자동 요약을 위한 웹 기사들의 유형 구분과 주연문맥 추출에 관한 연구

A Study on Classifying and Analyzing the News Form in the Web for Automatic Summarization

이태영, 전북대학교 문헌정보학과, hsl0001@unitel.co.kr

Lee Tae-Young, Dept. of Library and Information Science,
Chonbuk National University

웹 상의 기사들의 종류를 보도, 기획/해설, 인터뷰/대담, 논설, 신변잡기로 나누어 자동요약을 위한 프레임을 작성하였다. 각 기사 프레임들은 글 구조적으로 분석되어 "사건, 상황, 연관, 과정, 원인, 결과, 결론, 전망, 방안, 평가 등"과 같은 문단 슬롯이 부여되었고 문단 슬롯은 다시 문장 슬롯으로 세분되었다. 'if-needed' 패싯으로 육하원칙인 "주체, 객체, 시간, 장소, 원인, 방법"을 택하였다. 슬롯이나 패싯의 실제값들을 추출 표현하는 과정에서 문구의 수사적 역할과 단어 최상위 범주 및 줄거리 단위를 참조하였다. 기사의 유형과 문단 및 문장 슬롯을 판별하기 위해서 유형 판별 규칙과 슬롯 판별 규칙을 구비하였다.

1. 서론

웹에 출현하는 기사들은 다중문헌(multi document)적이며 멀티미디어(multimedia)적이다. 한 주제 또는 사건에 대해 동일 사이트 내에서 여러 가지 기사로 기록되며 더 나아가 여러 사이트에서 동시다발적으로 기록되고 기사 표현 양식이 멀티미디어적으로 다양하게 작성되어 있을 뿐만 아니라 그 양 또한 지식정보사회란 이름에 걸맞게 엄청나게 많다. 그러므로 많은 정보들 중에서 중요 정보 발췌와 그것들의 정리 요약이란 행위는 필연적이며 지식정보사회의 정보폭발 현상에서 매우 경제적 가치가 있는 IT 상품이라고 할 수 있다.

그동안 요약을 작성하는 방법론은 여러 가지 측면-코퍼스적 방법, 담화(수사)구조에 의한 방법, 언어적 지식을 이용한 방법-에서 모색되어 왔다. 이 중에서 담화·언어적 지식을 원용하려면 웹 기사도 학술잡지 논문기사처럼 글 구조(문단 또는 문장들)를 "연구목적, 방법, 결과, 결론 등"과 같이 파악해 놓을 필요가 있다.

따라서 본 연구는 웹 상에 출현하는 기사들을 담화·언어적으로 요약하는데 기초적으로 필요한 기사의 유형(이하 장르라고도 칭함)과 유형에 따른 글 구조의 정립을 목적으로 하였다. 아울러 유형구분과 유형구조를 파악하는 판별규칙을 제시하였다.

2. 웹 정보의 특징

현재 웹에 게재되는 정보자료들의 형태는 그림, 소리, 동영상, 글의 형식이 주종을 이루고 있고, 정보유형으로 1차정보(원 생산자가 작성한 정보), 2차정보(서지정보), 3차정보(1차정보를 각색한 정보) 모두를 망라하고 있으며, 정보내용에서 생활(소식, 물가, 금융, 마케팅, 알림광고)정보와 학술정보를 두루 포함하고, 주제별로는 정치, 경제, 사회, 문화 등 전 주제를 고루 다루고 있어 가히 전 방위적인 정보유통 매체라고 할만 하다. 이러한 웹 정보자료들에게는 다음과 같은 몇 가지 공통적인 특징이 존재한다.(Jepsen et al. 2004, 1239-40)

(1) 지리적, 언어적 다양화와 지수적인 증가

(2) 출판의 자유; 여과와 비평이 없는 출판의 자유가 있다.

