문장 클러스터링에 기반한 자동요약 모형
A Text Summarization Model Based on Sentence Clustering

정영미(Young-Mee Chung)*, 최상희(Sang-Hee Choi)**

초 록

본 연구에서는 문장 클러스터로부터 대표문장을 선정하여 요약문을 생성하는 자동요약 모형을 제시하고, 학습문서 집단을 이용하여 최적의 요약 환경을 구축한 후 요약 실험을 수행하였다. 학습 과정에서 문장의 클러스터링 기법으로는 7개의 계층적 기법들을 비교한 결과 클러스터를 구성하는 문장 수의 편차가 가장 적고 단일 문장 클러스터를 가장 적게 생성하는 센트로이드 기법이 선택되었다. 또한 각 클러스터를 대표하는 문장의 선정을 위해 용어 및 문장 가중치를 합산한 문장값과 클러스터-문장 벡터간 유사도의 두 기준을 비교한 결과 문장값 기준이 선택되었다. 용어 가중치로는 역문장빈도와 표제어 가중치, 그리고 문장의 위치 가중치가 자동요약 성능을 개선시키는 것으로 나타났으며, 적절한 요약문의 길이는 전체 문서의 1/3인 것으로 나타났다. 실험문서 집단으로는 문서의 길이와 특성이 다른 신문기사와 잡지기사의 두 집단을 이용하였다. 요약 모형의 검증 실험 결과 요약 정확률은 신문기사 집단에서는 53%, 잡지기사 집단에서는 47%인 것으로 나타났다. 두 실험 모두 랜덤하게 생성한 베이스라인 요약문보다 성능이 우수하였으나, 리드문장들로 구성된 베이스라인 요약문과의 비교에서는 짧은 길이의 신문기사의 경우 요약 모형의 성능이 오히려 떨어지는 것으로 나타났다.

ABSTRACT

This paper presents an automatic text summarization model which selects representative sentences from sentence clusters to create a summary. Summary generation experiments were performed on two sets of test documents after learning the optimum environment from a training set. Centroid clustering method turned out to be the most effective in clustering sentences, and sentence weight was found more effective than the similarity value between sentence and cluster centroid vectors in selecting a representative sentence from each cluster. The result of experiments also proves that inverse sentence weight as well as title word weight for terms and location weight for sentences are effective in improving the performance of summarization.

키워드 : 자동요약, 가중치, 용어 가중치, 위치 가중치, 문장값, 클러스터링

text summarization, term weight, location weight, sentence weight, clustering

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

** 연세대학교 문헌정보학과 시간강사

■ 논문 접수일 : 2001년 8월 20일

■ 게재 확정일 : 2001년 9월 11일

1 서 론

문서의 내용을 압축하여 표현하는 방법인 요약은 정보학 분야에서 다양한 각도로 연구되어 왔다. 요약의 목적은 원문을 읽지 않고서도 원문의 내용을 파악할 수 있도록 하거나 원문을 읽을 필요가 있는지 여부를 판단할 수 있도록 원문의 핵심내용을 간략하게 표현하는 데 있다. 문서의 자동요약에는 원문의 핵심내용을 나타낸다고 생각되는 문장을 추출하여 나열하거나 재구성하는 방법이 일반적으로 사용된다.

최근 들어 인터넷과 대용량 데이터베이스가 정보검색 환경을 주도하게 되면서 자동요약의 기능은 더욱 절실하게 요구되고 있다. 한 번에 수백 수천 건씩의 자료가 검색될 때 각 자료를 일일이 읽어볼 수 없으므로 요약을 보고 적합성을 판단하거나 원문 내용을 파악할 수밖에 없기 때문이다. 그러나 자동요약은 자동색인에 비하면 아직 다양하게 실용화되지 못한 편이다. 텍스트를 입력하면 주제어 리스트와 주요 문장을 추출하는 자동요약 프로그램인 Extractor나 보도자료를 요약하여 기사를 생성하는 시스템인 JASPER 시스템등이 개발되어 있으나 자동색인 시스템이 검색 엔진과 더불어 다양하게 상용화된 것에 비하면 그 효과는 미미한 것으로 보인다. 이는 자동요약이 복잡한 자연언어 처리과정을 수반하거나 원문의 규칙성에 의존하는 바가 크므로 대체적으로 제한된 분야를 대상으로 적용되었기 때문이다.

자동요약 기법에 대한 초기 연구들은 주로 1950년대-60년대에 이루어졌다. 이 시기의 대표적인 연구로는 통계적 기법을 제안한 룬(Luhn 1958)의 연구, 정보 소재지 기법을 제안한 박센데일(Baxendale 1958)의 연구, 그리고 주제어 기법, 표제어 기법, 소재지 기법 등에 의해 생성된 요약문들을 비교한 에드먼슨(Edmundson 1969)의 연구가 있다. 이 세 연구에서 시도된 방법은 현재 다양하게 접근하고 있는 자동요약 연구의 근간이 되었고, 현재도 기본적인 원문 분석 방법으로서 응용되고 있다.

초기의 자동요약 연구들이 용어의 출현빈도나 위치를 이용하여 단순히 요약문에 적합한 문장을 추출하였더면 이후 연구들은 자연스러운 요약문을 생성하기 위하여 문장을 수정하거나 구문분석 결과를 응용하여 새롭게 요약문을 생성하려는 시도를 하였다(Rush, Salvador, and Zamora 1971; Skorokhod'ko 1972). 또한 초기 자동요약 연구가 다양한 각도로 발전되어 가고 있을 때 새로운 접근방법인 지식기반 요약 기법을 제시한 연구들이 나타났다(Cullingford, 1981; Mckeown and Radev 1995; Paice and Jones 1993; Schank, Kolodner, and DeJong 1981).

요약하고자 하는 원문에 대한 이해는 효과적인 요약 생성의 기초가 된다는 인식에 기반하여 최근의 자동요약 연구는 원문에 대한 이해를 증진시키려는 환경에서 진행되고 있다. 특히 문서를 구성하는 각 문장간의 관계를 분석하여 문장 관계 지도를 구축하고 이를 활용하여 요약문을 작성하는 실튼 등(Salton, Singhal, Mitra, and Buckley 1997)의 연구나 학습 말뭉치를 이용하여 요약의 질을 높이고자 한 연구(Kupiec, Pedersen,

and Chen 1995; 장동현, 맹성현 1997) 등이 주목할 만하다.

