단어 관련성 추정과 바이트 페어 인코딩(Byte Pair Encoding)을 이용한 요약 기반 다중 뉴스 기사 제목 추출

유홍연[O*], 이승우[**], 고영중[*]
[*]동아대학교, [**]한국과학기술정보연구원
{hongyeon1408, pinesnow.lee, youngjoong.ko}@gmail.com

Summarization Based Multi-news Title Extraction Using Term Relevance Estimation and Byte Pair Encoding

Hongyeon Yu[O*], Seungwoo Lee[**], Youngjoong Ko[*]
[*]Dong-A University, [**]Korea Institute of Science and Technology Information (KISTI)

요 약

다중 문서 제목 추출은 하나의 주제를 가지는 다중 문서에 대한 제목을 추출하는 것을 말한다. 일반적으로 다중 문서 제목 추출에서는 다중 문서 집합을 단일 문서로 본 다음 키워드를 제목 후보군으로 추출하고, 추출된 후보를 나열하는 형식의 연구가 많이 진행되어져 왔다. 하지만 이러한 방법은 크게 두 가지의 한계점을 가지고 있다. 먼저, 다중 문서를 단순히 하나의 문서로 보는 방법은 전체적인 주제를 반영한 제목을 추출하기 어렵다는 문제점이 있다. 다음으로, 키워드를 조합하는 형식의 방법은 키워드의 단위를 찾는 방법에 따라 추출된 제목이 자연스럽지 못하다는 한계점이 있다. 따라서 본 논문에서는 이 한계점들을 보완하기 위하여 단어 관련성 추정과 Byte Pair Encoding을 이용한 요약 기반의 다중 뉴스 기사 제목 추출 방법을 제안한다. 평가를 위해서는 자동으로 군집된 총 12개의 주제에 대한 다중 뉴스 기사 집합을 사용하였으며 전문 교육을 받은 연구원들이 정성평가를 진행하여 5점 만점 기준 평균 3.68점을 얻었다.

주제어: 다중 문서 제목 추출, 요약 기반 제목 추출, 단어 관련성 추정, Byte Pair Encoding,

1. 서론

제목 추출(Title Extraction)이란 하나의 주제를 가지는 단일 혹은 다중 문서에 대한 제목을 추출하는 것을 말한다. 문서의 제목을 추출하는 것은 짧은 텍스트(Short Text)를 이용하여 핵심 내용을 간단히 표현할 수 있기 때문에 매우 중요하다. 특히, 다중 뉴스 문서에 대한 제목 추출을 이용하여 대량의 뉴스를 간단히 표현하는 것은 이슈 분석에 많은 도움이 된다.

최근 심층학습 기반 End-to-End 시스템이 높은 성능을 보여줌에 따라 각 문서의 제목을 이용하여 간단하게 학습 데이터를 구축할 수 있는 단일 문서 제목 추출에 대한 연구는 생성기반으로 많이 진행되어지고 있다[1]. 그러나 학습 데이터 구축이 비교적 어려운 다중 문서 제목 추출에는 지도학습 기반 기법을 적용하기 어렵기 때문에 다중 문서를 단일 문서로 보고 제목 후보를 선택한 뒤 각 후보의 스코어를 계산하여 스코어가 가장 높은 후보를 제목으로 선택하는 키워드 추출 기반 방법이 사용될 수 있다[2,3]. 하지만 이러한 방법은 크게 두 가지 한계점이 존재한다. 첫 번째, 다중 문서를 단순히 단일 문서로 보기 때문에 각 단어의 스코어 계산 시 추가적인 정보가 없다면 문서 집합에 대한 전체적인 주제에 대한 정보를 반영하기 어렵다. 두 번째, 제목의 후보들에 대한 단위를 선택할 때 대부분 N-gram이나 명사와 형용사의 조합을 사용하는데, 제목 후보군이 적절하지 못할 경우 제대로 된 제목을 추출할 수 없거나 추출이 되더라도 자연스럽지 못하다는 한계점이 있다.

