감성과학, Vol. 7, No. 2, pp. 133-140, June 2004

감성공학 문서 데이터의 지표 자동화를 위한 코퍼스 분석 기반 특성정보 추출

Extraction of Informative Features for Automatic Indexation of Human Sensibility Ergonomic Documents

배희숙*† · 곽현민** · 채균식** · 이상태**

Hee-Sook Bae*† · Hyun-Min Kwak** · Kyun-Sik Chae** · Sang-Tae Lee**

한국과학기술원 전문용어언어공학연구센터*

Korea Terminology Research Center for Language and Knowledge Engineering, KAIST

한국표준과학연구원 정보전산그룹**

Technical Information & Computing Group, KRISS

Abstract : A large number of indices are produced from human sensibility ergonomic data, which are accumulated by the project "Study on the Development of Web-Based Database System of Human Sensibility and its Support". Since the research in this field will be increased rapidly, it is necessary to automate the index processing of human sensibility ergonomic data. From the similarity between indexation and summarization, we propose the automation of this process. In this paper, we study on extraction of keywords, information types and expression features that are considered as basic elements of following techniques for automatic summarization: classification of documents, extraction of information types and linguistic features. This study can be applied to automatic summarization system and knowledge management system in the domain of human sensibility ergonomics.

Key words : human sensibility ergonomics, automatic indexation system, automatic summarization, information types, analysis of corp

요약 : 최근 대량으로 쏟아지는 감성공학 연구 결과와 논문들을 가치 있는 자료로 만들기 위해서는 감성 데이터가 산업 전반에 활용될 수 있도록 지표로 정리해야 한다. 본 논문에서는 "웹기반 감성 데이터 베이스 구축 및 보급에 관한 연구" 과제를 통해 작성된 감성 데이터 지표에 입각해서 앞으로 대량으로 출현할 감성공학 데이터의 지속적인 지표화를 위한 과정의 자동화를 제안한다. 문서 데이터의 지표화 작업이 자동요약과 유사하다는 점에 착안하여 자동지표화 시스템을 위한 기술들의 기초가 되는 정보유형 및 주요어 추출, 특성표현을 통한 정보문 추출에 대해 감성공학 코퍼스 분석을 통해 연구하고자 한다. 이는 감성공학 분야에서의 지식관리 시스템이나 자동요약 시스템에 활용될 수 있다.

주제어 : 감성공학, 자동지표화 시스템, 자동요약, 정보유형 추출, 코퍼스 분석

† 교신저자 : 배희숙(한국과학기술원 전문용어언어공학연구센터)

E-mail : elle@world.kaist.ac.kr

TEL : 042-869-8794

1. 서론

산업 전반에 인간의 감성을 반영하려는 요구가 높아지면서 감성공학 연구가 활발해지고 관련 연구도 쏟아지고 있다. 연구 결과로 얻어지는 많은 원시자료, 보고서, 논문들을 체계화 하여 실제로 산업 사회 전반에 가치 있는 정보로 활용하기 위해서는 전문적인 감성 데이터들을 지표로 정리하여야 한다.

지표란 원래 기호학적 용어로 Prieto는 지표를 "즉각적으로 지각 가능하지 않은 어떤 사실에 대하여 무엇인가를 알려주는 즉각적으로 지각 가능한 사실"이라 정의한다. 연기가 나면 불이 나고 있음을 알 수 있듯이 연기라는 결정적 지표를 통해 불에 대한 정보를 얻을 수 있다. 이와 같이 감성공학 문서 데이터의 지표란 해당 문서에 대해 정확하고 응집된 지식을 제공할 수 있는 정보 덩어리이다.

본 논문에서는 앞으로 대량으로 출현할 감성공학 데이터의 지속적인 지표화를 위해 과정의 자동화를 제안한다. 여기서 제안하는 지표자동화 시스템을 위해서는 문서의 자동 분류, 특성언어 추출, 정보유형 추출 및 문장 재구성이라는 여러 단계의 기술이 필요하며, 이 기술들의 기초가 되는 주요어, 정보유형, 정보문의 언어특성 파악과 같은 기초 과정을 다룰 것이다. 이를 위해 전문적으로 편집 작성된 지표와 해당 논문을 비교분석하고, 각 특성 정보들이 원문의 어느 위치에 어떤 표현으로 제공되는지 코퍼스 분석을 통해 알아본다. 이러한 특성정보가 지표 작성의 자동화를 위한 알고리즘으로 이어진다면 감성공학 분야 지식 관리의 한 단계로 간주될 수 있으며, 지표자동화뿐만 아니라 자동요약 시스템에도 활용될 수 있다.

