DEFINING ARTIFICIAL INTELLIGENCE COMPETENCES AND KNOWLEDGE BASED ON THE JOB MARKET ANALYSIS

Y. Demchenko[1], O. Chertov[2], J. Vreeburg[1]

[1]University of Amsterdam (NETHERLANDS)
[2]Igor Sikorsky National Technical University of Ukraine "Kyiv Polytechnic Institute" (UKRAINE)

Abstract

The growing use of Artificial Intelligence (AI) technologies and applications in almost all sectors of the economy and human activities creates demand for specialists in AI. Higher Education responded to this demand by establishing new AI focused undergraduate and graduate programs that attract the growing number of students. However, there is no established common academic curricula or AI competence framework, many AI programs are created based on existing Computer Science, Data Science or Machine Learning programs. This paper presents ongoing research and current results of applying combined bottom up (job market analysis) and top down (Analysis of top AI curricula) approach to defining job market demand for AI competences and knowledge and assessing what AI related disciplines comply with the market demand. The paper uses the EDISON Data Science Framework (EDSF) approach to analysing the job market data by matching the selected AI related vacancies against the multi-vector controlled vocabulary for AI related competences and knowledge topics that are identified based on the analysis of the leading AI curricula. The presented research investigates different text matching algorithms and produced reference dataset of the collected AI related job vacancies from indeed.com to help future research on defining community agreed AI competences and Body of Knowledge. The artefact is published on Zenodo and available on github.

Keywords: Artificial Intelligence, Artificial Intelligence Competences, Body of Knowledge, Artificial Intelligence Educations, Machine Learning, EDISON Data Science Framework (EDSF).

1 INTRODUCTION: DEMAND FOR AI PROFESSIONALS

The demand for skilled professionals in the Artificial Intelligence (AI) field is rising. To address this demand, universities have started offering dedicated undergraduate and graduate AI programs. However, there is no formally defined AI Body of Knowledge and competences framework, similar to those defined for Computer Science, Software Engineering or Data Science. Higher Education responded to this demand by establishing new AI focused undergraduate and graduate programs that attract growing number of students. Consequently, there is no established common academic curricula, many AI programs are created based on existing Computer Science, Data Science or Machine Learning programs and may not cover important knowledge areas.

This paper presents ongoing research and current results of applying combined bottom up (job market analysis) and top down (Analysis of top AI curricula) approach to defining job market demand for AI competences and knowledge and assessing what AI related disciplines comply with the market demand.

2 AI CURRICULA ANALYSIS AND IDENTIFIED COMPETENCES AND KNOWLEDGE AREAS

The initial set of AI-related competences and knowledge areas was identified based on the analysis of existing AI programs, in particular, the AI Bachelor and AI Master programs offered by the University of Amsterdam in the Netherlands [1], which is recognized internationally. It was later verified and confirmed with the job market analysis what was the focus of this research.
The following are the enumerated list of competence groups and knowledge areas identified from existing curricula analysis (not specific relevance order at this stage):

CKAI_01 Knowledge representation and reasoning

CKAI_02 Automated planning and scheduling

CKAI_03 Machine learning

CKAI_04 Natural language processing

CKAI_05 Machine perception

CKAI_06 Computer vision

CKAI_07 Speech recognition

CKAI_08 Robotics

CKAI_09 Affective computing

CKAI_10 Deep learning

CKAI_11 Information retrieval

CKAI_12 Computer science

CKAI_13 Causality

CKAI_14 Data mining

CKAI_15 Commonsense knowledge

CKAI_16 Intelligent agent

The main challenge was to create a reference dataset that could serve as a controlled vocabulary for text analysis algorithms, in conditions that there is no formal definition of the AI competence framework and Body of Knowledge. The initial set of reference text articles was created using corresponding Wikipedia articles with expert review and removal of insignificant information such as introduction, context, history, and similar. It was correlated with the curricula and academic subjects description. In the process of research, the reference dataset has been revised multiple times to ensure the best possible competences identification in analysing vacancies certainty and minimising data bias.

3 JOB MARKET ANALYSIS FOR DEMANDED AI COMPETENCES AND KNOWLEDGE

The project used the EDISON Data Science Framework (EDSF) approach [2, 3] to analyse the job market data by matching the selected AI related vacancies against the multi-vector controlled vocabulary for AI related competences and knowledge topics that identified based on the analysis of the leading AI curricula.
The following steps are used when analysing the job advertisement data

1) Collect data on required competences and skills

2) Extract information related to competences, skills, knowledge, qualification level, and education; translate and/or reformulate if necessary

3) Split extracted information on initial classification or taxonomy facets, first of all, on required competences, skills, knowledge; suggest mapping if necessary

4) Apply existing taxonomy or classification: for the purpose of this study, we used competences and knowledge groups as defined above.

