중요 문장추출 휴리스틱과 MMR을 이용한 질의기반 문서요약.

김동현[O] 이승우 이근배
포항공과대학교 컴퓨터공학과
{dhkim, pinesnow, gblee}@nlp.postech.ac.kr

Query-Based Document Summarization using Important Sentence Selection Heuristics and MMR.

DongHyun Kim[O] Seungwoo Lee, Gary Geunbae Lee
Dept. of Computer Science & Engineering, POSTECH

요 약

본 논문은 자연어 검색엔진에서의 검색결과에 대한 HIT LIST[6]와 검색 문서의 요약을 위하여 질의 기반의 3단계 문서요약을 제안한다. 첫째단계로 IR에 주어지는 질의를 유의어 DB를 통해 질의확장을 거친다. 둘째로 질의와 검색문서상의 문장의 유사도 계산을 통해 문장의 중요도 점수를 구한다. 좀더 정확한 요약을 위해 4가지 방법론을 적용하여 각 문장의 중요도를 ranking한다. 셋째로 MMR (Maximal Marginal Relevance)방식을 적용하여 요약시 중복요약이 되는 부분을 줄인다. 이때 요약 압축률을 임의로 조절할수 있다. 실험은 KORDIC의 신문기사로 구성된 문서요약 테스트 집합을 사용하여 좋은 요약결과를 얻었다.

1. 서론

인터넷에 산재한 URL만도 14억개가 넘고 여기에 포함된 문서는 셀수없이 많다. 이 문서들중에서 사용자가 필요로하는 문서를 찾아내는 일은 점점 더 어려워지고 있다. 이를 해결하기 위해서 여러 해결책들이 제시되고 있는데, 정보검색, 문서 분류, 정보 추출, 가시화, 질의 응답, 문서요약에서 다양한 접근이 시도되고 있다.

정보검색에 의해 검색된 문서에서도 필요한 정보를 찾으려면 다시 그 문서 전체를 살펴야하는 문제가 있다. 이런 문제점에 의해 정보검색엔진에 문서요약엔진을 결합할 필요성이 대두되고 있다.

정보검색과 문서요약은 기술적인 면에서 유사점이 많다. 차이가 있다면 기술의 적용 범위에 둘 수 있다. 즉

정보검색이 문서집합에서 사용자가 원하는 몇개의 적합한 문서를 찾아내는 것이라면, 문서요약은 문서의 내용을 대표하는 몇 개의 문장을 찾아내는 작업으로 생각할수 있다.

문서요약이란 문서의 기본적인 내용을 유지하면서 문서의 복잡도(문서의 길이)를 줄이는 작업이다[2]. 이러한 문서요약에는 생성요약(Abstract)과 추출요약(Extract)이 있으며 현재 NLP기술의 한계로 아직은 생성 요약에 대한 만족스러운 결과는 나와있지 않다.

추출요약은 생성요약에 비해 가독성(readability)와 응집도(cohesion)가 떨어지지만 비교적 생성요약에 비해 문제가 단순화 되어 처리된다. 본 논문에서는 추출요약에 의한 문서요약 방법으로, 특히 정보검색 결과에 대해서 주어진 질의의 내용에 가장 맞는 요약을 생성하고자 한다. 이는 PDA같은 정보단말기를 위한 정보검색 결과요약 자동 생성 등에 활용 될 수 있다.

요약은 기능과 제시방법, 문서의 개수에 따라 여러가지로 분류되기도 한다[5].

기능에 따른 분류에서는 지시적 요약(indicative summary)과 정보적 요약(informative summary)으로 나누어지는데 지시적 요약은 모든 문서의 내용을 포함할 필요가 없이 사용자가 적합성 판단을 내리는데 도움을 줄만큼의 지시적 기능을 갖추면 된

다. 본 논문에서 다루는 정보검색시스템의 요약이 지시적 요약에 해당된다. 그러나 정보적 요약은 원래 문서를 대치할 수 있을 정도의 주제의 내용을 포함하고 있어야한다. 제시되는 방법에 따른 분류에는 사용자 주도 요약(user-driven summary)과 포괄적 요약(generic summary)으로 나누어 지는데, 사용자 주도요약은 특정 사용자의 요구에 맞게 생성되는 요약을 말하고, 포괄적 요약은 특정사용자의 특정 질의를 대상으로하지 않고 일반적 문서전체에 대한 포괄적인 요약이다. 마지막으로 요약하는 문서의 개수에 따른 분류에서는 단일문서요약과 다중문서요약으로 나뉜다.

