단어공기정보를 이용한 자동화 문서 요약

류동원[0], 이종혁

포항공과대학교 컴퓨터공학과

{dwryu, jhlee}@kle.postech.ac.kr

Word Co-occurrence based Automatic Text Summarization

Dong-Won Ryu[0] Jong-Hyeok Lee

Dept. of Computer Engineering, Pohang University of Science and Technology

요약

본 연구는 문서를 구성하고 있는 각 단락들(paragraphs)간의 단어공기정보(word co-occurrence)를 이용해 이들간의 관계를 바탕으로 중요단락을 추출하여 문서의 요약을 한다. 이같은 접근법 문서요약의 성능은 단락들간의 정보추출방법과 추출된 정보에 의한 중요단락 선택방법에 크게 좌우된다. 본 논문에서는 중요단락에 대한 선택을 할 때 기존의 방법론에서 발생하는 요약문의 가독성(readability)을 높이면서 또한 성능의 향상도 꾀할 수 있는 방법론을 제시한다.

1. 서론

현대를 살고 있는 우리는 스스로가 처리할 수 있는 정보처리 능력을 벗어나는 엄청난 양의 정보속에서 살고 있다. 그러므로 이런 환경에서는 한정된 시간에 얼마만큼 유익하고 또 필요한 정보를 얻을 수 있느냐 하는 것은 중요한 문제이다. 이러한 문제의식에서 자동화 문서요약에 대한 동기를 부여할 수 있다.

문서요약이라 함은 단순히 원문의 내용만 줄이는 것이 아니라, 이와 동시에 원문이 나타내고 있는 정보를 손실없이 포함할 수 있어야 한다. 문서요약에서는 요약문을 생성하는 방식에 따라 추출(extract)과 요약(abstract)으로 나눌 수 있다. 본 논문에서는 단락레벨에서 정보검색 방법론을 도입한 단어공기정보(word co-occurrence)접근법을 통해 요약문을 추출하는 방법을 택했다.

이러한 방법론은 그 대상을 하나의 문서로 보고 이 문서를 구성하는 여러 개의 단락(paragraph)들 간의 관계나 중요도를 정보검색 방법론을 적용하여 구분해낸다. 이를 토대로 요약문을 생성하기 위한 중요단락을 추출한다.

2. 관련연구

문서요약은 그 접근법과 방법론에 따라 다양하고 많은 시도와 연구가 이루어졌고 또 현재 계속 진행되어 오고 있는 분야이다.

요약을 한다는 것은 우선 문장에 대한 이해, 중요한 주제의 선별, 요약문 생성의 세단계로 크게 볼 수 있다[1]. 특히 요약문의 생성방법에 따라 앞서 언급한 요약과 추출로 그 분류를 나눌 수 있는데, 최근 들어서는 추출쪽에 관련된 연구가 많이 이루어 지고 있다. 본 논문에서 언급하고 있는 방법론 또한 이러한 방법론중의 하나이다[2].

여기서 중요한 것이 바로 어떻게 중요한 단락에 대한 선택을 할 것인가 인데, 단순히 단어의 빈도수 같은 통계적 정보를 이용하거나[3], 또는 말뭉치(corpus)를 통해 얻은 핵심구(cue phrase), 위치(position), 제목(title) 등의 자질(feature) 정보를 이용하는 방법등이 있다[4]. 그러나, 이러한 코퍼스를 사용하지 않고 공기정보를 이용하는 방법도 G. Salton 에 의해 제시되었다[1][5][6]. 이는 단락간의 관계를 통계적 정보를 이용하여 구한 후 각 단락들 간의 유사도(similarity) 관계를 하나의 path 로 보고 이를 따라가면서 중요한 문장을 선택하는 방식이다. Salton 은 이러한 방법으로는 bush, depth-first, segmented bush path 이 세가지를 제시하였다[1][8].

3. 문서요약(Text Summarization)

3.1 분석

문서를 구성하고 있는 각 단락들을 그 기본 단위로 하여 각 단락들간의 단어 빈도수를 이용하여 문서에 있는 모든 단락들 쌍에 대해 각각의 유사도를 구한다.
문서의 분석을 위해 사용하고 있는 벡터 공간 모델의 각 단락들간 유사도 계산에서 주로 코사인 유사도 계산(cosine similarity)과 내적 유사도 계산(inner similarity)을 주로 사용하고 있다. 본 논문에서는 코사인 유사도를 사용하며, 용어 가중치(term weighting) 방식은 tf * idf 를 이용한다.

