전문용어 추출시스템

박정오[1], 황도삼
영남대학교 컴퓨터공학과
{jopark,dshwang}@nlp.yeungnam.ac.kr

A terminology extraction system

Jung-Oh Park, Do-Sam Hwang
Department of Computer Engineering Yeungnam University
and
Advanced Information Technology Research Center(AITrc)

요 약

현재, 과학기술, 정치, 사회, 문화의 급격한 변화와 발전에 따라, 전문분야마다 새로운 전문용어가 빈번히 생성되거나 소멸되고 있다. 이러한 전문용어를 포함한 문서를 정확히 해석하기 위해서는 전문용어 전자사전이 필요하다. 전문용어 전자사전을 개발하는데는 수시로 생성되는 전문용어 표제어를 정확히 추출하는 것이 무엇보다 중요하다. 본 논문에서는 이러한 전문용어 표제어를 컴퓨터를 이용하여 추출하는 시스템을 개발하였다. 기본적으로는 기존의 전문용어가 사용된 특정어구를 이용하여 전문용어를 추출한다. 또한, 전문용어의 어절 패턴을 이용하여 후보 전문용어를 추출한 후, 전문용어를 구성할 수 있는 단어의 위치정보를 이용하여 전문용어를 추출하는 방법을 제안한다. 기존 전문용어 사전에 없는 단어에 대해서는 시소러스를 이용하여 유사 단어의 위치정보를 이용하는 방법을 이용하였다.

1. 서론

일상 생활에서 사람들간의 대화는 일상적인 어휘만을 사용해도 의사 소통을 하는데는 어려움이 없다. 그러나 웹이나, 전문분야의 서적이나 신문 전문분야의 섹션 등에서 전문분야 어휘들을 접하게 될때, 누구나 한번쯤은 "이게 무슨 의미인가?" 하는 의문을 가진 경험이 있을 것이다. 이는 이제까지 알고 있던 분야의 용어가 아니기 때문이다. 다른 전문분야의 용어를 이해하기 위해서는 이런 용어들을 분류한 전문용어 사전이 필요하다. 또한 과학기술, 정치, 사회, 문화의 급격한 변화와 발전에 따라, 각 전문분야에서는 새로운 전문용어가 빈번히 생성되거나 소멸되고 있다. 이러한 전문용어를 포함한 문서를 정확히 해석하기 위해서는 전문용어 사전이 또한 필요하다. 전문용어 사전을 개발하는데는 수시로 생성되는 전문용어 표제어를 정확히 추출하는 것이 무엇보다 중요하다. 현재 필요에 의해 전문용어를 추출하는 경우 관련분야의 전문가가 수작업에 의해 전문용어를 추출한다. 이런 경우 객관적인 용어 추출이 어려우며 관련 전문가의 부족으로 용어 추출작업에 많은 경비나 시간이 소비된다는 단점이 있다. 본 논문은 이런 전문용어 추출의 어려움을 해결하고자 전문용어를 자동을 추출하는 시스템을 구성하였다. 컴퓨터 관련분야의 전문용어를 추출하였으며, 추출방식은 문서에서의 외래어 병기여부, 특정어구, 특수문자 사용과 같은 문서의 특징과 용어 사전을 사용하여 추출하였다.

2장에서는 전문용어의 추출방법으로 문장에서 전문용어 구문 패턴 및 시스템에 기술하고, 3장에서는 실험 및 평가에 대해 기술하고, 4장에서는 결론과 앞으로 연구방향에 대해 기술한다.

1) 본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

2. 전문용어 추출 방법

색인어는 문서에서 중요어만을 추출하지만, 전문용어는 문서에서 중요어로 사용되지는 않지만 관련분야를 이해하는데 도움이 되는 용어이다. 즉, 전문용어는 관련분야에서 사용되는 용어로서 그 분야에서만 통용되는 의미도 갖고 있다. 같은 용어라 할지라도 관련분야가 다르면 서로 다른 의미를 갖는 용어가 된다. 이러한 전문용어를 이해하기 위해서는 전문용어 사전이 필요하다. 이장에서는 문장에서 일반적인 전문용어의 출현패턴, 시스템의 구성, 추출방법, 추출에 이용한 재료들에 대해 기술한다.