따라서, 본 논문에서는 다중 문서 요약을 통해 추출된 요약에 대한 정보를 이용하여 문서 집합의 전체적인 주제 정보가 반영된 단어 관련성를 추정하고, Byte Pair Encoding (BPE)[4] 기반으로 중요 키워드를 추출 한 뒤 요약 문장에 적용하여 최종 제목 후보를 추출하는 방법을 제안하여 두 가지 한계점을 해결한다.

제안 시스템은 크게 다섯 단계로 다중 문서 요약 단계, 요약 기반 단어 관련성 추정 단계, BPE 기반의 중요 키워드 선택 단계, 추정된 단어 관련성와 선택된 중요 키워드를 이용한 제목 추출 단계, 후처리 단계로 나뉜다. 하지만 본 논문에서는 다중 문서 요약 단계는 다루지 않고 [5]에서 제안한 방법의 다중 문서 요약을 시행한 결과를 사용하였다. 제목 추출 결과에 대한 평가를 위해 자동으로 군집된 총 12개의 주제에 대한 다중 뉴스 기사 집합을 사용하였으며, 관련 전문 교육을 받은 연구원이 다중 문서 집합에 대한 제목으로 적절한지에 대한 정성 평가를 진행하였다.

본 논문의 구성은 2장에서 관련 연구를 소개하고, 3장에서는 단어 관련성 추정과 BPE를 사용한 요약 기반의 제목 추출 방법을 소개한다. 4장에서는 추출된 제목에 대한 결과를 평가하고, 마지막 5장에서 결론에 대해 기술한다.

본 연구는 2018년도 한국과학기술정보연구원(KISTI) 주요사업 과제로 수행한 것입니다.

2. 관련 연구

단일 문서에 대한 제목 추출 연구는 학술 논문, 뉴스 문서 등과 같이 문서와 제목이 1:1로 존재하는 경우 학습 데이터 구축이 용이하기 때문에 지도학습 기반의 연구가 많이 진행되고 있다[1]. 하지만 다중 문서 제목 추출 연구의 경우 학습 데이터 구축이 비교적 어렵다는 한계점이 있다. 따라서, 존재하는 대부분의 다중 문서 제목 추출에 대한 연구는 [2,6]와 같이 군집화 기법을 이용하여 군집된 각 문서 집합에 대한 제목을 추출하기 위해 군집 정보를 사용하는 연구가 대부분이다.

이와 같은 연구들은 단일 문서에서 키워드를 추출하는 [3,7]과 같은 키워드 추출 연구와 비슷하게 다중 문서 집합을 하나의 문서로 보고, 제목 후보를 선택한 뒤 각 후보의 스코어를 계산하여 스코어가 가장 높은 후보를 제목으로 선택한다. [3,7]에서는 키워드 추출에서 단어의 가중치를 계산하기 위해 웹 페이지의 가중치를 계산하는 PageRank[8] 알고리즘 기반 방법의 TextRank[3], ExpandRank[7]를 제안하였으며, 키워드 후보의 단위를 명사와 형용사의 조합으로 정의하였다. 이와 비슷하게 [2]에서는 단어 관련성 계산을 위해 LDA의 토픽 정보를 이용하고, 복합 명사를 제목으로 추출하였다. 하지만 이러한 연구들은 최종 결과가 단순한 명사의 나열이기 때문에 추출된 제목이 자연스럽지 못하다는 단점이 있으며, [2]와 같이 추가적인 토픽 정보가 없다면 [3,7]과 마찬가지로 일반적인 주변 단어와의 관계 정보에만 의존하는 등 전체 주제와 관련된 직접적인 정보를 사용할 수 없다.