본 연구의 목적은 일반인이 가장 많이 접근하는 정보인 신문과 시사잡지 기사의 요약을 작성하는 효과적인 요약 모형을 제시하는 것이다. 이 연구에서는 요약본을 구성하는 문장을 추출하는 방식으로 클러스터링 기법을 적용하였다. 이는 클러스터링 기법에 의해 기사를 구성하는 문장들을 군집화하게 되면 유사한 내용의 문장들이 동일 클러스터에 속하게 되고, 각 클러스터로부터 대표문장을 하나씩 선정하여 나열하면 전체 기사의 내용을 표현하는 요약문이 될 것이라는 가정을 전제로 한 것이다. 즉, 유사한 내용의 문장들을 하나의 클러스터에 속하게 한 후 각 클러스터를 대표하는 문장을 선정하여 요약문을 구성하는 방식을 요약 기법으로 채택하고, 수작업 학습을 통해 최적의 요약 환경을 설정한 후 요약 실험을 통해 성능을 평가하였다.

2 연구 설계

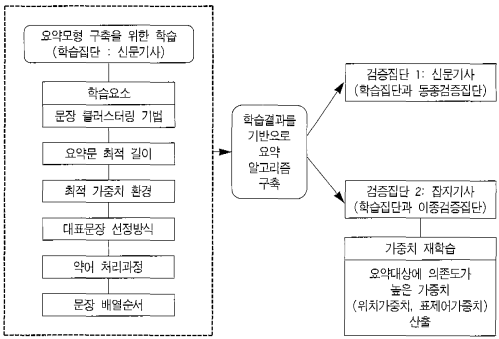
효과적인 요약 모형의 구축을 위하여 요약 대상물의 특성과 구조를 분석하여 자동화할 수 있는 규칙성을 도출하는 것이 필요하다. 이를 위해 학습집단으로부터 최적의 요약 환경을 수작업으로 학습하였는데, 학습요소는 요약문 길이, 클러스터링 기법, 대표문장 선정 방식 등이다.

요약문 길이는 대개 기사를 대표하는 주요 문장의 수를 의미하며, 기사 하나가 몇 개의 문장으로 요약될 수 있는가를 미리 학습할 필요가 있다. 기사를 대표하는 주요 문장의 수는 기사를 구성하는 문장들을 유사도에 기반하여 클러스터링할 때 생성할 클러스터의 수가 된다. 즉, 각 클러스터로부터 하나씩의 대표문장이 선정되기 때문에 클러스터 수가 요약문을 구성하는 문장 수에 해당하는 것이다.

클러스터링 기법으로는 비계층적 기법에 비해 속도는 떨어지지만 성능이 우수한 계층적 기법들이 고려되었다. 그 이유는 클러스터링 대상물이 하나의 기사를 구성하는 문장이므로 대상물의 수가 다른 응용에 비해 현저히 적어 클러스터링 속도에 큰 영향을 미칠 정도가 아니기 때문이다.

클러스터로부터 대표문장을 선정하는 방법으로는 각 문장의 중요도를 나타내는 문장 값을 산출하여 그 값이 가장 큰 문장을 선택하는 방법과 각 클러스터 센트로이드 벡터와 그 클러스터에 속한 각 문장 벡터를 비교하여 유사도가 가장 큰 문장을 선택하는 방법을 고려하였다. 문장값 기법에서는 각 문장을 구성하는 용어에 단어빈도와 역문장빈도 등의 출현빈도 가중치와 표제어 가중치가 부여되었고, 각 문장에는 다시 출현 위치에 따른 위치 가중치가 부여되어 최종 문장값을 산출하도록 하였다. 벡터간 유사도 기법에서 문장의 위치 가중치는 사용되지 않는다.

본 연구에서는 신문기사와 잡지기사의 두 가지 유형의 실험문서 집단을 사용하였다. 신문기사 집단은 모두 150건의 조선일보 기사로 구성하여 일차적으로 요약문 길이와 위치 가중치 학습을 위한 분석에 사용하였고, 여기서 학습문서 30건과 검증문서 30건을 추출하여 클러스터링 기법과 대표문장 추출 기법의



〈그림 1〉 요약 모형 구축 및 실험 과정

학습 및 검증에 사용하였다. 잡지기사 집단은 주간조선 기사 60건으로 구축하였으며, 각각 학습문서 30건과 검증문서 30건으로 구성하였다. 〈그림 1〉은 본 연구의 요약 모형 구축 및 실험 과정을 보여 준다.

3 요약 모형 구축을 위한 학습

3.1 요약문 길이

생성되는 요약문은 원문의 내용을 축약적으로 표현하면서도 원문보다는 훨씬 짧아야만 요약의 목적에 충실하다고 할 수 있다. 가장 적절한 요약문 길이를 학습하기 위하여 1500자 이하의 조선일보 기사 150개의 학습집단을 각각 500자 이하, 501-1000자, 1001-1500자의 세 그룹으로 나누어 전체 문장 수와 주요 문장 수의 비율을 조사하였다. 첫째 그룹인 500자 이하의 기사를 구성하는 문장 수는 3개부터 14개까지 다양하였지만 3-9개 사이의 문장으로 구성된 경우가 대부분을 차지하였다. 둘째 그룹인 501자부터 1000자 이하의 기사를 구성하는 문장 수는 7개부터 37개까지로 범위가 훨씬 넓었으나 기사는 8-18개 사이의 문장 수 범위에 집중적으로 몰려있는 경향을 보였다. 셋째 그룹인 1001자부터 1500자 기사를 구성하는 문장 수는 15개부터 44개까지 분포 범위가 넓었고, 15-30개 사이의 문장을 갖는 기사가 대부분이었다.

주요 문장은 기사의 내용을 이해하는 데 필수적인 기사의 주제 설명과 발생일시와 같

은 정보를 담고 있는 문장이다. 따라서 주요 문장 수는 기사 내용 전달에 필요한 최소한의 문장 수를 나타내며 또한 적절한 요약문의 길이에 해당한다고 볼 수 있다. 주요 문장은 두 명의 연구 참여자가 미리 주요 문장 선정 기준을 대략 조정한 후 각자 기사를 읽고 선정하였다.

그룹별로 기사의 평균 문장 수와 주요 문장 수를 비교한 것이 〈표 1〉에 나와 있다. 〈표 1〉에서 알 수 있듯이 500자 이하의 기사는 기사당 평균 6개 문장 가운데 주요 문장이 2개, 501-1000자 기사는 평균 14개 문장 중 주요 문장이 5개, 1001-1500자 기사는 평균 23개 문장 중 주요 문장이 8개였다. 즉, 대략 전체 문장 수의 30%에 해당하는 문장이 주요 문장으로 판정된 것이다. 이렇게 분석된 주요 문장의 비율은 요약 모형에서 생성할 요약문의 길이에 반영되었다. 즉, 전체 문장 수의 30%에 해당하는 수만큼의 클러스터를 생성한 다음 각 클러스터에서 대표문장을 추출하여 요약문을 구성하도록 한다.

