. 11	0	1	0	0	1	1	0
Snt. 12	0	1	0	0	1	1	0
Snt. 13	0	0	0	1	0	0	1

	Selected	Covered
Iteration 1	Snt. 8	Snt. 4, 8
Iteration 2	Snt. 3	Snt. 3, 11, 12
Iteration 3	Snt. 1	Snt. 1
Iteration 4	Snt. 13	Snt. 13

〈Figure 11〉 Process of Summary Generation

1, 3, 8, 13”의 4개 문장이 요약문에 포함되었으며, “Snt. 4, 11, 12”의 세 개 문장은 요약문에 이미 포함된 문장과 유사한 것으로 인정되어 요약문에 포함되지 않았다. 이러한 방식으로 해당 주제에 대해 완전성과 간결성을 충족하는 요약문을 작성할 수 있으며, 각 주제별 요약문을 통합하여 전체 요약문을 구성할 수 있다.

4. 실험

4.1 실험 개요

본 절에서는 3장에서 소개한 제안 방법론의 검증을 위한 실험 환경과 데이터를 소개한다. 실험을 위해 여행 정보 사이트인 TripAdvisor로부터 리뷰를 수집하고, 이들 리뷰를 문장으로 분해하여 총 50,000건의 문장 집합을 구성하였다. Word2Vec 학습에는 50,000건의 문장을 모두 사용하였으며, 리뷰의 대상이 된 호텔 중 가장 많은 수의 문장을 포함하고 있는 New York 소재 “E” 호텔을 선정하여 해당 호텔의 리뷰 문장 23,087개에 대한 요약을 수행하였다. 실험에서 토픽 모델링과 TF-IDF 가중치 도출은 SAS Enterprise Miner 14.2를 사용했으며, 용어 및 문장 간 유사도 도출, 그리고 문서 요약은 Python 3.6을 사용하였다.

4.2 주제별 용어 사전 구축 결과

본 절에서는 호텔 리뷰가 내포한 다양한 주제를 확인하고, 이를 바탕으로 주제별 용어 사전을 구축한 결과를 제시한다. 우선 호텔 리뷰가 내포한 주제를 파악하기 위해 사전 구축용 리뷰 데이터로부터 20개의 토픽을 도출하고, 각 토픽 별 주요 키워드를 확인하였다. <Figure 12>는 50,000개의 호텔 리뷰 문장으로부터 도출한 토픽 키워드를 나타낸다.

다음으로 리뷰 데이터를 구성하는 용어의 토픽 별 가중치를 확인하고, 각 토픽별로 상위 가중치를 갖는 100개의 후보 용어를 추출하였다. <Figure 13>은 토픽 별로 추출된 후보 용어의 일부를 보여준다. 이렇게 도출된 2,000개의 용어로부터 “Food”, “Location”, “Room”, “Service” 의 4개의 주제를 식별하였고, 각 주제별로 종합된 용어에 대한 선별 작업을 거쳐 해당 주제를 대표할 수 있는 10개의 시드 용어를 선정하였다. 주제별 시드 용어는 <Figure 14>에서 확인할 수 있다.

Topic ID	Topic Keywords	Topic ID	Topic Keywords
Topic 01	friendly, staff, accommodating, front, professional	Topic 11	noise, traffic, loud, noisy, lobby
Topic 02	breakfast, buffet, shower, bathroom, cheese	Topic 12	helpful, concierge, staff, pleasant, desk
Topic 03	desk, front, concierge, housekeeping, clerk	Topic 13	beds, double, queen, comfy, bed
Topic 04	bathroom, modern, bed, lobby, sink	Topic 14	subway, station, blocks, block, access
Topic 05	walk, station, avenue, district, mins	Topic 15	coffee, tea, breakfast, lounge, wine
Topic 06	bar, drinks, drink, food, lounge	Topic 16	manhattan, midtown, east, downtown, mid-town
Topic 07	bed, king, comfy, pillows, sofa	Topic 17	restaurant, breakfast, lobby, food, bar
Topic 08	walking, distance, station, places, modern	Topic 18	shower, bath, bathroom, tub, toilet
Topic 09	lobby, wifi, internet, modern, coffee	Topic 19	food, modern, breakfast, drinks, places
Topic 10	staff, front, desk, pleasant, accommodating	Topic 20	restaurants, nearby, bars, subway, block

