단락 자동 구분을 이용한 문서 요약 시스템
(Korean Summarization System using Automatic Paragraphing)

김 계 성 † 이 현 주 †† 이 상 조 †††
(Kye Sung Kim) (Hyun Ju Lee) (Sang Jo Lee)

요 약 본 논문은 단락의 자동 구분을 통해 중요한 문장을 추출하는 요약 시스템을 제안한다. 먼저 어휘의 재출현 여부를 파악하여 어휘의 일치도와 어휘의 역할 변화와 같은 재출현 어휘의 양상 정보를 수집하고, 이를 통하여 문장 간의 긴밀도를 정량적으로 계산한다. 다음으로 측정된 문장간 긴밀도를 이용하여 사용자의 추출 범위에 따라 단락을 구분하고, 각 단락의 대표 문장을 선정하여 최종 요약문을 추출한다. 제안한 방법은 문서 제목, 문장의 위치, 수사 구조 등의 정보를 이용하지 않기 때문에 수사 구조가 자주 발견되지 않는 문서에도 적용이 가능하다.

키워드 : 문장간 긴밀도, 단락 자동 구분, 문서 요약, 문장 추출

Abstract In this paper, we describes a system that extracts important sentences from Korean newspaper articles using automatic paragraphing. First, we detect repeated words between sentences. Through observation of the repeated words, this system compute Closeness Degree between Sentences(CDS) from the degree of morphological agreement and the change of grammatical role. And then, it automatically divides a document into meaningful paragraphs using the number of paragraph defined by the user's need. Finally, it selects one representative sentence from each paragraph and it generates summary using representative sentences. Though our system doesn't utilize some features such as title, sentence position, rhetorical structure, etc., it is able to extract meaningful sentences to be included in the summary.

Key words : Closeness Degree between Sentences(CDS), Automatic Paragraphing, Text Summarization, Sentence Extraction

1. 서 론

문서 요약은 주어진 문서나 문서 집합의 기본적인 내용을 유지하면서 문서의 복잡도, 즉 문서의 길이를 줄이는 작업이다. 다시 말해, 사용자 혹은 태스크에 맞는 요약본을 만들기 내기 위하여 원문으로부터 가장 중요한 정보들을 걸러주는 과정이라 할 수 있다[1]. 자동 문서 요약에 관한 연구는 이미 1950년대부터 시작되었지만, 최근 정보의 폭발적인 증가로 인하여 자동 요약에 대한 관심이 다시 증대되고 있다. 자동 문서 요약은 현재 온라인과 오프라인에 존재하는 수많은 문서 자료들을 보다 쉽고 빠르게 접근할 수 있도록 도와주며, 정보 검색 시스템과 더불어 사용자가 원하는 적확(的確)한 문서 검색을 가능하게 한다.

본 논문은 단락 자동 구분을 통한 한국어 문서 요약 시스템을 제안한다. 먼저 문서에 나타난 어휘의 재출현 양상을 분석하여 문장 간 긴밀도를 파악하고, 이를 통하여 단락을 자동 구분한다. 그리고 각 단락의 대표 문장을 선출함으로써 최종 요약문을 구성하고자 한다.

2장에서는 자동 요약에 관한 기존 연구를 살펴보고, 3장에서 단락 자동 구분을 이용한 문서 요약 시스템에 대해 알아본다. 그리고 4장에서 실험 및 평가를 하며, 5장에서 결론을 맺는다.

2. 관련 연구

문서 요약은 크게 생성 요약(abstact)과 추출 요약(extract)으로 나누어진다. 생성 요약은 문서를 이해하는

† 비 회 원 : 경일대학교 교양학부 교수
kskim@kiu.ac.kr
†† 비 회 원 : 경북대학교 컴퓨터공학과
hyunju@comeng.ce.knu.ac.kr
††† 종신회원 : 경북대학교 컴퓨터공학과 교수
sjlee@bh.knu.ac.kr
논문접수 : 2001년 8월 20일
심사완료 : 2003년 3월 17일

분석(analysis) 단계와 변형(transformation) 단계, 그리고 자연 언어로 문장을 생성하는 통합(synthesis) 단계를 거치는 전문적인(professional) 요약이다[2]. 이에 반해 추출 요약은 새로운 텍스트를 생성하는 일련의 과정을 포함하지 않기 때문에 생성 요약에 비해 적은 비용으로 비교적 쉽게 요약이라는 기술에 접근할 수 있다는 특징을 가진다.

