질의확장을 이용한 자동 문서요약

한경수[U] 백대호 임해창
고려대학교 컴퓨터학과
{leganza, daeho, rim}@nlp.korea.ac.kr

Automatic Text Summarization Using Query Expansion

Kyoung-Soo Han[U] Dae-Ho Baek Hae-Chang Rim
Dept. of Computer Science & Engineering, Korea University

요 약

문서요약이란 문서의 기본적인 내용을 유지하면서 문서의 복잡도를 줄이는 작업이다. 인터넷과 같은 정보기술의 발달로 정보의 양이 급증함에 따라, 정보 과적재(information overload) 문제의 해결을 위해 자동 문서요약시스템의 필요성이 대두되었다.
본 논문에서는 의사 적합성 피드백(pseudo relevance feedback)에 의한 질의확장(query expansion) 기법을 적용한 자동 문서요약 모델을 제안한다. 제안하는 모델의 특징은 질의를 분해함으로써, 적합성 피드백 과정에서 질의가 편향(bias)되어 요약이 잘못되는 문제를 방지할 수 있다는 것이다. 신문기사를 대상으로 평가한 결과 제안한 모델이 질의확장을 적용하지 않은 방법이나 하나의 질의만을 유지하는 일반적인 적합성 피드백 모델보다 더 좋은 성능을 보였다.

1. 서론

문서요약이란 문서의 기본적인 내용을 유지하면서 문서의 복잡도, 즉 문서의 길이를 줄이는 작업이다[7].
인터넷과 같은 정보유통시스템의 발달로 인해 정보의 양은 하루가 다르게 지속적으로 증가하고 있다. 이런 상황에서 요약 작업을 수동으로 하는 것은 비효율적인 일이 되어 자동 요약시스템의 필요성이 대두되고 있다.
방대한 정보들 중에서 원하는 정보를 찾기 위해서 정보검색시스템을 사용하지만 정보검색시스템이 제시하는 검색 결과는 사용자가 하나씩 읽어보면서 확인하기에는 너무나 많은 양이다. 따라서 정보 과적재(information overload) 문제는 정보검색시스템에서 해결해야할 과제로 남아 있다.
일반적인 정보검색시스템들은 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 정보 과적재 문제에 대해 효과적인 해결책을 제시해줄 수 있다[2].
요약은 기능에 따라 사용자의 적합성 판단에 도움을 주는 지시적 요약(indicative summary)과 문서의 중요 내용을 유지하여 그 문서의 대용으로도 사용될 수 있는 정보적 요약(informative summary)으로 나눌 수 있다. 또 요약이 제시되는 방법에 따라 문서 내용전체를 포괄하는 포괄적 요약(generic summary)과 사용자 질의에 따라 질의에 관련 있는 내용만을 포함하는 사용자 주도 요약(user-driven summary)으로 나누어 볼 수 있다.
본 논문은 정보검색 분야에서 사용되는 의사 적합성 피드백(pseudo relevance feedback)에 기반한 질의확장(query expansion) 기법을 문서요약에 적용하는 기법을 제안한다. 제안하는 문서요약 모델은 포괄적 요약과 사용자 주도 요약에 모두 사용될 수 있으며, 문장추출을 통한 지시적 요약문을 제시한다.
본 논문은 다음과 같이 구성되어 있다. 2장에서는 문서요약 분야에서 수행됐던 관련 연구들에 대해 살펴보고, 3장에서는 제안하는 문서요약 모델에 대하여 설명한다. 4장에서 실험 및 평가를 하고, 5장에서 결론 및 향후연구를 기술하여 끝을 맺는다.

