단어의 공기정보를 이용한 클러스터 기반 다중문서 요약

(Multi-document Summarization Based on Cluster using Term Co-occurrence)

이 일 주 † 김 민 구 ††
(Iljoo Lee) (Minkoo Kim)

요 약 대표문장 추출에 의한 다중문서 요약에서는 비슷한 정보가 여러 문서에서 반복적으로 나타나는 정보의 중복문제에 대해 문장의 유사성과 차이점을 고려하여 이를 해결할 수 있는 효율적인 방법이 필요하다. 본 논문에서는 단어의 공기정보에 의한 관련단어 클러스터링 기법을 이용하여 문장의 중복성을 제거하고 중요문장을 추출하는 다중문서 요약을 제안한다. 관련단어 클러스터링 기법에서는 각 단어들은 서로 독립적으로 존재하는 것이 아니라 서로 간에 의미적으로 연관되어 있다고 보며 주제별 문장클러스터단위의 단어 연관성(cohesion)을 이용한다. 평가용 실험문서인 DUC(Document Understanding Conferences) 데이타를 이용하여 실험한 결과 본 논문에서 제안한 문장클러스터단위의 단어 공기정보를 이용한 방법이 단순 통계정보와 문서단위 단어 공기정보, 문장단위 단어 공기정보에 의한 다중문서 요약에 비해 좋은 결과를 보였다.

키워드 : 다중문서 요약, 공기정보, 연관성

Abstract In multi-document summarization by means of salient sentence extraction, it is important to remove redundant information. In the removal process, the similarities and differences of sentences are considered. In this paper, we propose a method for multi-document summarization which extracts salient sentences without having redundant sentences by way of cohesive term clustering method that utilizes co-occurrence information. In the cohesive term clustering method, we assume that each term does not exist independently, but rather it is related to each other in meanings. To find the relations between terms, we cluster sentences according to topics and use the co-occurrence information of terms in the same topic. We conduct experimental tests with the DUC(Document Understanding Conferences) data. In the tests, our method shows better performance of summarization than other summarization methods which use term co-occurrence information based on term cohesion of document or sentence unit, and simple statistical information.

Key words : multi-document summarization, co-occurrence information, cohesion

1. 서 론

인터넷 등의 발전으로 정보가 대량으로 증가하고 있는 상황에서 사용자들은 정보를 효율적으로 관리하고 검색하는 문제가 중요하게 되었다. 이러한 문제에 대해 효과적인 해결책을 제시해줄 수 있는 방법으로 문서 요약이 요구된다. 문서 요약이란 문서의 기본적인 내용을 유지하면서 원문으로부터 가장 의미 있는 내용만을 추려내어 문서의 길이를 줄이는 작업을 의미한다[1]. 문서 요약은 요약대상의 수에 따라 단일문서 요약과 다중문서 요약으로 구분할 수 있다. 단일문서 요약은 한 문서의 내용을 요약하는 경우에 해당하며 다중문서 요약은 유사내용을 다루고 있는 문서 집합이나 연속적으로 입력되는 문서들을 요약한다. 최근의 요약시스템은 단일문서 요약에서 다중문서 요약으로 점차 발전하고 있는 추세이다. 이것은 대량의 정보 속에서 어떤 정보집합이 무엇에 관한 내용인지 판단할 수 있는 능력과 유사한 정보가 반복되는 중복정보(redundant information)를 제거할 수 있는 도구의 필요성이 요구되고 있기 때문이다.

† 정 회 원 : 동원대학 모바일컨텐츠과 교수
ijlee@tongwon.ac.kr
†† 종신회원 : 아주대학교 컴퓨터공학과 교수
minkoo@ajou.ac.kr
논문접수 : 2005년 7월 15일
심사완료 : 2005년 12월 14일

따라서 문서 간의 유사성과 차이점을 파악하여 중복정보 문제를 해결할 수 있는 다중문서 요약시스템의 개발이 필요하다[2].

본 논문은 단어의 공기정보를 이용한 관련단어 클러스터에 기반하여 중요문장을 추출하는 다중문서 요약시스템을 제안한다. 주제별로 관련단어들이 정확하게 분리될 수 있도록 문장클러스터단위의 공기정보를 기반으로 단어의 연관관계를 분석하여 대표단어 클러스터를 생성한다. 클러스터링된 단어들은 주제별로 다중문서의 문장들을 분리하는데 사용되며, 최종적으로 문장별 가중치에 의한 클러스터별 대표문장을 추출하여 요약문을 완성한다. 본 논문에서 제안한 방법은, 문서별로 공기정보를 이용하는 경우 주제분리가 정확히 되지 않아 서로 다른 내용의 문장들이 동일 클러스터를 형성하는 현상에 의해 클러스터별 대표문장 선정 시 중요한 문장이 탈락될 가능성을 줄인다. 또한 문장별 공기정보 이용 시 충분히 단어의 연관정보를 반영하지 못해 동일주제의 문장들이 서로 다른 클러스터를 형성하여 전체 요약성능이 떨어지는 현상을 극복하여 다중문서 요약에서 높은 성능향상을 낼 수 있다.

본 논문은 다음과 같이 구성된다. 2장에서 기존의 관련 연구에 대해 소개하며, 3장에서 관련단어 클러스터링을 위한 연관규칙에 대해 기술한다. 4장에서 본 논문에서 제안한 다중문서 요약을 설명하고, 5장에서 제안된 시스템의 성능을 평가한다. 그리고 6장에서 결론을 맺는다.

