산업 / 직업 분류 자동코딩 시스템

강 유 경*

요 약

많은 통계조사에서 사용되고 있는 산업/직업분류코드가 기존에는 사람에 의해 수동으로 부호화되어 왔는데, 이러한 작업은 시간과 인력면에서 고비용을 요구할 뿐 아니라 개인별 시각, 이해도의 차이 등으로 정확성에 많은 문제가 제기되어 왔다. 본 논문에서는 이러한 수동코딩 작업의 문제점을 해결하기 위하여 자동코딩 시스템을 개발, 이를 인구주택총조사와 사업체기초통계조사에 시험 적용하여 본 바를 바탕으로 향후 자동시스템으로의 전환 방향 등을 제시하고 있다.

Abstract

Korean standard industrial/occupational classification has been the basis of producing accurate statistical data related with our industrial structure and distribution of industry and occupation since 1960. But coding over several million records not only requires high cost in the aspects of time and manpower but also has many problems in accuracy and consistency. Therefore, we got to develop the automatic coding system in order to work out these problems of manual coding. This paper shows the structure of our system and the result of experiment over survey data of 2,000 Census.

*통계청 정보처리과 전산사무관

Ⅰ. 서 론

통계표준분류는 통계데이터의 정확성과 비교가능성을 확보하는 기초가 되는 것으로 우리나라에는 표준산업분류, 표준직업분류, 표준질병사인분류 등이 있다. 한국표준산업/직업분류는 1963년에 처음 제정되어 현재까지 각각 8번, 5번 개정되었으며, 우리나라 산업, 직업의 구조의 변화를 반영하여 왔다. 인구주택총조사 및 사업체기초통계조사를 비롯하여 각종 통계조사에서 산업분류 및 직업분류를 적용하여 결과자료를 산출하고 있으나, 산업/직업분류 결과자료가 여러 여건들로 인하여 정확성에 다소간 문제가 있어온 것이 사실이다.

이는 조사 자체의 한계에서부터 현재까지의 코드 부여작업이 다수의 비전문 인력들에 의한 수동코딩 작업에 주로 의존한 데에 기인한다. 조사 자체의 한계란 응답자들과 조사요원들의 산업/직업분류에 대한 무지로, 조사되는 내용이 분류 가능하도록 적절히 조사되는 것이 아니라, 너무 피상적이거나 혹은 서로 다른 분류의 내용들이 함께 조사되는 것을 뜻한다. 물론 조사요원들의 경우 교육을 통하여 이해도를 증진시킬 수 있겠으나, 인구주택총조사와 같은 대규모 조사는 단기간에 엄청난 수의 조사요원들을 임시로 고용하여 조사하게 되므로, 불과 1~2일간의 교육 실시만으로는 큰 효과를 기대할 수 없는 실정이다. 이는 해결책을 찾기가 매우 힘든 문제로, 개선책에 대한 계속적인 연구가 필요하다. 따라서 그간 코드 부정확성의 다른 한 요인이었던 코드부여 과정상의 문제에 초점을 맞추고자 한다. 현재까지 코드부여 작업은 사람에 의한 수동코딩작업이었다. 과거에는 임시채용된 심사요원들이, 최근에는 통계청 지방사무소 직원들이 코드를 부여하여 왔으나 코딩요원들간의 분류에 대한 이해도의 차이, 시각의 차이 등 다양한 원인으로 정확성에 문제가 있었고, 또한 시간과 인력 측면에서 고비용을 요구하기에, 산업/직업분류 코드화 과정을 자동화하는 작업을 시도, 2000년 인구주택총조사 등에 시범 적용하여 향후 자동코딩으로의 전환 가능성을 타진해 보았다.

Ⅱ. 관련연구

산업/직업분류 코딩 작업을 자동화하려는 움직임은 외국에선 1980년대부터 있어왔다. 대표적인 나라로 미국, 프랑스, 일본 등을 들 수 있다. 이들 나라의 공통점은 10년이 넘는 아주 장기적인 프로젝트로서 여러 분야의 전문가들이 추진해 왔는데, 그만큼 자동화 작업의 어려움을 시사한다고 하겠다.