〈Figure 12〉 Topic Keywords for 20 Topics

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	…	Term 98	Term 99	Term 100
Topic 01	friendly	staff	front	professional	accommodating	desk		housekeepers	reach	receptionist
Topic 02	breakfast	buffet	shower	bathroom	cheese	bed		pleasant	staff	professional
Topic 03	desk	front	concierge	clerk	housekeeping	bell		unprofessional	personalized	furnished
Topic 04	bathroom	modern	bed	lobby	sink	towels		channels	maid	twin
Topic 05	walk	station	avenue	district	mins	places		effort	museums	galleries
Topic 06	bar	drinks	drink	food	wine	lounge		sofa	cakes	noises
Topic 07	bed	king	comfy	pillows	sofa	tv		trains	mattresses	freezer
Topic 08	walking	distance	station	places	modern	district		showers	robes	bellman
Topic 09	lobby	wifi	internet	modern	coffee	shower		breakfasts	bistro	maid
Topic 10	staff	front	desk	pleasant	accommodating	attentive	…	cheese	district	village
Topic 11	noise	traffic	loud	noisy	lobby	beds		balcony	bathtub	direction
Topic 12	helpful	concierge	staff	desk	pleasant	restaurant		cordial	considerate	responsive
Topic 13	beds	double	queen	comfy	bed	pillows		train	greeting	maid
Topic 14	subway	station	blocks	block	access	lines		refrigerator	toiletries	services
Topic 15	coffee	tea	breakfast	lounge	tv	wine		king	sofa	avenue
Topic 16	manhattan	midtown	east	downtown	mid-town	modern		transportation	theaters	stadium
Topic 17	restaurant	breakfast	lobby	food	bar	eat		refrigerator	bread	screen
Topic 18	shower	bath	bathroom	tub	toilet	curtain		navigate	furnished	classic
Topic 19	food	modern	breakfast	drinks	places	eat		comfy	spotlessly	channels
Topic 20	restaurants	nearby	bars	subway	block	shops		destinations	vicinity	sofa

〈Figure 13〉 Seed Term Candidates

	Food	**Location**	**Room**	**Service**
Seed Terms	delicious	nearby	cozy	friendly
	breakfast	block	towels	staff
	eat	subway	bed	lobby
	restaurant	distance	shower	customer
	drinks	terminal	noise	housekeeping
	bread	access	deluxe	bell
	coffee	traffic	wifi	concierge
	buffet	midtown	tv	doorman
	menu	avenue	tub	desk
	dining	walk	sofa	attitude

〈Figure 14〉 Seed Terms of Each Subject

<Figure 14>에 나타난 4개 주제, 40개 시드 용어에 대해, Word2Vec을 이용한 용어 확장을 수행하여 주제별 용어 사전을 구축하였다. 50,000개의 문장에 포함된 용어 간의 코사인 유사도를 도출하고, 시드 용어 별로 높은 유사도를 갖는 50개의 용어를 추출한 뒤, 용어의 선별 과정을 통해 주제별로 100개씩의 용어를 포함하는 사전을 생성하였다. <Figure 15>는 시드 용어와 유사한 용어를 추출하는 용어 확장 과정의 일부이며, <Figure 16>은 이러한 과정을 통해 구축한 주제 용어 사전 전체를 나타낸다.

Subject	Seed Terms	Similar Terms					
		Top 1	Top 2	Top 3	…	Top 49	Top 50
Food	delicious	tasty	yummy	complementary		doeuvres	nibbles
	breakfast	buffet	continental	food	…	newspapers	cappuccino
	eat	eating	dine	drink		downstairs	anytime
Room	bed	beds	mattress	pillows		loungers	cabinet
	shower	tub	pressure	bath	…	stall	air-conditioning
	deluxe	superior	king	junior		connecting	rm
Service	friendly	polite	courteous	gracious		exceedingly	friendliness
	staff	doormen	staffs	employees	…	concierges	guys
	lobby	foyer	lounge	evenings		crowd	handy
Location	nearby	eateries	delis	diners		alternatives	plethora
	block	blocks	steps	yards	…	police	road
	subway	metro	subways	stops		hub	min

〈Figure 15〉 Term Expansion by Word2Vec

Room

air-conditioning	couch	linen	signal
airy	cozy	linens	sink
audible	curtain	loud	soap
balcony	curtains	mattress	soaps
bath	decoration	mattresses	sofa
bathrobes	deluxe	mirror	sofabed
bathroom	double	mirrors	sound
bathtub	dresser	modern	soundproofing
bed	dryer	noise	sounds
bedding	duvet	noises	spotlessly
bedroom	duvets	noisy	suite
beds	faucet	one-bedroom	table
blankets	flat-screen	pillows	television
chair	fluffy	plush	tidied
chairs	freezer	queen	toilet
channels	furnished	radiator	toiletries
classic	furnishings	refrigerator	towel
closet	hangers	replenished	towels
closets	hd	robe	tub
comforter	heat	robes	tv
comforters	internet	roomy	twin
comfy	king	screen	ventilation
compact	kingsize	sheets	wardrobe
connection	king-size	shower	wifi
cosy	king-sized	showers	wi-fi

