자동 발췌문/요약 시스템 구축에 관한 연구*

- 학술지 논문기사를 중심으로 -

A Study on the Construction of the Automatic Extracts and Summaries

- On the Basis of Scientific Journal Articles -

이 태 영(Tae-Young Lee)**

목 차

초 록

코퍼스 기반의 제 방법, 담화구조의 수사역할, 유사문장의 통합을 이용하여 발췌문과 기초적 요약문을 자동으로 작성하는 방법론을 구축하였다. 코퍼스에 따른 기법들의 효율적 한계치를 사전에 확인하였고 발췌/요약문의 신축적 작성을 위해서 요약문을 이루는 문장들의 수사역할을 목적, 배경, 방법, 결과, 결론 등으로 정하고 각각의 발췌기를 적용하였다. 발췌 성공률은 90%이었다. 수사역할별로 선정된 문장의 합성과 분리를 위하여 유사도 공식을 이용한 유사문장의 통합, 불필요한 의미의 수식절, 삽입절의 제거, 짧은 문장들과 연결이 가능한 문장들의 합성을 시도하였다. 높은 발췌 성공률을 바탕으로 문장의 수사역할, 절의 용언어미 표징, 단서적 어구와 소재를 가미한 문장 정리 시스템의 개발이 요망된다.

ABSTRACT

Various corpus-based approaches, rhetorical roles of discourse structure, and unifications of similar sentences were applied to construct the automatic Ext/Sums(extracts and summaries). Rhetorical roles of sentences like objective, method, background, result, conclusion, etc. for making elastic Ext/Sums were established and extraction engines according to respective role were prepared. The 90% of Success rate in extracting the important sentences of sample articles was accomplished. Rearranging the selected sentences, it used unification of similar sentences using the cosine coefficient equation, deletion of unnecessary modification and insertion clauses, junction of short sentences, and connection of sentences able to link. They suggest the methods applying rhetorical roles of sentences, meaning and signature of noun and verb in clauses, and cue words and location will be researched to construct the more effective Ext/Sums.

키워드: 단서, 발췌기법, 소재, 수사역할, 웹시스템, 의미범주, 자동발췌문, 자동요약, 중요어
Automatic Summaries, Extraction Methods, Location, Rhetorical Roles, Web

* 이 논문은 2004 년도 전북대학교 연구기반 조성 연구비로 이루어졌음.
** 전북대학교 인문대학 문헌정보학과 교수(taehyun@chonbuk.ac.kr)
논문접수일자 2005년 8월 15일
게재확정일자 2005년 9월 13일

1. 서 론

1. 1 연구 목적

초록은 원문의 내용을 축약하여 이용자에게 원문 대신 제공되거나 원문을 판단하게 하는 자료로 이용되는 정보이다. 문헌정보학 분야의 초록에 관한 저서들을 살펴보면 다수의 저자들이 초록은 지시초록, 정보초록, 비평초록, 발췌문(extract)으로 나누어진다고 언급하고 있으며, 색인과 마찬가지로 컴퓨터가 등장한 1950년대 이후 자동화에 대한 연구가 이루어져 왔음을 알 수 있다. 현재는 '요약(summary)' 또는 '자동요약(automatic summarization)'이란 표현이 학술잡지 기사에서 초록을 대신하여 많이 쓰이고 있고, Jones(1999, 5)같은 이는 요약도 지시요약, 정보요약, 비평요약으로 구분하였다.

이러한 초록, 발췌문, 요약에 대한 견해와는 달리 Chowdhury(1999, 144-146)는 초록, 발췌문, '해제(annotation)', 요약문이 각기 다르다고 하였다. 발췌문은 문헌 자체에 있는 문장들을 이끌어내어 만든 요약판인 반면에, 초록은 문헌 내에 있는 단어들을 사용하기는 하나 저자 문장의 직접적 이용이라기 보다는 초록자에 의해 창조된 글조각(a piece of text)으로 정의하였다. 해제는 문헌의 제목이나 서지적 사항들에 대해 코멘트나 설명을 한 주기(note)이며, 요약문은 문헌 내에 있는 현저한 발견점이나 결론에 대한 재서술(보통 말미에 옴)이라고 정의하였다.

Chowdhury의 견해는 "발견점이나 결론에 대한 재서술"인 요약이 논문의 일반적인 기술체계(목적, 결과 등)를 고려하여 핵심적 내용을 서술하는 초록과 같을 수 없으며, 창조적인 문장을 다루는 초록과 기존의 문장을 채택 나열하는 발췌문은 서로 성질이 다르다는 것으로 해석된다. 본 연구에서는 Chowdhury와 Jones의 의견을 반영하여, 본문을 축약하는 형식을 초록과 요약 및 발췌문으로 구분하고, 요약을 지시, 정보, 비평으로 세분하였다. 이와 같이 축약의 형식이 정리되었을 때, 발췌문과 초록 또는 요약에 자동화라는 명제를 대입하여 소위 축약 자동화 시스템을 구축하려고 한다면 먼저 손쉽게 선택할 수 있는 것은 발췌문과 지시적 요약이 될 수 있다.[1)]

오늘날 수많은 지식 및 정보 문서가 연구 또는 생활 속으로 쏟아져 들어오는 지식정보사회의 현상과 함께 여러 가지 발췌문/요약 시스템들이 출현하였지만, 수작업 초록과 같은 완성도를 갖춘 상용 자동초록 시스템이 없는 현실에서 나름대로의 효율성을 갖춘 자동 발췌문과 요약 시스템들은 다다익선인 것이다. 따라서 본고는 웹 사이트를 통해 자동으로 작성되는 발췌문과 지시적 요약 시스템을 구축하고자 한다. 또한 자동 발췌 및 요약의 기능을 향상시키기 위해서 여러 가지 중요문장 추출방법을 적용하고 간단한 문장정리를 통하여 보다 유연한 자동화 시스템을 구축하는 것을 목표로 삼았다.

---

1) "지시, 정보, 비평 초록" 중 자동화를 이루기 가장 쉬운 것은 지시초록이며 "초록, 요약, 발췌문" 중에서는 발췌문이다. 전자의 경우 정보초록은 지시초록에 비해 초록길이가 몇 배 정도 길고 본문의 대용구실을 하기 때문에 부정확성에 대한 부담이 있고 비평초록은 초록자의 비평을 첨가하는 노력이 부가되기 때문이다. 후자에서 초록과 요약은 발췌문에 비해 문장을 생성하여야 하는 어려운 작업이 기다리고 있다.

1. 2 연구 방법과 제한점

본 실험 시스템의 어휘사전과 문장 발췌 및 정리 규칙을 구축하고 성능을 평가하기 위해 문헌정보학 분야에서 4 페이지 정도의 논문 70개를 표본으로 추출하였다. 이 중에서 50 문헌은 시스템 설계를 위한 분석용으로 사용하였고, 나머지 20 문헌은 시스템 평가용으로 삼았다.[2] 평가 방법은 시스템이 만든 자동 발췌문을 수정된 저자초록과 대비하여 그 완성도를 정확률과 재현율로 나타냈다.

형태소의 원할한 분석을 위하여 고유명사와 이에 준하는 어구를 제외하고는 최소자립형태소 단위로 띄어쓰기를 하였다. 어휘분석을 위한 품사는 부사, 조사, 명사, 용언, 명용(명사와 용언으로 동시 사용되는 경우), 수사, 감탄사로 나누었다. 이 중에서 일반명사, 용언, 명용은 1음절, 2음절, 3음절 이상의 세 파일로 분리하였고, 고유명사와 관용어구 및 영어 단어는 별도로 한 파일에 저장하였다. 용언어미도 '하다'류 용언어미와 토종 용언어미로 나누어 작성하였다.

핵심적인 문장들의 발췌를 위해서 코퍼스(단서, 소재, 통계공식), 담화구조, 문맥적 지식기반의 방법들을 적용하였다.[3] 평가용 저자초록 표본들과 자동 발췌문 및 요약의 길이는 5문장 이하로, 문장의 길이는 23-27단어로 한정하였다.[4] 따라서 선택되어져 나온 문장들이 적당한 길이이며 5문장 이하이면 원문의 순대로 나열하여 발췌문이 되게 하고, 그렇지 않으면 문장들을 정리하여 요약이 되게 하였다.

문장의 정리는 적당한 초록 문장 길이인 23-27 단어를 많이 넘어가는 문장들은 분리(dejunction), 또는 불필요한 절을 삭제하여 적당한 길이로 만들었다. 짧은 문장들은 용언어미 '고'를 이용하여 연접시키고, 요약이 가능한 여러 개의 문장들은 문장의 수사역할 별로 정리하여 놓은 "-고 -니 -다"와 같은 문맥리스트를 이용하여 합성(conjunction) 하였다. 용언구들의 표현을 간략화하기 위해서 "-하는 것이 었다" 등은 '-하였다'와 같이 단어형으로 교체하였다.

코퍼스 기반의 모형들을 처리하기 위한 프로그램은 윈도우XP 체제의 퍼스널 컴퓨터에 웹서버 형태로 구현하였으며, 프로그램 언어는 주로 ASP를 사용하고 발췌 규칙들의 표현을 위하여 프롤로그의 혼 절(Horn Clause) 형식을 도입하였다.

2. 발췌문/요약 문장

2. 1 문장의 수사 역할

학술잡지 논문기사의 지시초록은 아는 바와 같이 4-5개의 문장으로 이루어지며 한 문장이 평균 23-27개 단어를 갖고 의미구조, 즉 수사역할 상 "연구목적, 방법, 결과, 결론" 등으로 구성

2) Llorens 등(2004, 848)은 코포스 분석을 위한 문헌 표본들은 요약, 참고문헌을 포함해서 똑같은 구조(섹션)를 가져야 하고 너무 방대하지 않아야 하지만 그 총수는 최소한 50문헌을 넘어야 한다고 하였다.

3) 본 논문 3장의 Aone, Mani, Hovy, Boguraev, Barzilay 등의 분류 참조

4) 미생물학분야의 논문기사의 초록 길이는 평균 4.5 문장들이고 문장 길이는 평균 20.8 단어들(복합어는 한단어로 취급) 이며 한 문장의 절 복합정도는 3.5 개 이었다.(이태영 1992, 60) 본 논문에 사용된 표본들을 분석한 결과 평균 문장길이가 25.4 단어이었다. 따라서 23~27단어를 적당한 길이로 삼았다.

되고 있다. 발췌문과 요약도 초록과 같은 기능을 발휘해야 하는 만큼 동일한 구조를 가져야 한다. Teufel과 Moens(1999, 162)가 조사한 바에 따르면 초록을 구성하는 문장들의 수사 역할 별 퍼센트는 <표 1>과 같다. 실제로 표본 논문 50편의 저자초록을 조사한 결과, '목적' 문장이 27%, '방법' 문장이 7%, '결과' 문장이 21%, '결론' 문장이 7%, '배경' 문장이 27%이었다. 그리고 설명 과정으로 끼어든 문장이 11%이었다.

