Proceedings of the First workshop on Language Technologies for Government and Public Admnistration (LT4Gov), pages 15–21
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Enhancing Job Searches in Mexico City with Language Technologies

Gerardo Sierra[1] **Gemma Bel-Enguix**[1] **Helena Gómez-Adorno**[1]
Juan-Manuel Torres-Moreno[2,3] **Tonatiuh Hernández-García**[1] **Julio V. Guadarrama-Olvera**[1,4]
Jesús-Germán Ortiz-Barajas[1] **Angela María Rojas**[4] **Tomas Damerau**[4]
Soledad Aragón Martínez[4]

[1]Universidad Nacional Autónoma de México, México, [2]LIA-Université d'Avignon, France,
[3]Polytechnique Montréal, Canada, [4]Secretaría del Trabajo y Fomento al Empleo (STyFE), México
{gsierram, gbele, thernandezg, jortizb}@iingen.unam.mx, helena.gomez@iimas.unam.mx,
juan-manuel.torres@univ-avignon.fr, {jvicente.go, angela.styfe} @gmail.com, {tdamerau, haragonm}@cdmx.gob.mx

Abstract

In this paper, we show the enhancing of the Demanded Skills Diagnosis (DiCoDe: Diagnóstico de Competencias Demandadas), a system developed by Mexico City's Ministry of Labor and Employment Promotion (STyFE: Secretaría de Trabajo y Fomento del Empleo de la Ciudad de México) that seeks to reduce information asymmetries between job seekers and employers. The project uses webscraping techniques to retrieve job vacancies posted on private job portals on a daily basis and with the purpose of informing training and individual case management policies as well as labor market monitoring. For this purpose, a collaboration project between STyFE and the Language Engineering Group (GIL: Grupo de Ingeniería Lingüística) was established in order to enhance DiCoDe by applying NLP models and semantic analysis. By this collaboration, DiCoDe's job vacancies system's macro-structure and its geographic referencing at the city hall (municipality) level were improved. More specifically, dictionaries were created to identify demanded competencies, skills and abilities (CSA) and algorithms were developed for dynamic classifying of vacancies and identifying terms for searches on free text, in order to improve the results and processing time of queries.

Keywords: Language Technologies for Citizens, Job Search, Information Retrieval

1. Introduction: Context of the job search in Mexico City

Mexico City (CDMX) has about 9 million inhabitants and it is the most populous federative state and city in the country with 5,967 inhabitants per square kilometer density (INEGI, 2020). In the fourth quarter of 2019, there were 4.5 million economically active inhabitants, of which 230 thousand are unemployed. Among the latter, about a third of them are young people (15-24 years old) and 38 percent young adults (25-44 y.o.), implying that the majority of the unemployed are young job seekers in full productive age (STyFE, 2019). Beyond unemployment, high and pervasive informality work rates intensify labor market (LM) inefficiencies and reduce decent job opportunities. Informal employment is defined as all paid work that is not regulated or protected by legal frameworks (OIT, 2020). One of the main problems between LM demand (LMD, which are employers) and LM supply (LMS, which are job seekers) is information asymmetry that affects job search effectiveness (Hart, 1983). Identifying the job profiles and skills required by LMD and comparing them with those brought along by LMS is one of the key goals of this project. This requires great efforts in systematizing a variety of data sources (CVs, job vacancies, training programs' contents, etc), among which the focus, so far, has been on LMD's job vacancies posted online job portals.

2. DiCoDe: Demanded Skills Diagnosis Project

Information is a valuable asset in the job search process for LMD and for LMS. For both types of actors the quality and kind of information regarding vacancies, salaries, skills, competencies, among others, are crucial for an efficient allocation of the jobs available in the labor market (Stigler, 1962). In a context of great heterogeneity of both jobs and workers (and where none has all the information), lowering the costs of finding work to achieve a "good match" can have effects on productivity (Mortensen, 2011).

On the other hand, achieving a good match between LMD and LMS can also result in wage segregation problems or access barriers for less qualified workers (David, 2001). Whereby, it is essential not only to facilitate an efficient assignment of positions but also to identify training lines in the most demanded skills so that more workers increase their chances of match.

In this context, DiCoDe arises as a system that tracks job vacancies that online job portals show for CDMX, downloading them in bulk and storing them systematically for subsequent analysis. In the Figure 1 we can see the amount of job offers published online in the year 2019-2020 in CDMX.

