단어 간 의미적 연관성을 고려한 어휘 체인 기반의 개선된 자동 문서요약 방법

(An Improved Automatic Text Summarization Based on Lexical Chaining Using Semantical Word Relatedness)

차준석*, 김정인*, 김판구**

(Jun Seok Cha, Jeong In Kim, Jung Min Kim)

요 약

최근 스마트 디바이스의 급속한 발달과 보급으로 인하여 인터넷 웹상에서 등장하는 문서의 데이터는 하루가 다르게 증가 하고 있다. 이러한 정보의 증가로 인터넷 웹상에서는 대량의 문서가 증가하여 사용자가 해당 문서의 데이터를 이해하는데, 어려움을 겪고 있다. 그렇기 때문에 자동 문서 요약 분야에서 문서를 효율적으로 요약하기 위해 다양한 연구가 진행 되고 있다. 효율적으로 문서를 요약하기 위해 본 논문에서는 텍스트랭크 알고리즘을 이용한다. 텍스트랭크 알고리즘은 문장 또는 키워드를 그래프로 표현하며, 단어와 문장 간의 의미적 연관성을 파악하기 위해 그래프의 정점과 간선을 이용하여 문장의 중요도를 파악한다. 문장의 상위 키워드를 추출 하고 상위 키워드를 기반으로 중요 문장 추출 과정을 거친다. 중요 문장 추출 과정을 거치기 위해 단어 그룹화 과정을 거친다. 단어그룹화는 특정 가중치 척도를 이용하여 가중치 점수가 높은 문장을 선별하여 선별된 문장들을 기반으로 중요 문장을 중요 문장을 추출하여, 문서를 요약을 하게 된다. 이를 통해 기존에 연구 되었던 문서요약 방법보다 향상된 성능을 보였으며, 더욱 효율적으로 문서를 요약할 수 있음을 증명하였다.

■ 중심어 : 텍스트 랭크 알고리즘, 문서 요약, 어휘 체인

Abstract

Due to the rapid advancement and distribution of smart devices of late, document data on the Internet is on the sharp increase. The increment of information on the Web including a massive amount of documents makes it increasingly difficult for users to understand corresponding data. In order to efficiently summarize documents in the field of automated summary programs, various researches are under way. This study uses TextRank algorithm to efficiently summarize documents. TextRank algorithm expresses sentences or keywords in the form of a graph and understands the importance of sentences by using its vertices and edges to understand semantic relations between vocabulary and sentence. It extracts high-ranking keywords and based on keywords, it extracts important sentences. To extract important sentences, the algorithm first groups vocabulary. Grouping vocabulary is done using a scale of specific weight. The program sorts out sentences with higher scores on the weight scale, and based on selected sentences, it extracts important sentences to summarize the document. This study proved that this process confirmed an improved performance than summary methods shown in previous researches and that the algorithm can more efficiently summarize documents.

■ keywords : TextRank algorithm, ducument summary, Lexical Chain

Ⅰ. 서 론

최근 인터넷 웹사이트의 발달과 보급으로 인하여 웹상에는 많은 문서의 데이터가 존재 하며, 대량으로 증가하고 있다. 대량으로 증가하는 문서로 인하여 사용자가 원하는 정보를 파악하기 위해 기존의 행하였던 방법으로는 해당 문서의 제목과 내용을 간략하게 보여 주여 해결하였다. 하지만 이러한 방법으로는 사용자가 원하는 문서의 주제를 파악하기에는 부족하며, 특히 모바일 디바이스의 경우 많은 내용을 한 화면에 담기 어렵고 선택한 내용이 사용자가 의도한 주제가 아닌 경우 사용자에게 추가적인 비용을 요구하는 경우도 있다. 또한 사용자가 원하는 문

* 학생회원, 조선대학교 소프트웨어융합공학과, 조선대학교 컴퓨터공학과
** 정회원, 조선대학교 컴퓨터공학과
이 논문(저서)은 2014년 교육부와 한국연구재단의 지역혁신창의인력양성사업(NRF-2014H1C1A1073115)과 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2016R1A2B4012638)

접수일자 : 2017년 03월 09일
수정일자 : 2017년 03월 16일
게재확정일 : 2017년 03월 29일
교신저자 : 김판구 e-mail : pkkim@chosun.ac.kr

서의 정보를 얻기 위해 많은 웹서비스에서 문서의 검색서비스를 찾을 수 있도록 되있지만, 일일이 문서를 읽고 스스로 정보를 확인하는 작업이 필요하다. 문서의 내용을 정확하게 표현하는 문서 내용의 형태를 취할 수 있다면, 사용자가 원하는 문서의 정보를 찾는 데 큰 도움이 될 것이다. 문서 요약이란 한 문서에서 핵심적인 내용을 추려 사용자에게 중요한 내용을 알려주는 것을 목표로 한다. 한 문서에서 담고 있는 핵심 의미를 유지하면서 문서의 크기를 효과적으로 줄여 문서의 내용을 쉽게 이해할 수 있어야 한다.

문서요약의 기본 시스템은 뉴스, 웹 페이지, 텍스트 마이닝 학회에서 제공해주는 데이터 셋을 문서요약 시스템에 입력한 뒤, 전처리 과정을 거쳐 키워드를 추출한다. 추출된 키워드를 문서요약시스템의 주요 방법론을 적용하여, 자동으로 문서를 요약하는 방법이다. 문서를 요약하기 위해서는 문서의 특징을 반영한 방법이 필요하다. 문서요약의 기본 목적은 원문을 읽지 않고서도 원문의 주제를 파악할 수 있도록 문서의 핵심 주제를 간략하게 압축 정리하는 것에 있다. 문서요약은 원문의 주제를 서술적으로 표현해야 하므로 복잡한 언어처리와 주제 분석 같은 고차원적인 문서 분석 기술을 필요로 한다. 본 논문에서는 문서에 존재하는 키워드를 상위 키워드와 하위 키워드로 추출하여 추출된 키워드를 기반으로 문서를 요약하게 된다. 키워드 추출을 위해 본 연구에서는 텍스트 랭크 알고리즘을 이용하여 해당 문장에 존재하는 문서의 문장을 간선그래프를 이용해 모델링 한 후, 각 키워드의 높은 점수의 키워드를 상위 키워드로 선정되며, 선정된 상위 키워드와 하위 키워드를 기반으로 중요 문장 추출 과정을 거친 후 문서를 요약하게 된다.