2. 2 문장의 정리

2. 2. 1 정보 부가

뉴스기사의 자동요약을 수행하는 STREAK에서는 기존의 문장에 새로운 정보를 부가하는데 있어, "부가, 연접, 흡착, 명사화, 인접"으로 지칭되는 다섯 가지의 다른 수정 도구를 적용하였다(McKeown, Robin, and Kukich 1995, 708). 예를 들어 처음의 문장이 "Hartford, CT--Karl Malone scored 39 points Friday night as the Utah Jazz defeated the Boston Celtics 118 94."이라고 하자. 여기서 아래의 <그림 1>과 같이 굵은 활자로 표시 된 것과 같은 정보 부가가 발생할 수 있다.

(가)

2. 2. 2 정보 합성

STREAK에서 사용하는 언어적 요약 도구는 네 가지-(1) 단일어로 복수의 사실을 표현, (2) 수식구 활용, (3) 연접, (4) 대명사로 대체 - 방법이 있다. 구체적으로 소개하면 다음의 <그림 2>와 같다 (McKeown, Robin, and Kukich 1995, 714-718).

<표 1> 수사 역할 별 초록문장 구조

수사적 역할	퍼센트	수사적 역할	퍼센트
배 경	6.3%	해결/방법	37.0%
논제	5.5%	결과	2.4%
관련 연구	4.3%	결론/요청	14.5%
목적/문제	30.0%		

1) 부가(adjunction) : "Hartford, CT--Karl Malone **tied a season high with** 39 points (가)"
2) 연접(conjoin) : "Hartford, CT--Karl Malone tied a season high with 39 points and **Jay Humphries added 24** (가)"
3) 흡착(adsorb) : "Hartford, CT--Karl Malone tied a season high with 39 points and Jay Humphries **came off the bench** to add 24 (가)"
(나)
4) 명사화(nominalization) : "(나) Friday night as the Utah Jazz handed the Boston Celtics **their sixth straight home** defeat 118 94."
5) 인접(adjoin) : "(나) Friday night as the Utah Jazz handed the Boston Celtics **their franchise record** sixth straight home defeat 118 94."

<그림 1> 정보 부가 예

2. 2. 3 정보 분리

긴 문장이 출현하였을 때 분리하는 방법에는 몇 가지가 있을 수 있다. 첫째로 용언어미 '~고'(영어의 경우 and, or)로 연결되어 문맥이 "A ~고, B ~다"인 문장은 "A ~다. B ~다."로 문장을 분리한다. 둘째로 "A ~다는 것을 말한다. B ~함을 ~~"와 같이 삽입절이 포함된 문장은 주절과 삽입절로 분리한다. 셋째로 불필요한 부분들을 삭제하고 다른 간단한 말로 수정하는 방법이 있다. 〈그림 3〉에서 이와 유사한 세 가지 경우 - 삭제, 일반화, 구성을 볼 수 있다.

3. 중요 정보의 발췌

3. 1 발췌방법 종류

초록 또는 요약의 후보문장을 발췌하는 방법으로 여러 가지 유형이 출현하였다. 그 방법들을 Aone 등은 빈도기반, 지식기반(또는 담화기반)으로 크게 나누었고(Alone, Okurowski, Gorlinsky, and Larsen 1999, 71), Hovy와 Lin(1999, 81)은 고전적(Older), 전통적 의미(Traditional Semantic) NLP, 그리고 IR 접근법으로 구분하였다. 또한 Boguraev와 Kennedy (1997, 100)는 템프릿 예시(template instantiation), 문구/문절 발췌(passage extraction) 법을, Barzilay와 Elhaadad(1997, 111)는 표층적(shallow) 언어분석(단어분포, 단서어구, 소재), 어휘체인(담화수준 임)을 주장하였다. Mani와 Maybury(1999)는 40여년간 출현했던 발췌방법들을 고전적 접근법(Classical Approaches), 코퍼스 기반적 접근법(Corpus-based Approaches), 담화구조 활용법(Expoiting Discourse Structure), 지식기반 접근법(Knowledge-rich Approaches)으로 대별하였다.

1) 단일어 사례 :
   ① Portland defeated Utah 101 97, in a tight game where the lead kept changing hands until late in the fourth quarter -〉 Portland outlasted Utah 101 97
   ② "Karl Malone scored 39 points. Karl Malone's 39 point performance is equal to his season high." -〉 Karl Malone tied his season high with 39 points.
2) 수식구 활용 사례 :
   ① "Jay Humphries scored 24 points. He came in as a reserve." -〉 Reserve Jay Humphries scored 24 points
3) 연접 사례 : 두 문장을 'and'나 'or' 로 연결한다.
4) 대명사 대체 사례 : 대명사인 "it, their, them"을 사용한다
   ① the Denver Nuggets and handing them their seventh straight loss"

〈그림 2〉 문장 요약 방법

Deletion : Peter saw a blue ball
(i.e., Peter saw a ball. The ball was blue.)
Generalization : Peter saw a hawk. Peter was a vulture =〉 Peter saw birds
Construction : Peter laid foundations, built walls, built a roof... =〉 Peter built a house

〈그림 3〉 문장 단순화의 예(van Dijk 1979)

본고에서는 위와 같은 결과를 참고하여 단어빈도와 단서어구 및 소재를 근거로 한 코퍼스 기반 방법, 담화(수사) 구조와 심층적 언어 지식에 근거한 담화 - 지식 기반 방법으로 나누어 전개한다.

3. 2 코퍼스 사례

3. 2. 1 구문/의미와 소재 적용

자동초록에서 구문/의미적 방법이 적용되는 근거는 문장의 구문적 정보와 단어나 어구의 의미적 정보가 한 문장의 중요성을 대변할 수 있다는 가정에서 출발한다. Edmundson(1969)은 자동초록 작성을 위하여 네 가지 방법을 제시하였는데, 그 중에서 ①단서, ②표제어, ③소재지 기법은 코퍼스적인 방법이다. '단서' 기법은 "그러므로, 결론적으로, 결과적으로, 중요한 부분으로 등"과 같이 '단서'를 포함하고 있는 문장이 다른 문장에 비해 그 만큼 중요도가 높다고 판단하는 방법을 말한다.[5)]

'표제어' 기법도 표제어에 출현한 의미어들이 그 만큼 중요하다고 판단한다. '소재지' 기법은 '서론', '결론'과 같이 특별한 부분에 나타나는 문장, 문헌의 첫 문단과 마지막 문단에 나타나는 문장, 그리고 문단의 첫 문장과 마지막 문장들이 중요한 문장이라고 판단하는 방법을 말한다. 이후 방법의 가감은 있으되 Kupiec과 Pedersen 및 Chen(1995, 56-57), Myaeng과 Chang (1999), Hovy와 Lin(1999) 등도 이를 언급하였다. 그리고 Rush(1971) 등의 WCL(Word Control List) 방법과 Earl(1970) 및 Paice (1990)의 방법들도 제시되었다.

한편 Llorens 등(2004)은 코퍼스 선정 과정에서 고려해야 할 사항들을 〈표 2〉와 같이 정리하였다. 이 과정은 영역 구별과 결정 과정에서 모아진 문헌들의 집합으로부터 완전하고 양질

〈표 2〉 코퍼스 선정시 고려사항

고 려 사 항	내 용
(1) 문헌표본 조건	①문헌의 디지털화, ②똑같은 섹션 구조 보유(목차, 요약, 참고문헌 등), ③같은 글쓰기 스타일 확보(저자수 제한), ④가능한 동질성(동일수준) 확보
(2) 문헌의 특정한 관점	①희귀성(N 출현빈도 이하), ②문단 당 단어의 수, ③부정구의 수, ④섹션 당 단어수, ⑤공식의 수, ⑥두자어와 약어의 수, ⑦참고문헌의 수, ⑧대명사의 수, ⑨가정법과 과거분사의 구의 수, ⑩미래와 조건 구의 수
(3) 참고 문헌	①가격색인(출판 전 5년 내의 참고문헌), ②저자참조, ③각 참고문헌의 가중치, ④s-r의 색인, ⑤관용어법 능력, ⑥Bradford 수(다른 저널의 수)
(4) 예비 디스크립터	①도표 설명 단어수, ②도표 내에 있는 디스크립터의 수, ③각 문단의 첫 문장에 있는 디그크립터의 수, ④각 구의 첫 동사 전에 있는 디스크립터의 수, ⑤요약 장에서 얻는 디스크립터의 수
(5) 쌍(밀집 성)	①집중성(centrality), ②동시인용, ③동시참조, ④디스크립터의 동시출현

5) Edmundson은 화학 논문기사 200개(each 100~3000 word를 가짐)를 가지고 실험하였다. $W(s) = \alpha C(s) + \beta K(s) + \Psi L(s) + \delta T(s)$ (여기서 CKLT는 각각 Cue, Keyword, Location, Title이고 $\alpha\beta\Psi\delta$는 대상 분야에서 경험적으로 파악된 가중치 비율이다.) Edmundson이 일찍이 실험한 바에 의하면 L이 가장 성능이 좋았고 K가 제일 나빴다. 무난한 것은 C-T-L의 결합이다. 물론 여러 가지 피쳐들의 중요성은 코퍼스에 따라 다양하다. 소재지는 장르에 따라 그 가중치가 다르다.

화된 정보의 코퍼스를 창조하도록 노력하였고, 코퍼스의 질을 확보하는 방법으로 표본의 각각의 문헌들에 계량서지(계량정보)학적인 방법을 강구하였다.

3. 1. 2 확률·통계 적용

Edmundson(1969)의 주요도는 각 문헌에서 출현한 단어의 출현빈도를 계산하여 이 빈도가 일정치 이상인 단어들을 주요어로 선택하여 주요어 사전에 빈도와 함께 수록한다. 이 기법에서는 출현빈도가 각 단어의 중요도를 나타낸다. 문장 선택 시 문장 내의 각 단어는 주요어 사전과 대조되어 가중치를 부여받고 다시 이 가중치를 더하여 문장의 중요도를 산출한다. Brandow와 Mite 및 Rau(1995)는 단어의 가중치를 내는 공식 **tf/idf**를 이용하여 단어들의 가중치를 산출한 후 일정한 한계치를 통과하면 징후어(signature word)로 선택하였고 또 표제어(headline word)도 징후어로 올렸다. 그리고 한 문장 내에 있는 징후어들의 가중치를 합산하여 그 문장의 경중을 가렸다. **tf**는 **tf** = Log_2(문헌집단에 있는 총단어 수/문헌집단에 출현한 용어 수)이고, **idf**는 **idf** = Log_2(문헌 내의 총단어 수/문헌 내의 용어의 출현 수)이다. 이러한 징후어 외에도 Brandow 등이 발췌문에 포함시켜야 할 조건으로 고려한 것은 문헌 내의 소재, 대용어(anaphora)의 존재, 발췌문의 길이, 발췌문의 종류 등이 있었다. 또한 요약문 크기의 기준치에 준하도록 문장들을 요약문에 추가하였는데, 허용 오차는 10 단어 정도였다.

