한국컴퓨터정보학회 동계학술대회 논문집 제19권 제1호 (2011. 1)

의미변화을 고려한 문서 요약 알고리즘 연구

이진관°, 장혜숙*, 이종찬*, 박상준*, 박기홍*
°*군산대학교 컴퓨터정보공학과
e-mail: {hs5486, leejinkwan, chan2000, lubimia, spacepark}@kunsan.ac.kr

A Study of Text Summarization Algorithm Using a Meaning Distortion

Jin-Kwan Lee°, Hae-Sook Jang*, Jong-Chan Lee*, Sang-Joon Park*, Ki-Hong Park*
°*Dept of Computer Information Engineering, Kunsan National University

● 요 약 ●

스마트폰과 같은 소형 이동단말기의 보급이 확산됨에 따라서 이동단말을 통한 웹 접속이 크게 증가하고 있다. 따라서 작은 화면에 웹문서의 내용을 표현하기 위해 문서요약이 필요하다. 형태소 치환에 의한 문서요약 방법은, 문장해석에서 의미변화와 단축처리에서 일부 단락에 치우치는 문제가 발생한다. 본 논문에서는, 의미변화의 문제는 의미변화율이 낮은 순서에 따라 요약 규칙을 분류하고 이 순위에 따른 요약 알고리즘을 제안하였다. 치우치는 문제는 요약처리가 문서전체에 똑같이 적용되는 새로운 기준을 정의해 요약 알고리즘에 도입하였다. 제안방법의 유효성은 20명의 피실험자로 실험한 결과에 의해 입증되었다.

키워드: 문서요약(text summarization), 의미변화(meaning Distortion)

I. 서 론

최근 현대인들은 웹 문서를 통해 필요한 정보를 가장 빠르고 편리하게 제공 받을 수 있는 것으로 인식되고 있다[1]. 특히 스마트폰과 같은 소형 이동단말기의 보급이 확산됨에 따라서 이동단말을 통한 웹 접속 또한 크게 증가하고 있다[2]. 따라서 작은 화면에 웹문서의 내용을 표현하기 위해 문서요약이 필요하다.

문서요약이란 한마디로 문서의 내용을 압축하는 것이라고 볼 수 있다. 즉, 본래 문서가 가지고 있는 기본적인 의미를 유지하면서 문서의 길이나 정보의 복잡도 등을 줄이는 작업이다. 컴퓨터 및 정보처리의 발달로 종래에 종이로 이뤄진 문서들이 워드프로세서나 DTP로 컴퓨터에 저장되게 되었다. 또한 인터넷 웹페이지로 문서의 형태가 바뀌고 있으며, 스마트폰이 활성화되면서, 인터넷 문서의 문서요약이 필수가 되고 있다. 형태소 치환에 의한 문서 요약은 문장 해석에 의미 변화가 생기는 경우가 있다. 따라서 의미변화가 적은 단축 후보를 우선 제시하는 것이 바람직하다. 또한 단축의 처리가 일부에 치우치는 현상이 발생하므로 문서 전체에 똑같이 실시하는 것이 바람직하다.

본 논문에서는 우선 요약 규칙에 의미변화율의 순위를 결정한다. 그리고 이 우선 순위에 따라 요약 방법을 제안해, 의미변화의 문제를 해결한다. 또, 일부 문서에 치우치는 문제는, 처리단락을 균일하게 하는 새로운 기준을 정의해, 이 기준을 요약 알고리즘에 도입해서 해결한다.

II. 관련 연구

1. 관련연구

스마트폰과 같이 화면이 작은 이동단말을 통하여 웹 브라우징을 할 경우 (그림 1)에서 나타내는 바와 같이 웹 페이지의 크기와 단말이 제공할 수 있는 화면 크기의 차이로 인하여 효율적인 브라우징이 이루어지지 않는 문제점이 발생한다[5].

