복수의 신문기사 자동요약에 관한 실험적 연구

An Experimental Study on Automatic Summarization of Multiple News Articles

김 용 광(Yong Kwang Kim)*
정 영 미(Young Mee Chung)**

초 록

이 연구에서는 복수의 신문기사를 자동으로 요약하기 위해 문장의 의미범주를 활용한 템플리트 기반 요약 기법을 제시하였다. 먼저 학습과정에서 사건/사고 관련 신문기사의 요약문에 포함할 핵심 정보의 의미범주를 식별한 다음 템플리트를 구성하는 각 슬롯의 단서어를 선정한다. 자동요약 과정에서는 입력되는 복수의 뉴스기사들을 사건/사고 별로 범주화한 후 각 기사로부터 주요 문장을 추출하여 템플리트의 각 슬롯을 채운다. 마지막으로 문장을 단문으로 분리하여 템플리트의 내용을 수정한 후 이로부터 요약문을 작성한다. 자동 생성된 요약문을 평가한 결과 요약 정확률과 요약 재현율은 각각 0.541과 0.581로 나타났고, 요약문장 중복률은 0.116으로 나타났다.

영 문 초 록

This study proposes a template-based method of automatic summarization of multiple news articles using the semantic categories of sentences. First, the semantic categories for core information to be included in a summary are identified from training set of documents and their summaries. Then, cue words for each slot of the template are selected for later classification of news sentences into relevant slots. When a news article is input, its event/accident category is identified, and key sentences are extracted from the news article and filled in the relevant slots. The template filled with simple sentences rather than original long sentences is used to generate a summary for an event/accident. In the user evaluation of the generated summaries, the results showed the 54.1% recall ratio and the 58.1% precision ratio in essential information extraction and 11.6% redundancy ratio.

키워드: 복수문헌 자동요약, 뉴스기사 자동요약, 템플리트, 슬롯 단서어, 의미범주
multi-document summarization, news article summarization, template, slot cue word, semantic category

* 연세대학교 문헌정보학과 대학원(ykkim@yonsei.ac.kr)
** 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr)

1. 서론

최근 개인의 인터넷 이용이 급증하고 생활화되면서 포털사이트나 신문사 홈페이지를 통해 뉴스를 검색하거나 브라우징하는 이용자가 많아지고 있다. 그러나 대규모 사건/사고의 경우 보도하는 기사가 다수 존재하기 때문에 이러한 기사들 중 필요한 기사를 검색하거나 브라우징하기 위해서는 이용자의 많은 노력이 필요하다. 따라서 하나의 사건에 대한 다수의 기사에서 핵심적인 정보만을 추출하여 요약문을 제시한다면 이용자의 이러한 노력을 상당 부분 감소시킬 수 있을 뿐 아니라 이 요약문 자체가 원문의 대용물(surrogate)이 될 수 있다.

효과적인 요약문을 작성하기 위해서는 원문의 핵심적인 정보를 파악하는 과정과 파악된 정보를 문장으로 생성하거나 적절한 순서와 형태로 구조화하는 과정이 필요하다. 요약문 작성을 위해 원문으로부터 핵심 정보를 추출하기 위한 방법으로 통계적인 방법부터 구문론이나 의미론, 화용론 등을 응용한 언어학적 방법까지 다양한 자동요약 기법이 제안되었다. 언어학적 방법은 텍스트를 좀 더 깊이 있게 분석할 수 있는 장점이 있지만 이를 위해서는 정교한 자연언어 분석 도구들이 필요하며 특히 한글의 경우 이러한 도구들이 많지 않다는 점에서 한글 텍스트에 적용하기가 쉽지 않다. 또한 정교한 자연언어 분석은 상당한 처리시간을 요구하기 때문에 인터넷과 같이 실시간으로 변화하는 환경에서는 적용하기 힘들다. 국외 연구의 경우 테러리즘(terrorism)이나 기업 합병과 같은 한정된 주제 분야의 뉴스기사를 대상으로 템플리트 채우기(template filling) 방법으로 뉴스기사를 자동요약하려는 시도가 있었지만(McKweon and Radev 1995; Rau, Jacobs, and Zernik 1989; DeJong 1982), 특정 주제 분야의 템플리트를 구축하기 위해서는 많은 비용과 시간이 소요되며 상당한 수준의 언어학적 분석이 필요하다는 점에서 한계가 있다.

따라서 실시간으로 뉴스가 제공되는 인터넷 환경에서 비교적 간단한 통계적 방법과 낮은 수준의 언어학적 방법을 도입하여 복수의 신문기사들을 효과적으로 요약할 수 있는 자동요약 기법에 대한 필요성이 커지고 있다. 동일한 사건이나 사고에 대해 작성되는 다수의 신문기사들로부터 하나의 요약문을 작성하는 것을 복수문헌 요약(muli-document summarization)이라고 하며, 복수문헌 용약 기법으로는 앞에서 언급한 템플리트 기법을 비롯하여 클러스터링 기법(Stein, Bagga, and Wise 2000; Radev, Jing, and Budzikowska 2000; Ando et al. 2000), MMR 기법(Goldstein, et al. 2000), 응집성 그래프 기법(Mani and Bloedorn 1999) 등이 있다(정영미 2005).

이 연구에서는 기존에 제안된 템플리트를 확장하여 보다 넓은 주제 분야에 적용하고자 하였다. 또한 최종 요약문 작성 시 텍스트 문장을 단일의 주어와 술어로 구성된

단순한 형태의 문장인 단문(simple sentence)으로 분리하여 요약문 단위로 사용함으로써 문장 내 불필요한 정보를 최대한 제거하고 핵심적인 정보만을 요약문에 제시하고자 하였다. 이로써 복수문헌 자동요약에서의 중요한 고려 사항 중 하나인 압축률(compression ratio)을 높이고 특정 사건에서 유사한 역할을 하는 문장들을 한데 모음으로써 요약문 전체의 구조적인 응집력(coherence)을 높이고자 하였다.

