Journal of the Research Institute of Industrial Technology, Vol 19, 2002

공기정보를 이용한 한국어 요약 시스템의 성능개선

박호진[1], 김재훈[1]

Performance Improvement of Korean Indicative Summarizer Using Collocation

Ho-Jin Park[1], Jae-Hoon Kim[1]

ABSTRACT

In this paper, we describe the performance improvement of Korean indicative summarizer using collocation information. We use two types of collocation information, compound nouns and syntactic relations, in which a noun collocates with a noun and a verb in a document, respectively. These informations keep more meanings of a sentence compared to a noun list, which is used in the existing system, then we use compound nouns and syntactic relations to represent the meaning of a sentence. In this paper, they are extracted from a document using t test.

Our experiments show that using collocations for indicative summarizers is very useful, particularly compound nouns. We observe that the collocations are useful in the high compression rate of a document and it is very difficult to extract correct collocations from a document, especially a small-size document, so a good method for extracting collocations should be develope. for further study.

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 정보압축이다. 문서요약은 일상적인 생활에서는 널리 사용되고 있는 방법이다. 최근 문서요약은 단순한 하나의 문서의 내용을 요약하는 것이 아니라 여러 문서의 내용을 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.

Salton 등[8]은 하나의 문서를 그래프로 표현하였으며, 이를 문서관계도이라고 한다[8]. 문서관계도

1) 한국해양대학교 컴퓨터공학과 첨단정보기술연구센터

에서 노드는 문장 혹은 단락이고, 링크는 의미적으로 관련된 노드들 사이의 관계를 나타낸다. 이 관계는 노드와 노드 사이의 유사도가 어떤 임계값 이상일 경우를 말한다. 무성도는 문서관계도에서 다른 노드와 연결된 링크 수(일명 부쉬경로), 즉 노드의 차수이며, 무성도가 높으면 높을수록 많은 다른 노드들과 연결되었음을 의미한다. 문서요약은 단락이나 문장을 무성도가 높은 순으로 재배치하는 것이다. 본 논문에서는 *Salton* 등의 무성도의 개념을 링크 수가 아닌 유사도의 합으로 정의한 도합유사도 방법[9]을 사용하여 요약을 하였다

[9]는 문장의 유사도를 계산할 때, 두 문장에 포함된 명사들이 얼마나 비슷한가를 이용하였다. 일반적으로 문장의 의미는 정확하게 표현하기 위해서는 자연언어처리 기법의 복잡한 과정을 이용하는데, 이 방법은 매우 복잡할 뿐 아니라 정확하게 분석한다는 것은 대단히 어려운 일이다. 그래서 일반적으로 추출 문서 요약에서는 문장의 단어 리스트과 같은 매우 간단한 방법으로 문장의 의미를 표현한다. 그러나 단어의 리스트만으로 문장의 의미를 표현하는 것은 그 문장이 가지고 있는 많은 정보를 잃어버리게 된다. 본 논문에서는 문장의 의미를 좀더 정확하게 표현하기 위해서 공기관계, 즉 복합명사와 구문구조를 문장의 의미 표현에 사용하고, 그것들이 한국어 문서요약 시스템에 미치는 영향을 살펴보고자 한다.

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

추출 문서 요약은 문서에 포함된 중요한 문장을 추출하여 그 문서의 요약으로 사용하는 것이다. 중요한 문장을 추출하기 위해서 한 문장을 벡터 공간의 한 점으로 표현하며, 이를 문장 벡터라고 한다. [9]에서는 문장 벡터를 명사 리스트로 표현하나, 본 논문에서는 문장의 의미를 좀더 정확하게 표현하기 위해서 명사뿐만 아니라 명사와 명사, 또는 명사와 용언과의 관계인 공기관계를 이용해서 문장 벡터를 정의한다. 본 절에서는 문장 벡터의 생성 방법에 대해서 기술한다.

