제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

명사-동사 공기패턴을 이용한 문서 자동 요약

남기종, 이창범, 강대욱*, 박혁로
전남대학교 전산학과
e-mail:{kjnam,cblee,hrpark}@dal.chonnam.ac.kr
dwkang*@chonnam.chonnam.ac.kr

Automatic Text Summarization using Noun-Verb Cooccurrence Pattern

Ki-Jong Nam, Chang-Beom Lee, Dae-Wook Kang*, Hyuk-Ro Park
Dept of Computer Science, Chonnam 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 위한 하나의 방법으로 문서 자동요약에 관한 연구가 활발히 진행되고 있다. 본 논문의 목적은 어휘 연관성 정보를 이용하여 한국어 문서를 자동으로 요약하는 효율적이며 효과적인 모형을 개발하는 것이다. 제안한 방법에서는 신문기사와 같은 특정 부류에 국한되는 단어간의 어휘연관성을 이용하여 명사-명사 공기패턴과 명사-동사 공기패턴를 구축하여 문서요약에 이용한다. 크게 불용어 처리 단계, 공기패턴 구축 단계, 문장 중요도 계산 단계, 요약 생성단계의 네 단계로 나누어 요약을 생성한다. 30% 중요문장 추출된 신문기사를 대상으로 평가한 결과 명사-명사 공기패턴과 빈도만을 이용한 방법보다 명사-동사 공기패턴을 이용한 방법이 좋은 결과를 가져 왔다.

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[7,9,11].

정보의 양은 기하 급수적으로 증가하고 있지만, 정작 필요한 정보를 얻는 것은 점점 더 어려워지고 있다. 유용한 정보와 불필요한 정보를 구분하고, 어디에 원하는 정보가 있는지 알아내는 것은 점점 힘들어 지고 있다. 이제는 선택의 문제를 넘어서 선택된 즉, 검색엔진이 결과로써 보여주는 문서 중에서 얼마나 빨리 그리고 정확하게 문서의 적합성을 판단할 수 있느냐 하는 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[1,9].

효율적인 정보 접근을 제공함과 동시에 정보 과적재를 해결하기 위한 하나의 방법으로 문서 자동 요약은 그 생성 방법에 따라 추출(extract)과 요약(abstract)으로 구분될 수 있다[6]. 추출은 문서에 나타나는 문장을 그대로 추출하여 요약에 사용하는데 비해, 요약은 문서 내 문장을 이용하여 새로운 문장을 만들어 요약을 한다는 점에서 추출 요약보다 훨씬 정교한 언어처리를 필요로 한다.

본 논문에서는 명사-동사 공기패턴을 이용한 문장 추출 요약 모델을 제안한다.

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

*본 연구는 한국과학재단 목적기초연구(R05-2001-000-01480-0)지원으로 수행되었음.

2. 관련연구

기존의 문서에 대한 요약 생성에 관한 연구들은, 크게 통계적 기법과 문맥 구조에 기반한 방법, 그리고 지식에 기반한 방법으로 구분할 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[3,4].

문맥 구조에 기반한 방법은 문장들 사이의 문맥 관계를 파악하여 요약문을 생성한다[10].

지식에 기반한 방법은 생성하고자 하는 문서와 관련된 배경 지식을 이용하여 요약문을 생성하는 방법이다[2,11].

어휘에 대한 지식베이스를 이용한 방법으로는 WordNet과 시소러스를 이용한 방법들이 있다.[9,11].

WordNet은 요약문의 길이 조절에 대한 처리를 하지 않았고, 시소러스는 고정밀도의 시소러스를 구성하는 것은 어렵다.

본 논문에서는 특정부류에 속하는 문서의 한 문장 안에 추출된 단어들은 일정한 공기패턴을 가지고 있고, 자주 나타타는 공기패턴 정보를 중요한 문장을 추출하는데 이용한다.

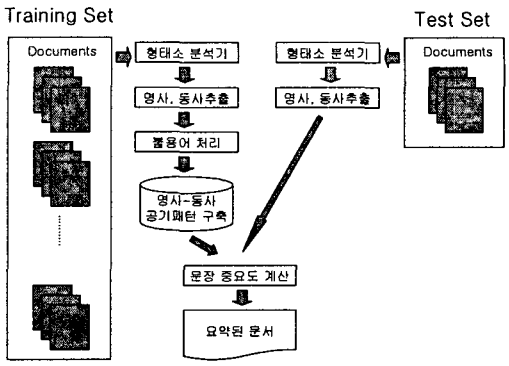
본 논문에서는 명사-동사 공기패턴을 이용하여 단어들의 연관성을 구성한 명사-동사 공기패턴의 정보를 토대로 문장 중요도가 높은 문장을 추출하는 문서 요약 방법을 제안한다.

