Generating E-commerce Related Knowledge Graph from Text: Open Challenges and Early Results using LLMs

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

[1]Institute of Computing, State University of Campinas, Campinas, Brazil

Abstract

E-commerce systems need to use and manage vast amounts of unstructured textual data. This poses significant challenges for knowledge representation, information retrieval, and recommendation tasks. This study investigates the generation of E-commerce-related Knowledge Graphs (KGs) from text. In particular, we explore using Large Language Models (LLMs). Our approach integrates ontology with text-based examples from existing KGs via prompts to create structured RDF triples. We outline a four-step method encompassing text classification, extracting relevant characteristics, generating RDF triples, and assessing the generated triples. Each step leverages LLM instructions to process unstructured text. We discuss the insights, challenges, and potential future directions, highlighting the significance of integrating ontology elements with unstructured text for generating semantically enriched KGs. Through case experimentations, we demonstrate the effectiveness and applicability of our solution in the E-commerce domain.

Keywords

Knowledge Graphs, Large Language Models, KG enhancement, text-to-triple

1. Introduction

The e-commerce has undergone a paradigm change, with an increase in the volume of textual data. The textual content in e-commerce environments presents unique challenges for knowledge extraction, necessitating advanced text analysis techniques tailored to the domain.

In this dynamic environment, the significance of Knowledge Graphs (KGs) has become increasingly pronounced. KGs play a key role in e-commerce [1] by enhancing information retrieval and knowledge representation and unlocking the potential for more advanced applications. By using KGs, the organization of e-commerce information reaches new levels of efficiency, fostering innovations such as personalized recommendations [2] [3] and semantic search [4].

The rapid expansion of textual data within the e-commerce domain presents a scaling challenge. As the amount of unstructured information proliferates, a need arises for effective methods to extract structured knowledge from this vast textual data. Traditional approaches

3rd International Workshop on Knowledge Graph Generation from Text, May 26–30, 2024, Hersonissos, Greece

[*]Corresponding author.

[†]These authors contributed equally.

andre.regino@students.ic.unicamp.br (A. G. Regino); jreis@ic.unicamp.br (J. C. d. Reis)

0000-0001-9814-1482 (A. G. Regino); 0000-0002-9545-2098 (J. C. d. Reis)

struggle to keep pace with the data explosion, necessitating novel solutions to harness the valuable insights present in e-commerce texts.

This study explores and proposes solutions for generating e-commerce-related KGs from textual data. In particular, we aim to leverage the capabilities of Large Language Models (LLMs) to transform unstructured e-commerce texts into RDF triples, laying the foundation for a more structured and actionable representation of e-commerce knowledge.

From a practical scenario in the context of a Latin American AI company that supplies e-commerce systems, our investigation maps and describes several open research challenges. The specific challenges addressed in our study consider specifically the following aspects:

1. **Multiple Intents**: in the diverse landscape of e-commerce, multiple intents within textual data, such as compatibility between products, product descriptions, and shipping details, poses a challenge;
2. **Multiple Text Sources**: e-commerce information is dispersed across various text sources, including product descriptions and user-generated question-and-answer sections. The integration of these diverse sources into a coherent KG is a non-trivial task;
3. **Multiple Domains**: The e-commerce domain spans various sectors, from automobiles to household goods. Integrating information from these disparate domains requires careful consideration and a nuanced approach in KG construction.

This article presents contributions concerning the e-commerce knowledge representation and text-to-triple generation. First, we **identify and formalize key major challenges inherent in transforming e-commerce texts into RDF triples**. This sheds light on the complexities of knowledge extraction from unstructured data and offers insights that can transcend the e-commerce domain, potentially inspiring advancements in text-to-triple generation across other domains.

Second, we propose a **systematic pipeline comprising four tailored steps for triple generation**, addressing the identified challenges. Our pipeline elucidates the interplay among text processing, ontology integration, and semantic validation, offering a comprehensive framework for navigating the complexities of knowledge representation in e-commerce. By integrating these steps, we elucidate the challenges that arise from their combination and provide a roadmap for overcoming them. This structured approach enhances the reproducibility and scalability of our solution, facilitating its adoption and adaptation in diverse e-commerce settings.

Third, we present a **novel solution to the challenges posed by e-commerce text-to-triple generation, leveraging LLMs [5] and tailored prompts** to extract meaningful knowledge from real-world e-commerce texts. This approach demonstrates the early effectiveness of LLMs in processing unstructured data and showcases the utility of prompts in guiding the model toward relevant information extraction. We explored real-world e-commerce cases in our experimental evaluation to showcase the practical applicability of our solution.

The remaining of this article is organized as follows: Section 2 presents related work on RDF triple generation from text; Section 3 describes the challenges and the text-to-triple generation pipeline; Section 4 presents our solution relying on LLMs; and Section 5 shows the application of our solution in two real-world cases. Section 6 discusses our findings, open issues, and future work; Section 7 draws conclusion remarks.

2. Related Work

We present literature studies using Deep Learning techniques to generate RDF triples from unstructured texts.

The AlimeKG framework [6] is the one focused on the e-commerce domain. Li *et al.* [6], introduced AlimeKG, a framework for KG construction in the e-commerce domain. AlimeKG integrates NLP components (named entity recognition (NER) and relation extraction (RE)), facilitating a semi-automated process for knowledge acquisition and validation. Their study focuses on pre-sales conversation scenarios. Our method also focuses on e-commerce but is not restricted to pre-sales conversation. Challenges include the substantial requirement for human annotations, alongside potential issues of reliability due to reliance on external knowledge sources to generate and populate the KG. Open challenges involved expanding the coverage of AlimeKG within the Alibaba e-commerce platform and exploring its applicability to other domains beyond e-commerce.

