과학기술 전문용어를 위한 정제 말뭉치 워크벤치 개발

이병희*, 정휘웅**, 정한민*, 성원경*
*한국과학기술정보연구원 차세대정보시스템연구실
**부산대학교 인지과학협동과정
e-mail : *{bhlee, jhm, wksung}@kisti.re.kr, **hwjeong@pusan.ac.kr

Development of the Corpus Refinement Workbench for Science & Technology Terminology

Byeong-Hee Lee*, Hwi-Woong Jeong**, Hanmin Jung*, Won-Kyung Sung*
*Information System Research Lab., KISTI
**Department of Cognitive Science, Pusan National University

요 약

본 논문에서는 효과적으로 문서를 정제할 수 있는 작업환경인 웹 기반의 정제 말뭉치 워크벤치 개발에 관하여 기술한다. 또한 정보검색의 효율성 향상, 전문용어의 자동추출, 전문용어가 쓰인 문맥의 파악 등을 위하여 정제된 문서에 포함된 과학기술 전문용어를 표시할 수 있게 하는 작업환경도 구축하였다. 이렇게 개발된 정제 말뭉치 워크벤치와 전문용어 태깅 툴을 이용하여 과학기술과 관련된 신문 기사에서 한국어 전문용어를 태깅하고, 논문의 제목과 초록에서 한영 전문용어 쌍을 태깅하는 작업을 진행하였다.

1. 서론

말뭉치(corpus)는 언어 연구에 있어 언중 즉, 언어 사용자들의 실제 언어 사용을 보여주는 귀중한 경험적 자료로 사용될 수 있다. 특히 전문 분야 말뭉치는 단어의 실제적인 쓰임새를 보여주는 기초 자료이다[5]. 이러한 전문 분야 말뭉치에서 말뭉치에 나타나는 전문용어를 추출할 경우, 전문용어의 생성, 성장, 사멸의 전문용어 라이프 사이클을 관리할 수 있는 정보를 추출할 수 있다.

한국과학기술정보연구원(이하 KISTI)은 과학기술 분야의 신속하고 정확한 정보 서비스를 위해 정보 처리 및 유통 환경을 구축해 오고 있다[3]. 본 논문에서는 KISTI 의 과학기술 분야 말뭉치 구축, 말뭉치 정제 지원 도구(tool) 및 전문용어 추출을 위한 작업 환경인 전문용어 워크벤치 개발에 대하여 기술하고자 한다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련 연구 및 과학기술 말뭉치 구축에 관하여, 3 장에서는 정제 말뭉치를 위한 기반 구조에 관하여, 4 장에서는 웹 기반의 정제 말뭉치 구축을 위한 환경 개발에 관하여, 마지막으로 5 장에서는 결론과 향후 방향에 대하여 기술한다.

2. 관련 연구 및 과학기술 말뭉치 구축

1964 년 Brown Corpus 에서 일반 분야의 말뭉치 구축에 컴퓨터를 이용하면서 말뭉치는 그 유용성을 더욱 인정받게 되었다[1]. 이후 LOB Corpus 나 Birmingham Collection 등의 말뭉치 구축이 이루어 졌고, 국내에서도 연세 말뭉치, 21 세기 세종계획 등에서 말뭉치가 구축되어 오고 있다[6].

이렇게 구축되는 말뭉치는 말뭉치 구성에 대해 해당 분야의 대표성을 고려하여 신중하고 엄격하게 구축되어야 한다. 가장 이상적인 말뭉치는 균형(balance)을 고려한 대용량의 다양한 분야 문서들로부터 적절한 샘플링을 통해서 이루어져야 한다.

그러나 이러한 조건을 만족시켜 말뭉치를 구축하기란 현실적으로 쉽지 않다[2]. 특히 말뭉치에서 추출한

전문용어를 가지고 전문용어 라이프 사이클을 관리할 날짜 정보를 추출하기에는 원시 문서에 날짜 정보가 없어 이용하기가 곤란한 경우가 발생하게 된다.

국내 과학기술 분야 말뭉치의 경우, 문화부의 21 세기 세종계획에서 1998 년부터 계속하여 구축되고 있다. 하지만 저작권과 관련하여 문제가 있어 어렵게 구축된 말뭉치를 배포하기는 어려운 실정이다.

이러한 상황에서 국내 연구보고서, 국내외 학술잡지, 국내 학위논문, 특허 등 과학기술 분야의 지식/정보 인프라를 구축하여 정보 서비스를 하고 있는 KISTI 에서는 이러한 여러 문서들과 신문의 과학기술면 기사를 이용하여 과학기술 분야 말뭉치를 구축하고 있다[4].

