육하원칙 활성화도를 이용한 신문기사 자동요약

윤재민[0] 강인수 권오욱 배재학 이종혁
포항공대 정보통신대학원[0] 포항공대 전자 및 컴퓨터공학부
포항공과대학교 전자컴퓨터공학부 컴퓨터공학과, 첨단정보기술 연구센터
{chiwoo[0], dbaisk, ohwoog, jhjbae, jhlee}@postech.ac.kr

An automatic extraction of newspaper articles using activation degree of 5W1H

Jae-Min Yoon[0] In-Su Kang Oh-Woog Kwon Jae-Hak Bae Jong-Hyeok Lee
Dept. of Graduate School for Information Technology, POSTECH[0]
Dept. of Computer Science and Engineering, Division of Electrical and Computer Engineering,
POSTECH, and Advanced Information Technology Research Center(AITrc)

요 약

본 논문은 신문기사에서 중요한 문장을 추출(Extract)하는데 있어서, 기존에 가장 우수한 방법인 전문기반 방법(Lead-based method)과 제목을 이용한 유사도 측정방법(Title-based method)의 문제점을 해결하기 위해서, 육하원칙 활성화도를 이용하여 신문기사를 효과적으로 요약할 수 있는 방법과 알고리즘을 제안하였다. 본 연구에서는 먼저, 제목(Title)과 전문(Lead)에서 중복출현하지 않는 육하원칙 구성성분을 결합하고, 본문의 각 문장에서 육하원칙 구성성분의 재사용성과 육하원칙 구성성분의 범주 중감을 파악하여 육하원칙 활성화도를 구하고, 전문기반 방법을 응용하여 각 문장의 상대적인 중요도에 따라 최종적인 가중치를 부여함으로써, 신문기사에서 중요한 문장을 효과적으로 추출할 수 있는 가중치 계산식을 제안하였다. 실험문서는 조선일보 웹사이트에서 제공하는 신문기사 100건을 대상으로 하였으며, 요약율이 30%일 경우 제안한 방법의 정확률은 74.7%로 기존의 전문기반(Lead-based method)방법보다 6.7% 향상되었다.

1. 서론

문서요약(Text Summarization)이란 대상문서에서 가장 중요한 정보를 증류해내는 과정[1]으로 크게 요약(Abstract)방법과 추출(Extract)방법으로 나눌 수 있다.

요약(Abstract)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(Abstract)방법에 비해 쉬운 접근 방법이며 통계적인 분석이나 확률적인 분석으로 구현되고 있다. 최근에는 추출(Extract)방법의 단점인 문장 가독률 저하를 방지하기 위해, 중요문장을 추출 후, 너무 긴 문장은 분리하고, 단문은 결합하면서, 추출된 문장들 사이에서 대용어 문제를 다루는 교정(Revision)방법이 연구되고 있다[1].

본 연구에서는 모바일(Mobile)이나 PDA(Personal Digital Assistants), 인터넷(Internet) 등을 통해 신문기사를 읽을 때, 공간과 시간적 제약으로 인해 원문을 다 볼 수 없는 경우의 문제점을 해결하기 위해, 신문기사에서 중요한 문장을 요약하기 위한 자동 추출(Extract)방법과 알고리즘을 제안한다. 2장에는 기존의 연구와 문제점을 살펴보고, 3장에서는 신문기사의 구조를 분석하고, 4장에서는 육하원칙 추출과 결합방법에 대해서 설명한다. 5장에서는 육하원칙 활성화도 평가와 가중치 조정문제에 대해서 다루며, 6장에서는 실험, 7장에서는 결론을 다룬다.

2. 기존 연구 및 문제점

일반적으로 신문기사는 여러 가지 요약방법을 테스트하기 위한 테스트용 코퍼스로 많이 사용되기 때문에 그 동안 다양한 요약 방법들이 적용되었다. 전문기반방법(Lead-based method), 제목을 이용한 유사도 측정방법(Title-based method), 문장간의 담화구조(Discourse Tree)를 이용한 방법이나 어휘사슬(Lexical Chain), 단어 공기정보(Co-Occurrence)를 이용한 방법, 또는 여러 가지 통계적인 방법을 통합한 방법 등이 그 예가 될 수 있다.

그러나, 이전의 신문기사를 요약하는 방법[2]들 중에, 전문기반방법과 제목을 이용한 유사도 측정방법이 가장 효율적이라고 알려져 있는데, 특히, Brandow[3]는 자신이 여러가지 통계적인 방법들을 이용해서 개발한 Hybrid 시스템보다 전문기반방법이 더 좋은 방법임을 강조하였다. 여기서 전문기반방법이란, 신문기사의 첫 문장부터 일정한 요약율에 의해 차례대로 선택하는 방법을 말한다. 그러나, 전문기반 방법의 문제점은 신문기사의 처음이나 둘째 부분, 또는 그 다음 부분이 중요하지 않은 문장이면서 상대적으로 긴 문장인 인용문이나 부가문 등이 포함될 경우, 요약률 저하를 유발할 수 있다. 또한, 문장길이가 너무 짧은 문장은 요약문에 포함되지 않는 경향이 있으나[7][8], 전문기반 방법은 중요한 내용을 내포했다고 볼 수 없는 짧은 문장이 주요문장으로 선택될 수도 있다. 여기서, 신문기사에서 인용문은 일반적으로 덜 중요한 문장이라고 알려져 있다[14].

제목을 이용한 유사도 측정방법의 문제점은 제목이 본문의 내용을 잘 내포하지 못하고 은유적으로 사용될 경우와 유사도를 측정하기 위해 필요한 어휘정보가 적을 경우 발생한다. 그리고, 이 방법은 문장길이가 비교적 길고, 중복된 단어가 많이 등장하는 문장을 중요 문장으로 선택하게 되는 문제점을 안고 있다.

신문기사 요약에 있어서, 최근에 신문기사의 작성 원칙인 육하원칙을 이용한 연구가 진행되었는데, [10]에서는 실험 대상 신문기사의 원문에서 문장의 유사도 관계를 이용하여 중심절을 추출하고, 추출된 중심절과의 유사도가 높은 절에서 휴리스틱 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, 휴리스틱 정보와 수사어구를 이용해서 추출된 절이 중심절의 사건, 또는 제목과 어떠한 연결고리가 있는지 의미적으로 파악하지 못했다. [4]에서는 [10]에서와는 달리 실험대상 신문기사의 원문에서 수사어구를 이용하여 육하원칙에 해당하는 정보를 추출

하지 않고, 신문기사 제목(Title)에서 육하원칙에 해당하는 정보를 추출하여, 시소러스를 이용, 각 육하원칙 구성단어를 만족시키는 상위레벨의 단어로 여러 개의 신문기사 제목을 분류(Classification)하여 요약하는 방법을 제안하였다. 그러나, 이 방법은 신문기사의 원문을 다루지 않기 때문에 제목이 원문의 내용을 충실하게 내포하지 못하는 경우는 잘못된 기사를 찾아줄 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 신문기사에 대한 구조적인 특징을 파악하거나 요약대상이 되는 문서의 특성을 고려하지 않고[2][3], 요약방법을 일률적으로 적용한 결과, 요약율이 상당히 저하되는 것을 알 수 있다[9]. 그리고, 신문기사의 구조를 분석할 때, 육하원칙과, 제목, 전문과의 관계를 심도있게 분석하지 않고, 육하원칙만 강조한 결과 전혀 의도되지 않은 결과을 생성해 낼 수 있으며[10], 육하원칙을 제목에만 적용함으로써, 피상적으로 분석하여 그 자체 내에 한계를 가지고 있다[4].

