creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

조선대학교
CHOSUN UNIVERSITY

[UCI]I804:24011-200000265978

2017年 2月
석사학위논문

단어 간 의미적 연관성을 고려한 개선된 문서요약 방법 연구

조선대학교 산업기술융합대학원

소프트웨어융합공학과

차 준 석

조선대학교
CHOSUN UNIVERSITY

2017년 2월 석사학위논문

단어간 의미적 연관성을 고려한 개선된 문서요약 방법 연구

차 준 석

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

단어 간 의미적 연관성을 고려한 개선된 문서요약 방법 연구

An Improved Automatic Text Summarization Using Semantical word relatedness

2017년 2월 24일

조선대학교 산업기술융합대학원

소프트웨어융합공학과

차 준 석

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

단어 간 의미적 연관성을 고려한 개선된 문서요약 방법 연구

지도교수 김 판 구

이 논문을 공학석사학위신청 논문으로 제출함.

2016년 10월

조선대학교 산업기술융합대학원

소프트웨어융합공학과

차 준 석

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

차준석의 석사학위 논문을 인준함

위원장 조선대학교 교수 정 일 용 (인)

위 원 조선대학교 교수 신 주 현 (인)

위 원 조선대학교 교수 김 판 구 (인)

2016년 11월

조선대학교 산업기술융합대학원

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

목 차

ABSTRACT

Ⅰ. 서 론 ····· 1
A. 연구 배경 및 목적 ····· 1
B. 연구 내용 및 구성 ····· 3

Ⅱ. 관련 연구 ····· 4
A. 생성 요약 기법 ····· 4
B. 추출 요약 기법 ····· 7

Ⅲ. 단어 간 연관성을 고려한 문서요약 방법 ····· 11
A. 시스템 구성도 ····· 11
B. 대표 키워드 추출 ····· 12
1. 전처리 과정 ····· 12
a. 토큰화(Tokenizing) ····· 12
b. 불용어 제거(Stopword) ····· 13
c. 어간 추출(Stemming) ····· 14
d. 품사 판별(part-of-speech tagging) ····· 14
e. 키워드 추출(Keyword Extraction) ····· 16
2. 간선 그래프 모델링 ····· 16
3. 텍스트 랭크를 이용한 대표 키워드 추출 ····· 18
C. 단어 간 연관성을 고려한 중요 문장 추출 ····· 22
1. 단어 그룹화 ····· 22
2. 문서 요약 ····· 25

Ⅳ. 실험 및 결과 ····· 30
A. 데이터 세트(Data Set) ····· 30

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

1. TAC(Text Analysis Conference) ····· 30
B. 실험 평가 방법 및 결과 분석 ····· 31
1. 실험 평가 방법 ····· 31
2. 결과 분석 ····· 32
a. 강한 문장 가중치 점수를 이용한 결과 분석 ····· 32
b. 비교 평가 실험 ····· 35

Ⅴ. 결론 및 제언 ····· 39

참고문헌 ····· 40

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

표 목차

[표 3-1] 전처리 과정 - 토큰화 ····· 2
[표 3-2] NLTK 라이브러리의 불용어 리스트 ····· 3
[표 3-3] 전치리 과정 - 불용어 제거 ····· 3
[표 3-4] 전처리 과정 - 어간 추출 ····· 14
[표 3-5] 영어 단어 품사 목록 ····· 15
[표 3-6] 전처리 과정 - 품사 판별 ····· 15
[표 3-7] 전처리 과정 - 키워드 추출 ····· 16
[표 3-8] 간선 그래프 구축 단계 ····· 17
[표 3-9] 상위 키워드 리스트 ····· 21
[표 3-10] 단어 그룹화 생성 단계 ····· 22
[표 3-11] 단어 그룹화 소스코드 ····· 23
[표 3-12] 중요 문장 가중치 수식 소스코드 ····· 26
[표 3-13] 강한 문장 선정 과정 ····· 27
[표 4-1] TAC 평가 자료 주제 ····· 30
[표 4-2] TAC 평가 자료 질의 ····· 30
[표 4-3] 강한 문장 가중치 점수 평가 결과 ····· 35
[표 4-4] 텍스트랭크 알고리즘 검증 결과 ····· 37
[표 4-5] 비교평가 실험 결과 ····· 38

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

그림 목차

[그림 2-1] 워드넷 계층구조 ····· 5
[그림 2-2] 질의 분해와 확장 ····· 8
[그림 3-1] 전체 시스템 구성도 ····· 11
[그림 3-2] 간선 그래프 ····· 18
[그림 3-3] 상위 키워드 추출 결과 ····· 21
[그림 3-4] 단어 그룹화 예시도 ····· 24
[그림 3-5] 단어 그룹화 생성 ····· 24
[그림 3-6] 문서 요약 결과 ····· 28
[그림 3-7] 전문가 문서 요약 결과 ····· 28
[그림 4-1] ROUGE-1 평가 결과 ····· 33
[그림 4-2] ROUGE-2 평가 결과 ····· 33
[그림 4-3] ROUGE-3 평가 결과 ····· 34
[그림 4-4] ROUGE-4 평가 결과 ····· 34
[그림 4-5] 강한 문장 가중치 점수를 이용한 평가 결과 ····· 35
[그림 4-6] 텍스트랭크 알고리즘 검증 결과 ····· 36
[그림 4-7] 비교 평가 실험결과 ····· 37
[그림 4-8] 실험 및 평가 실험결과 ····· 38

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

ABSTRACT

An Improved Automatic Text Summarization Using the Semantical word relatedness

JunSeok Cha
Advisor : Prof. Pankoo Kim, Ph.D.
Department of Software Convergence Engineering
Graduate School of Industry Technology Convergence, Chosun University

Along with the recent development and distribution of smart devices, data contained in documents appearing on the internet are increasing exponentially. As documents are increasing exponentially, only the titles and main points are shown to users as a solution to figure out information in a document they want. Only with brief contents, however, users may find it difficult to get the information of a document they want. If document wanted by a user is expressed in an accurate form, it will be helpful when trying to find out necessary information.

Document summarization is referred to eliminating redundancies and producing condensed information while maintaining consistency in collected documents. Automatic document summarization technology is to process a large amount of documents automatically and efficiently by extracting main sentences from a document using a computer and eliminating overlapped contents.

To summarize a document efficiently, the present study uses a text rank al gorithm. The text rank algorithm expresses sentences or keywords as a graph and employs the peaks or main lines of a graph to figure out semantic correl ation between words and sentences and understand the importance of sente

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

nces. Through this, it goes through a process to extract the superordinate keywords of sentences and extract main sentences based on those keywords. To go through a process to extract main sentence s, word grouping is done. For word grouping, a particular weighting scale is used to screen sentences having a high weighted value. Based on the screen sentences, main sentences are extracted, and the document is summarized. This study has proven that it shows higher performance than any of the document summarization methods previously examined and performs summarization more efficiently.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅰ. 서론

A. 연구 배경 및 목적

최근 인터넷의 발달과 보급으로 인하여 인터넷 웹상에서 등장하는 문서의 데이터는 대량으로 증가하고 있다. 대량으로 증가하는 문서로 인하여 사용자가 원하는 정보를 파악하기 위해 문서의 제목과 내용을 대략적으로 보여 주어 해결하였다. 하지만 이러한 해결점으로는 사용자가 원하는 문서의 적합성을 판단하기에는 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 원하는 문서의 정보를 얻기 위해 많은 웹 서비스에서 사용자가 원하는 문서의 검색 서비스를 통해 쉽게 찾을 수 있도록 되었지만, 일일이 문서를 읽고 스스로 정보를 확인하는 작업이 필요하다. 문서의 내용을 정확하게 표현하는 문서 요약의 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 빠른 시간 안에 이해할 수 있어야 한다. 문서 요약은 주로 뉴스, 신문기사, 웹페이지 등 여러 웹사이트에서 많은 연구들이 진행되고 있다[1,2,3]. 이러한 연구들은 해당 웹사이트의 특징을 반영한 요약 방법으로 문서 요약에는 적합하지 않다. 문서를 요약하기 위해서는 문서의 특징을 반영한 요약 방법이 필요 하다. 문서 요약의 기본 목적은 원문을 읽지 않고서도 원문의 주제를 파악할 수 있도록 문서의 핵심 주제를 간략하게 압축, 정리하는 것에 있다. 문서 요약은 원문의 주제를 서술적으로 표현해야 하므로 복잡한 언어처리와 주제분석 같은 고차원적인 문서분석 기술을 필요로 한다.

최근 문서 요약 연구에는 문서에 등장하는 단어와 문장 간의 일관성을 유지하면서 문장의 중복을 제거하고 응축된 정보를 생산함으로서 문서의 복잡도를 줄이는 연구가 있으며[4], 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 기법과는 달리 추출 요약은 원문서로부터 중요하다고 생각되는 문장

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

만을 선별하여 요약문으로 제공한다. 문서 요약 과정은 문서에 존재하는 주요 키워드를 파악하는 단계와 파악된 주요 키워드를 기반으로 문서의 주요 내용을 요약으로 문장을 추출 하는 단계를 포함하고 있다. 따라서 문서의 내용을 추론 가능한 의미적인 표현으로 변환하고, 이 표현으로부터 문장을 생성하는 것은 현재 자연어처리 기술로는 매우 어려운 작업이다.

추출 요약 기법은 생성 요약 기법의 주요 내용의 문장을 새롭게 생성하는 대신에 문서의 내용을 대표하는 핵심 문장만을 추출하여 요약으로 제시하는 방법이 진행되고 있다[5]. 사용자가 여러 문서에서 정보를 찾는 것은 정보 요구를 해결할 수 있는 답을 찾기 위한 것이다. 이러한 답을 찾기 위해 단어의 출현 빈도나 문장의 문맥 구조 등의 특성을 사용하여 문장 단락의 중요도를 계산하며, 그 가중치 값이 높은 문장이나 단락을 요약문으로 제시한다. 문장의 중요도를 평가하는 자질로는 단어의 출현 빈도, 문장의 단어나 문장에 존재하는 구 또는 절을 이용한다. 또한, 중요 문장을 추출하기 위해서는 단어의 연관성을 파악해야 한다. 단어의 연관성을 파악하기 위해 문서의 정보가 아니라 시소러스나 워드넷 등의 다른 정보 도구를 이용할 수 있다. 하지만 시소러스나 워드넷은 어떤 특정한 문서에서만 출현하는 단어에 대한 정보가 없거나 그 내용이 전혀 다른 내용일 수 있기 때문에 특정 문서에는 맞지 않은 정보가 표현될 수도 있으며, 워드넷을 이용한 문서요약 기법은 요약 과정 중 많은 계산 비용이 발생하여 성능이 떨어지는 단점을 가진다. 그러므로 실시간으로 정보가 생성되고 급속하게 퍼지는 인터넷 공간에서 문서 요약 기술은 반드시 필요 하다. 소셜 네트워크, 스포츠 뉴스 기사 같은 웹 서비스공간에서는 장문의 글보다는 단문 형태의 메시지를 통해 커뮤니케이션이 이루어지기 때문에 문서 요약을 통한 내용 전달이 무엇보다 중요하게 된다. 이러한 기존의 방법들은 같은 수의 공통 단어를 포함한 문장들이라도 개개의 문장이 포함한 단어 수나 문장의 출현 위치에 따라 유사도가 다르게 계산되며, 전혀 의미 없는 키워드에 의해 중요한 문장이 삭제되어 문서가 요약되는 단점이 있다.

본 논문에서는 문서에 존재하는 키워드를 상위 키워드와 하위 키워드로 추출하여 추출된 키워드를 기반으로 문서를 요약하게 된다. 키워드 추출을 위해 전처리 과정과 본 연구에서 제안하는 텍스트랭크 알고리즘을 이용한다. 텍스트 랭크 알고리즘은 그래프 기반의 랭킹 알고리즘으로 해당 문장에 존재하는 그래프 점수를 계산한 후 높은 점수의 키워드를 상위 키워드로 선정되며, 선정된 상위 키워드와 하위 키워드를 기반으로 중요 문장 추출 과정을 거친 후 문서를 요약하게 된다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

B. 연구 내용 및 구성

본 연구에서는 문서에 존재하는 핵심적인 내용을 추출하기 위해서 추출요약 기법 중 텍스트 랭크 알고리즘과 단어 간의 연관성을 고려한 문서요약 방법을 제안하고자 한다. 제안한 방법은 기존에 지속적으로 연구 되어 오던 문서 요약 연구 주제인 중요 문장을 추출함에 있어 단순히 각 문장에서 중요 단어의 출현 여부만을 이용한 방법을 보완하기 위해 텍스트 랭크 알고리즘을 개선하여 본 연구에 도입을 하였고, 또한 텍스트 랭크 알고리즘을 기반으로 단어 간의 의미적 연관성을 고려하여 중요 문장을 추출하는 방법을 제안한다.