2. 관련 연구

문서 요약에 관한 관련 연구를 살펴보면 접근방법에 따라 크게 언어학적 방법과 통계적 방법으로 나눌 수 있다. 언어학적 방법에서는 문장 간의 구조적 정보에 의한 담화트리(discourse tree) 및 수사구조트리(rhetorical structure tree), 단어 간의 의미관계에 의한 단어사슬(lexical chain) 등을 이용하여 문장들을 순위화한 후 요약문을 생성한다. 통계적 방법에서는 요약하고자 하는 문서에 나타난 단어의 빈도, 제목, 위치, 단서어 등의 통계적 자질을 이용하여 중요문장들을 선정한다. 통계적 정보에 기반한 문서 요약에서는 요약을 하기 위한 대상이 되는 문서를 전통적인 "Bag of Words" 가정에 기반하여 문서를 표현하게 된다. 이 경우 각 단어들은 서로 독립적으로 존재하게 되므로 원래 문서가 가지고 있는 정보의 의미를 손실한다는 단점이 있다[3]. 이 단점을 해결하기 위해 문서에 출현한 단어들을 독립적으로 보는 것이 아니라 서로 간에 의미적으로 연관되어 있다고 보고, 단어 간의 연관성(cohesion)을 사전적 의미로 구성한 영문 시소러스 WordNet에 기반한 문서 요약 방법이 있다[4]. WordNet을 사용하는 경우의 문제점은 사전에 등록되지 않은 단어에 대해서는 정의를 할 수 없다는 것이다. 따라서 고유명사 및 신조어, 전문용어가 많이 사용된 문서를 대상으로 연관관계를 추출하고자 할 때 오류가 발생한다[5].

단어 간의 연관성을 이용하여 문서를 표현하기 위해 WordNet과 같은 별도의 사전을 사용하지 않고 단어의 공기정보(Co-occurrence)를 이용하는 방법이 있다. 공기정보란 두 단어가 동일문서, 문장, 구 등에 같이 발생하는 현상을 말하며, 더 자주 발생할수록 두 단어가 밀접한 관계를 가지고 있다는 전제에 기반하고 있다[6]. Salton은 단어의 공기정보를 이용하여 문서집합 내의 단어들이 얼마만큼 서로 연관이 있는가에 대한 측정방법을 제시하였다[7]. 이 방법은 문서 내에서 두 단어의 동시출현 회수를 연관 값으로 계산한 후, 질의어 확장에 사용하기 위해 연관 값에 의한 유사한 관련관계를 갖는 단어들끼리 클러스터링한다. 그러나 문서단위별 관련단어 클러스터링 방법은 중요한 문제점을 가지고 있다. 문서는 몇 가지 서로 다른 주제내용을 포함할 수도 있는데, 문서단위로 공기정보를 계산하면 단어들이 서로 다른 주제영역에 포함되어 있는 경우에도 관련성이 있다고 판단하는 경우가 발생할 수 있다. 단어의 공기정보 계산 시 문장단위별 연관성분석도 고려할 수 있으나 이 경우 지나치게 많은 주제로 세분화되어 정확한 공기정보를 반영하지 못하게 된다. 즉, 동일주제에 해당되는 단어들이 서로 다른 주제영역으로 분리되는, 문서단위별 계산과는 정반대의 오류현상을 보일 수 있다. 따라서 다중문서 환경에서는 문장단위별 공기정보에 의한 요약은 큰 효과를 기대할 수 없다.

3. 단어 공기정보의 연관규칙

문서에서 대표단어들을 추출하는 경우 대부분 단어의 출현빈도와 역 문헌빈도에 의존하는 경우가 많다. 단어들 사이의 독립적 통계정보에 의해 선정된 대표단어들은 문서내용과 상관없이 일상적으로 자주 사용되는 단어나 단순히 출현빈도만 높은 단어가 추출되는 오류가 발생할 가능성이 높다. 문서의 단어들은 문서가 작성된 의도에 따라 서로 의미적으로 연관된 형태로 형성된다는 점에서, 출현횟수와 공기정보를 동시에 활용하여 대표단어들을 추출한다. 그리고 서로 관련 있는 단어들끼리 클러스터로 구성한다면 이 클러스터 단어들은 하위주제별로 문서의 의미를 나타낸다고 할 수 있다. 단어 공기정보의 연관성 분석은 하나의 문서나 문단, 문장, 절, 구에 포함되어 있는 단어들의 관련성을 파악하기 위하여 둘 이상의 단어들로 구성된 연관성 규칙을 이용하는 탐색적 자료 분석방법이다. 본 논문에서는 문장클러

스터별로 두 단어 간의 연관성 정도를 측정하여 연관성이 많은 단어들을 클러스터링한다. 표 1은 문장클러스터별 출현단어의 행렬구성이고, 표 2는 단어 간 동시발생 상관표이다.

- 연관성 규칙 : 어떤 단어의 존재가 다른 단어의 존재를 암시하는 것을 의미
  (단어 A) → (단어 B)
  (if A then B : 만일A가 일어난다면 B가 일어난다.)

표 1 문장클러스터-단어 행렬

문장클러스터 / 단어	C_1	C_2	⋯	C_n
T_1	tf_{11}	tf_{12}	⋯	tf_{1n}
T_2	tf_{21}	tf_{22}	⋯	tf_{2n}
⋮	⋮	⋮		⋮
T_m	tf_{m1}	tf_{m2}	⋯	tf_{mn}

표 2 단어 간 동시발생 상관표

단어 / 단어	T_1	T_2	⋯	T_m
T_1	-	cf_{12}	⋯	cf_{1m}
T_2	cf_{21}	-	⋯	cf_{2m}
⋮	⋮	⋮	-	⋮
T_m	cf_{m1}	cf_{m2}	⋯	-

n: 문장클러스터의 개수
m: 단어의 개수
$C_i(i=1,\cdots,n)$: i번째 문장클러스터
$T_j(j=1,\cdots,m)$: j번째 단어
tf_{ij}: j번째 문장클러스터에 나타난 i번째 단어의 빈도수
$cf_{ij}(i,j=1,\cdots,m)$: i번째 단어와 j번째 단어가 동일한 문장클러스터에서 동시에 나타난 빈도수

단어와 단어 사이의 의미적 관계를 찾아내는 문제는 두 단어 T_i와 T_j가 주어졌을 때 두 단어 사이의 성립 가능한 관계를 찾는 작업으로 정의한다. 동시출현 단어의 연관도는 공기빈도를 직접 사용할 수도 있고, 자카드 계수(Jaccard coefficient), 카이제곱 통계량(x^2 statistic), 상호정보량(Mutual information) 등을 이용하여 정규화할 수도 있다. 본 논문에서는 지지도 및 신뢰도, 향상도를 이용하여 의미관계의 성립여부를 정의한다.

