A controlled vocabulary for research and innovation in the field of Artificial Intelligence (AI)

This controlled vocabulary of keywords related to the field of Artificial Intelligence (AI) was built by SIRIS Academic in collaboration with ART-ER (the R&I and sustainable development in-house agency of the Emilia-Romagna region in Italy) and the Generalitat de Catalunya (the regional government of Catalonia, Spain), in order to identify AI research, development and innovation activities. The work was carried by consulting domain experts advice and it was ultimately applied to inform regional strategies on AI and research and innovation policy.

The aim of this vocabulary is to enable one to retrieve texts (e.g. R&D projects and scientific publications) featuring the concepts included in the present vocabulary in their titles and abstracts, assuming that these records have a certain contribution of applications, techniques and issues, in the domain of AI.

The present effort was carried out because, despite the high number of contributions and technological developments in the field of AI, there is no closed or static vocabulary of concepts that allow one to unequivocally define the boundaries of what should be considered “an Artificial Intelligence intellectual product” (or what should not). Indeed, the literature presents different definitions of the domain, with visions that could be contradictory. AI encompasses today a wide variety of sub-domains, ranging from general purpose areas such as learning and perception to more specific ones such as autonomous vehicle driving, theorem proving, or industrial process monitoring. AI synthesises and automates intellectual tasks, and is therefore potentially relevant to any area of human intellectual activity. In this sense, it is a genuinely universal and multidisciplinary field. AI draws upon disciplines as diverse as cybernetics, mathematics, philosophy, sociology and economics.

For the current definition of an AI controlled vocabulary, the initial set of terms setting the boundaries of AI were taken from different sub-domains of the *ACM Computing Classification System 2012*. Notably, although some relevant AI sub-domains have an independent category in the ACM taxonomy outside of AI, they have been included in the list of sub-domains. In order to align the ACM taxonomical definition with the Catalan Strategy of AI, *CATALONIA.AI*, the emerging area of AI Ethics has been included in the vocabulary, while some other categories which are not relevant for the objectives of this resource have been removed from the sub-domains list.

In short, AI subdomains considered in the present vocabulary are the following: (1) General, (2) Machine Learning, (3) Computer Vision, (4) Natural Language Processing, (5) Knowledge Representation and Reasoning, (6) Distributed Artificial Intelligence, (7) Expert Systems, Problem-Solving, Control Methods and Search and (7) AI Ethics.

Although a keyword rule-based approach suffers of the major two shortcomings of not being able to capture all the lexical and linguistic variants of a specific term, and of not capturing the context of the

terms (in other words, keyword-based approaches would miss relevant texts if the specific pattern is not matched during the search), the present vocabulary allowed us to obtain fairly good results, due to the specificity of the concepts describing the AI domain. Furthermore, an understandable and transparent controlled vocabulary allows a better control of the final results and the final definition of the domain borders. Also, a plain list of terms allows a much easier and interactive engagement of interested stakeholders with different degree of knowledge (such as, for instance, domain experts, policy-makers and potential users) who can make use of vocabulary to retrieve pertinent literature or to enrich the resource itself.

The vocabulary has been built taking advantage of advanced language models and resources from knowledge datasets such as arXiv, Dpedia, Wikipedia and Scopus. The resulting vocabulary comprises 599 keywords, annotated by AI sub-domain, and has been validated by experts from several universities in Emilia-Romagna and Catalonia.

The first version of this resource was developed by the *SIRIS Academic* in 2019 in collaboration with ART-ER, Emilia Romagna (Quinquillá et *al.*, 2020), and the current version was updated in 2020 in collaboration with the Generalitat de Catalunya.

The methodology for the construction of the controlled vocabulary is presented in the following steps:

1. An initial set of scientific publications was collected by retrieving the following records as a weakly-supervised (in the sense that records are linked to AI by their taxonomy and not by a manual label) dataset in the domain of Artificial Intelligence :
    a. Publications from Scopus with the keyword “Artificial Intelligence”
    b. Publications from arXiv in the category “Artificial Intelligence”
    c. Publications in relevant journals in the scientific domain of “Artificial Intelligence”
2. An automated algorithm was used to retrieve, from the APIs of DBpedia, a series of terms that have some categorical relationships (i.e. those that are indexed as “sub-categories of”, “equivalent to”, among other relations in DBpedia) with the Artificial Intelligence concept and with the AI categories in the ACM taxonomy. The DBpedia tree has been exploited down to the level 3, and the relevant categories have been manually selected (for instance: *Classification algorithms*, *Machine learning* or *Evolutionary computation*) and others were ignored (for instance: *Artificial intelligence in fiction*, *Robots* or *History of artificial intelligence*) because they were not relevant, or not specifically in the domain.
3. The keywords in publications in the dataset were extracted from the keyword sections and from the abstracts. The keywords with a higher *TF-IDF*, using an *IDF* matrix in the open domain, have been selected. The co-occurrence of keywords with categories in specific AI sub-domain and a clusterization of the main keywords has been used for a categorization of the keywords at the thematic level.
4. This list of keywords tagged by thematic category has been manually revised, removing the non-pertinent keywords and changing the wrong categorizations by fields.
5. The weak-supervised dataset in the domain of Artificial Intelligence is used to train a Word2Vec (Mikolov *et al.*, 2013) word embedding model (a machine learning model based on neural networks).
6. The terms’ list is then enriched by means of automatic methods, which are run in parallel:
    a. The trained Word2Vec model is used to select, among the indexed keywords of the reference corpus, all terms “semantically close” to the initial set of words. This step is

carried out to select terms that might not appear in the texts themselves, but that were deemed pertinent to label the textual records.