따라서 본 논문에서는 전제적인 문서 집합에 대한 주제 정보를 사용하기위해 요약 기반의 단어 관련성 추정 기법을 제안하고, 자연스러운 제목 추출을 위해 BPE 기반 키워드 추출 및 요약 기반 제목 후보 선택 방법을 제안한다.

3. 다중 뉴스 기사 제목 추출

본 논문에서는 주제가 같은 다중 뉴스 기사에 대한 제목을 추출하기 위해서 다중 문서 요약 기반 제목 추출 시스템을 제안한다. 제안 시스템은 크게 다섯 단계로 다중 문서 요약 단계, 요약 기반 단어 관련성 추정 단계, BPE 기반의 중요 키워드 선택 단계, 추정된 단어 관련성과 선택된 중요 키워드를 이용한 제목 추출 단계, 그리고 후처리 단계로 나뉜다. 전체적인 시스템 구성도는 [그림 1]과 같다.

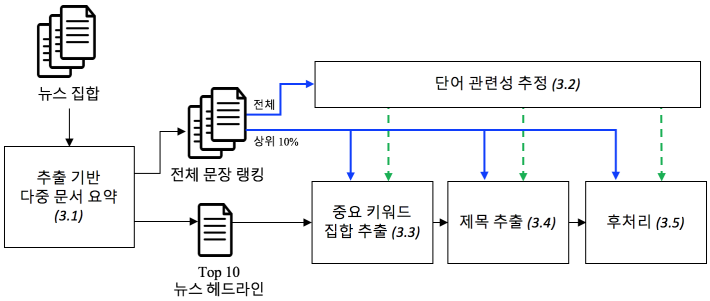

그림 1. 다중 뉴스 기사 제목 추출 시스템 구성도

3.1. 다중 문서 요약

가장 먼저 수행되는 다중 문서 요약 단계에서는 대량의 뉴스 기사가 입력되면 추출 기반의 다중 문서 요약을 시행한다. 본 논문에서는 [5]에서 제안된 추출 기반 요약 기법을 사용하였으며, 요약의 결과로 사용하는 것은 크게 두 가지로 전체 문장 랭킹과 Top 10 뉴스 헤드라인이다. 먼저, 전체 문장 랭킹은 추출기반 요약에서 사용되는 문장 스코어를 통해 전체 문장을 순위화한 것을 말한다. 이 전체 문장 랭킹을 이용하여 제목 추출을 위한 단어 관련성를 추정하고, 상위 10% 문장을 이용하여 시스템 전반적인 단계에서 사용한다. 다음으로, Top 10 뉴스 헤드라인은 전체 문장 랭킹에서 순위가 높은 문장이 포함되어 있는 뉴스 기사에 대한 헤드라인을 순차적으로 10개를 추출한 것이다. 이 뉴스 헤드라인은 중요 키워드 집합 추출 단계에서 사용된다.

3.2. 단어 관련성 추정

단어 관련성 추정은 각 단어에 대해 뉴스 집합의 전체적인 주제와 얼마나 관련 있는지 추정하는 것을 말한다. 단어에 대한 주제 관련성 추정을 위해 다중 문서 요약 단계의 결과인 전체 문장 랭킹 상위 30%를 주제 관련 문장으로 보고, 나머지 하위 70%를 비관련 문장으로 하여 각 집합에서의 빈도 계산을 통해 단어의 중요도를 측정한다. 단어 관련성 스코어 TR_d는 다음과 같이 계산된다.

$$TR_d(w) = \log\frac{p_d \times (1-q_d)}{(1-p_d)\times q_d} = \log\frac{(r_d+0.5)(S_d-s_d+0.5)}{(R_d-r_d+0.5)(s_d+0.5)}$$

p_d와 q_d는 전체 문장 집합 d에서 주제 관련 문장과 비관련 문장에서 단어 w가 나타날 확률을 말하며, R_d와 S_d는 전체 문장 집합 d에서 주제 관련 문장 수와 비관련 문장의 수를 나타낸다. 그리고 r_d와 s_d는 전체 문장 집합 d에서 단어 w를 포함하고 있는 주제 관련 문장의 수와 비관련 문장의 수를 말한다.

