술어기반 문형정보를 이용한 자동요약시스템에 관한 연구

A Study on an Automatic Summarization System Using Verb-Based Sentence Patterns

최인숙(In-Sook Choe)*, 정영미(Young-Mee Chung)**

초 록

본 연구에서는 특정 주제분야의 텍스트를 대표할 수 있는 단서술어를 추출하고 기본문형을 형성한 후 각 단서술어의 기본문형을 실례화하여 연결함으로써 요약문을 작성하는 자동요약시스템의 모형을 설계하고 구현하였다. 시스템은 학습과정과 요약과정으로 구분되며, 학습과정에서는 술어와 격조사의 출현빈도를 이용하여 주제분야 텍스트집단을 대표하는 단서술어와 필수격 조사를 추출한 뒤 단서술어가 이루는 문장의 기본문형을 형성한다. 요약과정에서는 실례화규칙을 요약 대상 문장의 구문 분석 결과에 적용하여 기본문형의 격조사와 결합될 논항을 찾아 단문을 생성하고 연결하여 요약문을 완성한다. '화재' 및 '강도'와 관련된 신문기사를 대상으로 실험을 수행하였으며, 작성된 요약문은 단서술어가 포함된 주요 문장에서 추출한 필수 정보항목과 술어를 중심으로 생성된 문장들로서 문장간의 연결이 자연스러울 뿐 아니라 텍스트의 전체적인 의미를 표현할 수 있었다. 또한, 통계적 기법을 이용한 학습을 통해 주제영역의 확장이 가능하였다

ABSTRACT

The purpose of this study is to present a text summarization system using a knowledge base containing information about verbs and their arguments that are statistically obtained from a subject domain. The system consists of two modules: the training module and the summarization module. The training module is to extract cue verbs and their basic sentence patterns by counting the frequency of verbs and case markers respectively, and the summarization module is to substantiate basic sentence patterns and to generate summaries. Basic sentence patterns are substantiated by applying substantiation rules to the syntactic structure of sentences. A summary is then produced by connecting simple sentences that are generated through the substantiation module of basic sentence patterns. Topics of 'fire' and 'robbery' in the daily newspapers are selected for a test collection. The system generates natural summaries without losing any essential information by combining both cue verbs and essential arguments. In addition, the use of statistical techniques makes it possible to apply this system to other subject domains through its learning capability.

키워드 : 자동요약, 단서술어, 기본문형, 기본문형 실례화, 문장 생성, 학습

automatic summarization, cue verbs, basic sentence patterns, sentence pattern substantiation, sentence generation, learning system

* 연세대학교 문헌정보학과 시간강사(ischoe@hananet.net)

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

■ 논문 접수일 : 2001년 11월 8일

■ 게재 확정일 : 2001년 12월 17일

1 서 론

요약문은 원문에 수록된 내용 중 가장 핵심적인 정보들이 문장 형식으로 표현되기 때문에 키워드 중심의 검색 방식을 보완하여 이용자의 정보 선별을 도와주고 검색의 정확률을 상승시킬 뿐 아니라 원문의 대용물이 될 수도 있다.

자동요약을 위한 초기의 연구들은 핵심문장 추출을 위한 기준으로서 단어의 출현빈도(Luhn 1958; Edmunson, and Wyllys 1961), 소재(Baxendale 1958), 단서어와 표제어(Edmunson 1969) 등 낮은 수준의 언어학적 방법을 주로 사용하였다. 이러한 기법은 오늘날에도 그 중요성을 인정받고 있지만 추출된 요약문은 문장과 문장간에 흐름이 단절되고 연결이 자연스럽지 못한 경우가 종종 있다.

텍스트를 구성하는 문장들간의 관계나 어휘간의 연결구조 등을 분석하여 텍스트를 새로운 구조로 표현한 후 핵심 구조에 대응하는 문장들만으로 요약문을 생성하는 연구들(Skorokhod'ko 1972; Taylor, and Krulee 1977; Salton et al. 1997; Morris, and Hirst 1991; Marcu 1996; Ono et al. 1994; 백혜승 1991)은 생성된 문장들 사이의 논리가 명확하며 주제 영역지식에 제약을 받지 않는다는 장점이 있으나, 문장들간의 관계 설정 방법이 명백하지 않고 표현양식이 단순한 텍스트에는 적용하기 어렵다는 단점이 있다.

다량의 학습데이터들로부터 요약문 추출시 필요한 정보를 학습한 후 이를 이용하여 요약문에 포함될 문장을 계산하는 말뭉치 기반 연구들(Kupiec, Pedersen, and Chen 1995; Hovy, and Lin 1977; 장동현 1997; 강상배 외 1997; 송인석, 박혁로 1997; 정영미, 최상희 2001)은 요약문을 일정 비율로 출력할 수 있다는 장점은 있으나, 출현빈도와 무관하게 결정되는 단어간의 의미적 관련성이나 문장간의 관계를 처리하기 어려운 점이 있다.

반면, 지식베이스에 포함된 상황정보를 이용하여 주제분야 텍스트를 분석하는 지식기반시스템은 문장들간의 연결을 원활히 하고 각 문장의 의미를 조합하여 텍스트의 총체적인 의미를 구성한다(Cullingford 1981; DeJong 1979; Paice, and Jones 1993; Mckeown, and Radev 1995; Rau 1987; Jacobs, and Rau 1990; 최인숙 1988). 이 부류의 시스템들은 해당분야의 지식과 문장의 문법적 구조를 기반으로 고품질의 자연스런 요약문을 작성하나 적용분야마다 각각 다른 영역지식이 필요하기 때문에 응용분야가 한정된다는 단점도 있다.

본 연구는 주제분야 지식을 스스로 수집, 표현하며 적절히 이용하여 요약문을 작성하는 시스템을 개발할 필요성에 따라 통계적 기법과 지식기반 기법을 결합하여 학습기능을 갖는 자동요약시스템을 제시하는 것을 목적으로 한다. 본 시스템의 특징은 텍스트에 출현한 단어의 빈도를 측정하는 통계적 기법을 이용하여 텍스트의 주요 내용을 암시하는 단서술어와 그 술어의 구문패턴정보를 자동으로 추출하고 요약문 작성에 이용함으로써 지식베이스 구축을 수작업에 의존하던 기존 시스템의 제한성을 극복할 수 있다는 점과, 의미 분석이나 완전한 구문 분석 등 복잡한 언어처리 과정을 거치지 않고 간단한 구문 분석기를 사용하여 기본적인 문장구조만 파

악해도 작업이 가능하다는 점이다.

실험집단(test collection)은 필수성분을 생략하는 표현이 드물고 주요 사건과 관련된 전형적인 정보항목들로 구성되는 신문기사로 정하고 '화재'와 관련된 기사를 국내 일간지의 사회면에서 선정하여 실험을 수행하였다. 그리고, 본 시스템의 기법이 타분야 텍스트에도 적용 가능한지 검토하고자 '강도'와 관련된 신문기사를 대상으로 검증실험을 하였다.

