Enabling Tabular Data Understanding by Humans and Machines through Semantic Interpretation

Marco Cremaschi

University of Milano - Bicocca, Milano, Italy
marco.cremaschi@unimib.it

Motivation and objectives of the research. In today's society, the value of information is considerably higher than any other sort of contribution. With the term "information" we mean the formula that binds a certain amount of data to a specific context. Digitisation has caused an explosion of data in every sector of our society that generates a huge amount of information, basically useless on its own because it is disorganised and complex. This huge amount of data, commonly defined as Big Data, can be categorised as structured (tables) or non-structured (text) and can originate not only from human beings, but also from technology (think about sensors and other tools of data acquisition).

The Web as it was conceived would contain resources mainly designed and produced for human consumption, rather than machines. As a matter of facts, reading and understanding a web page is very simple for a user, but it is not so easy for computers, which are able to distinguish different *markup* elements that define a page but cannot comprehend its meaning or how those elements are interrelated. The attempt to correct this weakness led to the creation of *Semantic Web* [1]. The idea behind this is precisely to transform web *documents* into *data* to which we assign a unique and well-defined meaning, adding information on how they can be linked together.

A huge source of data are the relational tables embedded into Web documents. When we happen to consult a table, some cognitive processes that get triggered inside our brains allow us to give table values -expressed in form of free text- a meaning (or a semantic), that can be more or less accurate depending on several factors, like our familiarity with the domain of the exposed data, the background knowledge, or the presence of contextual information. The free-text content of tables cannot be understood by machine if it does not refer to a Knowledge Graph (KG) that assigns well-defined meaning to terms and support the creation of links between them -the real added value of the Semantic Web. Therefore, the issue is to find the right association between values in tables and concepts/properties of a KG that describes real-world entities and their inter-relations, organised in a graph [11]. The attempt to solve this problem led to the birth of Semantic Table Interpretation (STI) techniques, which are processes through which, given a table and a KG as input, it is possible to interpret the structure and the semantic of the table by associating its content to semantic concepts taken from the KG.

The semantic interpretation of tables has different application fields of great importance, such as: (i) Data search, which is the option of making the relational data contained within tables accessible through human searches; (ii) Data enrichment, which is the option of completing and extend the table's content with other data and therefore with additional information that come from other sources; and (iii) KG construction/KG population, meaning the option of building or enrich a KG thanks to the semantic information found by the annotation process.

Moreover, there is an emerging application of STI, which is the translation of tables into natural-language sentences to help users understand the table content even if they are not familiar with the KG terminology, or use devices that cannot fully display tables (e.g. smartphones with audio interaction capability via chatbots).

In the thesis we address the Semantic Table Interpretation, the API composition using Semantic Annotations and Natural Language Generation of tabular data. In particular we mainly focus on the implementation of a semantic table interpretation approach, with the ultimate goal of making the data within the table more accessible, both to machines and humans. This leads to the following research questions:

- **Question 1**: What does it mean to annotate semantically a table? What are the table's elements that must be taken into consideration?
- **Question 2**: How can elements within a table be made unambiguous? How can the context to support disambiguation be created?
- **Question 3**: What enables the presence of semantic annotations with respect to the processing of data by a machine?
- **Question 4**: What make data more accessible for a user? Is it possible to create a representation of the tabular data in natural language? What is the most relevant information inside the table for the user?
- **Question 5**: What are the techniques for converting a RDF triple into natural language?

Major results. The contributions of the thesis can be grouped in three main results:

1. A fully automated approach to a complete STI;
2. A practical approach to services composition by exploiting light semantic annotations;
3. A pipeline for RFD lexicalization.

Description of the approach, highlighting its key aspects and the novelty with respect to the state of the art. Most of the the state-of-the-art approaches focus on individual aspects of the STI like, for example, the analysis of columns that contains entities or literal. Only recently few works attempting a comprehensive approach have been published. In particular, [14] proposed a very interesting and promising breakdown of the STI problem. We took inspiration from it to revise our work and deliver the novel comprehensive approach [5]. In

summary, our approach creates contexts using elements of the input table, i.e. headers of columns and cells in the same row in order to disambiguate the text elements within a cell [6,7]. More specifically, the approach calculates the similarity between the representation (union of contexts) of the element in the table and the representation of the candidate entities in the reference KG. This task is repeated several times until the correct annotation is identified (annotation with highest score).

