2단계 문장 추출방법을 이용한 자동 문서 요약

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

Automatic Text Summarization with Two Step Sentence Extraction

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

요 약

자동 문서 요약 시스템은 문서내에 담겨있는 정보를 최대한 표현하면서 문서의 크기를 줄이는 시스템이다. 본 논문에서는 문서 요약을 크게 2단계로 나누어서 수행한다. 문장내 요약본으로써의 불필요한 문장을 미리 제거하고 이에 더해 다양한 통계적 요약방법의 여러 장점들을 수용함으로써 보다 나은 성능 향상을 얻을 수 있었다. 비교시스템으로는 제목, 위치, 빈도, 도합유사도, 어휘 클러스터링을 이용한 시스템을 구축하여 사용하였으며 30%, 10% 문장요약에서 제안한 시스템은 모두 우수한 성능을 보였다.

1. 서 론

문서 요약을 위한 효과적인 방법은 사람이 문서를 요약할 때와 마찬가지로 언어적인 지식을 미리 학습하고 습득된 지식베이스(knowledgebase)에 따라 문서를 요약하는 방법이다. 기존의 Wordnet, 시소러스, 어휘 클러스터링 등이 이에 해당되며 기계가 사용할 수 있는 매우 좋은 자원이라 할 수 있다. 그러나 언어학적 방법의 단점은 구축하기 위해서 많은 시간과 비용이 소요된다는 것이다. 본 논문에서는 통계에 기반한 방법을 크게 2단계로 나누어서 통계적 자질들의 장점들을 확인하고 각 단계별로 문장의 중요도가 현저한 방법들을 선형결합함으로써 효과적인 문서요약 시스템을 구현한다. 본 논문은 다음과 같이 구성되어 있다. 2장에서 자동 문서 요약 관련 연구들에 대해 살펴보고, 3장에서 2단계 문장 추출 방법을 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. 이에 반해 어휘 클러스터링을 이용한 방법[2]은 Wordnet을 대신하여 유사 어휘들을 클러스터링을 함으로써 Wordnet과 같은 고비용의 의미 체계가 없이도 단어의 의미 관계를 사용할 수 있다. 담화 구조(discourse structure)에 기반한 방법[3]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.

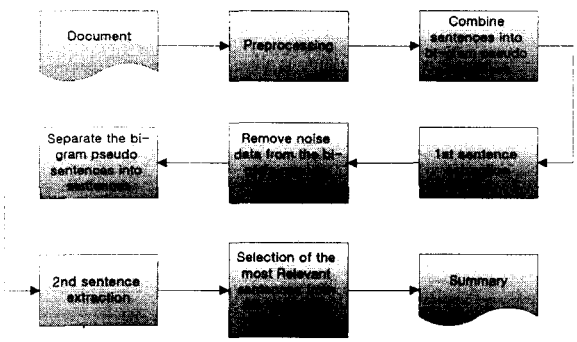
본 연구는 한국과학재단 목적기초연구 (R01-2003-000-11588-0) 지원으로 수행되었습니다.

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 이용하여 기계학습 기법을 적용함으로써 요약문을 생성한다[4][5]. 속도가 빠르고 구현이 쉬우나 제목이 없는 문서 등 통계적 자질이 충분하지 않은 도메인의 경우 적용하기 힘들다.

3. 2단계 문장 추출 방법을 이용한 문서요약

본 논문에서 제안하는 시스템은 통계에 기반한 접근방법이다. 비록 단어간의 의미관계가 고려되지는 않지만 지식베이스(knowledgebase)가 필요 없고 통계적인 장점을 최대한 살리면서 추출 방법의 특성을 고려함으로써 문서내의 자질들만으로 효과적인 문서요약을 가능하게 한다. 어떤 문서를 어느 정도 의미가 있는 단위로 나누어서 각 segment마다 통계적 score가 큰 문장들을 선택하고 이들을 비교하여 최종 요약문장을 선택하는 기법[6]은 비교적 효과적인 방법으로 알려져 있다. 본 논문에서 제안하는 방법은 2단계의 문장 추출과정을 사용한다. 우선 두 단계중 첫번째 단계에서 2개의 문장을 하나의 의미관계로 생각하여 요약문 작성을 위한 후보 문장들을 추출하고 추출된 후보 문장들을 다시 2차적으로 요약함으로써 최종 요약문을 얻는다.



[그림 1] 2단계 문장 추출을 이용한 자동 문서 요약

우선 문서가 주어지면 각 문장들은 bi-gram의 형태의 pseudo 문장들로 재조합되어진다. 중요한 문장들은 주제문의 위치와 상관 없이 일정 위치에 군집되어 있다는 가정과 한개의 문장에서 자질들을 추출하는 것 보다는 두개의 문장에서 자질을 추출함으로써 문장의 의미단위를 확장하여 더 많은 자질을 얻을 수 있게 된다. 1단계의 문장 추출 시도후 우리는 상당한 noise data의 제거 효과와 정답문서와 높은 관련성이 있는 후보 문장들을 얻어 낼 수 있게 된다.

