시소러스를 이용한 문서 자동 요약

이창범[U] 박혁로
전남대학교 전산학과
{cblee, hrpark}@cs.chonnam.ac.kr

Automatic Text Summarization Using Thesaurus

Chang-Beom Lee[U] Hyuk-Ro Park
Dept. of Computer Science, Chon-Nam National University

요 약

문서 자동요약은 입력된 문서에 대해 컴퓨터가 자동으로 요약을 생성하는 과정을 의미한다. 즉,컴퓨터가 문서의 기본적인 내용을 유지하면서 문서의 복잡도 즉 문서의 길이를 줄이는 작업이다. 효율적인 정보 접근을 제공함과 동시에 정보 과적재를 해결하기 하기 위한 하나의 방법으로 문서 자동요약에 관한 연구가 활발히 진행되고 있다.

본 논문에서는 의미기반 정보검색용 시소러스(thesaurus)를 이용한 문서 자동요약을 제안한다. 제안한 방법에서는 단어간의 연관 관계 즉, 동의어, 유의어, 상위어, 하위어 관계를 문서 요약에 이용한다. 크게 연관 사슬 형성 단계, 중심 문장 추출 단계, 요약 생성 단계의 세단계로 나누어 요약을 생성한다. 수동 요약된 신문기사를 대상으로 평가한 결과 평균66%가 일치하였다.

1. 서론

문서 자동요약은 입력된 문서에 대해 컴퓨터가 자동으로 요약을 생성하는 과정을 의미한다. 즉, 컴퓨터가 문서의 기본적인 내용을 유지하면서 문서의 복잡도 즉 문서의 길이를 줄이는 작업이다[9].

현재 우리들이 살고 있는 시대는 인터넷의 시대라 해도 과언이 아니다. 인터넷의 발달과 급속한 보급으로 쏟아지는 정보는 주체할 수 없을 정도이다. 넘쳐나는 정보 속에서 우리들에게 필요한 정보를 어떻게 선택하느냐하는 문제가 발생된다. 하지만, 이제는 선택의 문제를 넘어서 선택된 즉, 검색엔진이 결과로써 보여주는 문서 중에서 얼마나 빨리 그리고 정확하게 문서의 적합성을 판단할 수 있느냐하는 문제가 대두되고 있다.

일반적인 검색엔진들은 문서의 제목과 앞부분을 약간만 보여주어 이 문제를 해결하려 하지만, 이 정도의 정보는 사용자가 검색 결과 문서의 적합성을 판단하기에 부족하다. 자동 문서요약시스템은 사용자가 원하는 정보를 찾아내는데 걸리는 시간을 단축시킴으로써 정보과적재 문제에 대해 효과적인 해결책을 제시해 줄 수 있다[5].

효율적인 정보 접근을 제공함과 동시에 정보 과적재를 해결하기 위한 하나의 방법으로 문서 자동요약에 관한 연구가 활발히 진행되고 있다.

문서 요약은 그 생성 방법에 따라 추출(extract)과 요약(abstract)으로 구분될 수 있다[8].

본 논문에서는 문장을 추출하여 그 문서의 요약으로 사용한다. 문장을 추출하는데 있어 의미기반 정보검색용으로 사용되었던 시소러스(thesaurus)를 이용하였다. 먼저 시소러스를 이용하여 그 문서를 대표할 수 있다고 볼 수 있는 중심 문장을 추출한다. 그리고 나서 추출된 중심 문장과 다른 문장과의 연관 관계를 파악한 후 연관 관계순으로 정렬한다. 중심 문장을 포함하여 사용자가 원하는 비율만큼의 문장을 추출하여 요약문으로 제시한다. 중요 문장과 다른 문장과의 연관 관계를 파악할 경우에도 또한 의미기반 정보검색용 시소러스를 이용한다.

