2001년 한국정보처리학회 추계 학술발표논문집 제8권 제2호

자동 문서요약을 위한 중요문 추출 방법 설계

신성혁*, 김태완**,
*인제대학교 전산학과
e-mail:s9331218@cs.inje.ac.kr

A Design of Important Sentence Extraction Method for Automatic Text Summarization System

Sung-Hyuk Shin*, Tae-Wan Kim**
*Dept of Computer Science, In-Je University

요약

본 논문에서는 빠른 속도로 증가하고 있는 인터넷상의 정보와 서비스를 검색함에 있어서 기본적인 내용은 유지하면서 정보의 과부하(information overload)문제를 해결하기 위한 문서요약의 방법으로 통계적 접근 방법에서 Kupiec의 요약문이 가지는 특성을 이용하여 문서요약의 방법을 설계하였다. 요약문의 각 문장에 대하여 중요도에 따라 가중치를 부여 한 후, 주어진 임계값에 따라 가중치가 낮은 문장들을 제외한다. 제외 후 가중치 점수를 부여해서 요약문 문장의 개수를 조절하면서 중요문을 추출할 수 있다.

1. 서론

인터넷상의 정보와 다양한 서비스가 빠른 속도로 증가하고 있다. 이런 정보를 검색하는 작업이 점점 어려운 문제가 되고 있는 것은 명백한 사실이다. 이러한 문제를 해결하기 위해서 다양한 정보 검색 시스템이 개발되어서 사용자에게 필요한 정보를 검색하는 일을 돕고 있다. 검색시스템에서 정보를 검색하면 검색된 문서의 수가 1,000건이 넘는 경우도 많다. 이들을 모두 읽으면서 정보의 적절성을 판단하는 것은 거의 불가능하다고 할 수 있다. 이런 경우에 수많은 문서의 정보들을 종합하여 문서를 요약하여 제시한다면 제공된 문서의 적절성과 유용성과는 별개로 정보를 이용하고자 하는 사용자들은 문서의 내용을 파악하기 위해서 많은 시간을 소비하지 않아도 된다.

문서요약이란 문서의 기본적인 내용을 유지하면서 문서의 복잡도, 즉 문서의 길이를 줄이는 작업이다.[5]

본 논문에서는 통계적 접근 방법에서 요약문이 갖는 다섯 가지의 특성(feature)들을 이용하여 수행 순위를 지정한 후 순위에 따라 사용하지 않는 문장을 제외한다. 제외한 문장에 가중치 점수를 부여해서 요약문장의 개수를 선택하는데 이용할 수 있다.

본 논문의 구성은 2장에서는 관련연구를 살펴보고 3장에서는 중요문 추출방법에 대하여 살펴보고 4장에서는 결론 내린다.

2. 관련연구

문서요약 방법 중 통계적 접근 방법에서 Kupiec[5]은 요약문이 가지는 특성을 학습 코퍼스로부터 추출한 후 문서내의 각 문장에 대하여 요약문의 특성을 갖는 확률을 계산하여 일정한 값 이상이면 요약문에 포함시키는 접근을 시도하였다. 학습코퍼스로부터 추출한 문장의 특성은 5가지로 분류할 수 있는데 먼저 문장 길이에 따른 특성(Sentence Length Cut-off feature)으로 문서 내에서 길이가 짧은 문장은 요약문에 포함되지 않는 경향을 갖는다는 것이다. 일반적으로 5단어 이하로 구성된 문장을

말한다. 두 번째로는 상용구 특성(Fixed-Phrase Feature)으로 "결과적으로", "요약하면", "말하자면" 과 같은 단어가 포함된 경우 요약문이 될 가능성이 많다는 특성이다. 세 번째로는 단락 특성(Paragraph Feature)으로 문서에서 처음 10개와 마지막 5개의 단락을 대상으로 각 단락 내에서의 위치를 세 가지 (앞부분, 세 개 이상의 문장으로 이루어진 단락인 경우 중간 부분, 두 개 이상의 문장으로 이루어진 경우 끝부분)로 분류하여 특성 값을 부여한다. 네 번째로는 주제어 특성(Thematic Word Feature)으로 문서의 내용을 대표하는 단어 중에서 빈도수가 많은 단어를 주제어라고 정의하고 각 문장에 주제어가 포함되어 있는지에 대한 특성이다. 마지막으로 대문자 특성(Uppercase Word Feature)으로 대문자로 이루어진 단어가 요약문에 포함될 확률이 높다는 특성으로 이런 단어는 문서 내에서의 빈도수가 어느 정도 횟수 이상이어야 한다.[4] 본 논문은 Kupiec의 특성 값을 이용하여서 문서를 요약하는 방법과 유사하다. 하지만 각 특성 값을 이용하여 제외 할 수 있는 문장들을 중요문 후보에서 제외시키고 나머지 문장에 가중치를 부여한다는 점이 다르다고 할 수 있다.

