통계청
통계개발원

인공지능 기반 통계분류 자동화

통계개발원 통계방법연구실 임경민사무관

통계청
통계개발원

목 차

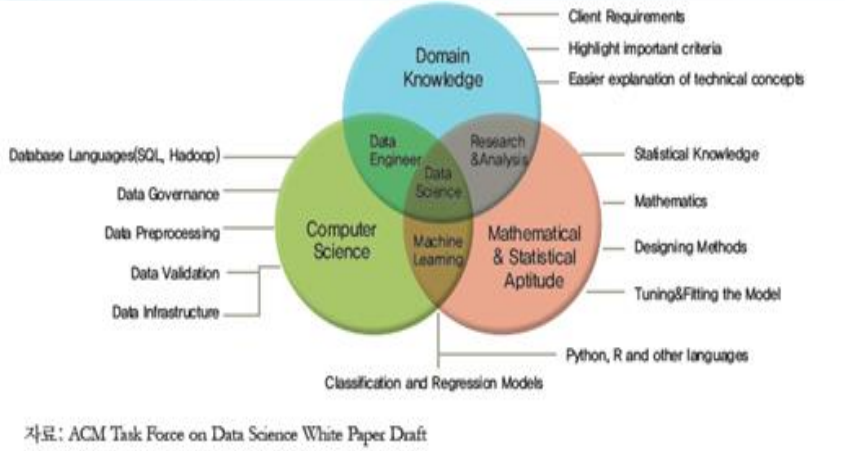
통계청
통계개발원

데이터과학 연구 추진경과

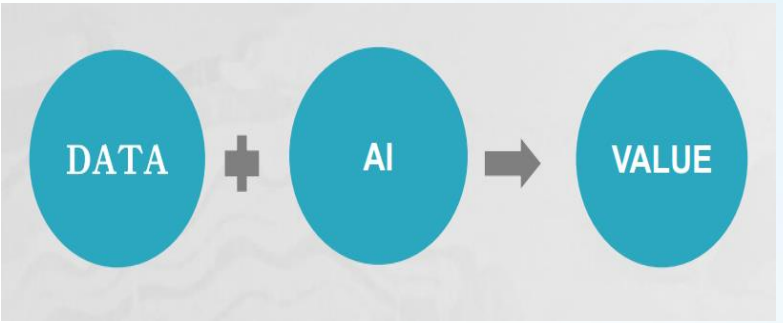
Ⅰ. 데이터과학 연구 추진경과

1-1. 데이터 과학(Data Science)

- 데이터 과학(Data science)이란 더 향상된 데이터 분석을 위해 통계학이 전산학과 융합하여 학습의 영역을 확장해가는 융합 학문이라고 볼 수 있음 <Willaim Cleveland>



자료: ACM Task Force on Data Science White Paper Draft



- 4차 산업혁명 시대를 맞아 다양한 대용량 자료를 IT 기술과 연계하여 신속·정확하게 정보를 창출하고 활용하는 **데이터과학 기법의 국가통계 적용** 요청이 증가하고 있음
  - 통계개발원은 미래 통계환경 변화를 위해 데이터 사이언스 관련 연구를 추진
    * 2020.2. 통계방법연구실에 데이터과학 연구팀 신설 (5급1명 6급1명 => 현재 4명으로 확대)

Ⅰ. 데이터과학 연구 추진경과

1-2. 데이터과학과 국가통계

AI 기술 국가통계 활용의 두 범주

구분	예상 쟁점
신규통계 생산	- AI 및 빅데이터 기반 분석결과에 대하여 보편적인 신뢰가 확보되었는가
	- 기존 통계 생산 프레임워크가 제공하는 모수적/인과적 설명이 가능한가
	- 재현불가능한 비결정론적 알고리즘 기반 결과를 국가통계체계로 관리 가능한가
기존 통계생산 프로세스 적용	- 기존의 방법과 대비하여 정확성을 객관적으로 입증할 수 있는가
	- 조사부터 공표까지 제한된 통계생산 기간을 AI 기술 적용으로 단축할 수 있는가

*자료:
김정민 外「AI 기술의 국가통계 활용사례 및 국내도입 촉진방안」
(소프트웨어정책연구소 ISSUE REPORT 2020.6.26. IS-098)

통계프로세스 적용 해외사례

■ **비정형 텍스트 분석을 통한 분류 자동화**

국가	활용 내용
미국	- (BLS) 업무상 상해 및 질병 설문조사 응답결과 자동 코드 분류
	- (Census) 미국 지역사회 설문조사의 산업 및 직업에 대한 코드 분류
독일	- 국내 법인에 대한 유럽국민계정체계 상 기관 분류 할당
	- 온라인 구인 공고 분류
캐나다	- 월별 소매거래 및 분기별 소매상품 조사에 스캔 데이터 활용
핀란드	- 경찰청 교통사고 보고 문건에 기반해 사고 내역 자동 분류
벨기에	- 구인 공고별 알맞은 경제 활동 통계 분류 코드 예측
룩셈부르크	- 민간 기업의 스캐너 데이터를 활용해 상품 분류 코드 추천
스위스	- 항공 이미지에 기반한 토지 피복 및 토지 사용 코드 분류 자동화

■ **다출처 자료 연계 및 보완**

국가	활용 내용
독일	- 데이터 연계를 통한 연방 고용청의 패널 데이터 보강
캐나다	- 인구 조사를 구성하는 이민 허가 변수에 대한 결측 값 대체
네덜란드	- 보건 실태조사 대상으로 한 혼합형 설문조사의 무응답 대체
오스트리아	- 행정 데이터와 SILC 융합을 통한 ICT 조사의 소득 문항 대체

Ⅰ. 데이터과학 연구 추진경과

1-3. AI 통계분류 가능성 시험분석(2020~2021)

지역별고용조사 산업분류(20)

분류 모델	분류	대분류 (21)	중분류 (77)	소분류 (232)	세분류 (495)
Baseline (FCNN), epoch=200, lr=0.05	정확도 (F1)	0.959	0.938	0.914	0.889
	학습시간 (s)	67.0	143.0	271.2	393.0
	분류 수행속도 (iter/s)	79691.9	81395.6	39805.3	21773.8
LightGBM, epoch=10, lr=0.5	정확도 (F1)	**0.941**	**0.867**	**0.729**	**0.653**
	학습시간 (s)	3987	6059	8089	10659
	분류 수행속도 (iter/s)	6009	5780	3546	1476
CNN Sentence Classifier, epoch=200, lr=0.05	정확도 (F1)	**0.954**	**0.930**	**0.909**	**0.870**
	학습시간 (s)	356	508	908	2067
	분류 수행속도 (iter/s)	26745	24897	13670	10245

- 지역별고용조사 과거 공표자료 기준 AI 산업분류 시험분석 결과 세분류 수준에서 88.9% 예측정확도 확인

건설업조사 공종/발주자분류(21)

평가 데이터	분류	정밀도 (Precision)	재현율 (Recall)	F-1 Score	정확도 (Accuracy)	비고
전체 데이터	공종	0.88	0.88	0.88	0.88	전체적인 정확도 성능
	발주자	0.88	0.88	0.88	0.88	
현 시스템 분류 결과 = 최종결과인 경우	공종	0.96	0.95	0.96	0.95	자동코딩 시스템과 유사성
	발주자	0.95	0.95	0.95	0.95	
현 시스템 분류 결과 없는 경우	공종	0.81	0.82	0.81	0.82	내검 작업 감소 정도
	발주자	0.83	0.82	0.82	0.82	
현 시스템 분류 결과 ≠ 최종결과인 경우	공종	0.67	0.62	0.63	0.62	오답율 개선 정도
	발주자	0.77	0.74	0.73	0.74	

- 건설업조사 과거 공표자료 기준 AI 공종/발주자분류 시험분석 결과 기존 규칙기반 전산연계 시스템보다 높은 수준의 예측정확도 확인

*자료: 임경민·김혜란(2020)「데이터과학 기법의 국가통계 시험적용 분석」통계개발원 연구보고서
오교중(2021) 「인공지능 기반 통계분류 자동코딩시스템 구축 연구」통계개발원 연구용역보고서

Ⅰ. 데이터과학 연구 추진경과

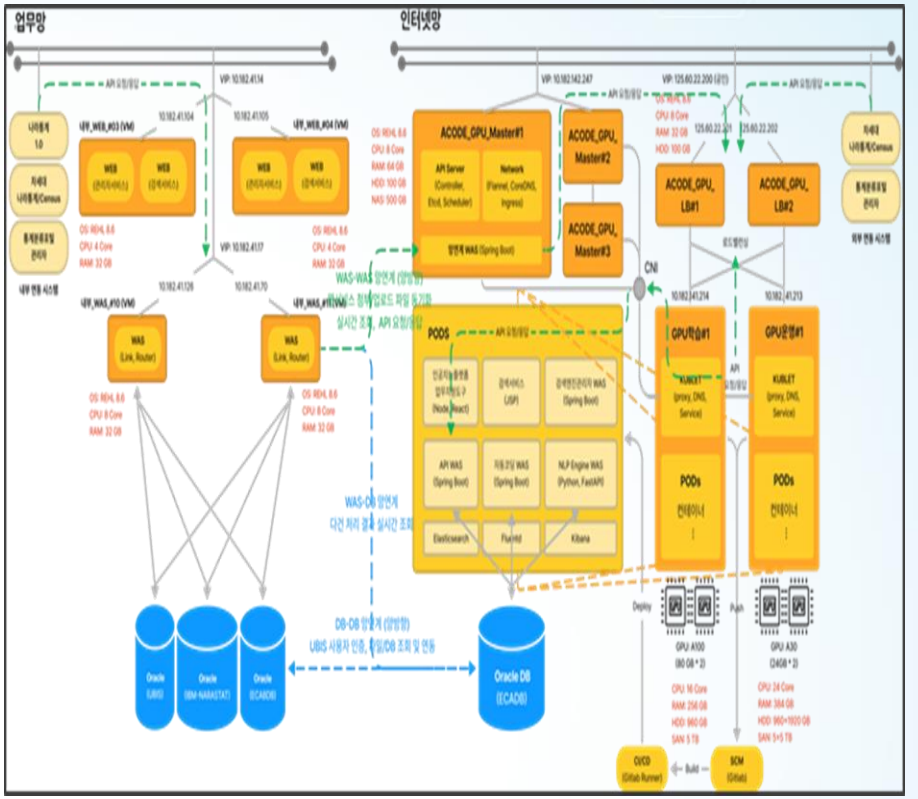
통계청 통계개발원

1-4. 인공지능 기반 통계분류 자동화 시스템 구축(2022)

개발예산 확보(15.5억)

- 선행연구를 토대로 범용 통계분류 자료처리 플랫폼인 인공지능 기반 통계분류 자동화 시스템 개발사업 추진
- ´22년 과기부 「디지털 공공서비스 혁신 프로젝트」 지원사업으로 선정되어 시스템 개발예산 확보
  * AI 등 첨단 ICT 기술을 공공분야에 선도적으로 적용하기 위해 사업공모를 통해 11개 과제를 선정하여 총 197억원 지원
- [사업범위] 대규모 5개조사 5종분류 대상으로 시스템 개발
  - 산업/직업분류, 공종/발주자분류, 상품군별분류
  - 인구총조사, 전국사업체조사, 지역별고용조사, 건설업조사, 온라인쇼핑조사
- [개발내용]
  - 자연어처리: XLM-RoBERTa 모델 사용 텍스트 분석
  - 기계학습: 공표자료 활용 통계분류 예측 및 확률값 제공
  - 최신 GPU서버(A100 2식) 사용으로 처리속도 향상

