도합유사도를 이용한 한국어 추출문서 요약

김재훈 ° 김준홍
컴퓨터공학과, 한국해양대학교
첨단정보기술연구센터

jhoon@hanara.kmaritime.ac.kr, rainmk@nlplab.kmaritime ac kr

Korean Indicative Summarization Using Aggregate Similarity

Jae-Hoon Kim° Jun-Hong Kim
Department of Computer Engineering, Korea Maritime University
and
Advanced Information Technology Research Center

jhoon@hanara kmaritime ac kr, rainmk@nlplab.kmaritime.ac kr

요 약

본 논문에서 문서는 문서관계도라고 하는 그래프로 표현된다. 노드는 문서의 구성요소인 문장을 표현하고, 링크는 노드들 간의 의미적인 관계를 나타낸다. 의미적 관계는 유사도에 의해서 결정되며, 문장의 중요도는 도합유사도로 나타낸다. 도합유사도는 한 노드와 인접한 노드들 사이의 유사도 합을 말한다. 본 논문에서는 도합유사도를 이용한 한국어 문서요약 기법을 제안한다.

실험에 사용된 평가용 요약문서은 정보처리 관련 분야에서 수집된 논문 100 편과 KORDIC 에서 구축한 신문기사 105 건을 이용하였다. 문서요약 시스템에 의해서 생성된 요약문서의 크기가 본문의 20%이고, 본문이 논문(서론과 결론)일 경우, 재현율과 정확률은 각각 46 6%와 76.9%를 보였으며, 또한 본문이 신문기사일 경우, 재현율과 정확률은 각각 30.5%과 42.3%를 보였다. 또한 제안된 방법은 상용시스템보다 좋은 성능을 보였다.

1. 서론

가상공간(cyberspace)이라고 하는 웹은 전세계를 통하여 많은 정보를 쉽게 얻을 수 있는 정보의 보고이다. 가상공간에 존재하는 정보들은 매우 다양하며, 그 양도 매우 빠른 속도로 증가하고 있다. 방대한 정보공간에서 유용한 정보를 찾기 위해 널리 사용되는 도구가 웹 정보검색 엔진이다. 일반적으로 웹 정보검색 엔진들은 너무 많은 정보를 검색해 주기 때문에 유용한 정보를 찾는 것은 그다지 쉬운 일이 아니다. 이와 같은 정보검색 환경에서 유용한 정보를 효과적으로 찾기 위해서는 자동문서요약 기술이 자주 사용된다[1-3].

문서요약은 원문서의 의미를 유지하면서 원문서의 길이나 정보의 복잡도를 줄이는 작업이다[2]. 즉, 문서요약은 정보압축(information compression)이다. 문서요약은 일상적인 생활에서는 널리 사용되고 있는 방법이다. 예를 들면, 헤드라인 뉴스, 각종 회의의 의사록, 책이나 CD 등의 논평 등이 일상적인 생활 속의 문서요약에 대한 예이다. 최근 문서요약은 단순한 하나의 문서의 내용을 요약하는 것이 아니라 여러 문서의 내용은 하나로 요약하기도 하고, 심지어는 문서가 아닌 이미지, 오디오, 비디오와 같은 멀티미디어 정보를 요약하기도 한다[4].

인터넷의 급속한 발전과 더불어 문서요약에 대한 관심이 고조되면서 문서요약에 대한 연구개발에 대한 투자도 꾸준히 증가하고 있다. 특히 문서요약은 상업분야,

통신산업 분야(British Telecom 의 Prosum[1]), 웹 정보검색 여과기(AltaVista Discovery 에 사용되는 Inxight 사의 LinguisticX[2]), 워드프로세서(Microsoft 의 AutoSummarize), 정보검색 색인기(National Research Council 의 Extractor[3]) 등 매우 다양한 분야에서 개발되고 있다.