2. 코퍼스

2.1 코퍼스 기반 연구

본 연구는 감성공학 코퍼스를 구성하고 이를 분석함으로써 코퍼스로부터 특성언어에 대한 정보를 추출하고 정보의 유형을 추출하여 문서의 지표화에 활용한다. 자연언어처리 분야에서 코퍼스 분석을 통한 각종 언어 정보를 활용하는 연구들이 많이 있으나 정보추출 및 텍스트 요약 시스템을 위해 코퍼스로부터 정보유형을 추출하여 활용하는 방법은 Saggion[6]의 연구에서 착안한 것이다. 그러나 감성공학 DB 구축 과정에서 지표화 작업을 수행하면서 문서요약과 지표자동화의 유사성을 발견하고 자동요약을 위해 [5]에서 적용하였던 방법을 활용하고자 한 것은 본 연구가 처음이다.

2.2 코퍼스 구성

코퍼스는 두 종류로 구성되었다. 표 1과 같이 보고서를 시각, 청각, 후각, 촉각이라는 네 가지 감각의 범주로 분류하여 구성한 코퍼스 A는 포괄적 정보유형에 대한 단서를 제공할 것이다.

표 1. 코퍼스 A 구성

감각	보고서명
시각	"색/조명환경 제시기술개발에 관한 연구", "색채감성을 적용한 디지털카메라 개발"
청각	"감성 인식 시스템을 위한 음성 DB 구축에 관한 연구", "음향·진동 환경 제시 기술"
촉각	"촉각 측정 및 질감 제시 기술 개발", "피부감각의 감성측정 기술 및 DB 개발"
후각	"후각/미각 감성 측정 기술 및 DB 개발", "후각환경 제시 기술 개발"

코퍼스 A는 한국과학기술원 형태소 해석기를 통하여 형태소 분석 및 품사 관련 홈페이지 http://morph.kaist.ac.kr/~morph에서 제공되는 형태소 분석 시스템을 통해 분석결과를 메일로 제공받을 수 있다. 분석된 자료를 다시 빈도순으로 정렬함으로써 감각별로 분류된 부분 코퍼스의 주요어와 정보유형 을 추출하였다. 고빈도어와 기존의 수동 지표화된 결과가 비교되면서 정보유형을 정리하였다.

한편, 구체적 정보유형과 그 특성을 추출하기 위해서 이미 지표화된 10편의 논문과 해당 지표에 의

해 코퍼스 B를 구성하였다. 이 코퍼스는 각 정보유형에 해당하는 문장들이 원문의 어느 위치에서 어떤 언어로 제시되는지 파악하는 데에 쓰일 것이다. 논문 목록은 표 2와 같다.

표 2. 코퍼스 B 구성

보고서	보고서명
1	정신지체장애아동의 기본형태와 제품형태에 대한 인지
2	생리량과 주관량에 대한 상관조사 시스템의 개발
3	디자인 과정에서 사고의 속성 파악
4	항등사상 모델을 응용한 다양한 해석 - 자동차 configuration 결정 지원 시스템의 구축
5	인간과 기계의 인터페이스와 디자인 요소와의 관계
6	실험도구에 의한 창조적 문제 해결과정 내용에 관한 연구
7	도구가 실험 참가자에게 주는 영향에 대한 정량적 연구
8	향 감각량 평가에 알맞은 흡착막 선택과 뉴럴 네트워크에 의한 인식
9	FFTA를 응용한 사무실 평가법에 대한 연구와 시각적인 면으로 본 연구
10	도시 경관의 색채 이미지 컨트롤을 위한 연구

3. 방법과 과정

그림 1은 연구의 전체적 흐름도이다. 점선 하단은 향후 연구가 될 것이다.

3.1 정보유형 및 주요어 추출

활용 가능한 자료의 축적, 그리고 축적된 자료의 활용 가능한 자료로의 전환은 정보 교환 측면에서 매우 중요하다. 문제는 전문적 논문들로부터 유용한 정보를 찾아내는 방법이다.