본 논문의 구성은 다음과 같다. 제2장에서는 문서 요약에 관련된 연구들을 살펴보고, 제3장에서는 제안하는 모델에 대해 설명한다. 그리고 제4장에서는 실험에 대해 기술한다. 마지막으로 제5장에서는 결론 및 향후 연구에 대해 기술한다.

2. 관련연구

기존의 문서에 대한 요약 생성에 관한 연구들은, 크게 통계적 기법과 문맥 구조에 기반한 방법, 그리고 지식에 기반한 방법으로 분류할 수 있다.

통계적 기법에서는 단어의 출현 빈도, 제목, 문장의 길이, 실마리 단어나 구(cue word or phrase)등을 자질(feature)로 사용하여 각 문장이나 문단의 중요도를 계산하여 그 값이 높은 문장이나 문단을 요약으로 제시한다[2,4].

문맥 구조에 기반한 방법은 문장들 사이의 문맥 관계를 파악하여 요약문을 생성한다[6].

지식에 기반한 방법은 생성하고자 하는 문서와 관련된 배경 지식을 이용하여 요약문을 생성하는 방법이다[1,3].

특히, [1]은 WordNet을 이용하여 같은 개념을 갖는 단어들을 사슬로 만들어, 즉 어휘 사슬(lexical chain)를 구성하여 강력한 사슬이 있는 문장을 선정하여 요약문을 생성한다. 하지만 이 방법에서는 긴 문장일수록 요약문에 포함될 가능성이 많으며, 요약문의 길이 조절에 대한 처리를 하지 않았다.

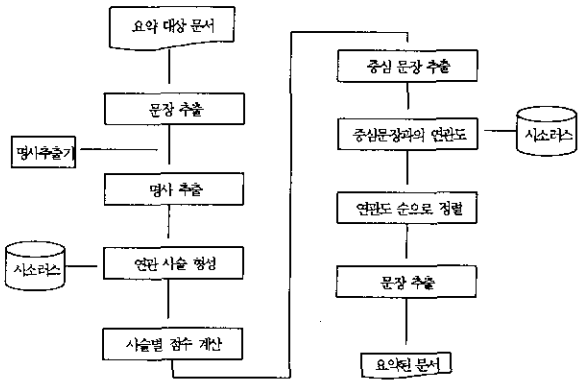
본 논문에서는 시소러스(thesaurus)를 이용하여 단어간의 연관 관계, 즉 동의어, 유의어, 상/하위어 관계를 이용하여 문서를 자동 요약하는 방법을 제안한다.

3. 시소러스 기반 문서 자동 요약

자연어 처리 시스템에서는 같은 주제라도 문헌 생산자나 색인 작성자, 이용자 간에 그 표현하는 용어가 달라질 수 있어 문헌의 분석이나, 색인 작성시에 많은 어려움이 야기된다. 따라서 필요한 정보를 찾으려고 하는 이용자는 하나의 검색어만으로 해당 주제를 전부 검색할 수 없으므로 그 검색어에 관련된 개념의 상위어, 하위어, 관련어 등을 모두 검색하여야 하는 번거로움이 있다. 이에 그 해당하는 주제 분야에서 필요한 모든 개념을 수집하여 이들에 대한 개념의 대소관계나, 동의어, 동형이의어, 관련어 등을 적절히 조절하여 정보시스템과 문헌 생산자, 색인 작성자, 이용자 간에 통일적으로 사용할 수 있도록 통제하여둔 용어통제어표를 시소러스(thesaurus)라 한다[10].

사람이 자연어 문장을 이해하는 경우에는 각자가 가진 상식이나 지식, 단어 개념 등의 지식베이스를 이용한다. 문서 요약에 이러한 지식베이스를 이용한다면 보다 자연스럽고, 요약 대상 문서의 주제에 더 접근할 수 있는 요약문을 생성할 수 있다. 예를 들어, 어떤 문서에 '국민', '민족', '종족', '인민', '국가', '공화국' 등의 단어가 나타난다고 하자. 이들 단어들을 개별적으로 이용하기 보다는 이들 단어들이 서로 연관이 있다는 정보를 이용한다면 그 문서의 주제를 파악하는데 더 유용하다. 사실, 이들 단어들은 사용하는 시소러스에서 유의어 관계로 서로 연결되어 있다.