(3) 학술적 자료들의 여과와 검색에 기여하는 구조적 정보 또는 메터데이터들이 거의 없다. 43% 미만의 사이트들이 최소한의 메터데이터를 키워드 또는 기술태그로 언급하고 있고

(4) 색인 정보를 등한시 한다. 1999. 2월 기준으로 16% 이상 색인된 검색엔진은 하나도 없었다. 소유주 중심의 환경하에서는 탐색엔진들이 이용자에게 왜곡되지 않은 정보(skewed selection)를 제공한다는 것을 믿을 수 없다.

(5) Allen 등의 연구에 의하면 검색 엔진인 'Northern.com'에서 검색한 결과 500 사이트 중 12-46%가 적합, 10-34%가 부정확하였다. 20-35%가 오류(misleading)로 판명되었다.

위의 특징으로 보았을 때, 웹 정보는 색인 요약문이 필요하며 그 곳에 진위 판별(비평) 및 보존기간을 명시하여 운영하는 것이 또한 요구된다.

3. 기사의 유형

기사의 종류는 매우 다양하고 서로 엄격히 구별하기 어려운 경우가 적지 않아 일목요연하게 분류하기는 쉽지 않은데, 본드(Bond)는 기사들을 크게 다섯개 유형 즉 인터뷰, 연설보도, 인간흥미, 사망(obituary) 및 스포츠 기사로 분류하고 있다. 그러나 Mandel은 8가지 즉 스팟뉴스(spot news), 인터뷰, 연설보도, 피처(feature), 인사동정, 논평과 칼럼, 공고(publicity), 논설 기사로 나누었다.

한편 Charnley는 뉴스보도, 인간흥미, 발굴보도(investigative reporting), 해설기사로 대별하는가 하면, Harris, Leiter와 Johnson은 주제에 따라 사망, 사고, 범죄, 일기예보, 행정, 정치, 경제, 교육과 학술과학, 종교, 가정, 스포츠, 연예오락, 논설, 칼럼, 해설, 발굴 기사 등 모두 17가지로 분류하였다. (http://terms.naver.com/item.php?d1id=6&docid=1210)

진행남(2002, 619)은 뉴스기사의 유형을 <표 1>와 같이 사전조사를 바탕으로 크게 5가지, 스트레이트, 기획/연재, 인터뷰, '신변잡기', 스트레이트+해설 등으로 구분하였다. 여기서 신변잡기란 '자신의 주변에서 일어나거나 몸소 겪은 일을 수필체로 쓴 기사'로 작위적 정의를 내렸다.

<표 1> 인쇄신문과 인터넷미디어의 뉴스 기사 유형 분석

매체명 / 기사유형	조선일보 비율 (건)	한겨레 비율 (건)	오마이뉴스 비율 (건)
스트레이트	57.1 (48)	35.3 (29)	37.5 (30)
기획/연재	7.2 (6)	9.8 (8)	10.0 (8)
인터뷰			11.2 (9)

신변잡기			31.3 (25)
스트레이트+해설	35.7 (30)	54.9 (45)	10.0 (8)
합 계	100% (84)	100% (82)	100%(80)

본고에서는 진행남의 구분을 중심으로 하여 기사의 장르를 "보도(스트레이트), 논설, 기획/해설, 인터뷰/대담, 신변잡기"로 아래의 <표 2>과 같이 정리하였다.

<표 2> 웹 기사의 유형구분

	내용
보도	논평이나 작성기자의 의견을 넣지 않고 어떤 사실을 있는 그대로 보도하는 기사 대부분의 보도기사는 이러한 스트레이트 **기사**이다
논설	사실의 진실성을 파악한 뒤 그 사실을 평가하며 그것에 대한 의견을 서술하는 기사.
기획/해설	어떤 사물이나 문제를 정리하여 보다 정확하게 독자들이 이해하도록 설명거나 신문 편집자(글쓴이)의 생각을 명확히 제시하면서, 기사에 살을 붙여 문제 의식을 가지고 사실을 파헤치는 기사.
인터뷰/대담	특정 인물이 보도의 대상이 될 때, 혹은 그 사람의 입을 통해 어떤 사실을 알아내려고 할 때, 그 사람의 대화로 얻어진 기사.
신변잡기	자신의 주변에서 일어나거나 몸소 겪은 일을 수필체로 쓴 기사

4. 유형별 프레임

4.1 프레임의 슬롯 조사

기사는 육하원칙(누가, 무엇을, 언제, 어디서, 왜, 어떻게)의 틀 속에서 FLUMP, SUMMONS, MUC-4 system 등 여러 시스템에서 정형화되어 왔다.

