TextRank 알고리즘을 이용한 한국어 중요 문장 추출

A Korean Important Sentence Extraction using TextRank Algorithms

저자 (Authors) 홍진표, 차정원
Jeen-pyo Hong, Jeong-won Cha

출처 (Source) 한국정보과학회 학술발표논문집 36(1C), 2009.6, 311-314(4 pages)

발행처 (Publisher) 한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE01219123

APA Style 홍진표, 차정원 (2009). TextRank 알고리즘을 이용한 한국어 중요 문장 추출. 한국정보과학회 학술발표논문집, 36(1C), 311-314

TextRank 알고리즘을 이용한 한국어 중요 문장 추출

홍진표 ○ 차정원

창원대학교 컴퓨터공학과

virnmer@cwnu.ac.kr, jcha@cwnu.ac.kr

A Korean Important Sentence Extraction using TextRank Algorithms

Jeen-pyo Hong ○ Jeong-won Cha

Dept. of Computer Engineering, Changwon National University

요 약

본 논문에서는 TextRank 알고리즘을 이용한 한국어 중요 문장 추출에 대해 소개한다. TextRank는 PageRank 알고리즘을 단어와 문장에 적합하도록 구성한 알고리즘을 말한다. 이 시스템은 기본적인 자연언어 처리 시스템이 없더라도 비지도 학습을 통해 중요 단어 혹은 문장을 추출할 수 있다. 여기서는 기존의 영어와 브라질어에 적용되었던 방법과 같은 방법으로 한국어에 TextRank 알고리즘을 적용을 했다. 그리고 두 언어와 달리 한국어의 경우 어절 내 축약과 같은 언어적 특성으로 인해 어근 추출기를 사용할 수가 없다. 그래서 본 논문에서는 어근 추출기 대신 한국어 품사 태거를 이용하여 어절 내 조사, 어미, 접사들에 정보를 제거하여 어근을 추출하도록 하였다. "조선일보"의 신문 기사 일부분을 모아 그 문서에 문장 중요도 정보를 태깅하여 만든 55문서에 대해 테스트를 해 본 결과, ROUGE 점수가 최대 0.6885로 다른 언어에 비해 월등히 좋은 성능을 기록했다.

1. 서론

최근 급속한 인터넷의 발전으로 텍스트 정보의 양은 기하급수적으로 증가했다. 이러한 정보의 증가로 인해 사용자는 원하는 정보를 찾는데 더욱 많은 시간을 투자하여 해당 정보를 일일히 확인하여 찾아야 하는 불편함이 가중되고 있다. 이러한 문제를 해결하기 위해 정보 검색, 텍스트 마이닝을 이용한 문서 요약에 대한 연구가 대두되고 있다.

문서 요약은 특정 문서에 대해 그 문서에서 말하고자 하는 내용을 보다 적은 양의 정보로 내용을 전달하는 데 그 목적이 있다[1].

이와 같은 문서 요약에는 생성요약과 추출요약으로 나눌 수 있다[2]. 생성요약은 원 문서로부터 중요한 단어 및 어휘를 추출하여 자연어 처리 기법을 이용해 요약문을 새로운 문장으로 만드는 것이다. 반면 추출요약은 원 문서로부터 중요하다고 생각되는 문장만을 선별하여 요약문으로 제공한다. 전자의 경우가 보다 적은 양으로 문서의 내용을 포괄하여 제공할 수 있으며 사람들이 생각하는 요약과 유사하다. 그러나 자연어 처리의 기술적 한계로 인해 상대적으로 접근이 쉬운 추출요약 쪽이 주류를 이루고 있다.

추출 요약 중 최근에 Google의 PageRank 알고리즘의 개념을 텍스트 문서에 대해 적용한 TextRank 알고리즘이 등장했다. 이 알고리즘은 기본 자연언어 처리 시스템 없이 비지도 학습을 통해 문서 요약이 가능하다[11].

본 논문에서는 이러한 언어적 제약을 덜 받는 TextRank 알고리즘을 이용해 한국어 문서에 대해 문장 추출 시스템을 구현했다. 그리고 [11]과 [12]에서 구현된 영어, 브라질어에 대해 적용한 시스템과 한국어에 적용된 시스템을 비교했다.

본 논문의 구성은 다음과 같다. 2장에서는 기존에 제안된 여러 중요 문장 추출 시스템에 대해 살펴보고 [11]에서 제안된 TextRank에 대한 기본 개념을 정리한다. 그리고 3장에서 본 논문에서 제안한 중요 문장 추출 시스템에 대해 자세히 설명한다. 그리고 4장에서 제안한 시스템에 대해 실험을 하고 각 언어권별로의 시스템과의 비교 후 그 결과를 정리한다. 마지막 5장에서는 본 논문의 결론 및 앞으로의 연구 방향에 대해 설명하고 논문을 마치도록 한다.