그러나 여러 연구보고서, 학술잡지, 학위논문, 신문 등의 문서를 수집 또는 입력을 할 때 문서에는 많은 오류가 있을 수 있다. 정확하게 쓰이지 않은 문서에서 검색 및 가공, 정보 추출을 하게 되면 당연히 원하는 결과를 얻을 수 없다(Garbage-In Garbage-Out). 즉, 문서에서 맞춤법 및 띄어쓰기 오류를 정제한 문서를 구축해야만 정보검색의 효율성 및 정확성을 높일 수 있다.

이를 위해 본 논문에서는 효과적으로 문서를 정제할 수 있는 작업환경인 웹 기반의 정제 말뭉치 워크벤치를 개발하였다. 또한 이렇게 정제된 문서에 포함된 전문용어를 표시해 두면 향후 정보검색의 효율성 향상, 전문용어의 자동추출, 전문용어가 쓰인 문맥의 파악 등에 효과적으로 쓰일 수 있도록 전문용어 추출을 위한 작업 환경도 구축하였다.

3. 정제 말뭉치를 위한 기반 구조

말뭉치를 언어적 관점에서 고찰할 경우 크게, 단일어로 구성되었는가, 이중어로 구성되었는가로 나눌 수 있다. 이중, 두 개 이상의 언어로 구성된 병렬 말뭉치(parallel corpus)는 오늘날 기계 번역, 전문용어 추출, 교육 시스템 기반자료 등에서 유용하게 사용될 수 있다. 본 논문에서는 이러한 상황을 지원하기 위하여 정제 말뭉치를 위한 기반 환경 역시 일반 코퍼스와 병렬 코퍼스로 나누어 기반 구조를 설계하였다.

그러나 병렬 말뭉치의 경우 그 용량 및 문장의 크기에 따라 특성이 또다시 구분되어야 한다. 단일 문장의 경우 번역 등가성이 확보될 수 있으나, 대용량 코퍼스를 고려할 경우 여러 문장 사이의 관계를 설명할 수 있어야 한다. 이 경우 하나의 문장이 다른 언어의 문장 하나와 반드시 번역 등가성을 확보한다는 보장이 없다. 따라서 이러한 문장 구조를 포함할 수 있는 새로운 형태의 문서 구조를 설계해야 하며, 이를 위해 본 논문에서 제시하는 정보 구조는 다음의 세 종류와 같다.

- **단일어 말뭉치 구조:** 단일어로서 다수의 문장과 문단으로 구분된 문서구조. 메타정보 및 전문용어를 지시.
- **병렬 단 문장 말뭉치 구조:** 제목 쌍이나 간단한 용어와 같이, 문장간에 번역이 겹치지 않는 독립된 형태의 단 문장 목록 구조.
- **병렬 복 문장 말뭉치 구조:** 초록 쌍이나 서술형 문서의 단일 문단 내 복수의 문장간 구조를 표현하는 구조.

단일어 말뭉치 구조는 하나의 XML 문서가 다수의 문서를 포함할 수 있도록 구성되어야 한다. 하나의 문서는 하나의 제목과 복수의 문단, 그 문단 내부를 구성하는 복수의 문장 정보를 저장한다. 각 문자는 세분화된 노드로 분리하여 문법정보와 전문용어 정보를 겹치게 표현할 수 있도록 하였다. 다음은 각 요소에 대한 설명이다.

- ARTICLES: 하나의 XML 문서에 포함되는 root element. 1 개 이상의 ARTICLE 요소를 포함한다.
- ARTICLE: 하나의 article 정보. 하나의 article 은 다수의 제목정보 혹은 제목 정보를 포함할 수 있다.
- TITLE: 제목 정보. 하나의 문서는 하나 이상의 제목 정보를 포함한다. 제목정보 혹은, 문단정보 모두 하나 이상의 SEN(sentence) element 를 포함한다.
- PARAGRAPH: 문단정보는 문장 정보를 기본 단위로 하는 element 의 set 으로 구성된다. 따라서 PARAPGRAH element 는 다수의 SEN(sentence) element 를 포함한다.
- SEN: 문장은 문서를 구성하는 기본 요소다. 문장 정보는 문서의 전처리기에 의해 시스템에 의해 지능적으로 분리될 수도 있으며, 사용자 인터페이스를 통해 문장으로 나뉠 수도 있다.
- C: optional selection. 만약 언어처리를 위한 추가적 인터페이스가 요구되는 경우 각기 독립된 문자정보를 저장하는 요소
- TERM: 추출 가능한 terminology 요소의 집합을 나타낸다. 하나 이상의 ITEM element 를 포함한다.
- ITEM: 하나의 전문용어 내에 저장되는 전문용어의 각기 다른 표기법