추출 요약에 대한 연구로 Edmundson의 방법과 코퍼스에 기반한 방법 등이 있다[2]. Edmundson은 4가지 자질(실마리어, 제목에 포함된 어휘, 키워드, 문장 위치)에 대한 가중치를 이용해 각 문장에 점수를 부여하고 이를 통해 문장을 추출하는 방법을 사용하였다. 코퍼스에 기반한 방법은 원문과 이미 만들어진 요약문을 학습시켜 요약문으로써 가치가 있는 문장들을 선택하는데 필요한 자질들을 추출하고 이를 이용한다. 이러한 추출 요약 방법들은 주로 문장들에 순위를 매기고 순위가 높은 문장을 중요 문장으로 추출하기 때문에 비슷한 내용의 문장들이 중복 추출되기도 하며, 문맥을 고려하지 않아 결합력이 약한(incoherent) 요약문을 구성한다는 문제가 발생한다. 이러한 문제들을 보완하기 위하여 개체(entity) 혹은 담화(discourse) 단계에서의 연구가 진행되고 있는데, Barzilay의 Lexical chains[3]이나 Marcu의 RST(rhetorical structure theory) 트리[4]가 대표적인 연구라 할 수 있다. 하지만 RST 트리는 수사 관계가 자주 등장하지 않는 문서에는 적용이 어려우며, Lexical chains 또한 한국어로 구성된 어휘 계층 사전의 부재로 한국어 문서에 그대로 적용하기가 어렵다.

본 논문에서는 추출 요약 방법을 사용하되, 단락의 자동 구분이라는 개념을 이용하여 중요 문장을 추출하고자 한다. 단락의 자동 구분은 토픽 분할(topic segmentation)의 개념을 축소시킨 것이라 할 수 있으며, Hearst의 TextTiling[5]이 담화의 덩어리로 토픽을 분할하는 대표적 연구이다. TextTiling은 잡지(magazine)와 같은 긴 문서를 대상으로 어휘적 응집(lexical cohesion) 관계를 이용하여 부논제(subtopic)의 변화를 인식하고 부논제의 이동에 따라 문서를 다수의 단락(multi-paragraph)으로 자동 분할하는 알고리즘이다.

하지만, 본 논문은 기사와 같은 비교적 짧은 문서를 대상으로 한다. 이러한 문서들은 대개 한 가지 토픽만을 다루고 있기 때문에, 본 논문에서는 내용의 흐름을 탐지하여 그 내용이 전환되는 곳에서 단락을 자동 분할하고 각 단락의 대표 문장을 요약문으로 구성한다. 따라서, 본 논문은 단순히 문장의 순위를 정해 중요 문장을 추출하는 방법에 비해 비슷한 내용의 문장이 요약문에 중복 추출되는 문제를 줄일 수 있으며, 문장 간의 관계를 살펴 요약문을 추출하기 때문에 담화 단계 분석을 위한 첫 단계라 할 수 있다.

3. 단락 자동 구분을 이용한 신문 기사 요약 시스템

텍스트는 문맥 위에서 의미적으로 연결된 문장들이 서로 관계를 이루며 나열되어 있다. 다시 말해, 관계없는 문장들을 아무렇게나 늘어놓아서는 하나의 텍스트가 될 수 없다. 따라서, 문장 사이에 존재하는 연결(connections)이라는 개념을 파악해야 하는데, 이를 위해 먼저 문장 내 어휘들 사이에 존재하는 연결 관계, 즉 응집성(cohesion)을 분석해야 한다[6].

