시간자질을 이용한 다중 문서요약

임정민 ○ 강인수 배재학+ 이종혁
포항공과대학교 전자컴퓨터공학부 컴퓨터공학과, 첨단정보기술 연구센터,
+울산대학교 컴퓨터 정보통신 공학부
{beuuett○, dbaisk, jhlee}@postech.ac.kr, +jhjbae@mail.ulsan.ac.kr

Multi-Document Summarization using Time Feature

Jung-Min Lim○ In-Su Kang Jae-Hak Bae Jong-Hyeok Lee
Dept. of Computer Science and Engineering, Division of Electrical and Computer Engineering
Pohang University of Science and Technology
and Advanced Information Technology Research(AITrc)
+School of Computer Engineering and Information Technology, University of Ulsan

요 약

시간에 종속적인 문서집합에서 사람이 만든 요약문은 시간에 따른 중요 내용의 분포를 보여준다. 본 논문은 다중 문서요약에 시간 자질을 이용한 문서의 분류와 시간별 문서집합에서 핵심문장과 부가문장을 선별하고, 문장간의 계층적인 클러스터링을 통해서 중요 문장을 선별하는 방법을 제안한다. 동일한 주제를 갖는 문서집합에서 사람이 선택한 중요 문장에 대해서 제안한 방법은 50% 정확률을 나타냈다.

1. 서론

전형적으로 문서요약은 크게 3 단계를 거치면서 이루어진다. 문서에서 중요한 내용을 찾고, 추출한 내용을 개념적으로 재구성하고, 요약문을 생성하는 단계를 거친다. 단일 문서요약에 적용되는 3 단계 과정은 요약대상 문서가 여러 문서인 다중 문서 요약에도 똑같이 적용된다. 그러나 다중 문서요약은 단일 문서요약과 달리 동일한 주제를 갖는 여러 문서에서 중요한 내용을 추출해야 하기 때문에 단일 문서에서 사용한 방법의 개선과 중복된 내용을 제거하는 방법이 요구된다.

시간에 종속적인 문서의 요약은 시간에 따른 중요 내용의 분포를 보여주기 때문에, 문서요약에서 시간 정보는 중요한 내용을 찾는데 자질로 사용될 수 있다.

본 논문에서는 다중 문서요약의 중요문장 추출시, 문서의 시간자질을 이용해서 문서를 분류하고, 각 시간별 문서집합에 Rhetorical Structure Theory (RST)의 nucleus 와 satellite 의 개념을 적용하여 중요문장의 추출 방법을 제안한다.

2. 관련 연구

현재 다중 문서요약에서 중요 내용을 추출하는 방법으로는 Query-based 인 요약의 경우 Query 와 문서간의 유사도, 문장이 속한 클러스터의 범위와 크기, 시간별 가중치, 문장의 통계적 언어학적 특성을 조합해서 가중치를 부여하는 Maximal Marginal Relevance Multi-Document(MMR-MD)방법[1]이 있다.

Cluster-based 방법으로는 입력문서들의 전체에 대해서 중요 단어들로 중심(Centroid)을 구성하고 중요 내용을 추출하는 Centroid-based 방법[2]이 있다.

중요 문장을 추출하기 위해서 문장에서 unigram, bigram trigrams 으로 단어를 조합하고, 각각의 단어 조합에 주제와 관련된 문서집합, 주제와 관련없는 문서집합을 통해서 확률값을 구하고, 이를 이용해서 중요 내용을 추출하는 방법[3]이 있다.

또한, 단어의 표층형태를 이용하지 않고 중요단어의 개념을 사용하여 동의어, 상위어, 하위어로 확장, 핵심 주제어의 개념집합을 이용해서, 중요 문장 및 내용을 추출하는 방법[4][5]이 있다.

위의 방법들은 기존에 단일 문서 요약에서 적용되는 방법을 다중문서 요약에 적용시키기 위해서 변형한 것이다. 또 다른 방법으로는 Text cohesion, Text coherence 을 바탕으로 Lexical Chaining 이나, RST 를 이용한 요약이다. 이 방법들은 아직 다중 문서요약에 적용하기에는 제약이 있다.

