공기정보를 이용한 한국어 요약 시스템의 성능개선

박호진[U] 김준홍 김재훈
한국해양대학교 컴퓨터공학과
한국과학기술원, 첨단정보기술연구센터
{hanwool, rainmk}@nlplab.kmaritime.ac.kr, jhoon@cs.unknown.ac.kr

Performance Improvement of Korean Indicative Summarizer Using Collocation

Ho-Jin Park[U] Joon-Hong Kim Jae-Hoon Kim Kim
Dept. of Computer Engineering, Korea Maritime University
and
AITrc, KAIST

요 약

본 논문은 공기정보를 이용하여 한국어 추출요약 시스템의 성능을 개선한다. 여기서 공기정보는 복합명사와 구문관계를 말하며, 복합명사는 인접한 명사들 사이의 공기관계이고, 구문관계는 인접한 명사와 동사 사이의 공기관계를 말한다. 본 논문에서 공기관계는 t test를 이용하였다.

공기정보를 이용한 시스템은 기존의 시스템보다 좋은 성능을 보였으나, 커다란 성능 향상을 가져오지 못했다. 복합명사는 거의 모든 환경에서 좋은 결과를 가져왔으나, 구문관계는 그렇지 못했다. 앞으로 공기정보의 추출방법을 좀더 개선한다면 좀더 좋은 성능을 기대할 수 있을 것이다.

1. 서론

가상공간(cyberspace)이라고 하는 웹은 전세계를 통하여 많은 정보를 쉽게 얻을 수 있는 정보의 보고이다. 가상공간에 존재하는 정보들은 매우 다양하며, 그 양도 매우 빠른 속도로 증가하고 있다. 방대한 정보공간에서 유용한 정보를 찾기 위해 널리 사용되는 도구가 웹 정보검색 엔진이다. 일반적으로 웹 정보검색 엔진들은 너무 많은 정보를 검색해 주기 때문에 유용한 정보를 찾는 것은 그다지 쉬운 일이 아니다. 이와 같은 정보검색 환경에서 유용한 정보를 효과적으로 찾기 위해서는 자동문서요약 기술이 자주 사용된다[1-3].

문서요약은 원문서의 의미를 유지하면서 원문서의 길이나 정보의 복잡도를 줄이는 작업이다[2]. 즉, 문서요약은 정보압축(information compression)이다. 문서요약은 일상적인 생활에서는 널리 사용되고 있는 방법이다. 최근 문서요약은 단순한 하나의 문서의 내용을 요약하는 것이 아니라 여러 문서의 내용은 하나로 요약하기도 하고, 심지어는 문서가 아닌 이미지, 오디오, 비디오와 같은 멀티미디어 정보를 요약하기도 한다[4].

한국어 문서요약에 대한 연구도 매우 활발히 진행되고 있다[2-3][6-7]. 그러나 아직은 성숙되지 않은 것 같다. 또한 대부분의 연구가 통계적 접근 방법을 채택하고 있으며, 여러 다양한 환경에서 평가되어 객관적으로 어떤 시스템이 좋은 성능을 보인다고 말할 수 없는 실정이다.

한국어 문서요약의 방법 중 도합 유사도를 이용한 문서요약[8]에서 있어서 공기관계가 성능에 미치는 영향에 대하여 기술하고 있다. 여기서 공기관계란 t test를 사용한 복합명사와 구문관계를 추출한 정보를 의미한다.

본 논문의 2장에서는 문장 벡터의 생성 방법에 대하여 기술하고, 3장에서는 문장 벡터를 이용한 도합유사도의 계산과 도합유사도를 이용한 요약에 대하여 기술하고, 4장에서는 실험 및 평가, 5장에서는 결론을 기술한다.

2. 문장 벡터의 생성

요약을 하기 위해서 텍스트 문서의 문장들을 벡터의 한점으로 표현한다. 이를 문장 벡터라고하며 이 벡터를 이용하여 요약을 하게된다[8]. 이전 논문에서의 벡터는 명사로만 이루어져 있지만[8] 본 본문에서 벡터라고 하는 것은 명사뿐만 아니라 t test를 이용한 명사와 명사의 관계인 복합명사, 명사와 용언과의 관계인 구문관계로 이루어져 있다.