그러나, 본 논문은 전문기반 방법을 보완하기 위해, 전문과 제목에서 육하원칙 관계를 파악하고, 제목과 전문에서 강조되고 있는 육하원칙의 구성성분이 본문에서 어떻게 재사용되고 있는지 분석하고, 육하원칙 구성성분의 활성화 정도를 측정하여, 본문에서 중요 문장을 추출하기 위한 방법과 알고리즘을 제안한다.

3. 신문기사 구조 분석

3.1 육하원칙과 신문기사의 이해

신문기사의 목적은 '누가', '언제', '어디서', '무엇을', '어떻게', '왜' 에 해당하는 구체적 사실을 독자와 시청자에게 전달하는 것이기 때문에[13], 좋은 신문기사는 정확성(Accuracy), 객관성(Objectivity), 공정성(Fairness) 등이 뒷받침되어야 한다[11]. 따라서, 복합문 보다는 단문위주로 문장을 간결하게 기술하고, 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 없는 장점이 있다[5][12]. 또, 피동태와 사동태는 거의 이용하지 않고, 능동태 문장을 주로 사용하며, 동작성이나 상태성을 지니고 있는 비실체성 보통명사가 목적어로 쓰였을 경우인 '농성을 하다', '수사를 벌이다' 등은 다음과 같이 '하다' 와 결합하여 '농성하다', '수사하다' 로 사용하는 것을 원칙으로 한다[12][13][14].

또한, 수사어구(Rhetorical marker)와 단서단어나 구(Cue Words or Phrase)가 거의 없기 때문에 담화트리(Discourse Tree)를 이용하거나 수사구조 파싱(Rhetorical Parsing)을 신문기사요약에 이용하기에는 부적절한 측면이 있다.

신문기사에서 전문이란, 일반적으로 본문에서 가장 상위에 위치하는 하나의 문장으로 구성되어 있는데, 상기 전문의 기술원칙은 육하원칙에 의하고, 본문에는 전문에 쓰인 육하원칙의 각 성분들이 본문에서 다시 재사용되면서 전문에서 주장하는 내용을 뒷받침하게 된다. 따라서, 본문에서 중요한 문장은 전문을 구성하는 육하원칙의 구성성분이 다시 재사용되면서 전문에서 주장하고 있는 내용을 보완하거나 전문에서 빠진 육하원칙 성분에 대해서 설명하는 문장이라고 할 수 있다.

3.2 전문과 역피라미드 구조

대부분의 신문기사구조는 중요도 순서에 따라, 제목(Title) -> 전문(Lead) -> 본문(Body)의 역 피라미드 구조를 가지는데 전문은 육하원칙에 의해 기술하고 본문에는 전문에서 강조한 것에 대해서 구체적으로 설명하거나 추가적인 정보를 기술한다[5][11][12][13]. 그리고, 전문은 신문기사에서 가장 중요한 핵심부분으로서, 기본 문형은 '주어+목적어+타동사' 의 구성을 원칙으로 한다[14][16].

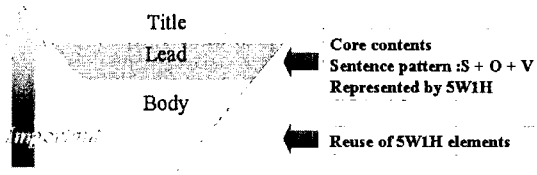


그림 1, 신문기사의 형태적 구성요소[11]

[그림 1]은 역피라미드(Inverted pyramid) 구조의 특징을 보여주는데, 이렇게 신문기사가 역피라미드 구조를 따라야 하는 이유는 현장에서 기자가 작성한 신문기사를 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 없고[5], 모든 사람들이 기사를 처음부터 끝까지 읽지 않고, 기사를 읽는 사람들 가운데 25% 정도는 중간 정도까지 읽고 중단하기 때문이다[6].

3.3 제목과 전문

신문기사의 전문은 역피라미드 구성의 첫 머리에 해당하는 만큼 기사의 가장 주요한 내용이 압축적으로 포함된다. 전문 한 줄만 써도 전체 기사의 내용을 짐작할 수 있을 정도가 돼야 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[13]. 따라서 육하원칙의 기사 구성요소가 기사 속에서 어떻게 제시되는가는 [그림 2]를 보면 알 수 있다.

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[13]

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

3.4 실험문서와 전문의 구성비율

일반적으로 기사는 그 내용과 성격에 따라 뉴스성 기사를 경성기사(Hard news), 해설성 기사를 연성기사(Soft news)로 분류하는데 연성기사는 논설이나 칼럼이 해당하고 경성기사는 국제, 정치, 경제, 사회, 문화 관련 기사로 나누어진다[11]. 본 논문

에서는 2002년 2월1일~30일 사이의 조선일보의 경성기사(뉴스성 기사) 100개를 무작위 추출해서 전문의 유무를 파악하기 위한 실험을 하였는데, 전문이 출현한 비율이 96%에 달했다.

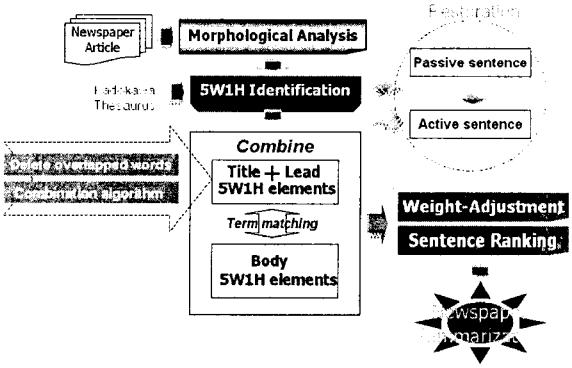
경관이 술집, 여관 불법경영[제목] - Who 강조
공무원법상 영리사업을 할 수 없는 *경찰관 중 상당수가* 경찰 단속 대상인 유흥업소, 숙박업소를 운영해 온 것으로 밝혀졌다. [전문]

대우, 산동에 대규모 시멘트 공장 설립[제목] - Where 강조
대우 그룹이 *중국 산동지역에* 시멘트 생산법인을 설립, 대규모 시멘트 공장을 세우기로 하고 8일 북경에서 기념행사를 가졌다. [전문]

그림 3, 전문과 본문[10]

4. 육하원칙 추출과 결합

본 연구에서 제안한 전체적인 시스템 구성은 [그림 4]과 같다.