Radev 는 단일 문서요약의 RST 처럼 다중 문서요약을 위한 교차 문서구조 이론(Theory of Cross-Document Structure, CST)[6]의 필요성을 제기하였고, 다중 문서집합을 나타내기 위한 기본 구조로 육면체 형태와 그래프 형태를 제시하였다.

또한, Document Understanding Conference(DUC) 2001[7], 2003[8]에서는 CST 를 문장 수준에서 연구하여 17 개의 CST 관계를 제안했다.

3. 시간 자질을 이용한 다중 문서요약

3.1 개요

Radev 가 제안한 CST 는 서로 다른 문서에 존재하는 문장간에 관계를 찾았다. 그러나 이 문서집합은 관련된 주제를 갖는 문서이지만, 각 문서마다 저자가 다르기 때문에, 교차 문서의 문장간에 관계를 설정하고 시스템을 통해서 찾기가 어렵다. 문서간의 관계를 찾기 이전에 시간자질을 이용한 문서의 분류는 문서간의 관계설정에 도움을 줄 수 있다.

시간 종속적 문서집합에서 사람이 중요 내용을 추출한 결과는 시간에 따른 중요 내용의 분포를 보여준다.[표 1] 따라서 시간 종속적인 문서집합을 요약할 경우, 먼저 시간 자질을 이용해서 문서를 분류, 시간별 문서집합으로 구분 짓고, 각 시간별 문서집합에서 핵심(nucleus)부분과 부가(satellite)부분으로 구분하여, 다중 문서요약에서 중요내용 선별과정에 사용하였다.

문서의 날짜별 구분		1	2	3	4	5	6	7	8	9	10
집합 1	문서개수	1	1	1	1	5	2	1	1	4	1
19 문서	중요문장	1		1	1	3				2	1
집합 2	문서개수	1	2	3	1	1	1	2	1		
12 문서	중요문장	1	1	1				1	2		
집합 3	문서개수	2	6	4	1						
12 문서	중요문장		3		3						
집합 4	문서개수	8	1	1	1						
11 문서	중요문장	4	1	1							
집합 5	문서개수	8	2	3							
13 문서	중요문장	3	1	2							

[표 1] 사람이 추출한 중요문장 및 문서의 날짜별 분류

3.2 시간별 문서구분 및 문서의 중요 내용 추출

시간 자질을 이용해서 문서를 분류한 후, 각 문서의 내용을 가장 잘 반영하는 문장들을 추출한다. 추출방법은 기존의 단일 문서 요약에서 전통적으로 사용되는 문장의 위치, 길이, 제목에 존재하는 단어에 가중치 부여, 인용문에 감점 부여, 등의 방법을 이용해서 문서에서 가장 점수가 높은 문장을 선택하였다.

$$S(S_i) = w_1 P_i + w_2 L_i + w_3 \sum_{t_k \in S_i \cap N} t_k \cdot W_n + w_4 \sum_{t_k \in S_i \cap H} t_k \cdot W_p$$

$S_i = \{t_1, \cdots, t_n\}$, P_i : 문장의 위치, L_t : 문장의 길이

N : 감점 단어 및 기호 집합, H : 제목에 있는 단어집합

W_p : 제목에 있는 단어 가중치, W_n : 감점 단어 페널티

[수식 1] 단일 문서에서 중요 문장 추출 계산식

3.3 핵심문장과 부가문장의 구분

시간별 문서집합에서 각 문서를 대표하는 문장들이 추출되면, 먼저 유사한 내용을 갖고 있는 문장들을 구분해서 중복 문장집합을 형성하고, 구성된 중복 문장집합들은 핵심(nucleus) 문장집합 혹은 부가(satellite)문장집합으로 구분된다.

요약 대상문서들의 핵심 주제는 여러 문서에 걸쳐서 공통적으로 나타날 것이기 때문에, 핵심 문장집합을 구별하는 방법은 중복 문장집합의 문장개수와 시간별 문장집합의 공통 단어들을 이용하였다. 그리고 각각의 문장에 존재하는 공통 단어들의 문장성분에 따른 가중치를 이용하여, 가장 높은 점수를 갖는 문장이 포함된 중복 문장집합을 핵심 문장집합으로 하고, 그 외 중복 문장 집합을 부가 문장집합으로 나눈다.