본 연구는 문장 사이에서 발견된 어휘의 재출현 양상 정보를 수집하고, 이를 기반으로 문장 사이의 연결을 탐지하고자 한다. 먼저, 어휘의 재출현 양상은 재출현한 어휘의 표면적 형태와 문법적 성분을 수집하는 것으로, 이를 통하여 두 문장 사이의 내용의 긴밀함을 분석한다. 본 논문에서 언급한 문장 간 긴밀도는 문장 내 어휘의 형태적 일치의 정도와 문법적 성분의 변화를 기반으로 하고 있다. 기존의 코사인 유사도를 이용한 문장 간 유사도는 두 문장 사이에서 중복 출현한 어휘의 개수에 기반을 두고 있기 때문에 두 문장 사이의 형태적 유사함을 살피는 일차적 방법이라 할 수 있다. 하지만 본 연구의 문장 간 긴밀도는 코사인 유사도를 기반하여 형태적 유사함의 정도를 분석하고, 동시에 문장 의미의 결정에 영향을 미치는 어휘의 문법적 성분까지를 고려한다. 왜냐하면, 무의미하게 나열된 어휘들 속에서, 단순히 겹치는 어휘의 개수가 많고 적음만을 가지고는 문장 사이의 내용의 긴밀함을 판단하기에 부족함이 있기 때문이다. 입력 문서의 문장 간 긴밀도 분석이 완료되면, 사용자의 추출 범위에 따라 단락을 동적으로 구분하고, 각

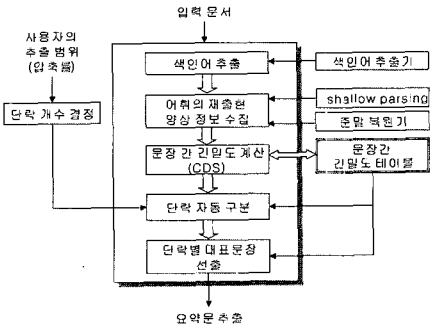

그림 1 시스템 구성도

단락의 대표 문장 선정을 통하여 요약문을 구성한다.

그림 1은 제안한 단락 자동 구분을 통한 신문 기사 요약 시스템의 전체 구성도이다.

3.1 어휘의 재출현 양상

글을 전개해 나갈 때는 동일한 형태의 어휘를 계속해서 반복 출현시키기보다는 비슷한 의미의 다른 어휘들로 바꾸어 내용을 전개시키는 일이 많다. 재출현 양상은 다시 어휘의 일치와 역할 변화로 구분한다.

3.1.1 어휘의 일치

어휘의 일치는 되풀이되는 어휘의 모습을 완전 일치와 부분 일치로 구분하여 수집한다. 완전 일치는 재출현한 어휘의 모습이 동일함을 의미하며, 부분 일치는 다시 중심어 일치와 수식어 일치로 나누어진다. 예를 들면, '외화 차입'은 '외화의 차입'으로, '차입조건'은 '차입에 대한 조건'으로 파악할 수 있고, 두 경우에서 보여지는 '차입'에 대한 역할이 서로 다르기 때문에 이를 구분하는 것이다. 또한 본 시스템은 문서상에 나타나는 준말 또는 축약형 어휘를 처리하기 위하여 [7]을 이용한다. 이것은 "도시계획위원회와 도시계획위", "서울대학교와 서울대", "IQ와 Intelligence Quotient" 등이 형태는 다르지만 서로 동일한 대상을 지칭하고 있기 때문에 이들을 동일 개체로 인식하기 위한 것이다. 따라서, 어휘의 일치는 내용 흐름상에 연결의 고리가 되는 어휘들이 새로운 복합 명사를 형성하거나 축약 또는 준말 처리되는 등의 이유로 인해 새로운 어휘로 등장하더라도 이들을 서로 별개의 어휘로 인식하지 않는다.

3.1.2 어휘의 역할 변화

어휘의 역할 변화는 되풀이되는 어휘의 문법적 역할을 수집하는 것으로, 크게 필수 성분으로의 이동과 수의 성분으로의 이동으로 나눈다. 본 논문에서 사용하고 있는 구문 분석기[8]는 문장의 의미적 중심 용언을 파악하여, 중심 용언에 대한 필수격(주어, 목적어, 필수부사어)만을 밝히도록 성능을 제한시킨다. 이는 문장 내에서 중심 역할을 담당하는 어휘의 일치 및 이동을 문장간 긴밀도 분석에 달리 반영하기 위함이다.

