한국어 영 대용어 처리를 통한 문서요약의 성능 향상

구상옥° 전명희 김미진 이상조
경북대학교 컴퓨터공학과 언어·정보연구실
tomato@sejong.knu.ac.kr°, sjlee@knu.ac.kr

Efficient Summarization Using Zero Anaphora Resolution

Sang-Ok Koo°, Myoung-Hee Jung, Mi-jin Kim, Sang-Jo Lee
Department of Computer Engineering, Kyungpook National University

요 약

본 논문에서는 보다 간결한 요약문을 생성하기 위하여, 문장 전체를 추출하는 것이 아니라 문장의 일부분을 요약으로 추출한다. 그런데 한국어의 경우 문장 구조상 반복되는 문장성분을 생략하는 영 대용 문제가 빈번하게 발생하기 때문에, 문장의 일부분 추출 시, 생략된 성분을 복원하지 않으면 요약문의 의미가 불완전하고 모호해 질 수 있다. 본 논문에서는 문서 안에서 중요한 부분을 추출한 뒤, 생략된 성분을 복원하여 요약문의 가독성을 높이는 방법을 제안한다. Luhn의 방법을 이용하여 문서내의 중요 클러스터를 추출하였고, 기존의 문장분할 및 영 대용어 복원 알고리즘을 사용하여 생략된 성분을 복원하였다. 본 논문에서 제안된 요약 방법은 신문기사와 같이 문장의 수는 많지 않고, 문장의 길이가 비교적 긴 문서를 짧은 문장으로 요약하는 데 효율적이다.

1. 서론

일반적으로 자연언어에서는 문장 구조상 반복되는 부분을 생략하거나 더 짧은 표현으로 대치하는 경향이 있다. 반복되는 부분을 생략하는 경우는 원래의 표현을 영(null) 표현으로 대치하는 것이라 생각하고, 이런 경우를 가리켜 영 대치라 한다. 그리고 이때 생략된 성분을 영 대용어라고 한다. 영 대용어를 복원하지 않고 문서의 일부분, 즉 몇 개의 문장 또는 몇 개의 절을 추출하여 요약하게 되면, 생략된 성분으로 인해 요약문의 내용이 매우 불완전해 지거나 의미가 모호해질 가능성이 크다. 따라서 보다 정확한 요약을 위해서 영 대용어 해결은 매우 중요하다고 할 수 있다.

영 대용어 처리 문제는 언어학적인 이론 연구 뿐만 아니라 전산학적으로도 활발한 연구가 이루어지고 있다. 국내 연구로는 중심화 이론과 개념 그래프를 사용하여, 체언의 대용과 생략을 시도하기도 했고[1][2], 중심화 이론이 한국어 대용과 생략을 해결하는데 적합한지 검토하고, *cf* 목록 순서를 한국어에 맞게 결정하는 기법도 있었다[3]. 국외의 경우는 국내보다 영 대용 연구가 활발하게 진행되고 있다. Slot grammar를 이용해 대용과 삭제 해결을 위한 세가지 알고리즘을 제시하고[5], 중심화 이론에 기반한 일본어의 담화 해석과 구문적 단서간의 연관성을 밝힌 연구[6][7][8] 및 문서구조와 의미 제약을 사용하여 일어 영 대명사를 해결한 연구도 있다[9].

이 논문에서는 문장 내에 생략된 영 대용어를 복원하여 요약 시스템의 성능을 향상시키고자 하였다. 영 대용어 복원은 기존에 연구된 바 있는 분해 및 복원 알고리즘을 이용하였고[4], 요약 시스템은 Luhn의 방법을 적용하여 구축하였다[10].

본 논문은 다음과 같이 구성된다. 2장에서는 영 대용어 복원과 요약에 관한 선행연구를, 3장에서는 본 논문에서 제안하는 문서 요약 방법을 명시하였다. 4장에서는 실험 결과를 통해 제안된 요약 시스템의 성능을 평가하고, 5장에서 결론을 맺는다.

2. 선행연구

2.1 한국어 영 대용어 처리

이 장에서는 기존에 제안된 한국어의 영 대용어를 해결하는 방안과 그 절차에 대해서 설명한다[4].

