Multilingual hierarchical classification of job advertisements for job vacancy statistics

Maciej Beręsewicz[1,2]

[1]Department of Statistics, Poznań University of Economics and Business
[2]Centre for the Methodology of Population Studies, Statistical Office in Poznań

This is a joint work with: Marek Wydmuch, Herman Cherniaiev and Robert Pater.

POZNAŃ UNIVERSITY OF ECONOMICS AND BUSINESS

UNIVERSITY of INFORMATION TECHNOLOGY and MANAGEMENT in Rzeszow, POLAND

Contents

1. Introduction
2. Polish Classification of Occupations and Specializations (KZiS)
3. Datasets
4. Proposed approach
5. Selected results
6. Summary
7. Literature

Contents

1. Introduction
2. Polish Classification of Occupations and Specializations (KZiS)
3. Datasets
4. Proposed approach
5. Selected results
6. Summary
7. Literature

Introduction

- Joint work with **Marek Wydmuch** (Snowflake AI Research; Poznan University of Technology), **Herman Cherniaiev** (Educational Research Institute – National Research Institute) and **Robert Pater** (University of Information Technology and Management in Rzeszów; Educational Research Institute – National Research Institute).
- Paper: **Multilingual hierarchical classification of job advertisements for job vacancy statistics** (in review in the Journal of Official Statistics).
- Link: `https://arxiv.org/abs/2411.03779`
- Funding:
  - Polish National Agency for Academic Exchange (NAWA) under the project number BPN/BEK/2023/1/00099;
  - National Science Centre, Poland within a project *OJALAB: Online job advertisements to study skill demand and job search patterns* [Grant number: 2024/53/B/HS4/01580].
- **OJALAB**: `https://ojalab.ue.poznan.pl`.
- **NCN-Foreigners**: `https://ncn-foreigners.ue.poznan.pl`.

Contents

1. Introduction
2. Polish Classification of Occupations and Specializations (KZiS)
3. Datasets
4. Proposed approach
5. Selected results
6. Summary
7. Literature

KZiS Classification Overview

- Polish Classification of Occupations and Specializations (KZiS).
- Based on ISCO-08 (International Standard Classification of Occupations).
- **6-digit hierarchical structure**: 1-digit major groups $\rightarrow$ 6-digit specific occupations
- **Coverage**: Over 2,500 occupations grouped into 10 major categories.
- Similar to European Skills, Competences, Qualifications and Occupations (ESCO).
- Hierarchy Levels:
  - 1-digit: Major groups (10 categories).
  - 2-digit: Sub-major groups (43 categories).
  - 3-digit: Minor groups (134 categories).
  - 4-digit: Unit groups (445 categories) – ISCO level.
  - 6-digit: Specific occupations (2,911 categories; 2,549 without *other/rest* categories).

Comparison with International Standards

Classification	Levels	Max Digits	Total Codes
ISCO-08	4	4	436
ESCO v1.2.0	5	Variable (5+)	3,039
KZiS (Poland)	5	6	2,911
SOC (USA)	4	6	867
O*NET (USA)	Variable	8	1,016

- KZiS extends ISCO-08 with detailed Polish occupations.
- Comparable granularity to ESCO framework.
- Enables detailed labor market analysis beyond ISCO limitations.
- Supports vocational education and regional policy decisions.

KZiS Structure by Major Groups

Code	Major Group	6-digit codes
0	Armed Forces Occupations	3
1	Public Authorities, Senior Officials and Managers	202
2	Professionals	789
3	Technicians and Associate Professionals	610
4	Clerical Support Workers	89
5	Services and Sales Workers	166
6	Skilled Agricultural, Forestry and Fishery Workers	63
7	Craft and Related Trades Workers	476
8	Plant and Machine Operators and Assemblers	387
9	Elementary Occupations	126
	Total	**2,911**

Contents

1. Introduction
2. Polish Classification of Occupations and Specializations (KZiS)
3. Datasets
4. Proposed approach
5. Selected results
6. Summary
7. Literature

Data sources

- **Online sources**: 1,805,967 ads from which 10,000 and 1,000 ads were sampled.
- **Expert validation**: Three PEO experts hand-coded 10,000 ads + extra 1,000 ads (210 were from CBOP).
- **Novel administrative source**: Central Job Offers Database (CBOP) – over 800k ads.
- **Other sources**: official dictionaries and Statistics Poland thesaurus.

Central Job Offers Database (CBOP)

- **Source**: All job vacancies submitted to Polish Public Employment Offices (PEOs).
- **Coverage**: 822,000+ fully labeled ads (2022-2023).
- **Expert coding**: PEO staff manually assign KZiS codes.
- **Structured format**: 174 JSON fields per advertisement.