2.1. 한국어 기준명사 추출

기준명사는 체언을 구성하는 가장 최소 단위를 말한다. 체언 중에서는 수사, 대명사, 의존명사를 제외한 보통명사와 고유명사의 최소 단위만을 본 논문에서는 기준명사라고 한다. 본 논문에서의 한국어 기준명사 추출 시스템의 구조는 그림 1과 같으며, 다음과 같은 절차에 의해서 기준명사를 추출하게 된다[10].

1. 사전을 이용해서 문장으로부터 수식언을 제거한다. 여기서 수식언은 부사, 관형사, 감탄사가 여기에 속한다.
2. 사전과 어미 집합을 이용해서 용언(동사, 형용사)을 제거한다. 몇몇 어절(예를 들면, "나는")은 명사구와 중의성이 발생되는데 이 중의성은 무시한다. 즉 체언과 용언의 중의성이 발생되면 용언을 선호하도록 하였다.
3. 식 (1)에 정의된 음절간의 상호정보를 이용해서 명사구에서 조사를 분리한다. 이 방법은 *Maosong* 등[14]에 의해서 중국어 단어분리 알고리즘을 약간 수정해서 사용하였다.
4. 수사는 오토마타를 이용해서 제거한다. 여기서 수사의 예를 들면, "1999년", "1천 2백원" 등을 의미한다.

$$mi(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \qquad \text{식 (1)}$$

5. 복합분리는 사전과 수정된 CYK 알고리즘[15]을 이용한다. CYK 알고리즘을 적용할 경우에 중의

성이 발생되는데, 이를 많은 수의 명사가 포함되는 경우를 우선하는 경험규칙을 사용하였다.

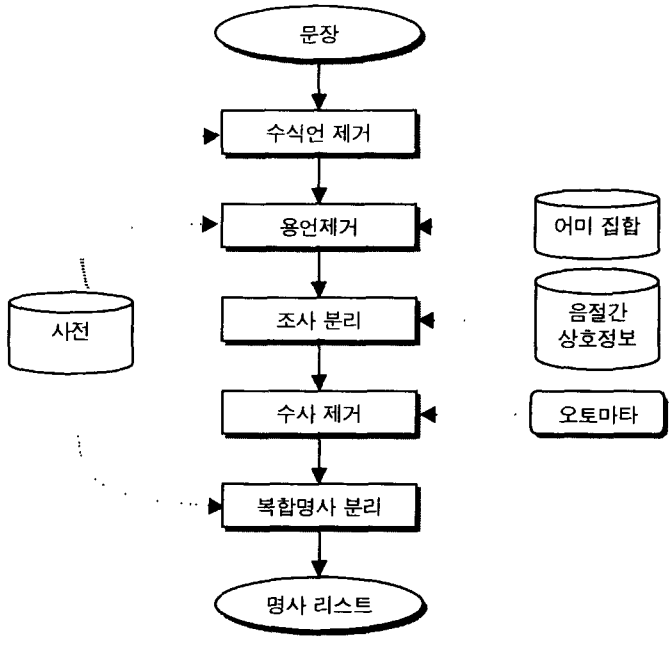

그림 1 한국어 명사추출 시스템의 구조

2.2. 공기정보 추출

공기정보란 두 개 이상의 단어로 구성되어 있는 표현이다[13]. 본 논문에서는 한국어 문서요약 시스템의 성능을 개선시키기 위하여 복합명사와 구문관계 공기정보를 사용하였다. 이 두 가지의 공기정보를 추출하기 위하여 t *test*를 사용하였으며, 식 (2)는 t 값을 계산하기 위하여 사용한 수식이다.

$$t=\frac{p(x,y)-p(x)p(y)}{\sqrt{\frac{p(x,y)}{N}}} \qquad \text{식 (2)}$$

식 (2)에서 x는 명사를, y는 명사 혹은 용언을 의미하며, $p(x)$는 x가 문서에서 나올 확률이다. 또한 $p(x,y)$는 x와 y가 함께 나올 확률이고, N은 문서에서의 명사나 용언의 개수이다. 만약 t 값이 어느 이상의 임계값을 넘을 경우, 두 단어 x, y가 공기관계를 가진다고 생각할 수 있다[13]. 본 논문에서는 이러한 공기관계를 이용하여 복합명사와 구문관계 공기정보를 추출한다.

