자동요약시스템 구축에 대한 연구*

- 웹 상의 보도기사를 중심으로 -

A Study on the Construction of the Automatic Summaries

- on the basis of Straight News in the Web -

이태영(Taeyoung Lee)**

초 록

웹의 보도기사에 관한 자동요약시스템을 구축하기 위하여 담화구조와 지식기반 기법을 적용한 글구조 프레임과 제 규칙들을 작성하였다. 프레임에는 문단과 문장 및 절의 역할, 문단과 문장의 성질, 역할을 구분하는 판별규칙, 주요문장 발췌규칙, 그리고 요약문작성규칙 슬롯이 포함되었다. 문맥정의, 고유명사 등을 안내하는 'if-needed'와 변화된 슬롯 값을 알려주는 if-changed 패싯도 구비되었다. 슬롯이나 패싯의 실제 값들을 추출 표현하는 과정에서 문구의 수사적 역할과 단어 최상위 범주 및 줄거리 단위를 참조하였다. 의미흐름의 연결성을 유지하면서 요약 문장들을 통합, 분리, 합성하는 재구성은 유사도공식, 구문정보, 담화구조와 지식기반 방법에서 도출한 제 규칙 및 문맥정의를 이용하였고 비평과 같은 새로운 문장을 생성하였다.

ABSTRACT

The writings frame and various rules based on discourse structure and knowledge-based methods were applied to construct the automatic Ext/Sums (extracts & summaries) system from the straight news in web. The frame contains the slot and facet represented by the role of paragraphs, sentences, and clauses in news and the rules determining the type of slot. Rearrangement like Unification, separation, and synthesis of the candidate sentences to summary, maintaining the coherence of meanings, was carried out by using the rules derived from similar degree measurement, syntactic information, discourse structure, and knowledge-based methods and the context plots defined with the syntactic/semantic signature of noun and verb and category of verb suffix. The critic sentence were tried to insert into summary.

키워드 : 문맥정의, 보도기사, 역할, 웹, 의미범주, 자동요약
automatic summarization, context plot, rhetorical role, semantic category, straight news, web

* 이 논문은 2005년도 전북대학교의 연구비에 의하여 연구되었음.
** 전북대학교 인문대학 문헌정보학과 교수 (taehyun@chonbuk.ac.kr)
■ 논문접수일자 : 2006년 11월 15일
■ 게재확정일자 : 2006년 12월 6일

1. 서 론

1.1 연구 목적

사람의 힘을 빌리지 않고 컴퓨터가 초록(이후 요약이라고 칭함[1]) 또는 색인을 자동으로 작성하여 주는 명제는 성공여부를 떠나서 오랫동안 매력적인 목표였다. 지금도 그 매력과 필요성은 여전하여 마이크로소프트 워드의 '도구' 패널에 장치되어 있는 '자동요약'처럼 여러 발췌문/요약[2] 시스템들이 출현하고 있다.

이러한 발췌문/요약 시스템들은 학술정보와 생활정보 전반이 인터넷 웹에 수용되어지는 현재의 지식정보사회의 환경에서 더욱 유용한 도구로 사용 될 수 있다. 현재 웹에는 학술논문, 학술 잡지기사, 일반 잡지기사, 신문·방송 뉴스, 일반 논(설)문, 특허, 보고서, 회의록, 탐색 결과물, 산문, 메모, 도표, 표, 공식, 그림, 동영상, 광고 등 다양한 형태의 자료들이 존재하며, 그것들은 다중문헌(multi document)적이며 다중미디어(multimedia)적이다. 한 주제 또는 사건에 대해 여러 사이트에서 기록하고 있으며 또한 표현 양식이 멀티미디어로 다양하게 작성되어 있고 그 양이 지식정보사회란 이름에 걸맞게 상상을 뛰어 넘을 정도로 많다.

Yahoo와 같은 세계적인 포털 사이트에서 단일 키워드로 질문을 하면 검색되어 출력된 문서나 사이트들이 1억 개를 넘는 경우가 비일비재하다. 그러므로 많은 정보들 중에서 중요한 정보의 발췌와 그것을 정리 요약하는 행위는 필연적으로 나타날 수밖에 없고 지식정보사회의 가치가 큰 IT 상품이라고 할 수 있다.[3]

그동안 발췌문/요약의 후보문장을 본문에서 선정하고 다듬는 방법론은 여러 가지 측면에서 모색되어 왔는데 그 내용을 정리하면 위치와 단서 및 단어의 출현 빈도에 근거한 코퍼스적 방법, 그리고 장르별 본문의 담화(수사) 구조에 의한 방법, 언어적 지식을 이용한 방법으로 크게 나누어진다(Mani 1999).

본고는 웹 문서들의 자동 요약 시스템을 구축하기 위한 시도로서 이와 같은 단서, 위치, 수사적 역할을 중심으로 웹 사이트에 게재되는 뉴스기사 유형 중 보도기사의 요약시스템을 정립하고자 하였다. 또한 Mani(2001)가 다루었듯이 자동요약에 있어 다중 문헌, 다중 미디어 문제를 고려한 비평문장을 제언하였다.

1.2 연구 방법과 제한점

본 연구를 위하여 한국어로 작성된 인터넷 웹 사이트인 "naver, empas, yahoo, daum, chol, nate, google, netian, unitel, dreamwiz"에서 10개의 제목을 선정한 후 각 제목별로 링크되어 있는 보도기사 5개씩을 표본으로 적출하였다. 표본들의 4/5는 통제집

1) 현재 '요약(summary)' 또는 '요약문(summary)'이란 표현이 학술잡지 기사에서 초록을 대신하여 많이 쓰이고 있으며 Jones(1999, 5) 같은 이는 요약도 지시요약, 정보요약, 비평요약으로 구분하였다.

2) 국내에 "웹요약1.1, 다이퀘스트, 모비코, 사이냅" 등의 웹문서 요약시스템이 있었다.

3) 미국 상무성(U.S Department of Commerce)은 IT 기술들을 현대 직업 시장에서 생존기술(survival skills)이라고 하였으며 영국의 교육·고용성(Department of Education and Employment)은 영국의 정보사회 전략이라고 하였다.(Talja 2005, 13) 자동요약도 현대 IT 기술의 한 부분이라 사려된다.

단, 1/5은 실험집단으로 나누었고 통제집단에 속한 표본들은 어휘사전, 글 프레임의 슬롯 종류와 패싯 종류, 그리고 제 판별 규칙을 작성하는 데에 사용하였다. 실험집단으로는 요약을 작성하여 문헌정보학 전공자 3인에게 5단계 리커드척도로 문장들의 응집(Cohesion)/일관성(Coherence), 내용파악의 편이성(간결성), 완성도(대변성)를 질문하였다.

프레임의 구조를 이루는 슬롯과 패싯들은 원문의 문장과 절을 발췌하고 분석하며 요약 문장으로 생성하는데 사용하였고 이 슬롯과 패싯을 파악하고 명명하기 위하여 담화, 언어지식에 기반한 여러 가지 방법론 중 어구의 수사적 역할, 최고 상위 범주어, 이야기의 줄거리 단위들을 중점적으로 사용하였다. 또한 주요 대목의 발췌 및 문장 재조립을 위하여 보도기사의 육하원칙과 '즉시성'과 같은 스토리 기본요소들을 활용하였다. 그리고 문장을 통합, 분리, 합성하는데 도움을 주고 요약문의 응집/일관성을 높이는데 기여하는 문맥정의 리스트를 도입하였다.

웹 기사들은 한 문장을 쓰고 줄을 바꾸어 다음 문장을 쓰는 형식이 많이 출현하여 문단을 구분하는데 어려움이 있는 관계로 본고에서는 줄바꿈을 문단 구분의 기준으로 삼았다. 또 웹에 출현하는 글들은 다중적이며 진위가 분명하지 않은 것들이 많기 때문에 비평적인 새로운 문장을 생성하기 위하여 미리 구조화되어 있는 비평정보를 제언하였다.

본고에서 적용한 명사와 용언의 의미 범주, 문장/문단의 역할 등 많은 부분이 주관적인 관점에서 진행되는 제한이 있었으며, 기호 ⇔ 는 두 사항이 서로 매칭 대상으로 올랐을 때, ㅗ는 상대 요소들에 영향을 미칠 때, /는 '또는', ㄱ는 '반대'를 나타낼 때 사용되었다.

2. 웹과 기사 정보

2.1 웹 정보의 특징

웹 사이트에 등재되어 있는 정보들은 지리와 언어적으로 다양화되어 있고 지수적인 증가(Lawrence and Giles, 1999, 107)를 하고 있으며 여과와 비평이 없는 출판의 자유로 인해 질, 안정(보존)성, 사이트 구조의 복잡성, 메터데이터의 포맷 문제를 안고 있다. 또한 학술 저작이 PostScript, PDF, MSWord로 출판되는데 이 학술자료들을 여과 검색하는 구조적 정보 또는 메터데이터들이 거의 없었다(Lawrence, Bollacker, and Giles 1999). 0.3%의 사이트들만이 더블린 코어 표준형식으로 메터데이터를 기록하고 있을 뿐이며 1999년 2월 기준으로 보유정보의 16% 이상을 색인한 검색엔진은 하나도 없었고(Lawrence and Giles 1999, 109) 그나마 키워드 또는 태그를 언급하고 있는 사이트는 43% 미만이었다(Drott 2002, 211).

현재 잘 가동되고 있는 Altavista와 Google 같은 상업적 검색엔진에서 웹 저작물의 획득과 색인작성 및 적합정보 서열화에 사용되는 절차와 알고리듬은 정보검색계에서 그대로 통용될 수 없는 소유주(시스템)의 방식이다. 이런 환경 하에서는 탐색엔진들이 이용자에게 왜곡되지 않은 정보(skewed selection)를 제공한다고 믿을 수는 없다(Jepsen et al. 2004,

1239-40). Allen(1999) 등의 연구에 의하면 'Northern.com' 검색엔진에서 탐색한 결과, 500 사이트 중 12-46%가 적합, 10-34%가 부정확, 그리고 20-35%가 오류(misleading)로 판명되었다.

2.2 보도기사 기술 특성

자동요약에서 기사들은 육하원칙(<표 1>의 밑줄 친 부분 참조)의 틀로 FLUMP, SUMMONS, MUC-4 system 등 여러 시스템에서 정형화되어 왔다(Moens 2000, 141). SUMMONS는 입력 기사의 요점 항목들을 담고 있는 템프릿 집합으로부터 메시지 이해시스템을 거쳐 요약을 생산한다. 이들 시스템은 입력되는 기사로부터 특정 정보 조각들을 발췌하였다. MUC-4 systems은 테러리스트 영역을 운영하면서 기사로부터 "가해자, 희생자, 사건 유형"과 같은 필드영역의 정보를 발췌하였는데 템프릿 당 25개 필드가 있었다(Radev and McKeown 1998).

