T2KG: Transforming Multimodal Document to Knowledge Graph

Santiago Galiano and **Rafael Muñoz,** and **Yoan Gutiérrez**
and **Andrés Montoyo** and **Jose I. Abreu**
Research Group of Language Processing and Information System. University of Alicante
sgs97@alu.ua.es, {rafael,ygutierrez,montoyo,jabreu}@dlsi.ua.es

L. Alfonso Ureña
University of Jaén
laurena@ujaen.es

Abstract

The large amount of textual information, in digital format available today, makes the knowledge extraction task unfeasible by manual means. It is therefore necessary to develop automatic tools that allow us to integrate this knowledge into a structure that is easy to use by both machines and humans. This paper presents T2KG, a framework that can incorporate the relevant information from several structured or unstructured documents into a semantic network. Structured documents are processed based on their annotation scheme. For unstructured documents, T2KG uses a set of Natural Language Processing sensors that identify relevant information to enrich the semantic network created by linking all the knowledge from different documents.

1 Introduction

Nowadays, the amount of information available on the web is available in multiple formats. Leveraging this data requires the design of software systems that can exploit the information, obtain relevant data, structure it in a specific format, and generate reports that help to evaluate this information. Software systems focused on performing all these actions are currently oriented to apply different natural language processing techniques. Many early developments were domain-dependent, so domain-specific resources, although costly in terms of time and expertise, were relatively easy to obtain. But recently, general-purpose, domain-independent systems are being developed. However, it would be difficult to imagine a system like ChatGPT incorporating a multi-domain ontology in real time. The trend in Natural Language Processing (NLP) today is text-to-text development that does not use manually curated semantic resources such as semantic networks. In other words, text-to-text oriented systems use self-generated resources without the need for external semantic resources. However, systems must take into account that the software created must be maintainable and extensible, using processes and methodologies that make all these aspects possible.

This work aims to present a framework capable of extracting knowledge from heterogeneous sources, structured such as comma-separated volumes or relational databases, or unstructured such as plain text from Wikipedia. Knowledge from different sources is integrated into an ontology. It also allows the user to query the knowledge in natural language while results are analyzed automatically to generate custom graphics or visualizations to ease its interpretation.

2 State of the Art

Knowledge representation is the process of modeling information in a way that enables effective reasoning, communication, and decision-making by computers. Given the increasing amount of digital data available, it has become more important than ever to conceive ways to represent it in a meaningful way to add knowledge to NLP systems. This knowledge has been used to improve the accuracy of tasks such as sentiment analysis, named entity recognition, and text classification (Gao et al., 2019; Peng et al., 2023). Two main tasks are involved in this process: knowledge extraction and knowledge integration from different sources. For knowledge extraction, it's necessary the development of sensors to extract pieces of relevant information (e.g. entities and relations) from unstructured documents. Knowledge integration needs to deal with linking entities, modeling uncertainty, or solving inconsistencies.

One of the challenges of multimodal representation is integrating information from different sources in a meaningful way. This requires the

Proceedings of Recent Advances in Natural Language Processing, pages 385–391
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_043

development of novel algorithms and techniques that can effectively capture the relationships between different types of data. Several approaches have been proposed for integrating text and image data, including deep neural networks for image and text (Gao et al., 2020; Zhang et al., 2020a).

In recent years, there has been significant progress in the field of knowledge representation from unstructured textual data, for example, processing scientific articles. Scientific articles contain a wealth of information, including structured data such as references and citations, as well as unstructured data such as text and figures. By representing this information in a structured way, it is possible to create a comprehensive knowledge graph that captures the relationships between different concepts and entities. One common approach is to use natural language processing techniques to extract structured data from the text of scientific articles (Zhang et al., 2020b; Dunn et al., 2022). For example, named entity recognition can be used to identify entities such as proteins, genes, and diseases in the text, while relationship extraction can be used to identify the relationships between these entities. Another approach is to use machine learning techniques to learn representations of entities and relationships in a knowledge graph directly from the text and image data (Liu et al., 2020). Also, it is possible to use image processing techniques to extract information from figures and tables (Zulkarnain et al., 2022). For example, optical character recognition (OCR) can be used to extract text from figures, while computer vision techniques can be used to identify patterns and relationships in the data.

There are different techniques for knowledge extraction:

1. Rule-based approaches involve the use of domain-specific expert-crafted rules that are designed to capture relevant information. Rule-based approaches can be effective in extracting structured information from scientific articles, but they are limited by the difficulty of designing rules that capture all the relevant knowledge or that apply to other domains (Atzmüller et al., 2008).