지지도(support)는 유효한 통계적 추론을 위해 필요한 실질적인 인스탄스의 수로서, 두 단어가 의미 있는 공기정보를 갖기 위해 전체 문장클러스터 중 두 단어의 동시출현 경우의 수가 식 (1)과 같이 일정수준 이상이여야 함을 의미한다. 지지도 계산 시 최소지지도를 설정할 수 있는데, 최소지지도는 적어도 어느 정도 이상의 출현수가 일어난 경우들만을 고려대상으로 삼는 것으로서 데이타의 상황에 따라 각기 다른 최소지지도가 선택될 수 있다. 최소지지도를 사용하는 이유는 전체 단어집합에 대하여 관심 있을 정도로 빈발하게 나타나는 단어만을 고려하기 위함이다. 최소지지도를 지지도의 의미로 사용하기 위해 식 (2)와 같이 두 단어의 동시출현 문장클러스터 수가 3인 경우는 지지도의 의미가 있다고 전제한다. 동시출현 문장클러스터 수가 2이하 인 경우에는 우연히 공기한 경우로서 의미가 없다고 가정한다. 이것은 출현빈도가 낮으면서도 문서의 핵심이 될 수 있는 단어들이 대표단어로 선정될 수 있는 기회를 부여하기 위해서이다.

- 지지도

$$Support(T_i, T_j) = \frac{\text{두 단어 } T_i\text{와 } T_j\text{가 동시 발생한 문장클러스터 수}}{\text{전체 문장클러스터 수}} \geq \text{임계값} \quad (1)$$

- 최소지지도

$$MinSupport(T_i, T_j) = P(T_i \cap T_j) \geq 3 \quad (2)$$
(두 단어 T_i와 T_j가 동시 발생한 문장클러스터 수)

신뢰도(Confidence)는 단어T_i를 포함하는 전체 문장클러스터 중에서 단어T_j가 포함된 문장클러스터의 비율을 말하며 조건부 확률 $P(T_i \mid T_j)$를 이용하여 신뢰도의 정도를 계산한다. 신뢰도 $Confidence(T_i \mid T_j)$는 단어T_j가 단어T_i와의 공기빈도수가 많으면 값이 커지므로 단어T_j가 단어T_i에 얼마나 종속적인가를 나타낸다. 반대로 $Confidence(T_j \mid T_i)$는 단어T_i의 단어T_j에 대한 의존도를 나타낸다. 식 (3)과 식 (4)의 두 가지 조건부 확률은 두 단어의 빈도수 크기에 따라 신뢰도가 서로 다르다는 점을 의미한다.

$$Confidence(T_i \mid T_j) = P(T_i \mid T_j) = \frac{P(T_i \cap T_j)}{P(T_j)} = \frac{\text{단어 } T_i\text{와 단어 } T_j\text{가 동시에 포함된 문장클러스터 수}}{\text{단어 } T_j\text{가 포함된 문장클러스터 수}} \quad (3)$$

$$Confidence(T_j \mid T_i) = P(T_j \mid T_i) = \frac{P(T_j \cap T_i)}{P(T_i)} = \frac{\text{단어 } T_j\text{와 단어 } T_i\text{가 동시에 포함된 문장클러스터 수}}{\text{단어 } T_j\text{가 포함된 문장클러스터 수}} \quad (4)$$

향상도(Lift/Improvement)는 단어T_i에 상관없는 단어T_j의 확률 대비, 단어T_i가 주어졌을 때 단어T_j의 확률의 증가비율로 이 값이 클수록 단어T_i의 출현여부가 단어T_j의 출현여부에 큰 영향을 미친다. 단어T_i와 단어T_j의 출현여부가 상호 관련이 없다면 $P(T_j \mid T_i)$와 $P(T_j)$가

같게 되어 향상도가 1이 된다. 향상도가 1보다 크다면 단어의 연관성을 예측하는데 우연적 기회보다 우수하고, 향상도가 1보다 작다면 연관성을 예측하는데 우연적 기회보다 나쁘다는 것을 의미한다. 반대로 단어T_j에 상관없는 단어T_i의 확률 대비, 단어T_j가 주어졌을 때 단어T_i의 확률의 증가비율 역시 값이 클수록 단어T_j의 출현여부가 단어T_i에 대해 큰 영향력을 나타낸다. 신뢰도와 마찬가지로 식 (5)와 식 (6)의 두 가지 조건부 확률은 두 단어의 빈도수 크기에 따라 향상도가 다르다는 점을 의미하며 두 가지 조건부 확률이 모두 1 보다 커야 두 단어의 공기정보가 의미 있다고 판단한다.

$$Lift(T_i \mid T_j) = \frac{P(T_i \bigcap T_j)}{P(T_i) \cdot P(T_j)} = \frac{P(T_j \mid T_i)}{P(T_j)} = \frac{Confidence(T_i \mid T_j)}{P(T_j)} \quad (5)$$

$$Lift(T_j \mid T_i) = \frac{P(T_j \bigcap T_i)}{P(T_j) \cdot P(T_i)} = \frac{P(T_i \mid T_j)}{P(T_i)} = \frac{Confidence(T_j \mid T_i)}{P(T_i)} \quad (6)$$