보도기사는 기사들 중에서 논평이나 작성기자의 의견을 넣지 않고 어떤 사실을 있는 그대로 보도하는 스트레이트(Straight) 기사가 대부분이다. 보도뉴스의 기사문장에는 특히 6하원칙의 요소가 강조되며 여러 가지 내용들을 자유스럽게 전개하지만 <그림 1>에서 암시하고 있는 바와 같이 첫 문장과 그 다음 문장들 간의 관계는 아래에 적은 (1)~(3)과 같은 특징을 갖고 있다.

(1) 통상 첫 문장/문단에는 기사를 이끌어 가는 요약적 리드(Summary Lead) 역할을 하는 문장 또는 문단이 두괄식 형태로 나타나고 이하 다음 문장/문단은 리드에 대한

<표 1> 기사문장에 포함된 육하원칙 요소 예

요소	문장의 육하원칙 요소 실제 값
Who	A smoke jumper extinguished a blaze and prevented a forest in Gallatin National Forrest, Wyo..yesterday by diverting a mountain waterfall over a burning tree.
What	A burning tree didn't become a fore fire in Gallatin National Forest. Wyo., yesterday because a smoke jumper diverted a small mountain waterfall.
When	Yesterday a smoke jumper prevented a forest fire in Gallatin National Forest. Wyo., when he diverted a small mountain waterfall over a blazing tree.
Where	In Gallatin National Forest, Wyo., a smoke jumper yesterday diverted a small mountain waterfall over a blazing tree.
Why	To prevent a forest fire in Gallatin National Forest, Wyo., a smoke jumper yesterday diverted a small mountain waterfall over a blazing tree.
How	By diverting a small mountain waterfall over a blazing tree in Gallatin National Forest, Wyo, yesterday, a smoke jumper prevented a forest fire.

(http://www.tpub.com/content/photography/14130/css/14130_28.htm)

SUMMARY LEAD	EGLIN AFB, FLA (NNS)-A Navy officer who had never before taken control of an aircraft brought an Air Force spotter plane in for a rough but successful landing recently.
FACT 1 (bridge)	The incident came about after the pilot died of a heart attack during a routine training mission over the Gulf of Mexico.
FACT 2	Lt. John G. Graf, USN, of Aurora, III., walked away from the emergency landing only "slightly shaken up." The incident occurred in an area 60 miles southwest of Eglin Air Force Base.
FACT 3	Graf took control of the single-engine plane and returned the aircraft to Eglin.
FACT 4	Presently assigned to Eglin as a Navy liaison officer, Graf reported to his present duty station last July.
FACT 5	A former enlisted man, the 39-year-old officer served as an aerial photographer for several years and his general familiarity with aircraft is credited with helping him land the plane.
(http://www.tpub.com/content/photography/14130/css/14130_34.htm)	

〈그림 1〉 스트레이트 기사 스토리의 도해

〈표 2〉 뉴스 스토리 기본 요소의 분석

뉴스요소	강도	정당화(Justification)
즉시성(Immediacy)	Strong	Accident occurred this morning. Story will be released this afternoon.
근접성(Proximity)	Strong	Accident occurred locally. Squadron and pilot are attached locally.
결과/결론 (Consequence)	Weak	Measures will undoubtedly be taken to prevent further recurrence of this type, but this one incident in itself does not affect a grate number of people.
투쟁(Conflict)	Weak	The pilot's struggle for survival is worth mentioning, but more details are necessary to make this element strong.
기이성(Oddity)	Very Strong	Northing like this has been recorded before.
성(Sex)	None	----------
감정(Emotion)	Very Weak	The reader will sympathize with the pilot, but not beyond the extent one human being sympathizes for another human being in an unfortunate situation.
현저(Prominence)	None	The pilot is not widely known.
의혹(Suspense)	Weak	Although the facts, as presented here, do not lend themselves to suspended interest, the story has a certain amount of drama and suspense.
진행(Progress)	Weak	Progress in aviation may eventually result from this situation, but there is nothing in the facts that will improve mankind's health, comfort or happiness.

(http://www.tpub.com/content/photography/14130/css/14130_21.htm)

다리(Bridge) 역할의 내용들이 출현한다.

(2) 동일한 의미를 갖는 문구들이 반복적으로 나타나는 경우도 있다.

(3) 응집성과 일관성이 결여된 문장/문단들

이 출현하여 대표격인 문장을 끄집어내어 요약을 하여야 할 경우도 생긴다.

이에 덧붙여 TRAMAN 시스템에서는 '뉴스 스토리 학습목표'(News Story Learning Objective)의 기본요소들로 〈표 2〉와 같이 10개의 범주를 제시하였다

3. 담화 · 지식 기반 요약

3.1 위치 기반 예

Hovy와 Lin(1999, 85)은 SUMMARIST에 장치할 위치모듈을 연구하면서 다음과 같은 것을 발견하였다. 컴퓨터에 관한 13,000 신문기사를 갖고 있는 Ziff-Davis 코퍼스에서 최적위치정책(Optical Postion Policy)은 [TI, P2S1, P3SI, P4SI, P1SI, P2S2, {P3S2, P4S2, P5SI, P1S2}, P6S1, ...] 순으로 되는 것을 알았다. 즉 표제가 가장 유사한 (많은) 논제들을 품고 있고, 다음으로 제 2 문단의 첫 문장, 그 다음은 제 3 문단의 첫 문장 순이었다. 반면에 the Wall Street Journal에서 최적위치정책은 [TI, P1S1, P1S2. ...]의 순이었다.

Strzalkowski(1999, 142) 등은 뉴스의 시작점에 가까운 문단일수록 보다 많은 내용을 싣고 있다고 하였다. 그러나 이 순서는 편집지향 뉴스 문일 때는 역으로 될 수 있다고 하였다.

3.2 담화-지식 기반 방법

Mani(2001, 16)는 언어분석 수준이 형태소론, 구문론, 의미론 담화론으로 크게 구별할 수 있다고 하였다. 본 장에서 서술하는 방법은 주로 구문과 의미론에 관련된다.

요약을 위해서 확장구의 분석과 반복어구를 처리하는 과정에서 Boguraev와 Kennedy(1997, 104)는 단어의 돋보이는 요인들을 "SUBJ(term)=100 iff term is a subject"와 같이 주어, 목적어, 보어, 소유격, 수식어구 등으로 구별하였다.

Marcu(1999, 124-125)는 Mann과 Thompson(1988)의 수사구조론(Rhetorical Structure Theory) RST가 담화구조에서 과거 10년간 가장 인기 있는 항목이었다고 말하고 이 RST를 이용하여 담화나무가지(discourse trees)를 개발하였다. 수사적 역할로는 "상술, 설명, 배경, 예, 양보, 대비(contrast), 증거, 원인, 대조(antithesis)" 등이 설정되었다.

그리고 자동요약에 담화구조적 방법을 시도하였던 Teufel과 Moens(1999, 160-164)는 수사역할 자질을 〈그림 2〉와 같이 16가지로 분류하였다.

한편 WordNet에서는 명사와 동사의 최상위 계층 범주어로 〈그림 3〉과 같은 어구들을 선정하였다(김영택 2001, 146-147).

Lehnert (1999, 178-180)는 〈그림 4〉와 같이 줄거리 단위(Plot Unit)로 이야기들을 요약하는 방법을 제시하였다. 그리고 기본 영향 상태-추론을 전제로 하지 않고 있는 그대로의 상황-를 '+'(긍정적 사건; 즐거운 사건),

배경, 논제, 관련연구, 목적/문제, 해결, 결과, 결론/요구, 해결-목적/문제, 해결-결론/요구, 목적/문제-결론/요구, 목적/문제-관련연구, 목적/문제-배경, 결론/요구-관련연구, 결론/요구-결과, 배경-관련연구, 가 있으며 어떤 특정 수사적 역할을 예견하지 못하는 구들을 위해서 제로 값을 준비하였음

〈그림 2〉 수사역할 자질 예

명사 : 동작, 동물, 인공물, 속성, 신체부위. 인지, 커뮤니케이션, 사건, 감정, 음식, 집합, 위 치, 동기, 자연물, 자연현상, 인물, 식물, 소유, 과정, 수량, 관계, 형상, 상태, 물질, 시간
동사 : 신체 기능과 치료, 변화, 커뮤니케이션, 경쟁, 소비, 접촉, 인지, 창조, 동작, 감정/심리, 상태, 지각, 소유, 사회 상호작용, 날씨

〈그림 3〉 WordNet 범주어 예

동기, 성공, 실패, 변심(Change of Mind), 손실(Loss), 잡다한 행운(Mixed Blessing), 인내(Perseverance), 해결, 숨은비밀 행운, 허용(Enablement), 부정 교환(Negative Trade-off), 복합 긍정 사건(Complex Positive Event), 문제, 긍정 교환, 복합 부정 사건

〈그림 4〉 줄거리의 단위들

'–' (부정적 사건; 불쾌한 사건), 'M' (마음상태; 중립적인 정서의 마음상태)과 같이 세 가지로 정하였다. 실례는 아래의 〈문맥 /문장 예1〉과 같다.

이들 줄거리 단위는 서로 영향을 미치는 상태와 인과적 연결을 가지고 있다. 예를 들면 '문제' 의 경우를 나타내고 있는 문장 "Your dog dies and you long for companionship." 은 아래 〈그림 5〉와 같이 '부정' 에서 '정신상태' 로 연결되어 가는데 '동기' 가 작용되었다는 표시이다. 기본 줄거리 단위들이 결합하여 많은 수의 복합 줄거리 단위를 만든다.

