논문 2004-41CI-2-5

단일 문서의 인위적 요약과 MMR 통계요약의 비교 및 분석

(Analyses and Comparisons of Human and Statistic-based MMR Summarizations of Single Documents)

유 준 현*, 변 동 률*, 박 순 철**

(Jun-Hyun Lyu, Dong-Rul Byun, and Soon-Chul Park)

요 약

웹과 같은 대량의 문서집단에서 단일 문서에 대한 자동 요약은 일반적으로 통계요약 방법을 이용한다. 그러나 단순한 통계요약 방법은 문서내의 빈도수가 높은 단어를 포함하는 문장들이 중복되어 나타날 확률이 높다. 이러한 단점을 보완하기 위하여 본 논문에서는 통계기반 요약방법에 MMR 기법을 적용하여 요약의 질을 향상시켰다(약λ=0.6에서 최고의 성능을 보임). 또한 본 논문에서는 인위적 요약을 수행하여 MMR 통계기반의 요약 결과의 성능을 평가하였다.

Abstract

The Statistic-based method is widely used for automatic single document summarization in large sets of documents such as those on the web. However, the results of this method shows high redundancies in the summarized sentences because this method selects sentences including words that frequently appear in the document. We solve this problem using the method MMR to raise the quality of document summary (The best results are appeared around λ=0.6). Also, we compare the MMR summaries with those done by human subjects and verify their accuracy.

Keywords: MMR, Statistic summarization, Text summarization

I. 서 론

인터넷의 확장과 이용자의 폭발적 증가는 정보량의 기하급수적인 증가를 가져왔다. 또한 많은 이용자들이 검색엔진을 이용하여 자신이 원하는 정보를 찾고 있다. 검색된 문서들의 내용이 사용자가 원하는 것인가를 확인하기 위해서 사용자가 직접 검색된 결과의 문서 내용을 전부 읽어보아야 한다. 그러나 문서의 내용을 사전에 요약하여 사용자에게 검색결과와 함께 제공함으로써 막대한 시간과 노력을 절감할 수 있다.

본 연구에서 제안하는 통계요약은 대량의 문서 집단에 존재하는 단일 문서내의 단어와 문장의 통계정보를 이용하여 중요 문장을 추출하는 것이다.[1-2] 일반적으로 통계요약에서 중요문장을 계산할 때 문서 내에 포함되어 있는 단어의 가중치의 합을 적용하게 된다. 따라서 가중치가 높은 단어를 포함하고 있는 문장들이 상대적으로 중요도가 높아지게 된다. 이 경우 단어의 가중치는 문서 내의 각 단어의 빈도수에 의해서 결정되기 때문에 상대적으로 중복된 내용의 문장이 추출될 확률이 높아진다. 이러한 단점을 제거하기 위하여 본 논문에서는 Maximal Marginal Relevance(MMR) 기법을 이용하여 중복된 문장을 제거함으로써 요약문서의 의미를 좀 더 분명하고 간결하게 할 수 있도록 했다.[3-4]

본 논문에서 제안한 MMR 통계요약을 검증하기 위하

* 학생회원, ** 종신회원, 전북대학교 전자정보공학부
(Dept. division of electorics and information eng., Chonbuk National Univ.)
접수일자: 2003년6월17일, 수정완료일: 2004년2월5일

여 전문가 및 본 프로그램 개발자가 직접 참여한 인위적 요약을 수행하였다. 인위적 요약은 문장 이해 기반 요약의 한 방법으로 요약 전문가가 문서의 내용을 파악하고 문서로부터 주제를 표현하고 있는 정보를 식별한 후 문장 생성과정을 통해 요약하는 것이다. 인위적 요약에 참여한 사람은 5명의 요약 전문가(전북대 인문대 대학원생)와 4명의 시스템 개발자로 이루어졌다.

본 논문의 구성은 다음과 같다. 다음 Ⅱ장은 단순 통계요약의 소개와 본 논문에서 사용한 MMR 통계 문서요약에 대한 설명이다. 제 Ⅲ장은 통계요약을 분석하기 위한 인위적 요약의 결과 및 분석과 통계요약을 비교한 실험결과이다. 마지막으로 Ⅳ장에서는 본 연구에 대한 결론과 향후연구에 대하여 언급하겠다.

Ⅱ. 통 계 요 약

본 절에서는 본 연구에서 사용된 요약 시스템의 구조와 MMR 통계요약에 대한 설명이다.