본 연구의 범위는 주제분야별로 예측 가능한 사건의 대표술어를 중심으로 사건별 정보항목들을 내포하는 문장을 생성함으로써 주제분야 지식베이스를 수작업으로 준비하지 않고도 요약문을 작성할 수 있는 가능성을 실험하는 것에 제한된다. 따라서, 술어정보 추출 작업의 기초가 되는 형태소 분석 및 구문 분석을 위해서는 이미 개발된 한국어 분석 모듈(Hangeul Analysis Module : HAM)(강승식 1999)을 활용하였다.

2 텍스트 요약을 위한 한국어 처리

텍스트의 내용을 이해하고 요약하기 위한 자연언어처리에 여러 가지 문법이 이용되고 있는데, 문장 내에서 구성 요소들이 위치에 따라 기능을 할당받는 영어, 불어와 같은 형상적 언어와는 달리 한국어, 일본어 같이 구성요소들이 조사에 의해 그 기능이 표시되며 문장 내 위치가 비교적 자유로운 비형상적 언어는 격문법으로 설명하기 적합하다고 할 수 있다. 특히, 한국어는 문장 내에서 술어가 중심적인 역할을 하므로, 술어와 필수격을 포함하는 축약된 문장만으로도 텍스트의 내용을 요약적으로 나타낼 수 있어 격문법을 활용한 한국어 처리는 매우 효과적이라 할 수 있다.

그러나, 언어학적인 지식만으로는 중의성이나 주제분야의 확장성 문제를 해결하기 어려운 한계에 부딪치게 되어 최근에는 말뭉치(corpus)라고 하는 대량의 언어자료로부터 언어정보를 획득하여 문제를 해결하고자 하는 시도가 생겨났다. 말뭉치에는 단어나 어절의 사용방법, 사용례가 담겨져 있으므로 어휘간의 관계정보 등을 자동으로 습득하여 자연언어처리시스템에 다양하게 활용할 수 있다.

따라서, 한국어와 같은 격언어를 컴퓨터로 분석하고 처리하기 위해서는 격문법을 활용하는 것이 유리하며 말뭉치에 기반한 통계적 방법론으로 보완하는 것은 매우 효율적이라고 할 수 있다.

한국어의 어순법칙에서 정상적인 어순에서는 종속되는 단어들이 언제나 주도적 단어의 앞에 온다는 규칙이 있다. 술어는 정상적인 어순에서는 언제나 문장의 제일 마지막에 위치하게 되는 지배적인 구성요소로서 술어의 지배를 받는 구성요소가 어떤 문법적 형태를 취해야 하는가, 어떤 구조적 단위가 되어야 하는가를 결정한다. 따라서, 한국어 기본문형의 설정은 반드시 술어를 중심으로 술어가 되는 단어들의 의미론적 특성을 중요하게 고려해야 한다(강은국 1994, 9-35).

술어와 문장성분들과의 관계구조인 기본문형은 술어의 종류에 따라 그 서술용언에

이끌리어 반드시 나타나야만 문장이 될 수 있는 최소한의 문장성분으로 이루어진 구조, 즉 격 프레임이다. 격 프레임을 자동으로 구축하기 위한 연구들은 주로 수작업 혹은 구문 분석기의 결과로 생성된 구문 분석정보가 부가된 말뭉치를 학습데이터로 사용하고 있다(송재관, 홍성웅, 박찬곤 1996; 류법모 외 1997; 김선호 1996; 윤준태 1997; 정후중 외 1998; 이휘봉, 강인수, 이종혁 1998).

본 연구에서는 특정한 주제분야의 텍스트로 구성된 학습데이터로부터 지배-의존관계를 갖는 〈술어-격조사〉쌍에 대한 공기정보를 추출하여 격 프레임 구성에 이용하고자 하며 수 개의 격 프레임은 모여서 특정 주제분야의 전형적 정보를 나타낼 수 있게 되는 것이다.

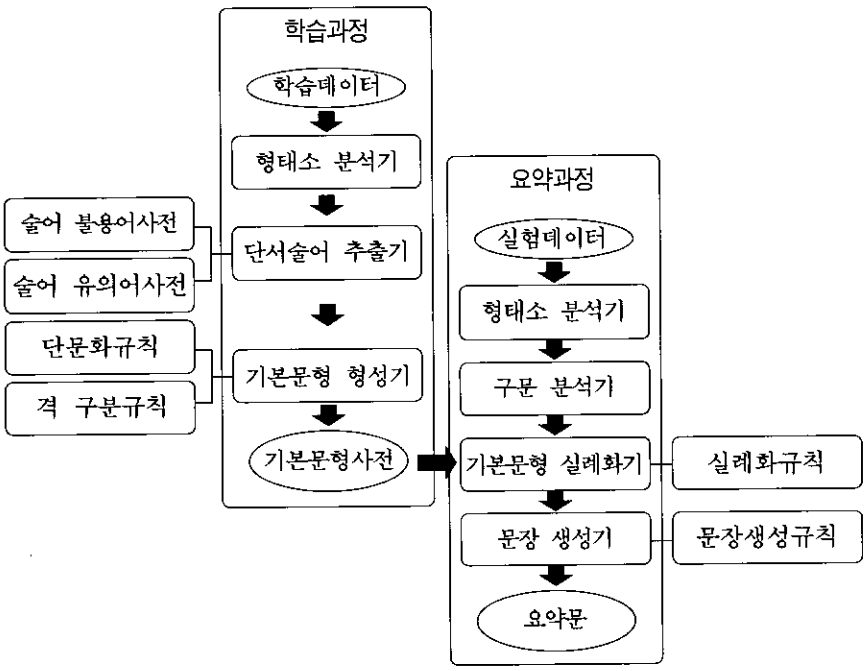
3 자동요약시스템 설계

3.1 시스템 개요

본 시스템은 〈그림 1〉의 시스템 구성도에서와 같이 학습과정과 요약과정의 두 부분으로 구분된다. 학습과정은 단서술어 추출기와 기본문형 형성기로 구성되며, 요약과정은 기본문형 실례화기와 문장 생성기로 구성되어 있다.

단서술어 추출기는 주제분야 텍스트집단을 대표하는 단서술어를 선정하기 위한 모듈이며 기본문형 형성기는 단서술어가 이루는 문장의 기본 틀을 추출하는 모듈이다.

기본문형 실례화기는 기본문형의 격조사와 결합될 체언, 즉 필수논항을 찾는 모듈이



〈그림 1〉 시스템 구성도

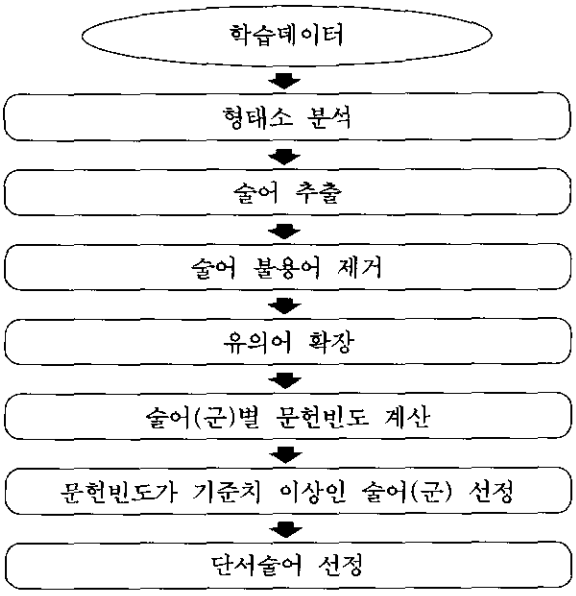
며 문장 생성기는 기본문형 실례화를 통해 생성된 결과, 즉 술어와 필수성분들을 포함하는 축약된 단문들을 연결하여 문장을 생성한 후 요약문을 출력하는 모듈이다.