표 3 향상도

향상도	의미
1	두 단어가 독립적
<1	두 단어가 음의 상관관계
>1	두 단어가 양의 상관관계

4. 다중문서 요약 시스템

본 논문에서 제안하는 다중문서 요약 시스템의 요약 대상은 문서집단의 주제적 특성에 독립적이며 관련단어 클러스터링에 의해 대표문장을 추출하는 요약 시스템이다. 그림 1과 같이 단일문서별로 모든 문장들을 하위주제에 의한 문장클러스터로 분리한 후 문장클러스터단위의 단어 연관성분석을 통해 대표단어 클러스터들을 생성한다. 이것은 문장 간의 유사도를 측정하여 주제별로 문장들을 클러스터링하는 방법 대신 단어들을 먼저 클러스터링한 다음 단어클러스터별로 유사문장들을 클러스터링하기 위함이다. 문서들은 몇 개의 서로 다른 주제들로 구성되어 있다고 가정하여 문서집합 내 문서평균 길이(문장 수)의 20%를 기준으로 클러스터의 개수를 결정한다. 즉, 클러스터의 개수가 하위주제의 개수이며 문서집합의 평균문장 수에 따라 클러스터의 개수가 변함을 의미한다. 이것은 일반적인 요약문서의 크기는 원문서의 1%에서 30%에 달하지만 추출요약의 경우에는 약 20% 정도는 되어야 원 문서의 의미를 전달할 수 있다[8,9]는데 기반한 것이다. 최종 요약문의 완성을 위해 클러스터 중요도에 의한 문장 중요도를 순위화하여 대표문장들을 추출한다. 대표문장 선정 시 단순히 가중치가 높은 값을 갖는 문장이 아닌 클러스터별 대표문장을 선정하는 이유는 문장의 중복성을 배제하고 문서의 흐름을 최대한 반영하기 위해서이다.

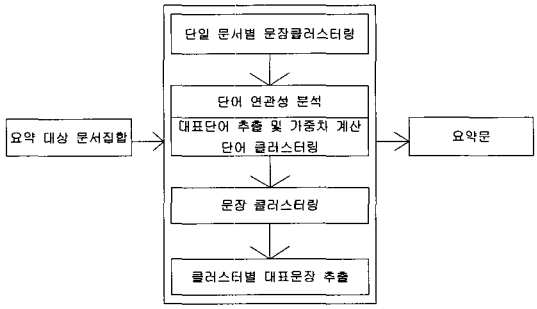

그림 1 다중문서 요약 시스템 구조

4.1 단일문서별 문장클러스터링

단어클러스터링을 위한 공기정보 계산단위인 문장클러스터를 생성하기 위해 우선 단일 문서별로 문장단위의 유사도를 측정하여 유사문장들의 클러스터링을 수행한다. 문장클러스터의 개수는 원문서의 내용을 전달하기 위한 최소비율인 문장길이(문장 수)의 20%를 기준으로 한다. 문장들은 식 (7)과 같이 단어의 가중치 값으로 표현 하였으며, 단어들의 가중치는 단어 출현빈도수를 전체 문장 수로 정규화 한 값인 단어빈도(tf) * 역 문장빈도(isf)로 표현한다. 두 문장 사이에 중복하여 출현한 단어들의 가중치 값을 이용한 코사인 유사도에 의해 식 (8)과 같이 문장 간 유사도 계산을 하여 주제별 문장클러스터를 생성한다. 표 4는 실험용 평가문서 중 "Hurricane Mitch"에 관한 문서집합 "d30002t"의 구성문서 "AWP19981027"에 대한 문장클러스터 구성 예이다. 문장 수는 30개이며 문장 클러스터의 개수는 6개이다.

• 단어 T_i의 가중치 값

$$Weight(T_i) = tf * isf \quad (7)$$

$isf = \log N/fS_i + 1$

N: 문서를 구성하는 전체 문장 수

fS_i : 단어T_i가 출현한 문장 수

• 문장 SA와 SB의 유사도

$$Sim(SA,SB) = \frac{\sum_{i=1}^{k}(SA_i) \cdot (SB_i)}{\sqrt{\sum_{i=1}^{k1}(SA_i)^2 \cdot \sum_{i=1}^{k2}(SB_i)^2}} \quad (8)$$

SA_i : 문장 SA의 단어 i에 대한 가중치

SB_i : 문장 SB의 단어 i에 대한 가중치

k : 문장 SA와 SB의 공통단어 개수
k_1 : 문장 SA의 전체단어 개수
k_2 : 문장 SB의 전체단어 개수

표 4 단일문서별 문장클러스터 예

문서집합	문서명	문장클러스터	
		클러스터 번호	관련 문장번호
d30002t	AWP19981027	1	2,4,5,8,26,28,30
		2	6,3,13
		3	1,16,27
		4	7,9,10,11,17,19,24,25
		5	12,18,20,21
		6	14,15,22,23,29

4.2 대표단어 추출 및 단어가중치 계산

전체 문서집합을 대표할 수 있는 대표단어는 공기정보에 의해 단어 간의 의미를 반영한 연관단어 형태로 추출된다. 공기정보의 계산은 두 단어 단위로 동일 문장 클러스터에서 동시 출현하는 것을 대상으로 한다. 단어의 공기정보를 이용하므로 단순히 단어의 출현횟수가 많은 단어라도 대표단어로 선정되지 못할 수 있으며 비록 출현횟수가 적더라도 대표단어로 선정될 가능성이 있다. 따라서 공기정보를 이용하면 단순 출현빈도 수가 많은 단어가 아닌, 문서 전개방향에 의해 유지되는 문서의 핵심 단어들을 대표단어로서 추출할 수 있다. 3장에서의 연관성 규칙을 적용하여 공기의 의미여부를 판단한 후 대표단어를 선정한다. 우선 공기정보를 갖는 단어쌍 들 중 식 (2)의 최소지지도를 갖지 않는 단어쌍들은 의미 있는 것으로 간주되지 않아 연관성 분석 대상에서 제외된다. 최소지지도를 만족하는 단어쌍들에 대해 식 (3)과 식 (4)의 신뢰도 계산을 한 후, 마지막으로 식 (5)와 식 (6)의 향상도 값이 모두 1보다 큰 단어쌍들을 두 단어의 공기정보가 의미가 있다고 판단하여 대표단어로 선정한다.