2. 이전 연구

2.1 중요 문장 추출

중요 문장 추출은 단어의 출현 빈도를 이용하는 방법, 문장 간 유사도를 이용하는 방법과 그래프를 이용하는 방법으로 나눌 수 있다.

단어의 출현 빈도를 이용하는 방법은 단순히 명사와 같이 문장에서 영향력을 가장 많이 행사하는 단어에

대해 단순히 출현 빈도만을 계산해 문서를 대표하는 단어 집합을 설정해 문장의 중요도를 정하는 방법[3]이 있다.

다음으로는 문장 간 유사도를 이용하는 방법이다. 이 방법은 문장 간의 단어나 문장의 위치 등 자질에 대해 코사인 유사도 등을 이용해 문장 간 유사도를 계산하고 이를 문장 중요도 계산식에 가중치로 반영하는 방법[2,4,5]이다.

그래프를 이용하는 방법은 기본적으로 문장 간의 연결 그래프를 만들어 요약에 이용하는 방법이다. 이러한 방법 중 첫번째로 문장 간의 공통 단어가 있을 경우 문장의 연결 그래프를 만든 다음 그래프의 이음새인 관절점이 되는 그래프의 노드인 문장을 찾아 그래프를 몇 개의 그룹으로 분할한 후 그룹별로 중요 문장을 추출하는 방법으로 이는 기존의 유사도에 기반한 방법에서 자질에 대해 민감하게 반영하는 문제점을 해결하기 위해 제안되었다[6,7,8]. 다른 방법으로는 Google의 PageRank 알고리즘의 개념[9]을 문장에 도입하여 중요 문장을 추출하고자 했다[10,11,12].

2.2 TextRank

Mihalcea와 Tarau는 TextRank를 이용해 중요 키워드 추출과 중요 문장 추출 방법을 제안했다[11].

PageRank는 기본적으로 중요한 사이트는 다른 많은 사이트로부터 링크를 받는다는 점에 착안하여 문서 A에서 가지고 있는 B의 링크가 B 사이트를 추천하는 한 표로 해석하여 이를 기준으로 중요도를 평가한다.

TextRank는 PageRank에서의 사이트를 단어 혹은 문장으로 생각하고 식 (1)과 같이 문장의 중요도를 계산한다.

$$TR(V_i) = (1-d) + d\sum_{j=0}^{N-1}\frac{w_{ij}}{\sum_{k=0}^{N-1} w_{jk}} \times TR(V_j) \quad \cdots (1)$$

식(1)에서 $TR(V_i)$ 는 단어 혹은 문장 i 에 대한 TextRank 값을 의미하며 w_{ij}는 i와 j 사이의 가중치를 의미한다. 값 d는 PageRank에서 해당 V_i에 대해서 사용자가 해당 웹페이지를 임의로 선택할 가능성을 나타내는 확률로 TextRank에서는 PageRank에서 제안한 값 0.85를 그대로 이용한다.

중요 키워드 혹은 문장 추출을 하기 위해, 각 단어 혹은 문장 V_i에 대해 가중치인 w_{ij}와 w_{ji}를 계산하고 각 V_i에 대해 그래프를 구성한다. 그 후, $TR(V_i)$를 임의의 값으로 초기화 한 후 이들 $TR(V_i)$의 값이 수렴하거나 임계값 만큼 식 (1)를 각 V_i 에 대해 반복적으로 적용한다. 그 결과 문장 추출의 경우 최종적으로 계산된 각 V_i에 대해 $TR(V_i)$의 값이 가장 큰 n개 문장만 결과로 보여준다. 중요 키워드 추출의 경우, 각 V_i 에 대해 $TR(V_i)$ 의 값이 가장 큰 n개의 단어들의 링크 정보와 전체 문장과 비교하여 문장에서 키어구로 등장하는 것들에 대해서만 키워드를 추출해낸다.

이 방법은 특별한 기본 자연언어 처리 시스템이 없더라도 비지도 학습을 통해 중요 단어 및 문장을 추출할 수 있다[11]. 또한, 시소러스와 같은 언어적 정보를 추가할 경우 문장 추출에 있어 보다 나은 성능을 얻을 수 있다[12].