3.3. 중요 키워드 선택

중요 키워드 선택 단계에서는 다중 문서 요약의 결과인 전체 문장 랭킹 상위 10% 문장과 Top 10 뉴스 헤드라

인을 이용하여 Byte Pair Encoding (BPE) 기반으로 제목 추출에 중요한 키워드를 추출한다. 중요 키워드 선택 단계는 총 세 가지 하위 단계로 체언 시퀀스 기반 BPE를 이용한 중요 키워드 후보 추출 단계, 중요 키워드 후보 순위화 단계, Top 10 뉴스 헤드라인 기반 중요 키워드 선택 단계로 이루어져있다.

3.3.1. BPE 기반 중요 키워드 후보 추출

먼저 기본적인 BPE는 [4]에서 단어를 하위 단위로 분할하는 단어 분할(Word Segmentation) 문제를 위해 [9]의 알고리즘을 변형하여 제안한 것으로, 적용하고자 하는 전체 데이터의 모든 단어를 최하위 문자 단위로 나누어 나열한 다음 가장 빈도가 높은 쌍을 반복적으로 결합하는 기법이다.

[표 1]은 전체 데이터에서 중복을 포함한 모든 단어 집합이 [‘low’ , ‘low’ , ‘lowest’ , ‘newest’] 로 4개 일 때, 3번 반복한 BPE 알고리즘의 각 단계별 예시이다. 먼저, 초기 단계에서 모든 단어를 최하위 단위로 모두 나눈 다음 단어 의 끝에 ‘*’ 심볼을 추가한다. 다음으로, 단계 1에서 ‘l’ 와 ‘o’ 쌍의 빈도가 가장 높기 때문에 ‘lo’ 이 새 심볼로 선택 되었으며, 모든 단어에서 결합되어 표현되었다. 이후 단계에서도 단계 1과 똑같은 방법이 반복되며 순차적으로 새 심볼이 선택된다.

표 1. 단어 분할에서 BPE 알고리즘 예시

단계	전체 단어 집합	새 심볼
초기	‘l o w *’, ‘ l o w *’, ‘l o w e s t *’, ‘n e w e s t *’	-
1	**‘lo** w *’, ‘ **lo** w *’, **‘lo** w e s t *’, ‘n e w e s t *’	‘lo’
2	**‘low** *’, ‘ **low** *’, **‘low** e s t *’, ‘n e w e s t *’	‘low’
3	‘low *’, ‘ low *’, ‘low e s **t***’, ‘n e w e s **t***’	‘t*’

위와 같은 BPE 알고리즘을 키워드 추출에 적용하기 위하여 알고리즘이 수행되는 최하위 단위를 영어의 문자(Character)에서 한국어의 형태소로 변경하고, 단어 단위는 영어의 단어(Word)에서 한국어의 체언의 시퀀스로 변경한다. 체언의 시퀀스는 각 문장 내에서 품사가 체언[1] 계열인 형태소의 최대 시퀀스를 말한다. 예를 들어, ’ 김철수 학생의 성적’ 에 대한 형태소 분석 결과가 ‘김철수/NNP 학생/NNG 의/JKG 성적/NNG’ 이라면 체언의 시퀀스 집합은 다음과 같다.

- 원본: ‘김철수 학생의 성적’
- 형태소 분석: ‘김철수/NNP 학생/NNG 의/JKG 성적/NNG’
- 체언의 시퀀스: ‘김철수 학생’ , ‘성적’

1) 일반명사, 고유명사, 의존명사, 수사, 대명사를 포함한다.