To do so, DiCoDe uses two type of bots, one to index all urls that contain a vacancy and the second one to visit and download its information. This is then stored, preserving its main text structure (html headings) such as name of the vacancy, location, date of publication, salary offered, time

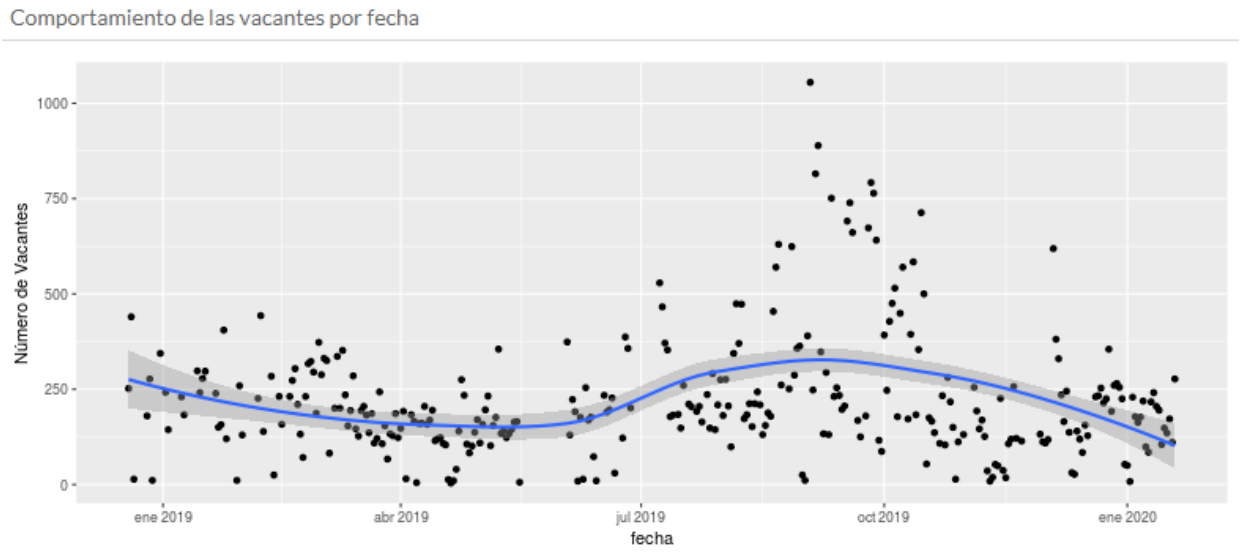

Figure 1: Number of vacancies published by year 2018-19 (screenshot from DiCoDe).

required, etc. Similar projects that take online job vacancies as a main source of information to analyze local labor markets have been developed in other countries of Latin America and the world (Altamirano et al., 2019; Amaral et al., 2018; Boselli et al., 2018). However, it is the first time that a project with these characteristics has been developed by the public sector in Mexico. It is important to emphasize that the DiCoDe System only uses data from vacancies offered online. Personal data of job applicants was not used for the system.

The main challenge is that each vacancy contains rich information, usually reported in an unstructured way (much of it in the "free text" section of each vacancy). Furthermore, employers use a variety of synonyms, ambiguous, redundant and imprecise language, which requires a detailed yet systematic NLP processing.

2.1. Collaboration academia and government

Faced with this challenge, STyFE contacted the Language Engineering Group (GIL) to find a solution based on the implementation of Language Technologies. The collaboration between academia and government allows the transfer of knowledge and technology for the benefit of citizens. The interdisciplinary work between the researchers and students of the GIL allowed the technological implementation to the challenge described.

We were asked to improve the performance of the DiCoDe system through the application of Language Technologies to: a) detect the macro-structure of job vacancies in CDMX and segment them by areas, b) improve the geographic location system by city hall, c) create dictionaries of the terms related to the competences, skills and abilities requested, d) obtain a classification of vacancies that allow structuring the database with categories reflecting CSA, e) improve the results and the processing time to perform text queries to the database.