1. 요약 시스템 구조

통계 문서요약의 기본 개념은 불용어가 아니면서 문서 내의 빈도수가 많은 단어에 높은 가중치를 부여하고 문장의 중요도를 계산하는 것이다. 본 논문에 사용한 요약의 알고리즘 역시 이 기본적인 개념을 포함하여 단어의 특성을 가중치에 적용한 것이다.

본 논문에서 구현한 요약시스템의 구조는 [그림 1]과 같다. 문서가 입력되면 형태소 분석기를 이용하여 문장 및 단어를 구분하여 추출한다. 이 때 불용어는 제외된다. 추출된 단어들은 문서 내의 통계정보 즉, 용어빈도수(f), 역 문서빈도수(idf), 단어의 특성 정보(P) 등을 계산하여 문장 단위로 수집한다. 그 후 수집된 통계정보를 이용하여 문장의 중요도를 계산한다. 요약문장은 중요도가 높은 문장 순서대로 추출하되 MMR 기법[3-4]을 이용하여 중복된 문장을 제거한다. 결과적으로 요약된 문장은 중요도가 높고 중복성이 적은 문장이 선택된다. 선택된 문장은 문서내 문장의 배열순으로 재 정렬하여 출력한다.

본 요약 시스템에 사용된 형태소 분석기는 국민대학교 강승식교수의 한국어 분석 모듈(HAM)을 이용했다.[5]

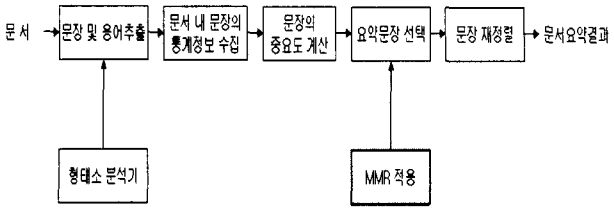


그림 1. 요약시스템의 구조.
Fig. 1. Structure of Summarization System.

2. 단순 통계요약

단순한 통계요약에서는 문서 내에 있는 단어들의 가중치를 계산하여 그 값을 기본으로 한 문장의 중요도를 결정하게 된다.[1-2] 이 방법에서는 일반적으로 문서 내에서 자주 등장하는 단어들을 포함한 문장들이 중요도가 높게 되어 요약문으로 선택되게 된다.

문서내의 문장에서 사용되는 단어의 가중치의 계산 방법은 식 (1)과 같다. 단어의 가중치, s_{ij}는 j문장에서 j 번째 있는 단어, w_{ij}의 가중치이다.

$$s_{ij} = f_{ij} \cdot idf(w_{ij}) \cdot P(w_{ij}) \tag{1}$$

여기서,

$$f_{ij} = \frac{freq_{ij}}{freq_{ij}+2},\quad idf(w_{ij}) = \log \frac{N}{n_{ij}},\quad P(w_{ij}) = \begin{cases} 2.0 & high \\ 1.5 & important \\ 1 & others \end{cases}$$

이다.

식(1)에서 $freg_{ij}$는 문서에서 단어 w_{ij}의 빈도수이다. f_{ij}는 $freg_{ij}$를 정규화한 w_{ij}빈도의 정도를 나타내며 0부터 1사이의 실수값을 갖는다. 역문서빈도수 $idf(w_{ij})$의 계산식 중 N은 본 시스템에서 문서의 총 수이다. n_{ij}는 단어 w_{ij}가 출현한 문서의 수이다. 요약할 문서가 다중 문서 공간에 있지 않은 경우에는 idf값을 1혹은 임의의 상수값으로 표시함으로써 계산을 단순화시킬 수 있다. 식(1)에서 P값은 단어의 특성(property)에 따라 임의 값으로 선택될 수 있다.

각 문서 내의 문장을 문장 벡터로 표현할 때 식 (2)와 같이 표현할 수 있다. 이 문장 벡터는 문장의 중요도를 계산할 때와 문장과 문장사이의 유사도를 계산할 때 사용되어진다.

$$\vec{S_i} = (s_{1i}, s_{2i}, \cdots, s_{ki}) \tag{2}$$

문장의 중요도는 문장 벡터 내에 있는 단어들의 평균값으로 표현된다. 식 (3)은 문장의 중요도 이다

$$|\vec{S_j}| = \frac{\sum_{i=1}^{size\ of\ \vec{S_j}} s_{ij}}{size\ of\ \vec{S_j}} \quad (3)$$

통계요약에서는 식 (3)의 방법으로 계산된 문장의 중요도를 바탕으로 문장들을 요약문장수에 맞게 선택하면 된다.