3.2 문장의 클러스터링

요약 모형에 적합한 클러스터링 기법을 선정하는 일차적인 기준으로 생성된 클러스터들의 크기를 고려하였다. 왜냐하면 클러스터가 너무 불균등하게 형성되는 경우 한 클러스터내 문장들이 대표문장으로 선별되기 위해 불공정한 경쟁환경에 처할 수 있기 때문이다. 즉, 한 클러스터가 특정적으로 커지면 대부분의 주요 문장들이 한 곳에 몰리게 되고 결과적으로 주요 문장들이 대표문장으로 선정될 기회가 줄어들게 될 것이다. 또한 단일 문장으로 구성된 클러스터가 많아지면 무의미한 문장이 어느 클러스터에도 속하지 못하다가 자연적으로 클러스터 대표문장이 되는 문제점이 발생할 수 있다. 따라서 클러스터 대표문장으로 요약문을 구성하는 경우 문장들을 가능한 한 클러스터에 고르게 분배하는 기법이 적절하다고 할 수 있다.

형성된 클러스터 크기로 각 기법의 특성을 살펴보기 위해서 그룹간 평균(between groups), 그룹내 평균(within groups), 완전연결(complete linkage), 단일연결(single linkage), 미디안(median), 센트로이드(centroid), 워드(ward) 등 7개의 계층적 클러스터링 기법을 사용하여 각 기사를 구성하는 전체 문장 수의 1/3에 해당하는 수의 클러스터를 생성하도록 하였다. 문장간의 유사도 측정에는 코사인 유사계수를 사용하였으며, 문장벡터를 구성하는 용어의 가중치로는 각 기사내 출현빈도인 단어빈도를 사용하였다.

클러스터링 결과 단일연결 기법은 클러스터내 구성원들간의 연결이 단계적으로 계속 이어져 결국은 하나의 클러스터로 통합되는

〈표 1〉 신문기사의 평균 문장 수와 주요 문장 수

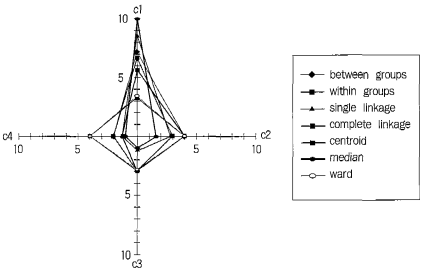
	500자 이하	501-1000자	1001-1500자
평균 문장 수	6.66	14.57	23.04
주요 문장 수	2	5	8

〈표 2〉 클러스터링 기법 비교

클러스터링 기법	표준편차	단일문장 클러스터 수
센트로이드	1.57	10
워드	1.58	12
그룹내 평균	2.51	70
완전연결	2.52	39
그룹간 평균	3.64	72
단일연결	4.42	101
미디안	4.70	118

성향을 보였고, 완전연결 기법은 클러스터내 모든 구성원이 상호 연결되기 때문에 크기가 작고 클러스터가 수가 커지는 경향을 보였다. 반면 그룹평균 기법은 클러스터나 클러스터 구성원간 유사도의 평균값을 이용하기 때문에 단일연결과 완전연결 기법의 중간정도의 특성을 나타냈다. 전체적으로 고르게 클러스터가 생성되는지 여부는 각 클러스터에 할당된 문장 수로 클러스터간 표준편차를 구하여 분석하였으며 그 결과가 〈표 2〉에 나와 있다.

클러스터링 결과를 표준편차로 분석한 결과 센트로이드 기법과 워드 기법이 가장 고르게 클러스터에 문장을 배분하는 것으로 밝혀졌고, 단일문장으로 구성된 클러스터도 가장 적게 생성하는 것으로 나타났다. 앞에서도 언급하였듯이 단일문장 클러스터는 요약 성능 평가에서 성능 저해 요인이 된다. 학습을 위한 실험 결과 표준편차가 가장 작고 단일



〈그림 2〉 클러스터 기법의 편향성

문장 클러스터를 가장 적게 생성하는 센트로이드 기법이 최적의 클러스터링 기법으로 선정되었다.

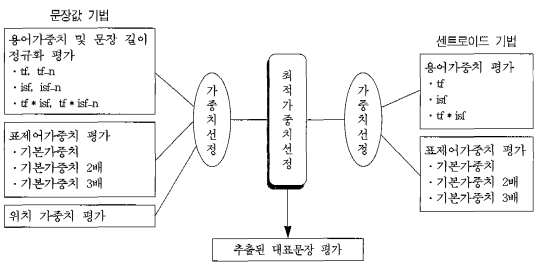
물론 클러스터 크기의 균일성에 중점을 두지 않고 생성된 클러스터 크기에 비례하여 대표문장 수를 조정하는 방법도 생각해 볼 수 있으나 이 경우 크기가 큰 클러스터의 주제가 그렇지 않은 클러스터에 비해 반드시 더 중요하다고 보기가 어렵기 때문에 이 방법은 채택하지 않았다.

학습문서의 클러스터링 결과 전체 기사의 절반 정도에 해당하는 기사가 4개씩의 클러스터를 생성하였는데 이러한 기사만을 대상으로 하여 클러스터 형성의 편향성을 살펴보았다. 〈그림 2〉는 4개의 클러스터(C1-C4)에 속한 문장 수를 기법별로 평균을 낸 것을 보여주는데 마름모 꼴에 가까울수록 클러스터에 고르게 문장을 할당하는 기법임을 의미한다. 이 분석에서도 센트로이드 기법과 워드기법이 가장 균형있게 클러스터를 생성하는 기법임이 확인되었다.

3.3 대표문장의 추출

클러스터로부터 추출한 대표문장으로 요약문이 구성되기 때문에 대표문장을 선정하는 방식은 본 요약 모형에서 매우 중요한 과정이라고 할 수 있다. 대표문장의 선정 기법으로 문장값과 클러스터-문장 벡터간 유사도의 두 기준을 평가하였다. 각 문장과 클러스터 센트로이드는 용어 벡터로 표현되는데 각 용어는 출현빈도 가중치와 표제어 가중치를 결합한 가중치를 갖는다. 출현빈도 가중치로는 단어빈도, 역문장빈도, 그리고 두 빈도의 결합 가중치가 사용된다.