3. Enhancing DiCoDe System

The enhancements imply the development of seven phases: 1) taxonomic information of job offers in CDMX, 2) design of a neural networks system to classify job offers, 3) implementation of an algorithm for improving the geographic location of the offers, 4) elaboration of a dictionary of vocabulary related to the topic, extracted form the data base, 5) building of a method based on regular expressions to automatically extract the features of the jobs, 6) implementation of dynamic filters for classification of offers, 7) identification of terms for free text searches.

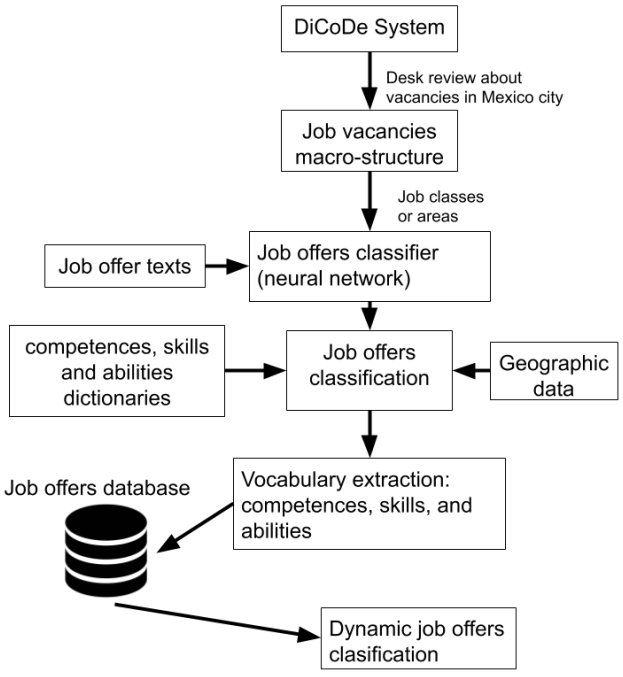

Figure 2: Structure of the system

3.1. Retrieving Taxonomic Information of job offers in CDMX

Determining the macro-structure of the set of job vacancies in CDMX is essential to know what types of occupations are offered to the inhabitants. A desk review suggests that some international occupational taxonomies (ISCO-08, 2012; SOC, 2018; ESCO, 2017) contain more occupations than actually apply in CDMX. Moreover, Mexico's official classification system SINCO (Hernández, 2018), developed by the national statistics office (INEGI) and that had been used by STyFE since DiCoDe's beginning, has the caveat that some occupational groups are not very frequent in the CDMX (e.g. coastal occupations, mining, etc), in addition, its classification system and vocabulary/jargon differs from the one used by LMD, or LMD does not post some type of occupations in job portals (e.g. fishermen).

The solution was to use the taxonomy of the Bumeran (Bumeran, 2020) job portal which groups 146 types of occupations into 23 main areas. The macro-structure of this site is built on the offers that employers publish with the geographical label "CDMX", therefore it represents the diversity of job vacancies found online.

This macro-structure of 23 categories grouping 146 types of occupations is an input for automatically classifying vacancies using a supervised learning algorithm. In the entry, the algorithm receives as input a set of job offers labeled with the categories of the Bumeran portal.

3.2. Neural Networks to classify job offers

We use a supervised learning approach with neural networks, using an LSTM architecture to build a job offers classifier based on the 23 classes mentioned above. These classes are: sales, human resources, technology, trades, administration, health, call center, legal, engineering, design, logistics, insurance, gastronomy, communication, secretary, finance, foreign trade, construction, marketing, production, education, management and mining.

3.2.1. Dataset

In order to train the classifier, we use a dataset with 979,956 examples, each one containing a job offer description text and its corresponding label. In Table 1, we show the details of the dataset:

Label	number of examples
Administration	124,349
Call center	152,820
Foreign trade	3,387
Communication	6,865
Construction	8,960
Design	9,897
Education	8,695
Financing	21,984
Gastronomy	20,995
Management	6,491
Engineering	14,521
Legal	10,004
Logistics	36,993
Marketing	31,519
Mining	635
Trades	56,931
Production	14,884
Human Resources	35,145
Health	21,501
Secretary	18,121
Insurance	0.9
Technology	87,105
Sales	277,197
Total	979,956

Table 1: Dataset details

3.2.2. Methodology

This section explains the processing that was carried out in the corpus to subsequently perform the classification task.

- Text normalization: Job offer texts were standardized to lowercase, and we put a dash if the text was empty.
- Stopwords removal.
- Punctuation symbols removal.
- Tokenization.