Meadow와 Boyce 및 Kraft(2000)는 문장을 선별하는 방법으로 "문장 내의 단어들의 수" n, "문장 내의 키워드 토큰수" k, "문장 내의 키워드 종수" kq와 같은 인자를 설정하고 이들 인자 간의 관계를 "①키워드 토큰 수/총 단어 수 k/n, ②키워드 종수/총 단어 수 kq/n, ③키워드 간의 평균거리 d"와 같이 설정하여 이 비례수로 의미 있는 문장과 없는 문장을 구별하였다.

Kupiec과 Pedersen 및 Chen(1995)은 어떤 문장이 요약문 안에 포함될 확률을 측정하는 공식을 아래의 공식 1과 같이 제시하였다. 이 공식은 베이지안 규칙(Mani 2001, 60)에 기초하고 있으며 피쳐에는 문장길이, 단서, 소재, 고정구, 주제어, 대문자 단어가 포함되었다.

공식 1:

$$P(s \epsilon E | F_1, \ldots, F_n) = \frac{\prod_{i=1}^{k} P(F_i | s \epsilon E) P(s \epsilon E)}{\prod_{i=1}^{n} P(F_i)}$$

여기서 P (s∈ E | F_1,...,F_k): 원문 내에 있는 문장 s가 중간 발췌문 E에 포함될 확률,

P(s ∈ E): 압축 비율(constant);

P(F_j | s ∈ E): 발췌문 안의 한 문장에서 발생하는 피쳐 F_i 의 확률

P(F_j): 원문 문장들의 코퍼스 내에서 일어나는 피쳐 F_i 의 확률

n: 피쳐의 수

F_j: j 번째 피쳐

3. 2 담화·지식 기반 사례

담화·지식 기반에 근거한 여러 가지 방법들이 발표되어 왔다. 그 중에서 학술 잡지 논문기사에 관하여 Barzilay와 Elhadad(1997, 111-121)

는 본문을 요약하기 위해서 중요한 문장을 뽑아 내는 수단으로 어휘 응집성과 연결성(cohesion and coherence)에 바탕을 둔 어휘 연쇄(lexical chaining) 정보를 이용하였다. 응집성이 강한 문장집단이 '요약' 문장의 발췌 후보로 등장하게 되는 것이다. 〈그림 4〉의 표본에서 H&S (Hirst & St-Onge 1998) 알고리듬에 의하면 'Mr.'란 단어의 연쇄가 먼저 생성된다. 이렇게 관련되는 단어들의 연결을 밝혀 놓은 연쇄리스트들을 '컴포넌트'(component)라고 정의한다. 'Mr.' 는 첫 연쇄리스트에 속하며 [lex "Mr.", sense {mister, Mr .}]와 같이 먼저 표현된다. 다음으로 'person'이란 단어는 "a human being"의 의미로서 'Mr' 와는 연관 강도가 중간이다. 이제 연쇄는 두개의 엔트리를 갖게 되어 [lex "Mr.", sense {mister, Mr.}] [lex "person", sense {person, individual, someone, man, mortal, human, soul}]와 같이 표현된다. 그 다음으로 접근하는 단어 'anesthetic'은 'Mr'와 의미적으로 상관이 없는 단어이기 때문에 현 연쇄에 추가되지 못하고 또 다른 연쇄를 갖게 된다. 이러한 수순으로[6] 연쇄가 만들어지고 연쇄 점수를 해석하게 된다.[7] 여기에 Barzilay와 Elhadad는 다음의 두 가지 사항을 보완하였다. ① 연쇄에 포함되는 단어로 명사 단일어(information, retrieval)와 명사 복합어(information retrieval)를 후보어(candidate)로 택하였다. ② 텍스트 세그먼트에 대한 Hearst(1994)알고리

"Mr. Kenny is the person that invented an anesthetic machine which uses micro-computers to control the rate at which an anesthetic is pumped into the blood. Such machines are nothing new. But his device uses two micro-computers to chieve much closer monitoring of the pump feeding the anesthetic into the patient."

Hirst, G., and ST-Onge, D. 1998[to appear]. Lexical Chains as representation of context for the detection and correction of malapropisms. In Fellbaum, C., ed., *WordNet: An Electronic Lexical Database and Some of its Applications*. Cambridge, MA: The MIT Press.

〈그림 4〉 문장 예

6) 위에서 다른 연쇄가 만들어진 그 다음으로 "machine"이 들어오는데 이 "machine"은 'WordNet'의 첫 의미로 "an efficient person"이므로 person의 holonym이 된다. 다른 말로 하면 "machine"과 "person"은 비록 틀린 결합이지만 강한 연관을 갖게 되는 것이다. Machine의 출현 후 나온 "micro-computer", "device", "pump"는 Machine과 매우 강한 연결 관계를 갖게 된다. 이렇게 연쇄들이 만들어지면 각 연쇄들의 점수를 해석하게 되는데 연쇄점수는 연쇄 사이에 있는 관계들의 수와 가중치에 의해 결정된다. 실험적으로 H&S에서는 반복과 동의어의 가중치는 10, 반의어는 7 등으로 주었다.

7) ① 점수 계산 방법 : 점수(연쇄) = 길이 * 동질성계수 (여기서 길이 : 연쇄 내에 있는 멤버(구성 단어)들이 출현한 수, 동질성계수 : 1 - 이종 멤버들의 수 / 총 멤버들의 출현 수)
② 예 : {baysian-system(2), system(2), baysian-net(2), network(1), baysian-network(5), weapon(1) } (여기서 network(6), net(2), system(4)은 같은 개념을 나타내는 단어임) 따라서 길이 = 12,
동질성계수 = 1 - 3/12, 점수 = 12 * 3/4 -> 9.0
③ 유의성(Strength) 기준 : 점수(연쇄) > 평균(점수) + 2 * 표준편차(점수)
(주류 연쇄(문단과 달리 비슷한 개념을 안고 있는 문장들의 모임)에서 유의성 있는 문장을 취택)

듬에 따라 세그먼트로 나누고 세그먼트에서 연쇄를 형성하였다. 그리고 다른 세그먼트들 사이의 연쇄들을 연결성만을 고려한 기준으로 합병하였다. 즉 두 연쇄에 같은 의미를 갖는 공통 단어가 있을 시에 합병한다.

Teufel과 Moens(1999, 160-164)는 코퍼스적인 방법과 담화구조 방법 등을 동원하여 자동 발췌 시스템을 종합적인 시각으로 접근하였다. 그는 Kupiec 등이 사용했던 자질-발췌요인 4개와 본인이 부가한 자질 3개 등 총 7개의 자질들을 사용하였는데 그 중 담화구조에 관련된 것은 아래의 (2)번 방법이고 (3)부터 (7)까지는 코퍼스 기반 접근법이다.[8)]

(1) 단서 성능 자질(Indicator Quality Feature): 단서성능자질은 본문 내에서 주제 문제가 아닌 이차적 식별 기능을 담당한다. 여기서는 모두 1728 단서어구 또는 공식적 표현(formulaic expressions)들을 다루었다. 〈표 3〉이 그 일부이다. 위의 점수는 5가지 척도(리커드 척도)로 '매우 높음'서부터 '매우 낮음'까지 5단계의 값을 갖는다. 예를 들면 "we argued"는 시제로 따져 "we have argued" 보다 높은 점수를 받아 +2가 된다.

〈표 3〉 단서어구 리스트 예

단서 어구	성능점수
we argued	2
we have argued	1
what we have argued is	1
This article	3
in this article	3
is an attempt to	1
I have attempted	2
our work attempts	2
supported by grant	-1

(2) 수사 역할 자질(Indicator Rhetorics Feature): 이 자질은 어구의 의미적(수사적 공헌도)인 측면 을 계상하며 16개 '클래스'로 구성되었다. 클래스들 중 7개는 단일적인 수사역할(rhetorical roles)로 그 종류는 "배경, 논제, 관련연구, 목적/문제, 해결, 결과, 결론/요구"이며 8개는 복합적인 것(confusion class)으로 "해결-목적/문제, 해결-결론/요구, 목적/문제-결론/요구, 목적/문제-관련 연구, 목적/문제-배경, 결론/요구-관련연구, 결론/요구-결과, 배경-관련연구"가 있다. 그리고 어떤 특정 수사적 역할을 예견하지 못하는 구들을 위해서 제로 값을 준비하였다. 그 예로 〈표 3〉의 'argue'는 '결론/요구'로 분류되고 'attempt'는 '목적/문제'에 속할 수 있는데 반해 두 번째와 네 번 째 줄은 제로 값을 받게 된다.

8) (3) 소재 자질(Relative Location Feature): 문헌의 첫 장(대부분 서론), 마지막 장(대부분 결론), 문헌 또는 장의 첫 문단과 마지막 문단 처럼 본문 내의 위치에 따라 중요도를 다르게 한다.
(4) 문장길이 자질(Sentence Length Feature): 일정치 이하의 문장 길이는 0을 그 이상은 1을 준다.
(5) 주제어 자질(Thematic Word Feature): 이것은 (tf.idf)로 계산된다. 먼저 상위 순위 10 안에 드는 단어를 고르고 이것들을 주제어로 규정한다. 그리고 각 문장들에 포함된 이 단어들의 가중치를 계산한 후 상위 40 문장 에게는 1을 40 이하는 0를 준다.
(6) 표제 자질(Title Feature): 평균적인 빈도수를 갖는 표제어(스톱워드는 제외함)를 포함하고 있는 문장들 중 상위 18 문장에게는 1을 아니면 0를 부여한다.
(7) 항목 유형 자질(Header Type Feature): 문헌의 장절을 수사적으로 구분되는 14개의 원형(prototypical)그룹을 나누고 각각의 장절 항목을 해당되는 그룹에 소속시킨다. 그리고 이 항목 단어를 갖는 문장들에게 그 항목이 속한 그룹의 값을 부여한다. 항목어가 없는 문장들은 '무원형'(non-prototypical)에 소속시킨다.

Li와 Wong 및 Yuan(2001)은 시간에 관련되는 정보를 추출하기 위해서 시간에 관련된 동사의 정보(동사의미, 시제 등)를 이용하였다.

4. 발췌방법의 사전 평가

4. 1 발췌 실험

3장의 내용을 참고하여 사전 실험에서 다루어야 할 방법을 다음과 같이 선정하였다. 코퍼스 사례에서 "단서, 소재, 표제, 중요어(주제어)[9]"를 이용한 방법과 "Edmundson, Brandow 등, Meadow 등"이 제시한 방법, 그리고 Kupiec 등의 확률모형을 원용해서 실험하였다. 담화 사례에서는 Barzilay와 Elhadad의 체인과 Teufel과 Moens의 수사적 역할을 응용 실험하였고, Llorens 등의 표에서 '(2)문헌의 특정한 관점' 중 "③부정구의 수, ⑨가정법과 과거분사 구의 수, ⑩미래와 조건구의 수", 그리고 Teufel과 Moens의 '(1) 단서성능자질'을 '용언 표징'[10]으로 규정하고 점검하였다. Kupiec 등과 Teufel과 Moens가 제시한 방법들은 종합적인 것이어서 단위적인 사전실험에서는 생략하고 5장의 실제 발췌방식에 적용시켰다. 각 기법과 실험 진행 절차는 아래의 〈표 4〉와 같고 표본 문헌의 입력은 최소자립형태소 단위로 띄어써서 입력하였다. 실험 대상문헌은 정보관리학회 학술대회 논문집에 실린 논문기사 중에서 Teufel과 Moens의 구조율을 참작하여 되도록이면 저자초록에 "목적, 방법, 결과, 결론, 배경"을 표현하는 문장을 갖고 있는 50개 문헌을 선정하였다.