2.1.1 문장 분할

영 대치는 복합문에서 주로 나타나는 현상이다. 따라서 영 대용어 분석을 위해서는 우선 주어진 문장이 복합문인지 여부를 판단해야 한다. 문장 내에서 서술어의 개수가 2개 이상이면 복합문으로 간주한다.

복합문에서는 하나의 서술어는 하나의 사건 EE(Elementary Event)를 나타낸다. 생략은 두개 이상의 EE가 접속되거나 내포되면서 발생하는 현상이므로, 생략된 성분 복원을 위해서는 한 문장 안에서 나타나는 두 개 이상의 EE를 각각의 EE로 분리하여 처리해야 한다. 이 과정이 복합문 분해이다. 김미진[4]에서는 한국어 복합문의 종류를 인용문, 명사화 내포문, 관형화 내포문, 접속문으로 나누었다. 복합문에서의 생략현상은 일반적으로 접속문에서는 후행절과 선행절, 내포문에서는 내포절, 인용문에서는 인용절에서 일어나기 때문에 문장에 내포된 절을 분리해 내어야 한다. 한국어에서 복합문의 특성은 어미에 잘 드러나므로 어미의 종류에 따라 복합문의 종류를 판별하고, 문장을 분할한다. 분할은 복합문의 종류에 따라 적용되는 알고리즘이 다르다.

2.1.2 영 대용어 탐색 및 복원

복합문을 활용어미에 따라 분해한 후 용언의 하위범주

정보를 이용하여 생략된 성분인 영 대용어를 찾는다. 탐색 된 영 대용어를 복원할 때에는, 접속문의 경우 접속 삭감 복원 규칙, 내포문의 경우는 관계명사구 탈락 복원 규칙 등을 사용하였다. 그리고 인용문의 경우는 주어 인칭 제약에 따른 동일 명사구 탈락규칙 등을 이용하였다.

2.2 문서 요약

2.2.1절에서는 해당 문서의 중요 단어를 추출하는 방법을, 2.2.2절에서는 중요 단어를 중심으로 문서내의 중요 문장을 추출하는 Luhn의 방법에 대해 설명한다.

2.2.1 중요 단어 추출

중요 단어(significant word)란 문장 내에서 비교적 높은 중요도를 가지는 명사들을 가리킨다. 단어의 중요도를 판단하기 위한 가중치 계산에는 널리 알려진 *TF/IDF* measure를 적용한다[11]. 즉 어떤 단어가 하나의 문서에서 차지하는 중요도는 그 문서 내에서의 단어의 빈도(TF)와 전체 문서집합에서의 그 단어의 역문헌빈도(IDF)의 곱이다. 이것이 의미하는 바를 예를 들어 설명하면, 어떤 단어가 A라는 문서 내에서는 출현빈도가 높고, 전체 문서집합에서는 출현빈도가 낮았다면 그 단어는 문서 A에서 매우 중요한 단어로 간주된다. 반면, 문서 A에서 출현빈도가 높은 단어일지라도, 문서집합 내의 모든 문서에 나타난다면, 그 단어는 낮은 가중치를 가질 것이다.

2.2.2 요약 문장 추출

Lunh은 중요한 단어들이 연속적으로 많이 나타나는 문장이 문서의 내용을 묘사하는데 매우 중요하다고 제안하고, 다음과 같은 중요 문장 추출 방법을 제시하였다[10].

먼저 문서 내의 모든 단어의 *TF/IDF* 가중치를 구한 다음, 가중치가 가중치한계 *W*보다 큰 단어를 그 문서의 중요 단어로 표시해 둔다. 그리고 나서, 다음의 조건을 만족하는 문장 내 클러스터(within-sentence cluster), 즉 각 문장내의 단어 열의 부분열을 찾는다.

조건 1) 클러스터는 중요 단어로 시작하여 중요 단어로 끝나야 한다. 중요 단어는 단어의 가중치가 가중치한계 *W*보다 큰 단어를 말한다.

조건 2) 중요 단어 사이에는 거리한계 *D* 보다 많은 수의 비중요 단어(insignificant word)가 올 수 없다.