The work by Xu *et al.* [7] presented a method for constructing a dynamic KG in the Traditional Chinese Medicine (TCM) domain. Their work contains BERT-based models and bootstrapping methods for resampling of features/texts. The dynamic proposed approach relates to the static nature of existing KG in the TCM domain and the idea of providing continuous growth through user interactions. Challenges include merging KGs and handling term ambiguity. In addition, the methodology's applicability is limited to English-based knowledge and the TCM domain.

Fei *et al.* [8] proposed a Perspective-Transfer Network (PTN) for constructing KG using few-shot Relational Triple Extraction (RTE). PTN offered a multi-perspective approach to constructing a KG, incorporating three perspectives: Relation, Entity, and Triple. The solution's strengths are its ability to handle a few training examples and its fully automatic nature, although it may require substantial hardware resources and can be sensitive to input data quality.

Li *et al.* [9] proposed STonKGs, a Transformer network for constructing KGs in the biomedical domain. STonKGs demonstrated adaptability to various domains and superior performance over reference models, with evaluations indicating strengths in classifying contexts and relationships. However, challenges include longer training times and more extensive pre-training data.

In addition, we identified alternative pipelines - other than LLM based - proposed in other domains. Scicero *et al.* [10] explored Transformer models to automatically extract entities from scientific texts and generate a KG. Tested on 6.7 million Computer Science articles, their work produced a KG with 10 million entities and demonstrated strong efficiency against a gold standard. In the scope of tourism, Chessa *et al.* [11] introduced a method for semi-automatically generating a Tourism Knowledge Graph (TKG) by leveraging the Tourism Analytics Ontology (TAO). Using data from Booking.com, Airbnb.com, DBpedia, and GeoNames, they produced more than 10 million triples detailing lodging facilities and reviews in Sardinia and London. This process was supported by an open-source pipeline and software. In the scope of social media, we found the Claims KG [12], a KG designed to store fact-checked claims and associated metadata. This fact-checking KG can be used to verify fake news in political elections. Using a semi-automated pipeline, data is acquired from popular fact-checking websites and annotated with entities from DBpedia. The pipeline then transforms this data into RDF format.

Unlike other approaches that address various domains or employ different methods such as bootstrapping, our method is tailored explicitly for the e-commerce domain, offering a four-step framework for KG construction and validation (cf. Section 4). By leveraging LLMs, our solution extracts, structures, and assesses knowledge from unstructured text, ensuring the accuracy and reliability of the resulting RDF triples. This focused approach enables targeted solutions for KG enhancement in e-commerce contexts, addressing specific challenges and requirements unique to this domain. To the best of our knowledge, there is no evidence of a fully LLM-based approach to enhance existing KGs with knowledge from e-commerce texts.

3. Mapping Open Challenges

This section presents challenges in populating existing e-commerce KGs from unstructured texts. A KG is a connected directed graph formed by V vertices and their directed E edges. The KG comprises RDF triples, a statement in the form of a subject-predicate-object. Tr_i is the set of triples (s, p, o) where s is a subject entity, p is a predicate (property), and o is an object entity.

The textual content analysis in e-commerce involves various characteristics, and we highlight three important ones for extracting knowledge: **intent**, **domain**, and **source**. The rationale for choosing them is grounded in their roles in understanding and extracting actionable knowledge from textual content. Intents provide insights into the goals and motivations of the text; text sources present an understanding of the origin and nature of the information; and the domain clarifies the thematic framework within which the information is situated. By focusing on these three characteristics, we address important dimensions of e-commerce text analysis, enabling more accurate, context-aware knowledge extraction and enhancing the effectiveness of e-commerce applications. To the best of our knowledge, these dimensions have been treated separately in the literature, emphasizing their individual significance and the need for specialized approaches to effectively incorporate them into KGs.

Figure 1 presents our four-step method for creating triples from texts and integrating them into an existing KG, based on the three challenges. The input for the method is a set of documents D. Section 4 describes how we implement the method using LLMs. The method's output is an enhanced version of the KG, with new triples and knowledge extracted from D. The challenges associated with the identified characteristics (intents, domains, or text sources) are reflected in each step. Step A involves identifying these characteristics in the text (*e.g.*, text related to Sports product, found in the product description), followed by Step B, where only relevant characteristics are extracted. Step C generates triples based on these characteristics (*e.g.*, only RDF triples related to Sports that describe a product), and Step D incorporates the validated triples into an existing KG.

For a comprehensive understanding of each challenge, we explore the details, formalization, and examples in Subsection 3.1, Subsection 3.2, and Subsection 3.3, revealing the challenges of multiple intents, multiple text sources, and multiple domains, respectively. Table 1 summarizes the challenges by each step.