3. 중요문 추출 방법

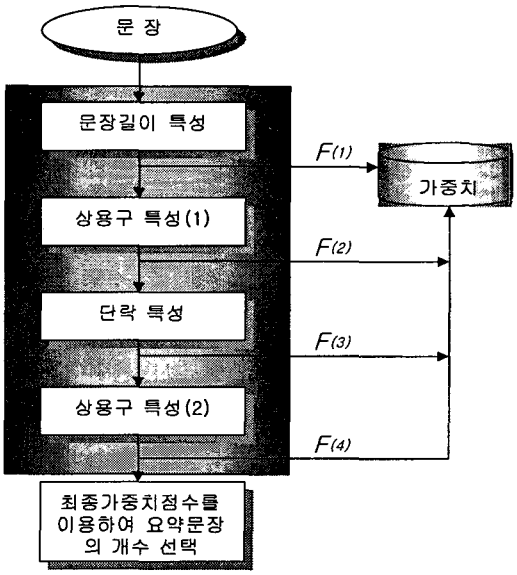
본 논문에서는 요약문에 포함될 확률이 낮은 문장을 추출하여 제외한 후 각 문장마다 가중치 점수를 부여하는 방법이다. 단순히 중요 문장만을 추출하는 방법[1,3,6,7]인 Kupiec의 통계적 접근 방법과 유사한 점이 있으나 이러한 방법들은 영어를 대상으로 한 방법이기 때문에 한국어에 적용하기 위해서는 수정이 필요하다. 영어에서는 문장 길이에 따른 특성인 5단어 이하의 문장은 중요 문장으로 선택되지 않는 경향이 있다. 하지만 한국어에서는 5단어 이하도 중요문장이 될 수 있는 가능성이 있으나 한 단어로만 사용된 경우는 거의 중요문장에 선택될 확률이 낮기 때문에 중요문 후보에서 제외하도록 한다. 이 같은 특성을 문장 길이 특성이라고 한다. 그리고 대문자 특성과 같은 경우는 한국에서는 대문자 소문자 구분이 없기 때문에 이 특성은 생각하지 않기로 한다. 요약문이 [그림 1]에서 볼 수 있듯이 문장길이 특성, 단락 특성 그리고 상용구 특성을 이용하여 수행 순위에 따라 사용하지 않는 문장들을 제외하고 나머지 문장에 가중치를 부여하고 부여한 가중치를 누적함으로써 효과적이고 적절한 결과를 만들 수 있다. 먼저 문장길이 특성을 이용해서 사용하지 않는 문장을 제거하는 방법을 수행 순위 첫 번째 순위에 두고 상용구 특성(1)을 두 번째 수행 순위, 단락 특성을 세 번째 수행 순위로 둔다. 위와 같이 수행 순위에 따라 사용하지 않는 문장을 제외한 후에 가중치를 부여하고 마지막으로 상용구 특성(2)을 한번 더 살펴서 상용구 특성(2)에 적합한 문장을 찾아내고 찾아낸 문장에 가중치를 부여한다.

3.1 문장길이 특성

요약 하고자 하는 문서에서 수행 순위를 생각하여서 사용하지 않는 문장을 살펴본다면 제일 먼저 문장길이 특성을 첫 번째 순서로 볼 수 있다. 예를 들어서 "발표자 : 홍길동", "서론"과 같은 단어들이 문서에 포함되어 입력된다면 이런 단어들은 요약문에 포함되지 않는 경향이 있다. 이런 특성을 이용하여 독립적으로 사용된 단어들을 사용하지 않는 문장으로 중요문 후보에서 완전히 제거한다. 이렇게 사용하지 않는 문장이 제거된 나머지 문장에 가중치 점수를 1점을 부여한다. 문장길이 특성에서 제거된 단어들은 완전히 제거하여 앞으로 가중치 점수를 부여하지 않는다.

3.2 상용구 특성(1)

입력받은 문서에서 요약 하고자 하는 문서의 두 번째 수행 순위로서는 상용구의 특성을 살펴 볼 수 있다.



[그림 1] 중요문 추출 구성도

문장을 서술하다가 부가적인 정보들을 나타내기 위해서 사용되는 즉 상용구 뒤에 오는 문장은 중요 문장으로 포함되지 않는 경향이 있는데 이러한 특성을 이용하여 사용하지 않는 문장을 찾아서 제외하는 방법을 선택한다. 예를 들어서 문장을 서술하다가 "예를 들어" "또한", "하지만"과 같은 상용구 뒤의 말들은 요약문이 될 가능성이 희박하다. 사용하지 않는 문장을 제외하고 남은 문장에 가중치 점수 2점을 부여한다.