3.2 학습과정

3.2.1 형태소 분석 및 단서술어 추출

텍스트로부터 술어를 추출하기 위해서는 각 어절의 기본단위를 분류하는 형태소 분석이 선행되어야 한다. 술어는 주체의 행위, 상태, 성질 등에 관하여 설명하는 말로서, HAM의 형태소 분석 결과에 따라 추출되었다.

형용사 술어는 하나의 논항, 즉 주어만을 취하는 것이 보통이고 주체의 상태나 성질을 나타낼 뿐 특정한 사건의 단서가 될 만한 행위를 나타내지 않으므로 본 시스템의 술어 추출 대상에서 제외하였고 주제분야와 무관하게 공통적으로 사용되는 고빈도 술어들은 술어 불용어사전에 등록하여 술어 추출과정에서 제외되도록 하였다. 술어의 유의어가 많은 경우 다양한 술어가 사용되므로 추출된 술어 각각의 빈도는 상대적으로 낮은 빈도를 보일 수 있다. 따라서, 술어 유의어사전을 참조하여 형태는 다르지만 뜻이 비슷한 유의어, 어근에 파생접미사인 '-되다, -하다, -이다'가 붙어서 이루어진 파생어, 어근에 ' , ' 가 붙어 이루어진 축약형, 주어의 직접 동작을 나타내는 동사어간에 피동과 사동의 파생접미사 '-이-, -히-, -리-, -기-, -우-, -구-, -추-'가 붙어 이루어진 파생어들은 통합하여 빈도를 계산하였다. 술어 유의어사전에는 학습데이터에 직접 출현하지 않은 술어들도 포함되어 있다.



〈그림 2〉 단서술어 추출과정

주제영역의 핵심적인 내용을 나타내는 술어들은 대부분의 주제 관련 텍스트에 출현할 것이다. 따라서, 술어는 다수의 텍스트에 출현할수록 높은 중요도를 갖는다고 보았으며 술어의 문헌빈도를 중요도 가중치로 부여하였다.

추출된 술어는 술어 유의어사전을 참조하여 확장한다. 유의어를 가진 술어는 술어군으로 확장되고, 유의어가 없는 술어는 단일술어로 남게 된다. 문헌빈도는 단일 술어나 술어군(이하 술어(군)으로 표기)을 하나의 단위로 하여 계산한다.

단서술어 추출과정은 〈그림 2〉와 같다.

3.2.2 기본문형 추출

기본문형은 주어, 술어, 목적어, 보어 등의 필수적 성분으로 이루어진 문장의 기본 틀로서, 하나의 술어가 어떤 문법적인 격을 얼마나 요구하는지를 반영하며 술어가 되는 동사의 어휘적 특성에 따라 필요로 하는 성분의 수효가 다르다.

본 시스템에서는 문법적인 필수성분보다는 텍스트의 내용 이해를 위해 필요한 요소들이 포함되도록 하였다. 기본문형 추출과정은 〈그림 3〉과 같다.

각각의 단서술어를 대상으로 〈단서술어-격조사〉쌍의 출현빈도를 계산하려면 대규모의 학습데이터가 필요하다. 그러나, 의미가 유사한 술어들은 대부분 기본문형도 유사하므로 본 시스템에서는 각각의 단서술어를 대상으로 기본문형을 추출하는 대신 술어(군)을 대표하는 대표술어의 문형을 추출하였고, 각 단서술어의 기본문형은 대표술어의 기본문형을 그대로 이용하게 하였다. 대표술어는 술어(군)에서 최고의 장서빈도를 갖는 술어로 하였다.

행위자격이나 목적격을 필요로 하는 술어

〈그림 3〉 기본문형 추출과정

의 경우 태(態, voice)에 따라 기본문형이 달라질 수 있으므로, 술어 유의어사전의 태정보를 참조하여 술어(군)을 두 종류로 구분한 후 각 그룹에서 최고의 장서빈도를 갖는 술어를 각각 대표술어로 선정하였다.

〈대표술어-격조사〉쌍의 출현빈도를 계산하기 위해서는 우선 학습데이터에서 대표술어를 포함하는 문장들을 추출하고 추출된 문장 중에서 지배-의존관계를 갖는 〈대표술어-격조사〉쌍을 추출하여야 한다. 본 시스템에서는 술어의 지배 범위를 단순화하기 위해 텍스트를 형태소분석한 후 하나의 술어를 갖는 단문을 단위로 하여 문장성분을 파악하는 방법을 채택하였다. 이는 기본문형 추출을 위한 정보의 양이 적더라도 지배-의존관계를 갖는 것이 확실한 〈대표술어-격조사〉쌍만 추출함으로써 잘못된 정보가 포함될 수 있는 가능성을 배제하려는 의도이다. 정확한 지배-의존관계를 갖지 않는 경우도 있을 수 있지만 대량의 올바른 데이터와 비교할 때 오류의 비율은 미미할 것으로 판단되었다.

제한된 주제분야 내의 텍스트에서 나타나는 술어이므로 의미의 갈래를 구별하지 않았으며, 논항의 자리에 오는 말이 어떤 경우에 어떤 성분과 주로 어울려 쓰이는지를 설명하는 참고정보도 생략하였다.

격문법을 이용한 시스템에서는 일반적으로 시스템이 처리하는 주제영역과 응용분야에 따라 다른 격범주가 설정되고 있다. 신문기사를 작성하려면 짧은 사건 보도의 경우라도 누가, 무엇을, 언제, 어디서, 왜, 어떻게 했다는 몇 가지 요건에 대한 자료를 갖추어야 한다. 따라서, 격조사는 신문기사 작성의 기준이 되는 6하 원칙과 학습데이터 내 출현빈도를 고려하여 행위자격, 목적격, 시간격, 처소격, 경과격, '의'격, '으로'격, '에'격 등으로 구분하였으며, 동일한 기능의 조사들은 통합한 후 빈도를 계산하였고 출현빈도가 낮은 조사들로서 이상의 8가지 범주에 속하지 않는 것들은 필수격 선정에 영향을 미치지 못하므로 무시하였다. 격을 예측할 수 없는 '은, 는, 도, 만' 등의 보조사는 모호성을 야기하므로 제외하였다.

대표술어의 필수격 산출의 기준이 되는 중요도 가중치는 〈대표술어-격조사〉쌍의 출현빈도로 하였다. 필수격조사는 대표술어 출현시 대부분 함께 나타날 것이므로 〈대표술어-격조사〉쌍의 출현빈도가 높을수록 필수성분이 될 가능성이 높다고 보아 출현빈도가 일정 기준치 이상인 것들을 필수격으로 선정하

〈표 1〉 단서술어의 기본문형사전 예

대표술어	단서술어	기본문형
나	나	나 - 처소격 - 행위자격
	발생하	발생하 - 처소격 - 행위자격
	일어나	일어나 - 처소격 - 행위자격
	일	일 - 처소격 - 행위자격

였다.