병렬 단 문장 말뭉치 구조는 다수의 문장 정보만을 저장하도록 구성되어야 한다. 하나의 SEN 정보는 언어정보로 구분된 나열형으로 저장되며, 각 문자는 C 요소를 통해 문법정보와 전문용어 정보 등을 겹쳐서 표현할 수 있도록 구성하였다.

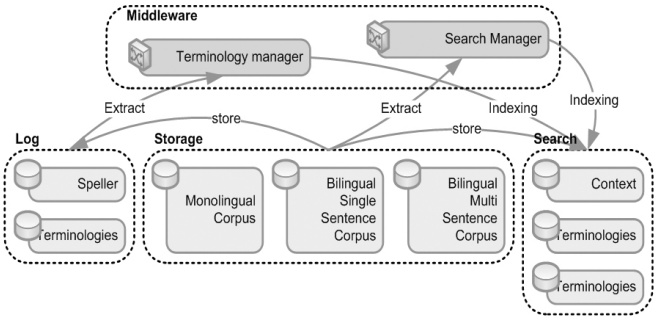
- ARTICLES: 단일어 말뭉치 구조와 동일.
- ARTICLE: 단일어 말뭉치 구조와 동일.
- SEN: 단일어 말뭉치 구조와 동일
- C: 단일어 말뭉치 구조와 동일
- TERM: 단일어 말뭉치 구조와 동일
- ITEM: 단일어 말뭉치 구조와 동일

병렬 복 문장 말뭉치 구조는 단일 문단에 대해 번역 등가성을 표현하는 것으로 전제한다. 따라서 각각의 번역 등가성은 언어정보(lang 과 같은 속성값을 설정)로 구분되는 다수의 문장 정보와 함께 하나의 SET 형태로 저장된다. 각 SET 내부에 저장되는 문장 정보는 나름의 일련번호를 저장하여 번역 등가성 및 순서가 뒤바뀌지 않도록 구성하였다.

- ARTICLES: 병렬 단 문장 말뭉치 구조와 동일.
- ARTICLE: 병렬 단 문장 말뭉치 구조와 동일.

- SET: 하나의 번역 등가성을 가지는 단위. 하나의 SET element 는 다수의 SEN 요소를 포함한다.
- SEN: 병렬 단 문장 말뭉치 구조와 동일.
- C: 병렬 단 문장 말뭉치 구조와 동일.
- TERM: 병렬 단 문장 말뭉치 구조와 동일.
- ITEM: 병렬 단 문장 말뭉치 구조와 동일.

정제 말뭉치를 효율적으로 저장하기 위한 구조를 위해서 본 논문에서는 XML 구조 형식으로 설계하였다. 이는 XML 이 계층화된 정보를 쉽게 표현할 수 있을 뿐만 아니라, 표준화된 형태로 각종 전산 시스템에 쉽게 응용될 수 있기 때문이다. 정제 말뭉치를 위한 기반 환경은 크게 데이터 저장 영역과 데이터를 자동처리할 수 있는 미들웨어 영역으로 구분할 수 있다.



<그림 1> 정제 말뭉치를 위한 데이터 레이어와 미들웨어 레이어

데이터 저장 영역은 크게 저장, 검색, 기록 세 영역으로 나뉜다. 저장은 단일어 말뭉치, 병렬 단 문장 말뭉치, 병렬 복 문장 말뭉치를 XML 형태로 파일 시스템에 저장하며, 검색을 위한 기반자료를 제공하는데 활용된다. 검색은 본문 검색, 메타정보(문서의 분류, 혹은 신문기사의 게재날짜 및 게재일 등) 검색, 전문용어 검색 정보를 저장하며, 저장 정보에 대해 참조할 수 있는 인덱스 테이블을 가진다. 기록 영역은 사용자에 의한 교정 내역 및 전문용어 추출에 대한 단순 기록 정보가 남는다. 검색 영역과 기록 영역의 전문용어 정보가 다른 점은, 기록 영역의 경우 전문용어 발생빈도수 및 참조정보를 찾기 위한 분석 성격이 강한 반면, 검색 영역은 빠른 정보 추출을 위한 성격이 강하다는 점이다.