본 논문에서는 단어간의 연관 관계 즉, 동의어, 유의어, 상위어, 하위어 관계를 문서 요약에 이용한다. 제안하는 모델의 전체 구성도는 [그림 1]과 같다. 크게 연관 사슬 형성 단계, 중심 문장 추출 단계, 요약 생성 단계로 나누어 볼 수 있다.



[그림 1] 시소러스 기반 문서 자동 요약 모델

3.1 연관 사슬 형성

요약 대상 문서의 주제를 찾아가는 단계라 할 수 있다. 문미기호 즉, '!','.','?' 등이 종결어미와 함께 나오는 경우를 문장으로 인식한다. 형태소 분석과 태깅 과정을 거친 후 각 문장에 대해 명사를 추출한다. 이렇게 추출된 명사들간의 연관 관계를 시소러스를 이용하여 파악한다.

만약, 명사들 간에 연관 관계가 형성된다면 그들 사이에 링크를 형성하고, 그렇지 않다면 다른 명사와 비교를 계속 한다. 추출된 모든 명사를 비교할 때까지 반복한다.

연관 관계는 단어의 반복, 동의어, 유의어, 상위어, 하위어로 제한하였다.

3.2 중심 문장 추출

요약 대상 문서의 주제를 가장 잘 표현한다고 볼 수 있는 중심 문장을 추출하는 단계이다.

형성된 연관 사슬의 중요도를 계산한다. 이를 위하여 연관 관계 사이에 다른 점수를 주었다. 가장 높은 점수를 획득한 사슬에서 가장 많이 발생하는 명사를 선택하여, 그 명사를 포함하는 첫 번째 문장을 중심 문장으로 선택한다. 연관 관계의 우선 순위는 다음과 같다.

반복 = 동의어 > 유의어 > 상위어 = 하위어

3.3 요약 생성

추출된 중심 문장과 다른 문장들 사이에 연관 관계를 파악하여 요약문을 생성하는 단계이다.

중심 문장에 포함된 명사와 다른 문장에 포함된 명사들 간에 연관도를 시소러스를 이용하여 계산한다. 그리고, 계산된 각 연관도를 문장의 길이 즉, 문장에 포함된 명사의 수로 나누어 주어 정규화 과정을 거친다. 이제 연관도 순으로 정렬하여 중심 문장을 포함해서 사용자가 원하는 비율 만큼의 요약문을 생성한다.

제안하는 모델은 단어간의 개념 관계를 이용하여 요약 대상 문서의 주제를 어느 정도 잘 표현하는 문장을 요약문으로 생성할 수 있다. 그리고 문장의 길이를 고려하였기 때문에 긴 문장의 선호도를 어느 정도 해소하였고, 또한 사용자가 원하는 비율만큼 요약을 생성할 수 있다.

4. 실험

의미 기반 정보 검색용 시소러스(thesaurus)를 실험에 사용했다. 이 시소러스는 단어(142,682건), 상위어(124,390건), 동의어(71건), 유의어(6,394건) 등으로 구성되어 있다. 하위어 관계는 상위어의 역으로 추정하였다[10].

실험에 사용한 데이터는 KISTI에서 제공되는 수동요약 테스트 컬렉션(test collection)를 사용하였다. 신문기사(1000건)에 대해 각각 10%, 30% 중요문장 추출, 10% 수동요약 결과로 구성되어 있다.

본 실험에서는 신문기사 30여건에 대해 30% 중요문장을 추출한 후 테스트 컬렉션에 있는 30% 요약 부분과 비교하였다. 실험에 사용된 문서는 평균 20.6개의 문장이다. 요약에 포함된 전체 문장수와 테스트 컬렉션의 30% 요약과 일치하는 문장수를 비교하였다. 본 논문이 제안하는 방법과 MS-Word의 자동요약을 비교한 결과는 아래의 표와 같다.