Ⅱ. 관련 연구

1. 생성요약 기법

최근 문서 요약 연구에는 문서에 등장하는 단어와 문장 간의 일관성을 유지하면서 문장의 중복을 제거하고 응축된 정보를 생산함으로서 문서의 복잡도를 줄이는 연구가 있으며[1], 그 내용의 구성 방법에 따라 생성요약과 추출 요약으로 구분할 수 있다. 생성 요약은 원문서로부터 중요한 단어들을 선별한 후 자연어처리 기법을 이용하여 새로운 문장을 구성한 후 요약문으로 제공하는 것이다.

생성 요약 기법은 문서 내 존재하는 중요 키워드와 문장을 파악하여 자연어 처리 기법을 이용한 문장조합기법, 재구축기법등을 이용하여 문서요약을 하는 연구이다. 또한 생성요약 기법은 문서에 존재하는 단어의 위치, 문장의 위치 같이 대상 문서에서 형태적으로 나타나는 정보를 이용하여 문서요약을 진행한다. 생성요약 기법은 요약 속도가 빠르고 시스템구형이 쉬우나 다의어, 유사어와 같은 문장의 의미 구분을 하지 못함으로서, 문서를 지나치게 단순한 통계 테이블로 간 경우가 빈번하게 발생하였기 때문에, 생성요약 기법에 관련된 초기 연구로는 문서의 주제를 표현하는 단어가 자주 사용된다는 직관에 의거하여, 가장 많이 사용되는 단어를 문서의 주제로 결정하는 연구가 진행 되었다[2]. 생성요약 기법은 주로 단어가 포함된 문장을 생성 및 추출을 함으로써 문서의 문장에 생성된다. 기존에 꾸준히 연구되어 오던 생성 요약 기법을 응용하여 문장의 위치에 따른 문서요약 연구가 있다. 문서상의 중요한 문장과 사용자가 개입된 질의 확장을 이용하여 문서를 요약하는 연구가 있으며[3], 문서의 구문형식에 포함된 정보인 주제, 용어의 빈도수 정보 등을 활용할 수 있는 문서 요약 방법이 제안되었으며[4], 생성 요약 기법은 주로 첫 번째 사용된 문장 또는 마지막에 사용된 문장이 같은 위치에 따라 문장의 중요도가 다르다는 연구가 발표되었다[5]. 아울러 같은 단서어도 중요 문장을 파악하기 위해 텍스트랭크 알고리즘을 이용하여 해당 문장에 중요한 역할을 할 수 있음을 보였으며, 이 외에 문서요약 연구에 기계 학습을 통해 문서의 중요도를 습득함으로써 문서요약 성능을 향상 시킨 연구도 있었다[6]. 하지만 이 방법들은 모두 자동 문서요약의 기초가 되는 역할을 하였으나, 지나치게 단순한 통계에 의존함으로써 문서의 주제를 파악하는데 한계를 보였다.

최근 새롭게 연구되고 있는 생성요약 연구에는 워드넷을 이용한 문서요약 기법이 활발히 진행 되고 있다. 다양한 문서를 분석하여 문서를 요약하기 위해서 각 문서마다 갖는 특징을 고려하여, 워드넷을 기반으로 한 문장의 분석이 이루어짐으로써 보다 정확한 문서요약이 이루어진다. 최근 담화이론을 이용한 생성요약 기법의 연구가 있다. 담화 이론에 따르면 문서는 크게 중심 부분과 주변 부분으로 구성된다는 가정 하에 두 부분 사이의 수사 관계를 이용하여 요약을 생성하는 문서요약 연구 기법이다[7]. 이 방법은 문서의 문장이 주어지면 수사구조 알고리즘에 의해 담화 트리를 생성하며, 담화 트리의 단말 노드는 문장의 구, 절, 문장 등이 되며, 내부 노드는 해당 자식 노드 사이의 수사관계를 표현하게 된다. 문서요약의 문장을 생성하기 위해서는 담화 트리의 각 노드에 대해 부모 노드가 자식 노드보다 높은 값을 갖도록 부여하여 그 값이 높은 순으로 정렬한다. 정렬 결과 순위가 높은 구, 절, 문장들을 요약으로 제시한다. 문서요약의 생성요약 기법은 문장의 수사관계, 문장의 모호성등 해결해야 할 문제점들이 많이 남아 있으며, 이를 위해 추출 요약 기법의 문서요약 방법이 대두 되고 있다.

2. 추출요약 기법

추출 요약 기법은 문서 내 문장들을 가지고 구, 절, 문장 등을 새롭게 분석하여 문서 내에 중요도를 판단 한 후 문서를 요약을

하는 기법이다.