직관적으로 문서의 주요 정보유형은 '왜', '무엇을', '어떻게', 그리고 기타 참조 사항이다. 코퍼스 분석을 통해 이러한 직관이 실제 문서에서 어떤 유형으로 처리되고 있는지 알아보자.

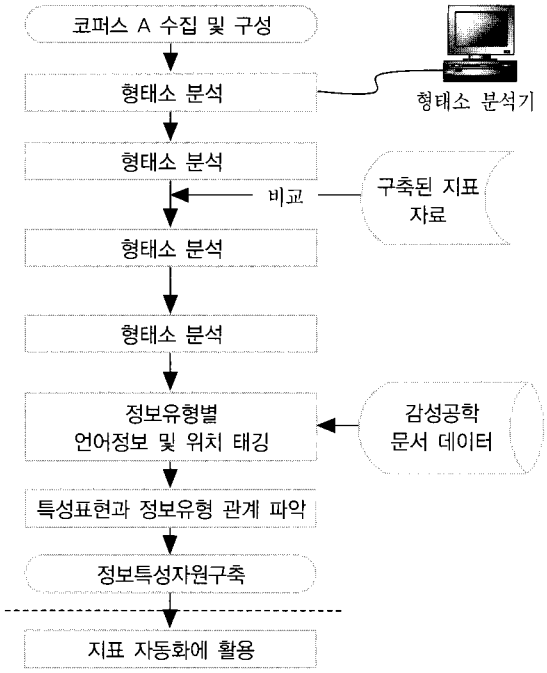

그림 1. 전체적 흐름도

3.1.1 어휘분포를 통한 주요어 추출

코퍼스 A를 한국과학기술원 형태소 분석기에 의해 분석하고 이 결과에 대한 형태소별 경우의 수, 상대빈도, 누적 상대빈도 등의 통계 처리를 하였다.

표 3. 기존 자료에서 주요어와 빈도와의 관계 파악

분야	누적 빈도 50% 이상의 실질어
시각	이다, 영상, 광원, 색채, 있다, 위하다, 대하다, 추정하다, 분광하다, 감성, 하다, 모니터하다, 값, 이용하다, 분포하다, 연구하다, 주위, 분석하다
청각	이다, 감정, 음성, 인식, 결과, 감성, 이용하다, 평가하다, 연구하다, 문장, 주관하다, 특징, 표, 위하다, 종류, 패턴, 대하다, 같다, 구축하다, 피치, 시스템, DB, 개발하다, 따르다, 지연되다
촉각	측정, 감성, 시스템, 직물, 연구, 표면, 특성, 질감, 대하다, 위하다, 개발, 제품, 보다
후각	감성, 향, 후각, 대하다, 있다, 분석하다, 측정하다, 위하다, 주관, 자극하다, 그림, 형용사, 연구, 지표, 개발, 경우, 감각, 평가
전체	있다, 감성, 대하다, 하다, 측정하다, 위하다, 연구, 분석, 이용, 결과, 향, 개발, 후각, 시스템, 기술, 평가, 영상, 사용, 주관, 광원, 보다, 따르다, 색채, 같다, 경우, 의하다

분야별 주요어가 대체로 누적빈도 50% 선에서 드러나는 것을 볼 수 있다[4].

표 4. 명사 누적빈도 상위 50%의 명사

시각	색채, 광원, 색, 영상, 주위
청각	감정, 감성, 음성, 인식, 피치, 문장, 종류, 패턴, 시스템
촉각	감성, 직물, 표면, 질감, 시스템, 특성, 제품
후각	감성, 향, 후각, 자극, 형용사, 지표, 감각

명사만 추려 보면 표 4와 같다. 이중에서 전체적으로 나타나는 감성, 시스템 등의 명사를 제거하면 문서의 주요어가 드러난다. 좀더 구체적인 사항은 3.2.1장에서 확인할 수 있다.

3.1.2 정보유형 추출

정확한 처리를 요구하는 지표화 작업을 자동화 한다는 것은 매우 어려운 일이다. 문서의 높은 전문성은 이러한 어려움을 배가시킨다. 까다로운 지표 자동화를 위해서는 주요 정보유형을 추출해야 하는데, 이를 위해 본 연구는 상위빈도 실질어를 기준으로 하여 유형의 범위를 좁힐 것을 제안한다. 표 5는 보고서 전체를 모아 분석한 결과에서 누적빈도 50% 이상에 있는 모든 실질어이다.

