문서 요약 시스템을 위한 대표 개념어 생성의 격틀 구성

김성규, 김미진, 이상조
경북대학교 컴퓨터공학과

A Caseframe Structure of Concept-Based Topic Fusion for Text Summarization System

Sung-Kyu Kim and Mi-Jin Kim and Sang-Jo Lee
Department of Computer Engneering, Kyungpook National University

요 약

대량의 정보를 빠르고 쉽게 검색하기 위한 많은 문서 자동 요약 시스템이 개발되고 있다. 현재에는 원문에서의 추출을 통한 방법 뿐 아니라 요약문의 생성에 초점을 두고 요약 시스템이 개발되고 있으며, 복잡한 분석을 통해 효과적인 요약문을 생성한다. 본 논문은 보다 나은 문서 자동 요약 시스템을 위해 대표 개념어 생성기를 위한 격틀 구성 방안을 제시한다. 격틀 구성을 위한 단계별 과정과 핵심어의 추출, 그리고 격틀 구성의 제한요건을 서술한다.

1. 서론

정보 검색의 효율을 향상시키기 위한 방법은 여러 가지가 있다. 보다 정확하고 많은 정보를 찾기 위하여 다양한 방법으로 접근을 시도하고 있으며 주로 질의어에 대한 분석이나 문서에 대한 정확률, 검색속도의 개선에 그 초점을 맞추고 있다. 그리고 자연어 인터페이스 방식의 채용으로 질의에 대한 사용자의 요구를 구체적으로 알아보는 시도가 있다.

이에 관한 연구로 문서 자동 요약 시스템이 있다. 요약문은 문서를 대표할 수 있는 문장들의 집합이라 할 수 있기 때문에 많은 양의 정보를 검색하는데 있어서 요약문의 존재는 검색속도를 줄이고, 검색 효율성을 증대시킨다.

본 논문은 요약 시스템을 위한 대표 개념어 생성 격틀 구성 방법을 제안한다. 대표 개념어란 글에 나타나는 상황이나 상태를 하나 또는 여러 개의 집약된 단어와 문장으로 나타내는 것을 말하며, 단순히 문서내에 존재하는 단어수준이 아니라 개념수준으로 끌어올려 일반화된 격틀을 만든다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 관련연구와 기존 요약 시스템에 대해서 살펴본다. 그리고 3장에서는 대표 개념어 생성을 위한 구성요소와 각각의 내용을 설명하며, 4장에서는 시스템의 구현과 실험 및 평가를 한다. 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

문서 요약 방법은 크게 두 가지 방법으로 나눌 수 있다. 단순히 중요한 문장만을 추출하는 방법[3,4,5,6]과 문서를 자연어처리 분석과정을 거쳐 요약문을 생성해 내는 방법[1,2]이다. 중요 문장을 추출하는 방법은 많은 분석과정을 거치지 않으므로 처리속도가 빠르다. 하지만 단순히 추출된 문장만을 열거하므로 부자연스럽고, 원문에 의존적이다. 따라서 앞으로의 문서 요약은 높은 수준의 자연어처리 과정으로 요약문의 생성에 그 초점을 맞추어야 한다.

Chin-Yew Lin[1]은 문서의 요약에 있어서 요약이란 3가지 즉, Topic Identification, Topic Interpretation, Summary Generation의 세 가지 과정을 거쳐야 한다고 말하고 있다. 이는 주제를 인지하고, 해석하여, 요약문을 생성함에 있어서 추출이 아닌 생성으로써의 요약을 말하고 있다.

3. 대표 개념어 생성

우선 대표 개념어가 무엇인지를 알아보기 위해 다음 예를 보자.

> 철수와 만수는 돈이 필요했다. 그들은 복면과 총을 구입하고, 차를 훔쳤다. 모든 준비를 한 다음 은행으로 들어가 단 몇 분만에 만원짜리가 가득 들어 있는 가방을 몇 개 가지고 나왔다. 그들은 은행을 나와 차를 타고 유유히 사라졌다. 그리고 결코 잡히지 않았다.