3.2.3. Neural network architecture

We use Keras library to build the model in a simple and fast way. Figure 3 shows a block diagram describing the neural network architecture:

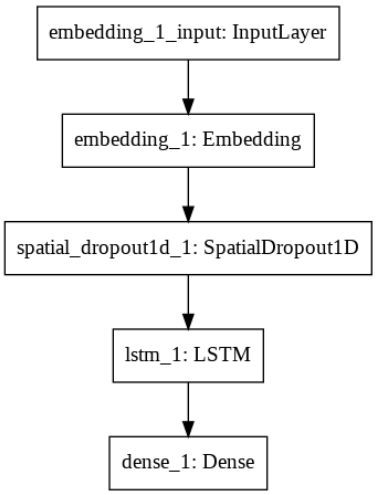

Figure 3: Neural network architecture

Embedding layer: It generates vector representations of words that capture semantic information from them.

Dropout layer: Set randomly a fraction of the input units to zero in each update during model training to avoid its overfitting.

LSTM layer: This layer allows the neural network to learn long-term dependencies.

Dense layer: This layer performs the activation function of the neural network, in this case, a softmax function.

Results are presented in Tables 2 and 3.

Label	Precision	Recall	F1	Support
Administration	0.95	0.94	0.94	24951
Call center	0.90	0.93	0.92	30578
Foreign trade	0.93	0.88	0.90	716
Communication	0.88	0.77	0.82	1396
Construction	0.94	0.92	0.93	1740
Design	0.96	0.93	0.94	2001
Education	0.92	0.93	0.92	1682
Financing	0.93	0.89	0.91	4396
Gastronomy	0.82	0.94	0.93	4313
Management	0.94	0.92	0.88	1321
Engineering	0.91	0.90	0.90	2073
Legal	0.89	0.95	0.92	2019
Logistics	0.92	0.95	0.94	7380
Marketing	0.93	0.92	0.93	6372
Mining	0.86	0.89	0.87	133
Trades	0.91	0.86	0.89	11331
Production	0.90	0.91	0.90	2943
Human Resources	0.95	0.94	0.94	6945
Health	0.95	0.93	0.94	4316
Secretary	0.89	0.89	0.89	3645
Insurance	0.91	0.90	0.90	2230
Technology	0.95	0.93	0.94	17537
Sales	0.94	0.94	0.94	55174

Table 2: Neural Network Results

Approach	Value
Cross Entropy	0.2399
Accuracy	0.9272
Precision	0.9437
Recall	0.9163
F1 measure	0.9298

Table 3: Neural Network Results (approaches)

3.2.4. Comparison experiments

Several experiments were carried out using different approaches in order to compare these results with the results from our proposed solution using F1-score as a metric evaluation. For all the experiments, we apply the same text preprocessing steps described in section 3.2.2.

3.2.5. Comparison results

In this section, we briefly describe the experiments and show the obtained results.

- FastText: This is a library created by Facebook, which is based on the Bag of Words model and it was improved using multilayer neural networks-based classifiers (Montañes Salas et al., 2017). Results obtained with FastText are shown in Table 4.
- Traditional machine learning algorithms: The performance of the job offers classifier was tested using the following algorithms: Support Vector Machine (SVM), Naive-bayes (NB), logistic regression (LR) and random forest (RF). Results obtained in these experiments are shown in Table 5.

Label	Precision	Recall	F1	Support
Administration	0.92	0.93	0.92	37318
Call center	0.88	0.92	0.90	45753
Foreign trade	0.94	0.79	0.86	1072
Comunication	0.89	0.67	0.76	2099
Construction	0.92	0.89	0.90	2643
Design	0.94	0.91	0.93	3040
Education	0.95	0.89	0.92	2592
Financing	0.92	0.86	0.89	6572
Gastronomy	0.92	0.92	0.92	6392
Management	0.82	0.80	0.81	1980
Engineering	0.91	0.86	0.88	4326
Legal	0.89	0.90	0.90	3039
Logistics	0.91	0.93	0.92	11044
Marketing	0.93	0.89	0.91	9553
Mining	1.00	0.56	0.72	190
Trades	0.88	0.84	0.86	17050
Production	0.89	0.86	0.88	4443
Human resources	0.94	0.92	0.93	10420
Health	0.94	0.91	0.92	6398
Secretary	0.91	0.85	0.88	5481
Insurance	0.92	0.85	0.89	3328
Technology	0.95	0.91	0.93	26206
Sales	0.91	0.94	0.92	83048