추출 요약 기법은 현재 다양한 연구를 통해 문장이 가지는 단어의 빈도수 및 가중치를 통해 문장과 단어 간의 관계를 분석하여 중요 문장을 추출하는 방식으로 이루어지고 있다. 하지만 자동으로 분석된 문장의 가중치가 기존 문서의 의미전달이 제대로 이루어지지 않는 경우 올바른 문장 요약이 이루어지지 않는다. 상대적으로 기존 생성 요약 기법보다는 구현이 쉽다는 장점 때문에 현대 추출 요약 기법을 이용한 문서 요약은 활발히 진행되고 있으며, 추출 요약 기법을 이용한 연구 중 PLSA(Probavilistic Latent Semantic Analysis) 알고리즘을 이용한 문서 요약 기법이 있다. 문서 내에 존재하는 단어들에 포함된 숨겨진 주제들 별로 클러스터들을 만들고 클러스터 된 단어와 문장과 관계를 유사도 측정을 수행하여 문장에 점수를 부여하는 기법을 제안하였다[8]. 기존의 PLSA 알고리즘을 이용한 문서 요약 연구를 개선시키기 위해 클러스터링 알고리즘을 이용하여 문서를 요약한 기법이 있다.

이 기법은 문서의 주제에 대한 클러스터들을 만들고, 링크분석(Link-Anaysis)기법인 HITS(Hypertext Induced Random Walk) 기법을 통한 문장들과 문서의 주제 클러스터들 간의 연관성을 분석하여 문장에 점수를 부여하는 기법을 제안하였다. 이러한 기법들은 비교적 우수한 결과 값을 얻을 수 있지만, 전처리 단계에서 사용되는 클러스터링 알고리즘과 분석단계에서 PLSA와 조건부 마르코프 알고리즘 또는 기계학습 알고리즘을 이용한 복잡한 분석을 수행하기 때문에 높은 계산 비용과 분석 시간이 요구된다는 문제점이 있다[9]. 기존의 문서 요약 기법과는 달리 문서의 중요 문장을 효율적으로 추출하기 위해 질의에 대한 확장과 질의 분해의 방법으로 문서를 요약한 연구가 있다. 이 방법은 문서 요약을 적합한 문장의 선택 작업으로 간주하여, 문서의 주요 정보를 검색하기 위해 사용하는 질의 확장 기법을 문서 요약에 적용한 것이다. 적합 문장을 이용하여 초기 질의를 확장할 때 적합 문장 전부를 초기 질의에 한꺼번에 적용하기 않고 적합 문장 각각의 개별적으로 질의 확장을 적용하여, 적합 문장 개수만큼 질의로 분해하는 방법이다[10]. 최근 추출 요약 기법의 동향으로는 문서를 요약하기 위해 그래프로 표현하며, 단어의 위치 정보와 더불어 소셜 네트워크를 이용한 단어의 의미적 연관성을 고려하는 문서요약 방법이 제안되어 연구 되고 있다[11].

사용자가 여러 문서에서 정보를 찾는 것은 정보 요구를 해결할 수 있는 답을 찾기 위한 것으로, 단어의 출현 빈도나 문장의 문맥구조 등의 특성을 사용하여 문장 단락의 중요도를 계산한다. 계산된 가중치 값 중 특정 노드의 중요도를 노드간의 의미적 유사성을 반영하지 못한다는 점에 착안하여 개선시킬 수 있다[12].

III. 본 론

본 논문에서는 추출 요약 기법을 기반으로 자동 문서요약을 하기 위한 문서요약 시스템을 제안하고자 한다. 그림 1은 제안하는 문서요약 시스템의 전체 시스템 구성도를 나타내고 있다

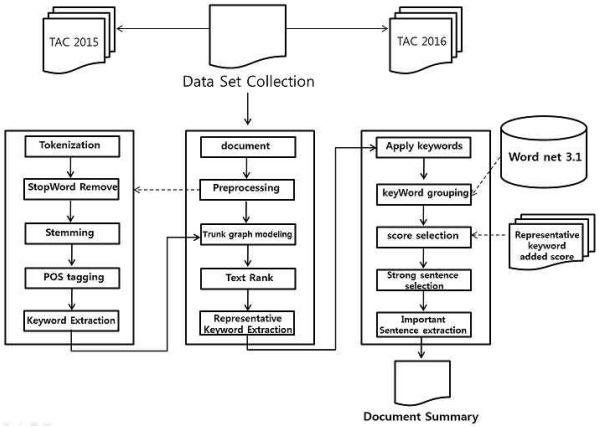


그림 1. 문서요약 시스템 구성도

그림 1 은 본 논문에서 제안한 문서 요약 시스템 구성도이다. 시스템 구성도는 크게 대표 키워드를 추출하는 단어 분석 단계와 각 단어의 연관성을 고려한 문서요약 단계인 중요 문장 추출 단계로 구성된다. 단어 분석 단계에서는 텍스트 랭크를 측정 하는 단계와 텍스트 랭크를 통해 추출된 핵심 키워드인 대표 키워드를 추출하기 위해 상위 키워드와 하위 키워드로 나누는 단계를 거친다. 텍스트 랭크를 측정하기 위해 먼저 키워드를 추출하는 단계에서 토큰화, 불용어 제거, 어간추출, 품사 추출, 키워드 추출 단계를 거치며, 추출된 키워드를 기반으로 본 논문에서 제시하는 텍스트 랭크 측정 방법을 통해 대표 키워드를 측정하여 상위 키워드와 하위키워드로 나누어지게 된다. 추출된 대표 키워드를 이용하여 문서요약을 하는 중요 문장 추출 과정을 거치게 된다. 문서요약을 하는 중요 문장을 추출하기 위해 단어와 문장사이의 연관성을 고려하여 각 문장 별 가중치 점수 값을 구하여 문서의 요약 방법을 제안한다.

1. 대표키워드 추출

문서 요약을 위해 본 장에서는 대표 키워드를 추출하는 단어 분석 단계를 거친다. 단어 분석단계에서는 전처리 과정을 거쳐 키워드를 추출 후 텍스트랭크 알고리즘을 적용하여, 상위 키워드와 하위 키워드를 추출하게 된다. 대표 키워드를 추출하기에 앞서 문서에 존재 하는 키워드를 먼저 추출해야 한다. 본 논문에서 제시하는 전처리 과정은 토큰화, 불용어 제거, 어간 추출, 품사 판별을 거쳐 키워드를 추출하게 되며, 키워드 추출을 위한 실험 데이터 셋은 TAC(Text Analysis Conference)에서 제

공해주는 데이터 셋 TAC 2015와 TAC 2016을 이용한다.

