육하원칙 활성화도를 이용한 신문기사 자동추출요약

(Automatic Extractive Summarization of Newspaper Articles using Activation Degree of 5W1H)

윤 재 민 † 정 유 진 †† 이 종 혁 †††
(Jae-Min Yoon) (You-Jin Chung) (Jong-Hyeok Lee)

요 약 육하원칙은 신문기사를 기술하는데 있어서 가장 기본적인 요소로서 기사 내용 파악에 핵심적인 역할을 수행한다. 본 논문은 이러한 육하원칙에 기반하여 기술되는 신문기사의 특성에 주목하여, 육하원칙 활성화도를 이용한 신문기사 요약 방법론을 제안한다. 제안하는 방법론은 기존의 요약 기법 중 가장 우수한 방법으로 알려진 두문 기반 기법(lead-based method)과 제목 기반 기법(title-based method)의 문제점을 극복하기 위해, 제목과 두문의 정보를 결합시켜 충분한 어휘정보를 확보하도록 하였다. 특히 육하원칙 활성화도, 육하원칙 범주 개수, 문장 길이, 문장의 위치 등과 같은 다양한 요소들을 문장 중요도 계산에 반영함으로써 보다 중요한 정보를 포함하면서도 가독성이 높은 문장들이 요약문으로 선택될 수 있도록 고려하였다. 제안된 방법론의 정확률은 74.7%로서 기존의 두문 기반 기법보다 우수한 성능을 보였으며, 신문기사를 자동 요약하는데 있어서 충분히 효과적으로 사용될 수 있는 방법론임을 실험을 통해 입증하였다.

키워드 : 신문기사, 문서요약, 육하원칙, 활성화도

Abstract In a newspaper, 5W1H information is the most fundamental and important element for writing and understanding articles. Focusing on such a relation between a newspaper article and the 5W1H, we propose a summarization method based on the activation degree of 5W1H. To overcome problems of the lead-based and the title-based methods, both of which are known to be the most effective in newspaper summarization, sufficient 5W1H information is extracted from both a title and a lead sentence. Moreover, for each sentence, its weight is computed by considering various factors, such as activation degree of 5W1H, the number of 5W1H categories, and its length and position. These factors make a great contribution to the selection of more important sentences, and thus to the improvement of readability of the summarized texts. In an experimental evaluation, the proposed method achieved a precision of 74.7% outperforming the lead-based method. In sum, our 5W1H approach was shown to be promising for automatic summarization of newspaper articles.

Key words : newspaper articles, summarization, 5W1H, activation degree

1. 서 론

문서요약(text summarization)이란 대상문서로부터 가장 중요한 정보를 추출하여 특정 사용자나 작업에 적합한 축약된 형태의 문서를 생성하는 작업으로서, 크게 생성 요약(abstract) 기법과 추출 요약(extract) 기법으로 나눌 수 있다. 생성 요약은 전체 문서의 내용을 압축하여 원문에 없는 새로운 문장을 생성해 내는 방법으로 구현하기 어렵고, 많은 지식 자원(knowledge resource)을 필요로 한다. 그러나 추출 요약은 원문에서 상대적으로 중요한 문장을 추출하여 요약문으로 제시하는 것으로서 생성 요약 기법에 비해 쉬운 접근 방법이며, 주로 통계적인 분석에 기반한 방법론들이 사용되고 있다. 최근에는 추출 요약 기법의 주요 단점인 문장 가독률의 저하를 완화시키기 위해, 중요 문장들을 추출한 후 문장

· 본 연구는 첨단정보기술연구센터를 통한 과학재단 및 2002년도 두뇌한국21사업에 의하여 지원 되었음
† 비 회 원 : (주) 얄리 대표이사
chiwoo@postech.ac.kr
†† 비 회 원 : 포항공과대학교 컴퓨터공학과
prizer@postech.ac.kr
††† 종신회원 : 포항공과대학교 컴퓨터공학과 교수
jhlee@postech.ac.kr
논문접수 : 2003년 6월 4일
심사완료 : 2004년 1월 6일

이 너무 길거나 짧을 때 문장 길이를 적절히 조정해 주거나 또는 추출된 문장들 안에 출현한 대용어들을 처리해 주는 교정 요약(revision) 기법도 연구되고 있다[1].

일반적으로 신문기사는 육하원칙에 의거하여 기술된다. '누가', '언제', '어디서', '무엇을', '어떻게', '왜'로 구성되는 육하원칙은 사건을 기술하는데 있어서 가장 핵심이 되는 요소들이므로, 하나의 신문기사로부터 그 기사가 강조하고자 하는 육하원칙 요소들을 추출하여 나열하면 그 기사에 대한 가장 이상적인 요약문을 생성할 수 있다. 본 논문은 육하원칙 요소들을 충실하게 기술하는 신문기사의 특성에 주목하여 육하원칙 활성화도(activation degree of 5W1H information)에 기반한 신문기사 요약 방법론을 제안한다.

제안하는 방법론은 추출 요약 기법 중 가장 우수한 방법으로 알려진 두문 기반 기법(頭文; lead-based method)과 제목 기반 기법(title-based method)의 문제점을 극복하기 위해, 제목과 두문의 정보를 결합시킴으로써 충분한 어휘정보를 확보하고 서로 부족한 부분을 보완하도록 구성하였다. 본 방법론에서는 우선 서로 결합된 제목과 두문으로부터 육하원칙 구성성분을 추출한 후, 제목과 두문에서 강조되고 있는 육하원칙 구성성분이 본문에서 어떻게 재사용되고 있는지를 분석하여 각 문장들의 육하원칙 활성화도를 계산하고, 그 문장에 사용된 육하원칙 범주의 개수, 문장의 길이 및 위치까지 반영시켜 최종적으로 그 문장의 중요도를 구한다. 그리고 이렇게 계산된 문장 중요도 수치에 기반하여 신문기사 본문으로부터 중요 문장들을 추출함으로써 요약문을 구성한다. 본 방법론은 각 언어별로 육하원칙 구성성분을 인식하기 위한 패턴만 추가하면 다른 언어를 대상으로도 방법론의 적용이 가능하다는 장점을 갖고 있다. 제안된 방법론의 정확률은 74.7%로서 신문기사를 자동 요약하는데 있어서 충분히 효과적으로 사용될 수 있는 방법론임을 실험을 통해 입증하였다.

본 논문의 구성은 다음과 같다. 2장에서는 문서 요약과 관련된 기존의 연구들과 문제점을 살펴보고, 3장에서는 본 연구의 대상이 되는 신문기사의 특징과 구조를 분석한다. 4장에서는 육하원칙 구성성분에 기반한 문서 요약 기법을 제안하며, 마지막으로 5장과 6장에서는 실험 결과 및 결론을 제시한다.

2. 기존 연구 및 문제점

신문기사는 그 내용상의 간결함과 명료성 때문에 그동안 문서 요약을 위한 많은 연구들에서 평가의 대상으로 사용되어 왔다. 지금까지의 연구 결과에 의해 두문 기반 기법(lead-based method), 제목 기반 기법(title-based method)[2,3], 문장간의 담화구조(discourse tree)를 이용한 기법[4-6]이나 어휘사슬(lexical chain)[7], 단어 공기정보(co-occurrence)를 이용한 기법[8-10], 그리고 기타 여러 가지 통계 정보를 이용한 기법[11,12] 등과 같은 다양한 접근법들이 제시되었으며, 그 중에서 두문 기반 기법과 제목 기반 기법의 성능이 가장 우수한 것으로 알려져 왔다. 특히, Brandow[13]는 여러가지 통계 정보들에 기반하여 구축한 자신의 방법론보다도 두문 기반 기법의 성능이 좋은 것으로 보고하고 있다.