본 논문의 구성은 다음과 같이 구성되어 있다.

본 장인 서론에 이어 2장 관련 연구에서는 본 연구의 배경이 되는 생성 요약 방법과 추출 요약 방법을 통해 효율적으로 해결하는 기존 연구들에 대해 살펴본다.

3장에서는 본 논문에서 제안하는 텍스트랭크 알고리즘 기반의 단어 간의 연관성을 고려한 문서요약 방법을 기술한다.

4장에서는 제안한 방법의 실험을 위한 데이터 수집방법과 다양한 실험을 통한 비교 측정 및 실험 성능을 평가한다.

마지막으로 5장에서는 결론과 향후 연구에 관해 서술하며, 마무리 한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

II. 관련 연구

최근 문서 요약에 대해 다양한 연구가 진행되고 있다. 문서 요약 기법은 대표적으로 생성요약(Abstract Summarization)기법과 추출요약(Extracrt Summarization) 기법으로 나눌 수 있다. 그중 추출 요약 기법은 다양하게 연구되고 있는 기법이며, 문서 내 존재하는 문장과 키워드를 추론하여 가중치를 통해 점수를 부여한다. 부여된 가중치 점수를 통해 문서에 존재하는 핵심 문장을 추출하는 문서 요약 기법이다. 본 논문에서는 추출 요약 기법을 이용하여 문서에 존재하는 대표적인 키워드를 추출하고, 추출된 키워드를 기반으로 단어 간 연관성을 고려하여 문서를 요약하고자 한다.

A. 생성 요약 기법

생성 요약 기법은 문서 내에 존재하는 중요 키워드와 문장을 파악하여 자연어 처리 기법을 이용한 문장조합기법, 재구축기법들을 통하여 문서요약을 하는 연구이다. 또한 생성요약 기법은 문서에 존재하는 단어의 위치, 문장의 위치 같이 대상 문서에서 형태적으로 나타나는 정보를 이용하여 문서요약을 시도한다. 생성 요약 기법은 요약 속도가 빠르고 시스템 구현이 쉬우나 다의어, 유사어와 같은 의미 구분을 하지 못함으로서, 문서를 지나치게 단순한 통계 테이블로 간 경우가 빈번 발생하였기 때문에, 생성 요약에 관련된 초기 연구로는 문서의 주제를 표현하는 단어가 자주 사용된다는 직관에 의거하여, 가장 많이 사용되는 단어를 문서의 주제어로 결정한 요약이 있다[6].

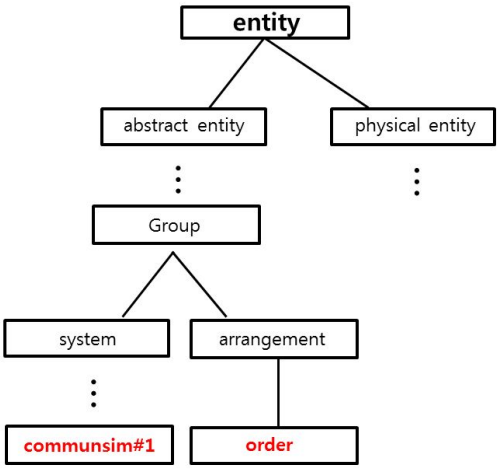
생성 요약 기법은 주요 단어가 포함된 문장을 생성 및 추출을 함으로써 생성된다. 기존에 꾸준히 연구되어 오던 생성 요약 기법을 새롭게 응용하여 문장의 위치에 따른 문서 요약 연구가 있다. 첫 번째 사용된 문장 또는 마지막에 사용된 문장이 같은 위치에 따라 문장의 중요도가 다르다는 연구가 있었으며[7], 아울러 ‘significant’, ‘hardly’, ‘impossible’과 같은 단서 어도 중요 문장을 파악하기 위해 텍스트 랭크 알고리즘을 이용하여 해당 문장이 중요한 역할을 할 수 있음을 보였다. 이 외에 기계 학습을 통해 문서의 중요도를 습득함으로써 문서 요약의 성능을 향상 시킨 연구도 있었다[8]. 이 방법들은 모두 자동 문서 요약 연구의 기초가 되는 역할을 하였으나, 지나치게 단순한 통계에 의존함으로써 문서의 주제를 파악 하

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

는데 한계를 보였다. 최근 새롭게 연구 되고 있는 생성 요약 연구에는 워드넷을 이용한 문서요약 기술이 활발히 진행 되고 있다. 다양한 문서를 분석하기 위해서 문서마다 갖는 특징을 고려하며, 워드넷을 구성하고 이를 기반으로 한 문장의 분석이 이루어짐으로써 보다 정확한 문서요약이 이루어진다.

워드넷은 단어 간의 유사도를 평가하기 위해 단어의 동의어 집합을 최상위 단어에서 단계가 내려 갈수록 구체적인 트리 형태의 계층구조로 설계하여 단어 간 짧은 거리 정도를 계산하는 방법도 사용되고 있다. 아래 [그림 2-1]은 워드넷의 단어 간 관계를 확장된 형태의 트리로 표현한 방법이다.



[그림 2-1] 워드넷 계층구조

기본적으로 워드넷의 계층구조는 해당 단어를 중심으로 상위어와 하위어로 구성되며, 상위어는 해당 단어를 포함하는 포괄적인 단어이고, 하위어는 단어가 포함하고 있는 구체적인 단어를 의미한다. 따라서 상위어와 하위어를 이용하여 단어에 대한 의미적 연관성을 알 수 있다[9].

담화 이론에 따르면 문서는 크게 중심 부분과 주변 부분으로 구성된다는 가정하에 두 부분 사이의 수사 관계를 이용하여 요약을 생성하는 문서 요약 연구 방법이다[10]. 이 방법에서는 문서의 문장이 주어지면 일단 수사 구조 알고리즘에 의해 해당 담화 트리를 생성한다. 담화 트리의 단말 노드는 구, 절, 문장 등이 되며, 내부 노드는 해당 자식 노드 사이의 수사 관계를 표현하게 된다. 요약을 생성하기 위해서는 담화 트리의 각 노드에 대해 부모 노드가 자식 노드보다 높은 값을 갖도

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

록 중요도 값을 부여하여 그 값이 높은 순으로 정렬한다. 정렬 결과 순위가 높은 구, 절, 문장들을 요약으로 제시한다.

문서 요약을 위해선 성능 높은 담화 분석기가 필요하며, 주어진 문서를 요약하기 위해 모든 문장의 수사 관계를 분석해야 하므로 시간이 많이 걸리고, 수사관계가 별로 없는 문서의 경우에는 성능에 한계를 보인다. 이 방법론이 전체 문서에 대한 수사 관계를 트리로 구성하는 데 반해 실제 요약은 전체 문서의 모든 수사 관계를 필요로 하지 않는다고 알려져 있다[11]. 문서 요약의 생성 요약 기법은 문장의 수사관계의 모호성 등 해결해야 할 문제들이 많이 남아 있다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

B. 추출 요약 기법

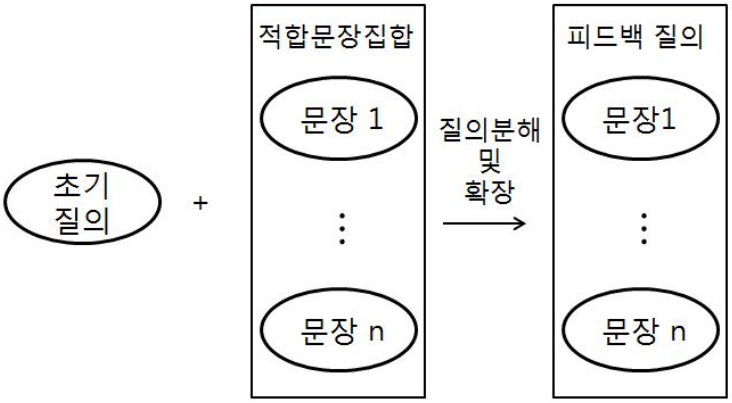
추출 요약 기법은 문서 내 문장들을 가지고 구, 절, 문장 등을 새롭게 분석하여 문서 내에 중요도를 판단 한 후 문서를 요약을 하는 기법이다. 추출 요약 기법은 현재 다양한 연구를 통해 문장이 가지는 단어의 빈도수 및 가중치를 통해 문장과 단어 간의 관계를 분석하여 중요 문장을 추출하는 방식으로 이루어지고 있다. 하지만 자동으로 분석된 문장의 가중치가 기존 문서의 의미전달이 제대로 이루어지지 않는 경우 올바른 문장 요약이 이루어지지 않는다. 상대적으로 기존 생성 요약 기법보다는 구현이 쉽다는 장점 때문에 현대 추출 요약 기법을 이용한 문서 요약은 활발히 진행되고 있다.

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[12].

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 분석시간이 요구된다는 문제점이 있다[13]. 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. 질의 분해를 이용한 방식은 [그림 2-2]과같이 n개의 적합 문장들에 대해 각각 개별적으로 질의 확장을 수행하여, 문서의 중요 문장을 추출하는 방식이다[14].

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 2-2] 질의 분해와 확장

이전 까지 수행해 왔던 연구들의 문제점을 보완하기 위해 다양한 연구가 진행되었다. 첫 번째로, 각 문장의 품사를 분석하여 문서를 요약하는 연구가 있다. 해당 문서에 존재하는 품사 중 명사, 형용사에 대하여 워드넷의 동의어, 상위어, 하위어 관계를 기반으로 의미 분석하여 랭킹을 정하고 가중치를 부여한 후, 문서 내에 핵심 단어가 포함되는 문장들을 추려내서 가중치가 부여된 단어들과 코사인유사도를 수행하여 문장에 점수를 주는 기법을 제안하였다[15].

두 번째로, 문서요약을 효율적으로 하기 위하여 주어진 문서의 평균 길이를 고려하여 핵심어 유사도를 나타내는 연결 그래프를 생성하여 분석 한 후 요약한 기법이 있다. 하지만 제안한 방법은 단지 다양한 문장 형태에 따른 문장 간의 연결 구조 파악만을 위한 방법으로 각 문장에 대한 의미상의 중요도 계산에 대한 고려하지 않는 문제점이 있다[16].

최근 추출 요약 기반의 문서요약에 관한 연구 들은 언어학적 접근 방법을 통해 문서 요약을 진행하고 있다. 언어학적인 접근 방법은 어휘체인(Lexical Chain)이나 워드넷 계층 구조 등을 활용하여 문서의 문장구조를 파악한 다음 중요 문장을 추출하여 제시하는 방법이다[17]. 그리고 중요 문장 추출을 위해 여러 가지 기법을 이용하여 최적화된 성능의 문서요약 시스템을 개발하고자 하였다. 대표적인 방법으로 텍스트랭크 알고리즘을 기반으로 추출 요약 기법을 이용한 연구가 있다. 텍스트랭크 알고리즘은 문서를 그래프로 표현하고, 그래프의 형태를 이용해 각 노드의

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

중요도를 결정하는 알고리즘이 있으며[18]. 이러한 텍스트 랭크 알고리즘을 기반으로 HITS(Hypertext Induced Random Walk)알고리즘과 같이 활용한 연구가 있다.

HITS란 본래 특정 웹페이지의 중요도를 산출할 때 쓰이는 링크 분석 알고리즘으로, 해당 페이지가 얼마나 가치 있는 내용을 포함하고 있는지, 중요도를 계산 하는 알고리즘이다. HITS를 통해 각 문서의 중요도를 계산 할 때 hub score로 텍스트 랭크를 사용 하였고, 중요한 문장들을 선정하여 요약문에 포함했다. 그리고, 요약문의 자연스러운 흐름을 위해 최단 거리 알고리즘을 사용하여 최단 거리 상에 있는 장들을 순서대로 연결하도록 하였다. 실험 결과 제안한 시스템이 자연스러운 요약문을 생성하는 데 강정을 보임을 확인한 연구였다[19].