[그림 4] 전체 시스템 구성도

신문기사가 입력되면, 형태소 분석 후, 가도카와 시소러스를 이용하여 제목과 전문, 본문에서 육하원칙 구성요소를 추출한다. 이때, 육하원칙 구성성분 중 'Who' 와 'What' 을 추출하는데 있어서 문제점을 해결하기 위해, 피동형 문장은 능동형 문장으로 복원한 뒤에 육하원칙 구성성분을 추출하게 된다. 그리고, 이렇게 전체의 문서에서 각 문장단위로 육하원칙 구성요소가 추출된 신문기사는 제목과 전문에서 사용된 육하원칙의 각 성분을 결합하고, 중복된 단어는 제거한다.

여기서, 제목과 전문을 결합한 육하원칙 구성성분과 본문의 각 문장에서의 육하원칙 구성성분을 비교하여, 육하원칙 구성성분의 활성화도를 파악한 후, 전문기반 방법을 응용하여 가중치를 조정하고, 가중치 합이 큰 문장을 중요문장으로 선택한다.

4.1 육하원칙 추출

먼저, 육하원칙이 무엇인가를 알아야 함으로, 국립국어연구원[15] 에서 정의하고 있는 육하원칙을 다음 [표 1]에 간단하게 나타내었다.

신문기사에서 육하원칙의 각 구성성분을 추출하기 위해 [표 1]에서 정의하고 있는 패턴정보를 [표 2]에 간단히 정리하였다. 여기서, 기존의 육하원칙 구성성분에 할당하기 어려운 정보들은 SE(Supplementary Element)에 할당하여 문장성분의 다양성을 확보함으로써, 중요한 문장의 우선순위를 세밀하게 계산할 수 있게 하였다.

그리고, 육하원칙 구성성분 중, 'How' 에 해당하는 패턴정보인 연결어미 '-아/-어/-며' 뿐만 아니라, 서술어까지 확장하여 'How' 에 할당[13]하였다

여기서, 'Who' , 'When' , 'Where' , 'What' 성분에서 육하원칙 구성성분을 추출하기 위해서 이용하는 유정명사(사람, 동물), 국적, 소속단체, 장소명사 등의 정보는 가도카와 시소러스를 이용하였다.

Who - 사람, 동물, 국적, 소속단체
► *연극계 최고의 미모 김종아는* 첫 마디부터 엄살연기를 펼친다.
When - 시간
► *지난달 24일*, 신한국당의 대표실에는 ~ ~
Where - 지명, 기관
► *수원의 신시가지 영통지구에* 상가개발이 활발하다.
What - 술어에 대한 목적어, 무정명사 주어
► *의료수가 인상*은 국민 의료비 부담으로 직결된다.
How - 연결어미 '-아/-어/-며'
► 민선으로 선출된 수협회장이 *자진* 사퇴했다.
Why - '기 위하여, -을 위해'
► *닭 값 안정을 위해* 올해 처음으로 닭고기를 수입한다.

표 1, 육하원칙 구성성분 정의

5W1H category		Example
Extended 5W1H	Who	사람,동물,국적,소속단체 + 주격조사(이/가)/보조사(은/는)
	When	숫자+년/월/일/시/분/초, 작년/올해/오전/오후/낮/밤 등
	Where	장소명사(구) + 에,에서 등
	What	(1)을 제외한 명사(구) + 주격조사(이/가)/보조사(은/는), 명사(구) : 목적격조사(을/를)
	How	용언(동사)
	Why	~위해서, 때문에, 이유로 등
	SE	명사(구)+보조사,공동격조사 등 예) *방송정책의 혼선에* 책임을 지고,

[표 2] 확장된 육하원칙 구성성분 분석

신문기사에서 중요하지 않은 성분들은 육하원칙 구성성분 추출시 제거하게 되는데, 대명사나 부사류, 관사류, 지시형용사 등과, 신문기사에서 자주 등장하는 상투어구인 '말했다, 발표했다, 밝혀졌다, 제안했다' 등이 해당한다.

본 연구의 핵심적인 아이디어는, 신문기사 작성의 기본 원칙인 하나의 문장에는 하나의 아이디어를 적는다는 원칙을 적용하여 제목과 전문에서만 각각 육하원칙 구성성분을 추출하지 않고, 본문의 각 문장에 대해서도 개별적으로 각각의 육하원칙 구성성분을 추출하는 것이다. 일반적으로 전문은 육하원칙 구성성분이 적절하게 배합되어 있으나, 전문에 모든 육하원칙 구성성분을 기술하는 것은 신문기사 작성에 있어서 금기시 되고 있다[13]. 따라서, 전문에서 기술하지 못한 육하원칙 구성성분은 본문에서 적절히 추가되는데, 이때, 전문에서 빠진 육하원칙 구성성분을 추출하기 위해서는 전문에서 출현하였던 육하원칙 구성요소를 일부 다시 재사용하면서, 전문과 비교하여 육하원칙 구성성분의 범주(Category)가 증가한 문장 위주로 추출해야 한다고 생각한다. 이때, 신문기사의 본문에서 각 문단(Paragraph)은 하나의 문장(Sentence)으로 구성되어 있기 때문에, 중요한 문장은 문장길이가 너무 길거나 짧은 문장은 배제되어야 한다.

사실, 신문기사는 사람이 직접적으로 읽어 보지 않으면, 그 신문기사가 주장하는 육하원칙 구성성분을 기계적으로 추출하기가 힘들기 때문에, 본 논문에서 제안한 방법이 신문기사에서 육하원칙 구성성분을 추출하는데 있어서 하나의 대안이 될 수 있다고 생각한다.

4.2 육하원칙 추출의 문제점

신문기사의 각 문장에서 육하원칙 구성요소를 추출하는데 있어서의 문제점은 육하원칙 구성성분의 하나인 'Who' 의 소속으로 언급하지 않은 명사가 의인화되어 주어로 쓰인 경우에 'Who' 와 'What' 성분을 구분하는 문제, 이중주어나 이중목적어가 쓰인 문장에서 'Who' 와 'What' 성분을 구분하는 문제, 피동문에서 'Who' 와 'What' 성분을 구분하는 문제이다.

4.2.1 의인화된 무정명사

다음 예문을 보면, '저기압이 비를 떨어뜨렸다' , '정부군 비행기는 아쿠엠 마을에 폭탄을 투하했다' 에서 주어로 쓰인 '저기압이' , '비행기는' 은 유정물은 아니지만 무정명사가 의인화된 것으로서 원칙적으로는 'Who' 성분에 할당해야 한다. 목적어를 수반한 타동사 구문의 이러한 문장표현은, 실험문서인 100개의 신문기사에서 출현한 968개의 문장 중, 약 15개의 문장에서 발생하였는데, 이것은 전체에서 약 2%에 해당하는 것이기 때문에 신문기사 요약에 미치는 영향이 작을 것으로 예상하고 본 연구에서는 무시하였다.

4.2.2 이중주어와 이중목적어

이중주어나 이중목적어가 쓰인 문장에서 'Who' 와 'What' 성분을 구분하는 문제에서, 이중주어는 전체 968개 문장에서 3개, 이중목적어는 전혀 나타나지 않으므로 이것도 역시 무시하였다.