두문 기반 기법은 신문기사의 첫머리에 위치하는 문장들(두문)이 가장 중요한 것으로 간주하여 이들만으로 요약문을 구성하는 방법이다. 그러나, 두문 기반 기법은 문장의 상위에 인용문이나 부가 설명문 등이 위치하고 있는 경우에는 요약 성능의 저하가 발생할 수 있으며[1], 또한 일반적으로 문장길이가 너무 짧은 문장들은 가독성을 위해 요약문에 포함시키지 않는 편이 좋으나[3, 15] 두문 기반 기법에서는 문장 길이에 대한 고려를 하고 있지 않기 때문에 중요한 정보를 내포했다고 볼 수 없는 짧은 문장을 요약문으로 선택하는 문제점이 발생한다.

제목 기반 기법은 제목 자체가 그 문서에 대한 핵심적인 요약이라는 관점에서 출발하는 방법론으로서, 제목에 출현한 단어들과의 유사도에 기반하여 문서 내의 중요 문장들을 추출하는 기법이다. 이 기법은 문서요약 문제를 정보검색 문제로 간주하여 접근하고 있는 방법론으로서, 정보검색에서 사용자가 입력한 질의문과 검색 대상이 되는 문서들과의 유사도 값에 의해 목표 문서를 찾는 것처럼, 제목 기반 기법은 제목을 하나의 질의문으로 보고 제목과 본문의 각 문장들을 비교하여 유사도 값이 큰 문장을 중요 문장으로 인식한다. 그러나 제목이 본문의 내용을 잘 내포하지 못하고 은유적으로 표현되어 있거나 또는 제목이 지나치게 짧아서 유사도 측정을 위한 충분한 어휘정보를 제공하지 못할 경우 성능 저하가 발생한다. 또한 이 방법은 일반적으로 문장길이가 비교적 길고, 중복된 단어가 많이 등장하는 문장을 중요 문장으로 선택하는 문제점을 안고 있다.

최근에는 신문기사의 작성 원칙인 육하원칙에 기반한 방법론도 제시되었다. [16]에서는 신문기사의 본문에 출현한 각 문장들 간의 유사도 관계를 이용하여 중심절을 추출한 후, 추출된 중심절과 유사도가 높은 절들 중에서 사건의 시간적, 공간적 배경, 사건의 원인과 결과에 해당하는 절을 휴리스틱 정보와 수사어구 정보를 이용하여 추출하는 방법을 제안하고 있다. 그러나, 이 방법은 추출된 정보를 어떻게 문장으로 구성하고 요약에 이용

1) 신문기사에서 인용문은 일반적으로 덜 중요한 문장이라고 알려져 있다[14].

할 것인가에 대한 해결책을 제시하지 못하였고, 휴리스틱 정보와 수사어구를 이용해서 추출된 절이 중심절의 사건, 또는 제목과 어떠한 연결고리가 있는지를 의미적으로 파악하지 못했다. [17]에서는 [16]과는 달리 실험대상 신문기사의 원문이 아닌 제목으로부터 육하원칙에 해당하는 정보를 추출하는 방법을 제안하였다. 그러나, 이 방법은 신문기사의 원문을 요약하는 방법이 아니라 단순히 신문기사의 제목을 구성하는 단어를 각 육하원칙 요소로 분류하기 위한 방법으로서 문서 요약 기법으로 보기는 힘들다.

이상에서 보는 바와 같이 기존의 연구들은 신문기사에 대한 구조적인 특징을 파악하거나 요약대상이 되는 문서의 특성을 고려하지 않고 요약방법을 일률적으로 적용한 결과, 성능이 상당히 저하되었음을 알 수 있다[13,18,19]. 또한 신문기사의 구조를 분석할 때, 육하원칙과 제목, 두문과의 관계를 심도있게 분석하지 않고 육하원칙 자체만을 강조한 결과 전혀 의도하지 않은 결과를 생성해 내기도 하였으며[16], 육하원칙을 제목에만 적용하여 피상적인 분석만 시도하는 등 방법론 자체에 한계를 가지고 있었다[17].

본 논문에서는 기존 연구에서의 문제점을 극복하기 위해 신문기사의 구조적 특성을 반영한 육하원칙 활성화도 계산 기법을 제안하고자 한다.

3. 육하원칙에 기반한 신문기사의 분석

3.1 신문기사의 특성

신문기사의 목적은 '누가', '언제', '어디서', '무엇을', '어떻게', '왜'에 해당하는 구체적 사실을 독자와 시청자에게 전달하는 것이기 때문에[20] 좋은 신문기사는 정확성(accuracy), 객관성(objectivity), 공정성(fairness) 등이 뒷받침되어야 한다[21]. 따라서, 복합문 보다는 단문 위주로 간결하게 기술되며, 단어들의 관계가 덜 복잡하다는 특징이 있다[14]. 그리고 하나의 문장에 하나의 아이디어를 쓰는 것을 원칙(one sentence, one idea)으로 하기 때문에 대부분 문장 하나가 단락 하나를 구성하고 있어서 단락을 구분할 필요성이 없다는 장점이 있으며[22, 23], 피동태는 거의 이용되지 않고 주로 능동태 문장 형태로 기술되는 경향이 있다. 또한 동작성이나 상태성을 지니고 있는 비실체성 보통명사가 목적어로 쓰였을 경우인 '농성을 하다', '수사를 벌이다' 등은 '-하다'와 결합하여 '농성하다', '수사하다'로 쓰는 것을 원칙으로 하고 있다[14,20,23].

3.2 신문기사의 구조

신문기사에서 두문(lead)이란 일반적으로 본문에서 가장 첫머리에 위치하는 문장을 가리킨다. 두문은 그 기사가 전달하고자 하는 핵심적인 육하원칙 내용을 중심으로 기술되며, 이러한 육하원칙의 각 성분들은 본문의 문장들에서 재사용되면서 두문에서 제시되었던 내용을 뒷받침하도록 구성된다. 따라서 신문기사에서 중요한 문장이란, 두문을 구성하는 육하원칙의 구성성분이 다시 재사용되면서 두문에서 제시된 내용을 보완하거나 두문에서 빠진 육하원칙 성분에 대해서 설명하고 있는 문장으로 볼 수 있다.

신문기사는 일반적으로 제목(title) > 두문(lead) > 본문(body)의 순으로 중요도가 점점 작아지는 경향을 갖고 있으며 이를 그림으로 나타내면 그림 1과 같이 역피라미드 구조로 표현할 수 있다. 이렇게 신문기사가 중요도에 기반한 역피라미드 구조로 구성되는 이유는 현장에서 기자가 작성한 신문기사를 편집실에서 편집할 때 시간적 또는 공간적인 제약으로 인하여 기사를 잘라내야 하는 경우 기사의 제일 끝부분부터 잘라내기 때문이다. 또한 기사를 읽는 사람들 가운데 25% 정도는 중간까지만 읽고 중단하기 때문에[24] 중요도가 높은 정보를 가장 앞쪽에 배치하고 뒷 부분으로 갈수록 상대적으로 중요도가 떨어지는 정보를 기술하도록 구성할 수 밖에 없다[22].