본 논문에서 제시하는 POSUM은 위의 분류에 의하면 지시적요약을 이용하여 정보검색에서의 결과의 요약을 제시하고 있으나 요약의 질을 높여, 정보적 요약으로 활용 될 수 있을 정도의 고정밀 요약을 지향한다. POSUM은 요약 대상을 하나의 문서로 보고 문서를 구성하는 여러 개의 문장들간의 관계나 중요도를 각 문장들간의 유사도를 이용하여 구분해낸다. 이 유사도 계산에는 여러가지 가중치가 사용되는데 이 가중치 휴리스틱은 각 문장에 골고루 적용된다. 이 유사도를 토대로 요약문을 생성하기 위한 중요 문장을 ranking하여 추출한다.
이후의 장은 다음과 같이 구성된다.
2장에서는 본 논문의 문서요약에서 수행되었던 요소기술 관련 연구내용을 기술하였고, 3장에서는 이 시스템의 수행환경인 POSNIR/K와 문서요약 시스템인 POSUM 시스템 전반적인 구조에 대해서 다룬다. 4장에서는 요약을 위한 다양한 문장 Weighting 휴리스틱과 MMR방식을 새롭게 제시한다. 5장에서는 KORDIC 문서요약 컬렉션을 이용한 Weighting방법들의 성능에 대한 실험을 기술하고. 6장에서는 본 연구의 결론과 향후 계획을 제시한다.

2. 문장 추출 요소기술에 대한 관련연구

문서에서 중요한 정보를 포함한 요약문을 만들어내기 위한 방법에는 특정 필드의 내용을 채워넣는 정보추출과 문장과 구의 추출에 따른 문서구성의 정보요약이 있다. 정보요약에서는 문장단위의 추출이 요구되는데 이 추출을 위한 방법에는 문서 안의 문장에 특정 요건에 의한 점수를 주는 방법이 있다. 문장에 주어진 점수로 랭킹을 결정하고 상위 몇 개를 취함으로서 요약이 만들어진다.
문서의 문장에 점수를 주기 위한 방법에는 문장을 벡터모델로 만들어 질의나 중요단어와의 관계를 유사도 계산법에 의해 구해 점수를 주는 방법과 적합문서 부적합문서에 나타나는 질의어의 통계 분포에 기초한 확률 모델이 있다. 이 논문에서는 정보검색엔진(POSNIR/K)에서는 확률모델을 사용하며, 요약엔진(POSUM)에서는 벡터모델과 확률모델을 혼합한 방식을 사용한다. 요약엔진에서 사용한 문장추출모델에 대해 관련된 연구를 살펴보기로 한다.

2.1 벡터 공간 모델의 유사도의 계산

벡터 공간 모델이란, 문서 D와 질의 Q를 n-차원의 벡터로 표현하고 그 내적의 합을 구하여 문서-질의간 유사도를 측정하는 모델이다.

요약에서는 문서단위가 아니라 문장단위로 문장-질의간의 유사도를 측정하므로 아래와 같은 식으로 계산가능하다.

$$S_j = (t_{1j}, t_{2j}, \ldots, t_{nj})$$
$$Q = (q_1, q_2, \ldots, q_n)$$
$$Sim(S_j, Q) = \sum_{i=1}^{n} t_{ij} \bullet q_i \quad (1)$$

여기서 S_j는 문서내의 한 문장을 말하며 여러단어들의 벡터로 나타내어진다.

이와같은 백터공간모델을 활용한 유사도 계산법에 대해서 알아보면,

summ_bin : 문장에서 질의 단어의 출현 유무로 몇단어가 나타났는지로 유사도를 계산하는 법으로 횟수는 적용되지 않는다.

$$\text{summ\_bin}: t_{ij} = \begin{cases} 1 (\text{질의단어가 문장} S_j \text{에 나타나면}) \\ 0 (\text{단어가 나타나지않으면}) \end{cases} \quad (2)$$
$$Sim(S_j, Q) = \sum_{i=1}^{n} (summ\_bin : t_{ij})$$

summ_tf : 문장에서 질의 단어의 출현 횟수를 세어서 유사도를 계산하는 방법이다.

$$\text{summ\_tf\_}: t_{ij} = freq_{ij} \quad (3)$$
$$Sim(S_j, Q) = \sum_{i=1}^{n} (summ\_tf : t_{ij})$$

summ_tf_norm : 문장에서 질의 단어의 출현 횟수를 세어서 이값에 정규화 값을 적용하는 것으로 유사도를 계산하는 방법이다.

$$\text{summ\_tf\_norm}: t_{ij} = \frac{freq_{ij}}{\max freq_{ij}} \quad (4)$$
$$Sim(S_j, Q) = \sum_{i=1}^{n} (summ\_tf\_norm : t_{ij})$$

본 논문에서 사용하는 4가지 문장추출 휴리스틱[3] 중 3가지인(Luhn's Cluster Method, Title Term

Frequency Method, Query Bias Method)는 summ_tf_norm의 변형형이다.