구분	유형별수	비율
총 문장개수	968	100%
이중주어	3	0.3%
이중목적어	0	0

표 3, 이중주어와 이중목적어 출현정보

4.2.3 피동문

피동문에서 육하원칙 'Who' 와 'What' 성분을 제대로 분석하기 위해서는 피동문이 능동문에서 어떻게 쓰이는가가 중요하다. 따라서, 피동문을 능동문으로 복원(*Restoration*)한 뒤에 능동문에서 육하원칙 구성성분을 추출하였다.

신문기사에서는 일반적으로 피동문보다 능동문을 사용할 것을 적극 권장하는데, [그림 5]에서 보는 바와 같이, 전체 968개 문장에서 나타난 4678개의 동사 가운데, 피동동사는 약 3.5%인 158개에 달했다. 이러한 피동동사를 능동태로 복원할 수 있는지 복원가능성을 테스트한 것이 [표 4]이다.

일반적으로, 피동동사는 150개의 타동사에 피동접미사 '이/히/리/기' 를 삽입하여 규칙적으로 피동문을 이루는 피동형과 용언에 '-어 지다, 되다, 받다, 당하다' 등을 첨가하여 이루어지는 피동표현으로 나눌 수 있는데, [표 4]에서 보는 바와 같이 피동문을 능동문으로 복원할 때에 약 5%정도 오류가 발생하였다.

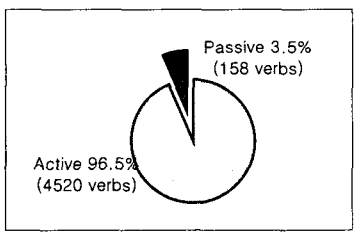


그림 5, 피동동사와 능동동사

피동문을 능동문으로 복원하는 과정에서 오류가 난 것은 '하객이 몰리다, 눈이 싸이다' 등의 문장을 '하객을 몰다, 눈을 쌓다' 로 복원하면 어색한데, 여기서 '몰리다, 쌓이다' 등은 각각 몰다, 쌓다의 피동형이 아니고 문장에서 자동사로 쓰인 경우에 발생하게 된다. 즉, 자동사와 그 피동형이 동일한 단어일 때, 문제가 발생하는데, 이런 동사 유형도 전체 동사에서 0.2%정도 밖에 나타나지 않기 때문에 무시하였다.

피동구분	피동유형	유형별수 (비율)	복원가능수	복원불가능수
피동형	이히리기	58(37%)	51	7
피동표현	-어 지다	75(47%)	75	
	-되다	5(3%)	5	
	-받다	12(7.6%)	12	
	-당하다	8(5%)	8	1
총 개수(비율)		158(100%)	151(95%)	8(5%)

표 4, 복원가능성 테스트

4.4 제목과 전문의 육하원칙 구성성분 결합

다음 [표 5]는 제목과 전문에 나타난 육하원칙 구성성분의 중복을 제거하면서 결합하는 알고리즘에 관한 것이다.

```
Combination algorithm
Begin
    Compared with Lead, delete overlapped words in Title; (Rule1)
  Repeat
     Pop one non-overlapped word in the Title
    If (the pattern of the non-overlapped word are "nation name, ~" ,
       "organization name, ~" or "+animate noun phrase,~ "
       then (Insert non-overlapped word at Who element in Lead); (Rule2)
    Else if (the pattern of the non-overlapped word is "-animate noun
phrase,~ "
       then (Insert non-overlapped word at What element in Lead); (Rule3)
    If (the extraction of 5W1H element in Title succeed) then (Rule4)
       (Insert non-overlapped word at the same 5W1H element in Lead);
    Else if (the pattern of the non-overlapped word are "-animate noun phrase
+ verb"
       then (Insert non-overlapped word at What element in Lead); (Rule5)
    Else if (the pattern of the non-overlapped word are "(-animate noun phrase)
+ CMCP"
    then {(Insert (-animate noun phrase) at What element in Lead),
       (Insert CMCP at How element in Lead)}; (Rule6)
    Else if (the non-overlapped word(T1) belongs to an NP in the body) then
      If (the other word(T2) in the NP exists at an 5W1H element in Lead) then
       (Insert the non-overlapped word(T1) at the other word(T2)
       belongs to an 5W1H element in Lead); (Rule7)
    Else
      break repeat;
  Until non-overlapped word in the Title is empty;
End
```

표 5, 제목과 전문 결합 알고리즘

제목을 이용한 유사도 측정방법은 제목을 구성하는 단어정보가 작기 때문에, 유사도를 비교하는데 있어서 편향된 결과를 제시할 수 있고, 중복된 단어가 많이 등장하는 긴 문장을 중요문장으로 선택하게 되는 문제점을 안고 있기 때문에, 본 연구에서는 단어정보 부족문제를 제목과 전문에 나타난 단어를 결합시킴으로써 해결하고, 중복된 단어가 많이 등장하는 문장을 걸러내기 위해서 단순한 단어매칭에 의해서 유사도를 구한다.

또한, 제목과 전문에 나타난 육하원칙 구성성분을 서로 결합하여 단일화 시킴으로써 서로 부족한 육하원칙 구성성분을 보완할 수 있기 때문에, 전문에서만 육하원칙을 추출할 때의 문제점도 보완할 수 있고, 제목에서 발생한 어휘정보만을 이용하는 제목을 이용한 유사도 측정방법에 비해 제목과 전문을 결합시킴으로써 충분한 어휘정보를 얻을 수 있는 장점이 있다. 그리고, 이렇게 결합시킨 육하원칙 구성성분이 다시 본문에서 어

떻게 재사용되는지, 얼마나 활성화되었는지에 따라 다르게 가중치를 부여하게 되므로, 단순한 단어기반이 아니라, 의미기반의 문서요약방법에 해당한다고 할 수 있다.

본 연구에서 제안하는 결합방법은 다음과 같은 규칙들로 구성되어 있다. 먼저, [규칙1]은 제목을 전문과 비교해서, 중복되는 단어는 제거하는 것으로, 100개의 실험문서 중 모두 82개의 문서제목에 발생하였고, 이 중에서 23개의 문서제목에 속한 단어가 모두 규칙1에 의해서 걸러져서, 다른 규칙 필요 없이 전문과 완전히 결합하였다.

[규칙2]와 [규칙3]은 제목에 남아 있는 단어의 패턴이 '국가명, ~', '조직명, ~', '유정명사, ~' 등이면, 국가명, 조직명, 유정명사에 해당하는 단어를 전문의 'Who' 성분에 삽입하고, 제목에 남아 있는 단어의 패턴이 '무정명사, ~' 이면, 무정명사에 해당하는 단어를 전문의 'What' 성분에 삽입하는 것이다.

위와 같은 [규칙2]와 [규칙3]에 해당하는 패턴이 100개의 실험문서에서 모두 14개 발생하였고, 이중에서 오류는 한 개 발생하였다. 정확한 유형은 '미국, 중동 평화안 환영' 에서처럼, 미국은 'Who' 성분에 할당될 수 있으나, 오류유형은 '프랑스, 의료계 총파업 비상' 에서처럼, '프랑스의 의료계' 인지 '프랑스에서' 인지 불분명할 때 발생하게 된다. 이런 오류유형은 프랑스와 의료계의 공기정보가 본문문장의 NP에 동시에 나타나는지를 조사하면 가능한 것으로 [규칙8]로 해결 가능하다.