그림1. 스마트폰에서의 웹문서
Fig. 1. Web Document of the SmartPhone

그러므로 사용자의 편리성 증대와 함께 필요한 정보를 보다 더 신속하고 정확하게 제공하기 위해 이동 단말에 적합한 웹 브라우징 기술을 개발하려는 시도가 있다. 현재 연구 되어 있는 이동단말을 위한 웹 브라우징 기술에는 Minimo[4], Opera Mobile[3], Power Browser[6][7], Collapse-to-zoom[5], Multiclient Collaborative Browsing[8] 등이 있다 그러나 이러한 웹 브라이징 기술은 근본적으로 내용을 압축하거나 요약하지 못하므로 가독성에 한계가 있다.

기존 웹기반의 텍스트 요약 방법으로는 Hierarchical Summarization 알고리즘[9], SmartView[10], 및 어휘 체인을 이용한 텍스트 요약 방법[11] 등이 제시되었다. Hierarchical Summarization 알고리즘에서는 입력된 문서에 대해 문서의 들여쓰기와 내어 쓰기의 정도를 이용하여 문장들에 대해 순위를 매기고 그 결과를 트리 구조로 확대 시켜 계층적 구조로 파악하였다. SmartView에서는 HTML 문서의 컨텐츠를 논리적인 부분으로 나누고 각각의 부분들을 독립적으로 선택하여 볼 수 있는 알고리즘을 제안하였다. 어휘 체인을 이용한 텍스트 요약 방법에서는 텍스트 주제

별 분할 기능, Tagging기능, 파싱기능을 통하여 선택적으로 명사를 삭제하거나, 서로 다른 단어들 사이의 의미적 연관성을 찾아 연관 짓는 기능을 수행한다 그 외 연관성 있는 텍스트 문장을 선택하여 어휘 체인 모듈을 통하여 얻어낸 단어들을 비교한 뒤 해당 문장의 중요도에 따라 순위를 매기는 과정을 통하여 텍스트 요약이 이루어진다[12].

III. 본 론

1. 문서요약 규칙의 분류

본 논문에서는, 의미변화율을 논의하기 쉽게, 문법적으로 유사한 정형 표현에 근거해 규칙을 확장하여 분류하였다. replace(x)는 정형표현x의 치환 후보를 나타내고, 명사를 나타내는 형태소를 N으로 표시한다.

1.1 관계 표현에 근거하는 규칙

조사를 단축할 수 있는 정형표현, 관계표현을 1-K와 1-S에 분류하고, 1-S의 광의적인 해석으로서 더욱 세분하여 1-T, 1-X, 1-Y, 1-Z를 제안한다.

【1-K】 활용어를 포함하지 않는 개개 표현 K
NK → N replace(K)
예) 사고가 원인으로 → 사고로

【1-S】 활용어를 포함한 관용적인 관계 표현 S (활용어의 의미가 약한 관용적 표현).
NSN' → N replace(S) N'
예) 문화에 관한 연구 → 문화의 연구

【1-T】 활용어를 포함한 관계 표현 S 이외 T.
NTN' → N replace(T) N'
예) 외국에 살고 있는 친구 → 외국의 친구

【1-X】 T의 관계 표현으로 활용어가 다음의 명사 N'도 포함해 명사구에 치환할 수 있는 정형 표현 X.
NX → N replace(X)
예) 기기를 사용하는 것 → 기기의 사용
화상이 아름다운 것을 → 화상의 아름다움을

【1-Y】 T의 정형 표현으로 N'이 명사 이외의 품사로, 그것도 포함해 치환 할 수 있는 정형 표현 Y.
NY → Nreplace(Y)
예) 전부를 희망했었어도 → 전부의 희망에 반해

【1-Z】 의문사절을 포함한 정형 표현 Z
NZ → N replace(Z)
예) 문제를 어떻게 대처 할까 → 문제의 대처방법

1.2 술어와 동사를 기반으로 하는 규칙

용언의 보조적인 의미를 부가하는 정형표현을 술어표현이라고 부르고, 첫 단어에서 변하는 것을 2-J로 분류한다. 동사는 실질적인 의미를 앞에 있는 명사에 의존하는 표현으로, 이것을 2-D로 분류한다. 용언은 M으로 나타낸다.