3.1. Intents

Understanding the diverse purposes of e-commerce text is required in the KG generation. An "intent" refers to the underlying purpose or goal expressed in a given text [13]. An "intent" within the realm of e-commerce text can be formally defined as the underlying semantic purpose

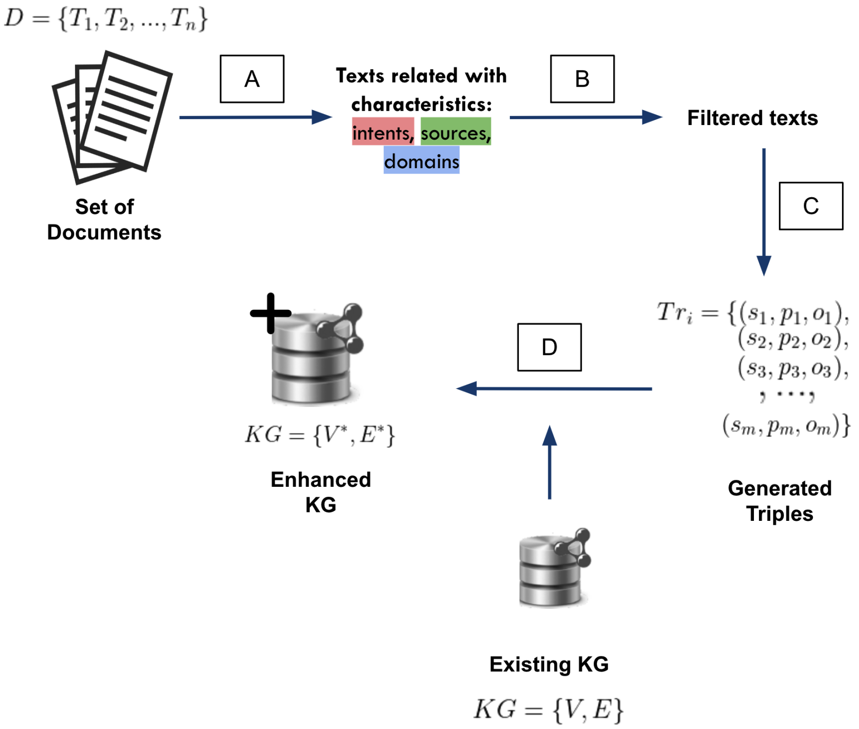

$D = \{T_1, T_2, ..., T_n\}$
A
Texts related with characteristics: intents, sources, domains
B
Filtered texts
Set of Documents
C
$Tr_i = \{(s_1, p_1, o_1), (s_2, p_2, o_2), (s_3, p_3, o_3), ..., (s_m, p_m, o_m)\}$
D
$KG = \{V^*, E^*\}$
Enhanced KG
Generated Triples
Existing KG
$KG = \{V, E\}$

Figure 1: Transforming a set of documents D in a set of RDF triples TR_i, that are appended in an existing KG. The letters inside the boxes are the steps of the method. **Step A** represents the identification of characteristics in the texts; **Step B** represents the extraction of both texts and characteristics; **Step C** stands for the generation of triples; **Step D** represents the evaluation of the generated triples and the addition of them to an existing KG.

Table 1
Summary of challenges and tasks across pipeline steps

Step	Intents	Text Sources	Domains
A	Identify intents using NLP techniques	Identify text source type	Identify various domains
B	Extract significant intents	Extract relevant information	Extract relevant domains
C	Generate RDF triples based on intents	Generate RDF triples based on source	Generate RDF triples from domain
D	Validate generated triples against intents	No specific use found in validation	Validate generated triples from a domain

or objective conveyed within a given textual segment. Mathematically, an intent I in a text T can be represented as a function $I = f(T)$.

Here, f is a mapping function that encapsulates the semantic meaning or purpose embedded in the text.

The challenge of intent identification involves the precise determination and classification of diverse intents present in e-commerce text. Given a corpus of text data D, containing individual instances T_i with varying intents I_i, the challenge can be expressed as:

$$\begin{gathered}\text{given } D = \{T_1, T_2, ..., T_n\} \\ \text{determine } I_i \text{ for } T_i\end{gathered} \tag{1}$$

In e-commerce, intents may range from establishing product compatibility and elucidating product features to specifying shipping details. Identifying and categorizing these varied intents is fundamental to constructing a meaningful KG. Consider the phrase "Compatibility with iOS and Android devices"; the intent is to convey product compatibility information. However, discerning such intents becomes intricate within the broader E-commerce textual landscape. As stated in Equation 1, one text may have one or more intents. These intents can or can not be interesting to be represented in a KG, depending on existing knowledge present in the KG and the decision of the ontology and KG maintainer.

The process of generating RDF triples from textual data (based on the intent of the texts) and, subsequently, integrating them into an existing KG involves a multi-faceted approach. This method is orchestrated through a sequence of steps, adapted from Figure 1, each presenting its own set of challenges:

- **Step A: Identification of Intents.** The challenge is two-fold. First, natural language is inherently context-dependent, requiring NLP techniques to discern intents accurately. Second, disambiguating between multiple intents within a single text snippet poses a challenge, necessitating context-aware methodologies to ensure precise intent extraction;
- **Step B: Extraction of Relevant Intents.** Once the intents are identified, the next step entails extracting them from the contextual nuances of the textual data. This step centers around identifying intents deemed significant for representation within the KG. This involves a filtration process, possibly assessing the temporal stability of intents. Intents subject to frequent fluctuations (*e.g.*, product price, product availability) may be less suitable for representation within the KG. The challenge here lies in establishing a robust criterion for intent relevance, one that considers both the dynamism of the E-commerce domain, the enduring nature of certain intents, and the necessity of the KG maintainer in keeping the knowledge related to the intent in the KG;
- **Step C: Generation of RDF Triples based on the Intents.** With the intents listed, the subsequent challenge lies in systematically generating RDF triples aligned with the identified intents. This involves transforming the extracted intent information into a structured format compatible with RDF. Ensuring semantic coherence and adherence to KG schema standards is relevant, demanding attention to detail during the triple generation process;
- **Step D: Validation of Generated Triples.** The final step involves validating the generated RDF triples against the initially identified intents. This step serves as a quality control mechanism, ensuring the transformed textual information is accurately represented within the KG. The challenge is establishing validation criteria that effectively verify the alignment between generated triples and the intended semantic meaning derived from the textual context.