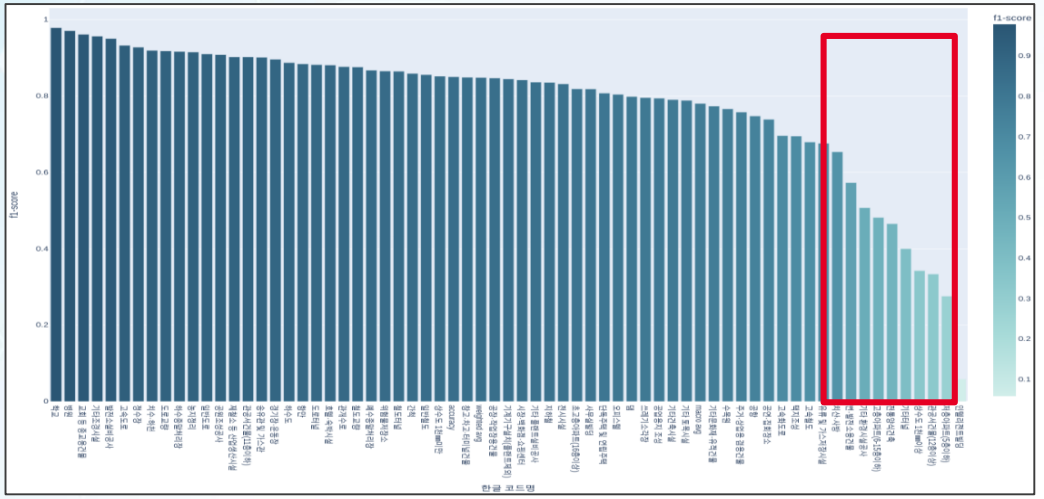
AI 통계분류 자동화 시스템 구성도



Ⅰ. 데이터과학 연구 추진경과

1-5. AI 통계분류 결과 실무 시험적용 및 검증(2023~)

- **AI 통계분류 예측결과를 활용하여 분류정확성·일관성 및 행정효율성 제고**



	text	prob
0	어린이집에서 보호자로부터위탁을받아 보호자로부터위탁을받아	0.898575
1	사업장에서 소매상 대상 소매상 대상	0.853393
2	영업장에서 고객의뢰를 받아 고객의뢰를 받아	0.602864
3	교회에서 기독교종교활동 기독교종교활동	0.334787
4	초등학교 초등교육서비스 초등교육서비스	0.572152

선택적 에디팅
→ 내용검토 우선순위 결정

1. 예측정확도 낮은 분류영역 내검
2. 예측정확도 낮은 개별사례 내검
3. 기존 자동코딩(사례사전) 결과와 불일치 사례 내검

내검결과 강화학습 및 재학습
→ AI 분류예측 정확도 제고 도모

Ⅰ. 데이터과학 연구 추진경과

통계청 통계개발원

[참고] AI 통계분류 관련 언론보도 등

'AI가 통계 분류를' 통계개발원, 산업분류 자동화 가능성 제시

파이낸셜뉴스 입력 :2022.03.03 10:31 수정 : 2022.03.03 10:31

자연어 처리기술·머신러닝 기반 알고리즘 응용해 연구

(통계개발원 홈페이지 캡처)

[파이낸셜뉴스] 통계청 통계개발원(SRI)이 인공지능(AI)을 활용해 자동으로 산업분류를 할 수 있는 가능성을 제시했다.

통계개발원이 3일 발간한 '데이터 리서치 브리프' 6호에는 이 같은 내용이 담긴 '인공지능 기반 산업분류 자동화 연구' 결과 보고서가 실렸다.

통계청, 통계데이터 이용한 인공지능 활용 대회 공모...'자연어 기반 AI 활용 산업분류 자동화'

NEWSIS 경제 > 경제일반

통계개발원, 2년간 AI 활용 통계분류 연구...정부 사업 발판

등록 2022.03.03 10:00:00

'데이터 리서치 브리프' 6호 발간

통계청 NEWSIS

[제종=뉴시스] 박영주 기자 = 통계청 통계개발원은 지난 2년간 진행한 인공지능(AI) 응용 통계분류 혁신 연구가 과학기술부 주관 '인공지능 기반 통계분류 자동화' 연구사업의 발판 역할을 했다고 3일 밝혔다.

'인공지능 기반 통계분류 자동화' 연구사업은 과학기술부 주관 '2022년 디지털 공공서비스 혁신 프로젝트'에 선정됐다. 올해 관련 사업비로 17억5000만원을 확보했다.

통계개발원은 '데이터 리서치 브리프' 6호를 통해 '인공지능 기반 산업분류 자동화 연구'가 과기부의 프로젝트 공모에 선정된 사업의 핵심이 됐다고 설명했다. 데이터 리서치 브리프는 통계개발원이 발간하는 계간지로서 민관학계와 소통을 강화시키는 역할을 한다.

연구진인 김혜란 사무관과 임경만 주무관은 '자연어 처리기술과 머신러닝 기반 분류 알고리즘 응용을 통해 인공지능 기반 산업 분류의 자동화 혁신 가능성을 제시했다. 한국표준산업분류를 중심으로 현재 통계 생산과정에 인공지능 기반 통계분류를 활용해 분류의 정확성을 높인 것이다.

전영일 통계개발원장은 "2년간 집중한 AI 연구는 정부 내 연구개발(R&D) 체신의 한계를 넘어 민관 협력을 통해 AI 혁신을 선도할 수 있는 가능성을 열었다"고 평가했다.

통계청, AI 기반 통계 자동화 등 적극행정 공무원 포상

파이낸셜뉴스 입력 :2023.06.07 14:00 수정 : 2023.06.07 14:20

통계청, 적극행정 공무원 선정 AI활용 최우수상...지표 및 프로그램 개선 등 4명 선정

통계청 Statistics Korea

(통계청 제공) /사진=뉴스1

[파이낸셜뉴스] 통계청의 '인공지능(AI) 기반 통계분류 자동화 시스템 구축' 사업 진행 등 통계 관련 적극행정 공무원을 선정하고 인센티브 등 포상에 나섰다. 적극행정 우수공무원은 온라인 국민투표와 사전심사를 참고해 '통계청 적극행정위원회'에서 최종 선정했다.

통계청은 7일 적극행정 공무원 4명을 선정하고 포상했다고 밝혔다. 적극행정 우수공무원과 우수부서에게는 인사상 인센티브, 포상휴가 등의 특전이 부여된다.

최우수상은 통계개발원 통계방법연구실의 우찬균 주무관에 주어졌다. 우 주무관은 '인공지능 기반 통계분류 자동화 시스템 구축' 사업을 진행하며 해외사례 및 선행연구를 토대로 기존의 산업·직업 자동코딩 결과보다 더 높은 분류 정확도를 확보했다. 인공지능 기반 통계분류 자동화 시스템은 기존의 자동코딩시스템을 대체하고 AI 알고리즘을 반영 등으로 업무 효율성 향상에 기여할 것으로 기대된다.

II 통계분류 자료처리 현황

Ⅱ. 통계분류 자료처리 현황

통계청
통계개발원

2-1. 통계분류의 중요성

시·도별	사업체 수				종사자 수			
	2020년	2021년	증감	증감률	2020년	2021년	증감	증감률
전 국	6,032,022 (100.0)	6,075,912 (100.0)	43,890 (0.0)	0.7	24,813,449 (100.0)	24,992,505 (100.0)	179,056 (0.0)	0.7
수도권 (서울, 인천, 경기)	2,972,805 (49.3)	2,976,201 (49.0)	3,396 (-0.3)	0.1	12,964,045 (52.2)	13,037,811 (52.2)	73,766 (-0.1)	0.6
비수도권	3,059,217 (50.7)	3,099,711 (51.0)	40,494 (0.3)	1.3	11,849,404 (47.8)	11,954,694 (47.8)	105,290 (0.1)	0.9
서 울	1,211,053 (20.1)	1,187,013 (19.5)	-24,040 (-0.5)	-2.0	5,868,926 (23.7)	5,816,141 (23.3)	-52,785 (-0.4)	-0.9
부 산	402,003 (6.7)	401,151 (6.6)	-852 (-0.1)	-0.2	1,537,281 (6.2)	1,543,764 (6.2)	6,483 (0.0)	0.4
대 구	283,033 (4.7)	279,727 (4.6)	-3,306 (-0.1)	-1.2	1,010,557 (4.1)	1,009,758 (4.0)	-799 (0.0)	-0.1
인 천	306,108 (5.1)	308,768 (5.1)	2,660 (0.0)	0.9	1,208,269 (4.9)	1,228,483 (4.9)	20,214 (0.0)	1.7
광 주	170,085 (2.8)	170,958 (2.8)	873 (0.0)	0.5	667,435 (2.7)	675,942 (2.7)	8,507 (0.0)	1.3
대 전	164,406 (2.7)	164,038 (2.7)	-368 (0.0)	-0.2	691,264 (2.8)	688,267 (2.8)	-2,997 (0.0)	-0.4
울 산	117,247 (1.9)	115,389 (1.9)	-1,858 (0.0)	-1.6	543,424 (2.2)	549,102 (2.2)	5,678 (0.0)	1.0
세 종	28,490 (0.5)	30,481 (0.5)	1,991 (0.0)	7.0	139,731 (0.6)	152,325 (0.6)	12,594 (0.0)	9.0
경 기	1,455,644 (24.1)	1,480,420 (24.4)	24,776 (0.2)	1.7	5,886,850 (23.7)	5,993,187 (24.0)	106,337 (0.3)	1.8
강 원	193,074 (3.2)	200,252 (3.3)	7,178 (0.1)	3.7	704,983 (2.8)	713,613 (2.9)	8,630 (0.0)	1.2
충 북	191,265 (3.2)	194,446 (3.2)	3,181 (0.0)	1.7	808,018 (3.3)	817,157 (3.3)	9,139 (0.0)	1.1
충 남	253,192 (4.2)	260,953 (4.3)	7,761 (0.1)	3.1	1,064,810 (4.3)	1,078,195 (4.3)	13,385 (0.0)	1.3
전 북	225,964 (3.7)	231,084 (3.8)	5,120 (0.1)	2.3	794,929 (3.2)	790,945 (3.2)	-3,984 (0.0)	-0.5
전 남	228,219 (3.8)	234,151 (3.9)	5,932 (0.1)	2.6	847,692 (3.4)	856,594 (3.4)	8,902 (0.0)	1.1
경 북	321,061 (5.3)	328,539 (5.4)	7,478 (0.1)	2.3	1,225,829 (4.9)	1,240,651 (5.0)	14,822 (0.0)	1.2
경 남	387,177 (6.4)	392,500 (6.5)	5,323 (0.0)	1.4	1,494,560 (6.0)	1,517,833 (6.1)	23,273 (0.0)	1.6
제 주	94,001 (1.6)	96,042 (1.6)	2,041 (0.0)	2.2	318,891 (1.3)	320,548 (1.3)	1,657 (0.0)	0.5

산업별	2020년	구성비	2021년	구성비	증감	증감률
전 체	6,032,022	100.0	6,075,912	100.0	43,890	0.7
A. 농림어업	12,707	0.2	11,409	0.2	-1,298	-10.2
B. 광업	2,205	0.0	2,097	0.0	-108	-4.9
C. 제조업	579,645	9.6	579,363	9.5	-282	0.0
D. 전기·가스·증기업	74,092	1.2	88,491	1.5	14,399	19.4
E. 수도·하수·폐기업	13,097	0.2	13,389	0.2	292	2.2
F. 건설업	471,217	7.8	484,909	8.0	13,692	2.9
G. 도·소매업	1,567,298	26.0	1,536,031	25.3	-31,267	-2.0
H. 운수업	593,274	9.8	616,884	10.2	23,610	4.0
I. 숙박·음식점업	865,333	14.3	862,544	14.2	-2,789	-0.3
J. 정보통신업	113,304	1.9	120,425	2.0	7,121	6.3
K. 금융·보험업	64,108	1.1	63,054	1.0	-1,054	-1.6
L. 부동산업	278,339	4.6	284,479	4.7	6,140	2.2
M. 전문·과학·기술업	221,460	3.7	222,951	3.7	1,491	0.7
N. 사업시설·지원업	151,264	2.5	142,151	2.3	-9,113	-6.0
O. 공공행정	12,480	0.2	12,578	0.2	98	0.8
P. 교육서비스업	234,741	3.9	248,928	4.1	14,187	6.0
Q. 보건·사회복지업	157,988	2.6	161,490	2.7	3,502	2.2
R. 예술·스포츠·여가업	143,161	2.4	143,454	2.4	293	0.2
S. 협회·기타서비스업	476,309	7.9	481,285	7.9	4,976	1.0