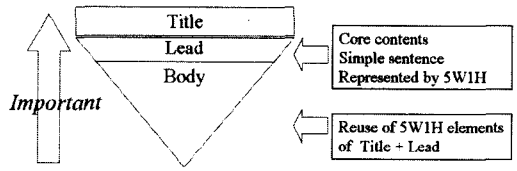


그림 1 신문 기사의 구조

신문기사의 두문은 역피라미드 구성의 첫머리에 해당하는 만큼 기사의 가장 주요한 내용이 육하원칙에 의거하여 압축적으로 기술되며, 두문 한 줄만 써도 전체 기사의 내용을 짐작할 수 있을 정도가 되어야 한다. 일반적으로 두문은 '주어+목적어+타동사'의 기본문형으로 구성되는 것을 원칙으로 하며[14,20,25], 본문에서는 두문에서 강조한 내용에 대해서 구체적으로 설명하거나 추가적인 정보를 기술한다[20-23].

3.3 육하원칙에 기반한 분석의 필요성

신문기사를 작성할 때는 육하원칙의 요소들인 '누가(who)', '언제(when)', '어디서(where', '무엇을(what)', '어떻게(how)', '왜(why)'에 해당하는 내용이 무엇인지를 분명히 밝혀야 한다. 그러나 신문기사의 구조가 중요한 문장의 순서대로 역피라미드 구조를 형성하고는 있지만 반드시 두문에 모든 육하원칙 요소들을 포함하고 있는 것은 아니다. 앞 절에서 기술했듯이 두문은 대개 '주어+목적어+타동사'의 구조로 구성되며, 이는 곧 두문에는 육하원칙의 요소 중 '누가(who)', '무엇을(what)', '어떻

게(how)'에 해당하는 내용들이 주로 기술됨을 의미한다. 두문에 모든 육하원칙 구성 성분을 기술하는 것은 일반적으로 신문기사 작성에 있어서 금기시 되고 있기 때문에[20], 두문에 기술되지 않은 나머지 육하원칙 요소들은 기사의 전체적인 구성을 고려하여 본문의 각 부분에 적절하게 배치되게 된다.

따라서 이렇게 두문 부분에 기사 요약에 가장 핵심적인 내용들이 모두 포함되어 있지 않으며 또한 가끔씩 인용문이나 부가 설명문이 포함되는 경우도 있기 때문에 두문에만 의존한 문서 요약은 한계가 있다.[2)]

보다 높은 요약 성능을 얻기 위해서는 두문 뿐만 아니라 본문의 모든 문장들까지 검토한 후, 그 기사에서 전달하고자 하는 육하원칙 요소에 해당하는 내용을 추출하여 요약문으로 제시하는 것이 바람직하다.

4. 육하원칙에 기반한 문서 요약 기법

기사에서 강조하고자 하는 내용에 해당하는 육하원칙의 각 성분들은 제목과 두문에 출현한 후, 본문의 문장들에서 재사용되면서 제목과 두문에서 주장하는 내용을 뒷받침하도록 구성된다. 따라서 본문에서 중요한 문장은, 제목과 두문을 구성하는 육하원칙의 구성성분이 다시 재사용되면서 제목과 두문에서 주장하고 있는 내용을 보완하거나 또는 제목이나 두문에 출현하지 않은 육하원칙 요소에 대해서 설명하고 있는 문장으로 볼 수 있다. 즉, 문장의 중요도는 그 문장 안에 기사가 강조하고자 하는 육하원칙 성분이 얼마나 활성화되었는지에 따라 판단될 수 있다. 본 논문의 목적은 제목과 두문에 출현한 육하원칙 성분들을 그 기사가 전달하고자 하는 육하원칙 요소라고 간주한 후, 비교 분석을 통해 본문 문장들로부터 제목과 두문에 출현하지 않은 육하원칙 요소들까지도 추출해 내는 데 있다.

본 연구에서 제안하는 전체적인 시스템 구성이 그림 2에 제시되어 있다. 신문기사가 입력되면 우선 품사 태깅과 구묶음(chunking)을 수행한 후, 전처리 단계로써 육하원칙 구성성분을 추출할 때 행위자를 보다 정확하게 인식할 수 있도록 피동형 문장을 능동형 문장으로 변환시키는 작업과 불필요한 단어(stop words)들을 제거하는작업을 수행한다. 다음으로 기 구축된 패턴 정보와 가도카와 시소러스(Kadokawa thesaurus)를 이용하여 제목과 두문, 본문의 각 문장들로부터 육하원칙 구성요소를 추출한다. 그리고, 결합된 제목과 두문으로부터 추출된 육하원칙 요소와 본문의 각 문장으로부터 추출된 육하원칙 요소를 비교하여 문장 가중치를 계산한 후,

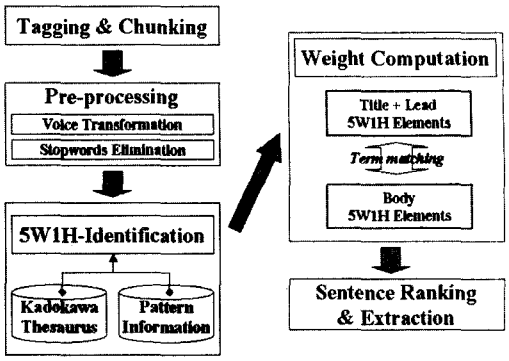


그림 2 시스템 구성도

가중치가 큰 문장을 중요 문장으로 선택한다. 이들 각 단계에 대한 세부 기술은 아래에 제시되어 있다.

4.1 전처리 작업 (Pre-processing)

육하원칙 성분 추출에 들어가기 전에 우선 먼저 피동문들을 능동문으로 변환시키는 작업을 수행한다. 피동문은 문장 상에서 행위주와 대상의 위치가 역전되어 있는 상태이므로 보다 정확한 육하원칙 성분의 인식, 즉 'WHO'와 'WHAT'에 해당하는 성분을 제대로 분석해내기 위해서는 피동문을 능동문 형태로 변환시키는 작업이 필수적이다.

피동문에는 타동사에 피동접미사 '-이/-히/-리/-기'를 삽입한 형태의 단형 피동과, '-어 지다, 되다, 받다, 당하다' 등의 표현으로 이루어진 장형 피동이 있다. 신문기사에서는 일반적으로 피동문보다 능동문을 사용할 것을 적극 권장하는데, 총 968 문장의 신문기사를 대상으로 조사한 결과 피동형으로 출현한 동사는 전체 4,678개의 동사 중 약 3.5%에 해당하는 158개였다. 이들 158개의 피동형에 대한 능동형으로의 변환 실험 결과가 표 1에 제시되어 있으며, 정확률은 95.6%(151개/158개)였다.

피동문을 능동문으로 변환시키는 과정에서 오류가 발생했던 문장은 '하객이 몰리다' 또는 '눈이 쌓이다' 등의 경우와 같이, 실제로는 자동사로 사용된 단어지만 '타동사+피동접사'의 형태로 분석되는 바람에 피동형으로 잘못 인식하여 '하객을 몰다'와 '눈을 쌓다'로 변형시키는

표 1 피동문의 능동문 변환 실험 결과

	개수	오류
피동접미사 (이/히/리/기)	58	7
-어 지다	75	0
-되다	5	0
-받다	12	0
-당하다	8	1
총 개수	158	8

2) [13, 18]의 연구 결과에 따르면 두문 기반 기법은 요약율 20%일 때 약 50%~60% 정도의 정확률을 나타내는 것으로 보고되어 있다.

