Semantic Structuring of E-commerce Texts: The QART Framework

André Gomes Regino[1], **Julio Cesar dos Reis**[1]

[1] Institute of Computing, University of Campinas (Unicamp) – Brazil.

andre.regino@students.ic.unicamp.br, jreis@ic.unicamp.br

Level: Doctorate

Admission: 03/2021 – **Qualification:** 03/2023 - **Defense:** 03/2025

Completed activities: Mandatory credits; qualification writing, qualification defense; framework big picture definition; Field Selection, Pre-Processing, Summarization, and part of Text Triplifying steps of the framework.

Future activities: Triple Validation step of the framework; thesis defense.

Publications: KEOD [Regino et al. 2022a], KEOD [Regino et al. 2022b], CCIS [Regino et al. 2023a], SNCS [Regino et al. 2023b], ESWC [Regino and dos Reis 2024]

Abstract. *The challenge of transforming natural language texts into structured knowledge representations is important to enhance data integration in e-commerce. We developed the QART framework to address this challenge of converting e-commerce questions and answers into RDF triples for integration into existing Knowledge Graphs (KGs). The QART framework consists of four main steps: field selection and pre-processing, text-to-text conversion, text triplifying, and RDF triple curation. These steps aim to manage the volume and complexity of e-commerce data while ensuring semantic correctness and consistency with predefined ontologies. Our evaluations demonstrated that intermediary steps, such as text summarization, produce competitive results and can improve the quality of the resulting triples.*

1. Introduction

The digital transformation of services, such as retail commerce transitioning to e-commerce, has led to challenges in managing vast amounts of data, including product visits, purchases, and abandoned carts. Enhancing user experience has been a significant research focus, employing techniques like eye-tracking, recommendation systems, and chatbots. Within this context, Knowledge Graphs (KGs) have emerged as a prominent method for representing computer-interpretable knowledge in e-commerce. Given their dynamic nature, KGs require continuous updates to remain relevant and accurate.

Populating KGs with error-free information is a challenging and time-consuming task traditionally performed by domain experts. Manual efforts can introduce inconsistencies [Wang et al. 2021], making using KGs effectively for e-commerce applications difficult. Therefore, automated approaches for populating KGs with reliable and consistent information are essential. Text-to-triple generation frameworks offer a promising solution by automatically extracting structured knowledge from unstructured

text, allowing domain experts to focus on verifying the accuracy of generated triples [Akter and Rahman 2019]. Despite existing methods for knowledge base completion and ontology learning, the challenge of transforming natural language text from e-commerce platforms into legitimate RDF triples persists.

The primary objective is to develop a methodology capable of inserting relevant information into e-commerce KGs, such as product compatibility, availability, and specifications. Specific objectives include identifying relevant parts of user texts for triple generation, summarizing user input texts into concise and unambiguous content, generating RDF triples based on summarized e-commerce sentences, and enabling the KG to answer user-generated questions using stored information. The overall aim is to create a scalable and efficient framework that reduces manual effort in data integration and enhances the accuracy and completeness of RDF triples, thereby improving e-commerce systems' functionality.

This research is relevant to the database community as it addresses the challenge of managing and integrating vast amounts of structured and unstructured data in e-commerce. By automating the generation of RDF triples from natural language texts, the proposed methodology enhances the accuracy and completeness of KGs, which are important for data integrity and usability. This advancement directly supports the core interests of the database community, such as efficient data processing and the development of innovative data management techniques.

This article is organized as follows: Section 2 presents related work. Section 3 describes our method to generate RDF triples. Section 4 shows evaluations and initial results. Section 5 draws conclusion remarks.

2. Related Work

The challenge of transforming natural language texts into structured knowledge representations has led to various approaches leveraging different methodologies. FRED [Gangemi et al. 2017] uses a combination of NLP tasks, such as Named Entity Recognition (NER) and Entity Linking (EL), to transform multilingual texts into large graphs composed of OWL and RDF specifications. Lodifier [Augenstein et al. 2012] was one of the first tools to convert NL text into triples by mapping relevant entities to DBpedia URIs and using statistical parsing and semantic construction to generate RDF graphs.

Seq2RDF [Liu et al. 2018] offers a machine-learning model to generate RDF triples using an encoder-decoder architecture trained on DBpedia. In contrast, Martinez-Rodriguez et al. [Martinez-Rodriguez et al. 2019] proposed a methodology involving feature extraction, entity extraction, and relation extraction steps. Their approach is similar to FRED but faces challenges in handling grammatical errors and is limited to named entities in the object. Rossanez and dos Reis [Rossanez and dos Reis 2019] developed a semi-automatic tool to build KGs from texts of Alzheimer's disease, using Semantic Role Labeling (SRE) for triple extraction and mapping concepts to a public domain ontology.

