ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2018 June: 24(2): 125~148 http://dx.doi.org/10.13088/jiis.2018.24.2.125

완전성과 간결성을 고려한 텍스트 요약 품질의 자동 평가 기법

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

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

다양한 스마트 기기 및 관련 서비스의 증가에 따라 텍스트 데이터가 폭발적으로 증가하고 있으며, 이로 인해 방대한 문서로부터 필요한 정보만을 추려내는 작업은 더욱 어려워졌다. 따라서 텍스트 데이터로부터 핵심 내용을 자동으로 요약하여 제공할 수 있는 텍스트 자동 요약 기술이 최근 더욱 주목을 받고 있다. 텍스트 요약 기술은 뉴스 요약 서비스, 개인정보 약관 요약 서비스 등을 통해 현업에서도 이미 활발하게 적용되고 있으며, 학계에서도 문서의 주요 요소를 선별하여 제공하는 추출(Extraction) 접근법과 문서의 요소를 발췌한 뒤 이를 조합하여 새로운 문장을 구성하는 생성(Abstraction) 접근법에 따라 많은 연구가 이루어지고 있다. 하지만 문서의 자동 요약 기술에 비해, 자동으로 요약된 문서의 품질을 평가하는 기술은 상대적으로 많은 진전을 이루지 못하였다. 요약문의 품질 평가를 다룬 기존의 대부분의 연구들은 사람이 수작업으로 요약문을 작성하여 이를 기준 문서(Reference Document)로 삼고, 자동 요약문과 기준 문서와의 유사도를 측정하는 방식으로 수행되었다. 하지만 이러한 방식은 기준 문서의 작성 과정에 막대한 시간과 비용이 소요될 뿐 아니라 요약자의 주관에 의해 평가 결과가 다르게 나타날 수 있다는 한계를 갖는다.

한편 이러한 한계를 극복하기 위한 연구도 일부 수행되었는데, 대표적으로 전문에 대해 차원 축소를 수행하고 이렇게 축소된 전문과 자동 요약문의 유사도를 측정하는 기법이 최근 고안된 바 있다. 이 방식은 원문에서 출현 빈도가 높은 어휘가 요약문에 많이 나타날수록 해당 요약문의 품질이 우수한 것으로 평가하게 된다. 하지만 요약이란 본질적으로 많은 내용을 줄여서 표현하면서도 내용의 누락을 최소화하는 것을 의미하므로, 단순히 빈도수에 기반한 “좋은 요약”이 항상 본질적 의미에서의 “좋은 요약”을 의미한다고 보는 것은 무리가 있다. 요약문 품질 평가의 이러한 기존 연구의 한계를 극복하기 위해, 본 연구에서는 요약의 본질에 기반한 자동 품질 평가 방안을 제안한다. 구체적으로 요약문의 문장 중 서로 중복되는 내용이 얼마나 적은지를 나타내는 요소로 간결성(Succinctness) 개념을 정의하고, 원문의 내용 중 요약문에 포함되지 않은 내용이 얼마나 적은지를 나타내는 요소로 완전성(Completeness)을 정의한다. 본 연구에서는 간결성과 완전성의 개념을 적용한 요약문 품질 자동 평가 방법론을 제안하고, 이를 TripAdvisor 사이트 호텔 리뷰의 요약 및 평가에 적용한 실험 결과를 소개한다.

주제어 : 요약문 품질 평가, 텍스트 요약, 텍스트 마이닝, 토픽 모델링

논문접수일 : 2018년 5월 29일 논문수정일 : 2018년 6월 24일 게재확정일 : 2018년 6월 25일
원고유형 : 일반논문 교신저자 : 김남규

* 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017S1A5A2A03067632)

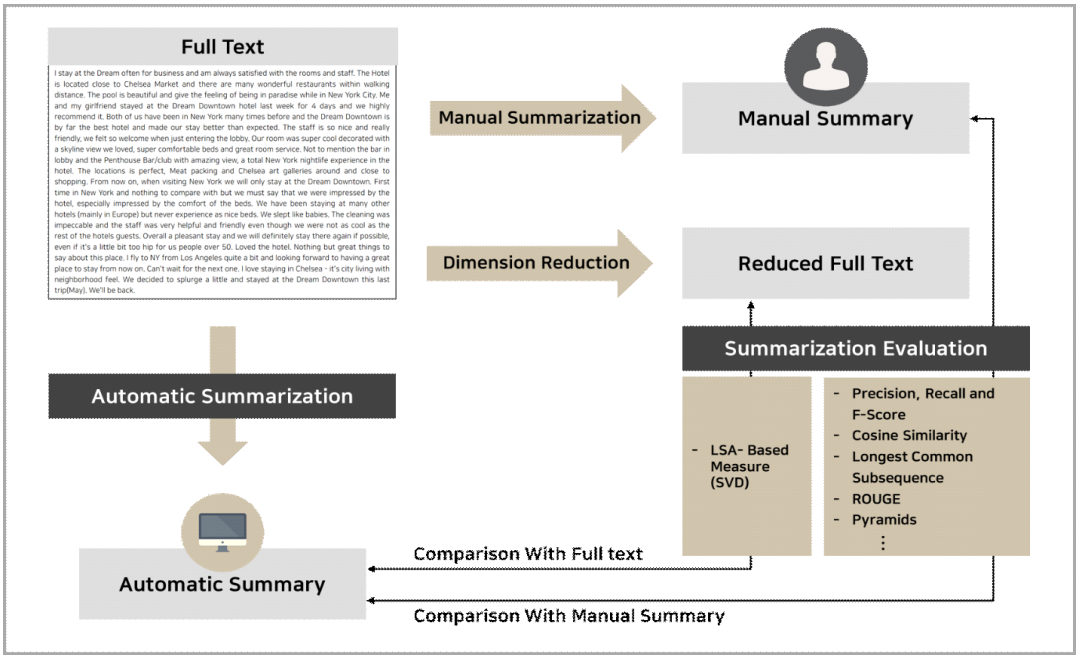
1. 서론

최근 빅데이터 분석에 대한 수요가 증가함에 따라 비정형 데이터를 분석하여 그 결과를 활용하는 사례도 증가하고 있다. 다양한 유형의 비정형 데이터 중에서도 텍스트는 거의 모든 분야에서 정보 전달의 수단으로 활용되고 있을 뿐 아니라, 데이터의 양이 매우 방대하고 다른 비정형 데이터 및 정형 데이터에 비해 수집이 상대적으로 용이하다는 측면에서 많은 분석가들의 관심의 대상이 되고 있다. 다양한 텍스트 분석 응용 중 문서를 사전에 정해진 카테고리로 분류하는 문서 분류(Document Classification), 다량의 문서로부터 주요 토픽을 추출하는 토픽 모델링(Topic Modeling), 텍스트에 포함된 감정이나 의견을 식별하는 감정분석(Sentiment Analysis) 또는 오피니언 마이닝(Opinion Mining), 그리고 하나의 문서 또는 여러 문서로부터 주요 내용을 요약하여 제시하는 문서 요약(Text Summarization)에 대한 연구가 매우 활발하게 이루어지고 있다.

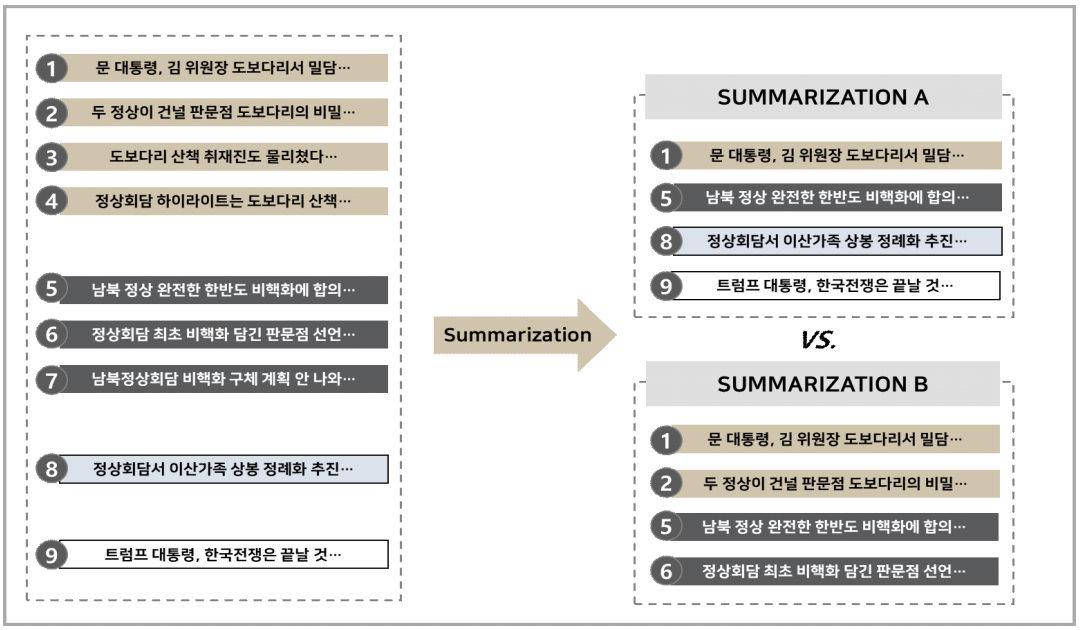
인터넷의 발전 및 스마트 기기의 보급률 증가에 따라 텍스트 데이터는 폭발적으로 증가하였지만 이로부터 간결하게 정리된 정보를 얻는 것은 더욱 어려워졌으며, 이에 따라 방대한 양의 문서로부터 중요한 내용을 정리해주는 문서 요약 기술의 중요성이 더욱 강조되고 있다. 이러한 문서 요약 기술은 국내 인터넷 포털 업체인 Daum과 Naver를 필두로 한 뉴스 요약 서비스, The Usable Privacy Policy Project 혹은 Polisis와 같은 개인정보 약관 요약 서비스, Google의 텐서플로우를 활용한 문서 요약 알고리즘 오픈소스 공개 등 여러 도메인에서 활발하게 개발 및 적용되고 있으며, 이와 관련한 연구도 학계를 중심으로 꾸준히 이루어지고 있다. 하지만 최근에는 문서 요약 기술 자체뿐 아니라 이렇게 요약된 내용을 어느 정도 신뢰할 수 있는지, 즉 요약 문서의 품질 측정에 대한 관심과 수요가 높아지고 있으며, 이러한 현상을 반영하여 요약 문서 품질 측정에 대한 연구도 그 중요성을 인정받고 있다.

