Annex A. Creating a mapping between the indicators of the demand and supply of skills: Using machine learning to bridge between the RTC and online job postings

This report uses a dataset of online job postings (OJPs) with monthly information between January of 2018 to June of 2022 to analyse Umbria's labour market trends. The data is collected, transformed and harmonised by Lightcast (formerly Emsi-Burning Glass Technologies). The data is composed of 6.8 million individual level job postings for Italy and 72 434 for Umbria. There are up to 70 different variables ranging from skill keywords contained in each job posting, qualifications and experience required to fill the job and its geographical location, as well as the type of contract (permanent, temporary) and, when available, the salary offered for the specific role advertised. The OECD further transformed the data to create yearly aggregates, cross tabulations and other statistics presented in the document. Furthermore, the raw text of the OJPs is used for analysis, which is explained in this Annex. Lightcast offers the unique possibility to investigate the text contained in each online job posting, which reveals an amount of information that cannot be matched by any other source.

The Regional Training Catalogue (RTC) by ARPAL Umbria contains information on 1649 different courses, that in have 23 652 training spots available. The variables within this dataset have been discussed in Table 2.1 in Chapter 2.

In order to analyse the text information contained in OJPs and in the RTC, this report leverages Natural Language Processing (NLP henceforth) techniques. NLP is a multi-disciplinary field that draws on techniques from computer science, linguistics, mathematics, and psychology. More precisely, NLP focuses on the interaction between human language and computers. It involves developing algorithms and computational models that can process and analyse natural language data, including text, speech, and images. NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment (IBM, 2022[1]). In a nutshell, NLP allows researchers to create a map from words and their complex semantic meanings to numbers that can be analysed through algebraic manipulations and the use of probability models.

This Annex provides technical guidance on the methods that have been used throughout this report to create a mapping between the dictionary of words used in the Regional Training Catalogue (RTC) in Umbria and the skill keywords extracted from online job postings (OJPs).

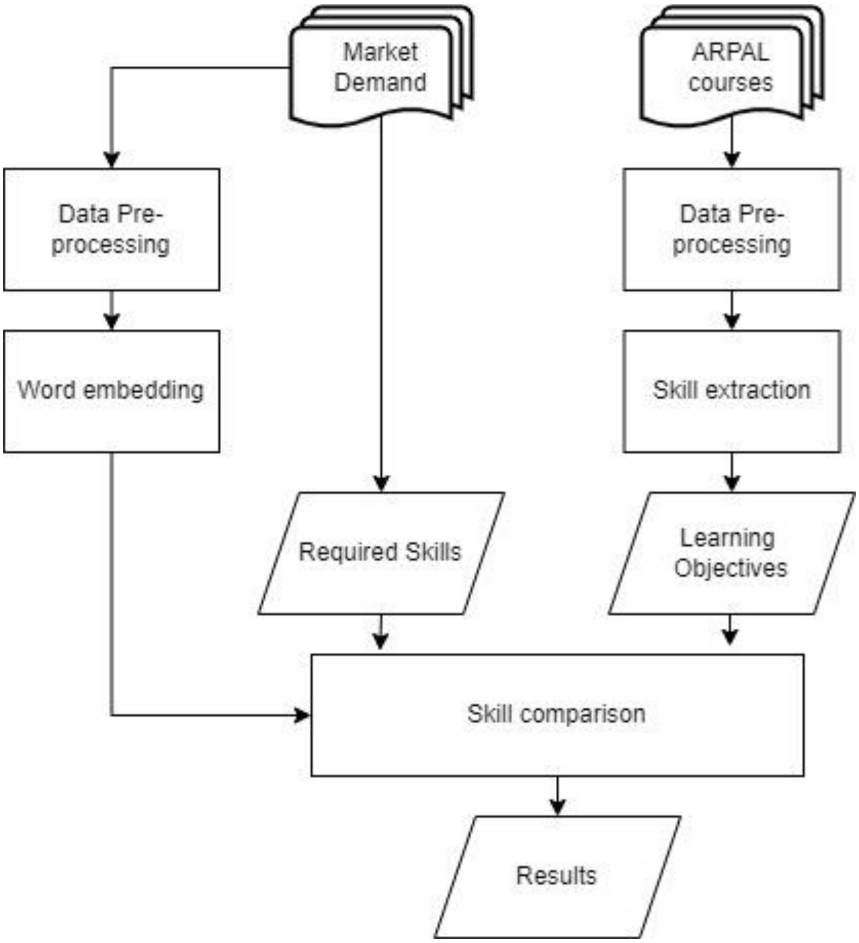
Figure A A.1 provides a graphical description of the conceptual framework adopted to create the mapping between the vocabulary used to describe the training courses in the RTC and the keywords used by employers in OJPs. Both data sources (RTC and OJPs), at the top of the picture, are initially treated as distinct entities, both requiring data pre-processing, that is, standardization of their textual content and the removal of unnecessary words (typically called 'stop words') that do not convey particular semantic meaning or useful information.