[예제 1]

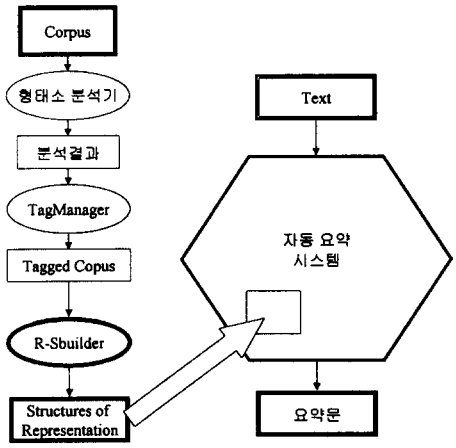
위의 예제는 시간적 구조를 가지는 서술적 성격을 가지며, 등장인물의 행위와 장소의 이동이 주된 내용이 된다. 이 글을 사람이 요약한다면 "철수와 만수가 은행에서 돈을 훔치다" 혹은 "철수와 만수가 도둑질하다" 정도로 요약할 수 있다. 그러나 글의 어디에도 중심이 되는 어구나 문장을 찾을 수 없고, 글의 주제가 되는 '도둑질하다' 라는 말도 문장에

서 찾을 수 없다. 이러한 글을 요약하기 위해서는 단순히 단어의 추출 이외에도 문장의 구조와 의미를 분석해야 한다.

본 논문에서는 위와 같은 글에 대한 대표 개념어를 생성하기 위해서 대표 개념어 격틀을 구성하는 것에 초점을 맞추고 있다. 여기서 말하는 격틀이란 하나의 단어를 설명하기 위해 만들어진 관련된 단어들이 시간적 순서를 가진 규칙적인 구조를 말한다. 한 단어가 여러 뜻을 가질 수 있으므로 대표 개념어도 여러 개의 격틀을 가질 수 있다.

3.1 요약 시스템

자동 요약 시스템에서의 대표 개념어 생성기의 위치는 그림 1과 같다. 왼쪽은 자동 요약 시스템의 한 부분으로서 대표 개념어 틀을 만들기 위한 전 단계를 나타낸다. R-Sbuiler는 이 논문에서 다루는 대표 개념어 격틀 생성기이고, Structure of Representation은 격틀 생성기의 결과이다.



[그림 1] 자동 요약 시스템에서의 대표 개념어 생성을 위한 격틀 구조

3.2 문서의 적합성

대표 개념어를 생성하는 데에는 코퍼스의 선택이 중요하다. 여기서는 위에서 살펴 본 예제에서와 같이 시간적 흐름의 구조를 가지고, 사건이 진행되는 형식의 글을 대상으로 한다. 대부분이 소설, 동화 등에서 이러한 구조를 발견할 수 있다. 이런 글들에 대한 형태적인 특징을 보면, 다음 예제에서 보면 동사의 사용이 현저하게 많은 것을 알 수 있다.

> 도덕적으로 훌륭한 삶을 산다는 것은 성인 군자나 학식이 많은 사람들만이 할 수 있는 일은 결코 아니다. 노력하면 누구나 실천할 수 있는 것이다. 즉, 어떻게 사는 것이 도덕적으로 훌륭한 삶을 사는 것인지 바르게 알아, 그렇게 살려고 최선을 다하면 되는 것이다.

[예제 2]

> 옛날 옛적 어느 곳에 한 어머니가 무남독녀 외동딸을 낳아 애지중지 키웠는데, 이 딸이 열두 살 먹었을 때 그만 병이 들어 죽었다. 그래서 새로 계모가 들어왔는데, 이 계모한테는 의붓딸이 눈에 가시인지라 그저 밤낮으로 구박을 하였다.

[예제 3]

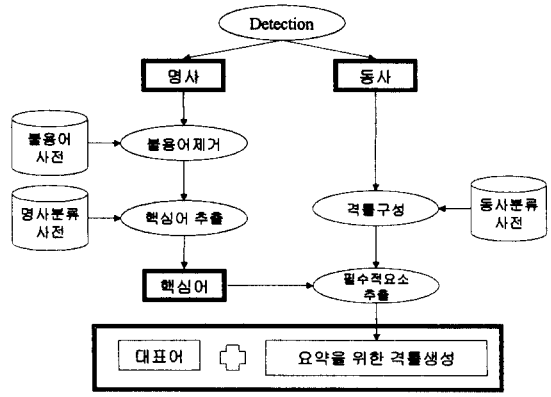
두 예제는 모두 단어나 문장에 있어서 비슷한 난이도를 가진다[7]. 그러나 동사의 사용 빈도는 현저하게 차이가 난다. 예제 2에서는 종결어휘가 모두 '~이다', '아니다'로 이루어져 있고, 예제 3에서는 거의 대부분이 행위를 나타내는 동사로 이루어져 있다. 하지만, 동사라고 해서 문서를 판별하는 데 모두 같은 가중치를 가지는 것은 아니며, '실천하다' 보다는 '죽다' 라는 어휘가 좀 더 대상에 가까운 글을 판별하는데 큰 작용을 한다. 문서가 대표 개념어를 생성할 수 있는지를 판별하기 위해 문서의 적합성을 계산한다.