예를 들어, 다음과 같은 문장 구조가 있다고 하자.

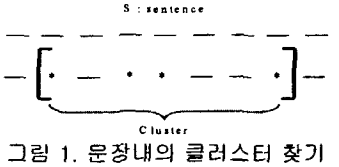

그림 1. 문장내의 클러스터 찾기

그림 1에서 — 는 비중요 단어 , * 는 중요 단어를 나타낸다. 만약 거리한계 D>2 라면, 중요 단어 사이에는 비중요 단어가 2개까지 올 수 있다. 따라서 대괄호 '[ ]'로 표시된 부분이 하나의 클러스터로 추출된다.

문서 내의 모든 클러스터를 구한 다음, 각 클러스터에 가중치를 부여하는데 각 클러스터의 가중치는 다음과 같이 계산한다.

$$W_c = \frac{n^2}{N} \qquad (1)$$

식 1에서 *N*은 클러스터 내의 전체 단어 개수, *n*은 클러스터 내의 중요 단어의 개수를 나타내며, 중요 단어의 개수의 제곱을 클러스터의 크기로 나누어 정규화 시킨 값 W_c이 클러스터의 가중치이다. 예를 들어, 그림 1에서 클러스터내의 전체 단어 개수는 7이고 그 중 중요단어의 개수는 4이므로 클러스터의 가중치는 $4^2/7$ 이 되어 16/7 이 된다.

한 문장 내에는 0개 또는 여러 개의 클러스터가 존재할 수 있는데, 문장 내에서 가중치가 가장 높은 클러스터의 가중치를 그 문장의 가중치로 정하고, 그 문서에서 가장 높은 가중치를 가지는 문장을 그 문서를 대표하는 중요 문장으로 결정한다.

3. 영 대용어 처리를 통한 문서 요약

문장의 길이가 짧고, 단문이 많은 문서는 몇 개의 중요한 문장을 추출하는 것만으로도 문서 요약이 가능하다. 그러나, 신문기사와 같이 문장의 길이가 대체로 길고 복합문이 빈번하게 나타나는 실용문의 경우, 한 문장 전체를 요약문으로 추출한다면, 압축적인 요약 결과를 얻지 못하는 경우가 많다. 특히 PDA나 휴대폰 같은 휴대용 기기는 소형 인터페이스라는 특성 때문에 더욱 간결하면서도 압축적인 형태의 요약을 필요로 한다. 따라서 우리는 높은 가중치를 가지는 문장 전체를 추출하는 것이 아니라, 높은 가중치를 가지는 클러스터 자체를 요약문으로 보고 그것을 완전한 절의 형태로 복원하여 요약문을 생성하는 방법을 제안한다.

본 논문에서 제안하는 요약문 생성방법은 다음과 같다. 먼저 대상 문서 내의 모든 명사의 가중치를 구하고, 명사의 가중치를 이용하여 Luhn의 방법대로 모든 문장 내 클러스터를 추출한다. 다음으로 가장 가중치가 높은 클러스터를 가지는 문장을 가져온 뒤, [4]에서 제안된 방법을 사용하여 그 문장의 문장분해정보와 영 대용어 탐색정보를 획득한다. 마지막으로 완전한 요약문 생성을 위하여, 문장 내의 클러스터의 크기를 완전한 의미를 표현할 수 있는 절로 확장한다. 확장된 절에서 생략된 성분이 있으면 이를 복원한다.

다음 예문들에 나타나는 문장이 가장 높은 클러스터를 포함하는 문장이라고 가정하자.

예문 1) 한국 투자신탁은 거래소 시장 활성화 대책이 *중소형 우량주*에 유리한 환경을 제공해 (*거래소 우량 중소형주들의 주가 상승*에) 도움이 될 것이라고 덧붙였다.

예문 2) 한국 투자신탁은 [[거래소 시장 활성화 대책이 중소형 우량주에 유리한 환경을 제공해][$Ø_{sub}$ (거래소 우량 중소형주들의 주가 상승에) 도움이 될 것이라]]고 덧붙였다.