경우가 대부분이었다. 즉, 자동사와 타동사의 피동형이 동일한 단어일 때 이들을 제대로 구분하지 못하기 때문에 발생하는 문제인데, 실험 결과 이러한 오류들의 빈도수가 그다지 높지 않으므로 본 연구에서는 무시하였다.

또한 기사의 내용 파악에 그다지 중요한 구실을 하지 못한다고 판단되는 단어들(stop words)은분석에 들어가기 전에미리 제거시킴으로써 이들이 육하원칙 요소로 추출되지 않도록 하였다. 대명사나 부사류, 관사류, 지시형용사 등과, 신문기사에서 자주 등장하는 상투어구인 '말했다, 발표했다, 밝혀졌다, 제안했다' 등이 이에 해당한다.

4.2 육하원칙 요소의 인식 및 추출

표 2는 기존 연구에서 사용된 육하원칙 범주에 대한 정의를 보이고 있다. 국립국어연구원[26]에서는 주어로 사람, 동물, 국가 등이 오게 되면 'WHO'로 할당하였으며, 'WHEN'은 시간, 'WHERE'는 지명 또는 기관, 'WHAT'은 술어에 대한 목적어, 'HOW'는 '~아/~어/~다/~며'와 같은 연결어미를 포함한 서술어, 그리고 'WHY'는 '~하기 위해서' 또는 '~을 위해서'가 붙는 어절이라고 설명하였다. 또한 윤만근[27]은 하나의 문장 안에서 '언제 어디에서 왜 어떻게 그가 그것을 했는가(When, where, why and how did he do?)'에 대한 답변에 해당하는 요소들을 각 육하원칙의 범주로 정의하였다. 그러나 이렇게 정의된 육하원칙의 범주만으로는 문장 내에 있는 모든 요소들에 육하원칙의 범주를 할당시키는데 한계가 있다. 예를 들어 아래와 같은 문장의 경우를 살펴보자.

예문) 정부는 닭값 안정을 위해 올해 처음으로
WHO WHY WHEN ?
닭고기를 수입한다.
WHAT ?

표 2에 제시된 각 육하원칙 범주들의 정의에 따라 위의 예문에 육하원칙 범주를 할당시키면 '정부'와 '닭값 안정을 위해', '올해', '닭고기를'은 각각 'WHO', 'WHY', 'WHEN', 'WHAT'에 해당하는 요소들임을 알 수 있다. 그러나 부사격 조사가 붙어있는 '처음으로'와 문장의 술어로 사용된 '수입한다'는 표 2에 제시된 육하원칙의 범주들 중 어느것에도 해당되지 않기 때문에, 위 문장에 출현한 모든 정보를 활용할 수 있도록 하기 위해서는 기존의 육하원칙 범주에 대한 정의를 확장할 필요성이 존재한다. 따라서 본 연구에서는 기존의 육하원칙 범주에 'SE(supplementary element)'와 'PE(predicate element)'라는 새로운 범주들을 추가한 '확장된 육하원칙(extended 5W1H) 범주'를 정의하여 사용하였다. 새로이 추가된 'SE'와 'PE' 요소는 기존의 육하원칙 범주에 할당되지 못하는 문장 성분들에도 범주 할당을 가능하게 함으로써 향후 단계에서 수행되는 육하원칙 활성화도 계산이 보다 세밀하게 이루어질 수 있도록 기여한다. 표 3은 본 연구에서 사용되는 의사 육하원칙 범주의 인식을 위한 패턴 정보를 정리한 것이다.

전처리가 끝나면 표 3과 같이 구축된 패턴 정보와 시소러스의 의미코드에 기반하여 제목과 두문, 그리고 본문의 문장들에 출현한 각 단어들에 육하원칙 범주를 할당시킨다. 본 연구에 사용되는 한국어 분석 사전은 한일기계 번역 시스템[3]을 위해 구축된 사전으로써, 사전에

표 3 확장된 육하원칙 범주 인식을 위한 패턴 정보

육하원칙 범주	패턴 정보
WHO	유정물(사람, 동물 등)/국가/조직 + 주격조사(이,가)/보조사(은,는)
WHEN	시간성 단어 (숫자+년/월/일/시/분/초, 작년, 올해, 오전, 오후, 낮, 밤 등)
WHERE	지명/기관/장소 + 에/에서/(으)로(부터) 등
WHAT	WHO에 해당되지 않는 명사구 + 주격조사(이,가)/보조사(은,는) 명사구 + 목적격조사(을,를)
HOW	방법, 수단 (~을 통해, ~로/을 이용하여, ~에 의해, ~ 과정으로/에서 등)
WHY	원인, 이유 (~기 위하여, ~을 위해, ~ 때문에, ~을 이유로, ~로 인해 등)
SE	위의 범주에 할당되지 못한 명사구 + 조사
PE	술어로 사용된 용언

3) 본 연구에서는 포항공대 지식 및 언어공학 연구실이 개발한 COBALT-KJ 시스템의 사전을 사용하였다.

표 2 기존 연구에서의 육하원칙 범주에 대한 정의

육하원칙 범주	국립국어연구원	윤만근
WHO	사람, 동물, 국가, 조직이 주어로 오는 경우	사람이 주어로 오는 경우
WHEN	시간	시간
WHERE	지명, 기관	장소
WHAT	술어의 목적어	술어의 목적어
HOW	연결어미 '-아/-어/-다/-며'를 갖는 술어	방법
WHY	'-하기 위해서', '-을 위해서'	이유

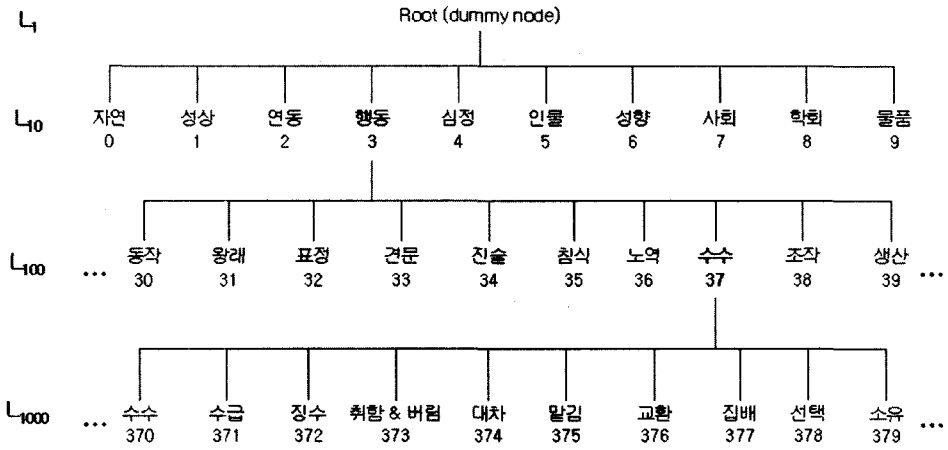


그림 3 가도카와 시소러스의 개념 계층 구조

수록된 모든 표제어들은 가도카와 시소러스[28]의 의미 코드가 부착되어 있다. 가도카와 시소러스는 총 1,110개의 개념과 4단계의 계층구조를 가지고 있으며, L1, L10, L100 레벨에 속해 있는 개념들은 각각 10개의 하위 개념들로 나뉜다(그림 3). 따라서 가도카와 시소러스의 의미코드에 의해 유정명사, 무정명사의 구분 뿐만 아니라 사람, 동물, 단체, 지명, 장소, 시간성 단어 등의 세분화된 의미 구분까지 가능하기 때문에 단어의 표층형태(surface form) 수준의 패턴 매칭이 아닌 의미코드 수준의 패턴 매칭을 수행함으로써 시스템의 성능을 높였다.