2.1 전문용어 출현 형태

명사의 용어 추출 구문 패턴

◇ 명사 + 강조·정의 구문
예) 자료 + 구조 + 이 + 다
추출 용어 : 자료 구조
◇ 명사 + 격조사 + 명사 + 격조사
예) 적합 + 하 + ㄴ + 사전머신 + 을
추출 용어 : 사전머신
◇ 명사 + 격조사 + 명사 + 동사형어미
병렬 + 수행 + 시스템 + 을 + 구현 + 하 + 다
추출 용어 : 병렬 수행 시스템

위의 형태로 전문용어를 추출한다. 복합명사의 처리는 명사 +

명사의 형태만을 복합명사로 처리하여 용어를 추출한다.

text(문서) 정보를 이용

◇ 명사 + (+ 명사 +)
도메인 + 전문가 + (+ domain + expert +)
◇ 복합어(외래어+명사)
Round-Robin스케쥴링 + 방식

위와 같은 형태로 용어를 추출한다.

2.2 시스템 구성

본 논문에서 제안하는 시스템 구조는 그림 1과 같다. 전문용어 추출에서 대상문서 분석에는 형태소 해석기 HAM[1]을 이용했다.

2.3 패턴정보 및 문서의 특징을 이용한 후보 용어추출

형태소분석된 결과에서 명사는 용어 사전을 사용해 등록어, 미등록어, 복합어로 전문용어 가능한 용어를 분석한다. 그런 다음 미등록어와 복합어로 전문용어 가능어를 패턴정보와 문서의 특징(외국어 병기단어, 특수문자)에 적용하여 용어의 앞뒷어절을 분석하여 추출한다. 복합어 처리는 2어절의 처리만 처리하였다. 잘못된 용어 추출을 방지하기 위해서는 용언사전을 이용하였다. 용어사전은 일반용어사전(전문용어가 될 수 없는 일반적인 용어), 관련분야 단어사전(복합명사로 구성시 전문용어로 사용될 용어로 구성된 용어사전), 특수용어사전(2음절로 구성된 용어로서 이 어절이 포함된 용어는 전문용어가 될 수 없는 용어들로 구성한 용어사전) 등을 사용한다. 추출된 명사와 형용사 중 특정어구와 관련단어사전의 용어정보와 패턴정보에 이용된 패턴정보의 적합한 용어를 전문용어로 추출한다.

2.4 시소러스 처리

관련분야의 최신 용어는 전문용어일 가능성은 많지만 대부분 미등록어이다. 이를 분류하면 첫째, 새롭게 생성된 용어 둘째, 기존 용어와의 합성에 의해 생성된 용어 셋째, 기존 용어와 복합어로 구성된 용어이다. 복합어로 구성된 미등록어의 분석율을 높이기 위해 시소러스를 사용하여 미등록 용어를 분석한다.

예) 용어 - 분산회선 교환망
시소러스 - 분산회선(존재하는 용어)
분산(상위어)
관련분야단어사전 - 분산

예를 살펴보면 '분산회선 교환망'이라는 용어를 단일어로 분리하여 '분산회선'이라는 용어를 시소러스와 비교한 후, 시소러스에 있는 용어일 경우 상위어를 조사하여 이를 관련단어사전과 비교하여 처리한다.

2.5 관련분야 단어사전

관련분야 단어사전이란, 복합어로 구성된 기존의 전문용어를 단일어로 분리하여 용어사전으로 구축한 것이다. 이를 사용하여 미분석된 복합어 분석에 이용한다.