문장값 기법은 문장을 구성하는 용어들의 가중치를 더한 다음 문장의 출현 위치에 따라 위치 가중치를 부여하여 최종 문장값을 산출한 다음 가장 값이 큰 문장을 선정한다. 클러스터-문장 벡터간 유사도는 각 클러스터



〈그림 3〉 대표문장 추출 기법 성능 평가 과정

센트로이드 벡터와 문장 벡터들을 비교하여 산출하며, 유사도 값이 가장 큰 문장을 대표 문장으로 선정한다.

〈그림 3〉은 대표문장 추출 방식과 최적의 가중치 환경을 학습하기 위한 과정을 보여 준다.

3.3.1 가중치 환경 설정

(1) 용어의 가중치

클러스터링 대상이 되는 기사의 각 문장은 용어들의 벡터로 표현된다. 이때 각 용어는 기사내 출현빈도에 근거한 가중치를 갖게 되는데 이 연구에서는 ① 단어빈도(tf), ② 역문장 빈도(isf), ③ 단어빈도 * 역문장빈도(tf * isf)의 세 가중치를 고려하였다. 역문장빈도는 스파크존스의 역문헌빈도를 응용한 것으로 다음 공식을 사용하였다. 이 공식에서 N은 기사를 구성하는 전체 문장 수, SFi는 용어 i가 출현한 문장의 수를 의미한다.

$$\text{역문장 빈도} = \log 2\frac{N}{SFi} + 1$$

제목은 본문의 내용을 함축적으로 표현하고 있는 경우가 일반적이다. 특히 기사인 경우에는 지면에 제한이 있기 때문에 짧은 길이의 어구로 기사를 명확하게 인지시킬 수 있는 제목을 갖는 경우가 많다. 제목을 구성하는 표제어는 명사형 단어로서 기사의 주제를 나타내는 주제어 역할을 하기도 한다. 본 연구에서는 표제어에 가중치를 주는 것이 기사의 주요 문장을 추출하는 성능을 얼마나 향상시키는지를 분석하고 최적의 표제어 가중치를 찾고자 하였다. 표제어 가중치는 기본 가중치, 2배 가중치, 3배 가중치로 변화시킴으로써 표제어 가중치가 주요 문장 추출에 미치는 효과를 평가하였다.

기본 표제어 가중치는 학습집단에서 표제어를 포함한 문장이 수작업으로 생성한 요약문에 포함될 확률과 일반 문장이 요약문에 포함될 확률의 비율로 산출하였다.

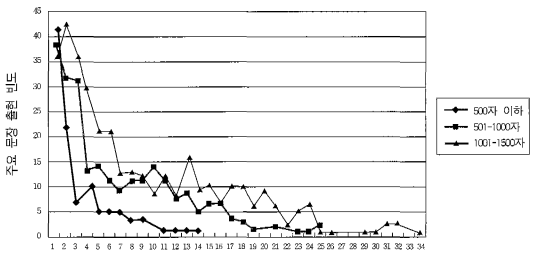
$$\text{표제어가중치} = \frac{\text{표제어가 포함된 문장이 요약문에 포함될 확률}}{\text{일반문장이 요약문에 포함될 확률}}$$

$$= \frac{0.65}{0.33} \approx 1.96$$

(2) 문장의 위치 가중치

자동요약 연구에서는 소제지 기법을 적용하여 문서의 특정 위치에 나타나는 문장에 가중치를 부여한 경우가 많았다(Edmundson 1969; Kupiec, Pedersen, and Chen 1995; 이태영 1992). 문서의 특정 위치와 주요 문장 출현과의 상관관계를 살펴보기 위하여 신문기사 학습집단을 각각 50건씩 500자 이하, 501-1000자, 1001-1500자의 세 그룹으로 나누어 주요 문장의 출현위치를 조사하였다. 이는 길이가 짧은 기사들은 보도기사에 해당하는 경우가 많고 길이가 긴 기사는 대부분 해설성 기사이므로 이러한 기사 유형에 따라 기사의 구조가 달라지며, 이에 따라 주요 문장 출현 위치가 달라질 수 있다는 신문기사의 구조적 특성을 고려한 것이다.

〈그림 4〉는 세 그룹의 기사에서 주요 문장이 어느 위치에 많이 출현했는가를 보여 준



〈그림 4〉 기사 자수별 주요 문장 출현 위치 분석

다. 이 그림은 전체적으로 기사의 도입부에 가장 많은 주요 문장이 출현하고 기사 중반 이후에서는 거의 출현하지 않는다는 점에서 공통점을 보이고 있다. 그룹별로 보면 500자 이하의 기사에서는 기사의 첫 문장이 주요 문장이 되는 경우가 대부분이었고, 세 번째 문장부터는 주요 문장이 되는 경우가 매우 적었다. 501-1000자 길이의 기사에서도 첫 문장이 가장 지배적인 출현 위치로 밝혀졌으나 두 번째 문장과 세 번째 문장이 주요 문장인 경우도 많았다. 1001-1500자 기사에서는 1-4번 문장이 주요 문장이었으며 다른 두 그룹과는 달리 두 번째 문장이 가장 주요 정보를 많이 전달하고 있는 것으로 분석되었다.

신문기사에서 주요 문장의 출현빈도와 문장 위치간 상관관계를 분석한 결과 기사의 주요 문장은 기사의 도입부인 1-3번 문장에서 집중적으로 나타남을 알 수 있다. 특히 500자 이하의 기사는 첫 문장이 주요 문장이 되는 경우가 80%에 해당하며, 500자 이하 기사의 평균 주요 문장 수가 2개인 점을 감안하면 1-2번 문장이 기사의 핵심 내용을 대부분 포함하고 있다. 따라서 500자 이하의 기사를 대상으로 자동요약 알고리즘을 적용하는 것은 무의미하며, 자동요약의 효율성이 나타나는 기사는 501-1500자 길이의 기사라고 할 수 있다.