3.3 단락 특성

문서요약을 위해서 앞에서 2가지 방법으로 사용하지 않을 확률이 높은 단어들과 상용구 뒤에 오는 사용하지 않는 문장들을 제외하였다. 이렇게 제외한 문서에서 다음 수행 순위인 단락 특성을 살펴보면, 한국어 문서에서 일반적으로 중요부분의 문장이 앞부분과 뒷부분에 대부분의 중요문장이 포함되어 있다. 즉 문단에서 귀납적인 방법이나 연역적인 방법을 사용하여 글의 주제를 나타내는 경우가 많다. 이러한 단락 특성을 이용하여서 위의 2가지 방법을 거쳐 추출된 문장 중에서 단락의 앞부분과 뒷부분의 문장의 일정 부분을 제외한 나머지 부분을 사용하지 않는 문장으로 중요문 후보에서 제외하면 된다. 이렇게 사용하지 않는 문장을 제외한 나머지 문장들에 가중치 점수를 3점을 부여한다.

3.4 상용구 특성(2)

앞에서는 사용하지 않는 문장 제외를 위한 상용구를 살펴보았으나 지금 말하고자 하는 상용구는 앞의 상용구와는 다르다. 앞에서 언급했던 상용구는 부가적인 정보를 나타내기 위한 것으로 제외를 하기 위하여 사용하지 않는 문장을 찾는 방법이었다. 그러나 여기서 사용 할 상용구는 제외하기 위한 것이 아니라 문서요약에 있어서 결정적인 문장이 될 수 있는 것을 찾고 그 문장에 가중치를 적용하는 방법이다. 예를 들어 '결과적으로', '요약하면', "따라서"와 같은 상용구 뒤에 이어지는 문장은 요약문이 될 가능성이 상당히 많다는 특성을 가진다. 앞에서 언급한 문장길이 특성, 상용구 특성(1) 그리고 단락특성과는 달리 이번 상용구특성은 앞의 특성들에 의해서 사용하지 않는 문장을 제외하는 방법이 아니라 상용구중에서 요약문이 될 가능성이 높은 문장을 추출하는 방법이다. 문장에서 가장 요약문이 될 가능성이 높은 상용구를 찾아서 추출한 후에 가중치 점수를 4점을 부여한다.

3.5 중요문 추출

앞에서 언급했듯이 각 특성들을 이용하여서 추출된 문장에 가중치 점수를 부여했다. 이 가중치 점수의 기준은 중요문 추출에 사용된다. 가중치 점수의 합을 이용하여서 최종 가중치 점수(S)를 구하고 요약문이 될 가능성이 높을수록 점수의 비중을 높이고 가능성이 낮을수록 점수의 비중을 줄인다. [그림 1]에서 볼 수 있듯이 각 특성들을 통과하면서 가중치 점수를 계속 부여하고 있는 것을 볼 수 있다. 최종 가중치 점수가 높을수록 그 문장은 요약문에 포함될 가능성이 높아지고 가중치의 합이 낮을수록 그 문장은 요약문에 포함될 가능성은 희박해진다.

$$(S) = \sum_{n=1} F(n)$$

여기서 n은 가중치를 부여하는 개수이다. 최종 가중치 점수에 대한 추출 가능한 값을 높게 잡는다면 요약된 문장의 개수가 작아지게 될 것이고 추출 가능한 값을 낮게 잡는다면 요약문장의 수가 그만큼 증가하게 될 것이다. 추출 가능한 값을 이용해서 문장의 개수를 조절할 수 도 있다. 이하 예문으로 설명하도록 한다.

[예제 1]

①5. 결론
②본 논문에서는 의미 자질과 결정 트리를 이용하여 한국어 부사격 조사의 의미격을 결정하는 방법을 제시하였다. ③이 방법에 대한 실험 결과, 한국어 부사격 조사 중 모호성이 가장 심하게 나타나는 조사 '왜' '와' '로'에 대하서, 각각 84.6%와 81.7%의 정확도를 보였다. ④따라서, 말뭉치로부터 추출한 <명사 자질-동사 자질_의미격>의 튜플로부터 결정 트리를 생성하여 의미격 결정 규칙을 생성하는 것이 타당함을 알 수 있다. ⑤하지만, 본 논문에서 제안한 방식은 학습 데이터의 부족 문제를 해결하기 위하여 의미 자질을 사용하였으나. 이 의미 자질의 객관성이 미약하고 단어 의미 중의성에 대한 해결책이 전혀 제시되지 못했다. ⑥또한, 결정 트리가 구분할 클래스의 수가 너무 많은 것도 문제가 된 것으로 보인다. ⑦따라서, 의미 자질 문제를 해결하고, 클래스의 수가 많을 때 구분문제를 학습할 수 있는 방법에 대한 연구가 필요하다.[2]