The information contained in the RTC comes in separated structured files that contain the description of the learning modules and of the skills that students will learn once enrolling in the course. To highlight the most important aspects of the course descriptions, specific skill keywords have been extracted using simple rules (see Step 1, below).

The OJPs data, on the other hand, is already mapped to the ESCO skill taxonomy when collected by Lightcast and requires minimal pre-processing and cleaning. The keywords extracted from the OJPs are much more abundant than those present in the vocabulary of the RTC as they are derived from the analysis of over a million OJPs gathered for the entire Italian labour market in 2019. These keywords are used to create a semantic representation, the so-called *embedding*, that describes to what extent each skill keyword in the OJPs is semantically associated with another (see Step 2, below).

The embedding created using the OJPs keywords can also be used to determine the extent to which each skill provided by a training program in the RTC is related to a skill mentioned in the OJPs. This is achieved by measuring the semantic similarity (cosine similarity) between the skills in the RTC and those extracted from the OJPs. This strategy allows to mapping each skill keyword in the RTC to a set of skills in the demand side (OJPs), effectively integrating them into the latter's dictionary. This mapping is then used for all subsequent analyses, as presented and described in chapters 2 and 3 of this report.

Figure A A.1. Diagram of the process



At the core of the whole procedure lies the creation of the map or graph: the word embedding. The main computational tool used to create this mapping is *fastText*, an algorithm able to infer a numerical representation of words by means of a prediction task (see Box A A.1). The main advantage of fastText is its focus on subwords, smaller portion of words, that are particularly helpful in creating a more accurate semantic structure for neo-Latin languages, like the Italian language.[1]

Box A A.1. FastText: A machine learning approach to the analysis of skill keywords in OJPs and RTC

The current study leverages fastText, an NLP algorithm developed by researchers Bojanowski et al. (2017[34]). This algorithm creates a mapping between the meaning (i.e. semantics) of subwords contained in text and mathematical vectors, so-called 'subword vectors' or 'embeddings'. Put differently, subword vectors are the mathematical representation of the meaning of the words.

An embedding contains the coordinates and hence the position that each skill has in a high-dimensional vector space, the so-called "graph". These coordinates make it possible to assess how close or distant every pair of skills are from each other.

As each skill is represented by a vector , the distance between two skills *A* and *B* is a measure of vector distance given by the cosine of the angle between the two, the *cosine similarity:*

$$distance(A,B) = \frac{A\ ^{\cdot}B}{\left||A|\right|\left||B|\right|}$$

where the denominator expresses product between the L2-norm (or Euclidean distance) of each n-dimensional vector.

The distance between vectors allows to rank skills from the closest to the farthest. In other words, this approach allows to rank the similarities between every skill vector by estimating their semantic closeness. Using this approach is, therefore, possible to assess whether the skill "Excel" found in the OJPs is semantically (and conceptually) close to "Electronic Spreadsheets", this latter a keyword found in the RTC.

The remainder of this Annex provides more details about the steps taken to map keywords in the RTC into the skill dictionary extracted from the OJPs.

Step 1: Extracting skills from the RTC

The information contained in the RTC does not conform to a pre-specified dictionary of keywords or a skill taxonomy. Instead, each training programme in the RTC is described by training providers by using words of very different nature and pertaining to a wide range of technical areas. While several of these words provide meaningful information about the course content (e.g. "using electronic spreadsheet", "baking pastries", "drawing technical designs"), other words convey little or no information (e.g "and", "or", "even", "the").

As the goal is to understand the degree of alignment between the skills provided by the training programs of the RTC and the skill-requirements of the demand side contained in OJPs, **the first step** requires to identify those useful keywords used in the description of each training course that describe the skills, technologies and tasks that are the core of the training programme.