3. 텍스트랭크를 이용한 한국어 문장 추출

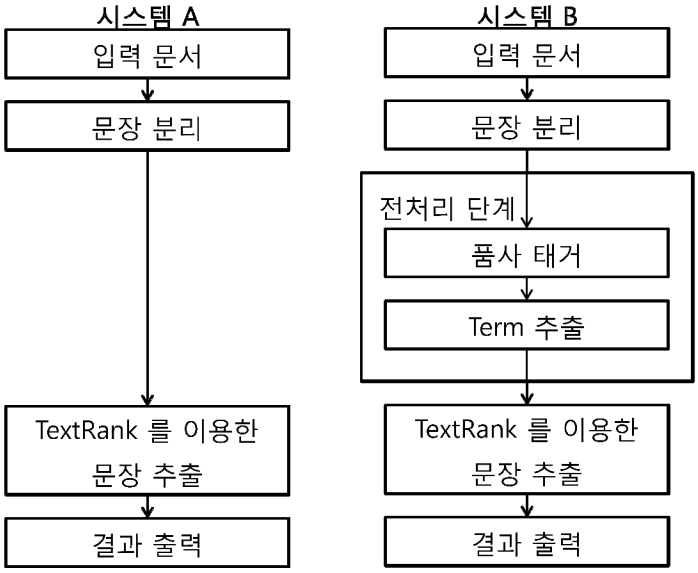

그림 1. 본 논문의 시스템의 구성도

본 논문에서 구성한 시스템들의 형태는 [그림 1]과 같다. 시스템 A는 단순히 어절 전체를 하나의 어근으로 간주한다. 반대로 시스템 B는 전처리 단계를 거쳐 한 어절 내에 조사, 어미, 접사들에 대한 정보를 제거한 나머지를 어근으로 추출한다.

[11]과 [12]와 달리 한국어에서는 한국어의 특성상 어절 내의 포함된 형태 정보들의 분리가 영어과 브라질어와 달리 어근 추출기로 불가능하기 때문에 품사 태거를 사용했다. 예를 들어, “박종만” 이라는 단어가 나타날 경우, 어근 추출기를 사용하게 되면 이 단어는 “박종” 이라는 형태가 되어 원래 의도했던 “박종만” 이라는 형태를 띄지 않게 된다. 또한 “난”과 같은 어절 사이에서의 축약 문제도 한국어에서 어근 추출기를 사용할 수 없게 하는 이유 중 하나이다.

3.1 중요 문장 추출

시스템 A의 문장 분리 과정과 시스템 B의 전처리 과정을 마치게 되면 TextRank를 이용한 중요 문장 추출 과정을 거친다.

이 과정은 모든 문장들을 엮어 그래프로 구성하고

이들에 대한 문장 간의 가중치 w_{ij}와 w_{ji}를 계산한다. 문장 간의 TextRank의 경우 방향성이 없기 때문에 w_{ij}와 w_{ji}가 같은 값을 가지게 된다. 그리고 이들 가중치는 문장 i와 문장 j 사이에 공통적으로 존재하는 단어들을 이용하여 유사도 식을 만들었으며 그 식은 식(2)와 같다.

$$\mathrm{Sim}(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \wedge w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)} \quad \cdots (2)$$

이 식에서 $|S_i|$는 문장의 길이를 말한다. 각 문장의 길이에 대해 log를 취해 나눠 준 이유는 문장의 길이가 길어질수록 유사도 식의 값이 커지는데 이를 막기 위해서 정규화를 해준 것이다.

최종적으로 실험에서 사용한 식은 식(3)과 같이 적을 수 있다.

$$\mathrm{TR}(V_i) = (1-d) + d \sum_{j=0}^{N-1} \frac{\mathrm{Sim}(S_i, S_j)}{\sum_{k=0}^{N-1} \mathrm{Sim}(S_j, S_k)} \times \mathrm{TR}(V_j) \quad \cdots (3)$$

5. 실험

본 논문에서는 실험에 사용된 말뭉치는 2006년 1월부터 2006년 4월까지의 "조선일보" 정치,경제 부분의 신문기사를 이용하였다. 이를 이용해 각 문장들을 분리하여 문장 중요도를 수작업으로 태깅하여 55문서를 구축했다. 구축한 문서의 정보는 [표 1]과 같으며, 해당 수치는 소수점 첫째 자리에서 올림하여 구했다.

표 1. 실험에 사용한 말뭉치 정보

평균 문장의 수	575/55 = 11
평균 어절의 수	9,204/55 = 168
평균 어근의 수	10,771/55 = 196

문장 추출 시스템의 성능평가는 ROUGE 평가 프로그램을 이용하여 평가했다[13]. 이 평가 방법은 N-gram 통계적인 방법에 기반을 둔 방법으로 사람이 평가하는 방법을 서로 높게 관련 시켜 찾아준다.

평가 방법은 [11]과 [12]의 실험 결과 자료를 비교하기 위해 동일한 방법으로 추출한 문장 요약의 결과로부터 100개의 어근을 뽑아 이를 ROUGE에서 Ngram(1,1)로 설정하고 95% 신뢰 수준을 가지도록 설정 했다.