전처리 과정의 토큰화는 문장 내에서 공백 또는 특정 단어를 기준으로 문장을 나누는 것을 의미하며, 문장을 어절 단위로 세분화 한다. 두 번째 불용어 제거 단계는 토큰화된 문장의 단어 중 불용어 리스트를 이용하여 불용어를 제거 시킨다. 불용어는 인터넷 검색을 할 시에 많이 쓰이지 않는 단어를 의미한다. 세 번째 어간 추출은 'es', 'ed' 같은 불필요한 어간을 추출하여, 단어의 원형으로 변경하는 것을 말한다. 마지막 품사 판별 단계는 명사, 형용사, 동사의 키워드를 추출하기 위한 이전 단계로서 각 단어의 품사를 추출하는 단계이다. 마지막 키워드 추출 단계에서는 품사가 정의된 단어들의 품사 중 명사, 형용사, 동사만을 추출한다. 전처리 과정을 통해 추출된 키워드들을 이용하여 텍스트랭크 알고리즘에 적용하기 위해서는 처리하고자 하는 문서의 문장을 간선 그래프 형태로 모델링해야 한다.

우선 그래프를 모델링하기 위해 그래프를 구성할 정점을 결정해야 한다. 정점은 전처리과정을 통해 모든 불용어를 제거한 문서의 단어들로 표현하며, 간선은 단어로 표현되는 정점들 간의 의미적 관계를 뜻한다. 간선 그래프를 기반으로 대표 키워드의 상위 키워드를 추출하는 경우 키워드 사이의 간선 형성의 기준이 될 수 있고, 간선의 가중치를 함께 등장하는 횟수로 결정할 수 있다. 하나의 문장에 각 정점과 간선으로 연결된 그래프에서 soul이라는 단어가 다음 문장에 존재한다면 soul을 기점으로 두 번째 문장이 간선 그래프로 연결되어 모델링 된다. 하지만 다음 문장에 기존 간선 그래프의 단어가 존재하지 않다면 간선그래프에 마지막으로 생성된 oak-line라는 단어부터 두 번째 간선 그래프가 생성된다.

구축된 간선 그래프를 기반으로 문서요약을 위해서 문장에 존재하는 대표 키워드를 추출하기 위해 텍스트랭크 적용 과정을 거치게 된다. 모델링된 그래프에 텍스트 랭크 알고리즘을 적용해야 하며, 간선 그래프를 기반으로 텍스트의 중요도를 결정하는 알고리즘의 종류는 다양하게 연구 되어오고 있다. 하지만 대부분 자연어처리에서 적합한 형태가 아니므로 자연어 처리에 적용하기 위해서 다소 변형과 응용을 거쳐야한다[13]. 이 때, 적용할 텍스트랭크는 세 가지 가중치 합을 기반으로 텍스트의 중요도 점수를 얻어 키워드의 랭킹을 산정 한다. 랭킹을 산정하기 위한 수식은 다음과 같다.

$$TR(V_i) = (1-d) + d\sum_{j=0}^{N-1} \frac{W_{ij}}{\sum_{k=0}^{N-1} W_{jk}} \times TR(V_j) \qquad (1)$$

수식 (1)의 TR(vi)는 단어 I에 대한 텍스트 랭크 값을 의미하며, Wij는 각 두 단어 i와 j사이에 존재하는 간선 그래프의 가중치 값을 의미한다. 값 d는 페이지 랭크 알고리즘에서 해당 vi에 대해 사용자가 해당 웹페이지를 임의로 선택할 가능성을 나타내는 페이지 랭크 알고리즘의 제동 계수(damping factor)로 0.85의 값을 갖는다.

n-1은 wij의 가중치 값을 구하기 위해 간선 그래프에 연결된 모든 단어들의 집합을 의미한다. 그 후 TR(vi)값이 계산되면 문서 내의 모든 단어들에 대한 점수 테이블을 생성할 수 있다. 생성된 점수 테이블은 문장 추출 단계를 거치기 위해 중요한 수식 값을 갖는다. 마지막으로 각 키워드의 점수를 내림차순으로 정렬하여 값이 가장 큰 n개의 단어의 링크 정보와 수집한 문서 데이터의 문장과 비교하여 해당 문장에서 텍스트 랭크 가중치 값의 평균 이상인 값들에서만 대표 키워드로 지정 된다. 표 2는 수집한 데이터 셋에서 추출된 대표 키워드의 랭크 값을 나타낸다.

표 1. 대표 키워드

keyword	score	keyword	score
entourage	0.17685	corruption	0.10270
protest	0.16612	economy	0.10140
presidents	0.14107	predecessor	0.09232
philippine	0.13026	parliament	0.09110
Economic	0.12647	Minister	0.09912
Cooperation	0.12013	Mahathir	0.08721
helicopters	0.11619	business	0.07512
Estrada	0.11470	Indonesia	0.07210
dismissal	0.11407	problems	0.07010
Ibrahim	0.10422	counterparts	0.60741
Southeast	0.10775	difficulty	0.60452