이러한 단순 통계요약의 경우, 문자의 중요도와 단어의 빈도수와의 역학관계로 인하여 선택된 문장들간의 유사도가 높은 것이 단점이다. 이러한 요약의 단점을 효율적으로 제거하는 것이 본 논문의 목적이다.

3. MMR 통계요약

MMR 통계요약 알고리즘은 웹 정보검색시스템의 검색결과를 재정렬[3-4]하거나 다중 문서의 요약[7-8]에서 중복된 내용의 가중치를 낮추는데 사용되고 있다. 그러나 아직 단일 문서요약에는 MMR 기법이 적용된 예를 찾기 힘들다.

본 논문에서는 이러한 MMR 알고리즘의 특징을 이용하여 단일 문서 내의 문장 집합에서 한정적인 문장을 추출할 때 중복된 내용의 문장의 가중치를 삭감하는 방법으로 응용했다. 따라서, 본 논문에서는 다수의 문서가 있는 문서공간에서 단일 문서를 요약할 때 요약 내용이 중복되지 않도록 했다.

기본적인 MMR 통계요약 알고리즘은 식 (4)와 같다. 즉 단어의 가중치에 기준 한 문장의 중요도를 계산하여 높은 순서대로 원하는 문장만큼 선택하는 것이다. 그러나 이 경우 가중치가 높은 단어를 포함하고 있는 문장들이 중복적으로 선택될 가능성이 높다. 이러한 단점을 보완하는 것이 본 논문에서 사용하는 MMR 기법이다.

식 (4)은 본 논문에서 사용된 MMR 알고리즘이다.

$$arg \max_{S_i \in R-A} \{(|\vec{S_i}| - \lambda \cdot \max_{S_j \in A} \cdot (sim(\vec{S_i}, \vec{S_j}))\} \quad (4)$$

여기서 A는 요약 문장으로 선정된 문장의 집합이다. R은 문장의 중요도에 의해 정렬된 리스트이다. 식 (4)의 유사도함수, $sim(S_i, S_j)$는 코사인 유사도를 사용하여 선택 가능한 문장과 이미 추출된 요약 문장 집합, A 에 포함된 문장과 비교하였다. 임의 두 문장, S_i 와 S_i 간의 코사인 유사도 계산은 식 (5)와 같으며 유사도 값은 0과 1사이의 실수이다.[9]

$$sim(\vec{S_i}, \vec{S_j}) = \frac{\vec{S_i} \cdot \vec{S_j}}{|\vec{S_i}| \times |\vec{S_j}|} = \frac{\sum_k s_{ki} \times s_{kj}}{\sqrt{\sum_i s_{ij}^2} \times \sqrt{\sum_m s_{mj}^2}} \quad (5)$$

식 (4)에서 값은 0과 1사이의 실수값을 갖는다. 이 0 이면 요약은 단지 문서의 통계정보에 따라 중요도가 높은 순으로 구성되면 값이 커질수록 요약에 포함된 문장 내용과 유사정도가 높은 문장이 제외된다. MMR 기법을 이용한 요약은 다음과 같이 수행된다.

1. 각 문장을 가중치에 따라 정렬한다.
2. 가중치가 가장 높은 문장을 요약의 첫 문장으로 선택하여 요약에 포함시킨다.
3. 첫 문장을 제외한 나머지 문장은 식 (3)에 의해서 선택된다. 선택된 문장은 요약문장에 포함된다.
4. 선택된 문장의 수가 만족되면 멈춘다.
5. 선택된 문장을 본문의 순서대로 정렬하여 읽기가 쉽도록 한다.

Ⅲ. 인위적 요약 및 결과

본 논문에서 인위적 요약에 사용된 실험 데이터는 2000년에서 20001년 사이의 코리아타임즈 신문에서 외국 뉴스 10개와 국내 뉴스 10개를 선택하였다. 사용된 총 문장수는 230개였으면 사용된 단어의 수는 3128개이다. 실험 문서의 형평성을 고려하여 문서 내의 문장수가 11에서 13 사이인 문장을 선택하였다. 각 문서 내의 문장은 평균 13.6개의 용어를 포함하고 있다.

인위적 요약에 참가한 사람은 전북대학교 인문대 대학원생 5명과 본 요약시스템 개발자를 포함한 전북대학교 공과대 대학원생 4명이다. 각 실험대상자는 기사를 먼저 읽고 각각의 기사에 대해 가장 중요한 문장을 1개 선택하고 그 다음 문장의 중요성에 따라 차기 문장을 선택하도록 했다. 이렇게 해서 선택된 5개의 문장을 중요도에 따라 1에서 5까지 각각 점수를 부여했다. [그림 2]는 인위적 요약방법에 따라 선택된 문장들의 평균 점수변화를 보인다. [그림 2]에서는 대부분 뉴스기사의 중요한 문장이 앞에 위치한다는 것을 보여준다.