1. 미국(Daniel, 1994)

미국은 1983년부터 자동화 알고리듬을 개발하기 시작하여 1990년 센서스의 산업/직업분류 코딩에 적용하였으며, 코딩 자동화 작업을 AIOCS, CACS, CATI/CAPI의 세 가지 영역으로 나누어 활용하였다. AIOCS(Automated Industrial/Occupational Coding System)는 주로 대규모의 배치작업에 활용하는 것을 목적으로 하였으며, CACS(Computer-assisted Clerical System)는 AIOCS가 코드부여에 실패한 데이터들을 처리하는 시스템이며, CATI/CAPI (Computer-Assisted Telephone Interviewing and Computer Assisted Personal Interviewing)는 조사시점에 활용하기 위한 시스템이다. AIOCS는 크게 KBS(Knowledge Base System), CS(Coding System), QMS(Quality Management System)로 구성되는데, KBS는 과거 정확한 사례들로 이루어진 데이터베이스이며, CS는 KBS를 이용하여 해답이 될 코드들을 찾는 역할을 하며, QMS는 일정한 점수 이상을 얻은 후보들 중 가장 최상위 후보를 되돌린다. 이 시스템을 1990년 센서스에 적용한 결과, 생성률이 산업분류 57%, 직업분류 37%이며, 정확률이 산업분류 93.8%, 직업분류 88.2%였다. 즉, 자동시스템에 의한 코드 부여율은 50% 내외이나, 부여된 코드의 정확도는 상당히 높다는 것을 알 수 있다. 미국에서는 이 시스템을 다른 경상조사 등에 활용 중이며, 생성률, 정확률을 높이는 시도가 계속되고 있다.

2. 일본

일본은 1992년부터 중·장기적으로 산업분류 자동화를 연구해왔다. 연구방향 설정에서부터 4년여에 걸쳐 시스템을 구축하고, 계속적인 시험적용 및 개선 등 8년 간의 장기 프로젝트를 진행해 왔다. 이 프로젝트는 사업소·기업통계조사의 산업분류에의 적용을 목표로 하여 여러 다양한 알고리듬 등을 구현, 성능을 비교하고 있는데, 1991년 사업소·기업통계조사의 실험 데이터 12,959 건에 대하여 최고 생성률 70%, 정확률(소분류 기준) 96.5%를 보이고 있다.

3. 프랑스(Pierrette, 1996)

프랑스에서는 1983년부터 QUID라는 소프트웨어를 코드 자동화에 이용해오다가, 1993년 QUID의 약점을 보완한 SICORE라는 새로운 소프트웨어를 개발·활용하기 시작하였다. 그 이용분야는 행정구역코드에서부터 직업코드에 이르기까지 다양하다. SICORE 시스템은 크게 학습과정과 코딩과정의 두 단계로 동작하는데, 학습과정에서 과거사례 등을 바탕으로 코드 트리를 형성, 형성된 트리는 코딩과정에서 코드부여 시 사용된다. 1990년 센서스 데이터에 적용하여 직업분류(세분류 기준)의 경우 생성률 66%, 정확률 90%를 얻었다.

Ⅲ. 산업/직업분류 자동화 시스템

앞에서도 언급하였듯이, 다수의 비전문 인력에 의한 코딩 작업의 문제점을 해결하고자 2000년 5월, 산업/직업분류 자동화 프로젝트에 착수하였다. 이 프로젝트는 통계청의 용역 의뢰로 고려대학교 팀에 의해 6개월간 수행되었다. 그림 1에서 보는 것과 같이 본 시스템은 크게 학습단계와 코딩단계의 두 가지 단계로 동작한다. 학습단계에서는 과거사례 등을 바탕으로 하여 지식베이스를 구축하고, 구축된 지식베이스를 활용하여 코딩단계에서 자동코드를 부여하게 된다. 시스템은 크게 형태소 분석기, 가중치 계산기, 코드 부여기, 지식베이스로 나뉘어진다.

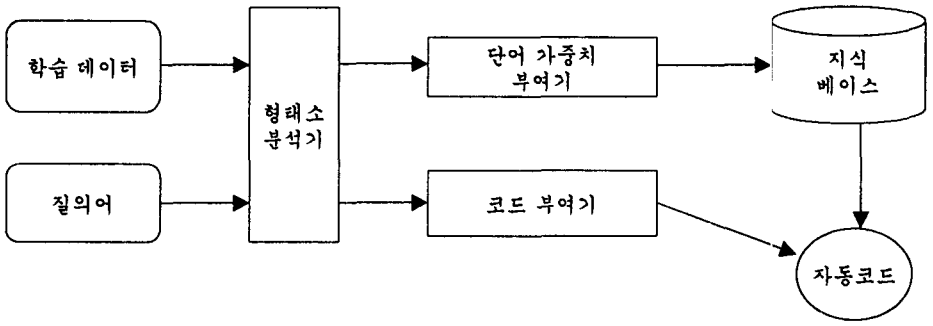

그림 4 자동화시스템 구조

1. 형태소 분석기

형태소 분석기의 역할은 문장 내에서 명사들을 추출하는 역할을 한다. 문장 내에서 중요단어들을 추출해 내는 능력은 전체 시스템에서 상당히 중요한 부분을 차지한다.