추출 요약 연구의 최근 동향으로는 문서를 요약하기 위해 그래프로 표현하며, 단어의 위치 정보와 더불어 소셜 폭소노미를 이용한 단어 간의 의미적 연관성을 고려하는 추출 요약 기법을 제안한 연구가 있다. 이를 위해 텍스트랭크 알고리즘을 사용하였다. 이는 텍스트 랭크 알고리즘의 한계점을 보완하기 위한 것으로, 기존의 알고리즘이 오직 문서상의 위치만을 고려하여 특정 노드의 중요도를 계산하기 때문에 노드 간의 의미적 유사성을 반영하지 못한다는 점에 착안하여 기법을 개선하였다. 제안한 기법의 성능을 평가하기 위하여 TAC 2008과 TAC 2009 데이터 세트를 이용하였으며, 기준 값으로 ROUGE 평가 시스템을 이용하였다. ROUGE-2, ROUGE-4 값의 정확률, 재현율, F-Score값을 각각 계산하여 베이스라인인 텍스트랭크 알고리즘을 사용한 경우가 기존베이스 라인보다 전체적으로 더 좋은 성능을 보였다[20].

마지막으로 기존에 연구 되었던 추출 요약 기법중 텍스트 랭크 알고리즘과 유사한 LexPageRank를 사용한 추출 요약 기법을 제안한 연구가 있다[21]. LexPageRank 알고리즘은 그래프로 표현된 문서에서 각 문장의 중요도를 계산하지만, 이때 중요도 기준으로 중심 값이 아닌 고유 벡터 값을 사용하여 해당 문장이 다른 문장들에게 얼마나 영향을 미치는지 반영하고자 하였다. 또한 키워드를 문장의 유사도 행렬과 코사인 유사도를 바탕으로 하여 추출하였다. 제안한 기법이 얼마나 유용한지를 평가하고자 DUC 2004 데이터 세트를 활용 하였고, 문서요약 평가 시스템인 ROUGE를 이용해 성능을 평가 하였다. 결과적으로 중심 값을 사용할 때보다 고유 벡터 값을 사용할 때 문서 요약 시스템의 성능이 우수 한 성능을 보인 연구였다.

LexPageRank 알고리즘을 이용한 또 하나의 연구는 문서의 문장에 주제를 이용

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

하여 문장들의 중요한 키워드를 추출하고 해당 문장의 정렬방법에 의한 다중 문서 요약 방법을 제안 하였다[22].

이 방법은 전처리 과정을 통해 문서의 주제를 추출하고, 문서의 문장을 요약하는 방법이며, 문장들과 단어들을 분석하여 문서를 요약하는 연구이다. 해당 문서에서 핵심이 되고 의미 있는 문장에 점수를 부여해야 하며, 높은 점수를 획득하는 문장을 기반으로 주요 문장을 추출하는 기법이다[23].

Collection @ chosun

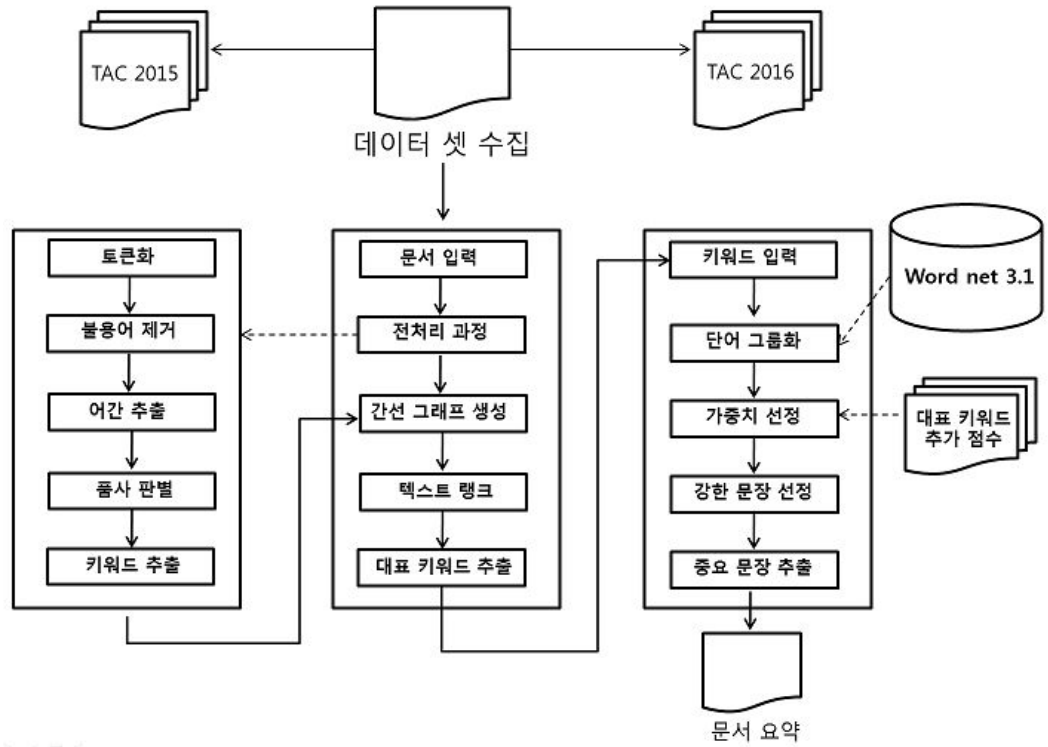
조선대학교
CHOSUN UNIVERSITY

III. 단어 간 연관성을 고려한 문서요약 방법

본 논문에서는 문서에 존재하는 핵심 키워드인 대표 키워드를 추출하여, 대표 키워드와 하위 키워드로 나누어 실험 데이터 세트를 구축한다. 구축된 실험 데이터 세트를 이용하여 중요 문장 추출 과정을 거치게 되며, 중요 문장을 추출하기 위해 단어와 문장 간의 연관성을 고려하여 각 문장 별 가중치 점수 값을 구한 문서 요약 방법을 제안한다.

A. 시스템 구성도

[그림 3-1]은 본 논문에서 문서 요약 시스템의 전체 구성도이다. 시스템 구성도는 크게 대표 키워드를 추출하는 단어 분석 단계와 각 단어의 연관성을 고려한 중요 문장 추출 단계로 구성된다.



[그림 3-1] 문서요약 시스템 구성도

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

B. 대표 키워드 추출

문서 요약을 위해 본 논문에서는 단어 분석 단계를 거친다. 단어 분석 단계에서는 텍스트랭크 알고리즘을 이용해 대표 키워드를 추출하게 된다. 대표 키워드 추출을 위해 본 절에서는 전처리 과정과 간선 그래프 모델링, 텍스트랭크 적용 단계를 거친다.

1. 전처리 과정

대표 키워드를 추출하기에 앞서 문서에 존재하는 키워드를 먼저 추출해야 한다. 키워드를 추출하기 위해 수집한 문서에서 불필요한 정보를 제거하고 필요한 정보를 추출 할 수 있는 전처리 과정이 필요 하다. 본 논문에서 제시하는 전처리 과정은 토큰화, 불용어 제거, 어간 추출, 품사 판별을 거쳐 키워드 추출하게 된다.

a. 토큰화(Tokenizing)

토큰화는 문장 내에서 공백 또는 특정 단어를 기준으로 나누는 것을 의미한다. CNN, ABC뉴스와 TAC에서 제공하는 데이터 세트를 수집하여 문서 요약 시스템 데이터베이스에 저장된 문서의 문장을 스페이스 바를 기준으로 하여 문장을 어절 단위로 세분화한다. 아래의 [표 3-1]은 원문에서 토큰화한 결과이다.

[표 3-1] 전처리 과정 - 토큰화

전처리 과정	뉴스 원문
원문	For the soul that yearns to be close to the cultural heart of Houston, the majestic oak-lined streets of the Museum District will quickly feel like home. Just north of the Texas Medical Center and minutes away from downtown. ……
토큰화	For the soul that yearns to be close to the cultural heart of Houston, the majestic oak-lined streets of the Museum District will quickly feel like home. Just north Texas Medical Center and minutes away from downtown, area claims an ideal. ……

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

b. 불용어 제거(Stopword)

불용어는 어절 단위로 세분된 단어 중 불용어 리스트를 이용하여 문장의 불용어를 제거 시킨다. 불용어는 인터넷이나 문서의 정보를 이용하여 실험을 할 시 단어로써 사용하지 않는 단어, 관사, 전치사, 조사, 접속사 등 검색 색인 단어로서 의미가 없는 단어이다. 다만 검색 엔진이나 수집한 문서의 문장에 해당 단어가 동일하지 않기 때문에 다를 수도 있다[24]. [표 3-2]는 Python NLTK(Natural Language Toolkit) 라이브러리에 포함된 불용어 리스트이다.

[표 3-2] 불용어 리스트

번호	불용어	번호	불용어	번호	불용어
1	I	13	his	25	who
2	me	14	himself	26	whom
3	my	15	she	27	this
4	myself	16	her	28	that
5	we	17	herself	29	these
6	our	18	it	30	am
7	ourselves	19	itself	31	is
8	you	20	they	32	are
9	your	21	their	33	was
10	youtself	22	themselves	34	were
11	he	23	what	35	be
12	him	24	which	……	……

[표 3-3] 전처리 과정 - 불용어 제거

전처리 과정	뉴스 원문
토큰화	For the soul that yearns to be close to the cultural heart of Houston, the majestic oak-lined streets of the Museum District will quickly feel like home. Just north Texas Medical Center and minutes away from downtown, area claims an ideal. ……
불용어 제거	soul yearns close cultural heart Houston, majestic lined streets Museum District quickly feel like home. Just north Texas Medical Center minutes away downtown, area claims ideal central location Inner Loop. ……

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

문서의 토큰화를 하면서 불용어를 제거하였고, 불용어는 특수문자를 제외한 특수문자들과 [표 3-2]에서 제시된 불용어 리스트를 이용하여 불용어를 제거한다. [표 3-3]은 불용어 리스트를 이용해 불용어를 제거한 결과를 보여주고 있다.

c. 어간 추출(Stemming)

어간 추출은 어형이 변형된 단어로부터 접사 등을 제거하고 그 단어의 어간을 분리해 내는 것을 의미한다. 여기서 어간은 반드시 어근과 같아야 할 필요는 없으며, 어근과 차이가 있더라도 관련이 있는 단어들이 일정하게 동일한 어간으로 맵핑되게 하는 것이 어간 추출을 목적이다. 예를 들어 'lined'이라는 단어가 있을 경우, 이 단어의 원형은 'line'을 나타낸다. [표 3-4]에서는 원본 데이터가 어간 추출 단계를 거치게 되면, 복수형, 과거형의 단어들이 단어의 원형으로 변경된 것을 볼 수 있다.

[표 3-4] 전처리 과정 - 어간 추출

전처리 과정	뉴스 원문
불용어 제거	soul yearns close cultural heart Houston, majestic lined streets Museum District quickly feel like home. Just north Texas Medical Center minutes away from downtown, area claims ideal central location Inner Loop. ……
어간 추출	soul **year** close cultural heart Houston, majestic **line** streets Museum District quickly feel like home. Just north Texas Medical Center **minute** away from downtown, area claims ideal central location Inner Loop. ……

d. 품사 판별(part-of-speech tagging)

키워드의 특징을 추출하기 전 각 단어의 품사를 구분하기 위해 품사 판별 과정을 거친다. 품사 판별은 형태소나 어절의 품사 중의성을 없애는 작업을 말한다. 형태소는 하나의 문법 단위로서 의미를 가지는 최소의 단위에 해당한다. 품사 판별을 위해 Python NLTK 라이브러리를 이용하여 명사, 동사, 형용사로 품사를 판별하였다. [표 3-5]는 품사를 파악하기 위해 사용한 품사 목록을 나타낸다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-5] 영어 단어 품사 목록

기호	품사	기호	품사
CC	등위 접속사	PRP$	소유대명사
CD	기수	RB	부사
DT	한정사	RBR	비교 부사
EX	유도부사	RBS	최상급 부사
FW	외래어	RP	불변화사
IN	전치사	SYM	기호
JJ	형용사	TO	to 부정사
JJR	비교 형용사	UH	감탄사
JJS	최상급 형용사	VB	동사(현재형)
LS	목록 항목	VBD	동사(과거형)
MD	법조동사	VBG	동사(동명사)
NN	단수명사	VBN	동사(현재 완료)
NNS	복수명사	VBP	동사(3인칭 단수 현재)
NNP	고유명사	VBZ	동사(3인칭 단수)
NNPS	고유명사 (복수)	WDT	한정 대명사
PDT	전치 한정사	WP	대명사
POS	소유격	WP$	소유대명사
PRP	인칭대명사	WRB	부사

[표 3-6]은 [표 3-5]의 품사 목록을 이용하여 어간 추출을 통해 나타난 데이터를 품사 python NLTK를 이용해 품사를 판별한 결과 이다.