4. 2 실험 결과

11가지 문장발췌 모형-코퍼스 기반에서 단서, 표제어, 소재지, 중요어, Edmundson 방법, Brandow 방법, 그리고 Meadow 방법, 담화/지식 기반에서 Kupiec 방법, Barzilay 방법, Teufel 방법, 그리고 용언 표징 방식-을 〈표 4〉에 기술되어 있는 절차대로 실험하여 1)~11)까지의 결과를 얻었다. 평가용으로 사용한 수정된 저자초록은 초록의 내용을 그대로 적용하되 저자초록의 문장이 너무 길 경우(약 45 단어 이상) 문장을 적절히 분리하였다.

---

9) 중요어의 선정은 실험 문서집단에 출현한 장서빈도와 문헌빈도가 각각 2개 이상인 단어를 선택하였고, 준 관용적 복합어를 우선시 하였다. 또한 관용적 복합어로서 주제표의를 강렬하게 나타내고 있으면 출현빈도에 관계없이 중요어리스트에 포함시켰다.

10) 용언은 "실험하다, 예상하다, 사려되다, 생각하다, 높았다, 낮았다, …"와 같이 이 단어들을 포함하고 있는 문장이 중요한 문장이라고 알리는 의미적으로의 표징과, "-기 때문에, -러한 -(으)로 보았을 때, -만을 위한 것이 아니고, -지를 -고자 한다, -리라 -된다(한다), …"과 같은 구문적인 표징이 있을 수 있다. 여기서 용언과 명사 및 조사의 표징적인 정보를 이용할 수 있다. 용언어미인 '하였으니까'와 명사/조사인 '~기 때문에'는 원인을 이야기하며 결과를 기대하는 표징을 나타낸다. 그리고 "내세우다, 걸어가다"는 "세우다, 걷다"를 보다 확실히 표현한 말이다. 조사 '만은'은 유일성을 나타내어 "가, 는, 을"들과는 다른 특별성을 준다. 따라서 이들 말과 같이 쓰인 문장은 보다 강한 상황적 의미를 지녔다고 할 수 있다.

〈표 4〉 사전 실험 방식별 설명 〈가중치는 특별한 명시가 없는 한 단어빈도이다.〉

방 식	실 험 절 차
단서	①실험집단 문서들에 있는 단서어구들의 리스트를 만든다. ②문장의 단어들을 리스트와 비교하여 출현한 단서어구들의 숫자를 합산한다. ③적정한 한계치를 설정하여 문장을 선별한다.
표제어	①실험대상인 문서의 표제에 있는 명사('연구' 제외)들을 의미어로 선정한다. ②문장의 단어들을 의미어들과 비교하여 출현한 의미어의 숫자를 합산한다. ③적정한 한계치를 설정하여 문장을 선별한다.
소재지	①서론(연구목적이 포함되어 있는 장/절)의 첫 문단과 마지막문단, 결론의 첫 문단과 마지막 문단을 선택한다. ②서론의 첫 문단의 첫 문장과 마지막문장, 마지막 문단의 첫 문장과 마지막 문장, 결론의 첫 문단의 첫 문장과 마지막 문장, 마지막 문단의 첫 문장과 마지막 문장을 선택한다.
중요어	①실험집단 문서들에 있는 중요어들의 리스트를 만들고 수작업으로 가중치를 부가한다. ②문장의 단어들을 리스트와 비교하여 출현한 중요어의 가중치를 합산한다. ③적정한 한계치를 설정하여 문장을 선별한다.
Edmundson	①실험집단 문서들에 있는 단어들의 장서빈가 적정한 한계치를 통과하면 주요어 리스트에 빈도와 함께 수록한다. ②문장의 단어들을 리스트와 비교하여 출현한 주요어의 가중치를 합산한다. ③적정한 한계치를 설정하여 문장을 선별한다.
Brandow	①일정한 (tf/idf) 한계치를 통과한 단어와 표제어를 출현빈도와 함께 징후어 리스트에 올린다. ②문장의 단어들을 사전과 비교하여 출현한 징후어의 가중치를 합산한다. ③적정한 한계치를 설정하여 문장을 선별한다.
Meadow	①실험집단에서 단어빈도가 적정한 수준인 단어들을 키워드로 선정한다. ②문장의 k/n(키워드 토큰수/총 단어 수), kq/n(키워드 종수/총 단어 수)을 계산한다. ③적정한 "(k/n) - (kq/n)"의 한계치로 문장을 선별한다.
Kupiec	①실험될 자질들을 '문장길이', '단서', '주제어', '소재지', '고유 및 관용어구'로 설정한다. ② 선험적 확률 값을 적용하여 각 문장 자질들의 값을 합산한다. ③적정한 한계치를 설정하여 문장을 선별한다.
Barzilay	①단어의 의미를 살펴 범주 클러스터를 형성시킨다. ②출현빈도로 가중치를 삼는다. ③문단의 문장 단위로 연쇄를 형성시킨다. ④가중치가 높은 연쇄를 갖는 문단을 선별한다.
Teufel	①실험 문헌의 문장들에서 수사적 역할(목적, 방법, 결과, 결론, 배경 등)을 하는 단어나 구가 있으면 선별한다.
용언 · 명사표징	①의미 표징에서 'Teufel'과 중복되는 것은 리스트에서 제외한다. ②의미표징과 구문표징에 해당하는 단어나 구를 갖고 있는 문장을 선별한다.

1) 단서 : 실헌집단 문헌들은 단서어(그림 5 참조)를 두개 이상 포함하고 있는 문장들을 평균 3개씩 가지 고 있었고 그 중 1.7 개는 평가용 수정된 저자초록에 수록되어 있는 요약 문장들이었다.(적중률 60%) 그리고 단서를 1개 이상 포합하고 있는 문장들은 평균 10 개였는데 '목적'과 '결론' 문장을 가지고 있었다.(이하 기술 되는 문장 숫자와 적중률은 표본 문헌 50 개의 평균 값임)

강점, 강조, 고로, 그것, 그래서, 그러나, 그러므로, 그리고, 결과, 결론, 본, 본고, 실제, 연구, 이론, 점, 고찰하다, 높다, 낮다, 단점, 문제점, 믿다, 보다, 사려하다, 사료하다, 살피다, 생각하다, 이것, 있다, 그러한, 이러한, 장점, 저것, 저러한, 해결, …

〈그림 5〉 단서어구 리스트

2) 표제어 : 실험집단에서 표제어를 하나 이상 가지고 있는 문장들은 20 개, 2 개 이상 가지고 있는 문장들의 수는 5 개 였다. 참고로 실험집단의 문헌들이 가지고 있는 문장 수는 90 이다. 따라서 2 개 이상을 기준으로 삼아 수정된 저자초록의 요약 문장들과 비교하여 50%의 적중률을 가짐을 알 수 있었다.

3) 소재지 : ①방법으로 수정된 저자초록에 수록된 요약 문장의 65%가 이곳에서 발견되었다. ②방법에서는 수정된 저자초록에 수록된 요약문장의 45%가 이곳에서 발견되었다(목적, 방법, 결과, 결론 문장 기준).

4) 중요어 : 중요어를 하나 이상 갖고 있는 문장들은 30 문장이었다. 선정 한계치를 10으로 높였을 때, 10 문장이 출현하였는데 그 중에 요약 문장이 다 들어 있었다. 선정 한계치를 높일수록 출현 문장은 적어지나 그 반대로 요약 문장으로의 적중률은 떨어졌다.

5) Edmundson 식 : 표본을 살펴본 결과 장서 빈도의 하한과 상한 한계치를 의미어들이 많이 포진하고 있는 20과 200으로 정하였다.(이 빈도 내에 중요하다고 생각되는 단어들이 70% 정도 모여 있음) 다음으로 문장 선정 가중치를 1000으로 입력하면 10 문장들이 출력된다. 요약 문장의 70%가 적중하였다.

6) Brandow 식 : 각 문헌에 적절하다고 판단된 tf/idf 값 하한 한계치 1과 상한 한계치 2에서 단어 빈도 가중치를 1000으로 정했을 때, 15개 문장이 출력되고 적중률 90%, 가중치를 2000을 하면 10개 문장 출력에 적중률 50%, 가중치를 3000으로 하면 5 개 문장 출력에 적중률은 30%에 그쳤다.

7) Meadow 식 : 각 문헌에 적절하다고 판단된 단어빈도가 5와 40 사이에서 문장선정 한계치를 0.1로 하였을 때, 6 문장이 출현하였고 적중률은 50%였다.

8) Kupiec 식 : 각 문헌에 적절하다고 판단된 확률 한계치를 0.1로 하였을 때, 6 문장이 출현하였고 적중률은 50%였다.

9) Barzilay 식 : 각 문헌에 적절하다고 판단된 문장선정 한계치를 500으로 하였을 때, 11 문장이 출현하였고 적중률은 70%였다.

10) Teufel 식 : 〈표 5〉의 수사역할 별 단어로 선출되어 나온 문장들이 저자초록에 있는 내역과 일치하였다. 단 출현 문장의 수는 10 이었다.

〈표 5〉 수사역할 별 용언 구

수사 역할	용언 구
연구목적	"논하다, 조사하다, 연구하다, 고찰하다, 설계하다, ~고자 하다, ~해 보다, ~데 있다, 알아 보다, 제시하다, 밝히다, 소개하다, 다루다 등"의 과거형과 현재형, (그) 목적이 있다, 목적이다 등
배경	~하/되/지고 있다, ~하게 되었다, ~이 있다, 이었다, 있는 현실/실정이다, 지정한다, ~로 한다, ~ㄴ 형편이다, ~되어 왔다, ~으로 채택한다, 로 한다, 어렵다 등
방법	"사용/이용하여, 모색하고, 혼용하여, 비교하여, 유도하여 등"의 현재/과거형, 문제를 분석하고, 있다 등
결과	제시하였다, 발견하였다, 나타났다, 할 수 있다, ~하게 하였다, 구현하였다, 설계하였다, 높았다, 아니었다, 없었다, 개발하였다, 요구하였다, 개설하였다, 명시하였다 등
결론	~(어)야 한다, ~야 할 것이다, ~할/될/시킬/질 것으로 본다, ~할/될/시킬/질 전망이다, 확인하였다, 제안한다, 요구된다, [중요함을] 제시한다, 사려된다 등

11) 용언 표징 방식: 〈그림 6〉과 같은 용언어구로 출현한 문장들은 중 평가용 수정된 저자초록의 요약 문장들과의 일치도는 60%이었다.

