1 Introduction

Over the past couple of decades there has been an exponential increase in the amount of data that is generated by the web and more recently the internet of things (IOT). This exponential increase has not always been accompanied by a similar increase in the extraction of information and knowledge from this data. The lag in information gain is mainly due to the large volume and unknown structure of the data that makes their analysis quite a difficult and complex task [1]. To overcome this issue, it is essential to develop the necessary tools that will help understand and manage the content of the data. These tools will allow the correct classification of the data and will incorporate data transformation, cleansing and standardisation as part of pre-processing steps.

On top of data mining and knowledge retrieval there are many cases where different services offered by different applications need to be integrated or data need to be transferred from one system to another (e.g. system migrations). In those cases, it is important to have an interface agreement whereby data from interacting systems need to be well curated and mapped. Oftentimes this is a cumbersome and manual task especially when no additional documentation or system specs are provided.

Within the space described above, the research question we will try to answer as part of this project is the following:

How can we enhance a set of data given as input (e.g. tabular data) with semantic meaning using existing knowledge graphs (e.g. DBpedia, WikiData) as reference?

This project is inspired by the Sem Tab challenge that has been organised annually since 2019. The scope of the challenge is organised in 3 separate but overlapping tasks listed below:

- Column-Type Annotation (CTA) Task: Assign a class from a KG to an entire column of a table
- Cell-Entity Annotations (CEA) Task: Assign an individual entity of a KG to each specific cell
- Column-Property Annotation (CPA) Task: Assign the relationship (i.e. object property) between 2 table columns

The above tasks are reflected in **Error! Reference source not found.**

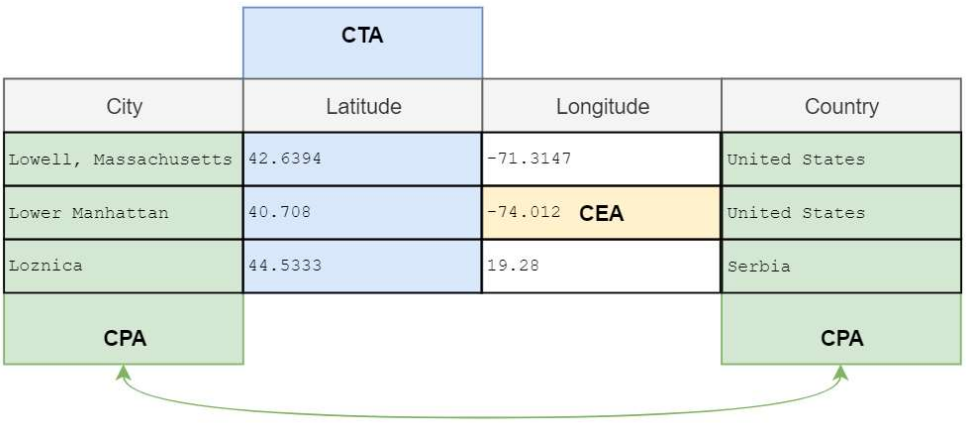

Figure 1. Matching tabular data to classes (CTA), entities (CEA) and properties (CPA)

2 Critical Context

As mentioned in the previous section, the scope of this project has been inspired by the Semantic Web Challenge on Tabular Data to Knowledge Graph matching (SemTab) challenge that was first introduced in 2019 [2]. This challenge aims to create a framework that would systematically evaluate proposed matching systems by automatically generating datasets of increasing complexity and suggesting metrics to assess the accuracy of the submitted solutions.