표 5. 누적빈도 50% 이상 실질어

서술어	있다, 대하다, 하다, 측정하다, 위하다, 연구하다, 분석하다, 이용하다, 평가하다, 사용하다, 주관하다, 따르다, 같다, 의하다
명사	결과, 향, 개발, 후각, 시스템, 기술, 영상, 광원, 색채, 경우

표 5에서 제시되는 어휘를 기반으로 다음의 정보유형을 도출할 수 있다.

표 6. 고빈도 실질어로부터 도출된 정보

고빈도 실질어	정보유형
있다, 이다	현상기술
위하다, 측정하다, 분석하다, 연구하다, 조사하다	연구목적
-을 이용하다, 사용하다, 의하다	연구방법
결과, 평가하다	연구결과

네 가지 정보유형 중에서 《현상기술》은 d보고서 전체에서는 고빈도로 나타나지만 감각별로 분류된 보고서 그룹에서는 불규칙적이다. 개별 보고서에도 변함없이 나타나는 정보유형은 목적, 방법, 결과에 해당하는 표현들이다.

표 7. 지표의 정보유형

주요어
설명
연구방법
소스데이터
활용분야
참고문헌
출처

한편, "웹기반 감성 데이터베이스 구축 및 보급에 관한 연구" 과제에서 주관적으로 정리한 주요 정보유형을 보면 표 7과 같이 일곱 가지이다.[2] 일곱 가지 정보유형 중에서 참고문헌, 출처, 소스데이터는 원문서로부터 바로 자동정렬이 가능하다. 정확한 자동 처리가 어려운 것은 《설명》, 《연구방법》, 그리고 《주요어》이다. 빈도 기준으로 추출된 정보유형과 비교하면 과제의 설명은 요약이며, 이 요약 부분에 《연구목적》, 《연구방법》, 《연구결과》가 포함된다. 《주요어》는 고빈도어나 제목(부제목)을 통해 추출할 수 있을 것이다.

3.1.3 정보유형별 특성 표현과 위치 파악

정보유형은 코퍼스 내 어휘 분포를 기준으로 추출되었다. 이제 각 정보유형에 해당하는 특성을 찾아내야 한다. 문제는 문서의 어느 위치에 정보특성이 제공되는지를 알아내는 일이다. 전문성이 매우 높은 특정 분야 논문들로부터 특성정보의 위치를 알아 내기 위하여 본 논문에서는 이미 과제로 축적된 지표와 논문을 대조함으로써 보편성을 이끌고자 하였다. 이를 위하여 한국표준과학연구원 인간공학그룹에서 수행한 "웹기반 감성 DB 구축 및 보급에 관한 연구"의 결과물로 축적된 열 개의 감성공학 분야 문서와 그 지표를 수동으로 분석하여 위에서 추출된 정보유형을 찾아내고 각 정보유형이 문서의 어느 위치에 있는지 표 8과 같이 일일이 표시하였다.

표 8. 정보유형에 해당하는 문장과 위치 표시 예

해당원문	원문위치	언어특성/정보유형
본 연구의 목적은 정신지체 장애 아동의 적응성, 지능, 사물의 기본 조형 인지력 등을 파악하는 것이다.	서론 3rd	본 연구의 목적은~, ~것이다. (목적)
기본 형태와 복합 형태의 제품을 대조시켜 복합 형태 제품에 해당하는 기본 형태를 피험자에게 질문하고 대답을 기록, 분석한다.	본문 3.2	실험방법 (방법)
본 연구의 조사 결과를 정리하면 다음과 같다.	결론	결과를 정리하면~ (결과)
이러한 배경으로부터 본 연구에서는 공업디자인에 있어서 디자이너의 사고과정을 디자인 실험으로 관찰 분석하여 그 속에 존재하는 보편적인 사고 특성을 밝히는 것을 목적으로 하였다.	서론 last	~을 목적으로 하였다.
디자이너의 사고과정을 관찰 기록하기 위해 프로토콜법을 사용하였다.	본문 2.2	-위해서, ~법을 썼다.
디자인 과정이 문제의 명확화, 아이디어 전개, 제안 평가라는 3단계로 구성된다는 것을 확인하였다.	본문 4	-을 확인하였다.
그래서 본 연구에서는 ( ) 최적해를 탐색하는 역추론 시스템을 개발하였다.	서론 last 1	본 연구에서는~
뉴럴 네트워크의 일종인 항등사상 모델을 이용하여 교사 신호에 대한 정보를 압축하고 그 압축된 해공간을 시각화하여...	서론 last 1	~을 이용하여, 시각화 하여
그 결과, 본 시스템에 의해 다양해가 탐색 가능하다는 것, 컨셉 항목과 configuration 항목 간의 교사 신호의 전체적인 경향에 있어서 항등사상 모델에 의해 정보가 압축되는 것을 추측할 수 있었다.	서론 last	그 결과
본 연구에 의해 다음과 같은 것이 확인되었다.	결론	~이 확인되었다.
연구 목적은 MMI 개발의 과정에서 디자인 결정을 하는 도중에 발생하는 다채로운 선택 부분에 대한 최적해를 얻기 위해 ( ) 장치에 대한 구체화를 꾀하는 것에 있다.	서론 2nd	연구목적은~
MMI 개발 지원을 위해 인체 모델의 생성을 고찰하여 컴퓨터 시뮬레이션에 의해 다섯 가지 주요 개념 중 네 번째 개념에 대한 개발지원 시스템에 대해 검토하였다.	서론 last	~을 위해, ~에 의해, ~을 검토하였다.
본 연구에 의해 얻어진 내용을 요약하면 (…)이 확인되었다.	결론 2nd	확인되었다.