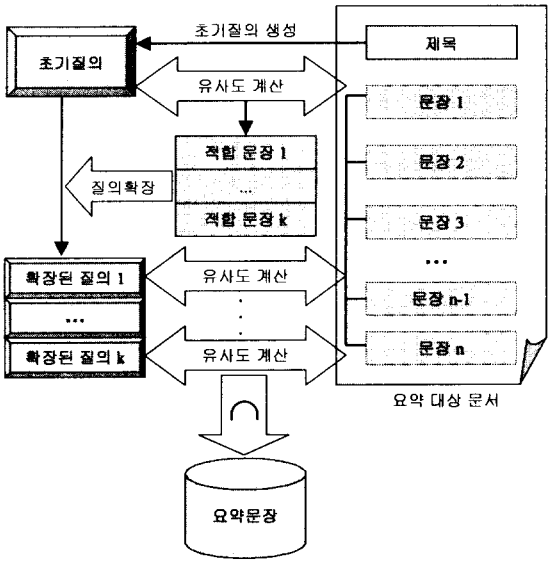
2. 관련연구

자동 문서요약에 관한 연구들은 연구방법론에 따라 크게 언어학적 접근방법과 통계기반 접근방법으로 나누어 볼 수 있다.

언어학적 접근방법은 어휘사슬(lexical chain)이나 담화트리(discourse tree) 등을 이용하여 문서의 담화구조(discourse structure)를 파악한 다음 요약을 제시하는 방법이다[3, 9]. 통계기반 접근방법은 단어의 빈도, 제목, 문장의 길이, 문장의 위치, 실마리 단어나 구(cue word or phrase) 등을 자질(feature)로 사용하여 각 문장이나 문단의 중요도 값을 구하여 그 값이 높은 문장이나 문단을 요약으로 제시하는 방법들이 있었다[5, 7]. 이 둘을 혼합한 접근방법들도 있다[4, 6].

질의확장 기법을 적용한 연구로는 [8]과 [6]이 있다. [8]은 INQUERY 검색 시스템의 지역적 문맥분석(local context analysis)을 이용하여 요약을 생성하였으나, 질의확장을 사용하지 않은 요약에 비해 성능 향상을 보이지 못했다. 이에 대해 [6]은 의사 적합성 피드백, 제목, 문서의 첫 문장 등을 이용하여 좀더 다양하게 질의확장을 적용함으로써, 질의확장이 성능 향상에 기여함을 보였다. 이 두 연구 모두 초기질의를 TREC 실험 문서집합에 주어진 질의를 사용하여 사용자 주도 요약을 생성하였다.

본 논문은 문서의 제목을 초기질의로 사용하여 의사 적합성 피드백을 통해 질의를 확장해 가면서 포괄적 요약을 생성하는 시스템을 제안한다.



[그림 1] 질의확장을 이용한 자동 문서요약 모델

3. 질의확장을 이용한 문서요약

정보검색이 문서집합에서 사용자가 원하는 몇 개의 적합한 문서를 찾아내는 것이라면, 문서요약은 한 문서, 즉 문장집합에서 그 문서의 내용을 대표하는 몇 개의 문장을 찾아내는 작업으로 생각할 수 있다.

정보검색에서는 자동으로 질의를 확장하기 위한 하나의 방법으로 처음 검색된 상위 문서를 적합 문서로 간주하고 그 문서의 단어를 질의에 추가하는 의사 적합성 피드백을 사용한다.

본 논문에서는 의사 적합성 피드백을 통한 질의확장 기법을 문서요약에 적용한다. 제안하는 모델의 구성도는 [그림 1]과 같다. 제안하는 시스템이 기존의 질의확장 기법을 적용한 방법들과 다른 점은 적합 문장을 이용하여 초기질의를 확장할 때 적합 문장 전부를 초기질의에 한꺼번에 적용하지 않고 적합 문장 각각을 개별적으로 질의확장을 적용하여 적합 문장 개수만큼의 질의로 분해한다는 것이다.

$$Q_i^{\text{new}} = Q_0 + S_i \quad (i = 1, \ldots, k)$$

Q_i^{new}는 새로이 확장되는 질의 벡터이고, Q_0는 초기질의 벡터, S_i는 i번째 적합 문장 벡터를, k는 적합 문장의 개수를 뜻한다.