선정된 〈대표술어-격조사〉쌍들을 조합하여 대표술어의 기본문형을 만들었고, 대표술어가 소속된 술어(군) 내 모든 단서술어들의 기본문형으로 삼아 기본문형사전을 구성하였다.

추출된 기본문형으로부터 구성된 기본문형사전은 〈표 1〉과 같다.

3.3 요약과정

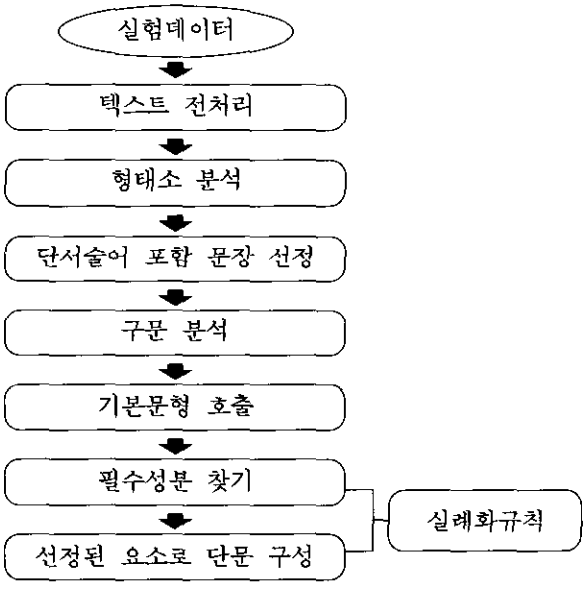
3.3.1 기본문형 실례화

기본문형 실례화는 기본문형의 격조사와 결합될 체언, 즉 필수논항을 찾는 과정이다. 텍스트에서 단서술어를 포함하는 문장들을 선별한 후 기본문형사전에서 단서술어별 기본문형을 참조하여 격조사와 결합될 요소를 찾는 것으로, 〈단서술어-격조사〉쌍을 실례화하면 단서술어는 문장의 필수성분들과 결합하여 단문을 이루게 된다.

본 모듈에서는 구문분석이 불필요한 부분과 구문분석 오류가 발생하는 부분에 대해 전처리한 후 형태소분석하여 단서술어가 포함된 문장을 선정하고, 구문분석한 결과에 실례화규칙을 적용하여 기본문형의 격조사와 결합되는 체언을 찾았으며, 체언의 관계정보를 보충하여 문맥을 자연스럽게 하기 위해 체언 앞의 어절까지 포함시켰다.

기본문형의 실례화과정은 〈그림 4〉와 같다.

실례화규칙은 단서술어의 필수성분을 선정할 뿐만 아니라 전후관계 이해에 도움이 되는 어절을 추가하여 선정함으로써 추출된 어절간의 관계정보를 보충하고 문맥이 자연스러운 문장을 만들기 위한 것이다. 경험적 지식을 기초로 정보추출규칙과 단문생성규칙으로 정의하였으며, 두 차례의 실험을 통해



〈그림 4〉 기본문형 실례화과정

규칙의 효율성을 알아보았다. 즉, 1차실험에는 최초로 규정한 정보추출규칙과 단문생성규칙을 적용하였으며, 2차실험에는 1차실험결과를 개선하기 위해 수정한 단문생성규칙을 적용하였다.

(1) 정보추출규칙

① 단서술어의 의존소 중 기본문형사전에 정의된 필수성분만을 선택한다. 필수성분인지의 판단은 필수격조사를 포함하고 있는지를 근거로 한다.

② 필수성분의 관계정보를 보충하는 앞의 어절이 필요하다고 보아 필수성분의 의존소를 모두 추가한다.

③ 단서술어의 의존소 중 기본문형이 요구하는 필수성분이 없는 경우, 단서술어와 의존관계에 있는 술어가 있다면 필수성분 대신 선택하고 그 술어의 의존소를 모두 추가한다.

④ 시간을 나타내는 어구인 '-일(또는 '시') … 명사 + 시간격 조사'와 장소를 나타내는 어구인 '-도(또는 '시') … 명사 + 처소격 조사'는 신문기사의 특성상 중요하다고 간주되므로 따로 추출해 두었다가, 해당 문장에서 선정된 다른 성분들과 함께 포함시킨다. 이를 위해 국명, 도명, 시명 등 사고발생 지역을 나타낼 수 있도록 분류한 지명사전을 준비하였다. 그러나, 해당 문장에서 연결될 술어가 선정되지 않은 경우 무시한다.

(2) 단문생성규칙

① 선정된 어절들은 각각의 단서술어를 중심으로 '체언 + 격조사 + … + 체언 + 격조사 + 단서술어' 형태의 단문을 생성하되, 원문에 출현한 형태와 순서대로 제시한다.

② 단서술어를 중심으로 생성된 단문에서 선두요소가 의존명사이거나 술어일 때, 의존명사나 술어의 관계정보를 보충하기 위해 앞의 어절 한 개를 추가해 주어 자연스러운 문맥을 만들어준다. 의존명사는 관형어의 선행을 필수조건으로 하는 명사로서 의존명사 앞에 있는 관형어는 필수적 성분의 기능을 하기 때문이며, 술어는 홀로 두면 문장이 부자연스럽기 때문이다.

③ 필수성분이나 의존관계의 술어가 하나도 선정되지 않아 단서술어만 있게 된 경우 문맥상 자연스럽지 못하므로 제거해준다. 다만, 앞에 위치한 단문의 마지막 어절과 연속되어 자연스럽게 연결되는 경우만 그대로 둔다.

(3) 수정된 단문생성규칙

① 단서술어를 중심으로 생성된 단문의 선두요소가 술어일 때 앞의 어절 한 개를 추가하는 대신 해당 술어를 제거해준다.

② 선정된 어절 중 의존명사가 있으면 앞의 어절 두 개를 추가해준다. 앞의 어절이 술어인 경우 술어의 의존소까지 포함하여 최소한 두개의 어절이 추가되어야 문맥이 자연스럽기 때문이다.

실례화의 예를 들면 다음과 같다.

〈원문〉

목포 문태고에 불…교실 등 20칸 전소

뉴 스 명 : 경향신문
등 록 일 : 1998/11/06
등록시간 : 08:17:28
크 기 : 888B

6일 오전4시28분께 전남 목포시 용당1동 문태고등학교에서 화재가 발생해 교무실과 교실을 태우고 출동한 소방관 등에 의해 1시간여만에 진화됐다.

1층 교무실, 화학실과 2층 교실 12칸 등 모두 20칸 총면적 2천여㎡가 불에 타 소방서 추산 1억2천만원 상당의 피해를 냈으나 학생들이 등교하지 않은 새벽이여서 다행히 인명 피해는 없었다.

소방관과 교사들이 생활기록부 등 중요한 서류는 모두 빼냈다.

정동민씨(35.교사)는 『당직 근무중 화재 경보기가 울려 나가보니 본관2층 통로에서 불길이 솟아오르면서 통로 부분이 붕괴되고 있어 신고했다』고 말했다.

목포소방서는 전기합선에 의해 불이 난 것으로 보고 정확한 화인을 조사중이다.

한편 문태고교측은 중학교와 고등학교에 여유교실11개가 있어 수업에는 지장이 없다고 밝혔다.