따라서, 중요 키워드 선택을 위해 문장 랭킹 상위 10% 문장에서 중복을 포함하는 체언의 시퀀스 집합을 추출하고, BPE 알고리즘을 수행한 뒤 새로 생성된 심볼의 집합을 키워드 후보로 선택한다. 체언의 시퀀스 집합이 [성적, 김철수 학생 성적, 김영희 학생 성적] 일 때, BPE 알고리즘의 2 단계까지의 결과는 [표 2]와 같으며, 추출된 키워드 후보 집합은 [(성적), (학생, 성적)] 이다.

표 2. BPE를 이용한 키워드 후보 추출 예시

단계	전체 단어 집합	새 심볼
초기	‘성적 *’, ‘ 김철수 학생 성적 *’, ‘김영희 학생 성적 *’,	-
1	**‘성적*’**, ‘ 김철수 학생 **성적*’**, ‘김영희 학생 **성적*’**,	‘성적*’
2	‘성적*’, ‘ 김철수 **학생성적*’**, ‘김영희 **학생성적*’**,	‘학생성적*’

3.3.2. 중요 키워드 후보 순위화

이번 단계에서는 중요 키워드 후보를 순위화하고, 하위 50%의 후보는 중요 키워드 후보에서 제거한다. 중요 키워드 후보 순위화를 위한 스코어는 다음과 같이 계산된다.

$$Score_{bpe}(key) = Step(key) + H\text{-}TR_d(key) + Len(key),$$

$$Step(key) = 1.0 - \frac{i-1}{N}, Len(key) = 1.0 - \frac{M-1}{L},$$

$$H\text{-}TR_d(key) = \frac{M}{\sum_{k \in key} \frac{1}{TR_d(k)}}$$

$Score_{bpe}(key)$는 총 세 가지 정보를 반영하여 스코어를 계산한다. 첫 번째로 BPE 알고리즘에서 키워드 후보 key의 선택 단계 정보를 반영하기 위한 $Step(key)$, 두 번째로 키워드 후보 내에 존재하는 형태소들의 단어 관련성 스코어의 조화 평균 $H\text{-}TR_d(key)$, 그리고 마지막으로 키워드 후보 내의 형태소 개수 정보를 반영한 스코어 $Len(key)$를 이용한다. 위의 수식에서 i는 BPE에서 key가 생성된 단계, N은 전체 키워드 후보 개수, M은 키워드 후보에 포함하고 있는 형태소 개수, L은 전체 키워드 후보 중 가장 많은 형태소를 포함하고 있는 키워드 후보의 형태소의 개수이다.

3.3.3. 헤드라인 기반 중요 키워드 선택

Top 10 뉴스 헤드라인은 전체 뉴스 문서 집합의 주제에 대해 중요하고 간결한 내용을 담고 있기 때문에 제목에 대한 중요 키워드를 추출하기 위해서 매우 중요한 정보가 될 수 있다. 이번 단계에서는 추출된 키워드 후보 집합과 Top 10 뉴스 헤드라인 집합의 패턴 매칭을 통해 최종 중요 키워드를 선택한다. 즉, 키워드 후보 집합에 존재하는 키워드 후보 중 Top 10 뉴스 헤드라인에서 한번

이라도 출현 했다면 최종 중요 키워드 집합에 포함시킨다.

3.4. 중요 키워드 기반의 제목 추출

최종 제목 추출 단계에서는 단순한 키워드 나열이 아닌 자연스러운 제목을 추출하기 위하여 요약의 결과인 상위 10%의 문장 집합과 중요 키워드 집합을 매핑하여 최종 제목 후보를 추출한다. 이때, 매핑된 단어 사이에 존재하는 형태소들의 품사 집합에 조사가 1개 이하라면, 매핑된 단어 사이의 형태소들을 모두 연결하여 제목 후보로 추출한다. 예를 들어, 중요 키워드 집합이 [(학생, 성적), (학생, 과제, 성적), (김영희)] 일 때, 키워드 매핑 결과와 제목 후보 선택 결과 예시는 [표 3]과 같다.