2.2 중요문장 선별 휴리스틱

본 논문에서는 [3]에서 제시된 다음과 같은 고려사항이 문장선택에 중요한 요소가 된다.

1) 위치관계 : 문서내에서의 문장의 위치
2) 단어빈도수 : 한 문서 전체에서의 단어의 발생횟수
3) 특정단어의 존재 :특정한 구나 단어의 존재유무
4) 문장의 상관관계 : 한 문서내의 한 문장과 다른 문장의 단어와 구의 관계

위의 네 가지 요소들로 가중치 점수를 부여하여 더한 값이 요약후보 문장추출에 대한 기준이 된다. 본 논문에서사용한 방법은 문장점수부여의 4가지 휴리스틱에 의해 문장의 위치관계, [3]에서 제시된 문장내의 단어빈도수가 고려되었으며 Maximal Marginal Relevance(MMR)[1]에 의해서 문장의 의미적 상관관계가 반영되어 있다.

3. 질의기반요약 시스템 개관

POSUM은 정보검색 결과를 위한 질의기반 문서요약[8] 시스템이다. 본 시스템은 POSNIR[14]와 연동하여 이루어진다. POSNIR/K의 색인모듈, 검색모듈과 POSUM의 요약모듈로 시스템이 구성된다. 시스템의 구조는 [그림 1]과 같다

본 논문에서는 정보검색에서 사용되는 기술을 변형하여 문서요약에 적용한다.

3.1 색인모듈

색인모듈은 HMM 기반의 품사 태거인 POSTAG/K를 사용한다. POSTAG/K[12] 크게 2개의 구성 요소로 이루어져 있다. 첫번째는 전처리 부분으로 이 부분에서는 문자열을 입력으로 받아 문장을 인식하고 한글의 단어 태깅 단위로 분리한다. 두번째는 형태소 분석/태깅 단계로, 이 단계에서는 사전에서 태깅 단위를 찾고 최적의 품사를 결정한다. 태거의 출력은 입력 문자열, 품사, 원형으로 이루어져 있다. 따라서 이 결과를 받아서 원하는 정보를 사용하면 된다. 다음은 POSTAG/K의 실행결과의 예를

```
입력 문장 : CAD 시스템에 대한 연구 조사를 하였다.

형태소 분석/태깅 결과 :
CAD      CAD/sm
시스템에  시스템/MC      에/j
대한      대하/D  ㄴ/eCNMG
연구      연구/MC
조사를    조사/MC 을/j
하였다.   하/MP   이/I    었/eGS  다/eGE  ./s
```

보여준다..

색인어 추출은 한글 형태소 분석 및 품사 태거의 결과로부터 명사와 형용사, 동사와 인용구를 대상으로 한다.

알파벳으로 구성된 영어단어의 경우에는 원형(lemma)을 색인 대상으로 하며, 원형이 사전에 등록 되어 있지 않은 키워드에 대해서는 색인어-질의어 간의 불일치를 해결하기 위해 Stemmer[4]를 사용하여 stemming결과를 색인 대상으로 한다.

색인어는 단일어 뿐만 아니라 복합명사 합성과 분할에 의해서도 추출된다. 복합명사 합성은 품사 태그 패턴의 학습을 통해 이루어지며[13] 복합명사 분할은 음절 길이 패턴과 상호정보(MI)에 의해 이루어 진다.

추출된 단일어 중에 불용어 리스트에 속한 키워드는 색인어에서 제외된다. 복합어의 경우, 복합어를 구성하는 단일어 중에서 제일 마지막 키워드가 불용어일 경우 복합어를 생성하지 않는다.

3.2 검색모듈

사용자가 요청한 문서는 POSNIR에 의해 검색되고, 사용자에게 제시되기전에 POSUM에 의한 요약과정을 거친다. POSUM은 사용자의 질의에 기반한 추출요약으로 POSNIR 검색에 사용된 질의를 그대로 이용한다.