Ⅱ. 통계분류 자료처리 현황

2-2. 통계분류 기준 (산업분류의 경우)

한국표준산업분류

- 연혁: 1958년 제정된 UN 국제표준산업분류에 기초
  현재 10차 개정 한국표준산업분류 체계 운영중
- 목적: 산업 관련 통계자료의 정확성, 비교성 확보
  세제감면, 중소기업지원 등의 정책 집행 기준
  (현재 150개 법령에서 산업분류 준용)
- 분류기준: 생산된 재화 또는 서비스 중에서
  부가가치가 가장 많이 창출되는 산업활동
  - 산출물의 특성: 물리적 구성,가공단계,수요처,용도 등
  - 투입물의 특성: 원재료,생산공정,생산기술,시설 등
  - 생산활동의 일반적인 결합형태
- 분류구조: 대분류(17)-중분류(77)-소분류(232)
  -세분류(495)-세세분류(1,196) 5단계

분류단계별 항목 수

대분류	중분류	소분류	세분류	세세분류
A 농업, 임업 및 어업	3	8	21	34
B 광업	4	7	10	11
C 제조업	25	85	183	477
D 전기, 가스, 증기 및 공기조절 공급업	1	6	5	9
E 수도, 하수 및 폐기물 처리, 원료재생업	4	6	14	19
F 건설업	2	8	15	45
G 도매 및 소매업	3	20	61	184
H 운수 및 창고업	4	11	19	48
I 숙박 및 음식점업	2	4	9	29
J 정보통신업	6	11	24	42
K 금융 및 보험업	3	8	15	32
L 부동산업	1	2	4	11
M 전문, 과학 및 기술서비스업	4	14	20	51
N 사업시설관리, 사업지원 및 임대서비스업	3	11	22	32
O 공공행정, 국방 및 사회보장 행정	1	5	8	25
P 교육서비스	1	7	17	33
Q 보건업 및 사회복지 서비스업	2	6	9	25
R 예술, 스포츠 및 여가관련 서비스업	2	4	17	43
S 협회 및 단체, 수리 및 기타 개인서비스업	3	8	18	41
T 가구 내 고용활동, 자가소비 생산활동	2	3	3	3
U 국제 및 외국기관	1	1	1	2

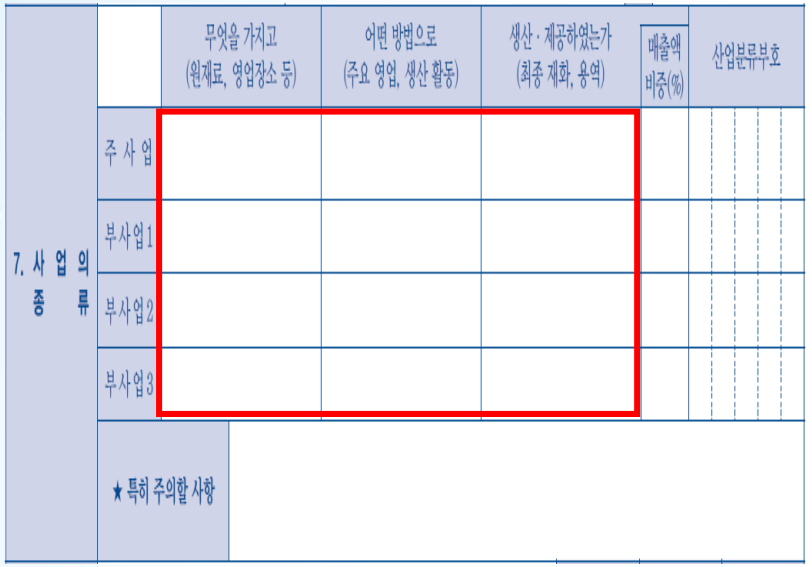
Ⅱ. 통계분류 자료처리 현황

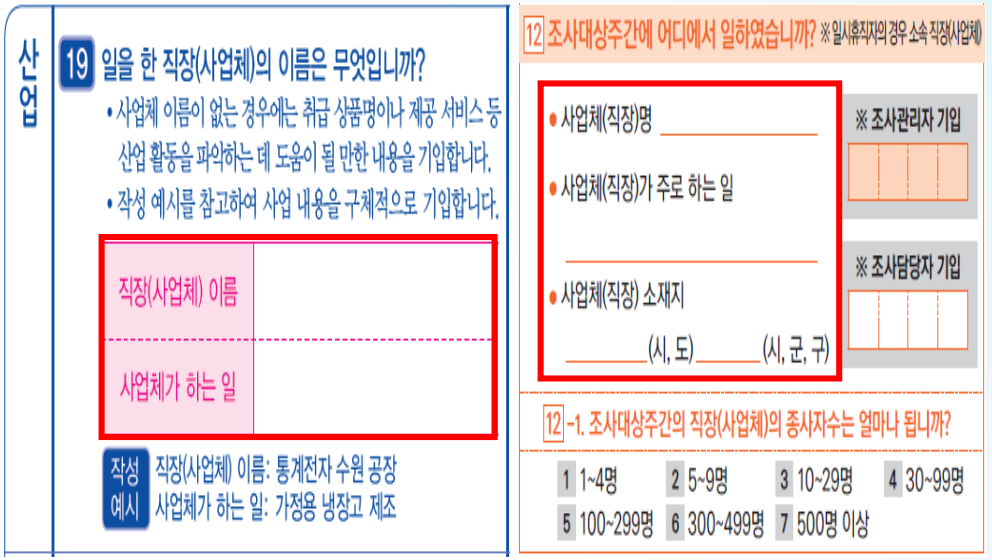
통계청
통계개발원

2-3. 통계분류 조사문항 (산업분류의 경우)

▪ [사업체 대상 조사] 산업분류 조사문항
① 무엇을 가지고(원재료,영업장소)
② 어떤 방법으로(주요 영업, 생산활동)
③ 생산 · 제공하였는가(최종 재화,용역)

▪ [가구 대상 조사] 산업분류 조사문항
① 사업체(직장)명
② 사업체(직장)가 주로 하는 일
(③ 사업체(직장) 소재지)





Ⅱ. 통계분류 자료처리 현황

2-4. 통계분류 조사문항 작성예시 (산업분류의 경우)

▪ 조사문항을 구체적으로 기입

	무엇을가지고 어디에서?	어떤 방법으로?	무엇을 생산·공급하는지?
광업	바다에서	염수를 증발시켜	천일염 생산
제조업	판유리를	가공하여	강화유리 제조

구분	조사해야할사항
음식료품	재료(해산물, 육지동물, 과실, 채소 등 구분), 제품 종류
섬유제품	제품의 종류(실, 원단, 의복, 천막, 이불, 융단, 장갑) - 모자, 장갑 등: 재료(섬유, 비닐, 가죽 등) - 의복: 재료(섬유, 가죽), 용도(정장, 한복, 평상복, 체육복)
종이제품	제품의 종류, 가공단계(펄프, 종이원지, 종이제품)
화학제품	제품의 종류, 제품의 용도 - 일반적 제품: 제품의 종류만 기입(비료, 농약, 의약품 등) - 공업용 제품: 제품의 정확한 명칭, 재료, 사용처, 종류
금속제품	재료(철,구리 등), 종류 및 용도, 가공방법(주조, 단조 등)
기계류	제품의 종류, 용도 및 사용처(건설용, 농업용, 가정용)
가구류	제품의 재료(나무, 플라스틱, 금속), 용도(주방용 등)
기계 부속품	명칭, 재료, 가공방법, 용도 및 사용처

Ⅱ. 통계분류 자료처리 현황

2-5. 현장조사 (산업분류의 경우)

사례	처리방법
나이키, 아디다스, 노스페이스 등 신발과 의류를 같이 판매하는 매장의 산업분류	- 신발(운동화)의 판매액이 큰 경우 47430(신발소매업)으로 분류 - 의류의 판매액이 큰 경우 47419 (기타 의복소매업)으로 분류 - 등산화, 등산용품의 매출이 큰 경우 47631(운동 및 경기용품 소매업)로 분류
예식장으로 분류되어 있는 사업체가 음식점업을 겸업하면서 예식장 사용료는 무료일 경우 산업분류	- 예식장 등은 설립목적에 따라 분류하므로 매출액과 관계없이 주된 사업인 96991 (예식장업)로 분류
장의사업의 산업분류	- 장의차량만 운영하는 경우 49233 (특수여객자동차운송업) - 장례식을 준비하거나 장례식장을 운영 96921(장례식장 및 장의관련 서비스업)
슈퍼마켓과 기타 음식료품 위주 종합소매업 구분	- 매장면적이 165~3,000㎡인 경우 47121(슈퍼마켓) - 매장면적이 165㎡미만인 경우 47129(기타 음식료품 위주 종합소매업)
전자담배 산업분류	- 전자담배 기기를 제조하면 28519 (기타 가정용 전기기기 제조업) - 전자담배 에센스를 제조하면 12000 (담배제품 제조업) - 전자담배를 도매 46333(담배도매업) - 전자담배를 소매 47222(담배소매업)
드론 교육학원 산업분류	- 스포츠 및 오락용 드론 조종학원 85613(레크레이션 교육기관) - 전문 직업인 양성학원 85669(기타 기술 및 직업훈련학원) - 산업용드론 조종 일반학원 85661(운전학원)
카카오톡 등에 이모티콘을 제작해서 판매하는 업체 산업분류	- 일반 대중이 사용할 수 있도록 이모티콘을 개발 공급할 경우 58222 (응용소프트웨어 개발 및 공급업) - 주문에 의해서 이모티콘을 개발 공급 62010(컴퓨터 프로그래밍 서비스업) - 이미지 및 시각적인 전달표현이 요구 되는 고도의 전문지식이 필요한 경우 73203(시각디자인업)

- [조사요원 교육]
  - 제한된 교육시간 내 산업분류 체계에 대한 이해 곤란
- [조사지침서 및 사례집]

	무엇을가지고 어디에서?	어떤 방법으로?	무엇을 생산·공급하는지?
광업	바다에서	염수를 증발시켜	천일염 생산
제조업	판유리를	가공하여	강화유리 제조

  - 부가가치가 가장 많이 창출되는 산업활동 조사
  - 다양한 특수사례 존재 ex) 예식사업 vs 장의사업
- [통계분류포털 검색 및 시스템 내검]
  - 정확한 키워드 검색 필요, 시스템부하로 인한 제한
- [콜센터 등 질의응답]
  - 현장조사 기간 중 문의의 상당부분이 통계분류 관련 사항

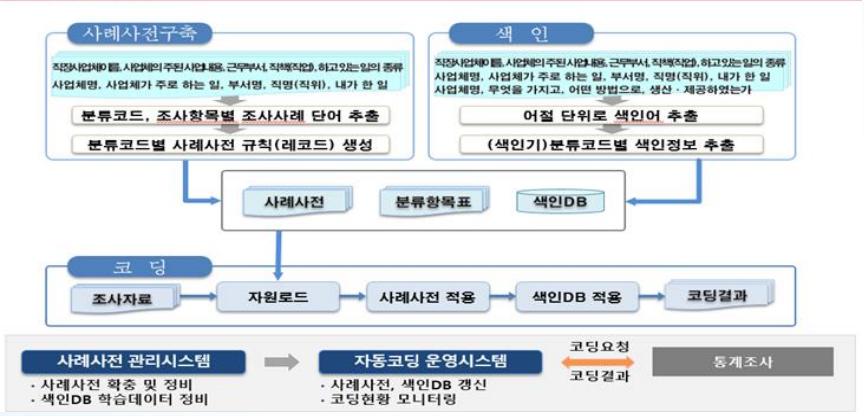
=> 1,196개 산업분류 체계에 대한 숙지 필요
현장조사 단계에서 정확한 통계분류코드 기입 어려움