표 3. 중요 키워드 기반 제목 추출 예시

번호	요약 결과 문장	선택
1	**학생들의 과제에 대한 성적**은 조교가 관리한다.	O
2	**학생들의 시험에 대한 성적**은 조교가 관리한다.	X
3	김철수 **학생의 성적**은 좋다.	O
4	**김영희** 학생 성적 또한 좋다.	O

예시의 2번 문장에서 매핑된 '학생들의 시험에 대한 성적' 이 제목으로 선택 되지 않는 이유는 키워드 (학생, 성적) 사이에 조사가 '의' 와 '에' 로 2개 이상이기 때문이며, 1번 문장에서 매핑된 '학생들의 과제에 대한 성적' 이 제목으로 선택된 이유는 (학생, 과제, 성적) 각 사이에 조사가 1개 이하이기 때문이다. 마지막으로, 추출된 제목 후보 중 최종 제목을 선택하기 위한 스코어는 다음과 같이 계산된다.

$$Score_{final}(key) = H-TR_d(key) + Len(key) + H-Josa(key),$$

$$Len(key) = 1.0 - \frac{M-1}{L},\ H-TR_d(key) = \frac{M}{\sum_{k \in key} \frac{1}{TR_d(k)}}$$

스코어 계산을 위한 $Len(key)$과 $H-TR_d(key)$는 3.3.2 장에서 사용된 것과 같다. $H-Josa(key)$는 제목 후보 내에 존재하는 조사들에 대한 스코어를 모두 계산한 뒤 조화 평균한 것을 말한다. 조사의 스코어는 Top 10 뉴스 헤드라인에서 어떤 조사들이 많이 출현했는지에 대한 빈도 정보를 반영한 결과를 사용한다.

3.5. 후처리

후처리 단계는 최종 추출된 제목이 매우 짧은 경우 정보량이 부족하다고 판단하여, 요약의 결과인 상위 10%의 문장에서 추가적인 정보를 간단하게 추출하여 추출된 제목에 추가하는 단계이다. 즉, 추출된 제목이 2개 이하의 명사로만 이루어진 경우 후처리 단계가 시행되며, 이때, 상위 10%의 문장에서 추출된 제목의 위치를 찾고, 그 뒤에 추가적인 명사가 존재하면 최대 길이로 연결한다. 예를 들어, 표 3에서 선택된 3개의 제목 후보 중 최종 제목으로 '김영희' 가 선택 된다면, '김영희' 뒤에 존재하는 최대 명사 시퀀스 '학생', '성적' 을 모두 연결하여 '김영희 학생 성적' 으로 제목을 추출한다.

4. 실험 및 평가

제목 추출 결과에 대한 평가를 위해 자동으로 군집된 총 12개의 주제에 대한 다중 뉴스 기사 집합을 사용하였으며, 관련 전문 교육을 받은 연구원 5명이 해당 다중 문서 집합에 대한 제목으로 적절한지에 대한 정성 평가를 진행하였다. 본 논문에서 정성평가는 1점부터 5점까지의 점수를 다음과 같은 기준으로 평가하였고, 평가 결과는 [표 4]와 같으며, 다중 문서 제목 추출을 위한 제안 시스템에서 추출된 각 주제별 제목은 [표 5]와 같다.

1: 주제가 완전히 틀린 제목
2: 주제가 정확하지는 않지만 비슷한 제목
3: 주제는 정확하나 정보량이 거의 없거나 너무 많은 제목
4: 주제는 정확하나 정보량이 부족하거나 조금 많은 제목
5: 주제가 정확하며 정보량이 충분히 많은 제목