질의어 추출 단계에서는 자연어 형태의 사용자 질의에서 형태소 분석, 태깅을 통해 명사와 형용사, 동사와 인용구 등을 질의어로 추출한다. 또한 이 과정 중에 색인 모듈의 색인어 추출과 마찬가지로 복합명사 합성 및 분할에 의한 키워드를 질의어 리스트에 추가한다.

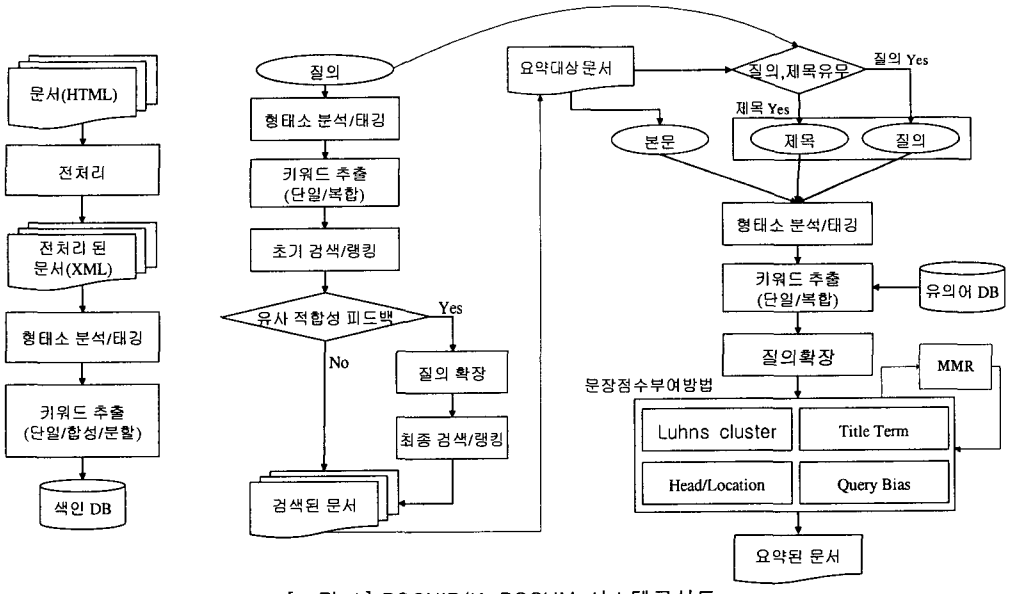
추출된 질의어 리스트에서 색인어 추출에서와 마찬가지로 불용어를 제거한다. '그것, 그러나, 의, 이것' 등과 같이 질의 문장에 속한 특수한 키워드를 추가로 제거한다. 이러한 질의 대상 불용어는 실제 질의로그에서 얻은 리스트이다.

검색모듈은 확률 분포에 기반한 Robertson의 2-포아송 모델을 사용한다. 본 시스템에서는, 문서의 출현 빈도(Term Frequency), 문서 길이(Document Length) 등을 고려한 Okapi BM25 함수[7][10][11]를 사용한다.

3.3 요약모듈

POSNIR에 의해 검색된 문서는 요약모듈에서 문장단위로 분할되며, 형태소 분석 및 태깅과정을 거쳐 문장단위로 키워드가 추출된다. 이렇게 추출된 단어들은 이전 단계에서 생성된 질의단어리스트와 문장단위의 유사도 계산에 사용되고, 이를 통해 요약문장을 추출한다.

질의의 유사도 관계 뿐만 아니라 요약에는 각 문장과의 유사도 관계, 제목의 유사도, 문서내에서의 중요한 단어와 문장과의 유사도, 문서에서의 중요문장의 위치관계를 이용한 문장중요도 값들을 함께 고려하여 이들 가중치의 합으로 문장의 랭킹이 결정된다. 랭킹에서 상위 10%~30%사이를 취하여



[그림 1] POSNIR/K, POSUM 시스템구성도.

요약의 정도를 조절할수 있으며 이렇게 추출된 문장을 원래 순서대로 정렬함으로써 요약결과가 얻어진다.

4. 문장추출 휴리스틱과 MMR.

본 연구의 중요문장 추출을 위해서 기존연구[3]에서 제시되었던 방법을 기반으로 휴리스틱 weight를 도출하고 문장 내용 중복을 방지하기 위해 새로히 MMR[1] 방식을 기존 휴리스틱에 도입한다. 다음은 적용된 휴리스틱과 MMR을 설명한다.