Our QART framework [Regino et al. 2023a] distinguishes itself by focusing on the e-commerce domain, converting questions and answers and other text sources into RDF triples. Unlike other methods, QART combines templates and text generation models to summarize texts and generate triples rather than solely relying on SRE. This novel

approach ensures semantic correctness and consistency with predefined ontologies, effectively managing the complexity of e-commerce data. One of the contributions of this doctorate research is a systematic literature review that examines how unstructured texts are transformed into RDF triples and integrated into existing KGs. Based on this review and to the best of our knowledge, QART is the first framework to transform natural language texts into RDF triples within a Q&A e-commerce context.

Table 1. Comparison of related methods/characteristics for transforming natural language texts into RDF triples. The lines represent scientific articles. The columns are: Named Entity Recognition/Entity Linking (NER/EL); Statistical Parsing (SP); Semantic Role Labeling (SRE); Machine Learning models (ML); Knowledge Graph Enhancement (KGE); Domain-specific (DS).

Article	NER/EL	SP	SRE	ML	KGE	DS
FRED [Gangemi et al. 2017]	X				X	
Lodifier [Augenstein et al. 2012]	X	X				
Seq2RDF [Liu et al. 2018]				X		
[Martinez-Rodriguez et al. 2019]	X					
[Rossanez and dos Reis 2019]			X			X
QART Framework [Regino et al. 2023a]	X			X	X	X

3. The QART Framework

In our doctorate research, we aim to develop the QART framework. The QART framework is designed to transform natural language texts into RDF triples, specifically within e-commerce. The framework begins with e-commerce questions and answers, converting them into RDF triples integrated into existing KGs. The methodology is organized into four primary steps, represented by the boxes in Figure 1: field selection and preprocessing, text-to-text conversion, and text triplifying, with an additional step of RDF triple curation currently under development.

The QART framework is a novel contribution from our work since it emphasizes the importance of intermediary processing stages, often overlooked, to enhance the quality of the final text-to-triple transformation output.

Step A: Field Selection and Pre-Processing. In this initial step, the framework starts with a comprehensive dataset containing various customer actions within an e-commerce environment, such as purchases, product evaluations, and questions and answers. The ontology maintainer selects relevant fields from this dataset to construct triples. The chosen fields undergo pre-processing to remove noisy data, such as stopwords, abbreviations, and punctuation, which do not contribute to the triple construction. The framework identifies entities and intents from these pre-processed fields. Entities relate to product specifications, while intents refer to actions described in sentences, such as purchase intent or product availability. These elements are essential for generating meaningful triples in subsequent steps.

Step B: Text-to-Text Conversion. Once the dataset is refined, the framework summarizes the most relevant fields into concise and factual texts. This process involves transforming selected fields into a single condensed field and generating a new dataset with summarized texts. The text summarization leverages templates and pre-trained machine learning models, which help produce clear and structured summaries that retain

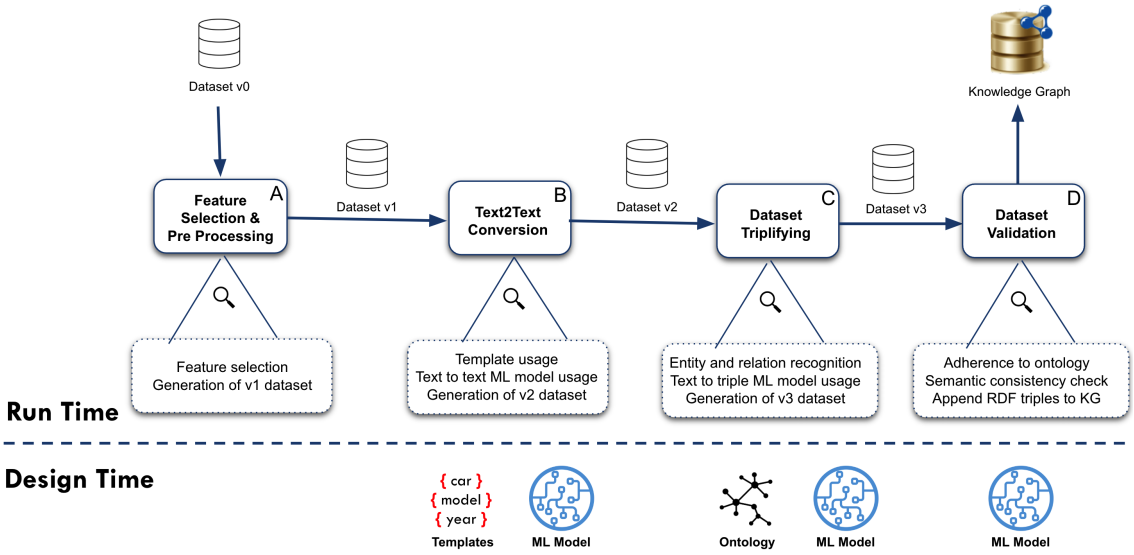

Figure 1. Method implemented by the QART framework to transform natural language texts in RDF triples and append them to an existing KG. The boxes represent the steps, from A to D. The icons represent the inputs and outputs used and generated among the steps.

the semantic essence of the original questions and answers. This transformation aims to facilitate the generation of RDF triples by providing a more straightforward and less ambiguous textual input for the next step.