이 연구에서 제시한 요약 기법은 문장을 템플리트의 각 필드 즉, 슬롯에 채우기 위한 슬롯 단서어를 선정하는 학습과정과 입력된 일련의 사건 기사를 범주화하고 그 요약문을 작성하는 과정으로 구분된다.

2. 자동요약 실험 설계

2.1 실험문헌집단

인터넷 환경에서 뉴스기사는 실시간으로 업데이트되는 특징을 가지며 일반적으로 다수의 신문기사를 요약해 주는 시스템은 단신의 사건보다는 시간에 따라 내용이 변화하고 여러 날에 걸쳐 게재되는 비교적 큰 규모의 사건에 적합하다. 그러나 현재 자동요약 연구에 사용할 수 있는 뉴스기사의 실험문헌집단은 대체로 단신의 사건이 많고 같은 사건을 보도하는 기사의 수는 많지 않아 본 실험연구에 적합하지 않다.

따라서 이 연구에서는 한국언론재단에서 운영 중인 뉴스검색 데이터베이스인 카인즈(KINDS)를 이용하여 2004년 1월부터 2005년 8월까지 발생했던 10개의 국·내외 대규모 사건/사고를 대상으로 실험용 기사를 수집하여 실험문헌집단을 새롭게 구축하였다.

실험문헌집단에 포함된 각 사건/사고에 대한 기본 정보 및 기사 수는 표 1과 같다.

<표 1> 실험문헌집단의 사건/사고 정보

	사건 번호	사건 / 사고명	기사수	평 균 문장수
학습집단	E1	양양/고성 산불	20	13.7
	E2	스페인 폭탄 테러	20	17.8
	E3	여중고생 성폭행	20	13.8
	E4	부천 초등학생 살해	18	12.1
	E5	런던 폭탄 테러	19	13.8
실험집단	E6	포천 여중생 살해	18	12.1
	E7	용천역 열차 폭파	16	15.8
	E8	남아시아 지진해일	14	14.1
	E9	동해 총기탈취	15	10.9
	E10	연천 GP 총기난사	15	12.3
총 175개 기사				13.6

실험문헌집단 중 E1부터 E5까지의 5개 사건, 97개 뉴스기사는 템플리트 슬롯 단서어를 선정하기 위한 학습문헌으로 사용하였고, 사건번호 E6부터 E10까지의 나머지 78개 뉴스기사는 실제 요약문을 작성하기 위한 실험문헌으로 사용하였다.

2.2 실험 개요

이 연구에서는 다수의 뉴스기사를 요약하기 위해 템플리트(template)를 이용하였다. 그러나 McKeown and Radev(1995)가 사용했던 템플리트가 핵심 정보 즉, 범인, 발생일, 장소 등의 정보(단어 혹은 구)를 추출하여 템플리트의 각 슬롯을 채운 반면 이 연구에서는 각 슬롯을 기사에서 선정된 문장으로 채웠다.

특정 사건이나 사고가 발생했을 때, 발생 장소나 시간 등의 정보나 피해상황, 정부나 기관의 대처방안, 수사과정 등은 해당 사건을 요약하는데 중요한 정보가 될 수 있다. 실제로 실험문헌집단에 대해 현재 문헌정보학과 석사과정에 재학 중인 두 명의 대학원생이 각 기사에서 요약문으로 작성될 수 있는 중요한 정보들을 담고 있는 것으로 판단되는 문장들을 추출하도록 하였다. 이 중 일치되는 문장만을 선정하여 이를 의미별로 분류하게 한 결과 중요 문장들은 6개의 상위 범주와 14개의 하위 범주로 구분되었다. 이들 각 의미범주를 템플리트 필드로 사용하였고 템플리트를 구성하는 각 필드와 그 의미는 표 2와 같다. 이 때, 학습기사로부터 식별된 의미범주는 템플리트의 각 슬롯을 이루게 되고, 실제 요약문 작성 과정에서 추출된 주요 문장은 특정 의미범주로 분류되어 템플리트의 슬롯에 채워지게 된다.

<표 2> 템플리트 구성 필드

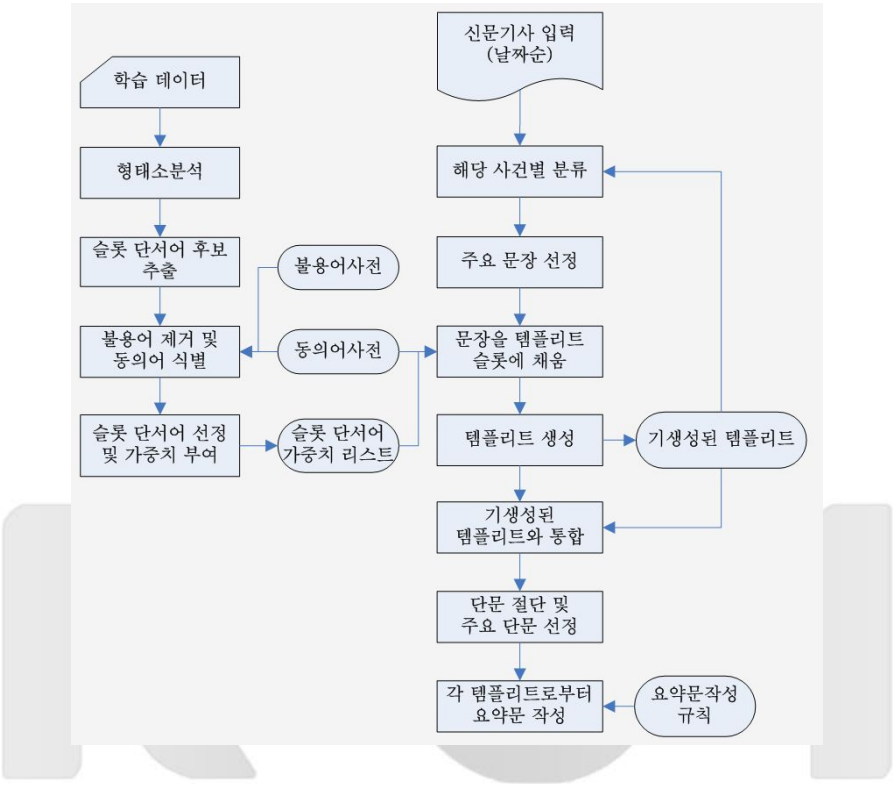
상위 필드	하위 필드	의 미
발생	발생	발생장소와 발생 시각, 발생 사실 등을 포함하는 사건의 기본 정보
원인	가해자	사건의 용의자를 포함한 가해자 정보
	직접원인	사건의 직접적인 원인 혹은 살인사건의 경우 사인
	간접원인	방비 부족 등의 사건의 간접적 원인
	사건의도	사건을 일으킨 가해자의 의도에 관한 정보
피해	인명피해	사건으로 인한 인명 피해 규모 및 피해자 정보
	재산피해	인명 피해 이외의 모든 재산 피해에 관한 정보
과정	조사과정	정부 기관의 사건 조사과정에 대한 정보
	진행상황	사고의 경우 시간에 따른 진행 상황
결과	법적결과	정부기관의 법적 대응(구속, 기소 등)
	대책	정부/기관의 공식적인 대책에 대한 정보
	기타결과	기타 사건으로 인한 영향 및 파장
반응	정부/기관의 입장표명	사건에 대한 정부/기관의 입장 표명
	사건에 대한 외부 의견	각종 분야의 전문가의 사건에 대한 의견이나 시민의 반응