To generate the benchmark dataset SemTab 2019 used DBpedia as the reference knowledge graph (and in fact its English annotations) however the approach is generic enough that the KG can be replaced by any other (e.g. Wikidata or other domain specific KG). The dataset creation methodology is summarised below:

- **Profiling**: Extracts a list of classes and properties of each class as well as a number of class instances that have the properties populated. The datatype and range is also extracted for the data and object properties respectively
- **Raw Table Generation**: Each identified class is extracted as a table with 3-7 columns randomly selected from the class properties. The process ignores tables with less than 5 rows. Moreover, to avoid large classes overwhelming the dataset there an upper limit (2000) rows that are randomly selected for each class. The process ensures that all values extracted are of the same data type and if multiple values exist of a property, only one of them is selected
- **Refinement**: With the dataset at hand this next step ensures that some noise is introduced to make the data more realistic and representative of a real life use case. This is done by identifying tables that more challenging to match and even within those identifies challenging rows and removes the rest from the dataset.
- **RDF Data**: Finally the process of converting the data to RDF format using column headers to identify class names and object properties while additionally the rdf:label predicate is used to keep the literal values of the cells in the dataset. As implied for the dataset to be converted to RDF it is necessary that tables have a header row.

To evaluate the proposed solutions the challenge used the traditional precision, recall and F1 scores for the CEA and CPA task whereas for the CTA task two different measures where used Average Hierarchical Score and Average Perfect Score in order to take into account the taxonomy of the classes.

In the following sections we present some of the proposed solutions for the SemTab challenge that will form the basis of our implementation

2.1 MTab

MTab [2] is the top performing proposal out of all participants in the 2019 challenge across all three tasks.

Prior to dealing with the tasks as hand the proposed system is performing some preprocessing step to clean the data and extract some metadata for the given tables. In summary the preprocessing is attempting to rectify incorrect encoding predict the language of the table cell values as well as metadata on their data types and entity type from another KG (OntoNotes 5) which is also manually matched to DBpedia. Finally, the system performs a lookup directly to DBpedia or indirectly through redirected links from Wikidata to retrieve a list of candidate entities considering the language parameters identified in one of the previous steps.

Following the prepossessing there is a two-phase approach whereby:

- in the first phase the system estimates the candidate according to relevance for each of the 3 tasks (entity, type and relation) and
- in the second phase those candidates are refined to come up with the final output of entity, property and class respectively. For entity re-estimation the algorithm is calculating the probability by combining the probability of the candidates from the first phase (entity, type and relation) with certain weights. Finally, it is using the entity with the highest probability to define the property and class.

The key strengths and differentiating elements of this approach are:

- the use of additional services, specifically DBpedia Lookup, DBpedia endpoint, Wikipedia and Wikidata.
- the language consideration
- the introduction of a fuzzy matching (e.g. Levenshtein distance) instead of exact term matching when looking up candidates

2.2 IDLab

Similar to MTab, IDLab is also proposing a solution for all three subtasks of the challenge. IDLab adopts a multistep approach trying to resolve the more specific subtasks of CEA and then using the results to infer the properties and classes.

To identify the candidate entities for CEA IDLab uses the cell value, following some basic text processing, to generate a URL in the dbpedia domain. If that URL is a valid one that entity is added to the pool. In parallel, the DBpedia Lookup is utilized with the cell value to produce more candidate and in case both of the above yield no valid results, DBpedia spotlight is used. Finally to derive the best match the system calculate the smallest Levenshtein distance between the rfs:label of each candidate entity in the pool and the cell value.

For the CTA tasks the class of the identified CEA entities are assessed taking into account a majority vote as well as the taxonomy that in inherent to DBpedia. This solution is trying to select the deepest node in the tree while at the same time also maintaining all the ancestors and equivalents as candidates for classes. This step is repeated near at the final step to refine the selection.

For CPA the system is also assessing all predicates of the (subject, object) pairs between columns and selects the most frequent one. Domain and range of column types is considered to break ties.

2.3 ColNet

Another approach to assign types (i.e. classes) to columns of tabular data has been proposed by ColNet [5]. ColNet doesn't assume that there are any metadata like column names or even entities of the cell values in the reference knowledge graph. Instead it uses the KG to automatically train a convolutional neural network CNN that would then predict types for columns not only based on the individual cell values but also the embedded semantics of the entire column. That way it also manages to address the presence of knowledge gaps in the KG (i.e. the instances where not all cell values from the tabular data have a corresponding entity in the KG).