요약 문서의 품질 측정에 대한 기존의 연구는 동향은 <Figure 1>과 같이 나타낼 수 있다. 우선 전문(Full Text)으로부터 여러 기법을 통해 자동 요약이 수행되며, 이 결과는 그림에서 자동 요약문(Automatic Summary)으로 나타나있다. 자동 요약문의 품질 측정을 위해 이상적인 요약 문서인 기준 문서(Reference Document)와의 비교가 이루어진다. 기준 문서는 크게 두 가지 방식으로 제공되는데, 가장 일반적인 방식은 사람이 수작업으로 이상적인 요약문을 작성하는 수동 요약(Manual Summarization) 방식이다. 이 방식은 요약문 작성 과정에서 사람의 개입을 필요로 하기 때문에 요약문 작성에 막대한 시간과 비용이 소요될 뿐 아니라, 요약자의 주관에 따라 평가 결과가 상이하게 나타날 수 있다는 한계를 갖는다. 따라서 이러한 한계를 극복하기 위해 사람의 개입 없이 요약 문서의 품질을 측정하기 위한 시도가 이루어지고 있다.

사람이 개입하지 않는 요약 문서 품질 평가의 가장 대표적인 방식은 자동 요약문과 전문의 유사성을 비교하는 방식이다. 다만 이 경우 요약과 전문은 길이 및 수록 어휘 수가 크게 상이하여 직접 비교가 어려우므로, SVD 등의 차원 축소를 통해 축약된 전문(Reduced Full Text)을 생성한 뒤, 이 축약된 전문과 자동 요약문의 유사도를 분석하는 방식으로 평가가 이루어진다. 이러한 방식은 사람의 개입 없이 자동으로 요약 문서의 품질을 측정한다는 점에서 그 기여가 인정되지만, 단순히 어휘의 빈도수에 기반하여 원문을 축



〈Figure 1〉 Traditional Approach of Evaluating Text Summary



〈Figure 2〉 Perspective of “Good Summary”

약한 문서를 품질 측정의 기준으로 사용한다는 본질적인 한계를 갖는다. 차원 축소는 기본적으로 어휘의 빈도수에 기반하여 이루어지므로, 자주 출현하는 어휘 또는 표현들이 축약된 전문에 주로 나타나게 된다. 즉 차원 축소에 기반한 자동 품질 측정 방식은 빈도가 높은 어휘가 다수 표현되고, 상대적으로 빈도가 낮은 어휘가 제거된 요약이 “좋은 요약”이라는 가정에 근거한 방식이라고 할 수 있다. 하지만 요약이란 본질적으로 방대한 내용을 줄여서 표현하면서도 정보의 누락을 최소화하는 것을 의미하므로, 단순히 빈도수에 기반한 “좋은 요약”이 항상 본질적 의미에서의 “좋은 요약”을 의미한다고 볼 수는 없다 (Figure 2).

“좋은 요약”에 대한 해석의 차이는 <Figure 2>를 통해 보다 자세히 설명될 수 있다. 그림의 좌측의 문서는 총 9개의 문장으로 구성되어 있으며, 유사한 어휘 및 주제를 가진 문장들이 총 4개의 군집으로 구분되어 나타나있다. 해당 문서가 Summary A와 Summary B의 두 가지 형태로 요약된 경우를 가정하자. 이 때 단순히 전문의 빈도수에 기반하여 요약문의 품질을 평가하는 경우에는, 자주 출현한 주제의 문장을 다수 포함한 Summary B를 “좋은 요약”이라고 평가할 것이다. 하지만 요약의 기본 원리, 즉 정보의 중복을 최소화하면서도 동시에 정보의 누락을 최소화한다는 관점에서 보면 서로 다른 주제로부터 문장을 하나씩 추출하여 포함하고 있는 Summary A가 “좋은 요약”이라고 평가할 수 있다.

이러한 관점에서 본 연구는 요약 문서의 자동 품질 측정에 대한 기존 연구의 한계를 극복하기 위해, 요약의 본질에 기반한 자동 품질 평가 방안을 제안하고자 한다. 구체적으로 요약문의 문장 중 서로 중복되는 내용이 얼마나 적은지를 나타내는 요소로 간결성(Succinctness) 개념을 정의하고, 원문의 내용 중 요약문에 포함되지 않은 내용이 얼마나 적은지를 나타내는 요소로 완전성(Completeness)을 정의하여 이를 평가에 활용하고자 한다. 또한 이 두 가지 척도의 조화 평균(Harmonic Mean)으로 F-Score를 측정하여, 간결성과 완전성 측면의 두 가지 측면에서 최적의 요약을 수행할 수 있는 균형점을 찾고자 한다.

본 논문의 이후 구성은 다음과 같다. 다음 장인 2장에서는 텍스트 요약 기법 및 요약 문서의 품질 평가에 대한 기존 연구를 요약하고, 3장에서는 본 연구에서 제안하는 방법론을 간단한 예시와 함께 소개한다. 다음으로 4장에서는 TripAdvisor의 호텔 리뷰 데이터에 대한 실제 실험 결과를 제시하고, 마지막 5장에서는 본 연구의 의의와 한계를 요약한다.

2. 관련 연구

2.1 문서 요약

문서를 자동으로 요약하는 접근법은 크게 추출(Extraction)과 생성(Abstraction)으로 구분된다. 추출 접근법은 문서 내에서 구, 절, 문장 등 특정 문서 요소의 중요도를 파악하여, 해당 요소를 그대로 발췌하여 사용하는 방법이다. 한편 생성은 원 문서에서 단어 혹은 문장을 발췌한 뒤, 자연어 처리 기법을 통해 원 문서의 요소를 조합하여 새로운 문장을 구성하는 과정이 반드시 포함된다는 특징을 갖는다. 자연어 처리에 대한 깊은 이해를 필요로 하는 생성 접근법에 비해 추출 접근법은 상대적으로 개념과 구현이 간단한 것으로 알려져 있으며, 이로 인해 문서 요약을 다루

는 많은 연구가 추출 접근법에 기반을 두어 이루어지고 있다.

추출 방식의 접근법은 Luhn(1958)에 의해 최초로 제안되었으며, 이후 지속적으로 발전하여 다양한 형태로 변형되어 왔다. 이들 중 가장 널리 사용되는 방법으로는 그래프 기반 접근법(Graph-based Approaches)과 토픽 표현 접근법(Topic Representation Approaches)을 들 수 있다. TextRank(Mihalcea and Tarau, 2004; Mihalcea and Tarau, 2005)는 그래프 기반 접근법의 가장 대표적인 기법으로, 각 문장을 하나의 노드(Node)로 간주하고, 문장 간 유사도에 따라 간선에 가중치를 부여하여 그래프를 구성한다. 이렇게 구성된 그래프에 대한 분석을 통해 PageRank 알고리즘으로 각 노드들의 중요도를 산출하고, 중요도가 높은 상위 N개의 노드를 선정하여 해당 문장들로 요약문을 구성한다.

한편 토픽 표현 접근법은 토픽을 정의하는 방법에 따라 TF-IDF 빈도 기반 기법(Gupta et al., 2007), 베이지안 토픽 모델(Bayesian Topic Models) 기반 기법(Daume III and Marcu, 2006; Haghighi and Vanderwende, 2009), LSA(Latent Semantic Analysis) 기반 기법(Gong and Liu, 2001; Steinberger and Jezek, 2004), PLSA (Probabilistic Latent Semantic Analysis) 기반 기법(Mnai, 2001) 등 매우 다양한 기법을 통해 구현되어 왔다. 특히 LSA 기반 기법은 완전하게 자동화된 수학적, 통계적 기법으로, SVD(Singular Value Decomposition)를 통해 전체 문서의 차원을 축소하고 단어의 맥락에 따라 의미를 추출하고 표현하는 방법이다. LSA 기법은 WordNet과 같은 어휘 자원을 사용하지 않고도, 전체 문서의 내용을 축약하여 문서의 중요 주제를 식별할 수 있다는 장점을 갖는다.

한편 문서 요약은 요약의 대상에 따라 단일 문서 요약(Single Document Summarization)(Ouyang, 2009)과 다중 문서 요약(Muti-Document Summarization)(Wan, 2007; Litak and Last, 2008)으로 구분될 수 있다. 단일 문서 요약은 하나의 문서를 요약하는 것이기 때문에, 해당 문서 내의 주요 문서 요소를 추출하는 방식으로 비교적 수월하게 수행된다. 하지만 다중 문서 요약은 대표 문서 요소를 추출하는 과정과 여러 문서에 각기 존재하는 다양한 주제들을 추출하는 과정의 두 가지 작업을 모두 필요로 하기 때문에 보다 난이도가 높은 것으로 알려져 있다. 하지만 여러 문서들이 하나의 주제를 다루고 있는 특수한 경우라면, 여러 문서를 하나의 문서로 취급하여 단일 문서 요약 방식을 그대로 적용할 수도 있다.

2.2 문서 요약 품질 평가

문서 요약 품질의 측정을 위해 가장 널리 사용되는 방법은 자동 요약문과 이상적인 요약 문서와의 유사도를 측정하는 내용(Content) 평가 방법이다. 이상적인 요약 문서로는 주로 사람이 수작업으로 작성한 Manual Summary가 사용되어 왔다. 내용 평가는 동시 선택(Co-selection) 방법과 내용 기반(Content-based) 방법으로 다시 세분화 되는데, 동시 선택은 Manual Summary와 자동 요약문의 문장이 얼마나 일치하는지를 확인하여 품질을 확인하는 방법이다. 구체적으로 사람이 작성한 이상적인 요약문과 기계가 자동으로 작성한 요약문을 비교하여, 두 문서 간 서로 완벽하게 일치하는 문장의 수를 집계하는 방식을 통해 상대적 유용성(Relative Utility) 등을 평가한다(Radev et al., 2004). 이러한 방식은 요약문이 기준 문서와 완벽하게 동일한 문장을 많이 포함할

수록 요약문의 품질이 높게 평가받지만, 형태가 다소 다르지만 내용이 매우 유사한 문장은 아무리 요약문에 많이 포함되더라도 품질 평가에 전혀 기여하지 못한다는 점에서 한계를 갖는다.