위의 학습결과에 근거하여 각 문장의 출현 위치에 따라 문장의 위치 가중치를 다음과 같이 산출하였다. 앞에서 요약문 길이는 전체 문장 수의 1/3로 학습되었으므로 일반 문장이 요약문에 포함될 확률은 0.33이 된다.

$$\text{위치가중치} = \frac{\text{특정위치의 문장이 요약문에 포함될 확률}}{\text{일반문장이 요약문에 포함될 확률}(=0.33)}$$

특정 위치별로 위치 가중치를 산출하기 위해 먼저 각 위치의 문장이 요약문에 포함될

〈표 3〉 문장의 위치 가중치

	요약문에 포함될 확률	위치 가중치
1그룹	0.7	0.7/0.33 = 2.12
2그룹	0.4	0.4/0.33 = 1.21
3그룹	0.2	0.2/0.33 = 0.6

확률을 계산한 결과, 유사한 확률을 나타내는 그룹이 3개로 나타났다. 즉, 1-3번 문장, 4-6번 문장, 7-16번 문장의 세 그룹으로 구분되었다. 각 그룹별로 위치 가중치를 산출한 결과 각 그룹에 속하는 문장의 위치 가중치는 〈표 3〉과 같이 1그룹과 2그룹이 각각 다음 그룹의 2배 정도의 값을 보였다. 이 분석결과 1그룹과 2그룹에만 위치 가중치를 적용하였다.

3.3.2 대표문장 추출 기법 선정

3.3.2.1 요약 성능 평가척도

최적의 대표문장 추출 기법을 찾아내기 위하여 먼저 각 기법별로 최적의 가중치를 찾아내는 실험을 실시하였다. 즉, 최적의 가중치 환경을 먼저 설정한 후 두 추출 기법을 비교하여 최적의 기법을 선정하고자 하였다.

대표문장을 추출한 후 요약문을 생성하는 요약 모형의 성능은 주요 문장 추출 정확률과 가독성의 두 척도를 사용하여 평가하였다. 정확률은 생성된 요약문에 미리 수작업에 의해 주요 문장으로 판정한 문장이 얼마나 포함되어 있는가를 평가하는 것이며, 가독성은 요약문의 문맥을 보고 요약문이 얼마나 자연스럽게 작성되었는가를 판정하는 것으로서 실험 참여자들이 직접 생성된 요약문을 읽고 0-3점 사이의 점수를 주어 평가하도록 하였다.

가독성 평가에 있어서 흥미있는 사실은 평가자가 원문을 먼저 읽었는지의 여부에 따라 점수가 달라진다는 것이다. 이 실험에서 두 사람이 가독성 평가에 참여하였는데 한 사람은 미리 원문을 읽고, 다른 사람은 읽지 않은 상태에서 요약문을 평가하였다. 그 결과 미리 요약문을 읽은 사람의 평균 점수는 1.7인데 비해 읽지 않은 사람은 2.5점의 높은 점수가 나왔다. 두 번째 평가자가 원문을 읽게 한 다음 재평가를 하게 한 결과 평균 점수는 1.9로 낮아졌다. 원문을 읽고 평가하는 경우 요약문에 포함되어 있는 문장보다 더 중요한 문장이 원문에 있다는 것을 이미 알고 있으므로 가독성을 더 낮게 평가하게 되는 것이다. 여기에서 주목할 점은 두 번째 평가자가 처음 요약문을 평가한 환경이 일반 이용자들이 요약문을 접하는 환경과 같다는 것이다.

본 연구에서 제안한 자동요약 기법의 성능이 어느 정도인가를 평가하기 위해 두 가지 베이스라인 요약문을 작성하였다. 하나는 난수기를 이용하여 무작위로 전체 분장 수의 1/3에 해당하는 문장을 선정하여 요약문을 생성한 랜덤베이스와 기사의 맨 앞부분의 문장을 1/3에 해당하는 수만큼 추출하여 작성한 리드베이스이다. 이 두 가지 베이스라인 요약문도 정확률과 가독성으로 평가되어 학습집단과 검증집단의 실험 결과와 비교하였다.

〈표 4〉 문장값 기법의 용어 가중치 비교

	tf	isf	tf * isf	tf_n	isf_n	tf * isf_n	b1_random	b2_lead
정확률	0.32	0.36	0.32	0.31	0.34	0.24	0.35	0.61
가독성	1.80	1.90	1.60	1.80	1.70	1.50	1.47	3.00

〈표 5〉 문장값 기법의 표제어 가중치 비교 평가

	기본가중치	기본가중치 2배	기본가중치 3배	b1_random	b2_lead
정확률	0.39	0.38	0.45	0.35	0.61
가독성	1.97	2.00	2.03	1.47	3.00

3.3.2.2 문장값 기법

(1) 단일 가중치의 사용 결과

문장값 기법의 최적화를 위하여 먼저 단어빈도, 역문장빈도, 단어빈도 * 역문장빈도의 세 가지 가중치와 문장 길이, 즉 문장을 구성하는 용어의 수로 정규화한 가중치를 사용하여 문장값을 산출한 후 그 성능을 평가하였다. 가중치를 정규화한 이유는 단순 가중치의 합을 내는 경우 많은 수의 용어로 구성된 문장이 중요도와 상관없이 문장값이 커지며 따라서 주요 문장으로 추출될 확률이 높아질 수 있기 때문이다.

〈표 4〉는 문장값 기법에 의해 주요 문장을 추출할 경우 각 가중치 유형에 따라 달라지는 요약 성능을 보여 주고 있다. 〈표 5〉에서 tf, isf, tf * isf는 각각 단순 가중치이고 tf_n, isf_n, tf * isf_n은 정규화한 가중치를 나타낸다. 또한 b1_random과 b1_lead는 베이스라인 요약문으로서 랜덤베이스와 리드베이스를 나타낸다.

〈표 4〉를 보면 정규화하지 않은 역문장빈도(isf)를 용어 가중치로 사용한 경우가 정확률 0.36으로 성능이 가장 좋게 나타났다. 단어빈도와 역문장빈도를 결합한 가중치는 정규화 여부에 상관없이 특별히 좋은 성능을 보이지 않았다. 모든 용어 가중치들이 리드베이스에 비하면 거의 절반 수준에 달하는 매우 낮은 성능을 보이고 있으며 무작위로 문장을 추출한 랜덤베이스와도 큰 차이가 없다. 평가 결과 거의 모든 문장이 기사의 내용과 비슷한 수준으로 표현하는 짧은 길이의 기사에서는 오히려 앞 부분의 몇 문장을 추출하는 것이 정확률과 가독성에서 훨씬 우수한 요약문을 작성할 수 있음을 알 수 있다.

리드베이스가 높은 성능을 나타내는 것은 앞의 위치 가중치 분석 실험에서 나타났듯이 신문기사에서는 앞의 3-4개 문장이 주요 문장이 될 확률이 70-80%에 달하기 때문이다. 가독성에 있어서는 리드베이스인 경우 문장의 흐름이 자연스러워 3점 만점을 받을 수 밖에 없다.