Table 4: FastText results

Classifier	Accuracy	Precision	Recall	F1-score
SVM	0.65	0.62	0.43	0.48
NB	0.49	0.32	0.15	0.16
LR	0.61	0.53	0.40	0.44
RF	0.58	0.66	0.29	0.36

Table 5: Traditional machine learning algorithm results

3.3. Improving the geographic location system

With the objective of geographically locating vacancies by city hall, we developed an algorithm to improve the geographic classification of job vacancies at the city hall level by tracking geographical elements that appear in the column of 'area' in the vacancy database. For this activity it was first necessary to conduct a study on CDMX's territorial demarcations. With this information a dictionary has been developed that has served as the basis for the algorithm design.

The denominations of each of the 16 different territorial demarcations of Mexico City have variants in everyday use. Among these variants can be cases of acronyms, abbreviations and apocopes.

To search for words that could be related to the city hall's offices, a dictionary composed of two elements was built:

1. A list of *colonias* (city halls' subdivisions) and neighborhoods with their zip codes.
2. A list composed of the different denominations of CDMX's city hall.

The algorithm uses the dictionary of *colonias*, neighborhoods and city halls. The dictionary searches the area within DiCoDe's database for each of the items in the dictionary and returns the city hall where it is when there are

coincidences. As a result the algorithm produces a file with the original fields plus that of the city hall's containing the algorithm's findings. Job offers that offer vacancies in different locations are treated as separate offers by location.
In Figure 4 we can see the result of the application of the algorithm, which allows us to know the number of vacancies published online in each city hall.

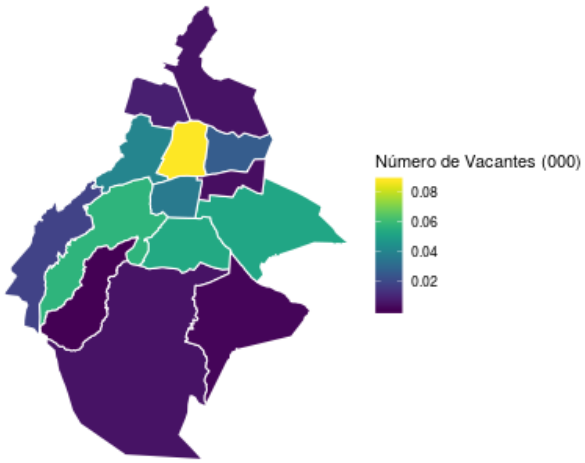

Figure 4: Vacancy number per City Hall (Map generated by DiCoDe System)

3.4. Extracting related vocabulary: competences, skills and abilities

The web portals that publish the job vacancies allow LMD actors to use a "free text" space for describing job duties, relevant skills, etc. In this field the specific requirements of the vacancy are detailed, without any restriction, so that each LMD actor writes the information differently, without following a preset template. Therefore, the names and distribution of the demanded CHD varies, resulting in the concepts being ambiguous and used interchangeably, so they are usually grouped into different items.
To enhance DiCoDe's robustness, Natural Language Processing (NLP) models are applied. To that purpose, the information found in each vacancy is classified into pre-established categories using a dictionary containing the competences, skills, and abilities, that was created ad-hoc to perform this task. It was constructed using the information found on the online job portals.
To this end, some basic concepts were defined in a first step: the dictionary, a semantic field and grammar. Then, the terms used by the OECD and the ILO on competence, dexterity, skill, requirement and capacity were revised, which allowed us to begin extracting dictionary entries. From here the categories were refined, which ended up being defined as follows:

Competence / skill: Knowledge of a specific activity or knowledge that encompasses both theory and practice and allows executing tasks. Ability to easily and accurately perform the tasks of an occupation that require physical movements.

Requirement: Competence, aptitude, knowledge or characteristic requested by an employer. It is generally a mandatory and a verifiable attribute and lacking it reduces the probability of getting the job.

Capacity: Cognitive process related to information management, accessible when carrying out a work task.

Experience: Practice acquired from the exercise of an activity during a certain period of time.

