IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 9, No. 3, September 2020, pp. 535~544
ISSN: 2252-8938, DOI: 10.11591/ijai.v9.i3.pp535-544

Towards a semantic integration of data from learning platforms

Khaoula Mrhar[1], Otmane Douimi[2], Mounia Abik[3], Naoual Chaouni Benabdellah[4]
[1,4]IPSS Research Team, FSR, Mohammed V University, Rabat, Morocco
[2]ENSIAS, Mohammed V University , Rabat, Morocco
[3]IPSS Research Team, ENSIAS, Mohammed V University, Rabat, Morocco

Article Info

Article history:

Received Oct 24, 2019
Revised Apr 4, 2020
Accepted Apr 18, 2020

Keywords:

Conditional random fields CRF
Data integration
E-learning
Long short term memory (LSTM)
MOOCs
Semantic labeling

ABSTRACT

Nowadays, there is a huge production of Massive Open Online Courses MOOCs from universities around the world. The enrolled learners in MOOCs skyrocketed along with the number of the offered online courses. Of late, several universities scrambled to integrate MOOCs in their learning strategy. However, the majority of the universities are facing two major issues: firstly, because of the heterogeneity of the platforms used (e-learning and MOOC platforms), they are unable to establish a communication between the formal and non-formal system; secondly, they are incapable to exploit the feedbacks of the learners in a non-formal learning to personalize the learning according to the learner's profile. Indeed, the educational platforms contain an extremely large number of data that are stored in different formats and in different places. In order to have an overview of all data related to their students from various educational heterogeneous platforms, the collection and integration of these heterogeneous data in a formal consolidated system is needed. The principal core of this system is the integration layer which is the purpose of this paper. In this paper, a semantic integration system is proposed. It allows us to extract, map and integrate data from heterogeneous learning platforms "MOOCs platforms, e-learning platforms" by solving all semantic conflicts existing between these sources. Besides, we use different learning algorithms (Long short-term memory LSTM, Conditional Random Field CRF) to learn and recognize the mapping between data source and domain ontology.

Corresponding Author:

Khaoula Mrhar,
IPSS Research Team,
FSR, Mohammed V University,
Ibn Batouta avenue, Rabat, Morocco.
Email: khaoula_mrhar@um5.ac.ma

1. INTRODUCTION

In the recent years, the number of MOOC (Massive Open online course) has been growing exponentially (more than 2,400 MOOCs exist in July 2015) [1]. Certainly, MOOCs has sparked a big revolution in higher education in the formal and non-formal learning curriculum [2]. Indeed, a growing number of universities (Mohammed V University Rabat, Sherbrooke University, University of Limoges...) began to produce their own MOOCs and integrate the traditional classroom to support face-to-face learning experiences in a blended format. Thus, MOOCs have attracted wide interest from students around the world and led them to explore other MOOCs offered in online platforms such as (Coursera, Open edx…). Therefore, pedagogical establishment face two issues: On the one hand, the majority of universities that adopt these learning strategies are unable to communicate their learning environments (e-learning platforms and MOOCs platforms) by reason of the heterogeneity of these platforms. On the other hand, all of what is

happening in the non-formal learning through the MOOCs remains imperceptible in the formal system. Accordingly, it is difficult to exploit the non-formal learners' feedbacks.

The exploitation of the feedback will improve the quality of the formal learning by adapting the future courses according to the new knowledge and skills acquired through the non-formal courses (MOOCs). In addition to that, through the feedback collected, the recommender system will help learners to target the MOOCs according to their profiles in order to increase their motivation and educational interests. The educational establishments can exploit their learners' data in their own MOOCs or other platforms MOOCs if these establishments are allowed, as part of a partnership, to retrieve their learner's data.

Hence, the need to have a unified system that collects all data related to learners and metadata courses from various educational heterogeneous platforms is asserted. Generally, data in educational platforms are stored in different format, and hosted in different platforms. Therefore, to meet our needs, we require building a consolidated system which integrates course metadata and learners' data and represent them in a suitable format for recommendation and personalization according to learner's profile. This can be difficult because these data sources are both distributed and heterogeneous. Each source has its own data format and its own structure. It also has its own data definition and vocabulary. Therefore, there is a need for flexible and efficient approaches to integrate information from various educational sources platforms.