먼저 복합명사의 공기정보를 추출하는 단계는 아래와 같다.

1. 명사추출기를 이용하여 명사를 추출한다.
2. 명사에 대한 빈도수를 계산한다.
3. 각각의 명사에 대한 확률을 계산한다.
4. t 값을 계산한다.

5. 임계값 이상을 가지는 x, y에 대하여 문장벡터에 $x\,y$를 추가한다.

구문관계 공기정보를 추출할 때에는 두 가지 경험규칙(heuristics)을 이용하였다. 첫 번째 경험규칙은 구문관계를 추출할 때 용언 앞의 명사 한 개만을 고려했다. 두 번째 경험규칙은 용언으로 추출된 것 중 1음절로 되어있는 용언은 제거하고 2음절 이상의 용언은 앞의 2음절만을 사용하여 구문관계를 추출하였다.

구문관계를 추출하는 단계는 아래와 같다.

1. 명사추출기를 이용하여 용언과 명사를 추출한다.
2. 경험규칙을 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.

3. 문서요약

본 논문에서는 문서요약 중에서 문장추출에 해당하며 통계적은 접근 방법을 사용한다. 문장은 명사 리스트로 표현되며, 문장 간의 유사도는 내적(*inner product*)을 사용한다. 문서요약의 알고리즘은 먼저 문장벡터간의 도합유사도를 계산하고 그 도합유사도가 높은 순으로 문장을 추출한다. 또한 본 논문은 원문서의 문장을 그대로 추출하여 요약문을 생성한다.

도합유사도의 계산은 [9]에서 사용한 방법을 그대로 사용하였는데, [9]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 유사도의 합으로 정의한다. 이를 도합유사도라고 한다[9]. 요약 문서에 적합한 문장을 추출할 때에는 문장의 중요도(도합유사도)에 따라 상위 순위에 해당하는 적정 수의 문장들을 추출하여 원문서에 나타난 순서대로 정렬시켜 요약문서를 생성한다[9]. 이와 같은 개념을 토대로 본 논문에서 제시한 한국어 문서 요약 시스템의 구조는 그림 2와 같다.

그림 2에서 전처리기는 입력문서를 문장 단위로 분리하고, 문장기호를 제거한다. 문장을 분리하는 기준은 기호 ".!?"가 있으면 문장으로 분리하였다. 이외에서 "1999. 12." 등과 같은 문자열에 대해서 오류가 발생하기 때문에 약간의 경험규칙을 사용하였다. 명사추출 방법과 공기관계 추출방법은 2절에서 구체적으로 설명하였다.

문장추출은 먼저 도합유사도가 높은 순으로 문장을 재정렬한다. 그리고 나서 앞에서부터 원하는 비율만큼의 문장을 추출하여 요약문서로 출력한다.