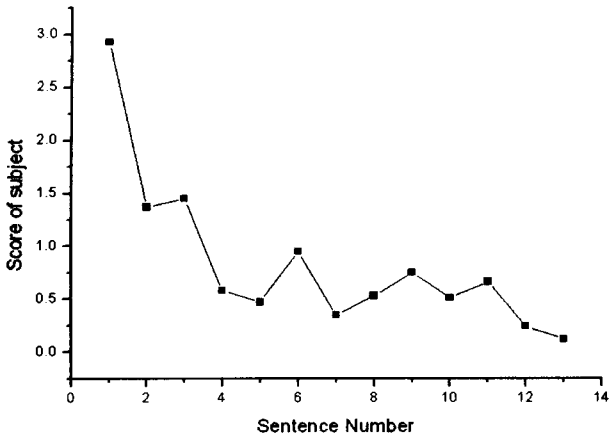


그림 2. 인위적으로 선택된 문장들의 점수변화.
Fig. 2. Average sentence score by human-subject.

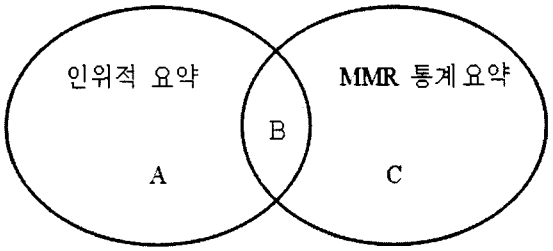


그림 3. 인위적 요약과 MMR 통계요약 관계.
Fig. 3. Relationship between summaries by humans and by MMR.

Ⅳ. 실험결과

본 실험을 인위적 문서요약을 기준으로 하여 통계 문서요약을 비교하였다.[10-13] 비교 값으로는 재현률과 정확율의 정보를 포함하고 있는 FScore를 사용하였다.

1. 인위적 요약과 통계요약의 비교

일반적으로 요약의 성능을 시험하기 위하여 전문가의 인위적 요약을 기준으로 한 재현율(recall)과 정확율(preciseion)을 이용한다. [그림 3]은 인위적 요약문서와 통계요약문서 사이의 관계를 표시한다. [그림 3]에서 A는 인위적 요약문서의 문장의 집합, C는 MMR 통계요약문서의 문장 집합, B는 두 요약문서 사이의 공통 문장의 집합을 나타낸다.

이 때 재현율과 정확률은 각 집합 내에 있는 문장 수에 의해서 결정이 되며 식 (6)과 같다.

표 1. 단순통계요약과 MMR통계요약의 비교.
Table 1. Comparisons of Simple and MMR Statistic Summarizations.

요약방법 / 요약문장수	단순 통계 요약	MMR (λ=0.4)	MMR (λ=0.6)	MMR (λ=0.8)
1	0.100	0.100	0.100	0.000
2	0.225	0.275	0.300	0.275
3	0.300	0.350	0.366	0.350
4	0.325	0.337	0.362	0.362
5	0.430	0.440	0.440	0.440

$$recall=\frac{|B|}{|A|} \qquad precision=\frac{|B|}{|C|} \tag{6}$$

본 논문에서는 요약의 성능을 비교하기 위하여 재현율과 정확률을 포함한 FScore 계산 값을 사용하며 FScore는 식 (7)로 표현된다.

$$\begin{aligned} FScore &= \frac{2*recall*precision}{recall+precision} \\ &= \frac{|CS\cap HS|}{No.\ of\ summaried\ sentences} \\ &= \frac{|B|}{No.\ of\ summariedsentences} \\ &= recall \\ &= precision \end{aligned} \tag{7}$$

여기서 CS는 MMR 통계요약의 문장집합이고 HS는 인위적 요약의 문장 집합이다.

본 실험에서는 비교를 용이하게 하기 위하여 인위적 요약문서의 문장수와 통계요약문서의 문장수를 같게 하였다. 즉, A+B와 B+C의 값을 같게 하였다. 따라서 재현율의 값은 정확율과 같게 되며, 아울러 FScore의 값도 재현율이나 정확율의 값과 같게 된다.