Data integration is the problem of regrouping data residing at different sources, and offering to the users a unified view of these data [3]. Data integration resolves the problem related to structural and semantic integration, heterogeneity and autonomy of data source. There are different integration approaches proposed to resolve these problems: basically virtual view approach and materialized view approach [4]. Many works present different approaches to integrate heterogeneous data sources using semantic technology [5]. The majority of the works are interested to solve our first challenge that is the problem of heterogeneousness between the learning platforms. Especially the integration of the courses metadata, namely, MOOCLink project [6], which is a web application that utilizes the Linked MOOC Data to allow users to discover and compare similar online course using the enhanced SPARQL search engine. In this project, they used semantic technology to create a semantic data model for educational data (MOOCs) and they published these data as linked data on the Web. The author in [7] proposed an Architecture based on Linked Data technologies for the Integration and reused Open Educational Resources (OER) in MOOCs Context, The framework provides an approach that allows MOOC designers to discover and access to open educational resources that are extracted from open distributed repositories. However, there is a lack of work which aims the integration of learners' interaction data for the exploitation of theirs feedbacks. Indeed, *the question is: how to integrate course metadata and learner's data from heterogeneous educational platforms in a unified system by solving all semantic conflicts?*

Our goal is to offer to universities and establishments of higher education a system that combines data existing in heterogeneous educational platforms (e-learning, MOOCs) with a unified view of these data sources. For this purpose, three steps are to follow: extracting, mapping and integrating.

After the introduction, a state of art of data integration approaches and a survey of information integration tools are presented. In section three, the integration system of extracting, modeling, and integrating data from different educational platforms is described and implemented using karma integration tool. Section four presents the limitations of the semantic labeling approaches which we have encountered with the databases of the learning platforms. The experimentation with the hybrid algorithm CRF and LSTM to improve the semantic types' recognition in Karma Tool is also exposed. We then conclude with a discussion and conclusion.

2. LITERATURE REVIEW

In this section, we briefly review main data integration approaches, and we present a comparison of a range of information integration systems based on several criteria and features.

2.1. Data integration approach

System integration allows the user to access via a unique interface to data stored in multiple and different data sources. The major problem encountered during the process of integration is the heterogeneity of data [8]. Generally, there are two main approaches to integrate heterogeneous data: the materialized approach and the virtual approach.

The materialized approach is the extraction of the useful data stored in heterogeneous sources consolidated and centralized physically in a data warehouse [9]. This approach allows sending direct requests to the warehouse without accessing to the heterogeneous data sources. The main advantage of this approach is the performance in term of time response. Therefore, it has certain limitations; the most important one is

the flexibility. Any change in the source can affect the whole integration, and the integrated data are not refreshed because it depends on the frequency of the update.

This virtual approach is the development of the application that acts as an interface between local data sources and applications of users. This architecture is based on two essential components: the mediator that executes and reformulates users' queries and the wrapper that establishes the link between the local source schema and the global schema [10].

An integrated schema is designed to describe the logic of the interface layer of a data integration system. Local schemas describe the logic of the data in the local data sources. Schema mapping refers to the transformations between objects in local sources and the integrated schema. To specify the correspondence between the schemas source and the global schema, there are many mapping alternatives.

Global as View (GAV) is the expression of the global schema as a function of the local schema. Local as View (LAV) assumes the existence of a global schema and defines the local schema of data sources to integrate as the views of the global schema [11]. The main advantage of this approach is the coherence, because it directly queries the data from sources and not a central database, which ensures more the flexibility and evolution. However, this approach requires the availability of sources in order to respond to user's queries.

GLAV mappings overcome the limitations of both GAV and LAV. In the query reformulation of the GLAV approach, each mapping rule is represented by a conjunctive query written in the global schema associated with a conjunctive one written in source schemas. In this section, the existing approaches of data integration are reviewed under two main categories: material and virtual. Different mapping approaches are cited. In the upcoming section we present a survey of data integration tools.

2.2. Data integration tools

Data integration system allows to share data between various and heterogeneous information sources in different domain (e-learning, bioinformatics, geospatial…) and exploits data from heterogeneous, distributed and autonomous sources. A comparison of a range of data integration tools based on different criteria presented in the following part is proposed.

2.2.1. Comparaison critera

The criteria taken into consideration are:

Mapping approach:

The data integration systems based on mediation approach uses a semantic mapping between the schema of data sources and the mediated schema to answer user queries. That's why a mapping approach is followed.

Integration technique:

The data integration tool uses a technique to integrate source data. Such as matching, rewriting and view creating. Matching is the linking concept in the global model with the data sources. View creating concept defines the global model as a collection of views sources. Rewriting concept is for a rewrite and for a queries translation.

Query language:

After the mapping process, we retrieve data from the sources indirectly by querying the global schema. It is the task of the mediator that consults the mappings to decide which data to retrieve from the sources and how to combine them appropriately in order to form the answer to the query.