This report uses two variables 'Competence Unit' and 'Learning Unit of Competence' that describe the course learning objectives (see Chapter 2), after having removed words that convey little information. To focus solely on the skill-related keywords, the text in "Competence Unit" and "Learning Unit of Competence" is broken down into its basic components using a simple rule based on verbs, which separates the sentences into these components. Table A A.1 provides an example of how the text in the

'Competence Unit' and 'Learning Unit of Competence' columns of the RTC is broken down into skill keywords using verbs as delimiters. The training program 20737 is for "Office Automation Operators". In the fourth column, the course description is "handle emails and retrieve information online". The verbs "handle" and "retrieve" are used as skill delimiters, resulting in the extraction of two skills: "handle emails" and "retrieve information online". These two skills are then considered as keywords in the RTC skill dictionary.

Table A A.1. Example of skill extraction

Training programme	Training programme title	Competence Unit (title of the learning objective)	Learning Unit of Competence (description of the training goals)	Extracted RTC skill
20737	Office Automation Operator	Internet and emails	Handle emails and retrieve information online	Handle emails
				Retrieve information online

Step 2: Creating the semantic representation

The data provided by Lightcast contain, for each job post, a list of skills required by the employer, identified directly by Lightcast using a variety of different skill taxonomies, including keywords in ESCO. The semantic representation, the word embedding, is created starting from the list of skills of all job postings collected in Italy in the year 2019[2] and leveraging the CBOW (*continuous bag-of-words*) algorithm (see Box A A.2). This step consists of two parts. Firstly, the list of skills in Italian job posts in 2019 is pre-processed by removing stop words and punctuation. Secondly, the resulting corpus is scanned to match words with a predetermined dictionary of skills, which is composed of RTC skills found in the first step, Lightcast skills provided directly by Lightcast, and ESCO skills.[3] The sequence of words that indeed match pre-coded skills coming from these three sources are coded as an *n-gram*.[4] All words that do not match pre-coded skills, therefore new to the dictionary, are instead selected to create entirely new n-grams; this method allows to expand and enrich the list of pre-determined skills provided by Lightcast. The reason for this choice is to preserve the nuances of the terminology used by training providers in their course descriptions in the RTC and by the employers in OJPs.

For computational reasons only words and bi-grams that appear more than 40 times are kept. It is important to notice, also, that the starting point of the creation of the embeddings is a set of n_grams, representing skills, made by the intersection of three different data sources: the RTC, Lightcast data and ESCO skills.

In the second step, *fastText*,[5] a python package is used to create the map between words and numbers; this map is technically referred to as word embedding. The algorithm used to do so is the CBOW, better described in Box A A.2.

Before describing the third step it's important to stress that after the second step each skill found in the RTC can be associated to skills coming from the Lightcast and ESCO data, via its cosine similarity described in Box A A.1.

Box A A.2. Continuous bag-of-words (CBOW)

CBOW is the mathematical procedure (technically the model architecture) that fastText implements to construct the semantic representation of skills. The advantage of using fastText over other algorithms is that it allows to focus on subwords. For instance, the skill “realizzare_prodotti_pasticceria” (English translation: “Making pastry products”) can be further decomposed into smaller windows of, for example, 3 characters, as follows:

[“rea”, “eal”, “ali”, “liz”, “izz”, “zza”, “zar”, “are”, “re_”, “e_p”, “pr”, “pro”, “rod”, “odo”, “dot”, “ott”, “tti”, “ti”, “i_p”, “_pa”, “pas”, “ast”, “sti”, “tic”, “icc”, “cce”, “cer”, “eri”, “ria”].

During training, that is, during the phase where the algorithm learns the mathematical representation of the skills, the embedding vectors are learned by the model to capture the semantic and syntactic meaning of each of the subwords. The result is that each of the 29 subwords above will be associated with an embedding vector. For example, the embedding vector for the subword “rea” might look like:

[0.25, -0.33, 0.10, 0.75, -0.05, 0.12, -0.01, 0.07, 0.42, -0.18].

A vector of dimension ten, in this particular example. The original skill, “realizzare_prodotti_pasticceria” will be represented as 29 vectors of dimension 10, each associated to the specifc subwords that make up the skill. FastText will repeat the same refined representation of words into subwords, for each skill present in the dictionary.

Whenever fastText will attempt to understand the meaning of the skill “realizzare_prodotti_pasticceria” on the basis of its context, that is, of the skills that surround it in job posts, it will use as input not the skills themselves but their subwords embeddings, adding a level of refinement that is particularly suited for the analysis of neo-latin languages.