Data Quality Characteristics

- Average description length: 50 words
- 2,468 unique occupation codes represented
- Long-tail distribution: 145 codes with single example
- Most frequent: Sales Assistant (522301), Warehouse Operator (432103)

Sampling Strategy: Stratified sample by KZiS code and description length $\rightarrow$ 167,244 ads (20% of full dataset)

Data Sources Overview

Source	Records	Description
Official Dictionary	9,200	KZiS occupation descriptions
CBOP (Administrative; 20% sample)	167,244	Public Employment Office ads
Hand-coded (10k)	9,992	Expert-coded online ads
Hand-coded (1k)	1,035	IT-focused expert-coded ads
ESCO Linkage	2,114	KZiS-ESCO matched descriptions
Civil Service	2,941	Government job postings
GUS Thesaurus	7,007	Statistics Poland synonyms
Total	**200,875**	**Polish dataset**
Multilingual	**3.4M+**	**24 EU languages**

Expert Coding Quality Assessment – 10,000 dataset

Expert Pair	1 Digit	4 Digits	6 Digits	Kappa
Before Clerical Review				
1 & 2	87.7%	78.7%	74.2%	85.4%
1 & 3	86.2%	74.7%	66.2%	83.5%
2 & 3	85.9%	75.6%	68.5%	83.1%
All three	79.9%	66.1%	59.2%	–
After Clerical Review				
1 & 2	92.6%	84.4%	80.5%	91.2%
1 & 3	90.2%	84.7%	77.5%	88.3%
2 & 3	86.8%	80.8%	72.9%	84.2%
All three	84.8%	75.6%	68.5%	–

- Strong agreement (>80% Kappa) at major group level
- CBOP vs. Expert agreement: 79% at 1-digit, 61% at 6-digit

Quality Assessment – CBOP dataset

Table 1: Point and 95% interval estimates of the rate of agreement of expert coding at 1, 4 and 6 digits for CBOP ads

Expert	Count	1 Digit	4 Digits	6 Digits	Kappa
1	75	78.2 (65.9, 88.3)	63.2 (50.7, 74.9)	56.3 (43.7, 68.5)	73.0 (59.2, 86.8)
2	66	79.0 (67.5, 88.5)	72.2 (60.3, 82.7)	65.6 (53.5, 76.7)	74.3 (61.8, 86.8)
3	69	81.0 (70.3, 89.7)	71.6 (59.8, 82.1)	62.7 (50.1, 74.5)	77.6 (66.2, 88.9)
All	210	79.3 (73.0, 85.0)	68.8 (62.0, 75.2)	61.4 (54.4, 68.1)	75.2 (67.9, 82.4)

Multilingual Dataset

- **Translation pipeline**:
  1. Polish → English (Google Sheets 'trick', i.e. `GOOGLETRANSLATE()` function).
  2. English → 22 EU languages (Argos Translate).
- **Final size**: 3.4M+ training records
- **Language coverage**: All 24 EU official languages.

	Train			Test		
Source	Cases PL	Cases ML	# Codes	Cases PL	Cases ML	# Codes
Official	9,200	22,143	2,911	-	-	-
Thesaurus	1,342	2,687	1,338	-	-	-
CBOP	116,879	280,461	2,213	50,365	181,648	2,468
ESCO	1,531	3,642	557	583	2,318	557
Hand 10k	6,720	16,081	708	3,272	12,050	1,039
Hand 1k	632	1,504	115	403	1,523	226
Info	5,004	12,010	996	2,003	7,012	996
KPRM	2,058	4,926	12	883	3,170	12
All	143,366	343,454	2,911	57,509	207,721	2,625

Note: Cases PL refers to Polish dataset, Cases ML refers to multilingual (24 languages) dataset

Contents

1. Introduction
2. Polish Classification of Occupations and Specializations (KZiS)
3. Datasets
4. Proposed approach
5. Selected results
6. Summary
7. Literature

Hierarchical Multi-Class Classification

Problem Definition

- Select single path from tree root to leaf.
- Respect hierarchical constraints.
- Provide probability estimates.
- Handle long-tail distribution.

Mathematical Constraints

- Parent probability = Sum of children probabilities.
- Level probabilities sum to 1.
- Coherent predictions across hierarchy.

Key Challenge: Long-tail distribution: 1/3 of classes have <10 examples, only 250 classes have >100 examples

Two Modeling Approaches – basic idea

Bottom-Up Approach

- Train classifier only on leaf nodes (6-digit codes).
- Use categorical cross-entropy loss.
- Reconstruct parent probabilities by summing children.
- Simple but ignores hierarchical structure during training.

Top-Down Approach

- Sequential decisions at each hierarchy level.
- Chain rule: $Pr(\text{code}|x) = \prod_{i=1}^{6} \Pr(\text{digit}_i|\text{prefix}, x)$.
- Trains separate classifiers for each level.
- Naturally respects hierarchy constraints.

Models Transformer Architecture

- **Baseline**: Linear models with TF-IDF features.
- **HerBERT**: Polish-specific BERT
  - Base: 110M parameters.
  - Large: 336M parameters.
- **XLM-RoBERTa**: Multilingual
  - Base: 279M parameters.
  - Large: 561M parameters.
- Trained with `napkinXC` package in Python.
- The code and weights are publicly available (see the links at the end of this presentation).