일반적인 문서요약 시스템은 문서분석(document analysis), 문서변환(document transformation), 문서생성(document synthesis) 단계를 거친다[5]. 문서분석은 주어진 문서를 분석해서 명사를 추출하거나 빈도수 추출하는 등의 작업을 수행하는 단계이고, 문서변환을 분석된 정보를 토대로 본문을 요약문서로 변환하는 단계이며, 문서생성은 요약문서의 가독성을 높이기 위해서 자연스러운 문장을 생성하는 단계이다.

문서요약 기법은 크게 문장이해를 기반으로 하는 인어학적 접근 방법과 단어 빈도수 등과 같은 통계정보를 기반으로 하는 통계적 접근 방법으로 나눌 수 있다. 후자는 문서 중에서 중요한 문장 혹은 단락을 추출하고 중요도에 따라 재배치하는 방법으로 문서를 요약한다. 후자에 의해 생성된 요약문서는 전자의 것에 비해 가독성이 떨어지며 최근에 가독성을 높이기 위한 연구들이 막 진행되고 있다[6].

한국어 문서요약에 대한 연구도 매우 활발히 진행되고 있다[2-3][6-7]. 그러나 아직은 성숙되지 않은 것 같다. 또한 대부분의 연구가 통계적 접근 방법을 채택하고 있으며, 여러 다양한 환경에서 평가되어 객관적으로 어떤 시스템이 좋은 성능을 보인다고 말할 수 없는 실정이다.

Salton 등은 하나의 문서를 그래프로 표현하였으며, 이를 문서관계도(text relationship map)이라고 한다[8]. 문서관계도에서 노드(*node*)는 문장(*sentence*) 혹은 단락이고, 링크(link)는 의미적으로 관련된 노드들 사이의 관계를 나타낸다. 이 관계는 노드와 노드 사이의 유사도(similarity)가 어떤 임계값(threshold) 이상일 경우를 말한다. 무성도(Bushiness)는 문서관계도에서 다른 노드와 연결된 링크수(일명 부쉬경로(bushy path)), 즉 노드의 차수(degree of a node)이며, 무성도가 높으면 높을수록 많은 다른 노드들과 연결되었음을 의미한다. 문서요약은 단락이나 문장을 무성도가 높은 순으로 재배치하는 것이다. 본 논문에서는 Salton 등의 문서관계도에서 무성도 개념[8]을 이용하였다. 그러나, 무성도는 링크수가 아니라, 유사도의 합으로 정의하였다 본 논문에서는 이를 도합유사도(aggregate similarity)라고 한다.

본 논문의 구성은 다음과 같다. 2 절에서 한국어 명사 추출 시스템에 대해서 기술하고, 3 절에서는 도합유사도에 대해서 자세히 기술한다. 4 절에서는 한국어 문서 요약 시스템을 설명한다. 5 절에서 실험 및 평가에 대해서 기술하고, 6 절에서 기존의 제안된 방법들과 비교하고 분석한다. 마지막으로 7 절에서 결론을 맺고 앞으로의 연구 방향에 대해서 논의한다.

2. 한국어 명사 추출

본 논문에서 문장은 벡터공간의 한 점으로 표현되고 이를 문장벡터라고 한다. 문장벡터는 각 명사의 빈도수에 의해서 표현된다. 따라서 문장벡터를 구성하기 위해서는 문장에 포함된 명사를 추출해야 한다. 한국어 명사추출에 관한 많은 연구는 한국어 정보검색 분야에서 많이 수행되었다[9-10]. 본 논문에서 명사추출 시스템의 구조는 그림 1 과 같으며, 다음과 같은 절차에 의해서 명사를 추출한다.

1. 사전을 이용해서 문장으로부터 수식언을 제거한다 여기서 수식언은 부사, 관형사, 감탄사가 여기에 속한다.
2. 사전과 어미 집합을 이용해서 용언(동사, 형용사)을 제거한다. 몇몇 어절(예를 들면, 나는)은 명사구와 중의성이 발생되는데 이 중의성은 무시한다. 즉 체언과 용언의 중의성이 발생되면 용언을 선호하도록 하였다.
3. 식 (1)에 정의된 음절간의 상호정보를 이용해서 명사구에서 조사를 분리한다 이 방법은 Maosong 등에 의해서 중국어 단어분리 알고리즘[11]을 약간 수정해서 사용하였다.

$$mi(x:y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

여기서 x 와 y 는 한국어의 음절이며, 여기에는 어절의 시작과 끝을 나타내는 EOW 도 포함된다. p(x)는 x 가 일어날 확률이다.