In summary ColNet comprises three key steps: lookup, prediction, and ensemble. In the lookup step entities in the KG that are matched to the column cell values are retrieved and their classes are added to the list of candidate classes for the column type annotation. The matched entities form a set that is called particular entities and all entities of the candidate classes for a set of general entities. These two sets are used to form positive and negative training sample to train the CNN.

After training the CNNs for each candidate class with the positive and negative samples that have been converted to vectors using a word representation model like word2vec ColNet tries to predict each column type. In the final step, ensemble, the CNN predictions are combined with the entities retrieved by the lookup in a way that rejects classes supported by few cells while accepting classes supported by a large part of the cells

3 Approaches: Methods & Tools for Design, Analysis & Evaluation

Instead of following a bottom up approach whereby we first address the CEA task and then we infer the solution for the CTA and CPA the proposed approach will try to identify the types of the columns first and that will narrow down the options for entities and properties. The above logic is not fully applicable to columns with literal values that will not have a type or corresponding entity.

The steps provided below will not be followed in a strictly waterfall approach but the lifecycle will be more of an agile style where each round will analyse, design, implement and evaluate a certain component and perhaps revisit it once again in case the results are not satisfactory before moving on to the next component. However, for the purposes of this report the approach for each of these steps will be presented once with details covering the entire project.

3.1 Analysis

As part of the analysis step, we will undertake a more extensive literature review on papers or books that have done similar work as the one proposed for this project. The initial literature review undertaken for the purposes of the RMPI proposal was more focused on trying to understand the space in which the project is so it was more wide than deep.

The literature review to be undertaken during the project will focus more in detail of specific parts implemented by others working on this field and will also consider the code (where available) to either reuse as an external library (referenced) or modified to fit the solution that will be implemented by the project itself.

Although not technically considered as analysis, in this step we will also explore the data we will use as input to the system. This data will be taken from the SemTab 2019 challenge that has proposed 4 datasets of increasing complexity that are tagged so that they can be evaluated and can also be used as a benchmark of the system against other suggested approaches like the top three systems of MTab [3], IDLab [4] and Tabularasi [6].

3.2 Design

The design of the system is based on a modular structure that will allow for various components to be enable/disabled at will to allow for testing the impact they have in the final outcome. A high-level view of the components can be seen in Figure 2.

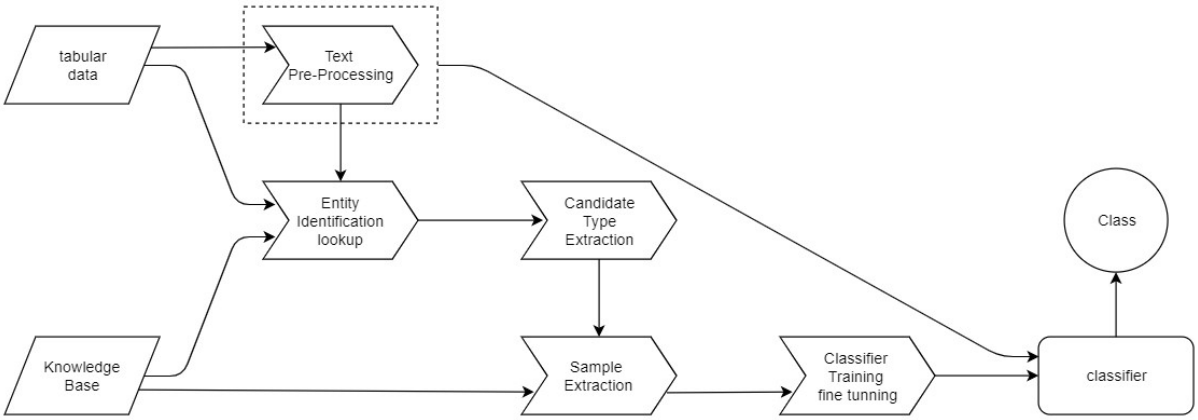

Figure 2. High level design of the proposed solution

As shown in Figure 2 the design assumes as input a set of tabular data and a reference knowledge graph. At the point of writing this proposal it has not yet been decided which knowledge bases will be used however we know that DBpedia will be the first point of reference and then we may test the solution for additional KB like in DAGOBAH [7]. Other potential references include Wikidata API and Wikipedia API provided we can evaluate the outcome.