◐ 경향신문 11월6일자 ◑

〈선별된 문장〉

6일오전4시28분께 전남 목포시 용당1동 문태고등학교에서 화재가 발생해 교무실과 교실을 태우고 출동한 소방관 등에 의해 1시간여만에 진화됐다.

1층 교무실 화학실과 2층 교실 12칸 등 모두 20칸 총면적 2천여가 불에 타 소방서 추산 1억2천만원 상당의 피해를 냈으나 학생들이 등교하지 않은 새벽이여서 다행히 인명 피해는 없었다.

목포소방서는 전기합선에 의해 불이 난 것으로 보고 정확한 화인을 조사하고 있다.

두 번째 문장을 구문분석한 예에서 단서술어 '타, 냈으나'의 필수성분을 찾고 필수성분의 관계정보를 보충하는 어절(이탤릭체)을 다음과 같이 찾아내었다.

타: 1층 교무실 화학실과 2층 교실 12칸 등 모두 20칸 총면적 2천여가. 불에 냈으나: 소방서 추산 1억2천만원 상당의 피해를

parse: 1층 교무실 화학실과 2층 교실 12칸 등 모두 20칸 총면적 2천여가 불에 타 소방서 추산 1억2천만원 상당의 피해를 냈으나 학생들이 등교하지 않은 새벽이여서 다행히 인명 피해는 없었다.

```
없[F]                    없/V 었/f 다/e
 인명 피해[U]            피해/N 은/j
 다행히[A]               다행히/Z
 새벽이[V]               새벽이/N 이/c 어서/e
    않[K]                않/V 은/e
      등교[V]            등교/N 하/t 지/e
        학생[S]          학생/N 들/s 이/j
          내[V]          내/V 었/f 으나/e                              (단서술어)
            피해[O]      피해/N 을/j                    ('냈으나'의 의존소로서 필수성분)
              소방서 추산 1억2천만원 상당[G] 상당/N 의/j   ('피해를'의 의존소로서…관계정보)
            타[V]        타/V 어/e(단서술어)
              불[U]      불/N 에/j                      ('타'의 의존소로서 필수성분)
              등 모두 20칸 총면적 2천여[S] 2천여/N 이/j   ('타'의 의존소로서 필수성분)
                2층 교실 12칸[N]        12칸/N    ('등 모두 20칸…'의 의존소로서…관계정보)
                1층 교무실 화학실[&]    화학실/N 과/j('등 모두 20칸…'의 의존소로서…관계정보)
...
```

선정된 요소들을 조합하여 아래와 같은 단문을 생성하였다.

...

1층 교무실 화학실과 2층 교실 12칸 등 모두 20칸 총 면적 2천여가 불에 타

소방서 추산 1억2천만원 상당의 피해를 냈으나

...

3.3.2 문장 생성

기본문형 실례화를 통해 형성된 결과는 단서술어와 필수성분 외에 관계정보 표시 어절들까지 포함하는 단문으로서, 문장 생성기는 이들 개별 단문들을 연결하여 문장을 생성하고 요약문을 출력한다.

(1) 문장 생성과정

선정된 요소들은 원문에서의 각 문장 단위로 모았다. 원문 중의 한 문장이 하나의 단서술어를 포함하고 있었다면 실례화된 하나의 기본문형, 즉 하나의 단문이 생성될 것이고 여러 개의 단서술어를 포함하고 있었다면 실례화된 여러 개의 기본문형들, 즉 여러 개의 단문들이 생성될 것이다. 각각의 단문은 원래의 문장에서 술어가 출현했던 순서대로 연결되어 하나의 문장을 생성하게 된다.

생성된 문장 중 4어절 이내의 짧은 것은 제거한다. '화재가 발생했다. 가능성이 큰 것으로 보았다' 와 같이 특별한 내용이 없는 형식적인 문장이 대부분이었기 때문이다.

문장에 포함된 단서술어는 원문 중에 나타났던 형태를 그대로 추출한 것이므로 문맥상 자연스럽지 못한 부분이 있을 수도 있다. 따라서, 단서술어의 어미를 수정하기 위한 규칙을 적용하였다.

어미수정규칙은 경험적 지식에 의해 다음과 같이 정의해 주었다.

① 문장의 후미에 오는 단서술어는 과거시제의 평서형 종결어미와 마침표를 갖도록 하였다. 기타 단서술어들은 원문에 나타난 형태대로 두었으나, 용언의 관형형으로 표시된 경우(K) 바로 뒤 어절과 연속하여 나타나지 않으면 연결형 어미를 갖도록 수정하였다.

② 단서술어 '의하', '인하'는 이유를 의미하며 그 이유는 추정된 상태가 대부분이다. 따라서, 문장의 후미에서는 '의한 것으로 보았다./인한 것으로 보았다.'로 변형시켰으며, 나머지는 원문에 나타난 형태대로 두었다.

어미수정규칙을 적용한 후 전처리과정에서 보관해둔 요소들을 복귀시키고 출력형식에 맞추어 요약문을 완성하였다.

〈 요약문의 예 〉

목포 문태고에 불…교실 등 20칸 전소(경향신문, 1998/11/06)

6일 오전4시28분께 전남 목포 용당1동 문태고등학교에서 화재가 발생해 교무실과 교실을 태우고 출동한 소방관 등에 의해 1시간여만에 진화됐다.

1층 교무실, 화학실과 2층 교실 12칸 등 모두 20칸 총면적 2천여㎡가 불에 타 소방서 추산 1억2천만원 상당의 피해를 냈다.

전기합선에 의해 난 것으로 보았다.

4 자동요약 실험

4.1 실험 개요

본 실험에서는 신문기사를 실험집단으로 정하고 '화재'에 관련된 기사를 국내 일간지의 사회면에서 선정하였다. 단서술어와 기본문형 추출을 위한 학습과정에는 211건의 기사를 학습데이터로 사용하였으며 요약과정에는 30건의 기사를 실험데이터로 사용하였다. 실험데이터로 사용된 기사 원문은 대부분 짧게 기술되고 있었는데, 3~10문장, 56~190어절로 구성되었으며 평균 문장 수는 4.7개, 평균 어절 수는 94개, 문장 당 어절 수는 평균 19.92개였다. 또한, 타주제분야 텍스트에의 적용가능성을 검증하고자 '강도'와 관련된 기사 210건과 30건을 각각 학습데이터와 실험데이터로 하여 추가실험을 하였다.

성능평가는 학습과정과 요약과정에 대하여 각각 실시하였다. 시스템이 추출한 결과를 수작업으로 생성한 결과와 비교하는 방식으로 학습과정에서는 단서술어와 기본문형정보를 누락하지 않고 찾아내었는지 평가하였고 요약과정에서는 이용자가 요구하는 정보를 포함한 논항을 제대로 찾았는지, 작성된 요약문의 구문적/의미적 오류는 없는지 평가하였다. 평가자는 본 연구자와 석사학위과정에 재학중인 대학원생 1인으로 하였다.

자동요약시스템은 Unix OS 상에서 구현하였으며, 단서술어 추출기 및 기본문형 형성기는 C언어를 사용하여 개발하였고 기본문형 실례화기와 문장 생성기는 Perl을 사용하여 개발하였다.