4. 수사는 오토마타를 이용해서 제거한다. 여기서 수사의 예를 들면, "1999 년", "1 천 2 백원" 등을 의미한다.
5. 복합분리는 사전과 CYK 알고리즘[12]을 이용한다. CYK 알고리즘을 직용할 경우에 중의성이 발생되는데, 이를 많은 수의 명사가 포함되는 경우를 우선하는 경험규칙을 사용하였다.

3. 도합유사도

Salton 등은는 문서를 문서관계도라고 하는 그래프로 표현하였다[8]. 본 논문에서도 문서관계도를 이용한다. 문서관계도에서 노드(node)는 2 절에서 설명한 문장벡터 S_i 을 나타내고, 링크(link)는 의미적으로 관련이 있는 노드(문장벡터)들 사이의 관계를 나타낸다. 의미적 관계는

[1] http://transend.labs.bt.com
[2] http://www.inxight.com
[3] http://extractor.iit.nrc.ca/

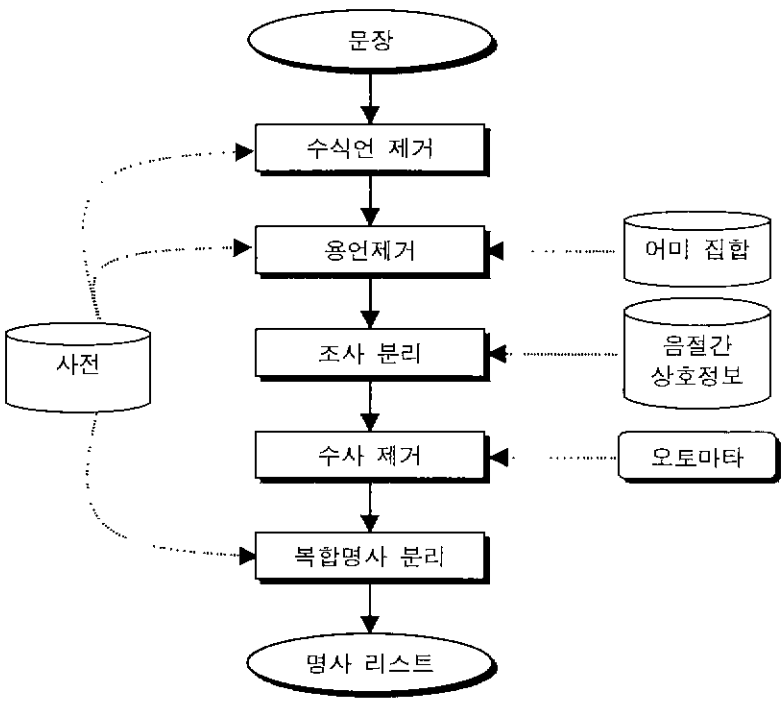

그림 1. 한국어 명사추출 시스템의 구조

유사도에 의해서 결정되며, 두 문장벡터 S_i와 S_j 사이의 유사도는 식 (2)와 같다.

$$sim(i,j) = \sum_{k=1}^{n} s_{i,k} s_{j,k} \qquad (2)$$

여기서 문장에 포함된 명사의 수는 n 개라고 할 때, 문장벡터 S_i는 $(s_{i,1}, s_{i,2}, \ldots, s_{i,n})$이고, $s_{i,k}$는 i 번째 문장에서 명사 k 의 빈도수이다. 문장의 중요도는 각 노드에 인접한 노드들과의 유사도의 합으로 정의한다. 이를 본 논문에서는 도합유사도(aggregate similarity)라고 하며, i 번째 문장에 대한 도합유사도 asim(i)는 식 (3)과 같이 정의된다. 그림 2 는 m 개의 문장으로 구성된 문서에서 도합유사도의 개념을 도식화한 것이다.