중복된 문장을 찾는 방법은 문장의 의미를 갖는 단어에 가중치를 부여하여 중복도를 계산하는 방법[9]을 사용하였다.

$$N(S_i) = w_1 R_k + w_2 \sum_{t_k \in S_i \cap D_t} t_k \cdot Df_k + w_3 \sum_{t_k \in S_i \cap D_t} W_g(t_k, S_i)$$

R_k : S_i가 속한 중복 문장집합의 문장 개수

D_t : 시간별 문장집합의 공통 단어집합

Df_k : 단어의 빈도수

$W_g(t,S)$: 문장에서 단어 t의 문법적 기능에 따른 가중치

[수식 2] 핵심문장 선별을 위한 문장점수 계산식

3.4 중요 문장 선택과 정렬

핵심문장과 부가문장을 구별한 후, 각각의 시간별 문장집합에서 핵심문장만을 선택해서 문서전체에 대한 핵심 단어집합을 구축한다. 이 핵심 단어집합과 시간별 문장집합에서 찾은 핵심문장과 부가문장에 따른 가중치 및 [수식 2]의 값을 이용해서, 각 문서의 대표 문장들에서 요약결과에 포함될 문장을 선택한다.

주어진 요약 대상문서는 작성된 날짜에 따라서 시간별 문서집합에서 문서개수의 차이가 있기 때문에, 많은 문서를 포함하고 있는 문서집합에서 문장이 선택될 가능성을 부여 하였고, 최근에 작성된 문서가 요약문에 새로운 정보를 부여할 가능성이 많으므로 시간에 따른 가중치도 부여하였다. 또한, 3.3 의 핵심문장 선별처럼 각각의 문장에 존재하는 핵심 단어의 문장성분에 따른 가중치도 부여하였다.

$$M(S_i) = w_1 T_i + w_2 \frac{|D|}{N} + w_3 I(S_i) + w_4 N(S_i) + w_5 \sum_{t_k \in S_i \cap M} t_k \cdot Mf_k + w_6 \sum_{t_k \in S_i \cap M} W_g(t_k, S_i)$$

T_i : 문장 Si 의 시간 가중치

$|D|$: 시간별 문장집합에 속하는 문서 수

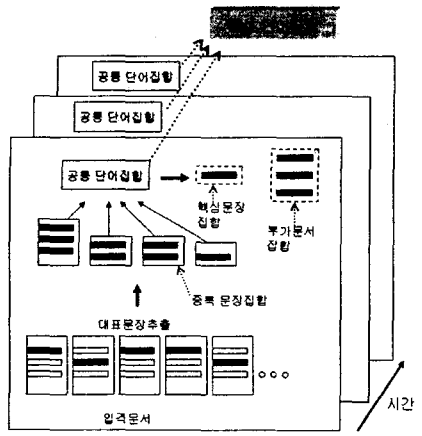
N : 전체 문서 수, $I(S_i)$: 핵심문장 가중치

$N(S_i)$: 핵심문장 선별에 사용된 점수

M : 핵심 단어집합, Mf_k : 핵심 단어의 빈도수

[수식 3] 중요 문장 추출을 위한 문장점수 계산식

[수식 3]을 통해서 문장들에 점수를 부여하고 요약 길이에 따라서 선택할 문장의 수를 결정하고, 문서의 시간 정보를 이용하여 재배열한다.



[그림 1] 시간 자질을 이용한 다중 문서요약

4. 실험 및 결과

제안한 방법을 검증하기 위해서, 실험 시스템에서 선택한 결과를 사람이 만든 요약결과 및 Base-line 시스템과 비교하였다.