$$\frac{N_v}{N_e} > C_1$$

Ne는 종결어미의 개수이고, Nv는 종결어휘로 쓰인 동사의 개수이다. 이 값이 임계치(C_1)보다 크면 대표 개념어를 생성하기 위한 문서로서의 적합성을 가진다.

3.3 대표 개념어 격틀의 생성

본 논문에서 제시하고 있는 대표 개념어 생성을 위한 격틀 구조 생성기는 그림 2와 같다. 기본 루틴은 크게 두 부분으로 나눌 수 있는데, 먼저 명사부의 분석을 통해서 글의 중심이 되는 사람, 사물, 장소를 추출하여 핵심어로 삼는다. 그리고 각 동사에 대해서 동사와 그 동사의 필수 논항을 추출하게 되는데 핵심어와 비교하여 필요없는 부분을 제외한 나머지를 요약을 위한 격틀로 만든다.



[그림 2] 격틀 구조 생성기

핵심어 선별요소는 명사의 사용빈도와 분산도를 이용하고, 선별을 위해 다음 식이 사용된다.

$$\log(af+bV) > C_2$$

식에서 보는 바와 같이 명사의 빈도와 분산도의 가중치를 곱한 값이

임계치(C_2)보다 크다면 핵심어로 선별된다. 분산도는 글 전체에 대해 고루 분포하는 핵심어, 즉 등장인물을 판별하기 위해서 사용된다. 명사의 f는 빈도, V는 분산도를, a와 b는 각각 명사의 빈도와 분산의 가중치를 나타낸다. 가중치는 명사 분류 사전의 값에 따라 달라진다.

여기서는 3가지의 사전을 사용한다. 불용어 사전은 형태소 분석결과로 추출된 명사 중 분석에 불필요한 명사를 모아놓았고, 명사 분류 사전은 명사의 분류를 크게 4가지로 사람, 사물, 장소, 기타로 분류하고 있다. 사람은 행위의 주체이고, 사물은 행위의 대상이며, 장소는 주체의 이동을 의미한다. 이는 이렇게 분류하는 것만으로도 핵심어 추출과 격틀 구성을 할 수 있으며, 구현에 있어서도 용이함을 보인다. 동사 분류 사전은 각 동사가 가져야 하는 필수 논항의 정보를 가진다. 예를 들어 '가다'는 "누가, 어디로"의 필수 논항을 가져야 한다. 이러한 정보는 한국어 동사 구문 사전[8]을 이용한다. 그림 하나의 예를 들어 보자.

> 옛날에 어떤 선비가 과거를 보러 갔다. 가다 보니 길가 버드나무에 까치 둥지가 있는데, 구렁이가 까치를 잡아먹으려고 까치둥지로 슬슬 기어올라갔다. 선비가 까치를 살리려고 구렁이를 활로 쏘았다. 그래서 구렁이는 화살을 맞아 죽고 까치는 살았다.

[예제 4]

[예제 4]는 "까치의 보은"이라는 동화의 첫 부분을 발췌한 것으로서 "선비가 까치의 목숨을 구하다"의 대표 개념어를 생성하기 위해 격틀을 구성한다. 우선 글 전체를 대상으로 명사부분 분석을 통해 핵심어인 선비, 까치, 구렁이를 추출하고, 격틀을 구성하여 필수요소를 추출한다. 추출된 결과는 다음과 같다. 여기서 마지막 요소는 문장상의 동사의 성격을 규정하고 있으며, 동사 뒤에 붙은 어미로 구분한다. 핵심어이지만 추출되지 못한 논항은 구문 분석기의 불완전성에 기인한다.