SUMMONS는 입력 기사에서 보고되는 요점 사실들을 포함하는 템프릿 집합으로부터 메시지 이해시스템에 의해 요약을 생산하였다. 이들 시스템은 주어진 기사로부터 특정 정보 조각들을 발췌하였다.

MUC-4 systems은 테러리스트 영역을 운영하기 위해 템프릿 당 25개 필드를 설정하였고 기사로부터 "가해자, 희생자, 사건유형" 등에 해당하는 정보를 발췌하여 동일 필드영역에 채워 넣었다

위의 시스템들과 같이 웹에 올라오는 기사들을 잘 요약하려면 장르별로 기사 글들의 특징 및 구조를 파악하여 그 특성에 맞게 요약이 수행되어야 한다.

자동요약에 담화구조적 방법을 시도하였던 Teufel과 Moens(1999, 160-164)는 수사역할 자질을 <그림 1>과 같이 16가지로 분류하였다.

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 1> 수사역할 자질 예

한편 WordNet에서는 명사와 동사의 최상위 계층 범주어로 <그림 2>와 같은 어구들을 선정하였다(김영택 2001, 146-147에서 인용).

명사: 동작, 동물, 인공물, 속성, 신체부위. 인지, 커뮤니케이션, 사건, 감정, 음식, 집합, 위치, 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 2> WordNet 범주어 예

언어지식기반 방법으로 자동요약을 시도하였던 Lehnert (1999, 178-182)는 줄거리 단위(Plot Unit)로 이야기들을 요약하는 방법을 제시하였다. 그는 기본적인 줄거리 단위를 아래의 <표 3>과 같이 15개 만들었다.

<표 3> 줄거리 단위

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

이와 같은 역할자질과 최상 범주어 및 줄거리 단위를 응용하여 본 연구에서는 <표 4>에 각 장르 프레임이 내포하고 있는 문단과 문장 슬롯들이 가지고 있는 역할 자질들을 분석하여 기재하였다.

4.2 슬롯과 문장생성

Lehnert가 말한 줄거리단위는 각각 영향상태와 인과적 연결을 가지고 있다. 예를 들면 '문제'의 경우를 나타내고 있는 문장 "Your dog dies and you long for companionship."은 다음의 <그림 3>과 같이 분석 표현이 되는데 이 표식은 앞 문장이 '부정'에서 '정신 상태'로 연결되어 가는데 '동기'가 작용되었다는 것이다.

이러한 기본 줄거리 단위들이 결합하여 많은 수의 복합 줄거리 단위를 만들었다.

문제
- → M
m

<그림 3> 줄거리 분석.

본 연구에서 요약 문장들의 생성은 도입→활동→마무리, 원인→결과→전망 등과 같은 줄거리들이 적용되며 항상 <표 4>에서 밑줄이 쳐진 도입 사상을 기준으로 전개하였다.

<표 4> 기사 글틀구조

유형 프레임	문단역할 슬롯	문장역할 슬롯
보도	사건, 과정, 연관, 원인, 결과, 전망, 방안, 평가,	상황, 과정, 원인, 결과, 방법, 전망 등
논설	문제(목적), 상황, 증명, 전망, 결과, 결론, 제시	목적(문제제기), 원인, 결과, 증명, 전제, 방법, 전망, 명제, 조건, 제언, 결론
기획/해설	대상, 상황, 배경, 과정, 결과, 결론, 목표, 평가	발제, 설명, 원인, 결과, 결론, 평가 등
인터뷰/대담	소개, 주제, 상황, 문제, 방안, 해석, 목적, 반박, 결론, 전망	상황, 문제, 원인, 결과, 결론, 해명, 권유, 동의, 방법, 과정 등
신변잡기	글감, 배경, 회상, 원인, 결과, 해설, 해석/상상, 활동/행위, 대화, 결과, 결론, 체득, 방안, 과정, 마무리	대상, 도입, 회상, 경험, 과정, 감정, 비유, 판단, 설명, 전달, 행위, 원인, 결과, 마무리 등