Step C: Text Triplifying. This is the core step of the QART. The summarized texts from Step B are converted into RDF triples. For instance, the summarized text can combine a customer's question and a human attendant's answer from an e-commerce.

The QART identifies characteristics of the texts used as input – intents, sources, and domains – to fill the LLM prompt with critical information, facilitating the generation of RDF triples. In Step C.1, the LLM identified texts based on these characteristics, using a carefully constructed prompt to categorize the intents, domains, and sources present in the texts. This foundational categorization enabled us to filter relevant text parts in Step C.2, aligning with the characteristics crucial to the existing KG.

Further, in Step C.3, we generated RDF triples by guiding the LLM with a list of important ontology classes and properties, ensuring the produced triples were coherent and meaningful. Step C.3 involves extracting relevant information from an existing ontology, including classes, properties, and triples, to define how the resulting RDF triples should be structured. The summarized text and extracted ontology elements are then used to construct a prompt for the language model, which generates RDF triples based on this input. The language model uses the prompt to generate one or more RDF triples. Any unforeseen elements in the generated triples are identified and removed to ensure consistency with the existing ontology.

Step D: RDF Triple Curation. Before the newly generated RDF triples are added to the KG, they undergo a curation process to ensure their semantic correctness and adherence to the ontology's structure. This step is important for maintaining the integrity and reliability of the KG, though it is still under development. It comprises four veri-

fications: syntactic consistency, URI uniqueness, adherence to classes and properties of the ontology, and semantic consistency. By incorporating this final validation phase, the QART framework aims to produce high-quality triples that enrich the existing knowledge base.

4. Summary of Results

The framework hypothesizes that intermediary steps, such as text summarization, between the input NL texts and the RDF triples can enhance the quality of the resulting triples. To validate this hypothesis, we conducted evaluations demonstrating the effectiveness of these intermediary steps within the QART process. In Subsections 4.1 and 4.2, we evaluate this hypothesis and show the results of using each step of QART.

4.1. Intent/Entity Identification and Templates/Language Models to Text Summarization

The first experiment assessed the QART framework's ability to transform e-commerce questions and answers into summarized texts by identifying intents and entities. Using a dataset of 200 real compatibility questions and answers from a Brazilian e-commerce platform, the framework was evaluated against a gold standard dataset. The results demonstrated that QART achieved 89.5% accuracy in intent classification and 86% accuracy in entity identification. Specifically, the framework correctly identified 179 out of 200 compatibility intents (c.f Table 2) and 516 out of 600 entities (c.f Table 3), with the highest accuracy for brand identification at 91.5% and the lowest for year identification at 82.5%.

Table 2. Results of comparing the intents in the gold standard and the predicted values from QART, as assessed by a set of evaluators, for a Brazilian e-commerce platform. The comparison is presented in three categories: fits, does not fit, and no compatibility.

		Gold Standard			
		fits	**does not fit**	**no compatibility**	**Total**
QART	**fits**	135	7	5	147
	does not fit	4	44	5	53
	Total	139	51	10	200

Table 3. Results of the comparison between 3 entities (brand, model and year) in gold standard and predicted values from QART. [Regino et al. 2022b]

	Brand	**Model**	**Year**	**Total**
Total	183	168	165	516
%	91,5	84,0	82,5	86,0

The second experiment evaluated four machine learning models—T5-small, PTT5, GPT-Neo, and Bloom—for their text summarization capabilities in e-commerce. Using the EcommercePR dataset, which includes 800 rows of product-related questions, answers, and their summarized versions, the models were tested on their ability to produce accurate and concise summaries. Bloom consistently achieved the highest scores

across quantitative metrics (c.f Table 4), indicating superior performance, while GPT-Neo performed well. T5 showed the lowest performance, struggling with the specific task. Qualitative analysis highlighted Bloom's effectiveness in assertivity, conciseness, and coherence, making it the most effective model overall for summarizing e-commerce texts.

Table 4. Evaluation metrics for the four evaluated models. The values represent the median of the f-measure. Values closer to 1 represent more similar summaries produced by each model.