3.2 문장 간 긴밀도

앞서 언급한 어휘의 일치와 역할 변화에 대한 재출현 양상이 파악되면, 이를 기반으로 문장간 내용의 긴밀도를 계산한다. 긴밀도는 두 문장의 형태적 일치와 역할 변화를 이용하며, 형태적 일치는 정보 검색 분야에서 많이 알려진 코사인 유사계수를 기반으로 구한다. 어휘의 겹침(overlap) 정도를 이용하여 문장들을 비교하는 척도로는 코사인 유사계수, dice's coefficient 등이 있다[9].

문장 간 긴밀도(CDS)는 두 문장의 내용이 얼마나 긴밀한 지를 살피는 척도로, 다음과 같이 계산한다.

$$CDS(x,y) = \alpha \times move(x,y) + \beta \times sim(x,y)$$

$$move(x,y) = \sum_{i=1}^{N1} f(x_i, y_i)$$

$$sim(x,y) = \frac{\sum_{i=1}^{N2}(x_i y_i) \times \gamma}{\sqrt{\sum_{i=1}^{N2}(x_i)^2 \times \sum_{i=1}^{N3}(y_i)^2}}$$

$$f(a,b) = \begin{cases} constant\ c & \text{if } a = \text{필수성분 and } b = \text{필수성분} \\ 0 & else \end{cases}$$

x_i는 문장 x의 어휘 i에 대한 단어빈도(tf : term frequency) 가중치를 말하며, y_i는 문장 y에서의 어휘 i에 대한 단어빈도 가중치를 의미한다. N1은 문장 x와 y의 공통 어휘 개수이며, N2와 N3는 각각 문장 x, y에 대한 전체 어휘 개수이다. 또한 sim은 두 문장 사이에 출현한 어휘의 일치 정도를 의미하며, $move$는 문장 사이에서 재출현한 어휘의 역할 변화 점수를 말한다. (α, β는 가중치이며, γ는 완전일치와 부분 일치에 대한 점수이다.)

출현 어휘의 각 양상에 대한 가중치는 다음의 조건을 고려하여 달리 설정한다.

- 어휘의 일치에 대한 가중치
  완전일치 > 중심어 일치 > 수식어 일치
- 어휘의 역할 변화에 대한 가중치
  필수성분으로의 이동 > 수의성분으로의 이동

다음은 문장 간 긴밀도의 예이다.

a.	(a1) 인간은 누구나 언어를 사용할 줄 안다. (a2) 언어 사용이야말로 인간과 다른 동물을 구별할 수 있게 하는 가장 큰 특징이다.	*CDS(a1,a2)* = 0.6837
b.	(b1) 인간은 누구나 언어를 사용할 줄 안다. (b2) 언어는 인간 사회에서 어떤 개념을 특정한 소리를 사용하여 지시하자는 약속이다.	*CDS(b1,b2)* = 0.5418

a와 b의 각 두 문장은 "인간", "언어", "사용"이라는 세 어휘가 동일하게 중복된다. 하지만 a의 긴밀도가 b의 긴밀도보다 높게 나타남을 볼 수 있는데, 이는 a의 두 문장이 b의 두 문장보다 내용상 더 긴밀함을 의미하는 것이다. 즉, 문장 간 긴밀도는 어휘의 일치와 어휘의 역할 변화 점수에 의해 계산되는 값으로, 재출현 어휘가 많을수록 그리고 그 형태가 완전히 일치하면서 문장의 필수 성분으로 이동한 경우에 가장 높은 점수를 받게 된다.

본 논문은 글의 흐름을 전제로 한다. 따라서 Si에서 Sj사이의 긴밀도만을 분석대상으로 설정한다(i<j ; i,j는 문장번호이다). 다시 말해, 문장 S1->S2, S1->S3, ..