특히 형태소 분석기와 관련하여 주의해야 할 점은 띄어쓰기에 상당히 민감하다는 것이다. 이는 비단 본 형태소분석기만의 문제는 아니다. 거의 모든 형태소분석기의 경우 같은 문장이나 구를 사용하더라도 띄어쓰기 정도에 따라 추출되는 단어들이 달라진다. 복합명사의 경우 이를 붙여쓰면 복합명사를 이루는 각 명사와 복합명사가 함께 추출되나, 띄어쓰게 되면 구성명사만이 추출된다. 자동코딩 시스템의 성능 개선을 위해서는 이러한 형태소분석기의 특성을 고려하여 조사·입력 시 띄어쓰기가 반드시 이루어질 수 있도록 해야 한다.

2. 가중치 계산기

형태소분석기를 거친 학습데이터는 가중치 계산기를 거치게 되는데, 가중치 계산기는 단어들의 상대적 중요도를 반영하는 가중치를 부여한다. 일반적으로 가중치 계산은 문서검색이나 정보검색 시 꼭 필요한 구성요소 중의 하나로, 주로 *tf×idf* 알고리듬을 사용한다. *tf×idf* 알고리듬에서 tf는 단어빈도수(*term frequency*)를 의미하는 것으로, 한 문서 내에 많이 등장하는 단어일수록 그 문서를 대표한다고 보아 가중치를 높게 부여한다. idf는 문서역빈도수

(*inverse document frequency*)로 전체 문서라이브러리에서 그 단어가 등장하는 문서 수가 적을수록 특정문서를 대변하는 단어라고 판단, 높은 가중치를 부여한다. 본 시스템에서도 가중치 계산기 알고리듬으로 보편적인 *tf×idf* 알고리듬을 사용하였다.

3. 코드부여기

코드부여기는 학습과정에서 생성된 지식베이스를 사용하여 들어온 질의(적용대상 데이터)에 대한 코드를 생성하는 역할을 한다. 질의문을 구성하는 단어들을 포함하는 코드(문서)들이 후보가 되는데, 이들의 적합도를 p-norm 모델에 의하여 계산하여 가장 점수가 높은 코드를 되돌린다.

4. 지식베이스

지식베이스는 학습과정의 산출물로서, 학습데이터에 따라서 지식베이스가 결정된다. 학습과정에서 사용되는 학습데이터의 양이 많으면 많을수록, 또 그 정확도가 높으면 높을수록 지식베이스의 양과 질이 높아진다. 전체 시스템의 성능을 좌우하는 것은 지식베이스이고, 이 지식베이스를 결정하는 것이 바로 학습데이터이다. 즉, 학습데이터의 질과 양이 바로 전체 시스템의 성능을 좌우한다고 해도 과언이 아니다. 그러나 양질의 학습데이터를 구한다는 것은 상당한 기간과 비용을 요구한다. 95년 인구센서스와 같은 과거사례들은 많이 있으나, 정확도를 보장하지 못하므로 표준분류 전문가에 의한 검토가 반드시 뒤따라야 한다. 이러한 이유로, 많은 양의 양질의 학습데이터를 처음부터 구축하지 못하고, 계속해서 지식베이스를 확장해나가는 방식을 택하게 되었다.

5. 구축 시 성능 테스트

시스템 구축 단계에서 수행한 자동화 시스템 성능 테스트 실험결과가 표1에 나와 있다.

표 1 지식베이스 구축 시 성능테스트 결과

실험	학습데이터	테스트데이터	생성률	일치율 (1위)	일치율 (5위안)
1차	한국표준산업분류책자 산업분류색인표	99년시험조사 7,119건	99.2%	34.6%	60.1%
2차	한국표준산업분류책자 산업분류색인표 99년시험조사 7,119건	95년인구주택총조사 17,181건	98.1%	68.7%	81.3%
3차	한국표준산업분류책자 산업분류색인표 99년시험조사 7,119건 95년 인구주택총조사 46,762건	95년 인구주택총조사 4,676레코드	99.4%	72.8%	91.7%

표1에서 보듯이 1차실험의 결과는 다른 실험결과와는 사뭇 다르다. 물론 지식베이스 구축 시 사용된 학습데이터가 매우 적으므로 일치율이 다른 실험에 비해 낮게 나올 수 있지만 그 정도가 아주 크다. 이는 학습데이터와 테스트데이터 특성차이에서 그 이유를 찾을 수 있다. 학습데이터로 사용된 한국표준산업분류책자나 산업분류색인표는 문어체나 정형화된 표현을 사용하는 반면, 테스트데이터는 실제 조사된 내용들이므로 구어체나 비정형화된 표현이 많다. 이러한 차이로 인하여 일치율이 아주 낮은 것이다.