표를 관찰하면, 정보유형에 직접적으로 해당하는 30개 문장 중 77%가 서론과 결론에 위치하고 있다. 본문 내에 위치한 경우에는 제목으로 정보유형에 해당하는 표현을 내포하고 있었다. 따라서 요약을 위해서는 서론과 결론으로부터 각 정보유형에 해당하는 문장이 포함하는 언어표현을 찾아 탐색하는 방법을 사용할 수 있을 것이다.

정보가 서론과 결론이 아닌 본론에 위치할 경우는 주로《방법》이었으며, 이 정보유형에 해당하는 정보문의 위치는 특성표현을 통해 자동 탐색할 수 있다. 특성표현에 의해 일차적으로 걸러진 문장들은 다시 서술어와 논항의 언어 관계를 통해 가장 적합한 문장 추출로 이어질 것이다.

지금까지 고빈도 어휘 분포에 의해 정보유형을 추출하고, 정보유형별 표현 특성과 원문에서의 위치를 수동으로 표시함으로써 정보유형과 언어 특성, 그리고 그 관계를 찾아보았다. 표 9에서 그 결과를 정리한 것이다.

표 9. 정보유형과 특성표

정보유형	특성표현
연구목적	〈-의 목적은 이다〉, 〈-을 목적으로 한다〉, 〈본 연구에서는 -을 개발하였다〉, 〈-을 얻고자 한다〉, 〈-이 필요하다〉, 〈-을 제안한다〉
연구방법	〈-을 이용하다〉, 〈-을 사용하다〉, 〈-에 의하다〉, 〈-으로써〉, 〈-법을 쓰다〉
연구결과	〈-결과, -을 밝히게 되었다〉, 〈-이 가능하게 되었다〉, 〈-을 구축하였다〉 〈-을 확인하였다〉, 〈이상의 결과로부터 -이 고찰되었다〉, 〈결과를 정리하면, -〉, 〈-이 얻어졌다〉

지표 자동화를 위해 코퍼스로부터 추출한 정보유형과 특성표현이 얼마나 유용한지, 주요어가 고빈

도 어휘나 제목에 얼마나 분포하고 있는지 제안된 방법의 타당성을 알아보자.

3.2 실험

좀더 방대한 양의 코퍼스를 토대로 타당성과 보편성에 대한 실험을 해야 하지만 본 논문에서는 기초적인 두 가지 실험에 만족하였다. 주요어가 고빈도 어휘로부터 구성 가능한지 알아보기 위해 논문 제목, 고빈도어, 지표 작성자가 수동으로 정리한 주요어를 비교하였다. 또한 정보를 포함하고 있는 문장을 원문에서 자동으로 추출하기 위해 특성표현을 통해 정보문이 얼마나 걸러지는지 조사하였다.

3.2.1 주요어 매칭율

다음 표는 고빈도어와 제목에서의 주요어 매칭율을 조사한 결과이다.