Data source type:

To ensure integration process, we must have the ability to access to several data source such as database system, flat files, web services, xml files.

2.2.2. Comparaison data integration systems

Based on the criteria explained before, a range of data integration tools are compared.

*Agora***:**

Agora [12] presents an architecture based on the LAV mapping to integrate relational databases and structured documents. Thus, for the query evaluation process, Agora uses XML as a user interface format. Queries are posed in Xquery, which is a standard XML query language developed by the W3C.

AutoMed:

In the AutoMed project [13] developed, the first implementation of a data integration technique is called Both-As-View (BAV). It uses a BAV mapping to integrate relational database, XML file and flat files. AutoMed uses AIQL languages to generate queries.

KARMA:

Karma [14] is a web application that enables users to perform data-integration tasks. It provides support for extracting data from variety of sources for cleaning and normalizing data, modeling it according to a

vocabulary of the user's choice. It allows the integration of multiple data sources, building a model or semantic description of each source and publishing in a variety of formats (CSV, KML, and RDF).

PICSEL:

PICSEL [15] is a semantic data integration approach that uses a logical formalism to represent both the domain of application and the contents of data sources. It uses CARIN language to mix the LAV and GAV approaches in order to avoid the query reformulation problem.

TSIMMIS:

TSIMMIS [16] is one of the first system that supports semi-structured data.It offers a data model and a common query language MSL or LOREL. It is a mediator data integration approach that uses many mediator with their independent logical integration schema, it uses GAV approach for schema mapping. Table 1 summarizes the features of different data integration tools presented in this section.

Table 1. Comparison information integration tools

Information Integration Tools	Integration technique	Mapping Approach	Query Language	Resources Type	Automatic semantic labeling
AGORA	Rewriting	LAV	Xquery	XML Relational	No
AutoMed	Matching	BAV	AIQL	Relational, XML, flat files	No
KARMA	Matching	GLAV	SPARQL	Spreadsheets, relational databases, web services CSV, JSON, XML	Yes
PICSEL	View creating	LAV	CARIN	Services	No
TSIMMIS	View creating	GAV	MSL/LOREL	Semi structured	No

In this section, we presented different integration approach and a comparison of a range of information integration systems. In the next part, we will present our integration system.

3. PROPOSED MODEL: A SEMANTIC INTEGRATION SYSTEM OF DATA FROM LEARNING PLATFORMS

The integration system is responsible for integrating data from different heterogeneous learning platforms (MOOCs and e-learning platforms). In this section, we present firstly our motivation and secondly the architecture of our integration system.

3.1. Motivation

The Learners' profile in our system is the bridge which links the formal to the non-formal learning. Indeed, by enriching the learner's profiles with the information emanating from the learners' interaction with MOOCs, the pedagogical establishment will be able to improve the quality of learning by adapting their curriculum according to their profile. To do so, the integration architecture of this system must be able to integrate all data related to learners' profile "skills and knowledge acquired, progression in activities and learners interaction" from various educational heterogeneous and distributed platforms. Unfortunately, this data is dispersed across several platforms, so it is difficult to have a complete learner's profile. For setting up the system above, educational data are needed to be integrated (MOOCs, E-learning) by offering heterogeneous platforms and data related to learner's profile saved in various platforms in the unified framework. This data is spread across several heterogeneous platforms and is represented differently, so an efficient and flexible integration system is required to ensure the following tasks:

- Collecting and retrieving learners' data and courses metadata from different MOOCs and e-learning platforms.
- Modeling learners' data and courses metadata collected in the specific format and enriched and updated learner profile in real time.
- Collecting metadata MOOCs and e-learning courses in real time to recommend to learners new MOOCs according to their profile updated.

In our previous work [17], we presented a federating environment for MOOCs FEM. The main objective of this environment is to provide to the formal learning environment a recommender system of MOOCs. FEM is composed of an integration layer and a recommendation layer of MOOCs. In this paper, the integration architecture is used in the integration layer of FEM environment.

3.2. Architecture of our integration system

The integration system is responsible for integrating data from different heterogeneous educational platforms. These platforms store its data in different heterogeneous databases and in different

format (Json, XML…). Therefore, to facilitate the regular access to the data sources, the proposed integration system is based on virtual semantic integration approach. The tasks of the integration system are: collecting and retrieving data of learners and data related to courses from different platforms, then modeling data of learners and courses in a unified format to facilitate the response to user request and resolve all semantic conflicts.