3.2. Text Sources

A “text source” in e-commerce refers to distinct channels or sections within textual data that contribute diverse facets of information. The challenge of multiple text sources lies in accurately identifying and categorizing disparate channels within e-commerce textual data. Given a document D containing varied sources S_i, the challenge can be expressed as:

$$\begin{aligned} \text{given } D = \{S_1, S_2, ..., S_n\} \\ \text{determine } S_i \text{ for } D \end{aligned} \tag{2}$$

The complexity arises from the diverse information housed in different text sources. Each source contributes unique perspectives, requiring precise identification to facilitate the integration of comprehensive insights into the KG. Addressing this challenge involves developing methods to discern and classify distinct text sources within the e-commerce data.

Consider an E-commerce product webpage comprising product descriptions, customer reviews, and a question-answer section. In this scenario, the text sources would be classified as:

S = {Product Description, Customer Reviews, Q&A Section}

Some examples of text sources that can be represented in a KG:

- **Product Data:** product descriptions, reviews, specifications, ratings, user manuals, product comparisons, recommendations;
- **Store Data:** store policies, order e-mails, FAQs, newsletters, social media comments;
- **Customer Data:** customer feedbacks, surveys, chat logs, complaints;
- **Sale Data:** promotional texts, price information, holiday sale.

The method of generating RDF triples based on the text source consists of several steps, each posing its own set of challenges:

- **Step A: Identification of Text Source Type.** The initial step focuses on discerning the type or source of the text. Text sources vary widely, encompassing product descriptions, customer reviews, and question-answer sections. The challenge lies in precisely classifying the type of source, as accurate identification forms the foundation for subsequent information extraction;
- **Step B: Extraction of Relevant Information from the Text Source.** Following identifying the text source type, the next challenge is pinpointing the specific information within that source that merits inclusion in the Knowledge Graph. For example, extracting details about the store’s opening and closing hours may be irrelevant if the source is a product recommendation. The challenge is to establish a robust criterion for relevance, ensuring that only pertinent information is extracted for subsequent triple generation;
- **Step C: Generation of RDF Triples Based on Identified Information.** Once relevant information is identified, the subsequent step involves the systematic generation of RDF triples aligned with the specified type and content of the text source. The source serves as additional information to the intents and domain, clarifying how the input text should be analyzed and impacting the resulting triples.

- **Step D: Validation of Generated Triples from a Specific Text Source.** The final step refers to the validation of the generated RDF triples. Our proposed method did not find a particular use of the text source in the validation process.

3.3. Domains

In e-commerce, a "domain" pertains to distinct categories or sectors of products, each characterized by specific attributes or features. A domain Do can be represented as a set $Do = \{P_1, P_2, ..., P_n\}$. Here, P_i represents individual products within a given category, collectively forming the domain.

The challenge of multiple domains lies in accurately identifying and categorizing diverse product categories within the E-commerce textual data. Given a set of products P_i within a document, the challenge can be expressed as: given P_i, determine Do.

The complexity arises due to the varied nature of products and the diverse terminology employed across different domains. Figure 1 presents the method to generate the RDF triples based on the domains presented in the text, which poses some challenges as follows:

- **Step A: Identification of Domains.** The initial step involves identifying the domains encapsulated within the e-commerce textual data. In this context, domains represent specific categories or sectors of products (, Food, Pet, Sports). The challenge is accurately classifying the text based on the diverse array of products it covers;
- **Step B: Extraction of Relevant Domains.** Following domain identification, the subsequent challenge is to filter and discern which domains are relevant for the stakeholders (people involved with the e-commerce KG, like the seller, the store owner, and the KG maintainer). Commercial considerations such as the volume of products, customer inquiries, and sales play a pivotal role in determining the commercial viability of a domain. The challenge here is to establish effective criteria for relevance, ensuring that resources are strategically allocated to domains with commercial significance;
- **Step C: Generation of RDF Triples from the Domain.** With relevant domains identified and filtered, the subsequent step involves systematically generating RDF triples aligned with the characteristics of the specified domain. This process demands the transformation of extracted information into a structured format compatible with RDF;
- **Step D: Validation of Generated Triples from a Specific Domain.** The final stage centers around assessing the generated RDF triples specific to the identified domain. This validation step serves as a crucial quality control mechanism, ensuring the accuracy and relevance of the information incorporated into the KG from that particular domain. The challenge is establishing rigorous validation criteria that effectively verify the alignment between the generated triples and the intended semantic meaning derived from the identified domain.

4. LLM-based solutions for Overcoming the Challenges

This section shows how LLMs serve as a tool to solve the challenges inherent in generating RDF triples from e-commerce texts and appending them to a KG. Leveraging our four-step method outlined in Figure 1 and guided by the insights from the three characteristics found in Section 3, we navigate the complexities of textual data to construct a coherent and enriched KG[1].