이 연구의 실험 과정은 크게 요약 템플리트를 채우기 위해 사용되는 슬롯 단서어를 선정하는 과정과 실제 요약문 작성과정으로 나뉜다.

첫 번째 과정은 요약문을 작성을 위한 템플리트의 각 슬롯에 주요 문장을 채우기 위해 필요한 슬롯 단서어를 선정하여 가중치를 부여하는 과정이다. 이 과정에서는 특정 사건을 보도하는 다수의 뉴스기사로부터 주요 문장을 수작업으로 추출하여 각 문장을 의미범주로 태깅한 학습문헌을 이용한다. 우선 학습문헌을 형태소 분석한 후 특정 품사의 단어들을 추출한 후 슬롯 단서어를 선정하고 몇 가지 가중치 기법을 이용하여 각 단서어에 가중치를 부여하여 슬롯 단서어 가중치 리스트를 작성한다. 이 때 선정된 단서어의 가중치 부여 방법에 따라 특정 문장이 각 슬롯에 삽입되는 정확도를 비교한 후 최적의 가중치 부여 방법을 이용하였다.

두 번째 과정에서는 입력되는 뉴스기사를 사건별로 클러스터링하고 각 기사에서 주요 문장을 추출한 후 첫 번째 과정에 작성한 단서어 가중치 리스트를 이용하여 추출된 문장을 템플리트 슬롯에 채워 각 기사의 템플리트를 생성한다. 이때 동일한 사건의 기존의 템플리트가 존재할 경우 템플리트를 통합한다. 이 과정에서 새롭게 생성된 템플리트의 각 슬롯에 존재하는 문장과 기존 템플리트의 문장을 비교하여 중복 문장을 제거한다. 마지막으로 템플리트가 통합된 후 각 문장을 단문으로 분리하고 의미없는 단문을 제거하여 템플리트를 최종 수정하고, 각 사건 템플리트로부터 요약문을 작성한다.

실험 과정의 전체적인 개요를 정리하면 그림 1과 같다.



<그림 1> 자동요약 실험 개요도

2.3 슬롯 단서어 선정 과정

이 과정은 요약 템플리트 작성을 위한 과정으로 각 사건의 요약문을 작성과 병행되는 과정이 아니라 요약문 작성을 위한 이전 단계로 볼 수 있다. 실험집단에서 수작업으로 주요 문장을 추출한 후, 해당 의미범주로 할당한 문장을 학습문헌으로 이용하여 그림 1에서와 같이 네 단계의 과정을 거쳐 슬롯 단서어 가중치 리스트를 생성한다.

첫 번째 단계에서는 “21세기 세종계획”에서 개발한 지능형형태소분석기를 이용하여 형태소 분석을 하고 두 번째 단계에서 형태소 분석 결과를 이용하여 동사, 고유명사, 의존명사, 일반명사를 대상으로 슬롯 단서어 후보를 추출하였다. 다음 단계에서는 수작업을 통해 불용어를 제거하고 동의어 사전을 작성하여 동의어를 식별하였다. 마

지막 단계에서는 슬롯 단서어를 선정하고 가중치를 부여하여 슬롯 단서어 가중치 리스트를 작성하고 이는 요약문 작성 시 각각의 기사로부터 추출된 주요 문장을 슬롯에 채워 템플리트를 생성할 때 이용되었다.

슬롯 단서어 후보를 추출하고 유의어를 식별한 후 단어의 문장 내 출현빈도와 의미범주를 이용하여 슬롯 단서어를 선정하고 가중치를 부여하게 된다. 특정 문장을 의미범주로 분류하는데 적합한 단서어는 여러 문장에 공통적으로 출현하는 단어가 될 가능성이 높다. 예를 들어 다음과 같은 문장을 '인명 피해'라는 의미범주로 분류한다고 가정해 보자.

용천 폭발사고로 현장에서 2천~3천명이 죽고 다친 사람은 7천~8천명에 이른다.

위 문장에서 '용천', '폭발' 등의 단어는 용천의 기차 폭발 사건을 구분하는데는 좋은 단어가 되지만, 이 문장을 '인명 피해'라는 의미를 갖게 하는 단어들은 '명', '죽다', '다치다', '사람'과 같은 일반적인 단어들이 될 것이다. 따라서 각 슬롯을 대표하는 슬롯 단서어를 선정하고 가중치를 부여하기 위한 기본적인 가정은 다음과 같다.

첫째, 많은 문장에 출현하는 일반적인 단어일수록 좋은 단서어가 될 것이다.
둘째, 특정 의미범주에 많이 출현한 단어일수록 좋은 단서어가 될 것이다.
셋째, 특정 의미범주의 문장에만 출현하는 단어는 그렇지 않은 단어보다 좋은 단서어가 될 것이다.

