2006년 한국산학기술학회 춘계 학술발표논문집

표준 통계 분류 코드 자동 생성

임희석
한신대학교 컴퓨터정보소프트웨어학부
e-mail:limhs@hs.ac.kr

Automatic Generation of Standard Classification Code

Heui Seok Lim
Div. of Computer, Information, and Software
Hanshin University

요 약

본 논문은 수동 코드 분류 규칙과 예제기반의 자동 학습을 이용하는 한국어 표준 산업/직업 코드 자동 분류 시스템을 제안한다. 제안된 시스템은 산업과 직업에 대하여 설명하는 자연어를 입력받아 해당 산업/직업 분류 코드를 생성하는 시스템으로 수작업으로 구축된 규칙을 적용한 후 규칙이 적용되지 않는 레코드는 예제 기반의 학습을 이용한 자동 분류 시스템에 의해서 해당 코드를 할당한다.

1. 서 론

표준산업분류 코드와 표준 직업분류 코드는 조사원이 가구 조사에서 얻은 사업체명, 사업체의 주된 사업 내용, 직책, 그리고 직무에 대한 자연어로 기술한 설명에 근거하여 수작업으로 작성된다. 표준산업분류코드와 직업분류코드 분류 코드 할당을 위한 이러한 수작업은 코드 분류 전문가의 경험과 지식에 의존함으로 인하여 다음과 같은 문제점을 초래한다.

① 수작업을 수행하기 위한 작업자 교육 및 활용에 많은 비용이 소요
② 막대한 수작업 량과 고비용 발생
③ 코딩된 작업 결과의 일관성 결여

위와 같은 수작업에 의한 표준 코드 분류 작업의 문제점을 극복하기 위한 방법은 가구 조사에서 얻은 자연어의 응답을 표준 분류 코드로 분류할 수 있는 자동 코드 분류 시스템을 개발하여 활용하는 것이다.

2. 제안 시스템

본 논문이 제안하는 시스템은 조사원들로부터 획득된 '근무 사업체명', '사업체의 주된 업무', '직책', 그리고 '직무'에 대한 내용을 자연어로 입력받아 입력된 내용에 해당되는 직업/산업 표준 코드를 생성한다. 제안하는 시스템은 크게 학습 모듈과 자동 코드 생성 모듈로 구성된다. 학습 모듈(learner)은 수작업으로 정확한 분류 코드가 할당된 학습 데이터를 입력받아 입력 데이터의 띄어쓰기 오류를 수정하는 띄어쓰기 교정 모듈, 색인어 추출 모듈, 2-포아송 모델에 의하여 색인어의 가중치를 계산하여 역화일 형식의 색인어 DB를 구성하는 kNN(k-nearest neighbors)기반의 학습 모듈로 구성된다. 자동 코드 생성 모듈(automatic code generation module)은 띄어쓰기 모듈, 색인어추출 모듈, 전문가에 의하여 작성된 규칙을 적용하는 수동 규칙 적용 모듈, 수동 규칙에 의해서 분류가 되지 않은 레코드들에 대하여 입력 레코드와 유사한 예제를 검색하고 검색된 결과의 유사도 값을 이용하여 분류 코드의 유사도를 계산하는 코드 자동 할당 모듈, 그리고 시스템이 자동으로 잘못 분류한 코드의 올바른 코드에 대한 피드백을

입력받고, 이를 이용하여 학습데이터의 신뢰도를 재조정하는 신뢰도 개선 모듈(refiner)로 구성된다.

3. 학습 및 코드 분류

3.1 자동학습

본 논문은 코드 자동 분류기의 학습을 위하여 kNN 방식의 학습 방법을 사용하였는데, 이 방법 사용하는 이유는 다음과 같다. 첫째, 코드 자동 분류 작업을 위한 입력 레코드의 길이는 30어절에서 50어절 크기로 매우 짧아 레코드의 검색이 매우 용이하다. 둘째, 이전의 통계 조사 때 수작업으로 분류된 대량의 예제 데이터를 활용할 수 있다. 셋째, 분류하여야 할 범주의 수가 매우 많다.

3.2 수동규칙

본 시스템은 자동 코드할당 이외에도 사용자가 정의한 규칙에 맞는 데이터의 경우에는 수동 규칙에 따라 코드를 할당할 수 있는 매커니즘을 제공한다. 수동 규칙은 산업/직업 코드 분류를 수행하는 통계청의 전문가들에 의해서 수작업으로 구축된 규칙으로서 '조건-행위'의 형식으로 구축되어 있다. 제안하는 시스템은 코드 분류의 정확도를 높이기 위하여 입력 레코드를 수동 규칙에 적용을 하여 수동 규칙에 적용되면 규칙에 의해서 코드를 할당하며 규칙이 적용되지 않는 레코드의 경우 자동 코드 할당 모듈에 의해서 분류 코드를 할당하도록 한다.

현재 구축되어 있는 수동 규칙은 총 2,655개이며 수동 규칙을 6,000개의 실험 레코드에 적용한 결과 약 98.9%의 정확도를 얻을 수 있었다.

3.3 자동 코드 생성

코드 자동 할당 과정은 입력 레코드와 유사한 코드를 검색하고 검색된 코드와 입력 레코드와의 유사도를 이용하여 최종적인 출력 코드를 계산하는 방식으로 이루어진다. 제안된 시스템에서 사용하는 검색 모델은 2-포아송 모델로 TREC-8의 Okapi 시스템 [3]이 사용하였던 BM25 방법에 의하여 계산한다.