표 10. 주요어 점유율 〈실험 1〉

논문번호	고빈도어	제목
1	0.750	0.750
2	0.857	0.444
3	0.833	0.571
4	0.500	0.375
5	0.889	0.625
6	0.250	0.500
7	0.400	0.600
8	0.800	0.667
9	0.333	0.667
10	0.571	0.667

전문 지표 작성자가 뽑아 놓은 주요어 목록을 기준으로 논문의 고빈도어 목록과 제목을 각각 탐색하여 얻은 매칭율이다. 프로그램이, "tool"이 "툴"로 되어 있거나 "Man Machine Interface"가 "MMI"로 표시된 경우를 인지하지 못하여 생긴 오류들은 수동으로 후처리하였다. 이는 본 연구의 코퍼스의 양이 적기 때문에 가능한 것이며 이후 연구에서는 후처리까지 자동화하기 위한 알고리즘을 짜야 한다. 본 논문 과정에서 발견하는 문제점들이 바로 이 후처리 알고리즘 구성에 활용될 것이다.

고빈도어 매칭율은 총 6.183이고 제목 매칭율은 5.866으로 고빈도어에서 조금 더 높게 나타났다. 또한 고빈도어에서 매칭되지 않는 어휘가 제목으로 보완될 수 있을지 조사하였으나 비매칭된 어휘 중 단 세 가지만이 제목에서 제시되었다. 제목이나 고빈도어에서 제시되지 않은 나머지 주요어는 "발화사고법", "역문제", "평정값", "프로토콜법", "발화량", "퍼지" 등과 같이 실험방법에 저빈도로 쓰인 구체적이고 전문적 용어이거나 "기호성", "환경", "감각검사", "스케일", "컨트롤" 등과 같이 일반적 어휘들이었다. 이 사실에서 출발하여 나머지 누락된 주요어들에 대한 보완이 이루어져야 할 것이다.

3.2.2 정보문장 추출

각 정보유형에 적합한 정보문을 특성표현을 통해 얼마나 추출할 수 있을까? 이를 알아 보기 위해 두 번째 실험을 하였다. 정보유형 《연구목적》, 《연구방법》, 《연구결과》 중에서 서론과 결론에서 다루어지지 않고 본론 속에 제시되는 주 정보유형이 《연구방법》이라는 사실에 근거하여 이를 실험 대상으로 삼았다.

용례추출기와 같이 특성표현을 조건으로 넣어 문서들을 탐색하고, 특성언어를 기준으로 앞뒤 5 어절씩을 뽑았다. 실험적으로 세 개의 문서에서 정보유형 《방법》에 해당하는 특성표현이 들어 있는 문장들을 모두 추출하여 관찰한 결과 각 특성표현에 동일 문장이 중복되어 나타났다. 이러한 문제점을 보완하기 위해 특성표현 중에서 세 가지 문서에 모두 나타났던 "방법", "으로써", "이용", "의하다"만을 가지고 다시 실험하였다.

실험 대상 문서로는 표현특성을 추출하는 데 사용한 코퍼스(B)와 임의의 여섯 편의 논문으로 구성한 실험코퍼스를 사용하였다.

하나의 특성표현에 의해 추출된 정보문 중에서 지표에 제시된 방법을 기술하는 문장 정보가 들어

있으면 성공으로 처리하고 그렇지 않으면 실패로 처리하였다.

표 11. 특성 표현을 통한 정보문 추출 성공률

특성언어	해당문장의 적합성	
	코퍼스 B	실험코퍼스
방법	1.000	0.667
-으로써	0.200	0.667
이용	0.600	0.833
의하다	0.900	0.667

이 마지막 도표가 보여주는 의미의 미약함을 인정해야 할 것이다. 사실, 특성표현에 의해 추출된 문장 중에 적합한 정보를 담은 문장이 있는지 알아내는 기초적 확인일 뿐이기 때문이다. 그러나 단 네 개 표현으로 이끌어낸 결과임을 감안한다면 방법의 효과는 주목할 만하다.

4. 결론 및 향후 연구

지금까지 감성공학 문서 데이터의 지표 작성을 자동화 하기 위하여 코퍼스 분석을 통한 특성 정보 추출 가능성에 대해 조사하였다. 비교적 긴 문서인 감성공학 과제 보고서로 코퍼스 A를 구성하여 분석함으로써 감성공학 전문분야 문서의 정보유형을 추출하였다. 10페이지 내외의 짧은 감성공학 관련 논문으로 구성한 코퍼스 B의 분석은 정보유형을 구체화하게 하였으며, 나아가 정보유형별 특성표현과 원문 내에서의 위치를 파악하도록 하였다. 정보유형에 해당하는 문장들은 77%가 서론과 결론에 분포하였고 나머지 23%만 본문에서 제시되었는데, 이들 대부분은 방법에 대한 기술이었다.