이 연구에서 슬롯 단서어를 선정하는 것은 문헌을 범주화하기 위한 자질의 선정과 유사하다. 즉, 문헌 범주화의 문헌은 문장이 되며 범주는 템플리트의 필드를 구성하는 의미범주가 된다. 따라서 이 연구에서는 텍스트 범주화에서 이용하는 자질 선정 방법을 수정하여 사용하였다. Yang and Pedersen(1997)의 연구에서 비교적 좋은 성능을 나타냈던 문헌빈도(Document Frequency)와 카이제곱 통계치(chi square statistics)를 사용하여 슬롯 단서어를 선정하였다. 추가로 특정 단어가 특정 의미범주에 속할 조건 확률을 슬롯 단서어 를 선정하고 가중치를 부여하였다. 이와 같이 세 가지 기법을 이용하여 성능을 비교하여 최적의 가중치 부여 기법을 요약문 작성에 이용하였다.

첫 번째로 사용한 가중치 공식은 문장빈도*역범주빈도이다. 통계적 기법을 이용한 자동요약 연구에서 문헌빈도는 문장빈도(Sentence Frequency)로 수정되며 따라서 문장빈도란 특정 단어가 출현한 문장의 수가 된다(정영미 2005). 사전 실험결과 문장빈도가 2 이상인 단어를 단서어로 선정하였을 때 최적의 성능을 보였다. 범주빈도란

특정 단어가 출현하는 범주의 수로 이 실험에서는 그 값이 작을수록 좋은 단서어가 될 수 있다. 따라서 이 연구에서는 다음과 같이 문장빈도*역범주빈도를 이용하여, 단서어 t의 의미범주 c_i에서의 슬롯 단서어 가중치 w(t, c_i)를 산출하였다.

$$w(t,c_i) = \log(1+sf(t,c_i)) \times \log(\frac{NC}{cft(t)}+1)$$

위 공식에서 sf(t , c_i)는 특정 의미범주 c에 속하는 단어 *t*의 문장빈도이며, cft(t)는 단어가 출현하는 의미범주의 수 즉 범주빈도, NC는 의미범주의 수로 이 연구에서는 14이다. 또한 문장빈도가 큰 특정 단어의 영향력을 감소시키기 위해 로그를 취해 각 가중치의 크기 차이를 감소시켰다.

카이제곱 통계치는 특정 단어 t와 범주 c간의 의존성을 측정하는 값으로 값이 클수록 특정 단어 t가 범주 c와 연관될 확률이 커지며 다음과 같은 공식을 이용하여 산출된다(Yang and Pedersen, 1997).

$$\chi^2(t,c) = \frac{N \times (AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

A : t가 특정 범주 c에 출현한 회수
B : t가 출현하지 않은 c의 개수
C : t가 c 이외의 범주에 출현한 회수
D : t도 c도 출현하지 않은 회수
N : 전체 문장 수

카이제곱 통계치는 특정 단어가 각각의 범주와 연관된 정도를 계산하는 것으로 특정 단어의 자질선정을 위해서는 특정 단어의 각 범주별 카이제곱 통계치의 최대값 혹은 평균값을 사용한다. 이 연구에서는 최대값이 기준치 이상인 단어를 단서어로 선정하였고 사전실험 결과 기준치를 4로 하였을 때 최적의 성능을 보였다. 슬롯 단서어 가중치 w(t, c_i)는 다음과 같다.

$$w(t,c_i) = \frac{\chi^2(t,c_i)}{\chi^2_{max}(t)}$$

마지막으로 사용한 슬롯 단서어 가중치 부여 방법은 각 단어가 각 의미 범주에 속

할 조건 확률로 이 실험에서는 별도의 단서어 선정 과정 없이 모든 단서어 후보를 사용하였다. 특정 단어 t가 특정 범주 c에 속할 조건 확률 $P(c_i|t)$는 다음과 같은 공식으로 구할 수 있다(Myaeng and Jang, 1999).

$$w(t, c_i) = P(c_i|t) = \frac{P(t|c_i)P(c_i)}{P(t)}$$

위 공식에서 $P(c_i)$는 특정 의미범주가 발생할 확률로 여기서는 의미범주가 14개이므로 1/14의 상수값을 가지며 P(t)는 전체 중요 문장에서 용어 t가 출현할 확률로 t의 출현빈도를 전체 용어의 출현빈도로 나눈 값이 된다. $P(t|c_i)$는 단어 t가 의미범주 c_i에 출현한 횟수를 c_i에 출현한 단어의 총수로 나누어 계산할 수 있다.

이렇게 선정된 슬롯 단서어과 그 가중치는 요약문 작성 과정에서 주요 문장을 각 템플리트의 슬롯에 채우기 위한 사전 정보로 이용된다. 이는 표 3과 같은 슬롯 단서어 가중치 리스트로 저장된다.

<표 3> 슬롯 단서어 가중치 리스트의 예

상위필드	하위필드	단어	품사	가중치
피해	재산피해	타	동사	0.683
피해	재산피해	채	의존명사	0.479
피해	재산피해	집계	명사	0.582
…	…	…	…	…
발생	발생	일어나	동사	0.653
발생	발생	발견	명사	0.552
발생	발생	시	의존명사	0.611

2.4 요약문 작성과정

실제로 요약 대상의 뉴스기사가 날짜순으로 입력되면 우선 <지능형형태소분석기>를 이용하여 형태소 분석을 수행한 후 다음의 여섯 단계에 걸쳐 요약문을 작성한다.

① 기사가 입력되면 해당 기사가 속하는 사건의 클러스터를 식별한다.
② 입력된 기사에서 주요 문장을 추출한다.
③ 선정된 문장을 템플리트 슬롯 단서어 리스트를 이용하여 템플리트 슬롯에 채워 템플리트를 생성한다.

④ 생성된 템플리트로부터 ①번 단계에서 식별된 사건 템플리트의 각 슬롯에 있는 문장들과 비교하여 중복 문장을 제거하고 앞서 생성된 사건 템플리트와 통합한다.
⑤ 템플리트의 각 슬롯에 있는 문장을 단문으로 분리하여 의미 없는 단문을 제거한다.
⑥ 최종적으로 사건 템플리트로부터 해당 사건의 요약문을 작성한다.