〈문맥/문장 예1〉 :

– the car won't start
M John wants to get it started
+ Paul gets it started

문제
– → M
m

〈그림 5〉 줄거리 연결 예

일반 작문에서 문단은 한 소 주제를 나타내기 위하여 소 주제문, 큰 뒷받침문장들, 작은 뒷받침문장들이 서로의 의미를 연결하고 있는 것으로 파악된다. 문단을 이루는 문장들을 연결시켜 주는 자질은 응집성(Cohesion)과 일관성(Coherence)인데 응집성은 문장들의 긴밀한 연결관계를 제대로 설명해 주지 못하는 단점이 있어 글을 이루는 문장들의 의미 내용의 연결성은 일관성을 지표로 한다.(이삼형 1994, 26)

따라서 초록이나 본문의 일관성을 형성하는 관계의미의 종류는 관점에 따라서 무한에 가까울 수 있지만 기본적인 문단의 관계의미로 압축할 수 있다. 이삼형(1994, 63-90)이 설정한 종류는 "수집, 부가, 공제, 인과, 이유, 비교/대조, 상세화, 문제/해결, 초담화(meta-discourse)" 와 같다. 그리고 절을 합성할 때 길잡이 역할로 문맥정의 리스트(이태영 2005,

프레임 : 보도기사

클래스 : 기사

인스턴스 :

구성소 :

슬롯	패싯	값
리드	IN-L1	육하원칙 요소
문단역할		리드, 과정, 연관, 원인, 결과, 전망, 방안, 평가
문장역할		상황, 과정, 원인, 결과, 방법, 전망
문단성질		수집, 부가, 공제, 인과, 이유, 비교/대조, 상세화, 문제/해결, 초담화
문장성질		즉시, 근접, 결과, 투쟁, 기이, 성, 감정, 현저, 의혹, 진행
단어역할		물질, 추상, 동작, 형용 등
단어범주	IC-L1	물질, 추상, 동작, 형용 등
판별규칙		문단, 문장 절 판별규칙
발췌규칙		발췌규칙
요약규칙		조립규칙, 합성규칙

〈그림 6〉 보도기사 프레임 구조

159)를 사용하기도 하였다.

4. 프레임과 슬롯 구조

4.1 프레임 구조

보도기사의 프레임에 속할 내용들을 알기 위해 '3.2' 장에서 거론된 담화 · 지식기반 방법의 지식을 가지고 표본으로 선택된 40 개의 보도기사를 분석하였다. 보도기사 프레임은 〈그림 6〉과 같은 '인스턴스'와 〈그림 7〉과 같은 '클래스'의 두 가지 구조로 작성되었다. 기사 프레임의 인스턴스인 보도기사 프레임은 슬롯 구성소로 "리드, 문단역할, 문장역할, 문단성질, 문장성질, 문장요소, 단어범주, 판별규칙"이 등록되어 있고 보조적인 정보로서 'if-needed'

프레임 : 보도기사

클래스 :

인스턴스 : 보도, 논설, 기획/해설, 인터뷰/대담, 신변잡기

구성소 : 인스턴스의 슬롯 종류와 동일함

〈그림 7〉 기사 프레임

와 'if-changed' 패싯을 소유하였다.

〈그림 6〉에서 패싯 란에 있는 'IN-L1, IC-L1'은 "if-needed, if-changed" 패싯을 지칭하는 것으로 'IN'은 'if-needed'를, 'IC'는 'if-changed'를 뜻한다. '-' 이하의 기호들은 조기성 있는 구별번호들이다. 보도기사 프레임은 〈그림 7〉의 기사프레임에 종속되어 있지만 자체 내에서 발생하는 정보를 추가로 보유한다.

'문장 역할'에 출현한 "상황, 과정, 원인, 결과, 방법, 전망"은 각 '문단역할'에 있는 항목(문단프레임에서는 슬롯)에 공통적으로 적용되었다.

4.2 슬롯과 패싯 정보

〈그림 6, 7〉 보도기사 및 기사 프레임에서 나타난 슬롯과 패싯의 값 중 내용을 규정할 필요가 있는 항목 및 규칙을 따로 〈표 3〉과 〈표 4〉에 기술하였다. 이 항목과 규칙은 문단이나 문장 프레임에서는 〈그림 8〉처럼 슬롯이나 패싯으로 적용된다. 여기서 패싯에 IN 이나 IC가 표시될 때, '값'은 슬롯이 아니라 패싯의 값이 된다.

문장에서 "시간, 장소, 원인/이유, 방법" 등을 나타내는 문구를 식별하고 문장의 역할을 정의하기 위해서는 단어(구) 사전에 각 단어(구)들의 의미범주가 〈그림 3〉 WordNet 형식으로 부기되어 있어야 한다.

그리고 문장에서 '주체, 객체'를 알려면 조사의 격이 있어야 한다. 이것들은 아래의 〈표 5〉와 같은 어휘사전을 통하여 확인할 수 있다. 어휘사전에는 문장에 대한 구문 및 의미 분석을 하기 위하여 "품사, 역할, 범주, 유의, 단서, 시제, 존칭, 상위, 하위, 동의"를 표기하였다. 본 시스템에서 채택한 품사는 편의상 명사, 용언, 부사, 조사, 수사, 관형사, 대명사로 결정하였다. '역할'은 절/문장/문단의 수사적 역할을 나타내는 것으로 〈그림 6, 8〉, 〈표 3〉에서 보이듯이 "리드, 상황, 원인, 결과, 결론, 전망" 등이 있다.

'범주'는 단어의 성질과 소속관계를 나타내

〈표 3〉 적용된 문장과 문단 슬롯의 정보 예

슬롯	식별성 범주와 실제단어(구) 내용
상황	~고 있다, 상태, 상황 등 〈표 6,7〉 참조
과정	과정, 진행, 이와 관련 등 〈표 6,7〉 참조
원인	원인, - 때문에, -면서, 등 〈표 6,7〉 참조
결과	결과, 이에 따라 등 〈표 6,7〉 참조
방법	방법, 기법, 방안, 등 〈표 6,7〉 참조
전망	전망, 예측, 내다보다 등 〈표 6,7〉 참조

〈표 4〉 적용한 패싯의 정보 예

패싯	내용
IN-L1	문맥정의
IN-L2	고유명사
IN-L3	단서 용언
.	.

프레임 : 문단/보도기사

슬롯	패싯	값
리드	IC-L5	상황, 과정, 원인, 방법, 결론
과정	IC-L11	과정, 원인, 결과
.		

프레임 : 문장/리드/보도기사

슬롯	패싯	값
상황		〈표 3〉과 동일
원인		.,

〈그림 8〉 문단과 문장 프레임

〈표 5〉 명사·용언 어휘사건 예

단어	품사	역할	범주	유의	단서	시제	존칭	장르	상위	하위	동의	결합
김치	명사		음식									
목표	명사	목적	논증/증거	상	0							
바뀐	용언		상황			현재						수식
사건	명사		문제/해결	상	0			보도				
사려	명용	결론	문제/해결	상	0							
습니다	어미						0					
왔	용언					과거						
의	조사	소유격										명사
척추동물	명사		동물						동물	개,고양이	등뼈동물	

〈표 6〉 명사와 용언의 의미 범주

종 류		항 목 요 소
명사	물질	물질 : 인물, 기관, 속종(종족, 속명, 종명), 장소, 상품, 금전, 글, 의복, 음식, 건물/구조물, 도로, 도구, 컴퓨터, 운송도구, 육지, 바다, 식물, 동물, 무기, 매체, 작품, 시민, 기계, 유물/유적, 자연, 인공(도시, 마을 등) 등
	추상	추상 : 감각(시각, 청각, 환청 등), 이론/실제(사상 등), 문제/해결(논제, 의논, 의견, 토론, 외교 등), 방법/규칙(법,법칙,방법,제한), 논증/증거(목적, 결론, 결과, 원인 등), 환경/배경, 학문, 과학/기술, 언어, 역사, 문화, 예술, 전쟁/평화, 영토/영역, 언론, 정부, 기업, 기관/단체, 스포츠, 가격/값, 서비스, 시간, 과정(과정,진행,진화), 사건 등
용언	동작	동작 : 통신(부르다, 말하다, 손짓, 전화, 전달 등), 운동/이동(움직이다, 경기, 가다, 긁다, 때리다 등), 식사/배설, 인식(보다, 맡다, 알다, 깨닫다, 반응 등), 강의/학습, 표현(나타내다, 쓰다, 그리다, 울다, 웃다, 쓰다, 발표, 비난 등), 질문/대답(묻다, 응답 등), 시작/마감(열다, 닫다 등), 요청/부여(투정, 요구, 빌다, 빌리다 등), 작심(결심, 맹세, 선정, 선택 등), 처리(계산, 작성 등), 의논(토의, 음모 등), 설명/이해, 양보/강탈, 비교/대조, 생산(만들다 등), 논증(결론, 제시, 판단 등), 조정(정치, 화해 등), 수렴(여론조사, 투표, 보도 등), 기대(희망, 소원 등), 구경(관광 등), 전달(전염, 전파 등), 구축, 전망(예측, 예단, 추측, 추정, 전망), 진전(발달, 개발, 개척, 발전 등), 변화(바뀌다 등) 등
	형용	형용 : 존재(있다, 없다, 살다, 죽다 등), 색깔(빨갛다 등), 맛(맵다 등), 모습(귀엽다, 등), 모양(둥글다, 등), 냄세(구수하다 등), 마음(기쁘다, 아프다, 분노하다, 희열하다 등), 상황(좋다, 어렵다 등), 상태(높다, 중요하다, 맞다 등), 속도(빠르다 등), 선(가늘다 등) 등

는데 절/문장/문단의 역할을 판별하고 문장을 조립하는데 사용된다. 대표적으로 육하원칙에 해당하는 "시간, 장소, 원인(이유), 방법"이 있다(<표 6 참조>). '유의'는 "상, 중, 하"로 표시되는 그 단어의 중요도에 대한 가중치를 말한다. '단서'는 문장의 역할을 구분할 때 단서가 되는 뜻이나 표징을 갖는 단어를 구별하는 표식이고, '시제'는 "과거, 현재, 미래"로 구분되며 '존칭'으로 존대말을 구별하였다.

시소러스에서 표시되는 계층관계의 상위어는 '상위'에 기록하고, 하위어는 '하위'에, 동의관계에 있는 동의어는 '동의'에 기록하였다. 명사 개념간의 연결을 꾀하는 조사는 격을 구분하여 '역할' 란에 기재하였다.

5. 판별과 발췌 규칙

5.1 유형 판별

요약시스템에서 새로 입력되는 기사의 유형 판별은 전장에서 정하였던 문단의 슬롯을 이용하였다. 기사 글에 나타나는 문단들의 역할자질을 파악하여 글 프레임의 슬롯으로 규정한 후, 이 슬롯(문단 역할자질)들의 구성관계를 분석하여 기사의 유형을 정하였다. 이 유형 식별 과정에서 따옴표로 표시되는 대화체 문장을 문단슬롯과 함께 유형 판별의 중요한 근거로 삼았다. 기사의 유형을 결정하는 기준은 다음의 <유형 판별 규칙 예>와 같이 생성규칙으로 작성되었다. 생성규칙에서 'IF' 조건문 안에 있는 굴림체로 쓰여진 '목적' 등은 '문단 역할자'이며 '첫 문단'과 같이 바탕체로 쓰여진 것은 실제 값을 나타내고 'THEN' 이하 결과문의 **'보도'**와 같이 진한 굴림체로 쓰여진 것은 기사 유형을 나타냈다.