육하원칙 성분 추출 시 발생할 수 있는 오류의 유형으로는 의인화된 무정명사로 인한 경우와 이중 주어, 이중 목적어의 경우가 있다. 먼저 의인화된 무정명사의 경우를 살펴보면, '저기압이 비를 떨어뜨렸다'라든지 '정부군 비행기는 아쿠엠 마을에 폭탄을 투하했다'와 같은 문장에서 주어로 쓰인 '저기압'과 '비행기'는 무정물이기 때문에 'WHAT'으로 할당되게 된다. 그러나 이러한 경우, 이들 단어들은 비록 무정물이지만 의인화된 행위주들이기 때문에 'WHO' 성분으로 할당되는 것이 맞다. 목적어를 수반한 타동사 구문의 이러한 문장표현은 실험 대상으로 선택된 100개의 신문기사에서 출현한 968개의 문장 중 전체의 약 1.5%에 해당하는 15개의 문장에서만 발생하였으며, 이로 미루어 볼 때 전체 시스템의 성능에 미치는 영향이 미미할 것으로 판단하고 본 연구에서는 무시하였다. 이중 주어와 이중 목적어 또한 'WHO'와 'WHAT' 성분을 할당하는 데 있어서 애매성이 존재하므로 오류를 발생시킬 가능성이 크지만, 조사 결과 이중 주어는 전체 문장에서 3개, 이중 목적어는 전혀 출현하지 않았으므로 이것도 역시 무시하였다.

4.3 제목에서의 육하원칙 범주 할당 및 제목과 두문의 육하원칙 요소 통합

표 3에 제시된 패턴 정보에 기반하여 육하원칙 요소를 추출할 때 두문 및 본문의 문장들에 출현한 육하원칙 요소들은 대부분 잘 인식되지만, 제목의 경우는 일반적으로 조사를 생략한 채 쓰는 경우가 빈번하기 때문에 표 3에 기술된 패턴 정보만으로는 잘 인식되지 않는다. 따라서 제목에서의 육하원칙 성분 인식을 위해서는 부가적인 처리과정이 필요하게 된다. 본 연구에서는 제목에 출현한 단어들 중 조사 정보의 부재로 인해 육하원칙 범주 할당에 실패한 단어들을 대상으로 다음과 같은 규칙들이 순서대로 적용되도록 설정하였다.

규칙 1 : 서로 인접한 단어들을 묶어서 명사구를 만들어 본 후, 그 명사구가 본문에 실제로 출현하는지 체크하여 명사구를 설정한다. 예를 들어, 기사 제목이 '세계 언어 절반 사멸 위기'였다면 '세계 언어', '언어 절반', '절반 사멸', '사멸 위기' 등으로 단어들을 묶은 후 본문의 문장들을 검색한다. 그리고 본문에서 '사멸 위기'라는 명사구가 실제로 출현했다면 이후의 단계에서는 '사멸'과 '위기'는 개별적으로 처리되지 않고 동일한 육하원칙 범주로 할당되도록 한다.

규칙 2 : 두문에도 동일한 단어가 출현해 있다면 그 단어의 육하원칙 범주로 할당한다.

규칙 3 : 단어가 유정명사, 국가, 조직이면 'WHO'로 할당한다.

규칙 4 : 단어가 무정명사이면 'WHAT'으로 할당한다.

규칙 5 : 단어가 서술성 명사이면 'PE'로 할당한다.

규칙 6 : 위의 조건에 하나도 일치하지 않으면 'SE'

로 할당시킨다.

실험 대상으로 선정된 100개의신문기사에서 제목으로 출현한 총 456개의 단어를 대상으로 육하원칙 범주 할당 성능을 평가한 결과, 잘못 할당된 단어는 총 67개로써 85.3%(389개/456개)의 정확률을 기록하였다.

제목과 두문, 본문에 출현한 각 단어들에 대한 육하원칙 범주 할당이 끝난 후, 다음으로 제목과 두문에 나타난 육하원칙 요소들을 서로 통합시키는 작업을 수행한다. 제목과 두문의 정보를 통합시키게 되면 서로 부족한 육하원칙 구성성분을 보완할 수 있기 때문에, 이들 각각을 독립적으로 사용하는 제목 기반 기법과 두문 기반 기법에 비해 보다 충분한 어휘 정보를 얻을 수 있다는 장점이 있다. 제목과 두문의 정보를 통합시킬 때 동일한 단어가 중복되어 출현한 경우에는 한번만 포함시킨다. 단, 이렇게 중복 출현된 단어는 육하원칙 활성화도 계산에 있어서 한번만 출현한 단어에 비해 좀 더 높은 가중치를 부여받게 된다. 이러한 통합 과정의 예가 그림 4에 제시되어 있다. 이상과 같이 제목+두문과 본문의 각 문장으로부터 추출된 육하원칙 요소들은 그림 5와 같이 단어 리스트(term list) 형태로 구축된다.

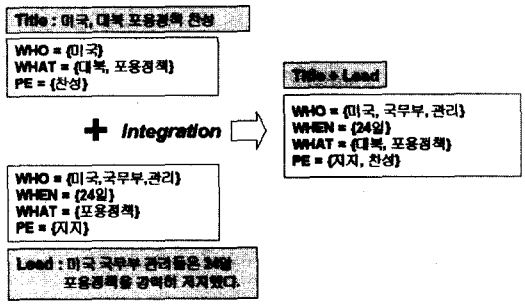


그림 4 제목과 두문의 육하원칙 요소 결합 예

4.4 육하원칙 활성화도에 기반한 문장 가중치 계산

본 연구에서는 우선 가장 중요한 문장인 두문을 디폴트로 선택한 후, 요약문을 구성할 나머지 문장들은 육하원칙 활성화도에 기반하여 계산된 문장 가중치 순위에 따라 추가시키는 방식으로 요약문을 구성한다. 이 때 문장의 중요도 평가에 사용되는 문장 가중치 수식 (1)은 아래에 제시된 네가지 요소를 반영하여 계산되도록 설정하였다.

$$W_{sentence} = W_{position} \times (W_{activation} + W_{vm\ category} + W_{length\ penalty}) \quad (1)$$

첫째, 제목과 두문에 출현한 육하원칙 구성성분이 현재의 문장에서 얼마나 활성화되었는지의 정도를 고려한다. 제목과 두문은 신문기사에서 가장 강조하고자 하는 내용을 배치한 것이므로 제목과 두문에서 출현했던 육하원칙 요소가 다시 출현한 문장일수록 중요도가 높다 ($W_{activation}$).

둘째, 신문기사의 모든정보는 육하원칙에 의거하여 기술된다. 따라서 문장 안에 다양한 범주의 육하원칙 정보를 포함하고 있는 문장일수록 정보량이 높은 문장이므로 중요도가 높다($W_{num\_category}$).

