자동 지표화를 위한 감성공학 분야 코퍼스 분석
- 전문적 문서의 특성 정보 추출 -

배희숙, 김관웅, 곽현민, 이상태
{elle, kkw, hyenmin, stlee}@kriss.re.kr
한국표준과학연구원 인간정보그룹

Analysis of Human Sensibility Ergonomic Corpora for Automatic Indexation - Extraction of informative features -

Bae Hee-Sook, Kim Kwan-Wung, Kwak Hyun-min, Lee Sang-Tai
Korea Research Institute of Standards and Science, Human Information Group

Abstract

본 논문은 감성공학 데이터의 지속적인 지표화를 위해 과정의 자동화를 제안하며 자동 지표화가 문서의 자동 요약과 유사하다는 점에 착안하여 문서 자동분류, 정보유형 추출, 특성언어 추출 및 문장 재구성이라는 단계별 기술의 기초가 되는 정보유형 및 핵심어, 그리고 특성표현을 통한 정보문 추출 방법에 대해 연구하였다. 감성공학 코퍼스 분석을 통한 본 연구는 감성공학 분야에서의 지식 관리 시스템과 자동 요약 시스템에 활용될 수 있다.

Keywords: human sensibility ergonomics, automatic indexation, analysis of corpus, informative features

1. 서론

감성공학 연구 결과로 얻어지는 많은 원시자료(raw data), 보고서, 논문들을 체계화 하여 실제로 산업 사회 전반에 가치 있는 정보로 활용하기 위해서는 전문적인 감성 데이터들을 지표[1]로 정리하여야 한다.

본 논문에서는 앞으로 대량으로 출현할 감성공학 데이터의 지속적인 지표화를 위해 과정의 자동화를 제안한다. 이를 위해 문서의 자동 분류, 특성언어 추출, 정보유형 추출 및 문장 재구성이라는 여러 단계의 기술의 기초가 되는 핵심어와 주요 정보유형, 그리고 정보유형의 언어특성에 대해 연구하고자 한다. 전문적으로 편집 작성된 지표와 원문서인 논문들을 비교·분석하고, 각 특성 정보들이 원문서의 어느 위치에 어떤 표현으로 제공되는지를 코퍼스 분석을 통해 알아보고자 한다. 이러한 연구는 감성공학 분야에서의 지식 관리의 한 단계로 간주될 수 있다.

2. 코퍼스

분석을 위해 두 종류의 코퍼스를 구성하였다. 시각, 청각, 후각, 촉각이라는 네 가지 감각

[1] 지표라는 용어 정의에 대해서는 Prieto(1968: 20) 참조.

의 범주로 분류하여 구성한 코퍼스A로부터 포괄적인 정보유형에 대한 단서를 얻을 것이다.

표 1 코퍼스 A 구성

감각분야	보고서명
시각	“색/조명환경 제시기술개발에 관한 연구”, “색채감성을 적용한 디지털카메라 개발”
청각	“감성 인식 시스템을 위한 음성 DB 구축에 관한 연구”, “음향·진동 환경 제시기술”
촉각	“촉각측정 및 질감제시기술 개발”, “피부감각의 감성측정 기술 및 DB개발”
후각	“후각/미각 감성 측정 기술 및 DB개발”, “후각환경제시 기술 개발”

코퍼스 A를 구성하는 보고서들은 감각별로 분류되어 한국과학기술원 형태소 해석기를 통하여 형태소 분석 및 품사 태깅[2]되었다. 분석된 자료를 다시 빈도 순으로 정렬함으로써 감각별 핵심어와 정보유형 추출을 위해 분석하였다. 이 정보를 기존의 수동 지표화된 결과와 비교함으로써 정보유형을 정리하였다.