5.2 슬롯 판별 규칙

5.2.1 문단 슬롯 판별 규칙

2.2장의 <그림 1>에서 보는 바와 같이 보도기사는 첫 문단/문장에 대부분 'summary lead'가 출현하였으며 그렇지 않을 경우는 두세 번째 문단에서 개요적 정보를 발견할 수 있었다. 그리고 마지막 문단에서 결론적 마무리 언급이 출현하곤 하였다. 이러한 현상은 3.1장의 Hovy와 Lin(1999), Strzalkowski(1999)의 연구에서 기술되었듯이 문단 및 문장의 역할을 파악하는데 '위치'도 좋은 길잡이로 작용할 수 있음을 보여 주는 것이다.

<유형 판별 규칙 예>

판별규칙 유1 : IF (첫 문단 = '리드') THEN **'보도'**
판별규칙 유2 : IF ('목적' ∧ '증명' ∧ '결과' ∧ '결론') THEN **'논설'**
판별규칙 유3 : IF ('기행' ∨ '정서' ∨ '감상' ∨ '회상') THEN **'신변잡기'**
판별규칙 유4 : IF (('주제' ∧ '인사' ∧따옴표 문장)∨('주제' ∧ '인사'))
THEN **'인터뷰/대담'**
판별규칙 유5 : IF ('대상' ∧ '현상' ∧ '전망')∧따옴표문장 THEN **'기획/해설'**

〈슬롯 판별 규칙 예1〉

* 문단 예
판별규칙 단1-1 : IF ((첫 문장) ∩ (육하원칙 ≥ 4)) THEN '리드'
판별규칙 단1-2 : IF ((첫 문단) ∩ (육하원칙 ≥ 6)) THEN '리드'
판별규칙 단2-1 : IF ((주절용언범주| '리드') = (주절용언범주|다음문장)) ∩ (((육하원칙| '리드') ∩ (육하원칙|다음문장)) ≥ 1) THEN '리드/보완'
판별규칙 단2-2 : IF (((주절용언범주| '리드') = (주절용언범주|다음문단)) ≥ x) ∩ (((육하원칙| '리드') ∩ (육하원칙|다음문단)) ≥ 1) THEN '리드/보완'
판별규칙 단3 : IF (((육하원칙| '리드') ∩ (육하원칙|다음문단 · 문장)) ≥ 3) ∩ (**'상황'** ∩ ('이다' 용언상당어구)) THEN '상황'
판별규칙 단4-0 : IF (**'방법'**) = 1 THEN '방안'
판별규칙 단4-1 : IF (주체|육하원칙) ∧ (방법∨규칙∨도구|객체|육하원칙) ∧ 통신 THEN '방안'
판별규칙 단4-2 : IF ((주체|육하원칙) ∧ (방법∨규칙∨도구|객체|육하원칙) ∧ 통신) ≥ 3 THEN '방안'
판별규칙 단5 : IF **'방법'** ∩ **'결과'** THEN '방법 · 결과'
판별규칙 단6 : IF (**'원인'** ∩ **'결과'**) THEN '문제해결'
(한문장이 한문단을 나타낼 때의 규칙은 -1에 속하는 것들임)

〈표 7〉 용언어미의 종류

종류	연결 어미	종류	연결 어미
+가정	-면, -거든, -(ㅏ)아도, -(ㅓ)어도, -라도	+과상	들고 있었다(과진)/들려 있었다(과상)
+과정1	-여, -여서, -해, -해서, 〈동사의 경우〉; -ㅓ(서), -ㅏ(서) -어(서) -아(서)	+과정2	-ㄴ(는)데, -다가, -(ㄴ)는 중에
+과정3	-게	+나열1	-고
+나열2	'고'로 대체될 수 있는 것; -며, -고서, ~하는가 하면,	+대립	-지만, -거늘, -되, -(으)나, -는데 불구하고, ~했으나, ~뿐만 아니라, ~ㄴ 반면
+목적	-려고, -ㄹ, -러, -도록, -라, -도록, -려면	+반반	-지~지, -던(든)지~-던(든)지
+불문	-(이)든(지) 간에, ~하더라도	+비교	-보다, -ㄴ(는) 것 보다, ~ㄴ 만큼,
+삽입	-ㅁ(이,을), -음(이,을), -기(가,는,를,로), -느냐가, -는 것(이,은 등), -(을)듯, -기 만을, -(하,라)고(는), -라는	+상태	-ㄴ 채(로, 로는 등), -는 데(에, 에 있어 등), -ㄴ(는) 거(것이)-, - 결과, -ㅣ(이,리)
+선택	-거나, -든지, -는가 / or	+수식	-ㄴ, -는
+시간	-기 전, -ㄴ(는) 후, -ㄹ 때(까지), ~에 앞서	+염려	-까봐
+원인1	-(으)니까, -기 때문에, -므로, -기에, 있는 만큼, ~함으로써	+원인2	-(다)니, -자, -면서, 〈형용사의 경우〉; -어(서)(는), -아(서)(는), -ㅓ(서)(는), -ㅏ(서)(는),
+의문	-면서도, -고서도, ~ㄹ 뿐더러, 있기나 한지	+전제	-아야, -어야, -(는)다고, -해야(만), -건만, -ㄹ 망정, -는 한
+종결	-다, -까, -오, -세, -게나	+현상	들고 있다(현진)/들려 있다(현상)
+동의	~가 까지	+동조	~하기도,
+반박	있지 않은, ~하지 않은	+누가	-ㄹ수록

문단슬롯 즉 문단의 역할자질은 일차적으로 문단을 구성하고 있는 문장들의 역할자질인 문장슬롯과 문장의 용언, 체언의 범주 및 의미로 결정하고 이차적으로 위치가 고려되었다. 문단슬롯의 예는 아래의 〈슬롯 판별 규칙 예1〉과 같이 작성되었다. 규칙 안에 있는 바탕체의 '첫 문장'과 같은 것은 단어 실제 값을 나타내고 굴림체의 '리드'는 문단 역할, 옥수수체의 **'원인'** 등은 문장 역할, 필기체의 '명승지' 등은 단어 범주를 나타내었다. 그리고 견고딕의 **'주제'** 등은 육하원칙의 "시간(언제), 장소(어디서), 주체(누가), 객체(무엇), 이유(왜), 방법(어떻게)"을 표시하였다.

5.2.2 문장과 절 슬롯 판별 규칙

보도기사 문장들이 글 내부적으로 표방하고 있는 바를 끌어내기 위하여 문장을 구성하고 있는 어휘 및 구들의 범주 및 의미를 도구로 삼았다. 예를 들면 "~~ -을 보여준다"와 같은 문장은 '결론'과 같이 마무리를 지을 때에 흔히 사용되는 표현이므로 '결론'이란 역할자를 받을 수 있다. 보도기사의 '문장 역할자'는 "상황, 과정, 원인, 결과, 방법, 전망"과 같이 간단히 유별하였다. 그 이유는 일반 기사들은 논문기사와 달리 역할 구분이 명확하지 않는 데에서 비롯되었다. 문장슬롯 즉 문장의 역할자질은 문장을 구성하고 있는 표징성이 강한 단어

〈슬롯 판별 규칙 예2〉

* 문장 예
판별규칙 장1 : IF (전망|주절 용언) THEN **'전망'**
판별규칙 장2-0 : IF **원인** ∩ **나열** ∩ (운동 · 이동|주절용언) THEN **'상황'**
판별규칙 장2-1 : IF **원인** ∩ **나열** ∩ **결과** THEN **'상황'**
판별규칙 장3 : IF (-이다|주절 용언상당어구) THEN **'상황'**
판별규칙 장4 : IF (방법/규칙|객체) ∧ 통신|주절 용언) THEN **'방법'**
판별규칙 장5 : IF (-부터|시간) ∧ (방법|주절 용언) THEN **'방법'**
판별규칙 장6 : IF (~을 보여준다) THEN **'결론'**
판별규칙 장7 : IF (원인 ∨ 결과 ∨ 결론) THEN **'상황'**
판별규칙 장8 : IF (도시명 ∨ 인물 ∨ 명승지 ∨ 강 ∨ 산 ∨ 유적) (~과로 이었다) THEN **'전달'**

* 절 예
판별규칙 절1 : IF (+과정1|용언어미) THEN **'과정'**
판별규칙 절2 : IF (+원인2|용언어미) THEN **'원인'**
판별규칙 절3 : IF (+나열2|용언어미) THEN **'나열'**
판별규칙 절4 : IF (+수식|용언어미) THEN **'수식'**
판별규칙 절5 : IF (-이다) THEN **'상황'**
판별규칙 절6 : IF 시간 ∨ (방법+나열2) THEN **'방법'**
판별규칙 절7 : IF (방법|주절용언) THEN **'방법'**
판별규칙 절8-0 : IF (형용|용언) THEN **'결과'**
판별규칙 절8-1 : IF (형용|&현진) THEN **'결과'**
판별규칙 절9-0 : IF (주체| '리드' ∨ 객체| '리드') ∧ (용언&과거) THEN **'결과'**
판별규칙 절10 : IF (~때문에) THEN **'원인'**

인 용언, 체언의 범주 및 단어 의미에 의해 아래의 〈슬롯 판별 규칙 예2〉의 문장 예와 같이 결정하였고 절은 〈표 6〉과 〈표 7〉 및 〈표 8〉을 참조하여 〈슬롯 판별 규칙 예2〉의 절 예와 같이 작성되었다. 여기서 '**얹인**' 같이 샘물체로 쓰여진 단어는 절 역할을 나타내고 있다.

