

# 확장된 6하원칙을 이용한 신문기사 자동요약

윤재민<sup>0</sup> 강인수 권오욱 배재학 이종혁  
포항공대 정보통신대학원<sup>0</sup> 포항공대 전자 및 컴퓨터공학부  
(chiwoo<sup>0</sup>, dbaisk, ohwoog, jhjbae, jnllee)@postech.ac.kr

## An automatic extraction of newspaper articles using expanded 5W1H

Jae-Min Yoon<sup>0</sup> In-Su Kang Oh-Woog Kwon Jae-Hak Bae Jong-Hyeok Lee  
Dept. of Graduate School for Information Technology, POSTECH<sup>0</sup>  
Div. of Electrical and Computer Engineering, POSTECH

### 요 약

본 논문은 신문 기사를 추출(Extract)하는데 있어서, 신문기사의 구조적인 특징을 분석한 후, 6하원칙(5W1H)과 전문(Lead)과의 관계를 유지하여 신문 기사를 효과적으로 요약하기 위한 방법과 알고리즘을 제안하였다. 본 연구에서는 먼저 신문 기사에서 가장 중요한 단락인 전문에서 6하원칙의 각 구성성분을 추출하고, 본문에서는 전문에 나타난 6하원칙의 각 성분이 어떻게 재사용되며 강조되고 있는가를 파악하기 위해 제안된 방법과 알고리즘을 이용하여 중요한 문장을 추출하였다. 실험문서는 조선일보 웹사이트에서 제공하는 신문 기사 100건을 대상으로 하였으며, 요약율이 20%와 30%일 경우 제안한 방법의 정확률은 각각 82.4%와 74.1%로 기존의 전문기반(Lead-based)방법보다 3.6%와 6.3% 향상되었다.

### 1. 서 론

자동요약(Automatic Summarization)이란 대상문서에서 가장 중요한 정보를 종류해내는 과정[1]으로 크게 요약(Extract)방법과 추출(Extract)방법으로 나눌 수 있다. 요약(Extract)방법은 원문에 없는 새로운 문장을 생성해 내는 방법으로 구현하기 어렵고, 많은 지식 자원(Knowledge Resources)을 필요로 한다. 그러나 추출(Extract)방법은 원문에서 상대적으로 중요한 문장을 추출해 내는 것으로서 요약(Extract)방법에 비해 쉬운 접근 방법이며 통계적인 분석이나 확률적인 분석으로 구현되고 있다. 최근에는 추출(Extract)방법의 단점인 문장 가독성 저하를 방지하기 위해, 중요문장을 추출 후, 너무 긴 문장은 분리하고, 단문은 결합하면서, 추출된 문장들 사이에서 내용이 문제를 다루는 교정(Revision)방법이 연구되고 있다[1].

본 연구에서는 모바일(Mobile)이나 PDA(Personal Digital Assistants), 인터넷(Internet) 등을 통해 신문 기사를 읽을 때, 공간과 시간적 제약으로 인해 원문을 다 볼 수 없는 경우의 문제점을 해결하기 위해, 신문기사의 구조적인 특징인 6하원칙에 근거한 자동 추출(Extract)방법을 제안한다. 2장에서는 기존의 연구와 문제점을 살펴보고, 3장에서는 신문기사의 구조적인 특징인 전문과 6하원칙의 관계에 대해서 기술한다. 그리고 4장에서는 3장에서 언급한 특징을 이용할 수 있는 방법과 추출 알고리즘에 대해서 설명하고, 5장에서는 실험, 6장에서는 결론을 다룬다.

### 2. 기존 연구 및 문제점

일반적으로 신문기사는 다양한 요약 방법들의 실험문서(Test Document)로 이용되어 왔다. 이러한 신문 기사 분야에 대한 요약 연구들은 문장간의 담화구조(Discourse Tree)를 이용한 방법이나 어휘사슬(Lexical Chain), 단어 공기정보(Co-Occurrence)를 이용한 방법, 또는 여러 가지 통계적인 방법을 통합한 방법 등을 이용하였다. [2][3]의 연구에서는 전문기반(Lead Based) 추출방법이 기타 다른 신문 기사 요약방법에 비해 성능이 월등히 좋다고 보고되었다. 여기서 전문기반 추출방법이란, 신문기사의 처음 문장부터 일정한 요약율에 의해 차례대로 선택하는 방법을 말한다. 그러나, 전문기반 방법의 문제점은 신문기사의 처음이나 둘째 부분, 또는 그 다음 부분이 중요하지 않은 문장이 있을 경우 요약률 저하를 유발할 수 있다는 문제점이 있다.