대표단어별 가중치는 각 노드에 인접한 노드들과의 의미유사도의 합인 도합유사도[10]로 표현한다. 본 논문에서 노드는 대표단어를 의미하며 연관성이 있다고 판단된 다른 대표단어와 많은 연결선을 가진 단어를 중요한 단어로 인식한다. 문장클러스터단위로 단어와 다른 단어들 사이의 연결개수를 식 (9)와 같이 의미유사도로 계산한 후 의미유사도 합인 식 (10)의 도합유사도 값으로 단어의 가중치를 표시한다. 단어의 연결개수를 그대로 사용하면 낮은 빈도의 연결수를 갖는 단어와 높은 연결수를 갖는 단어사이의 가중치가 과도한 차이를 보이며 이 현상을 막기 위해 로그를 취한 값을 사용한다. 예를 들어, "fuel"이라는 단어가 "food"와 2개, "relief"와 3개, "medicine"과 2개의 연결개수를 갖는다면, 이 단어의 가중치는 0.69와 1.09, 0.69의 합인 2.47이 된다.

• 두 단어 T_i와 T_j의 의미유사도

$$Sim(T_i, T_j) = \log(Frequency(T_i, T_j)) \quad (9)$$

$Frequency(T_i, T_j)$: 두 단어 T_i와 T_j의 연결개수 (동시 출현 문장클러스터 수)

• i 번째 단어의 도합유사도

$$asim(T_i) = \sum_{j=1}^{n} Sim(T_i, T_j) \quad (n: \text{연관단어의 개수, i} \neq \text{j}) \quad (10)$$

4.3 단어 클러스터링

공기정보 계산에 의해 선정된 대표단어들을 대상으로 관련단어 클러스터링을 수행하면 단어들은 각 클러스터에 속하게 된다. 관련단어 클러스터 생성 시에는 서로 직접적인 연관도가 높은 단어들로 클러스터를 형성하기 때문에 단어클러스터 사이의 주제분리가 명확하게 이루어진다. 단어클러스터의 개수는 문서집합 내의 각 단일 문서별로 생성된 문장클러스터 개수의 평균으로 한다. 클러스터 생성 시 클러스터의 크기(단어 개수)가 중요한 요소가 된다. 만약 클러스터의 크기 차이가 많게 되면, 각 문장들을 클러스터링하기 위한 조건이 공평하게 조성되지 않게 된다. 이 경우 많은 중요문장이 어느 한 클러스터에 집중하게 되면서 중요문장 대신 의미 없는 문장이 대표문장으로 선정될 가능성이 높아진다. 단어를 클러스터링하는 기법은 클러스터 센트로이드 간의 유클리드 거리를 최소화하는 방법에 의해 되도록이면 같은 크기로 클러스터를 생성시키려는 경향을 보이는 WARD 기법을 이용한다. WARD기법은 계층적 클러스터링 방법 중 하나로 클러스터를 구성하는 모든 대상들의 측정치의 분산을 기준으로 클러스터링을 수행한다. 클러스터 결합 때마다 가장 유사한 클러스터들, 즉 측정치의 분산이 가장 작은 쌍을 결합해 나가는 방식이다.

최초의 관련단어 클러스터링 시 연관도가 가장 높은 두 단어를 연결하고, 한 번 클러스터에 할당된 단어들은 다음 클러스터 형성에서 제외된다. 나머지 단어들 중 가장 연관도가 높은 단어를 선택하여 이미 형성된 클러스터와 값을 비교하여 추가 혹은 새로운 클러스터를 형성한다. 단계별로 단어클러스터를 만들어 가는 과정에서 모든 가능한 결합 중 두 클러스터의 결합으로 인한 거리의 증분이 최소가 되도록 클러스터를 결합한다. 크기가 n_1과 n_2인 두 단어클러스터 N_1과 N_2를 결합할 때 생기는 E의 증분은 식 (11)과 같다. d는 유클리디안 거리, $(\overline{X_1}, \overline{X_2})$는 단어클러스터의 대표점이며, $E(N_1, N_2)$를 두 클러스터N_1과 N_2의 거리로 정한다.

$$E(N_1, N_2) = \frac{n_1 n_2}{n_1 + n_2} d^2 \ (\overline{X_1}, \overline{X_2}) \quad (11)$$

표 5 문장클러스터-대표단어 행렬 예

문장클러스터 / 단어	C_1	C_2	C_3
hurricane	1	2	0
doctor	0	1	1
fuel	2	1	3
coast	1	3	1
medicine	3	0	2

관련단어를 클러스터링하기 위한 과정을 예를 통해서 설명한다. 표 5와 같이 문장클러스터별 대표단어의 행렬을 구성한다면, 식 (11)을 이용하여 다음과 같이 단어 간의 거리를 계산할 수 있다. 처음에는 모든 클러스터가 하나의 단어로 이루어져 있으므로 크기는 모두 1이 되며 문장클러스터에 나타나는 단어의 출현빈도가 최초의 대표점이 된다.

$$E(hurricane,doctor)=(\frac{1\cdot 1}{1+1})\,d^2((1,2,0),(0,1,1))$$

$$=\frac{1}{2}\{(1-0)^2+(2-1)^2+(0-1)^2\}=\frac{3}{2}$$

$$E(hurricane,fuel)=(\frac{1\cdot 1}{1+1})\,d^2((1,2,0),(2,1,3))$$

$$=\frac{1}{2}\{(1-2)^2+(2-1)^2+(0-3)^2\}=\frac{11}{2}$$

$E(hurricane,coast)=1$, $E(hurricane,medicine)=6$,

$E(doctor,fuel)=4$, $E(doctor,coast)=\frac{5}{2}$,

$E(doctor,medicine)=\frac{11}{2}$, $E(fuel,coast)=\frac{9}{2}$,

$E(fuel,medicine)=\frac{3}{2}$, $E(coast,medicine)=7$

	doctor	fuel	coast	medicine
hurricane	$\frac{3}{2}$	$\frac{11}{2}$	1	6
doctor		4	$\frac{5}{2}$	$\frac{11}{2}$
fuel			$\frac{9}{2}$	$\frac{3}{2}$
coast				7