5.3 발췌 규칙

보도기사 원문에서 요약문에 포함될 후보문장들을 발췌할 때 위치와 절/문장/문단 역할이 사용되었고, 2.2 장의 〈표 2〉 스토리 기본요소 중 "즉시, 근접, 기이, 현저, 의혹"을 나타내는 문구도 발췌의 단서로 활용하였다. 그리고 이에 덧붙여 고유명사와 관용 명사구 및 복합 용언구도 고려의 대상에 포함하였다. '대한무역진흥공사'와 같은 고유명사나 '-지 않으면 안된다'와 같은 복합 용언구들은 전자는 의미의 특정성에서, 후자는 내용을 강조함으로써 소속 문장의 유의수준을 높이는 역할을 하여 중요 문장 선정에 영향을 미친다. 고유명사와 관용 명사구 및 복합 용언구는〈표 8〉과 같이

〈표 8〉 복합 용언구 및 고유·관용 명사구

구분	명칭	용언구	명칭	용언구
첨어 (보조용언과 조동사 사용)	&미래	-것이다, -려고 한다	&당위	-야〈만〉 한다, -지 않으면 안된다
	&현진	-고 있다, -는 중이다, -가다	&습관	-곤 하였다, -기도 하였다
	&현완	-〈해〉버렸다, -적이있다, -고 있다	&희망	-고 싶다, -기를 원한다
	&부정1	-지〈도〉 않(못)하-, -지도	&권유	-는 것이 좋다, -ㄹ 필요가 있다
	&가능	-ㄹ 줄 알다(모르다), -ㄹ 수〈가〉 있다(없다)	&가정	-면 -ㄹ 것이다, -ㄹ 지도 모른다, -ㅆ을지도 모른다, -텐데
	&추측	-는 듯 하다, -는 것 같다, -ㄹ 것이다	&양태	-게 되다, -게 하다, -게 시키다, -게 지다 (곱게지다, -해 지다(깨끗해 지다)
	&허가	-〈하여, 해〉도 좋다, -어도 된다	&전달	-라고 한다
부정구 (부정사와 동명사 및 용언구)	&부정2	-기 위하여, -ㄴ 것	&상태1	-게 되다(하다, 시키다, 지다), -해 지다, -면 된다, -기도한다, ㄴ다고 한다, -하기만 하다
	&동명	-는 것을	&상태2	-단다, -잖아, -는 구나, 자구요, -는거(야, 지증), -는 건(가 등)
	&형동	("동+동, 형+형, 동+형, 형+동"일 때)-게~-다, -며~-다, -야만~-겠-	&종첩	-고~-다(기다리고 기다리다), -다고 ~생각하다(이에 준하는 용언), -면 -ㄹ수록
관용 용어구	&정당	-는 것이 당연하다	&것뿐	-ㄴ 다는 것 뿐이-
	&록할	-도록 할 수 있-	&밖없	-할 수 밖에 없다
	&뿐아	-ㄹ 뿐만 아니라	&않것	-지를 않는 것이-
	&않없	-하지 않을 수 없다	&적뿐	-ㄴ 적이 없(었)을 뿐만 아니라
	&지없	〈아무리〉 강조하여(해)도 지나침이 없다	&최것	최대한 발휘하는 것이다
	&할것1	-게 할(될) 것(거)이-	&할것2	-야 할(될) 것(거)이-
	&척뿐	-ㄴ(한) 척하는 것 뿐이-	&하안	-하여(해)서는 (하면, 하는 것은) 안된다(곤란하다)
	&할없	-할 것 없-	&목하	~을 목적(목표)으로 하(삼)
고유명사	인명, 국명, 상품명, 기관명, 단체명, 동물명, 식물명, 화학명...			
관용 명사구	관용구, 용언구, 명사구, 중요(주요, 핵심)문장, 정보검색, 초등학교, 중학교, 고등학교, 대학교, 고유명사,.....			

(여기서, 현진 → 현재명사, 현완 → 현재완료, 가능 = 불가, '형' → 형용사, '동' → 동사, 할것1 = 될것1, 할것2 = 될것2

〈요약 후보문장 발췌 규칙 예〉

발췌 규칙1-0 : IF '리드' THEN 3순위 후보
발췌 규칙1-1 : IF '리드' ∧ (길이 ≤ 23) THEN 2순위 후보
발췌 규칙1-2 : IF '리드' ∧ (길이 ≤ 23) ∧ (육하원칙 ≥ 4) THEN 1순위 후보
발췌 규칙2-0 : IF ('결과' ∨ '전망') THEN 3순위 후보
발췌 규칙2-1 : IF ('결과' ∨ '전망') ∧ '방안' THEN 2순위 후보
발췌 규칙2-2 : IF ('결과' ∨ '전망') ∧ '원인' THEN 2순위 후보
발췌 규칙2-3 : IF ('결과' ∨ '전망') ∧ 단서어 THEN 2순위 후보
발췌 규칙2-4 : IF ('결과' ∨ '전망') ∧ '방안' ∧ 단서어 THEN 1순위 후보
발췌 규칙2-5 : IF ('결과' ∨ '전망') ∧ '원인' ∧ 단서어 THEN 1순위 후보
발췌 규칙2-6 : IF ('**방법**' ∨ '**원인**') ∧ '**결과**' THEN 2순위 후보
발췌 규칙2-7 : IF ('**방법**' ∨ '**원인**') ∧ '**결과**' ∧ 단서어 THEN 1순위 후보
발췌 규칙3-0 : IF ('원인' ∨ '평가') THEN 4순위 후보
발췌 규칙4-0 : IF ('과정' ∨ '연관') THEN 5순위 후보
발췌 규칙5-0 : IF 단서어 THEN 3순위 후보
발췌 규칙6-0 : IF (고유명사(관용어구) ∨ 복합 용언구) THEN 3순위 후보
발췌 규칙7-0 : IF (원인 ∨ 결과 ∨ 결론) THEN 4순위 후보

정리하였다.

요약문에 포함될 후보문장들을 발췌하는 규칙은 〈요약 후보문장 발췌 규칙 예〉와 같다.

6. 요약 문장 작성

6.1 요약 문장의 특징

보도기사의 요약 문장은 아래의 〈그림 9〉에서 기술한 내용을 그대로 준수하여 작성하도록

1) 초록은 흔히 논제 또는 리드 문장으로 시작한다. 이 첫 문장이 표제의 반복이서는 안된다.
2) 초록자는 문구의 반복을 회피하고 되도록 이면 긴 문장의 사용을 저지해야 한다.
3) 모든 초록은 예외적인 길이로 만들어 질 수밖에 없는 것을 제외하면 한문단으로 기록한다.
4) 초록문에서 문장번호 매김과 한 문장 내에서 각 리스트(문구)를 구별하는 방식을 사용할 수 있다.
5) 모호한 단어와 용어는 회피하여야 한다. 모호성은 이용자의 지식배경에 달려 있는데 대체로 약어, 두자어, 상표명, 주제전문어(은어), 등은 잠재적인 혼란을 가져 올 수 있다.
6) 문장의 간결성을 최우선으로 한다.

예:	부적합	적합
	Ss in this study were	Ss were
	in order to	to
	in a similar fashion	similarly

7) 아이디어들의 표현형식과 순서는 큰 수정이 없는 한 저자의 서술을 따른다.

〈그림 9〉 초록문장의 특징

한다. 특히 간결성을 염두에 두어야 한다. 예를 들면, (1)술어를 표현하는데 있어서 "하지 않으면 안된다, 할지도 모른다"를 "하여야 한다, 할 것이다"로 표현을 단순화시켜야 한다. 그리고 (2) 직접, 간접화법을 사용하여 늘어나는 문장의 길이를 평서문을 사용하여 길이를 줄여야 한다. 예를 들어 아래의 〈화법 예〉에 있는 (1) 문장은 (2) 문장과 같은 형식으로 고쳐 써야 한다.

〈화법 예〉

(1) 기상청은 "아직까지 북태평양 고기압의 영향권 아래 놓여있어 무더위와 열대야가 이어지고 있다"며 "16일 이후 기압골이 한반도 상공을 한차례 지나고 나면 더위가 식을 것"이라고 밝혔다.

(2) 연일 찜통더위와 열대야가 이어지는 가운데 오는 16일쯤 전국적으로 비가 내린 뒤 더위가 한풀 꺾일 전망이다

6.2 문장의 통합과 분리

Mani(2001, 174)는 문서 내에 존재하는 중복적 글 요소를 네 가지 경우로 요약하였다. 여기서 글 요소를 문장이라고 가정하여 기술하면 다음과 같다.

(1) 의미적 대등(semantically equivalent) : 문장들이 서로 의미적으로 같은 내용을 말할 때,

(2) 단어 동일(string identical) : 문장들이 서로 똑같은 단어로 기술되어 있을 때,

(3) 정보적 대등(informationally equivalent) : 문장들이 서로 같은 내용의 정보를 포함하고 있다고 판단될 때,

(4) 정보적 함의(informationally subsumes) : 어느 한 문장이 다른 문장의 정보를 포함하고 있다고 판단될 때.

요약 후보문장으로 발췌된 문장들에서 이러한 중복 요소들이 출현하면 중복요소를 갖는 문장들을 통합하여야 하고 역으로 중복 요소들이 출현하는 문장들을 요약의 후보문장으로 상정하고 문장 통합을 수행하여야 한다. Salton(1996)과 Moens(1999), Schutze(1998)의 유사도 측정 공식 모형을 이용하여 두 문장의 유사도를 비교하고 유사도가 한계치를 넘은 문장들은 가중치가 큰 문장으로 통합하였다. 아래의 〈표 9〉에서 통합되는 예를 보여주고 있다.

통합의 반대로 분리도 고려하지 않으면 안된다. 요약 후보문장의 길이가 큰 것은 문장을 분리해야 한다. 문장이 비대해지는 것은 수식구(절), 대등구(절)-동격인 어휘-가 많거나 논리적인 절의 연결이 많아지기 때문이다. 따라

〈표 9〉 통합 규칙 예('+' → 통합, '|' → 둘중의 하나 채택, '^' → 둘다 채택)

	포함	포함	부가	부가/선택	선택	연접
①	A B	A B	A B	① A B C	①A B C D E	① A B
②	B	A	B C	② B C D	② F F G H I J	② C D
①+②	A B	A B	A B C	ABCD, (① | ②)	(①^②), (①|②)	① '고' ②

서 문장을 간소화하기 위해 이 원인들을 제거하거나 절들을 분리하였다.

(1) "그러나, 그리고" 등의 접속사 제거하고 부사와 관형사도 되도록 삭제한다.
(2) 수식 단어와 구 및 절을 제거한다.
(3) 숫자 및 수사를 제거한다.
(4) "'한다'고 말하였다" 류의 용언화 또는 명사화 삽입절은 첫 주어와 "다, ㅁ, 음, 기, 지" 이후를 생략한다.
(5) 출현빈도가 낮은 단어로 구성된 수식절과 과정절은 생략한다.
(6) "것, 척, "등으로 표현된 불완전 명사를 제거한다.
(7) 분리용 특정 용언어미 리스트에 있는 특정 용언어미가 출현할 때 분리시킨다.
예) 40단어 이상 출현한 "~고, ~며, ~서" 등으로 이어지는 종속절과 "~~, 즉, ~~ㄴ 반면"과 같이 인접형 이거나 '~~것 또한' 등 흡착내지 부가형 문장을 분리한다.
(8) "원인절/결과절, 긍정절/부정절, 방법절/결과절"로 묶여진 문장을 각각 원인과 결과 문장으로 독립시킨다.