2.1. 한국어 기준 명사 추출

문장 벡터를 생성하려면 우선 문장에서 명사를 추출[1]해야 한다. 한국어 명사 추출에 관한 많은 연구는 한국어 정보검색 분야에서 많이 수행되었다[10, 11]. 본 논문에서의 한국어 기준 명사 추출은 아래와 같은 단계로 명사를 추출하고 있다.[9].

1. 사전을 이용한 수식언 제거
2. 사전과 어미 집합을 이용한 용언 제거
3. 음절간의 상호정보를 이용한 명사구와 조사의 분리
4. 오타마타를 이용한 수사 제거
5. 사전과 CYK알고리즘을 이용한 복합명사 분리

1) 본 논문에서 사용하는 명사추출기는 [8]에서 사용한 명사추출기의 성능을 개선한 것이다.

2.2. 공기정보 추출

공기정보란 두 개 이상의 단어로 구성되어 있는 표현이다[12]. 본 논문에서는 한국어 문서요약 시스템의 성능을 개선시키기 위하여 복합명사와 구문관계 공기정보를 사용하였다. 이 두 가지의 공기정보를 추출하기 위하여 t test를 사용하였다. 아래의 수식은 t 값을 계산하기 위하여 사용한 수식이다.

$$t = \frac{p(x,y) - p(x)p(y)}{\sqrt{\frac{p(x,y)}{N}}}$$

위 수식에서 x는 명사를, y는 명사 혹은 용언을 의미하고, $p(x)$는 x가 문서에서 나올 확률이다. 또한 $p(x,y)$는 x와 y가 붙어 나올 확률을 의미한다. 여기에서 N은 문서에서의 명사나 용언의 개수를 의미한다. 만약 t 값이 어느 이상의 임계치를 넘을 경우 두 단어 x, y가 믿을만한 공기관계를 가진다[12].

본 논문에서는 이러한 공기관계를 이용하여 복합명사와 구문관계 공기정보를 추출한다. 먼저 복합명사의 공기정보를 추출하는 단계는 아래와 같다.

1. 명사추출기를 이용하여 리스트를 추출한다.
2. 명사에 대한 빈도수를 계산한다.
3. 각각의 명사에 대한 확률을 계산한다.
4. t 값을 계산한다.
5. 임계치 이상을 가지는 x, y에 대하여 문장벡터에 $x\,y$를 추가한다.

구문관계 공기정보를 추출할 때에는 두 가지 휴리스틱을 이용하였다. 첫 번째 휴리스틱은 구문관계를 추출할 때 용언 앞의 명사 한 개만을 고려했다. 두 번째 휴리스틱은 용언으로 추출된 것 중 1음절로 되어있는 용언은 제거하고 2음절 이상의 용언은 앞의 2음절만은 사용하여 구문관계를 추출하였다.

구문관계를 추출하는 단계는 아래와 같다.

1. 명사추출기를 이용하여 용언과 명사를 추출한다.
2. 휴리스틱을 이용한 용언과 명사의 빈도수를 계산한다.
3. 각각에 대한 확률을 계산한다.
4. t 값을 계산한다.
5. 임계치 이상을 가지는 x, y에 대하여 문장벡터에 $x\,y$를 추가한다.

실제 구현할 때에는 명사추출기에서 용언을 미리 표시한 뒤, 위와같은 방법으로 구문관계를 추출하였다.

3. 문서요약

문서요약의 알고리즘은 먼저 문장벡터간의 도합유사도를 계산하고 그 도합유사도가 높은 순으로 문장을 추출한다. 또한 본 논문은 원문서의 문장을 그대로 추출하여 요약문을 생성한다.