-려고 한다, -고자 한다, ~하기 때문이다, -는 것이 좋다, -ㄹ 필요가 있다, -하여야 한다,-야만- -겠다, -다고 생각하다, -는 것이 당연하다, -ㄴ 다는 것뿐이-, -도록 할 수 있-, -할(될,질,시킬) 수밖에 없다, -ㄹ 뿐만 아니라, -지를 않는 것이-, 하지 않을 -려고 한다, -고자 한다, ~하기 때문이다, -는 것이 좋다, -ㄹ 필요가 있다, -하여야 한다,-야만- -겠다, -다고 생각하다, -는 것이 당연하다, -ㄴ 다는 것뿐이-, -도록 할 수 있-, -할(될,질,시킬) 수밖에 없다, -ㄹ 뿐만 아니라, -지를 않는 것이-, 하지 않을 수 없다, -ㄴ 적이 없(있)을 뿐만 아니라, 〈아무리〉 강조하여(해)도 지나침이 없다, 최대한 발휘하는

것이다, -야 할(될) 것(거)이-, -ㄴ(한) 척하는 것 뿐이-, -하여(해)서는(하면, 하는 것은) 안된다(곤란하다), -할 것 없-, (할,일,될,시킬) 뿐이다, -지 않을 뿐 이다(더러), -ㄴ 〈것〉일 뿐이다, -는 것으로 파악(알려, 나타, 예상, 보, 조사)-, 것도 무리는 아니다, 것이라고 맞서(주장, 대응)- 있다, 가장 타당(효과적, 타당,자극적)-, -는 점에서 충격적이다, -는 것도 무리는 아니다, -는 사실이다, -는 것이나 마찬가지다, -야 하는 까닭이 여기에 있다, -ㄹ 수밖에 없지 않겠느냐, -를 하고 있는 것이다, -경향마저도 적지 않았다.

〈그림 6〉 용언 표징을 위한 용언구

앞에서 도출된 결과에서 몇 가지 기본적인 사항을 짚어 볼 수 있다. 첫째, 11 가지 모든 방법이 출력 문장 수를 늘릴수록 평가용 초록에 있는 문장들과 일치하는 수를, 즉 적중률을 증가시키며 문장 수를 줄일수록 적중률을 감소시킨다. 다시 말해서 문장 수를 늘릴수록 재현율은 높아지고 문장수를 줄일수록 재현율이 낮아진다는 것이다. 그렇다면 문장 수를 늘릴수록 정확률은 감소하고 문장 수를 줄일수록 정확률은 증가할 것인가 하는 의문이 남는다.(7장 시스템 평가에서 확인함)

둘째, 실험결과에서 단서, Barzilay 식의 범주(이하 범주로 칭함), 소재지, 중요어, Teufel 식은 알고리듬이 간단하면서도 좀더 복잡한 Edmundson 식, Brandow 식, Meadow 식, 그리고 Kupiec 식에 비해 효율이 대동소이하거나 높았다. 따라서 이와 같이 효율이 높지 않은 복잡한 식은 본 실험시스템에서 도입할 필요가 없을 것이다.

셋째, 소재지 방법은 특히 서론과 결론의 문단 수와 문단 내의 문장 수에 영향을 받아 문단 수와 문장 수가 많으면 많을수록 발췌문의 길이가 늘어나는 특징이 있다. 그러므로 상황에 따른 영향을 최소화할 수 있는 완충 해법(다른 방법과의 결합)이 필요하다.

넷째, 단서와 중요어, 범주는 사전 수작업 결과에 따라 영향을 받는다. 즉 단서와 중요어, 및 범주를 어떻게 구성하느냐에 따라 작업수행 결과가 달라질 수 밖에 없다. 따라서 단어선정 및 가중치 부여에 대한 일반화된 방법의 개발이 요구된다. 본 실험시스템에서는 주제 전공자 5인의 의견을 수렴하였다.

다섯째, Teufel 식 수사역할에 의한 방법은 사전 선정 문제에도 불구하고 초록의 구조에 맞는 문장들을 다른 방법에 비해 보다 완전하게 발췌하였다. 그리고 연구 '목적'과 '결론' 문장은 앞에서 기술한 5 문장 발췌를 기준으로 삼았을 때, '방법'이나 '결과' 문장 또는 다른 서술적 문장에 비해 평균 두 배 정도 높게 발현되었다. 이것은 '목적'이나 '결론' 문장이 문장 형식 또는 태생 상 단서, 중요어, 용언표징 등을 중첩하여 보유하고 있기 때문이라 사려된다.

여섯째, 용언 표징 방식은 '-고자 한다'(수사역할의 '목적')의 예에서 보듯이 수사역할의 용언들과 중복되는 점이 있어 정확한 효능을 예측하기가 어렵다. 그럼에도 불구하고 수사역할이 발휘한 힘과 같이 "방법, 결과" 문장을 발췌는 것과 문장을 정리하는 데에 보조적 역할을 할 수 있을 것으로 사료된다.

일곱째, 소재지 기법에서 단락들을 선정하게 되는 경우가 있는데, 이 단락은 이웃 글이 중요한데도 불구하고 다른 기법에서는 탈락되어 보강의 기회를 놓치는 경우를 만회할 수 있다. 중요 문장 앞뒤에 짧은 문장이 있어 중요문장을 보강하는 경우가 종종 있다.

이러한 기본적인 사항들을 감안하여 본 실험

에서는 비경제적인 복잡한 공식은 제거하고, 문장수를 지시초록에 적당한 문장 수 5 개 이하로 한정시켰다. 결과가 뚜렷한 수사역할과 소재지 방법은 적극 활용하고, '방법'과 '결과' 문장 추출을 보강하기 위해 "단서, 중요어, 범주"와 같은 기법들을 추가하였다.

5. 실험 시스템 구축

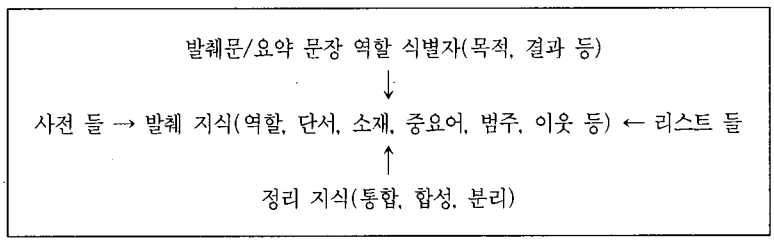
5. 1 시스템 개요

본 실험 시스템은 2, 3 장에서 소개한 초록문장의 구조와 중요문장의 발췌 방법 및 문장의 합성과 분리 범례를 그대로 적용하거나 또는 변형시킨 모형을 적용하여 만들었다. 그 개요는 〈그림 7〉과 같다. 이 시스템의 처리과정은 ① 입력되는 문장을 최소자립형태소별로 분리→ ②역할 식별자로 목적, 방법, 결과, 결론 등 문장 역할 식별(용언으로 판단) → ③문장 발췌기로 중요한 문장을 선정(이웃에 있는 짧은 문장과 '그러나' 등으로 이어지는 문장도 함께 선정) → ④유사도 공식으로 유사문장 통합 또는 문맥지식으로 문장의 분리 및 합성 → ⑤ 최대 5문장 까지 본문의 순서대로 나열로 이루어진다.

5. 2 문장 역할 식별

보다 정밀한 요약 후보문장의 발췌를 위해서 문장의 수사적 역할을 4장의 사전 실험 때와는 다르게 〈표 6〉에서 보는 바와 같이 역할의 종류에 따라 명사의 표징을 추가하였다. 그러므로 각 역할 자질들은 주절의 명사(주어, 목적어, 보어 순으로 하나만 택함)와 용언의 표징으로, 두 개의 표징 변인이 존재하게 된다. 〈표 6〉에 나와 있는 역할 자질들인 "목적/논제, 방법, 결과, 결론/제언, 배경" 등은 논문의 초록이 흔히 "목적, 방법, 결과, 결론"을 나타내는 문장들을 중심으로 작성된다는 점과 3.2장의 Teufel의 (2)와 (7) 번 항목을 절충하여 선정하였다. 용언 및 명사들의 실제 값은 표본 논문기사를 분석하면서 발췌하여 낸 것들이다.

사전 실험의 결과에서 보듯이 문장의 수사적 역할은 용언의 종류에 따라 대부분 밝혀질 수 있는 것이다. 그런데 '분석하다'의 경우처럼 한 용언이 여러 가지 역할을 대변하는 경우를 볼 수 있어 명사의 표징을 식별 변인으로 더하였다. 그래도 미흡한 것은 용언/명사의 출현 소재를 또 하나의 변인으로 삼아 모호성을 상쇄시킬 수가 있다. 다음은 문장 역할 별로 출현하는 소재 위치를 열거하였다.



〈그림 7〉 실험시스템 개요도

〈표 6〉 수사역할 자질과 소속 단어

역할	명사의 종류	용언의 종류
목적(/논제)	목적, 목표, 본고, 본 연구, 본 논문, 논제, 명제 등	"논하다, 조사하다, 연구하다, 고찰하다, 설계하다, ~고자 하다, ~해 보다, ~데 있다, 알아 보다, 제시하다, 밝히다, 소개하다, 다루다 등"의 과거형과 현재형, (그) 목적이 있다, 목적이다, 측정하다, 시도하다, ~게 하다, 고찰하다,
방법	가설, ~법, ~규칙, 법칙, ~론, 수단 등	"사용/이용하여, 모색하고, 혼용하여, 비교하여, 유도하여 등"의 현재/과거형, 문제를 분석하고, 있다, 통하다, 분석하다, 구분하다, 전개하다, 의하다, 대하다, 관하다, 위하다
결과	'목적'의 주어, 목적어 명사, 결과 등	제시하였다, 발견하였다, 나타났다, 할 수 있다, ~하게 하였다, 구현하였다, 설계하였다, 높았다, 아니었다, 없었다, 개발하였다, 요구하였다, 개설하였다, 명시하였다, 낮았다, 있었다, 증가하다
결론(/제언)	'목적'의 주어, 목적어 명사, 결론 등	~(어)야 한다, ~야 할 것이다, ~할/될/시킬/질 것으로 본다, ~할/될/시킬/질 전망이다, 확인하였다, 제안한다, 요구된다, [중요함을] 제시한다, 사려된다, 추정하다, 생각되다, 알다, 판명되다, 밝히다, 제언하다, 제시하다, 요구되다,
배경	배경, 환경, 상태, 상황, 모습, 성질 등	~하/되/지고 있다, ~하게 되었다, ~이 있다, 이었다, 있는 현실/실정이다, 지정한다, ~로 한다, ~ㄴ 형편이다, ~되어 왔다, ~으로 채택한다, 로 한다, 어렵다, 소개하다, 실정이다, 도입하다, 넣다, 주사하다, 따르다, 다르다, 가지다, 일어나다

1) 목적문장 식별 : 소재 = '서론' 또는 '서론'의 '연구목적', '연구의 필요성'에서 출현
2) 방법문장 식별 : 소재 = '서론' 또는 '서론'의 '연구방법', '연구방법 및 제한'에서 출현하고 본론 부분에서 재출현됨
3) 결과문장 식별 : 소재 = 본론 부분에서 출현하는데 특히 결과 분석, 평가와 같은 장에서 출현하고 결론에서 재확인될 수 있다.
4) 결론문장 식별 : 소재 = '결론'에서 마지막 단락 부분에서 출현한다.
5) 배경문장 식별 : 소재 = '서론'과 '이론적 배경'에서 나타나며 통상 2장 이하에서 많이 출현한다.