신문 기사 요약에 있어서, 전문기반 방법 대신에 최근에 신문기사의 작성 원칙인 6하원칙을 이용한 연구가 진행되었는데, [8]에서는 실험 대상 신문기사의 원문에서 문장의 유사도 관계를 이용하여 중심절을 추출하고, 추출된 중심절과의 유사도가 높은 절에서 유리스틱 정보와 수사어구 정보를 이용해서 사건의 시간적, 공간적 배경, 사건의 원인과 결과에 해당하는 절을 추출하는 방법을 제안하고 있다. 그러나, 추출된 정보를 어떻게 문장으로 구성하고 요약에 이용할 것인가에 대한 해결책을 제시하지 못하였고, 유리스틱 정보와 수사어구를 이용해서 추출된 절이 중심절의 사건, 또는 제목과 어떠한 연결고리가 있는지 의미적으로 파악하지 못했다. [4]에서는 [8]에서와는 달리 실험 대상 신문기사의 원문에서 수사어구를 이용하여 6하원칙에 해당하는 정보를

추출하지 않고, 신문 기사 제목(Headline)에서 6하원칙에 해당하는 정보를 추출하여, 시소러스를 이용, 각 6하원칙 구성단어를 만족시키는 상위레벨의 단어로 여러 개의 신문기사의 제목(Headline)을 분류(Classification)하여 요약하는 방법을 제안하였다. 그러나, 이 방법은 신문기사의 원문을 다루지 않기 때문에 제목이 원문의 내용을 충실하게 내포하지 못하는 경우는 잘못된 기사를 찾아볼 수 있으며, 또한 신문기사의 원문을 요약하는 방법이 아니고 신문기사의 제목을 분류하기 위한 요약방법이기 때문에 그 자체로 제한적인 요약이라고 할 수 있다.

이상에서 보는 바와 같이 기존의 연구는 신문 기사에 대한 구조적인 특징을 파악하거나 요약대상이 되는 문서의 특성을 고려하지 않고 [2][3], 요약방법을 일률적으로 적용한 결과 요약율이 상당히 낮은 것을 알 수 있다 [7]. 그리고, 신문기사의 구조를 분석할 때, 6하원칙과, 제목, 전문과의 관계를 심도있게 분석하지 않고, 6하원칙만 강조한 결과 전체 의도할 수 없는 결과를 생성해 낼 수 있으며 [8], 6하원칙을 적용한 적용함으로써, 피상적으로 분석하여 그 자체 내에 한계를 가지고 있다 [4]. 따라서, 본 논문은 전문기반 방법을 보완하기 위해, 전문과 제목에서 6하원칙 관계를 파악하고, 제목과 전문에서 강조되고 있는 6하원칙의 구성성분이 본문에서 어떻게 재사용되고 있는지를 분석하고, 본문에서 중요 문장을 추출하기 위한 방법과 알고리즘을 제안한다.

### 3. 신문 기사 구조 분석

#### 3.1 6하원칙과 신문기사의 이해

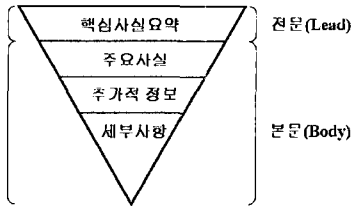
기사의 사실을 구성하는 요소는 6하원칙(5W1H)이고 '누가', '언제', '무엇을', '어디서', '왜', '어떻게'에 해당하는 구체적 사실을 독자와 시청자에게 전달하는 것이 기사이다 [11]. 따라서, 좋은 신문기사는 정확성(Accuracy), 객관성(Objectivity), 공정성(Fairness) 등이 뒷받침되어야 한다 [9]. 이러한 이유로 신문기사는 문장이 간결하고, 수사어구(Rhetorical marker)와 단서단어나 구(Cue Words or Phrase)가 거의 없기 때문에 담화트리(Discourse Tree)를 이용하거나 어휘사슬(Lexical Chain) 등을 신문 기사 요약에 이용하기에는 부적절한 측면이 있다.

#### 3.2 전문과 역피라미드 구조

대부분의 신문기사는 제목(Headline), 전문(Lead), 본문(Body)의 역 피라미드 구조를 가지는데 전문은 6하원칙에 의해 기술하고 본문에는 전문에서 강조한 것에 대해서 구체적으로 설명하거나 추가적인 정보를 기술한다 [5][9][10][11]. [그림 1]은 역피라미드(Inverted pyramid) 구조의 특징을 보여주는데, 이렇게 신문기사가 역피라미드 구조를 따라야 하는 이유는 모든 사람들이 기사를 처음부터 끝까지 읽지 않고, 기사를 읽는 사람들 가운데 25% 정도는 중간 정도 읽고 중단하기 때문이다 [6]. 또한, 현장에서 기자가 작성한 신문 기사를 편집실에서 편집할 때, 시간적, 또는 공간적인 제약으로 인하여 기사를 잘라내게 되는데, 이때 기사의 제일 밑부분부터 잘라내기 때문에 신문기사는 기사의 마지막 부분으로 갈수록 상대적으로 중요도가 떨어지는 정보를 기술할 수 밖에 없다 [5].