표 4. 각 주제별 다중 문서 제목 추출 정성 평가 결과

Topic #	평가자별 정성 평가 결과					평균
	A	B	C	D	E	
1	3	3	3	3	3	3.00
2	4	5	4	5	5	4.60
3	2	2	2	4	3	2.60
4	5	4	4	5	5	4.60
5	4	5	4	5	4	4.40
6	3	2	3	3	3	2.80
7	3	5	1	4	4	3.40
8	5	4	5	5	4	4.60
9	5	5	4	4	5	4.60
10	3	4	3	4	3	3.40
11	2	4	2	4	3	3.00
12	3	5	1	4	3	3.20
평균	**3.50**	**4.00**	**3.00**	**4.16**	**3.75**	**3.68**

표 5. 각 주제별 다중 문서 제목 추출 결과

Topic #	추출된 다중 문서 제목
1	밥 딜런
2	백남기 씨 시신에 대한 부검
3	부산 지하철 2호선 시립 미술관역
4	갤럭시 노트 7 단종 사태 이후 삼성전자
5	태풍 '차바'로 큰 피해를 입은 지역을 특별 재난 지역으로 선포
6	대통령 탄핵
7	트럼프가 당선
8	대통령의 퇴진을 요구하는 시국선언
9	서울 도심에서 열리는 박근혜 정권 규탄 대규모 촛불 집회
10	대통령의 퇴진을 요구
11	최순실 결탁 의혹
12	최순실 관계 입증

5. 결론

본 논문에서는 다중 문서 제목 추출에서의 한계점을 해결하고자 단어 관련성 추정과 Byte Pair Encoding을 통한 요약기반의 제목 추출 방법을 제안하였다. 먼저, 요약 기반의 단어 관련성 추정을 통해 단어의 스코어가 다중 문서 집합의 전체적인 주제에 대한 정보를 반영하도록 하였으며, Byte Pair Encoding를 이용하여 중요한 키워드를 추출하는 새로운 방법을 제안하였고, 요약 문장 기반의 중요 키워드 매핑을 통해 자연스러운 제목 추출이 가능하도록 하였다. 평가를 위해서는 자동으로 군집된 총 12개의 주제에 대한 다중 뉴스 기사 집합을 사용하였으며 전문 교육을 받은 연구원들이 정성평가를 진행하여 5점 만점 기준 평균 3.68점을 얻었다.

참고문헌

[1] 이현구, 김학수, “주의집중 및 복사 작용을 가진 Sequence-to-Sequence 순환신경망을 이용한 제목 생성 모델”, *정보과학회논문지*, Vol. 44, No. 7, pp. 674-679, 2017.

[2] 한규열, 안영민, “LDA로 형성된 한국어 문서 클러스터의 자동 제목 생성”, *한국정보과학회 학술발표 논문집*, pp. 616-618, 2013.

[3] Rada Mihalcea and Paul Tarau, “TextRank: Bringing order into text”, In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404-411, 2004.

[4] Rico Sennrich, Barry Haddow, Alexandra Birch, “Neural Machine Translation of Rare Words with Subword Units”, In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715-1725, 2016.

[5] 유홍연, 이승우, 고영중, “실시간 이슈 분석을 위한 뉴스 군집화 및 다중 문서 요약”, *제30회 한글 및 한국어 정보처리 학술발표 논문집*, 2018.

[6] 김태현, 맹성현, “계층구조를 이용한 문서 클러스터 제목의 자동생성”, *한국정보과학회 언어공학연구회 학술발표 논문집*, pp. 163-170, 2001.

[7] Xiaojun Wan and Jianguo Xiao, “Single Document Keyphrase Extraction using Neighborhood knowledge”, In *Proceedings of the 2008 American Association for Artificial Intelligence*, pages 855-860, 2008.

[8] Lawrence Page, Sergey Brin, Rajeev Motwani, andTerry Winograd, “The Pagerank Citation Ranking: Bringing Order to the Web”, *Technical report, Standford Digital Library Technologies Project*, 1998.

[9] Philip Gage. “A New Algorithm for Data Compression”, *C Users Journal*, Vol. 12, No. 2, pp.23-38, 1994.