2. Statistical approaches use statistical models to identify and extract knowledge from scientific articles. These models are trained on large datasets of annotated texts to identify patterns corresponding to different types of knowledge. Statistical approaches can be effective in extracting knowledge from large volumes of unstructured data, but they can be limited by the quality of the training data (Momtazi and Moradiannasab, 2019).

3. Machine learning-based approaches involve using machine learning algorithms to automatically learn patterns in the data that correspond to different types of knowledge. These algorithms are typically trained on large datasets of annotated scientific articles to identify complex patterns in the data that are difficult to capture using rule-based or statistical approaches. Machine learning-based approaches can be highly effective in extracting knowledge from texts but require large amounts of high-quality training data (Tiddi and Schlobach, 2022).

Knowledge graph construction involves the creation of a structured representation. A knowledge graph consists of a set of entities representing objects or concepts and a set of relationships between them. There are different techniques for constructing knowledge graphs:

- Ontology-based approaches use pre-defined ontologies to structure knowledge extracted from scientific articles. These approaches typically involve mapping entities and relationships from the text to concepts defined in the ontology. They can be effective for building knowledge graphs consistent with domain-specific knowledge, but they can be limited by the availability and quality of the ontology (Krötzsch, 2017).

- Co-occurrence-based approaches leverage statistical techniques to identify relationships between entities. Typically they compute the frequency of entities appearing together, connecting them based on this information. These approaches can be adequate for constructing knowledge graphs that capture the co-occurrence relationships between entities, but not for more complex relationships (Heist, 2018).

- Machine learning-based approaches use large annotated corpora to learn to identify entities and relationships from the text. They can spot

complex patterns. Machine learning-based approaches can be highly effective in constructing knowledge graphs that capture complex relationships between entities, but they require large amounts of high-quality training data (Neelakantan, 2017).

Multimodal knowledge extraction and representation have promising applications in healthcare and biomedicine. By representing medical data in a structured way, it is possible to create a more comprehensive understanding of diseases and to develop more effective treatments. For example, knowledge graphs have been used to identify new drug targets for diseases (Sang et al., 2018; Gao et al., 2022).

In healthcare, knowledge representation techniques are being used to extract valuable insights from electronic health records (EHRs). EHRs contain a wealth of information, including patient demographics, diagnoses, and treatments. By applying knowledge representation techniques to EHRs, researchers can extract valuable insights into disease risk factors, treatment efficacy, and patient outcomes (Liao et al., 2010). In biomedicine, knowledge representation techniques are being used to extract knowledge from large volumes of scientific literature. The aim is to create a comprehensive biomedical knowledge graph that can be used to facilitate drug discovery, disease diagnosis and personalised medicine. Knowledge graphs constructed from the biomedical literature can thus capture complex relationships between genes, proteins and diseases, which can be used to identify potential drug targets. (Yuan, 2020).

3 T2KG Framework

We present T2GK, a framework for managing (i.e., extracting, storing and retrieving) knowledge from heterogeneous sources. Section 3.1 describes how align structured and unstructured data into an unified schema. Next, the data mining process is described in the section 3.2. Subsequently, in section 3.3 the extracted pieces of information are integrated into the Knowledge Graph. Finally, the section 3.4 presents the retrieval and visualisation of the data, as well as the evaluation of the platform.

3.1 Standard Annotation

The system works with both structured and unstructured data. The structured data follows an internal organization that can be used to label the information it contains. For example, an Excel sheet with a column called ”city” will label the rest of the elements in that column with that label or as the database name field. There are many different structured formats, the first action is to manage and standardize these representations for internal use. We use subject-verb-predicate triples to link the relevant information, according to the scheme shown in Figure 1.

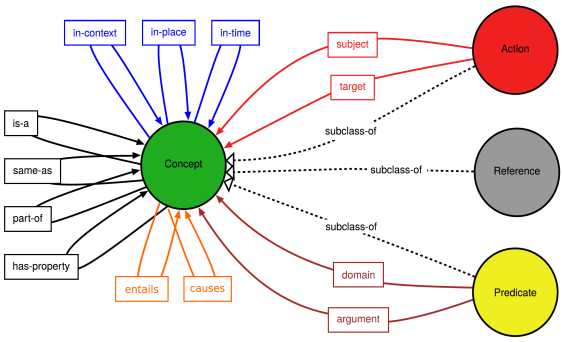

Figure 1: Conceptual schema

On the other hand, unstructured content lacks a predefined structure of concepts and relations. Hence, the stage for processing unstructured data is designed as a text-mining pipeline through which simple concepts are processed and transformed into more complex ones.