도합유사도의 계산은 [8]에서 사용한 방법을 그대로 사용하였다. [8]에서 사용한 방법은 먼저 문서를 문서관계도라고 하는 그래프로 표현한다. 문서관계도에서 노드는 각 문장을 뜻하며, 링크는 의미적으로 관련이 있는 노드들 사이의 관계를 나타낸다. 각 노드의 중요도는 둘러싼 다른 노드들과의 유서도의 합으로 정의한다. 이를 도합유사도라고 한다[8].

요약에 사용될 문장을 추출할 때에는 문장의 중요도(도합유사도)에 따라 상위 순위에 해당하는 적정 수의 문장들을 추출하여 원문서에 나타난 순서대로 정렬시켜 요약 문서를 생성한다[8].

4. 실험 및 평가

본 논문은 기존의 한국어 요약시스템에서 복합명사와 구문관계가 요약에 미치는 영향을 검증하는 논문이므로, 기존 시스템의 요약 말뭉치와 평가 방법으로 실험을 하였다[8]. 평가용 말뭉치는 기존 논문과 같이 3개를 사용하였는데, C1은 논문 전체를 의미하고, C2는 신문기사, 그리고 C3는 C1의 논문에서 초록과 결론 부분에 속하는 말뭉치이다.

본 논문에서는 세 가지의 실험을 하였다. <표1>은 복합명사에 대하여 최적의 임계값을 찾는 실험이고, <표2>는 실험은 구문관계 임계값에 대한 실험, 마지막으로 <표3>은 이전 두 실험의 최적 임계값을 가지고 복합명사와 구문관계를 같이 적용했을 때의 성능이다. 또한 이 실험의 결과는 10% 요약에서의 성능이다.

요약율	신뢰도 (임계값)	정확률			재현율		
		C1	C2	C3	C1	C2	C3
10%	이전 시스템	32.2	42.4	82.3	57.0	17.9	32.5
	85% (1.282)	33.3	43.8	82.9	58.3	17.9	32.4
	90% (1.645)	33.3	44.4	83.9	58.3	18.0	33.0
	95% (1.960)	33.3	43.6	81.9	58.4	17.9	32.1

<표 1> 10% 요약에서의 복합명사

요약율	신뢰도 (임계값)	정확률			재현율		
		C1	C2	C3	C1	C2	C3
10%	이전 시스템	32.2	42.4	82.3	57.0	17.9	32.5
	85% (1.282)	32.6	42.4	83.2	56.7	17.3	32.5
	90% (1.645)	32.7	42.4	83.2	56.7	17.3	32.5
	95% (1.960)	32.8	42.4	83.2	57.0	17.3	32.5

<표 2> 10% 요약에서의 구문관계

요약율	신뢰도 (임계값)	정확률			재현율		
		C1	C2	C3	C1	C2	C3
10%	이전 시스템	32.3	42.4	82.3	57.0	17.9	32.5
	제안된 시스템	33.5	44.4	83.9	58.8	18.0	33.0

<표 3> 10% 요약에서의 성능

위에서 볼 수 있듯이 복합명사에서는 임계값 1.645, 즉 신뢰도가 90%인[12] 것이 제일 좋은 성능을 보였으며, 구문관계에서는 임계값 1.960(신뢰도 95%)이 제일 좋은 것으로 나타났다. <표3>에서는 복합명사와 구문관계를 같이 사용하였는데, 그때의 임계값은 이전 실험에서 좋은 성능을 보인 2개의 값을 각각 사용하였다.

<표4>는 10%와 20%일 때의 성능을 보였다. 20% 요약에서는 복합명사의 임계값이 1.282(신뢰도 85%), 구문관계의 임계값은 1.645(신뢰도 90%)일 때 좋은 성능을 보였다. 10% 요약에서 보다 20% 요약에서 신뢰도가 낮아지며, 성능향상에도 많은 영향을 주지 못했다. 따라서, 많은 요약문을 추출할 때에는 복합명사와 구문관계가 성능에 영향을 미치지 못한다는 것을 알 수 있다.