처음 입력되는 기사는 ①번 단계와 ④번 단계를 생략하고 최종적으로 하나의 기사에 대한 요약문이 생성된다.

2.4.1 입력 기사의 사건별 클러스터링

첫 번째로 입력된 기사는 하나의 사건 클러스터를 생성하게 되고 다음 단계들을 걸쳐 하나의 템플리트로 생성된다. 이 후 새로운 기사가 입력되면 입력된 기사로부터 문헌벡터를 생성하고 기존에 생성된 사건 템플리트의 벡터와 유사도를 산출한 후 기준치 이상이면 유사도가 가장 큰 사건 클러스터로 식별하고 기준치 이하이면 이 기사는 새로운 사건 클러스터를 생성한다. 이 때 템플리트 벡터는 템플리트의 슬롯에 채워진 문장에 출현하는 단어들로부터 생성된다. 벡터간 유사도는 코사인 유사도를 이용하여 산출하였고(정영미 2005), 사전 실험 결과 유사도 임계치를 0.38로 설정하였을 때 분류재현율과 분류정확률이 각각 0.893과 0.896으로 가장 좋은 결과가 나타났다.

2.4.2 주요 문장 추출

뉴스기사의 경우 일반적으로 기사의 제목과 첫 문장에 기사의 핵심 정보가 나타나는 경우가 많다. 따라서 해당 사건을 보도하는 모든 기사의 제목과 첫 문장에 출현하는 용어를 벡터로 생성하고 각 문장벡터와의 유사도를 계산하여 각각 상위 10%, 20%, 30%, 40%의 문장을 추출한 후 주요 문장 추출 정확도를 비교하였다. 이 연구의 예비 실험에서 몇 가지 용어가중치를 사용하여 주요 문장을 추출한 결과 용어가중치 간 성능에 유의미한 차이는 없었다. 따라서 이 실험에서는 가장 간단한 용어가중치인 이진단어빈도(binary term frequency in sentence)를 사용하고 벡터 유사계수는 코사인 유사계수를 사용하였다(정영미 2005).

2.4.3 템플리트 채우기

두 번째 단계에서 추출된 주요 문장을 슬롯 단서어 가중치 리스트를 이용하여 템플리트의 각 슬롯에 삽입한다. 이 때 각 문장이 템플리트의 각 슬롯에 포함될 가능성의 점수를 투표 점수(voting score) 방식으로 계산하였다. 중요 단문 s_k가 c_i에 포함될 점수 즉, 문장점수는 다음과 같이 문장에 포함되는 슬롯 단서어의 가중치의 합으로 계산된다.

$$score(s_k, c_i) = \sum_{t \in s_k} w(t, c_i)$$

여기서 $w(t, c_i)$는 슬롯 단서어 가중치 테이블에 포함된 각 슬롯 단서어의 가중치이며 각 문장 s_k는 문장점수 $score(s_k, c_i)$의 최대값을 갖는 슬롯에 삽입된다. 주요 문장이 모두 템플리트의 해당 슬롯에 채워지게 되면 하나의 템플리트가 생성된다.

2.4.4 중복 문장 제거 및 기생성된 템플리트와 통합

이전 단계에서 입력된 기사의 템플리트가 생성되면 템플리트 슬롯에 입력된 문장 간의 유사도를 산출하여 임계치 이상의 유사도를 갖는 문장을 제거한다. 문장 간 유사도는 이진단어빈도와 코사인계수로 산출하였고(정영미 2005), 사전 실험 결과 임계치를 0.7로 했을 때 최적의 성능을 보였다. 중복 문장이 제거되면 같은 클러스터에 포함되는 기사를 앞서 생성된 템플리트와 통합한다.

2.4.5 단문 분리 및 의미 없는 단문의 제거

각 사건/사고의 템플리트가 생성되면 템플리트 슬롯에 입력된 각 문장을 단문으로 분리한 후 의미 없는 단문을 제거한다. 우선 각 문장을 단문으로 분리하기 위해 형태소 분석 결과와 다음과 같은 간단한 규칙을 이용하였다.

1) 동사(VV)나 형용사(VA)가 선어말어미(EP), 종결어미(EC), 연결어미(EC)를 만나면 단문으로 분리한다.
2) 명사(NN) + 지정사(VC) 또는 명사(NN) + 동사파생접미사(XSV) 혹은 형용사파생접미사(XSA)를 만나면 단문으로 분리한다.
3) 그러나 1)과 2)와 같은 경우라도 명사형전성어미(ETN)나 관형형전성어미(ETM)가 뒤따르면 단문으로 분리하지 않는다.
4) 1)과 같은 형태가 오고 다음 어절에 보조용언(VX)이 나타나면 보조용언 다음에 분리한다.

이와 같은 규칙으로 실제로 다음과 같이 문장을 단문으로 분리할 수 있다.

< 문장 > : 천년 고찰 낙산사의 주요 건물이 불타버린 것은 강한 바람 탓도 있지만, 당국의 안이한 대응도 한몫했다는 지적이 일고 있다.

<형태소 분석 결과>

천년 천년/NNG
……
불타버린 불타/VV+아/EC+버리/VX+ㄴ/ETM …… ㉠
것은 것/NNB+은/JX
강한 강하/VA+ㄴ/ETM …… ㉡
바람 바람/NNG
탓도 탓/NNG+도/JX
있지만, 있/VA+지만/EC+,/SP …… ㉢
……
한몫했다는 한몫/NNG+하/XSV+았/EP+다는/ETM
지적이 지적/NNG+이/JKS
일고 일/VV+고/EC …… ㉣
있다.있/VX+다/EF+./SF …… ㉤

㉠과 ㉡에서 VV와 VA가 EC가 나타나지만 ③번 규칙에 의해 단문으로 분리되지 않고 ㉢에서 ①번 규칙에 의해 분리된다. ㉣에서 VV와 EC가 나타나지만 다음 어절에 VX가 뒤따르므로 규칙 ④번에 의해 “있다.” 다음에 단문으로 분리되어 다음과 같은 두 개의 단문이 생성된다.