【2-J】 첫 단어에 의한 술어표현 J
MJ → M replace(J)
예) 써있는 것이다 → 써있다

【2-D】 동사 표현 D에 의한 규칙.
MD → M replace(D)
예) 검색을 실행 한다 → 검색 한다

1.3 복합명사에의 요약 규칙

복합명사의 문법 규칙으로부터 요약규칙을 정의한다.

【3-B】 N와 N'의 사이의 정형표현 B의 소거.
NBN' → NN'
예) 수험을 위한 자격 → 수험자격

1.4 동의어에 의한 단축규칙

【4-R】 정형표현 R를 동의어에 의해 요약되는 경우로, 동일 품사에 의한 치환, R자체의 삭제 등이 있다.
R → replace(R)
예) 그렇지만 → 그러나(접속사의 동의어)
이러한 소프트웨어는 → 이것들은(명사의 생략)

2. 의미변화의 순위

단축처리로 생기는 의미변화를 조사하기 위해서, 코퍼스 등을 이용하여, 이하에 표시된 요약후보 약1,000개를 피실험자 20명에게 실험을 실시했다.

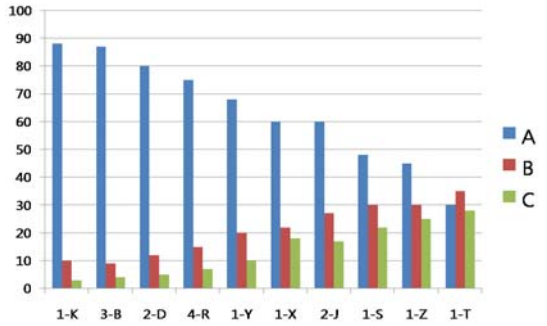

그림 2. 의미변화의 실험 결과
Fig. 2. Experimental results of meaning transformation

【단축후보의 예】 이 10년간에 있어서의[의(S)], 퍼스널 컴퓨터의 보급에는 눈부신 것이 있었다[눈부셨다(J)].

밑줄 친 부분을 대괄호([])내의 표현으로 단축했을 때, 의미변화 없음은 A, 의미는 조금 변하지만 문장의 해석에는 영향이 적은 경우는 B, 의미변화에 의해 문장의 해석에 문제가 생기는 경우는 C로 평가 했다. 이 실험(그림 2)에서 의미변화의 우선순위는 A의 비율이 높은 순서로 결정했다. 게다가 순위가 가장 낮은 분류 1-T에 대해 T의 전후의 명사쌍(Nl, N2)의 의미생성의 쌍을 분석하여 애매성이 적은 순위를 다음으로 결정했다.

순위1:(직업, 인간);순위 2. (인간, 선물);
순위3:(조직, 인간), 순위 4. (의복류, 인간);
순위5:(재료, 제작물);순위6:(목적, 수단);
순위7:(장소, 인간);순위8:(인간, 구상물).

순위1의 예 "의사의 형"에서는, "의사가 된 형", "의사인 형"등 추측이 한정되지만, 순위8의 예 "어머니의 회"에서는, 구매(어머니가 산 회), 소유(어머니가 가지고 있는 회), 작성(어머니가 그린 회) 등 애매하기 때문에 우선순위를 낮게 해야 한다. 덧붙여 분류 1-S에도 이 순위를 적용할 수 있다.

3. 문서 요약 알고리즘

실험 결과 순위를 응용하면, 같은 종류의 규칙이 치우쳐 사용되므로, 의미변화가 적어도 치우치는 문제가 발생한다. 따라서 다음의 3단계의 그룹으로 나누어, 이 치우치는 문제를 해결하였다.