6.3 문장의 합성

문장 합성은 한 줄거리 단위에 속한 여러 문장에서 문장 요소들을 발췌하여 그 요소들을 문장형식과 결합규칙에 맞게 조립하여 문장을 만드는 작업을 말한다. 문장 합성은 첫째, 절 단위 합성과 둘째, 단어 및 구 중심 합성을 도입하였다.

6.3.1 절 단위 합성

절 단위 합성은 서로 응집성이 강한 문장들 내에서 각 문장 절들을 구문/의미적 역할로 정리하고 가장 중요한 절들을 절 역할 순서에 맞게 결합시키는 방법이다. 예를 들면 〈그림 10〉에서 (5)와 (6) 번 문단/문장들은 '방법 · 결과'를 말하는 것들이며 샘물체로 쓰여진 **'원인·결과'** 등의 절을 갖고 있다.

이 문장들에서 ①{(서울경찰청은 이날 오전 4시20분부터 노들길~여의도진입로의 통행을 제한했으며) (서울 청계천 산책로는 이날 오전 비가 계속 내린다는 소식에 여전히 통행이 제한되고 있다.)}

②{(노들길~여의도진입로와) (한강시민공원 개화6란문과) (서울 청계천 산책로는) 통행이 제한되고 있다.}

③{(판당댐에서 초당 1만1천t을 방류하면서) (한강시민공원 반포·강서·망원지구는 완전히 침수됐다)}와 같이 ① '~ -고(며), ~', ② '-와(과), ~', 또는 ③ **'원인결과'**와 같은 합성을 유도할 수 있다.

6.3.2 어구 단위 합성

문장들이 입력되면 각 문장들을 최소자립형 태소까지 분해한 다음 그 형태소를 가지고 원문의 뜻을 살려 문장을 재조립하는 방법이다. 재조립이 원문의 내용을 대변하려면 분해될 때에 원 정보를 확실히 소유해야 한다. 즉 글 속에서 맡은 ① 문단, 문장의 역할을 파악하고, ② 문장에서 절이 맡은 역할을 용언과 용언어미를 통해 분석하고, ③ 단어의 역할(격)과 의미(범주) 및 타 단어와의 연결관계를 보존하였다. 그

연합뉴스) 성혜미 기자 =

(1) 28일 서울 · 경기 · 강원 · 충청 등 중부지역에 또 다시 많은 비가 쏟아질 전망이다. ← **전망**

리드

(2) 기상청은 중국에서 우리나라 중부지방으로 다량의 수증기가 유입되고 있어 (과정) 오전부터 장마전선이 다시 활성화돼 (과정) 서울 · 경기 · 충남북 · 강원지역에 50~100㎜, 많은 곳은 150㎜ 이상, 전북 · 경북 · 서해5도 · 울릉도 · 독도에 20㎜~60㎜이상, 경북에는 최고 100㎜ 이상 비가 내릴 것으로 내다봤다. (전망) ← **전망**

보완/리드

(3) 장마전선이 남하하면서 (원인) 이날 오전 0시부터 오전 5시 현재까지 충북 제천 81㎜, 안성 80.5㎜, 이천 87.5㎜, 충주 엄정면 71㎜의 비가 왔으며 (결과+나열) 전날 많은 비가 왔던 홍천은 12㎜, 양평 5㎜, 서울 4.5㎜, 동두천 0.1㎜가 내렸다. (결과) ← **상황** 27,28일 누적강수량은 홍천 257.5㎜, 서울 199.5㎜, 동두천 146.6㎜, 양평 145㎜ 등이다. (상황) ← **상황**

상황

(4) 기상청은 서울 · 인천 · 경기 · 강원(평창 · 횡성 · 홍천 · 춘천) · 충남 당진에 호우경보를, 서해5도 · 강원 · 충남(천안 · 아산 · 태안 · 서산) · 충북(충주 · 제천 · 진천 · 음성)에 호우주의보를 발령했다. ← **방법**

방안

(5) 서울경찰청은 이날 오전 4시20분부터 노들길~여의도진입로의 통행을 제한했으며 (방법) 잠수교와 한강시민공원 개화6관문의 통행도 전날 오후부터 금지했다. (방법) ← **방법** 팔당댐에서 초당 1만1천t을 방류하면서 (원인) 한강 잠수교의 수위는 현재 8.23m로 계속 높아지고 있으며 (결과+나열) 중랑천 월계1교의 수위는 15.45m로 (위험 수위인) 18.85m에 가까워지고 있다. (결과) ← **상황** 또는 **결과**

방법 · 결과

(6) 27일 오전부터 시민들의 접이 금지된 (방법+수식) 서울 청계천 산책로는 이날 오전 비가 계속 내린다는 (상황+수식) 소식에 여전히 통행이 제한되고 있다. (방법) ← **방법** 한강시민공원 반포 · 강서 · 망원지구는 완전히 침수됐으며 (결과+나열) 이촌 · 광나루 지구도 부분적으로 물에 잠겼다. (결과) ← **결과**

방법 · 결과

〈그림 10〉 보도기사 분석 표본 예1

리고 ④ 〈그림 11〉 리스트에 있는 '문맥 정의'와 '절 재조립 규칙'으로 문장을 정리하였다.

'문맥정의'는 본문 문장을 분석하고 다시 만드는데 도움을 주는 도구인데 〈그림 11〉처럼 절 단위로 이루어졌다. 보도기사가 갖는 대표적인 특징인 육하원칙의 요소들로 이루어진 명사와 용언의 범주를 주축으로 하고 역할별로 군집된 〈표 7〉의 용언어미 종류와 〈표 8〉의 복합용언구를 보조축으로 활용하여 작성되었다.

상기한 ①, ②, ③ 세 절차로 〈그림 12〉의 (1), (2), (3) 문장들을 분석하고 5.3 장의 '발췌규칙'에 따라 역할, 단서어(깜짝, 결과, 무조건, 전면), 리드의 의미어 빈도(정부(3), 신도시(4), 개발(2), 발표(3), 집값(2), 안정(2), 선정(2) 등)-5.3장의 예시에는 미출현-정보를 적용하여 주요 개념 을 선정한 후 위의 ④ 절차를 실행하였다. 그 결과 "㉠정부가 신도시 계획을 발표하-, ㉡집값 안정과 무관하-, ㉢수요 적합성을 무시하-, ㉣신도시 개발방식은 재검토해야 하-"가 발현될 수 있었다. 이 절들

〈문맥 정의 예〉

문맥정의 1-0 : {주제[은/는, (이)가], '용언범주' +[ ]&[ ]}
문맥정의 1-1 : {주제[은/는, (이)가], 계량[이다]+[ ]&[ ]}
문맥정의 1-2 : {주제[의] 주제[은/는, (이)가], 계량[이다]+[ ]&[ ]}
문맥정의 2-0 : {객체[을/를], '용언범주' +[ ]&[ ]}
문맥정의 2-1 : {객체[을/를], 장소/인공물[에/에서], '용언범주' +[ ]&[ ]}
문맥정의 3 : {주제[(이)가], 객체[을/를], '용언범주' +[ ]&[ ]}
문맥정의 4 : {주제[(이)가], 객체[을/를], 시간[부터/까지/에], '용언범주' +[ ]&[ ]}
문맥정의 5 : {주제[(이)가], 객체[을/를], 장소[에], '용언범주' +[ ]&[ ]}
문맥정의 6 : {주제[(이)가], 객체[을/를], 장소[에], 이유[를 위해/때문에], '용언범주' +[ ]&[ ]}
문맥정의 7 : {주제[(이)가], 객체[을/를], 장소[에], 시간[부터/까지/에], '용언범주' +[ ]&[ ]}
문맥정의 8 : {주제[(이)가], 객체[을/를], 장소[에], 시간[부터/까지/에], 이유[를 위해/때문에], '용언범주' +[ ]&[ ]}
.....
〈절 재조립 규칙 예〉

조립규칙 절1 : IF (a|A절), (b|A절), (c|B절), (d|A절), (e|C절) THEN a+b+d ← 가장 가까운 이웃끼리(원문의 같은 절) 재조립함
조립규칙 절2 : IF (a|A절) [의] (b|A절) THEN a+[의]+b
조립규칙 절3 : IF (주제/객체/보체)+(술어+ㅅ 식)+(체언) THEN (체언)+(주제/객체/보체)+(술어+종 결)
조립규칙 절4-0 : IF (X), (값), (술어) THEN X+값[이다]
조립규칙 절4-1 : IF (계량), (값), (술어) THEN 계량+값[이다]
조립규칙 절5 : IF (사건), (결과 1), (결과 2) THEN (사건)+(결과 1) ← 한 사건이나 대상에 대해 여러 가지 사실이 기록되어 있으면 첫 사실만 기술한다.
조립규칙 절6 : IF (a|A범주), (b|A범주), (c|A범주), (d|A범주), (e|A범주) THEN a+b+c [등]
조립규칙 절7 : IF (A절|원인) THEN A절+원인1/원인2

〈그림 11〉 문맥정의와 규칙 리스트

서미숙 기자 = {정부가 {집값 안정을 위해 (방법)} 발표한} (과정+ㅅ 식) {신도시 개발 계획이 도마 위에 올랐다.} (상황) ← **상황 ; A**

리드

{이번 신도시 발표는 처음부터 추병직 건설교통부 장관의 '깜짝 발표'로 물의를 빚더니}, (원인2) {{집값 안정과 무관한 곳에 신도시를 선정했다는} (과정+ㅅ 식) {비난까지 받고 있다.}} (상황) ← **상황 ; B**

보완/리드

{전문가들은 {"정부가 물량 확보에만 치중한 결과} (원인) {수요 적합성은 무시한 것 같다"}며} (결과+나열) {{"무조건 신도시를 지정하고 보자는} (방법+ㅅ 식) {현행 신도시 개발 방식은 전면 재검토해야 한다"} (방법) 고 지적한다.} (결론) ← **상황 ; C**

보완/리드

리드

.....