$$asim(i) = \sum_{\substack{j=1 \\ j \neq i}}^{n} sim(i,j) \qquad (3)$$

4. 한국어 문서요약 시스템

본 논문에서는 문서요약 중에서 문장추출에 해당하며, 통계적인 접근 방법을 사용한다. 문장은 명사 리스트로 표현되며, 문장 간의 유사도는 내적(inner product)을 사용한다. 문장의 중요도는 Salton 의 부쉬경로와 비슷한 도합유사도를 이용한다. 문장생성은 문장의 중요도가 높은 순으로 정렬한다. 또한 본 논문은 특별한 학습기법을 전혀 사용하지 않는다. 이와 같은 개념을 토대로 본 논문에서 제시한 한국어 문서 요약 시스템의 구조는 그림 3 과 같다.

그림 1 에서 전처리기는 입력문서를 문장 단위로 분리하고, 문장기호를 제거한다. 문장의 분리하는 기준은 기호 "!?"가 있으면 문장으로 분리하였다. 이외에서 "1999. 12." 등과 같은 문자열에 대해서 오류가 발생하기 때문

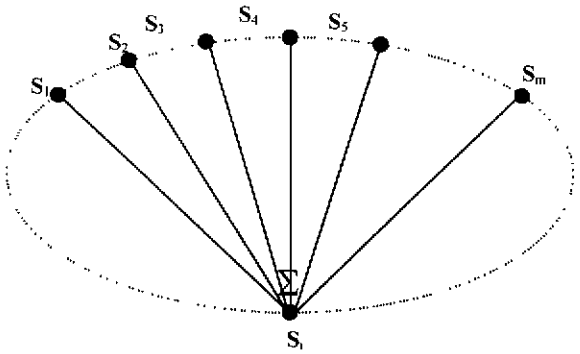

그림 2. 도합유사도의 개념도

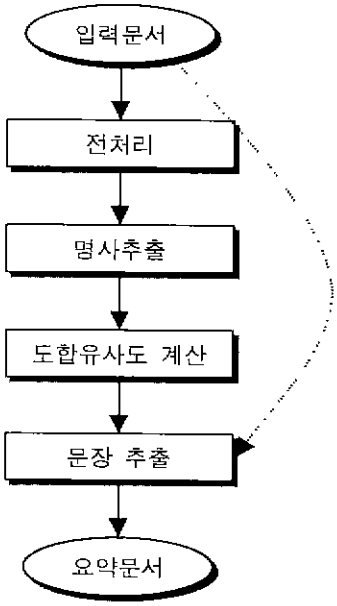

그림 3. 한국어 문서요약 시스템의 개요

에 약간의 경험규칙(heuristics)을 사용하였다. 명사추출 방법과 도합유사도의 계산 방법은 각각 2 절과 3 절에서 구체적으로 설명하였다.

문장추출은 먼저 도합유사도가 높은 순으로 문장을 재정렬한다. 그리고 나서 앞에서부터 원하는 비율만큼의 문장을 추출하여 요약문서로 출력한다.

5. 실험 및 평가

5.1. 실험 말뭉치

본 논문에서는 평가를 위해 두 종류의 평가용 말뭉치를 구축하였다. 하나는 정보처리 관련 분야의 논문 100 편(PAPER)으로 구성되었다. 이 말뭉치의 구축 방법은 Teufel 과 Moens 이 제안된 방법[13]과 비슷하게 원저자의 초록(abstract)을 이용하였다. 즉, 논문의 전체(PAPER-ALL) 혹은 서론과 결론 부분(PAPER-IntroConcl)에 속하는 문장들 중에서 원저자의 초록에 속하는 문장에 가장 유사한 문장들을 추출하여 평가용 말뭉치를 구축하였다. 이때 평가요약문서의 크기는 원저자의 초록의 크기와 같고, 유사도로 내적을 사용하였다. 또 다른 하나는 KORDIC 말뭉치의 신문기사 105 건(NEWS)이고, 이 말뭉치의 평가용 요약문서는 요약전문가에 의해서 구축되었다[14]. 이들 두 말뭉치의 통계치는 표 1 과 같다. PAPER-IntroConcl 과 NEWS 의 통계치들은 서로 비슷하며, 장르가 서로 다르다.