이와 달리 내용 기반 방법은 이상적인 요약 문서와의 완벽한 일치도가 아닌, 얼마나 유사한지 여부를 측정하는 것에 초점을 맞춘 방법이다. 내용 기반 방법은 방식에 따라 자동 요약된 내용을 Manual Summary와 비교하는 경우와 Full Text와 비교하는 경우로 나뉜다. 현재까지 널리 사용되고 있는 기법은 대부분 요약문을 Manual Summary와 비교하는 방식을 채택하고 있으며, ROUGE(Recall-Oriented Understudy for Gisting Evaluation) (Lin and Hovy, 2003; Lin, 2004), 피라미드(Pyramids) (Nenkova and Passonneau, 2005) 등이 그 대표적 예이다. 특히 ROUGE는 전문가가 직접 요약한 문서와 자동으로 요약된 시스템 문서를 비교하여 평가하는 방법으로, 정확률(Precision)과 재현률(Recall)을 산출한 뒤 최종적으로 F-Score 결과를 측정하는 방식으로 수행된다. 하지만 이상의 기존 방법들은 모두 Manual Summary 작성 과정에서 사람의 개입을 필요로 하기 때문에, 기준 문서 작성에 막대한 시간과 비용이 소요될 뿐 아니라 요약자의 주관에 의해 평가 결과가 다르게 나타날 수 있다는 점에서 한계를 갖는다.

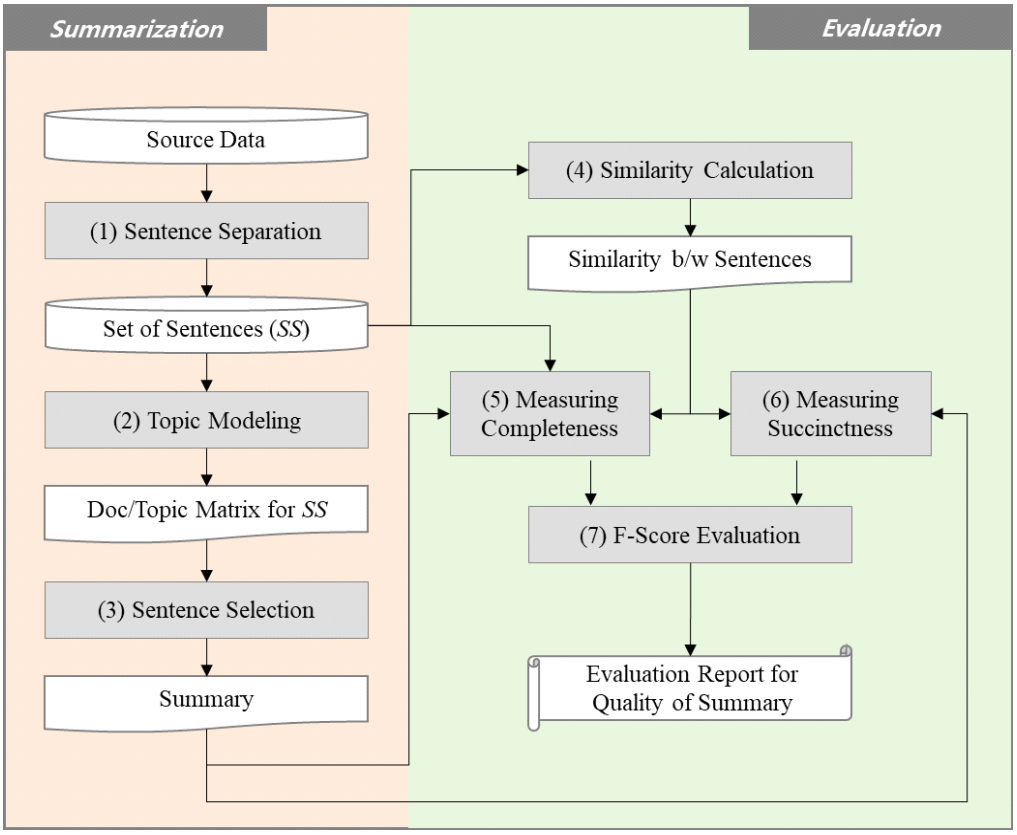
이러한 한계를 극복하기 위해 고안된 LSA 기반 Measure(Steinberger and Jezek, 2004) 방식은 요약문의 품질 측정을 위한 기준 문서로 Manual Summary가 아닌 Full Text를 사용한다. 즉 요약문과 전문의 유사도를 품질의 척도로 사용하는데, 요약문과 전문의 길이 및 수록 어휘 수가 크게 상이하여 직접적인 비교를 수행하기에는 어려움이 있다. 따라서 원문에 대한 차원 축소를 수행한 뒤, 축약된 원문과 요약문의 벡터 계산을 통해 주요 토픽 유사도(Main Topic Similarity)와 용어 의미 유사도(Term Significance Similarity)를 산출한다. 하지만 이러한 방식은 기본적으로 어휘의 빈도수에 기반한 비교를 수행하기 때문에, 전술한 바와 같이 단순히 빈도수에 기반한 “좋은 요약”이 항상 본질적 의미에서의 “좋은 요약”을 의미한다고 볼 수는 점에서 한계를 갖는다.

3. 제안 방법론

3.1 연구 모형

본 장에서는 완전성과 간결성을 정의하고, 이 두 가지 관점에 근거하여 요약문의 품질을측정하기 위한 방안을 제시한다. 본 연구에서 요약문의 완전성은 요약문에 포함된 문장들이 전문의 내용을 얼마나 포함하는지, 그리고 요약문의 간결성은 요약문에 포함된 문장들 간 얼마나 중복이 없는지를 나타내는 개념으로 정의된다. 또한 요약문의 품질 평가를 위해 자동으로 생성된 요약문이 필요하므로 본 장에서는 토픽 모델링에 기반을 두어 요약문을 생성하는 과정을 소개하고, 이렇게 생성된 요약문의 품질을 제안 방법론에 따라 평가하는 과정을 간단한 예를 통해 설명한다. <Figure 3>에서 좌측의 Summarization 부분은 요약문 생성 과정을, 우측의 Evaluation 부분은 본 연구의 핵심인 요약문의 품질을 평가하는 과정을 나타낸다.

우선 원본 문서를 각 문장 단위로 분리하여 문장의 집합으로 구성된 데이터(SS)로 재구성한다(1). 다음으로 SS에 대한 토픽 모델링을 하여 미리 정해진 개수의 토픽을 추출한다(2). 다음으로



〈Figure 3〉 Research Overview

각 토픽을 구성하는 문서 중 가장 높은 문서/토픽 가중치를 갖는 문장을 식별하고, 이들을 추출하여 조합함으로써 요약문을 구성한다(3). 이와 더불어 문장 단위로 구성된 원본 데이터에 대해 Sentence2Vec 알고리즘을 적용하여 각 문장 간 유사도를 도출한다(4). 다음으로 SS의 전체 문장 각각에 대해 특정 임계값 이상의 유사도를 갖는 문장이 요약문에 얼마나 포함되었는지를 측정하여 완전성을 산출하고(5), 요약문의 전체 문장 각각에 대해 특정 임계값 이상의 유사도를 갖는 문장이 요약문에 얼마나 포함되지 않았는지를 측정하여 간결성을 산출한다(6). 마지막으로 완전성과 간결성의 조화 평균으로 F-Score를 산출하여, 대상 요약문의 품질 평가 결과를 도출한다(7). 본 장의 이후 절에서는 제안 방법론의 주요 과정을 가상 예를 통해 자세히 설명하고, 다음 장인 4장에서는 제안 방법론을 실제 데이터에 적용한 실험 결과를 소개한다.

3.2 문서 요약

본 절은 <Figure 3>의 좌측에 나타난 (1) ~ (3)에 해당하는 과정, 즉 원본 문서를 문장 단위로

분리하고 이에 대한 토픽 모델링을 진행하여, 각 토픽 별로 추출된 대표 문서를 조합하여 요약문을 구성하는 과정을 소개한다. 문서 요약에는 LSA 기반의 토픽 모델링을 활용하였으며, 대상 원문으로는 숙박 업체에 대한 리뷰를 사용하였다. <Figure 4>는 리뷰 4건을 15개의 문장으로 분리한 예를 보이고 있으며, 이 단계 이후로는 분석의 최소 단위로 문장을 사용한다.

다음으로 <Figure 4>의 하단 테이블의 각 문장을 분석 최소 단위로 설정한 뒤, 이에 대한 토픽 모델링을 수행함으로써 <Figure 5>의 문서/토픽 행렬(Doc/Topic Matrix)과 토픽 정보를 얻을 수 있다. 토픽 모델링은 텍스트 분석 연구에 이미 많이 활용되어 왔으며 기존 연구에서 충분히 상세하게 소개되었으므로, 본 연구에서는 토픽 모델링에 대한 자세한 개념 설명은 생략하고 제안 방법론에서 채택한 주요 과정만을 설명한다(Blei et al., 2003; Deerwester et al., 1990; Kim et al., 2017). 토픽 모델링을 통해 산출되는 결과물은 크게 용어 별 토픽에 대한 참여 정보를 나

ReviewNo	Review Contents
1	We stayed at The Belvedere Hotel in May for 7 nights and really enjoyed our stay. Room was spacious and clean and the bed is very comfortable. Most of the staff are friendly and helpful. Close to the subways, walking distance to Central Park, Times Square,Broadway, Port Authority, etc.
2	The hotel is not a premium hotel. This hotel is in a great location and good price hence why we picked it. When I took the elevator up to our floor, the elevator door opened right in front of our room. The room was clean and there was a microwave and a coffee maker and a little fridge.
3	The Hotel was spotlessly clean and the staff were helpful and friendly. Each time we called front desk and asked for something it was delivered to our room promptly. The room was spacious and clean, the bathroom was ok.
4	This hotel is at a great location near Times Square and a subway stop. Be aware that it is an older hotel, so the elevators are a bit small and slow. Also, shower water temperature is not consistant. We stayed there 6 nights and those are my only complaints.