Function: Activity that a person performs within a job.

In this way, the final categories, their definitions and the analysis of semantic features allowed the concrete and real determination of what Mexican employers consider when referring to each of the categories.

3.5. Sorting job offers to structure the database

At this stage we perform a class and category detection algorithm through the search for patterns in the free text of job vacancies. For this activity we mainly use regular expressions. In NLP tasks the use of regular expressions for the patterns' detection in text is very frequent for subsequent analysis and treatment. The extraction of information is the process of locating portions of a text given that they contain information relevant to the needs of a user and provide such information in a manner appropriate to their process.
To enhance DiCoDe's efficiency and accelerate its classification process we have decided to use regular expressions in Perl 6.0, which contains a standardized and standard set of regular expressions, quite powerful and easy to understand. We decided to implement our solution in Perl's modules, because it's open source, free and portable.

ID Job offer identifier (unique sequential integer)

GENDER Requested gender: male | female | indistinct, if it exists

SALARY Salary offered (if any)

AGE Required age (if any)

HOURS Working hours: in hours, full-time | part-time, days of the week

SCHEDULE Field that sometimes contains additional salary information, schedule

DEDICATION Field that sometimes informs about full time, part time, etc.

TEXT Field that contains the information of the company and the offer, if it exists.

For the extraction of the fields corresponding to competence / ability, capacity, experience, requirements and functions a python program was used in which the following process applies:

- Preprocessing of the text: the free text corresponding to job offers is put into lowercase, accents are removed with the aim of avoiding repetitions of words and possible writing errors and vacancies are rearranged in a line.

- Search of fields of interest: from a collection of regular expressions, the areas mentioned above are searched in the previously processed text, these are: competence / ability, capacity, experience, requirements and functions.

- Writing output file: once the fields of interest are extracted from the free text of the offers, they are stored in a csv file with the following columns: text, competence / skill, ability, experience, requirements and functions.

3.6. Dynamic vacancy classification

We perform a filtering algorithm that allows a dynamic classification of job vacancies, creating a structured database for DiCoDe. The algorithm allows identifying those fields mentioned above.
To carry out the above, the following is done:

1. Take the first of the variables and display a list with the different elements it contains numbered by their respective indexes.

2. Ask the user to indicate the number of items he wants to take, from the list displayed, to add to the filter.

3. Based on the number of items selected by the user, you are asked to add the index number of the first item, of the subsequent ones.

4. The console takes the following variable and will display a list with the different elements it contains numbered by their respective indexes.

5. The process is repeated until the filters are created for all the variables

6. Finally, indicate the number of vacancies selected with the filters and the base where they were stored.

3.7. Identification of terms for free text searches

We build an algorithm for identification of terms for free text searches on job offers present in the DiCoDe database. The developed algorithm allows to obtain all those requirements belonging to an area of the macro-taxonomy used for the classification of job offers; This means that it is possible to look for the competences, desired experience, requirements, functions to perform, requested sex, age range and salary belonging to a certain area.
The algorithm receives as input two things: the area on which the search will be conducted and a database where an initial filtering will be carried out to obtain only the job offers that correspond to the area of interest.
Then the values for the columns of sex, salary, age, schedule, competence / ability, ability, experience, requirements and functions of each job offer belonging to the area of interest are obtained, and adds them to a list where there are no repeated elements .
Finally, a file in a standard .xlsx format is returned that contains the values of sex, salary, age, schedule, competence / ability, ability, experience, requirements and functions of each offer in the selected area.