4.1 Luhn's Keyword Cluster Method.

첫번째는 Luhn의 클러스터 방법[9]이다.

문서에서 요약을 위한 문장을 결정하기 위해서, 모든 문장들의 정보나 내용이 분석되어야 한다. Luhn은 단어의 출현횟수가 문장의 상대적 위치관계만큼이나 중요하다고 보고 이를 사용하였다.

Luhn의 방법론은 먼저 본문에서 후보단어를 생성하고 문서내에서의 단어의 출현횟수의 내림차순으로 단어를 정렬한다. 그리고 여기서 한 문서에서만 많이 나타나고 다른 문서에서는 전혀 나타나지 않는 단어는 중요하지 않은 단어로 분류하고, 또한 모든 문서에서 골고루 많이 나오는 단어도 불용어로 처리하고 리스트에서 지운다. 여기서 중요단어 선별을 위한 기준은 문서의 특징에 따라 달라지겠지만 7회이상의 발생이면 중요단어로 선정한다.

그리고 40문장 이상이거나 25문장 이하일때에는 중요단어선정을 위한 기준이 아래의 식에 의해 조절된다.

NS : 문서내의 문장의 수.

문서에서 NS < 25 일때
ms = 7 + [0.1*(L-NS)]
문서에서 NS > 40 일때
ms = 7 + [0.1*(NS-L)]

여기서 ms는 중요단어임을 나타내는 값
L = Limit (25 for NS < 25 and 40 for NS > 40)

중요단어의 선택 기준

이렇게 구해진 문서의 중요단어들로 각 문장에 블록을 만든다. 이 블록은 중요단어로 선택된 단어가 문장내에서 처음나오는 곳과 마지막으로 나오는 곳을 묶어서 만든다. 그리고 그 블록안의 단어의 수를 TW, 중요단어의 수를 SW로 하고 아래의 식으로 각 문장에 점수를 부여한다.

$$SS1 = SW^2 / TW$$

SW : 블록안의 중요단어의 수,
TW : 블록안의 단어의 수

4.2 Title Term Frequency[3]

제목은 문서의 주제와 연관이 있다. 따라서 제목은 일반적으로 문서내의 주요한 단어라고 할 수 있

는 단어를 포함하고 있다. 제목의 단어수를 TTT라고 하고 제목에서 추출한 단어가 문장에서 나타나는 횟수를 TTS라고 하고 아래의 식을 적용하여 주제와 연관된 문장에 점수를 부여한다.

$$SS2 = TTS / TTT$$

TTS : 문장안에서 주제단어의 나타나는수
TTT : 주제의 단어의 수.

4.3. Location/Header Method[3]

한 문서 내에서 첫째 문장과 둘째 문장은 대부분 주제와 관련되거나 문서안에서 중요한 역할을 한다. 이러한 사실을 반영해 문서내의 문장의 수를 정규화 값으로 하는 아래의 식을 첫째와 둘째 문장에 적용하여 점수를 부여한다.

$$SS3 = 1 / NS$$

NS : 문서안에서의 문장의 수

4.4. Query-Bias Method[3]

정보검색 기반의 문서요약에서는 사용자 질의가 항상 존재한다. 사용자의 의향을 반영하는 이 질의는 요약에 있어서도 중요한 요소가 될 수 있다. 질의에 나타나는 단어수를 nq로하고 질의 키워드가 문장에 나타나는 횟수를 tq로 하여 아래의 식으로 각 문장에 점수를 부여한다.

$$SS4 = tq^2 / nq$$

tq : 현재문장안에서 질의단어가 나타나는 수
nq : 질의의 단어의 수

위의 네가지의 문장 가중치 식을 각 문장에 적용하여 합한 값이 그 문장의 중요도가 된다.

SS = SS1 + SS2 + SS3 + SS4

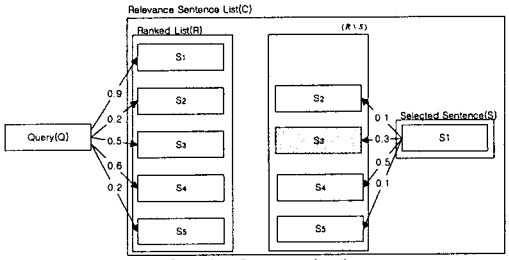
여기서 파라메타($\alpha, \beta, \tau, \delta$)값을 조절하여 어느 관점에서 요약을 할것이냐를 정할수 있다.

$$SS = \alpha SS1 + \beta SS2 + \tau SS3 + \delta SS4$$

4.5 MMR(Maximal Marginal Relevance)[1]

대부분의 IR 검색 시스템은 사용자의 질의에 의해 순서가 결정된 랭킹리스트를 생성한다. 그런데 이 리스트에는 중복된 결과가 포함될 가능성이 높다. MMR은 이런 문제를 줄이기 위한 방법론[1]인데, 이를 문서단위대신 문장단위에 적용하면 문서요약에서도 사용할수 있다.