표 1. 평가용 말뭉치의 통계치

	PAPER - ALL	PAPER-IntroConcl	NEWS
총 문서 수	100	100	105
평균 문서 당 문장 수	113.2	20.6	21.5
평가요약문서의 평균 문장 수	5.6	5.6	6.0

5.2. 성능 평가

성능평가의 측도는 정보검색 분야에서 널리 사용되고 있는 정확률(precision)과 재현율(recall) 그리고 F 측도(f-measure)를 사용하였으며, 이들은 각각 식 (4)와 (5) 그리고 (6)과 같이 정의된다[15].

$$P = \frac{N_R}{N_S} \tag{4}$$

$$R = \frac{N_R}{N_C} \tag{5}$$

$$F = \frac{2PR}{P+R} \tag{6}$$

여기서 N_S는 문서요약시스템이 제시한 전체 문장 수이고, N_R은 N_S 중에서 평가요약문서에 속한 문장 수이고, N_C는 평가요약문서에 속한 문장 수이다. 표 2 는 본 논문에서 제안한 성능이다.

표 2. 제안된 문서요약 시스템의 성능 평가

요약문서의 크기	말뭉치	P	R	F
10%	PAPER-ALL	34.3	74.2	46.9
	PAPER-IntroConcl	85.2	33.8	48.4
	NEWS	46.8	13.3	20.7
20%	PAPER-ALL	19.7	83.3	31.9
	PAPER-IntroConcl	76.9	46.6	58.0
	NEWS	42.3	30 5	35.4

PAPER-ALL 은 본문의 크기가 크기 때문에 시스템이 생성한 요약문서의 크기도 다른 시스템에 비해서 비교적 크다. 이와 같은 이유로 다른 말뭉치와는 반대로 재현율

이 높은 결과를 가져왔다. PAPER-IntroConcl 과 NEWS 는 서로 비슷한 말뭉치의 특성을 가지고 있으나 성능은 크게 차이를 보인다. 이 원인은 크게 두 가지로 요약될 수 있다. 하나는 평가요약문서의 구축 방법의 차이이다. PAPER-IntroConcl 은 자동으로 구축되었고, NEWS 는 요약전문가에 의해서 구축되었다. 자동으로 구축된 PAPER-IntroConcl 의 평가용 요약문서는 문서요약 시스템의 특성이 일부 반영되었다고 볼 수 있다. 다른 하나는 장르의 차이이다. 기술논문에는 반복적인 표현들이 자주 발생된다는 특징을 잘 반영하는 것으로 추정된다. 좀더 전확한 원인 분석을 위해 좀 더 많은 연구가 필요하다.

5.3. 왜 내적 유사도를 사용하는가?

본 논문에서 문장 간의 유사도로 내적을 사용하였다. 내적 유사도를 사용한 특별한 이유는 없다. 다만 일반적으로 정보검색 시스템에서 널리 사용되는 코사인 유사도와의 성능을 비교했을 때, 더 좋은 결과를 가져왔기 때문이다. 표 3 은 시스템에 의해서 생성된 요약문서의 크기가 본문의 10%일 때, 내적 유사도와 코사인 유사도의 성능을 비교한 것이다. 내적 유사도는 코사인 유사도에 비해 계산이 용이할 뿐 아니라 문장의 길이에 대한 특성을 반영하고 있다

표 3 내적 유사도 대 코사인 유사도

말뭉치	유사도	P	R	F
PAPER-IntroConcl	내적	85.2	33.8	48.4
	코사인	77.6	22.3	34.6
NEWS	내적	46.8	13.3	20.7
	코사인	31.9	8.7	13.7