ReviewNo	SentenceNo	Review Contents
1	1	We stayed at The Belvedere Hotel in May for 7 nights and really enjoyed our stay.
	2	Room was spacious and clean and the bed is very comfortable.
	3	Most of the staff are friendly and helpful.
	4	Close to the subways, walking distance to Central Park, Times Square,Broadway, Port Authority, etc.
2	5	The hotel is not a premium hotel.
	6	This hotel is in a great location and good price hence why we picked it.
	7	When I took the elevator up to our floor, the elevator door opened right in front of our room.
	8	The room was clean and there was a microwave and a coffee maker and a little fridge.
3	9	The Hotel was spotlessly clean and the staff were helpful and friendly.
	10	Each time we called front desk and asked for something it was delivered to our room promptly.
	11	The room was spacious and clean, the bathroom was ok.
4	12	This hotel is at a great location near Times Square and a subway stop.
	13	Be aware that it is an older hotel, so the elevators are a bit small and slow.
	14	Also, shower water temperature is not consistant.
	15	We stayed there 6 nights and those are my only complaints.

〈Figure 4〉 Sentence Separation

	Topic1	Topic2	Topic3	Topic4	Topic5
Doc1	0.048	0.018	0.001	**0.631**	0.048
Doc2	**0.624**	-0.053	0	-0.035	0.192
Doc3	-0.054	0.719	0	0	-0.077
Doc4	0	0	0.288	0	-0.013
Doc5	0	0.048	0.155	0.040	0.381
Doc6	0	0.020	0.610	0.001	-0.048
Doc7	0.107	0.008	0.000	0.005	0.020
Doc8	0.285	0.007	0	0.005	0.104
Doc9	0	0.017	0.265	0.001	0.076
Doc10	0.115	0.143	0.017	0.065	0.098
Doc11	0.035	**0.722**	0.018	0	0.059
Doc12	0.017	0.068	**0.633**	0	0.146
Doc13	-0.243	0.092	-0.059	-0.057	**0.393**
Doc14	0.105	0.023	0.066	0.072	0.182
Doc15	0.121	0.007	0	0.004	0.016

	Topic Information
1	room, comfortable, bed, clean, small
2	staff, helpful, friendly, concierge, front
3	location, great, hotel, subway
4	night, stay, time, trip
5	elevator, small, time, floor

〈Figure 5〉 Doc/Topic Matrix and Topic Information

타내는 용어/토픽 행렬(Term/Doc Matrix)와 문서별 토픽에 대한 참여 정보를 나타내는 문서/토픽 행렬로 나뉜다. 각 행렬의 셀 값은 각각 특정 토픽에 대한 용어와 문서의 부합 정도를 나타내는 용어 가중치(Term Weight)와 문서 가중치(Document Weight)를 의미하며, 문서/토픽 행렬의 경우 <Figure 5>의 좌측 도표와 같은 형태로 나타낸다.

<Figure 5>는 토픽 모델링을 통해 5개의 토픽을 도출한 예를 나타내며, 동일 토픽에 속한 문서들은 유사한 내용을 다루고 있는 것으로 간주된다. 전술한 바와 같이 본 분석에서 각 문서는 개별 문장을 나타내므로, 각 토픽에서 가장 높은 문서 가중치를 갖는 문서를 추출함으로써 해당 주제를 다루고 있는 대표 문장을 선별할 수 있다. 유사한 내용을 다루고 있는 복수의 문장을 추출하는 경우 요약문의 간결성을 해치게 되므로, 각 토픽 별로 하나의 문장만을 추출하도록 한다. 예를 들어 <Figure 5>에서는 Doc2, Doc11, Doc12, Doc1, Doc13의 문장이 추출되며, 이들 5개의 문장을 조합하여 요약문을 구성할 수 있다.

3.3 문서 요약 품질 측정

본 절은 <Figure 3>의 (4) ~ (7) 해당하는 과정을 간단한 예를 통해 소개한다. 전체 과정은 SS에 속한 각 문장 간 유사도를 계산하여 유사도 행렬을 도출하고, 이를 활용하여 완전성, 간결성, 그리고 F-Score를 산출하는 과정으로 요약된다. 문장 간 유사도를 계산하기 위해 우선 각 문장은 수치로 변환되어야 하며, 본 연구에서는 이를 위해 문장의 구조화에 널리 사용되고 있는 Sentence2Vec 기법을 적용한다. 구체적으로는 Sentence2Vec을 사용하여 각 문장을 100차원 벡터로 변형한 뒤, 각 벡터 간 코사인 유사도(Cosine Similarity)를 산출하여 이를 문장 간 유사도로 사용한다. 유사도는 -1부터 1까지의 값을 가지며, 두 문장이 유사할수록 1에 가깝고 유사하지 않을수록 -1에 가깝게 나타난다. 따라서 완전히 동일한 두 문장의 유사도는 1로 나타난다.

<Figure 3>의 (5) Measuring Completeness는 원문의 내용이 요약문에 얼마나 누락없이 표현되었는지를 측정하는 과정을 나타낸다. 본 연구에서는 내용의 누락이 없는 정도를 나타내는 개념으로 완전성을 정의하였으며, 이는 다음의 식에 의해 계산된다.

$$Completeness = \frac{\#\ of\ Covered\ Sentences}{\#\ of\ Sentences\ in\ Full\ Text}$$

위 식에서 분모의 *# of Sentences in Full Text*는 원문에 포함된 전체 문장 수를 의미하며, 분자의 *# of Covered Sentences*는 원문의 문장 중 해당 문장 자체, 또는 해당 문장의 내용과 유사한 내용의 문장이 요약문에 나타난 문장의 수를 의미한다. *Covered Sentences*의 개념은 <Table 1>을 통해 보다 자세히 설명할 수 있다.

<Table 1>에서 각 행은 원문에 포함된 15개의 문장을, 각 열은 요약문에 포함된 5개의 문장을 나타내며, 각 셀의 수치는 해당 두 문장 벡터 간의 코사인 유사도를 나타낸다. 맨 우측의 Max Value는 원문의 각 문장에 대해 산출되며, 해당 문장이 요약문의 5개 문장 각각에 대해 갖는 유사도 중 가장 큰 값을 나타낸다. 예를 들어 Sentence 1은 그 자신이 요약문에 포함되어 있으므로 Max Value의 값이 1이 된다. 한편 Sentence 4는 그 자신이 요약문에 포함되어 있지는 않지만, 유사도가 0.611로 상대적으로 높게 나타나는 문장인 Sentence 12가 요약문에 포함되어 있음을 알 수 있다. 반대로 Sentence 10의 경우 유사도가 높은 문장이 요약문에 포함되어 있지 않아서, 가장 유사한 문장과의 유사도인 Max Value 값이 불과 0.078로 낮게 나타나고 있다. 이 때 여기서 언급한 Sentence 1, Sentence 4, Sentence 12에 대해, Sentence 1과 Sentence 4의 경우 완전히 동일하거나 유사한 내용이 요약문에 포함되어 있으

〈Table 1〉 Covered Sentences and Completeness Measure (*Similarity Threshold = 0.3*)

		Sentences in Summary					Max Value
		Sentence 2	Sentence 11	Sentence 12	Sentence 1	Sentence 13	
Sentences In Full Text	Sentence 1	-0.083	0.080	0.207	1	0.105	***1***
	Sentence 2	1	0.703	0.050	-0.083	-0.360	***1***
	Sentence 3	0.243	0.287	0.060	0.010	0.016	0.287
	Sentence 4	-0.133	-0.090	0.611	-0.395	-0.461	***0.611***
	Sentence 5	0.036	-0.051	0.072	0.037	0.437	***0.437***
	Sentence 6	-0.022	-0.083	0.611	0.048	-0.356	***0.611***
	Sentence 7	-0.206	-0.103	-0.094	-0.008	0.290	0.290
	Sentence 8	0.303	0.392	0.358	-0.090	-0.302	***0.392***
	Sentence 9	-0.091	0.347	0.047	0.260	0.074	***0.347***
	Sentence 10	-0.207	0.078	-0.236	-0.091	0.074	0.078
	Sentence 11	0.703	1	-0.005	0.080	-0.273	***1***
	Sentence 12	0.050	-0.005	1	0.207	-0.196	***1***
	Sentence 13	-0.360	-0.273	-0.196	0.105	1	***1***
	Sentence 14	-0.177	-0.309	-0.218	0.178	0.366	***0.366***
	Sentence 15	0.110	0.100	0.150	0.451	-0.036	***0.451***

며 Sentence 12의 경우 유사 내용이 요약문에서 누락되어 있다고 해석할 수 있다. 하지만 이러한 판단은 어느 정도의 유사도를 유사성이 높다고 판단할지 여부에 따라 달라지게 되므로, 일관적인 판단 기준을 마련하기 위해 유사도의 임계값(Similarity Threshold)을 명시적으로 설정할 필요가 있다. 예를 들어 <Table 1>은 유사도의 임계값을 0.3으로 설정했을 경우 임계값 이상의 Max Value 값을 밑줄로 구분하여 표시하고 있으며, 전체 15개 문장 중 임계값 이상의 값을 갖는 문장인 *Covered Sentences*는 12개인 것으로 파악된다. 따라서 임계값이 0.3인 경우 <Table 1>의 요약문의 완전성은 0.8 (=12/15)로 나타나며, 임계값을 높일수록 완전성은 낮아짐을 알 수 있다.