2. 문서 요약

표 1의 대표 키워드는 entourage, Protests, president, Philippine 순으로 키워드의 데이터가 내림차순으로 저장되며, 이를 기반으로 상위 키워드로 분류가 되며, 대표키워드로 측정되지 않은 이외의 키워드를 하위 키워드로 분류가 되어 문서를 요약하기 위한 문장 분석 단계를 거치게 된다. 텍스트 랭크를 통해 추출된 대표 키워드를 기반으로 중요 문장 추출 과정을 거친다. 중요 문장 추출을 위해 본 논문에서는 어휘 체인을 이용한다. 어휘 체인은 텍스트에서 문법적인 장치를 제외한 어휘적 의미만을 이용하여 해당 텍스트를 분석하고 각 의미적 연관성이 있는 단어들끼리 그룹화 하는 시스템이다. 본 논문에서 제안한 어휘 체인을 이용하기 위해 워드넷 계층 구조를 이용하였다. 어휘 체인의 단어 그룹화를 하기 위해 워드넷이 제공하는 개념 간의 계층 관계는 두 단어의 의미가 서로 얼마나 밀접한가를 측정하는데 매우 중요한 척도로 사용될 수 있다. 단어 그룹화를 하기 위해 워드넷 상에서 의미가 정의된 단어의 상·하위어 관계, 동의어 관계를 이용한다. 그림 2는 단어의 의미적 연관성을 찾기 위한 어휘 체인의 예시도이다.

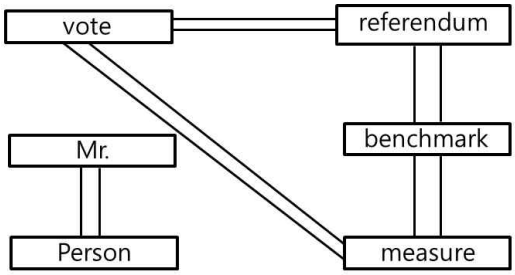


그림 2. 어휘 체인 예시도

그림 2에서 보는 바와 같이 vote의 의미가 “투표, 표결”로 사용되었다면 ‘referendum’의 뜻인 “국민 투표, 총선거”라는 뜻으로 해당 단어의 의미적 연관성이 성립된다. 하지만 ‘vote’가 “투표하는 사람, 채택하는 사람”의 의미로 사용되었다면 ‘referendum’과 ‘vote’사이에서는 단어 간의 의미적 연관성이 성립되지 않는다. ‘referendum’과 ‘benchmark’, ‘measure’ 사이에서는 의미가 같은 유의어 관계이므로 각 단어 간의 의미적 연관성이 성립되어 어휘 체인이 생성된다. 생성된 어휘 체인을 기반으로 중요 문장을 추출하기 위해 강한 문장 가중치 수식을 적용하여 각 문장별로 점수를 선정해야 한다. 이를 위한 수식은 다음과 같다.

$$Score(chain) > Avg(Score) + 2 \times StandardDeviation(Score) \quad (2)$$

수식 2는 본 장에서 언급한 어휘체인을 이용해 각 문장별로 가중치 점수를 부여하여 문서에 핵심이 되는 문장을 추출하게 된다. 두 키워드 간의 단어 간 연관성이 성립되지 않은 경우에도 중요 문장 추출 과정에 있어 가중치 점수를 부여 받게 된다. 수식 2의 Avg(Score)는 각 수집한 문서의 문장에 단어 그룹화로 가중치 점수를 부여하게 된다. 하지만 그림 2에서와 같이‘vote’와 ‘referendum’가 단어의 의미적 연관성이 성립되지 않고, 단어 그룹화가 생성 되지 않았으나, 수집한 문서의 문장에 해당 단어가 존재 할 경우 0.5점을 부여 받게 된다. ‘vote’와 ‘referendum’의 의미가 같아 단어 그룹화가 생성 되었을 경우는 1.0점을 부여받게 된다. 생성된 단어 그룹화에 텍스트 랭크 알고리즘을 통해 추출된 상위 키워드가 존재 하여, 추가 점수를 받게 된다. 상위 키워드로 이루어진 단어가 문장에 존재 할 경우 추가 점수 1.0점을 받아 강한 문장 가중치 점수 2.0점을 받게 되며, 단어 그룹화에 상위 키워드와 일반 키워드로 생성 되었을 경우 0.5점의 추가 점수를 받으며, 1.5점을 부여 받게 된다. 단어 그룹화가 생성 되지 않았고, 해당 단어가 상위 키워드로 이루어진 경우 0.5점의 추가 점수를 받아 1.0점의 가중치 점수를 받게 된다. 수식 2의 Standard Deviation(Score)는, 각 문장 별로 이루어진 가중치 점수 선정 한 뒤 가중치 점수의 표준 편차를 구하게 되며, 표 2는 어휘 체인의 가중치 점수를 이용하여 강한 문장 선정 과정을 위한 가중치 점수 평균을 도출한 결과이다.

표 2의 가중치 점수 평균 결과에서 [‘problems’, ‘difficulty’]는 어휘 체인이 되어 해당 문장에 존재 할 경우 1.0점의 가중치 점수를 받으며, ‘problems’과 ‘difficulty’는 상위 키워드에 속하므로 추가 점수 1.0점을 받아 2.0점의 가중치 점수를 받게 된다. 표 3에서 보는 바와 같이 가중치 점수 평균 이상인 문장들이 중요 문장으로 추출된다. 가중치 점수 평균 결과 1.5점 이 나왔으며, 문장 1번과 문장 2번이 평균 이하의 문장이므로 해당 문서에서 삭제되어 문서가 요약 된다.