Food

appetizers	cakes	espresso	plates
apples	cappuccino	food	porridge
ate	cereal	fries	potatoes
bacon	cereals	fruit	pub
bagel	cheese	fruits	pubs
bagels	chocolate	grill	restaurant
bagles	cider	ham	restaurants
baked	cocktail	iced	salad
bar	cocktails	juice	sausage
bars	coffee	juices	sausages
beer	cooked	liquor	scrambled
beverages	cookies	meal	snack
biscuits	croissants	meals	snacks
bistro	delicious	meats	soda
boiled	dining	menu	tasty
bread	dishes	milk	tea
breads	donuts	minibar	teas
breakfast	doughnuts	muffins	toast
breakfasts	drink	non-alcoholic	vegetarian
brunch	drinks	oatmeal	waffle
buffet	eat	omelets	waffles
burger	eateries	pancake	wine
burgers	eating	pancakes	yogurt
cafe	egg	pastries	yogurts
cafes	eggs	pastry	yummy

Service

24hrs	concierges	helpful	professionalism
accommodating	considerate	helpfulness	prompt
accommodations	cordial	hops	reception
accomodating	courteous	hospitable	receptionist
accomodations	customer	hospitality	receptionists
answering	dedicated	housekeeper	responsive
approachable	dedication	housekeepers	rude
assistance	desk	housekeeping	security
attention	detail	info	services
attentive	doorman	information	sincere
attitude	doormen	informative	sincerely
bell	effort	kindness	skills
bellboys	employee	knowledgeable	smiling
bellhop	employees	lobby	staff
bellhops	focused	lounge	staffs
bellman	foyer	maid	supervisor
bellmen	friendliness	maids	support
caring	friendly	maintenance	understanding
cleaning	front	manner	unfailingly
clerk	frontdesk	obliging	unfriendly
clerks	genuine	owner	unhelpful
commitment	genuinely	personalized	unprofessional
competent	gracious	pleasant	warmth
conceirge	greeted	polite	welcoming
concierge	greeting	professional	workers

Location

access	hike	port	supermarket
accessible	hill	proximity	terminal
ave	landmarks	railroad	theater
avenue	lines	railway	theaters
avenues	locale	reach	theatres
bike	locations	riverside	timesquare
block	mall	road	traffic
blocks	manhattan	rockerfeller	train
bus	map	route	trains
buses	markets	routes	transit
centrally	meters	seaport	transport
chinatown	metres	shops	transportation
close-by	metro	sites	tunnel
closest	midst	soho	underground
crosstown	midtown	south	uptown
cvs	mid-town	squares	venues
destinations	mins	stadium	vicinity
direct	museums	station	village
direction	navigate	stations	walk
distance	nearby	steps	walkable
district	nearest	stops	walked
downtown	north	stores	walking
east	pharmacy	stroll	walks
emporium	places	subway	west
galleries	plaza	subways	yards

〈Figure 16〉 Subject Dictionaries for “Room”, “Food”, “Service”, and “Location”

4.3 리뷰의 주제별 재구성 결과

본 절에서는 23,087개의 리뷰 문장에 대한 주제별 TF-IDF 가중치를 도출하고, 가중치의 정규화를 통해 각 문장에 주제를 할당한 결과를 소개한다. 400개의 용어를 리뷰 문장 집합에 적용하여 용어의 TF 값과 IDF 값을 도출하였고, 이를 바탕으로 각 문장의 주제별 TF-IDF 값을 도출하였다. 전체 23,087개의 문장 중 주제 관련 용어를 하나도 포함하지 않은 문장을 제거하고 총 9,985개의 문장에 대해서 4개의 주제별 TF-IDF값을 도출하였고, 정규화를 통해 각 가중치를 0에서 1 사이의 값을 갖는 상대적인 비율로 변환하였다. 각 문장에 대한 정규화된 가중치 산출 결과의 일부가 <Figure 17>에 나타나 있다.

	Food	Location	Room	Service
Snt. 1	0	0	0	1
...	...	...	...	...
Snt. 9108	0	0.473	0.527	0
Snt. 9109	0.476	0	0.524	0
Snt. 9110	0	0.477	0	0.523
Snt. 9111	0.479	0	0	0.521
Snt. 9112	0	0.48	0	0.52
Snt. 9113	0	0	0.48	0.52
Snt. 9114	0	0	0.481	0.519
Snt. 9115	0.522	0.478	0	0
Snt. 9116	0	0	0.481	0.519
Snt. 9117	0.481	0	0	0.518
...	...	...	...	...
Snt. 9985	0	0.885	0.115	0

〈Figure 17〉 Normalized Sentence Weight for Each Subject

<Figure 17>의 값을 근거로 9,985개의 문장을 주제별로 재구성하였다. 임계값은 0 초과로 설정하였으며, 한 문장이 0이 아닌 값을 여러 개 갖는 경우 해당 문장은 여러 주제에 동시에 포함시켰다. 그 결과 “Food”, “Location”, “Room”, 그리고 “Service”의 주제에 대한 문서 집합은 각각 1,558개, 1,957개, 4,119개, 그리고 3,853개의 문장으로 구성되었다. <Figure 18>은 이들 문서 집합 중 “Room”에 관련된 4,119개 문장의 일부를 나타내며, 해당 주제와 대응되는 정도에 따른 우선 순위 기준으로 정렬되어 있다.

