신문기사에서 육하원칙 중심의 정보 추출*

이현주 김계성 구상옥 이상조
경북대학교 컴퓨터공학과
{hyunju, kskim, tomatoo}@sejong.knu.ac.kr sjlee@knu.ac.kr

Information Extraction from newspaper article by recognizing 5W1H elements.

Hyun-Ju Lee Kye-Sung Kim Sang-Ok Ku Sang-Jo Lee
Dept. of Computer Engineering, Kyungpook Nat'l University

요 약

본 논문은 신문 기사문에 특정적인 정보 추출의 내용과 방법을 제안한다. 신문 기사에서 이용자가 원하는 정보 추출의 내용으로 육하원칙을 중심으로 한 다섯 가지 정보를 제시하였으며, 이를 추출하기 위해 통계적인 기법을 주로 이용하고 부분적으로 언어적 지식을 이용하였다. 본 논문에서는 비교적 문서의 길이가 짧은 신문기사문을 요약 대상으로 하므로 단락이나 문장이 아닌 절 이하 단위로 추출하며, 중심절을 추출한 뒤 그 결과의 관계를 통해 나머지 정보들을 추출함으로써 추출되는 내용이 유사하거나 산만하지 않기 때문에 이 추출 정보로 요약문을 생성할 경우에 긴밀한 요약문을 생성할 수 있다.

1. 서론

인터넷의 발달로 우리는 정보의 홍수시대에 살고 있으며, 이렇게 많은 정보들 속에서 자신이 필요로 하는 정보들을 찾고 처리하는 것이 현대사회를 살아가는 개인에게 중요한 과제가 되었다. 그러나 제한된 시간에 정보를 습득하고 이용하기란 쉬운 일이 아니다. 따라서 정보 이용자에게 정보의 전체 내용을 압축된 상태로 보여 준다거나, 아니면 적어도 그 정보가 이용자에게 유용한 것인지에 대한 판단을 짧은 시간에 할 수 있도록 해 주는 자동 문서요약 기술이 현대사회의 중요한 과제로 부각되고 있다.

문서의 종류는 실용문에서부터 문학작품에 이르기까지 다양하며 이런 문서의 종류에 따라 글의 구성 형식과 문장 형태, 어휘의 종류 등이 다르게 나타나고, 이용자가 원하는 요약의 내용도 달라질 수 있으므로 모든 종류의 문서에 일반적인 요약 방법을 찾는 것은 매우 어려운 작업이다. 그러므로 이용자가 원하는 정보를 포함한 요약이 되기 위해서는 그 문서의 종류에 기반한 요약 방법을 찾는 것이 효과적이라고 본다.

본 논문은 대중적 실용산문인 신문기사를 대상으로 하여 요약을 위한 정보추출을 목적으로 한다. 추출하고자 하는 정보는 신문기사라는 장르에 특정적인 다섯 가지 정보로써 이 정보들의 내용은 육하원칙을 위주로 구성된 것이다. 추출방법은 추출하고자 하는 정보에 따라 달라지는데, 중심절을 추출할 때는 통계 기반 접근 방법의 하나인 유사도 관계를 절 단위에 적용시켜 추출하였으며, 다른 정보들은 중심절과의 유사도와 관계를 고려하여 추출하였다.

* 본 연구는 정보통신연구진흥원의 대학기초연구지원사업 과제 "Web 상에서 적확한 검색을 위한 문서의 대표 개념어 생성 및 요약 시스템"의 일부로 수행되었음.

2. 관련 연구

문서요약은 요약문을 생성하는 방법에 따라 추출(extract)과 요약(abstract)으로 나뉘는데, 요약을 하기 위해서는 가독성이 높은 자연스러운 문장을 생성하는 단계를 거쳐야 하므로 어려움이 따른다. 따라서 최근 국내의 연구들은 추출에 관한 것이 대부분이다.