S2->S3 등의 관계는 긴밀도 분석의 대상이 되지만, 문장 S2->S1, S3->S2, S4->S1 등의 관계는 긴밀도 분석에서 제외한다는 것이다. 왜냐하면, 일반적으로 글은 내용의 흐름에 따라 일관성 있게 전개되기 때문이다.

3.3 단락 자동 구분과 단락별 대표 문장 추출

단락의 자동 구분은 문맥의 개념을 중요 문장 추출에 최대한 반영하기 위한 것이다. 본 논문은 어휘의 재출현 양상을 기반으로 문장 사이의 연결을 파악하고 있다. 앞서 언급한 문장 간 긴밀도는 문장 사이의 내용의 긴밀함을 수치로 측정한 값이며, 이를 기반으로 단락을 자동 구분하게 된다. 이 때, 나누어지는 단락의 개수는 사용자가 원하는 요약문의 추출 범위에 따라 동적으로 변한다. 예를 들어, 11개 문장으로 구성된 문서에서 20%의 중요 문장을 추출하고자 한다면, 두 개의 문장이 추출되어야 하므로 긴밀도가 가장 느슨한 두 문장 사이에서 단락을 구분하고, 각 단락의 대표 문장을 추출하여 요약문으로 구성하는 것이다.

먼저, 문서의 길이(문장수)와 요약문의 추출 범위(압축률)에 따라 구성해야 하는 단락의 개수를 결정한다. 다음으로 단락의 분리점을 판단해야 하는데, 이 때는 인접한 문장 간 긴밀도만을 이용한다. 다시 말해, 문서의 내용이 전개되어 나가는 방향에서 가장 낮은 점수를 가진, 즉 내용상으로 가장 느슨한 긴밀도를 가지는 두 문장 사이에서 단락이 1차 분리된다. 계속해서 그 다음으로 느슨한 긴밀도를 가진 두 문장 사이가 2차, 3차, ... 분리점이 되는 방식으로 단락이 나누어진다.

단락을 동적으로 자동 구분한 후에는 인접한 두 문장 사이의 긴밀도뿐만 아니라 이미 계산된 문장간의 긴밀도를 모두 이용하여 각 단락의 대표 문장을 추출한다. 본 시스템은 단락 내 대표 문장의 후보가 되는 각 문장들 중에서, 타단락의 모든 문장들과 가장 높은 긴밀도를 가지는 문장을 그 단락의 대표 문장으로 선택한다. 다시 말해, 단락 내 대표 문장의 후보가 되는 각 문장들에 대해서 그 자신과 타단락에 있는 모든 문장들 사이의 긴밀도 점수를 모두 합하고, 그 중에서 가장 높은 점수를 가지는 문장을 그 단락의 대표 문장으로 선출한다. 그런 후, 각 단락에서 추출된 대표 문장들을 가지고 요약문을 구성한다.

4. 실험 및 평가

요약 시스템 평가는 크게 내적 평가와 외적 평가로 나누어진다. 본 논문은 사람이 추출한 이상적인 요약문과 시스템이 추출한 요약문을 분석하여 시스템의 성능을 직접 평가하는 내적 평가를 수행한다. 실험은 30건의 신문 기사를 대상으로 하며, 이들은 모두 8문장 이상으로 구성된 문서이다. 실험 문서의 문장 수는 평균 9.4개이다.

4.1 실험 문서에 대한 재출현 어휘의 개수

요약 시스템 평가에 앞서, 실험 대상으로 삼은 신문 기사 문장의 재출현 어휘 수를 분석해 보았다. 이는 본 논문에서 제안한 문장 간 긴밀도가 재출현 어휘의 개수에 영향을 받을 수 있기 때문이다. 표 1은 문장 내 어휘수에 따른 재출현 어휘의 평균 개수이다.