셋째, 비록 문장 안에 많은 양의 정보를 담고 있더라도 지나치게 긴 문장은 요약문으로 적절하지 않다. 따라서 지나치게 길거나 짧은 문장은 음의 가중치(penalty)를 부여하여 중요한 문장으로 선택되는 것을 배제한다 ($W_{length\_penalty}$).

넷째, 앞서 기술했듯이신문기사는 중요한 문장일수록

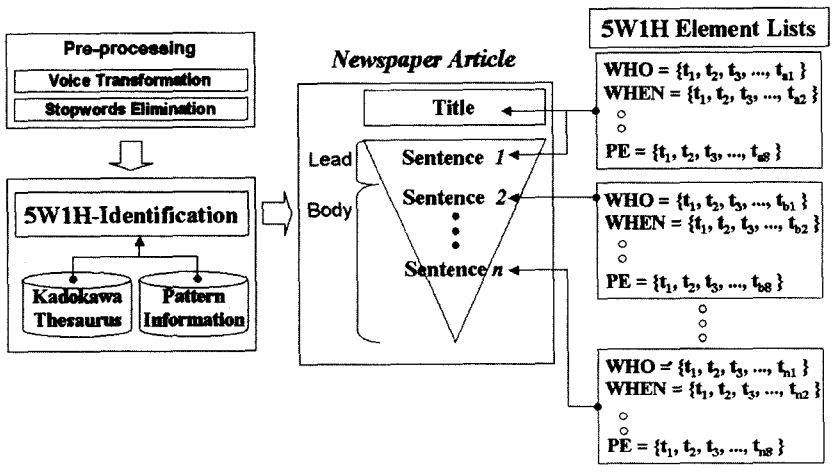


그림 5 육하원칙 요소의 추출

기사의 상위에 배치시키는 경향이 있다. 따라서 전체 기사 상에서 현재 문장이 배치되어 있는 위치가 앞부분일수록 문장의 중요도가 높다($W_{position}$).

이후의 절에서는 이들 각각의 가중치 요소에 대해 상세히 기술한다.

4.4.1 육하원칙 요소의 활성화도

기사에서 강조하고자 하는 내용에 해당하는 육하원칙의 각 성분들은 제목과 두문에 출현한 후, 본문의 문장들에서 재사용되면서 제목과 두문에서 주장하는 내용을 뒷받침하도록 구성된다. 따라서 본문에서 중요한 문장은, 제목과 두문을 구성하는 육하원칙의 구성성분이 다시 재사용되면서 제목과 두문에서 주장하고 있는 내용을 보완하거나 또는 제목과 두문에서 기술되지 않은 육하원칙 요소에 대해 설명하고 있는 문장으로 볼 수 있다. 즉, 문장의 중요도는 그 문장 안에 기사가 강조하고자 하는 육하원칙 성분, 다시 말하면 제목과 두문에 출현한 육하원칙 구성성분이 현재 문장 안에서 얼마나 활성화되었는지의 정도에 따라 판단될 수 있다. 현재 문장에서의 육하원칙 활성화도 $W_{activation}$은 그 문장에 속한 모든 육하원칙 요소들의 활성화도의 합으로 계산된다.

$$W_{activation} = \sum_{i=1}^{k} ActWeight_i \qquad (2)$$

위의 수식에서 k는 현재의 문장 안에 존재하는 육하원칙 요소들의 개수를 나타내며, $ActWeight_i$는 i번째 육하원칙 요소의 육하원칙 활성화 가중치를 의미한다. 이 때, 제목+두문에서와 동일한 육하원칙 범주로 본문의 문장에서 출현한 경우에는 그 육하원칙 요소가 특별히 강조된 것으로 간주하여 보다 높은 가중치를 부여하도록 하였다. 또한 제목과 두문은 신문기사에서 가장 중요한 부분이기 때문에 제목과 두문에 중복해서 출현한 육하원칙 성분이라면 그 또한 강조된 단어라고 볼 수 있으므로 높은 가중치를 부여한다. 각 세분화된 경우에 대한 *ActWeight* 수치는 반복적인 실험을 통해 경험적으로 결정되었으며, 그 설정된 값이 표 4에 제시되어 있다. 표에서 'Same Category'는 현재의 육하원칙 요소가 제목+두문에서 동일한 육하원칙 범주로 출현했는가의 여부를 나타내며, 'Overlapped'는 현재의 육하원칙 요소가 제목과 두문에서 중복해서 출현했는가의 여부를 나타낸다. 만약 제목+두문에서 출현하지 않은 요소라면 *ActWeight*는 0으로 할당된다.

표 4 육하원칙 활성화도 가중치의 설정

$ActWeight_i$		Overlapped	
		Yes	No
Same Category	Yes	1.7	1.4
	No	1.2	1.0

그림 6은 각 경우별 육하원칙 활성화 가중치 계산의 예를 보이고 있다. 먼저 1번의 경우, '괴한'은 제목과 두문에서 한번만 출현한 단어이며 현재 문장(즉, i번째 문장)에서는 'WHAT' 성분으로 출현하였으나 제목+두문에서는 'WHO' 성분으로 출현했으므로 동일한 육하원칙 범주로 출현한 경우도 아니다. 따라서 육하원칙 활성화 가중치는 1.0이 부여된다. 2번의 '총기'는 제목과 두문에서 중복해서 출현한 단어이며 범주도 'WHAT'으로써 동일하다. 따라서 가장 높은 가중치인 1.7을 부여받게 된다. 마찬가지로 3번 경우의 단어 '에루살렘'은 중복해서 출현한 단어지만 육하원칙 범주는 변경되었으므로 1.2의 가중치를 부여받는다.

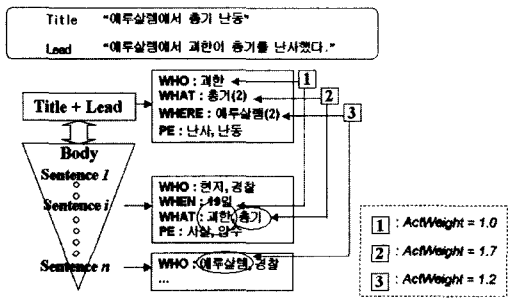


그림 6 육하원칙 활성화도 가중치 계산의 예

4.4.2 육하원칙 범주의 개수

문장에 포함된 육하원칙 범주의 개수가 많다는 것은 하나의 문장에 행위자(WHO), 장소(WHERE), 시간(WHEN), 대상(WHAT) 등에 대한 다양한 정보들을 포함하고 있다는 의미이므로 중요한 문장으로 간주될 수 있다. 이와 비슷한 개념으로 육하원칙 범주의 개수가 아닌 육하원칙 요소 개수의 많고 적음에 따라 문장의 중요도를 판별할 수도 있겠지만, 포함하고 있는 육하원칙 요소의 개수가 많더라도 이들이 모두 'WHO'에 해당하는 정보라든지 또는 'WHEN'에 해당하는 정보일 경우 등 한두 범주에 치우친 정보만을 제공하는 경우도 존재할 수 있기 때문에 현재 문장이 제공하는 정보의 양에 대한 객관적인 기준으로 사용되기엔 한계가 있다. 따라서 본 연구에서는 육하원칙 요소의 개수가 아닌 범주의 개수로 문장이 제공하는 정보량을 평가하였다. 아래의 수식에서 Ncat은 현재 문장에 포함된 육하원칙 범주의 개수를 의미하며, 분모에 사용된 8은 육하원칙 범주의 총 개수이다(WHO, WHEN, WHERE, WHAT, WHY, HOW, SE, PE).