이때 추출 단위는 단락 전체를 추출하는 경우[1]가 있고 문장을 추출하는 경우[2,3]도 있는데, 단락 전체를 추출하는 경우, 길이가 긴 문서에는 적절하나 신문기사와 같이 비교적 길이가 짧은 문서에는 불필요한 정보가 많이 포함되거나 정보의 손실이 커질 가능성이 높다. 그리고 웹 문서에서 개인이 올린 글이나 신문기사의 경우, 단락이 나뉘어져 있지 않거나 의미적으로 올바르게 나뉘어진 단락인지에 대해서도 확신하기 어렵다. 이 점을 고려하여 문장들을 클러스터링 하거나[4], 문단을 자동 구분한 후 대표 문장을 추출하여 요약으로 삼는 경우[3]도 있다. 그러나 이렇게 문장들을 추출하는 경우에는 문장들 간의 긴밀

성이 떨어져 가독성이 낮아진다는 단점이 있다. 또한 한국어는 어미가 잘 발달되어 있어 여러 개의 문장이 어미에 의해 연결되어 문장이 길어지는 경우가 많은데, 이때에는 실질적으로 단락 추출과 유사한 결과를 가지게 된다.

본 논문에서는 비교적 짧은 길이의 문서인 신문기사를 대상으로 하므로 짧은 요약 안에 정보의 손실률을 최소화하기 위해 단문이나 구를 추출단위로 삼았다.

추출기법으로는 크게 통계기반 방법[1,2]과 언어학적 접근법[5,6]으로 나누어진다. 통계기반 방법에서는 주로 단어의 출현 빈도를 측정하고 이를 반영하여 문장이나 문단의 중요도 값을 구한 뒤 그 값이 큰 문단이나 문장을 요약문으로 추출한다. 이 방법은 문장들 사이의 관계를 고려하지 못하기 때문에 일관성 있는 요약문을 생성하기 어렵다. 언어학적 접근법은 문서의 담화 구조를 파악하여 요약을 하는 방법으로, 문장의 관계를 객관적으로 설정하기 어렵고 문장 간의 관계를 요약을 위한 문서 전체의 구조로 확장하는 것도 쉽지 않다.

본 논문은 통계적인 방법을 이용하여 중심절을 추출한 뒤 이 절과의 관계를 토대로 언어적 지식을 이용하여 다른 정보들을 추출함으로써 이 정보들로 요약문을 생성할 경우 긴밀한 요약문이 생성될 수 있다.

3. 육하원칙 중심의 정보추출

3.1 신문기사에서의 추출 정보

본 논문은 요약 생성을 위한 정보추출을 목표로 하며, 추출 내용과 방법 면에서 신문기사라는 장르에 특정적이다.

신문의 기사를 구성하는 내용요소로는 육하원칙과 취재원이 있다. 육하원칙은 기사가 나타내고자 하는 사건에 대해 '누가, 무엇을, 언제, 어디서, 왜, 어떻게'의 형식으로 나타낸 것이다. 이외에 사건의 성격, 현황, 전망 등을 서술하는 부분인 해설이 덧붙여지기도 한다. 어떤 종류의 기사이든 이 여섯 가지 원칙을 중심으로 하여 쓰이지만 이 여섯 가지 원칙이 모두 나타나지는 않는다. 그리고 전문(*Lead*)에 이 육하원칙이 모두 나타나지도 않는다[7].

따라서 본 논문에서 추출하고자 하는 정보는 신문기사의 내용요소인 육하 원칙을 기본으로 한 다음의 다섯 가지 정보들로 규정하였다.

① 중심 사건(누가 무엇을 하다, 무엇이 어떻게 되다.)
② 사건의 시·공간적 배경(언제, 어디에서)
③ 사건의 세부 정황
④ 사건의 원인(왜)
⑤ 사건의 결과

3.2 추출 방법

신문기사에서 위의 다섯 가지 정보를 추출하기 위한 본 논문의 전체 시스템 구성은 다음 그림 1과 같다.

우선 태깅된 문서를 입력으로 하여 중심사건절을 추출하기 위한 전처리기를 거친다. 전처리기는 세 부분으로 구성되는데, 인용구 제거는 신문기사에 상투적으로 나타나지만 기사의 내용에 크게 기여하지 않는 보조용언과 인용구를 제거하는 것이다. 그