[표 1]는 단순 통계요약과 MMR 통계요약의 FScore 실험결과를 보인다. 한 문장 요약일 경우에 MMR을 적용할 수 없기 요약방법에 따른 결과가 같았다. λ=0.6일 때 MMR을 적용한 요약의 성능이 가장 좋았다. 2~3문장 요약일 경우 상대적인 요약 결과가 단순 통계요약보다 좋았고 요약 문장수가 커질수록 요약성능의 차가 줄어들었다.

[그림 4]에서는 표 2의 실험결과를 그래프로 표현하

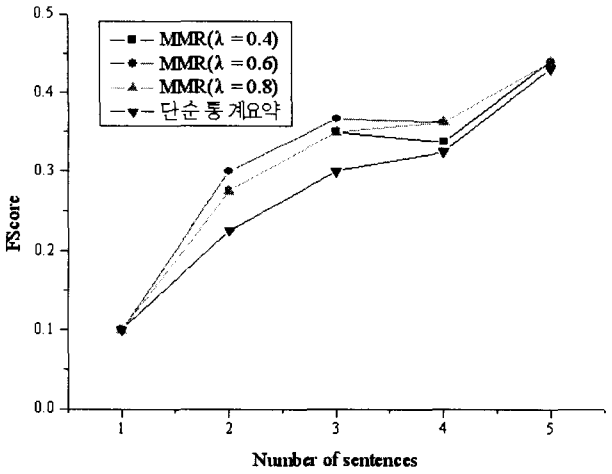


그림 4. 요약문장 수에 따른 FScore.
Fig. 4. FScore vs. number of summarized sentence.

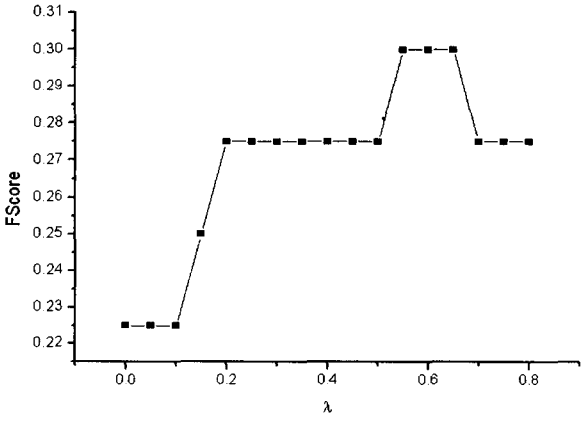


그림 5. λ에 따른 FScore (문장수가 2인 경우).
Fig. 5. FScore vs. λ(when number of sentences is 2).

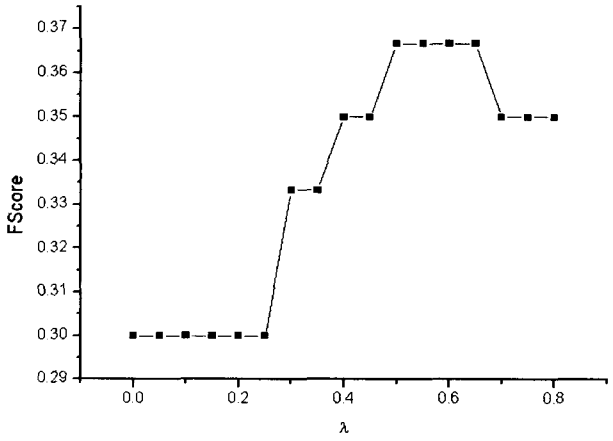


그림 6. λ에 따른 FScore (문장수가 3인 경우).
Fig. 6. FScore vs. λ(when number of sentences is 3).

였다. 요약 문장수가 증가함에 따라 FScore값이 선형적으로 증가하는 것을 볼 수 있다. 이것은 요약문장의 수가 증가하면 인위적 요약문서의 내용과 MMR 통계 문서요약 내용의 같은 부분이 따라서 증가하기 때문이다. 선택한 문장이 원문의 문장수와 같게 되면 두 방법 모두 원문의 모든 내용을 포함하기 때문에 FScore는 1값을 갖게 된다.

[그림 5]와 [그림 6]는 MMR 기법이 얼마나 요약에 영향을 미치는지를 보여준다. [그림 5]는 요약문장의 수가 2일 때 값의 변화에 따른 FScore값을 보여준다. 이 경우 FScore는 λ가 약 0.6일 때 가장 높다. λ가 0에 가까울수록 FScore의 값이 낮아지는데 이 경우 대부분의 결과가 MMR 기법을 제외한 통계정보만 유지되기 때문이다. λ가 매우 높고 1에 가까울 때 FScore의 값이 다소 떨어지는데 그 이유는 λ가 커지면 중요도가 높은 문장이더라도 이미 선택된 문장과의 유사도가 높은 경우 요약 문장에서 배제되기 때문이다. 따라서 λ 값의 설정은 요약의 질을 높이는 중요한 변수가 된다.