4.2 학습과정 평가

4.2.1 단서술어 추출 결과와 평가

시스템이 선정한 술어(군)은 출현확률(술어의 문헌빈도÷텍스트 수)에 따라 다음의 세 범주로 구분하였다.

① 출현확률이 50% 이상인 술어: 문헌빈도 121~175회인 술어(군) 5종

② 출현확률이 20~49% 사이의 술어: 문헌빈도 43~74회인 술어(군) 3종

③ 출현확률이 10~19% 사이의 술어: 문헌빈도 25~40회인 술어(군) 11종

211건의 텍스트집단에서 시스템이 추출한 술어(군)의 종류는 총 1,070종이었으며, 출현확률이 10% 이상인 술어(군)은 19종으로 종수로 볼 때 전체 술어집단의 1.7%에 불과했으나 각각의 문헌빈도를 모두 누적한 빈도는 1,257회로서 전체술어의 문헌빈도 총합인 3,331회의 37%에 달하였다. 이는 출현확률이 10% 이상인 술어(군)의 종류는 많지 않지만 텍스트집단 내에서 사용되는 비율은 술어 2.7개 중 1회일 정도로 텍스트집단에 대한 대표성을 갖고 있음을 보여주는 결과이다.

성능평가척도는 추출재현율과 추출정확률을 사용하였으며 공식은 다음과 같다.

$$\text{추출재현율1} = \frac{\text{시스템이 추출한 적합한 단서술어(군)의 수}}{\text{수작업으로 선정한 단서술어(군)의 수}}$$

$$\text{추출정확률1} = \frac{\text{시스템이 추출한 적합한 단서술어(군)의 수}}{\text{시스템이 추출한 단서술어(군)의 수}}$$

성능평가 결과는 〈표 2〉에 제시하였다.

4.2.2 기본문형 추출 결과와 평가

시스템이 선정한 조사는 출현확률(대표술어가 출현한 단문 내 조사의 빈도 ÷ 대표술어가 출현한 단문의 수)이 50% 이상인 조사, 20~49% 사이의 조사, 10~19% 사이의 조사

들로 구분할 수 있었다.

시스템이 추출한 조사는 대표술어에 따라 1종에서 21종까지 다양했으며 그 중 출현확률이 10% 이상인 조사는 대표술어에 따라 1~3종이 선정되었다. 출현확률이 10% 이상인 조사의 누적빈도가 전체 조사의 누적빈도 중 차지하는 비율은 대표술어에 따라 49~92%였다. 누적빈도가 80% 이상인 경우의 대표술어는 4종, 70% 이상인 경우의 대표술어는 3종, 60% 이상인 경우의 대표술어는 8종, 50% 이상인 경우의 대표술어는 4종이었다.

이는 출현확률이 10% 이상인 조사가 종수는 많지 않지만 대표술어와 함께 사용되는 비율은 49~92%에 달하여 대표술어의 필수격이 될 자격이 있음을 보여주는 결과이다.

기본문형 형성기의 성능평가척도는 추출재현율과 추출정확률을 사용하였으며 공식은 다음과 같다.

$$\text{추출재현율2} = \frac{\text{시스템이 추출한 대표술어의 적합 필수격 수}}{\text{대표술어의 필수격 수}}$$

$$\text{추출정확률2} = \frac{\text{시스템이 추출한 대표술어의 적합 필수격 수}}{\text{시스템이 추출한 대표술어의 필수격 수}}$$

〈표 2〉는 단서술어와 필수격의 선정 기준치를 10%, 20%, 50%로 했을 때 단서술어 추출기와 기본문형 형성기의 성능을 평가한 것이다.

4.3 요약과정 평가

실험은 정보추출규칙과 단문생성규칙의 효율성을 알아보고자 두 차례에 걸쳐 수행되었다. 즉, 1차실험에는 최초로 규정한 정보추출규칙과 단문생성규칙을 적용하였으며, 2차실험에는 1차실험 결과를 개선하기 위해 수정한 단문생성규칙을 적용하였다.

4.3.1 기본문형 실례화 결과와 평가

단서술어와 필수격 선정 기준치를 모두 20%로 하여 1차 실험한 결과 작성된 요약문의 어절 수는 원문의 39.8%로서 평균 3.13개의 문장을 포함하고 있으며, 화재일시, 소재지, 진화일시(혹은 경과시간), 사망자, 부상자, 소실물, 피해액, 화재원인 등을 나타내는 전형적인 문장형식에 따라 표현되고 있었다.

2차 실험 결과 작성된 요약문의 어절 수는 원문의 36.00%로 1차실험에 비해 3.8%의 축소효과를 보였다. 요약문은 평균 3.0개의 문장을 포함하고 있으며, 1차실험에 비해 0.13개 줄어드는 효과를 보였다. 포함된 정보는 1차실험과 마찬가지로 화재일시, 소재지, 진화일시(혹은 경과시간), 사망자, 부상자, 소실물, 피해액, 화재원인 등이 전형적인 문장형식에 따라 표현되고 있었다.

기본문형 실례화 결과에 텍스트의 주요 특

〈표 2〉 단서술어 추출기와 기본문형 형성기의 성능평가

출현확률	추출재현율1	추출정확률1	추출재현율2(평균)	추출정확률2(평균)
10%	100%	42%	92%	64%
20%	100%	100%	92%	78%
50%	62%	100%	82%	91%

성들이 누락되지 않고 나타났는지 평가하기 위한 성능평가척도로 요약재현율과 요약정확률을 사용하였으며 공식은 다음과 같다.

$$\text{요약재현율} = \frac{\text{시스템이 적합하게 실례화한 정보항목수}}{\text{수작업 요약문의 정보항목수}}$$

$$\text{요약정확률} = \frac{\text{시스템이 적합하게 실례화한 정보항목수}}{\text{시스템이 실례화한 정보항목수}}$$

성능평가는 30개의 텍스트에 대한 평균치를 이용하였고 〈표 3〉에 제시하였다.

4.3.2 문장 생성 결과와 평가

단서술어와 필수격 선정 기준치를 모두 20%로 하여 1차 실험한 결과 작성된 요약문은 평균 3.13문장을 포함하고 있었으며 그 중 구문적/의미적으로 오류가 있는 문장은 평가자에 따라 평균 0.73문장, 평균 0.93문장이었다.

2차 실험 결과 작성된 요약문은 평균 3.0문장을 포함하고 있었으며 그 중 구문적/의미적으로 오류가 있는 문장은 평가자에 따라 평균 0.53문장, 평균 0.67문장이었다.

요약문 작성에 필요한 주요 정보가 포함되지 않은 문장, 구문적 오류나 문법적 오류가 있는 문장, 내용이 중복되는 문장을 점검하였으며 각각의 요약문으로부터 산출된 잡음율의 평균치를 구하는 공식은 다음과 같다.

$$\text{잡음율} = \frac{1}{m}\sum_{i=1}^{m}\frac{\text{오류있는 문장수}}{\text{시스템이 생성한 문장수}}$$

(m: 텍스트의 수)

〈표 3〉은 단서술어와 필수격의 선정 기준치, 적용된 실례화규칙에 따라 요약과정의 성능을 평가한 것이다.

실험 결과 축약률이 높아질수록 요약재현율과 요약정확률은 낮아지고 잡음율은 높아지는 것으로 밝혀졌다. 따라서, 이용 목적에 맞추어 적절한 요약문을 선택적으로 제공할 수 있는 것으로 드러났다.