<단문 1> : 천년 고찰 낙산사의 주요 건물이 불타버린 것은 강한 바람 탓도 있지만,
<단문 2> : 당국의 안이한 대응도 한몫했다는 지적이 일고 있다.

일단 슬롯에 입력된 문장이 단문으로 분리되면 각 단문이 전달하는 정보의 양은 다를 수 있다. 따라서 이 과정에서 의미 없는 단문을 삭제했을 경우 요약 문장의 질을 높일 수 있고 요약의 압축률 또한 높일 수 있다. 각 단문의 중요도는 크게 두 가지 기준으로 측정될 수 있는데, 하나는 단문이 해당 사건의 정보를 얼마나 많이 담고 있는가이고 다른 하나는 단문이 해당 템플리트의 슬롯 즉, 의미범주에 얼마나 적합한가이다. 전자는 해당 단문의 벡터와 단문이 속한 템플리트의 단어로 구성된 벡터와의 유사도로 측정될 수 있으며 이를 단문의 중심성으로 정의한다. 후자는 해당 단문의

슬롯 단서어 가중치의 투표점수인 단문점수로 측정될 수 있다. 이 두 기준을 모두 고려하기 위해 단문의 중심성과 단문점수에 의해 각각의 순위를 산출하고 각 순위를 더한 순위점수가 작은 단문을 선택하고 순위점수가 임계치 이상의 단문은 삭제하였다. 임계치는 1.6×(슬롯 내 단문의 개수)로 정하였다. 이는 최대 순위의 80% 지점을 의미한다.

2.4.6 요약문 작성

최종적으로 템플리트의 각 슬롯에 삽입된 문장(의미 없는 단문이 삭제된 경우에는 단문)을 이용하여 최종적으로 요약문을 작성한다. 이를 위한 요약문 작성 규칙은 크게 요약문 배열 규칙과 문장 생성 규칙으로 구별된다.

이 연구에서 사용한 요약문 배열 규칙은 다음과 같다.

1) 요약문의 전체적인 구조는 템플리트의 상위 필드인 '발생', '피해', '원인', '과정', '결과', '반응' 순으로 한다. 또한 상위 필드는 문단을 구분하는 기준이 된다.
2) 상위 필드 안의 단문들은 하위 필드를 기준으로 정렬하고 해당 단문이 속한 기사의 게재일 순으로 정렬한다.
3) 기사의 날짜가 같은 경우 단문점수의 순위와 단문의 중심성 순위 합계의 역순으로 문장을 정렬한다.
4) 하나의 문장에 속한 단문은 차례로 배열한다.

요약문 배열 규칙에 의해 요약문의 순서가 정해지면 문장생성 규칙에 의해 적절한 문장으로 작성된다. 이 규칙은 요약문의 가독성을 확보하기 위한 여러 가지 규칙을 포함한다. 다음은 문장생성 규칙 중 몇 가지 사례이다.

1) 단문 첫 머리에 관형사(MM)나 부사(MA*)가 나오면 삭제한다. 이 때 부사(MA)는 일반부사(MAG)와 접속부사(MAJ)로 구분된다.
2) 종결어미(EF)로 끝나지 않는 단문의 경우 술어의 기본형에 'ㄴ 것으로 알려졌다.'를 추가하고 '하'나 '되'와 같이 동사파생접미사(XSV)나 형용사파생접미사(XSA)로 끝나는 경우 각각 '하였다.', '되었다.' 등으로 수정한다.

3. 자동요약 실험 결과

3.1 주요 문장 추출 결과

주요 문장을 자동으로 추출했을 때의 성능을 나타내는 정확률과 재현율을 측정하기 위해 수작업으로 주요 문장을 추출한 결과와 비교하였다. 이 때 수작업 추출 문장은 중복된 문장을 모두 포함한 것이며, 정확률과 재현율은 시스템에 저장된 문헌번호와 문장번호의 직접적인 비교를 통해 산출하였다.

우선 문장의 추출비율에 따른 주요 문장의 추출 성능을 알아보기 위해 추출비율을 10%, 20%, 30%, 40%로 변화시켜 반복 실험하였다. 표 4는 실험집단에 속한 모든 기사에 출현한 문장을 대상으로 주요 문장을 추출한 결과 산출된 정확률과 재현율이다.

<표 4> 추출 비율에 따른 정확률과 재현율

추출 비율	정확률	재현율	F척도
10%	0.617	0.439	**0.513**
20%	0.427	0.591	**0.496**
30%	0.289	0.556	0.369
40%	0.258	0.622	0.365

표 3을 살펴보면 추출 비율이 10%일 때 가장 좋은 성능을 나타낸다. 그러나 10%를 추출 비율로 했을 경우 정확률은 높아지는 반면 재현율 값이 급격하게 떨어지므로 정보의 손실량이 너무 많아지게 되는 단점이 있다. 따라서 재현율이 어느 정도 확보되고 정확률이 높은 20% 지점이 추출비율의 가장 적당한 지점이라고 판단할 수 있다.

표 5는 추출 비율을 20%로 했을 때의 각 사건별 정확률과 재현율이다. 여기서의 사건번호는 시스템이 생성한 임의의 번호로 표 1의 사건번호와 일치하지 않는다.

<표 5> 사건별 주요 문장 추출 정확률과 재현율

사건번호	정확률	재현율	F척도
1	0.250	0.474	0.327
2	0.381	0.552	0.451
3	0.448	0.565	0.500
4	0.571	0.762	0.653
5	0.484	0.600	0.536
평 균	**0.427**	**0.591**	**0.496**

3.2 슬롯 단서어 가중치에 따른 실험 결과

입력되는 뉴스기사로부터 주요 문장을 추출한 후 학습문헌에서 추출한 슬롯 단서어와 그 가중치 정보를 이용하여 각 문장을 의미범주로 분류하여 각 템플리트 슬롯을 채우게 된다. 연구에서는 최적의 슬롯 단서어를 선정하고 가중치를 부여하기 위해 문장빈도*역범주빈도, 카이스퀘어 통계치, 조건 확률의 세 가지 가중치 기법을 이용하여 가중치를 부여한 후 의미범주로의 분류 정확률을 비교하였다.