[표 3-6] 전처리 과정 - 품사 판별

전처리 과정	뉴스 원문
어간 추출	soul year close cultural heart Houston, majestic line streets Museum District quickly feel like home. Just north Texas Medical Center minute away from downtow n, area claims ideal central location Inner Loop.……
품사 판별	soul/NNP, year/NNS, close/NNP, cultural/JJ, heart/NN, Housto n/NNP, majestic/JJ, line/NN, streets/NNS, Museum/NNP, Distr ict/NNP, feel/VBP, home/NN. Just/RB, north/RB, Texas/NNP, Medical/NNP, Center/NNP, mi nute/NN, downtown/NN. ……

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

명사와 동사, 형용사만을 고려하여 품사 판별을 한 이유는 어간 추출 과정을 통해 키워드의 어근만을 고려하여 키워드의 형태가 혼전하지 않았기 때문에 명사, 동사, 형용사를 고려하였다.

e. 키워드 추출(Keyword Extraction)

마지막 키워드 추출은 품사가 정의된 단어들을 이용하여 문장에 존재하는 전치사, 조사, 접속사 등 필요 없는 단어를 제거하며, 형용사, 동사, 명사를 키워드로 추출한다. [표 3-7]은 최종적으로 전처리 과정을 수행한 결과이다.

[표 3-7] 전처리 과정 - 키워드 추출

전처리 과정	뉴스 원문
품사 판별	soul/NNP, year/NNS, close/NNP, cultural/JJ, heart/NN, Houston/NNP, majestic/JJ, line/NN, streets/NNS, Museum/NNP, District/NNP, quickly/RB, feel/VBP, like/IN, home/NN. Just/RB, north/RB, Texas/NNP, Medical/NNP, Center/NNP, minute/NN, away/RB, from/IN, downtown/NN. ……
키워드 추출	soul, year, close, cultural, heart, Houston, majestic, line, streets, Museum, District feel, home. north, Texas, Medical, Center, minute, downtown. ……

2. 간선 그래프 모델링

전처리 과정을 통해 추출된 키워드들을 이용하여 텍스트랭크 알고리즘에 적용하기 위해서는 처리하고자 하는 문서의 문장을 그래프 형태로 모델링 해야 한다. 우선 그래프를 모델링 하기 위해 그래프를 구성할 정점을 결정해야 한다. 정점은 전처리과정을 통해 모든 불용어를 제거한 문서의 단어들로 표현하며, 간선은 단어로 표현되는 정점들 간의 의미적 관계를 뜻한다.

간선은 단어와 문장 간의 관계를 표현해주는 중요한 역할을 하게 되며, 간선 그래프의 정점과 간선 사이의 특징을 텍스트랭크 알고리즘의 응용 목적에 따라 새롭게 바뀔 수 있다. 간선을 그리기 위해 co-occurrence 관계를 이용해서 정점과 간선을 연결한다. 연결된 간선을 기반으로 간선 그래프를 모델링 해야 한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

간선 그래프를 모델링하기 위해서는 4단계를 거쳐 그래프를 모델링 하게 된다. 간선 그래프의 모델링 순서는 [표3-8]과 같다.

[표 3-8] 간선 그래프 모델링 단계

단계 1. 간선 그래프를 구성할 때, 각 단어의 정점과 단위를 결정한다. 단계 2. 정점과 간성의 사이를 연결 시킨다. 정의된 관계에 의해서 정점 사이에 간선을 생성하여 그래프를 모델링 하게 된다. 단계 3. 각 그래프의 랭킹 알고리즘을 수렴할 때까지 반복 적용한다. 단계 4. 마지막으로 정점을 텍스트 랭크 알고리즘 기반의 결과로 부여된 스코어를 적절히 이용하여, 해당 문서 키워드의 대표 키워드를 선정하게 된다.

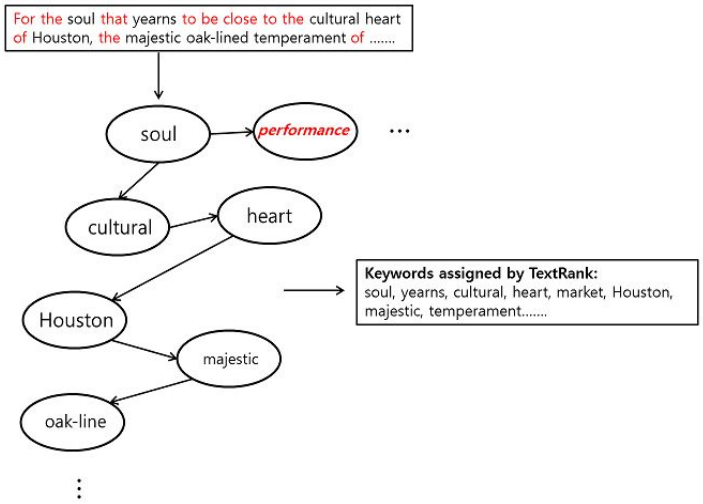
[표 3-8]의 방법을 이용해 키워드를 이용하여 정점을 구성할 때, 구성할 단위와 정점 사이의 관계는 텍스트랭크 알고리즘의 목적에 따라 결과가 바뀔 수 있다. 간선 그래프의 정점은 N-Gram의 uni-gram또는 bi-gram으로 만들 수 있다.

Uni-gram의 경우 하나의 키워드가 하나의 정점을 구성하며, bi-gram으로 정점을 구성할 경우 연속적으로 등장하는 두 개의 키워드가 하나의 정점을 구성하게 된다. 본 실험에서는 uni-gram을 기준으로 정점을 구성하게 된다. 구성된 점점의 관계에 의해서 정점 사이의 간선을 생성하여 그래프를 모델링하게 되며, 모델링을 하기 위해 [표 3-8]의 단계 3과 같이 텍스트랭크 알고리즘의 스코어 값이 수렴할 때까지 반복 적용한다.

간선 그래프에 포함될 정점들이 결정되면 정점들 사이의 관계를 이용해서 간선을 그리는 단계로 넘어간다. 상위 키워드를 추출하는 경우 키워드 사이의 co-occurrence가 간선 형성의 기준이 될 수 있고, 간선의 가중치를 함께 등장하는 횟수로 결정할 수 있다. [표 3-8] 방법을 이용하여 [그림 3-2]와 같이 간선 그래프를 모델링 하였다. 그 후 구축된 그래프를 기준으로 각 정점의 스코어 선정 과정을 거친다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 3-2] 간선 그래프 모델링

[그림 3-2]는 하나의 문장에서 불용어를 제거하고 간선을 이용해 모델링 된 그래프이다. soul이라는 단어가 다음 문장에 존재한다면 soul을 기점으로 두 번째 문장이 간선 그래프로 연결되어 모델링 된다. 하지만 다음 문장에 기존 간선 그래프의 단어가 존재하지 않다면 oak-line라는 단어부터 두 번째 간선 그래프가 생성된다. 간선 그래프의 정점은 형용사 혹은 명사, 동사가 될 것이고, 만약 핵심 문장을 추출하는 경우 정점은 하나의 문장이 된다. [표 3-8]의 과정을 통해 구축된 간선 그래프를 기반으로 문서 요약을 위해 문장에 존재하는 핵심 키워드를 추출하기 위해 대표 키워드 추출 과정을 거치게 된다.

3. 텍스트랭크를 이용한 대표 키워드 추출

대표 키워드 추출은 주어진 텍스트를 간선 그래프로 모델링 한 후, 모델링된 그래프에 텍스트 랭크 알고리즘을 적용하여 각 단어별 스코어를 측정한다. 각 단어별 스코어를 측정 한 후 측정된 스코어를 기반으로 해당 문서의 핵심 키워드를 추출해야한다. 그래프에서 중요도를 결정하는 알고리즘의 종류는 굉장히 다양하다. 하지만 대부분 자연어처리에 적합한 형태가 아니므로 자연어처리에 적용하기 위해서는 다소 변형과 응용을 거쳐야한다[22]. 앞서 언급한 바와 같이 주어진 키워드로부터 생성된 간선 그래프에 텍스트 랭크 알고리즘을 적용하여 각 정점의 스코어를

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

얻어낸다. 이 때, 적용할 텍스트랭크는 세 가지 규칙을 고려하여 스코어 점수를 얻어 키워드의 랭킹을 산정 한다.

첫 번째 규칙은 그래프를 통해 나타난 정점과 간선을 이용한 방법이다. 하나의 문서에서 나타나는 인접한 단어는 서로 의미적 연관성이 존재 한다는 규칙 하에 각 정점과 간선의 가중치를 위해 수식 (1)을 이용 한다.

$$W_{co-occurrence} = \begin{cases} 1, co-occur \\ 0, otherwise \end{cases} \quad (1)$$

수식 (1)은 co-occurrence 관계를 이용한 방법이다. 첫 번째, 규칙은 상위 대표 키워드 추출을 위해 수집한 문서를 기반으로 모델링 할 때, 함께 등장 하는 키워드의 횟수를 가중치로 설정 한다. 예를 들어 “Galaxy”라는 단어와 “Samsung”라는 단어가 해당 문서에 등장하는 횟수가 10이라고 할 때, 가중치 점수는 10점이 된다. 하지만 “Samsung”과 “apple” 이라는 단어가 문서에 등장 횟수가 10이지만, 한 문장에 공통적으로 나타나는 횟수가 2라고 할 때, 가중치 점수는 2점이 된다. 그러므로 “Galaxy”라는 단어와 “Samsung”이라는 단어 사이의 관계, 즉 “Samsung” 이라는 단어와 “Apple” 이라는 단어 사이의 관련성은 똑같이 적용된다.

두 번째, 규칙은 문장 내 키워드 알파벳 유사성을 이용해 키워드와 문장 간의 연관성 가중치를 적용하는 방법이다. 수식 (2)는 두 키워드 사이에 공통적으로 존재하는 부분의 길이에 대한 비율을 의미하는 NLCS(Nomailzed Longest Common Subsequence)[25,26] 기법을 통해 계산하였다.

$$W_{NLCS} = \frac{2 * longest\,common\,substring}{length\,of\,word1 + lengthword2} \quad (2)$$

수식 (2)의 word1과 word2는 연결된 간선 그래프의 단어의 개수를 의미하며 longest common substring은 각 두 단어가 문서에 공통적으로 나타나는 간선의 개수를 의미한다. 수식 (2)는 모든 가중치의 합이 간선 그래프의 간선 개수가 평균 n개 이상이 되도록 재계산 해준다.

마지막 규칙은 태그 클러스터로부터 단어 간 연관성을 추출하여 이를 가중치로 부여 하는 방법이다. 각 단어와 단어 간의 정점을 결정하고 가중치를 구해야 한다. 가중치를 구하기 위한 수식은 아래와 같다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

$$W_{folksonomy} = \begin{cases} 2, [w1 \in TC_{w2} \text{and} W2 \in TC_{w1}] \\ 1, [W1 \in TC_{w2}] \\ 1, [W2 \in TC_{w1}] \\ 0, else \end{cases} \quad (3)$$

수식 (3)의 w1과 w2는 각 정점의 두 단어를 의미하고 TC는 태그 클러스터를 의미한다. w1의 태그 클러스터에 w2가 존재하면서 w2의 태그 클러스터에도 w1이 존재한다면 가중치 점수 2가 되며, 단일 방향으로만 존재한다면 가중치 점수는 1점이 된다. 하지만 어느 방향으로도 존재 하지 않는 다면 가중치 점수 0점이 된다. 이 수식 역시 가중치의 합이 모든 간선의 개수가 문서의 평균 개수를 되도록 계산이 되어야 한다. 이렇게 계산된 세 가지 규칙을 이용한 값은 서로 다른 가중치를 가지고 합해진다. 각 규칙들이 단어와의 관계에 기여하는 정도가 다를 수 있기 때문이다. 앞서 언급한 세가치 규칙의 가중치 합을 이용하여 각 단어의 순위를 계산한다. 아래의 수식 (4)는 대표 키워드를 선정하기 위한 텍스트 랭크 알고리즘의 수식이다.

$$TR(V_i) = (1-d) + d\sum_{j=0}^{N-1} \frac{W_{ij}}{\sum_{k=0}^{N-1} W_{jk}} \times TR(V_j) \quad (4)$$

수식 (4)의 TR(vi)는 단어 i에 대한 텍스트랭크 값을 의미하며 wij는 수식 (1),(2),(3)에서 구한 모든 단어들의 가중치 합이다. 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 키어구로 등장하는 것들에서만 대표 키워드로 지정한다. [표 3-9]는 수집한 데이터 세트에서 추출된 대표키워드의 랭크 값을 나타낸다.