3.3 제목과 전문

신문기사의 전문은 역피라미드 구성의 첫 머리에 해당하는 만큼 기사의 가장 중요한 내용이 압축적으로 포함된다. 전문 한 줄만 써도 전체 기사의 내용을 짐작할 수 있을 정도가 되어 한다. 본문은 전문을 부연 설명하는 과정에서 실타래가 풀리듯이 쓰여지기 때문이다. 일반적으로 전문의 기본 문장구성은 주어, 목적어, 동사로 이루어져 있다. 전문은 반드시 주어로 시작해서 그 다음에 목적어가 나오고 동사로 끝나야 한다[11]. 따라서 6하원칙의 기사 구성요소가 기사 속에서 어떻게 제시되는가는 [그림 2]를 보면 알 수 있다.



[그림 1] 기사의 형태적 구성요소[11]

미, 대북 포용정책 지지 [제목]  
 미국 국무부 고위 관리들은[누가:Who] 24일[언제:When] 한국 김대중 대통령의 대북 포용정책을[무엇을:What] 강력히 지지한다[어떻게:How]는 뜻을 밝혔다. [전문]  
 제임스 루빈 국무부 대변인은 .... [본문]

[그림 2] 기사의 내용적 구성요소[11]

전문을 작성할 때 가장 먼저 생각해야 하는 점은 6하원칙에서 무엇을 가장 강조해야 할 것인가를 정하는 문제다. [그림 3]는 제목에서 6하원칙의 구성성분 중 무엇을 강조하는 가에 따라 강조하는 부분이 전문에 모두 녹아 있는 것을 보여준다[10].

경관이 슬집, 여관 불법경영[제목] - Who 강조  
 공무원법상 영리사업을 할 수 없는 경찰관 중 상당수가 경찰 단속 대상인 유흥업소, 숙박업소를 운영해 온 것으로 밝혀졌다. [전문]  
 대우, 산동에 대규모 시멘트 공장 설립[제목] - Where 강조  
 대우 그룹이 중국 산둥지역에 시멘트 생산법인을 설립, 대규모 시멘트 공장을 세우기로 하고 8일 북경에서 기념행사를 가졌다. [전문]

[그림 3] 전문과 본문[10]

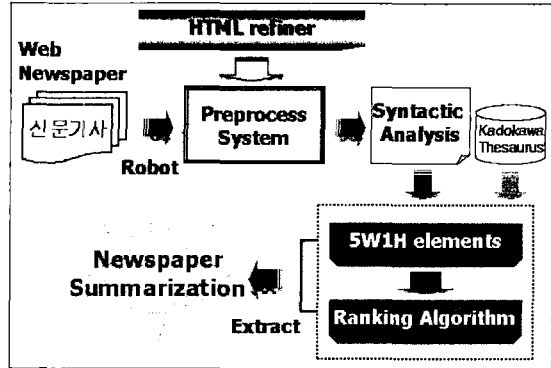
3.4 전문의 구성비율

일반적으로 기사는 그 내용과 성격에 따라 뉴스성 기사를 경성기사, 해설성 기사를 연성기사로 분류하는데 해설성 기사는 논설이나 칼럼에 해당하고 뉴스성 기사로는 국제, 정치, 경제, 사회, 문화 관련 기사로 나누어진다[9]. 본 논문에서는 뉴스성 기사를 대상으로 하여 연구하였으며, 야후의 뉴스기사(Newswire)와 조선일보의 신문기사(Newspaper)에서 각각 100개의 기사를 무작위 추출해서 전문의 유무를 파악하기 위한 실험을 하였다. 야후 뉴스기사의 경우, 전문이 차지하는 비율이 99%이고, 신문기사의 경우 95%였다

4. 확장된 6하원칙과 순위부여 알고리즘

본 연구에서 제안한 전체적인 시스템 구성은 [그림 4]과 같다. 실시간으로 신문기사 요약하기 위해, 해당 웹페이지에서 로봇으로 신문기사를 가져와서 전처리 시스템(Preprocess system)을 거치게 되는데 이때, 웹페이지를 구성하고 있는 각종 태그들이 제거되고, 신문기사 원문을 복구하게 된다. 그리고 구문분석을 통하여 문장의 성분을 분석하고, 분석된 정보를 바탕으로 동의어, 유사어 정보를 얻기 위해 가도카와 시소러스를 이용해서 어휘를 확장한 다음, 6하원칙의 각 구성성분을 분리하고, 본 논문이 제안하는 순위부여 알고리즘을 바탕으로 중요문장을 추출하게 된다. 여기서, 6하원칙의 각 구성성분을 추출하기 위해서는 [표 1]과 같은 패턴정보 테이블이 필요하고, 구체적인 6하원칙 성분으로 기술할 수 없는 부분을 객체(Object)로 할당하여 문장 성분의 다양성을 확보하였다. 일반적으로 제목 부분에서 언급한 내용이 전문부분에 대부분 기술되지만, 전문에는 없이 제목에만 출현한 중요단어는 전문에서 추출한 6하원칙 구성성분과 동일한 가치를 부여한 후 두 부분을 합쳐서 중요문장을 추출하기 위한 알고리즘에 이용한다.