   b. Further, terms that are mentioned in the texts of the reference corpus and that are valued by the trained Word2Vec model as "semantically close" to the initial set of words are also retained. This step is performed to include in the controlled vocabulary a series of terms that are related to the focus of the SDGs and which are used by practitioners.

7. The final list produced by steps 2-6 is manually revised.

The definition of the vocabulary does not, per se, allow to identify STI contributions to AI: this activity in fact boils down to actually matching the terms in the controlled vocabulary to the content of the gathered STI textual records. To successfully carry out this task, a series of pattern matching rules must be defined to capture possible variants of the same concept, such as permutations of words within the concept and/or the presence of null words to be skipped. For this reason, we have carefully crafted matching rules that take into account permutations of words and that allow words within concept to be within a certain distance. Some relatively ambiguous keywords (which may match unwanted pieces of text), have a set of associated "extra" terms. These "extra" terms are defined as further terms that must co-appear, in the same sentence, together with their associated ambiguous keywords. Finally, each keyword in the vocabulary was assigned one or more AI sub-domains, so that the vocabulary can also be used to tag collections of texts within narrower AI sub-domains.

The final controlled vocabulary has been evaluated with an external test set, proposed by (Dunham *et al.,* 2020). The test set consists of the abstract of 10,606 papers published in the arXiv repository, of which 1,076 within the Artificial Intelligence subcategories and 9,530 in arXiv categories other than Artificial Intelligence. Evaluating the controlled vocabulary on this data set, we observe accuracy of .94. However, because the pertinence of these publications to the field of AI is based solely on their taxonomic classification (i.e., on whether they are classified in the arXiv within Artificial Intelligence and not on a manual labeling), this evaluation can only yield an orientative performance assessment.

The AI controlled vocabulary has been applied in two practical cases, which have the purpose of identifying skills, stakeholders and capabilities, of a specific research ecosystem at the regional level. See the following references:

- Quinquillá, Arnau, Duran-Silva, Nicolau, Massucci, Francesco Alessandro, Fuster, Enric, Rondelli, Bernardo, Bologni, Leda, … Moretti, Giorgio. (2020). Text mining to identify skills, stakeholders and capabilities: the case of Artificial Intelligence in Emilia-Romagna. Zenodo. http://doi.org/10.5281/zenodo.3606342. Poster presented at: World Open Innovation Conference 2019 (WOIC); 11th desember 2019, Rome, Italy.

- Bigas, E., Duran, N., Fuster, E., Parra, C., Fernández, T. (2021): "Anàlisi de l'especialització en intel·ligència artificial". Col·lecció Monitoratge de la RIS3CAT, Generalitat de Catalunya http://catalunya2020.gencat.cat/web/.content/00_catalunya2020/Documents/estrategies/fitxers/analisi-especialitzacio-intelligencia-artificial.pdf

Acknowledgements

- Tatiana Fernández (Direcció General de Promoció Econòmica, Competència i Regulació, de la Generalitat de Catalunya),
- Daniel Marco, Daniel Santanach i Eduard Balbuena (Departament de Polítiques Digitals i Administració Pública, de la Generalitat de Catalunya)
- Albert Sabater (Observatori d'Ètica en Intel·ligència Artificial i Universitat de Girona)
- Leda Bologni, Lucia Mazzoni and Giorgio Moretti (Art-ER)

Bibliography

Bigas, E., Duran, N., Fuster, E., Parra, C., Fernández, T. (2021): "Anàlisi de l'especialització en intel·ligència artificial". Col·lecció Monitoratge de la RIS3CAT, Generalitat de Catalunya http://catalunya2020.gencat.cat/web/.content/00_catalunya2020/Documents/estrategies/fitxers/analisi-especialitzacio-intelligencia-artificial.pdf

Dunham, J.W., Melot, J., & Murdick, D. (2020). Identifying the Development and Application of Artificial Intelligence in Scientific Text. ArXiv, abs/2002.07143. Available at: https://arxiv.org/abs/2002.07143

Mikolov, Tomas & Corrado, G.s & Chen, Kai & Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. 1-12.

Quinquillá, Arnau, Duran-Silva, Nicolau, Massucci, Francesco Alessandro, Fuster, Enric, Rondelli, Bernardo, Bologni, Leda, ... Moretti, Giorgio. (2020). Text mining to identify skills, stakeholders and capabilities: the case of Artificial Intelligence in Emilia-Romagna. Zenodo. http://doi.org/10.5281/zenodo.3606342. Poster presented at: World Open Innovation Conference 2019 (WOIC); 11th december 2019, Rome, Italy.