용어의 출현빈도가 아니라 해당 용어가 표제어로 사용된 경우에 가중치를 부여하는 표

〈표 6〉 문장값 기법의 가중치 결합 결과

성능 \ 가중치	tf/표제어/위치	tf_n/표제어/위치	isf/표제어/위치	isf_n/표제어/위치
정확률	0.48	0.50	0.49	0.53
가독성	2.30	2.10	2.50	2.46

제어 가중치에 의해 문장값을 산출하여 성능을 평가해 보았다. 즉 한 문장 속에 표제어로 사용된 용어가 많을수록 그 문장값은 커지는 것이다. 표제어 가중치는 앞에서 제시한 기본 가중치의 1배, 2배, 3배를 각각 가중치로 하여 요약 성능을 평가하였다. 평가 결과 〈표 5〉와 같이 기본 가중치의 3배를 사용하였을 경우 성능이 가장 우수하였고, 출현빈도 가중치만을 사용한 경우보다 성능이 높게 나타났다.

문장의 위치 가중치만을 적용하여 문장값을 산출한 결과 정확률은 0.48, 가독성은 2.27로 나타났다. 이 성능은 출현빈도 가중치만을 사용한 경우보다 훨씬 좋고 표제어 가중치만을 사용한 경우보다도 높다.

(2) 가중치의 결합 결과

문장값 기법의 성능을 높이는 최적의 가중치를 선정하기 위하여 성능이 좋게 나타난 가중치만을 결합하여 주요 문장을 추출하여 보았다. 빈도기반 용어 가중치에서는 tf * isf가 tf와 isf를 단독으로 사용한 경우에 비해 성능이 높지 않았고 특히 정규화한 결과 성능이 현저하게 낮아졌으므로 두 빈도를 결합한 가중치는 추가 실험에서 제외하였다.

표제어 가중치는 앞의 평가 결과에 따라 3배 가중치를 채택하였고, 위치 가중치는 보정하지 않은 값을 그대로 사용하였다. 빈도가중치, 표제어 가중치, 위치 가중치를 결합하여 문장값을 산출한 결과가 〈표 6〉에 나와 있다. 가중치 결합결과 〈표 6〉과 같이 각 가중치를 단독으로 사용하였을 때보다 성능이 높아진 것을 볼 수 있다. 특히 정확률은 정규화한 역문장빈도가 0.53, 가독성은 정규화하지 않은 역문장빈도가 2.5로 각각 가장 높게 나타났다. 따라서 문장값 기법에 의해 대표문장을 추출하는 요약 실험의 최적 가중치 환경으로는 정규화한 역문장빈도 가중치, 3배 표제어 가중치, 위치 가중치 등 세 가중치의 결합가중치를 선정하였다.

3.3.2.3 벡터간 유사도 기법

벡터간 유사도 기법에서는 룰러스터의 센트로이드 벡터와 각 문장 벡터를 구성하는 용어에 부여할 최적의 가중치 환경을 찾아내는 일이 필요하다. 각 용어에는 tf, isf, tf * isf의 세 가지 빈도기반 가중치와 표제어 가중치가 적용되었다. 벡터들을 비교하여 유사도를 산출하는 이 기법에서는 위치 가중치는 적용되지 않으며, 빈도기반 용어 가중치에 있어서도 문장을 구성하는 용어의 수가 미치는 영향이 크지 않으므로 문장길이 정규화는 적용하지 않았다.

실험 결과 〈표 7〉에서와 같이 가장 성능이 좋게 나타난 출현빈도 가중치는 역문장빈도 가중치로서 다른 가중치에 비해 월등히 좋은 성능을 보이고 있다. 표제어 가중치도 문장값

〈표 7〉 벡터간 유사도 기법의 가중치 비교 평가

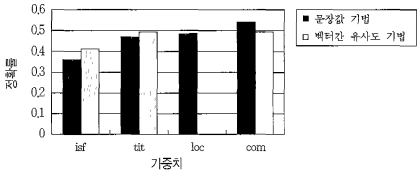
	tf	isf	tf * isf	표제어 기본가중치	표제어 기본가중치 2배	표제어 기본가중치 3배	b1_random	b2_lead
정확률	0.29	0.41	0.32	0.39	0.46	0.49	0.35	0.61
가독성	1.63	2.23	1.83	2.20	2.30	2.40	1.47	3.00

기법과 동일하게 기본 가중치를 1, 2, 3배하여 성능을 평가한 결과 기본 가중치의 3배 값이 가장 좋은 성능을 보였다. 출현빈도가중치 중 가장 성능이 좋은 역문장빈도 가중치와 3배 표제어 가중치를 결합하였을 경우의 성능은 정확률이 0.49, 가독성이 2.4로서 단일 가중치를 사용하였을 때에 비해 성능이 향상되었다.

3.3.2.4 대표문장 추출 기법 비교 평가

〈그림 5〉는 각 대표문장 추출 기법에서 적용한 다양한 가중치별 요약 성능을 보여 주고 있다. 즉, 문장값 기법에서는 역문장빈도 가중치(isf), 표제어 가중치(tit), 위치 가중치(loc), isf-tit-loc 결합 가중치(com)를 사용했을 때의 성능을 나타내며, 벡터간 유사도 기법에서는 역문장빈도 가중치, 표제어 가중치, isf-tit 결합 가중치를 사용했을 때의 성능을 나타낸다.

문장값에 의한 대표문장 추출 실험에서 발견한 최적의 가중치 환경은 정규화 역문장빈도 가중치, 표제어 가중치, 위치 가중치를 결합한 것이며, 벡터간 유사도 기법에서는 역문장빈도 가중치와 표제어 가중치를 결합한 경



〈그림 5〉 가중치별 대표문장 추출 성능 비교

〈표 8〉 대표문장 추출 성능 평가

기법 / 성능	문장값 기법 (isf_n / 표제어 3배 / 위치 가중치)	벡터간 유사도 기법 (isf_n / 표제어 3배)	b1_random	b2_lead
정확률	0.53	0.49	0.35	0.61
가독성	2.46	2.40	1.47	3.00

우로 나타났다. 각각 최적의 가중치 환경에서 두 기법의 성능을 비교한 결과 〈표 8〉과 같이 문장값 기법이 정확률 0.53, 가독성 2.46으로 백터간 유사도 기법에 비해 정확률이나 가독성에서 모두 더 좋은 성능을 보이고 있음을 알 수 있다.

문장값 기법은 평가 기준이 되는 베이스라인 요약본인 랜덤베이스보다는 월등히 좋은 성능을 보이고 있으나 아직도 리드베이스에는 미치지 못하고 있다. 그러나 각 가중치를 단독으로 사용하였을 때는 리드베이스에 비해 성능이 현저히 낮았지만 최적의 가중치 환경에서는 성능이 크게 향상되어 리드베이스에 가까운 성능을 보이는 것으로 나타났다.