4.1. Step A: Identification of Texts with Characteristics

We utilize LLMs to identify texts enriched with characteristics such as intents, domains, and sources. We achieve this by constructing a prompt, denoted as **Prompt 1**, instructing the LLM to categorize intents, domains, and sources of e-commerce texts. This prompt comprises three key components:

1. **Task Description:** We specify the task, instructing the LLM to categorize the intents, domains, and sources of e-commerce texts;
2. **Few-shot Examples:** We provide a series of examples, each comprising an input text and its corresponding output, which includes the identified intents, domains, and sources. Notably, we do not specify specific intents or domains but aim to capture all potential intents and domains present in the text, along with the single source;
3. **Input:** The texts.

This step receives the set of documents and outputs a set of parts of the texts and their corresponding domain, source, and intents.

4.2. Step B: Extraction of Texts and Characteristics

In this step, we refine the extraction process by determining the text's relevant parts and corresponding characteristics. Unlike Step A, where we retrieved all parts of the texts and their characteristics, we filtered the most relevant ones. To apply this filter effectively, we manually create a list comprising the desired intents and domains. This list should be based on the existing KG and the important knowledge we want to add.

For instance, if our KG describes compatibility relations between products and also stores reviews of products within the gardening domain, the list of important characteristics should include Compatibility and Review intents, the gardening domain, a question-and-answer section, and reviews as text sources. We discard texts with intents not listed in this filter.

We introduce **Prompt 2** to materialize this filtering process. This serves as a guideline for the LLM to restrict and refine the extracted phrases and their corresponding characteristics based on important characteristics. This comprises three main parts:

1. **Task Description:** The task involves filtering and restricting phrases and their corresponding characteristics based on important characteristics. This process ensures that only relevant information aligned with the specified criteria is retained. The output is a list of filtered phrases;
2. **Few-shot Examples:** These examples illustrate the input texts, their corresponding characteristics, the set of important characteristics, and the filtered phrases. This provides a clear understanding of how the filtering process operates in practice, demonstrating the refinement of information based on the specified criteria;

[1] You can access supplementary materials online: https://colab.research.google.com/drive/1tobyFXvTGuj-WfPydmZ2LHElljb_6HWu?usp=sharing

3. **Input:** The unfiltered texts and characteristics and the list of important characteristics.

4.3. Step C: Generation of RDF Triples

In this step, our focus shifts to creating RDF triples based on the list of filtered texts generated in Step B. To generate these triples, we incorporate additional inputs, specifically a list of important classes and properties derived from an existing ontology. This list serves as a guideline for the LLM to discern which knowledge should be represented in RDF format. By providing this input, along with the pertinent ontology elements, the LLM can effectively discern and represent relevant knowledge in RDF format, enhancing the richness and coherence of the KG. For instance, if a text contains information about a product voltage but is not structured in the ontology or the KG, it should be discarded, and no triples should be produced based on this voltage knowledge.

To facilitate the RDF triple generation process, we designed the **Prompt 3** with three main parts:

1. **Task Description:** Create RDF triples based on the texts while adhering to the list of important classes and properties of the ontology. This ensures that only relevant knowledge aligned with the specified ontology is represented in RDF format;
2. **Few-shot Examples:** These examples demonstrate the input comprising the filtered texts, the set of important classes and properties of the ontology, and the corresponding output of RDF triples generated. This provides a clear illustration of how the texts are transformed into structured RDF format while respecting the constraints imposed by the ontology;
3. **Input:** Filtered texts from Step B, set of important classes and properties of the ontology.

4.4. Step D: Evaluation and Addition to Existing KG

In this step, our method evaluates the generated RDF triples to ensure their validity before adding them to an existing KG. The LLM may produce erroneous RDF triples, including hallucinations and incorrect representations of knowledge not present in the original texts. Therefore, it is imperative to verify that the generated RDF triples adhere to specific criteria:

- **No Redundancies:** Ensure that there are no duplicate RDF triples in the generated set;
- **No Semantic Inconsistencies:** Verify that the generated RDF triples do not introduce semantic inconsistencies, such as assertions of disjoint classes.
- **Respect the List of Classes and Properties:** Eliminate RDF triples containing properties and classes not present in the predefined set of important classes and properties derived from the ontology.
- **Correct Order of Triples:** Guarantee that the order in which the RDF triples are added to the KG does not produce errors, such as attempting to assert properties of entities before the entities themselves are added.

To operationalize the validation criteria, we designed the **Prompt 4**, comprising three parts:

1. **Task Description:** Validate the generated RDF triples using specific criteria to ensure their correctness before addition to the KG;
2. **Few-shot Examples:** These examples illustrate the input, comprising the set of generated RDF triples, the set of important classes and properties from the ontology, and the validation criteria. The output consists of the list of validated RDF triples that adhere to the specified criteria, ensuring their suitability for addition to the existing KG;
3. **Input:** Set of generated RDF triples from Step C, set of important classes and properties from the ontology, validation criteria (a) No redundancies, (b) No semantic inconsistencies, (c) Respect the list of classes and properties, (d) Correct order of triples.

5. Experimental Evaluation

This section presents the application of our method in transforming an e-commerce text into a set of RDF triples while ensuring adherence to the three characteristics identified: intents, text sources, and domains. We implemented all the prompts and steps described in Section 4 to understand the benefits and weaknesses of using LLMs in our context. We selected two cases with very different characteristics, as our method's rationale is to deal with multiple characteristics. As a general procedure, we chose two real-world e-commerce cases, executed the prompts described in Section 4 using the HuggingFace API[2] to utilize the open-source LLM called Bloom [14], and obtained the results that can be observed in Section 5.1 and Section 5.2. We chose Bloom because it is an open-source pioneer in multilingual LLMs, which is important for our Portuguese language testing scenario, and for the ease of using the method through a free online API via Hugging Face.