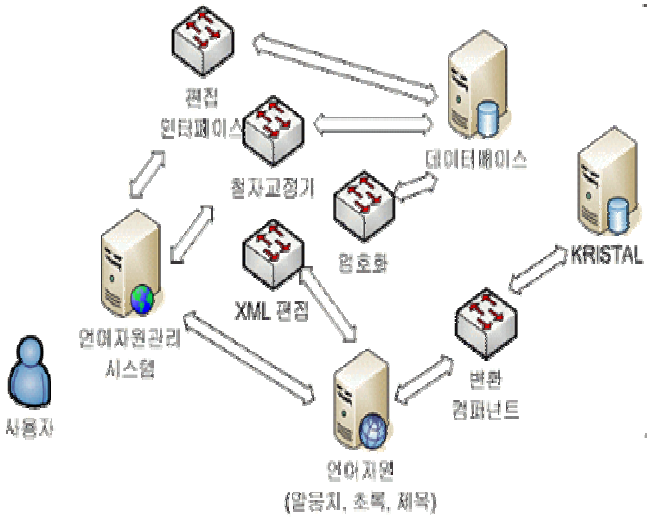
또한 추출되는 전문용어의 관리를 위한 전문용어 관리기(terminology manager)와 문서 작업의 히스토리(history) 관리를 위한 로그 관리기(log manager)는 컴퍼넌트와 데이터베이스 저장 프로시져로 구성되어 각 데이터 레이어 요소간 원활한 정보 유통이 이루어지도록 하였다.

4. 웹기반 정제 말뭉치 구축 환경 개발

본 논문에서 개발한 워크벤치는 다음과 같은 사항을 제공하도록 개발되었다.

- 언어자원관리 시스템의 요구분석, 기존 기술 동향 분석을 바탕으로 시스템을 설계. 분산환경을 고려하여 Miscrosoft.NET 기반으로 개발
- 정보검색관리시스템인 KISTI 의 KRISTAL 과 연동할 수 있는 기능 제공
- 전문가의 개입이 필요한 작업은 분산/협동에 의한 작업이 가능하도록 구축

<그림 2>는 전체 시스템 구성도이다.



<그림 2> 전체 시스템 구성도

4.1 언어자원 구축용 통합 편집기 개발

- **일반 문서 정제 환경 구축:** 철자교정기, 전문용어 관리기와 연동하여 띄어쓰기 및 철자법을 교정하고 전문용어를 marking 할 수 있는 환경.
- **병렬 말뭉치 정제 환경 구축:** 병렬 말뭉치의 가장 큰 어려움은 80%에 달하는 1 대 1 문장 대역이 아니라 1 대 n 혹은 n 대 1 의 문장 대역이다. 이러한 문제를 효율적으로 관리하고 정제할 수 있는 환경.
- **전문용어 markup 환경 구축:** 전문용어 markup 인터페이스는 문서정제 환경뿐만 아니라 병렬 말뭉치 환경에도 동일하게 적용될 수 있도록 표준형 인터페이스.

4.2 언어자원 구축 및 서비스용 통합 뷰어 개발

- **정제 문서 뷰어:** 과학분야 정제된 문서를 브라우징, 검색 환경과 연동된 사용자 인터페이스.
- **전문용어 기반 정보 뷰어:** 전문용어의 사용주기 및 예문 정보를 검색할 수 있는 인터페이스.
- **병렬 말뭉치 기반 뷰어:** 전문용어와 연동된 병렬 말뭉치 환경을 브라우징 할 수 있는 환경.

4.3 정보검색 및 관리 시스템 개발

- **정보저장 및 관리:** 효율적인 정보 저장이 가능하도록 인덱싱 파일 생성에 적합한 형태로 미들웨어 구축.
- **정보검색기능:** 문장 및 병렬 말뭉치 검색이 가능한 미들웨어 환경 구축(뷰어와 연동).
- **자료 수집 환경:** 특정 정보 추출 및 저장을 위한 미들웨어 및 API 환경 구축.