본 시스템은 통계에 기반한 방법들을 사용하기 때문에 일반적인 통계적 기법들의 성능을 [표 1]과 같이 비교해 보았다. 위치와 제목을 사용하는 방법이 가장 높은 성능을 보이고 있다. 그러므로 성능이 가장 낮은 빈도를 이용한 방법을 제외한 3가지 방법을 본 시스템의 문장 추출 방법에 맞게 선형 결합(linear combination)하여 각 통계적 요약 방법들의 장점들을 모두 사용하였다.

[표 1] 실험 결과

방법	30%	10%
위치	49.4	46.6
제목	48.8	43.5
도합 유사도	40.6	23.9
빈도	31.0	14.8

1차 문장 추출에 사용된 방법은 제목과 위치를 선형결합(linear combination)하여 사용하였고 식(1), 2차 문장 추출에서는 기존의 식에 도합유사도를 식(2)을 추가 하였다. 이는 1차 문장 추출과정에서 중요도가 떨어지는 문장이 어느 정도 제거 되고 2차 문장 추출과정에서 도합유사도를 추가 함으로써 전체 문장과의 유사도를 고려하는 도합유사도가 문장의 score를 높이는 역할을 할 수 있기 때문이다.

1차 문장 추출

$$Score(S_i) = sim(S_i, Q) + (1 - \frac{i-1}{N}) \quad (1)$$

2차 문장 추출

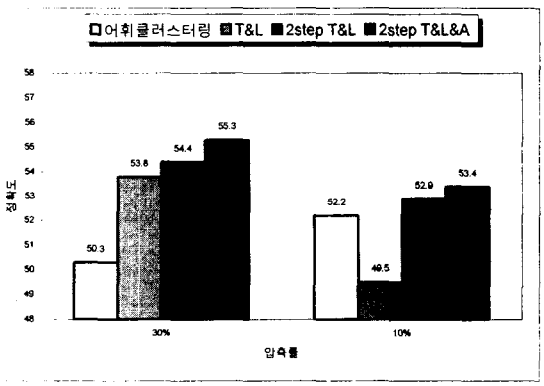
$$Score(S_i) = sim(S_i, Q) + (1 - \frac{i-1}{N}) + w_a asim(S_i) \quad (2)$$

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[7], 중복된 문서와 제목이 없거나 요약이 없는 문서를 제외한 841 개를 사용하였다. 압축률을 고정한 경우(10%, 30%)에 binary matrix를 이용하여 실험하였다. 비교시스템으로는 언어학적인 방법과 통계적인 방법을 고려한 어휘클러스터링을 이용한 시스템[2]과 bi-gram을 사용하지 않는 제목과 위치가 선형 결합(linear combination)된 시스템을 사용하였다. 성능평가의 척도로는 다음과 같은 F1 값을 사용하였다.

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



[그림 2] 실험결과 : T(제목), L(위치), A(도합유사도)

기본적으로 제목과 위치(T&L)를 선형 결합(linear combination)하여 사용 할 경우 30% 요약의 경우 어휘클러스터링을 사용한 시스템[2]보다 높은 성능을 보이나 10% 요약의 경우 낮은 성능을 보인다. 그러나 이를 2단계 문장 추출 방법(2step T&L)에 적용하면 4.1%, 0.7%의 성능 향상을 보인다. 또한 2단계를 거치면서 noise가 제거된 문장들에 서로간에 유사도를 고려하는 도합유사도 방법(2step T&L&A)을 추가 할 경우 관련성이 높은 문장들이 후보문장으로 남게 되므로 5%, 1.2%의 성능 향상을 얻을 수 있다.

5. 결론 및 향후 과제

중요한 문장들은 주제문의 위치와 상관 없이 일정 위치에 군집되어 있다고 고려하여 문장의 bi-gram으로 보다 많은 자질을 추출하여 자질의 부족함을 보완하고 적절한 문장 요약 방법들을 선형 결합(linear combination)하여 후보문장들을 선정하고 도합유사도를 추가하여 요약 방법을 다양화 함으로써 효과적인 문장 scoring을 가능하게 하였다. 이는 Wordnet과 같은 언어적인 지식이 불필요 할 뿐만 아니라 통계적인 장점들을 최대한 사용하기 때문에 문서 요약에 있어서 도메인에 종속적이지 않고 적은 량의 정답문서만으로도 시스템 구현이 가능하다. 향후 연구로는 제목이 없이 문서내 본문의 내용을 가지고 요약하는 방법에 대한 연구가 필요하다. 또한, 본 논문에서 제안한 2단계 자동 문서 요약 방법을 다중 문서 요약에 적용하여 단일 문서 요약과의 차이점을 연구하고 중복되는 문장의 효과적인 제거 방법에 대한 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] 김건오, 고영중, 서정연, "어휘 클러스터링을 이용한 자동 문서 요약", 제29회 한국정보과학회, 2002.

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

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

[5] 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.

[6] J. Larocca Neto, A.D. Santos, C.A.A. Kaestner, A.A. Freitas, Generating Text Summaries through the Relative Importance of Topics." , IBERAMIA-2000 (7th Ibero-American Conf. on Artif. Intel.)

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