The architecture is presented in Figure 1, it is composed of three layers: data gathering, data modeling, and data mapping. Based on this integration architecture, we implement the integration process using KARMA which is one of the information integration tools presented in section 2.2. We choose this tool because it uses the GLAV mapping approach which overcomes the limitations of both GAV and LAV. Besides, KARMA is based on ontology to solve semantic conflicts and it has an ability to learn and recognize the mapping of data to ontology based on learning algorithm. We will dedicate the section 3.3 to discuss the learning algorithm of semantic labeling.

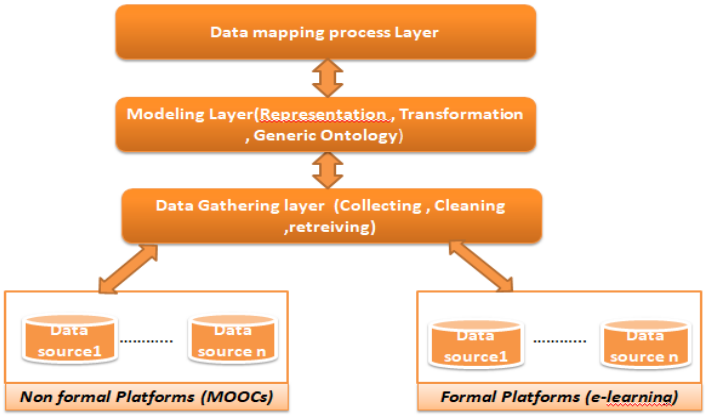

Figure 1. The integration architecture of heterogeneous educational data

3.2.1. First layer: data gathering

In this part, we present how data from various MOOC platforms and e-learning platforms are gathered. We aim to identify, select and collect data from different platforms to demonstrate the efficiency of the proposed system. Three MOOCs platforms OpenEdx, Canvas and the Learning Management System (LMS) Moodle are considered as examples. These open sources platforms are chosen in purpose to add other important platforms (Coursera, Udacity …) when we had an access to its database. These data are related to the:

- Learners' profile: General information such as (the names, email, the levels of education, country…).
- Information related to the progress of the learner in the MOOC and e-learning in order to define the degree of accomplishment of a course such as (scores, grade…)
- Information concerning the MOOCs and online courses such as the name, the description of the course, the start date and end date…

The integration system Karma can enable users to quickly and easily integrate data from a variety of data sources. This means that Karma provides a support for extracting data from a variety of sources (relational databases, CSV files, JSON, and XML). In our case study, we can make a connection with different types of databases of MOOCs and e-learning Platforms. For example: to make a connection with edxapp table or JSON file in open edx platform and with mdl_course database in moodle is possible.

3.2.2 Second layer: modeling layer (representation)

The ultimate goal at this step is to convert heterogeneous data into a unified format. Each data source has its own structure and vocabulary. Namely, the courses data in open Edx platform are stocked in Mongodb database which is a Nosql database, and course data in Moodle platform stocked in mysql database, and it is possible to have other format such as json file or xml file.

This heterogeneity causes several structural and semantic conflicts, such as: the name of conflicts that appears when different terminologies are used in organizations and structural conflicts lies when different choices of modeling construct or integrity constraints are adopted.

To overcome these problems of heterogeneity and conflicts, this work is based on semantic solution where the ontology has an important role in providing conceptual knowledge and the semantic vocabularies

that make the domain available to exchange and to read information in the system. The generic ontology is proposed for aligning the extracted data. Figure 2 shows the structure of the generated generic ontology. The subclass session, section and organization are used to model courses' data. Similarly, the class person represents teachers and learners' data and their progression in course.