5. 3 중요문장 발췌

5. 3. 1 목적문장 발췌규칙

논문기사의 연구목적 문장은 보통 서론 또는 서론의 연구목적, 연구필요성의 말미 문단에 출현한다. 그리고 그 표징도 뚜렷하여 쉽게 찾을 수 있다. 보통 목적문장은 문장 첫머리에 "본고, 본 연구는, 이 연구는" 등과 같은 '본고'류의 단서 어구를 갖고 있으며 그 문장 중에 "본고의 목적은, 목적으로 한다."와 같이 '목적' 단어가 출현한다. 흔히 한 문장으로 만들어지나 그 이상의 문장들로 이루어질 때도 있다. 사용되는 변인들은 수사적 역할 상의 명사와 용언의 표징과 소재 및 단서가 적용되었다. 목적문장의 특징과 발췌규칙은 다음과 같다.[11]

11) 여기서 "PurpoSent1() → 목적문장 발췌규칙 1, noun|[본고류] → '본고'類로 한정되는 명사, yongeun|purpose → 목적을 나타내는 용언, section|mainchapter → 서론과 결론을 제외한 장, adverb|cue[방법] → '방법'으로 한정되는 단서를 갖는 부사, word(x, y) → x와 y 조건을 갖는 단어, Sentence(x) → x 조건을 갖는 문장, MethodBasic1 → 결과 기초규칙 1"의 뜻을 나타낸다.

특징: 주어 -> '본고'류; 주절 용언->
목적형; 소재-> 서론 마지막 문단;

발췌규칙: PurpoSent1() :- Sentence
(word(noun|[본고류],
yongeun|purpose)), location
(section|introduction)

5. 3. 2 방법문장 발췌규칙

방법문장이 출현하는 양상은 한 문장 전체가 연구 방법을 표현하는 문장으로 등장하는 경우와 방법과 결과를 동시에 기재하는 혼합형 문장의 경우가 있다. 주목할 사항은 과정/설명을 말하는 절이나 문장과 같은 장소에서 많이 출현하고 있다. 그리고 연구목적 문장의 주어 목적어 명사들이 이 문장에서 반복 출현하는 상호참조적 현상을 띄운다. 방법문장의 특징과 영향요인 및 발췌규칙은 아래와 같다.

특징: 명사 -> 목적문장 중의 명사와 -법, -론 등의 명사; 주절 또는 부절 용언 -> 방법형; 소재 -> 서론의 '연구방법', 본론, 결론 부분; 에서 출현;
영향요인: 문장길이; 단서; 범주; 중요어;
발췌규칙:
MethodBasic1 :-
word(yongeun|method)
MethodBasic2 :- word(yongeun|method,
noun|pursose)
MethodBasic3 :- word(yongeun|purpose,
adverb|cue[방법])
MethodBasic4 :- word(yongeun|purpose,
noun|pursose, adverb|cue[방법])
MethodBasic5 :- word(noun|category)
MethodBasic6 :- length(sentence)
MethodSentRule1() :-
sentence(MethodBasic1,
word(adverb|cue))
MethodSentRule2() :-
sentence(MethodBasic2,
word(adverb|cue))
MethodSentRule3() :-
sentence(MethodBasic3),
location(section|(method,
mainchapter, conclusion))
MethodSentRule4() :-
sentence(MethodBasic4),
location(section|(method,
mainchapter, conclusion))
MethodSentRule5() :- MethodBasic2,
MethodBasic5, MethodBasic6
MethodSentRule6() :- MethodBasic4,
MethodBasic5, MethodBasic6

5. 3. 3 결과문장 발췌규칙

결과문장은 주로 본론의 '결과 분석' 또는 '~평가' 부분에 일단 출현하고 결론에서 다시 반복하여 출현하게 된다.(흔히 방법절과 함께 출현함) 그리고 연구목적 문장의 주어 목적어 명사들이 반복 출현되는 양상을 띈다. 결과문장의 특징과 영향요인 및 발췌규칙은 아래와 같다.

특징: 명사 -> 목적문장 중의 명사와 -법, -론 등의 명사; 부절 용언 -> 결과형, 방법형; 주절의 용언 -> 결과형; 소재 -> 서론의 '연구방법', 본론, 결론;

부분에서 출현,
영향요인: 범주; 중요어; 문장길이; 단서;
발췌규칙:

```
ResultBasic1 :- word(yongeun|result)
ResultBasic2 :- word(yongeun|result,
    noun|pursose)
ResultBasic3 :- word(yongeun|purpose,
    adverb|cue[결과])
ResultBasic4 :- word(yongeun|purpose,
    noun|pursose, adverb|cue[결과])
ResultBasic5 :- word(noun|category)
ResultBasic6 :- length(sentence)
ResultSentRule1() :-
    sentence(ResultBasic1,
    word(adverb|cue))
ResultSentRule2() :-
    sentence(ResultBasic2,
    word(adverb|cue))
ResultSentRule3() :-
    sentence(ResultBasic3),
    location(section|(mainchapter,
    conclusion))
ResultSentRule4() :-
    sentence(ResultBasic4),
    location(section|(mainchapter,
    conclusion))
ResultSentRule5() :- ResultBasic2,
    ResultBasic5, ResultBasic6
ResultSentRule6() :- ResultBasic4,
    ResultBasic5, ResultBasic6
```

5. 3. 4 결론문장 발췌규칙

결론문장들은 연구목적 절과 함께 결론 섹션의 후반부에 출현하는 경향이 강하다. 표본조사 결과에서 연구목적 문장에 출현한 일부 절이나 어구가 결론문장에 다시 나타나는 선험 확률은 단어나 구의 경우, 90%에 이르고 있다. 그리고 결론 섹션의 후반부에 결론문장이 포함될 선험 확률은 80%이었다. 결론문장의 특징과 영향요인 및 발췌규칙은 아래와 같다.

특징: 명사 -> 목적문장 중의 명사, 부절
' 용언 -> 목적형, 소재 -> 결론의
마지막문단 부분에서 출현,
영향요인; 범주; 중요어; 문장길이; 단서;
발췌규칙:

```
ConcBasic1 :-
    word(yongeun|conclusion)
ConcBasic2 :- word(yongeun|conclusion,
    noun|pursose)
ConcBasic3 :- word(yongeun|purpose,
    adverb|cue[결론])
ConcBasic4 :- word(yongeun|purpose,
    noun|pursose, adverb|cue[결론])
ConcBasic5 :- word(noun|category)
ConcBasic6 :- length(sentence)
ConcSentRule1() :-
    sentence(ConcBasic1,
    word(adverb|cue))
ConcSentRule2() :-
    sentence(ConcBasic2,
    word(adverb|cue))
ConcSentRule3() :-
    sentence(ConcBasic3),
    location(section|conclusion)
ConcSentRule4() :-
```

```
    sentence(ConcBasic4),
    location(section|conclusion)
ConcSentRule5() :- ConcBasic2,
    ConcBasic5, ConcBasic6
ConcSentRule6() :- ConcBasic4,
    ConcBasic5, ConcBasic6
```

5. 4 문장 정리

선정된 문장들의 정리를 위하여 2장에서 살펴보았던 정보의 부가, 합성, 분리에 대한 사항들을 원용하였는데, 선정된 문장들 중에 유사문장이 있을 경우를 대비하여 Moens와 Uyttendaele 및 Dumortier(1999), Schutze(1998)들이 제시한 문장간 유사도 측정 공식을 문장통합 함수로 장치하였다. 본문에서 발췌가 이루어져 역할별로 선정되어 나온 문장들의 수가 역할 유형별로 허용되는 기준치(숫자)를 초과하였을 때 이 유사도 공식을 적용하여 문장의 수를 적당하게 조정할 수 있다. 또한 문장의 길이가 표준길이 보다 10단어 이상 넘어갈 때, 반대로 문장의 길이가 표준길이에 비해 10단어 이상 짧을 때는 원래의 문장들의 의미를 유지하면서 적당한 길이를 갖는 문장들로 개조(재조립)하였다. 다시 말하면 유형별 문장들의 수가 기준치를 넘었을 때 수를 조정하는 방법에 유사문장 통합과 문장 개조(분리와 합성)를 병행하여 시행할 수 있다. 따라서 문장정리는 ①유사문장 통합, ②긴 문장 분리, ③짧은 문장 합성으로 이루어졌으며 아래에 그 방법과 실용 예들의 표본을 나열하였다.

1) 유사 문장의 통합 : 문장가중치는 명사, 용언 등 의미어의 숫자로 합산한다.
   (1) 동일한 역할 속성을 갖는 문장들이 여러 개 존재할 때 통합을 시도한다.
   (2) 유사성의 판단은 Moens 등과 Schutze의 유사도 공식으로 수행한다. [12]
   (3) 유사하다고 판단된 문장들 중에서 아래의 선택 예와 같이 문장가중치가 큰 것 순으로 선택한다.

선택 예)

문장1 : B C D(갖고 있는 단어나 절 종류와 수)-->탈락
문장2 : A B C D(〃)-->선택

2) 긴 문장의 분리 : "연접과 인접 및 흡착, 부가"와 "용언 및 명사화 삽입절 삭제'와 '불필요 부분 삭제'를 활용하였다.
   (1) '고'로 연접되어 있으면 〈수정 예1〉과 같이 '고' 전후를 분리한다.

<수정 예 1> 이러한 지식관리시스템은 기존의 도서관 서비스의 확장된 영역으로 T 대학의 특성

12) Moens 등과 Schutze는 각각 텍스트의 단락과 단어벡터의 유사성을 측정하는데 다음과 같은 코싸인계수를 제시하였다. 여기서 Vi는 단락 또는 단어 벡터 V, Wi는 단락 또는 단어 벡터 W를 뜻하며 따라서 같은 의미어(명사, 용언, 부사)의 수가 많으면 많을수록 유사성이 높아진다. 그리고 유사 기준치를 넘어 같은 문장이라고 판정을 받으면 문장의 가중치를 측정하여 낮은 가중치를 갖는 문장들을 탈락시킨다.