표 2. 성능 평가 결과 ROUGE 점수 – Ngram(1,1)

	Basic	Stemmed
한국어	0.6885	0.6514
영어[11]	0.4708	0.4904
브라질어[12]	0.4963	0.5426

[표 2]는 제안한 시스템에 대한 실험 결과와 기존 영어[11]과 브라질어[12]의 시스템의 결과이다.

[표 2]에서 한국어의 Basic 시스템은 3장에서 언급한 시스템 A를 적용했으며, Stemmed 시스템은 시스템 B를 적용했다.

결과를 보면, 다른 언어의 결과에 비해 한국어에서 ROUGE 점수가 월등히 높게 보이고 있다. 이는 다른 언어들에 비해 실험 말뭉치의 수가 적어 섣불리 판단할 수는 없지만 TextRank 알고리즘을 이용한 문장 요약이 한국어가 다른 언어들에 비해 중요 문장 추출에 매우 효과적이라 말할 수 있다. 다만, 중요 문장 추출에서 조사, 어미, 접사를 제거할 경우 성능이 증가하는 것이 일반적인데 본 논문의 실험에서는 오히려 성능이 다른 시스템의 증가폭 만큼의 수치가 감소한 것으로 볼 수 있다. 이는 제외한 어근 외에 성능에 영향을 줄 수 있는 어근이 존재하거나 그 외의 가능성이 있다고 보여진다.

6. 결론 및 향후 연구 방향

본 논문에서는 TextRank를 한국어에 적용을 해보았다. 적용 결과, 기본적인 TextRank 알고리즘을 적용하여 중요 문장을 추출했는데도 불구하고 월등한 성능을 보였다.

[12]에서 시소러스를 적용한 경우, TextRank에 어근 추출기를 적용한 것에 비해 약 0.02 가량 성능을 향상을 보이는 점을 감안한다면 지금의 성능 이상으로 한국어에 대해 성능 향상을 기대해볼 수 있을 것으로 생각된다. 다만, 그전에 있어 일반적으로 조사, 어미, 접사를 제거할 경우 성능 향상을 보여야 하는데 이 점에 대해 앞으로 보다 깊이 있는 실험을 통해 원인이 분석되어야 할 것이다.

그리고 실험에 사용된 말뭉치가 특정 한 분야에만 적용을 했다. 앞으로 말뭉치를 늘이는 과정에 있어 이들 분야 외 다른 분야에 대해서도 실험을 하여 보다 일반화 시킬 필요가 있다.

참고 문헌

[1] Inderjeet Mani, Automatic Summarization, Kohn Benjamins Publishing Co., 2001.

[2] Ohm Sornil, Kornnika Gree-ut, "An Automatic Text Summarization Approach using Context-Based and Graph-Based Chacteristics", IEEE Conference on Cybernetics and Intelligent Systems, pp.1-6, 2006.

[3] K.McKeown, J.Robin, and K.Kukich, "Generating

Concise Natural Language Summaries", Advances in Automatic Text Summarization, MIT press, pp.233-264, 1999.
[4] Daniel Mallett, James Elding, Mario A. Nascimento, "Information-Content Based Sentence Extraction for Text Summarization", IEEE International Conference on Onformation Technology: Coding and Computing, Vol.2, pp.214-218, 2004.
[5] Takaharu Takeda, Atsuhiro Takasu, "UpdateNews: A News Clustering and Summarization System Using Efficient Text Processing", International Conference on Digital Libraries, pp.438-439, 2007.
[6] Il joo Lee, Minkoo Kim, "Document Summarization Based on Sentence Clustering Using Graph Division", Journal of Korea Information Processing Society, Vol.13-B, No.2, pp.149-154, 2006.
[7] Philipp Cimiano, Ontology Learning and Population from Text, Springer, 2006.
[8] 송원문, 김영진, 김은주, 김명원, "동적 연결 그래프를 이용한 자동 문서 요약 시스템", 정보과학회논문지: 소프트웨어 및 응용, 제 36권 1호, pp.62-69, 2009.
[9] S.Brin and L.Page, "The anatomy of a large-scale hypertextual Web search engine", Computer Networks and ISDN Systems, vol.30, no.1-7, pp.107-117, 1998.
[10] G.Erkan, and D.R. Radev, "LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research, vol.22, no.2004, pp.457-479, 2004.
[11] Rada Mihalcea, Paul Tarau, "TextRank: Bringing Order into Texts", Proceedings of the European Conference on Artificial Intelligence (ECAI 2004), Valencia, Spain, August 2004.
[12] Daniel S.Leite, Lucia H.M.Rino, Thiago A.S. Pardo, Maria das Graças V.Nunes, "Extractive Automatic Summarization: Does more linguistic knowledge make a difference?", In Proceedings of Human Language Technology Conference (HLT-NAACL 2003), Rochester, NY, USA, p.17-24, April, 2007.
[13] C.Y.Lin and E.H.Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics", In Proceedings of Human Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada, May, 2004.