어휘 클러스터링을 이용한 자동 문서 요약

김건오[0] 고영중 서정연
서강대학교 컴퓨터학과 자연어처리연구실
(kono@nlprep.sogang.ac.kr[0], kyj@nlpzodiac.sogang.ac.kr, seojy@ccs.sogang.ac.kr)

Automatic Text Summarization with Lexical Clustering

Kono Kim[0] Youngjoong Ko Jungyun Seo
NLP Laboratory, Dept. of Computer Science, Sogang University

요 약

자동 문서 요약 시스템은 문서내 담겨있는 정보를 최대한 표현하면서 문서의 크기를 줄이는 시스템이다. 본 논문에서는 어휘를 자동으로 클러스터링하여 문서 대표어를 찾고, 이를 제목과 조합하여 요약을 수행하는 시스템을 제안한다. 특히 이 시스템은 제목이 없는 문서도 요약을 수행할 수 있는 장점이 있다. 비교시스템으로는 제목, 위치, 빈도를 이용한 시스템을 구축하여 사용하였으며 30%, 10%, 그리고 4문장 요약에서 제안한 시스템은 모두 우수한 성능을 보였다.

1. 서 론

정보의 양이 많아짐에 따라 자동 문서 요약의 필요성이 커지고 있다. 사람이 문서 요약을 할 때는 언어적 지식을 가지고 문서를 이해하여 요약문을 생성하는데 반해, 기계가 요약을 수행하도록 하기 위해서는 이런 언어적 지식을 대체할 지식베이스(knowledgebase) 구축이 필요하다. 기존의 Wordnet, 시소러스 등은 기계가 사용할 수 있는 언어적 지식으로 매우 좋은 자원이나, 구축하기 위해서는 많은 시간과 비용이 소요되는 단점이 있다. 본 논문에서는 단어간의 공기정보를 이용하여 자동으로 언어적 지식을 구축하고 이를 이용하여 요약을 수행함으로써 기존의 통계적 방법과 언어학적 방법의 장점들을 모두 취했다.

본 논문은 다음과 같이 구성되어 있다. 2장에서 자동 문서 요약 관련 연구들에 대해 살펴보고, 3장에서 어휘 클러스터링을 이용한 시스템을 제안한다. 4장에서 실험 및 평가를 하고, 5장에서 결론 및 향후과제를 기술한다.

2. 관련 연구

자동 문서 요약에 대한 기존 연구들은 크게 두가지로 분류된다. 단어간의 의미관계, 문장내 구나 절의 구조적 정보 등을 이용한 언어학적 방법과 제목, 단어의 빈도, 문장의 위치, 단서어 등의 통계적 정보를 주로 사용하는 방법이다.

2.1 언어학적 방법

이 방법은 문서에 대한 직접적인 이해를 시도한다. 어휘 사슬(lexical chain)을 이용한 방법[1]은 Wordnet을 이용하여 단어의 의미관계를 파악하여 어휘 사슬을 만들고, 강한 어휘 사슬을 중심으로 요약을 수행한다. 담화 구조(discourse structure)에 기반한 방법[2] 은 각 문장의 의미와 문장간의 관계 분석을 통한 문맥 구조의 파악을 바탕으로 이루어진다. 이런 언어학적 방법은 고품질의 요약문을 생성할 수 있지만, 속도나 확장 가능성 면에서 아직 많은 개선이 필요하다.

2.2 통계에 기반한 방법

통계 기반 접근 방법은 학습과정에서 요약에 참고할 통계적 자질들을 추출한다. 이러한 통계적 자질로는 특정 단어의 빈도, 제목, 문장의 길이, 문장의 위치, 단서어(clue word) 등이 있다. 이러한 자질이 추출되면, 문서 내의 각 문장이나 문단의 중요도 값을 구하여 그 값이 높은 문장이나 문단을 요약문으로 제시하거나 주어진 자질들을 이용

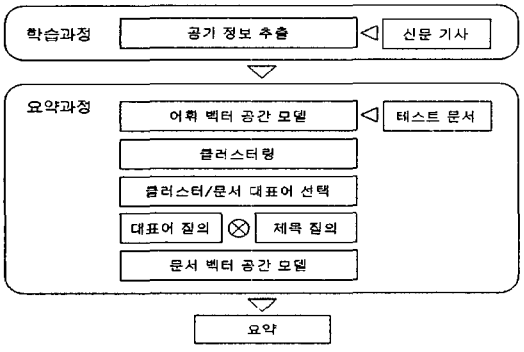
하여 기계학습 기법을 적용함으로써 요약문을 생성한다 [3][4]. 속도가 빠르고 구현이 쉬우나 제목이 없는 문서 등 통계적 자질이 충분하지 않은 도메인의 경우 적용하기 힘들다.