구체적 정보유형과 그 특성을 추출하기 위해서 이미 지표화 된 열 편의 논문과 해당 지표를 가지고 코퍼스 B를 구성하였다. 이 코퍼스는 각 정보유형에 해당하는 문장들이 원문서의 어느 위치에서 어떤 언어로 제시되는지 수동으로 일일이 표시하였다. 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
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 연구
10	도시 경관의 색채 이미지 컨트롤을 위한 연구

정리하면, 코퍼스 A는 거시적 정보유형에 대한 아이디어를, 코퍼스 B는 미시적 정보유형 및 표현 특성을 찾는데 사용되었다.

3. 방법과 과정

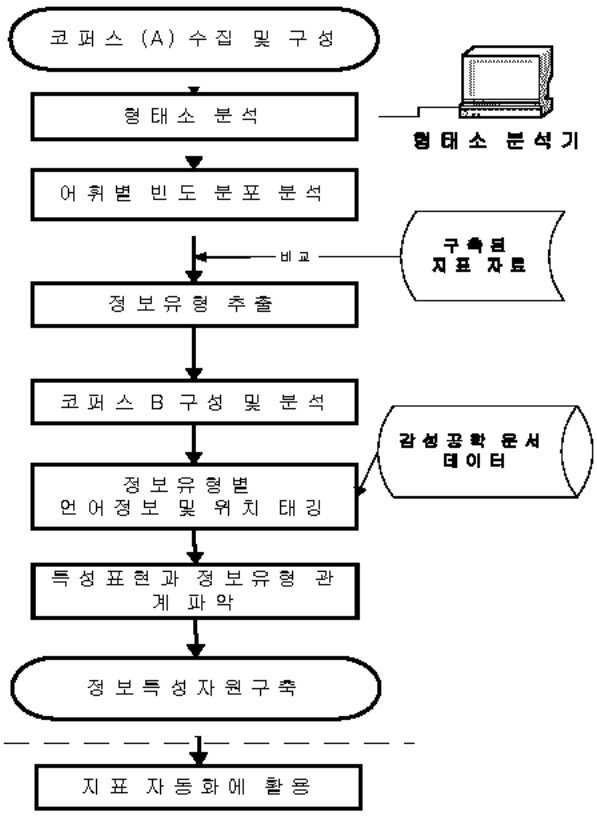

그림1 전체적 흐름도

그림 1은 연구의 전체적 흐름도이다. 점선 하

[2] http://morph.kaist.ac.kr/~morph에서 제공되는 형태소 분석 시스템은 사용자가 웹을 통해 원하는 문서를 입력함으로써 형태소 분석결과를 메일을 통해 직접 제공 받을 수 있다.

단은 향후 연구가 될 것이다.

3.1 정보유형 및 핵심어 추출

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

직관적으로 문서의 주요 정보유형은 《왜》, 《무엇을》, 《어떻게》, 그리고 기타 참조 사항으로 판단된다. 코퍼스 분석을 통해 이러한 직관이 실제 문서에서 어떤 유형으로 처리되고 있는지 알아 보자.

3.1.1 어휘분포 및 핵심어 추출

코퍼스 A를 한국과학기술원 형태소 분석기에 의해 분석 및 태깅하고 이 결과에 대한 형태소별 경우의 수, 상대빈도, 누적 상대 빈도 등의 통계 처리하였다.

문서 분류를 위한 분야별 핵심어가 대체로 누적빈도 50% 선에서 드러나는데, 이는 제목 텍스트와 고빈도어 목록에서의 핵심어 점유율 실험을 통해 구체화 될 것이다.(본 논문 3.2.1)

표 3 명사 누적빈도 상위 50%에 속하는 명사

시각	색채, 광원, 색, 영상
청각	음성, 인식, 피치
촉각	직물, 표면, 질감
후각	향, 후각, 자극

3.1.2 정보유형 추출

정확한 처리를 요구하는 지표화 작업을 자동화 한다는 것은 매우 어려운 일이다. 더욱이 원문서의 높은 전문성은 이러한 어려움을 배가시킨다. 까다로운 지표 자동화를 위해서는 주요 정보유형을 단순화 해야 하며, 본 연구는 문서 전체의 상위빈도 실질어를 기준으로 한 단순화 방법을 제안하는 바이다. 보고서 전체에서 누적빈도 50% 이상에 있는 모든 실질어는 다음과 같다.