6. 관련 연구

6.1. 도합유사도와 부쉬경로

Salton 등은 문장의 중요도를 부쉬경로(bushy path)로 측

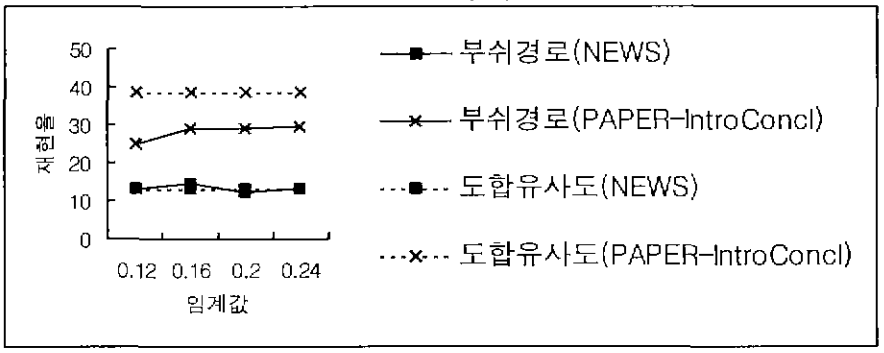

그림 4. 도합유사도와 부쉬경로의 재현율 비교

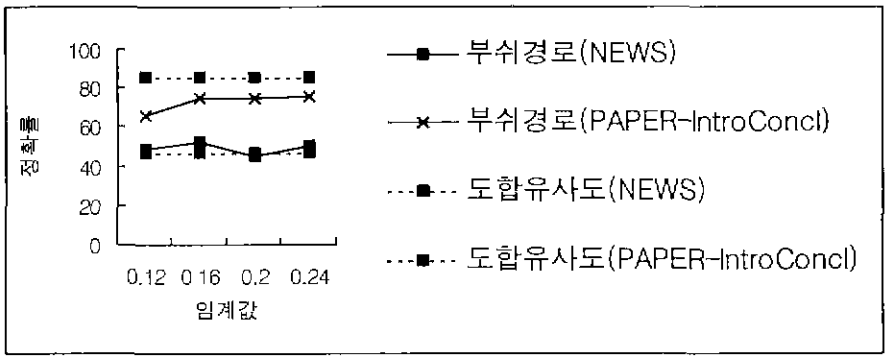

그림 5. 도합유사도와 부쉬경로의 정확률 비교

정한다[6]. 1 절에서 언급했듯이 부쉬경로는 문서관계도에서 링크 수를 나타내며, 문장 간의 유사도가 어떤 특정 임계값 이상일 때, 링크가 설정된다. 그림 4 와 5 는 생성된 요약문의 크기가 본문의 10%일 때. 임계값의 변화에 따른 부쉬경로 방법의 재현율과 정확률을 그래프로 보이고 있다 NEWS 말뭉치에서 도합유사도와 부쉬경로는 비슷한 성능(재현율과 정확률 모두)을 보였으며, PAPER-IntroConcl 에 대해서는 도합유사도가 훨씬 더 좋은 성능을 보이고 있다. 도합유사도는 부쉬경로의 임계값과 같은 매개변수를 가지고 있지 않기 때문에 실용적인 환경에 더 잘 적응할 수 있을 것이다.

6.2. 제안된 시스템과 상용시스템

독립적으로 상용되는 한국어 문서요약 시스템은 없으나, 워드프로세서(Microsoft Word 와 훈민정음) 내의 자동요약 도구를 가지고 있다. 객관적인 성능을 비교하기 위해서 본 논문에서는 이들 상용시스템과의 성능을 평가해 보았다. 표 4 는 생성된 요약문서의 크기가 본문의 10%인 경우, 제안된 시스템과 상용시스템과의 성능을 비교한 결과이다 제안된 시스템이 재현율과 정확률 모든 면에서 좋은 결과를 보였다.