표 2. 어휘 체인 가중치 부여 결과

	영어 문장	
원문	leaders difficult because of his concerns about the arrest of Malaysia’s former deputy prime minister, a Thai newspaper reported Sunday.……	
추출된 키워드	leaders, difficult, arrest, former, deputy, prime, minister, newspaper, report, Sunday, Asia, pacific, schedule, Last, week, philippine, economy, ideology, rivalry, genocide, organization ……	
상위 키워드	entourage, protests, presidents, Philippine, President, Economic, cooperation, helicopters, Estrada, Singapore, dismissal, Ibrahim, Southeast, concerns, treatment, Panorama, interview, President, corruption, economy, predecessor, parliament, Minister, Mahathir, business, Indonesia, problems, counterparts ,difficulty. minister.	
어휘체인	['nepotism'], ['interview'], ['President'], ['Mahathir'], ['corruption', 'corruption'], ['treatment'], 'buildings'], ['parliament'], ['Ngezayo'], ['minority'], ['business'], ['pastures'], ['servants'], ['Indonesia'], ['problems', 'difficulty'], ['paycheck'],……	
문장별 가중치 점수	문서의 문장 집단1	1.4
	문서의 문장 집단2	1.2
	문서의 문장 집단3	1.6
	이하 생략	이하생략
가중치 평균	1.56728052427	

VI. 실험 및 결과

1. 실험 평가 방법

실험의 평가는 ROUGE(Recall-Oriented Understudy of Gisting Evaluation)평가 시스템을 이용하였다[14]. ROUGE 평가 시스템은 전문가가 직접 요약한 문서와 자동으로 요약된 시스템 문서를 비교 평가 하는 방법으로 본 논문에서는 ROUGE의 시스템 중 ROUGE-N을 이용하여 논문의 실험을 평가 하였다.

ROUGE-N은 n-gram을 이용하여 재현률, 정확률을 이용해 최종적으로 F-Score의 결과를 측정하는 방법이다. n-gram은

n개의 어절 또는 음절을 연쇄적으로 분류해 문서의 빈도를 측정 하는 방법이다[15]. n=1 일 때는 uni-gram, n=2 일 때는 bi-gram, n=3는 tri-gram, n=4 일 때는 four-gram을 기준으로 하며, ROUGE-N을 측정하기 위한 방법은 다음과 같다.

$$ROUGE-N=\frac{\sum_{S\in[ReferencesSUmmary]}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{S\in[ReferencesSUmmary]}\sum_{gram_n\in S}Count(gram_n)} \quad (3)$$

$gram_n$은 n-gram의 길이이며, $Count_{match}(gram_n)$은 전문가 요약문과 문서요약 시스템의 요약문이 동시에 발생한 최대 n-gram(1,2,3,4)의 개수이다.

2. 실험 평가 방법

본 절에서는 문서요약 시스템의 효율성 검증을 위해 강한 문장 가중치 점수를 이용한 결과 분석과 문서요약 시스템의 성능 검증을 위해 베이스 라인 실험을 진행한다.

첫 번째 결과 분석은 강한 문장 가중치 점수를 이용한 결과 분석이다. 본 논문에서 제시하고 있는 방법 중 텍스트랭크 알고리즘의 대표 키워드의 추가 점수를 부여하지 않고 어휘 체인의 각 문장별 가중치 점수만을 추가적으로 점수만을 올린 방법과 본 논문의 텍스트 랭크 알고리즘의 대표 키워드 추가 점수를 주는 방법의 효율성 평가를 하였으며, 먼저 각 문장별 가중치 점수만을 올리는 방법은 각각 0.2, 0.4. 0.6점만을 올린 방법 중에서 가장 좋은 결과 나온 점수를 기반으로 본 논문에서 제시하고 있는 실험 방법과 비교하여 결과 분석을 하였다. 또한 0.2점을 기준으로 비교 실험 한 이유는 0.1점씩 실험을 진행 하였을 때 해당 실험에 대한 결과가 미비하게 나올 수 있기 때문에 0.2점씩 기준으로 정하여 실험을 진행 하였으며, 또한 0.1점씩 실험을 진행 하더라도 같은 결과가 나왔으며, 0.2점을 기준으로 하였을 때, 해당 실험에 대한 차이가 명확하고 자세한 결과가 나왔기 때문에 0.2점씩 차이를 두어 실험을 진행 하였다. 강한 문장 가중치 점수만을 이용한 실험 결과는 다음과 같다.

그림 3과 그림 4는 강한 문장 가중치 점수만을 추가적으로 점수를 주어 나타난 결과 이며 ROUGE-N의 uni-gram, bi-gram, tri-gram, four-gram을 기준으로 평가를 하였다. 평가 결과로 미루어 볼 때 각 문장별 가중치 점수 선정 시 가중치 점수 0.2점을 추가한 방법이 Bi-gram을 기준으로 하였을 때 재현율과 F-Score의 측면에서 가장 좋은 성능을 나타내었으며, uni-gram과 tri-gram, four-gram을 기준으로 한 방법역시 각각 재현율, 정확률, F-Score의 값이 효율성이 좋은 결과로 나타낸 것을 알 수 있었으며, 강한 문장 가중치 점수 0.2점만을 이용한 결과가 좋은 성능을 나타나는 것을 확인 할 수 있었다.