본 시스템에서 작성한 요약문은 원문의 크기에 따라 일정 비율로 축소된 형태가 아니라 포함되는 정보항목 수에 따라 비교적 일정한 길이를 유지하고 있다. 요약문에 포함되는 정보항목과 압축률에 영향을 미치는 요소는 단서술어와 필수격의 선정 기준치인 출현확률이다. 선정 기준치를 높게 잡을 경우 전형적인 단서술어와 그의 필수논항만을 포함하는 간략한 요약문이 생성되나 이때 누락되는 정보항목이 생기게 된다. 선정 기준치 조정에 따른 성능변화를 측정한 결과, 선정 기준치를 20%-20%에서 50%-50%로 변화시킴에 따라 요약문의 크기는 4.8% 축소효과를 나타낸 반면 누락된 정보항목으로 인해 재현율은 약 11%정도 하락하고 잡음율은 약 3%

〈표 3〉 시스템의 성능평가

선정 기준치		1차 실험				2차 실험			
단서 술어	필수격	요약문의 크기	요약 재현율	요약 정확률	문장생성 잡음율	요약문의 크기	요약 재현율	요약 정확률	문장생성 잡음율
20%	20%	39.80%	93.53%	96.00%	21.47%	36.00%	94.07%	96.72%	15.55%
20%	50%	38.70%	91.00%	96.00%	21.47%	34.90%	91.00%	96.00%	15.55%
50%	50%	35.00%	82.60%	95.20%	24.77%	31.20%	83.00%	96.00%	18.85%

상승하였다.

압축률에 영향을 미치는 또 한가지 요소는 실례화규칙이다. 실례화규칙 중 단문생성규칙의 수정에 따른 성능변화를 측정한 결과 수정된 실례화규칙으로 인해 요약문은 3.8% 축소되는 효과를 보였고, 잡음율이 5.92% 하락되었으며 추출된 정보항목의 성능에는 큰 영향이 미치지 않은 것으로 드러났다.

본 시스템은 주제와 형식이 제한된 텍스트를 요약하기 위한 것이므로 방대한 분량의 학습데이터를 이용하지 않아도 단서술어 추출과 기본문형 추출이 가능할 것으로 예상하였다. 그러나, 실험에서 사용한 학습데이터의 크기는 단서술어 추출에는 아무 문제가 없었으나 기본문형 추출을 위한 목적으로는 규모가 작았던 것으로 드러났다. 출현빈도가 39회 이상인 대표술어의 경우 필수격 추출에 큰 문제가 없었으나, '보이, 끄, 입히' 등의 술어는 출현빈도가 19회, 7회, 3회에 불과하여 올바른 필수격조사를 찾아내기 어려웠다.

시스템의 성능 향상 방안을 모색하기 위해서 요약문에 정보항목이 누락된 경우와 구문적/의미적으로 오류가 있는 문장들을 대상으로 그 원인을 분석한 결과 전형적 표현방식을 벗어난 문장은 통계적 방법만으로 학습한 단서술어와 기본문형으로 처리하기 어려웠으며, 구문 분석기가 술어와 의존소간의 관계분석을 잘못한 경우 본 시스템에서도 오류를 유발할 수밖에 없었고 언어학적인 처리와 주제분야의 특성을 고려한 규칙이 더 필요한 것으로 판단되었다.

'강도' 분야의 텍스트를 대상으로 한 추가실험 결과 학습과정에서 추출된 단서술어와 기본문형사전이 불완전했고, 문장이 길면서 핵심정보 외의 어절이 많이 사용되었기 때문에 정확한 정보를 이끌어내기 어려워 성능이 다소 낮게 나타나기는 했지만, 본 시스템의 기법은 '화재' 분야 텍스트 외의 타분야 텍스트에도 적용가능한 것으로 밝혀졌다. 학습데이터의 크기를 조절하여 보다 완전한 기본문형사전을 구성하면 요약문의 성능은 훨씬 향상될 것이다.

5 결 론

본 연구에서는 지식베이스 구축의 부담을 덜고 주제영역의 확장을 용이하게 하기 위하여 학습기능을 갖는 자동요약시스템을 제시하였다. 주제분야 지식을 스스로 수집하고 표현하며 이 지식을 적절히 이용하여 요약문을 작성하는 시스템은 지식기반 접근방식의 장점을 살리면서 사전작업이라는 부담을 덜어줄 수 있는 대안이 될 것이다. 또한, 주제영역의 확장이 쉬워지며 보다 많은 양의 정보를 처리하는 효과를 얻게 된다.

본 연구를 통해 밝혀진 사실은 다음과 같다.

첫째, 술어의 출현확률을 이용한 통계적 방법에 의거하여 주제분야를 대표하는 단서술어를 효과적으로 추출할 수 있었으며 단서술어 선정 방법을 체계화할 수 있었다. 그리고, 단서술어 선정 기준치를 조정함으로써 요약문의 크기와 정보항목의 요약재현율, 문장의 구문적/의미적 잡음율을 변화시킬 수 있는 것으로 드러났다.

둘째, 조사의 출현확률을 이용한 통계적 방법에 의거하여 문장의 주요 성분을 파악할 수

있었으며 기본문형 구성 방법을 체계화할 수 있었다. 그리고, 필수격 선정 기준치를 조정함으로써 요약문의 크기와 정보항목의 요약재현율을 변화시킬 수 있는 것으로 드러났다.

셋째, 구문 분석 결과에 경험적 지식에 의거하여 정의한 실례화규칙을 적용함으로써 단서술어의 필수논항을 파악할 수 있었으며 기본문형 실례화 방법을 체계화할 수 있었다. 실례화규칙을 조정함으로써 요약문의 크기, 잡음율을 변화시킬 수 있었으며, 추출된 정보항목의 성능에는 큰 영향이 미치지 않은 것으로 드러났다.

넷째, 축약률이 높아질수록 요약재현율과 요약정확률은 낮아지고 잡음율은 높아지는 것으로 밝혀졌다. 따라서, 이용 목적에 맞추어 적절한 요약문을 선택적으로 제공할 수 있는 것으로 드러났다.

다섯째, 기본문형 실례화 결과 생성된 단문들을 연결하여 문장을 생성할 때 어미수정규칙을 적용하여 대부분의 문장을 문맥이 어색하지 않게 수정하였으며 자연스러운 요약문을 작성할 수 있었다.

본 실험의 결과는 중요문장을 발췌하여 나열하는 요약방식이나 템플릿을 형성한 뒤 키워드를 추출하여 삽입하는 방식과 큰 차이를 나타내고 있었다. 작성된 요약문은 단서술어가 포함된 주요 문장에서 추출한 필수 정보항목과 술어를 중심으로 생성된 문장들로서 문장간의 연결이 자연스러울 뿐 아니라 텍스트의 전체적인 의미를 표현할 수 있었다. 또한, 통계적 기법을 이용한 학습을 통해 주제영역의 확장이 가능하였다. 타분야에의 적용성 검토를 통해 보완할 사항을 일부 발견하였으며 이는 보다 다양한 영역의 후속연구를 통해 망라적으로 정리되어야 할 것이다.