〈그림 12〉 보도기사 분석 표본 예2

로 문장을 만들 때, 원칙 "ⓐ원문에 출현한 순서대로 열거(수식절 예외), ⓑ앞 절의 사실을 반대하는 뒷 절이 출현하면 앞 절을 '-나'로 연결, ⓒ앞 절과 뒷 절이 동질이면 앞 절을 '+나열'로 연결, ⓓ뒷 절의 용언에 '-해야 한다'가 출현하면 바로 앞 절은 '-여'로 연결"을 따르면 위의 ㉠~㉤ 절들은 "정부가 신도시 계획을 발표하였으나 집값 안정과 무관하고 수요적합성을 무시하여 재검토해야 한다"와 같은 요약 문장으로 합성될 수 있다.

6.4 요약 실제

6.4.1 절 단위 요약 실제

절 단위 요약문의 작성을 〈그림 10〉을 대상으로 진행하였다. (1)4.2장 〈표 5〉에 소개된 바의 어휘사전으로 단어 속성을 파악하였다. (2) 5.2장의 슬롯 판별 규칙으로 절, 문장, 문단의 역할을 〈그림 10〉과 같이 파악하였다. (3)5.3장의 발췌규칙 '1-1'과 '2-6'을 적용하여 "리드 + 결과" 요약 방식으로 요약문을 작성하였다. (4)절 합성 절차는 〈절 단위 합성규칙 예〉를 사용하여 진행하였다.

> 28일 서울 · 경기 · 강원 · 충청 등 중부지역에 또 다시 많은 비가 쏟아질 전망이다. 노들길 · 여의도진입로와 잠수교와 한강시민공원 개화6관문과 청계천 산책로는 통행이 제한되고 있다. 팔당댐에서 초당 1만1천t을 방류하면서 한강시민공원 반포 · 강서 · 망원지구는 완전히 침수됐으며 이촌 · 광나루 지구도 부분적으로 물에 잠겼다.

〈그림 13〉 요약문 예1

〈그림 10〉에 보면 (1) 번과 (2) 번 문단/문장이 '리드'에 속하고 (5) 번과 (6) 번 문단/문장이 '결과'에 속한다. 리드에 속하는 (1)과 (2) 문장에서 '발췌규칙1-1'에 따라 문장의 길이가 23 단어를 넘지 않는 간결하면서 육하원칙 요소 4개 이상을 담고 있는 (1) 번 문장을 리드 문장으로 선정하였다. 그리고 결과에 속하는 (5)와 (6) 문단은 리드의 '의미어'와 '방법'과 '결과' 문장을 공히 갖고 있으므로 〈절 단위 문장 합성 규칙〉을 통해 합성을 시도하였다. 그러나 (6)문단이 단서어를 갖고 있기 때

〈절 단위 문장 합성규칙 예〉

합성 규칙 절1-0 : IF (((a절=**원인**)|(A문=**결과**)) ⇔ ((b절=**결과**)|(B문=**결과**))) THEN a + b
합성 규칙 절1-1 : IF (((a절=**원인**)|(A문=**결과**)) ⇔ ((b절=**결과**)|(A문=**결과**)) ⇔ ((c절=**결과**)|(단서어)|(B문=**결과**))) THEN a + c
합성 규칙 절1-2 : IF (((a절=**원인**)|(A문=**결과**)) ⇔ ((b절=**결과**)|(B문=**결과**)) ⇔ ((c절=**결과**)|(단서어)|(B문=**결과**))) THEN a + c
합성 규칙 절2-0 : IF (((a절, b절)|**x절역할**|**Y문장역할**) THEN a-고 + b-다
합성 규칙 절2-1 : IF (((a절, b절, c절)|**x절역할**|**Y문장역할**) THEN a-고 + b-며 + c-다
합성 규칙 절3-0 : IF (((a절=**결과**) ⇔ (b절=**결과**)) ⊥ 용언공통) THEN (체언부|a)와(과) + (체언부|b) + 공통용언부
합성 규칙 절4-0 : IF (((a절=**결과**) ⇔ (b절=**결과**)) ⊥ 체언공통) THEN 공통체언부 + (용언부|a)-고 + (용언부|b)
(여기서 '문' = 문장 또는 문단)

문에 (6)문단의 결과와 결합시켰고 요약문은 〈그림 13〉과 같이 생성되었다.

6.4.2 어구 단위 요약 실제

전 절 '6.4.1'에서 합성한 요약문장 중에서 "팔당댐....잠겼다"는 문장은 〈그림 10〉의 (5)번 문장에서 "한강 잠수교의 수위는 현재 8.23 m로 계속 높아지고 있으며 중랑천 월계1교의 수위는 15.45m로 (위험 수위인) 18.85m에 가까워지고 있다"라는 부분을 생략하고 있다. 따라서 불만족스러운 요약문의 소지가 있으므로 보완을 위해서 어구 단위 요약을 시도하였다. 먼저 '6.4.1' 절의 (1)~ (3)을 실행하고 (4) 문장들의 단어를 〈단어 정보 예〉와 같이 분석하였다.

이와 같이 분석된 단어들을 가지고 '절 재조립 규칙'을 적용하고 '문맥정의'를 참조하여 신조된 절들을 만들어 나간다. 예를 들면 〈단

〈단어 정보 예〉

(단어	역할,	범주,	절 역할,	문장 역할,	문단 역할,	빈도,	절 식별자,	문장 식별자)
(서울경찰청	주체,	정부,	방법,	방법,	결과,	1,	A;B,	A)
(여의도	객체,	장소,	방법,	방법,	결과,	1,	A,	A)
(잠수교	객체,	도로,	방법,	방법,	결과,	1,	B,	A)
(팔당댐	객체,	구조물,	원인,	결과,	결과,	1,	A,	B)
(방류하	방법,	이동,	원인,	결과,	결과,	2,	A,	B)
(접근	주체,	이동,	방법,	방법,	결과,	1,	A,	C)
(통행	주체,	이동,	방법,	방법,	결과,	3,	B,	C)
(수위	주체,	계량,	결과,	결과,	결과,	2,	B;C	B)
(한강	주체,	장소,	결과,	결과,	결과,	1,	B,	B)
(중랑천	주체,	장소,	결과,	결과,	결과,	1,	C,	B)
(8.23m	보체,	계량,	결과,	결과,	결과,	2,	B,	B)
(월계1교	주체,	도로,	결과,	결과,	결과,	2,	C,	B)
(15.45m	보체,	계량,	결과,	결과,	결과,	2,	B,	B)
(1만1천t	객체,	계량,	원인,	결과,	결과,	2,	A,	B)

〈어구 단위 문장 합성 규칙 예〉

합성규칙 어1 : IF ((A절|원문) , (B절|원문), (C절|원문), (D절|원문)) THEN (A+B+C+D|요약문) ← 원문에 출현한 순서대로 열거한다. 삽입절(수식절)은 예외이다.
합성규칙 어2 : IF ((A절|앞 절), (ㄱA절|뒷 절)) THEN (A절[-나] + B절)
합성규칙 어3 : IF ((A절|앞 절), (B절|뒷 절)) THEN (A절+나열1/나열2+ B절)
합성규칙 어4 : IF ((A절|앞 절), (B절|뒷 절)|[-해야 한다]) THEN (A절[-여] + B절)
합성규칙 어4 : IF ((A절|A절역할|A문장역할), (B절|A절역할|A문장역할), (C절|A절역할|A문장역할), (D절|A절역할|A문장역할)) THEN (A절+나열1/나열2+ B절[다.]) + (C절+나열1/나열2+ D절[다.])

28일 서울 · 경기 · 강원 · 충청 등 중부지역에 또 다시 많은 비가 쏟아질 전망이다. 노들길~여의도진입로와 잠수교와 한강시민공원 개화6관문과 청계천 산책로는 통행이 제한되고 있다. 팔당댐에서 초당 1만1천t을 방류하면서 한강 잠수교의 수위는 8.23m이고, 중랑천 월계교의 수위는 15.45m이다. 한강시민공원 반포 · 강서 · 망원 지구는 완전히 침수됐으며 이촌 · 광나루 지구도 부분적으로 물에 잠겼다.

〈그림 14〉 요약문 예2

어정보 예〉에서 '팔당댐'과 '1만1천t' 및 '방류하-' 단어는 '절 A', '문장 B'를 갖는 아주 가까운 이웃이며 '원인' 절이다. '문맥정의 2-1'과 조립규칙 절7 '로 절을 만들면 "1만1천t을 팔당댐에서 방류하 면서/니까" 등으로 정리될 수 있다. (5)번 문단의 생략 부분은 "①한강 잠수교의 수위는 8.23m이-, ②중랑천 월계교의 수위는 15.45m이-"와 같이 재조립이 되었다.

여기서 〈어구 단위 문장 합성규칙 예〉를 적용하여 〈그림 14〉와 같은 요약문을 생성하였다.

7. 평가 및 제언

문헌정보학 전공자 3인에게 문장 합성 시에 작성된 10개의 요약을 원문과 비교 분석하게 한 후, "1: 간결성, 2: 대변성, 3: 응집성, 4: 일관성, 5: 길이"를 5단계 리커드 척도(A:아주 충족, B:충족, C:보통, D:부족, E:아주 부족)로 평가토록 하여 〈표 10〉과 같은 결과를 얻었다.

〈표 10〉을 살펴보면 첫째, 다섯 가지 부분의 만족도는 평균적으로 보통 단계이었으나 일관성은 '부족'이었고, 둘째. 간결성이 높으면 대변성이 낮아지고, 셋째, 대변성은 길이와 상관이 없고, 넷째, 길이가 클수록 응집성과 일관성

〈표 10〉 합성된 요약의 평가

	1	2	3	4	5	A, B, D, E에 대한 분석
문헌1	B	D	C	C	B	1. 두 문장이며 각 문장이 육하원칙 요소가 네 개 이상이고 읽기가 편함, 2. 중간 연결 내용이 빠졌음, 5. 각 문장이 25 단어 이내에 들어 옴.
문헌2	C	D	C	D	C	2. 문단 역할 부여의 미비로 주요 '결과' 문단이 제외되었음, 4. '상황', '원인', '결과', '결론' 중에서 '결과' 부분이 제외되었음.
문헌3	C	B	B	D	D	2. 다섯 문장이 등장하여 본문을 대변함, 3. 핵심 내용이 모여 있음, 4. 각기 다른 역할 문단에서 발췌되어 의미흐름이 어색함, 5. 두 문장이 25 단어를 넘음.
문헌4	B	E	C	D	B	1. 문헌1의 1과 같음, 2. '리드' 문단이 빠졌음, 4. 문헌3의 4와 같음, 5. 문헌1의 5와 같음.
.						
.						
문헌10	D	C	C	C	B	1. ~했으며, ~시키고, ~하면 등의 절 어미연결이 과도하게 사용되었음, 3. 한 문장이면서 요약으로 적합한 길이임.