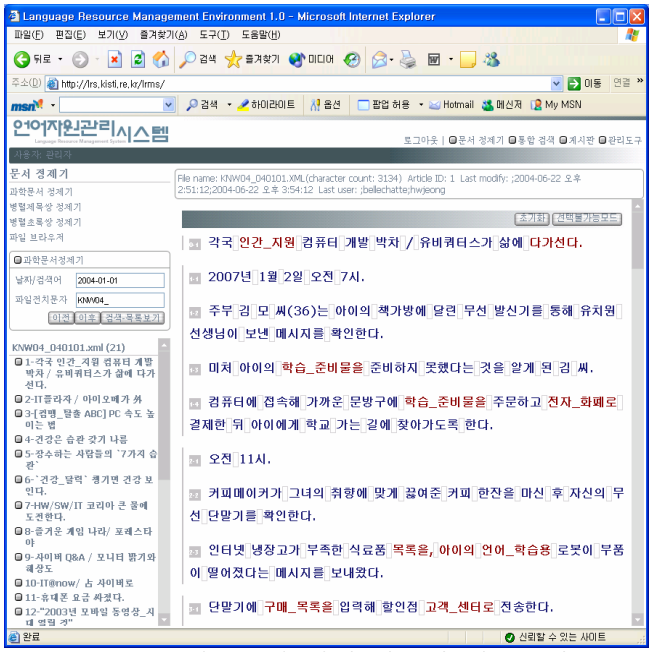
4.4 언어자원 관리용 형태소 분석기 작업

- **API 호환용 COM component 구축:** 언어자원 관리용 형태소 분석기는 정보의 입력 및 출력이 컴퍼넌트 형태로 구축되도록 하여 웹서버 및 일반 어플리케이션 형태에서도 손쉽게 연동되도록 구축.
- **multiple scanning 이 가능하도록 지원:** parser 에서 진행하는 한 번의 철자교정 및 띄어쓰기 구성이 아니라 통계적 기법에 기준한 guessing 알고리즘을 적용하여 다중 철자법 검색 및 적용이 가능.
- **반환 정보의 프로토콜 기반 구조화:** API 환경에 의해 전송되는 자료뿐만 아니라 인터넷 환경에 적합하게 연동되도록 프로토콜 정보를 정교화, 계층 구조로 구성.

4.5 언어정제 및 가공을 위한 미들웨어군

- **원시자료 처리기:** 원시 말뭉치 정보가 구축될 사용자 인터페이스에 적합하도록 구성하고 연동.
- **전문용어 관리기:** 전문용어의 범주를 관리하고, 이를 일반 문서 정보에 형태소 정보를 고려하여 연동하는 인터페이스.
- **철자 교정기 연동:** 철자교정기를 바탕으로 하여 띄어쓰기 및 잘못된 한글 병기, 외래어 병기 정보를 교정하고 이를 사용자 인터페이스와 연동하여 구성.
- **용어추출 및 관리기:** 기존 전문용어 라이브러리와 말뭉치 정보 및 철자 검색기 라이브러리가 함께 연동되어 구동될 수 있도록 시스템 환경 및 기반 구조를 구성.

<그림 3>은 이와 같은 점을 고려하여 개발된 전문용어 정제 말뭉치 워크벤치의 예이다.



<그림 3> 전문용어 정제 말뭉치 워크벤치

5. 결론

지금까지 본 논문에서는 문서를 정제할 수 있는 작업환경인 웹 기반의 정제 말뭉치 워크벤치 개발과 정제된 문서에 포함된 과학기술 전문용어를 표시할 수 있게 하는 작업 환경 구축에 대하여 알아 보았다.

향후 전문용어를 자동으로 추출할 수 있는 연구가 더욱 필요하며 전문용어 평가 및 검증에 관한 체계의 확립이 절실히 필요하다.

참고문헌

[1] 곽용진, “합목적적 말뭉치(Corpus) 자동 구축”, 언어정보와 사전편찬, 한국문화사, pp.31-68, 2003.

[2] 배희숙, 백혜승, 김재호, 서충원, 최기선, “코퍼스 분석을 통한 분야 특성 서술어 선정과 용어의 의미제약 연구”, 전문용어연구 4, pp.53-69, 2002.

[3] 이병희, 박동인, 류범종, “과학기술 분야 언어자원 구축 기반 조성에 관한 연구”, HCI 2004 논문 발표집, pp.298-303, 2004.

[4] 이병희, 정휘웅, 윤애선, “국내 주요 일간지의 과학기술 기사와 전문용어”, The 6th East Asia Forum on Terminology Proceedings, pp.59-65, 2003.

[5] 조은경, 서상규, “전문용어와 전문언어 말뭉치”, 전문용어연구 2, 한국문화사, pp.201-229, 2000.

[6] 최기선, 2004 년도 21 세기 세종계획 전문용어의 정비 중간보고서, 2004.