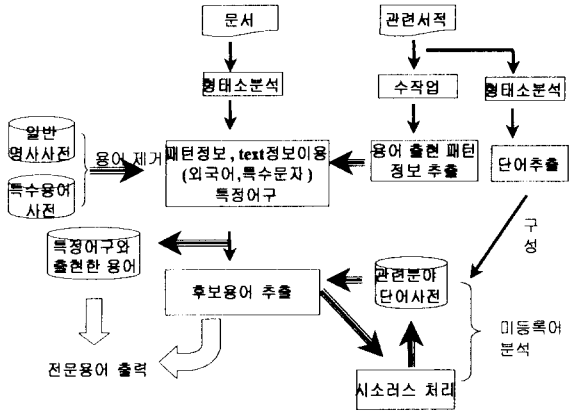

그림 1 전문용어 추출시스템

예) 복합어 구성시 전문용어 가능어 위치정보 빈도수
네트워크 /12 / 20
노드 /12 / 3
데이타 /12 / 36

복합어의 경우 단일어로 분류하여 각 단일어를 관련분야 단어사전과 비교하여 이 용어 사전에 있을 경우 위치정보를 비교하게 된다. 이 중 일치하는 용어를 전문용어로 추출하는데, 빈도수 조사는 하였지만 아직 빈약하여 실제 정확성을 높이는데 크게 도움이 되지 않아 현재는 이용하고 있지 않다.

3. 실험 및 평가

3.1 실험대상

본 논문에서는 대상 문서를 한국통신에서 작성한 코퍼스 KT-SET를 이용하여 실험하였다.

원시 문서만을 이용하여 처리 할 수 없어 용어의 분석을 위해 형태소해석기를 이용했다[1]. 미등록어 처리를 위해 세가지의 용어 사전을 이용하였다. 하나는 일반어휘, 즉 전문용어로 전혀 사용되지 않는 용어는 51,145개의 단어로 구성되어 있으며, 관련분야에서 복합어로 구성할때, 전문용어가 될 수 있는 용언은 1988개의 단어이다. 또한 미등록어 처리를 위해 합성어로 구성될 경우 전문용어가 될 수 없는 용어는 100개의 단어로 구성했다.

3.2 실험 결과

대상 자료를 이용해서 전문용어 추출을 실험하였다. 용어가 추출된 실험1의 결과는 [표 1]과 같다. 형태소분석된 라인의 수는 27만 정도이다. 형태소분석된 결과를 사용하여 전문용어를 추출하는 시스템을 구성하였다. 전문용어로 추출한 용어의 어절 수는 1어절 용어와 2어절 용어만이 처리가 되었기 때문에 3어절 이상의 전문용어는 추출에서 제외했다. 추출된 용어 중

중복된 용어를 제외한 용어는 2656개이다. 추출된 용어를 분석한 결과, 용어 중 기존 전문용어를 이용하여 제거된 용어는 436개이다. 추출된 용어는 2656개의 용어가 추출되었는데 그 중 적합한 용어는 1672개(62.9%) 추출되었다. 부적합 용어는 979(37.1%)이다.

표 1 대상자료 및 용어의 추출 결과

대상 라인의 수	형태석해석후 라인의 수	추출된 용어
18,650	273,882	2,656

표 2 용어 추출 결과

최종추출된 용어	부적합 용어		적합 용어
	잘못된 용어	해석 및 잘못된 용어	
2,656	826	153(5.7%)	1,672(62.9%)

실험 2는 일부 문서에 대한 평가이다. 너무 방대한 자료를 대상으로 하면 정확한 결과를 분석하기 어려우므로 문서의 일부분만을 수작업 및 전문용어 추출 시스템을 이용하여 추출하였다.