(여기서 '25' 단어는 보도기사 문장들의 평균길이 임)

이 작아졌으며, 다섯째, '절 단위 합성'이 '어구 단위 합성' 보다 일관성에서 성능이 좋았다.

응집성과 일관성을 높이기 위하여 절과 문장 및 문단의 역할에 대한 섬세하고 정확한 분류가 필요하며 간결성 있고 의미흐름이 유연한 요약을 생성하기 위한 어구단위 문장합성 규칙의 정제가 필요하다. 그리고 웹 기사는 진위가 분명하지 않은 점이 많기 때문에 비평이 필요하고 한 사건에 대해 여러 사이트에서 동시 다발적으로 다루기 때문에 비평 작성이 그 만큼 쉬울 수 있다. 다음에 비평문장을 작성하는 알고리듬을 〈비평 문장 작성 예〉와 같이 간략히 제언하였다.

〈비평 문장 작성 예〉

(1) 출현한 문장들에 정보를 간단히 덧붙인다.

①동일 사항의 결과들이 다르게 서술되었으면 "ㅆㅅ[이지만, 고하나 등] 확인을 요한다"를 덧붙이고 결과들이 유사한 사항이면 "와, 과, 또는, ~고"로 통합한다.

예1)

{한계리의 도로가 상당 부분 유실되었다. 한계리의 도로는 안전하였다} ->

{한계리의 도로가 상당 부분 유실되었다고 하나 확인을 요한다}

예2)

{한계리의 도로가 상당 부분 유실되었다. 한계리의 도로가 일부 유실되었다} ->

{한계리의 도로가 상당 부분 또는 일부 유실되었다}

②동일 사항의 결과들이 중복 서술되었으면 "(누차) 강조하였(되었)다"를 덧붙인다.

예3)

{한계리에 내린 비의 양은 800㎜이었다. 한계리에 쏟아진 폭우의 양은 800㎜라고 하였다}->

{한계리에 쏟아진 폭우의 양은 800㎜라고 누차 강조하였다}

(2) 적립된 문장 줄거리를 사용한다. 즉 수해, 비, 장마 등에 대한 지식베이스를 참조하여 비평을 작성한다. 예를 들어 생성규칙이 "IF (비, 장마, 유실) THEN 수해"라고 작성되어 있을 때, "비, 장마, 유실"에 대한 기사를 입력하면 미리 적립되어 있던 수해의 원인, 복구, 대비, 구호 등에 대한 문맥들이 작동한다.

①원인문맥 : ㉠'장ㅁ'의 '원인'을 제거하여야 한다

예4)

{한계리 도로의 급경사를 제거하여야 한다}

②복구문맥 : ㉠금번에 '장ㅁ'의 '주체'(인물)에게 당국이 적정한 장비와 비용을 지원해야 할 것이다

예5)

{금번에 강릉의 '수재민'에게 당국이 적정한 장비와 비용을 지원해야 할 것이다}

③대비문맥 : ㉠새로운 '방법'이 필요하다, ㉡'개체'를 보강하여야 한다
예6)
{새로운 물관리 시스템이 필요하다}
{제방을 보강하여야 한다}
④구호문맥 : ㉠'장수'는 작년에 이어 피해를 받은 지역으로 지원대책을 마련한다
예7)
{한계리는 작년에 이어 피해를 받은 지역으로 지원대책을 마련한다}

8. 결 론

웹 상에 올라오는 보도기사들을 요약하기 위하여 먼저 웹 기사와 보도기사의 특징을 파악하고 기존의 요약 기법들을 살펴보았다. 요약 후보문장을 발췌하는데 쓰이는 단서문구, 위치 기반 방법, 발췌와 요약에 사용되는 담화-지식기반의 수사적 역할, WordNet의 범주어, 줄거리 단위 및 연결, 문단/문장의 성질 등이 조사 분석되었다. 원문의 분석과 요약문의 생성에 기본이 되는 단어/절/문장/문단의 구문/의미적 역할, 단어 의미범주, 단서어구, 단어 출현빈도, 용언어미 범주, 복합 용언구, 관용 명사구, 고유명사를 정리하고 응용하였다.

실제로 요약을 진행하는 방식으로 통합, 분리 또는 삭제, 합성의 세 가지 방법론을 정립하였는데 그 중에서 합성에 중점을 두어 요약문 작성을 실험하였다. 문장들의 합성에서는 절 단위 합성, 어구 단위 합성으로 각각 진행하였다. 합성을 통한 요약문 작성을 위하여 장르 판별규칙, 문단/문장/절 슬롯(역할) 판별규칙, 주요 절/문장/문단 발췌규칙, 절 재조립 규칙, 절 및 어구 단위 합성 규칙, 문맥정의 리스트를 작성 사용하였다.

합성으로 작성된 요약문에 대한 "간결성, 대변성, 응집성, 일관성, 길이"를 평가하여 본 결과 보통 정도의 수준인 것으로 평가되었고 일관성은 보통 수준 이하였다. 앞으로 상기의 방법론 보다 정밀한 기법이 개발되어 수준 높은 자동요약을 생산하여야 하며 또한 웹 기사들의 특징상 진위를 가리는 비평적 문장을 포함시키기 위한 정밀한 논리가 개발되어야 할 것이다.

참 고 문 헌

이삼형. 1994. 『설명적 텍스트의 내용 구조 분석방법과 교육적 적용 연구』, 박사학위논문, 서울대학교 대학원 .

이태영. 2005. "자동 발췌문/요약 시스템 구축에 관한 연구-학술지 논문기사를 중심으로", 『한국문헌정보학회

지』, 39(3) : 139-163.

Allen, E.S., J.M. Burke, M.E. Welch & L. H. Rieseberg. 1999. "How reliable is science information on the Web?." *Nature*, 402, 722. quoted in E.T. Jepsen, P. Seiden, P. Ingwersen, & L. Bjorneborn. 2004. "Characteristics of Scientific Web Publications : Preliminary Data Gathering and Analysis", *JASIST*. 55(14) : 1239-1249.

Boguraev, and C. Kennedy. 1999. "Salience-based Content Characterization of Text Documents", quoted in I. Mani and M.T. Maybury(eds.). 1999. *Advanced in Automatic Text Summarization*. Cambridge, Massachusetts : the MIT Press.

Drott, M. 2002. "Indexing aids at corporate Websites : The use of robots.txt and META tags." *Information Processing and Management*, 38 : 209-219. quoted in E.T. Jepsen, P. Seiden, P. Ingwersen, & L. Bjorneborn. 2004. "Characteristics of Scientific Web Publications : Preliminary Data Gathering and Analysis", JASIST. 55(14) : 1239-1249.

Hovy, E. and C. Lin. 1999. "Automated Text Summarization in SUMMARIST." quoted in I. Mani and M.T. Maybury(eds.). 1999. *Advanced in Automatic Text Summarization*. Cambridge, Massachusetts : the MIT Press.

Jepsen, E.T., P. Seiden, P. Ingwersen, & L. Bjorneborn. 2004. "Characteristics of Scientific Web Publications : Preliminary Data Gathering and Analysis". *JASIST*. 55(14) : 1239-1249.

Jones, K. S. 1999. "Automatic summarizing : factors and directions." quoted in I. Mani and M.T. Maybury(eds.). 1999. *Advanced in Automatic Text Summarization*. Cambridge, Massachusetts : the MIT Press.

Lawrence, S. & C.L. Giles. 1999. "Accessiblity of information on the Web. Nature", 400, 107-109. quoted in E.T. Jepsen, P. Seiden, P. Ingwersen, & L. Bjorneborn. 2004. "Characteristics of Scientific Web Publications : Preliminary Data Gathering and Analysis." *JASIST*, 55(14) : 1239-1249.

Lawrence, S., K. Bollacker, & C.L. Giles. 1999. "Indexing and retrieval of scientific literature", quoted in E.T. Jepsen, P. Seiden,

P. Ingwersen, & L. Bjorneborn. 2004. "Characteristics of Scientific Web Publications : Preliminary Data Gathering and Analysis." *JASIST*, 55(14) : 1239-1249.

Lehnert W.G., 1999. "Plot Unit : A Narrative Summarization Strategy", quoted in I. Mani and M.T. Maybury(eds.). 1999. *Advanced in Automatic Text Summarization*. Cambridge, Massachusetts : the MIT Press.

Mani, I and M.T. Maybury(eds.). 1999. *Advanced in Automatic Text Summarization*, Cambridge, Massachusetts : the MIT Press.

Mani, I. 2001. *Automatic Summarization*, Amsterdam : John Benjamins Publishing Company.

Mann W. & S. Thompson. 1988. "Rhetorical structure theory : toward a functional theory of text organization." *Text* 8(3) : 243-281 quoted in D. Marcu. 1999. "Discourse trees are good indicators of importance in text." quoted in I. Mani and M.T. Maybury(eds.). 1999. *Advanced in Automatic Text Summarization*. Cambridge, Massachusetts : the MIT Press.

Marcu, D. 1999. "Discourse trees are good indicators of importance in text." quoted in I. Mani. 2001. *Automatic Summarization*. Amsterdam : John Benjamins Publishing Company.

Moens, M-F., C. Uyttendaele, and J. Dumortier. 1999. "Abstracting of Legal Cases : The Pontential of Clustering Based on the Selection of Representative Objects." *JASIS*, 50 : 151-161.

Moens, M-F. 2000. *Automatic Indexing and Abstracting of Document Texts*, Boston : Kluwer Academic Publishers.

Radev, D.R. and K. R. Mckeown. 1998. "Generating Natural Language Summaries from Mutiple On-line Sources." *Computational Linguistics*, 24 : 469-500.

Salton, G., J. Allen, and A. Singhal. 1996. " Automatic text decomposition and structuring." *Information Processing & Management*, 32 : 127-138.

Strzalkowski, T., G. Stein, J. Wang, B. Wise. 1999. "A Robust Practical Text Summarizer", quoted in I. Mani and M.T. Maybury(eds.). 1999. *Advanced in Automatic Text Summarization*. Cambridge, Massa

chusetts : the MIT Press.

Schutze, H. 1998. "Automatic word sense discrimination." *Computational Linguistics*, 24 : 97-123.

Talja, S. 2005. "The Social and Discourse Construction of Computing Skills." *JASIST*, 56(1) : 13-22.

Teufel, S. and M. Moens. 1999. "Argumentive classification of extracted sentences as a first step towards flexible abstracting." quoted in I. Mani and M.T. Maybury(eds.). 1999. *Advanced in Automatic Text Summarization*. Cambridge, Massachusetts : the MIT Press.