요약 작성과정을 개요적으로 설명하면 다음과 같다.

1) 각 장르의 글 도입부를 의미하는 목적적, 발의적 문단 슬롯인 "목적, 사건, 대상, 주제, 글감"에 담긴 내용을 요약문의 제1대상으로 올린다.

2) 제1대상이 요약의 조건을 만족시키면 요약을 멈춘다. 요약의 조건에는 '육하원칙의 내포', '문장의 개수(1 문장 또는 3 문장 등)',

'비평포함' 등이 제시될 수 있다.

3) 제1대상이 요약의 조건을 충족하지 못하면 다른 문단 슬롯의 문장을 다음과 같은 순서로 첨가하여 요약문의 골격을 완성한다.

보도기사 : 사건 +(①상황→②원인→③결과→④결론→⑤전망→⑥평가)

논설기사 : 목적 +(①상황→②증명→③원인→④결과→⑤결론→⑥제시)

기획/해설 : 대상 +(①상황→②배경→③결과→④결론→⑤목표→⑥평가)

인터뷰/대담 : 주제 +(①상황→②문제→③방안→④해석→⑤목적→⑥반박→⑦결론→⑧전망)

신변잡기 : 글감 +(①배경→②회상→③원인→④결과→⑤해설→⑥결론→⑦체득→⑧마무리)

4) 보도기사로 예를 들어 요약작성 과정을 추단하면 (1) '사건' 문단 슬롯에 기재된 내용으로 요약문을 충족시키는 경우, (2) '사건'에다 3)에서 분석된 다른 슬롯의 문단이나 문장들을 보완하여야 할 경우가 있다. 그리고 (2)는 (가) 발췌된 문단의 문장(한 문장이 한 문단일 적도 있음)을 그대로 요약에 삽입할 수 있는 경우와 (나) 다듬어서 사용하는 경우로 나누어진다.

(1) '사건' 문장 하나로 충족되는 경우: <그림 4>의 ① 문장은 육하원칙의 "**주체, 객체, 시간, 장소, 원인, 방법**"을 다 가지고 있어 요약의 구실을 할 수가 있다.

(2)의 (가)에 해당하는 것은 <그림 4>의 ③과 ⑤ 및 ⑦ 문단이다.

(2)의 (나)에 해당하는 것은 <그림 4>의 ⑥ 문단의 둘째 문장이며 이 문장은 "**방법, 원인, 결과**" 문장으로 분리될 수 있다.