5.1. Assessement 1: Electronic Compatibility Q&A Text

Scenario: A customer visits a Brazilian E-commerce platform's product page for a VGA cable. They inquired about the compatibility of the cable with their specific monitor, a Samsung T3 monitor with a 27-inch display. In addition, the customer queries the store's operating hours, specifically inquiring if the store is open on Sundays. The vendor responds affirmatively regarding the cable's compatibility with the customer's monitor and states that the store operates from Monday to Friday, from 8:00 AM to 6:00 PM.

Step A: In the initial step of our method, Step A, we identify texts based on their inherent characteristics. Utilizing **Prompt 1**, we discern that the provided text *D*, consisting of a question and an answer, originates from the question and answer section of the product ($S = \{Q\&ASection\}$). Moreover, our analysis reveals that the domain pertains to electronics, indicating the subject matter of the inquiry *Do*. In addition, the intents inferred from the text encompass compatibility queries regarding product usage and inquiries about store information ($I = \{Compatibility, StoreInformation\}$). Table 2 summarizes our characteristics and corresponding values. Through this analysis, we understand the text's context and attributes, facilitating subsequent processing in the RDF triple-generation process.

[2] https://huggingface.co/bigscience/bloom

Table 2
List of characteristics identified in *D*.

Characteristic	Identified Value
Text Source	Question and Answer Section
Domain	Electronics
Intents	Compatibility, Store Information

Step B: In the next phase of our method, Step B comes into play. We retrieve a list of key characteristics previously registered, taking into account the most important attributes of the current KG. In this case, the existing KG stores information about product compatibility, product specifications, and user reviews. It contains knowledge about electronics, furniture, home decor, and construction. Table 3 lists the important characteristics.

Given that the inquiry is related to compatibility and falls within the electronics domain, we prompt the LLM to filter the text accordingly, ensuring that only phrases containing the allowed characteristics are retained. Consequently, the phrase, which encompasses multiple intents, including store information, is segmented, as store information is not among the key characteristics. As a result, the LLM returns the segmented question and answer "The product is compatible with your monitor, a 27-inch Samsung monitor" as the filtered list of phrases.

Table 3
List of key characteristics that the KG accepts as valid knowledge. The characteristics in blue color are the ones identified in *D*, in our assessment case.

Characteristic	Value
Domain	Electronics
	Furniture
	Home Decor
	Construction
Intent	Compatibility
	Specification
	Review

Step C: In Step C, we generate RDF triples based on the filtered phrases containing intents, domains, and sources. Assuming we have filtered phrases, we aim to generate triples that accurately represent the input data.

To facilitate this process, we manually identify the ontology elements previously registered as important elements of the intent and domain in our case. These elements are compiled into a list of classes and properties (cf. Table 4). For example, some of the classes identified include *ConsumerItem*, *Product*, *FullCompatibility*, and *NoCompatibility*, while properties such as *has:Voltage*, *has:Brand*, and *has:Model* are also included.

All these classes and properties have been manually registered in advance for the specified domain and intent. The list of important classes and properties, along with the question and answer and examples, is input for **Prompt 3**. The output of Step C is a set of RDF triples, as follows:

Table 4
Important classes and properties of the ontology. The generated RDF triples should adhere to the classes and properties.

Ontology Element	Value
Classes	*ConsumerItem*
	Product
	FullCompatibility
	NoCompatibility
Properties	*has:Voltage*
	has:Brand
	has:Model

- *<ConsumerItem/monitor-samsung-27> rdf:type onto:ConsumerItem*
- *<ConsumerItem/monitor-samsung-27> onto:hasBrand ”samsung”*
- *<ConsumerItem/monitor-samsung-27> onto:hasModel ”T3”*
- *<ConsumerItem/monitor-samsung-27> onto:hasDimensions ”54x58”*
- *<Product/store_name/cabo-vga-150cm> rdf:type onto:Product*
- *<FullCompatibility/store_name/cabo-vga-150cm/monitor-samsung-27> rdf:type onto:FullCompatibility*
- *<Product/store_name/cabo-vga-150cm> onto:hasCompatibility <FullCompatibility/store_name/cabo-vga-150cm/monitor-samsung-27>*
- *<FullCompatibility/store_name/cabo-vga-150cm/monitor-samsung-27> onto:compatibleWith <Product/store_name/cabo-vga-150cm>*