$$W_{vm\ category} = N_{cat}/8 \qquad (3)$$

4.4.3 문장의 길이

요약문의 선택시 문장의 길이를 고려하는 이유는 본

문에서 요약문으로 뽑힐 중요한 문장을 선택할 때, 너무 길거나 짧은 문장은 음의 가중치(penalty)를 부여하여 가능한 한 중요한 문장으로 선택되지 않도록 하기 위해서이다. 본 연구에서는 인간에 의한 신문기사 요약 실험을 통해, 일반적으로 인간에 의해 중요한 문장으로 판단되어 선택된 문장들은 대부분 10~30어절 정도의 길이를 갖는다는 것을 경험적으로 학습하였으며 이러한 실험 결과를 문장의 가중치 계산에 반영하고자 하였다. 문장의 길이에 따른 페널티는 다음과 같이 설정하였다. 아래의 수식에서 N_{eojeol}은 현재 문장이 갖고 있는 어절의 개수를 의미한다.

$$W_{length\ penalty} = \begin{bmatrix} N_{eojeol}/10 - 1 & , & N_{eojeol} < 10 \\ 0 & , & 10 \le N_{eojeol} \le 30 \\ 3 - N_{eojeol}/10 & , & N_{eojeol} > 30 \end{bmatrix} \quad (4)$$

4.4.4 문장의 위치

앞서 3.2절에서 기술했듯이 신문기사는 중요도에 따라 역피라미드 형태로 문장을 배치하기 때문에 기사의 앞부분일수록 중요한 문장일 가능성이 높다. 따라서 이러한 신문기사의 특성을 문장 가중치 함수에 반영하기 위해 문장의 위치에 따른 가중치 함수를 다음과 같이 설정하였다.

$$W_{position} = \frac{2}{2N_{sentence} - 1}(2N - i), \quad 1 \le i \le N_{sentence} \quad (5)$$

위의 수식에서 $N_{sentence}$는 전체 문장 개수를 나타내며, i는 현재 문장의 위치를 가리킨다. 이렇게 구성된 문장 위치에 따른 가중치 함수는 첫번째 문장, 즉 두문일 때 최대값 2.0을 가지며 마지막 문장일 때 최소값 1.0을 갖게 되어, 첫번째 문장부터 마지막 문장까지 차등적으로 가중치를 부여하는 역할을 한다. 그러나 제안된 방법론에서 요약문 생성시, 두문은 디폴트로 선택되기 때문에 실제로는 두번째 문장부터 가중치 계산이 적용되는 것으로 볼 수 있다.

신문 기사에 있어서 가장 중요한 특성은 중요 문장의 역피라미드형 배치이며, 이는 기존 연구들에서 두문 기반 기법의 성능이 가장 우수했다는 점으로 충분히 입증되었다. 따라서 기존의 두문 기반 기법의 특성을 최대한으로 반영하기 위해, 문장 가중치 계산 수식 (1)에서 다른 가중치들이 서로의 합으로 구성되는 것에 비해 문장 위치에 따른 가중치는 전체 수식에 곱셈으로 적용되도록 함으로써 문장 위치에 따른 가중치가 가장 큰 영향력을 미치도록 수식을 구성하였다.

5. 실험 결과

본 논문에서 사용한 실험데이타는 조선일보 웹사이트에서 제공하는 경성기사[4] 중에서 총 100건의 신문기사를 무작위로 선택하였다. 선택된 신문기사는 평균 약 9.7개의 문장으로 구성되어 있으며, 그 중 두문이 존재하는 기사는 총 96건이었다. 정답으로 사용될 요약문은 기사 요약율을 30%로 설정하여 연구원 3명의 합의에 의해 각 기사로부터 중요 문장 3개를 추출하는 방식으로 구축하였다.

기존 시스템과의 성능의 비교를 위해 본 논문에서 제안하는 육하원칙 활성화도를 이용한 방법 외에 두문 기반 기법과 제목 기반 기법, 그리고 마이크로소프트사의 MS Word에서 제공하는 문서요약 시스템 각각에 대해 동일한 실험 문서 집합을 이용하여 요약 성능을 평가하였다. 일반적으로 요약 성능의 평가 척도로는 정확률과 재현율, 그리고 F-measure가 주로 사용되는데, 본 연구에서는 사람에 의해 구축된 요약문의 개수와 위에서 언급한 각각의 방법들에 의해 추출될 문장의 개수를 3개로 동일하게 설정했기 때문에 정확률과 재현율, F-measure 값은 동일하다. 표 5에 각방법론들에 의한 요약실험 결과가 정리되어있다. 실험 결과, 본 논문에서 제안하는 방법은 74.7%로써 가장 높은 정확률을 기록했으며, 다음으로 두문 기반 기법과 제목 기반 기법이 우수한 성능을 보였다. 그러나 상용 소프트웨어인 MS Word는 다른 방법론들에 비해 월등하게 저조한 성능을 보여 대조를 이루었다.

또한 문장 가중치 함수의 구성에 따른 성능 비교를 위해 아래와 같은 세 종류의 문장 가중치 함수를 설정한 후 이들 각각의 성능을 평가하였으며, 그 결과가 표 6에 제시되어 있다. Case 1은 앞서 제시된 문장 가중치 함수를 그대로 사용한 경우이며, Case 2는 가중치 계산 시 문장 위치에 의한 영향을 배제시킨 경우, 마지막으로 Case 3은 단지 육하원칙 활성화도만을 가중치에 반영한 경우이다. 실험 결과, 육하원칙 활성화도만을 이용해도 (Case 3) 두문 기반 기법의 성능(70.0%)에 근접하는 좋은 요약 성능을 보였으나(69.7%), 설정된 모든 요소를 가중치 함수에 반영했을 때(Case 1) 가장 좋은 성능을 보임을 알 수 있었다.

Case 1 :

$$W_{sentence} = W_{position} \times (W_{activation} + W_{vm\ category} + W_{length\ penalty})$$

Case 2 :

$$W_{sentence} = W_{activation} + W_{vm\ category} + W_{length\ penalty}$$

Case 3 :

$$W_{sentence} = W_{activation}$$

4) 일반적으로 신문기사는 그 내용과 성격에 따라 경성 기사(hard news)와 연성 기사(soft news)로 분류된다. 정치, 경제, 사회 등에 관한 뉴스성 기사들은 경성 기사에 해당되며, 논설이나 칼럼과 같은 해설성 기사는 연성 기사에 해당된다.

표 5 각 방법론들의 요약 성능 비교

맞은 문장 개수	224	210	204	131
정확률	74.7%	70.0%	68.0%	43.7%

표 6 문장 가중치 함수 구성에 따른 요약 성능 비교

맞은 문장 개수	224	217	209
정확률	74.7%	72.3%	69.7%

6. 결 론

본 논문에서 제시된 육하원칙 활성화도를 이용한 방법은 다음과 같은 장점을 가지고 있다.