(평균)

	생성 문장수	포함 문장수	비율(%)
제안한 모델	5.8	3.8	66%
MS-Word	7.1	3.4	48%

비록 포함 문장수는 비슷하지만 그 의미는 차이가 있다. "임기중 개헌 없다/김대통령 취임 100일 회견"이라는 제목의 실험 문서를 예로 들어 보자. 이 문서에는 '대통령', '국가', '국민'이 각각 4번, 3번, 2번이 나타난다. 하지만 제안한 모델에서는 '국가'와 '국민'이라는 단어가 유의어 관계임을 사용하기 때문에 '대통령'이라는 단어보다는 더 중요하게 이용된다. 여기에서 제안한 모델과 MS-Word의 자동요약과 차이가 있었으며, 제안한 모델이 테스트 컬렉션과 더 일치함을 보여준다.

또한 기존 연구[1]와 비교하여 긴 문장 선호도를 어느 정도 해소하였고, 사용자가 원하는 만큼의 요약을 생성할 수 있도록 보완하였다.

5. 결론 및 향후 연구

본 논문에서는 의미기반 정보 검색용 시소러스를 이용하여 문서를 요약하는 방법을 제안하였다. 문서 전체의 주제를 표현하는 중심 문장을 추출하는데 시소러스를 이용하였다. 그리고, 그 중심 문장과 다른 문장과의 연관 정도도 역시 시소러스를 이용하여 측정하였다. 사람이 수동요약한 요약문과 평균 66%가 일치를 하였다. 또한 긴 문장 선호도를 해소하고자 했고, 사용자가 원하는 정도의 비율만큼 요약을 생성할 수 있도록 하였다.

요약 대상 문서의 주제를 파악하기 위해서 단어(명사) 반복이나 그 단어의 관련어만을 이용하였다. 여기에 동사나 형용사등의 용언까지 그 범위를 확장할 수 있는 연구가 필요하다. 또한 같은 단어이지만 다른 의미로 쓰일 경우 즉, 단어의 의미 정보를 이용할 수 있는 연구가 필요하다.

참고 문헌

[1] Regina Barzilay, Michael Elhadad, "Using Lexical chains for Text Summarization", proc. Association for Computational Linguistics, pp.10-17, 1997

[2] J.Kupiec, J.Pedersen, F.Chen, "A Trainable Document Summarizer", Proc. 18th ACM-SIGIR Conf., 1995

[3] Eduard Hovy and Chin Yew Lin, "Automated Text Summarization in SUMMARIST", Proc. Association for Computational Linguistics, pp.18-24, 1997

[4] H. P. Edmundson, "New Methods in Automatic Extracting", Journal of the Association for Computing Machinery, Vol.16,No.2,pp. 264-285, 1969

[5] Anastasios Tombros and Mark Sanderson, "Advantages of Query Biased Summaries in Information Retrieval", Proceedings of ACM-SIGIR'98, pp.2-10, 1998

[6] 양기주, "수사구조에 기반한 한국어 요약문 생성", 연구개발정보센터, 1997

[7] 박혁로, 신중호, "검색/요약/필터링을 위한 텍스트 이해 모형 및 처리 기술 개발", 연구개발정보센터 연구보고서, 1999

[8] 류동원, 이종혁, "단어공기정보를 이용한 자동화 문서 요약", 제27회 정보과학회 봄 학술발표 논문집(B), 제27권, 1호, pp.339-341, 2000.

[9] 장동현,맹성현, "자동 요약 시스템", 정보과학회지 제15권 제10호, pp.42-49, 1997

[10] 박혁로, 이현민, 전남열, 최선화, 정경석, "Answer Set 구축 지원도구 개발에 관한 연구", 한국전자통신연구원 연구보고서, 2000