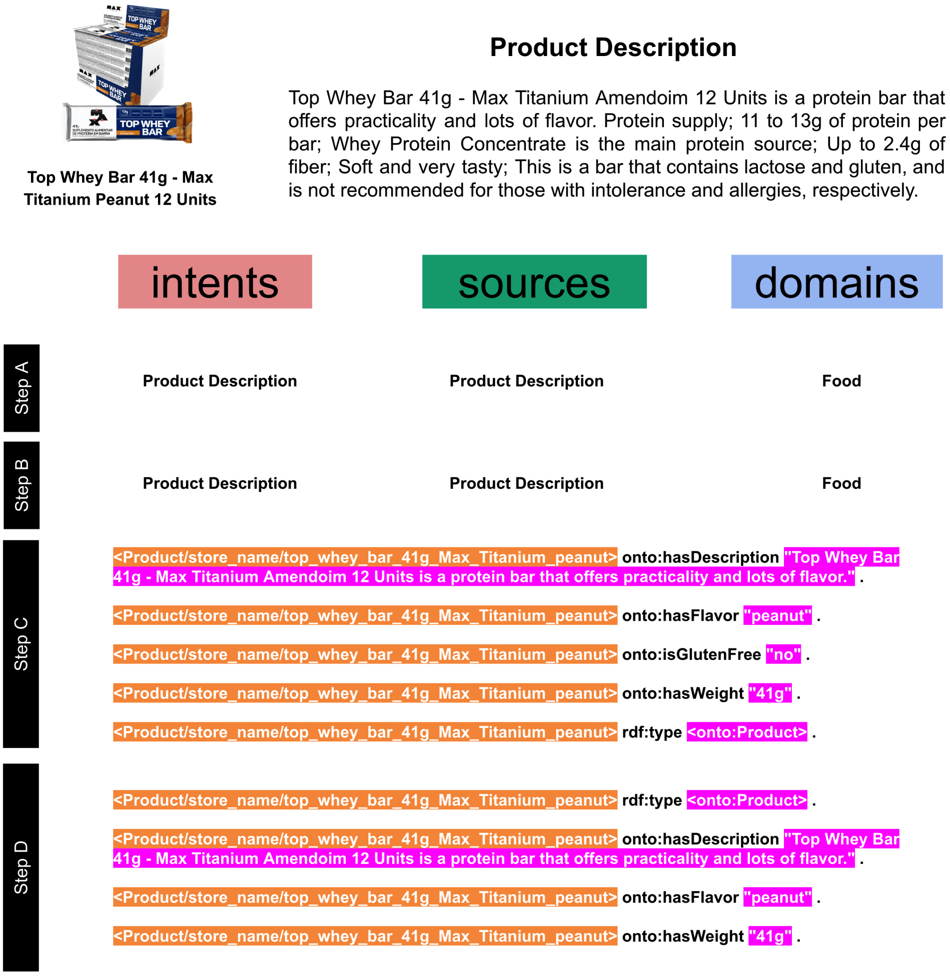
Step D: Step D validates the generated RDF triples. This process involves an automated investigation to detect any classes or properties used that were not listed as elements of the ontology. The corresponding triple containing the unrecognized element is discarded if such instances are found.

For instance, in our testing case, an additional triple stating the dimensions of the monitor as 54x58 was generated, utilizing the property *has:Dimensions*. However, this property was not listed among the elements of the ontology. Therefore, this triple must be removed from the set of generated triples to ensure adherence to the ontology.

The validation process includes checks for the order of triples, redundancies, and semantic inconsistencies. These checks are crucial for ensuring the integrity and coherence of the generated RDF triples before their addition to the KG. All these validations are performed through **Prompt 4**, orchestrating the automated investigation and verification process. The RDF triples are now ready to be appended to the existing KG.

5.2. Assessement 2: Food Product Description Text

Scenario: A store sells fitness food products and wants to represent it in a KG once it enables a product recommendation feature when populated with the products. One of the products being sold is a kit with 12 units of a protein bar. The ontology maintainer wants to represent the product description automatically in a KG format, respecting the ontology properties and classes. Figure 2 shows the product image, title, and description.



Top Whey Bar 41g - Max Titanium Peanut 12 Units

Product Description

Top Whey Bar 41g - Max Titanium Amendoim 12 Units is a protein bar that offers practicality and lots of flavor. Protein supply; 11 to 13g of protein per bar; Whey Protein Concentrate is the main protein source; Up to 2.4g of fiber; Soft and very tasty; This is a bar that contains lactose and gluten, and is not recommended for those with intolerance and allergies, respectively.

	intents	sources	domains
Step A	Product Description	Product Description	Food
Step B	Product Description	Product Description	Food

Step C

<Product/store_name/top_whey_bar_41g_Max_Titanium_peanut> onto:hasDescription "Top Whey Bar 41g - Max Titanium Amendoim 12 Units is a protein bar that offers practicality and lots of flavor." .

<Product/store_name/top_whey_bar_41g_Max_Titanium_peanut> onto:hasFlavor "peanut" .

<Product/store_name/top_whey_bar_41g_Max_Titanium_peanut> onto:isGlutenFree "no" .

<Product/store_name/top_whey_bar_41g_Max_Titanium_peanut> onto:hasWeight "41g" .

<Product/store_name/top_whey_bar_41g_Max_Titanium_peanut> rdf:type <onto:Product> .

Step D

<Product/store_name/top_whey_bar_41g_Max_Titanium_peanut> rdf:type <onto:Product> .

<Product/store_name/top_whey_bar_41g_Max_Titanium_peanut> onto:hasDescription "Top Whey Bar 41g - Max Titanium Amendoim 12 Units is a protein bar that offers practicality and lots of flavor." .

<Product/store_name/top_whey_bar_41g_Max_Titanium_peanut> onto:hasFlavor "peanut" .

<Product/store_name/top_whey_bar_41g_Max_Titanium_peanut> onto:hasWeight "41g" .

Figure 2: Summary of important elements of the assessment 2. At the top of the figure, we show the product name, image, and description. Below is a table with columns representing the characteristics and rows representing the steps of the pipeline. The cells represent the result output of each step based on the corresponding characteristic.