Once a table has been semantically annotated, it can be used to populate the semantic description of informative services that consume the data of the table. In the second part of the thesis, the STI is used to (semi) automatically create semantic descriptions that correlate API's properties at a semantic level [10,3,4]. To enhance interoperability we propose to enrich OpenAPI specification[1], a standard that is getting popular in the domain of services. By exploiting semantic descriptions, machines can automatically invoke services and process the results. Moreover, semantic descriptions are enabling the definition of automatic procedures for discovery and composition of services.

The Resource Description Format (RDF) output of a STI process can be difficult to exploit by human readers, hence a more human friendly representation is desirable. The third part of the thesis proposes an approach aimed at converting groups of triples into natural-language textual descriptions [2]. The conversion takes place by applying a Neural Machine Translation (NMT) supervised technique: a neural network model receives a set of RDF triples in input, and produces a textual description containing the information present in the triples. The use of a supervised approach, however, requires the construction of a training dataset composed of a set of triple-text pairs to train the model. For this reason, we propose a pipeline for the creation of a new training dataset. This pipeline uses a combination of state-of-the-art tools to make alignments between triples and text.

Another important contribution of the thesis are the tools that implement the proposed theoretical approaches:

1. STI approach - MantisTable
2. STI validation tool - STILTool
3. Semantic Description for API - AutomAPIc
4. Pipeline for RDF-text alignment - SeaLion
5. Tool that creates soccer articles automatically for RDF - GazelLex[2]

Further information and links to tools are available at zoo.disco.unimib.it.

Description of the evaluation methods used to validate the results. Related to our STI approach, for the first part of the experiments, we use two Gold Standards: T2Dv2[3] and Limaye200 [9]. In addition, we extend the experi-

[1] swagger.io/specification/

[2] The project was funded by a Digital News Innovation Fund call (newsinitiative.withgoogle.com/dnifund/) and commissioned by an italian newspaper publisher; currently the code can't be released, but the description of the project is in [2].

[3] webdatacommons.org/webtables/goldstandardV2.html

ments and run MantisTable on additional tables proposed by the challenge "Tabular Data to Knowledge Graph Matching"[4]. Thanks to the good results achieved in the challenge, the approach obtained the "Outstanding Improvement" award during the ISWC 2019 conference.

To verify the validity of the composition approach, we collected a set of APIs for the creation of a *benchmark* with characteristics that cover all possible cases. The chosen APIs comes from various domains, including public transport, films, books, music and events.

Regarding the validation of the RDF lexicalization approach, we decided to apply the techniques in a specific domain. We developed GazelLex tool (Gazette Lexicalization) [2], a prototype that covers several steps of Natural Language Generation, in order to create soccer articles automatically, using data from Knowledge Graphs. The project was partly commissioned by an Italian newspaper publisher. GazelLex, through the use of deep learning techniques, implements a NMT approach to generate articles (sentences) starting from data composed by RDF triples. To the best of our knowledge, our prototype is the first to provide an all-in-one integrated approach to Natural Language Generation (NLG) with RDF triples in the context of helping journalist in writing articles.

Significance of the work, open issues, and future directions of work. Current STI approaches perform well on table-to-KG annotations with multiple solutions. The problem of *out-of-KG annotations* is a field that has not been deeply investigated and is gaining attention in the Semantic Web community. The emerging challenges are related to the (i) identification and annotation of tabular data consisting of entity aliases already in a KG and (ii) tabular data representing entities not yet present in a KG (i.e. they represent "novel" information). A third challenge would be to automatically extend existing KGs in order to ensure continuous improvement of completeness. A very preliminary work to address these problems has been presented in [12]. In this paper a statistical method has been proposed to identify errors or missing entity and concept inside the KG. Some more recent works instead use feature-based methods [13] and embeddings [8]. The approach proposed by [13] takes advantage of Word2vec models. In particular, it uses the cosine similarity to discover topical space similarity and the Levenshtein distance to discover novel entities annotations. These approaches show greater recall with respect to table-to-kg annotations approaches but the quality of the annotations is on average low. For this reason, the goal of this research proposal is a deep investigation of the above issues to define a new STI approach with a particular focus on the management of information not yet present in a KG (out-of-Knowledge-Graph information).