2차실험은 1차실험 시 테스트데이터로 사용된 99년 시험조사 자료를 추가로 학습시킨 후, 95년 인구주택총조사 자료 17,181건을 테스트한 결과이며, 3차실험은 2차실험 시의 테스트데이터를 포함한 95년 인구주택총조사 자료 46,762건을 추가로 학습시킨 후, 학습에 사용되지 않은 4,676건의 데이터를 테스트한 결과이다.

위 실험을 통하여 학습데이터로는 실제 조사된 과거사례들이 지침서나 색인표와 같이 정형화된 표현들로 이루어진 데이터들보다 더 적합하며, 과거사례 양이 많으면 많을수록 성능이 좋아짐을 확인할 수 있다. 또한 학습되지 않은 사례들에 대하여 약 70%의 일치율을 보이고 있음을 알 수 있다. 그러나 테스트데이터로 사용된 양이 너무 적으므로 방대한 양의 사례에 적용했을 때 유사한 성능을 보이리라 확신하기는 어려웠다.

현재까지 구축된 지식베이스에 사용된 데이터들은 다음과 같다. 한국표준 산업/직업분류 책자와 99년 인구주택총조사 시험조사 자료 7,100여건, 95년 인구주택총조사 50,000여건, 그 외 최신 직업리스트와 1999년 기준 사업체 기초통계조사 20,000여 자료이다.

Ⅳ. 적 용

본 시스템은 인구주택총조사와 사업체기초통계조사와 같은 대규모 조사에 적용하기 위하여 개발되었다. 대규모 조사의 경우 단기간 내에 방대한 양의 데이터를 처리해야 하므로 엄청난 인원이 필요하게 되는데, 이로 인한 비용을 줄이면서 좀더 정확하게 코드를 부여하려는 취지에서였다.

1. 2000년 인구주택총조사

인구주택총조사는 크게 전수와 표본조사로 나뉘어지는데, 표본조사중 경제활동에 관한 항목중에는 취업자의 산업·직업에 대한 것이 있다. 「사업체명(직장명)」, 「주 사업내용」, 「부서 및직책」, 「주로 하는 일」의 4가지 항목이 그것인데, 주로 앞의 두 항목으로 산업코드를, 뒤의 두 항목으로 직업코드를 부여한다.

분류는 계층적인 구조를 가지며 분류 정도에 따라 산업·직업 구조와 분포의 세밀도가 결

정된다. 산업·직업분류 모두 5단계(대, 중, 소, 세, 세세분류)로 나뉘어져 있는데, 분류정도를 높이기 위해서는 4가지 항목에 대한 조사가 아주 구체적으로 이루어져야 한다. 과거 인구주택총조사 시에는 소분류 기준으로 부여했으나, 2000년 인구주택총조사에서는 세분류 기준으로 그 분류 정도를 높였다. 또한 부호화 작업은 수동·자동코딩을 병행하기로 하였다. 이는 자동코딩 시스템이 아직 실험단계에 있으며 구축 시 성능테스트를 거쳤다고는 하나 테스트 데이터의 양이 적어 방대한 양의 데이터에 동일한 성능을 나타내리라 기대하기 어려우므로 단독적용에는 무리가 따를 것으로 판단, 기존의 수동코딩의 문제점을 보완하는 차원에서 자동코딩 시스템을 적용키로 하였다. 먼저 2000년 인구주택총조사 자료에 대하여 수동코딩과 자동코딩을 각각 실시하고 코드결과가 서로 일치하지 않는 자료에 한하여 다시 재확인·재코딩 작업을 실시함으로써 수동코딩의 오류가능성을 최소화하고자 하였다.