3. 명사-동사 공기패턴을 이용한 문서 자동 요약

본 논문에서 제안하는 모델은 신문기사와 같은 특정부류에 속하는 문서들은 단어들간의 어느 정도 일정한 공기패턴이 있다고 볼 수 있다. Training Set의 한 문장 내에서 발생하는 단어들의 연관된 정보를 이용한 패턴을 공기패턴이라 하겠다. 예를 들어 "잠-잤다" "꿈-꾸었다" "먹이-먹다" "셈-세다" "마개-막다"와 같은 명사-동사 패턴을 이용하여 명사-동사 공기패턴을 구축하였다.

특정한 부류에서 단어들간의 공기패턴은 발생 빈도가 높다. 자주 발생하는 단어들의 연관된 공기패턴은 그 부류에서 중요함을 내포하고 있다. 명사-동사 공기패턴을 이용하여 Test Set의 각 문장에 있는 명사-동사를 추출하여 명사-동사 공기패턴을 적용하여 각 문장들의 중요도를 주게 되면 중요도가 높게 측정된 문장은 그 문서를 대표 할 수 있다.

전체 논문에서 구현한 자동 요약 시스템의 구성도는 [그림3.1]과 같다. 자동 요약 과정은 크게 불용어 처리 단계, 공기패턴 구축 단계, 문장 중요도 계산 단계, 요약 생성 단계 등 4단계로 구성된다.



[그림3.1] 명사-동사공기패턴을 이용한 문서자동요약

3.1 불용어 처리 단계

형태소 분석과 태깅 과정을 거친 후 각 문장에 대해 명사와 동사를 추출하고 불용어 처리를 한다.

Training Set에서 너무 빈도가 높은 단어는 변별력이 좋지 않다. 문장 중의 의미 없는 단어이거나 자주 출현하는 단어를 파악하여 공기패턴 구축 시에는 이러한 단어들을 제외시켰다. 불용어의 예는 [표2.1]과 같다.

[표2.1] 불용어 처리 예제

	예제
불용어	"것", " ", "데", "군", "느", "하", "게", "라", "듯", "느", "하", "낸", "눈", "블".etc

3.2 공기패턴 구축 단계

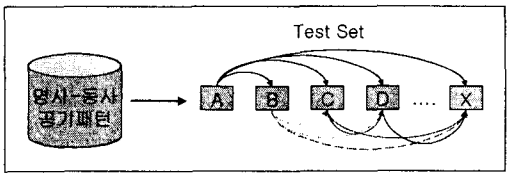
요약 대상 문서의 연관성을 파악할 수 있는 공기패턴을 구축하는 단계이다. 추출한 명사와 동사들 간의 연관관계를 이용하여 공기패턴을 구축한다.

추출된 명사와 동사들간의 연관 관계를 공기패턴을 이용하여 파악한다. 만약, 명사와 동사들 간에 연관 관계가 형성된다면 그들 사이에 링크를 형성하고, 그렇지 않다면 다른 명사와 동사들 간을 비교한다. 모든 Training Set 문서들의 모든 명사, 동사가 처리될 때까지 반복하여, Training Set 문서들에 대한 명사-동사 공기패턴을 완성한다.

3.3 문장 중요도 계산

요약 대상 문서의 문장 중요도를 계산하는 단계이

다. 명사-동사 공기패턴에 있는 패턴정보를 이용하여 Test Set에 있는 각 문장에 명사와 동사의 중요도를 계산하였다. 문장 중요도 계산방법은 [그림3.2]와 같다.



[그림3.2] 문장 중요도 계산방법

$$Total\ Score = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} f(Wi+Wj) \quad \left\{ \begin{array}{l} f(\,) = x \ \ .....(1) \\ f(\,) = 0 \ \ .....(2) \end{array} \right\} \text{식(3.1)}$$

식(3.1)에서
$f(Wi, Wj)$ 값은
Wi, Wj 가 명사-동사 공기패턴에 있으면 (1)
Wi, Wj 가 명사-동사 공기패턴에 없으면 (2)
※ Wi, Wj = 한 문장에 있는 noun_verb
※ x = 명사-동사 공기패턴에 빈도수

다음과 같은 계산방법으로 각 문장에 대한 Total Score를 계산 한다.

3.4 요약 생성

각 문장의 중요도를 계산한 후 상위 30% 문장을 파악하여 요약문을 생성하는 단계이다.

긴 문장의 선호도를 해소하기 위해 각 문장의 길이에 비례하여 나누어 주었다. 요약 생성된 문장을 원문서의 맞게 정렬을 시켰다.

4. 실험

실험에 사용한 데이터는 한국과학기술정보연구원(KISTI)에서 제공되는 수동요약 테스트 컬렉션(test collection)를 사용하였다.[7] 신문기사(1,000건)에 대해 각각 10%, 30% 중요문장 추출, 10% 수동요약 결과로 구성되어 있다.

본 실험에서는 신문 기사 중 정형화된 971건을 사용하였다. 971건 중 871건은 명사-동사 공기패턴을 구축하는데 사용하였고, 나머지 100건은 Test Set으로 사용하였다.

구축한 명사-명사 공기패턴과 명사-동사 공기패턴의 통계적인 특성은 [표4.1]과 같다.