Snt. No.	Contents	Priority
Snt. 4942	Wifi was complimentary, the bed and linens were super comfortable, the shower was powerful, the bathroom was well-stocked with plush towels and soaps/shampoos/lotions, and the flat-screen TV provided the news I needed at the end of my long days	1
Snt. 39	Although the mini suite has some functional bathroom problems, like a tub faucet that is unaccessible unless you get into the tub and a shallow surface mounted lavatory with a quirky faucet control that is likely to get your clothes wet util you learn to be very very careful when you turn it on; the room is ample and luxuriously comfortable if you overlook the here and there unglued wall paper joints	2
Snt. 4895	Just enough space for 2 double beds but really it was the bathroom that was ridiculous - the door is just shy of touching the toilet, the sink faucet is so low it practically touches the sink basin making it tough to wash your hands, and there is no counter space to put any toiletries, which makes it ridiculously tough to get ready in the morning without risking your belongings falling in the toilet	3
...	...	...
Snt. 9059	Excellent location, very welcoming, efficient staff, very pleasant decor, delicious in-hotel restaurants with very good service and excellent food; room pleasant and only problem was window caulking missing so cold air came pouring in, also city noise	4117
Snt. 8820	The staff is excellent, the room was clean and comfortable, the Lobby Lounge is welcoming and there is a Roof Top Lounge, a wonderful restaurant, a pool (seasonal) and a fitness facility	4118
Snt. 9985	Comfortable Hotel in great location, block from Central Park, actually directly across the street from Lincoln Center--always a cab immediately available as it is situated on a one-way, short block, low traffic street	4119

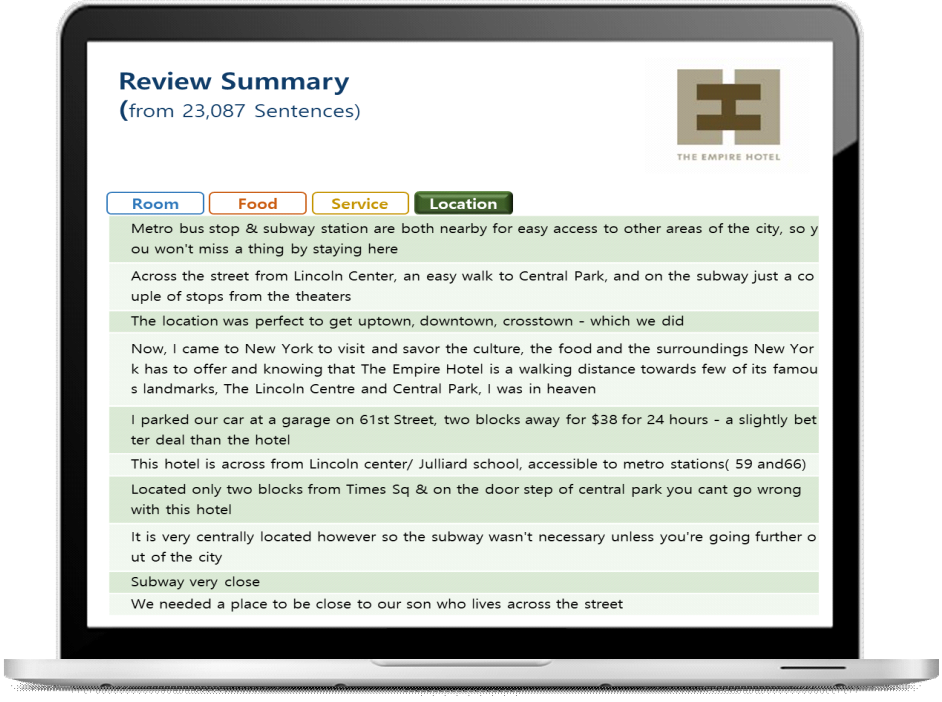
〈Figure 18〉 Sentence Set in Subject “Room”

4.4 주제 균형 요약문 생성 결과

본 절에서는 주제별 문서 집합으로부터 주제별 요약문을 도출한 결과를 제시한다. 주제별 문장을 Sen2Vec을 이용해 문장 벡터로 변환하고, 벡터 간 코사인 유사도를 계산하여 문장의 유사도 행렬을 도출하였다. 또한 이를 이진 유사도 행렬로 변환한 뒤 완전성과 간결성 측면의 문서 요약을 진행하였다. 요약문이 포함하는 문장의 수는 유사도 임계값에 따라 달라지며, 임계값이 높을수록 요약문의 문장 수도 증가하는 경향을 보인다. 유사도 임계값에 따른 각 주제별 요약문의 문장 수 및 각 주제별 구성 비율은 <Table 1>과 같다. 또한 이들 유사도 임계값 0.2 기준으로 도출한 주제별 요약문 중 "Location" 주제의 요약문 전체가 <Figure 19>에 제시되어 있다.