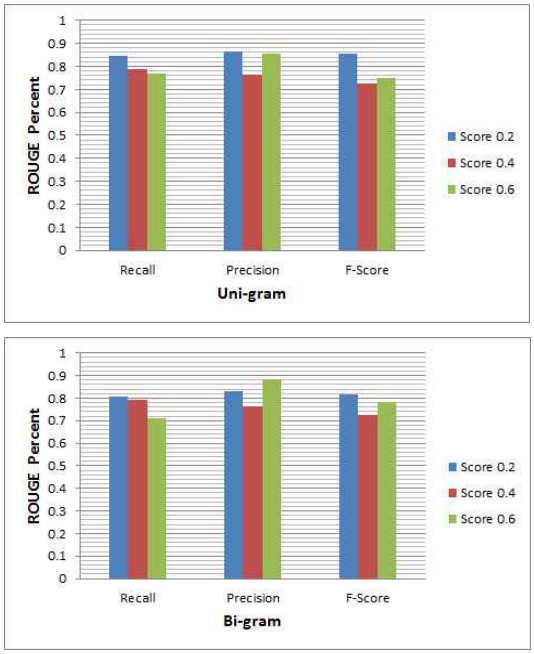


그림 3. Uni-gram 및 Bi-gram 가중치 선정 결과

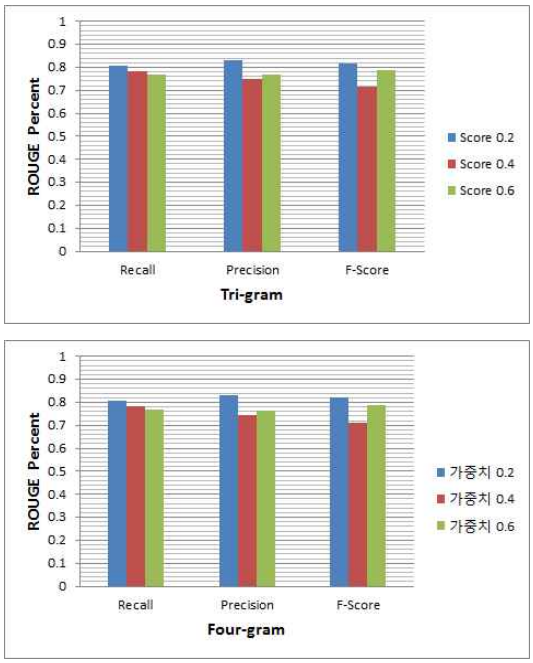


그림 4. Tri-gram 및 Four-gram 가중치 선정 결과

이를 기반으로 본 실험에서는 텍스트랭크 알고리즘을 추가하지 않고 가중치 점수 0.2점만을 추가한 방법과 본 논문에서 제시하는 문서요약 방법을 비교 실험을 진행 해보았다.

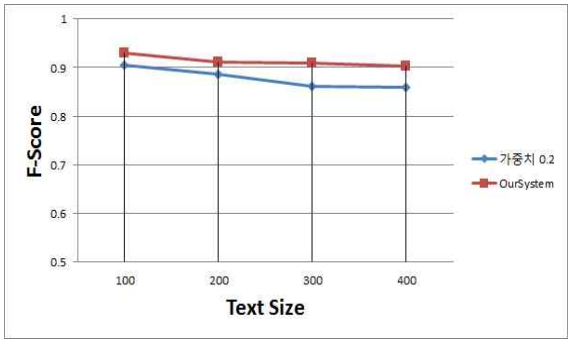


그림 5. 강한 문장 가중치를 이용한 비교 실험 결과

본 실험의 텍스트 랭크 알고리즘 기반의 상위 키워드 추가 점수 성능 검증을 위해 TAC에서 제공하는 TAC 2015와 TAC 2016의 문서 중 400문서를 이용하여 비교 평가를 진행 하였으며, 그림 5는 전체적인 ROUGE의 F-Score평균을 기반으로 나타난 결과이다.

그림 6에서 보는 바와 같이 단순히 점수를 올리는 방법보다 가중치 점수를 다양하게 선정 하는 방법이 기존에 가중치 점수만을 추가한 방법 보다 효율적으로 좋은 성능이 나타나는 것을 알 수 있었다. 두 번째 결과 분석은 기존에 연구 되었던 문서요약 연구인 추출요약 기법 기반의 다중 문서 요약 시스템과 비교 평가 실험을 진행 하였다[16]. 비교 평가할 문서 요약 시스템은 본 논문의 초점인 추출요약 기반으로 진행된 연구이며, 비교 평가 결과는 다음 그림 6과 같다.

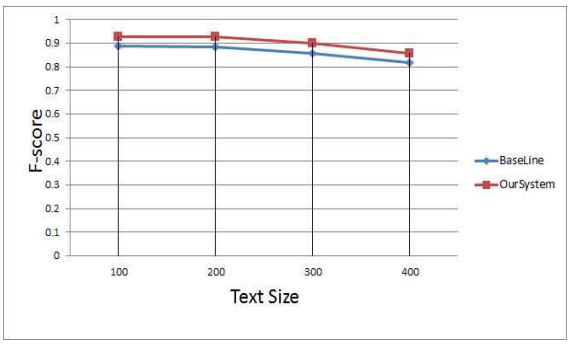


그림 6. 비교 평가 실험 결과

TAC에서 제공하는 데이터 세트 문서 중 400문서를 이용해 비교 평가한 결과이며, 비교 결과를 ROUGE-N의 전체적인 평균의 F-Score값을 보여주고 있다. 100문서를 기준으로 하였을 때, F-Score값이 베이스 라인 시스템보다 0.92723값이 도출되어 좋은 결과가 나왔다. 200문서를 기준으로 F-Score값을 도출 하였을 때는 0.92589값이 나왔으며, 300문서를 기준으로 실험을 도출 하였을 때는 0.90045의 값이 나왔다. 마지막으로 400문서의 F-Score값을 도출한 결과는 0.85889의 값이 나왔다. 그림 6 에서 보는 바와 같이 기존의 연구 되었던 문서요약 방법보다 본 논문에서 제안하는 문서요약 시스템을 사용했을 때와 비교해 성능 차이가 있다는 것을 확인 할 수 있었다.