Ⅱ. 통계분류 자료처리 현황

2-6. 본청내검 (산업분류의 경우)

▪ 산업분류 자동코딩시스템 전산연계

[시스템 구성]



사례사전 이용 코딩

사례사전(산업/직업분류를 한정할 수 있는 단어들의 조합으로 이뤄진 규칙베이스)에 따라 분류코드 부여

구 분	수용조건						
	사업체명		사업내용		부서/직위	하는일	
	1	2	1	2		1	2
산 업	초등학교		교육기관		교사:선생	학생지도	
직 업	초등학교		교육기관		교사:선생	학생지도	

구 분	배타조건	결과		
	배타조건	사전 번호	우선 순위	분류 코드
산 업	유아교육:유치원	1004	1	8512
직 업	교감:교장:방과후:보육:보조:유치원:특수	108	1	2522

색인 DB 이용 분류

조사자료에서 어절단위로 추출한 색인어를 색인DB에서 검색하여, 검색된 색인어의 빈도 정보로 검색점수를 계산

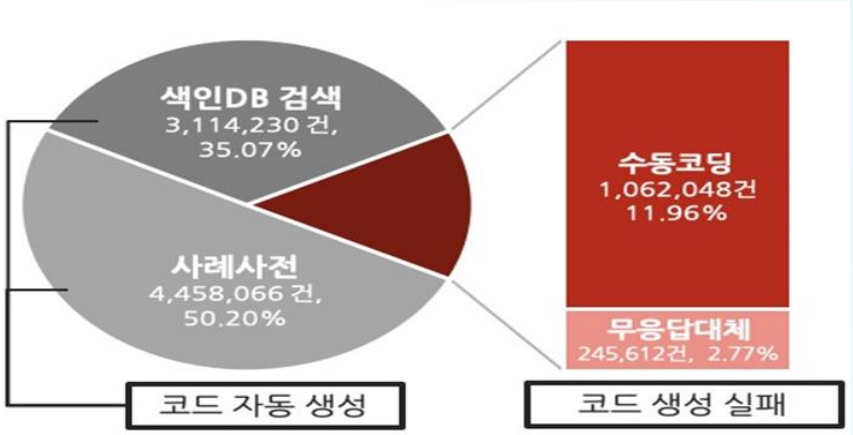
사업체명	사업내용	근무부서	직책	하는일의 종류
○○초등학교	초등 교육기관	교무실	교감	학생관리

어절 단위 색인어 추출	초등학교, 초등, 교육기관, 교무실, 교감, 학생관리
검색점수 계산	초등학교 (26.2) + 초등(14.1) + 교육기관 (8.7) + 교무실(14.7) + 교감(12.6) + 학생관리(11.0)] × 가중치 (1.0)
색인DB 검색 결과	산업분류 8512, 빈도수 값 87.3

▪ [규칙(사례사전) 기반 전산연계]

- 업무담당자의 전문적인 판단에 의한 내검규칙 적용
- 높은 정확도 but 커버리지 문제

ex) 가구 대상 조사의 경우
세분류(495개) 기준 약 34,000개 사례사전 구비



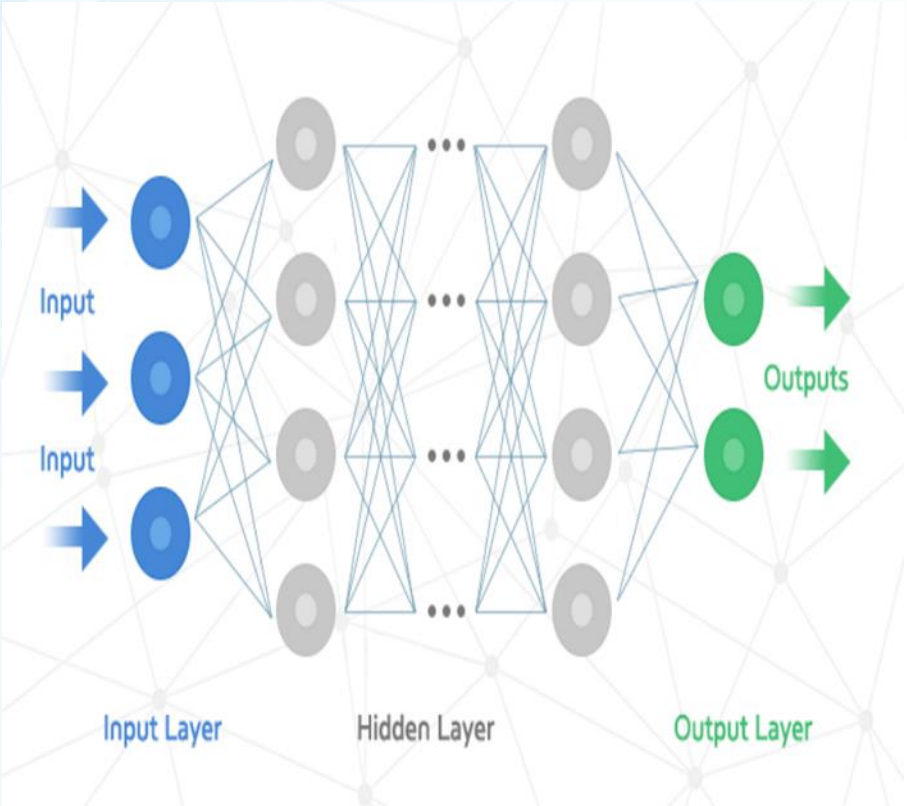
=> 내검원에 의한 내용검토 업무량 증가 문제 심화
(총조사 및 대규모 행정자료 이용의 경우)

AI 기반 통계분류 자동화 연구

Ⅲ. AI 기반 통계분류 자동화 연구

3-1. 인공지능(Artificial Intelligence)

인공신경망



분류(classification)

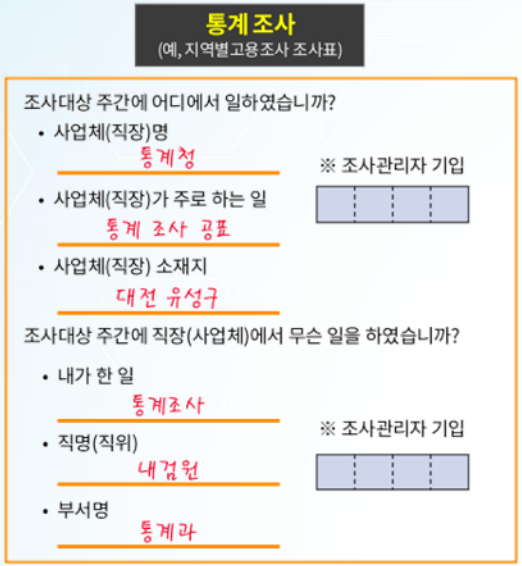
Ⅲ. AI 기반 통계분류 자동화 연구

통계청 통계개발원

3-2. AI 기반 통계분류 자동화 방법론 개요

해결할 문제

통계조사표
(텍스트 입력)



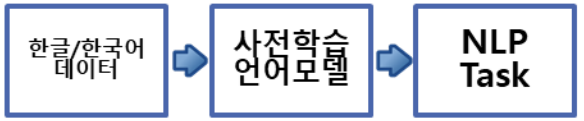
통계조사 자료 분석 및 전처리

활용 방법론

데이터 마이닝
(비정형 데이터 처리)

기계학습 모델링

비지도학습



전이학습

지도학습

자연어입력 **임베딩** **분류결과**

업무 적용 방향

시스템 화
(통계생산 프로세스)

내재화 **최적화**

한국표준산업코드

8	4	1	2

정부기관 일반 보조 행정

한국표준직업코드

3	9	1	0

통계관련 사무원

나라통계 2025 Census

Ⅲ. AI 기반 통계분류 자동화 연구

통계청 통계개발원

3-3. AI 기반 통계분류 자동화 방법론(1): 자연어처리

어휘 분절 방법론 선정

선정 기준	
▪ **특정 형태소 분석기에 의존적이지 말 것** – 충분한 어휘를 갖는 분절 방법은 형태소 분석과 비슷한 결과를 냄 – 라이센스 및 시스템 유지보수 상의 문제	**형태소 단위 분절 X**
▪ **한국어 서브워드 분절을 지원 할 것** – 복합어 등을 분해하는데 이용 – 학습되지 않은 표현(OOV)에도 강건함 – 응용 도메인에 특화하여 최소 분절 어휘 구축	**BPE** (Byte-pair encoding) 방법론
▪ **특정 사전학습 언어모델에 특화된 분절 방법은 피할 것** – 향후 사전학습 모델 교체, 개선 등 고려 – XLM-RoBERTa, KoELECTRA, GPT 등에서 적용	**WPM** (WordPiece Model)

사전학습 언어모델 선정

선정 기준	
▪ **한국어 서브워드 분절을 지원 할 것** – 복합어 등을 분해하는데 이용 – 분절의 등장 n-gram/빈도/확률에 따라 분리	**W2V, Globe X**
▪ **공개 사전학습 언어모델을 이용할 것** – 향후 사전학습 모델 교체, 개선 등 고려	**BERT 등**
▪ **공개 모델의 분절 어휘 수가 적거나 특정 사전학습 언어모델에 특화된 분절 방법은 피할 것** – SK KoBERT: 어휘 개수 8,002개로 적음 – ETRI KorBERT: 자체 어휘 분절 라이브러리 제공	**XLM-RoBERTa,** KoELECTRA

예시) 원문장: “오늘의 연합뉴스 기사입니다”
분절결과: ['_오늘', '의', '_연합뉴스', '_기사', '입니다']

임베딩

원문장: “오늘의 연합뉴스 기사입니다”
임베딩 결과: [6.7092e-02, -5.0252e-02, 1.1564e-01, … 1.5695e-01]

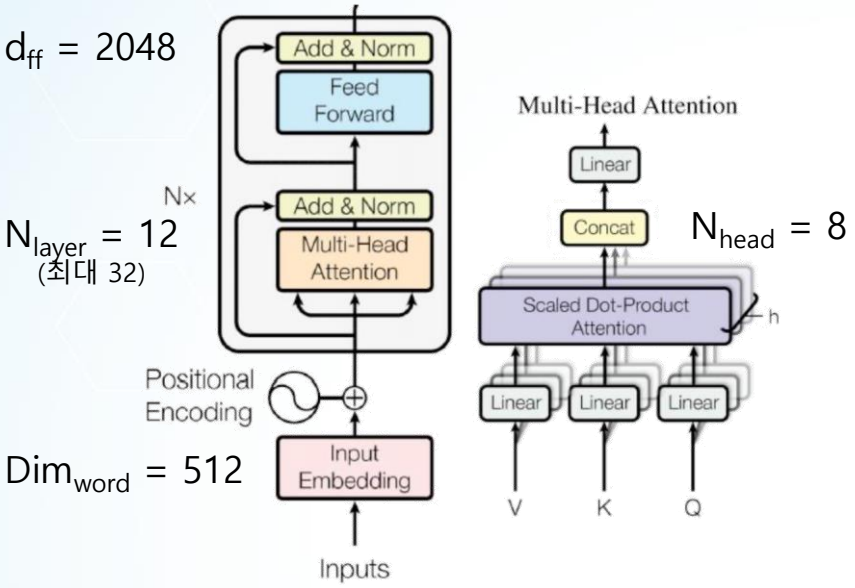
768 차원 실수 벡터

Ⅲ. AI 기반 통계분류 자동화 연구

3-3. AI 기반 통계분류 자동화 방법론(1): 언어모델 선정

▪ BERT [Jacob Devlin, et al. 2018]

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- 양방향 Transformer 구조에서 Attention 메커니즘을 통해 관련있는 단어(토큰)들 간의 연결 관계 정보를 학습



▪ RoBERTa [Yinhan Liu, et al, 2019]

- 모델의 하이퍼 파라미터가 일반적인 태스크 성능에 영향을 끼치는지에 대해 실험하여 최적화 방법론을 찾음
- 다양한 NLP Task 학습데이터를 통해 Multi-task Learning
- BERT 모델의 학습 말뭉치 및 하이퍼 파라메터 설정에 따라 종류가 파생

- 파생 모델 (시스템 적용 모델)
  - 다국어병렬 말뭉치: xlm-RoBERTa [Google. 2020]
  - 한국어 말뭉치: KLUE-RoBERTa [박성준 외, 2021]

- 하이퍼 파라메터에 따른 분류

Model	Embedding size	Hidden size	# Layers	# Heads
KLUE-BERT-base	768	768	12	12
KLUE-RoBERTa-small	768	768	6	12
KLUE-RoBERTa-base	**768**	**768**	**12**	**12**
KLUE-RoBERTa-large	1024	1024	24	16

Ⅲ. AI 기반 통계분류 자동화 연구

[참고] 인공지능 자연어처리

- 사람의 언어를 컴퓨터가 알아듣도록 처리하는 기술과 일련의 과정

  (사전학습) 언어모델 → 임베딩 → 자연어처리 Task

계산 가능한 형식으로 변환 필요

사람의 언어 (자연어)

컴퓨터의 언어 (숫자, 벡터)

자연어 이해 　　연산 또는 처리

Up-stream Task

- BERT, Electra, T5, GPT

Down-stream Task

- 분류, 기계 독해, 번역, 요약

(사전학습) 언어모델

- 특정 과제(task) 학습에 앞서 자연어를 잘 임베딩 하는 모델
- (ex: GPT, BERT, Electra, T5, etc...)