3.2 단락간의 병합(merge)을 통한 traverse 방식

Bush path, depth-first path 등과 같은 기존의 traverse 방법들은 모두 긴 단락 수를 가지는 문서를 그 대상으로 삼고 있었기 때문에, 짧은 단락 수, 즉 대략적으로 20 개 미만의 단락을 갖는 문서에 대해서는 그 성능에 대한 고찰이나 실험이 없었다. 그러므로 단락수가 적은 문서에 대해서는 7~80 개 이상의 단락으로 이루어진 문서에서 나타나는 결과 만큼의 성능을 보장할 수 없다. 이러한 원인은 단락이 짧거나 혹은 개수가 작으면 주제와는 무관한 단어들의 중복만으로도 유사도 값의 상승을 가져와 그만큼 단락간의 유사도에 신뢰성을 떨어뜨리게 되는 결과를 가져오기 때문이다.
따라서, 본 논문에서는 기존의 방법을 사용하면서 동시에 짧은 문서에 대한 요약에서도 긴 문서에 적용하였던 기존 traverse 방식의 결과와 같거나 좀더 나은 성능을 나타낼 수 있도록 하기 위해 몇 가지 개선된 방법을 제안한다. 그리고, 추출에 의해 요약을 수행하는 시스템에 있어 요약한 문장의 정확성 뿐만 아니라 이러한 문장의 가독성도 고려 대상에서 빼놓을 수 없는 부분이기도 하다. 기존 방법론들 중 bush 와 depth-first 방법들은 서로 상호 보완적인 장단점을 가지고 있으므로, 본 논문에서 제시할 방법론을 통해서 성능의 저하 없이 어느 정도 이 두 가지 방법론의 장단점 차이를 좀더 극복해 보고 장점을 더 부각 시키는 방법론으로서 본 연구의 의미를 부여하고자 한다.

1) adaptive bush path

기존의 bush path 는 문서에서 다른 단락들과의 상호 연관관계 횟수가 높은 순으로 단락의 순위를 결정하여 추출하는 방법으로 이러한 방법은 추출된 단락들간의 내용 연결성이 결여된다[1]. 즉, 이로인한 문서의 가독성이 문제가 된다. 본 방법은 유사성의 신뢰성 뿐만 아니라 기존bush path 의 단점의 하나인 추출된 단락간의 내용연결성의 부족 문제를 해결하기 위해 d)단계에서는 병합된 단락들과 유사도가 높은 문장들로 나머지 요약문을 선택하도록 했다.

a) 기존의 bush path 를 통해 원하는 요약문 길이의 50% 정도의 단락을 얻는다.
b) Bush path 를 통해 얻어진 단락들을 하나의 단락으로 병합한다.
c) 얻어진 단락을 새로운 단락으로 하여 이와 나머지 단락들과의 새로운 유사도를 구한다.
d) 여기서 얻어진 유사도 값을 기준으로 나머지 요약문의 크기(length)를 채운다.

2) adaptive depth-first path

기존의 depth-first path 방식에서는 중심 단락을 선택한 후 다음 단락을 선택할 때 중심단락과의 응집력(cohesion)정도와 bushiness 만을 고려한다. 이러한 방법은 문서의 내용연결성은 좋을지 모르나 문서에 여러 주제를 담고 있을 경우 모든 주제를 요약문에 포함시킬 수가 없다[1]. 그러나 adaptive depth-first path 에서는 앞서와 같이 병합을 하는 방식으로 traverse 를 수행하지만, 다른점은 adaptive bush 방식과는 달리 기존 depth-first 방식의 장점인 응집력을 해치지 않기 위해 유사도와 그 단락의 위치를 함께 고려하여 각 단락의 가중치를 구하였다. 또 만약 중심단락의 앞쪽에 있더라도 중요 단락이고 다른 많은 단락들과의 관련도를 가진다면 이러한 단락들도 요약 단락의 후보가 될 수 있도록 하여 depth-first 방식의 단점을 보완했다. 기존에는 이러한 단락들은 요약문의 응집력을 해친다고 하여 그 후보로서 제외되던 것들이었다.

a) 기존의 depth-first path 를 통해 원하는 요약문 길이의 50%정도의 단락을 얻는다.
b) 이를 통해 얻어진 단락들을 하나의 단락으로 병합한다.
c) 얻어진 단락을 새로운 단락으로 하여 이와 나머지 단락들과의 새로운 유사도를 구한다.
d) 새로 얻은 유사도와 여기에 중심단락과의 근접성을 고려하여 각 단락들의 새로운 가중치을 구한다.
e) 이러한 가중치 부여 후 높은 가중치를 가지는 순으로 나머지 단락을 선택하여 원문의 순서대로 나열한다.

d)단계에서는 중심 단락과의 거리가 멀수록 가중치가 낮아 지도록 하고 중심 단락 앞쪽은 가중치 값을 매우 낮게 설정함으로써 기존의 depth-first 방식에서 단락간의 응집력을 고려했던 부분을 구현하였다. 여기서 구하여진 가중치와 유사도의 곱으로써 그 최종 가중치를 계산하였다.

4. 실험

본 실험에서는 기존의 방법론 뿐만 아니라 새롭게 제시한 방법론에 대해 모두 실험을 하였고, 여기다 유사도의 임계값이 실제적으로 어떻게 요약의 성능에 영향을 미치는지에 대한 실험도 행하였다.
실험에 사용된 문장은 과학기술 관련 기사들로서 평균

15 개의 단락으로 이루어져 있으며 총 25 개의 문서를 그 대상으로 했다. 이러한 방법론은 요약하고자 하는 문서의 장르나 영역에는 어느 정도 무관하다는 장점도 있다.