표 4 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 얻은 어휘를 기반으로 다음의 정보유형을 도출할 수 있다.

표 5 고빈도 실질어로부터 도출된 정보유형

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 보고서 전체에서는 고빈도로 나타나지만 감각별로 분류된 보고서 그룹에서는 불규칙적이다. 개별 보고서에도 변함없이 나타나는 정보유형은 목적, 방법, 결과에 해당하는 표현들이다.

표 6 지표의 정보유형

주요어
설명
연구방법
소스데이터
활용분야
참고문헌

한편, "웹기반 감성 데이터베이스 구축 및 보급에 관한 연구" 과제에서 주관적으로 정리한 주요 정보유형을 보면 표 7과 같다. 과제에서는 이 여섯 가지 정보 외에 출처가 있다.[3] 일곱 개 정보유형 중에서 참고문헌, 출처, 소스데이터는 원문서로부터 바로 자동 정렬이 가능하다. 정확한 자동처리가 어려운 것은 《설명》, 《연구방법》, 그

[3] 김진호 외 (2001: 2) 참조.

리고 《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 알아 내는 일이다. 전문성이 매우 높은 특정 분야 논문들로부터 특성정보의 위치를 알아 내기 위하여 본 논문에서는 이미 과제로 축적된 지표와 논문을 대조함으로써 보편성을 이끌고자 하였다. 열 개의 전문적 문서와 그 지표를 수동으로 분석하여 위에서 추출된 정보유형을 찾아내고 각 정보유형이 문서의 어느 위치에 있는지 일일이 표시하였다.

표 7에서 태깅 결과를 관찰하면, 정보유형에 직접적으로 해당하는 30개 문장 중 77%가 서론과 결론에 위치하고 있다.[4] 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 연어 관계를 통해 가장 적합한 문장 추출로 이어질 것이다.

표 7 정보문의 언어특성

사무실 평가를 위해 office planner의 지식을 정리하는 평가법을 제안한다. 특히 시각 평가를 위해 퍼지 집합을 받아들인다.	서론 last	제안한다
퍼지 확률을 응용한 정량적 분석 방법, FFTA 아이디어를 사용함으로써, FT에 의해	본론 3.2	방법 -에 의해, 으로써
이와 같이 FT에 의해 전체를 간단히 파악할 수 있도록 대책을 세우기 쉽게 되었다. 가장 중요한 점으로는 FFTA 아이디어를 사용함으로써 시각적 평가가 애매한 경우의 평가가 가능하게 되었다.	결론 1	쉽게 되었다, 가능하게 되었다

지금까지 고빈도 어휘 분포에 의해 정보유형을 추출하고, 정보유형별 표현 특성과 원문에서의 위치를 수동으로 표시함으로써 정보유형과 언어 특성, 그리고 그 관계를 찾아 보았다. 표 8에서 그 결과를 정리한 것이다.

표 8 정보유형과 특성표현

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 고빈도 어휘나 제목에 얼마나 분포하고 있는지 제안된 방법의 타당성을 알아보자.

[4] 이와 같은 수치는 Saggion(2001)의 결과와 많이 다르지 않다. "We found that 72% of the information for the analytical stage comes from the following structural parts of the parent document: the title of the document, the first section, the last section and the subtitles and captions."

3.2 실험

타당성과 보편성에 대한 많은 실험이 필요하지만 본 논문에서는 기초적인 두 가지 실험으로 국한 시켰다. 핵심어가 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 얼마나 걸러지는지 실험하였다.