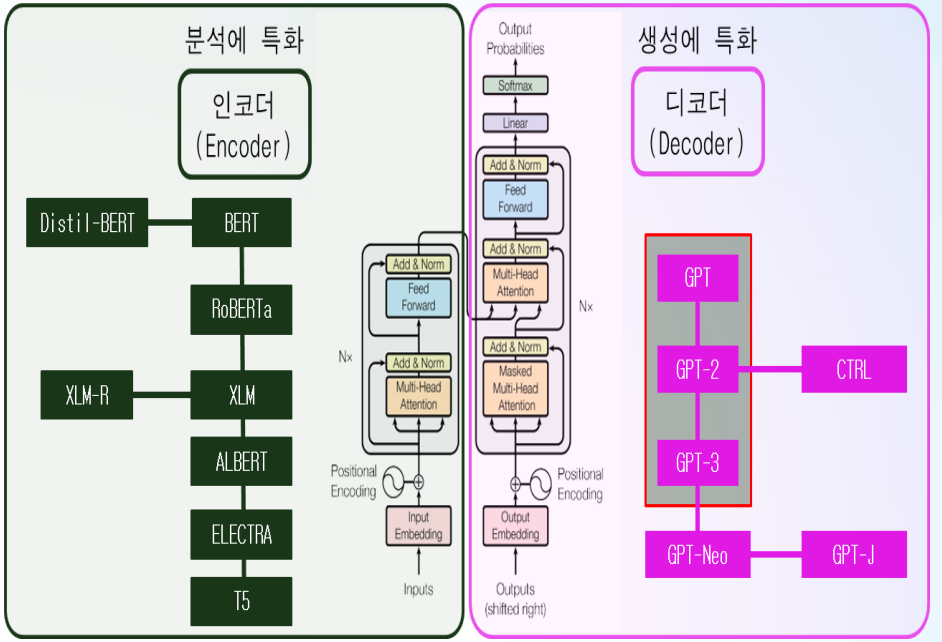
임베딩

- 자연어를 숫자의 나열인 벡터로 바꾼 결과 혹은 그 일련의 과정

- **트랜스포머 모델[Ashish Vaswani, et al.2017]**

  특정 표현에 대한 고유한 정보로 단어 간의 상호 등장(Co-occurrence) 관계 정보를 사용

  → 문장 내 표현의 선행 단어(토큰)로부터 다음에 등장할 단어(토큰)의 확률을 학습/예측



Ⅲ. AI 기반 통계분류 자동화 연구

[참고] 적용 언어모델 성능 벤치마크(1)

시스템 적용 기반 사전학습 언어모델 1: KcBERT-Large

- **한국어 자연어처리 태스크에 따른 성능 벤치 마크**
  - 학습 데이터: Korean Comments (KC)
    - 출처: [Release Train Data Release: v2022.3Q · Beomi/KcBERT · GitHub (2023년 09월 현재)
    - 대상:
      - 한국어 위키, 뉴스 기사, 책 등 잘 정제된 데이터
      - 온라인 뉴스에서 댓글과 대댓글 추가
    - 전체 데이터 수(공백열 제외): 345,452,030 (3.4억건)

KcBERT: Korean comments BERT

** Updates on 2022.11.07 **

- KcELECTRA v2022 학습에 사용한, 확장된 텍스트 데이터셋(v2022.3Q)를 공개합니다.
- https://github.com/Beomi/KcBERT/releases/tag/v2022.3Q
- 기존 11GB -> 신규 45GB, 기존 0.9억건 -> 신규 3.4억건으로 기존 v1 데이터셋 대비 약 4배 증가한 데이터셋입니다.

** Updates on 2022.10.08 **

- KcELECTRA-base-v2022 (구 dev) 모델 이름이 변경되었습니다.
- 기존 KcELECTRA-base(v2021) 대비 대부분의 downstream task에서 ~1%p 수준의 성능 향상이 있습니다.

** Updates on 2022.09.14 **

- emoji의 v2.0.0 업데이트됨에 따라 Preprocessing 코드가 일부 변경되었습니다.

출처: https://github.com/Beomi/KcELECTRA
(2022년 09월 공개)

KcELECTRA Performance

- Finetune 코드는 https://github.com/Beomi/KcBERT-finetune 에서 찾아보실 수 있습니다.
- 해당 Repo의 각 Checkpoint 폴더에서 Step별 세부 스코어를 확인하실 수 있습니다.

	Size (용량)	NSMC (acc)	Naver NER (F1)	PAWS (acc)	KorNLI (acc)	KorSTS (spearman)	Question Pair (acc)	KorQuaD (Dev) (EM/F1)
KcELECTRA-base-v2022	475M	**91.97**	87.35	76.50	82.12	83.67	95.12	69.00 / 90.40
KcELECTRA-base	475M	91.71	86.90	74.80	81.65	82.65	**95.78**	70.60 / 90.11
KcBERT-Base	417M	89.62	84.34	66.95	74.85	75.57	93.93	60.25 / 84.39
KcBERT-Large	1.2G	90.68	85.53	70.15	76.99	77.49	94.06	62.16 / 86.64
KoBERT	351M	89.63	86.11	80.65	79.00	79.64	93.93	52.81 / 80.27
XLM-Roberta-Base	1.03G	89.49	86.26	82.95	79.92	79.09	93.53	64.70 / 88.94
HanBERT	614M	90.16	87.31	82.40	80.89	83.33	94.19	78.74 / 92.02
KoELECTRA-Base	423M	90.21	86.87	81.90	80.85	83.21	94.20	61.10 / 89.59
KoELECTRA-Base-v2	423M	89.70	87.02	83.90	80.61	84.30	94.72	84.34 / 92.58
KoELECTRA-Base-v3	423M	90.63	**88.11**	**84.45**	**82.24**	**85.53**	95.25	**84.83 / 93.45**
DistilKoBERT	108M	88.41	84.13	62.55	70.55	73.21	92.48	54.12 / 77.80

출처: https://github.com/Beomi/KcELECTRA
(2023년 09월 현재)

Ⅲ. AI 기반 통계분류 자동화 연구

[참고] 적용 언어모델 성능 벤치마크(1)

시스템 적용 기반 사전학습 언어모델 2: Klue-RoBERTa-Large, xlm-RoBERTa-Large

- **자연어처리 Task 에서의 성능 벤치 마크**
  - 다양한 자연어처리 태스크에 전체적 성능이 우수
  - 다국어 병렬 말뭉치로 학습한 언어모델로 다양한 언어 대응 가능
  - 출처: fairseq/examples/roberta/README.md at main · facebookresearch/fairseq · GitHub (2023년 9월 현재)

Results

GLUE (Wang et al., 2019) *(dev set, single model, single-task finetuning)*

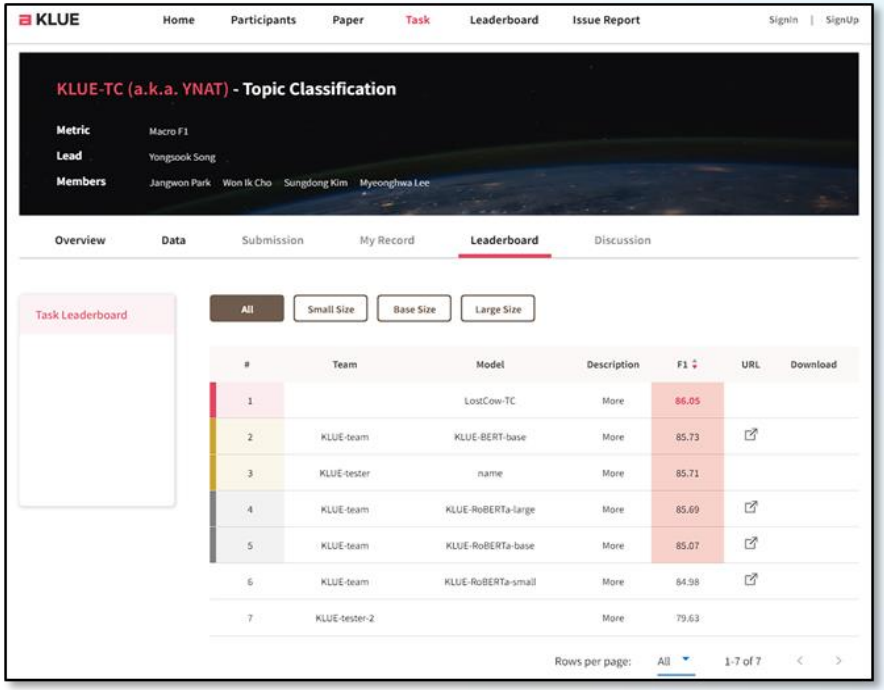
Model	MNLI	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B
roberta.base	87.6	92.8	91.9	78.7	94.8	90.2	63.6	91.2
roberta.large	90.2	94.7	92.2	86.6	96.4	90.9	68.0	92.4
roberta.large.mnli	90.2	-	-	-	-	-	-	-

SuperGLUE (Wang et al., 2019) *(dev set, single model, single-task finetuning)*

Model	BoolQ	CB	COPA	MultiRC	RTE	WiC	WSC
roberta.large	86.9	98.2	94.0	85.7	89.5	75.6	-
roberta.large.wsc	-	-	-	-	-	-	91.3

- **한국어 텍스트 분류(KLUE-TC) 에서의 성능**

  - 출처: https://klue-benchmark.com/tasks/66/leaderboard/task (2023년 9월 현재)