2000년 인구주택총조사 1,938,257 건에 대한 수동코딩 작업은 통계청의 지방사무소·출장소 직원들에 의하여 한 달간 진행되었다. 자동코딩 1순위와의 일치율은 산업분류 56%, 직업분류 42%였으며, 일치되지 않은 레코드에 대한 재확인 작업이 각 통계사무소·출장소 직원들에 의해 두 달간 진행되었다. 재확인 작업을 실시함으로써, 수동코딩과 자동코딩의 성능(생성률, 정확률)을 측정할 수 있다. 여기서 정확률은 엄격한 의미에서 계산된 정확률이 아니다. 즉 정확률이란 정답을 알고 있을 경우에만 계산될 수 있는데, 여기서 계산된 정확률은 다음의 두 가지 가정을 전제로 하였기 때문이다. 첫째, 수동과 자동코드가 일치하는 레코드의 경우, 코드가 잘못 부여되었을 가능성이 있음에도 불구하고 정확하게 부여되었다고 추정한다. 둘째, 불일치 레코드에 한하여 재확인 작업을 하게 되는데, 재확인 작업의 오류 가능성을 무시하고 재확인 시 부여된 코드가 정확한 것으로 역시 추정한다. 이러한 가정 하에 자동코딩과 수동코딩의 성능을 평가한 결과, 자동시스템의 생성률과 정확률은 각각 산업분류 98%, 60%, 직업분류 98%, 48%였으며, 수동코딩의 생성률과 정확률은 각각 산업분류 100%, 89%, 직업분류 100%, 76%였다.

2. 2000년 기준 사업체기초통계조사

사업체 기초통계조사는 300만이 넘는 전국의 모든 사업체를 대상으로 매년 실시되고 있다. 2000년부터 조사시 출력조사표를 사용하는데, 출력조사표란 사업체명, 대표자명, 주소, 사업내용 등 사업체의 기본사항이 이미 조사표 상에 기재되어 있는 것을 말한다. 따라서 각 조사원들은 변경·추가된 항목이나 전년도와 비교하여 변경된 내역만을 조사하게 된다. 사업체기초통계조사에서는 「사업체명」, 「주 사업내용」, 「주 영업품목」내용을 바탕으로 산업분류코드를 세세분류 기준으로 부여하는데, 사업체의 주 사업내용이나 영업품목이 변경되지 않는 이상 전년도의 코드를 그대로 사용한다. 여기에 부여된 코드 역시 조사원(지방자치단체 직원 또는 임시고용된 조사원)에 의해 매겨진 것으로 분류전문가에 의한 것은 아니다. 따라서 오류

의 가능성은 여전히 존재한다. 시간과 인력이 허락한다면 전체 자료에 대하여 점검하는 것이 바람직하나 이는 거의 불가능하므로, 분류가 잘못되었을 가능성이 높은 자료들을 대상으로 점검한다면 정확성을 크게 향상시킬 수 있을 것이다. 그럼 전체 자료 중에서 어떤 자료가 잘못 분류되었을 가능성이 높은 지 어떻게 알 수 있는가? 자동코딩 시스템을 활용함으로써 가능하다. 즉 자동코드와 수동코드가 일치한다면 코드가 정확하게 부여되었을 가능성이 높고, 불일치할 경우 코드가 잘못 부여되었을 가능성이 높을 것이다. 310만건에 이르는 사업체를 대상으로 자동코딩을 적용한 결과, 수동코딩 결과와 71% 일치하였다. 물론 일치한 71%가 100% 정확하다고 단언할 수는 없지만, 이중 오류 확률은 미미할 것이다. 또한 일치하지 않은 29%에 대해서도 전문가에 의한 재확인 작업을 거치지 않고서는 정확한 코드를 알 수 없으나, 29% 내에는 분명 수동코드가 정확한 경우도 많이 있을 것이므로 적어도 산업분류코드의 70% 이상은 정확하다고 미루어 짐작할 수 있다. 일치하지 않은 29%에 대한 분류전문가들의 확인작업이 필요할 것이다.

Ⅴ. 결과 분석

위 적용에서 우리는 세 가지 특이점을 발견할 수 있다. 첫째, 인구주택총조사와 사업체기초통계조사 적용 결과 일치율의 차이가 꽤 크다는 점이다. 둘째, 외국의 경우와 비교했을 때 상당히 높은 생성률을 갖는 반면에 정확률은 낮다는 점이다. 셋째, 수동코딩과 비교했을 때 생성률 및 정확률 모두 낮다는 점이다. 이는 자동화 프로젝트를 시작할 때, 정확성을 높이고자 한 취지와는 분명 다른 결과이다. 여기서는 위 세 특이점에 대해 분석하고, 자동화 시스템이 가지고 있는 문제점을 알아보고자 한다.