3.2.1 핵심어 매칭율

다음 표는 고빈도어와 제목에서의 핵심어 매칭율 조사 결과이다.

표 9 핵심어 점유율

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

핵심어 목록을 기준으로 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 후처리 하였다. 이는 본 연구의 코퍼스의 양이 적기 때문에 가능한 것이며 이후 연구에서 보완되어져야 한다.

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 개만이 제목에서 제시되었다. 제목이나 고빈도어에서 제시되지 않은 나머지 핵심어는 “발화 사고법”, “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 콘코던스 프로그램과 같이 특성 표현을 조건으로 넣어 이미 형태소 분석된 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 임의로 구성한 실험코퍼스를 사용하였다. 하나의 특성표현에 의해 추출된 정보문 중에서 지표에 제시된 방법을 기술하는 문장 정보가 들어 있으면 성공으로 처리하고 그렇지 않으면 실패로 처리하였다.

표 10 특성 표현을 통한 정보문 추출 성공율

특성언어	해당문장의 적합성	
	코퍼스B	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 고백해야 할 것이다. 사실, 특성표현에 의해 추출된 문장 중에 적합한 정보를 담은 문장이 있는지 알아내는 기초적 확인일 뿐이기 때문이다. 그러나 단 네 개 표현으로 이끌어 낸 결과임을 감안한다면 방법의 효과는 주목할 만하다.

4. 결론 및 향후연구

감성공학 과제 보고서로 구성된 코퍼스(A) 분석을 통하여 감성공학 전문분야 문서의 정보유형을 추출하고, 감성공학 관련 논문으로 구성된 코퍼스(B) 분석에 입각하여 정보유형을 구체화하고 정보유형별 특성표현과 원문 내에서의 위치를 파악하였다. 정보유형에 해당하는 문장들은 77%가 서론과 결론에 분포하였고 나머지 23%만 본문에서 제시되었는데, 이들 대부분은 방법에 관한 내용이었다.

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.

특성표현을 통하여 정보유형에 적합한 문장들을 찾아 낸다는 제안의 타당성을 살펴 보기 위해 비교적 본론에서 다루어 지는 비율이 높은 《방법》을 대상으로 특성표현을 통해 추출하고 전문가가 구성한 지표에서의 방법 기술과 비교하였다. 이로써 주요 정보유형, 정보유형별 특성표현, 핵심어 추출 방법의 타당성과 특성 정보에 입각한 지표 자동화의 가능성을 확인하였다.

그러나 제안된 방법의 보편성에 대해서는 좀더 연구가 필요하다. 실험 코퍼스의 크기가 지나치게 작았고, 비정보문 대비 정보문의 비율에 대한 조사와 더불어 보완해야 할 점이 있다. 이는 차후 연구를 통해 점진적으로 보완할 것이다.

참고문헌

[1] 곽현민, 조해성, 이상태, 「웹 기반 감성지표 DB 구축에 관한 연구」, 한국감성과학회, 2002 춘계학술대회 및 한일 국제 감성공학 심포지움 논문집, pp.79-85.

[2] 김진호, 이동춘, 박민용, 임좌상, 박수찬, 윤정선, 임현균, 김경택, 「웹기반 감성지표 개발 및 보급에 관한 연구」, 한국감성과학회 학술대회논문집, 2001.

[3] 박길환, 임은영, 박민용, 「Web 기반 감성 데이터베이스 구축을 위한 사용성 관련 감성지표 개발」, 감성과학회 학술대회논문집, 2001.

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

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

[6] Prieo, "Semioloie", dans *Le Langage*, La Pl旭ade, 1968[1997].

[7] Saggion, H. and Lapalme, G., "Where does Information come from? Corpus Analysis for Automatic Abstracting", Proceedings of *RALI*, Canada, 2000.

[8] Wright, S. and Budin, G. *Handbook of Terminology Management*. Vol. I. John Benjamins Publishing Company. Amsterdam /Philadelphia, 1997