The first module to be built is one that will pre-process the data from the table cells to either extract some metadata on the types (e.g. string, number and range) or cleanse and standardize data by removing noise and perhaps strip words down to their stem/lemma. This step will be one of those that can be turned on and off to investigate if the lexical information from the word suffixes is important in the type/entity identification.

The cell data (pre-processed or in its raw form) will then be used to extract the candidate classes from the knowledge base that are linked to the identified entities. This step will be done for all columns in the table apart from the ones that our textual analysis has identified as containing literal values. Following that, entities belonging to the candidate classes will be extracted from the knowledge base in order to form a set of training samples with both positive and negative results. These samples will be used to train a set of classifiers (the details of which are yet unknown and will start formalising post the analysis step). The idea is to somehow introduce a distance between classes so that classes that are the domain/rage of object properties can be considered better candidates for a pair of columns compared to classes that are not linked. The classifiers will then be presented with the previously unseen data and try to predict the correct types for the columns.

With the column types predicted we will then try to address the remaining two tasks by lexically matching the cell values yet again but only to the entities of the given class or any of its parents. This is envisaged to provide a better accuracy than performing the lookup across the entire KB.

Finally given that we've successfully managed to identify the entities and even the types, the system should be able to infer the relationships between the columns.

3.3 Implementation

The proposed solution will be implemented in Python using existing libraries that are well documented and widely used in this space. The code will be structured in a way that will enable users to determine the parameters of the execution via a configuration file (i.e. selection of KB,

enable disable components, etc.). Any code snippets implemented by others and reused will be well documented to make the contributions from this particular project clear.

3.4 Evaluation

The proposed solution should both be able to produce high quality results but at the same time do so in a reasonable timeframe. To measure the accuracy of the system we will use the traditional metrics of precision, recall and F1 scores whereas we will also take into account the measures proposed in [1] for average and hierarchical score when evaluating the column types.

The evaluation will both compare the accuracy for different setting combinations of the proposed solution as well as compare the best version with the other submissions on the SemTab 2019 challenge for the relevant tasks.

Finally, we will make sure that the solution has been built in an efficient way, so the processing time and memory usage are such that allow for it to be widely used.

4 Risks

ID	Risk	Likelihood	Impact	Mitigation
1	Computational Power	50%	Medium	Use of cloud providers like Amazon Web Services or Google Cloud
2	Hardware Failure / Loss of work	10%	High	Use of a Git repo to do frequent saves after every key change
3	Feasibility of project completion within the given timelines	20%	High	Approach in modular way so that different areas can have an MVP that is succeeded on time can then be improved further
4	Unplanned work due to unknown implementation issues	30%	Medium	Have intermediate milestones to access the proposed approach and propose scope amendments in agreement with the supervisor
5	Use existing code / libraries as a starting point may be hard to reproduce	40%	Medium	Reverse engineer the code and perhaps built simpler solution where there is ambiguity
6	Unpredicted workload may eat up time originally dedicated to project work	60%	Low	Try to stick to the work plan milestones and supplement with additional days of unpaid leave when needed to catch up on progress
7	Failure to compare with other proposed systems	40%	Low	Use the SebTab evaluation results as benchmark

Table 1. Risk that the candidate may have to cope with and ways to mitigate them

5 Work Plan

Table 2 presents the work plan that will be followed for the completion of the individual project. Although the tasks are presented in a waterfall way there is scope to revisit previously completed tasks once more information is available in order to fine tune them. The duration of each of the tasks has incorporated that logic. Finally, it is assumed that work will kick off after the announcement of the exam results which for now is estimated at the beginning of June. If that date slips, then the plan will need to shift accordingly.