표 5. 제안된 시스템 대 상용시스템

말뭉치	시스템	P	R	F
PAPER-IntroConcl	제안된 시스템	85.2	33 8	48.4
	Microsoft Word	42.1	18.8	20.8
	훈민정음	66.3	31.0	42.2
NEWS	제안된 시스템	46.8	13.3	20.7
	Microsoft Word	29.3	12.5	17.5
	훈민정음	20.5	11.3	14.6

6.3. 기존의 한국어 문서요약 시스템

본 절에서는 제안된 문서요약 시스템과 기존의 한국어 문서요약 시스템들과의 비교하고자 한다. 각 시스템들은 매우 다양한 실험 환경 때문에 객관적인 성능을 비교하는 것은 어려운 일이다. 표 6 은 기존 시스템들의 주요 특징을 나타낸 것이다. 대부분의 방법은 통계적인 방법을 이용하고 있으며, 기술문서의 일종인 논문을 중심으로 평가했으며, 평가용 요약문서의 수가 20 ~ 30 으로 비교적 작은 규모로 실험하였다 본 논문은 객관성을 높이기 위해서 약 100 여개의 문서에 대해서 논문 뿐아니라 신문기사에 대한 실험을 수행하였다.

또한 대부분의 시스템들이 학습모델을 이용하고 있다. 학습모델을 사용할 경우에는 학습되지 않은 영역에 쉽게 적용할 수 없다는 문제를 가지고 있다. 본 논문에서 제안한 방법은 학습모델을 사용하지 않으며 주어진 문장에 대한 명사의 빈도수만 필요하다.

본 논문에서 제안한 한국어 문서요약 시스템은 기존의 다른 시스템에 비해서 모델이 단순하며 쉽게 구현할 수 있고 실용적으로 사용할 수 있도록 설계 구현되었다. 그렇지만, 성능면에서도 기존의 다른 시스템보가 떨어지지 않았으며, 객관적인 성능 비교는 어려울지 모르지만 재현율과 정확률에 대해서 조금 더 좋은 성능을 보였다.

7. 결론

본 논문은 도합유사도를 이용한 문서요약 시스템을 제안하였다. 한 문장의 도합유사도는 본문 내에 있는 다른 문장들과의 유사도 합을 의미하며, 유사도로 내적을 사용하였다. 본 논문의 문서요약 방법은 단순한 모델을 사용하고 있으며, 구현이 용이하고, 쉽게 실용적으로 사용할 수 있다는 장점을 가지고 있다.

제안된 시스템을 평가하기 위해 두 종류의 말뭉치(논문, 신문기사)를 사용하였다. 시스템이 생성한 요약문의 크기가 본문 크기의 20%이고, 본문이 논문(서론과 결론)일 경우, 재현율과 정확률은 각각 46.6%와 76.9%를 보였으며, 또한 본문이 신문기사일 경우, 재현율과 정확률은 각각 30 5%과 42.3%를 보였다 또한 제안된 방법은 상용시스템보다 좋은 성능을 보였다.

통계적인 접근 방법을 이용하는 대부분의 시스템에서 생성된 요약문서는 가독성이 좋지 않다. 이를 위해서는 문서계획과 문장생성에 관한 연구가 필요로 하다. 문서요약 기술은 정보검색의 색인이나 문서분류 등의 분야에 적용하여 더욱더 질 좋은 문서검색 시스템을 구현할 수 있을 것이다.

8. 감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단과 지원을 받았으며, 또한 과학기술부 STEP2000 프로젝트에 의해 지원되고, 전문용어언어공학연구센터에 의해 수행중인 "대용량 국어정보 심층처리 및 품질관리 기술개발" 연구과제의 일환으로 수행되었습니다.

참고문헌

[1] Cowie, J., Mahesh, K., Nirenburg, S. and Zajac. R. (1998). MINDS – multilingual interactive document summarization. in *Working Notes of the AAAI Spring Symposium on*

Intelligent Text Summarization, Spring, pp. 131-132.