〈Table 1〉 The Number of Sentences in Summary (Our Approach)

		Food	Location	Room	Service	Total
Original		1,558	1,957	4,119	3,853	11,487
Summary	W0.2	10 (18.52%)	10 (18.52%)	21 (38.89%)	13 (24.07%)	54 (100%)
	W0.5	44 (12.87%)	57 (16.67%)	118 (34.50%)	123 (35.96%)	342 (100%)
	W0.8	403 (9.68%)	722 (17.34%)	1,683 (40.43%)	1,355 (32.55%)	4,163 (100%)



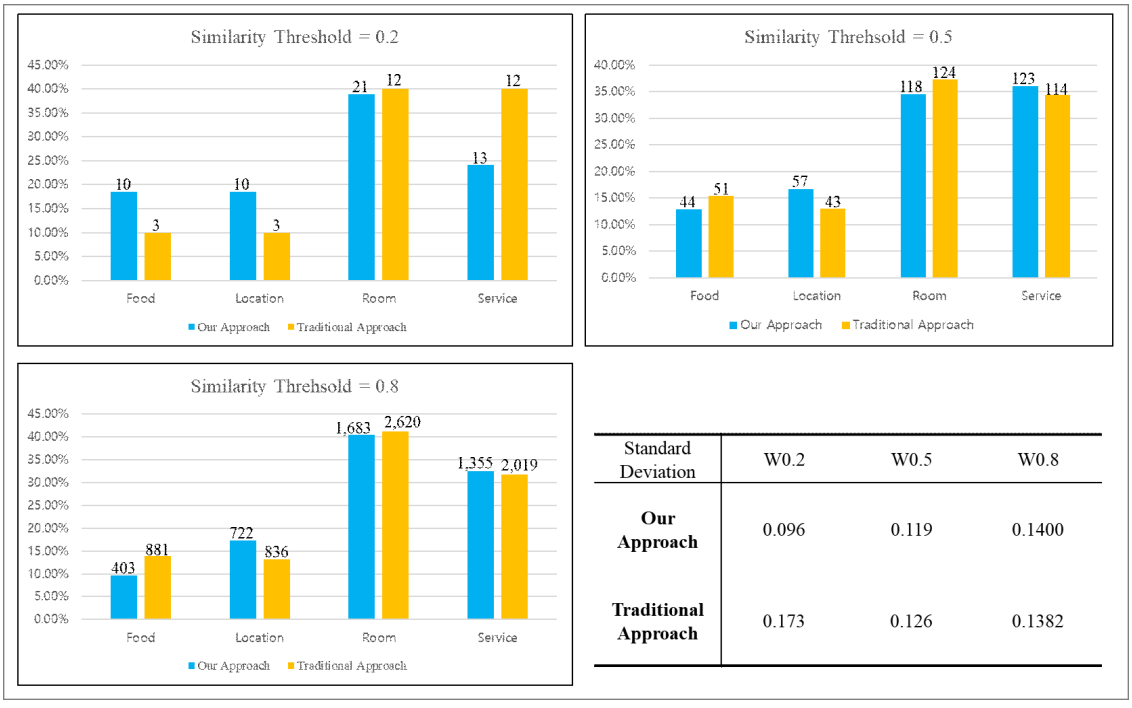
〈Figure 19〉 Summary Report of Subject "Location" for Hotel "E"

4.5 요약문의 주제별 구성 비율 평가

본 절에서는 제안 방법론을 통해 도출한 요약문의 주제별 균형 정도를 평가하기 위한 분석을 수행한다. 구체적으로는 각 주제별 문장의 구성 비율을 비교하고자 하며, 각 주제별 문장 수의 표준 편차도 함께 비교한다. 비교를 위해 주제에 대한 고려 없이 한꺼번에 요약문을 도출하는 방식의 실험도 수행하였으며, 유사도 임계값은 <Table 1>과 마찬가지로 0.2, 0.5, 그리고 0.8을 적용하였다. 비교 실험의 유사도 임계값에 따른 각 주제별 요약문의 문장 수 및 각 주제별 구성 비율은 <Table 2>과 같다. 또한 <Table 1>과 <Table 2>의 내용을 통합하여 그래프로 비교한 결과가 <Figure 20>에 제시되어 있다.