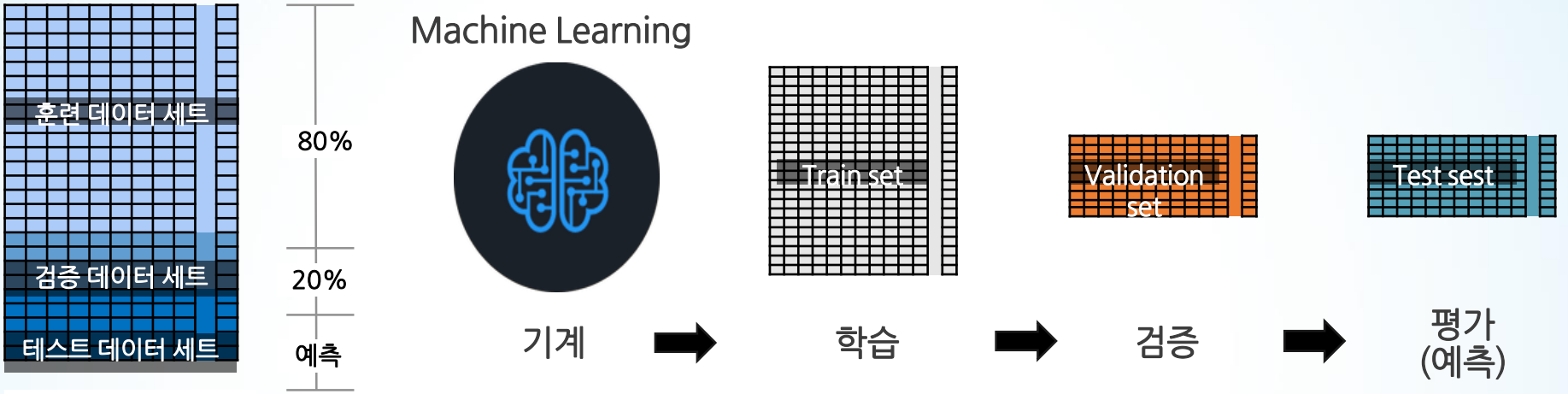


Ⅲ. AI 기반 통계분류 자동화 연구

3-4. AI 기반 통계분류 자동화 방법론(2): 지도학습

Human Learning

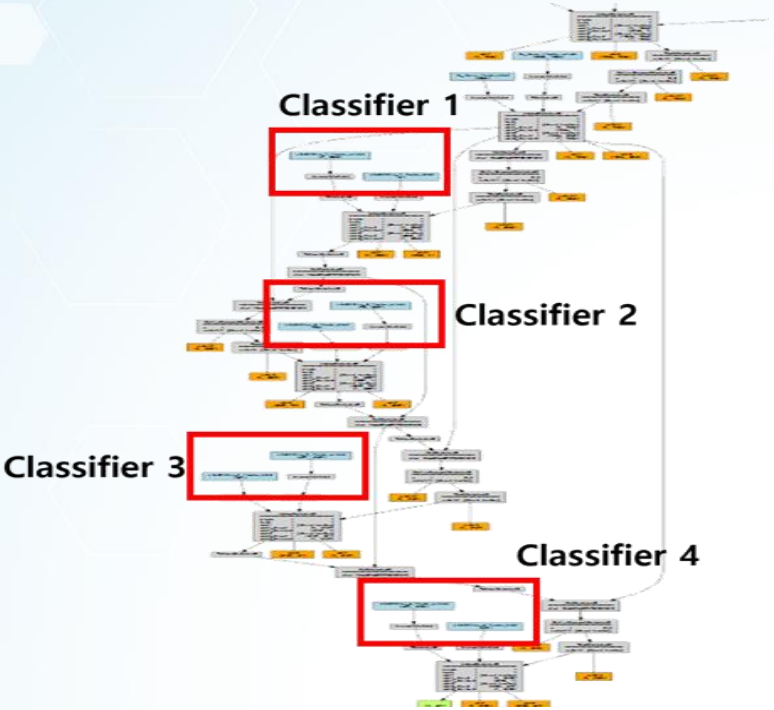
Machine Learning



Ⅲ. AI 기반 통계분류 자동화 연구

3-4. AI 기반 통계분류 자동화 방법론(2): 지도학습 방법 결정

Hierarchical Text Classification(Hi-TC) 모델



예) 산업/직업 분류 모델 구조
(세분류 기준)

▪ **기본 구성**
- 텍스트 분류 태스크의 미세조정(Finetuning)을 위한 통계자료 처리 전용 지도학습 모델
- Sequential Classifier 로 이루어진 Text Classification 파이프라인을 계층적으로 결합한 구조

▪ **모델 특징**
- BERT 계열의 여러 사전학습 언어모델 적용 가능
- 단순 자연어 입력 뿐만 아니라 상위 계층(대 > 중 > 소 > 세 > 세세)의 분류 결과를 참고하여 다음 단계의 분류 코드를 학습 및 예측

▪ **특장점**
- 여러 계층으로 구성되는 분류 체계를 사용하는 Real-Problem 분류(통계청 표준분류 등) 문제에 특화
- 조사 별로 사용하는 분류 체계에 따라 적용 수준을 조절하여 계층형 분류 결과 학습 및 예측 가능

Ⅲ. AI 기반 통계분류 자동화 연구

3-5. 분류모델 구축(5종분류, 5개조사) 및 성능평가 결과

분류 구분	관련 부서	적용조사 (주기)	데이터양 (추이/예측)	기초 데이터 이수 건수	데이터 전처리	데이터 정제	데이터 분리	학습/평가 품질 (Accuracy / F-1)	구축횟수 (최종학습일)
한국표준 산업 (10차)	고용 통계과	지역별 고용조사 (연 2회)	약 43만 (6개월 간격)	총 347만 (44~51회 총 8회차 19~22 4년 자료)	결측: 109만 중복: 32만	122만	학습: 110만 검증: 12만 평가: 43만	검증: **92.74%** / **0.9271** 평가: **94.17%** / **0.9414**	4회 (12/14)
	경제 총조사과	전국사업체 조사 (연 1회)	약 800만 (1년 간격)	총 2,594만 (18~21년)	결측: 484만 중복: 553만	834만	학습: 751만 검증: 83만 평가: 723만	검증: **85.04%** / **0.8508** 평가: **86.48%** / **0.8623**	9회 (01/11)
	인구 총조사과	인구총조사 (5년)	약 1,000만 (5년 간격, 전가구원 종합)	총 981만 (2020 인구총조사 기초데이터)	처리: 571만 중복: 40만 결측: 65 개	410만	학습: 369만 검증: 41만	검증: **86.01%** / **0.8610** 평가: **85.60%** / **0.8551**	6회 (11/06)
한국표준 직업 (7차)	고용 통계과	지역별 고용조사 (연 2회)	약 43만 (6개월 간격)	총 347만 (44~51회 총 8회차 19'~22' 4개년 자료)	결측: 109만 중복: 32만	122만	학습: 110만 검증: 12만 평가: 43만	검증: **91.44%** / **0.9141** 평가: **92.67%** / **0.9259**	5회 (12/14)
	인구 총조사과	인구총조사 (5년)	약 1,000만 (5년 간격, 전가구원 종합)	총 981만 (2020 인구총조사 기초데이터)	처리: 571만 중복: 40만 결측: 65	410만	학습: 369만 검증: 41만	검증: **79.52%** / **0.7915** 평가: **79.25%** / **0.7896**	6회 (11/06)
건설업 공종	산업 통계과	건설업 조사 (연 1회)	평균 180만 (1년 간격)	총 883만 (17~21년 전수)	중복: 250만	633만	학습: 192.1만 검증: 21.3만 평가: 92.1만 (21년, 내검 대상)	검증: **84.22%** / **0.8411** 평가: **85.03%** / **0.8482**	8회 (01/04)
건설업 발주자	산업 통계과	건설업 조사 (연 1회)	평균 180만 (1년 간격)	총 883만 (17~21년 전수)	중복: 294만 결측: 5만	589만	학습: 192.6만 검증: 21.4만 평가: 92.2만 (21년, 내검 대상 Y)	검증: **91.81%** / **0.9177** 평가: **89.32%** / **0.8937**	6회 (01/04)
물품통관 분류	서비스업 동향과	온라인쇼핑 동향 (월 1회)	평균 230만 (1월 간격)	총 9,730만 (21 ~ 22년 2년 자료)	중복: 8,315만 (21' 01.~22' 08.)	상품군: 93만 (22년 04월) HS코드: 1,182만 (21' 01.~22' 08.)	학습: 26만 검증: 2.9만 평가: 9.3만 (상품군, 22년 04월)	검증: **96.54%** / **0.9653** 평가: **96.86%** / **0.9685**	4회 (12/02)

Ⅲ. AI 기반 통계분류 자동화 연구

통계청 통계개발원

[참고] AI 분류 예측 정확도 평가 기준

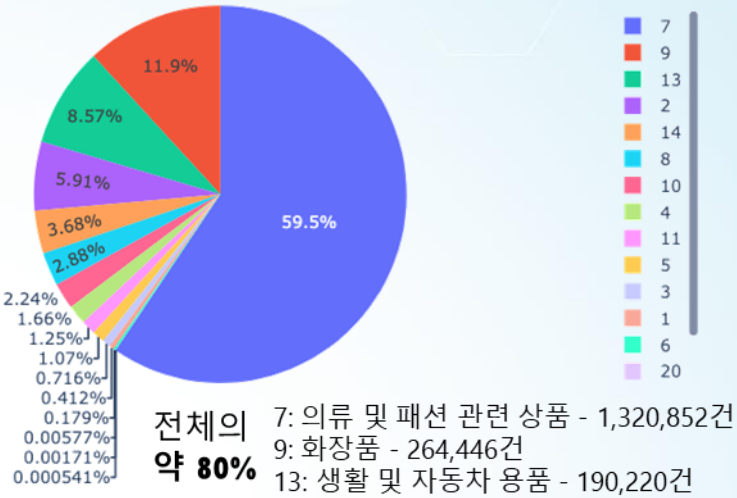
- **정분류율 (=정확도, Accuracy)**
  - 전체 개수 중 정답을 맞춘 개수 → 내검 대상 추출 및 업무 할당에 활용
    - 다중 분류 문제에서는 Weighted(=Micro) 재현율(Recall)과 같음
  - 데이터의 분포에 따라(=불균형이 심한 경우) 수치를 신뢰할 수 없음
    - 데이터 수가 많은 항목의 점수가 높을 수록 전체적인 수치가 높아짐

예) 건설업 발주자 분류 결과

code	precision	recall	f1-score	support
84	0.995169	0.976303	0.985646	211.0
85	1.000000	0.989733	0.994840	487.0
86	0.960000	1.000000	0.979592	24.0
87	0.992659	0.990581	0.991619	1911.0
accuracy	0.915626	NaN	NaN	NaN
macro avg	0.866879	0.843880	0.853297	218906.0
weighted avg	0.916240	0.915626	0.915331	218906.0

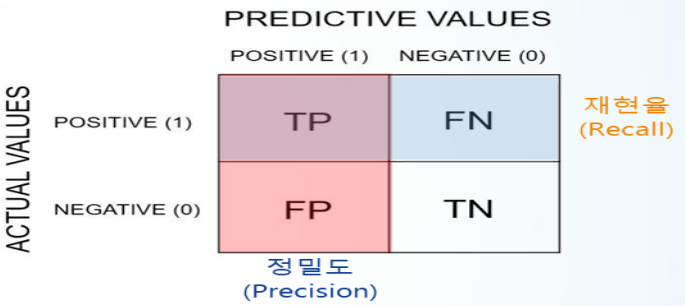
- **F-1 점수 (= 분류 항목별 예측율과 재현율의 조화평균)**
  - 정밀도(Precision): 모델 예측 결과 중 정답(GS)과 일치율
  - 재현율(Recall = 민감도): 실제 정답(GS) 중 예측 결과 일치율
  - 항목(=코드) 별로 정밀도와 재현율의 조화 평균을 구함
    - Macro 평균: 데이터 분포 고려 하지 않고 전체 항목 개수로 나눔
    - Weighted(=Micro) 평균: 데이터 표본 분포를 고려해서 가중치 반영
      = 데이터 분포 고려 → 모델 전체적인 성능 측정에 활용

예) 해외직접구매 상품군 분류 데이터 분포



전체의 약 **80%**

7: 의류 및 패션 관련 상품 - 1,320,852건
9: 화장품 - 264,446건
13: 생활 및 자동차 용품 - 190,220건



Ⅲ. AI 기반 통계분류 자동화 연구

통계청 통계개발원

[참고] 기존 시스템과 AI 통계분류 시스템 성능비교

미분류율

- 사례사전 (산업 40.06%, 직업 66.88%)
- 색인 DB (산업 0.29%, 직업 0.46%)
- **AI모델 (0%)**

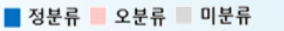
정밀도
(미분류 제외 정확도)