코퍼스 분석을 통해 얻은 결과를 토대로 정보유형별 특성표현을 분류하였으며, 특성표현을 통한 정보문 추출의 타당성을 살펴보기 위해 논문의 본론에 잘 나타나는 《방법》을 대상으로 실험하였다. 방법 기술 문장들을 특성표현을 통해 추출하고 전문가가 구성한 지표의 방법 기술과 비교하였다.

논문의 핵심 정보가 되는 주요어의 경우, 지표 작성자가 선별한 주요어의 62%가 논문의 고빈도 어휘로 구성되어 있었다. 제목과의 비교에서는 주요어의 59%가 제목을 구성하는 명사와 일치하였다. 고빈도어와 제목이 커버하지 못하는 어휘들은 주로 구체적인 방법을 기술하는 고도의 전문적 용어이거나 지나치게 일반적인 어휘였다. 이러한 경우에는 요약문과 《방법》 정보문을 통해 해결 방안을 모색할 수 있을 것이다.

주요 정보유형, 정보유형별 특성 표현, 주요어 추출에 대한 방법상의 타당성 조사는 코퍼스 특성 정보에 입각한 지표 자동화의 가능성을 확인하게 하였다.

그러나 도표와 그림에 대한 설명 처리에 대한 보완 연구가 필요하며, 제안된 방법의 보편성에 대한 실험과 연구도 보완되어야 한다. 비정보문 대비 정보문의 비율에 대한 조사와 함께 훈련 코퍼스의 크기도 늘려야 할 것이다.

이러한 취약점들을 점진적으로 보완한다면 지표 자동화에 직접적으로 활용될 수 있으며 일반 분야 및 전문 분야 문서요약 시스템, 혹은 지식관리 시스템에 중요한 자료로 활용될 수 있을 것이다. 또한 코퍼스 기반 특성정보의 축적은 그 자체로서도 가치 있는 것으로 특성정보의 전문 분야별 가중치 결정에 기여할 것이다.

참고문헌

[1] 곽현민, 조해성, 이상태 (2002). 웹 기반 감성 지표 DB 구축에 관한 연구, 한국감성과학회, 2002 춘계학술대회 및 한일 국제 감성공학 심포지움 논문집, 79-85.

[2] 김진호, 이동춘, 박민용, 임좌상, 박수찬, 윤정선, 임현균, 김경택 (2001). 웹 기반 감성지표 개발 및 보급에 관한 연구, 한국감성과학회 학술대회 논문집.

[3] 박길환, 임은영, 박민용 (2001). 웹 기반 감성 데

이터베이스 구축을 위한 사용성 관련감성 지표 개발, 감성과학회 학술대회논문집.

[4] 배희숙 외 (2003). 문서 자동요약을 위한 말뭉치 기반 언어정보 추출, 계량언어학 2집, 도서출판 박이정, 35-52.

[5] Bae, H. S. (1997). "Structures lexicales, syntaxiques et phontiques dans deux pices de J. Tardieu", Thse de Doctorat, Universit de Strasbourg.

[6] Bae Hee-Sook, Paik Haeseung, Seo Chung-Won, Kim Jae-Ho, & Choi Key-Sun (2002). On the Semantic Constraints of Terms through Characteristic Predicates Selection in Domain-specific Corpus, TKE (Terminology and Knowledge Engineering) International Conference.

[7] Dragomir R. Radev, & Kathleen R. Eckeown. (1998). Generating natural language summaries. In Computational Linguistics, 24(3): 469-500.

[8] Prieo (1968). Semioloie, dans Le Langage, La Plade.

[9] Saggion, H., & Lapalme, G., (2000). Where does Information come from Corpus Analysis for Automatic Abstracting, Proceedings of RALI, Canada.

[10] Saggion, H., & Lapalme, G. (2000). Concept Identification and Presentation in the Context of Technical Text Summarization. Proceedings of NAACL-ANLP2000 Automatic Summarization Workshop. Canada.

[11] Wright, S., & Budin, G. (1997). Handbook of Terminology Management. Vol. I. John Benjamins Publishing Company. Amsterdam /Philadelphia.