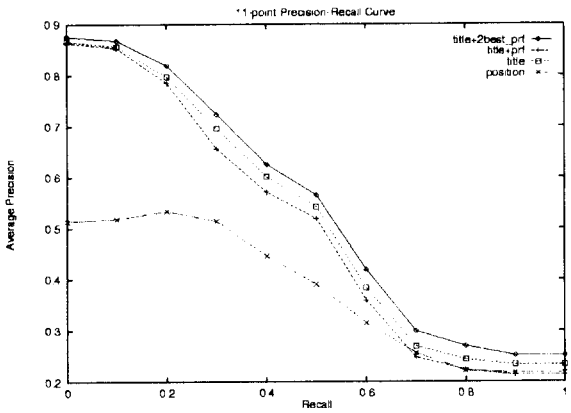
이 모델은 제목과 문서내의 문장들에서 모두 품사 태거를 사용하여 명사만 추출한 후 제목벡터와 각 문장벡터를 생성한다. 문서의 제목은 해당 문서의 내용을 대표한다는 가정 하에, 제목을 초기질의로 설정한다. 이 초기질의와 각 문장 간의 코사인 유사도를 계산하여 상위 k개의 문장을 선택한다. 유사도가 같을 경우에는 문서에서 먼저 나오는 문장을 선호한다. 선택된 k개의 문장을 모두 초기질의 하나를 확장하는데 사용하지 않고, k개의 문장을 각각 개별적으로 초기질의의 확장에 이용하여 k개의 질의를 생성한다. 각 k개의 질의에 대해 문서의 문장들과 유사도를 계산하여 k개의 적합 문장 집합을 생성한다. 이 집합들에 출현하는 문장을 빈도순으로 정렬하여 고빈도 문장을 요약문으로 제시한다. 출현 빈도가 같을 경우 문서내에서 먼저 나오는 문장을 선호한다.

의사 적합성 피드백의 특성상 처음 찾아내는 유사도가 가장 큰 문장이 모델의 성능에 큰 영향을 미치게 된다. 제목이나 문서내 문장의 정보가 불충분하여 유사도 값이 동일한 문장이 다수 존재하는 등의 이유 때문에 유사도가 큰 문장이 부적합한 경우가 발생한다. 이 경우 상위 k개의 적합 문장을 한꺼번에 질의확장에 적용하면 너무 포괄적인 질의로 확장이 되어 불필요한 노이즈 문장이 요약문에 포함될 수 있다. 이 문제를 해결하기 위해 k개의 문장을 초기질의의 확장에 각각 사용하여 k개의 확장된 질의를 생성한다. 이렇게 질의를 분해하여 노이즈 문장에 대해 좀더 배타적인 질의를 생성할 수 있다.

제안하는 모델은 초기질의를 k개의 질의로 분해함으로써, 질의확장 과정에서 질의가 편향(bias)되어 요약이 잘못 생성되는 문제를 완화시킬 수 있다.

4. 실험 및 평가

실험에 사용된 데이터는 KORDIC에서 소프트과학 프로젝트의 일환으로 수집한 신문기사 문서집합 중 486개의 문서를 사용하였다[1]. 이 문서 집합은 각 문서마다 10%



[그림 2] 11-포인트 정확률-재현율 곡선

추출요약(extract)과 30% 추출요약이 표시되어 있고, 수동으로 작성한 요약이 첨부되어 있다. 이 중에서 본 실험에서는 30% 추출 요약만을 사용하였다. 이 실험 문서들은 평균 16.46개의 문장으로 구성되어 있다.

요약의 길이는 문서의 길이와는 무관하므로, 압축률(compression ratio)을 고정하여 실험하는 것보다는 요약의 길이를 고정해두고 실험하는 것이 더 적당하다[6]. 본 실험에서는 요약문장의 개수를 4로 고정하여 실험하였다.