- 사례사전 (산업 **94.90%**, 직업 **93.18 %**)
- AI 모델 (산업 84.20%, 직업 79.64%)
- 색인 DB (산업 69.44%, 직업 63.42%)

정분류율
(미분류 건수 포함(=재현율) 정확도)

- **AI 모델 (산업 84.20%, 직업 79.64%)**
- 색인 DB (산업 69.24%, 직업 63.13%)
- 사례사전(산업 56.89%, 직업 30.86%)

분류	분류 방식	데이터 수 (건)	분류 (건) 정분류	분류 (건) 오분류	미분류 (건)	정/오/미 분류 비율 비교
산업분류	AI	409,901	**84.19%** (345,102)	15.81% (64,799)	0% (0)	
	색인DB		69.24% (283,806)	30.47% (124,908)	0.29% (1,187)	
	사례사전		56.88% (233,153)	3.06% (12,536)	40.06% (164,212)	
직업분류	AI		**79.64%** (326,446)	20.36% (83,455)	0% (0)	
	색인DB		63.13% (258,779)	36.41% (149,247)	0.46% (1,875)	
	사례사전		30.86% (126,508)	2.26% (9,261)	66.88% (274,132)	



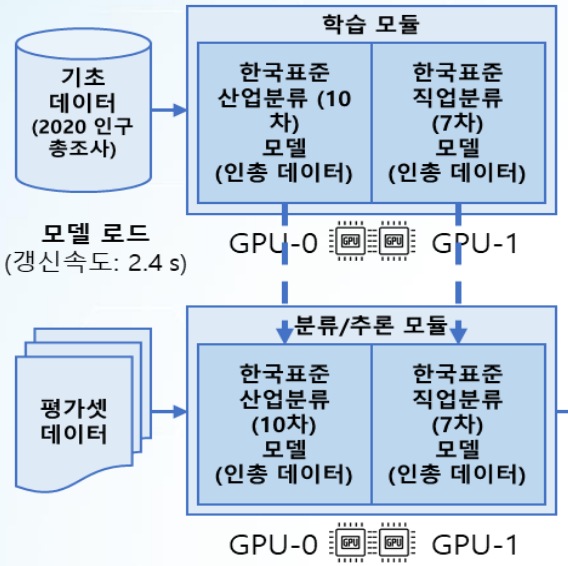
* 2020년 인구총조사 산업/직업분류 자료 분석결과

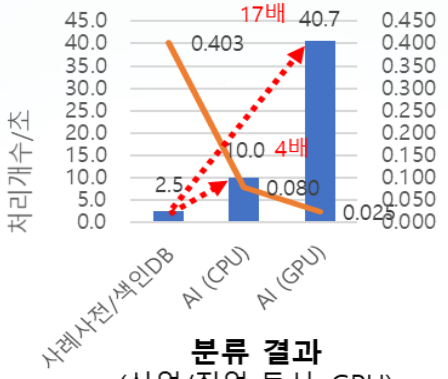
Ⅲ. AI 기반 통계분류 자동화 연구

통계청 통계개발원

3-6. 분류모델 처리속도 평가

- **2020 인구총조사 (약 20% 샘플링) 시스템 별 처리 성능 분석**
  - **현재 RMI 서버 단건 기준, CPU 4배, GPU 17배 향상**
    - 개발계 환경) NVIDIA RTX 3090(2식, 24GB), AMD Ryzen 9 5950X 16-Core(32 Thread) Processor, RAM 126GB, NVME(2식, 500GB, 2TB), SSD (500GB), HDD (2TB)





분류 결과
(산업/직업 동시, GPU)
단건: 40~41 개/s
0.0246 s/개
배치: 150 개/s
0.0067 s/개

	처리량 (개/초)	단위처리 시간 (s/개)	비율
사례사전/색인DB	2.5	0.403	
AI (CPU)	10.0	0.080	5.04
AI (GPU)	**40.7**	**0.025**	**16.39**

- **40만 개 처리 시 약 80분/모델 별 소요**
  - 산업/직업 분류 동시 처리 시 2배 소요
  - 사례사전/색인DB RMI 서버 대비 (약 100ms/건, 단 동시 처리)
    약 3~4 배 (네트워크 속도 포함 시),
    약 4~5 배 (네트워크 속도 제외) 개선

통계청
통계개발원

Ⅳ 향후 연구 추진 계획(안)

Statistical Research Institute

Ⅳ. 향후 연구 추진계획

통계청 통계개발원

4-1. AI 통계분류 결과 실무 시험적용 및 검증(1)

< 5종 조사 5개 분류 정확도 시험분석 결과 >

	기존(As-Is)	개선(To-be)
(1) 인구총조사	- 산업분류 69.2%, 직업분류 63.1%	- 산업분류 84.2%, 직업분류 79.6%
(2) 전국사업체조사	- 산업분류 73.5%	- 산업분류 84.1%
(3) 지역별고용조사	- 산업분류 87.4%, 직업분류 84.4%	- 산업분류 94.2%, 직업분류 92.0%
(4) 건설업조사	-	- 공종분류 85.0%, 발주자분류 89.3%
(5) 온라인쇼핑조사	-	- 상품군별분류 96.8%

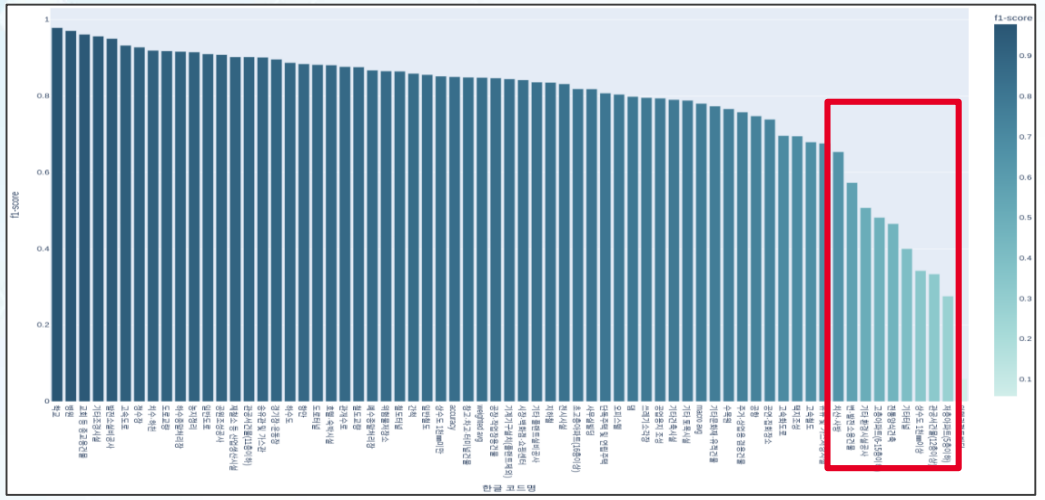
- AI 통계분류 예측결과 실무 시험적용
  - 5개분류: 산업/직업분류, 공종/발주자분류, 상품군별분류
  - 5종조사: 인구총조사, 전국사업체조사, 지역별고용조사, 건설업조사, 온라인쇼핑조사
- 기존 통계분류 자료처리와 AI 통계분류 예측결과 비교 분석
- AI 통계분류 예측 정확도 향상을 위한 학습데이터 관리방안 연구

Ⅳ. 향후 연구 추진계획

통계청 통계개발원

4-1. AI 통계분류 결과 실무 시험적용 및 검증(2)

- **AI 통계분류 예측결과를 활용하여 내검효율성 제고**



	text	prob
0	어린이집에서 보호자로부터위탁을받아 보호자로부터위탁을받아	0.898575
1	사업장에서 소매상 대상 소매상 대상	0.853393
2	영업장에서 고객의뢰를 받아 고객의뢰를 받아	0.602864
3	교회에서 기독교종교활동 기독교종교활동	0.334787
4	초등학교 초등교육서비스 초등교육서비스	0.572152

선택적 에디팅
→ 내용검토 우선순위 결정

1. 예측정확도 낮은 분류영역 내검
2. 예측정확도 낮은 개별사례 내검
3. 기존 자동코딩(사례사전) 결과와 불일치 사례 내검

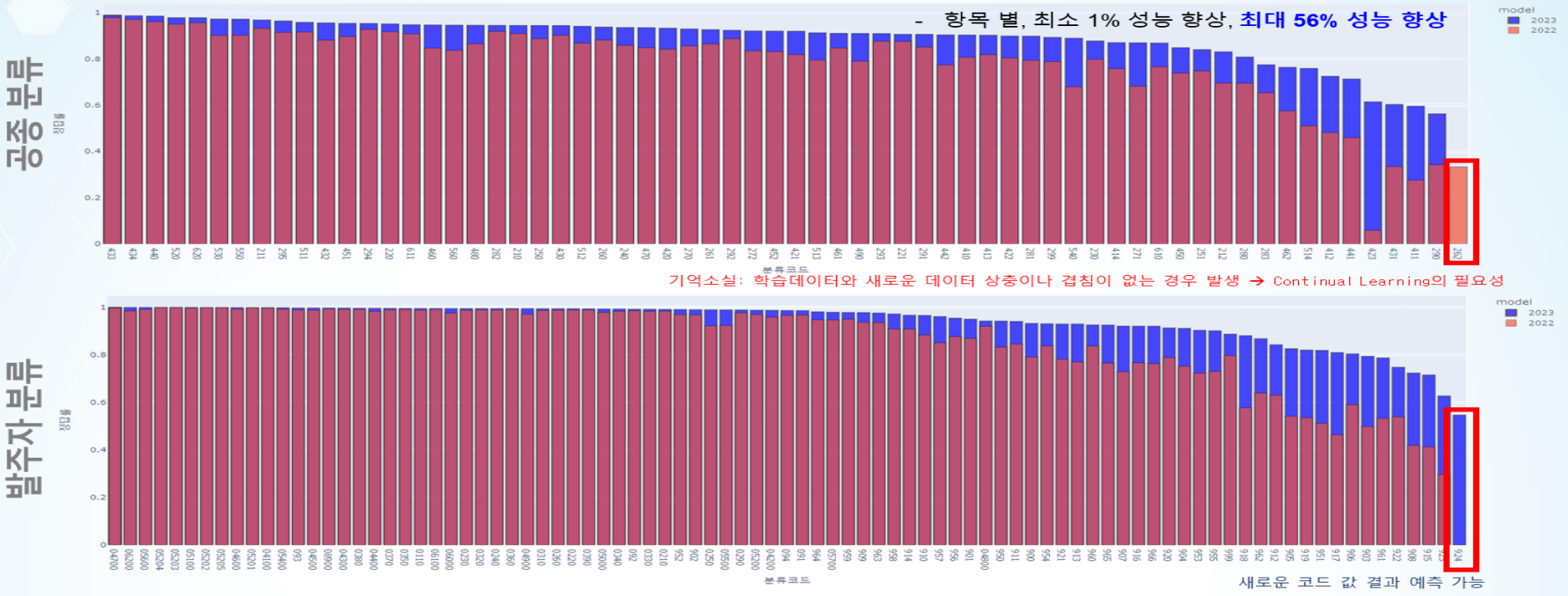
내검결과 강화학습 및 재학습
→ AI 분류예측 정확도 제고 도모

Ⅳ. 향후 연구 추진계획

통계청
통계개발원

4-1. AI 통계분류 결과 실무 시험적용 및 검증(3)

- **AI 통계분류 예측 정확도 향상을 위한 학습데이터 관리**



- 지속적 학습(Continual Learning) 최적 방법론 검토
  - 과거 학습데이터 + 신규 데이터셋을 사용하는 iCaRl [Sylvestre-Alvise Rebuff et al, 2017] 및 Rehearsal Methods 적용

Ⅳ. 향후 연구 추진계획

4-2. AI 통계분류 자동화 적용 확대(1)