3.2 Knowledge Discovery

This stage presents a machine-learning pipeline for the automatic annotation of entities and relations in raw text. This pipeline is trained on manually annotated sentences and applied to the remaining corpus. Figure 2 shows a high-level overview of the pipeline, which comprises the following steps:

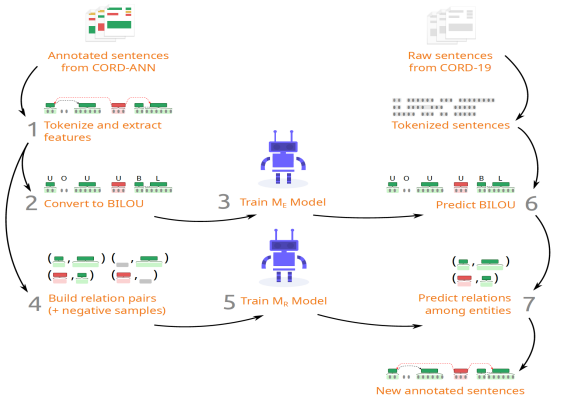

Figure 2: Illustrative representation of the text-mining pipeline used

1. Sentences are tokenized, computing syntactic and morphological features for each token (using the spaCy [1] library).

2. Training data is manually annotated for entities using BRAT [2]. Then BRAT format is converted to BILOUV encoding (i.e., Begin, Inside, Last, Out, Unit, and oVerlap) for entities.

3. An Entity Model (EM) is trained on the token features to predict the BILOUV encoding. For experiments, we use a Conditional Random Fields (CRF) [3] model.

4. Training data is manually annotated for relations. Each relation pair is converted to a set of aggregated features, and negative relation pairs are randomly sampled.

5. A Relation Model (RM) is trained using the relation features. We use a logistic regression model [4].

6. The EM is applied to unlabeled sentences.

7. The RM is executed on the pairs of entities predicted in the previous step.

For the entity model, the syntactic and morphological features include lemma, coarse and fine-grained part-of-speech, dependency labels, general purpose entity labels (e.g., PERSON, LOCATION, etc.), word shape, and several flags for specific patterns such as emails, numbers, and URLs. For the relation model, the aggregated features correspond to those from the tokens that comprise the two entities that participate in the relation, as well as the features of all the tokens in the smallest sub-tree of the dependency tree that contains both entities.

The ultimate purpose of these models is to automatically extract relevant knowledge from the unlabeled pool of sentences. Taking into account the complexity of this natural language processing task, there is always a trade-off between extracting as much knowledge as possible (i.e. maximizing recall) versus extracting knowledge as accurately as possible (i.e., maximizing precision). However, this trade-off can be explicitly controlled by measuring a degree of uncertainty in the model predictions, and only outputting the elements (i.e., entities and relations) whose uncertainty is below a given threshold. For the EM, the raw marginal probabilities provided by the CRF model are a possible measure of uncertainty. In the case of the RM, the logits provided by the logistic regression model can be used.

3.3 Knowledge Graph Integration

The knowledge graph discovered from each input document should be merged with the knowledge previously extracted by the system. Each of these knowledge graphs represents a collection of knowledge assets from a particular domain or a general domain. Some of them may overlap, containing the same knowledge facts, even if labeled as different entities or relations. Others may have contradictions or inconsistencies, either within themselves or with one another. For that reason, this stage is required to be able to undertake a matching among entities, relations, and instances in two or more graphs that are deemed similar. The result of this process is a unified knowledge graph integrating knowledge from different sources.

3.4 Case Study and Evaluation

After the new knowledge graph is created, this step provides quality evaluation metrics that assert the reliability, completeness, or soundness of the new knowledge. These metrics are based on comparing the new knowledge graphs with the existing knowledge.

This section shows the use of the T2KG system through a practical scenario that involves the processing of both unstructured and structured data sources. We use publicly available data, being the main reason for not designing this experiment with biomedical and health content.

The case study includes data about the geolocations of schools, hotels, restaurants, and bars in the province of Alicante, Spain. Also, structured population data from the Spanish Institute of Statistics (INE) was used. Unstructured data contains comments on social networks. The first step consists of obtaining the statistics data in CSV format and mapping it to a knowledge graph. The CSV file[5] contains information about Alicante's neighborhoods and their residents. The next step involves the processing of a continuous stream of Twitter messages. These are obtained through the standard Twitter query API.