성능 평가의 척도로는 정확률과 재현율을 사용하였다.

$$\text{정확률} = \frac{\text{모델이 제시한 올바른 요약문 개수}}{\text{모델이 제시한 요약문의 총 개수}}$$

$$\text{재현율} = \frac{\text{모델이 제시한 올바른 요약문 개수}}{\text{올바른 요약문의 총 개수}}$$

[그림 2]는 선형 보간법을 이용한 11-포인트 정확률-재현율 곡선이다. 기존의 검색시스템에서 주로 사용하는 방법인 문서의 앞부분을 요약으로 제시하는 방법(position)은 성능이 가장 낮았고, 제안하는 시스템(title+2best_prf)의 성능이 가장 높았다. 이때 사용된 파라미터 값은 k=2이다. 제목을 초기질의로 사용하고 초기질의와의 유사도 상위 첫 문장에 대해서 일반적인 적합성 피드백을 사용한 경우(title+prf)는 제목만을 가지고 유사한 문장을 요약으로 제시하는 방법(title)보다 오히려 성능이 저하됨을 알 수 있다. 이 문제는 의사 적합성 피드백을 적용하는데 있어서 잘못 편향되었기 때문에 발생한다. 제안하는 모델에서는 질의를 분해하여 이 문제를 완화시킬 수 있었다.

5. 결론 및 향후 연구

본 논문에서는 정보검색 분야에서 사용되는 질의확장 기법을 이용한 문서요약 모델을 제안하였다. 초기질의는 문서의 제목으로 시작하고, 질의확장 과정에서 질의를 분해하여 잘못 편향되지 않도록 한다. 실험 결과, 제안하는 방법으로 요약을 생성하는 경우가 질의확장을 이용하지 않는 경우나 질의를 분해하지 않고 일반적인 적합성 피드백을 적용하는 방법 보다 더 좋은 성능을 보였다.

제안하는 모델은 포괄 요약뿐만 아니라 사용자 주도 요약에도 적용할 수 있는 모델이다. 사용자가 입력한 질의를 초기질의로 시작하여 질의확장 과정을 거치면 사용자 주도 요약을 생성할 수 있다.

질의 분해를 통해 문서요약 시스템의 성능이 개선된다는 것은 요약 문장들이 하나의 밀집된 클러스터를 이루는 것이 아니라 여러개의 소규모 클러스터로 이루어져있다는 의미가 된다. 이 사실을 뒷받침하기 위해 요약 문장들간의 분포에 대한 연구가 더 필요하다.

참고 문헌

[1] 김태희, 박혁로, 신중호, "검색/요약/필터링을 위한 텍스트 이해 모형 연구", 제3회 소프트과학 워크숍, 1999.

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

[3] Daniel Marcu, "Discourse trees are good indicators of importance in text", Advances in Automatic Text Summarization, pp.123-136, The MIT Press, 1999.

[4] Eduard Hovy and Chin-Yew Lin, "Automated Text Summarization in SUMMARIST", Advances in Automatic Text Summarization, pp.81-94, The MIT Press, 1999.

[5] H. P. Edmundson, "New Methods in Automatic Extracting", Advances in Automatic Text Summarization, pp.23-42, The MIT Press, 1999.

[6] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell, "Summarizing Text Documents: Sentence Selection and Evaluation Metrics", Proceedings of ACM-SIGIR'99, pp.121-128, 1999.

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

[8] Mark Sanderson, "Accurate User Directed Summarization from Existing Tools", Proceedings of the 7th International Conference on Information and Knowledge Management, pp.45-51, November, 1998.

[9] Regina Barzilay and Michael Elhadad, "Using Lexical Chains for Text Summarization", Advances in Automatic Text Summarization, pp.111-121, The MIT Press, 1999.