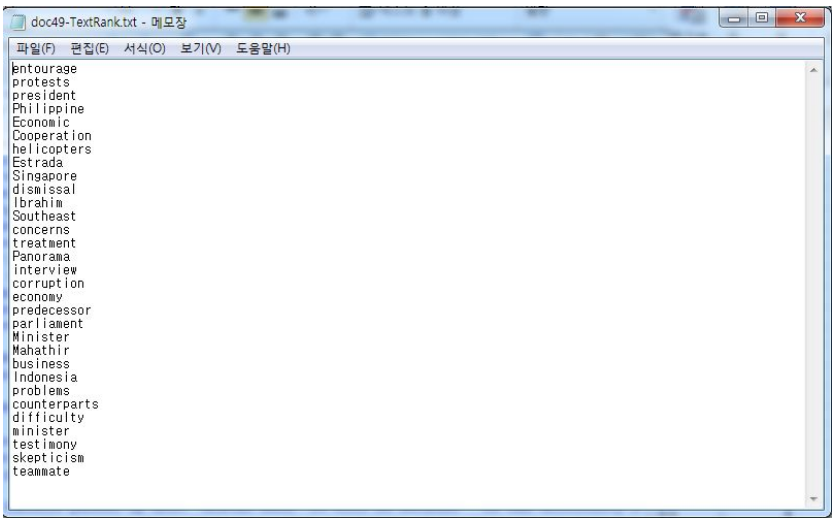
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-9] 대표 키워드 리스트

단어	스코어	단어	스코어
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
concerns	0.10629	minister	0.60512
treatment	0.10521	testimony	0.60481
Panorama	0.10510	skepticism	0.60323
interview	0.10508	teammate	0.60096

대표 키워드는 entourage, Protests, president, Philippine 순으로 [그림 3-3]과 같이 키워드의 데이터가 내림차순으로 저장되며, 이를 기반으로 상위 키워드와 하위 키워드로 나누어 문서를 요약하기 위한 문장 분석 단계를 거친다.



[그림 3-3] 대표 키워드 추출 결과

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

C. 단어 간 연관성을 고려한 중요 문장 추출

앞 절에서는 문서요약을 위해 상위 키워드와 하위 키워드로 추출하는 방법을 제안 하였으며, 제안된 방법을 통해 추출된 키워드를 본 절에서는 단어와 문장 간의 연관성을 파악하여 문서를 요약하는 방법을 제안하고자 한다.

1. 단어 그룹화

중요 문장을 추출하기 위해 본 논문에서는 단어 그룹화 과정을 거쳐야한다. 단어 그룹화는 문법적인 장치를 제외한 어휘적 의미만을 이용하여 수집한 문서를 분석하고 단어 간 연관성이 있는 키워드들끼리 그룹화를 하는 방법이다.

추출된 키워드의 단어 그룹화를 하기 위해 워드넷을 기반으로 한다. 워드넷은 영어의 의미 목록으로 워드넷은 영어 단어를 'synset'이라는 유의어 집단으로 분류하여 간략하고 일반적인 정의를 제공하며, 워드넷은 어휘 목록 사이의 다양한 의미관계를 보여준다. 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. [표 3-10]에서 보는 바와 같이 단어 그룹화는 3단계를 거친다.

[표 3-10] 단어 그룹화 생성 단계

단계1. 전처리 과정을 통해 후보 키워드를 추출한다. 후보 키워드는 텍스트랭크 알고리즘을 통해 추출된 상위 키워드와 하위 키워드를 의미한다. 단계2. 각 후보 키워드들의 단어 그룹화를 하기 위해 수집한 문서에 있는 각 키워드의 상·하의어, 동의어를 탐색한다. 단계3. 탐색 결과 각 키워드가 상·하위어, 동의어에 존재 한다면, 해당 키워드들을 그룹화 한다.

의미적 연관성이 있는 키워드 끼리 단어 그룹화가 생성 되면 강한 문장 선정 과정을 거치게 된다. [표 3-11]은 단어 그룹화를 생성하는 과정을 나타내는 소스코드 이다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

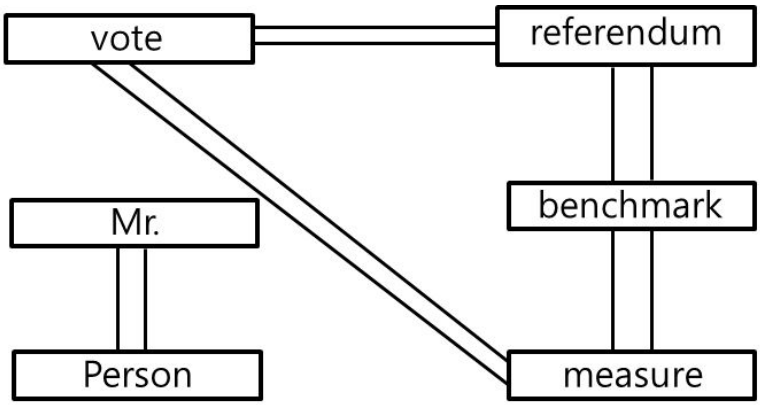
[표 3-11] 단어 그룹화 소스코드

```
def find_chain(all_nouns, key_nn):
    chain_train = []

    for each in key_nn:
                                        ⋮

        for n in all_nouns: ##키워드 명사 탐색
            if each == n:
                temp_chain.append(n)
            ##수집한 문서 키워드를 읽기 위해 append(n) 함수를 이용한다.
            elif (chk_hypo(each, n) == '1'):
                temp_chain.append(n)
            ##워드넷 상위어 탐색
            elif (chk_hyper(each, n) == '1'):
                temp_chain.append(n)
            ##워드넷 하위어 탐색
            elif (chk_coordinate(each, n) == '1'):
                temp_chain.append(n)
            ##워드넷 동위어 탐색
           elif (chk_synset(each, n) == '1'):
                temp_chain.append(n)
            ##워드넷 키워드 신셋 넘버 탐색
            else:
                pass

        chain_train.append(temp_chain)
        ##print temp_chain

    return chain_train
```

[표 3-11]는 단어 그룹화를 하기 위해 키워드가 입력되면 두 키워드가 워드넷의 상위어, 하위어, 동의어를 탐색하여 해당 키워드가 존재할 경우 단어 그룹화가 된다. [그림 3-4]는 [표 3-11] 에 대한 결과로 각 의미적인 연관성을 지닌 단어 그룹화를 예시도로 표현 해보았다.

Collection @ chosun

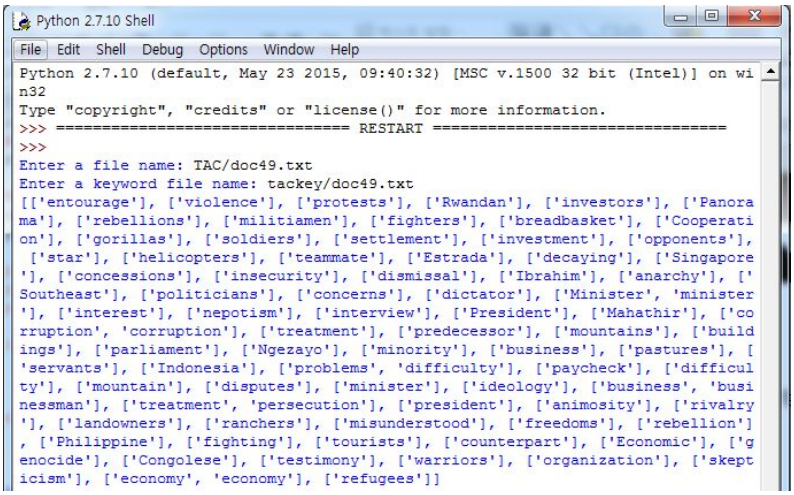
조선대학교
CHOSUN UNIVERSITY



[그림 3-4] 단어 그룹화 예시도

[그림 3-4]의 단어 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 단어 그룹화가 생성된다.

[그림 3-5]는 [표 3-11]에 대한 결과로 각 의미적인 연관성을 지닌 상위 키워드와 하위 키워드가 단어 그룹화가 된 결과를 보여준다.



[그림 3-5] 단어 그룹화 생성

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[그림 3-5]에서 보는 바와 같이 ['treatment', 'persecution'], ['economy', 'economics']등 단어 그룹화를 생성되었다. 그리고 문서 내 중복되는 단어 또한 단어 그룹화로 생성되기 때문에 ['President', 'President']가 생성 되었다.

2. 문서 요약

단어 그룹화 단계를 거쳐 문서 요약을 위해 중요 문장 추출 과정을 거치게 된다. 단어 그룹화를 이용해 강한 문장 가중치 수식을 적용하여 각 문장 별로 점수를 선정한다. 이를 위한 수식은 다음과 같다.

$$Score(chain) > Avg(Score) + 2 \times StandardDeviation(Score) \quad (5)$$

수식 (5)는 앞 장에서 진행한 단어 그룹화를 이용해 각 문장별로 가중치 점수를 부여 하여, 문서에 핵심이 되는 문장을 추출하게 된다. 두 키워드간의 단어 간 연관성이 성립되지 않은 경우에도 중요 문장 추출 과정에 있어 가중치 점수를 부여받게 된다. 수식 (5)의 Avg(Score)는 각 수집한 문서의 문장에 단어 그룹화로 가중치 점수를 부여하게 된다. [그림 3-4]와 같이 'vote'와 'mr'가 단어의 의미적 연관성이 성립되지 않고, 단어 그룹화가 생성 되지 않았으나, 수집한 문서의 문장에 해당 단어가 존재 할 경우 0.5점을 부여 받게 된다. 하지만 'vote'와 'referendum'의 의미가 같아 단어 그룹화가 생성 되었을 경우 1.0점을 부여받게 된다. 생성된 단어 그룹화에 텍스트 랭크 알고리즘을 통해 추출된 상위 키워드가 존재 할 경우 추가 점수를 받게 되며, 상위 키워드로 이루어진 단어가 문장에 존재 할 경우 추가 점수 1.0점을 받아 강한 문장 가중치 점수 2.0점을 받게 된다. 단어 그룹화에 상위 키워드와 일반 키워드로 생성 되었을 경우 0.5점의 추가 점수를 받으며, 1.5점을 부여받게 된다. 단어 그룹화가 생성 되지 않았고, 해당 단어가 상위 키워드로 이루어진 경우 0.5점의 추가 점수를 받아 1.0점의 가중치 점수를 받게 된다. 수식(5)의 Standard Deviation(Score)는, 각 문장 별로 이루어진 가중치 점수 선정 한 뒤 가중치 점수의 표준 편차를 구하게 된다. 다음 [표 3-12]는 이에 대한 과정을 나타내는 소스 코드이다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-12] 강한 문장 가중치 수식 소스코드

```
def score_chain(chain, key_nouns, rel_nouns):
    num_chain = len(chain)
    store_scr = []
                                ⋮
        ##강한 문장 가중치 수식 소스코드
        for i in each:
            for w,t,v in key_nouns:
            ## w: 키워드 t: 상위 키워드 v: 점수
                if (i == w) and (v == 1):
                    chain_scr += 1
                ##단어 그룹화가 생성 되어 문장에 존재 할 경우1점을 받는다.
                elif (i == t) and (v  == 1.5):
                    chain_scr += 1.5
                ##생성된 단어 그룹화에 하나 키워드가 상위 키워드일 경우
                    0.5점의 추가 점수를 받아 1.5점을 받는다.
                elif (i == t) and (v  == 2):
                    chain_scr += 2
               ##단어 그룹화에 텍스트 랭크 알고리즘의 상위 키워드로 구성 되어
                   있다면 각각 1.0점의 추가 점수를 받아 2점을 받는다.
                else:
                    pass

            for w,t,v in rel_nouns: ## 그룹화 되지 않은 키워드
                if (i == w) and (v == 0.5):
                    chain_scr += 0.5
                ## 단어 그룹화가 되지 않은 키워드일 경우 문장에 존재 한다면,
                     0.5점의 가중치 점수를 받는다.
                elif (i == t) and (v == 1):
                    chain_scr += 1
                ##  키워드가 상위 키워드 일 경우 0.5점의 추가 점수를 받
                     아 1점을 받는다.
                else:
                    pass
        store_scr.append(chain_scr)
    get_the_chain = zip (chain, store_scr)

    return get_the_chain
```


Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-12]는 본 절에서 제안한 강한 문장 가중치 수식 소스코드이며, 이를 이용해 수집한 문서에서 중요 문장을 추출 하여 문서를 요약 하게 된다. 아래의 [표 3-13]은 [표 3-12]의 과정을 거쳐 가중치 점수의 평균을 도출한 결과이다.