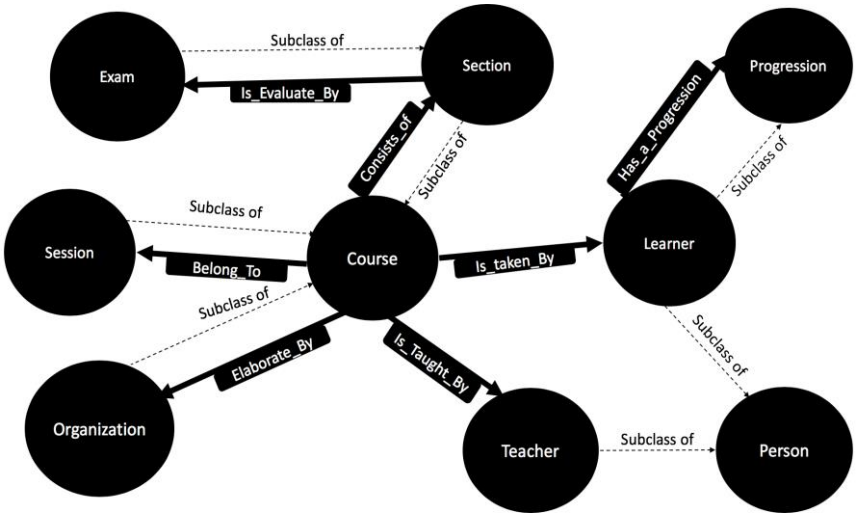

Figure 2. Generic ontology of learning platforms

3.2.3. Third layer: data mapping process

To ensure a correspondence between the generic ontology and sources platforms MOOCs and e-learning (OpenEdx, canvas, moodle…), we align the extracted data by defining a mapping.

The mapping is based on the GLAV approach which overcomes the limitations of both GAV and LAV and it is recommended for queries over the ontology. The mapping process in karma tool consists of the four steps: the assignment of semantic types, the specification of relationships, the generation of source descriptions and the generation of RDF document [18]. The input of the mapping process are: an OWL generic ontology , the data sources that we want to map to generic ontology, and a database of semantic types that the system has learned to recognize based on prior uses of the tool [18]. The output is a RDF triples that represents the content of the sources aligned to generic ontology.

To assign the semantic types, karma proposes the semi-automatic process that is based on user's guidance. Karma assigns the types automatically based on the data's values in each column it is also based on a set of the learned probabilistic models that is based on conditional random fields CRF algorithm constructed from assignments done in prior sessions. After the type's assignment, we can construct a subgraph that connects all nodes with all columns in the tables. Karma in this stage uses a Steiner tree algorithm to compute a minimal subgraph or set of sub graphs that connects the nodes and present them to the user. If karma proposes incorrect semantic types or inappropriate sub graphs, the user can modify them.

During our implementation with karma tool, we noticed that the assignment types are incorrect after the "cold start" and after many prior sessions. It needs several learning experience to have better results. To solve this problem and to improve the detection of the semantic types the hybrid algorithm is used between CRF and LSTM which gives better result in Named Entity Recognition. We present in the upcoming section, our experimentation for semantic labeling for learning platforms data sources.

3.3. Semantic labeling for relational data source

The integration process of different heterogeneous data sources must follow two main steps. Firstly, the semantic labeling step which is the assignment of semantic types to data attributes in data sources. Secondly, the specification of the relationship between the semantic types is made. The semantic types specify the mapping between attributes in diverse data sources with different schema and classes, properties in the corresponding domain ontology.

To finalize the semantic labeling stage, the manual method is very exhaustive, for this reason, several works propose approaches to automate the semantic labeling process. However, it is difficult to have

a high accuracy for the automatic or semi-automatic semantic labeling process because people represents the data in different ways (similar label with different data or different label with similar data).

The majority of works are interested to solve these challenges, such as karma. They proposed a semi-automatic process using Conditional Random Fields (CRF) to learn the assignment to semantic types to columns in data source from users provided assignments [19]. Besides, the machine learning approaches for semantic labeling is categorized into unsupervised and supervised technique. The authors in the reference [20] proposed a benchmark with an evaluation strategy. It compare different approaches for supervised semantic labeling such as: Data INTegrator (DINT), two Deep learning CNN architectures, Multi-Layer Perceptron (MLP). The main conclusion of this comparison is that each semantic labeling approach has its strengths and weaknesses, and the choice of an approach depends on the use case. On the other hand, in the same paper the DSL approach gives a good precision by leveraging information about labeled instances from other domains. Furthermore, The DSL approach [21] learns a matching function to assign the semantic label for data depending on the learned similarity metrics.

Moreover, a notorious limitation of this approach, especially for textual data semantic labeling are:

Firstly, the similarity metric is based on vector space model, the main disadvantage of this method is that it is used in the lexical level and not in the semantic one. The reason is that it ignores the semantic relationship among words and treats words independently. Thus, if two columns use different collections of words to represents the same attributes, they can be assigned to different semantic label for the words.

Secondly, in case of the existence of a multi-lingual data sources there are many limitations. In our case, MOOCs are an alternative model for education in the developing countries and one potential challenge for global use of MOOCs is to offer MOOCs in different language. Indeed, there are a huge MOOCs in different languages emanating from different platforms, such as French MOOC platform FUN, Arab MOOC platforms Edraak. Assigning the semantic label of columns from data sources written in various languages, such as in the MOOC information database where the description of course column is written in various languages according to the language of the courses. As appeared in Table 2, the semantic labeling may not give a good results if the similarity method doesn't support the cross lingual data similarity. A possible way to resolve this problem is to unify the language space by using machine translation between languages [22]. Or enriching data representation with knowledge background like Wikipedia and using their inter-language links.