Model	Rouge-1	Rouge-2	Rouge-L	BLEU
T5	0.283	0.145	0.265	0.05
PTT5	0.683	0.656	0.682	0.38
GPT Neo	0.784	0.672	**0.797**	0.61
Bloom	**0.802**	**0.722**	0.777	**0.65**

4.2. LLMs to RDF Triple Generation

This qualitative experimental evaluation focused on transforming e-commerce text into RDF triples while maintaining identified characteristics such as intents, text sources, and domains. Using two real-world cases, we executed prompts via the HuggingFace API with the Bloom LLM, selected for its open-source nature and multilingual capabilities. The first assessment involved a customer query about VGA cable compatibility and store operating hours. We identified the text's characteristics, filtered relevant phrases, generated RDF triples, and validated them against an ontology. This process ensured only allowed characteristics were included, and the triples adhered to the ontology.

In the second assessment, we analyzed a product description of a protein bar kit from a fitness food store. Following a similar approach, we identified the text's domain and intents, filtered the text to retain pertinent phrases, and generated RDF triples based on these elements. Validation ensured the removal of unlisted properties and verified the triples' order and consistency. This evaluation demonstrated the method's capability to accurately represent product descriptions in RDF triples, ready for integration into a KG.

5. Conclusion

Our research addresses the challenge of transforming natural language texts into structured knowledge representations for improved data integration in e-commerce. The developed QART framework converts e-commerce questions and answers into RDF triples, facilitating integration into (KGs). Our approach manages data complexity through field selection, text summarization, RDF triple generation, and curation and ensures semantic accuracy with predefined ontologies. Evaluations confirm the effectiveness of intermediary steps. Future work will refine RDF triple curation and expand the framework's applicability across diverse e-commerce domains.

Acknowledgments

This study was financed by the National Council for Scientific and Technological Development - Brazil (CNPq) process number 140213/2021-0. This research was partially funded by the São Paulo Research Foundation (FAPESP) (grants #2022/13694-0, #2022/15816-5 and #2024/07716-6).

References

Akter, Y. A. and Rahman, M. A. (2019). Extracting rdf triples from raw text. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–4. IEEE.

Augenstein, I., Padó, S., and Rudolph, S. (2012). Lodifier: Generating linked data from unstructured text. In *Extended Semantic Web Conference*, pages 210–224. Springer.

Gangemi, A., Presutti, V., Reforgiato Recupero, D., Nuzzolese, A. G., Draicchio, F., and Mongiovì, M. (2017). Semantic web machine reading with fred. *Semantic Web*, 8(6):873–893.

Liu, Y., Zhang, T., Liang, Z., Ji, H., and McGuinness, D. L. (2018). Seq2rdf: An end-to-end application for deriving triples from natural language text. In *CEUR Workshop Proceedings*, volume 2180. CEUR-WS.

Martinez-Rodriguez, J. L., Lopez-Arevalo, I., Rios-Alvarado, A. B., Hernandez, J., and Aldana-Bobadilla, E. (2019). Extraction of rdf statements from text. In *Iberoamerican Knowledge Graphs and Semantic Web Conference*, pages 87–101. Springer.

Regino, A. G., Caus, R. O., Hochgreb, V., and dos Reis, J. C. (2022a). Knowledge graph-based product recommendations on e-commerce platforms. In Aveiro, D., Dietz, J. L. G., and Filipe, J., editors, *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2022, Volume 2: KEOD, Valletta, Malta, October 24-26, 2022*, pages 32–42. SCITEPRESS.

Regino, A. G., Caus, R. O., Hochgreb, V., and dos Reis, J. C. (2022b). QART: A framework to transform natural language questions and answers into RDF triples. In Aveiro, D., Dietz, J. L. G., and Filipe, J., editors, *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2022, Volume 2: KEOD, Valletta, Malta, October 24-26, 2022*, pages 55–65. SCITEPRESS.

Regino, A. G., Caus, R. O., Hochgreb, V., and dos Reis, J. C. (2023a). From natural language texts to rdf triples: A novel approach to generating e-commerce knowledge graphs. In Coenen, F., Fred, A., Aveiro, D., Dietz, J., Bernardino, J., Masciari, E., and Filipe, J., editors, *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 149–174. Communications in Computer and Information Science.

Regino, A. G., Caus, R. O., Hochgreb, V., and Reis, J. C. d. (2023b). Leveraging knowledge graphs for e-commerce product recommendations. *SN Computer Science*, 4(5):689.

Regino, A. G. and dos Reis, J. C. (2024). Generating e-commerce related knowledge graph from text: Open challenges and early results using llms. In *TEXT2KG @ ESWC (accepted for publication)*.

Rossanez, A. and dos Reis, J. C. (2019). Generating knowledge graphs from scientific literature of degenerative diseases. In *SEPDA@ ISWC*, pages 12–23.

Wang, X., Chen, L., Ban, T., Usman, M., Guan, Y., Liu, S., Wu, T., and Chen, H. (2021). Knowledge graph quality control: A survey. *Fundamental Research*, 1(5):607–626.