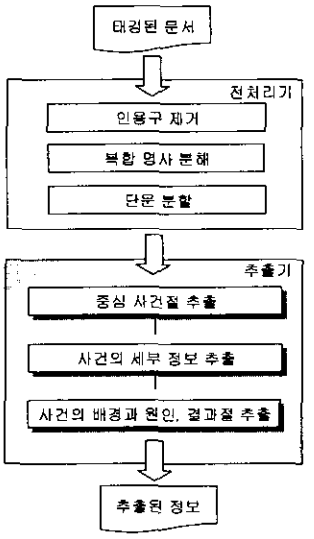

그림 1. 시스템 구성도

리고 복합 명사 분해는 중심사건절 추출을 위한 유사도 계산에서 보다 높은 정확도를 위한 것이다. 복합 명사 분해시 문서의 단일명사를 근거로만 분해하므로 완전한 복합명사 분해가 이루어지지는 않는다. 본 논문은 단문이나 구를 추출단위로 하기 때문에 단문 분할을 해야 하는데 이때 연결어미를 기준으로 접속문을 단문으로 분할하며 내포문 분리는 하지 않는다. 단문으로 분할하고 난 뒤 뒷 절의 생략된 주어는 앞 절의 주어를 보고 복원시킨다.

추출기에서는 위의 다섯 가지 정보를 추출하게 되는데, 먼저 중심절을 추출한 뒤 이 중심절과의 유사도나 관계로 다른 정보들을 추출하므로 중심절 추출의 정확도가 전체 추출된 정보의 정확도에 큰 영향을 미친다.

중심절 추출은 문서에서 유사 관계를 가지는 다른 단문들의 개수와 그 유사도를 이용하여 추출한다. 단문들 간의 유사도 계산에서는 코사인 유사도 계산을 이용하고, 단어 가중치 계산은 정보 검색에서 사용하는 역문헌 빈도 가중치[4]를 이용하여 계산한다.

사건의 세부 정보 추출 단계에서는 앞에서 추출된 중심 사건절을 뼈대로 하여 이 절과 유사도가 높은 문장들에서 사건절 각각의 성분에 수식어 등을 가져와서 보다 상세한 내용의 절로 만드는 것이다.

사건의 배경과 원인, 결과절은 중심절이나 중심절과의 유사도가 높은 절에서 휴리스틱 정보와 수사어구를 이용하여 추출한다. 각 정보를 추출하기 위한 휴리스틱 정보와 수사어구의 예를 보이면 다음 <표 1>과 같다.

이들 단서 정보에서 배경에 해당하는 정보들은 그 단서가 들어 있는 구를 추출하며, 사건의 원인과 결과 정보는 절 단위로 추출한다.

<표 1> 정보추출을 위한 단서 정보

추출정보	추출을 위한 단서 정보의 예
시간적 배경	올/지난, ... 작년/재작년/올해/오전/오후, ... 숫자 + 년/월/일/시/분/초/ + (께/쯤/중), ...
공간적 배경	지명, 공간명사 + -에서
사건의 원인	조사(-로, ...) 어미(-아/어서, -니까, -므로, -기에, ...), 구(-기 때문에, -ㄴ 바람에, -로 인해, ...)
사건의 결과	이로 인해, 결과적으로, 이에따라, -ㄴ 셈이다, ...

4. 정보추출 예

실험 대상 문서의 원문과 정보 추출 결과를 보이면 다음과 같다.

<원문>

중소형 우량주가 3월장세 이끈다...한국투신

<위경환>
우량 중소형주가 주도종목군으로 부상해 3월의 장세를 이끌 가능성이 높다는 분석이 나왔다.
한국투자신탁 (www.kitc.co.kr)은 3일 향후 장세는 수급악화와 대외 경제여건 악화등으로 대형주의 주가는 당분간 조정기간을 거칠 것으로 예상된다고 지적했다.
이에따라 주식시장은 박스권에서 등락을 거듭하는 장세가 전개될 가능성이 높다고 전망했다.
그러나 이같은 장세 속에서도 코스닥시장의 차별화가 완화되면서 거래소내에 중소형주 중 첨단 기술력을 갖춘 기업과 낙폭이 과대한 실적 우량주, 우량 유가증권을 보유한 신자산주등 우량 중소형주들이 시장의 주도주로 부상할 가능성이 높다고 분석했다.
거래소 시장과 코스닥 시장간의 차별화 현상은 정보의 비효율성 등으로 인한 것으로 장기적으로 시장의 효율성이 반영되어 차별화 현상이 해소될 것으로 전망했다.
시장간 차별화 해소는 코스닥 시장 벤처기업 주가가 하락하기보다는 거래소 시장의 첨단 기술 기업주가 상승하는 방향으로 진행될 것으로 내다봤다.
또 거래소시장 활성화 대책은 중소형 우량주에 유리한 환경을 재공해 거래소 우량 중소형주들의 주가 상승에 도움이 될 것이라고 덧붙였다.
한국투신은 이같은 판단에 따라 3월에 장세를 주도할 가능성이 높은 중소형 우량주를 선정 발표했다.
한국투신이 발표한 중소형 우량주는 한섬 삼양통상 삼양사 신세계백화점 중외제약 종근당 유한양행 성미전자 ＬＧ애드 우신산업 대한페인트 율촌화학 한국포리올 한국화인케미칼 태평양 한솔케미언스 팬택 성미전자 금호전기 희성전선 메디슨 대덕산업 삼화전자 삼화콘덴서 코리 아써키트 케이씨텍 태영 국동 도시가스 한일시멘트 평화산업 한라공조 세아제강 이구산업 동원산업 농심 삼양제넥스 등이다.