Table 2. Sample data from course open EdX platform

ID	Name	Start-Date	Overview
CS50X/2018	Introduction to Fintech	2013-05-15	Over the past decade emerging technologies…
ER22x/2016	Gestion de projets de développement	2016-02-15	Jour après jour les gouvernements, institutions publiques…
Ph207X/2013	Android: Introducción a la Programación	2013-01-03	Android es la plataforma libre….

Consequently, in our case we suggest to use a cross-lingual similarity method to give a better precision in similarity metric used in the training algorithm.

In addition, karma uses a probabilistic graphical model to solve the problem of semantic labeling. It assigns semantic types to every value in an attribute and then combines these semantic types to infer the semantic type for the whole attribute. In our implementation, we use a model for semantic types recognition based on a combination between CRF and LSTM that takes advantages from both generative and discriminative model in order to improve the accuracy of semantic type recognition.

3.3.1. LSTM-CRF

Recurrent neural networks (RNNs) are a family of neural networks that operate on sequential data. They take as input a sequence of vectors (x1, x2, ..., xn) and output a sequence of class posterior probabilities, (y1, y2, ..., yn) . An intermediate layer of hidden nodes (h1, h2, ..., hn) is also part of the model.

Moreover, long short-term memory (LSTM) was introduced by [23], it is a special architecture of RNN, capable of learning long-term dependencies. LSTM replaces hidden units in RNN architecture with units called memory blocks. Each block contains one or more self-connected memory cells and three multiplicative units - the input, output and forget gate [23].

Conditional random fields (CRF) is a probabilistic model for structured prediction introduced by Lafferty [24]. It became more and more popular models during the last decade for sequence modeling because they are discriminative models and they do not rely on the same restrictive assumptions.Structured output prediction aims at building a model that predicts accurately a structured output vector y = {y0, y1, .., yT } for any input sentence x = {x0, x1, .., xT }. The inputs and outputs

are directly connected, as opposed to LSTM and LSTM networks where memory cells/recurrent components are employed.

The combination of a LSTM network and a CRF network is used in [25]. This network can efficiently use past input features via a LSTM layer and sentence level tag information via a CRF layer. Characters of each word in a sentence are fed into a LSTM network to catch word character-level. Then these character-level vectors are concatenated with word embedding as word representation and put them into LSTM network. Then the outputs of the LSTM network will be fed into the Conditional Random Fields (CRF) layer. The parameters of LSTM layers (weight matrices, biases, word embedding matrix) and transition matrix of CRF layer are tuned during training stage by back propagation algorithm with stochastic gradient descent. Then, they add the dropout training into input and output layers during the LSTM training.We apply this combination between LSTM and CRF network in semantic types recognition step in integration process in karma, in order to compare the results with CRF model.

4. EXPERIMENTS AND RESULTS

In this section, we aim to compare the accuracy of semantic type's recognition between CRF model and LSTM-CRF model. For our experimentation, we trained our model LSTM-CRF on different dataset from multiple domains and based on the nature of semantic labels to be assigned in the data sources. We choose datasets that contain different types of named entities and that we can found it in database in learning platforms for resolving the cold start in karma. Then we proposed to users the data that they want to integrate in the learning platform according to the model to facilitate the automatic process for semantic labeling. Some dataset [26] that we used are: Name (person name, hotel name), Location (Cities, countries,...), Organizations (Universities, companies, establishments,…) , description courses (Description of content of a courses and their pedagogical objectives in different domains), Topics Date, Time.

The model was tested by using four databases: open edx learners, open edx courses, moodle courses and moodle learner's data bases. Two experiments were made, 1) CRF model was applied in karma to label each sources attributes to semantic types, 2) LSTM- CRF model to label each sources attributes to semantic types was applied too. The objective of this test is to compare the model used by karma CRF and LSTM CRF model in the assignment and recognition of semantic types

Using this proposed integration system of heterogeneous educational data, the evaluation with training model for semantic type identification is executed by experimenting four tables, learner's profile, and course tables in both open edx platform and Moodle platforms. We compared the correct semantic type recognition obtained that we don't need to users actions (menu choices to select correct semantic types if it's incorrect) between CRF and CRF-LSTM models. As shown in Table 3, karma with CRF model was able to accurately infer the semantic types for 62.2% columns and require manual assignment for the remaining columns. LSTM-CRF model was able to accurately infer the semantic types for 80.3% columns and require manual assignment for the remaining columns.