[그림 6]는 3문장 요약에 대한 것을 보여준다. 이 경우 FScore값은 λ가 약 0.5~0.65일 때 가장 높다.

[그림 5]와 [그림 6]이 계단형으로 나타나는 이유는 λ의 변화가 요약 문장 수에 따라 변화를 보이기 때문이다.

2. 실험 예

본 실험에서 MMR 통계요약과 인위적 요약 방법으로 요약된 결과와 원문이 [그림 7-10]에 나타난다.

[그림 7]은 코리아타임즈의 사설을 번역한 원문 내용이다. 원문에서 '보테로'가 11번, '그림'은 6번, '콜롬비아'는 5번, '전시'는 4번등으로 나타났다.

[그림 8]의 인위적 요약은 전문가를 포함한 9명이 원문을 읽고 원문의 내용의 대표할 수 있는 3개의 문장을 추출한 결과이다. 총 12개의 문장 중 앞부분에 있는 1번째와 2번째, 2문장이 추출되어 전형적인 두괄식 형태의 뉴스문장이다. 추출된 모든 문장이 빈도수가 가장 많은 '보테르'를 포함하고 있으며 '그림' 혹은 '콜롬비아'를 적당히 포함하고 있다.

[그림 9]는 MMR을 적용하지 않은 통계요약문, 즉 λ = 0인 경우이다. 인위적 요약과 다르게 '보테로'와 '그림'이 있는 11번째 문장이 선택되었다. 그것은 빈도수가 높은 단어를 포함하고 있고 상대적으로 짧은 문장에 중요도가 높아지는 통계요약이 특성이 있기 때문이다.

콜롬비아의 폭력을 그리는 보테로

보고타, 콜롬비아 (AP) — 통통한 인물들을 그려서 뉴욕과 마드리드를 비롯한 도시들을 묘사해온 화가 페르난도 보테로가 자신의 어마어마한 작품들을 가지고 고향인 콜롬비아로 올 것인데, 그는 콜롬비아에서 사는 것이 안전하지 않다고 느끼고 있다.

남미에서 현존하는 가장 유명한 화가인 보테로는 메들린시에 유화 70점과 십여점 이상의 조각을 기부하고 있다.

차분하고 낙관적인 작품으로 잘 알려진 보테로가 기부한 작품 중에는 새로운 경향의 그림이 몇 점 있는데 그것은 그의 조국을 파멸시키고 있는 폭력을 소재로 한 그림이다.

피해자와 가해자 모두 뚱뚱한데 이것은 큰 것에 대한 그의 집착을 반영하는 그의 작품의 표식이다.

한 그림은 경찰의 총탄 세례를 받고 죽어가는 뚱뚱한 마약왕 파블로 에스코바르의 모습을 보여준다.

메들린 코카인 조직의 지도자는 7년 전 보테로의 상설 전시회장에서 멀지 않은 곳에서 피살되었는데, 그의 전시회는 한 공원과 인접한 박물관 두 군데서 매우 수요일 공식적으로 열린다.

보테로는 수도인 보고타에 있는 한 박물관에 자신의 작품 100점을 추가로 기부할 것인데 여기에는 대량 학살 장면을 묘사한 그림이 포함되어 있다.

"이 그림은 콜롬비아가 겪고 있는 광란과 치욕의 순간을 극명하개 보여준다," 보테로는 10월 7일 가진 인터뷰에서 이렇게 말했다.

비록 일부 작품의 주제는 죽음이지만 그의 전시회로 인해 보테로가 68년전에 태어난 메들린시의 일부 지역이 말끔해졌다.

쇠락해가던 한 도심 지역은 이제 치우다드 보테로 즉 보테로시로 알려지게 되었으며 실물 보다 큰 그의 상들이 공원을 따라 줄지어 있다.

보수한 식민지 시대 건물 두 채가 치우다드 보테로의 양쪽을 장식하며 그의 그림들을 돋보이게 한다.

미술 평론가인 에두아르도 세라노는 "여러가지 다양한 작품을 야외에 전시해놓으니 예술이 이 도시 삶의 중심부분이 되었다" 고 말했다.

그림 7. 코리아타임즈의 사설 원문.
Fig. 7. Original documentation of Korea Times.