계산에 의한 증분값 중에서 $E(hurricane,coast)$가 가장 작으므로 단어 "hurricane"과 "coast"를 하나의 클러스터로 결합한다. 새로운 단어클러스터 (hurricane, coast)의 크기는 2, 대표점은 $(\frac{1+1}{2},\frac{2+3}{2},\frac{0+1}{2})=(1,\frac{5}{2},\frac{1}{2})$가 되며 다음 클러스터 계산 시 새로운 값을 적용한다.

$$E((hurricane,coast),doctor)=(\frac{1\cdot 2}{1+2})d^2((1,\frac{5}{2},\frac{1}{2}),(0,1,1))$$

$$=\frac{2}{3}\left\{(1-0)^2+(\frac{5}{2}-1)^2+(\frac{1}{2}-1)^2\right\}=\frac{7}{3}$$

$$E((hurricane,coast),fuel)=\frac{19}{3},$$

$$E((hurricane,coast),medicine)=\frac{25}{3}$$

	doctor	fuel	medicine
(hurricane,coast)	$\frac{7}{3}$	$\frac{19}{3}$	$\frac{25}{3}$
doctor		4	$\frac{11}{2}$
fuel			$\frac{3}{2}$

계산결과 $E(fuel,medicine)$의 값이 가장 작으므로 단어 "fuel"과 "medicine"을 새로운 클러스터로 결합하며 목표 클러스터의 개수만큼 클러스터가 형성될 때까지 클러스터링을 위한 계산을 반복한다.

4.4 문장 클러스터링 및 클러스터별 대표문장 추출

클러스터 내의 단어들이 나타난 각 문장들은 하나의 단어클러스터에 속하게 된다. 문장 클러스터링을 하기 위해 문장과 클러스터 단어들 간의 유사도 값을 계산하여 문장이 포함되어야 할 단어클러스터를 결정한다. 문장과 클러스터 간의 유사도 비교 시 적용되는 공식은 가장 일반적으로 적용되는 계수인 코사인 계수를 이용한다. 문장 S와 클러스터 C가 벡터길이 k일 때, 단어i의 가중치 값을 이용한 문장과 클러스터벡터 간 유사도를 식 (12)의 코사인 계수에 의해 나타낸다.

$$Sim(s,c)=\frac{\sum_{i=1}^{k} w(s_i)\cdot w(c_i)}{\sqrt{\sum_{i=1}^{k}(w(s_i))^2\cdot\sum_{i=1}^{k}(w(c_i))^2}} \quad (12)$$

$w(s_i)$ = 문장S에 나타난 단어i의 가중치

$w(c_i)$ = 클러스터C에 나타난 단어i의 가중치

전체 문장들을 클러스터링하면 유사한 내용의 문장들이 동일 클러스터에 모이게 되고, 각 클러스터별로 가장 연관도가 높은 하나의 문장을 클러스터 대표문장으로 추출한다. n개의 단어 $w_1,w_2,\cdots,w_n$으로 구성된 k번째 문장 $Sentence_k$가 존재하고 클러스터의 구성단어가 총C개 존재할 때 $Sentence_k$의 문장 값 $Score_{sentence}$는 식 (13)과 같다. 이 계산 값에 의해 클러스터 1개당 k개의 구성문장 중에서 연관도가 가장 높은 1문장을 선택한다. 표 6은 단어클러스터와 문장 클러스터의 결과 예이며, 각각의 단어가중치와 문장 값을 보여주고 있다.

$$Score_{sentence}(Sentence_k)=\sum_{i=1}^{k}(\sum_{i=1}^{c} asim(w_j)) \quad (13)$$

문장의 중복성을 배제하기 위해 클러스터별로 가장 중요한 문장 하나를 선택하여 대표문장으로 추출하며, 이 과정에서 일정크기에 의한 최종 요약문 완성을 위해

표 6 단어클러스터와 문장클러스터의 결과 예

단어클러스터	문장클러스터
coast[2.08] hurricane[8.15] hondura[2.08] Mitch[3.30] caribbean[1.39]	The entire coast of Honduras was under a hurricane warning and up to 15 inches (38 centimeters) of rain was forecast in mountain areas.[12.31] The strongest hurricane to hit Honduras in recent memory was Fifi in 1974, which ravaged Honduras' Caribbean coast, killing at least 2,000 people.[19.77] **At its peak, Mitch was the fourth-strongest Caribbean hurricane in this century, behind Gilbert in 1988, Allen in 1980 and the Labor Day hurricane of 1935.[20.99]**
fuel[2.48] relief[1.39] food[5.66] president[0.69] cliton[1.39] medicine[2.08]	**In Washington on Thursday, President Bill Clinton ordered dlrs 30 million in Defense Department equipment and services and dlrs 36 million in food, fuel and other aid be sent to Honduras, Nicaragua, El Salvador and Guatemala.[10.22]** Mexico was sending 700 tons of food, 11 tons of medicine, at least 12 helicopters, four cargo planes and 475 soldiers to help in relief operations.[9.13]

대표문장들은 클러스터별 중요도에 의해 순위화된다. 이를 위해 클러스터 내의 문장 개수를 기준으로 한 방법과 클러스터를 구성하고 있는 단어들의 가중치 값을 이용하는 방법을 비교한 결과, 문장 개수를 기준으로 한 방법이 더 효율이 높은 것으로 나타났기 때문에 클러스터별 중요도는 클러스터 내의 문장 개수를 기준으로 한다.