〈Table 2〉 The Number of Sentences in Summary (Traditional Approach)

		Food	Location	Room	Service	Total
Original		1,558	1,957	4,119	3,853	11,487
Summary	W0.2	3 (10%)	3 (10%)	12 (40%)	12 (40%)	30 (100%)
	W0.5	51 (15.36%)	43 (12.95%)	124 (37.35%)	114 (34.34%)	332 (100%)
	W0.8	881 (13.86%)	836 (13.15%)	2,620 (41.22%)	2,019 (31.77%)	6,356 (100%)



Standard Deviation	W0.2	W0.5	W0.8
Our Approach	0.096	0.119	0.1400
Traditional Approach	0.173	0.126	0.1382

〈Figure 20〉 Comparison of Subject Distribution

<Figure 20>의 그래프에서 Y축은 요약문에서 각 주제의 문장 수가 요약문 전체 문장 수에서 차지하는 비율을 나타내며, 그래프 위에 표시된 수치는 요약문에 포함된 각 주제의 문장 수를 나타낸다. 두 기법 모두 “Food”와 “Location”에 비해 “Room”과 “Service”의 문장 수가 높게 나타나며, 이는 원문에서의 주제별 분포의 영향을 받은 것으로 보인다. 또한 기존 기법의 경우 임계값의 변화에 따른 주제별 구성 비율의 표준 편차가 특정한 양상을 보이지 않는 반면, 제안 방법론의 경우는 임계값의 증가에 따라 주제별 구성 비율의 표준 편차도 증가하는 양상을 보였다. 이로 인해 임계값이 0.2인 경우는 주제별 구성 비율의 표준편차가 0.096 < 0.1732로 제안 방법론이 현저히 낮게 나타나는 반면, 임계값이 0.8인 경우는 표준편차가 0.1400 > 0.1382로 제안 방법론이 오히려 높게 나타났다. 하지만 임계값이 0.8인 경우의 요약문의 문장 수가 원문의 문장 수의 거의 절반에 해당되므로, 이렇게 큰 임계값을 사용하여 요약이 이루어지는 경우는 실제로 존재하기 어렵다. 이에 비해 임계값이 0.2인 경우는 요약문의 문장 수가 두 기법 각각 54개와 30개로 현실적인 수준의 요약이 이루어진 것으로 보인다. 따라서 현실적인 수준의 요약에서 제안 방법론이 기존 방법론에 비해 주제의 구성 비율이 고르게 나타남을 알 수 있다.

5. 결론

최근 방대한 양의 텍스트 데이터를 자동으로 요약하는 기술에 대한 연구가 활발하게 수행되고 있으며, 특히 다양한 관점에서 요약문의 품질을 향상시키기 위한 시도가 최근 이루어지고 있다. 이에 본 연구에서는 주제별 용어 사전 구축을 통해 문서가 포함한 다양한 주제를 식별하고, 대상 문서들을 주제별 문서 집합으로 분할한 후, 각 주제별 요약문을 도출하고 이를 통합함으로써 주제별로 균형이 이루어진 요약문을 도출하는 문서 요약 방법을 새롭게 제시했다. 또한 제안 방법론을 실제 호텔 리뷰 데이터에 적용한 실험을 통해, 제안 방법론을 통해 도출한 요약문의 주제별 구성 비율이 기존 방법론에 의해 도출된 요약문에 비해 고르게 나타남을 확인하였다.

본 연구의 기여는 다음과 같다. 우선 본 연구는 낮은 빈도로 인해 소외될 수 있는 주제도 요약문에 포함시킬 수 있는 자동 요약 기법을 새롭게 제안하였으며, 이는 본 연구의 학술적 기여로 인정받을 수 있다. 또한 이를 위해 완전성과 간결성이 높은 요약문을 생성하는 기법을 구체적으로 제시한 점도 새로운 시도로 인정받을 수 있을 것이다. 또한 주제별 용어 사전을 구축하는 과정에서 Word2Vec을 적용하여 비교적 수월한 방법으로 양질의 사전을 구축하였으며, 이러한 방식은 유사한 주제를 다루는 후속 연구에서도 충분히 활용할 수 있을 것이다.

하지만 본 연구는 다음과 같은 측면에서 향후 보완이 필요하다. 우선, 제안 방법론은 문장 간 유사도에 기반을 두어 동작하기 때문에, 유사도의 임계값이 결과에 미치는 영향은 절대적이라고 할 수 있다. 따라서 향후 연구에서는 유사도 임계값의 변화가 요약문의 품질에 어떤 영향을 미치는지에 대한 엄밀한 검토가 이루어져야 한다. 이와 관련하여 본 논문에서는 제안 방법론을 통해 도출된 요약문의 품질을 요약문의 주제별 구성 비율 측면에서만 평가하였다. 하지만 최근 국내외의 여러 연구를 통해 요약문의 품질 측정 방법이 소개되고 있으므로, 다양한 평가 기준에

의해 제안 방법론에 대한 성능 평가를 수행할 필요가 있다.

참고문헌(References)

Bingham, E. and H. Mannila, "Random projection in dimesionality reduction: applications to image and text," *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, (2001), 245~250.