먼저, 인구주택총조사는 세분류 기준으로, 사업체기초통계조사는 세세분류 기준으로 코드를 부여했음에도 불구하고 일치율은 사업체기초통계조사가 15% 이상 높다. 이는 조사실태와 관련된다. 인구주택총조사의 경우 학생에서부터 주부에 이르기까지 아주 다양한 이력의 사람들이 조사원으로 활동하는 반면, 사업체 기초통계조사의 경우 사업체 수가 적은 곳에서는 주로 읍·면·동 직원들을 직접 조사원으로 투입, 그 외 지역에서는 임시조사원들을 대동토록 하고 있으며, 산업분류색인표를 마련하여 조사원들이 조사시점에 참고할 수 있는 지침으로 사용토록 하고 있다. 그러므로 사업체기초통계조사시 조사되는 내용들에는 지침서에 나와있는 내용들과 유사하거나 표준화되고 정형화된 표현들이 많다. 또한 인구주택총조사의 경우 산업/직업분류 관련 항목들에 관하여 실제 그 산업이나 직업에 종사하는 사람이 아닌 그 가족이 응답하는 경우가 많으므로 정확하고 자세하게 조사되지 않는데 반해, 사업체기초통계조사의 경우 그 사업체에 근무하는 직원들이 응답자가 되므로 보다 정확하고 구체적으

로 조사된다. 이러한 차이가 15% 이상의 일치율의 차이를 이끌어낸 것으로 판단된다. 조사되는 형태가 자동코딩 시스템의 성능에 많은 영향을 미친다는 것을 여기서 알 수 있다.

둘째, 생성률은 높지만 정확률이 낮은 이유는 시스템 구현 알고리듬과 관련된다. 미국이나 프랑스, 일본의 경우 정확률은 90% 이상이나 생성률은 50% 정도인데, 이는 생성률이 100%에 가깝고 정확률이 50~60% 정도인 우리나라의 경우와 분명 반대되는 현상이다. 이는 임계값을 사용하지 않는 시스템 특성에 기인한 것이다. 외국의 경우 후보들을 평가한 후 일정 임계값 이상이 되는 후보들만 다시 선별하고 그 중 가장 평가값이 높은 후보를 되돌린다. 즉 모든 후보가 일정 임계값을 넘지 않으면 시스템은 어떠한 결과도 되돌리지 않으므로 생성률이 떨어지는 것이다. 임계값이 높으면 높을수록 생성률은 낮아지고 정확률은 높아진다. 즉 생성률과 정확률은 역(trade-off)의 관계에 있는 것이다. 본 시스템의 경우 임계값이 0이므로 생성률은 그만큼 높고 정확률은 낮은 것이다.

셋째, 수동코딩보다 자동코딩이 정확률 측면에서 30% 정도 낮은 이유에 대해 생각해 보자. 물론 여기에는 시스템 차원의 문제를 비롯한 여러 다양한 요인들이 복합적으로 얽혀있을 것이나, 크게 두 가지로 나누어 설명하도록 한다.

1. 가정의 오류 가능성

앞서 언급했듯이 정확률을 계산할 때, 두 가지 가정을 전제했었다. 이 두 가정이 거짓이 되면 계산된 정확률의 신빙성이 떨어진다. 두 가정 중 하나가 재확인 작업 결과가 100% 정확하다는 것이었는데, 여기서 주의해야 할 점은 재확인 작업이 표준분류 전문가에 의해 이루어진 것이 아니라는 것이다. 표준분류 전문가의 손을 거치기에는 전국적으로 불일치하는 데이터의 양이 너무도 방대하다. 부득이 재확인 작업은 수동코딩을 담당하였던 직원들이 다시 한 번 더 참여할 수밖에 없었다. 더욱이, 2달 여의 재확인 작업기간에 다른 업무를 병행하면서 해야 하는 등 어려운 여건 하에서 이루어진 작업이라 그 질을 보장할 수가 없다.

다음은 충북 1개동 1,814건 자료에 대한 표준분류 전문가가 분석한 결과이다. 1,814건에 대한 수동·자동코드(제1위로 선택된 코드) 일치율은 산업 54%, 직업 32%이며, 불일치 레코드들을 대상으로 분석해본 결과, 조사내용 부실로 세분류 단계까지 코딩이 불가능한 비율이 전체 1,814건 기준으로 산업 8%, 직업 4%이며 수동·자동코드 모두 다 틀린 비율이 산업 3%, 직업 15%였다. 수동코드의 정확률은 산업 76%, 직업 58%이고, 자동코드의 정확률은 산업 63%, 직업 34%이다. 물론 이 분석은 2,000여건 데이터에 대한 분석이므로 전체 200만건에 이르는 데이터에 대한 결과값과 동일하다고 볼 수는 없지만 유사한 양상을 보일 것으로 생각된다. 이는 일반 직원들에 의한 재확인작업 결과와는 많은 차이가 있음을 알 수 있다.