V. 결 론

본 논문에서는 효율적으로 문서 요약을 하기 위해 문서에 존재하는 대표 키워드를 추출하며, 추출된 대표 키워드를 상위 키워드와 하위 키워드로 나누는 방법을 제안하였다. 이 방법은 문서에 존재하는 문장과 키워드 간의 연관단어를 파악하여 기존에 연구되었던 문서 요약 방법보다 문서에 존재하는 핵심 주제를 추출하기 위해 제안한 방법이다. 대표 키워드를 이용하여 문서를 요약하기 위해 전처리 과정을 이용하였으며, 전처리 과정을 통해 추출된 일반 키워드를 텍스트 랭크 알고리즘을 이용하여 대표 키워드를 추출하였다. 추출된 대표 키워드를 상위 키워드로 지정하였으며, 상위 키워드 이외의 키워드는 하위 키워드로 지정하였다. 추출된 상위 키워드와 하위 키워드를 이용하여 문서 요약을 하기 위해 단어 그룹화라는 방법을 이용하였으며, 단어 그룹화를 통해 중요 문장을 추출하는 방법을 제안하였다. 기존에 연구되었던 가중치를 이용한 문서 요약 방법에서는 가중치 점수를 1.0, 0.5, 0점을 주어 문서를 요약하였지만 가중치 점수를 해당 문장에 0점을 부여하였기 때문에 핵심 주제를 추출하는 데에 있어 정확도가 떨어졌었다.

본 연구에서는 0점을 부여하지 않고, 상위 키워드와 하위 키워드로 나누어진 키워드를 단어와 문장 간의 가중치 점수를 줄 수 있었으며, 단어 그룹화에 상위 키워드가 존재할 경우 가중치 점수의 추가적인 점수를 부여하여 중요 문장 추출의 정확도를 올려 기존에 연구 하였던 문서 요약 방법보다 핵심적인 주제를 추출하여 요약할 수 있었다. 향후 연구로는 보다 다양한 연관단어를 파악할 수 있는 키워드 태그 클러스터를 구축하여 문장과 단어 사이의 관계에 대해서 정확도를 높이기 위한 문서 요약 방법을 제안하고자 한다.

References

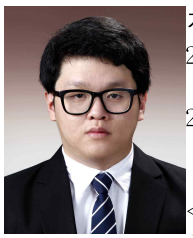
[1] Ohm Sornil, Kornnika Gree-ut, “An Automatic Text Summarization Approach using Content-Based and Graph-Based Characteristics,” *In Proceedings of IEEE Conference on Cybernetics and Intelligent Systems*, pp. 1-6, 2006.

[2] 이창범, 김민수, 이기호, 이귀상, 박혁로, "주성분 분석을 이용한 문서 주제어 추출," *정보과학회논문지 : 소프트웨어 및 응용*, pp. 747-754, 2002.

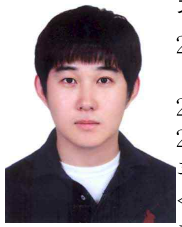
[3] D.D. Lewis, S.K. Jones, "Natural language proces

sing for information retrieval," *Communications of the ACM,* Vol. 39, No.1, pp. 92-101, 1996.
[4] E.D. Liddy, S.H. Myaeng, "DR-LINK's: linguistic-comceptual approach to document and detection, " *The First Text REtreival Conference*, pp. 113-129, 1993,
[5] Mihalcea, Rada, Paul Tarau. "TextRank: Bringing order into texts." *Association for Computational Linguistics*, 2004.
[6] J. Kupiec, J. Pedersen, and F. Chen, "A Trainable Document Summarizer." *Proceedings of 18th ACM-SIGIR Conference*, pp.68-73, 1995.
[7] Chanback Jeong, Taehwan Kim, Hochul Jeon, Joongmin Choi, "A News Recommendation System based on Document Clustering Using WordNet," *Department of Computer Science & Engineering Hangyang University*, 2008.
[8] I. Mani, "Automatic Summarization," *John Benjamins Publishing Company*, pp. 114-125, 2001.
[9] Henning, Leonhard, "Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis", *Proceedings of the International Conference RANLP'09*, 2009
[10] Xiaojun Wan, Jianwu Yang, "Multi-Document Summarization Using Cluster-based Link Analysis," *Proceedings of the International Conference (SIGIR'08)*, 2008.
[11] K.S. Thakkar, R. V. Dharaskar, and M. B. Chandak, "Graph-based algorithms for text summarization",*In 2010 3rd International Conference on Emerging Trends in Engineering and Technology*, pp. 516-519, 2010.
[12] 조형락, 김성진, 이동호. "의미기반 텍스트 랭크 알고리즘을 이용한 다중문서 요약. *한국정보과학회 2015년도 동계학술발표회 논문집,* 756-758쪽, 2015
[13] S. Harabagiu, L. Finley "Topic Themes for Multi Document Summarization," *In proceeding of ACM SIGIR*, pp. 202-209, 2005.
[14] C.Y.Lin, E.H.Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics", *In Proceedings of Human Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada*, May, 2004.
[15] P. McNamee, J. Mayfield, "Character N-Gram Tokenization for European Language Text Retrieval", *Information Retrieval*, Vol 7, No. 1-2, pp. 73-97, 2004.
[16] J. Heu, Y. Joo, D. Lee, "Multi-Document Summarization Technique using Semantic Analysis between Tags," *Journal of KIISE*, 2012.

저 자 소 개



차준석(학생회원)
2015년 조선대학교 컴퓨터공학과 졸업(공학사).
2017년 조선대학교 산업기술융합대학원 소프트웨어융합공학과 석사 졸업(공학석사).
<주관심분야 : 텍스트 마이닝, 정보검색, 자연어처리>



김정인(학생회원)
2011년 조선대학교 컴퓨터공학과 졸업(공학사)
2014년 조선대학교 박사수료.
2017년 현재 조선대학교 석·박사 연계과정.
<주관심분야 : 소셜 네트워크, 정보처리, 시맨틱 웹과 온톨로지>

김판구(정회원)
1988년 조선대학교 컴퓨터공학과 학사 졸업(공학사).
1990년 서울대학교 컴퓨터공학과 석사 졸업(공학석사).
1994년 서울대학교 컴퓨터공학과 박사 졸업.
<주관심분야 : 정보검색, 시맨틱웹, 자연어처리, 빅데이터>