아래의 식이 나타내는 것은 질의의 결과 리스트에서 중복을 없애는 과정인데, 실제 수행은 사용자 질의와 문장리스트 간의 가장 유사도가 높은 문장을 선택하고 선택된 문장을 이미 요약에 포함된 다른 문장들 리스트와 유사도를 계산하여 가장 낮은 유사도를 가지는 문장을 선택을 하면 문서에서 중복되지 않는 요약이된다. 따라서 각 단계에서 이미 선택된 문서는 문서 리스트에서 뺀다.



[그림 2] MMR의 예

$$MMR = \arg\max_{S_i \in R \setminus S}[\lambda(Sim_1(S_i, Q) - (1-\lambda)\max_{S_j \in S} Sim_2(S_i, S_j))] \quad (5)$$

식에서
C 는 문장의 집합 (또는 문장 스트림)
Q 는 유저의 질의
$R = IR(C, Q, \theta)$ R은 요약시스템에 의해 Ranking 된 문장 리스트, C와 Q와 적합성 기준치인 θ
S 은 R에서 이미 선택된 문장들.
$R \setminus S$ 는 Ranking 된 문장 리스트에서 이미 선택된것을 뺀 리스트
Sim_1 는 문장들과 질의에 대해 4가지 휴리스틱을 사용한 유사도값이다.
Sim_2 는 질의에 의해 선택된 문장(S_i)과 문서내의 이미 선택된 다른 문장(S_j)과의 유사도 값이며 역시 4가지 휴리스틱을 사용한다.

그림 2는 MMR의 동작을 그린 그림으로 두번째 단계에 해당한다.

첫번째 단계에서 질의(Q)와 각 문장(S1,S2…)에 4가지 휴리스틱에 의한 유사도 값이 매겨지고 여기서 큰 값을 가진 문장S1이 선택된다. 두번째 단계에서는 문장S3가 선택되는데 이는 식(5)에 의해 질의와의 유사도 값과 S1과의 유사도 값의 차가 다른 문장에 비해서 가장 크기 때문에 선택된다. 질의와 S3와의 유사도는 0.5이고 S1과 S3와의 유사도는 0.5이므로 이때의 유사도의 차이는 0.2이다. 이 값은 다른 문장(S2, S4, S5)의 값에 비해 가장 큰 값을 갖는다.

이와 같은 방법으로 $R \setminus S$ 가 공집합이 될 때까지 실행된다.

5. 실험 및 평가

문서요약에서는 요약 결과에 대한 객관적 평가가 난해하다. 자연언어 생성의 다양성으로 인해 사람이 직접평가에 개입해야하며 이렇게 개입하더라도 사람에 따라 다른요약이 나오며, 동일인이 요약을

하더라도 시간이 지나고 하면 조금씩 틀려지는 등의 문제가 있다.
그리고 가독성에 대한 평가뿐만아니라 원본에 대한 충실도도 평가에 반영하는 등의 다양한 평가가 병행되어야 한다.

본 실험에는 KORDIC의 신문기사 문서요약 셋을 사용하였다. 문서요약셋의 구성은 제목을 나타내는(#T), 내용을 나타내는(#S) 문장추출에 의한 10% 요약(#A), 30%요약(#B), 그리고 수동으로 한 생성요약(#C) 으로 구성되어 있다. [그림3]

이 요약셋은 총 998개의 요약문서셋으로 구성되어 있는데 이중에서 제목이 없는 것, 요약부분이 잘못된 것을 전처리 과정을 거쳐 정재하고 그중에서 60 개를 임의 추출하여 실험에 사용하였다.
질의 기반 요약을 위해 POSNIR/K시스템에서 인덱싱을 하였으며 질의기반 실험을 할수 있게 했다.

평가의 척도로는 정확률, 재현률, F-MEASURE를 사용하였다.