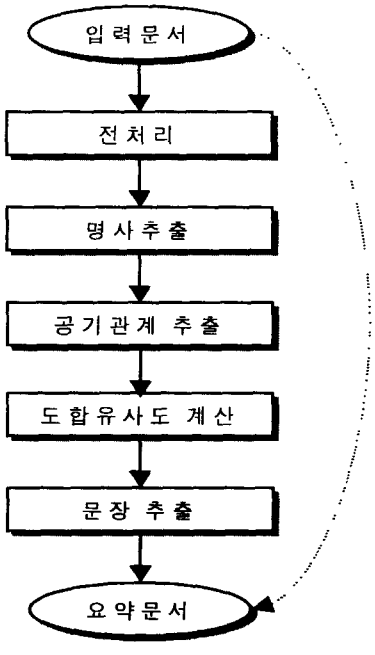

그림 2 한국어 문서요약 시스템의 개요

4. 실험 및 평가

본 논문은 기존의 한국어 요약시스템에서 복합명사와 구문관계가 요약에 미치는 영향을 보이기 위하여 기존시스템이 사용한 요약 말뭉치와 평가 방법으로 실험을 하였다[9]. 평가용 말뭉치는 기존의 논문과 같이 세 부류로 나누어 사용하였는데, C1은 논문 전체를 의미하고, C2는 신문기사, 그리고 C3는 C1의 논문에서 초록과 결론 부분에 속하는 말뭉치이다.

성능평가의 측도는 정보검색 분야에서 널리 사용되고 있는 정확률과 재현율을 사용하였으며, 이들은 각각 식 (3), (4)와 같이 정의된다[13].

$$P=\frac{N_r}{N_s} \qquad \text{식 (3)}$$

$$P=\frac{N_r}{N_c} \qquad \text{식 (4)}$$

여기서 N_s는 문서요약시스템이 제시한 전체 문장 수이고, N_r은 N_s 중에서 평가요약문서에 속한 문장 수이고, N_c는 평가요약문서에 속한 문장 수이다.

본 논문에서는 세 가지의 실험을 하였다. 첫 번째 실험은 복합명사 공기관계가 요약시스템의 성능 향상에 어떠한 영향을 주는가에 대한 실험이고, 두 번째 실험은 구문관계 공기관계가 성능에 미치는 영향을 실험하였다. 각각의 실험 결과는 <표1 >, <표 2>에 나타나있다.

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

<표 1>과 <표 2>에 나타나 있는 것과 같이 복합명사에서는 임계값 1.645, 즉 신뢰도가 90%일 때, 가장 좋은 성능을 보였으며, 구문관계에서는 임계값 1.960(신뢰도 95%)일 때, 가장 좋은 성능을 나타났다. 마지막 세 번째 실험은 이전 두 실험의 최적 임계값을 가지고 복합명사와 구문관계를 같이 적용했을 때의 성능이다. <표 3>은 세 번째 실험의 10% 요약일 때의 성능이다.

요약률	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

<표 4>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 및 향후 연구방향

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

제안된 방법은 기존의 방법보다 약 1%~2% 정도의 성능 향상을 보였다. 또한 10% 요약보다 20% 요약에서의 성능 향상이 비약했다. 물론 많은 성능의 향상은 아니지만, 복합명사와 구문관계가 요약 시스템에서 고려해야할 한 대상이라는 것을 확인한 계기가 되었다.

향후 연구방향은 공기관계 추출에 있어서 단순한 모델이 아닌 다양한 방법이 필요하다. 본 논문에서는 단순한 모델을 사용함으로서 구현이 용이하고 빠르다는 장점이 있지만 정확한 공기관계 추출이 어렵다. 또한 공기관계에 가중치를 부여하여 문장 간의 유사도를 높이는 방법도 고려하면 좀 더 나은 성능의 요약시스템을 구현할 수 있을 것이다.

6. 감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 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. "MINDS multilingual interactive document summarization". in Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization, Spring, pp. 131-132. 1998

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

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

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

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

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

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

[8] Salton, G., Singhal, A., Mitra, M. and Buckly, C. "Automatic Text Structureing and Summarization. in Mani, I. and Maybury, M. T"., editors, Advances in Automatic Text Summarization, pp. 61-70. 1999

[9] 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.

[10] 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.

[11] 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.

[12] 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

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

[14] Maosong, S., Dayang, S. and Tsou, B. K. "Chinese word segmentation without using lexicon and hand-crafted training data". in Proceedings of COLING-ACL 98, pp. 1265-1271. 1998

[15] Aho. V. A. and Ullman, J. D *The Theory of Parsing, Translation, and Compiling*, Prentice-Hall. 1973z