실험 결과, 사건별로 다소의 차이는 있지만 특정 단어가 각 의미범주에 포함될 조건 확률을 가중치로 사용했을 때 분류정확률이 0.825로 가장 좋은 성능을 보였다. 표 6은 각각의 가중치 기법에 따른 평균 분류정확률을 나타낸다.

<표 6> 슬롯 단서어 가중치 기법에 따른 분류정확률

사건 번호	카이제곱	조건 확률	문장빈도*역범주 빈도
1	0.737	0.842	0.684
2	0.724	0.759	0.690
3	0.870	0.826	0.696
4	0.714	0.857	0.762
5	0.720	0.840	0.680
평균	0.753	**0.825**	0.702

3.3 요약 성능의 평가

이 실험에서 자동으로 생성된 요약문을 평가하기 위한 사용한 평가 척도로는 요약문의 정보성을 측정하는 요약 정확률 및 요약 재현율, 생성된 요약문에서의 중복도를 측정하는 요약 중복률, 단문을 문장으로 생성하면서 연결이 자연스럽지 않은 문장의 정도를 나타내는 문장생성 오류율을 사용하였다. 마지막으로 요약 압축률을 산출하였다.

3.3.1 요약 정확률과 요약 재현율

요약 정확률과 요약 재현율을 측정하기 위해 네 명의 이용자에게 자동으로 생성된 요약문을 수작업 요약문과 비교하도록 하여 같은 정보를 나타내는 문장을 표시하도록 하였고, 네 명중 세 명의 이용자 이상 적합문장으로 평가한 문장을 적합문장으로 판단하였다. 평가 결과, 5개 사건에 대한 평균 요약 정확률은 0.541, 요약 재현율은 0.581로 나타났다.

표 7은 각 사건에 대한 요약 정확률과 요약 재현율이다.

<표 7> 자동 요약문의 요약 정확률과 재현율

사건번호	요약 정확률	요약 재현율	F척도
1	0.421	0.471	0.445
2	0.433	0.481	0.456
3	0.565	0.565	0.565
4	0.636	0.824	0.718
5	0.650	0.565	0.605
평균	**0.541**	**0.581**	**0.560**

이 연구에서 제시한 요약 기법의 요약 성능은 통계적 기법을 이용한 다른 연구의 요약 성능에 비해 비교적 높게 나타났다. 초기 연구자인 Edmundson(1959)은 요약 정확률을 약 0.4로 보고하고 있으며, Myaeng and Jang(1999)이 제시한 요약 기법에 의하면 이 연구와 가장 유사한 재현율 수준인 0.532에서 0.395의 요약 정확률을 보였다.

3.3.2 요약 문장 중복률

다음으로 자동으로 생성된 요약문 내 정보의 중복 정도를 측정하기 위해 요약 문장 중복률을 산출하였다. 평가 결과, 전체 자동으로 생성된 요약문의 평균 22.8개의 문장 중 2.55개의 문장이 중복된 정보를 포함하고 있었다.

표 8은 사건별 요약 문장 중복률을 나타낸다.

<표 8> 자동 요약문의 요약 문장 중복률

사건 번호	자동 요약문 문장수	이용자별 중복 문장수 평균	요약 중복률
1	19	2.75	0.145
2	30	2.5	0.083
3	23	2.25	0.098
4	22	1.75	0.080
5	20	3.5	0.175
평균	22.8	2.55	**0.116**

3.3.3 문장생성 오류율

이 연구에서는 최종적인 요약문을 작성할 때 완전한 문장이 아닌 단문을 이용하기 때문에, 간단한 문장생성 규칙에 의해 단문의 어미를 수정하여 요약문을 작성하였다. 따라서 생성된 문장이 불완전하거나 원래의 의미를 다르게 전달하는 경우가 발생할 수 있기 때문에 문장생성 오류율을 측정하였다. 이 때, 문장생성 오류는 어법이나 문법에 오류가 있거나, 의미 없는 단문을 삭제하면서 주어나 목적어 등 문장의 필수적인 요소들이 누락되어 문장의 의미를 명확히 알 수 없는 경우로 정의하였다. 최종적으로 작성된 요약문 내 114개 중 문장생성 규칙에 의해 생성된 문장은 총 32개의 문장이고, 이 중 15개의 문장이 오류로 평가되어 0.469의 오류율을 보였다.

이와 같은 결과는 낮은 수준의 텍스트 분석에 의한 한계를 잘 보여주고 있다. 실제로 형태소 분석 수준에서 문장의 필수적인 요소를 식별하는데 매우 제한적이기 때문에 위와 같은 오류가 발생하였다.

3.3.4 단문 제거에 의한 원문 압축률의 변화

이 연구에서는 불필요한 정보를 제거하고 요약문의 압축률을 높이기 위해 문장을 단문으로 분리하여 의미 없는 단문을 제거하였다. 따라서 이러한 단문 제거의 유용성을 알아보기 위해 단문을 제거했을 때와 그렇지 않았을 때의 압축률의 변화를 측정하였다.

한편 일반적으로 문장의 길이는 매우 다양하기 때문에 문장의 수를 이용하여 압축률을 산출했을 경우 실제 요약문의 길이와 실제 압축률은 차이가 있을 수 있다. 따라서 편차가 작은 어절의 수를 이용하여 압축률을 산출하였다. 즉, 압축률은 생성된 요약문의 어절수를 원문의 뉴스기사의 총 어절수로 나눈 값이 된다. 표 9는 단문 제거에 따른 각 사건별 압축률의 변화를 나타낸다.