The number of skills that define the context, and the dimensionality of the vector that represents each skill, are chosen ex-ante and set, in this case, to 100.

Step 3: Creating a mapping between keywords in the RTC and the OJPs

Each RTC skill found in step 1 is matched with up to five of the most correlated Lightcast skills that share a minimum threshold of cosine similarity of 0.7 with the RTC skill. Table A A.2 and Table A A.3 give for two skills found in the RTC (in the first step) the four and five most similar skills found in Lightcast and ESCO. In the first example, the RTC skill reads “diagnosis_energetic_buildings” and it’s matched with four Lightcast skills: “efficiency_energetic_buildings”, “security_industrial_buildings”, “renovation_system_buildings” and “technology_monitoring_system_building”. The cosine similarity between the RTC skill and the four Lightcast skills spans from 0.81 to 0.7. Despite the possibility of having up to five Lightcast skills associated with an RTC skills the fifth highest match had a similarity below 0.7, and hence was not considered.

The second example instead, shows a case in which five Lightcast skills are matched with the RTC skill “make_products_pastry”. The five skills are “prepare_products_pastry”, “make_products_pastry_base_chocolate”, “being_machinery_products_pastryshop”, “prepare_products_bakery” and “decorate_products_pastry_events_special”. All five Lightcast skills share a similarity above 0.7 with the RTC skill, spanning from 0.91 to 0.79.

Table A A.2. Example (1/2) of RTC – Lightcast/ESCO most similar match

RTC skill	Lightcast/ESCO skill	Cosine Similarity
diagnosi_energetica_edifici	rendimento energetico_edifici	0.81
	sicurezza_edifici_industriali	0.75
	ristrutturare_impianti_edifici	0.7
	tecnologica_monitoraggio_impianti_edificio	0.7

Table A A.3. Example (2/2) of RTC – Lightcast/ESCO most similar match

RTC skill	Lightcast/ESCO skill	Cosine Similarity
realizzare_prodotti_pasticceria	preparare_prodotti_pasticceria	0.91
	realizzare_prodotti_pasticceria_base_cioccolato	0.85
	essere_apparecchiature_prodotti_pasticceria	0.83
	preparare_prodotti_panetteria	0.81
	decorare_prodotti_pasticceria_eventi_speciali	0.79

The mapping from skill keywords found in one corpus of text to another is, hence, built using similarity measures to match skills by their semantic proximity to each other in the skill space. This is to say that, if two skills have a similar meaning, they are also found closer to each other in the embedding and in the graph (i.e. the vector space). Such mapping allows the analysis to bridge between the dictionary of keywords in the OJPs with those used in the RTC, de facto allowing to create indicators of the alignment of the training offer with the demand of employers in the local labour market.

The accuracy of the mapping was checked by manually examining the most relevant RTC skills and their five most similar Lightcast skills to ensure they made sense together. Any mismatches were flagged and corrected. In particular, the most relevant RTC skills[6] were manually checked. In order to refine the incorrect associations, data trained and classified on the Italian version of Wikipedia[7] were used to suggest feasible alternative matches.

Table A A.4 presents an example of an initially mismatched pair of terms and the proposed replacement by a alternative suggestion. The two RTC skills to be matched in this case were "evaluate_quality_productive_process_breeding" and "evaluate_quality_productive_process_baking" both wrongly associated with the Lightcast skill "evaluate_quality_productive_process_pharmaceutical". While the skill is related to evaluate the productive process, it refers to clearly different context. The suggestion of the general-embedding model in this case was to opt for, unsurprisingly, a more general description of both skill, "analize_productive_process_improve", shown in the last column.

Table A A.4. Example (2/2) of RTC – Lightcast/ESCO wrong match replaced with a different skill

RTC skill	Removed Lightcast Skill	New match
valutare_qualità_processo_produttivo_allevamento	valutare_processo_produzione_farmaceutico	analizzare_processi_produttivi_migliorarli
valutare_qualità_processo_produttivo_panificazione	valutare_processo_produzione_farmaceutico	analizzare_processi_produttivi_migliorarli

Step 4: K-means clustering

In an additional step, a cluster analysis is carried out using a k-means algorithm with each element of the embedding vectors as a feature for the Lightcast skill classification. This helps to simplify the large amount of data and identify main areas of skill categories. To describe each cluster, two features are used: the most commonly occurring words in each cluster and the skill that is at the centre of the cluster, which is used to give the cluster a name.