예문 3) 한국 투자신탁은 [[거래소 시장 활성화 대책이 중소형 우량주에 유리한 환경을 제공해][($Ø_{sub}$ 거래소 우량 중소형주들의 주가 상승에 도움이 될 것이라)]]고 덧붙였다.

예문 4) 한국 투자신탁은 [[거래소 시장 활성화 대책이 중소형 우량주에 유리한 환경을 제공해][(거래소 시장 활성화 대책이 거래소 우량 중소형주들의 주가 상승에 도움이 될 것이라)]]고 덧붙였다.

예문 1에서 이탤릭체로 강조된 명사들이 중요 단어들이며, 소괄호로 표시된 부분이 클러스터이다. 그러나 이 클러스터만으로는 완전한 요약문이라고 하기 어렵다. 그래서 문장분해정보를 바탕으로 클러스터의 처음과 끝에서

가장 가까운 분해 마크가 있는 곳까지 클러스터를 확장한다. 예문 2는 문장분해정보 및 영 대용어 탐색정보를 보여주고, 예문 3은 예문 2의 분해정보를 바탕으로 클러스터를 완전한 절의 형태로 확장한 결과를 보여준다. 그러나 클러스터가 절의 형태를 가지게 되었다 하더라도, 주어가 생략되었기 때문에 예문 3의 클러스터만으로는 정확한 의미를 파악하기가 어렵다. 영 대용어 복원규칙 중 접속삭감복원규칙에 의해 생략된 주어성분을 복원하면 예문 4와 같이 하나의 문장으로서 완전한 의미와 형태를 가지는 클러스터를 구할 수 있다. 마지막으로, 최종 클러스터의 맨 끝에 나타나는 서술어의 어미를 종결어미 '다'로 변환하는 등의 후처리를 통하여 요약문 "거래소 시장 활성화 대책이 거래소 우량 중소형주들의 주가 상승에 도움이 될 것이다"를 얻을 수 있다.

4. 실험 및 평가

본 논문에서 제안한 요약시스템을 이용하여 130편의 신문기사에 대해 문서 당 단 하나의 요약문 만을 추출하였다. 대상 문서는 매일 경제, 한겨레 신문에서 증권, 금융, 부동산, 정치, 사회, 문화 별로 발췌한 것이다. 실험문장은 전체 1328문장 중 단문이 302문장이고 복합문이 1026문장이었다. 이는 한 문서 당 평균 10개 이상의 문장을 가지고 있으며, 이 중 약 8개가 복합문이라는 것을 말해준다. 뿐만 아니라, 복합문은 단문보다 비교적 문장의 길이가 길기 때문에, 단문보다 중요 클러스터를 포함할 가능성이 크다.

요약시스템의 성능 비교를 위해 본 연구에서는 단순히 첫 문장만을 추출하는 방법, 하나의 중요 문장을 추출하는 방법, 하나의 중요 클러스터만 추출하는 방법, 그리고 추출된 중요 클러스터를 제안된 방법으로 확장하고 생략된 성분을 복원하는 방법에 관해 모두 실험을 하였다.

요약결과에 대한 성능 평가는 2명의 평가자에 의해 다음과 같은 방법으로 평가되었다. 평가자 A는 신문의 제목과 위의 4가지 방법에 의해 얻어진 요약결과를 보고, 10점을 만점으로 하여 각 요약에 대해 점수를 부여한다. 평가자 B는 제목, 문서 전체 그리고 4가지 요약결과를 모두 보고, 마찬가지로 10점을 만점으로 하여 각 요약결과에 대해 점수를 부여한다. A, B 둘 다 요약의 간결성과 의미의 완결성을 기준으로 평가하도록 하였고, B의 경우, 특히 제목의 내용과 중복되지 않으면서도, 중요한 의미를 담고 있는 요약에 좋은 점수를 주도록 당부하였다. 표 1은 요약 결과에 대한 정확도 평가 결과 결과를 보여준다. A의 점수와 B의 점수는 평가자들이 130개 문서 각각의 요약결과에 부여한 점수의 평균값이다.