평가 방법
POSUM의 결과와 KORDIC요약셋에 이미 제시된 요약 결과 #A, #B, #C를 비교하여 정확률과 재현률을 구한다.
정확률은 KORDIC에서 제시된 요약결과문장 #A, #B를 올바른 요약문으로 간주하고 이와 일치하는 POSUM 이 추출한 문장과 POSUM이 추출한 전체 요약문장 의 비를 통해 정확률을 구한다.
재현율은 #A, #B와 일치하는 POSUM 이 추출한 문장과 실제로 올바른 요약문과의 비를 통해 구한다.

정확률 (Precision):

$$\frac{\text{시스템이제시한올바른요약문}}{\text{시스템이제시한요약문}}$$

재현율(Recall):

$$\frac{\text{시스템이제시한올바른요약문}}{\text{올바른요약문}}$$

F 값 (F-MEASURE) : 2PR/(P+R)

10%요약과 30%요약을 수행하며, 앞장에서 제시했던 네가지 방법(Luhn cluster, Title, Location, Query Bias)의 실행유무에 따른 정확률, 검색률, F 값을 구하여서 평가에 반영한다.

첫 번째 실험에서는 4가지 휴리스틱을 사용하였을 때와 MMR을 적용하였을 때를 비교하여 문서의 정확도, 재현율, 그리고 F-값을 구하였다. 이 결과는 표1에서 보여주고 있는데, 테스트 문서의 길이(평

#T
미-북한 「막후교감」 암시/북한 「NPT복귀 조짐」 언저리
#S
◎미,체제불간섭 등 약속가능성/「특별 사찰」 근본과제는 그대로/우리정부,북핵해결 새 일정표 고심
북한은 지난 5월 이후 줄곧 유엔 안보리가 자신들의 핵문제를 이유로 한 대북제재 결의안을 채택한다면 이를 「선전포고」로 간주하겠다고 공언해 왔다.
북한은 현재 유엔 안보리에 결집해 있는 국제사회의 요구가 무엇인지를 잘 알고 있다. 그러나 북한은 이 요구에는 묵묵부답이다. 정말 북한은 밖으로 비쳐지는 것처럼 국제사회와의 정면충돌을 각오한 것일까.
이를 1차 가늠할 수 있는 시험대가 10일 뉴욕에서 열리는 미-북한 3차 고위급 회담이다. 미-북한 양측은 이미 6월 돌어서만도 2차례의 회담을 가졌다.
○회담장 안팎 다른 태도
그러나 북한은 다시한번 반전을 시도하고 나섰다. 미국측에게 한번 더 협상을 갖자고 요구한 것이다.이같은 태도는 두가지로 해석할 수 있다.
○「면죄부」 될순 없어 당초 한-미양국이 구상했던 미-북한 회담은 2차로 국한됐었고, 실제 미국 수석대표인 갈루치 국무부 정치-군사담당 차관보는 2차 회담이 끝난 뒤 북한측의 3차회담 요청에 대해 「NPT 복귀가 전제되어야 한다」고 분명히 언급했기 때문이라는 것이다.이 경우 미-북한간에는 북한의 NPT 복귀라는 약속과 미국의 대북 정치적 약속 몇가지가 포함되어 있을 가능성이 높다.
.....
<파리=박두식기자>
#A
북한은 지난 5월 이후 줄곧 유엔 안보리가 자신들의 핵문제를 이유로 한 대북제재 결의안을 채택한다면 이를 「선전포고」로 간주하겠다고 공언해 왔다.
...
#B
북한은 지난 5월 이후 줄곧 유엔 안보리가 자신들의 핵문제를 이유로 한 대북제재 결의안을 채택한다면 이를 「선전포고」로 간주하겠다고 공언해 왔다.
북한은 현재 유엔 안보리에 결집해 있는 국제사회의 요구가 무엇인지를 잘 알고 있다. 그러나 북한은 이 요구에는 묵묵부답이다. 정말 북한은 밖으로 비쳐지는 것처럼 국제사회와의 정면충돌을 각오한 것일까.
이를 1차 가늠할 수 있는 시험대가 10일 뉴욕에서 열리는 미-북한 3차 고위급 회담이다.
지난 4일 2차 회담후 미국측은 「결렬」을 선언하지 않았을 뿐이지, 내심 협상은 사실상 끝났다고 판단하고 있었다.
.....
#C
10일 뉴욕서 열리는 미-북한 3차 고위급 회담, 북한 의도 알아 낼 수 있는 시험대. 2차 회담 결렬후 3차 회담 제의는 북한의 지연전 술로 해석됨. 2차 회담과 3차 회담 사이 6일 동안에 미-북 간의 모종의 교감이 있었던 것으로 추측됨. 북한의 NPT복귀가 북한 핵개발 계획 전체에 대한 면죄부가 될 수는 없다.
우리정부, 10일 미-북한 접촉후 예상되는 북한 핵문제 해결의 새로운 일정표 작성에 착수.