Table A A.5 and Table A A.6 display an example of the cluster assigned by the k-means algorithm to the same skills shown in Table A A.2 and Table A A.3. It is important to stress that RTC skills are less numerous and typically more generic than Lightcast skills and, therefore, they can be mapped (See step 3) to up to 5 different Lightcast skills. Each Lightcast skills, however, is mapped to one (and only one) cluster, but the corresponding RTC keyword (being broader in meaning) can be mapped to one or more clusters. In Table A A.5, for instance, each Lightcast skill is assigned to the same cluster "Technical knowledge in the design and installation of air conditioning systems". In the second example (Table A A.6), instead, the RTC keyword is mapped to Lightcast skills that fall into two different clusters.

The data-driven clustering was finally also checked manually. In particular, this led to the clusters "Art, cinema and writing", "Aesthetics" and "Skills in management of vehicles and mobile machinery" all initially grouped into one cluster to be regrouped into three different classes of skills.

Table A A.5. Example (1/2) of RTC – Cluster association

RTC skill	Lightcast/ESCO skill	Cosine Similarity	Cluster Name
diagnosi_energetica_edifici	rendimento energetico_edifici	0.81	Conoscenze tecniche nella progettazione e installazione di sistemi climatici
	sicurezza_edifici_industriali	0.75	Conoscenze tecniche nella progettazione e installazione di sistemi climatici
	ristrutturare_impianti_edifici	0.7	Conoscenze tecniche nella progettazione e installazione di sistemi climatici
	tecnologica_monitoraggio_impianti_edificio	0.7	Conoscenze tecniche nella progettazione e installazione di sistemi climatici

Table A A.6. Example (2/2) of RTC– Cluster association

RTC skill	Lightcast/ESCO skill	Cosine Similarity	Cluster Name
realizzare_prodotti_pasticceria	rendimento energetico_edifici	0.91	Competenze nella ristorazione e preparazione alimenti
	realizzare_prodotti_pasticceria_base_cioccolato	0.85	Conoscenze tecniche nella produzione del settore alimentare, tessile e della moda
	essere_apparecchiature_prodotti_pasticceria	0.83	Conoscenze tecniche nella produzione del settore alimentare, tessile e della moda
	preparare_prodotti_panetteria	0.81	Competenze nella ristorazione e preparazione alimenti
	decorare_prodotti_pasticceria_eventi_speciali	0.79	Competenze nella ristorazione e preparazione alimenti

Reference

IBM (2022), *What is natural language processing?*, https://www.ibm.com/topics/natural-language-processing (accessed on April 2023). [1]

Notes

[1] As noticed by the creators of fastText other techniques "[…] ignore the internal structure of words, which is an important limitation for morphologically rich languages […]. For example, in French or Spanish, most verbs have more than forty different inflected forms" (paper link https://arxiv.org/pdf/1607.04606.pdf).

[2] The number of OJPs in 2019 in Italy is 1 308 156.

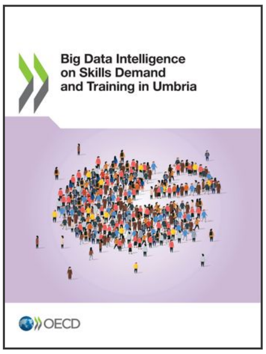
[3] downloaded from https: \\esco.ec.europa.eu\it\classification\skill_main.

[4] An n-gram is a series of words concatenated and considered as a unique term. A general example is "new" and "york" usually collapsed in the bi-gram "new_york".

[5] *fastText* documentation is available at https://fasttext.cc/docs/en/python-module.html.

[6] That is, those skills that appeared in the RTC's training programs descriptions with a frequency higher than 30.

[7] In greater detail, the third step was repeated using, again, a *fastText* model algorithm but this time already trained with Italian Wikipedia data with the intent of giving the numerical representation of words a more general domain of reference. The pre-trained model used is hosted at source: https: \\fasttext.cc\docs\en\crawl-vectors.html.



From:

Big Data Intelligence on Skills Demand and Training in Umbria

Access the complete publication at:

https://doi.org/10.1787/4bbbbfd6-en

Please cite this chapter as:

OECD (2023), “Creating a mapping between the indicators of the demand and supply of skills: Using machine learning to bridge between the RTC and online job postings”, in *Big Data Intelligence on Skills Demand and Training in Umbria*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/52e6006b-en

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Extracts from publications may be subject to additional disclaimers, which are set out in the complete version of the publication, available at the link provided.