Chen, Y. and M. Bansal, "Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, (2018), 675~686.

Chorpa, S., M. Auli and A. Rush, "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks," *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (2016), 93~98.

Eduard, H., and C. Lin, "Automated text summarization and the SUMMARIST system," *Proceedings of a workshop*, (1998), 197~214.

Eduard, H., *The Oxford Handbook of Computational Linguistics 2nd edition*, Oxford University Press, Oxford, 2015.

Erk, K. and S. Pado, "A structured vector space model for word meaning in context," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (2008), 897~906.

Gao, J., Y. He, X. Zhang and Y. Xia, "Duplicate Short Text Detection Based on Word2Vec," *2017 8th IEEE International Conference on Software Engineering and Service Science*, (2017), 33~38.

Goldstein, J., M. Kantrowitz, V. Mittal and J. Carbonell, "Summarizing text documents: sentence selection and evaluation metrics," *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, (1999), 121~128.

Gong, Y. and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, (2001), 19~25.

Gupta, V. and G. Lehal, "A Survey of Text Summarization Extractive Techniques," *JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE*, Vol.2, No.3(2010), 258~268.

Joel, L. N., A. Alex and C, Kaestner, "Automatic Text Summarization Using a Machine Learning Approach," *Brazilian Symposium on Artificial Intelligence*, (2002), 205~215.

Kageback, M., O. Mogren, N. Tahmasebi and D. Dubhashi, "Extractive Summarization using Continuous Vector Space Models," *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality*, (2014), 31~39.

Kim, J., J. Kim and D. Hwang, "Korean Text Summarization Using an Aggregate Similarity," *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, (2000), 111~118.

Ko, E. and N. Kim, "Automatic Quality Evaluation with Completeness and Succinctness for Text Summarization," *Journal of Intelligence and Information Systems*, Vol.24, No.2(2018), 125~148.

Li, W., X. Xiao, Y. Lyu and Y. Wang, "Improving Neural Abstractive Document Summarization with Structural Regularization," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (2018), 4078~4097.

Marco, B., D. Georgiana and K. German, "Don't count, predict! A Systematic comparison of context-counting vs. context-predicting semantic vectors," *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, (2014), 238~247.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality," *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Vol.2, (2013), 3111~3119.

Mittal, N., B. Agarwal, H. Mantri, R. Goyal and M. Jain, "Extractive Text Summarization," *International Journal of Current Engineering and Technology*, Vol.4, No.2(2014), 870~872.

Mohamed, A. F. and R. Fuji, "GA, MR, FFNN, Pnn and GMM based models for automatic text summarization," *Computer Speech & Language*, Vol.23, (2009), 126~144.

Nallapati, R., B. Zhou, C. Santos, C. Gulcehre and B. Xiang, "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond," *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, (2016), 280~290.

Nenkova, A. and K. Mckewon, "A Survey of Text Summarization Techniques," *Mining Text Data*, (2012), 43~76.

Omer, L., Y. Goldberg and I. Dagan, "Improving Distributional Similarity with Lessons Learned from Word Embeddings," *Transactions of the Association for Computational Linguistics*, Vol.3, (2015), 211~225.

Rachit, A. and B. Ravindran "Latent Dirichlet Allocation and Singular Value Decomposition Based Multi-document Summarization," *2008 8th IEEE International Conference on Data Mining*, (2008), 713~718.

Ramiz, M. A., "A new sentence similarity measure and sentence based extractive technique for automatic text summarization," *Expert Systems with Applications*, Vol.36, (2009), 7764~7772.

Salton, G., A. Wong and C. S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, Vol.18, No.11 (1975), 613~620.

Singhal, A., "Modern Information Retrieval: A Brief Overview," *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, (2001), 35~43.

Sonawane, S., A. Ghotkar and S. Hinge, "Context-Based Multi-document Summarization," *Contemporary Advances in Innovative and Applicable Information Technology*, (2018), 153~165.

Tan, A. "Text Mining: The state of the art and the challenges," *Proceedings of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases*, (1999), 65~70.

Wan, X. and J. Yang, "Multi-document summarization using cluster-based link analysis," *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, (2008), 299~306.

Wen, Z., T. Yoshida and X. Tang, "A comparative study of TF*IDF, LSI and multi-words for text classification," *Expert Systems with Applications*, Vol.38, No.3(2011), 2758~2765.

Yeh, J. Y., H. Ke and W. Yang, "iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network," *Expert Systems with Applications*, Vol.35, (2008), 1451~1462.

Zhang, F., J. Yao and R. Yan, "On the Abstractiveness of Neural Document Summarization," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (2018), 785~790.

Zhang, P. and C. Li, "Automatic text summarization based on sentences clustering and extraction," *2009 2nd IEEE International Conference on Computer Science and Information Technology*, (2009), 167~170.