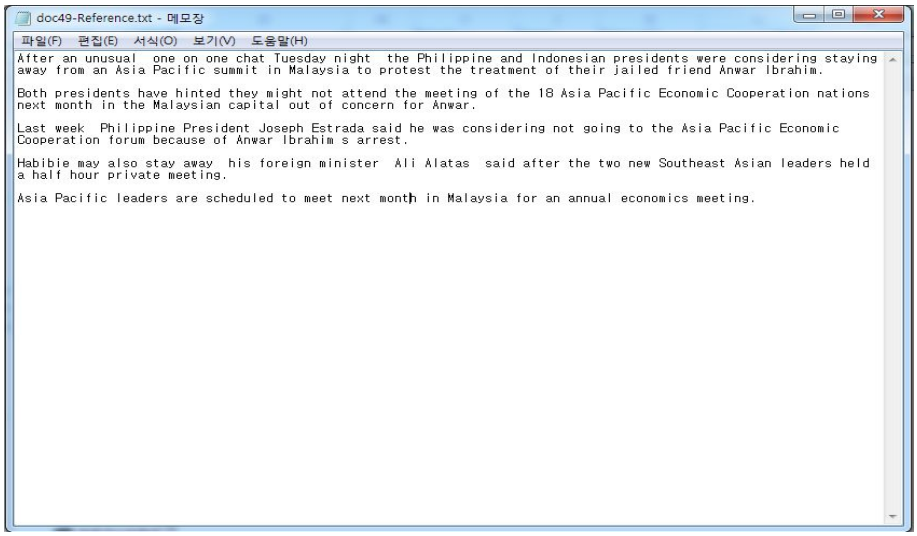
[표 3-13] 강한 문장 선정 가중치 결과

	영어 문장
원문	leaders difficult because of his concerns about the arr est of Malaysia's former deputy prime minister, a Thai newspaper reported Sunday. Asia-Pacific leaders are s cheduled to meet next month in Malaysia for an annua l economics meeting. Last week, Philippine President J oseph Estrada said he was considering not going to t he Asia-Pacific Economic Cooperation forum because of Anwar Ibrahim's arrest. ……
하위 키워드	leaders, difficult, arrest, former, deputy, prime, ministe r, newspaper, report, Sunday, Asia, pacific, schedule, Last, week, philippine, economy, ideology, rivalry, gen ocide, organization ……
상위 키워드	entourage, protests, presidents, Philippine, President, E conomic, cooperation, helicopters, Estrada, Singapore, dismissal, Ibrahim, Southeast, concerns, treatment, Pa norama, interview, President, corruption, economy, pre decessor, parliament, Minister, Mahathir, business, Ind onesia, problems, counterparts ,difficulty. minister.
단어 그룹화	['entourage'], ['violence'], ['protests'], ['Rwandan'], ['investors'], ['Panorama'], ['rebellions'], ['militiamen'], ['fighters'], ['breadbasket'], ['Cooperation'], ['gorillas'], ['soldiers'], ['settlement'], ['anarchy'], ['Southeast'], ['politicians'], ['concerns'], ['dictator'], ['Minister', 'minister'], ['interest'], ['nepotism'], ['interview'], ['President'], ['Mahathir'], ['corruption', 'corruption'], ['treatment'], ['predecessor'], ['mountains'], ['buildings'], ['parliament'], ['Ngezayo'], ['minority'], ['business'], ['pastures'], ['servants'], ['Indonesia'], ['problems', 'difficulty'], ['paycheck'],……
가중치 점수 평균	1.56728052427

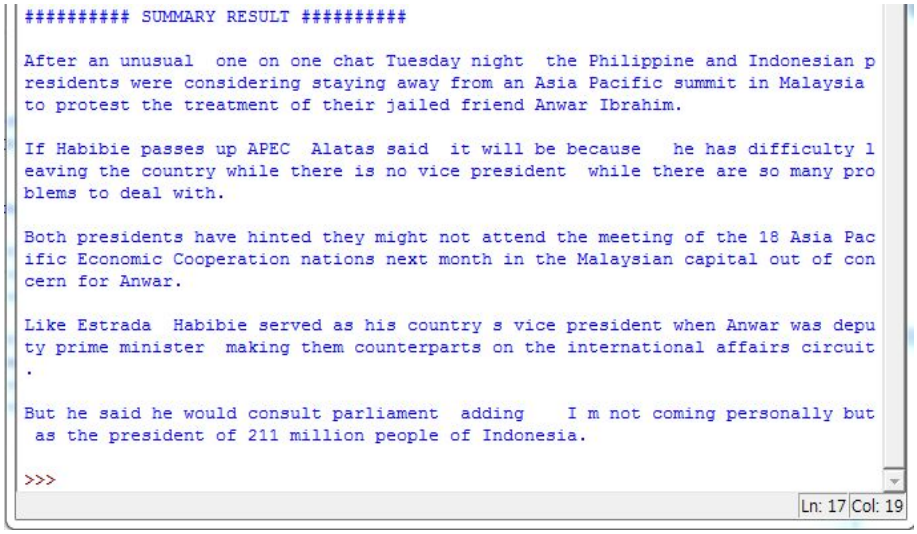
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-13]의 가중치 점수 평균 결과에서 [‘problems’, ‘difficulty’]는 단어 그룹화가 되어 해당 문장에 존재 할 경우 1.0점의 가중치 점수를 받으며, ‘problems’과 ‘difficulty’는 상위 키워드에 속하므로 추가 점수 1.0점을 받아 2.0점을 받게 된다. 아래의 [그림 3-7]은 [표 3-12]의 과정을 기반으로 문서가 요약된 결과를 보여주고 있으며, [그림 3-6]은 같은 문서를 전문가가 요약한 결과이다.



[그림 3-6] 전문가 문서 요약 결과



[그림 3-7] 문서 요약 결과

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[그림 3-7]에서 보는 바와 같이 [표 3-13]의 가중치 점수의 평균에서 보는 바와 같이 가중치 점수 평균 이상인 문장들은 중요 문장으로 추출된다. 문서 요약의 결과는 가중치 점수 평균 1.56728052427점 이상 나온 문장들을 추출하여 요약한 결과이며, 각 문장별 가중치 점수가 1.56728052427점의 평균 이하로 나온 문장들은 문서에서 제외 되어 문서가 요약 되었다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅳ. 실험 및 결과

본 장에서는 Ⅲ장에서 제안한 문서요약 방법에 대한 효율성 평가를 위해 TAC(Text Analysis Conference)에서 제공해주는 데이터 세트 수집하여 실험을 진행 하였으며, 결과 분석을 위해 평가 기준을 2가지로 나누어 진행 하였다.

A. 데이터 세트(Data Set)

1. TAC(Text Analysis Conference)

TAC 평가 자료는 문서요약 평가 자료의 표준으로 사용 되고 있다. TAC는 전문가들이 작성한 요약문과 제안된 시스템이 만든 요약문을 비교하여 각 문서요약 시스템 성능을 평가하는 국제회의다. 본 실험에서 쓰이는 TAC는 주로 여러 방면의 주제와 그 주제에 관련된 자료의 문서들을 군집으로 이루어져 있으며, 본 실험에서는 TAC 2015와 TAC 2016을 이용하였다. 다음 [표 4-1]은 TAC평가 자료에 포함된 레이블과 주제의 예시를 나타내며, [표 4-2]는 평가를 위하여 수집한 문서의 군집 주제의 질의의 예시를 나타낸다.

[표 4-1] TAC 평가 자료 주제

군집	군집 주제
d100a	New research studies are providing valuable insight into the probable causes of schizophrenia
d101a	Million soldiers fight flood that have already killed 1,268
d102a	With Snoopy looking on, Charles Schultz says farewell to readers.
……	……

[표 4-2] TAC 평가 자료 질의

군집	질의
d100a	A family history of schizophrenia remains the best predictor of whether a person will develop the illness, ……
d101a	China's military has ordered soldiers to fight to the death to ensure that waterlogged dikes holding back tported Monday. …….
……	……

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

B. 실험 평가 방법 및 결과 분석

1. 실험 평가 방법

실험의 평가는 ROUGE(Recall-Oriented Understudy of Gisting Evaluation)평가 시스템을 이용하였다[27]. ROUGE 평가 시스템은 전문가가 직접 요약한 문서와 자동으로 요약된 시스템 문서를 비교하는 방법으로 본 논문에서는 ROUGE-N을 이용해 논문의 실험을 평가하였다. ROUGE-N은 n-gram을 이용하여 재현율, 정확률을 이용해 최종적으로 F-Score의 결과를 측정하는 방법이다.

n-gram은 n개의 어절 또는 음절을 연쇄적으로 분류해 문서의 빈도를 측정 하는 방법이다[28]. n=1 일 때는 uni-gram, n=2 일 때는 bi-gram, n=3는 tri-gram, n =4 일 때는 four-gram을 기준으로 한다.

ROUGE의 재현율은 문서 요약 시스템을 통해 추출된 n-gram의 개수와 전문가가 요약한 문서와 직접 요약한 문서의 n-gram 집합 개수를 이용하여 수식 (6)에서 보는 바와 같이 재현율을 측정한다.

$$Recall = \frac{|SystemSummary| \cap |ReferenceSummary|}{|SystemSummary|} \qquad (6)$$

정확률은 전문가가 직접 요약한 문서의 n-gram 개수와 문서 요약 시스템을 통해 추출된 n-gram의 개수 그리고 전문가가 요약한 문서의 개수를 이용하여 정확률을 측정하게 되며, 수식 (6), 수식 (7)을 이용해 수식 (8)에서 보는 바와 같이 F-Score를 측정하게 된다.

$$\mathrm{Precision} = \frac{|SystemSummary| \cap |ReferenceSummary|}{|ReferenceSummary|} \qquad (7)$$

$$F-Score = \frac{2*\mathrm{Precision}*Recall}{\mathrm{Precision}+Recall} \qquad (8)$$

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

앞 절에서 각각 재현율, 정확률, F-Score를 측정하기 위해 ROUGE-N을 측정하기 위한 방법은 다음 수식 (9)와 같다.

$$ROUGE-N = \frac{\sum_{S \in [ReferencsSUmmary]} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in [ReferencsSUmmary]} \sum_{gram_n \in S} Count(gram_n)} \quad (9)$$

수식 (9)의 $gram_n$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. 결과 분석

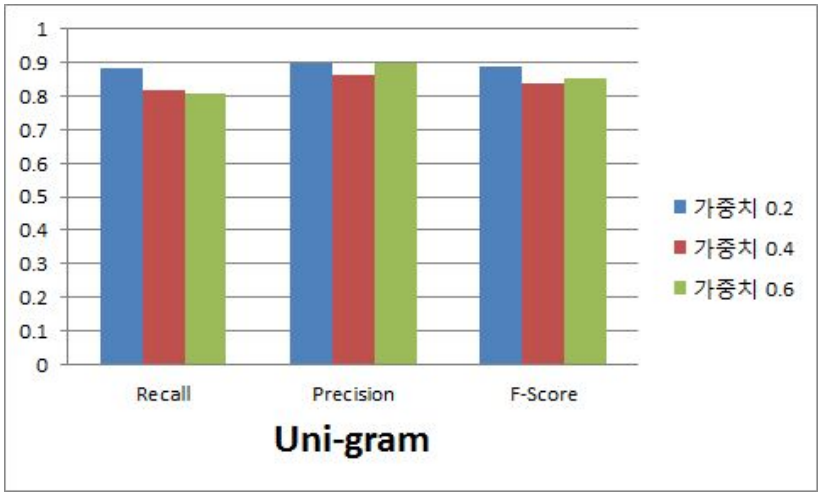
본 절에서는 문서요약 시스템의 효율성 검증을 위해 강한 문장 가중치 점수를 이용한 결과 분석과 비교 평가를 위해 베이스 라인 실험을 진행 하였다.

a. 강한 문장 가중치 점수를 이용한 결과 분석

본 절에서는 제안한 방법의 정확한 평가를 위해 기존에 연구 되었던 문서요약 방법과 비교 평가를 진행 하였으며[29], 이를 위해 강한 문장 가중치 점수를 이용한 결과 분석을 2가지로 나누어 실험을 하였다. 첫 번째 결과 분석은 본 논문에서 제시하고 있는 방법에서 텍스트랭크 알고리즘의 상위 키워드의 추가 점수를 부여하지 않고 강한 문장 가중치 점수만을 변형하여 성능 평가를 하였으며, 강한 문장의 가중치 변형한 실험과 본 논문에서 제시하는 방법과 비교 평가를 하기 위해 먼저 강한 문장 가중치 점수에서 각각 0.2, 0.4, 0.6점 점수를 이용하여 가장 좋은 성능이 나온 점수를 기반으로 본 논문에서 제시하고 있는 실험 방법과 비교하여 성능 검증을 하였다. 그리고 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.