[규칙4]는 제목에서 남아있는 단어들에 대해, 육하원칙 구성성분의 추출이 성공하면, 그 남아 있는 단어를, 제목의 육하원칙 구성성분 위치와 동일한 전문의 육하원칙 구성성분에 삽입하는 것으로, 100개의 실험문서 중, 26개의 실험문서에서 발생하였으며, 이 중에서 한 개가 다른 규칙 필요 없이 전문과 완전히 결합하였다.

[규칙5]은 제목에 남아 있는 단어의 패턴이 무정명사+동사이면, 그 남아 있는 단어를 전문의 What요소에 삽입하는 것으로, 전체 100개의 신문기사 중 2개의 문서제목에서 발생하였다.

[규칙6]은 제목에 남아 있는 단어의 패턴이 (무정명사)+서술성 보통명사이면, 이때 무정명사는 'What' 에 삽입하고, 서술성 보통명사는 육하원칙 구성성분 중 'How' 요소에 삽입하는 것으로, 전체 100개의 문서 중 32개 문서에서 나타났으며, 이 중 3개의 문서에서 오류가 발생하였으며, 한 개의 문서만 다른 규칙 필요 없이 전문과 완전히 결합하였다. 오류유형은 '큰 폭 증가' 에서 원래는 '큰 폭으로 증가' 이므로, '폭' 이 무정명사로 사용되었지만, 이때에는 'What' 의 구성요소에 삽입하면 안 된다는 것을 알 수 있다.

[규칙7]은 제일 마지막으로 실행되는 것으로서, 만약 중복되지 않은 단어(T1)가 본문의 어떤 NP에 속하고, 그 NP의 다른 한쪽 단어(T2)가 전문의 어떤 육하원칙 구성성분에 속해 있으면, 그 중복되지 않은 단어(T1)을 다른 한쪽 단어(T2)가 속한 전문의 육하원칙 구성성분에 삽입하는 것을 말한다. 즉, 제목에서 중복되지 않고 남아 있는 단어인 '사멸' 은 '절반 사멸', '사멸 위기' 등이 본문에 재사용되었는지 검토하고 본문에 서로 공기한 NP가 있을 경우, 전문에서 일부 남아 있는 '위기' 에, 본문에서 찾은 '사멸 위기' 로부터 '사멸' 을 전문에 삽입하는 과정을 거친다. 이 과정은 이전의 규칙에 의해서 걸러주지 못한 단어에 대해 최종적으로 행하는 과정으로서, 19개의 신문기사 제목에서 7개만이 본문의 NP에서 공기한 단어를 찾을 정도로 오류(63%)가 많았다.

마지막으로 위에서 언급한 결합규칙에는 포함되지 않지만, 복합어문제 때문에 100건의 신문기사 중 10개의 문서제목에서 오류가 발생하였다.

즉, 제목에 쓰인 복합어 '토막살해, 대지진, 복제소' 등은 다시 본문에 재사용되면서 '토막시체, 대규모 지진, 복제된 첫소' 등으로 조금씩 단어의 형태가 서로 바뀌는데, 제목에 쓰인 단어가 정확하게 전문이나 본문에서 재사용되지 않을 경우, 제목에 쓰인 단어를 육하원칙 성분에 할당하는데 어려움이 따른다.

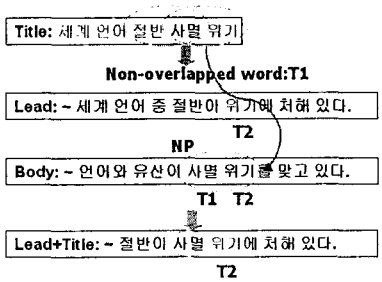


그림 6, 규칙8의 적용예제

최종적으로, 본 논문에서는 위와 같은 결합 알고리즘에 의해서 100건의 실험문서 중 25건이 하나의 알고리즘에 의해서 완벽하게 결합하였고, 2가지 이상의 알고리즘이 이용되어 완벽하게 결합한 실험문서는 총 33건이다. 따라서, 총 58건의 제목이 전문과 정확하게 결합한 것을 알 수 있다. 오류가 생긴 나머지 42건은 제목에 출현한 단어가 모두 결합하지 못한 오류가 발생한 것이 아니라, 각 문서마다 한 두개의 단어가 결합하지 못한 것을 말하는 것으로, 100건의 실험문서 중 제목에 출현한 총 단어 456 중, 67개의 단어가 제대로 결합하지 못하였으므로, 결합 정확율은 85.3%이다.

5. 육하원칙 활성화도 평가와 가중치 조정

본 연구의 목적은 신문기사의 구조적인 특징을 파악한 후, 신문기사를 요약하기에 가장 최선의 요약방법을 찾는 것으로써, 본 연구에서 제시하는 요약문장 구성은 가장 중요한 문장으로 전문을 디폴트로 선택하고, 요약문장을 구성하는 나머지 중요한 문장은 다음과 같은 3가지의 가중치 결합방법을 통하여 구한다.

첫째로, 제목과 전문을 결합한 육하원칙 구성성분과 본문의 각 문장에서의 육하원칙 구성성분을 비교하여, 육하원칙 구성성분의 재사용여부에 대한 Term matching관계를 파악한다. 여기서, 제목과 전문을 차별화하여 제목에 나타난 단어가 가장 중요하기 때문에, 제목에 나타난 어떤 범주에 속한 육하원칙 구성성분이 본문에서 같은 범주에 속할 경우, 특별히 강조하는 성분이므로 가중치를 높게 부여하고, 또한 제목과 전문을 결합할 때, 서로 중복되는 성분도 높은 가중치를 주었다.

둘째로, 제목과 전문을 결합한 육하원칙 구성성분의 범주개수와 본문의 각 문장에서의 육하원칙 구성성분의 범주개수의 증감을 구하여 육하원칙 범주의 활성화 정도를 계산하고, 이와 동시에, 문장 길이와의 상관관계도 파악한다.

셋째로, 전문기반 방법을 변형하여 위에서 구한 육하원칙 활성화도와 결합하여 효과적인 가중치 계산식을 유도하고, 최종적으로 가중치 합이 큰 문장을 중요문장으로 선택한다.

5.1 육하원칙 구성성분의 재사용성

먼저, 제목과 전문을 차별화하여 제목에 나타난 단어가 가장 중요하다고 보고, 제목의 어떤 범주에 소속된 단어가 본문에서 자신이 제목에 속한 육하원칙 범주와 똑같은 범주에 속할 때, 많은 가중치를 할당한다. 이 말은 역할이 바뀌는 것보다는 자

신의 역할을 그대로 유지하면서 본문에서 재사용된 육하원칙 구성성분을 포함한 문장이 중요하다는 것이다.