	Jun-21					Jul-21				Aug-21				Sep-21				Oct-21					Nov-21				Dec-21			
	31-May	07-Jun	14-Jun	21-Jun	28-Jun	05-Jul	12-Jul	19-Jul	26-Jul	02-Aug	09-Aug	16-Aug	23-Aug	30-Aug	06-Sep	13-Sep	20-Sep	27-Sep	04-Oct	11-Oct	18-Oct	25-Oct	01-Nov	08-Nov	15-Nov	22-Nov	29-Nov	06-Dec	13-Dec	20-Dec
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Preparation Phase																														
Literature review on NLP (stemming/lemmatising)																														
Literature review on converting text to vectors																														
Literature review on classification for KG embeddings																														
Set a programming environment																														
Design Phase																														
Text Processing Feature																														
Integrate with KB to lookup entities extract classes																														
Design Classifier																														
Design a solution for CEA																														
Design a solution for CPA																														
Implementation Phase																														
Implement Text Processing Feature																														
Implement KB Integration																MVP														
Implement Classification for CTA																														
Test and Experiment run																														
Experiment with other KGs																														
Implement CEA																														
Implement CPA																														
Additional experimentation for CEA and CPA																														
Code Refactoring																														
Reporting																														
Write up initial Draft																														
Send for review																														
Incorporate review feedback																														
Submit thesis																														

Table 2. Suggested work plan for the individual project

6 Ref

[1] A. Lausch, A. Schmidt, L. Tischendorf, Data mining and linked open data – New perspectives for data analysis in environmental research, Ecological Modelling 295, 2015, p. 5-17

[2] Jimenez-Ruiz, E. , Hassanzadeh, O., Efthymiou, V., Chen, J. and Srinivas, K. (2020). SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. In: The Semantic Web. ESWC 2020. Lecture Notes in Computer Science. (pp. 514-530).

[3] Nguyen, P., Kertkeidkachorn, N., Ichise, R., Takeda, H.: MTab: Matching Tabular Data to Knowledge Graph using Probability Models. SemTab, ISWC Challenge (2019)

[4] Steenwinckel, B., Vandewiele, G., Turck, F. and Ongenae F.: CSV2KG: Transforming Tabular Data into Semantic Knowledge. SemTab, ISWC Challenge (2019)

[5] Chen, J., Jimenez-Ruiz, E., Horrocks, I. and Sutton, C. (2019). ColNet: Embedding the Semantics of Web Tables for Column Type Prediction. Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, 33, pp. 29-36

[6] Thawani, A., Hu, M., Hu, E., Zafar, H., Teja Divvala, N., Singh, A., Qasemi, E., Szekely, P. and Pujara, J.: Entity Linking to Knowledge Graphs to Infer Column Types and Properties. SemTab, ISWC Challenge (2019)

[7] Chabot, Y., Labbe, T., Liu, J., Troncy, R.: DAGOBAH: An End-to-End Context-Free Tabular Data Semantic Annotation System. SemTab, ISWC Challenge (2019)

[8] F. Kalloubi, E. H Nfaoui, O. El Beqqali, Micro blog semantic context retrieval system based on linked open data and graph-based theory, Expert Systems With applications 53, 2016, p. 138-148

[9] T. Di Noia, V. C. Ostuni, J. Rosati, P. Tomeo, E. Di Sciascio, R. Mirizzi, C. Bartolini, Building a relatedness graph from Linked Open Data: A case study in the IT domain, Expert Systems With Applications 44, 2016, p. 354-366

7 PART A: Ethics Checklist.

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/*	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/*	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study?	**NO**

	Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/	
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? *For example, students studying on a particular course or module.* *If yes, then approval is also required from the Head of Department or Programme Director.*	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**NO**