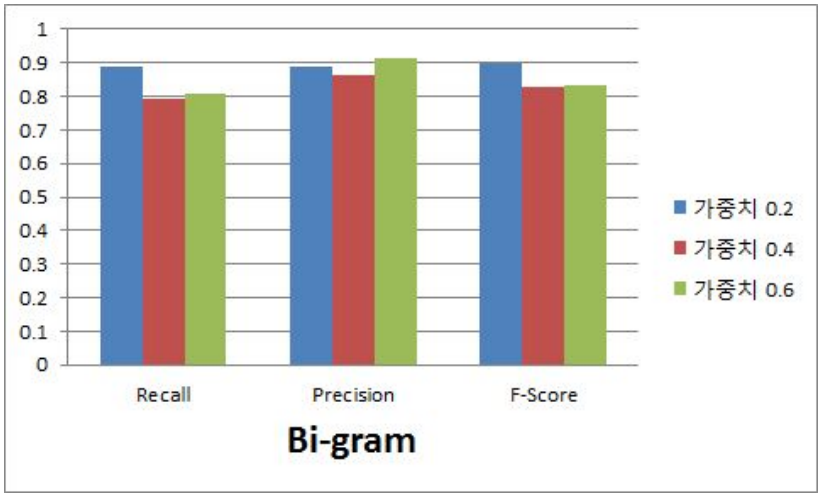
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 4-1] ROUGE-1 평가 결과

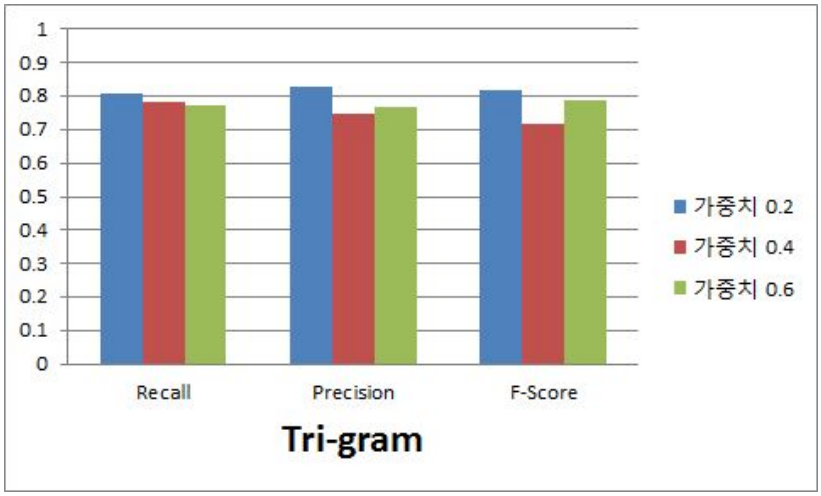
[그림 4-1]는 강한 문장 가중치 점수만을 추가적으로 점수를 주어 나타난 결과이이며, ROUGE-N의 uni-gram기준으로 평가한 결과이다. Uni-gram을 기준으로 하였을 때, 재현율과 F-Score의 측면에서 가장 좋은 성능을 나타낸 결과는 가중치 점수 0.2점으로 확인 할 수 있다. 가중치 점수 0.2점의 재현율은 0.88472점, 정확률은 0.869624점, F-Score는 0.89048점이 나왔다.



[그림 4-2] ROUGE-2 평가 결과

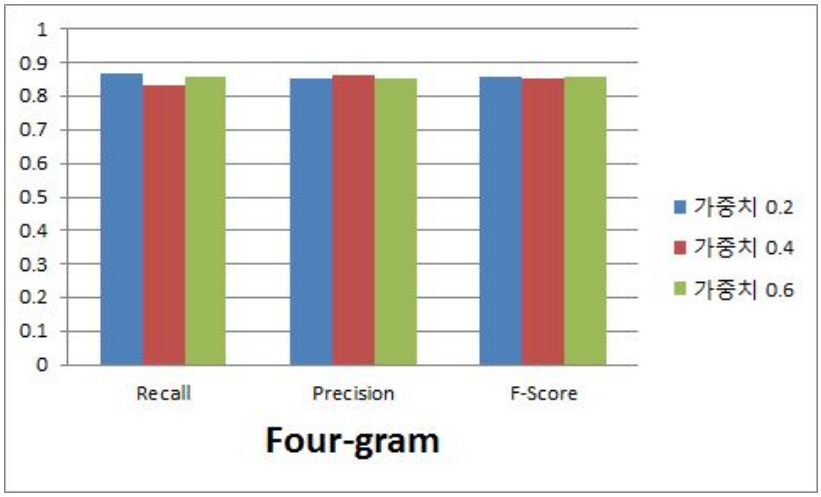
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 4-3] ROUGE-3 평가 결과

[그림 4-2]와 [그림 4-3]의 평가 결과는 ROUGE-N의 bi-gram과 tri-gram을 기준으로 평가를 하였으며, [그림 4-2]는 재현율과 F-Score값 부분에서 가중치 점수 0.2점이 가장 효율 적으로 나왔다. 재현율은 0.88769점이 나왔으며, F-Score는 0.89818점이 나왔다. [그림 4-3]은 마찬가지로 가중치 점수 0.2점을 주었을 때 재현율이 0.87795점이 나왔으며, 정확률은 0.88993 점으로 가중치 점수 0.2점만을 추가한 결과가 더 좋은 결과가 나왔다.

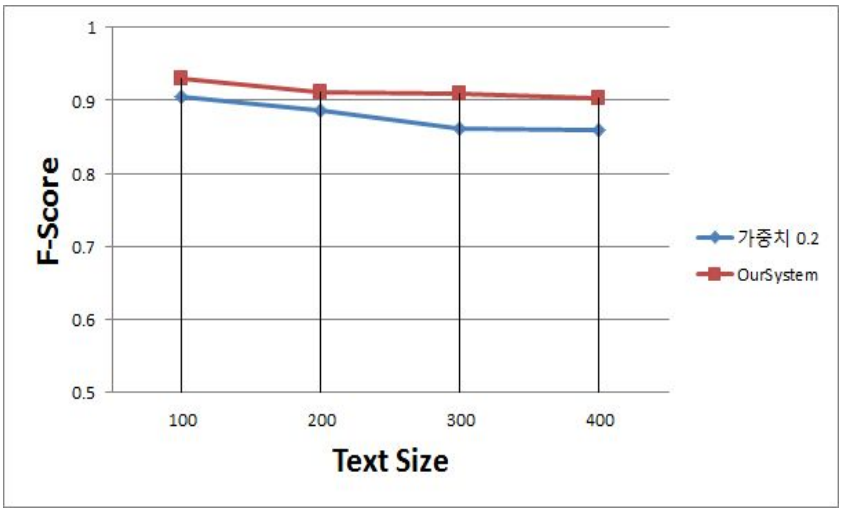


[그림 4-4] ROUGE-4 평가 결과

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

마지막으로 [그림 4-4]는 four-gram을 기준으로 평가하기 때문에 기존에 수집하였던 문서의 양을 2배로 늘려서 실험을 진행 하였다. four-gram을 이용한 방법 역시 가중치 점수 0.2점을 이용한 결과가 각각 재현율, 정확률, F-Score의 값이 0.86822, 0.85099, 0.85994점이 나오는 것을 확인 할 수 있었으며, 강한 문장 가중치 점수 0.2점만을 이용한 결과가 좋은 성능을 나타나는 것을 확인 할 수 있었다. 이를 기반으로 본 논문에서 제시하는 방법과 가중치 점수 0.2점만을 추가한 방법과 비교 실험을 해보았다.



[그림 4-5] 강한 문장 가중치 점수를 이용한 평가 결과

본 논문의 텍스트 랭크 알고리즘 기반의 상위 키워드 추가 점수 성능 검증을 위해 TAC에서 제공하는 TAC 2015 문서 중 400문서를 이용하여 비교 평가를 하였다. [그림 4-5]는 전체적인 F-Score평균을 기반으로 나타난 결과이며, 텍스트랭크 알고리즘 기반의 문서요약 방법이 기존에 가중치 점수만을 추가한 방법보다 효율적으로 좋은 성능이 나타나는 것을 알 수 있었다. [그림 4-5]의 평가 결과는 [표 4-3]과 같다.

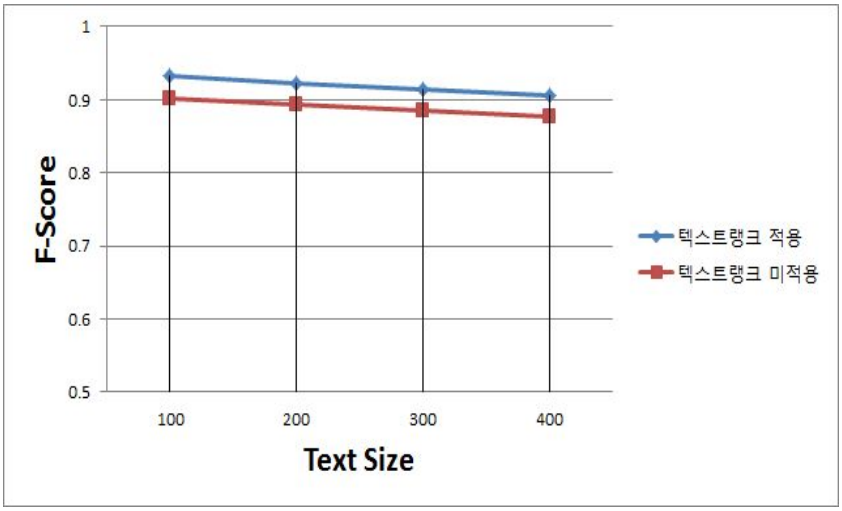
[표 4-3] 강한 문장 가중치 점수 평가 결과

Text Size	가중치 0.2	OurSystem
100	0.90614	0.92974
200	0.88711	0.91181
300	0.86214	0.91015
400	0.85984	0.90254

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

두 번째 결과 분석은 앞서 실험한 평가 방법에서 반대되는 예를 들어 결과 분석을 해보았다. 가중치 점수 0.2점만을 이용한 방법과 본 논문에서 제시하는 추출요약 기법을 적용하여 단어 그룹화에 상위키워드로만 이루어진 경우 0.4점의 추가 점수를 받아 1.4점의 가중치 점수를 얻는다. 그리고 상위키워드와 하위 키워드로 이루어진 경우 1.2점의 가중치 점수를 받으며, 또한 하나의 키워드로 이루어진 단어일 경우 0.7점의 가중치 점수를 얻는다. 반면 본 논문에서 제시한 방법에서 텍스트 랭크를 적용하지 않고 1.0점의 점수를 추가 하였으며, 단어 그룹화로 이루어지지 않는 경우 0.5점의 추가 점수를 주어 각각 2.0점과 1.0점의 가중치 점수를 받고 해당 문장에 단어가 존재 하지 않을 경우 0점을 주는 가중치 점수를 추가적으로 주어 기존 연구와 반대되는 예로 실험을 진행 하였다.