<추출결과>

추출 정보	추출된 내용
중심사건절	우량 중소형주가 주도종목군으로 부상해
세부정보	거래소내에 중소형주 중 첨단 기술력을 갖춘 기업과 낙폭이 과대한 실적 우량주, 우량 유가증권을 보유한 신자산주 등 우량 중소형주가 시장의 주도종목군으로 부상해
사건의 배경	
사건의 원인	이같은 장세 속에서도 코스닥시장의 차별화가 완화되면서
사건의 결과	우량 중소형주가 3월의 장세를 이끈다

본 실험문서는 증권 관련 기사로, 사건의 배경 정보는 나타나지 않았으므로 추출되지 않았다.

5. 결론

본 논문에서는 신문기사문에 특정적인 정보 추출의 내용과 방법을 제시하였다. 정보 추출의 내용으로 신문기사의 내용 요소인 육하원칙을 중심으로 한 다섯 가지 정보를 제시하였으며, 추출방법으로는 추출 정보에 따라 통계적인 방법과 언어적 지식을 함께 사용하였다. 본 논문에서는 단락이나 문장이 아닌 절 이하 단위의 추출을 함으로써 비교적 짧은 문서를 대상으로 하는 요약에서 추출되는 양에 비해 이용자가 필요로 하는 정보의 양을 최대화할 수 있다. 그리고 중심절을 추출한 뒤 그 절과의 관계를 통해 나머지 정보들을 추출함으로써 동일하거나 유사한 내용이 반복되지 않고 이 추출 정보로 요약문을 생성할 경우에 긴밀한 내용의 요약문을 생성할 수 있다.

현 시스템에서는 중심절 추출의 결과가 전체 정보 추출 결과에 큰 영향을 미치므로 중심절 추출의 정확도를 높이기 위한 방안을 모색 중이며, 특히 화제가 둘 이상일 경우에 대한 중심절 추출 방법을 모색하고 있다.

그리고 앞으로 이렇게 추출된 정보들을 가지고 자연스러운 요약문을 생성하는 일이 향후 과제로 남아 있다.

6. 참고문헌

[1] 류동원, 이종혁, "단어 공기 정보를 이용한 자동화 문서 요약", 정보과학회 봄 학술발표 논문집(B), 제 27권 1호, pp.345-347, 2000

[2] 강상배, "한국어 문서의 통계적 정보를 이용한 문서 요약 시스템 구현", 부산대학교 대학원 전자계산학과 석사학위논문, 1998

[3] 김계성 외, "단락 자동 구분을 통한 중요 문장 추출", 제12회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.233-237, 2000

[4] 박혁로, 신중호, 김태희, "검색/요약/필터링을 위한 텍스트 이해 모형 및 처리기술 개발", 연구 개발 정보 센터 연구 보고서, 1999

[5] 정준호, "수사구조를 이용한 문서 요약 시스템", 경북대학교 컴퓨터공학과 석사학위논문, 1999

[6] Daniel Marcu, "The Rhetorical Parsing, Summarization, and Generation of Natural Langauge Texts", Ph.D dissertation, University of Toronto, Canada, 1997.

[7] 이종근, 올바로 써야 기사가 된다, 전국언론노동조합연맹, 1997