5. 실험 및 평가

제안시스템의 실험은 요약 연구자들이 모여 요약시스템을 비교하는 DUC(Document Understanding Conferences)에서 사용하고 있는 ROUGE(Recall Oriented Understudy for Gisting Evaluation)를 이용하였다. DUC에서는 요약전문가들이 미리 작성한 이상적인 요약문을 기반으로, 각 시스템들이 자동으로 생성한 요약문을 대상으로 시스템성능을 측정하고 있다[11]. DUC의 요약평가분야는 표 7과 같이 총 5개의 Task로 이루어져 있으며, 본 논문에서는 다중문서 요약 평가를 위한 Task2 분야에 대해 실험하였다. ROUGE는 ISI/USC의 Chin-Yew이 요약 성능을 평가하기 위해 개발한 재현율에 기반한 평가시스템이다[12]. ROUGE-n gram은 문자열의 선두에서부터 한 문자씩 옮기면서 n문자 단위(n=1,2,3,4)로 일치 여부를 측정하는 방법이다. ROUGE-L은 두 문장에서 공통으로 나타나는 단어 개수를 순서에 상관없이 측정하며 ROUGE-W는 두 문장에서 공통으로 나타나는 단어 개수에 대해 연속적인 일치여부를 고려하여 측정한다.

• ROUGE-N =

$$\frac{\sum_{S\in ReferenceSummaries}\ \sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in ReferenceSummaries}\ \sum_{gram_n\in S} Count(gram_n)} \quad (14)$$

n: $n-gram, gram_n$ 의 길이

$Count_{match}(gram_n)$: 전문가 요약문과 시스템 생성 요약문에서동시 발생한 $n-gram$수

• $ROUGE-L$: Longest Common Subsequence with maximum length

• $ROUGE-W$: Weighted Longest Common Subsequence

다중문서 요약을 위한 제안시스템과의 비교평가 실험을 위한 대상은 단어의 공기정보를 이용하지 않고 단순히 단어의 출현횟수에 따른 방법과 문서별 단어공기정보를 이용하는 방법, 문장별 단어의 공기정보를 이용하는 방법이다.

표 8은 제안시스템에 대한 비교실험결과이다. 실험결과 문서별 공기정보와 문장클러스터별 공기정보를 이용하는 경우 단순 통계정보에 의한 방법보다 향상된 결과를 보이고 있으며, 제안시스템의 성능이 가장 좋은 것으로 평가되었다. 이것은 문장클러스터에 의한 단어의 공

표 7 DUC Tasks

Tasks	Summary	Data
Task 1	single document summaries (<=75 bytes)	50 English TDT clusters (10 documents/cluster)
Task 2	multiple document summaries(<=665 bytes)	50 English TDT clusters (10 documents/cluster)
Task 3	cross-lingual single document summaries	25 TDT Arabic document clusters
Task 4	cross-lingual multiple document summaries	25 TDT Arabic document clusters
Task 5	Short summaries by Question (Who is X ?)	50 TREC English document clusters

표 8 대상 요약시스템의 비교 실험

Method	TF	공기정보		
		문서별	제안시스템 (문장클러스터별)	문장별
Rouge 1	0.33748	0.34385	0.34955	0.33324
Rouge 2	0.05898	0.06135	0.06434	0.05742
Rouge 3	0.01407	0.01575	0.01636	0.01431
Rouge 4	0.00446	0.00503	0.00508	0.00523
Rouge L	0.31202	0.32006	0.32229	0.30979
Rouge W	0.10709	0.11010	0.11072	0.10635

기정보 이용 시 정확한 주제분리가 가능하여 중복성 제거에 효율적임을 의미한다. 따라서 전체 문서집합의 중요의미를 충분히 반영할 수 있어 대표문장 추출에 의한 다중문서 요약에서 좋은 결과를 보인 것이다. 문서별 단어의 공기정보를 이용하여 단어를 클러스터링하는 경우 주제분리가 명확하지 않다는 문제점에도 불구하고 단순 통계정보를 이용하는 경우보다는 좋은 결과를 보이고 있다. 단어의 공기정보를 이용하는 방법 중 문장별 공기정보를 이용하는 경우가 실험대상 방법 중에서 가장 저조한 결과를 보이고 있다. 이것은 지나치게 세분화된 공기정보 측정 단위로 인해 다중문서 집합에서 문장별 단어의 공기정보가 충분히 단어 간의 정보를 반영하지 못하고 있음을 의미한다. 또 다른 원인으로서, 공기정보 행렬구성 시 전체 문서집합에 나타나는 단어의 총수는 매우 많은데 비해 하나의 문장 안에 포함되는 단어의 수는 상대적으로 적을 수밖에 없게 된다. 따라서 단어-문장 행렬의 요소는 0인 곳이 대부분이고 0이외의 요소를 갖는 것은 소수가 된다. 이와 같은 상관행렬 구성의 특성으로 인한 잡음(noise)현상 때문에 요약결과가 부정확한 것으로 평가된다.

표 9는 DUC2004에 참여한 타 요약시스템들과의 비교실험 결과이다. Rouge 1인 경우 전체 35개 참가시스템 중에서 15번째에 해당하는 순위로 평가 되었으나 나머지 경우에는 중하위의 순위 결과를 보이고 있다. 이것은 평가의 방법이 전문가에 의한 생성요약 결과물과 본 논문에서 제안한 방법인 대표문장 추출에 의한 단순한 단어 재현율 비교이기 때문에 갖는 특성으로 보인다.

재현율을 높이기 위해서는 생성요약을 수행하거나 전체 문장을 추출하지 않고 문장 중에서 일부 중요한 구나 절을 추출하는 방법이 있는데, 이 경우 가독성을 위한 자연스런 문장 연결방법이 별도로 필요하다는 과제가 있다. 이 과제를 해결하기 위한 기술적인 어려움은 있지만 궁극적으로 자동요약의 성능향상을 위해서는 반드시 극복을 해야 할 문제라고 판단되기 때문에 지속적인 연구를 통해 요약시스템을 보완할 필요성이 있다.