그 다음, 제목과 전문을 결합한 육하원칙 구성성분과 본문의 각 문장에서의 육하원칙 구성성분을 비교하여, 본문의 각 문장에서 다시 어느 정도 재사용되는가를 파악하고, 특히, 제목과 전문에서 중복된 단어는 특별히 강조된 부분이므로 이러한 단어에도 높은 가중치를 할당하였다. 위에서 설명한 내용은 다음 식으로 간단히 표현할 수 있다.

$$\omega_{5W1H_i} = \sum_{t \in Matched\ Terms} (\omega_t^1 \times \omega_t^2) \text{----------(1)}$$

여기서, i는 본문에 속한 어떤 임의의 문장을 말하고, t는 제목과 전문을 결합한 구성성분에 나타난 단어가 본문의 i번째 문장에서 다시 재사용된 단어의 개수를 나타내고, ω_t^1는 제목의 어떤 범주에 소속된 단어가 본문의 i번째 문장에서 자신이 제목에 속한 육하원칙 범주와 똑 같은 범주에 나타났을 경우에 대한 가중치, ω_t^2는 제목과 전문을 결합할 때, 중복해서 등장한 단어가 다시 본문의 i번째 문장에 나타났을 경우에 대한 가중치를 각각 나타내고, 제목과 전문을 결합한 성분에 속한 단어가 다시 본문의 문장에서 재사용되었을 때는 가장 낮은 가중치를 할당한다.

5.2 활성화된 육하원칙 범주

신문기사는 육하원칙에 의해 작성되는 것으로서, 제목과 전문에서 나타난 육하원칙 요소가 많이 재사용되면서 범주(Category)도 많이 포함하고 있는 문장이 그렇지 않은 문장보다 기사의 내용을 보다 충실하게 표현한다고 볼 수 있기 때문에 육하원칙 범주의 개수가 중요하고, 육하원칙 범주가 많은 문장을 중요한 문장이라고 볼 수 있다.

다음은 제목과 전문을 결합한 육하원칙 요소(5W1H elements)의 범주개수와 본문의 각 문장에서의 육하원칙 구성성분의 범주개수 증감을 구하여 육하원칙 범주 활성화정도를 계산하는 방법에 대해서 설명하고, 활성화된 대상 문장이 긴 문장인가 짧은 문장인가를 분석하여, 너무 길거나 짧을 경우 음의 가중치를 부여하였다.

제목과 전문을 결합한 육하원칙 구성성분의 범주는 Who, When, Where, What, How, Why, SE의 7가지 경우로 나누었다. 제목과 전문에서 강조한 육하원칙 구성성분 이외에, 다시 본문에서 강조하고 있는 육하원칙 구성성분을 파악하기 위해서는 본문의 각 문장에서 재사용되는 육하원칙 구성성분의 개수와 육하원칙 구성성분의 범주 개수가 몇 개인가가 중요하다.

또한, 긴 문장은 육하원칙 구성성분의 범주를 대부분 많이 내포하기 때문에, 육하원칙 요소가 많다고 중요한 문장이라고 할 수 없다. 즉, 육하원칙 범주의 활성화정도를 파악할 때는 해당 문장의 문장길이가 지나치게 길거나 짧은지도 고려해야만 한다.

신문기사의 평균문장길이를 계산하기 위해서 140만 어절의 신문기사 코퍼스를 이용해서 평균문장길이를 구하였다. 이 코퍼스는 2000년도 조선일보 웹사이트 신문기사를 포항공대 지식 및 언어공학연구실에서 자체 수집하여 구축하였다. 위의 코퍼스에서 구한 신문기사의 평균 길이는 문장 당 15.92 어절로 이 어절 길이를 중심으로 너무 길거나 짧은 문장은 Bell MF(Generalized bell MF)을 이용하여 음의 가중치를 부여하여, 너무 길거나 짧은 문장이 중요한 문장으로 선택될 가능성을 줄였다.

이상에서 설명한 내용은 다음과 같은 식으로 표현될 수 있다.

$$\omega_{Activated\_Category_i} = \frac{Bc_i}{Cc} + \alpha \cdot \left(\frac{1}{1 + \left| \frac{x_i - c}{a} \right|^{2b}} - 1 \right) \text{--- (2)}$$

여기서, Bc_i는 본문의 임의의 문장에서 육하원칙 구성성분의 범주개수, Cc는 제목과 전문의 결합성분에 나타난 육하원칙 구성성분의 범주개수를 나타내고, α는 조정인자, c는 MF(Membership function)의 중심을 표현하고, a는 MF의 폭, b는 MF의 교차점(Crossover point)에서의 기울기를 결정하는데 이용된다. 식(2)에서 x_i는 각 문장에서 육하원칙 구성성분의 개수를 나타내는데, 실험문서의 문장 평균길이에 해당하는 c값(15.92 어절) 주위로 폭 a만큼은 영향을 거의 미치지 않으므로, 가중치를 1로 고정시키고, 너무 긴 문장이거나 짧은 부분은 음의 가중치를 할당함으로써, 중요도 순위를 낮추게 된다. 본 연구에서는 a=15, b=3의 값을 사용하였다. 이상과 같이 제안한 방법에 의하면 문장길이가 적절하고 육하원칙 범주가 많이 활성화된 문장이 중요한 문장으로 선택되어진다.

5.3 육하원칙 활성화도

다음은 위에서 언급한 육하원칙 구성성분의 재사용성과 육하원칙 범주의 활성화 정도를 이용하여 육하원칙 활성화도는 다음 식(3)과 같다.

$$\omega_{Activation\_degree_i} = \omega_{5W1H_i} + \omega_{Activated\_Category_i} \text{------ (3)}$$

식(3)의 물리적인 의미는 육하원칙 구성성분의 재사용성(ω_{5W1H_i})이 높지만 문장의 길이가 지나치게 길면, 육하원칙 범주의 활성화 정도($\omega_{Activated\_Category_i}$)에 의해서 전체적인 가중치를 낮추어 주며, 육하원칙 구성성분의 재사용 정도가 같고, 적절한 문장길이를 가지는 경우, 육하원칙 범주가 더 많이 활성화된 문장을 중요한 문장으로 선택할 수 있게 해준다.

5.4 전문기반 방법 응용

전문기반 방법은 문서의 첫 문장들로만 요약문을 구성하는 방법이다. [표 6]에 보는 바와 같이, 인간이 추출한 중요한 문장과 비교하여, 요약율 10%일때, 즉, 10개의 문장으로 구성된 신문기사에서 중요한 문장을 하나 선택했을 경우, 첫번째 문장이 선택될 확률이 96%이고, 요약율 20%일 때, 두번째 문장이 선택될 확률은 54%, 요약율 30%인 경우, 세번째 문장이 중요문장으로 선택될 확률은 38%였다. 따라서, 본 논문의 실험문서에서는 첫번째 문장, 즉, 전문이 선택된 확률이 96%이므로 전문은 디폴트로 중요한 문장으로 선택한다. 그리고, 맨 처음 문장부터 마지막 문장까지 중요문장으로 선택될 확률이 차등적으로 분포한다는 것을 알 수 있으므로, 이전에 구한 육하원칙 활성화도에 전문기반 방법을 응용하여 문장가중치를 할당하면 최종적으로 식 4와 같은 가중치 조합을 얻을 수 있다.