첫째, 제목과 두문에 출현한 육하원칙 요소들을 서로 결합시킴으로써 문서 요약의 중요한 단서가 될 수 있는 충분한 어휘 정보를 확보할 수 있도록 하였으며, 이를 통해 기존의 두문 기반 기법과 제목 기반 기법이 갖는 한계점을 보완하였다.

둘째, 제목과 두문에 출현한 육하원칙 요소들과 본문의 각 문장에서 출현한 육하원칙 요소들을 비교하여, 육하원칙 요소의 활성화 정도를 파악함으로써, 제목과 두문에서 강조하고 있는 육하원칙 요소가 다시 본문에서 재사용되면서 제목과 두문에서 주장하고 있는 내용을 뒷받침하는 중요 문장을 선택하도록 하였다.

셋째, 문장 내에 포함된 육하원칙 범주의 개수를 고려함으로써 보다 다양한 정보를 갖고 있는 문장이 중요한 문장으로 선택될 수 있도록 문장 가중치 함수에 반영하였다.

넷째, 지나치게 짧거나긴 문장이 중요 문장으로 선택되는 경우를 배제하기 위해, 문장 길이에 기반한 패널티 값을 문장 가중치 함수에 반영함으로써 요약문의 가독성을 높이고자 하였다.

다섯째, 기사의 상위에위치한 문장일수록 높은 가중치를 부여받게 함으로써, 중요한 내용일수록 기사의 앞부분에 배치시키는 신문기사의 특성을 적극 반영할 수 있도록 하였다.

여섯째, 각 언어별로 육하원칙 구성성분을 인식하기 위한 패턴만 추가하면 다른 언어를 대상으로도 방법론의 적용이 가능하다.

제안된 방법론의 정확률은 74.7%로서 기존의 두문 기반 기법보다 우수한 성능을 보였으며, 신문기사를 자동요약하는데 있어서 충분히 효과적으로 사용될 수 있는 방법론임을 실험을 통해 입증하였다.

참 고 문 헌

[1] Mani, I., Automatic summarization, John Benjamin Publishing Company, 2001.

[2] Edmundson, H. P., "New Methods in Automatic Extracting," Journal of the ACM, Vol.16, No.2, pp.264-285, 1969.

[3] Teufel, S. and Moens, M. "Sentence Extraction as a Classification Task," In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pp.58-65, 1997.

[4] Marcu, D., "Building Up Rhetorical Structure Trees," In Proceedings of the 13th National Conference on Artificial Intelligence, Vol.2, pp.1069-1074, 1996.

[5] Marcu, D., "The Rhetorical Parsing of Natural Language Texts," In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics(ACL'97/EACL'97), pp.96-103, 1997.

[6] Marcu, D., "Discourse trees are good indicators of importance in text," In Inderjeet Mani and Mark Maybury, eds, Advances in Automatic Text Summarization, pp.123-136, The MIT Press, 1999.

[7] Brazilay, R. and Elhadad, M., "Using Lexical Chains for Text Summarization," In Inderjeet Mani and Mark Maybury, eds, Advances in Automatic Text Summarization, pp.111-121, The MIT Press, 1999.

[8] Salton, G. and Singhal, A., "Automatic Text Theme Generation and the Analysis of Text Structure," Cornell U. Technical Report TR 94-1438, 1994.

[9] Salton, G. et al., "Automatic Text Decomposition Using Text Segments and Text Themes," '96 ACM Conference on Hypertext, pp.53-65, 1996.

[10] Salton, G. et al., "Automatic Text Structuring and Summarization," Information Processing and Management, Vol.33, No.2, pp.193-207, 1997.

[11] Lin, C. Y. and Hovy, E., "Identifying Topics by Position," In Proceedings of the 5th Conference on Applied Natural Language Processing(ANLP'97), pp.283-290, 1997.

[12] Hovy, E. and Lin, C. Y., "Automated Text Summarization in SUMMARIST," In Inderjeet Mani and Mark Maybury, eds, Advances in Automatic Text Summarization, pp.81-94, The MIT Press, 1999.

[13] Brandow, R., Mitze, K. and Rau, L. F., "Automatical condensation of electronic publications by sentence selection," Information Processing and Management, Vol.31, No.5, pp.675-685, 1995.

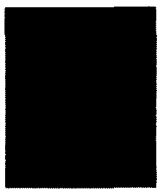
[14] 고혜련, 신문 취재와 기사작성, 중앙M&B, 2001.

[15] Kupiec, J., Pedersen, J. and Chen, F., "A Trainable Document Summarizer," In Proceedings of ACM-SIGIR'95, pp.68-73, 1995.

[16] 이현주, 김계성, 구상옥, 이상조, "신문기사에서 육하원칙 중심의 정보추출", 한국정보과학회 춘계 학술발표

논문집, pp.361-363, 2001.
[17] Okumura, A., Ikeda, T. and Muraki, K., "Text Summarization based on Information Extraction and Categorization Using 5W1H," Journal of Natural Language Processing, Vol.6, No.6, pp.27-44, 1999.
[18] Marcu, D., "Improving Summarization through Rhetorical Parsing Tuning," In Proceeding of the COLING ACL Workshop on Very Large Corpora, Montreal, Canada, 1998.
[19] 김재훈, 김준홍, "도함유사도를 이용한 한국어 추출문서 요약", 제10회 한글 및 한국어 정보처리 학술발표 논문집, pp.238-244, 2000.
[20] 이행원, 취재보도의 실제, 나남출판, 1999.
[21] 김지용, 현장신문론, 도서출판 쟁기, 1996.
[22] Hohenberg, J., The Professional Journalist, Henry Holt and Company Inc., New York, 1960.
[23] 윤석홍, 김춘옥, 신문방송, 취재와 보도, 나남출판, 2000.
[24] Brooks, B. et al., The Missouri Group : News Reporting and Writing, St. Martin's Press, 1996.
[25] 조용철 외, 취재와 기사작성, 도서출판 양지, 1999.
[26] 국립국어연구원, 한국신문의 문체, 1997.
[27] 윤만근, Chomsky 생성문법의 변천, 경진문화사, 2001.
[28] Ohno, S. and Hamanishi, M., "New Synonym Dictionary," Kadokawa Shoten, Tokyo, 1981 (Written in Japanese)

윤 재 민
2001년 2월 부산대학교 정밀기계공학과 석사. 2003년 2월 포항공과대학교 정보통신학과 석사. 2003년 4월~현재 (주) 알리 대표이사. 관심분야는 Agent Technology



정 유 진
1998년 8월 포항공과대학교 컴퓨터공학과 학사. 2000년 2월 포항공과대학교 컴퓨터공학과 석사. 2000년 3월~현재 포항공과대학교 컴퓨터공학과 박사과정. 관심분야는 기계 번역, 의미 중의성 해소

이 종 혁
1980년 2월 서울대학교 수학교육과 학사 1982년 2월 한국과학기술원 전자계산학과 석사. 1988년 8월 한국과학기술원 전자계산학과 박사. 1989년 11월~1991년 1월 일본 NEC C&C 정보연구소 초청연구원. 1998년 8월~1999년 7월 미국 New Mexico State University/CRL 방문연구원. 1991년 2월~1996년 9월 포항공과대학교 컴퓨터공학과 조교수. 1996년 10월~2003년 2월 포항공과대학교 컴퓨터공학과 부교수. 2003년 3월~현재 포항공과대학교 컴퓨터공학과 정교수. 관심분야는 자연언어처리, 한국어처리, 기계번역, 정보검색