[그림3] KORDIC 요약셋의 구성

균 17.24문장)가 짧음에도 MMR에 의한 성능향상을 확인할 수 있었다.

	4heuristic	4heuristic+MMR
AVG.Precision	0.4497	0.4782
AVG.Recall	0.5495	0.5902
F-Measure	0.4946	0.5283

[표 1] 4 Heuristic과 MMR 비교 실험(30%)

두 번째 실험에서는 MMR을 사용한 상태에서 10%와 30% 요약시의 성능을 비교하였다.

	10%	30%
AVG.Precision	0.5833	0.4782
AVG.Recall	0.6481	0.5902
F-Measure	0.6139	0.5283

[표 2] 압축률에 따른 실험 (4heuristic+MMR)

표2에 따르면 10%에서 성능이 높음을 알 수 있다. 이는 짧은 요약을 필요로 하는 검색엔진에 적용시 높은 정확도의 HIT LIST를 나타낼 수 있음을 보여준다. 30%의 요약에서도 비교적 높은 성능을 나타냄으로서 일반적인 요약에도 적절함을 보여준다.

6. 결론

문장 추출을 위한 기존의 네가지 heuristic과 MMR이라는 IR적 접근법을 문서요약에 적용하여 기존의 정보검색엔진과 연계하여 효과적으로 동작하는 요약엔진을 개발하였다. 이 엔진은 질의를 받아서 사용자 주도의 요약뿐만아니라 일반적인 요약까지 가능하여 좀더 범용적인 시스템으로 확장할수 있게 염두에 두었다.

정보검색시스템과의 결합을 통해 단순히 문서의 앞부분만 보여주는 기존의 정보검색 시스템의 결과 제시를 대치해 줄 수 있다.

차후의 연구 분야로는 Portable Device에서의 요약과 다중문서요약을 고려해볼 사항이다.

7. 참고 문헌

[1]. Jaime Carbonell, Jade Goldstein. *The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries,* SIGIR'98, 335-336, 1998.

[2] Julian Kupiec, Jan Pedersen, and Francine Chen, *A Trainable Document Summarizer,* In Proceedings of ACM-SIGIR'95, pp.68-73, 1995.

[3]. Adenike M. Lam-Adesina, Gareth J.F.Jones. *Applying Summarization Techniques for Term Selection in Relevance Feedback.* SIGIR'01, September 2001

[4] Porter M.F. *An algorithm for suffix stripping.* Program 14,130-137. 1980.

[5] Inderjett Mani, David House, Gary Kein, Lynette Hirschman, and Leo Obrst. *The TIPSTER SUMMAC Text Summarization Evaluation Final Report,* Technical Report MTP98W0000138, MITRE, 1998.

[6] Dragomir R. Radev. Weiguo Fan. *Automatic summarization of search engine hit lists.* ACL Workshop on Recent Advances in NLP and IR , Hong Kong, October 2000.

[7] Rocchio and J.J. *Relevance Feedback in Information Retrieval. The Smart System-experiments in automatic document processing,* 313-323. Prentice Hall, 1961.

[8]. Anastasios Tombros and M. Sanderson. *Advantages of Query Biased Summaries in Information Retrieval.* SIGIR'98, 1998.

[9] Anastasios Tombros and M. Sanderson. *Reflecting user information needs through query biased summaries.* SIGIR'98, 1998.

[10] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. *Okapi at TREC-3. In Overview of the Third Text Retrieval Conference (TREC-3),* 109-126, 1995.

[11] Robertson S.E. and Sparck Jones K. *Relevance Weighting of Search Terms. Journal of the American Society for Information Science,* 129-146, 1976.

[12] 심준혁, 김준석, 차정원, 이근배. *통계와 규칙을 이용한 강인한 품사태거.* 제 1 회 MATEC99 평가대회 논문발표집, 전자통신연구소, 1999.

[13] 원형석, 박미화, 김지협, 이근배. *복합명사 처리를 위한 통합 다단계 색인모델.* HCI'99 학술발표대회 논문집, pp 80-87, 1999.2.

[14] 조봉현, 이창기, 안주희, 이근배. *확률적 정보검색 모델에서의 유사 적합성 피드백 실험.* 한글 및 한국어 정보처리 논문집, p183~190, 2001.