표 1 문장내 어휘수에 따른 재출현 어휘의 개수

문장의 어휘수	재출현 어휘의 평균 개수
1~9	2.31
10~15	2.47
16~20	2.57
21이상	3.22

그 결과, 두 문장 사이에서 재출현할 어휘는 평균 2.5개였으며, 신문 기사의 한 문장은 평균 14개의 어휘(명사류)로 구성됨을 볼 수 있었다. 따라서, 신문 기사가 다른 종류의 문서에 비해 비교적 긴 문장을 많이 사용하고 있음을 알 수 있다.

4.2 평가자 간 일치도

평가자 사이의 일치도는 percent agreement를 사용하여 측정한다[10]. Percent Agreement는 과반수의 판단에 대하여, 가능한(possible) 일치와 관찰된 일치 사이의 비율을 말한다. 본 실험에서는 3명 이상의 평가자들이 선택한 문장을 과반수의 판단으로 본다. 표 2는 30개의 실험 문서에 대해서 5명의 평가자들이 각각 20%의 요약문을 추출하였을 때 측정된 평가자 간 일치도이며, 표 3은 단락 분할에 대한 평가자 간 일치도이다.

표 2 요약문에 대한 평가자 간 일치도

Length	Avg.Agreement	Max	Min
20%	74.50%	100%	50%

표 3 단락 분할에 대한 평가자 간 일치도

Avg.Agreement	Max	Min
74.36%	100%	67%

4.3 요약 시스템 평가

평가는 이상적인 요약문에 대해서 본 시스템이 추출한 요약문과 MS-word가 추출한 요약문을 각각 비교함으로써 수행하였다. 성능 평가를 위한 척도로는 정보 검

표 4 시스템 평가

length	MS-Word			본 시스템		
	Precision	Recall	F measure	Precision	Recall	F-measure
20%	0.39	0.33	0.36	0.64	0.49	0.56

색 시스템을 평가하는데 널리 이용되는 정확률과 재현율, 그리고 F-점수를 사용한다. 다음은 두 시스템의 평가 결과이다.

각 문서에서 20%의 요약문을 추출하였을 때, MS-Word의 F-점수는 36%, 본 시스템의 F-점수는 56%였다. 정확률이나 재현율에 있어서도 본 시스템이 MS-Word에 비해 우수함을 볼 수 있다. 이는 단락 자동 구분을 이용하여 중요 문장을 추출하는 본 시스템이 이를 이용하지 않는 타시스템에 비해 나은 성능을 보이고 있음을 의미한다.

그러나 아직까지 요약시스템이 그다지 만족스럽지 않다. 이는 몇 가지 요인으로 나누어 살펴볼 수 있다. 먼저, 이상적인 요약문의 문장과 비슷한 의미를 가지지만 표면적으로 서로 다른 문장으로 판단되어 요약문 구성에서 누락된 경우이다. 그러나, 이 문제는 요약 시스템이 어느 정도 만족할 만한 성능을 가진 차후에 논의되어야 할 것으로 생각된다. 다음으로, 문장 간 응집력 형성에 영향을 미치는 요소들을 생각해 볼 수 있다. 본 시스템은 어휘의 재출현 양상 정보를 이용하고 있는데, 여기에 수사어구, 동의어, 대용어 등에 관한 연구를 추가하여 문장 사이의 긴밀도를 보다 정확히 판단할 수 있다면 제안한 시스템의 성능 향상에 기여할 것이다. 마지막으로, 요약문을 보다 객관적으로 평가할 수 있는 외적 평가 방법의 도입을 고려해본다면, 보다 나은 평가 결과를 기대해 볼 수 있을 것이다.

여기서, 제안한 시스템의 단락 구분이 어느 정도의 성능을 보이는지를 살펴보고자 한다. 단락 분할 실험은 사람이 수작업으로 구성한 단락과 시스템이 구성한 단락을 비교함으로써 측정되었다.

표 5에서 보듯이, 제안한 시스템의 단락 분할에 대한 F-점수는 63%였다. 이 결과는 문장에 드러난 수사어구 등을 고려하지 않은 현 시점에서는 만족할 만한 것이라 할 수 있다. 그러나, 앞서 언급한 바와 같이, 문장 간 응집력 형성에 영향을 주는 여러 요인들이 단락 형성에까지 영향을 미치기 때문에, 이들과 함께 단락 자동 구분을 위한 담화 단계에서의 연구가 진행되어진다면 단락 분할 단계의 성능을 향상시킬 수 있다.