[1] https://spacy.io/
[2] http://brat.nlplab.org/
[3] https://sklearn-crfsuite.readthedocs.io
[4] https://scikit-learn.org/stable/
[5] download from www.ine.es

Metric	Value	Percent
Correct matches	532	**41.95**
Correct mismatch	49	**3.86**
Matching error	297	23.42
Extraction error	405	31.94
Knowledge error	27	2.13
Context missing	4	0.18
Total errors	697	54.96

Table 1: Results of the knowledge discovery process

Processing the structured data, the location of different entities can be matched to the neighborhoods where they are located. From the unstructured data, we can know the emotions conveyed by the comments mentioning the entities.

A total of 532 instances were matched, which indicates a 41.95% of accuracy for the Twitter entity extractor. A manual review of the 1268 recognized instances was performed, to evaluate the reasons for the mistakes. All entities appearing on Twitter were searched in Google and the first result was used as ground truth. Table 1 summarizes these results.

The following figures show a set of screenshots of the application that visualizes the knowledge extracted. Figure 3 shows information related to the neighborhoods of Alicante. Figure 4 shows how the knowledge extracted from comments on social networks about hotels, restaurants, and bars in Alicante is incorporated. It is possible to see the distribution by each hotel or by stars. Finally, Figure 5 shows the graph of the knowledge retrieved by a query about the hotels in Alicante.

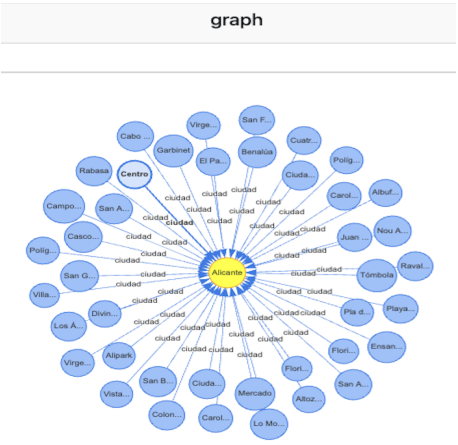

Figure 3: Alicante's neighbourhoods

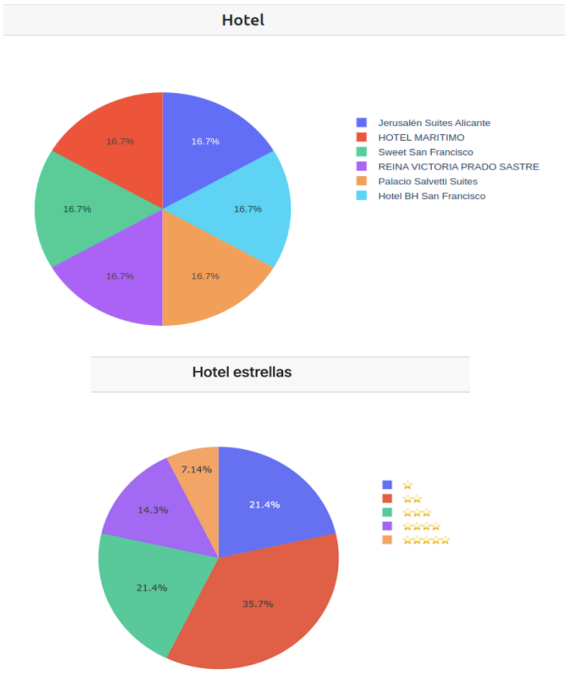

Figure 4: Alicante's Hotels identify from social network

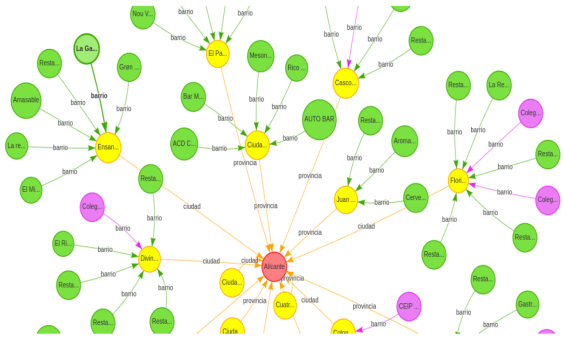

Figure 5: Knowledge Graphs from Alicante's

4 Conclusions

The manual construction of ontologies involves a human effort that sometimes cannot be tackled in domains that need to incorporate knowledge immediately. The use of knowledge graphs can alleviate this lack of resources that hinder the development of tools based on general-purpose knowledge resources. In this work, the aim was to design and implement a framework for automatic knowledge discovery from different data sources. This framework has been designed as a modular set of stages that perform specific tasks that communicate with each other. In future lines of development, we will pursue the implementation of health domain sensors (e.g. medicines, diseases, treatments, substances, etc.), and more complex mechanisms for knowledge integration (e.g., ontology merging and

mapping processes). Another line for future research is related to context mismatch and recognition. This process is necessary for accurately matching portions of unstructured text to sections of an already stored ontology. We also aim to develop a full application for the analysis of scientific articles from the biomedical and health domain.