According to the evaluation, LSTM-CRF method improves the accuracy of semantic type's recognition more than CRF method. The use of LSTM-CRF model is recommended in the assignment semantic type step in the integration process to improve the accuracy of the semi-automatic assignments semantic types and for the mapping of data sources column to a node in the ontology.

Table 3. Comparison evaluation result between CRF karma and LSTM-CRF model

Sources	Table Name	Correct semantic types recognition	
		Karma CRF Model	LSTM-CRF Model
Open Edx	Learner profile	66.7%	83.3%
	Course	61.5%	84.6%
Moodle	Learner Profile	57.1%	71.4%
	Course	63.6%	81.8%
		Total= 62.2%	Total= 80.3%

5. DISCUSSION AND CONCLUSION

A critical challenge of educational data integration is its distribution and heterogeneity. Indeed, each educational resources are hosted in different platforms "MOOCs and e-learning platforms" and every platform has its own format and structure.

In this paper, a semantic data integration system is proposed. Item powers pedagogical establishment to rapidly extract their data and semantically map and integrate them from various

heterogeneous sources. Three steps are considered: the first is to collect and to extract data from various educational platforms, the second is to create a generic ontology for educational data, and the third is to align and map the generic ontology to extract data to resolve all semantic conflict. This system is implemented within an information integration tool called Karma that is chosen based on a comparison with others according to several criteria. The integration process in karma follows two steps: firstly, the semantic labeling step which is the assignment of semantic types to data attributes in data sources; and secondly, the specification of the relationship between the semantic types. Based on CRF model, karma proposes a semi-automatic approach that generates a mapping from the data source into the ontology. Since the precise mapping is sometimes ambiguous, the user is allowed to interactively refine the mappings.

To improve the accuracy of semantic labeling in the integration of data emanating from different learning platforms the model for semantic labeling is used. It is based on hybridization between CRF and LSTM that takes advantages of both generative and discriminative model and already trained on datasets of the existing data in the learning platforms. Our preliminary experimentation showed that LSTM-CRF model gives better result in automatic assignments of semantic types than CRF model.

We plan in future works to apply this integration system in our university Mohammed V by bridging their e-learning Moodle platform and MOOCs platform open Edx in a consolidated system which contains all data related to courses and learners. We also consider integrating additional MOOCs platforms by resolving all possible semantic problems. Another future work will be to enhance the semantic labeling for multilingual relational data source, by proposing a semantic labeling approach based on semantic similarity metric as features and support the cross lingual similarity.

REFERENCES

[1] T. Brahimi, A. Sarirete. "Learning outside the classroom through MOOCs". *Computers in Human Behavior*, vol. 51, pp. 604-609. 2015.

[2] K. Jordan. "Initial trends in enrolment and completion of massive open online courses". *The International Review of Research in Open and Distributed Learning*, vol. 15, no. 1. 2014.

[3] Lenzerini M. "*Data Integration: A Theoretical Perspective*". PODS '02 Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2002.

[4] Tatbul N, Karpenko O., Convey C., Yan J. Data Integration Services. Brown University, Computer Science. 2001.

[5] Cheatham M., Pesquita C. Semantic Data Integration. In: Zomaya A., Sakr S. (eds) Handbook of Big Data Technologies. Springer, Cham. 2017.

[6] Sebastian Kagemann, Srividya K. Bansal. "MOOCLink: Building and Utilizing Linked Data from Massive Open Online Courses". *IEEE 9th International Conference on Semantic Computing (ICSC)*, pp. 373-380, February, Anaheim, USA. 2015.

[7] N. Piedra, J. Chicaiza, J. López and E. Tovar. "An Architecture based on Linked Data technologies for the Integration and reuse of OER in MOOCs Context", *Open Praxis*, vol. 6, no. 2. 2014.

[8] Sheth, A., "Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics". *Interoperating Geographic Information Systems*, pp. 5-29. 1999.

[9] Kermanshahani S. "Semi-materialized framework: a hybrid approach to data integration". ACM. 2008.

[10] G. Wiederhold, "Mediators in the architecture of future information systems", *Computer*, vol. 25, no. 3, pp. 38, 1992.

[11] Bayardo et al., InfoSleuth. "*Semantic Integration of Information in Open and Dynamic Environments*", Proceedings of the 1997 ACM International Conference on Management of Data (SIGMOD), Tucson, Arizona, May 1997, http://www.mcc.com/projects/ 18 infosleuth.