(보다, 선비가, 과거를, #서술)
(있다, 둥지가, #상태)
(잡아먹다, 구렁이가, 까치를, #의도)
(살리다, 선비가, 까치를, #의도)
(쏘다, 구렁이를, 활로, #서술)
(맞다, 구렁이가, 화살을, #서술)
(살다, 까치가, #서술)

격틀을 구성하기 위해 사용된 제한요건은 다음과 같다. 첫째, 주어 성분은 반드시 핵심어여야 한다. 둘째, 문장에서 필수 논항이 하나라도 없는 동사는 제외한다. 셋째, 대화체의 내용은 제외한다.

이렇게 구성된 격틀은 단순히 하나의 상황에 대한 예에 불과하다. 따라서 격틀을 정제하여 보다 객관적이고, 표준화된 격틀이 요구된다. 정제된 격틀은 요약 시스템에서 대표 개념어를 생성하기 위한 주된 요소가 된다.

4. 실험

대표 개념어 생성 격틀을 위한 적합한 전자문서의 획득과 분석, 그리고 사전 구성에 어려움이 있었다. 그래서 분석의 용이함을 위해 단어와 구문이 쉬운 동화[9]를 기본으로 선택하여 실험하였다.

전체 1657개의 대표 개념어 격틀 생성을 위해 동화 60539개의 어절이 사용되었으며, 그림 3은 예제 4에 대한 격틀 구성을 보이고 있다.

A가 C의 목숨을 구하다
보다 / A(s) / 과거(o) / 서술
있다 / 둥지(s) / 상태
잡아먹다 / B(s) / C(o) / 의도
살리다 / A(s) / C(o) / 의도
쏘다 / B(o) / 활(z) / 서술
맞다 / B(s) / 화살(o) / 서술
살다 / C(s) / 서술
선비:A / 구렁이:B / 까치:C

[그림 3] 격틀 구성의 예

위의 구성된 격틀은 "목숨을 구하다"라는 대표 개념어를 구성한 것이다. 첫 줄에는 대표 개념어를, 다음 줄부터는 이를 설명하기 위해 구성된 격틀, 대표 말을 들고 있다. 이는 하나의 상황 자체를 동사와 그 동사의 필수 논항으로 구성하고, 해당되는 단어에 대해 선비, 구렁이, 까치 대신에 A, B, C로 대치하고 있다. s는 주어, o는 목적어 z는 보어를 나타낸다.

전체 나타난 격틀 중에 대화체에 나온 경우를 뺀 나머지에서 동사의 성격이 '서술'로 사용되는 비중이 전체의 57.1%로 가장 많은 비중을 차지하고 있었다. 그리고, 하나의 대표 개념어가 둘 이상의 격틀을 가지는 경우도 2.14%를 차지하였다.

5. 결론 및 향후 연구 과제

본 연구에서는 자동 요약 시스템을 위한 대표 개념어 생성의 격틀 구성 방안을 제시하였다. 대표 개념어 격틀 구성을 위한 핵심어를 추출하고, 격틀을 구성, 필수 요소를 추출하였다. 이렇게 만들어진 격틀은 요약 시스템에 쓰이며, 단어의 의미구조를 밝히는 데 기초자료가 될 수 있다. 향후 연구 과제로는 동사에 따른 격틀의 표준화와 의미망을 이용한 격틀의 확장이 필요하다.

참고문헌

[1] Chin-Yew Lin, "Assemnly of Topic Extraction Modules in SUMMARIST", Information Science Institute, 1998

[2] Chin-Yew Lin, "Robust Automated Topic Identification", Ph.D. Thesis, August 1997

[3] Simone Teufel, Marc Moens, "Sentence extraction and rhetorical classification for flexible abstracts", AAAI Spring Symposium on Intelligent Text summarization, Stanford, March 1998.

[4] Daniel Marcu, "The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts", Ph.D dissertation, University of Toronto, Canada, 1997

[5] 장동현, 맹성현, "문서 구조 정보를 이용한 확률 모델 기반 자동요약 시스템", 제9회 한글 및 한국어정보처리 학술대회, 1997

[6] 강상배, 조혁규, 권혁철, 박재득, 박동인, "한국어 문서의 통계적 정보를 이용한 문서 요약 시스템 구현", 제9회 한글 및 한국어정보처리 학술대회, 1997

[7] 황미향, 한국어 텍스트의 계층구조와 결속표지의 기능 연구, 경북대학교 박사 학위 논문, 1998

[8] 홍재성, 한국어 동사 구문 사전, 동아출판사, 1997

[9] 동화작가 김문기의 홈페이지
http://myhome.netsgo.com/hipen/default.htm