[그림 10]은 MMR(λ=0.6)을 적용한 요약문을 나타낸다. [그림 10]에서는 [그림 9]의 11번째 문장이 제거되고 1번째 문장이 선택됨으로써 인위적 요약과 결과가 같아졌다. 이는 11번째 문장이 8번째 문장과의 유사성이 높고 또한 두 문장 중 중요 단어가 8번째 문장에 더 많이 포함되었기 때문이다. 따라서 11번째 문장이 제거되고 대신이 1번째 문장이 선택되었다.

보고타, 콜롬비아 (AP) — 통통한 인물들을 그려서 뉴욕과 마드리드를 비롯한 도시들을 묘사해온 화가 페르난도 보테로가 자신의 어마어마한 작품들을 가지고 고향인 콜롬비아로 올 것인데, 그는 콜롬비아에서 사는 것이 안전하지 않다고 느끼고 있다.

차분하고 낙관적인 작품으로 잘 알려진 보테로가 기부한 작품 중에는 새로운 경향의 그림이 몇 점 있는데 그것은 그의 조국을 파멸시키고 있는 폭력을 소재로 한 그림이다.

"이 그림은 콜롬비아가 겪고 있는 광란과 치욕의 순간을 극명하게 보여준다," 보테로는 10월 7일 가진 인터뷰에서 이렇게 말했다.

그림 8. 인위적요약.
Fig. 8. Summary by human subject.

차분하고 낙관적인 작품으로 잘 알려진 보테로가 기부한 작품 중에는 새로운 경향의 그림이 몇 점 있는데 그것은 그의 조국을 파멸시키고 있는 폭력을 소재로 한 그림이다.

"이 그림은 콜롬비아가 겪고 있는 광란과 치욕의 순간을 극명하게 보여준다," 보테로는 10월 7일 가진 인터뷰에서 이렇게 말했다.

보수한 식민지 시대 건물 두 채가 치우다드 보테로의 양쪽을 장식하며 그의 그림들을 돋보이게 한다.

그림 9. 단순 통계요약.
Fig. 9. Simple Statistic Summary.

보고타, 콜롬비아 (AP) — 통통한 인물들을 그려서 뉴욕과 마드리드를 비롯한 도시들을 묘사해온 화가 페르난도 보테로가 자신의 어마어마한 작품들을 가지고 고향인 콜롬비아로 올 것인데, 그는 콜롬비아에서 사는 것이 안전하지 않다고 느끼고 있다.

차분하고 낙관적인 작품으로 잘 알려진 보테로가 기부한 작품 중에는 새로운 경향의 그림이 몇 점 있는데 그것은 그의 조국을 파멸시키고 있는 폭력을 소재로 한 그림이다.

"이 그림은 콜롬비아가 겪고 있는 광란과 치욕의 순간을 극명하게 보여준다," 보테로는 10월 7일 가진 인터뷰에서 이렇게 말했다.

그림 10. MMR. 통계요약(λ=0.6).
Fig. 10. Summary with MMR (λ=0.6).

V. 결론 및 향후 연구

본 논문에서 제시한 MMR 통계요약 방법은 최근에 정보검색 시스템의 reranking 알고리즘과 다중문서 요약시스템에서 내용의 중복성을 감소시키기 위하여 사용되고 있는 MMR 기법을 이용했다. MMR 기법은 구현이 상대적으로 쉽고 수행속도가 빠른다. 이러한 MMR

기법의 특성을 이용하여 웹 같은 대용량 문서집단에 있는 문서들의 요약에 알맞은 알고리즘은 구현하였다.

또한 본 논문에서는 실험을 통하여 문장간의 유사도의 영향(λ)을 변화시키면서 MMR 기법을 이용한 요약의 최적조건을 찾았다. 실험결과에서 λ 값이 약 0.6일 때 가장 좋은 결과를 보였다.

본 방법을 다양한 문서에 적용하여 문서들의 특성 조사 함께 거기에 따른 요약기법의 연구가 이루어져야 한다. 본 논문에서 문서요약 실험데이터로 코리아타임즈의 뉴스 중 사설을 이용하였다. 인위적 요약의 실험결과에서 대부분 요약문장이 문서 앞부분의 문장으로 선택되었다. 이러한 특징은 뉴스가 두괄식으로 쓰인다는 통념을 보이는 현상이다. 그러나 이러한 특징은 특정 문서집단에 따라 다를 것으로 예측된다. 향후 연구는 요약하고자 하는 문서집단의 문서 특성을 조사하여 요약 기법에 적용함으로써 요약의 질을 높이는 것이다.