한가지 흥미로운 사실은 여전히 자동코드보다는 수동코드의 정확률이 우수하나, 수동코드와 정답가능성 1위인 자동코드가 일치하지는 않지만 추가로 생성된 상위 5위안에 정답이 있을 비율이 산업 84%, 직업 68%로 상당히 높다는 것이다. 이는 순위부여 메커니즘을 잘 조정

하면 정확률을 상당수준 향상시킬 수 있음을 뜻한다.

또한 한가지 짚고 넘어가야 할 점은 조사내용의 부실로 전문가조차 분류 불가능한 데이터의 비율도 산업 8%, 직업 4%로 꽤 높다는 것이다. 아무리 좋은 학습데이터를 사용하여 지식베이스를 잘 구축하고 시스템을 잘 설계하더라도 조사 자체가 부실하면 일정 성능 이상의 정확률을 얻을 수 없다는 것을 여기서 잘 알 수 있다.

2. 시스템 차원의 문제점

시스템 차원의 문제점은 시스템을 이루는 각 구성요소별로 나누어 살펴보자.

(1) 형태소 분석기

형태소분석기의 가장 큰 문제점은 띄어쓰기에 민감하다는 것이다. 이는 비단 본 시스템에서 사용된 형태소분석기만의 문제점은 아니다. 문제는 조사표상에 조사내용이 기재되거나 전산입력되는 과정에서 띄어쓰기가 무시된 채로 기재·입력된다는 데에 있다. 띄어쓰기가 완전히 무시된 채 입력될 경우 중심단어가 추출되기보다는 엉뚱한 단어가 추출될 확률이 높아져 성능저하를 초래한다.

(2) 가중치계산기

각 조사항목에 기술된 단어들이 분류에 있어 똑같은 중요도를 갖지는 않는다. 단어에 따라서 매우 중요한 단어일수도, 그다지 영향을 주지 않는 단어일 수도 있다. 각 단어의 중요도를 반영하고자 가중치를 계산하게 되는데, 이때 *tf*×*idf* 알고리듬을 사용하였다. 단어의 가중치 계산 시, *tf*×*idf* 알고리듬이 보편적으로 사용되나 *idf* 속성은 계층적 구조를 가지는 문제에 적합하지 않은 점이 있다. '양말 제조'라고 조사된 경우를 예로 들자. '제조'라는 단어는 산업분류의 대분류를 결정짓는 역할을 한다. 즉 산업분류는 대분류 '제조업' 내의 특정 세분류에 속할 것이다. 그러나 전체 문서라이브러리에서 볼 때, 제조라는 단어는 대분류 '제조업'에 속하는 모든 세분류코드에 사용되고 있으나 '양말'이라는 단어가 등장하는 코드는 몇 되지 않는다. 따라서 *idf* 속성에 의해 '양말'이 '제조' 보다 더 높은 가중치를 얻게 되는 문제가 발생한다.

(3) 지식베이스의 양과 질

지금 현재 구축되어 있는 지식베이스의 양은 매우 적다. 지식베이스에 포함될 학습데이터의 질과 양을 높이는 일은 많은 시간과 노력을 필요로 하나, 성능의 기본이 되는 아주 중요한 작업이므로 계속해서 확장시켜 나가야 한다.

또한 현 지식베이스의 경우, 각 코드에 속하는 사례들의 양이 균등하지가 않고 편중되어 있는 편이다. 이러한 편중성은 자동코딩 시스템의 성능에 영향을 미친다. 즉 사례를 많이 포

함하고 있는 코드들이 그렇지 못한 코드들에 비해 상대적으로 유리한 입장에 있다. 사례를 많이 포함하고 있다는 것은 그만큼 다양한 표현력을 가지고 있음을 뜻하며, 이는 질의어에 나타나는 표현들을 포함하고 있을 가능성이 높음을 뜻하기 때문이다. 현재 지식베이스의 규모가 적은 편이므로 이러한 편중성을 허용하고 있으나, 향후 계속해서 지식베이스를 확장시켜나갈 경우에는 반드시 이러한 편중성을 제거해야 할 것이다.

Ⅵ. 향후과제

앞서 살펴본 것 같이 현 자동화 시스템의 경우 아직 실험단계에 있어 단독으로 적용하기에는 많은 문제점들을 내포하고 있으므로 기존의 수동코딩을 부분 보완하는 쪽으로 시범적용하였다. 비록 수동코딩과의 병행으로 인력·시간면에서의 비용은 줄어들지 않았으나 방대한 양의 자료 중에서 재확인 작업의 필요성이 있는 자료들을 추출하는데 활용했다는 점도 큰 의의를 갖는다. 특히 인구주택총조사의 경우 재확인 작업을 거침으로써 수동코딩의 상당부분(산업분류 11%, 직업분류 24%)에 수정이 가해져 정확성을 한층 더 높일 수 있었다.