코드 자동 분류를 위하여 새로 입력된 레코드를 질의로 하여 기분류된 레코드들을 검색한 유사도 값을 이용하여 랭킹한 후 상위 k번째까지 이웃들의 유사도 값을 참고하여 (식 1)에 의해서 최종적으로 레코드 r_i가 코드 c와 갖는 스코어 값을 계산한다.

$$CSV_c(r_i) = \sum_{t_j \in kNN} sim(r_i, t_j)\, y(t_j, c)\, conf(t_j) \quad \text{(식 1)}$$

4. 실 험

한국 표준 산업 분류와 직업 분류 코드는 1수준에서부터 5수준까지 계층적으로 분류되어 있으며 각 수준의 분류 코드의 수는 [표 1]과 같다.

표 1. 한국표준산업(직업) 분류 코드 분류 체계

코드 \ 수준	1	2	3	4	5
산업분류	20	63	194	442	1,121
직업분류	11	46	162	447	1,404

본 논문은 직업 코드 분류에서는 4수준 코드로의 분류를 실험하였고, 산업 코드 분류에서는 5수준의 코드 분류 실험을 하였는데, 이는 일반적으로 통계 조사 시 직업 코드 분류는 4수준의 결과를 산업 코드 분류에서는 5수준의 결과를 많이 사용하기 때문이었다. 즉 직업 코드는 447개의 코드가 분류 대상이었으며 산업 분류 코드는 1,404개의 코드가 분류 대상이었다. 자동 코드 분류 시스템의 학습 데이터와 실험 데이터는 [표 2]와 같이 구성하였다.

표 2. 학습/실험 데이터

항목	학습 데이터	실험 데이터
직업분류코드	400,000	10,000
산업분류코드	400,000	35,697

시스템의 자동 분류의 성능을 평가하기 위한 평가척도로는 N-best 정확도를 이용하였다. N-best 정확도란 실험에 사용된 전체 레코드의 수에 대해, 시스템이 출력한 상위 N개의 레코드 중에 정답이 포함된 레코드의 비율로 계산하였다.

아래 [표 3]과 [표 4]는 데이터 색인 시 바이그램, 명사추출기, 그리고 형태소 분석기 방법을 사용한 경우의 직업/산업 직업 코드 분류의 성능을 나타낸 것이다. 두 표에서 각 행은 상위 1개부터 10개까지의 결과를 출력했을 때의 N-best 정확도를 나타낸다.

실험 결과에 따르면, 색인어 추출 방식 중에서는 바이그램 색인 방법이 가장 높은 성능을 보였으며

그 다음은 명사 추출, 형태소 분석을 이용한 방법이었다. 가장 단순한 바이그램 색인 방법의 성능이 가장 우수했다는 것은 띄어쓰기 문제와 미등록어 처리에 가장 견고했음을 의미하는 것으로 추측된다. 색인 방법에 상관없이 직업 코드 분류 결과가 산업 코드 분류 결과보다 낮은 정확도를 보였는데, 이러한 실험 결과는 직업 코드 분류가 산업 코드 분류보다 난이도가 높다는 것을 의미한다.

표 3. 직업분류코드 실험 결과

	바이그램	명사추출	형태소 분석
1위	**42.80**	41.86	40.88
2위	**56.82**	55.74	55.34
3위	**63.85**	62.76	62.31
4위	**68.22**	66.94	66.89
5위	**71.07**	69.70	69.70
6위	**72.98**	71.99	71.55
7위	**74.38**	73.45	72.95
8위	**75.42**	74.46	74.08
9위	**76.10**	75.28	74.72
10위	76.63	75.89	75.49

표 4. 산업분류코드 실험 결과

	바이그램	명사추출	형태소 분석
1위	**95.84**	94.79	95.63
2위	**98.49**	97.95	98.40
3위	**99.14**	98.69	98.97
4위	**99.38**	99.05	99.21
5위	**99.51**	99.25	99.34
6위	**99.60**	99.34	99.42
7위	**99.63**	99.37	99.47
8위	**99.65**	99.42	99.50
9위	**99.66**	99.43	99.51
10위	**99.68**	99.44	99.51

6. 결 론

본 논문은 인구통계조사를 통하여 수집된 산업/직업 분류에 관한 자연어로 기술된 내용을 입력받아 해당 표준 코드로 분류하는 산업/직업 코드 자동 분류 시스템을 제안하였다. 제안된 시스템은 코드 분류에 관한 전문가가 이전에 수작업으로 분류한 데이터를 활용할 수 있도록 메모리 기반 학습의 일종인 kNN 학습 기법을 이용하여 학습 및 자동 생성하였다.

제안된 시스템을 40만개의 학습 데이터를 이용하여 실험한 결과 바이그램을 이용한 색인어 추출 기법을 사용하였을 때 가장 높은 성능을 보였다. 10-best 성능 평가 결과 직업분류 데이터에 대해서 76.69%, 산업분류 데이터에 대해서는 99.68%의 정확도를 보였다.

제안한 시스템은 코드 분류를 위하여 수작업을 전혀 사용할 필요가 없는 완전 자동화 시스템으로 사용하기 위해서는 아직 성능이 떨어지는 상황이지만 10-best의 결과 중 올바른 코드를 할당하도록 함으로써 코드 분류를 하는 전문가의 작업을 경감시킬 수 있는 반자동 도구나 수작업자의 코드 분류 결과를 검증할 수 있는 도구로서는 충분히 활용될 수 있다고 판단된다.

참고문헌

[1] Apeel, M. V. and Hellerman, E., Census Bureau Experiments with Automated Industry and Occupation Coding, Proceedings of the American Statistical Association, 32-40, 1983.

[2] Rowe, E. and Wong, C., An Introduction to the ACRT Coding System, Bureau of the Census Statistical Research Report Series No. RR94/02 (1994)

[3] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3", in the Proceedings of Text REtrieval Conference (TREC-3), 1995.