표 5 단락 분할에 대한 평가

Precision	Recall	F-measure
0.73	0.55	0.63

5. 결 론

본 논문에서는 단락의 자동 구분을 통한 문서 요약 시스템을 제안한다. 먼저 재출현 어휘의 양상을 어휘의 일치도와 어휘의 역할 변화로 구분하여 분석하고, 이를 문장 사이의 내용이 얼마나 긴밀한가를 판단하기 위한 척도로 이용하였다. 문장 사이의 긴밀도 분석을 통해서 사용자의 원하는 요약문의 추출 범위에 따라 단락을 나누고 각 단락의 대표 문장을 선출함으로써 최종 요약문을 추출하였다. 실험한 결과, 제안한 시스템이 MS-Word에 비해서 성능이 어느 정도 향상되었음을 볼 수 있었다. 하지만, 아직까지도 요약문 평가에 대한 어려움은 계속 남아 있다.

앞으로 문장 간 긴밀도와 단락 형성에 영향을 미치는 요소들에 대한 연구를 계속 진행시켜야 하며, 단락 분할을 위한 학습 방법의 도입이나 문장 사이의 결합 관계 등을 접목시켜 시스템의 전체적인 성능 향상을 모색해 보려 한다. 그리고 대용어나 생략 등으로 인해 저하되는 추출 요약 문장들 사이의 결합력, 가독성(readability) 등을 높이기 위한 후처리 방안도 추후 진행되어야 할 것이다.

참 고 문 헌

[1] Inderjeet Mani and Mark T. Maybury, Advances in automatic text summarization, The MIT Press, 1999.

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

[3] Regina Barzilay, "Lexical Chains for Summarization," M.Sc. degree of Ben-Gurion University of the Negev, 1997.

[4] Daniel Marcu, "Discourse trees are good indicators of importance in text," In I. Mani and M. Maybury editors, Advances in Automatic Text Summarization, pages 123-136, The MIT Press, 1999.

[5] Marti A. Hearst, "Multi-paragraph segmentation

of expository text," In Proceedings of the 32nd Annual Meeting of the Association for Computational Lingustics(ACL), Las Cruces, NM, June 1994.
[6] 담화 연구의 기초, 이원표 역, 한국문화사, 1999.
[7] 김상수, 김계성, 노태길, 이상조, "문서 요약을 위한 조응대용 해결", 제29회 정보과학회 추계학술발표논문집(B), 2002.
[8] 정영규, 이현주, 이상조, "신문기사 요약문 생성을 위한 구문 분석기 구현", 제28회 정보과학회 춘계학술발표논문집(B), 2001.
[9] Gerard Salton, Michael J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983.
[10] Gale William, Kenneth W.Church, and David Yarowsky, "Estimating upper and lower bounds on the performance of word-sense disambiguation programs," In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics(ACL-92), pages 249-256, 1992.

김 계 성

1996년 부산여자대학교(현:신라대학교) 전자계산학과 졸업(이학사). 1998년 경북대학교 대학원 컴퓨터공학과(공학석사) 2002년 경북대학교 컴퓨터공학과 박사과정 수료. 2003년~현재 경일대학교 교양학부 수업전담전임강사. 관심분야는 문서요약, 정보 검색

이 현 주

1995년 경북대학교 국어국문학과 졸업(문학사). 1997년 경북대학교 대학원 국어국문학과(문학석사). 1999년 경북대학교 대학원 국어국문학과 박사과정 수료 2001년~현재 경북대학교 대학원 컴퓨터공학과 석사과정 재학중. 관심분야는 문서요약, 구문분석

이 상 조

1974년 경북대학교 수학교육과(이학사) 1976년 한국 과학기술원(이학석사). 1994년 서울대학교 컴퓨터공학과(공학박사) 관심분야는 자연어 처리, 기계번역, 운영체제, 프로그래밍 언어, 데이타베이스