주제와 형식이 제한된 텍스트를 요약하기 위한 본 실험에서 방대한 분량의 학습데이터를 이용하지 않아도 단서술어 추출이 가능함을 알 수 있었다. 그러나, 기본문형 추출을 위한 학습데이터는 더 큰 규모로 확대할 필요가 있는 것으로 나타났다. 학습데이터의 크기를 조절하여 보다 완전한 기본문형사전을 구성하면 요약문에 포함되는 정보항목의 성능은 훨씬 향상될 것이다.

본 연구에서 제시한 자동요약기법을 보다 일반적인 분야로 확대 적용할 수 있기 위해서는 다양한 형식과 주제분야에서 다양한 규모의 학습데이터를 이용한 후속연구가 계속되어야 할 것이며, 본 연구의 결과를 토대로 제안할 수 있는 성능 향상 방안은 다음과 같다.

첫째, 단서술어와 필수격을 보다 정확하고 망라적으로 준비하기 위해 단순한 빈도의존 방법 이외의 기법을 추가하는 것이 바람직하다.

둘째, 단서술어의 의미구분, 사용상의 특성을 정의하여 단서술어사전을 보완할 뿐 아니라, 필수성분의 의미제한을 둘 수 있도록 기본문형사전을 보완할 필요가 있다.

셋째, 분야별로 고유한 표현방식이나 언어학적, 형태적인 특성을 고려한 규칙을 추가하여 보다 정확하고 자연스러운 문장을 생성하는 것이 바람직하다.

참 고 문 헌

강상배, 조혁규, 권혁철, 박재득, 박동인. 1997. 한국어문서의 통계적 정보를 이용한 문서요약시스템구현. 『제9회 한글 및 한국어 정보처리 학술대회 발표자료집』: 28-33.

강승식. 1999. HAM: 한국어 분석 모듈. 〈http://ham.hansung.ac.kr/〉.

강은국. 1993. 『조선어 문형 연구』. 서울: 서광학술자료사.

김선호. 1996. 『통계정보를 기반으로 한 어휘 관계 예측』. 석사학위논문, 연세대학교 대학원.

류법모, 장명길, 박수준, 박재득, 박동인. 1997. 구문구조부착 말뭉치를 이용한 술어의 하위범주화 정보 구축. 『제9회 한글 및 한국어 정보처리 학술대회 발표자료집』: 116-121.

백혜승. 1991. 한국어문서축약시스템의 설계. 『제3회 한글 및 한국어 정보처리 학술대회 발표자료집』: 238-246.

송인석, 박혁로. 1997. 텍스트이해모델에 기반한 정보검색시스템. 『제9회 한글 및 한국어 정보처리 학술대회 발표자료집』: 1-6.

송재관, 홍성웅, 박찬곤. 1996. 기계 번역을 위한 한국어 문장 패턴에 관한 연구. 『제8회 한글 및 한국어 정보처리 학술대회 발표자료집』: 308-312.

윤준태. 1997. 『공기 관계 기반 어휘 연관도를 이용한 한국어 구문 분석』. 박사학위논문, 연세대학교 대학원.

이휘봉, 강인수, 이종혁. 1998. 개념패턴과 통계정보를 이용한 한국어 미지격의 구문관계 결정 방법. 『제10회 한글 및 한국어 정보처리 학술대회 발표자료집』: 261-266.

장동현. 1997. 『효과적인 정보제시를 위한 문서요약시스템의 개발』. 석사학위논문, 충남대학교 대학원.

정영미, 최상희. 2001. 문장 클러스터링에 기반한 자동요약 모형. 『정보관리학회지』, 18(3): 159-177.

정후중, 황영숙, 곽용재, 박소영, 임해창. 1998. 구문 분석에서의 중의성 해소를 위한 일반화된 어휘정보의 자동 구축 및 적용. 『제10회 한글 및 한국어 정보처리 학술대회 발표자료집』: 269-275.

최인숙. 1988. 『자동초록을 위한 지식기반시스템 설계에 관한 연구』. 석사학위논문, 연세대학교 대학원.

Baxendale, P. B. 1958. "Machine-Made Index for Technical Literature-An Experiment." *IBM J. of Research and Development*, 2(4): 354-361.

Cullingford, R. 1988. "SAM." In *Inside Computer Understanding*. Edited by Schank, R. C., and C. K. Riesbeck. Hillsdale: LEA Publishers.

DeJong, G. F. 1979. "Prediction and Substantiation: A New Approach to

Natural Language Processing." *Cognitive Science*, 3: 251-273.

Edmunson, H. P. 1969. "New Methods in Automatic Extracting." *J. of the ACM*, 16(2): 264-289.

Edmunson, H. P., and R. E. Wyllys. 1961. "Automatic Abstracting and Indexing - Survey and Recommendation." *Communications of the ACM*, 4(5): 226-234.

Hovy, E., and C. Y. Lin. 1997. "Automated Text Summarization in SUMMARIST." In *Proceedings of a Workshop on Intelligent Scalable Text Summarization*: 18-24.

Jacobs, P. S., and L. F. Rau. 1990. "SCISOR: Extracting Information from On-line News." *Communications of the ACM*, 33(11): 88-97.

Kupiec, J., J. Pedersen, and F. Chen. 1995. "A Trainable Document Summarizer." In *Proceedings of 18th ACM-SIGIR Conference*: 68-73.

Luhn, H. P. 1958. "The Automatic Creation of Literature Abstracts." *IBM J. of Research and Development*, 2(2): 159-165.

Marcu, D. 1996. "Building up Rhetorical Structure Trees." In *Proceedings of 13th National Conference on AI*, 2 : 1069 - 1074.

McKeown, K. R., and D. R. Radev. 1995. "Generating Summaries of Multiple News Articles." In *Proceedings of the eighteenth Annual International ACM SIGIR Conference on Research and Development in IR*. [cited 1998]. 〈http://www.dcs.shef.ac.uk/~gael/Articles/Summarization/Webpage/McKeown.ps〉

Morris, J., and G. Hirst. 1991. "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of the Text." *Computational Linguistics*, 17(1): 21-45.

Ono, K., K. Sumita, and S. Miike. 1994. "Abstract Generation based on Rhetorical Structure Extraction." In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, 1: 344-348. [Kyoto, Japan]. [cited 1998]. 〈http://www.dcs.shef.ac.uk/~gael/Articles/Summarization/Webpage/McKeown.ps〉

Paice, C. D., and P. A. Jones. 1993. "The Identification of Important Concepts in Highly Structured Technical Papers." In *Proceedings of 16th ACM-SIGIR*: 69-78.

Rau, L. F. 1987. "Knowledge Organization and Access in a Conceptual Information System." *Information Processing and Management*, 23(4): 269-283.

Salton, G., A. Singhal, M. Mitra, and C.

Buckley. 1997. "Automatic Text Structuring and Summarization." *Information Processing and Management*, 33(2): 193-207.

Skorokhod'ko, E. F. 1972. "Adaptive Method of Automatic Abstracting and Indexing". In *IFIP Congress*, 71 : 1179-1182. [Amsterdam: North_holland]

Taylor, S. L., G. K. Krulee. 1977. "Experiments with an Automatic Abstracting System." In *Proceedings of the American Society for Information Science Annual Meeting*, 14. [Chicago].