다음으로 <Figure 3>의 (6) Measuring Succinctness는 요약문의 각 문장이 내용의 중복 없이 얼마나 간결하게 표현되었는지를 측정하는 과정을 나타낸다. 본 연구에서는 내용의 중복이 없는 정도를 나타내는 개념으로 간결성을 정의하였으며, 이는 다음의 식에 의해 계산된다.

$$Succinctness = \frac{\#\ of\ Unique\ Sentences}{\#\ of\ Sentences\ in\ Summary}$$

위 식에서 분모의 *# of Sentences in Summary*는 요약문에 포함된 전체 문장의 수를 의미하며, 분자의 *# of Unique Sentences*는 요약문의 문장 중 해당 문장이 요약문 내의 다른 문장과 일정 수준 이상으로 유사하지 않은, 즉 유일한 내용을 담고 있는 것으로 인정될 수 있는 문장의 수를 의미한다. *Unique Sentences*의 개념은 <Table 2>을 통해 보다 자세히 설명할 수 있다.

<Table 2>에서 각 행과 열은 요약문에 포함된 5개의 문장을 나타내며, 각 셀의 수치는 해당 두 문장 벡터 간의 코사인 유사도를 나타낸다. 맨 우측의 Max Value는 요약문의 각 문장에 대해 산출되며, 해당 문장이 요약문의 5개 문장 중 자신을 제외한 4개 문장 각각에 대해 갖는 유사도 중 가장 큰 값을 나타낸다. 예를 들어 Sentence 2는 자신을 제외하고 Sentence 11과의 유사도가 0.703으로 가장 높게 나타나므로, Max Value의 값이 0.703이 된다. 즉 Sentence 2와 Sentence 11은 서로 간에 유사도가 0.703로 상대적으로 높게 나타나므로, 요약문 내에 이들 두 문장이 모두 포함된 것은 불필요한 중복으로 해석할 수 있다. 한편 Sentence 13의 경우 요약문 내에 자신과 유사도가 높은 문장이 존재하지 않아서, 가장 유사한 문장과의 유사도인 Max Value 값이 불과 0.105로 낮게 나타나고 있다. 이러한 관점에서 Sentence 2와 Sentence 11은 유사한 내용이 요약문 내에 중복으로 존재하며, Sentence 1, Sentence 12, Sentence 13은 유사 내용이 요약문 내에 존재

〈Table 2〉 Unique Sentences and Succinctness Measure (*Similarity Threshold = 0.3*)

		Sentences in Summary					Max Value
		Sentence 2	Sentence 11	Sentence 12	Sentence 1	Sentence 13	
Sentences In Summary	Sentence 2	1	0.703	0.050	-0.083	-0.360	0.703
	Sentence 11	0.703	1	-0.005	0.080	-0.273	0.703
	Sentence 12	0.050	-0.005	1	0.207	-0.196	***0.207***
	Sentence 1	-0.083	0.080	0.207	1	0.105	***0.207***
	Sentence 13	-0.360	-0.273	-0.196	0.105	1	***0.105***

하지 않는다고 해석 할 수 있다. 하지만 이러한 판단 역시 어느 정도의 유사도를 유사성이 높다고 판단할지 여부에 따라 달라지게 되므로, 일관적인 판단 기준을 마련하기 위해 유사도의 임계값을 명시적으로 설정할 필요가 있다. 예를 들어 <Table 2>은 유사도의 임계값을 0.3으로 설정했을 경우 임계값 이하의 Max Value 값을 밑줄로 구분하여 표시하고 있으며, 전체 5개 문장 중 임계값 이하의 값을 갖는 문장인 *Unique Sentences*는 3개인 것으로 파악된다. 따라서 임계값이 0.3인 경우 <Table 2>의 요약문의 간결성은 0.6(=3/5)으로 나타나며, 임계값을 높일수록 간결성은 높아짐을 알 수 있다.

마지막으로 <Figure 3>의 (7) F-Score Evaluation은 완전성과 간결성의 두 가지 측면을 얼마나 동시에 만족시켰는지를 측정하는 과정을 나타낸다. 전술한 바와 같이 유사도의 임계값이 높아질수록 완전성은 낮아지고 간결성은 높아지게 된다. 따라서 본 연구에서는 완전성과 간결성을 동시에 고려하여 요약문의 품질을 향상시키기 위해, 두 가지 척도를 통합한 F-Score를 사용하여 유사도 임계값의 최적 지점을 찾고자 한다. F-Score는 다음과 같이 조화 평균에 의해 계산된다.

$$F-Score = \frac{2 \cdot Completeness \cdot Succinctness}{Completeness + Succinctness}$$

유사도 임계값의 변화에 따른 완전성, 간결성, 그리고 F-Score의 변화 양상이 <Figure 6>에 나타나있다.

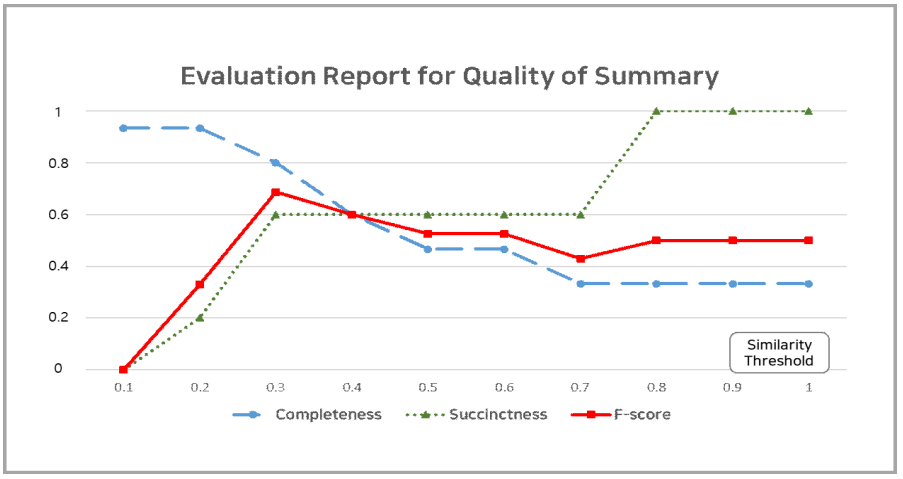
유사도 임계값에 따라 완전성, 간결성, F-Score 세 가지 척도의 값을 산출한 결과의 예가 <Figure 6>의 우측에 제시되어 있다. 유사도 임계값을 0.3으로 설정했을 경우 F-Score는 0.686으로 최댓값을 보이며, 이 때 완전성의 값은 0.8, 간결성의 값은 0.6로 나타난다. <Figure 6>에서

	Topic1	Topic2	Topic3	Topic4	Topic5	Max Value
	Sentence 2	Sentence 11	Sentence 12	Sentence 1	Sentence 13	
Sentence 1	-0.083	0.080	0.207	1	0.105	1
Sentence 2	1	0.703	0.050	-0.083	-0.360	1
Sentence 3	0.243	0.287	0.060	0.010	0.016	0.287
Sentence 4	-0.133	-0.090	0.611	-0.395	-0.461	0.611
Sentence 5	0.036	-0.051	0.072	0.037	0.437	0.437
Sentence 6	-0.022	-0.083	0.611	0.048	-0.356	0.611
Sentence 7	-0.206	-0.103	-0.094	-0.008	0.290	0.290
Sentence 8	0.303	0.392	0.358	-0.090	-0.302	0.392
Sentence 9	-0.091	0.347	0.047	0.260	0.074	0.347
Sentence 10	-0.207	0.078	-0.236	-0.091	0.074	0.078
Sentence 11	0.703	1	-0.005	0.080	-0.273	1
Sentence 12	0.050	-0.005	1	0.207	-0.196	1
Sentence 13	-0.360	-0.273	-0.196	0.105	1	1
Sentence 14	-0.177	-0.309	-0.218	0.178	0.366	0.366
Sentence 15	0.110	0.100	0.150	0.451	-0.036	0.451

	Topic1	Topic2	Topic3	Topic4	Topic5	Max Value
	Sentence 2	Sentence 11	Sentence 12	Sentence 1	Sentence 13	
Sentence 2	1	0.703	0.050	-0.083	-0.360	0.703
Sentence 11	0.703	1	-0.005	0.080	-0.273	0.703
Sentence 12	0.050	-0.005	1	0.207	-0.196	0.207
Sentence 1	-0.083	0.080	0.207	1	0.105	0.207
Sentence 13	-0.360	-0.273	-0.196	0.105	1	0.105

Threshold	Completeness	Succinctness	F-score
0.1	0.93	0	-
0.2	0.93	0.2	0.329
0.3	0.80	0.6	0.686
0.4	0.60	0.6	0.600
0.5	0.47	0.6	0.525
0.6	0.47	0.6	0.525
0.7	0.33	0.6	0.429
0.8	0.33	1	0.500
0.9	0.33	1	0.500
1	0.33	1	0.500

〈Figure 6〉 Evaluation Report for Quality of Summary with Three Measures



〈Figure 7〉 Completeness, Succinctness, and F-Score in accordance with Similarity Threshold

유사도 임계값의 증가에 따라 F-Score는 점차 증가하다가 다시 감소함을 알 수 있다. 이러한 결과는 <Figure 6>을 그래프로 도식화한 <Figure 7>를 통해서도 동일하게 확인할 수 있다.

본 장에서는 제안 방법론에 따른 요약문의 품질 측정 과정을 간단한 예를 통해 소개하였으며, 실제 데이터에 대해 본 방법론을 적용한 실험 결과는 다음 장에서 소개한다. 실제 실험에서는 토픽 수를 변화시켜가며 서로 다른 문장 수를 갖는 요약문을 다수 생성한 후, 각 요약문의 품질을 평가하여 요약문의 길이에 따른 요약 품질의 변화 양상도 함께 분석한다.