4 요약 모형 검증 실험

4.1 신문기사 검증집단

4.1.1 요약 모형

학습집단을 대상으로 한 실험 결과 문장 클러스터링 기법으로 센트로이드 기법이 선정되었고, 대표문장 추출에는 억문장빈도 가중치, 표제어 가중치, 위치 가중치 등 세 가중치의 결합 가중치를 사용한 문장값 기법이 선정되었다. 이 요약 모형의 적용 과정은 〈그림 6〉과 같으며, 이 모형을 검증하기 위하여 신문기사 30건과 잡지기사 30건이 각각 검증집단으로 사용되었다.

4.1.2 요약 모형의 검증

구축된 요약 모형을 검증하기 위해 조선일보 정치/경제면 기사 중 500-1500자에 해당하는 기사 30건을 처리하여 결과를 평가하였다. 이 범주에 해당하는 기사 가운데 글머리형 기사와 인터뷰/대담 기사는 글머리표로 세부사항이 나열되거나 인터뷰 형식으로 질문과 응답이 진행되기 이전인 기사 도입부의 2-3 문장이 기사의 내용을 요약한 주요 문장인 경우가 90% 이상이었다. 따라서 이런 기사 유형은 특수한 기사 유형으로 처리하여 도입부의 문장을 요약 문장으로 추출해내는 것이 바람직하므로 요약 대상에서 제외하였다.

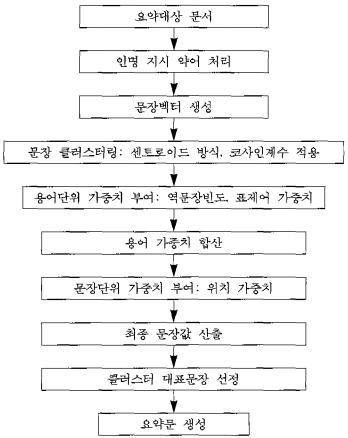
신문기사 검증집단을 대상으로 실험한 결과 요약성능은 〈표 9〉에 나와 있는 것과 같이 정확률 0.53, 가독성 2.28로서 리드베이스의 성능을 능가하지는 못하지만 랜덤베이스보다는 우수한 것으로 나타났다. 이 결과는 학습집단에 대한 실험결과와 매우 유사하다.

4.2 잡지기사 검증집단

신문기사의 경우 시작 부분의 문장들이 요

〈표 9〉 신문기사 검증집단 실험 결과

	신문기사	b1_random	b2_lead
정확률	0.53	0.41	0.60
가독성	2.28	1.40	3.00



〈그림 6〉 요약 모형 실행 과정

약문을 대치할 수 있을 정도로 기사의 핵심 내용을 전달하는 경우가 많았는데 기사의 길이가 짧을수록 이러한 특성이 현저하게 드러났다. 반면 길이가 긴 기사인 경우 기사의 핵심 내용을 나타내는 주요 문장이 기사의 여러 부분에 분산되는 현상이 나타났다. 즉, 길이가 짧은 기사인 경우는 앞부분만 읽어도 기사의 내용을 파악할 수 있지만 길이가 긴 기사는 기사 전체를 읽어야 기사의 내용을 제대로 파악할 수 있으므로 길이가 짧은 기사보다 긴 기사의 자동요약에 대한 필요성이 더 크다고 할 수 있다.

요약문 작성에 있어서 리드부분의 문장에 의존도가 적은 문서집단을 대상으로 요약 모형을 검증하기 위하여 시사잡지인 주간조선의 기사를 60건 선정하였다. 수집한 기사 중 30건은 문서 특성에 영향을 많이 받는 위치 가중치와 표제어 가중치를 학습하는 학습집단으로 사용하고 나머지 30건은 요약문 생성을 위한 검증집단으로 사용하였다. 또한 적절한 요약문의 길이가 실험집단의 특성에 따라 변화하는지 여부도 확인하였다.

4.2.1 요약문 길이 및 가중치의 재학습

본 연구의 요약 모형 구축을 위해 사용한 신문기사 학습집단에서 나타난 기사당 평균 주요 문장 수는 총 문장 수의 1/3에 해당하였다. 잡지기사 학습집단 분석 결과 평균 주요 문장 수는 총 문장 수의 2/7로 나타났다. 이 길이는 신문기사의 1/3 길이보다 약간 짧은 길이지만 큰 차이는 보이지 않고 있다. 따라서 앞에서 구축한 요약 모형에서 최적화한 요약문의 길이는 적절한 것으로 판명되었다.

표제어 가중치와 위치 가중치는 학습집단에서 특정 문장이 요약문에 포함될 확률을 기반으로 한 것으로서 학습대상 집단이 달라졌을 경우 확률 값이 달라질 수 있으며 또한 가중치 값도 달라지게 된다. 따라서 잡지기사에 적합한 표제어 가중치와 위치 가중치를 산출하려면 잡지기사로 구성된 학습집단을 이용하는 것이 합리적일 것이다. 물론 가중치 값의 산출에는 이미 구축한 요약 모형에서 사용한 방식을 그대로 적용한다. 그리고 기존의 실험에서 도출된 최적의 가중치 환경도 동일하게 적용하여 구축된 요약 모형의 타당성을 확인하고자 하였다.

잡지기사의 학습집단에서도 신문기사 학습집단과 동일한 방식으로 표제어 가중치를 재학습한 결과 기본 가중치가 1.91로 나타나 신문기사에서 학습된 1.96과 큰 차이가 없었다. 잡지기사의 요약 실험에서는 신문기사와 마찬가지로 기본 가중치를 3배한 표제어 가중치를 역문장빈도 가중치와 결합하여 용어 벡터를 구성하였다.

신문기사에서는 도입부의 문장이 주요 문장이 될 확률이 다른 위치에 비하여 명백하게 높은 것으로 나타났으나 잡지기사에서는 구조 분석 결과 주요 문장이 집중적으로 출현하는 지배적인 위치를 발견할 수 없었다. 잡지기사내 문장의 위치를 3개 그룹으로 분할하여 주요 문장의 출현 확률을 계산한 결과 그룹간의 차이가 3-7% 정도에 지나지 않았다. 따라서 이와 같이 주요 문장이 특정위치에 지배적으로 나타나지 않는 긴 길이의 잡지기사에서는 위치 가중치를 부여하는 것이 의미가 없는 것으로 나타났다.