유사도 측정 공식: $corr(V_i, W_i) = \frac{\sum_{i=1}^{n} V_i \cdot W_i}{\sqrt{\sum_{i=1}^{n} V_i^2} \cdot \sqrt{\sum_{i=1}^{n} W_i^2}}$

화를 달성할 수 있는 전문화된 서비스를 제공할 수 있고 전환교육에 대한 형식적 지식과 함께 조직 구성원의 암묵적 지식을 축적함으로써 최신성 있는 데이터베이스를 유지할 수 있다. ->

① 이러한 지식관리시스템은 기존의 도서관 서비스의 확장된 영역으로 T 대학의 특성화를 달성할 수 있는 전문화된 서비스를 제공할 수 있다. ② 전환교육에 대한 형식적 지식과 함께 조직 구성원의 암묵적 지식을 축적함으로써 최신성 있는 데이터베이스를 유지할 수 있다.(결과)

(2) "~~, 즉", "~~ㄴ 반면"과 같이 인접형 이거나 '~~것 또한' 등 흡착내지 부가형 문장은 〈수정 예2〉와 같이 분리한다.

<수정 예2> 대학도서관이 수행해야 할 보다 적극적이고 활동적인 서비스 형태로 이용자 학술지원을 위한 지식관리시스템의 모형, 즉 T 대학의 특수교육분야 중에서 장애인의 전환교육에 관련한 지식관리시스템 모형을 설계하였다. ->

① 대학도서관이 수행해야 할 보다 적극적이고 활동적인 서비스 형태로 이용자 학술지원을 위한 지식 관리시스템의 모형을 설계하였다. ② T 대학의 특수교육분야 중에서 장애인의 전환교육에 관련한 지식 관리시스템 모형을 설계하였다.

(3) "'한다'고 말하였다" 類의 용언화 또는 명사화 삽입절은 첫 주어와 "다, ㅁ, 음, 기, 지" 이후를 생략한다.

(4) 출현빈도가 낮은 단어로 구성된 수식절과 과정절은 불필요 부분으로 〈수정 예3〉과 같이 생략한다.

<수정 예3> 본 연구는 초등학교 교사들이 개인홈페이지를 개설하고 필요에 따라 데이터베이스를 구축하며 더 나아가 데이터베이스에 있는 정보의 내용(본문)을 분석할 수 있게 문장 분석에 필수적인 어휘사전을 구축하는데 있어서 용언의 어간, 어미와 명사구에 대한 처리방식을 모아쓰기 형태로 살펴보고자 한다. ->

① 문장 분석에 필수적인 어휘사전을 구축하는데 있어서 용언의 어간, 어미와 명사구에 대한 처리방식을 모아쓰기 형태로 살펴보고자 한다.

3) 짧은 문장들의 합성 : 연접, 명사화, 인접, 부가, 흡착, 단일어, 수식구(절)들을 응용한다. 그리고 합성에 필요한 단어들의 의미범주를 〈표 7〉과 같이 설정하였으며 문맥정의 리스트는 〈그림 8〉과 같이 작성되었다.

(1) 동일한 역할과 범주를 갖는 주어나 목적어(조사로 구분)를 포함한 비슷한 크기의 문장들은 〈수정 예4〉와 같이 "'~고, ~며'(앞 문장), '~다'(뒷 문장)"와 같이 연접 합성을 한다(〈그림 8〉의 ⑦참조).

<수정 예4> 전자저널을 구독하면 이용자는 도서관을 직접 방문하지 않고도 웹을 통해 원문을 이용할 수 있다. 또한 전자저널 플랫폼이 갖고 있는 강력한 검색기능을 통해 다양하게 접근할 수 있다 ->

전자저널을 구독하면 이용자는 도서관을 직접 방문하지 않고도 웹을 통해 원문을 이용할 수 있고 전자저널 플랫폼이 갖고 있는 강력한 검색기능을 통해 다양하게 접근할 수 있다.

〈표 7〉 단어의 의미 범주 예

구분	범주 종류
물질	건물, 공간, 공기, 글통, 대상, 도구(운송도구, 예술도구, 측정도구[단위]), 동물, 매체, 무기, 무기물, 물, 상품, [생]산물, 속종, 시간, 식물, 신체/부분, 물, 유기물, 음식, 의복, 인/인명, 인공물, 자연물, 재화, 컴퓨터,
추상	가치, 감각, 계급/계층, 과정, 관계, 기술, 기관/기업, 단체/그룹, 문제/해결, 문화, 방법, 보조, 사건, 상태, 서비스, 수량, 스포츠, 예술, 언어, 역사, 영토, 이론/실제, 정부, 정신/의사, 지칭, 질병, 표준, 표징, 학문, 현상, 환경/배경/체제, 그림
동작	동작: 강의/학습, 거래, 계속, 기록, 논증/의논, 변화, 보존/투쟁, 비교/대조, 생산/소멸, 설명/이해, 수렴, 시도/마감/완성, 식사/배설, 양보/강제, 운동/이동, 요청/부여, 인식, 정리, 조정, 증감, 질문/응답, 처리, 통신, 공간시간/포용, 표현, 합성/분리
형용	형용: 존재, 색깔, 맛, 모습, 모양, 마음, 상황, 상태

(2) 동일한 역할과 범주를 갖는 주어나 목적어를 포함한 문장들 중 크기가 현저히 다르면 〈수정 예5〉와 같이 "'~는데' (앞문장 용언) '~ -다' (뒷문장)", "~ (앞문장), 즉 ~ (뒷문장)" 과 같이 인접 합성을 한다(〈그림 8〉 의 ⑱참조).

<수정 예5> 주제 범주 벡터는 학습집단 웹문서 250건을 자동으로 클러스터링한 후 수작업으로 재조정하여 생성 하였다. 최종 선정된 센트로이드 벡터는 모두 21개이다 ->

주제 범주 벡터는 학습집단 웹문서 250건을 자동으로 클러스터링한 후 수작업으로 재조정하여 생성 하였는데 최종 선정된 센트로이드 벡터는 모두 21개이다.

(3) 앞 문장의 역할이 '방법'이고 뒷 문장의 역할이 '결과'였을 때 방법문장(앞)과 결과 문장(뒤)의 연결 문맥을 문맥 리스트(〈그림 8〉의 ⑥참조)를 참고하여 〈수정 예6〉과 같이 앞문장의 용언을 '~하여' 등 흡착형으로 전환하고 뒷문장은 그대로 쓴다.

<수정 예6> 본 연구에서는 설문조사와 발견적 평가를 사용하였다. 이 연구결과를 토대로 도서관 웹사이트에 맞는 웹페이지와 웹문서의 디자인에 관한 개괄적인 원리들과 표준화된 사용성 평가법 및 구체적인 지침을 개발하고자 한다. ->

본 연구에서는 설문조사와 발견적 평가를 사용하여 도서관 웹사이트에 맞는 웹페이지와 웹문서의 디자인에 관한 개괄적인 원리들과 표준화된 사용성 평가법 및 구체적인 지침을 개발하고자 한다.

(4) 이웃 문장이면서 뒷 문장이 "그러나, 반면"과 같은 상반 접속부사를 가진 경우는 〈수정 예7〉과 같이 앞 문장 용언을 "지만, ㄴ 반면에" 등으로 수정한 인접 연결을 한다. (〈그림 8〉의 ⑮참조)

<수정 예7> 인터넷으로 인하여 인간은 시간과 공간을 초월하여 전자우편, 채팅, 정보검색, 가상교육 등을 자유롭게 할 수 있게 되었다. 그러나 인터넷은 장소의 한계를 완전히 극복하지는 못하였다 ->

인터넷으로 인하여 인간은 시간과 공간을 초월하여 전자우편, 채팅, 정보검색, 가상교육 등을 자유롭게 할 수 있게 되었지만 장소의 한계를 완전히 극복하지는 못하였다.

① ~ -을 위하여 ~ -을 하고자 한다〈목적〉 ② ~ -을 위한 ~ -을 하고자 한다〈목적〉
③ {본고는, 본 연구는, 본 논문은} ~ 것{이, 을} 목적{이다, 으로 한다} 〈목적〉
④ {본고{는, 에서는}, 본 연구는, 본논문은} ~ -을 목표로 {한다, 삼는다, 삼고자 한다}〈목적〉
⑤ ~ -때, ~ -를 ~ -고자 한다〈목적〉 ⑥ ~ -하여 (이 연구결과를 토대로) ~〈방법결과〉
⑦ ~ -고 (또한) ~ 〈일반〉 ⑧ 그 결과 ~ -가 있다는 것을 -다〈결과〉
⑨ [방법절] ~ -이 -다 〈결과〉 ⑩ ~ -되는 것{을, 이} -{하였, 되었]다 〈결과〉
⑪ ~ -다고 사{려, 료}된다 〈결론〉 ⑫ ~ -을 위해서는 보다 ~ -하다고 사료된다〈결론〉
⑬ ~ -하고 ~ -위한 ~ -이 필요할 것이다 〈결론〉 ⑭ -{하다고, 된다고} ~ -되었다〈결론〉
⑮ ~ -지만 (그러나) ~ 〈일반〉 ⑯ [범주] ~ (그러나 [범주]+[조사]) ~ 〈일반〉
⑰ ~ㄴ 반면에 ~ 〈일반〉 ⑱ [범주] ~ -는데 ~ [범주] ~ 〈일반〉

〈그림 8〉 문장의 문맥리스트 예

인쇄저널은 별도의 이용교육이 필요없고 브라우징이 편리하며 지적소유권 문제도 크게 신경 쓸 것이 없다. 반면 전자저널은 DB사의 정책에 따라 전자저널로 발행될 수도 있고 중단될 수도 있어 매우 유동적이다 -〉

인쇄저널은 별도의 이용교육이 필요없고 브라우징이 편리하며 지적소유권 문제도 크게 신경 쓸 것이 없는 반면 전자저널은 DB사의 정책에 따라 전자저널로 발행될 수도 있고 중단될 수도 있어 매우 유동적이다.

상기한 예와 같이 응집성을 갖는 여러 문장이 나열될 때 목적, 방법, 결과, 결론 등 〈그림 8〉과 같은 문맥정의 리스트에서 적당한 문맥형식을 참조하여 축약케 한다.

6. 시스템 평가

본 시스템이 작성한 발췌문/요약의 성능을 평가하는데 있어서 전거로 사용하는 것은 수정된 저자초록이다. 표본 문헌의 저자초록의 구조에 따라 자동 발췌문/요약의 구성도 달라진다. 다시 말하면 저자초록의 구조가 "목적(1문장), 방법(1문장), 결과(1문장), 결론(1문장)"이면 자동 발췌문/요약도 "목적(1문장), 방법(1문장), 결과(1문장), 결론(1문장)" 문장들이 구성된다. 또는 "배경, 목적, 결과"와 같은 구조이면 자동 발췌문/요약도 그와 같은 역할 유형 문장들로 구성한다.