[2] Jang. D. and Myaeng, S.-H. (1997) Automatic text summarization systems. Korea Information Science Society Review, vol. 15, no. 10, pp.42-49.

[3] Kang, S.-B. (1997) Implementation of a summarization system using statistical information of Korean documents. Master's thesis, Department of Computer Science, Pusan National University.

[4] Mani, I. and Maybury, M. T. (1999) Advanced in Automatic Text Summarization, The MIT Press

[5] Sparck Jones, K. (1999). Automatic summarizing: factors and directions, in Mani, I. and Maybury, M. T., editors, Advances in Automatic Text Summarization, pp. 1-12. The MIT Press.

[6] Lee, M.-H., Park, M.-S., Kim, M.-J., and Lee, S.-J. (1999) Sentence extraction using document features and heading. in *Proceedings of KIPS,* vol. 6. no. 2. pp. AI41-AI45.

[7] Ryu, D.-W. and J.-H. Lee. (2000). Word co-occurrence based automatic text summarization,. in *Proceedings of KISS*. vol. 27. no. 1, pp. 345-347.

[8] Salton, G., Singhal, A., Mitra, M. and Buckly, C. (1999). Automatic Text Structureing and Summarization. in Mani, I. and Maybury, M. T., editors, Advances in Automatic Text Summarization, pp. 61-70

[9] Won, H., Park, M. and Lee. G. (2000). Integrated indexing method using compound noun segmentation and noun phrase synthesis. *Journal of KISS· Software and Applications*, vol. 27, no. 1, pp. 84-95.

[10] Yun, B.-H., Cho, M.-J. and Rim, H.-C. (1997). Segmenting Korean compound noun using statistical information and a preference rule. *Journal of KISS(B). Software and Applications*, vol. 24, no. 8, pp. 900-909.

[11] Maosong, S., Dayang, S. and Tsou, B. K (1998). Chinese word segmentation without using lexicon and hand-crafted training data. in *Proceedings of COLING-ACL* 98, pp. 1265-1271.

[12] Aho. V. A. and Ullman, J. D. (1973) "The Theory of Parsing, Translation, and Compiling, Prentice-Hall.

[13] Teufel. S. and Moens, M. (1999). Argumentative classification of extracted sentences as a first step towards flexible abstracting, in Mani, I. and Maybury, M. T., editors, Advances in Automatic Text Summarization, pp. 155-171. The MIT Press.

[14] Kim, T.-H , Park, H.-R., Shin, J.-.H. (1999). A study on text understanding model for retrieval / summarization / filtering. in Proceedings of the Workshop on Softscience.

[15] Manning, C. D. and Shűtze, H. (1999). *Foundation Statistical Natural Langauge Processing*, The MIT Press.

[16] Myaeng, S. H. and D. Jang (1999) Development and evaluation of a statistically based document summarization system. in Mani, I. and Maybury, M. T., editors, Advances in Automatic Text Summarization, pp. 61-70, The MIT Press

표 6. 한국어 문서요약 시스템들의 특징

시스템	접근 방법	실험 환경	성능		비고 (요약문서의 크기)
			R	P	
[3] (Kang, 1997)	통계적 방법 학습함. 문장 유사도 문장단위로 추출	논문(서론과 결론) 평가 문서 수: 25	51.08	42.4	본문 크기의 20%
[6] (Lee et al., 1999)	휴리스틱 방법 통계적 방법 학습함. 문장 유사도 문장단위로 추출	논문(서론과 결론) 평가 문서 수: 20	66.8.0		본문 크기의 30%
[7](Ryu and Lee, 2000)	통계적 방법 부쉬경로 코사인 유사도 단락 단위로 추출. 가중치: tf * idf	논문(서론과 결론) 평가 문서 수: 25		35.0	본문 크기의 30%
[15] (Myaeng and Jang, 1999)	휴리스틱 방법 통계적 방법 학습함 문장단위로 추출	논문(서론과 결론) 평가 문서 수: 30	53.19	39.53	5 개의 문장