[12] I. Manolescu, D. Florescu, D. Kossmann. "*Answering XML queries over heterogeneous data sources*". Proc. of the 27th Int. Conf. on Very Large Data Bases (VLDB 2001). 2001.

[13] Boyd M., Kittivoravitkul S., Lazanitis C., McBrien P., Rizopoulos N. "AutoMed: A BAV Data Integration System for Heterogeneous Data Sources". In: Persson A., Stirna J. (eds). Advanced Information Systems Engineering. CAiSE. Lecture Notes in Computer Science, vol 3084. Springer, Berlin, Heidelberg. 2004.

[14] Gupta, S., Szekely, P., Knoblock, C. A., Goel, A., Taheriyan, M., and Muslea, M. (2015). "*Karma: A System for Mapping Structured Sources into the Semantic Web*". The Semantic Web: ESWC 2012 Satellite Events: ESWC 2012 Satellite Events, Heraklion, Crete, Greece, May 27-31, 2012.

[15] F. Goasdoué, V. Lattès and M. Rousset. "The use of Carin Language and Algorithms for Information Integration: The Picsel System", *International Journal of Cooperative Information Systems*, Vol. 09, No. 04, Pp. 383-401. 2000.

[16] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y.Sagiv, J. Ullman, V. Vassalos, J. Widom. (997). The TSIMMIS approach to mediation: data models and languages. *Journal of Intelligent InformationSystems*, 8 (2) 117-132.

[17] K. Mrhar, Zary, N., et Abik, M. "*Making MOOCs matter in formal education through a federating environment*". In Proceedings of the European Conference on e-Learning, ECEL, vol. 2010-October, p. 557-565. 2017.

[18] Szekely, P.A., Knoblock, C.A., Gupta, S., Taheriyan, M., & Wu, B. "Exploiting semantics of web services for geospatial data fusion". *GIS-SSO*. 2011.

[19] M. Taheriyan, C. Knoblock, P. Szekely and J. Ambite. "Learning the semantics of structured data sources". Web Semantics: Science", *Services and Agents on the World Wide Web*, vol. 37-38, pp. 152-169. 2016.
[20] Rümmele, N., Tyshetskiy, Y., & Collins, A. "Evaluating approaches for supervised semantic labelling". *CoRR, abs/1801.09788*. 2018.
[21] Gad-Elraba, M. H., Stepanovaa, D., Urbanib, J., and Gerhard Weikuma. "Exception-enriched Rule Learning from Knowledge Graphs". ISWC, 9981(PART 1), 33-48. 2016. https://doi.org/10.1007/978-3-319-46523-4
[22] D. Oard. A comparative study of query and document translation for cross-language information retrieval. *Machine Translation and the Information Soup, Springer*, pp. 472-483. 1998.
[23] Sepp Hochreiter, Jrgen Schmidhuber. "Long Short-Term Memory". *MIT Press*, Vol. 9, No. 8, 1735-1780. 1997.
[24] Lafferty, J. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". *ICML* pp. 282-289. 2001.
[25] Huang, Z.; Xu, W.; Yu, K. "Bidirectional LSTM-CRF models for sequence tagging", *arXiv*; arXiv:1508.01991. [Google Scholar] 2015.
[26] Erik F. Tjong Kim Sang and Fien De Meulder. "*Introduction to the conll-2003 shared task: Languageindependent named entity recognition*". In Proc. CoN. 2003.

BIOGRAPHIES OF AUTHORS

Khaoula Mrhar is currently Ph.D student at IPSS research team in Mohammed V University Rabat, Morocco. Her background includes a degree in mathematics and computer science. Her research interest contains Formal and Non Formal learning, Artificial intelligence, Natural language processing, Deep Learning, Data integration, Text mining and recommender system.

Otmane Douimi is a final year master's degree student, He received his Bachelor degrees in mathematics and computer science from Hassan II University Casablanca, Morocco. He started his early research career at Mohammed V university Rabat, Morocco. His research mainly focuses on Natural language processing, Deep Learning.

Mounia Abik I received a PhD from the National High School for Computer Science and Systems Analysis (ENSIAS) in 2009 and an Habilitation to Drive Research (HDR) from Mohammed V University of Rabat in 2014. My main research interests focus on e-Learning, Knowledge Extraction from Social Networks, Semantic Web and Cyber-violence.

Naoual Chaouni Benabdellah is a scientific researcher who received a PhD in Computer Science in 2015 from the faculty of Science, University Mohammed V of Rabat, Morocco. She is specialist in e-learning. Her research area interests are: computational intelligence, machine learning and e-learning.