그러나 자동코딩 시스템의 궁극적인 활용목적은 인력과 시간측면에서의 비용을 줄이면서 정확성을 한결 높이는 데 있을 것이다. 이를 위해서는 시스템차원에서의 연구뿐 아니라 다방면에서 계속적인 연구가 필요하다.

1. 조사 차원

모든 통계조사가 그렇듯이 통계의 정확성은 조사과정에서 일차적으로 결정된다. 조사가 잘못되면 자료처리를 아무리 잘 하더라도 소용없다. 특히 산업/직업분류 관련항목들에 관해서는 많은 응답자들이 응답하기를 꺼려할 뿐 아니라, 응답을 하더라도 두루뭉실하게 응답하는 경향이 짙다. 이럴수록 조사원들의 역량이 중요하므로 조사원들에 대한 관련교육이 철저히 이루어져야 할 것이다.

조사과정에서 언급해야 할 중요한 사항 중 하나가 조사되는 내용을 표현하는 방식에 따라서 자동코딩의 성능이 달라진다는 것이다. "길을 떠돌아다니며 장사함"이라고 표현하기 보다는 "행상"으로 표현하는 것이 자동코딩에 더 적합한 표현방식인 것이다. 이러한 표현형태에 대한 다양한 연구가 이루어져야 하며, 이것이 조사원 교육 시 반드시 반영되어야 할 것이다. 또한 현재 조사·입력되는 형태를 보면 띄어쓰기가 완전히 무시되고 있는데, 자동코딩 시스템의 성능을 위해서는 띄어쓰기가 제대로 되어야 한다. 이를 위해서는 조사표 설계 당시부터 띄어쓰기를 유도할 수 있는 방법들을 고려하여야 할 것이다.

2. 산업/직업분류 전문가 전담팀 구성

외국의 경우와 비교했을 때, 프로젝트 진행 시의 미비점은 단연 프로젝트 팀의 구성원에 있다고 할 것이다. 미국이나 프랑스, 일본의 경우 관련 분야의 전문가들이 함께 진행한 데 반하여, 우리나라는 전산시스템 측면에서만 접근했다는 것이다. 산업/직업분류 자동화에 있어 가장 중요한 작업이 정확한 과거 사례를 바탕으로 한 지식베이스구축에 있다고 할 때, 프로젝트 추진팀에 있어 빠져서는 안될 사람은 산업/직업분류 전문가들이다. 우리나라의 경우 프로젝트 추진팀에 산업/직업분류 전문가들이 주축이 아닌 협조라인 상에 있어, 업무사정상 그들의 전폭적인 협조를 구하기가 사실상 어려웠다. 이들이 프로젝트에 함께 참여하여 주체적으로 일을 추진할 때 가장 기본이 되는 틀을 제대로 형성할 수 있을 것이다.

3. 시스템 차원

현 시스템은 형태소분석기나 가중치 부여 알고리듬 등에 많은 개선의 여지가 남아 있다. 이들 구성요소 각각에 대해 다양한 알고리듬들을 개발·적용하여 최적의 알고리듬을 선별해야 할 것이다. 또한 향후 적용을 위해서는 정확률보다 생성률이 높은 현 체제보다는 외국의 경우와 같이 정확률이 90% 이상 높은 것이 바람직하다. 이를 위해서는 임계값을 도입하여 일정수준 이상의 정확률을 확보해야 할 것이다. 이렇게 함으로써 자동코딩을 단독으로 활용할 수 있는 기반을 마련할 수 있다. 물론 생성률이 100%에 이르지 않는 이상 자동코딩만으로 답을 알 수 없는 일부자료들에 대해서는 수동코딩이 불가피하다. 이에 대해서는 정확률을 확보한 연후에 정확률을 떨어뜨리지 않으면서 생성률을 높이는 방법들에 대한 연구가 뒤따라야 할 것이다. 아울러 활용방법면에 있어서도 자동코딩 시스템을 방대한 양의 배치작업에만 활용할 것이 아니라, 상위 5위안에 있는 자동코드의 정확률이 높은 점을 감안할 때 이를 CAPI 등과 연계하여 조사원들이 조사시점에 후보코드들 중에서 직접 코드를 선택하게 하는 등 여러 방법으로 활용방안을 확대해 나가야 할 것이다.

참고문헌

Daniel W. Gillman and Martin V. Appel, 1994. "Automated Coding Research At the Census Bureau," Research Report

Pierrette Schuhl, 1996. "SICORE, The INSEE Automatic Coding System,"