4. Conclusion and perspectives

By tracking and analyzing vacancies posted on job portals, DiCoDe promises to contribute to evidence-based policy making, specially those aspects regarding training and skills development. To that end, the collaboration between STyFE and GIL has led to fruitful preliminary results. The classification task approached with a neural network has retrieved very satisfactory results with 1,000.000 records and 23 classes, taken from one of the portals.
Preliminary results also suggest that this methodology can be easily adaptable to classify vacancies to both other macro-strcutures and to other geographical areas beyond CDMX. These results have contribute to reduce computing time that, given an accumulated stock of about 5.6 million vacancies, has proven crucial. Moreover, the developed methodology and algorithms not only contribute to overcome the challenges faced by DiCoDe but also to extend some of the know-how to other activities undertaken by STyFE.
An appropriate evaluation should be performed to test both efficiency and correctness of classifications into macrostructures and CSA categories, while also robustness and sensitivity analysis to varying inputs is also pending.
With the implementation of language technologies on the job vacancies data, DiCoDe is expected to reduce information asymmetries in different analytical categories (occupation types, CSA, etc) thereby making information more accessible to job seekers, employers and other actors, including STyFE itself. In this sense, it is expected to foster a more efficient job search and match, while also allowing an enhanced identification of training needs to allow, for more equitable job placement opportunities, inter alia.
It could be interesting in the future to implement hybrid strategies in order to make a best match between the job vacancies and candidates (Kessler et al., 2012).
Finally, we also contemplate in the future the use of Automatic Text Summarization techniques, which could generate relevant syntheses (groups) of groups of job vacancies (Torres-Moreno, 2014).

5. Acknowledgements

The project was supported by an agreeement (IISGCONV-084-2019) UNAM-STyFE (Mexico). J.-M. Torres-Moreno was partially financed by the *Laboratoire Informatique d'Avignon* (LIA), Université d'Avignon (France).

6. Bibliographical References

Altamirano, A., Azuara, O., González, S., Ospino, C., Sánchez, D., and Torres, J. (2019). Tendencias de las ocupaciones en américa latina y el caribe 2000-2015. Technical Report IDB-TN-1821, Banco Interamericano de Desarrollo.

Amaral, N., Eng, N., Ospino, C., Pagés, C., Rucci, G., and Williams, N. (2018). How far can your skills take you. Technical Report IDB-TN-01501, Banco Interamericano de Desarrollo.

Boselli, R., Cesarini, M., Marrara, S., Mercorio, F., Mezzanzanica, M., Pasi, G., and Viviani, M. (2018). Wolmis: a labor market intelligence system for classifying web job vacancies. *Journal of Intelligent Information Systems*, 51(3):477–502.

Bumeran. (2020). Bumeran: empleos destacados en méxico. https://www.bumeran.com.mx/. Accessed: 2020-02-16.

David, H. (2001). Wiring the labor market. *Journal of Economic Perspectives*, 15(1):25–40.

ESCO. (2017). European skills/competences, qualifications and occupations. Technical report, European Union. https://ec.europa.eu/esco/portal.

Hart, O. D. (1983). Optimal labour contracts under asymmetric information: An introduction. *The Review of Economic Studies*, 50(1):3–35.

Hernández, C. (2018). Ocupaciones laborales: clasificaciones, taxonomías y ontologías para los mercados laborales del siglo xxi. Technical report, Observatorio laboral: http://www.observatoriolaboral.gob.mx/.

INEGI. (2020). Encuesta nacional de ocupación y empleo (enoe), población de 15 años y más de edad. https://www.inegi.org.mx/programas/enoe/15ymas/default.html#Tabulados. Accessed: 2020-02-16.

ISCO-08. (2012). International standard classification of occupations. isco-08. Technical report, International Labor Office.

Kessler, R., Béchet, N., Roche, M., Torres-Moreno, J.-M., and El-Bèze, M. (2012). A hybrid approach to managing job offers and candidates. *Information Processing & Management*, 6(48):1124–1135.

Montañes Salas, R. M., del Hoyo Alonso, R., Vea-Murguía Merck, J., Aznar Gimeno, R., and Lacueva-Pérez, J. (2017). FastText as an alternative to using Deep Learning in small corpus.

Mortensen, D. T. (2011). Markets with search friction and the dmp model. *American Economic Review*, 101(4):1073–91.

OIT. (2020). Centro interamericano para el desarrollo del conocimiento en la formación profesional. https://www.oitcinterfor.org/taxonomy/term/3366. Accessed: 2020-02-16.

SOC. (2018). Standard occupational classification. Technical report, US Bureau of Labor Statistics.

Stigler, G. J. (1962). Information in the labor market. *Journal of political economy*, 70(5, Part 2):94–105.

STyFE. (2019). Análisis de las intersecciones entre la oferta y la demanda laboral en la Ciudad de México. Technical report, Secretaría del Trabajo y Fomento al Empleo. https://www.trabajo.cdmx.gob.mx.

Torres-Moreno, J.-M. (2014). *Automatic Text Summarization*. Wiley, London.