3. 어휘클러스터링을 이용한 문서요약

본 논문에서 제안하는 시스템은 언어학적 방법과 통계에 기반한 방법을 혼합한 방법이다. 통계에 기반한 방법은 단어간의 의미관계가 고려되지 않고 통계적으로만 접근함으로써 문서의 내용을 충분히 반영하지 못한다. 또한 언어학적 방법은 Wordnet 과 같은 지식이 구축되어있지 않으면 사용할 수 없는 단점이 있다. 본 논문에서는 단어간의 의미관계를 고려하기 위해 문서 내 각 단어들을 어휘 벡터로 표현하였다. 각 어휘 벡터는 대용량의 말뭉치(corpus)에서 빈도(term frequency)가 높은 기준 단어 450개를 추출하고 이 기준단어와의 공기정보(co-occurrence)값으로 표현된다. 공기정보를 구하는 식은 다음과 같다.

$$\text{공기정보} = \frac{\text{문장내 동시에 사용된 횟수}}{\text{단어}A\text{의 사용 횟수} \times \text{단어}B\text{의 사용 횟수}} \quad (1)$$

단어간 의미유사도는 각 어휘 벡터간의 내적으로 계산한다. 이렇게 벡터로 표현된 어휘들은 k-Means 알고리즘으로 클러스터링된다. 이렇게 생성된 여러 클러스터들 중 핵심 클러스터를 선택하여, 그 핵심 클러스터를 대표하는 단어를 문서 대표어로 뽑고, 그 문서 대표어를 포함한 문장들과 제목에 사용된 단어를 포함한 문장들을 추출한다. 요약은 이렇게 추출된 문장 중에서 중요한 문장을 선택하여 생성된다. 제목이 없는 문서의 요약은 문서 대표어를 포함한 문장만으로 생성한다.



[그림 1] 어휘 클러스터링을 이용한 자동 문서 요약

기준 단어와 공기정보 추출을 위해 조선일보 1996 ~ 1997 년 신문기사 (약 1,660만 어절)를 이용하였다. 일반명사와 고유명사를 대상으로 하였으며, 저빈도 단어(term frequency < 3) 와 관련 없는 단어쌍(mutual information < 3)을 제외한 후 식(1)에 의해 두 단어의 공기정보 값을 구한다. 기준 단어는 빈도를 기준으로 상위 450 개를 사용한다. 학습과정에서 생성한 공기정보값과 기준 단어는 요약과정에서 주어진 문서를 어휘 벡터 공간으로 만드는데 사용된다.

	시장	업체	계획	미국	기업	공간
가게	★	★		★	★	
가격	★	★	★	★	★	
가상		★			★	★

★ = 계산된 공기정보값

[그림 2] 문서 벡터 = {가게, 가격, 가상}, 기준단어= { 시장, 업체, 계획, 미국, 기업, 공간} 일 경우 어휘 벡터 공간

요약 대상 문서가 주어지면 [그림2] 와 같이 문서에 사용된 어휘들을 각각 벡터로 표현하고 이를 클러스터링한다. 이 때 클러스터링은 k-Means를 사용한다. k 의 개수는 문서에 사용된 단어의 개수가 많을수록 증가시키는데, 매 50 단어 마다 하나씩 증가시킨다. 각 어휘 클러스터는 비슷한 의미를 가지는 어휘들로 구성되며, 이 중 어떤 클러스터가 정보가 많은 핵심 클러스터인가를 판별하기 위해 다음과 같은 식(2)에 의해 점수를 매겨 최상위 클러스터와 최상위 클러스터와 점수 차이가 10% 이내인 클러스터를 핵심 클러스터로 선택한다.