요약의 질을 비교하기 위하여 정확한 인위적 요약 결과가 필수적이다. 본 실험에서 인위적 요약을 위한 실험에 참여한 사람은 총 9명이다. 전문성을 높이기 위해 문서의 분석 능력이 뛰어난 인문대 대학원생 5명이 요약에 참여하였다. 그러나 실험의 객관성과 정확도를 높이기 위해서는 좀더 훈련된 요약 전문가가 다수 참여한 실험이 향후 연구에서 요구된다.

본 논문에서 제시한 MMR 통계요약은 실제로 웹 검색엔진, 콘도르(http://irlab.chonbuk.ac.kr 혹은 http://search.searchline.info)에 사용되어서 그 기능과 성능을 검증받았다.

참 고 문 헌

[1] 김영택 외 공저, 자연언어처리, 생능출판사, 2001.9.

[2] 강상배, 한국어 문서의 통계적 정보를 이용한 문서요약 시시스템 구현, 부산대학교, 전자계산학과 , 석사 학위 논문, 1998. 2.

[3] A. Leuski and J. Allan. Improving interactive retrieval by combining ranked lists and clustering. In Proceedings of RIAO'2000, pages 665--681, April 2000.

[4] Jaime Carbonell and Jade Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998.

[5] http://nlp.kookmin.ac.kr/ 국민대학교 강승식 교수, 한국어 분석 모듈(HAM)

[6] Kathleen McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin, Towards Multidocument Summarization by Reformulation: Progress and Prospects, In Proceedings of AAAI'99, Orlando, FL, July 1999.

[7] W. Kraaij, M. Spitters, and M. van der Heijden. Combining a mixture language model and naive bayes for multi-document summarisation. In Working notes of the DUC2001.

[8] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell, Summarizing Text Documents: Sentence Selection and Evaluation Metrics, In Proceedings of ACM-SIGIR'99, Berkeley, CA, August 1999.

[9] Inderjeet Mani and Eric Bloedorn, Summarizing Similarities and Differences Among Related Documents, Information Retrieval 1 (1-2), pages 35-67, June 1999.

[10] Thrse Hand. A Proposal for Task-Based Evaluation of Text Summarization Systems, in Mani, I., and Maybury, M., eds., Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, July 1997.

[11] Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad, Summarization Evaluation Methods: Experiments and Analysis, In Working Notes, AAAI Spring Symposium on Intelligent Text Summarization, Stanford, CA, April 1998.

[12] Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Leo Orbst, Thrse Firmin, Michael Chrzanowski, and Beth Sundheim. The TIPSTER SUMMAC text summarization evaluation. Technical Report MTR

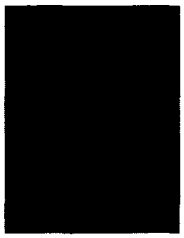
98W0000138, MITRE, McLean, Virginia, October 1998.

[13] Inderjeet Mani and Mark Maybury. Advances in Automatic Text Summarization. MIT Press, 1999.

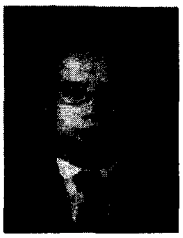
[14] 오형진, 변동률, 이신원, "클러스터링 중심 결정 방법에 따른 문서클러스터링 성능 분석", 대한전자공학회, 2002.

[15] 유준현, 변동률, 박순철, "MMR, 클러스터링, 완전연결기법을 이용한 요약방법비교", 대한전자공학회 2003.

저 자 소 개



유 준 현(학생회원)
2002년 전북대학교 공학사
2004년 전북대학교 정보통신 석사
<주관심분야 : 정보검색, 데이터베이스>



변 동 률(학생회원)
1998년 전북대학교 공학사
2003년 전북대학교 정보통신 석사
2001년~2002년 (미) 카네기멜론대학 언어기술연구소 방문연구
<주관심분야 : 정보검색, 멀티미디어 정보검색, 데이터 베이스>

박 순 철(종신회원)
1979년 인하대학교 (공학사)
1991년 (미) 루이지아나 주립대학 전산학 박사
1991년~1993년 한국전자통신 연구원
1993년~현재 전북대학교 전자정보공학부 부교수
1999년~2001년 전북대학교 정보검색시스템센터 센터장
2001년~2002년 미국 카네기멜론대학 언어기술연구소 방문연구
<주관심분야 : 정보검색, 영상정보검색, 데이터구조 및 알고리즘>