4. 실험

4.1 실험 개요

본 절에서는 제안 방법론의 적용 실험을 위한 환경 및 데이터에 대해 간략하게 소개한다. 문서 집합으로부터 주요 토픽을 추출하기 위한 토픽 모델링은 SAS Enterprise Miner Workstation 14.1을 통해 수행하였으며, 리뷰의 문장 단위 분리 및 Sentence2Vec은 Python 3.6.3을 통해 수행하였다. 분석을 위해 TripAdvisor 사이트에서 세 개의 호텔을 선정하고, 이들 호텔에 대해 2010년 2월부터 2016년 8월 사이에 작성된 리뷰를 수집하여 요약의 대상이 되는 원문으로 사용하였다. 사용된 원문은 총 3,020건의 리뷰에 대한 문장 29,671개로 구성되어 있다.

4.2 실험 결과 및 해석

4.2.1 문서 요약

본 부절에서는 3.2에서 소개된 과정에 따라 실제 TripAdvisor의 리뷰 데이터로부터 요약문을 생성하는 과정 및 결과를 소개한다. 우선 각 호텔 별로 리뷰를 통합한 뒤 이를 다시 문장 단위로 분리하였다. 또한 각 호텔 별로 토픽 모델링을 실시하였으며, 토픽의 개수가 요약문의 품질

에 미치는 영향을 확인하기 위해 토픽의 수를 5, 10, 15, 20, 100, 그리고 320개로 변경해 가며 실험을 수행하였다. 즉 3개의 호텔에 대해 각 호텔별로 6번, 총 18번의 토픽 모델링을 실시하였으며, 이렇게 도출된 토픽의 결과 중 일부가 <Figure 8>에 나타나있다.

다음으로 토픽 모델링 결과 중 하나인 문서/토픽 행렬을 사용하여 토픽 별 가중치가 가장 높은

Hotel A

Topic 5	Topic Information
Topic1	room, comfortable, bed, clean, small
Topic2	staff, helpful, friendly, concierge, front
Topic3	location, great, great location, hotel, subway
Topic4	belvedere, night, stay, time, trip
Topic5	hotel, elevator, small, time, floor

Topic 10	Topic Information
Topic1	bed, comfortable, clean, comfortable bed, comfy
Topic2	staff, helpful, friendly, concierge, front
Topic3	location, great, great location, price, good
Topic4	belvedere, stay, night, time, trip
Topic5	hotel, good, price, old, breakfast
Topic6	microwave, fridge, small, coffee, maker
Topic7	distance, walking, walk distance, subway, restaurant
Topic8	room, clean, spacious, floor, bathroom
Topic9	elevator, floor, small, time, slow
Topic10	night, desk, front, front desk,day

Hotel B

Topic 5	Topic Information
Topic1	central, park, location, subway, distance
Topic2	room, bed, comfortable, clean, spacious
Topic3	staff, friendly, helpful, front, desk
Topic4	night, hotel, room, time, day
Topic5	hotel, good, location, price, good location

Hotel C

Topic 5	Topic Information
Topic1	staff, friendly, helpful, hotel, comfortable
Topic2	room, bed, comfortable, small, bathroom
Topic3	hotel, night, club, lobby, place
Topic4	great, location, great location, restaurant, good
Topic5	desk, front, front desk, room, door

〈Figure 8〉 Topic Information (Part)

Hotel A - Topic 5

	Topic1	Topic2	Topic3	Topic4	Topic5
Doc1	0.000	0.017	0.205	0.015	0.139
Doc2	0.000	0.000	0.000	0.000	0.000
Doc3	0.169	0.008	0.000	0.038	0.045
Doc4	0.119	0.017	0.036	0.013	0.140
Doc5	0.014	0.000	0.000	0.000	0.012
Doc6	0.073	0.016	0.035	0.013	0.099
Doc7	0.105	0.023	0.066	0.072	0.182
Doc8	-0.029	0.518	0.000	0.000	-0.029
⋮			⋮		
Doc10549	0.000	0.102	0.021	0.000	0.034

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Max Weight Document No	541	9736	9698	5738	9997

Summary

Room was spacious and clean and the bed is very comfortable
The Hotel was spotlessly clean and the staff were helpful and friendly
This hotel is at a great location near Times Square and a subway stop
We stayed at The Belvedere Hotel in May for 7 nights and really enjoyed our stay
I can't believe that all the rooms in this hotel were this disgusting or the hotel would have been empty

Hotel A - Topic 10

	Topic1	Topic2	Topic3	Topic4	Topic5	···	Topic10
Doc1	0.008	0.000	0.015	0.010	0.184		-0.006
Doc2	0.000	0.000	0.000	0.000	0.000		0.000
Doc3	0.025	0.000	0.000	0.021	0.000		0.082
Doc4	0.017	0.000	0.010	0.006	0.118	···	-0.020
Doc5	0.000	0.000	0.000	0.000	0.000		0.027
Doc6	0.009	0.000	0.009	0.006	0.114		0.011
Doc7	0.013	0.000	0.013	0.057	0.163		0.045
Doc8	0.000	0.515	0.000	-0.020	0.000		0.167
⋮			⋮				
Doc10549	0.000	0.095	0.024	0.000	0.022		0.075

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	···	Topic 10
Max Weight Document No	7383	765	1417	5738	1884	···	5270

Summary

The bed was very comfortable and clean
Most of the staff are friendly and helpful
This hotel is in a great location hence why we picked it
We stayed at The Belvedere Hotel in May for 7 nights and really enjoyed our stay
The hotel is not a premium hotel
The room was clean and there was a microwave and a coffee maker and a little fridge
Close to the subways, walking distance to Central Park, Times Square, Broadway, Port Authority, etc
The room was spacious and clean, the bathroom was ok
When I took the elevator up to our floor, the elevator door opened right in front of our room
Each time we called front desk and asked for something it was delivered to our room promptly

〈Figure 9〉 Summary Generation in Accordance with the Number of Topics

문장을 추출하고, 해당 문장으로 요약문을 구성한 결과의 일부를 <Figure 9>에 제시하였다. 각 토픽 마다 하나의 문장을 추출하였기 때문에, 토픽 수의 변화에 따라 요약문의 문장 수는 5, 10, 15, 20, 100, 320개로 다르게 나타난다. 예를 들어 <Figure 9>의 상단에 나타난 요약문은 5개의 문장으로 구성되며, 하단의 요약문은 10개의 문장으로 구성된다.

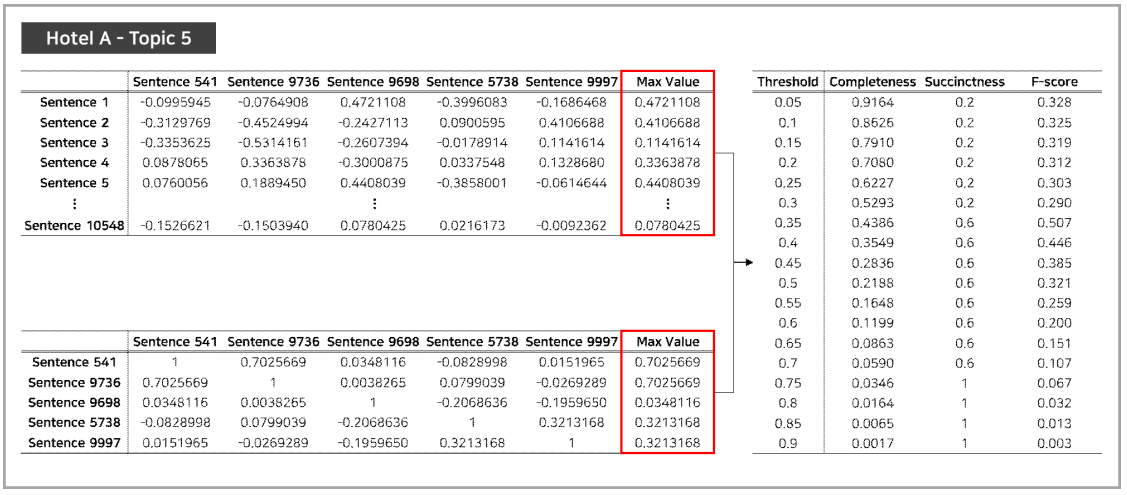
4.2.2 문서 요약 품질 측정

본 부절에서는 3.3절에 소개된 과정에 따라, 앞에서 도출한 리뷰 요약문의 품질을 측정한 결과를 제시한다. 첫 단계로 Sentence2Vec을 활용하여 각 문장을 벡터화하고, 이들 벡터 간 코사인 유사도를 계산하여 각 문장 간 유사도를 행렬로 도출하였다(Table 3).

다음으로 앞의 부절에서 생성한 요약문과 <Table 3>의 유사도 행렬을 토대로 유사도 임계값에 따른 완전성, 간결성, F-Score의 값을 산출한다. 전체 문장과 요약 문장 간의 유사도에서는 자기 자신에 대한 유사도, 즉 1을 포함하여 최댓값을 도출하여 완전성을 계산한다. 반면 요약 문

〈Table 3〉 Similarity Matrix between Sentences

	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5	···	Sentence 10548
Sentence 1	1	0.1411104	0.1293644	-0.1676280	0.1113393		-0.1054437
Sentence 2	0.1411104	1	0.4095090	0.0950363	-0.1441542		0.0017545
Sentence 3	0.1293644	0.4095090	1	-0.2174963	-0.2106007		0.0651208
Sentence 4	-0.1676280	0.0950363	-0.2174963	1	0.3021876	···	-0.0945512
Sentence 5	0.1113393	-0.1441542	-0.2106007	0.3021876	1		0.1132104
⋮			⋮				
Sentence 10548	-0.1054437	0.0017545	0.0651208	-0.0945512	0.1132104		1



〈Figure 10〉 Evaluation Report for Quality of Summary (Hotel A, # of Sentences = 5)

장 간의 유사도에서는 자기 자신에 대한 유사도를 포함하지 않고 최댓값을 도출하여 간결성을 계산한 후, 완전성과 간결성의 조화 평균을 통해 F-Score를 산출한다. Hotel A에 대한 실험 중 토픽의 수를 5개로 설정한 요약문의 품질 평가 결과가 <Figure 10>에 나타나있다.