6. 결론 및 향후 연구과제

본 논문에서는 단어의 공기정보를 이용한 관련단어 클러스터 기반 다중문서 요약을 제안하였다. 중요문장 추출에 의한 다중문서 요약에서는 정보의 중복문제가 매우 중요한데, 정보의 중복성 제거를 위해 문서집합 내 단어들의 연관성을 반영하여 유사문장들을 클러스터링 함으로써 좋은 요약결과를 보였다. 이는 단어의 공기정보를 이용하면 관련단어 클러스터링 수행 시 주제분리의 정확률을 향상시킬 수 있다는 것을 의미한다. 관련단어 클러스터링을 위한 단어의 공기정보 계산 시 어떤 단위를 적용하느냐에 따라 성능의 차이를 보였다. 실험결과 문서, 문장클러스터, 문장 중 문장클러스터단위로 공기정보를 계산할 경우 가장 우수한 것으로 나타났다. 관련단어 클러스터링 시 각 단어들은 여러 개의 클러스터에 중복되게 하지 않고 단지 하나의 클러스터에 속하도록 수행했지만, 하나의 단어가 복수개의 클러스터를

표 9 DUC2004 참가 시스템과 비교 실험

Method	제안시스템	Rank	Best System	Worst System	Best Human	Worst Human
Rouge 1	0.34955	15	0.38224	0.24190	0.41828	0.38902
Rouge 2	0.06434	24	0.09216	0.01876	0.10654	0.08595
Rouge 3	0.01636	27	0.03529	0.00277	0.03574	0.02427
Rouge 4	0.00508	28	0.01658	0.00078	0.01280	0.00783
Rouge L	0.32229	26	0.38950	0.27630	0.43380	0.40631
Rouge W	0.11072	26	0.13378	0.09358	0.14804	0.13805

• Rank : 전체 참가시스템 35개 중 해당순위
• Best System : 참가 시스템 중 최고 성능 값
• Worst System : 참가 시스템 중 최저 성능 값
• Best Human : 요약 전문가에 의한 최고 성능 값
• Worst Human : 요약 전문가에 의한 최저 성능 값

형성 했을 경우의 결과에 대해서도 비교연구가 필요하다.

앞으로 반드시 개선해야 할 연구과제는 단어의미의 다양성에 관한 것이다. 현재는 단어의 공기정보 계산 시 이형동의어와 동형이의어를 고려하지 않았지만 이 문제를 보완하여 향후 요약시스템에 적용하면 더 정확한 결과가 나올 것으로 예상된다. 또한 본 논문에서는 문장의 중요도 순에 의해 최종요약문을 생성하였지만, 요약문의 가독성을 위해 요약문장들을 대상으로 시간과 주제흐름에 맞도록 문장을 다시 정렬(sentence ordering)하는 문제도 반드시 해결되어야 할 과제이다.

참 고 문 헌

[1] Julian Kupiec, Jan Pedersen, and Francine Chen, "A Trainable Document Summarizer," In Proceedings of ACM-SIGIR'95, pp.68-73,1995.

[2] Mani and Inderjeet, Automatic Summarization. Amsterdam:John Benjamina Publishing Co. 2001.

[3] 박성배, 장병탁, "Co-Trained Support Vector Machines을 이용한 문서분류," 한국정보과학회 봄 학술발표 논문집 (B), 제29권 1호, pp.259-261, 2002.

[4] Barzilay, Regina and Michael Elhadad, "Lexical Chains for Text Summarization," Master's thesis, Ben-Gurion University, 1997.

[5] 장두성, 최기선, "단서 구문과 어휘 쌍 확률을 이용한 인과관계 추출", 제15회 한글 및 한국어 정보처리 학술대회, 2003.

[6] C. J. van Rijsbergen., "A Theoritical Basis for the Use of Co-occurrence Data in Information Retrieval," Journal of Documentation, Vol.33:106-119, 1977.

[7] Salton.G., Automatic text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, 1989.

[8] Sparck Jones, K., "Automatic summarizing:factors and directions," In Mani and Maybury, (eds), Advances in Automatic Text Summarization, pp. 1-12. The MIT Press. 1999.

[9] Morris. A.H.. Kasper and G.M, Adams. D.A., "The effects and limitations of automated text condensing on reading comprehension performance," Information systems Research. pp. 17. 35, March 1992.

[10] 김재훈, 김준홍, "도합유사도를 이용한 한국어 문서요약 시스템", 한국 인지과학회 논문지 제12권 제1·2호, pp.35-42, 2001.

[11] http://www-nlpir.nist.gov/projects/duc/index.html

[12] http://www.isi.edu/~cyl/ROUGE/

[13] Salton.G., Singhal.A., Mitra.M. and Buckly.C., "Automatic text structuring and summarization: Information Processing and Management. Vol. 33, no.2. 1997.

[14] Lin, Chin-Yew and E.H. Hovy., Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada, May 27-June 1, 2003.

[15] Buckley,C.,Singhal,A.,Mitra, M. and Salton, G.,:"New retrieval approaches using SMART:TREC4", Proceedings of the Fourth Text Conference(TREC-4), pp.25-48, 1996.

이 일 주

1988년 아주대학교 전자계산학과(공학사) 1994년 한양대학교 전자계산학과(공학석사). 2002년 아주대학교 컴퓨터공학과(박사수료). 1989년-1998년 현대미디어시스템/현대정보기술 책임. 1998년-현재 동원대학 모바일컨텐츠과 조교수. 관심분야는 정보검색 시스템, 모바일프로그램

김 민 구

1977년 서울대학교 계산통계학과(이학사). 1979년 KAIST 전산학과(공학석사) 1989년 펜실베니아 주립대학교 전산학과(박사). 1999년-2000년 루지애나 대학 연구과학자. 1981년-현재 아주대학교 정보 및 컴퓨터공학부(교수). 관심분야는 지능형 정보검색 시스템, 인공지능(지식표현 추론), 온톨로지 자동구축