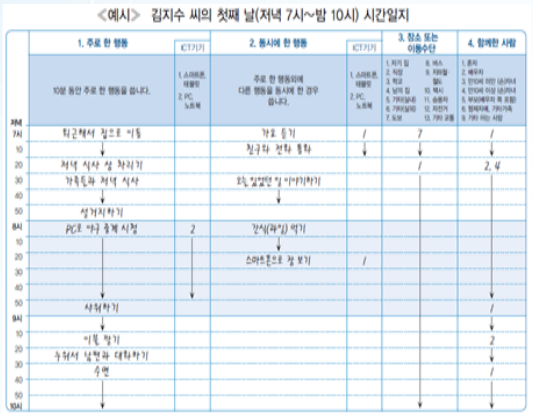
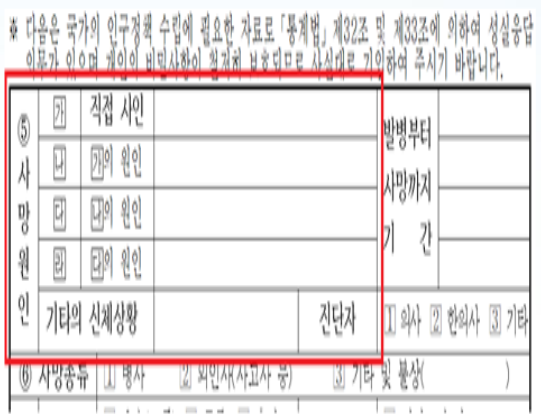
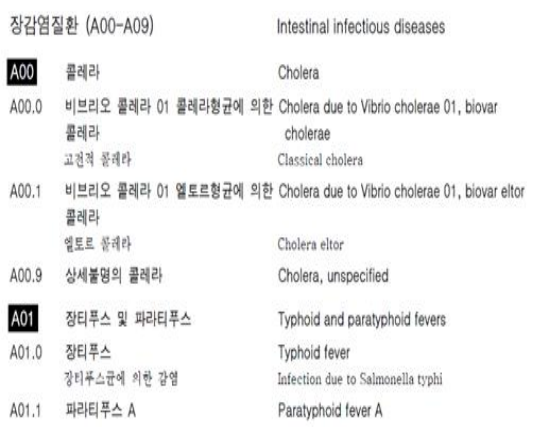
산업분류 조사 통계청 작성통계

구분	조사명	공표 주기	조사 대상	산업분류 조사항목
사업체	경제총조사	5년	390만	[세세분류단위] - 무엇을 가지고 (원재료, 영업장소 등) - 어떤 방법으로 (주요 영업, 생산 활동) - 생산·제공하였는가 (최종 재화, 용역)
	전국사업체조사	1년	490만	
	광업제조업조사	1년	7만5천	
	도소매서비스업조사	1년	10만	
	프랜차이즈업조사	1년	2만5천	
	소상공인실태조사	1년	4만4천	
가구	인구총조사	5년	전체가구	[세분류(인총·지역별고용) /중(대)분류단위] - 사업체(직장) 이름 - 사업체가 하는 일 - 사업체 소재지
	지역별고용조사	반기	23만	
	경제활동인구조사	월	3만5천	
	가계금융복지조사	1년	2만	
	가계동향조사	분기	7천	
	이민자체류실태조사	1년	2만5천	
	사회조사	1년	2만7천	
	생활시간조사	5년	2만9천	
농림어업	농림어업총조사	5년	120만	[대분류 일부단위] - 주 종사 분야 선택 (농업·임업·어업·제조업·건설업· 도소매업·숙박음식업·기타)
	농림어업조사	1년	4만5천	
	농가경제조사	1년	3천	
	어가경제조사	1년	1천	

Ⅳ. 향후 연구 추진계획

통계청
통계개발원

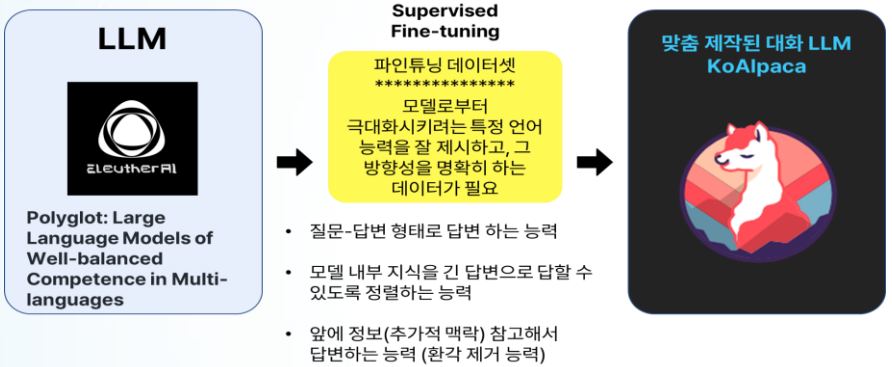
4-2. AI 통계분류 자동화 적용 확대(2)

	생활시간조사	사망원인통계	가계동향조사
조사표			
분류 코드			

Ⅳ. 향후 연구 추진계획

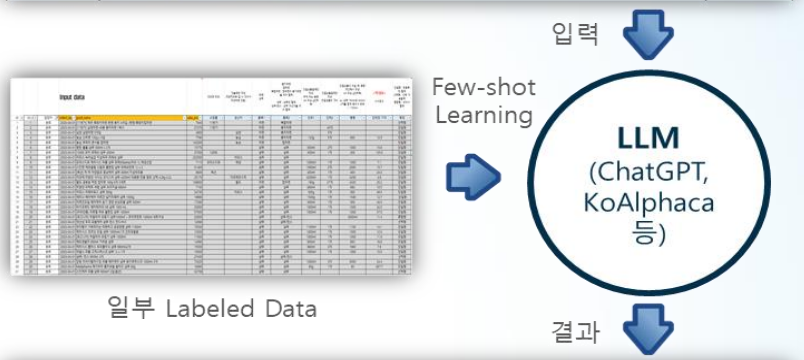
4-3. 통계분야 텍스트 분석 범용 플랫폼으로 확대 발전 검토

- **통계도메인 업무에 특화된 LLM 기술 활용 방안 모색**
  - 사용 비용에 대해 걱정을 덜 하면서 (GPT-4 토큰 1000개에 0.04 달러)
  - 도메인 특화 지식을 추가
    - 교육, 법률, 의료, 등 세부 도메인 지식은 LLM학습이 안되어 있는 경우가 많음 (OpenAI에서 미세조정은 가능하지만 커스텀 작업시 2,3배 비싼 비용 요구)
  - 데이터 유출 가능성이 낮으면서도 내부망에서 서빙이 가능한 구조
    - 외부 웹/API 요청이 아닌 내부망에서 서비스



- **메타러닝(Meta Learning) 가능성 검토**
  - 보다 적은 학습 데이터(One/Few-shot Learning)로 목적 태스크를 달성할 수 있는 방법론 정립
  - 데이터 라벨링(Labeling) 자동화 및 증강(Augmentation) 등에 활용

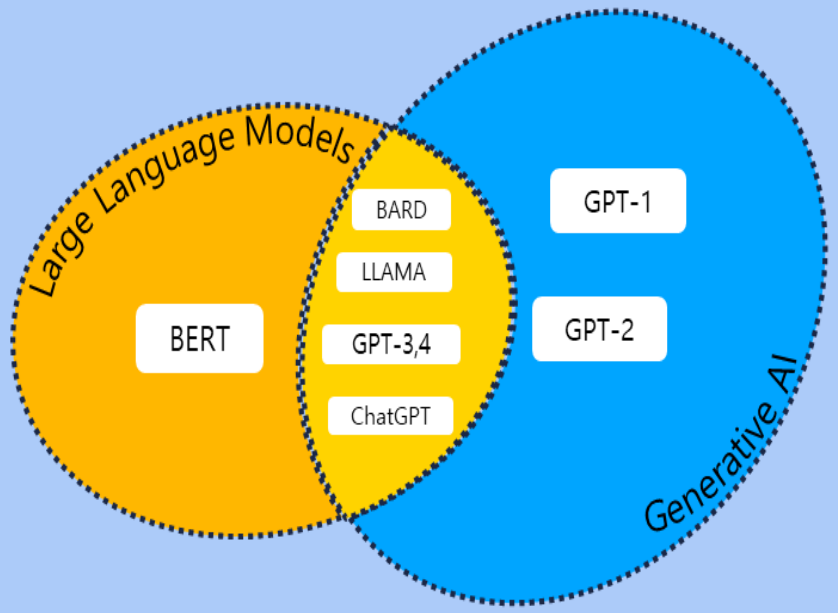
collect_day	good_name	sales_price
45017	뚬양꿍라면 컵라면용기 67g x 3개	7710
45017	[닛신] 돈베이 UFO 야끼소바 모음 / 키츠네 소바 카레 유부우동 / 일본 컵라면	2660
45017	[][][] 8585 진라면 순한맛 5개 5개입	3040
45017	[]농심 신라면 120g 10개 + 얼큰한 너구리 120g 5개 + 올리브 짜파게티 140g 5개	19680
45017	[11번가] [] 맛있는 라면 비건 100g	4480




```
{
    "instruction": "input에서 쇼핑몰, 생산자, 품목1, 품목2, 단위1, 단위2, 특징을 파싱해줘.",
    "input": "[11번가] 제주 흑돼지라면 돗멘 봉지 4개입 /돗맨/흑돼지컵라면",
    "output": "쇼핑몰 : 11번가, 품목1 : 라면, 품목2 : 복합라면, 특징 : 선택형"
},
```

Ⅳ. 향후 연구 추진계획

[참고] LLM(Large Language Model)



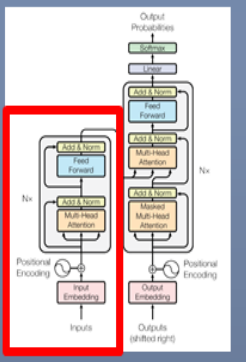
BERT 계열 언어모델

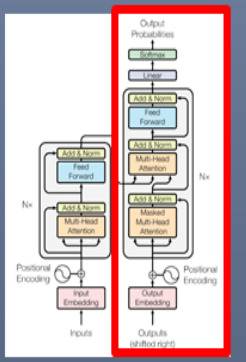
- 트렌스포머의 인코더 부분
- NLP TASK(목적)에 따른 전이학습 모델 파인튜닝(미세조정)에 주로 이용
- 정확도 개선을 위해서 새로운 도메인에서 언어모델 추가 학습 필요 (시간 비용 증가)
- 생성 능력 없음

GPT 계열 언어모델

- 트렌스포머의 디코더 부분
- Meta Learning 능력
- Zero/Few-shot Learning 기법으로 새로운 도메인에 바로 적용 가능
- 문장 생성 능력
- 프롬프트 엔지니어링을 통한 기능 추가 가능

트랜스포머 구조





Ⅳ. 향후 연구 추진계획

통계청 통계개발원

[참고] GPT 발전 역사와 모델별 정보

	Fine-tuning (미세조정)	NLP–Task (자연어처리 작업)
GPT-1 • 학습 데이터 : 5 GB • Parameter : 1억 1750만	• 태스크별로 미세조정 실시	• 문장생성, 요약, 분류, 독해, 등..
GPT-2 • 학습 데이터 : 40 GB • Parameter : 15억	• 태스크별 미세조정 실시 안함 • 일반화된 모델이라고 생각	• 문장생성, 요약, 분류, 독해, 등.. • Prompt로 작업 : Few, one, zero shot으로 자연어처리 작업 실시
GPT-3 • 학습 데이터 : 753.4 GB • Parameter : 1750억	• 태스크별 미세조정 실시 안함 • 일반화된 모델이라고 생각	• 문장생성, 요약, 분류, 독해, 등.. • Prompt로 작업 : Few, one, zero shot으로 자연어처리 작업 실시
GPT-3.5 • 학습 데이터 : > 754 GB • + 텍스트 • + 코드	• 프롬프트(Prompt)를 이용 → InstructGPT • 사람처럼 대화를 잘 하는 모델을 만들자	
GPT-4 • 학습 데이터 : UNK • + 텍스트 • + 코드 • + 이미지	강화학습+사람 피드백 → Chat-GPT 탄생	

감사합니다.

통계청
통계개발원