결과의 평가는 사람이 직접 문서에서부터 선택한 요약문과 기계가 자동으로 추출한 요약문 사이의 중복정도(overlapping)로 하였다. 본 실험에서는 정확률 뿐만 재현율까지도 결과에 포함하였다.

또한, 요약문 생성시 원문의 20%와 30% 두 가지의 고정길이의 요약문을 생성 시켜서 이 두개의 서로 다른 요약문에 대해 그 결과를 비교하였다.

<표 1> 20% 길이의 요약문 결과

(정확률)

방법론 / threshold	Bush Path	DF Path	Bush_A	DF_A
0.13	0.27	0.20	0.23	0.28
0.20	0.21	0.20	0.21	0.21
Random	0.15			

<표 2> 30% 길이의 요약문 결과

(정확률)

방법론 / threshold	Bush Path	DF Path	Bush_A	DF_A
0.13	0.35	0.31	0.27	0.34
0.20	0.27	0.25	0.32	0.26
Random	0.26			

여기서 사용한 random 은 문서에서 임의로 단락을 추출해낸 것으로 이는 다른 방법론의 결과에 대한 기준이 될 수 있다. 본 실험에서는 고정길이의 요약문을 생성함으로 재현율은 언급하지 않는다. 본 실험에서는 유사도의 임계치를 넘는 요약 후보단락수가 부족하여 원하는 길이의 요약문을 채우지 못할경우에는 원하는 길이의 요약문이 생성될 때 까지 계속 임계치의 feedback 을 주어 고정길이 요약문을 생성하도록 하고있다. 이는 오히려 정확률의 저하를 유발시키는 요인으로 보인다.

요약문의 가독성에 대한 평가를 위해 본 실험에서는 서로 다른 방식으로 추출한 요약문을 제시하고 각 개인에게 이들의 가독성이나 문맥의 흐름에 대한 채점을 하도록 했다. 각 문서당 최고 3 점에서 최하 0 점 사이로 범위를 한정하고, 총 9 명에 대해 본 실험을 실시했다. 실험에 사용한 요약문은 임계치가 0.13, 30% 길이의 bush path 와 adaptive-bush path 에 의해 요약된 두 요약문서을 그 대상으로 했다.

Bush path 방식 요약문의 경우 평균 1.86 의 점수를 기록했고, 본 논문에서 제안한 adaptive-bush path 의 경우는 2.02 의 수치가 나왔다. 이를 각 문서단위로 보면, 실험에 사용된 총 25 개의 요약문 중 18 개의 문서에서 adaptive-bush path 방법이 bush path 보다 높은 점수를 나타냈다.

5. 결론 및 향후 계획

본 논문에서 제시한 새로운 방법론을 적용했을 경우 앞의 표-1, 표-2 의 결과에 나타나듯이 Depth-first 방법의 경우 성능의 향상이 이루어졌다. 그러나 bush 방법의 경우는 임계치 0.2 일 경우를 제외하곤 만족스럽지 못한 결과를 나타내었다. 앞서도 언급 했듯이 adaptive bush path 방법은 주로 bush path 의 성능의 향상보다는 좀더 가독성을 높이고자 하는 것이 주 목적이 있다. 앞서의 가독성에 대한 평가를 실시하였으나, 실제로 문서의 가독성을 수치적으로 평가한다는 것이 상당히 문제가 있고 또 그 평가방법도 모호 하지만, 이 방법을 통한 요약문이 bush path 방법론의 요약보다는 실험적인 데이터를 통해서도 나타났듯이 가독성이 개선되었다고 할 수 있겠다.

게다가, 앞서의 문제는 1 차적으로 병합을 하기 위해 선택한 단락들의 비율을 정할 때 본 실험에서 사용한 50%가 적절하지 못한 이유로 여겨진다. 이 값의 적절한 조절을 통해서도 bush path 방법 이상의 성능 향상을 기대할 수 있을 것이다.

앞으로 이러한 연구를 바탕으로 요약문을 통한 Full Text Retrieval 을 하고자 한다.

참고문헌

[1] G. Salton, A. Singhal, C. Buckley, M. Mitra "Automatic Text Structuring and Summarization", Information Processing & Management, 1997

[2] Mani and Maybury "Advances in Automatic Text Summarization", MIT Press, 1999

[3] H.P. Luhn "The Automatic Creation of Literature Abstracts", IBM Journal of Research & Development, 1958

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

[5] G. Salton, A. Singhal, C. Buckley, M. Mitra "Automatic Text Decomposition Using Text Segments and Text Themes", '96 ACM Conference on Hypertext, 1996

[6] G. Salton, A. Singhal "Automatic Theme Generation and the Analysis of Text Structure", TR, 1994

[7] R. Brandow, K.Mitze, L F. Rau "Automatic Condensation of Electronic Publications by Sentence Selection", Information Processing & Management, 1995

[8] G. Salton, J. Allan, "Selective Text Utilization and Text Traversal", Hypertext '93 Proceedings, 1993