시스템 성능 평가대상은 ①중요문장 발췌율과 ②발췌된 문장들이 정리된 수준, ③내용일치 정도인 세 가지를 설정하였다. 첫째, 발췌율의 평가는 문헌에서 역할별로 출력되어 나오는 문장들이 수정된 저자초록의 역할별 문장들과 내용상 일치하는가를 점검하여 그 일치된 숫자로 정확률과 재현율을 측정하였다. 둘째, 문장정리 수준의 평가는 응결성(Coherence)을, 내용일치 정도의 평가는 결속성(Cohesion)을 기준으로 삼았는데, 양자 모두 문헌정보시스템 분야에 종사하는 전문가의 판단에 의존하여 5단계 리커드척도로 집계하였다.

1) 발췌율: 발췌율 평가 종류는 사전 실험결과 뚜렷한 효과를 보였던 '역할'과 '소재' 두 경우를 합쳐 A방법으로 삼고, 여기에

'단서', 중요어', '범주'를 첨가하여 보다 정제한 것을 B방법으로 잡았다. 실험결과 〈표 8〉에서 보는 바대로 20 문헌 평균 재현율 A방법: 90%, B 방법: 92%, 평균 정확률 A방법: 77%, B방법: 86% 이었다.

2) 문장정리 수준 : 아래의 〈표 9〉에서 보는 바대로 문장정리 수준은 약 50% 정도의 만족도를 나타냈다.

3) 내용 일치 정도 : 아래의 〈표 9〉에서 보는 바대로 문장정리 수준은 약 50% 정도의 만족도를 보였지만 내용일치 면에서는 약 87.5%의 만족을 얻어 고무적이었다.

7. 결 론

이 논문에서 구축된 실험시스템의 특징은, (1) 코퍼스 기반의 각종 발췌기법들을 사전 실

〈표 8〉 발췌율 표본 (20 문헌 중에서)

문헌 번호	출현문장수		출현적합문장수		비출현적합문장수		정확률(%)		재현율(%)	
	A	B	A	B	A	B	A	B	A	B
1	5	4	4	4	0	0	80	100	100	100
3	4	5	4	4	1	1	100	80	80	80
5	5	4	3	3	1	1	60	73	73	80
7	5	4	3	3	0	0	60	73	100	100
9	5	4	4	4	0	0	80	100	100	100
11	5	4	4	4	0	0	67	80	100	100
13	4	4	3	4	1	0	73	100	80	100
14	5	5	4	4	1	1	80	80	80	80
17	5	5	4	4	0	0	80	80	100	100
19	5	4	4	4	1	1	80	100	80	80
평균	4.8	4.4	3.7	3.8	0.4	0.35	77	86	90	92

(유사하여 통합당하거나 정리를 당한 문장들은 합산에서 제외하였고 평균은 20 문헌에서 계산하였음)

〈표 9〉 문장정리와 내용일치 수준 (20문헌 중에서)

문헌 번호	100% 수준		75% 수준		50% 수준		25% 수준		0% 수준	
	정리	일치	정리	일치	정리	일치	정리	일치	정리	일치
1				0	0					
3				0			0			
5		0	0							
7				0	0					
9		0					0			
11		0			0					
13				0	0					
15		0			0					
17		0			0					
19			0	0						

(B방법으로 점검하였음)

험을 거쳐 선택된 방법을 본 시스템의 발췌기에 적용한 점, (2) 문장 역할별로 발췌기를 마련하고 발췌알고리듬은 수사 '역할'과 '소재'로 발췌를 시작(A방법)한 후, '단서', '중요어', '범주' 등의 발췌 영향요인을 적용하여 여과한 점(B방법), (3) 긴문장은 분리 및 삭제를 하고 짧은 문장들은 단어 의미범주를 포함한 문장역할 별 문맥리스트를 활용하여 합성한 점 등이다.

실험 시스템을 가동하여 중요문장 발췌율은 A방법이 재현율 90%, 정확률 77%, B방법이 재현율 92%, 정확률 86%인 것으로 측정되었다. 그 결과로 볼 때 "단서, 중요어, 범주" 등 발췌 영향요인들은 정확률을 상승시키는 효과적인 수단임을 알 수 있다. 한편 '역할'과 '소재' 방법을 똑같이 적용한 본 실험에서 재현율이 사전 실험에서 보다 저조한 것은 본 실험이 문장 발현 숫자를 5개 이하로 제한하였기 때문이다. 비출현 적합문장은 90%가 '배경'에 관한 문장들이었다. 그리고 정확률과 재현율이 동시에 상승한 것을 볼 수 있는데 이것은 정확률과 재현율은 반비례 관계에 있다는 통설을 깬 것이다. 그 이유는 통합을 당하거나 정리를 당한 문장들을 본문의 적합문장 수에서 제외시켰기 때문에 재현율이 낮아질 수가 없었다.

문장정리의 응결성은 50%의 만족도를, 내용일치의 결속성은 87.5%의 만족도를 나타내었다. 내용일치의 만족도가 87.5%인 것은 발췌 정확률과 재현율이 각각 86%와 92%인 점을 감안하면 그 정도의 평가를 예측할 수 있는 것이다. 여기서 문장정리의 응결성이 보통인 것은 아직 문맥정의 리스트의 정리 문맥이 완전하지 않다는 방증이며 또한 인간과 같은 문장 작성 추리력을 갖는 알고리듬의 개발을 필요로 하는 증거이다.

참 고 문 헌

이재윤. 1993. 『동적 시소러스의 구축에 관한 실험적 연구』, 연세대학교 대학원 석사학위 논문.

이태영. 1992. 『한국어 초록 작성의 자동화에 대한 연구-미생물학분야 학술지의 논문을 대상으로-』, 연세대학교 대학원 박사학위 논문.

최상희. 2004. 『질의응답을 위한 복수문서 요약에 관한 실험적 연구』, 연세대학교 대학원 박사학위논문.

최인숙. 2000. 『술어기반 문형정보를 이용한 자동요약시스템에 관한 연구』, 연세대학교 대학원 박사학위 논문.

Alone, C., M. E. Okurowski, J. Gorlinsky, and B. Larsen. 1999. "A Trainable Summarizer with Konwledge Acquired from Robust NLP Techniques", quoted in I. Mani and M.T. Maybury (eds.). 1999. *Advanced in Automatic Text Summarization*. Cambridge, Massachusetts: the MIT Press.

Barzilay, R. and M. Elhaadad. 1997. "Using Lexical Chains for Text Summarization". In *Proceedings of the Work-*

shop on Intelligent Scalable Text Summarization at the ACL/EACL Conference, 2-9. Madrid, Spain.

Boguraev, B. and C. Kennedy. 1997. "Salience-based Content Characterization of Text Documents", In *Proceedings of the Workshop on Intelligent Scalable Text summarization at the ACL/EACL Conference, 2-9.* Madrid, Spain.

Brandow, R., K. Mite, and L. Rau. 1995. "Automatic condensation of Electronic Publications by Sentence Selection." *Information Processing & Management.* 31(5): 675-685.

Chowdhury, G. G. 1999. *Introduction to Mordern Information Retrieval.* London: Library Association Publishing.

Earl, L. L. 1970. "Experiments in Automatic Extracting and Indexing." *Information Storage & Retrieval,* 6(4): 313-334. quoted in F. W. Lancaster. *Indexing and Abstracting in Theory and Practice.* London: 1998. 270.

Edmundson, H. P. 1969. "New Methods in Automatic Extracting." *Journal of ACM,* 16(2): 377-391. quoted in F. W. Lancaster. *Indexing and Abstracting in Theory and Practice.* London: 1998. 269.

Hirst, G. and D. ST-Onge. 1998[to appear]. "Lexical Chains as representation of context for the detection and correction of malapropisms". In Fellbaum, C., ed., *WordNet: An Electronic Lexical Database and Some of its Applications.* Cambridge, MA: The MIT Press.

Hovy, E. and C. Lin. 1999. "Automated Text Summarization in SUMMARIST", In *Proceedings of the Workshop on Gaps and Bridges in NL Planning and Generation, 53-58. ECAI Conference.* Budapest, Hungary.

Jones, K. S. 1999. "Automatic summarizing: factors and directions", quoted in I. Mani and M.T. Maybury(eds.). 1999. *Advanced in Automatic Text Summarization.* Cambridge, Massachusetts: the MIT Press.

Kupiec, J., J. Pedersen, and F. Chen. 1995. "A Trainable document summarizer". *Proceedings of the Eighteenth Annual International ACM Conference on Research and Development in Informaton Retrieval (SIGIR), 68-73.* seattle, WA.

Li, W., K-F. Wong, and C. Yuan. 2001. "Toward Automatic Chinese Temporal Information Extraction." *JASIST,* 52(9): 748-62.

Llorens, J., M. Velasco, A. de Amescua, J. A. Moreiro, and V. Martinez. 2004. "Automatic Generation of Domain Representations Using Thesaurus Structures", *JASIST,* 55(10): 846-858.

Mani, I. and M. T. Maybury(eds.). 1999.

Advanced in Automatic Text Summarization. Cambridge, Massachusetts: the MIT Press.

Mani, I. 2001. *Automatic Summarization*. Amsterdam: John Benjamins Publishing Company.

McKeown, K., J. Robin, and K. Kukich. 1995. "Generating Concise Natural Language Summaries", *Information Processing and Management: an International Journal*, 31(5): 703-733.

Meadow, C. T., B. R. Boyce, and D. H. Kraft. 2000. *Text Information Retrieval Systems*. San Diego: Academic Press. 208-211.

Moens, M-F., C. Uyttendaele, and J. Dumortier. 1999. "Abstracting of Legal Cases: The Pontential of Clustering Based on the Selection of Representative Objects." *JASIS*, 50: 151-161.

Myaeng, S. H. and D. H. Jang. 1999. "Development and Evaluation of a Statistically-based Document Summarization System", quoted in I. Mani and M.T. Maybury(eds.). 1999. *Advanced in Automatic Text Summarization*. Cambridge, Massachusetts: the MIT Press.

Paice, C. D. 1990. "Constructing Literature Abstract by Computer : Techniques and Prospects." *Information Processing & Management*, 26(1): 171-186.

Rush, J. E. et al. 1971. "Automatic Abstracting and Indexing. II. Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria." *JASIS*, 22(4): 260-274.

Salton, G., J. Allen, and A. Singhal. 1996. "Automatic text decomposition and structuring." *Information Processing & Management*, 32: 127-138.

Salton, G., Singhal, A., Mitra, M., Buckley, C., 1997. "Automatic text structuring and summarization." *Information Processing & Management*, 33: 193-207.

Schutze, H. 1998. "Automatic word sensec discrimination." *Computational Linguistics*, 24: 97-123.

Teufel, S. and M. Moens. 1999. "Argumentive classification of extracted sentences as a first step towards flexible abstracting", quoted in I. Mani and M.T. Maybury(eds.). 1999. *Advanced in Automatic Text Summarization*. Cambridge, Massachusetts: the MIT Press

van Dijk, T. A. 1979. "Recalling and Summarizing Complex Discourse". In W. Burchart and K. Hulker(eds.), *Text Processing Science*. 49-93, Berlin: Walter de Gruyter. quoted in I. Mani. 2001. *Automatic Summarization*. *Amsterdam*, John Benjamins Publishing Company, 139-142.