요약율	선택문장	선택비율
10%	First sentence	96%
20%	Second sentence	54%
30%	Third sentence	38%

표 6, 전문기반 방법에서 중요문장 추출

$$\omega_{SENTENCE_i} = (\omega_{Activation\_degree_i}) \cdot \alpha_i \text{ ---------- (4)}$$

여기서, α_i 는 본문의 문장가중치를 나타내는데, 본문의 최상위에 위치한 첫번째 문장부터 차등적으로 가중치를 부여하여 계산하였다.

일반적으로, 신문기사의 중요한 문장은 제목과 전문에서 빠진 육하원칙 범주가 재출현하여 전문에서 주장하고 있는 내용을 뒷받침하거나, 전문에서 기술한 육하원칙 구성 성분를 부가적으로 잘 설명한 문장이라고 할 수 있다. 이상에서 설명한 바와 같이, 이러한 문장들은 육하원칙 구성성분의 재사용성과 육하원칙 범주의 활성화 정도를 이용하여 선택적으로 골라낼 수 있음이 분명하다.

이러한 방법론과 알고리즘에서는 제목을 이용한 유사도 측정방법에서 문제가 되는 문서정보 부족문제를 해결하기 위해서, 제목과 전문에서 출현한 육하원칙 구성성분을 서로 결합하고, 이때, 중복되어 등장하는 단어는 제거함으로써, 중요문장 결정시 단순히 단어빈도와 문장길이를 이용하는 것이 아니라 신문기사의 제목과 전문에서 강조하고 있는 육하원칙 구성 성분들이 어떻게 재사용되었는가, 그리고 적절한 문장길이를 가지면서 육하원칙 범주가 얼마나 활성화되었는가를 파악하는데 중점을 두었다.

6. 실험 결과

본 논문에서 사용한 실험데이터는 조선일보 웹사이트에서 제공하는 경성기사로 100건의 신문기사를 무작위로 선택하였다. 실험에 사용된 각 신문기사는 평균 약 9.7개의 문장으로 구성되어 있으며, 여기서 하나의 신문기사에서 중요한 문장순서대로 문장 3개까지 대학원생 3명이 각각 추출해서 비교한 뒤, 최선의 문장을 중요도에 따라 순서대로 3개까지 선택하였다.

다음 [그림 7]은 제목을 이용한 유사도 측정방법과 전문기반 방법, 그리고, 육하원칙 활성화도를 이용한 방법의 정확율을 비교한 그래프로써 본 논문에서 제안한 방법이 가장 효율적인 방법임을 알 수 있었다.

육하원칙 활성화도를 이용한 방법에서 제목과 전문을 결합하지 않고, 제목만 적용한 결과와 전문만 적용한 결과를 보여주고 있는데, 전문만으로 적용한 방법이 제목만으로 적용한 방법보다 더 우수함을 알 수 있었다. 이 결과는 제목을 이용한 유사도 측정방법과 비교해보면 서로 상반된 성능을 보여주고 있는데, 제목을 이용한 유사도 측정방법을 응용한 방법(QE)에서 제목과 전문을 합하여 유사도를 측정한 방법은 중복된 단어가 많이 등장함으로써, 성능을 저하시켰다고 생각한다. 여기서, QE는 Query Expansion의 약자이다.

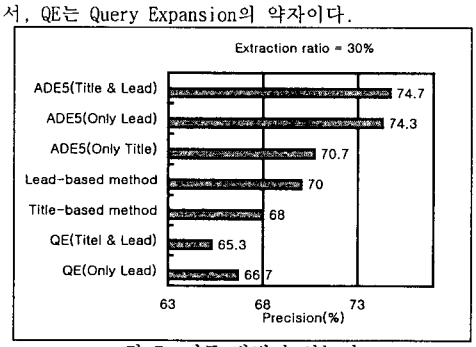


그림 7, 기존 방법과 성능비교

[그림 8]은 단어중복이 정확율에 미치는 영향을 평가한 것으로서, 제목과 전문에 출현한 단어를 결합하면서 중복된 단어를 제거하느냐, 그대로 사용하느냐에 따라 서로 다른 결과를 보여주고 있다. 제목과 전문에 출현한 단어를 결합하면서 중복을 허용하지 않은 방법이 중복을 허용한 방법보다 우수함을 알 수 있었다. 그리고 육하원칙의 활성화도를 사용하는 방법이 전문기반방법보다 우수함을 알 수 있었고, 또한, 육하원칙을 확장함으로써 역할을 다변화한 것이 정확율 성능개선에 효과적임을 알 수 있었다.

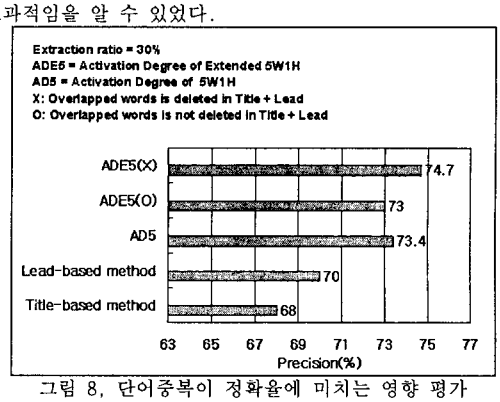


그림 8, 단어중복이 정확율에 미치는 영향 평가

다음 [그림 9]는 육하원칙 구성성분의 변화와 유지에 대한 중요도를 평가하였는데, 육하원칙 구성성분의 변화란, 제목과 전문을 결합한 구성성분에 쓰인 'Who' 의 구성요소가 본문에서는 'What' 의 구성요소로 바뀌는 것을 육하원칙 구성성분의 변화라 하고, 유지되는 것을 육하원칙 구성성분의 유지라고 한다.

아래의 그래프에서 'Transfer' 는 육하원칙 구성성분의 변화(역할변화)만으로 중요문서를 평가한 그래프이고 'Maintenance' 는 육하원칙 구성성분의 유지(역할유지)만으로 평가한 그래프이다. 원래의 정확율을 나타내는 'All' 에 비교해볼 때, 구성성분 유지의 중요성이 구성성분 변화보다 크다고 할 수 있다. 이 말은 제목과 전문의 육하원칙 구성성분이 본문에서 재사용되면서 계속 같은 성분에 위치하는 문장이 더 중요하다는 것을 암시하는 것으로써, 역할변화가 발생한 문장은 종속절이나 인용절 등의 덜 중요한 문장성분으로 이동하면서 중요도가 줄어드는 경향이 발견되었다.

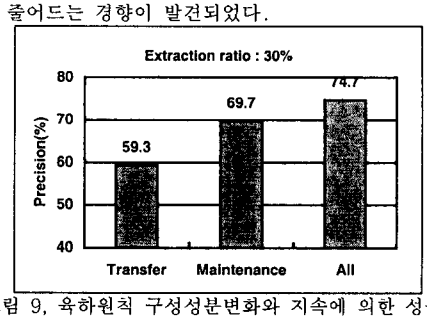


그림 9, 육하원칙 구성성분변화와 지속에 의한 성능변화

다음 [그림 10]은 육하원칙의 각 구성성분의 범주가 정확율 향상에 미치는 영향을 평가한 것으로써 막대그래프는 'Who' , 'Where' , 'What' , 등을 각각 제거한 뒤에 구한 정확율을 나타낸다. 대체로 'Who' , 'What' , 'How' 을 제거하

여 성능을 측정했을 때, 정확율이 가장 낮으므로, 이들이 미치는 영향이 가장 크다는 것을 알 수 있었다.