To allow better conversion of tabular data into text, currently we are adopting eye-tracking techniques to better understand how a user reads and summarises the contents of a table. Preliminary results have shown that users with low familiarity with the domain of the table tend to analyse more the columns on the left, on the other hand, users with a higher level of familiarity tend to focus more on the rightmost columns as they contain more specific information.

[4] www.cs.ox.ac.uk/isg/challenges/sem-tab/

These analyses will help us to correctly select the data and annotations to be lexicalised, possibly extending them using the data present in the KG, according to the user's information needs.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284**(5), 34–43 (2001)
2. Cremaschi, M., Bianchi, F., Maurino, A., Pierotti, A.P.: Supporting journalism by combining neural language generation and knowledge graphs. In: Sixth Italian Conference on Computational Linguistics 2019. CEUR-WS.org (2019)
3. Cremaschi, M., De Paoli, F.: Toward automatic semantic api descriptions to support services composition. In: De Paoli, F., Schulte, S., Broch Johnsen, E. (eds.) Service-Oriented and Cloud Computing. pp. 159–167. Springer International Publishing, Cham (2017)
4. Cremaschi, M., De Paoli, F.: A practical approach to services composition through light semantic descriptions. In: Kritikos, K., Plebani, P., de Paoli, F. (eds.) Service-Oriented and Cloud Computing. pp. 130–145. Springer International Publishing, Cham (2018)
5. Cremaschi, M., Paoli], F.D., Rula, A., Spahiu, B.: A fully automated approach to a complete semantic table interpretation. Future Generation Computer Systems **112**, 478 – 500 (2020)
6. Cremaschi, M., Rula, A., Siano, A., De Paoli, F.: Mantistable: A tool for creating semantic annotations on tabular data. In: Hitzler, P., Kirrane, S., Hartig, O., de Boer, V., Vidal, M.E., Maleshkova, M., Schlobach, S., Hammar, K., Lasierra, N., Stadtmüller, S., Hose, K., Verborgh, R. (eds.) The Semantic Web: ESWC 2019 Satellite Events. pp. 18–23. Springer International Publishing, Cham (2019)
7. Cremaschi, M., Rula, A., Siano, A., De Paoli, F.: Semantic table interpretation using mantistable. In: The Fourteenth International Workshop on Ontology Matching 2019. CEUR-WS.org (2019)
8. Efthymiou, V., Hassanzadeh, O., Rodriguez-Muro, M., Christophides, V.: Matching web tables with knowledge base entities: From entity lookups to entity embeddings. In: d'Amato, C., Fernandez, M., Tamma, V., Lecue, F., Cudré-Mauroux, P., Sequeda, J., Lange, C., Heflin, J. (eds.) The Semantic Web – ISWC 2017. pp. 260–277. Springer International Publishing, Cham (2017)
9. Limaye, G., Sarawagi, S., Chakrabarti, S.: Annotating and searching web tables using entities, types and relationships. Proc. VLDB Endow. **3**(1-2), 1338–1347 (2010)
10. Lucky, M.N., Cremaschi, M., Lodigiani, B., Menolascina, A., De Paoli, F.: Enriching api descriptions by adding api profiles through semantic annotation. In: Proc. of the 14th ICSOC 2016. pp. 780–794. LNCS Springer (2016)
11. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web **8**(3), 489–508 (2017)
12. Paulheim, H., Bizer, C.: Improving the quality of linked data using statistical distributions. Int. J. Semant. Web Inf. Syst. **10**(2), 63–86 (2014)
13. Zhang, S., Meij, E., Balog, K., Reinanda, R.: Novel entity discovery from web tables. In: Proceedings of The Web Conference 2020. p. 1298–1308. WWW '20, Association for Computing Machinery, New York, NY, USA (2020)
14. Zhang, Z.: Effective and efficient semantic table interpretation using tableminer+. Semantic Web **8**(6), 921–957 (2017)