<표 9> 단문 제거에 따른 압축률 변화

사건 번호	원문 어절수	단문 제거전 요약문 어절수	단문 제거후 요약문 어절수	단문 제거전 요약문 압축율	단문 제거후 요약문 압축률	압축률 차이
1	4,233	420	385	0.099	0.091	0.008
2	4,167	570	425	0.137	0.102	0.035
3	2,877	425	328	0.148	0.114	0.034
4	3,157	552	384	0.175	0.122	0.053
5	3,198	520	413	0.163	0.129	0.034
평균	3,526.4	497.4	387.0	**0.144**	**0.112**	**0.032**

의미 없는 단문을 제거한 결과, 약 110개의 어절이 삭제되었고 압축률은 14.4%에

서 11.2%로 약 3.2%가 향상되었다. 따라서 의미 없는 단문을 제거하는 것이 요약문의 압축률을 높이는데 어느 정도 유용한 것으로 나타났다.

4. 결론

인터넷을 통해 뉴스의 정보를 얻는 이용자가 많아지면서 뉴스 검색 서비스가 급증하고 있다. 그러나 해당 사건을 보도하는 기사가 다수 존재할 경우 이용자에게 요약된 형태의 정보를 제공한다면 이용자가 손쉽게 해당 사건을 개관할 수 있다.

이를 위해 우선 수작업을 통해 기사로부터 요약 문장을 추출하여 각각의 문장의 의미범주를 식별하고 이 의미범주를 필드로 하는 템플리트를 이용하여 요약문을 작성하는 기법을 제안하였다. 이로써 해당 사건의 중요한 정보인 사건 발생일이나 장소, 사건 피해자, 가해자 정보, 수사 과정, 사건의 영향 등을 나타내는 문장을 한 곳에 모음으로써 요약문 내용의 응집성을 높이고, 추출된 문장을 단문으로 분리하여 의미 없는 단문을 제거함으로써 압축률을 높이고자 하였다.

이 연구에서 제시한 요약 기법의 성능을 최대화하기 위해, 문장빈도*역범주빈도, 카이제곱 통계치, 조건 확률의 세 가지 슬롯 단서어 가중치 기법을 이용했을 때 실제 문장을 템플리트 슬롯에 삽입할 때의 정확률을 비교하였다. 그 결과 조건 확률을 이용하여 가중치를 부여했을 때 0.825의 정확률로 가장 좋은 성능을 나타냈다.

또한 기사로부터 주요 문장을 추출하기 위하여 각 기사의 제목과 첫 문장을 질의벡터로 이용하여 기사 내 각 문장과의 코사인 유사도를 이용하였다. 실험 결과 상위 20%의 문장을 추출했을 때 0.591의 재현율을 확보하면서 0.427의 비교적 높은 정확률 성능을 보였다.

최종적으로 자동 생성된 요약문을 이용자가 평가한 결과, 요약 정확률과 재현율은 각각 0.511과 0.568로 나타났고, 요약 문장 중복률은 0.116으로 나타났다. 한편 요약문 작성 규칙에 의해 생성된 문장 중 약 46%의 문장에서 오류가 발생하였다. 또한 요약문 압축률은 약 0.112로, 각 기사로부터 20%의 문장을 추출한 후 중복문장 제거 결과 약 6%의 문장이 제거되고 의미 없는 단문을 제거한 결과 추가적으로 약 3%가 제거된 것으로 나타났다. 이로써 단문 제거가 압축률 측면에서 유용한 것으로 나타났다.

이 연구에서 제시한 요약 기법의 특징은 다음과 같다.

첫째, 모든 유형의 사건/사고로부터 공통적인 단서어를 선정하여 요약문 작성에 이용함으로써 특정 주제영역에 상관없이 모든 사건/사고에 적용이 가능하다.

둘째, 유사한 정보를 전달하는 문장들을 한곳에 모음으로써 이용자가 원하는 정보를 손쉽게 찾을 수 있고 요약문에서의 응집성을 높일 수 있다.

셋째, 다른 통계적 기법을 이용한 요약 기법에 비해 비교적 높은 요약 성능을 보인다.

그러나 이 요약 기법은 문장 생성 오류율이 상당히 높게 나타났으며, 중복된 정보를 제거하는 데 있어서도 한계를 보였다. 이와 같이 자연언어 텍스트 처리의 불완전성으로 인해 발생한 문제점들은 면밀한 오류 사례 분석을 통해 해결해야 할 것으로 보인다.

참 고 문 헌

정영미. 2005. 『정보검색연구』. 서울: 구미무역(주) 출판부.

Ando, R. K., Boguraev, B.K., Byrd, R.J., and Neff, M.S. 2000. "Multi-document Summarization by Visualizing Topical Content." In Proceedings of the Workshop on Automatic Summarization, 79-88. New Brunswick, New Jersey: Association for Computational Linguistics.

DeJong, G. 1982. "An overview of the FRUMP system". In W.G. Lehnert and M.H. Ringle, eds. Strategies for Natural Language Processing. 149-176.

Edmunson, H.P. 1969. "New methods in automatic extracting." Journal of the ACM, 16(2): 264-285.

Goldstein, J., Mittal, V.O., Carbonell, J.G., and Kantrowitz, M. 2000. "Multi-document summarization by sentence extraction." In Proceedings of the Workshop on Automatic Summarization. 40-48.

Mani, I. and Bloedorn, E. 1999. "Summarizing similarities and differences among related documents." Information Retrieval, 1: 35-67.

McKeown, L. and Radev. D. 1995. "Generating summaries of muliple news articles." In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 74-82.

Myaeng, S.H. and Jang, D-H. 1999. "Development and evaluation of a statistically-based document summarization system." In I. Mani and M.T. Maybury, eds. Advances in Automatic text Summarization. Cambridge, MA: The MIT Press. 61-70.

Radev, D.R., Jing, H., and Budzikowska, M. 2000. "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies." NAACL/ANLP Workshop on Automatic Summarization. 21-30.

Rau, L.F., Jacobs, P.S., Zernik, U. 1989. "Information extraction and text summarization using linguistic knowledge acquisition." Information Processing and Management, 25(4): 419-428.

Stein, G.C., Bagga, A., and Wise, B. 2000. "Multi-document summarization: methodologies and evaluations." In Proceedings of Conference TALN 2000.

Yang, Y. and Pedersen, J.O. 1997. A comparative study on feature selection in text categorization. In Proceedings of the 14th International Conference on Machine Learning.