<Figure 10>는 5개의 문장으로 구성된 요약문에 대해 유사도의 임계값을 변화시켜가며 세가지 품질 척도를 분석한 것으로, 임계값이 높아질수록 완전성은 낮게, 간결성은 높게 나타남을 확인할 수 있었다. 한편 F-Score는 임계값의 증가에 따라 점차 증가하다가 다시 감소하는 양상을 보였으며, 유사도의 임계값이 0.35일 때 F-Score가 최댓값인 0.507을 나타냈다.

한편 요약문을 구성할 때 토픽의 수가 많아질수록 요약문이 포함하는 문장의 수도 증가하며, 이로 인해 요약문의 완전성, 간결성, 그리고 F-Score도 영향을 받을 것으로 예상된다. 토픽 수의 증가에 따른 세 가지 척도의 변화를 분석하기 위해 실험 결과를 그래프로 도식화하였으며, 그 결과를 <Figure 11>에 제시하였다.

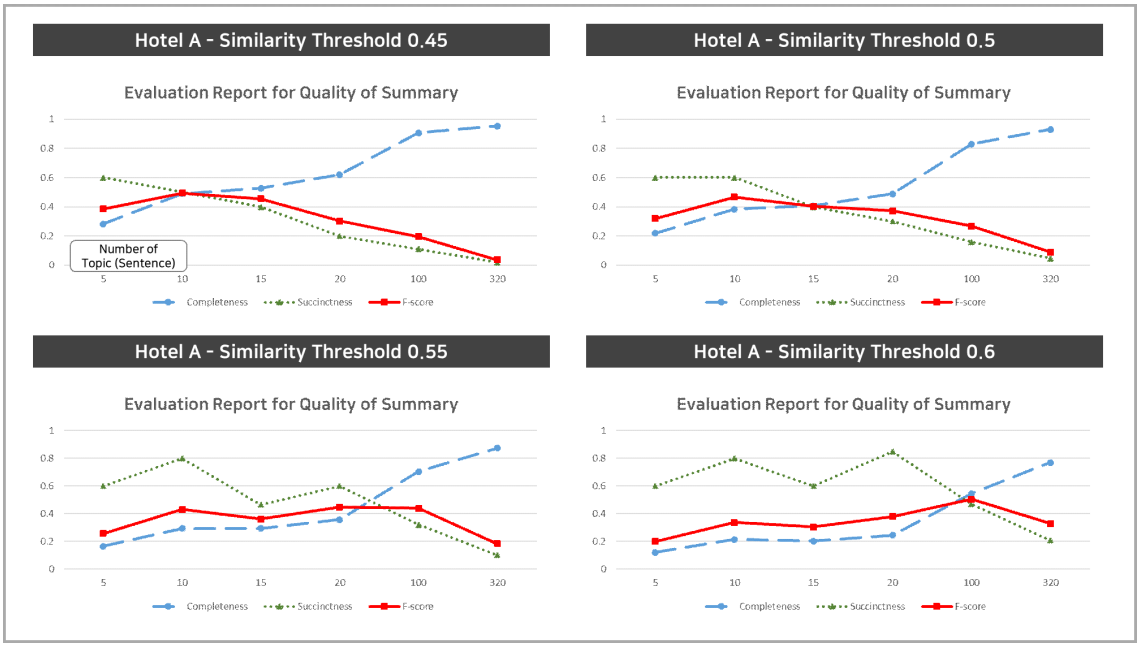
<Figure 11>의 네 가지 그래프에서, 유사도의 임계값이 동일하더라도 토픽 수가 증가함에 따라 완전성은 점차 증가하며 간결성은 점차 감소하는 양상을 보인다. 이러한 현상은 유사도의 임계값이 높아질수록 유사한 문장으로 판단하는 기준이 바뀜에 따라 *Covered Sentence*와 *Unique Sentence*의 개수가 변화하고, 이로 인해 요약문의 완전성, 간결성, 그리고 F-Score가 영향을 받기 때문에 야기된 것으로 파악된다. 한편 F-Score가 최대로 나타나는 유사도의 임계값은 토픽 수의 증가에 따라 점차 증가함을 알 수 있다. 즉 토

〈Figure 11〉 Evaluation Report for Quality of Summary (Hotel A, # of Topic=5, 20, 100, 320)

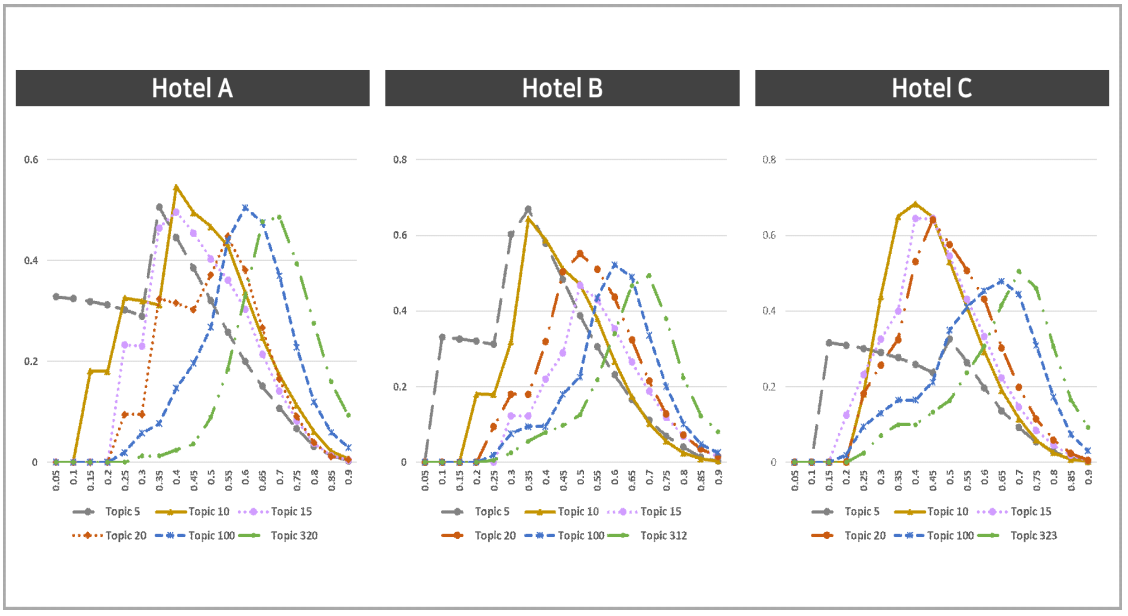
픽 수가 5개일 때는 F-Score의 최댓값이 임계값 0.35에서 나타났지만, 토픽 수가 20, 100, 320개일 때는 F-Score의 최댓값이 각각 임계값 0.55, 0.6, 0.7에서 나타났다. 토픽 수에 따른 세 가지 척도의 변화 패턴은 <Figure 12>에서 보다 명확하게 파악할 수 있다.

<Figure 12>에서 토픽의 수, 즉 요약문에 포함된 문장의 수가 증가함에 따라 완전성은 함께 증가하고 간결성은 감소하는 현상이 확인되었다. 한편 F-Score의 경우 일정 수준까지는 토픽 수의 증가에 따라 함께 증가하지만, 특정 지점 이후에는 오히려 감소하는 추세를 보였다. 또한 F-Score가 최대가 되는 토픽의 수는 유사도의 임계값에 따라 서로 다르게 나타나며, 임계값이 높을수록 토픽의 수가 많을 때 F-Score가 최댓값을 가짐을 알 수 있었다. 예를 들어 임계값이 0.45일 때는 토픽 수, 즉 요약문에 포함된 문장의 수가 10일 때 F-Score가 최댓값을 가졌으나, 임계값이 0.6일 때는 토픽 수가 100일 때 F-Score가 최댓값을 갖는 것으로 나타났다.

이상의 결과를 종합하면 토픽의 수와 유사도의 임계값은 문장의 유사성 여부를 판가름하는 데 영향을 미치기 때문에, 요약문을 평가하는 세 가지 척도가 두 가지 조건의 영향을 동시에 받음을 알 수 있다. 따라서 완전성과 간결성의 통합 지표인 F-Score 관점에서 최고의 품질을 갖는 요약문을 구성하기 위해선, 토픽의 수와 유사도의 임계값을 동시에 변화시키면서 최적 조합(Optimal Combination)을 탐색할 필요가 있다. 세 가지 호텔 각각의 리뷰 요약문에 대한 최적 조합을 찾기 위한 실험 결과가 <Figure 13>에 제시되어 있다.



〈Figure 12〉 Evaluation Report for Quality of Summary
(Hotel A, *Similarity Threshold* = 0.45, 0.5, 0.55, 0.6)



〈Figure 13〉 Investigating Optimal Combination of # of Topics and Similarity Threshold

〈Table 4〉 Three Quality Measures at Optimal Setting

	Opt. # of Topics	Sim. Threshold	F-Score	Completeness	Succinctness
Hotel A	10	0.4	0.547	0.604	0.5
Hotel B	5	0.35	0.667	0.696	0.6
Hotel C	10	0.4	0.684	0.597	0.8

<Figure 13>은 Hotel A, Hotel B, 그리고 Hotel C의 리뷰 요약문에 대해 토픽 수와 유사도 임계값의 최적 조합을 찾기 위한 실험 결과를 비교하여 보이고 있다. 세 가지 경우 모두 최적 토픽의 수는 5 ~ 10개 사이에서 나타났으며, 유사도의 임계값은 0.35 ~ 0.4의 구간에서 F-Score가 높게 나타나는 현상을 보였다. 구체적으로 Hotel A, Hotel B, 그리고 Hotel C는 각각 (토픽 수, 유사도의 임계값)이 (10, 0.4), (5, 0.35), 그리고 (10, 0.4)일 때 F-Score가 최댓값을 가졌으며, 이 때의 세 가지 품질 척도의 값이 <Table 4>에 요약되어 있다.

본 장에서는 제안 방법론에 따라 리뷰 요약문의 품질을 측정하고, 토픽의 수와 유사도 임계값의 최적 조합을 탐색하여 가장 좋은 품질을 갖는 요약문을 생성하는 과정을 실험을 통해 소개하였다.