표 1. 요약 정확도 평가 결과

요약 방법	A의 점수	B의 점수	A, B의 평균
첫 문장 추출	8.52	6.05	7.29
중요 문장 추출	5.05	4.75	4.90
클러스터만 추출	3.24	1.68	2.46
제안된 방법	6.87	8.96	7.92

표 1의 결과를 보면, 제안된 방법이 나머지 방법들보다 좋은 점수를 얻었으며, 특히 중요 문장만을 추출하는 방법에 비해 약 62%의 높은 성능향상을 보였다. A의 평가에서 첫 문장을 요약으로 사용하는 방법이 높은 점수를 얻었는데, 이는 첫 문장에서 제목의 문구가 그대로 반복되는 신문기사의 특성 때문인 것으로 보여진다.

5. 결론

본 논문에서는 영 대용어 복원을 통하여 문서 요약의 품질을 향상시킬 수 있음을 보여주었다. 본 논문에서 제안하는 요약 시스템에서는 보다 간결한 요약문을 획득하기 위하여 요약문을 추출할 때, 문장 대신 문장의 일부분인 문장 내 클러스터를 추출한다. 문장의 일부분이 형식적으로나 의미적으로 불완전한 경우가 많으므로, 이를 보완하기 위하여 문장분할 및 영 대용어 탐색 정보를 이용한다. 특히, 추출된 클러스터에 생략된 성분인 영 대용어가 있을 경우, 이를 복원함으로써 요약문 자체가 의미적으로 보다 완결성을 가지도록 하였다. 신문기사를 대상으로 한 실험 결과 제안된 방법이 간결성과 의미의 완결성 면에서 비교적 좋은 성능을 보이는 것으로 평가되었다.

앞으로 요약 시스템을 보다 많은 양의 문서에 적용해 보고, 신문기사 뿐 아니라, 다른 종류의 문서에도 적용해 보고자 한다. 아울러, 보다 좋은 성능을 위한 중요 단어 또는 중요 클러스터의 결정 방법 및 가중치한계와 거리한계를 조율(tuning)하여 최적의 수치를 얻기 위한 방법도 향후 연구해야 할 과제이다.

참고문헌

[1] 한승연, 송만석, " 개념 그래프를 이용한 대용어 해결 시스템", 한국정보처리학회 추계 학술발표논문집, 제2권 제2호, pp. 844-851, 1995.

[2] 한승연, " 지식 기반을 이용한 대용어 해결 시스템", 연세대학교 전산과학과 석사학위논문, 1995

[3] 차건희, 송도규, 박재득, " 한국어 대용과 생략 해결을 위한 센터링 이론의 적용", 한글 및 한국어정보처리발표논문집, pp347-352,1997.

[4] 김미진, 박미성, 구상옥, 강보영, 이상조, " 한국어 복합문에서의 제로 대용어 처리를 위한 분해 알고리즘과 복원규칙", 정보과학회논문지 : 소프트웨어 및 응용 제 29권 제10호, pp. 737-746, 2002.

[5] Lappin, S. and McCord, M., " Anaphora Resolution in Slot Grammar", Computational Linguistics, Vol 16, No.4 pp.197-212, 1990.

[6] Walker, M., " Centering, Anaphora Resolution, and Discourse Structure," Centering in Discourse, eds. Marilyn A. Walker, Aravind K. Joshi and Ellen F. Prince, Oxford Univirsty Press, 1997.

[7] Walker, M. A., Iida, M., and Cote, S., " Centering in Japanese Discourse," Proceedings of the 13th International Conference on Computational Linguistics (Coling 90), 1990.

[8] Walker, M. A., Iida, M., and Cote, S., " Japanese Discourse and the process of centering," computational Linguistics, 20, pp. 193-232, 1994.

[9] Nakaiwa, H., " Automatic Extraction of Rules for Anaphora Resolution of Japanese Zero Pronouns from Aligned Sentence Paris," Proc. Of ACL-97/EACL-97 Workshop on Operational Factors in Practical, Robust, Anaphora Resolution for Unrestricted Texts, Madrid, Spain, pp. 22-29, 1997.

[10] Luhn, H. P., *The automatic creation of literature abstracts*, IBM J. Res. Dev. 2(2), 159-165, 1958.

[11] Salton, G., *Automatic Text Processing*, Addison-Wesley, 1989.