5) Identify competences and skills groups that do not fit into the initial/existing taxonomy and create new competences and skills groups

6) Do clustering and aggregations of individual records/samples in each identified group

7) Verify the proposed competences groups definition by applying to originally collected and new data

8) Validate the proposed competence framework via community surveys and individual interviews.

The outlined above process up to step (7) has been applied to the collected information, and the obtained results are ready for community discussion in step (8).

4 DATA PREPARATION AND ANALYSIS

The job vacancies were collected from the indeed.nl website (using available subscription) using the webscraping process and tools (see artefact description). The collected data were analysed by creating a vector matrix with sentence embedding models, which displays compact knowledge about AI job vacancies and can classify to what extent a job vacancy is represented in a specific competence group. Data preprocessing included textual information retrieval from the specific fields in the job vacancy description (which is standardized for the majority of job search sites), removal of stopwords; additionally, stemming and lemmatisation was used to improve accuracy.
For the vacancies analysis the project tested few text similarity algorithms based on frequency statistics and using vector matrix constructed by sentence embedding. The following algorithms were tested and compared: TF/IDEF, ScaCy, TensorFlow, Bert, MP_net. MP_net showed best competences recognition and selectiveness, which is known as the best performing accuracy score of pre-trained sentence embedding.

5 CONCLUSION AND FURTHER DEVELOPMENTS

The developed code and the data computed by the used algorithms are stored in an online GitHub repository [4], dataset is published on Zenodo [5].

The presented results can advise AI program coordinators and advise on the selection of individual courses. The analysis may also help companies and Human Resource departments in defining the AI related vacancies profiles and candidates selection.

The main goal of this paper is to initiate discussion and cooperation on further defining AI competences and AI Body of Knowledge as well as exchange of information on AI curricula design and implementation experience. The authors believe that the EDISON Data Science Framework methodology using in this research will be also helpful in the further steps with the AI professional domain definition.

REFERENCES

[1] Artificial Intelligence, Study Programme, University of Amsterdam [online] https://www.uva.nl/shared-content/programmas/en/masters/artificial-intelligence/study-programme/study-programme.html

[2] The Data Science Framework, A View from the EDISON Project, Editors Juan J. Cuadrado-Gallego, Yuri Demchenko, Springer Nature Switzerland AG 2020, ISBN 978-3-030-51022-0, ISBN 978-3-030-51023-7

[3] EDISON Data Science Framework (EDSF). [online] https://github.com/EDISONcommunity/EDSF

[4] Artificial Intelligence Job Market Analysis Project, Artefact, 2021 [online] https://github.com/atomcracker/Competence_analysis.git

[5] Artificial Intelligence: Professional reference dataset of Artificial Intelligence professional competences analysis based on the job market, by J. Freeburg, Y,Demchenko [online] https://zenodo.org/record/6402152

APPENDIX A. ARTEFACT DESCRIPTION AND RESULTS

The Python code used in this research and data both collected and produced are available at the project github repository https://github.com/atomcracker/Competence_analysis.git

The reference dataset comprising reference textual articles for each competence is available in the project directory ‘Data_Indeed/Field_text_new’

The job vacancies dataset is published on Zenodo https://zenodo.org/record/6402152

A.1. Artefact Functionality and Components

This *section* describes the functionality and components provided by the artefact and includes three groups: extracting information from the web, preprocessing and preparing data, and analysing data.

A.1.1 Data collection: Web scraping

Webscraping can be broken down into three steps. First, a list of the corresponding uniform resource locator (URL) containing online job vacancies needs to be obtained; this is achieved by searching for the 'Artificial Intelligence' keywords on the Indeed.com website. Data extracted from HTML using BeautifulSoup Python library. All extracted text files are given a unique Job ID.

A.1.1.2. Preprocessing data

Stopwords are removed from the dataset to improve the accuracy of the similarity scores. Stemming and lemming are applied to improve accuracy further.

A.1.1.3. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) breaks a multidimensional vector space down to a two-dimensional observable space. PCA is used in combination with clustering to identify data points (in our case reference CKAI articles) that have more correlation than other data points.

In this project, PCA was used to assess the consistency and selectiveness of the competences definition and the reference dataset was analysed and improved using PCA, what shown that the following CKAIs has close similarity:

Cluster 1: CAI_05 Machine Perception, CAI_06 Computer Vision, CAI_09 Affective Computing

Cluster 2: CAI_04 Natural Language Processing, CAI_11 Information Retrieval, CAI_14 Data Mining

Cluster 3: CAI_01 Knowledge Presentation, CAI_15 Commonsense Knowledge

A.1.2. Analysing data with similarity algorithms

The similarity algorithms used in this project can be divided into two different sections: algorithms based on a frequency statistic and the second one using a vector matrix constructed by sentence embedding.