1단계 : 1-K, 3-B, 2-D, 4-R
2단계 : 1-Y, 1-X, 2-J
3단계 : 1-S, 1-Z, 1-T

이 단계를 전역변수를 이용하여, 1단계부터 처리한다. 또, 각 단락 P에는 해석 위치를 전역 변수Loc(P)로 나타내, 미처리량의 비율이 높은 단락을 우선으로 처리한다. Len(P)는 전체 단락의 길이일 때, 다음과 같은 기준을 사용한다.

NEW_PRI(P) = Len(P) / Loc(P)

4. 평가

네 가지 종류의 문서에 대하여 피험자 20명 실험결과를 (표1)에 나타내었다. 요약행수는 20명의 평균치이며, 요약율은 전행수에 대한 요약행수의 비율(%)이다. 본 논문에서는, 형태소치환 방법보다 규칙을 추가하였으므로 요약율은 향상했다. 규칙의 적용회수는 20명의 선택총수이다.

표 1. 문서요약 실험결과
Table 1. Experimental results of text summarization

	문서1	문서2	문서3	문서4
문서정보 형태소수 단락수	2,995 38	2,530 33	3,903 42	4,682 36
요약결과 요약행수	28.3	47.6	35.1	47.1

Ⅳ. 결 론

본 논문에서 스마트폰 환경에서 보다 효율적인 브라우징을 제공하기 위한 문서요약 알고리즘을 설계하고 평가하였다. 평가를 통해 문서요약으로 생기는 의미변화와 일부 단락에 치우치는 문제의 해결법을 제안해 그 유효성을 입증했다. 향후는 구문 의미 해석을 이용한 보다 향상된 문서요약 규칙을 확장하여 개선 발전 시킬 예정이다.

참고문헌

[1] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma, “VIPS: a Vision based Page Segmentation Algorithm,” Microsoft Technical Report, MSR-TR-2003-79, pp.28, Nov., 2003.
[2] http://www.pewinternet.org
[3] http://www.opera.com/products/mobile
[4] http://www.mozilla.org/projects/min
[5] Patrck Baudisch, Xing Xie, Chong Wang, and Wei-Ying Ma, “Collapse-to-Zoom: Viewing Web Pages on Small Screen Devices by Interactively Removing Irrelevant Content,” 17th Annual ACM Symposium on User Interface Software and Technology(UIST 2004), TechNote, Sante Fe, NM, Oct., 2004.
[6] O. Buyullplten, H. Garcia-Molina, A. Paepacke , and T. Winograd, “Power Browser: Efficient Web Browsing for PDAs,” in Proc. of the Conf. on Human Factors in

Computing System, CHI'00, pp.430~437, 2000.

[7] O. Buyullplten, H. Garcia-Molina, A. Paepacke, “Accordion Summarization for End-Game Browsing on PDAs and Cellular Phones,” In Proc. of the Conf. on Human Factors in Computing System, Washington, CHI'01, 2001.

[8] Z. Hua, Lu, H., “Web Browsing on Small-Screen Devices: A Multiclient Collaborative Approach,” IEEE Pervasive Computing 5(2), pp.78~84, 2006.

[9] R. Dragomir, Radev, Omer kareem, and Jahna Otterbacher, “Hierarchical text summarization for WAP-enabled mobile devices,” SIGIR, pp.679, 2005.

[10] Natassa Milic-Frayling, Ralph Sommerer, “Smart View : Enhanced Document Viewer for Mobile Devices,” Microsoft Research Technical Report 2002-114, 2002.

[11] M. Brun, Y. Chali, and C. Pichak, “Text Summarisation using lexical chains,” In Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001. New Orleans, Louisiana, 2001.

[12] Ji-Eun Cha, Seung-Man Chun, Jong-Tae Park, “Design and Implementation of Web-based Text Summarization System for Mobile Device.