위의 예제에서 먼저 수행 순위에 의해서 첫째, 문장길이 특성을 살펴볼 수 있다. 문장길이 특성에 의해서 ①과 같은 단어는 중요 문장에 포함되지 않는 경향이 있으므로 제거한다. 그리고 나머지 모든 문장에 가중치 점수 1점을 부여한다. 둘째, 상용구 특성(1)을 살펴보면 위의 예제에서 "하지만", "또한"과 같은 상용구 뒤에 나오는 문장은 중요문장에 포함되지 않는 경향이 있으므로 ⑤, ⑦ 문장은 제외하고 가중치 점수 2점을 부여한다. 셋째, 단락 특성을 이용하여 문장을 제외한다. 여기서는 첫 문장과 마지막 문장은 중요 문장에 포함될 확률이 높으므로 ①, ⑧을 제외하고 나머지 문장들은 제외하고 가중치 점수 3점을 부여한다. 넷째, 상용구 특성(2)을 이용한다. 상용구 특성(1)과 달리 사용하지 않는 문장을 제외하는 것이 아니라 "결과적으로", "요약하면", "따라서"와 같은 상용구 뒤의 문장은 중요문장으로 사용될 가능성이 높다. ④, ⑧이 상용구 특성(2)에 부합된다. 이들 두 문장에 가중치 점수 4점을 부여한다.

문장번호	최종 가중치 점수
①	0
②	6
③	3
④	7
⑤	1
⑥	1
⑦	10

[표 1] 가중치 점수 부여의 결과

[표 1]에서 나온 최종 가중치 점수를 이용해서 추출 가능한 문장을 선택하는데 추출 가능한 문장의 최종 가중치 점수 값을 높게 잡으면 문장의 수가 줄게 되고 추출 가능한 문장의 최종 가중치 점수 값을 낮게 잡으면 문장의 수가 늘게 되어 있다.

(6점 이상 : ②, ④, ⑦ #3개의 문장 추출)
(7점 이상 : ④, ⑦ #2개의 문장 추출)
(8점 이상 : ⑦ #1개의 문장 추출)

4. 결론 및 향후 연구 과제

본 논문에서는 자동 문서요약을 위한 중요문 추출 방법에 대하여 제시하였다. 통계적 접근 방법에서 요약문이 가진 특성들의 수행 순위를 지정한 다음 수행 순위에 맞게 사용하지 않는 문장들을 제외한 후 가중치를 부여하여서 최종 가중치 점수가 높은 것을 선택하여 문서 요약을 하는 방법을 선택하였다.

각각의 특성들에 따라서 사용하지 않는 문장을 제외한 후에 가중치 점수를 부여하는 방법으로 특성들의 분류가 더 많이 되어 있다면 가중치 점수가 세분화되어서 문서 요약에 있어 더 좋은 결과를 가져 올 수 있을 것이다. 향후 연구 과제로는 이러한 문제를 해결하기 위해서 요약문이 가진 특성들의 분류를 세분화 할 필요가 있다.

참고문헌

[1] 강상배, 조혁규, 권혁철, 박재득, 박동인, "한국어 문서의 통계적 정보를 이용한 문서 요약 시스템 구현", 제9회 한글 및 한국어정보처리 학술 대회, 1994

[2] 박성배, 김영택, "한국어 부사격 조사의 의미격 결정" 정보과학회, 1998

[3] 장동현, 맹성현, "문서 구조 정보를 이용한 확률 모델 기반 자동요약 시스템", 제9회 한글 및 한국어 정보처리 학술대회, 1997

[4] 장동현, 맹성현, "자동 요약 시스템" 정보과학회지 제 15권, 1997

[5] Julian Kupiec, Jan Pederson and Francine Chen, "A Trainable Document Summarizer", Proc. of 18th ACM-SIGIR Conference, pp. 68-73, 1995

[6] Daniel Marcu, "the Rhetorical Paring, Summarization, and Generation of Natural Language Text", Ph.D dissertation, University of Toronto, Canada, 1997

[7] Simone Teufel, Marc Moens, "Sentence Extraction and rhetorical classification for flexible abstracts", AAAI Spring Symposium on Intelligent Text summarization, Stanford, March 1998