4.2.2 요약 모형의 재검증

잡지기사 30건을 자동요약하기 위하여 학습 결과를 반영하여 요약 모형을 적용하였다. 즉, 센트로이드 기반 클러스터링과 결합 가중치에 의한 문장값 기법을 적용하였는데 신문기사 요약과 다른 점은 요약 대상물인 잡지기사의 구조적 특성에 따라 위치 가중치를 도입하지 않았다는 것이다.

잡지기사에 대한 요약 실험에서도 랜덤베이스와 리드베이스의 두 가지 베이스라인 요약문을 작성하였다. 두 베이스라인 요약문의

〈표 10〉 잡지기사 검증집단 실험 결과

	잡지기사	b1_random	b2_lead
정확률	0.47	0.29	0.34
가독성	2.46	1.20	3.00

성능은 신문기사에 비해 낮았으며, 특히 리드베이스의 성능은 랜덤베이스와 큰 차이가 없었다. 요약 모형에 의해 요약문을 작성한 결과 정확률은 0.47, 가독성은 2.46으로 베이스라인 요약문에 비해 월등히 우수한 성능을 보이고 있다.

5 결 론

본 연구에서 구축한 요약 모형은 문장 클러스터링을 통해 유사한 내용의 문장들을 집단화한 다음 각 클러스터의 대표문장을 추출하여 요약문을 작성하는 과정으로 구성된다. 최적의 클러스터링 기법과 대표문장 추출 기법은 학습집단을 대상으로 한 사전 실험을 통해 발견하였으며, 문장 벡터를 구성하는 용어의 가중치는 학습을 통해 최적화하였다.

사전 실험 결과 문장 클러스터링에 가장 적합한 계층적 클러스터링 기법은 센트로이드 기법이며, 대표문장 추출 기법으로는 문장 값을 사용하는 것이 벡터간 유사도를 사용하는 것보다 성능이 좋은 것으로 나타났다. 또한 최적의 요약 성능을 가져오는 빈도기반 용어 가중치는 역문장빈도 가중치였으며, 단일 가중치를 사용하는 것보다 역문장빈도 가중치와 표제어 가중치, 문장의 위치 가중치를 결합하는 것이 가장 좋은 성능을 가져왔다.

신문기사를 대상으로 한 요약 실험에서는 정확률 0.53, 가독성 2.28로서 랜덤하게 구성된 베이스라인 요약문의 성능(0.41/1.40)에 비해 우수했으나 리드 문장들로 구성된 요약문에 비해서는 성능이 낮았다. 반면에 잡지기사를 대상으로 한 요약 실험 결과는 정확률 0.47, 가독성 2.46으로서 랜덤베이스(0.29/1.20)와 리드베이스(0.34/3.00)에 비해 성능이 우수했다.

본 연구에서 밝혀진 주요 사실은 신문기사와 같은 짧은 텍스트는 에써 복잡한 요약 알고리즘에 의해 요약문을 생성하는 것보다는 기사의 앞 부분에 오는 몇 개의 리드 문장으로 요약문을 작성하는 것이 더 효율적이라는 것이다. 신문기사 250건을 대상으로 실험한 다른 연구에서도 기사 첫머리에 오는 5개 문장으로 구성한 리드베이스 요약문이 자동요약 알고리즘에 의해 생성된 요약문에 비해 좋은 성능을 보인 바 있다(Brandow, Mitze, and Rau 1995).

그러나 잡지기사의 경우에는 이 연구에서 구축한 클러스터링 기반 요약 모형이 모든 베이스라인 요약문에 비해 좋은 성능을 보이고 있다. 또한 신문기사나 잡지기사 집단에서 모두 0.5 수준의 정확률을 보임으로써 다른 요약 알고리즘들에 비해 상대적으로 우수한 성능을 나타내는 요약 모형임이 입증되었다.

요약문 작성에 있어서 특정한 위치의 문장이 주요 문장이 되는 성향이 있는 문서에서는 그 위치의 문장을 추출하는 것이 단순하고 효율적인 요약 기법이 될 수 있다. 자동요약에서는 이와 같이 특정 위치의 문장으로 작성한 베이스라인보다 성능이 좋은 요약문을 생성하는 것이 알고리즘 개발의 목표가 될 수 있을 것이다. 특히 신문기사 자동요약에 있어서는 리드문장들로 구성되는 리드베이스보다 높은 정확률을 얻을 수 있도록 요약 모형을 개선하는 것이 향후 연구과제가 되어야 할 것이다.

참 고 문 헌

이태영. 1992. 『한국어 초록 작성의 자동화에 관한 연구: 미생물학 분야 학술지의 논문을 대상으로』 박사학위 논문. 연세대학교 대학원. 문헌정보학과.

장동현, 맹성현. 1997. 자동요약시스템. 『정보과학회지』. 15(10): 42-49.

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

Brandow, Ronald, Karl Mitze, and Lisa F. Rau. 1995 "Automatic Condensation of Electronic Publications By Sentence Selection." *Information Processing & Management*. 31(5): 675-685.

Cullingford, R. E. 1981. "Sam." In *Inside Computer Understanding: Five Programs Plus Miniatures* Edited by Schank, R. C. and C. K. Riesbeck. Hillsdale, NJ: Lawrence Erlbaum.

Edmundson, H. P. 1969. "New Methods in Automatic Extracting." *Journal of the Association for Computing Machinery*. 16(2): 264-289.

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

Mckeown, K and D. Radev. 1995. "Generating Summaries of Multiple News Articles." In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 74-82).

Kupiec. J., J. Pedersen, and F. Chen. 1995. "A Trainable Document Summarizer." In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 68-73).

Paice, C. D., and P. A. Jones. 1993. "The Identification of Important Concepts in Highly Structured Technical Papers." In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 69-78).

Rush, J. E., R. Salvador and A. Zamora. 1971. "Automatic abstracting and indexing. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria." *Journal of American Society for Information Sciences*. 22(4): 260-274.

Salton, Gerard, Amit Singhal, Mandar Mitra, and Chris Buckley. 1997. "Automatic Text Structuring and Summarization."

Information Processing & Management, 33(2): 193-207.

Schank, R. C., J. L. Kolodner, and G. DeJong. 1981. "Conceptual information retrieval." In *Information Retrieval Research* Edited by Oddy, R. N., S. E. Robertson, C. J. van Rijsbergen, and P. W. William. London, Great Britain: Butterworth.

Skorokhod'ko, E. F. 1972. "Adaptive Method of Automatic Abstracting and Indexing," In *Proceedings of IFIP Congress* (pp. 1179-1182).