요약율	신뢰도 (임계값)	정확률			재현율		
		C1	C2	C3	C1	C2	C3
10%	이전 시스템	32.3	42.4	82.3	57.0	17.9	32.5
	제안된 시스템	33.5	44.4	83.9	58.8	18.0	33.0
20%	이전 시스템	20.4	42.0	76.6	71.8	26.2	46.0
	제안된 시스템	20.4	43.5	76.6	72.1	27.1	46.2

<표 4> 10%와 20% 요약에서의 성능

5. 결론

본 논문에서는 기존의 요약 시스템에 복합명사와 구문관계를 사용하여 요약 시스템의 성능을 향상시켰다. 또한 복합명사와 구문관계를 추출 방법은 단순한 모델을 사용함으로서, 구현이 용이하며, 쉽게 사용할 수 있다는 장점이 있다.

제안된 방법은 기존의 방법보다 약 1%~2% 정도의 성능 향상을 보였다. 물론 많은 성능의 향상은 아니지만, 복합명사와 구문관계가 요약 시스템에서 고려해야할 한 대상이라는 것을 알게 되었다.

앞으로 복합명사와 구문관계를 추출하는 방법에 있어서 본 논문에서 사용한 모델과 다른 모델을 같이 사용한다면 요약 시스템의 성능에 좀더 낳은 결과를 기대할 수 있을 것이다.

6. 감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단과 지원을 받았으며, 또한 과학기술부 STEP2000 프로젝트에 의해 지원되고, 전문용어언어공학연구센터에 의해 수행중인 "대용량 국어정보 심층처리 및 품질관리 기술개발" 연구과제의 일환으로 수행되었습니다.

7. 참고 문헌

[1] Cowie, J., Mahesh, K., Nirenburg, S. and Zajac, R. (1998). MINDS multilingual interactive document summarization. in *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*, Spring, pp. 131-132.

[2] Jang. D. and Myaeng, S.-H. (1997) Automatic text summarization systems. Korea Information Science Society Review, vol. 15, no. 10, pp.42-49.

[3] Kang, S.-B. (1997) Implementation of a summarization system using statistical information of Korean documents. Masters thesis, Department of Computer Science, Pusan National University.

[4] Mani, I. and Maybury, M. T. (1999) Advanced in Automatic Text Summarization, The MIT Press.

[5] Sparck Jones, K. (1999). Automatic summarizing: factors and directions, in Mani, I. and Maybury, M. T., editors, Advances in Automatic Text Summarization, pp. 1-12. The MIT Press.

[6] Lee, M.-H., Park, M.-S., Kim, M.-J., and Lee, S.-J. (1999) Sentence extraction using document features and heading. In *Proceedings of KIPS*, vol. 6. no. 2. pp. AI41-AI45.

[7] Ryu, D.-W. and J.-H. Lee. (2000). Word co-occurrence based automatic text summarization,. in *Proceedings of KISS*. vol. 27. no. 1, pp. 345-347.

[8] Kim, J.-H., Kim, J.-H. and Hwang, D.-S., Korean text summarization using an aggregate similarity, In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, Hong Kong, pp. 111-118, 2000.

[9] Kim, J.-H., Kim, J.-H., and Park, H.-J., Korean noun extraction with filtering and segmentation, In *Proceeding of the 1st International Conference on East-Asian Language Processing and Internet Information Technology (EALPIIT2000)*, Northeastern University, Shenyang, China, pp. 107-112, 2000.

[10] Won, H., Park, M. and Lee. G., Integrated indexing method using compound noun segmentation and noun phrase synthesis. *Journal of KISS: Software and Applications*, vol. 27, no. 1, pp. 84-95., 2000.

[11] Yun, B.-H., Cho, M.-J. and Rim, H.-C., Segmenting Korean compound noun using statistical information and a preference rule. *Journal of KISS(B): Software and Applications*, vol. 24, no. 8, pp. 900-909. 1997

[12] Manning, C. D. and Schutze, H., *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.