[표4.1] 패턴구축용 문서집합의 통계적인 특성

대상 영역	명사-명사 공기패턴	명사-동사 공기패턴
문서 개수	871건	871건
단어 수	25,563(명사)	41,974(명사&동사)
문장 수	14,735	14,735

테스트컬렉션의 문서집합 중에서 100건에 대해서 실험을 하였다. 실험한 Test Set 100건의 통계적인 특성은 [표4.2]와 같다.

[표4.2] 실험대상 Test Set의 통계적인 특성

대상영역	명사-명사 공기패턴	명사-동사 공기패턴
문서개수	100건	100건
문서의 평균길이	19.9	19.9
30%요약의 평균길이	6.3	6.3
단어평균수	114.4	168.3

Test Set에서 30% 중요문장을 추출한 후 테스트 컬렉션에 있는 30% 중요문장 부분과 비교하였다. 요약에 포함된 전체 문장수와 테스트 컬렉션의 30% 요약과 일치하는 문장수를 비교하였다. Test Set 100개의 문서에 대한 본 논문이 제안하는 방법과 명사-명사 공기패턴만을 가지고 이용한 방법과 명사와 동사의 빈도수 2회 3회만을 비교한 결과는 아래의 표와 같다.

[표4.3] 871건의 명사-명사 공기패턴을 가지고 테스트한 결과

	명사-명사 공기패턴	Frequency 2	Frequency 3
합계	148	140	143
평균	1.48	1.4	1.43

[표4.4] 871건의 명사-동사 공기패턴을 가지고 테스트한 결과 (제안한 모델)

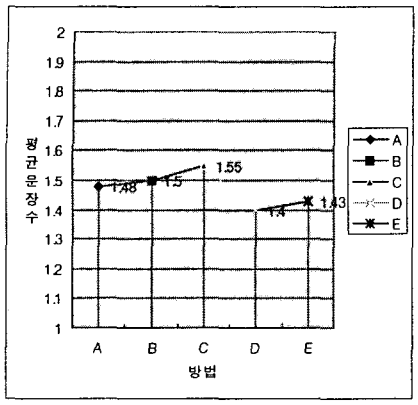
	명사-동사 공기패턴	Frequency 2	Frequency 3
합계	150	148	149
평균	1.5	1.48	1.49

[표4.5] 871건의 명사-동사 공기패턴의 빈도수를 가지고 테스트한 결과(제안한 모델)

	명사-동사 공기패턴 빈도수 포함	Frequency 2	Frequency 3
합계	155	148	149
평균	1.55	1.48	1.49

[그림4.1]은 각 실험 방법의 평균 일치 문장수를 나

타낸다.



[그림4.1] 각 실험 방법의 평균 일치 문장수

단순히 단어의 출현 빈도만을 고려하거나 명사만을 기반으로 요약문을 생성하는 방법과 본 논문이 제안하는 모델을 비교하기 위하여 다음과 같은 방법으로 실험을 하였다.

- A : 명사-명사 공기패턴을 이용한 주제어 선택
- B : 명사-동사 공기패턴을 이용한 주제어 선택(제안한 모델)
- C : 명사-동사 공기패턴의 빈도수를 이용한 주제어 선택(제안한 모델)
- D : 출현빈도 2번의 명사를 주제어로 선택
- E : 출현빈도 3번의 명사를 주제어로 선택

실험결과 단어의 출현 빈도만을 고려한 것과 명사-명사 공기패턴을 이용한 방법보다 명사-동사 공기패턴이 좋은 결과를 보였다. 명사-동사 공기패턴에서 빈도수를 포함한 실험을 해본 결과 더욱 향상된 결과를 보였다.

5. 결론 및 향후 연구

본 논문에서는 명사-동사 공기패턴을 이용하여 문서를 요약하는 방법을 제안하였다. 문서 전체의 주제를 표현하는 중심 문장을 추출하는데 명사-동사 공기패턴을 이용하였다. 제안한 모델에서 30% 중요 문장을 추출한 결과를 기준으로 명사-명사 공기패턴을 구축한 결과와 빈도수 2회, 3회 비교를 하였을 때 본 논문에서 제안하는 방법이 좋은 결과를 가져왔다. 명사-동사 공기패턴만을 이용하는 것보다 명사-동사 공기패턴의 빈도수를 포함한 실험을 한 결과 더 좋은 결과를 가져 왔다.

본 논문에서는 요약 대상 문서의 주제를 파악하기 위해서 명사-동사 공기패턴만을 이용하였다. 그러나 871개 문서에 대한 41,974건의 단어만을 가지고 공기패턴을 구축한 결과 공기패턴이 작다는 단점이 있었다. 더 많은 공기패턴을 구축하여 실험할 필요성이 있다. 또한 "것"과 같은 의미 없는 단어가 자주 사용되는 경향이 있었다. 불용어 처리를 더욱 개선한다면 더 좋은 결과를 가져올 것이다.

참고문헌

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

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

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

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

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

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

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

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

[9] 이창범, 박혁로, "시소러스를 이용한 문서 자동 요약", 한국정보과학회 제28권 1호, pp352-354, 2001.

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

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