[그림 5]는 전문에서 나타난 6하원칙의 구성성분에 의해 본문의 중요한 문장을 분석하는 방법을 설명하는 그림이다. 여기서 다루는 문장들은 전문에서 기술한 6하원칙 중에서 빠진 부분을 포함한 문장을 분석하는 문장, 전문의 6하원칙의 각 성분에 대해서 일부 덜 포함되거나 역할 변화된 것을 분석하는 문장, 그리고, 일반적인 신문기사의 문장형식인 6하원칙 구성성분의 재사용과 역할 변화가 동시에 나타나는 복합문장을 분석하는 방법에 대해서 설명하고 있다.



[그림 4] 전체 시스템 구성도

		Example
Expanded 5W1H	Who	명사(구) + 주격조사, 보조사(은,는)
	When	숫자+ 년/월/일/시/분/초, 작년/올해/오전/오후/낮/밤 등
	Where	장소명사(구) + 예,에서 등
	What	명사(구) + 목적격조사
	How	용언(동사)
	Why	~위해서, 때문에, 이유로 등
	Object	명사(구) + 보조사, 접속조사 등

[표 1] 확장된 6하원칙 구성성분 분석

- 전문에서 기술한 6하원칙 구성요소보다 많이 포함된 문장(내포문)  
 [전문문장]: 알 카에다 대원이 드디어 체포되었다.  
 E5W1H=>Who(알카에다대원),How(체포)  
 [본문문장]: 알 카에다 대원이 카불에서 체포된 것은 어제 저녁이었다.  
 E5W1H=>When(어제),How(저녁),Who:(Who(알카에다대원), Where(카불),How(체포))  
 [Ranking=>W(1:2)R(-1,-1)S(+2)]
- 전문 6하원칙 성분이 일부 재사용된 문장  
 [전문문장]: 이스라엘의 예루살렘에서 무장 폭탄이 행인들에게 충격을 안겼다.  
 E5W1H=> Who(무장폭탄),Where(예루살렘(이스라엘)),What(충격),How(안사하다),Object(행인)  
 [본문문장]: 현지 경찰은 무장 폭탄이 죽인 부상자 가운데 6명은 중태라고 말했다.  
 E5W1H=>Who(현지경찰),Who(부상자(가운데)),How(중태),Object(죽이다),Object(중태),How(말하다)  
 [Ranking:W(1:1)R(-1,0)S(-2)]
- 전문 6하원칙 성분이 역할변화 한 문장  
 [전문문장]: 주가가 연속 상승하며 연초 수준인 720선을 회복했다.  
 E5W1H=>Who(주가),What(720선(연초수준)),How(연속상승) (회복하다)  
 [본문문장]: 외국인인 11일만에 순매수로 전환하며 주가 상승을 이끌었다.  
 E5W1H=>Who(외국인),What(주가상승),How(전환하다)(이끌다),Object(11일(순매수)),  
 [Ranking:W(1:2)R(0,2)S(+1)]
- 6하원칙 구성성분의 재사용과 역할 변화가 동시에 나타나는 복합문장  
 [전문문장]: 여성은 장신의 유전자와 비슷하고 또 아버지의 체질을 느끼게 하는 남성을 선호한다는 사실이 밝혀졌다.  
 E5W1H=>Who(여성),What(남성:What(체취(아버지)),How(느끼다)),How(비슷하다)(선호한다),Object(유전자(자신))  
 [본문문장]: 미국 시카고대학의 마사 메들린 박사는 영국의 의학전문지에 발표한 연구보고서에서 임신한 적이 없는 미혼여성 49명을 대상으로 인공적 배정이 다른 남자 6명의 와이셔츠를 맡고 선호도를 평가하게 한 결과 이 같은 사실이 밝혀졌다고 말했다.  
 E5W1H=>Who(마사메들린박사(미국시카고대학))What(와이셔츠냄새(남자:Who(배정(인공))),How(다르다))(6명))(선호도),How(말다)(평가하다),Object(연구보고서:Object(의학전문지(영국)),How(발표하다))(대상:What(미혼여성:How(없다