Abstract

Subject-Balanced Intelligent Text Summarization Scheme

Yeoil Yun* · Eunjung Ko** · Namgyu Kim***

Recently, channels like social media and SNS create enormous amount of data. In all kinds of data, portions of unstructured data which represented as text data has increased geometrically. But there are some difficulties to check all text data, so it is important to access those data rapidly and grasp key points of text. Due to needs of efficient understanding, many studies about text summarization for handling and using tremendous amounts of text data have been proposed. Especially, a lot of summarization methods using machine learning and artificial intelligence algorithms have been proposed lately to generate summary objectively and effectively which called "automatic summarization". However almost text summarization methods proposed up to date construct summary focused on frequency of contents in original documents. Those summaries have a limitation for contain small-weight subjects that mentioned less in original text. If summaries include contents with only major subject, bias occurs and it causes loss of information so that it is hard to ascertain every subject documents have. To avoid those bias, it is possible to summarize in point of balance between topics document have so all subject in document can be ascertained, but still unbalance of distribution between those subjects remains. To retain balance of subjects in summary, it is necessary to consider proportion of every subject documents originally have and also allocate the portion of subjects equally so that even sentences of minor subjects can be included in summary sufficiently.

In this study, we propose "subject-balanced" text summarization method that procure balance between all subjects and minimize omission of low-frequency subjects. For subject-balanced summary, we use two concept of summary evaluation metrics "completeness" and "succinctness". Completeness is the feature that summary should include contents of original documents fully and succinctness means summary has minimum duplication with contents in itself. Proposed method has 3-phases for summarization. First phase is constructing subject term dictionaries. Topic modeling is used for calculating topic-term weight which indicates degrees that each terms are related to each topic. From derived weight, it is possible to figure

* College of Business Administration, Kookmin University
** Graduate School of Business IT, Kookmin University
*** Corresponding Author: Namgyu Kim
College of Business Administration, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 136-702, Korea
Tel: +82-2-910-5425, Fax: +82-2-910-5209, E-mail: ngkim@kookmin.ac.kr

out highly related terms for every topic and subjects of documents can be found from various topic composed similar meaning terms. And then, few terms are selected which represent subject well. In this method, it is called “seed terms”. However, those terms are too small to explain each subject enough, so sufficient similar terms with seed terms are needed for well-constructed subject dictionary. Word2Vec is used for word expansion, finds similar terms with seed terms. Word vectors are created after Word2Vec modeling, and from those vectors, similarity between all terms can be derived by using cosine-similarity. Higher cosine similarity between two terms calculated, higher relationship between two terms defined. So terms that have high similarity values with seed terms for each subjects are selected and filtering those expanded terms subject dictionary is finally constructed. Next phase is allocating subjects to every sentences which original documents have. To grasp contents of all sentences first, frequency analysis is conducted with specific terms that subject dictionaries compose. TF-IDF weight of each subjects are calculated after frequency analysis, and it is possible to figure out how much sentences are explaining about each subjects. However, TF-IDF weight has limitation that the weight can be increased infinitely, so by normalizing TF-IDF weights for every subject sentences have, all values are changed to 0 to 1 values. Then allocating subject for every sentences with maximum TF-IDF weight between all subjects, sentence group are constructed for each subjects finally. Last phase is summary generation parts. Sen2Vec is used to figure out similarity between subject-sentences, and similarity matrix can be formed. By repetitive sentences selecting, it is possible to generate summary that include contents of original documents fully and minimize duplication in summary itself.

For evaluation of proposed method, 50,000 reviews of TripAdvisor are used for constructing subject dictionaries and 23,087 reviews are used for generating summary. Also comparison between proposed method summary and frequency-based summary is performed and as a result, it is verified that summary from proposed method can retain balance of all subject more which documents originally have.

Key Words : Document Summarization, Review Summarization, Text Mining, Topic Modeling, Word Embedding

Received : January 3, 2019 Revised : January 3, 2019 Accepted : May 6, 2019
Publication Type : Conference(Fast-track) Corresponding Author : Namgyu Kim

저 자 소 개

윤 여 일
현재 국민대학교 비즈니스IT전문대학원 석사과정에 재학 중이다. 국민대학교 경영정보학부에서 학사 학위를 취득하였으며, 주요 관심분야는 텍스트 마이닝, 데이터 마이닝, 데이터 처리 등이다.

고 은 정
현재 한국스마트카드에 재직 중이다. 경영정보학부에서 학사 학위를, 국민대학교 비즈니스IT전문대학원에서 석사 학위를 취득하였다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 모빌리티 데이터 분석 등이다.

김 남 규
현재 국민대학교 경영정보학부 교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국지능정보시스템학회 부회장, 한국정보기술응용학회 부회장, 한국경영학회 상임이사, 한국경영정보학회 이사, 한국인터넷정보학회 이사, 한국CRM학회 이사를 역임하였다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝, 데이터 모델링 등이다.