5. 결론

최근 문서 자동 요약 기술의 수요 증가 및 텍스트 분석 기술의 고도화에 따라 다양한 방법의 문서 요약 기술이 개발되고 있으며, 개발된 기술은 실제로 현업의 여러 도메인에서 다양하게 적용되고 있다. 이처럼 문서 요약의 중요성과 활용성이 증가함에 따라, 요약 기술 자체뿐 아니라 요약된 내용의 품질을 측정하는 기술에 대한 관심이 급증하고 있다. 이에 본 연구에서는 내용 중복의 최소화 및 내용 누락의 최소화의 두 가지 관점에서 요약문의 품질을 측정할 수 있는 평가 방법론을 새롭게 제안하였다. 또한 제안 방법론의 실제 적용 가능성을 평가하기 위해 TripAdvisor의 호텔 리뷰로부터 29,671개의 문장을 추출하여 각 호텔 별로 리뷰를 요약하고, 요약된 리뷰에 대해 제안 방법론에 따라 품질 평가를 수행한 실험 결과를 소개하였다.

본 연구의 학술 및 실무적 기여는 다음과 같다. 우선 학술적 측면에서 본 연구의 가장 큰 기여는 요약의 본질에 기반하여 요약문의 품질을 평가하기 위해 완전성과 간결성을 새롭게 정의하고, 이를 산출할 수 있는 방법을 제안하였다는 점이다. 또한 상충 관계(Trade off)에 있는 완전성과 간결성을 F-Score로 통합하여, 문장 유사도의 임계값을 변화시켜가며 최적의 요약을 수행할 수 있는 방안을 제시하였다는 점도 제안 방법론의 큰 특징 중 하나이다.

이와 더불어 실무적 측면에서는 제안 방법론을 통해 자동 요약문의 품질 평가를 사람의 개입 없이 자동으로 수행함으로써, 자동 요약을 제공하는 다양한 요약 서비스에서 객관적이고 효율적인 요약문 품질 관리가 가능해질 것으로 기대한다. 또한 제안 방법론은 자동 요약의 기법에 상관없이 요약문의 품질 측정이 가능하기 때문에 실제 다양한 요약 서비스의 품질을 측정할 때 활용도가 매우 높을 것으로 예상된다.

하지만 본 연구는 향후 다음의 측면에서 보완이 필요하다. 본 연구는 제안 방법론에 따라 요약문의 품질을 평가하기 위해, 본 연구에서 자체적으로 구현한 방식에 따라 문서 요약을 수행하고 그 결과에 대한 품질 평가를 수행하였다. 향후 이미 알려진 다양한 문서 요약 기법에 따라 요약을 수행한 뒤, 이들 요약문의 품질 평가에 본 방법론을 적용함으로써 방법론의 견고성을 높일 필요가 있다. 또한 본 연구에서 정의한 완전성과 간결성 척도는 개념 자체로는 요약의 본질을 충실히 반영하고 있지만, 완전성과 간결성의 측정 산식은 향후 다양한 관점에서 더욱 정교화 될 필요가 있다. 마지막으로 제안 방법론에 의한 요약문의 품질 평가와 사람이 작성한 수동 요약에 기반을 둔 품질 평가와의 비교를 통해, 제안 방법론의 우수성과 신뢰도를 높일 필요가 있다.

참고문헌(References)

Blei, D. M., A. Y. Ng and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research,* Vol.3, (2003), 993~1022.

Daume III, H. and D. Marcu., “Bayesian Query-Focused Summarization,” *Proceeding of the International Conference on Computation Linguistics and the annual meeting of the Association for Computational Linguistics*, (2006), 305~312.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, “Indexing by

Latent Semantic Analysis," *Journal of the American Society for Information Science,* Vol.41, No.6(1990), 391~407.

Gong, Y. and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," *Proceeding of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2001), 19~25.

Gupta, S., A. Nenkova and D. Jurafsky, "Measuring Importance and Query Relevance in Topic-Focused Multi-Document Summarization," *Proceeding of the Annual Meeting of the Association for Computational Linguistics*, (2007), 193~196.

Haghighi, A., and L. Vanderwende, "Exploring Content Models for Multi-Document Summarization," *Proceeding of Human Language Technologies: The 2009 Annual Conference of the North American Chpter of the Association for Computational Linguistcs*, (2009), 362~370.

Kim, N., D. Lee, H. Choi and W. X. S. Wong, "Investigations on Techniques and Applications of Text Analytics," *The Journal of Korean Institute of Communications and Information Sciences,* Vol.42, No.2(2017), 471~492.

Lin, C. Y. and E. Hovy, "Automatic Evaluation of Summaries Using n-Gram Co-Occurrence Statistics," *Proceeding of HLT-NAACL*, (2003), 71~78.

Lin, C. Y., "Rouge: A Package for Automatic Evaluation of Summaries," *Proceeding of the Workshop on Text Summarization Branches Out*, (2004), 74~81.

Litvak, M. and M. Last, "Graph-based keyword extraction for single-document summarization," *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization. Association for Computational Linguistics*, (2008).

Luhn, H. P., "The Automatic Creation of Literature Abstracts," *IBM Journal of Research Development*, Vol.2, No.2(1958), 159~165.

Mani, I., "Automatic Summarization," *John Benjamins Publishing Company*, (2001), 114~125.

Mihalcea, R. and P. Tarau, "TextRank – Bringing Order Into Texts," *Proceeding of the Conference on Empirical Methods in Natural Language*, (2004), 8~15.

Mihalcea, R. and P. Tarau, "An Algorithm for Language Independent Single and Multiple Document Summarization," *Proceeding of the International Joint Conference on Natural Language*, (2005), 19~24.

Nenkova, A. and R. Passonneau, "Evaluating Content Selection in Summarization: The Pyramid Method," *Proceedings of HLT-NAACL*, (2004), 145~152.

Radev, D., H. Jing and M. Budzikowska, "Centroid-Based Summarization of Multiple Documents," *Information Processing & Management*, Vol.40, (2004), 919~938.

Ouyan, Y., W. Li and Q. Lu, "An Integrated Multi-Document Summarization Approach based on Word Hierarchical Representation," *Proceedings of the ACL-IJCNLP Conference Short Papers*, (2009), 113–116.

Steinberger, J. and K. Jezek, "Text Summarization and Singular Value Decomposition," *Lecture Notes for Computer Science*, Vol. 2457, (2004), 245~254.

Wan, X., “Timed TextRank: Adding the Temporal Dimension to Multi-Document Summarization,” *Proceeding of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2007), 867~868.

Abstract

Automatic Quality Evaluation with Completeness and Succinctness for Text Summarization

Eunjung Ko* · Namgyu Kim**

Recently, as the demand for big data analysis increases, cases of analyzing unstructured data and using the results are also increasing. Among the various types of unstructured data, text is used as a means of communicating information in almost all fields. In addition, many analysts are interested in the amount of data is very large and relatively easy to collect compared to other unstructured and structured data. Among the various text analysis applications, document classification which classifies documents into predetermined categories, topic modeling which extracts major topics from a large number of documents, sentimental analysis or opinion mining that identifies emotions or opinions contained in texts, and Text Summarization which summarize the main contents from one document or several documents have been actively studied.

Especially, the text summarization technique is actively applied in the business through the news summary service, the privacy policy summary service, ect. In addition, much research has been done in academia in accordance with the extraction approach which provides the main elements of the document selectively and the abstraction approach which extracts the elements of the document and composes new sentences by combining them. However, the technique of evaluating the quality of automatically summarized documents has not made much progress compared to the technique of automatic text summarization. Most of existing studies dealing with the quality evaluation of summarization were carried out manual summarization of document, using them as reference documents, and measuring the similarity between the automatic summary and reference document. Specifically, automatic summarization is performed through various techniques from full text, and comparison with reference document, which is an ideal summary document, is performed for measuring the quality of automatic summarization. Reference documents are provided in two major ways, the most common way is manual summarization, in which a

* Graduate School of Business IT, Kookmin University
** Corresponding Author: Namgyu Kim
School of MIS, 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

person creates an ideal summary by hand. Since this method requires human intervention in the process of preparing the summary, it takes a lot of time and cost to write the summary, and there is a limitation that the evaluation result may be different depending on the subject of the summarizer. Therefore, in order to overcome these limitations, attempts have been made to measure the quality of summary documents without human intervention.

On the other hand, as a representative attempt to overcome these limitations, a method has been recently devised to reduce the size of the full text and to measure the similarity of the reduced full text and the automatic summary. In this method, the more frequent term in the full text appears in the summary, the better the quality of the summary. However, since summarization essentially means minimizing a lot of content while minimizing content omissions, it is unreasonable to say that a "good summary" based on only frequency always means a "good summary" in its essential meaning. In order to overcome the limitations of this previous study of summarization evaluation, this study proposes an automatic quality evaluation for text summarization method based on the essential meaning of summarization. Specifically, the concept of succinctness is defined as an element indicating how few duplicated contents among the sentences of the summary, and completeness is defined as an element that indicating how few of the contents are not included in the summary.

In this paper, we propose a method for automatic quality evaluation of text summarization based on the concepts of succinctness and completeness. In order to evaluate the practical applicability of the proposed methodology, 29,671 sentences were extracted from TripAdvisor 's hotel reviews, summarized the reviews by each hotel and presented the results of the experiments conducted on evaluation of the quality of summaries in accordance to the proposed methodology. It also provides a way to integrate the completeness and succinctness in the trade-off relationship into the F-Score, and propose a method to perform the optimal summarization by changing the threshold of the sentence similarity.

Key Words : Text Summarization Evaluation, Text Summarization, Text Mining, Topic Modeling

Received : May 29, 2018 Revised : June 24, 2018 Accepted : June 25, 2018
Publication Type : Regular Paper Corresponding Author : Namgyu Kim

저 자 소 개

고은정
현재 국민대학교 비즈니스IT전문대학원 석사과정에 재학 중이다. 국민대학교 경영정보학부에서 학사 학위를 취득하였으며, 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 등이다.

김남규
현재 국민대학교 경영정보학부 부교수로 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 Text Mining, Data Mining, Data Modeling 등이다.