$$\text{클러스터점수} = \frac{\sum \text{클러스터에 속한 단어의 빈도}}{\text{단어의 종류}} \quad (2)$$

즉 단어의 종류가 적으면서 문서내 사용 빈도가 높은 단어가 포함되어 있는 클러스터가 핵심 클러스터로 선택된다. 이렇게 선택된 클러스터에서 최상위 빈도 단어와 최상위 빈도 단어와의 빈도 차이가 10% 이내인 단어가 문서 대표어이다. 요약문은 제목과 문서 대표어를 가지고 각각 질의를 수행하여 선택된 문장 중에서 공통으로 뽑인 문장을 먼저 뽑고 제목 사용 질의와 문서 대표어 사용 질의 결과가 일치하지 않을 경우 문장 위치가 문서내 상위인 문장이 선택된다.

4. 실험 및 평가

실험 데이터로는 연구개발센터(KORDIC) 문서요약 집합(신문기사)을 사용하였다. 이 문서집합은 제목과 본문, 10% 요약과 30 % 요약 그리고 수동요약으로 나뉘어져 있다. 이 문서집합은 1,000 개라고 보고되었으나[5], 중복된 문서와 제목이 없거나 요약이 없는 문서를 제외한 816 개를 사용하였다. 압축률을 고정한 경우(10%, 30%) 와 길이를 고정한 경우(4문장)를 모두 실험하였다. 비교시스템으로는 제목을 질의로 사용한 시스템(제목)과 문서 내 위치가 상위인 문장으로 요약문을 제시하는 시스템(위치), 그리고 문장에 사용된 단어들의 빈도 합계가 높은 문장을 요약문을 제시하는 시스템(빈도)을 사용하였다. 성능평가의 척도로는 다음과 같은 F1 값을 사용하였다.

$$F_1 = \frac{2(\text{재현율} \times \text{정확율})}{\text{재현율} + \text{정확율}} \qquad (3)$$

[표 1] 실험 결과

방법	30%	10%	4문장
제안 시스템	51.1	51.2	53.6
제목	48.6	43.3	51.6
위치	49.4	46.6	51.6
제안 시스템※	44.4	39.6	47.1
빈도	35.9	14.8	38.4

제안시스템※ 은 제목이 없는 경우를 가정하고 문서 대표어 만으로 요약을 생성한 경우이다.

[표 2] 실험 결과 분석

문서 개수	816 개
문서의 평균 길이	16.37 문장
문서의 평균 단어 종류	195.94 개 (명사)
평균 클러스터 개수	문서당 2.98 개
평균 핵심 클러스터 개수	문서당 1.38 개
대표어 평균 단어 수	문서당 1.79 개

제안 시스템은 압축율을 고정하거나, 길이를 고정한 경우 모두 비교 시스템에 비해 나은 성능을 보였다. 위치가 성능이 좋은 이유는 신문기사가 대부분 두괄식으로 구성되어 있기 때문인 것으로 보인다. 따라서 신문기사가 아닌 도메인에서 제목없는 문서의 요약을 할 경우에 문서 대표어 만으로 요약을 수행한 제안시스템※ 이 매우 유용할 것이다.

5. 결론 및 향후 과제

공기정보를 이용하여 어휘 클러스터링을 하고, 핵심 클러스터를 찾아서 이를 문서 요약에 이용함으로써 다른 모든 비교 시스템에 비해 나은 성능을 보였다. 제안한 방법은 문서 내 단어들의 의미관계를 파악하는데 Wordnet 과 같은 수동 구축한 언어 지식 자원을 사용하지 않고도 가능하다는 장점이 있으며, 제목이 없는 문서나, 요약문이 위치와 상관없는 도메인에서도 사용될 수 있다. 향후 연구로 문서 대표어가 다른 자질들과 어떻게 같이 조합될 경우에 성능이 향상되는지와 각 문서별 특징에 따라 클러스터 개수가 자동으로 생성되도록 하는 연구가 필요하다.

참고 문헌

[1] Barzilay, R. and M. Elhadad, "Using Lexical Chains for Text Summarization." In Proceedings of the TIPSTER Text Phase III Workshop, 1998.

[2] Marcu, D., "Building up Rhetorical Structure Tree", In Proceedings of the 13th National Conference on Artificial Intelligence, Vol. 2, pp. 1069~1074, 1996

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

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

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