실험 시스템은 문장단위로 중요내용을 추출하기 때문에, 사람이 만든 정답은 전체문서에서 요약문 작성에 필요한 내용을 포함하는 문장을 선택해서 문서번호와 문장번호로 구성하였다. 실험에 사용한 문서집합은 1999 년 한국일보 신문기사에서 ' 태풍올가' , ' 씨랜드 참사' 등 5 개의 주제별로 각각 10-18 개의 신문기사로 구성하였다. 정답문장은 대학원생 4 명이 각각의 신문기사를 읽고, 9 문장을 추출 하였고, 4 명이 선택한 문장의 빈도수에 따라서 문서개수가 13 개 이하인 경우 6 문장, 14 개 이상인 경우 9 문장의 최종 정답문장을 선택하였다. 사람이 추출한 중요문장은 신문기사의 특성을 반영하듯, 주로 첫번째 문장을 선택하는 경향이 있었다.

문서 집합	전체 문서수	정답 문장수	첫번째 문장수	그외 문장수	정답에서 첫번째 문장의비율
1	19	9	8	1	89%
2	12	6	5	1	83%
3	12	6	5	1	83%
4	11	6	3	3	50%
5	13	6	6	0	100%

[표 2] 사람이 추출한 중요 문장에서 첫번째 문장의 비율

[표 2]를 근거로 Base-line 시스템은 Lead method 을 적용해서 모든 문서에서 첫 문장만을 추출, 시간에 따라서 정렬한 후, 가장 오래된 문장부터 추출하는 방법(Base1)과, 가장 최근 문장부터 추출하는 방법 (Base2)을 적용하였다.

	Base 1	Base 2	제안한 방법
Precision	34.6%	39.8%	50%

[표 3] 제안한 방법과 Base-line 시스템의 성능 비교

실험 결과를 통해서 중요 문장은 최근 문서에 존재할 가능성이 많음을 나타냈고, 제안한 방법은 50%의 정확률을 보였다.

5. 결론 및 향후 연구

다중 문서요약의 중요내용 추출시, 시간 자질을 이용한 문서의 분류와 이를 바탕으로, 계층적 클러스터링을 통한 중요 내용의 추출은 사람이 만든 정답에 50% 의 정확률을 보였다. 그러나 Lead-method를 적용한 방법에 비해서 월등한 성능향상을 나타내지 못했다.

향후 연구로는 단일 문서요약에서 시간자질의 특성 [10]분석 및 시간자질을 이용하지 않은 방법과 비교 실험을 통해서, 다중 문서요약에서 시간 자질의 특성 분석이 필요하다. 또한 현재 시간자질로 분류한 문서집합에서 핵심과 부가의 관계만을 설정했지만, 보다 체계적인 다중 문서요약을 위해서는 문서 집합에서 문서간의 관계 혹은 문장간의 관계를 설정하고 찾는 방법[6]의 연구가 계속 되어야 할 것이다.

감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았습니다.

6. 참고 문헌

[1] J. Goldstein, V. Mittal, J. Carbonell, M. Kantrowitz, Multi-Document Summarization By Sentence Extraction, ANLP/NAACL Workshop, 2000

[2] D. R. Radev, H. Jing, M. Budzikowska, Centroid-Based Summarization Of Multiple Documents, ANLP/NAACL Workshop, 2000

[3] C. Y. Lin, E. Hovy, Form Single to Multi-document Summarization: A Prototype System and its Evaluation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), p.457-464, 2002

[4] K. McKewon, J.L. Klavans, V. Hatzivassiloglou, R. Bazilay, E. Eskin, Towards Multidocument Summarization by Reformulation, AAAI, 1999

[5] B. Schiffman, A. Nenkova, K. McKeown, Experiments in Multidocument Summarization, HLT, 2002

[6] D. R. Radev, A Common Theory of Information Fusion from Multiple Text Sources Step One : Cross-Document Structure, ACL SIGDIAL Workshop, 2000

[7] D. R. Radev, S. Blair-Goldensohn, Z. Zhang, Experiments in Single and Multi-Document Summarization Using MEAD, DUC, 2001

[8] D. R. Radev, S. J. Otterbacher, H. Qi, D. Tam, MEAD ReDUCs: Michigan at DUC2003, DUC, 2003

[9] 임정민, 강인수, 배재학, 이종혁, 다중 문서요약에 문장의 중복도 측정방법 개선, 한국정보과학회 가을 학술논문발표집, 2003

[10] J. Allan, R. Gupta, V. Khandelwal, Temporal Summaries of News Topics, SIGIR 2001