Contents

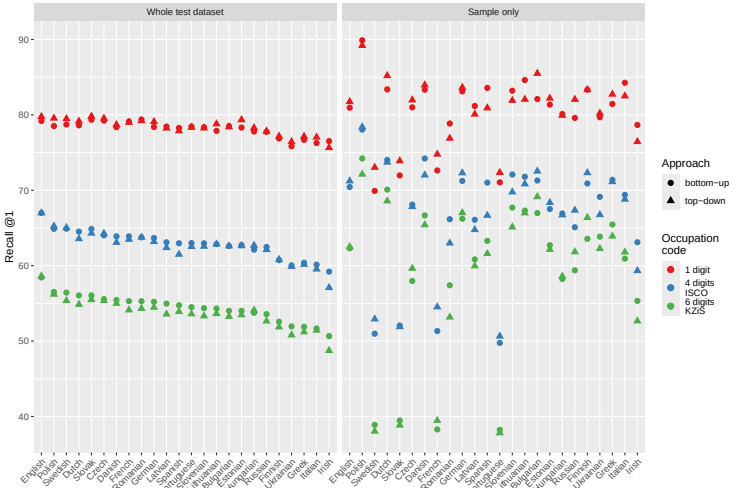
1. Introduction
2. Polish Classification of Occupations and Specializations (KZiS)
3. Datasets
4. Proposed approach
5. Selected results
6. Summary
7. Literature

Selected results – multilingual model

Table 2: Recall@1 metric for the multilingual test data by dataset and transformer approach

Dataset	Hierarchy	1 digit	2 digits	ISCO (4 digits)	KZiS (6 digits)
Overall	bottom-up	84.37	80.87	73.67	64.42
	top-down	84.40	80.70	73.25	63.74
CBOP	bottom-up	84.38	80.88	73.77	64.41
	top-down	84.43	80.75	73.46	63.90
Hand-coded	bottom-up	77.82	72.95	62.37	53.51
	top-down	78.59	73.62	62.46	53.15

Selected results – multilingual model (translated hand-coded data)



Contents

1. Introduction
2. Polish Classification of Occupations and Specializations (KZiS)
3. Datasets
4. Proposed approach
5. Selected results
6. Summary
7. Literature

Summary

- The effect of open-source translators such as Argos is unfortunately unknown. We plan to use more powerful alternatives, e.g., Qwen-MT (`https://huggingface.co/spaces/Qwen/Qwen3-MT-Demo`).
- The model was trained almost two years ago on an old classification (it requires an update).
- We did not compare to existing LLMs or fine-tuned smaller models (yet)...
- ...but our goal is also to include uncertainty (i.e., via probabilities), so we aim to use custom classifiers (with embeddings).

Software and contact

- **Models**: `https://repod.icm.edu.pl/dataset.xhtml?persistentId=doi:10.18150/OCUTSI`.
- **Code**: `https://github.com/OJALAB/job-ads-classifier`.
- **Tutorial**: Google Colab notebook available.
- **Contact**:
  - `https://ojalab.ue.poznan.pl`
  - Maciej Beręsewicz: maciej.beresewicz@ue.poznan.pl
  - Robert Pater: rpater@wsiz.edu.pl

Figure 1: https://ojalab.ue.poznan.pl

Figure 2: https://github.com/BERENZ

Contents

1. Introduction
2. Polish Classification of Occupations and Specializations (KZiS)
3. Datasets
4. Proposed approach
5. Selected results
6. Summary
7. Literature

Literature (selected)

- Beręsewicz, M. and Pater, R. (2021). Inferring Job Vacancies from Online Job Advertisements. Publications Office of the European Union.
- Beręsewicz, M. E., Białkowska, G., Marcinkowski, K., Maslak, M., Opiela, P., Katarzyna, P., and Pater, R. (2021). Enhancing the demand for labour survey by including skills from online job advertisements using model-assisted calibration. Survey Research Methods, 15(2):147-167.
- Dembczyński, K., Kotłowski, W., Waegeman, W., Busa-Fekete, R., and Hüllermeier, E. (2016). Consistency of probabilistic classifier trees. In Frasconi, P., Landwehr, N., Manco, G., and Vreeken, J., editors, Machine Learning and Knowledge Discovery in Databases, pages 511-526, Cham. Springer International Publishing.
- Jasinska-Kobus, K., Wydmuch, M., Dembczynski, K., Kuznetsov, M., and Busa-Fekete, R. (2020). Probabilistic label trees for extreme multi-label classification. arXiv preprint arXiv:2009.11218.
- Sun, A. and Lim, E.-P. (2001). Hierarchical text classification and evaluation. In Proceedings of the 2001 IEEE International Conference on Data Mining, ICDM '01, page 521-528, USA. IEEE Computer Society.

Thank you and feel free to test our software!