위에서 테스트한 결과를 최종적으로 정리하면, [표 7]과 같다. 여기서 보는 바와 같이, 본 논문에서 제안한 육하원칙 활성화도를 이용한 방법의 성능이 가장 좋고, 그 다음으로 전문기반방법, 제목을 이용한 유사도 측정방법의 성능 순서를 가진다. 또한, 이러한 방법들은 상용 소프트웨어인 MS Word 보다 월등히 높은 성능을 가진다는 것을 확인할 수 있었다.

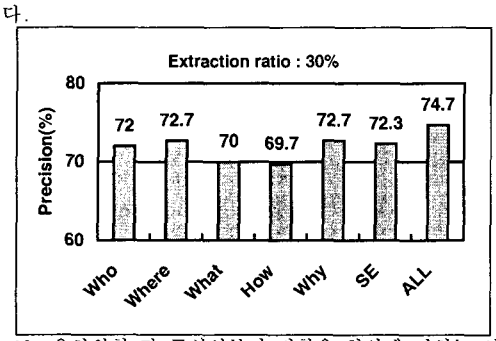


그림 10, 육하원칙 각 구성성분이 정확율 향상에 미치는 영향

Method	Precision(%)
	Extraction ratio : 30%
Activation degree of Extended 5W1H	74.7
Lead-based	70.0
Title	68.0
MS Words	43.7

[표 7] 최종 실험결과비교

본 논문에서 제안한 육하원칙 활성화도를 이용한 방법과 전문을 이용한 방법, 제목을 이용한 유사도 측정 방법은 중요한 문장에 따라 순서대로 3개까지 추출하여 인간이 추출한 요약결과와 비교하여 정확률을 계산하였다. 일반적으로 요약률의 성능평가의 척도는 정확률과 재현율, 그리고 F값을 비교하는데, 본 연구에서는 사람이 선택한 중요문장의 총 개수와 위에서 언급한 각각의 방법들에 의해 추출된 문장의 개수가 같기 때문에 정확률과 재현율, F값은 동일하다.

7. 결론 및 향후 연구

신문기사 요약에 있어서, 본 연구에서 제안한 육하원칙 활성화도를 이용한 방법은 기존의 전문기반 방법보다 더 우수한 방법임을 확인하였다.

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

첫째로, 제목과 전문을 결합한 육하원칙 구성성분과 본문의 각 문장에서의 육하원칙 구성성분을 비교하여, 육하원칙 구성성분의 재사용 정도를 파악함으로써, 제목과 전문에서 강조하고 있는 육하원칙 구성성분이 다시 본문에서 재사용되면서 제목과 전문에서 주장하고 있는 내용을 뒷받침하는 문장을 선택할 수 있었다.

둘째로, 제목과 전문을 결합한 육하원칙 구성성분의 범주개수와 본문의 각 문장에서의 육하원칙 구성성분의 범주개수를 구하여 제목과 전문에서 사용된 육하원칙 구성성분의 범주가 본문의 문장에서 다시 재사용되면서 상대적으로 얼마나 증가했는지 감소했는지를 파악함으로써, 육하원칙 범주의 활성화 정도를 계산할 수 있었고, 이와 동시에, 문장 길이와의 상관관계도 파악할 수 있었다.

셋째로, 육하원칙 구성성분의 재사용 정도와 육하원칙 범주의 활성화 정도를 더해서 육하원칙 활성화도를 구할 수 있었고, 이렇게 함으로써, 육하원칙 구성성분의 재사용성이 높고, 적절한 문장길이를 가지면서 육하원칙 범주가 많이 활성화된 문장을 중요한 문장으로 선택할 수 있었다.

넷째로, 신문기사요약에서 가장 효율적인 방법인 전문기반 방법을 변형하여, 육하원칙 활성화도에 문서의 상위에 위치한 문장부터 아래로 가중치를 차등적으로 적용하여 결합하는 방법을 제안함으로써, 신문기사에서 상위에 위치한 문장일수록 선택될 확률을 높였다.

향후, 신문기사의 각 문장에서 육하원칙 구성요소의 추출효율을 높이고, 특히, 제목과 전문을 결합할 때, 복합명사 문제를 해결할 수 있는 방법을 연구할 생각이다.

또한, 본 연구에서는 유사한 단어를 구별할 때, 가도카와 시소러스를 이용하여 단어확장을 시행한 결과 오히려 성능이 저하되었는데, 예를 들어, '남성' 과 '남자' 는 같은 의미범주에 속하지만, '남성' 과 '여성' , '시청' 과 '우체국' 등도 각각 같은 의미 범주에 속함으로써, 너무 어휘를 확장하여 전체 요약성능을 저하시키는 요인으로 작용하였다. 따라서, 시소러스를 이용하여 단어를 확장할 때, 유사한 단어를 구별할 수 있는 방법을 연구할 예정이다.

[감사의 글]

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

8. 참고문헌

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

[2] D. Marcu, "Improving Summarization through Rhetorical Parsing Tuning" , Proceeding of the COLING ACL Workshop on Very Large Copora, Montreal, Canada, 1998

[3] R. Brandow, K. Mitze, and L.F. Rau, "Automatical condensation of electronic publications by sentence selection" , Information Processing and Management, 31(5):675-685, 1995

[4] Akitoshi Okumura, Takahiro Ikeda, and Kazunori Muraki, "Text Summarization based on Information Extraction and Categorization Using 5W1H" , Journal of Natural Language Processing, 6(6):27-44, 1999

[5] J. Hohenberg, "The Professional Journalist" , Henry Holt and Company Inc., New York, 1960

[6] S. Brian, Brooks et al., "The Missouri Group : News Reporting and Writing" , pp.48, 1996

[7]Julian Kupiec, Jan Pedersen, and Francine Chen, A Trainable Document Summarizer, In Proceedings of ACM-SIGIR' 95, pp.68-73, 1995

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

[9] 김재훈, 김준홍, 도합유사도를 이용한 한국어 추출문서 요약, 제10회 한글 및 한국어 정보 처리, pp.238-244, 2000

[10] 이현주, 김계성, 구상옥, 이상조, 신문기사에서 육하원칙 중심의 정보추출, 한국정보과학회 봄 학술발표 논문집, pp.361-363, 2001

[11] 김지용, 현장 신문론, 도서출판 쟁기, 1996

[12] 윤석홍, 김춘옥, 신문방송, 취재와 보도, 나남출판, 2000

[13] 이행원, 취재보도의 실제, 나남출판, 1999

[14] 고혜련, 신문 취재와 기사작성, 중앙M&B, 2001

[15] 국립국어연구원, 한국신문의 문체, 1997

[16] 조용철외, 취재와 기사작성, 양지, 1999