5) 같은 사건의 기사들에서 서로 같은 내용이거나 또는 다른 내용이 등장하였을 때 '**비평**' 란을 만들고 요약이 만들어지는 전 기사

① 장마전선이(주체) 활성화되면서(원인) 27일(시간) 서울에(장소) 시간당 30mm가(방법) 넘는 폭우가(객체) 쏟아지는(원인) 등 서울과 경기, 강원 지방에(장소) 많은 비가(객체) 내려(원인) 침수, 교통통제 등의 피해가(방법) 잇따랐다.(결과) '사건'
② ..
③ 이번 폭우의 원인은 27일 중부지방에 머물던 장마전선에 제5호 태풍 개미로부터 많은 양의 수증기가 서남풍을 타고 공급된 데 있었다. '원인'
④ 폭우로 가장 큰 추가 피해가 우려되는 강원지역에서는 27일 오전부터 인제와 평창, 양양 등 3개 군 31개 마을 주민 4522명이 안전지대로 대피했으나 저녁 무렵 대부분 돌아가 270가구 720명으로 줄었다. '결과'
⑤ 경기도에서 가장 많은 비가 내린 김포지역에서는 대곶면 54ha, 김포1·2동 34ha 등 모두 100ha의 농경지가 물에 잠겼다. '결과'
⑥ 한강수력발전처는 오후 10시 30분 현재 청평댐 수문 13개를 38m까지 열고 초당 3730t을 흘려보내 서울의 한강시민공원 반포지구 전체와 강서, 망원, 여의도, 이촌지구 일부가 침수됐다.(결과) 서울에서는 이날 폭우로 잠수교와 영동1교 등 서울 시내 일부 도로의 교통이 통제됐으며(방법) 청계천 일대와 한남대교 고양시 방향~강변북로 진입램프 등 시내 곳곳에서 차량이 시속 20km 이하의 '거북 운행'을 해(원인) 퇴근길 교통정체가 빚어졌다.(결과) '결과'
⑦ 막바지 장맛비는 중부지방의 경우 토요일인 29일 오전에야 그칠 전망이다. '전망'
홍수영 기자 gaea@donga.com

<그림 4> 표본 예

에 보존기간을 명시하는 란을 추가한다.

5. 유형과 슬롯 판별 규칙

5.1 유형 판별 규칙

요약시스템에서 새로 입력되는 기사의 유형 판별은 전장에서 사전에 정하였던 문단의 슬롯을 이용하였다. 기사 글에 나타나는 문단들의 역할자질을 파악하여 글 프렘임의 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 같았다.

<유형 판별 규칙 예>

규칙1: IF (첫문단 = '사건') THEN '보도'
규칙2: IF ('목적' ∧ '증명'∧ '결과' ∧ '결론') THEN '논설'
규칙3: IF ('기행' ∨ '정서' ∨ '감상' ∨ '회상') THEN '신변잡기'
규칙4: IF (('주제' ∧ '인사' ∧ 따옴표 문장) ∨ ('주제' ∧ '인사')) THEN '인터뷰/대담'
규칙5: IF ('대상' ∧ '현상' ∧ '전망') % 따옴표문장 THEN '기획/해설'

5.2 슬롯 판별 규칙

문단슬롯 즉 문단의 역할자질은 문단을 구성하고 있는 문장들의 역할자질인 문장슬롯과 문장의 동사, 명사의 범주 및 특징에 의해 결정된다. 이와 마찬가지로 문장슬롯 즉 문장의 역할자질은 문장을 구성하고 있는 표징성이 강한 단어인 동사, 명사의 범주 및 특징에 의해 아래의 <슬롯 판별 규칙 예>와 같이 결정된다.

<슬롯 판별 규칙 예>

규칙1: IF ((첫문장) ∧ (육하원칙 ≥ 4)) THEN '사건'
규칙2: IF (도서명 ∨ 인물 ∨ 명승지 ∨ 강 ∨ 산 ∨ 유적) ∧ (~라고 하였다) THEN '전달'
규칙3: IF ('원인' ∧ '결과') THEN '증명'

6. 결론

웹상의 기사 자동요약 시스템 구축의 일환으로 기사의 장르들과 장르에 따른 글 구조들-문단역할 및 문장역할 슬롯을 조사 분석하였다. 그 슬롯을 이용하여 요약 문장을 생성하기 위한 개념적 시도를 하고, 유형과 슬롯을 결정하는 판별규칙에 대한 예를 작성하여 보았다. 그리고 웹 정보들이 현재 갖고 있는 특징에서 요약 연구의 당위성을 찾았고 진위판별을 할 수 있는 '비평'과 '보존기간'을 명시하는 것이 시급함을 알았다.

참고 문헌

김영택 외. 2001, 『자연언어처리』, 서울; 생능출판사, 421p.

진행남. 2002, 인터넷미디어의 뉴스 영역 및 유형에 관한 연구, 『한국언론학보』, 제46권 2호, 606-632.

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.

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.

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

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