	라인의수	추출된 용어수	전문용어	부적합용어
수작업	651		213	
자동색인		184	126	58

$$\text{재현율} = \frac{\text{검색된 적합문헌의 수}(A)}{\text{적합한 전체 문헌의 수}(A+C)}$$

$$\text{정확률} = \frac{\text{검색된 적합문헌의 수}(A)}{\text{검색된 전체 문헌의 수}(A+B)}$$

A : 검색된 적합문헌의 수
B : 검색된 부적합 문헌의 수
C : 검색되지 않은 적합 문헌의 수

$$\text{재현율} = \frac{126}{126+87} = 59.1\%$$

$$\text{정확율} = \frac{126}{184} = 68.4\%$$

3.3 평가

KT-SET문서를 이용한 전문용어 추출 실험에서 실험1에서는 정확률이 62.9%이고, 실험2에서는 68.4%결과를 보였다. 두 실험에서 결과를 정확률이 낮은 편인데 결과를 분석해 보면 미등록어를 전문용어로 추출한 결과로 인해 이런 결과가 발생했다. 이를 해결하기 위해서는 용어사전을 이용할 수 있으나, 모든 일반용어 추가는 불가능하다. 그러므로 같은 용어를 사용하지 않는 일반용어 분석 방법이 필요하다.

추출된 용어 중 다른 분야에서는 전문용어로 이용되는 용어
- 금속 전극
- 무선 부문
- 무선 분야
- 저압화학기상증착법
- 저온 공정

추출된 용어를 분석하면, 용어 중 컴퓨터 분야의 용어는 아닌 공학 분야의 용어가 많이 추출되었다. 실제 이 용어들이 각 관련분야 전기·전자·건축 등 관련분야에서는 전문용어로 사용되지만 전산 분야에서 전문용어로 사용이 되지 않은 용어들이 추출되었다. 이를 상세히 분류할 수 있는 방법이 있어야 한다.

형태소 분석의 오류
- virtual reality등
- 균일도가 증가히
- 통신용 IC들
- 통한 기술예측결과

또한 형태소 해석 오류에 의해 발생한 오류들을 살펴보면 첫째, 영문자의 경우 현재의 국내 형태소해석기로는 형태소분석의 어려움이 있다. 일반적으로 한글과 같이 사용되면 명사로 사용이 된다. 하지만 영어의 경우 모든 용어를 명사로 해석을 한다. 둘째, 명사 분류 및 띄어쓰기 오류로 인한 형태소 분석의 실패의 경우이다. 이런 경우를 해결하기 위해서는 형태소 해석의 오류를 수정해 줄 수 있는 시스템의 구성이 필요하다.

IV. 결론

기존의 전자사전은 전문용어에 대해 용어의 개념 정립이 없었기 때문에 전문적인 서적 등 전문분야의 내용을 이해하기 위한 참고 자료로는 부족한 점이 있었다. 이를 해결하기 위해서는 관련분야의 전문용어로 구성된 전문용어사전이 필요하다.

향후 문서뿐만 아니라 이를 웹으로 확장하여 처리하여야 할 것이다. 요즘의 추세는 문서에 비해 웹에서 새로운 용어나 새로운 기술의 보급이 빨리 이루어지기 때문에 검색 에이전트와 결합하여 웹에서 검색하여 문서 처리를 할 수 있으면 최신의 정보에서 전문용어의 추출을 할 수 있을 것이다.

참고문헌

[1] 강승식, "형태소 해석기 HAM", http://ham.hansung.ac.kr/, 1998.

[2] 문화관광부, "전문용어 표준화를 위한 기반 조성", 한국과학기술원, 1998.

[3] 한성현, "구문해석을 이용한 색인어 자동 추출 시스템의 설계와 구현", 석사학위논문, KAIST, 1990.

[4] 정진성, "단일 문서내에서의 언어 및 통계정보를 이용한 자동색인", 석사학위논문, KAIST, 1992.

[5] 남영준, "코퍼스를 이용한 정보검색용 전자사전구축에 관한 연구", 한글 및 한국어정보처리 학술 발표논문집, pp430-440, 1996.

[6] 유춘식, "자연어 처리, 통계적 기법, 적합성 검증을 이용한 자동색인 시스템에 관한 연구", 한국정보처리학회 논문지, pp1552-1561, 1998.

[7] 양성현, "명사의 연어 정보와 서술성 명사의 공기 정보를 활용한 복합명사 분석 및 자동색인", 한글 및 한국어 정보처리 학술발표 논문집, pp59-64, 1997.