Acknowledgments

This work has been partially supported by the Valencian Agency for Innovation through the project INNEST/2022/24, "T2Know: Platform for advanced analysis of scientific-technical texts to extract trends and knowledge through NLP techniques", partially funded by the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) through the following projects NL4DISMIS: TLHs for an Equal and Accessible Inclusive Society (CIPROM/2021/021) and partially supported by the Project MODERATES (TED2021-130145B-I00) for Spanish Government.

References

Martin Atzmüller, Peter Klügl, and Frank Puppe. 2008. Rule-based information extraction for structured data acquisition using textmarker. In *LWA*.

Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S. Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. 2022. Structured information extraction from complex scientific text with fine-tuned large language models.

Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A Survey on Deep Learning for Multimodal Data Fusion. *Neural Computation*, 32(5):829–864.

Junyang Gao, Xian Li, Yifan Ethan Xu, Bunyamin Sisman, Xin Luna Dong, and Jun Yang. 2019. Efficient knowledge graph accuracy evaluation. In *Proceedings of the VLDB Endowment, Vol. 12, No. 11*, pages 1679–1691.

Zhenxiang Gao, Pingjian Ding, and Rong Xu. 2022. Kg-predict: A knowledge graph computational framework for drug repurposing. *Journal of biomedical informatics*, 132:104133.

Nicolas Heist. 2018. Towards knowledge graph construction from entity co-occurrence. In *Proceedings of the EKAW Doctoral Consortium 2018 co-located with the 21st International Conference on Knowledge Engineering and Knowledge Management (EKAW 2018), Nancy, France, November 13, 2018*, volume 2306 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Markus Krötzsch. 2017. Ontologies for knowledge graphs? In *Proceedings of the 30th International Workshop on Description Logics (DL 2017)*, volume 1879 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Katherine P. Liao, Tianxi Cai, Vivian S. Gainer, Sergey Goryachev, Qing Zeng-Treitler, Soumya Raychaudhuri, Peter Szolovits, Susanne E. Churchill, Shawn N. Murphy, Isaac S. Kohane, Elizabeth W. Karlson, and Robert M. Plenge. 2010. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care & Research*, 62.

Zhiyuan Liu, Yixin Cao, Liangming Pan, Juanzi Li, Zhiyuan Liu, and Tat-Seng Chua. 2020. Exploring and evaluating attributes, values, and structures for entity alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6355–6364, Online. Association for Computational Linguistics.

Saeedeh Momtazi and Omid Moradiannasab. 2019. A statistical approach to knowledge discovery: Bootstrap analysis of language models for knowledge base population from unstructured text. *Scientia Iranica*, 26(Special Issue on: Socio-Cognitive Engineering):26–39.

Arvind Ramanathan Neelakantan. 2017. Knowledge representation and reasoning with deep neural networks.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. 2023. Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*.

Shengtian Sang, Zhihao Yang, Lei Wang, Xiaoxia Liu, Hongfei Lin, and Jian Wang. 2018. Sematyp: a knowledge graph based literature mining method for drug discovery. *BMC bioinformatics*, 19:1–11.

Ilaria Tiddi and Stefan Schlobach. 2022. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, 302:103627.

Jin Z. Guo H. et al Yuan, J. 2020. Constructing biomedical domain-specific knowledge graph with minimum supervision. *Knowl Inf Syst*, 62:317–336.

Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. 2020a. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493.

Xingyao Zhang, Linjun Shou, Jian Pei, Ming Gong, Lijie Wen, and Daxin Jiang. 2020b. Exploring and evaluating attributes, values, and structures for entity alignment. In *Proceedings of the 28th International Conference on Computational Linguistics, (COLING)*, pages 51–61. Association for Computational Linguistics.

Izuardo Zulkarnain, Rin Rin Nurmalasari, and Fazat Nur Azizah. 2022. Table information extraction using data augmentation on deep learning and image processing. In *2022 16th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, pages 1–6.