[그림 4-6] 텍스트랭크 알고리즘 검증 결과

[그림 4-6]은 기존에 연구 되었던 0.2점만을 추가한 방법에서 텍스트랭크를 적용한 방법과 본 논문에서 제시한 방법의 텍스트랭크를 미적용한 결과이다. 가중치 점수를 무조건 크게 주는 것보다 본 논문에서 제시하는 추출 요약 기법을 적용하여 가중치 점수를 다양하게 주는 것이 [그림 4-6]에서 보는 바와 같이 전체 적인 재현율, 정확률, F-Score 평균으로 보았을 때, 가중치 점수 0.2점에서 텍스트랭크를 적용한 결과가 효율적으로 나오는 것을 확인 할 수 있었다. 이러한 결과를 통해 본 논문에서 제시하는 텍스트랭크 알고리즘을 통한 상위 키워드의 추가 점수를 적용한 기법이 효율성이 있었음을 결과 분석을 통해 알 수 있었으며, [표 4-4]는 [그림 4-6]에 대한 결과이다.

Collection @ chosun

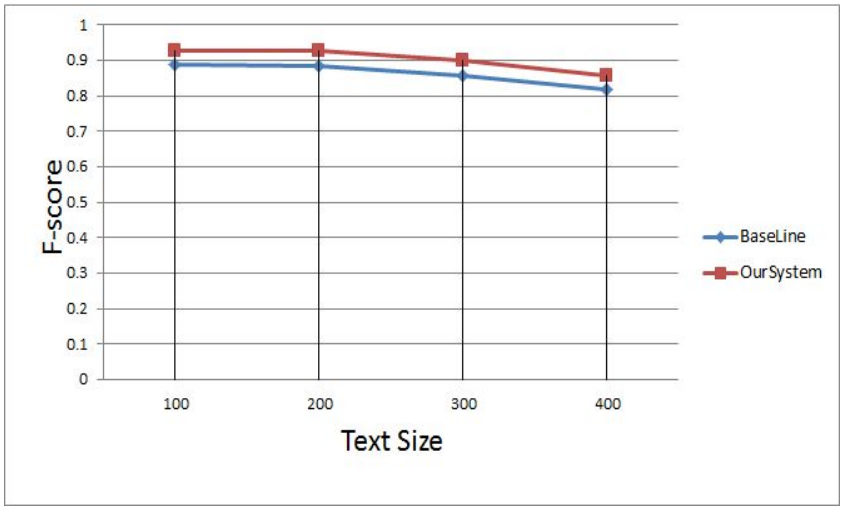
조선대학교
CHOSUN UNIVERSITY

[표 4-4] 텍스트랭크 알고리즘 검증 결과

Text Size	텍스트랭크 미적용	텍스트랭크 적용
100	0.90190	0.92310
200	0.89312	0.91252
300	0.88442	0.91442
400	0.87571	0.90558

b. 비교 평가 실험

본 절에서는 본 논문의 성능 검증을 위해 기존에 연구 되었던 문서 요약 연구와 비교 평가 실험을 진행 하였다[30]. 비교 평가 실험 방법 역시 본 연구의 초점인 추출 요약 기반으로 진행된 논문이다.



[그림 4-7] 비교 평가 실험결과

[그림 4-7] 결과는 TAC에서 제공하는 TAC 2016 문서 중 400문서를 이용해 비교 평가한 결과이며, 비교 결과를 ROUGE-N의 전체적인 평균 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 4-7]의 평가 결과는 [표 4-5]와 같다.

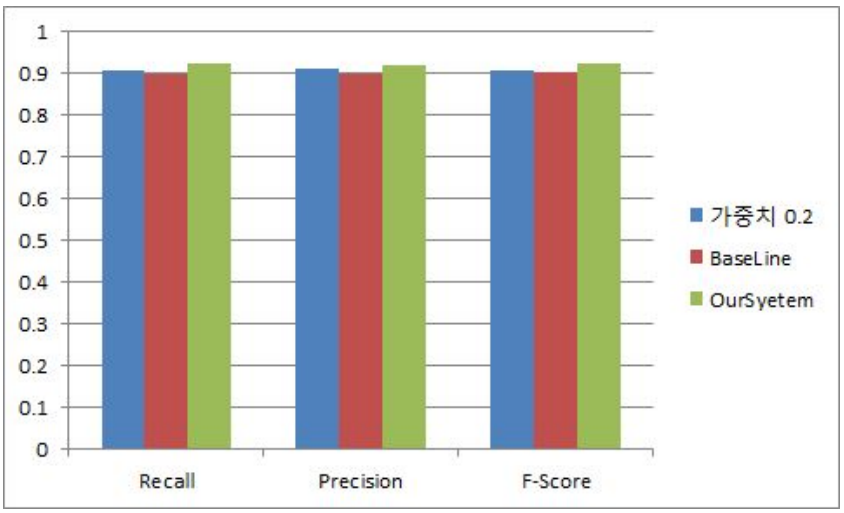
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 4-5] 비교평가 실험 결과

Text Size	BaseLine	OurSystem
100	0.88723	0.92723
200	0.88412	0.92589
300	0.85889	0.90045
400	0.81844	0.85889

[표 4-5]에서 보는 바와 같이 기존의 연구 되었던 문서요약 방법 보다 본 연구에서 제안하는 문서요약 시스템을 사용했을 때와 비교해 성능 차이가 있다는 것을 확인 할 수 있었다. 아래의 [그림 4-8]은 본 실험에서 비교 평가한 가중치 점수 0.2점만을 준 방법과 비교 실험 평가 방법, 본 논문의 방법들을 재현율, 정확률, F-Score의 전체 적인 평균을 보여 주고 있다.



[그림 4-8] 실험 및 평가 결과

이와 같이 본 논문에서 제시하는 문서요약 방법을 통해 사용자가 원하는 문서를 요약함으로써 해당 문서의 원하는 정보를 쉽게 찾을 수 있도록 해결 했다는데, 큰 의미가 있으며, 이를 통해 문서요약 방법의 체계가 갖춰진다면 스마트 디바이스나 웹사이트를 통해 중요 이슈나 사회적 현상 등 다양한 분야에서 문서 검색 서비스의 다양한 분석이 가능 할 것으로 기대된다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

V. 결론 및 제언

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 문장에 0점을 부여하였기 때문에 핵심 주제를 추출하는 데에 있어 정확도가 떨어졌었다. 그렇기 때문에 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.

제안한 방법에 대해 성능을 평가하기 위해 본 실험에서는 2가지로 나누어 실험을 진행하였다. 첫 번째 평가 방법은 본 논문에서 제안한 방법과 제안한 방법에서 상위 키워드의 추가 점수를 부여하지 않고 단순히 가중치 점수를 부여하는 방법에 있어 0.2점만을 추가적으로 올려 비교 실험을 평가한 결과 90.5%로 기존의 연구 방법보다 본 연구에서 제안한 실험 방법이 더 좋다는 것을 알 수 있었으며, 두 번째 평가 방법은 기존에 연구되어오던 문서 요약 방법과 비교 실험 평가한 결과 92.2%로 기존 문서 요약 방법보다 핵심적인 내용을 담고 있는 문장을 추출하며, 더욱 효과적인 문서를 요약할 수 있었다는 것을 검증할 수 있었다. 향후 연구로는 보다 다양한 연관 단어를 파악할 수 있는 태그 클러스터를 구축하여 문장과 단어 사이의 관계에 대해서 정확도를 높이기 위한 문서 요약 방법을 제안하고자 한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

참고문헌

[1] http://likms.assembly.go.kr/record/index.html

[2] Inderjeet Mani, Automatic Summarization, John Benjamins Publishing Company, 2001.

[3] J.Kupiec, J.Pedersen, F.Chen,“A Trainable Document Summarizer”, Proceedings of 18th ACM-SIGIR Conf., pp. 68-73, 1995.

[4] 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.

[5] 류동원, 이종혁, "단어공기정보를 이용한 자동화 문서 요약",제27회 정보과학회 봄 학술발표논문집(B), pp. 339-341, 2000.

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

[7] Mihalcea, Rada, Paul Tarau. "TextRank: Bringing order into texts." Association for Computational Linguistics, 2004.

[8] Julian Kupiec, Jan Pedersen, and Francine Chen, "A Trainable Document Summarizer." Proceedings of 18th ACM-SIGIR Conference, pp.68~73, 1995.

[9] C. Aone, M. E. Okurowski, J. Gorlinsky, and B. Larsen, “A Scalable Summarization System using Robust NLP”, Proceedings of the Workshop on Intelligent Scalable Text Summarization (ACL/EACL’97), pp. 66-73, 1997.

[10] Chanback Jeong, Taehwan Kim, Hochul Jeon, Joongmin Choi Department of Computer Science & Engineering Hangyang University, A News Recommendation System based on Document Clustering Using WordNet, 2008.

[11] D. Marcu, "The Rhetorical Parsing of Natural Language Texts.", Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pp. 96~103, 1997.

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

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[13] Henning, Leonhard, "Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis", Proceedings of the International Conference RANLP'09, 2009,

[14] Xiaojun Wan, Jianwu Yang, "Multi-Document Summarization Using Cluster-based Link Analysis",Proceedings of the International Conference(SIGIR'08), 2008.

[15] 한경수, "질의분해를 이용한 적합성 피드백 기반 자동문서요약", 고려대학교 컴퓨터 학과 석사학위논문, 2001.

[16] Dang, C and Luo, X, "WordNet-based Document Summarization", 7th WSEAS Int. Conf. on Applied Compuster & Applied Computational Science (ACACOS'08), 2008.

[17] WonMoon Song, YoungJin Kim, EunJu Kim, MyungWon Kim, "A Document Summarization System Using Dynamic Connectino Graph", Journal of KIISE : Software and Applications, vol.36, no.1, pp.62-69, Jan. 2009.

[18] Mihalcea, R., & Tarau, P. “TextRank: Bringing order into texts. Association for Computational Linguistics. 2004.

[19] Thakkar, K. S., Dharaskar, R. V., & Chandak, M. B. 2010. “Graph-based algorithms for text summarization”. In 2010 3rd International Conference on Emerging Trends in Engineering and Technology : 516-519.

[20] 조형락, 김성진, 이동호. “의미기반 텍스트 랭크 알고리즘을 이용한 다중문서 요약. 한국정보과학회 2015년도 동계학술발표회 논문집: 756-758.

[21] Erkan, G., & Radev, D. R. “ LexRank: Graph-based lexical centrality as salience in text summarization”. Journal of Artificial Intelligence Research, 22: 457-479. 2004.

[22] Harabagiu, S. Finley L. "Topic Themes for Multi Document Summarization," In proceeding of ACM SIGIR, pp. 202-209, 2005.

[23] Inderjeet Mani, Automatic Summarization, Kohn Benjamins Publishing Co. 2001.

[24] http://terms.naver.com/entry.nhn?docId=857191&cid=50371&categoryId=50371

[25] Hirschberg, Daniel. S. "Algorithms for the longest common subsequence

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

problem." Journal of the ACM (JACM) Vol. 24. No. 4. pp.664-675, 1977.

[26] Bergroth, Lasse, Harri Hakonen, and Timo Raita. "A survey of longest common subsequence algorithms." String Processing and Information Retrieval, 2000. SPIRE 2000. Proceedings. Seventh International Symposium on. IEEE, pp.39-48, 2000.

[27] 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.

[28] http://terms.naver.com/entry.nhn?docId=862633&cid=42346&categoryId=42346

[29] JunSeok Cha, PanKoo Kim. “The Automatic Text Summarization Using Semantic Relevance And Hierarchical Structure Of Wordnet.” 11th International Conference on Broadband and Wireless Computing, Communication and Applications 2016.

[30] Jee-Uk Heu, Young-Do Joo, Dong-Ho Lee, Multi-Document Summarization Technique using Semantic Analysis between Tags, 2012.

Collection @ chosun