A.1.2.1 Frequency algorithms

Term Frequency-Inverse Document Frequency (TFIDF) uses frequency statistics to determine the weighted importance of every word in contrast to the overall corpus. On top of the TFIDF algorithm, NLTK tokenisation and stemming is used.

A.1.2.2 Sentence embedding models

Sentence embedding models analyse a sentence about the grammar construction; this ensures that the meaning of each exact word is better understood. Sentence embedding models especially come into play when words have lots of synonyms.

Multiple popular sentence embedding models have been tested:

SpaCy: SpaCy is a free, open-source library for advanced Natural Language Processing (NLP) in Python.

TensorFlow: Google's TensorFlow network contains a pre-trained model for Universal Sentence Encoding (USE), trained for single sentences or small paragraphs in mind (CER, 2018). The model converts text into 512-dimensional embeddings. This model is based on the Transformer architecture and uses an 8000-word piece library.

Bert: Google's Bidirectional Encoder Representations from Transformers (BERT) model computes a vector space using the BERT Transformer encoder architecture (Reimers, 2019).

MP_net: MP_net v2 model (Song, 2020) is, according to SBERT (2022), the best performing accuracy score of pre-trained sentence embedding. The model is intended to be used as a sentence and short paragraph encoder, which computes a vector space given an input text.

A.2. Example results with MP_net algorithm

MP-net algorithm is identified as providing best results for AI jobs analysis.

The github repository provides the code, reference and collected dataset, as well as results obtained with testing different similarity algorithms which should be sufficient for the reproducibility of this research.

Table A.1 below illustrates the results of assessing the competences and knowledge areas (CKAI#) relevance to market demand. Scores are normalized to the highest score obtained by CKA03 Machine Learning, followed by CKAI16 Intelligent Agents, and CKAI09 Affective Computing.

Figure A.1 illustrates CKAI scores with MP_net algorithm modified to process extended text fields up to 512 words.

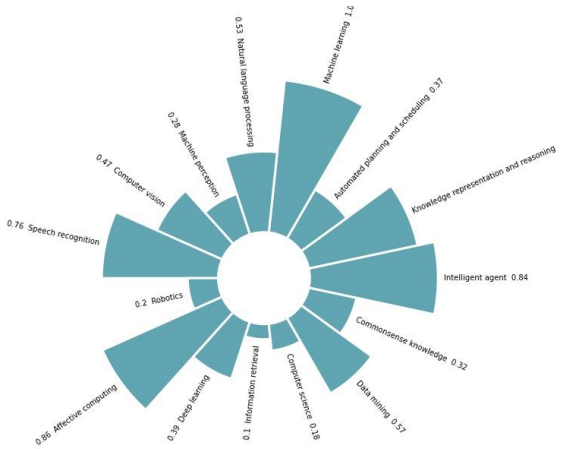

Figure A.1 Example CKAI scores obtained with MP_net algorithm

Table A.1 CKAI relevance to the collected AI vacancies

a) Ordered by CKAI_#

Comp/ Know ID	Competence/Knowledge Topic	MPnet score
CAI_01	Knowledge representation and reasoning	0.62
CAI_02	Automated planning and scheduling	0.57
CAI_03	Machine learning	1.00
CAI_04	Natural language processing	0.35

b) Ordered by job market relevance

Comp/ Know ID	Competence/Knowledge Topic	MPnet score
CAI_03	Machine learning	1
CAI_16	Intelligent agents	0.87
CAI_09	Affective computing	0.78
CAI_01	Knowledge representation and reasoning	0.62

CAI_05	Machine perception	0.60
CAI_06	Computer vision	0.50
CAI_07	Speech recognition	0.40
CAI_08	Robotics	0.45
CAI_09	Affective computing	0.78
CAI_10	Deep learning	0.43
CAI_11	Information retrieval	0.40
CAI_12	Computer science	0.32
CAI_13	Causality	0.10
CAI_14	Data mining	0.62
CAI_15	Commonsense knowledge	0.10
CAI_16	Intelligent agents	0.87

CAI_14	Data mining	0.62
CAI_05	Machine perception	0.6
CAI_02	Automated planning and scheduling	0.57
CAI_06	Computer vision	0.5
CAI_08	Robotics	0.45
CAI_10	Deep learning	0.43
CAI_07	Speech recognition	0.4
CAI_11	Information retrieval	0.4
CAI_04	Natural language processing	0.35
CAI_12	Computer science	0.32
CAI_13	Causality	0.1
CAI_15	Commonsense knowledge	0.1