Step A: By running **Prompt 1**, we discern that the provided text D, consisting of a product description, originates from the description section of the product ($S = \{ProductDescriptionSection\}$). Our analysis reveals that the domain pertains to food. The intent inferred from the text is product description ($I = \{ProductDescription\}$). The row with the identifier "Step A" of Figure 2 shows the intents, sources, and domains found in the text.

Step B: At Step B our method retrieves a list of key characteristics, previously registered, taking into account the most important attributes of the current KG. In our case, the existing KG stores triples of product compatibility and description. It contains knowledge about a single domain: food. Given that the inquiry is related to product description and falls within the food domain, we prompt the LLM to filter the text accordingly, ensuring that only phrases containing the allowed characteristics are retained. Consequently, the whole text is retrieved, constituting the filtered list of phrases. The list of filtered characteristics is available in the row with the identifier ”Step B” in Figure 2.

Step C: In Step C, we manually identify the ontology elements previously registered as important elements of the intent and domain in our case. These elements are compiled into a list of classes and properties, summarized in Table 5. Based on these elements, our method generates the RDF triples in the row with the identifier ”Step C” in Figure 2.

Table 5
Important classes and properties of the ontology. The generated RDF triples should adhere to the classes and properties.

Ontology Element	Value
Classes	*ConsumerItem*
	Product
	FullCompatibility
	NoCompatibility
Properties	*rdf_type*
	hasWeight
	hasDescription
	hasFlavor

Step D: Our method detects properties used that were not listed as elements of the ontology: an additional triple stating that the product is not gluten-free was generated, utilizing the property *isGlutenFree*. This triple was removed from the set of generated triples to ensure adherence to the ontology. In addition, the validation process included checks for the order of triples, redundancies, and semantic inconsistencies. It identifies that the triple containing *rdf:type* should be added before the other triples. The final list of RDF triples generated by our method is shown in the row with the identifier “Step D” in Figure 2. After the four steps, the RDF triples are ready to be appended to the existing KG.

6. Discussion

This section discusses the implications and findings of our method. We examine the strengths, limitations, and potential avenues for future research.

Our method offers a systematic approach to transforming unstructured e-commerce texts into structured RDF triples, thereby enriching KGs with valuable information. By leveraging LLMs, we effectively extract intents, domains, and text sources from the input text, facilitating the generation of relevant RDF triples.

Our solution is novel because it integrates ontology elements and examples from the existing KG. Our method ensures semantic rigor in the generated triples by employing a set of ontology statements rather than generating triples without any predefined format, entity, or property. This integration of structured ontology elements with unstructured text represents a significant contribution, allowing for creating triples that accurately represent the underlying information. Ultimately, this approach enables us to generate triples with semantic precision by incorporating aspects of the ontology within the prompt for the LLM.

In Step A of our method, we utilized a LLM to classify the characteristics present in the text. However, it is worth noting that a full-fledged LLM may not be necessary for this task. A simpler model like BERT can suffice for this classification task. A less complex model would offer a more efficient solution because the primary goal is to categorize text rather than generate it.

Furthermore, the validation (Step D) ensures the coherence and adherence of the generated triples to the ontology, enhancing the quality of the KG. We proposed validating the generated triples to ensure their integrity. However, an additional consideration for validation involves leveraging a reasoner. This would enable validation against the existing KG, ensuring no inconsistencies, such as duplicate triples or incoherent statements. By employing a reasoner, we can enhance the validation process, ensuring the overall coherence and integrity of the KG.

Despite its effectiveness, our method faces challenges. One significant issue is identifying intents, domains, and text sources, particularly in complex e-commerce texts with diverse content. The reliance on manually curated lists of important classes and properties may introduce biases and overlook emerging concepts not covered in the ontology. The validation process, while essential for ensuring the integrity of the generated triples, may introduce computational overhead, particularly for large-scale KGs.

For future work, we plan to refine the process of characteristics identification (Step A) by eliminating noise from input texts, making them more factual and succinct. This involves removing extraneous elements such as greetings and irrelevant sections of the text that do not contribute to the extraction of intents and domains. Given the inherent noise in natural language texts, this refinement can improve the accuracy of the generated triples. Furthermore, leveraging user feedback and iterative refinement processes enhances the ontology and validation criteria, ensuring their relevance and adaptability to evolving E-commerce domains.

7. Conclusion

In e-commerce, the exchange of information between humans and machines encounters a significant challenge in bridging the gap between human-friendly language and machine-understandable data structures. Our approach presented a systematic framework for converting unstructured e-commerce texts into structured knowledge representations, paving the way for advanced applications like personalized recommendations and semantic search. The primary difficulty lied in automating the generation of RDF triples from natural language questions in e-commerce to connect them with existing ontologies. Complexities are multifactorial, ranging from traditional NLP challenges, like synonymy and polysemy, to machine learning challenges, like intent and domain classification. Our findings underscored the significance of integrating ontology elements with unstructured text to generate semantically enriched KGs with coherence via prompts in LLMs. Despite encountering challenges like precise characteristic identification and validation, our method provided valuable contributions to knowledge representation and information retrieval in e-commerce.

Acknowledgments

This study was financed by the National Council for Scientific and Technological Development - Brazil (CNPq) process number 140213/2021-0. In addition, this research was partially funded by the São Paulo Research Foundation (FAPESP) (grants #2022/13694-0 and #2022/15816-5). The opinions expressed in this work do not necessarily reflect those of the funding agencies.

References

[1] D. T. Sant'Anna, R. O. Caus, L. dos Santos Ramos, V. Hochgreb, J. C. dos Reis, Generating knowledge graphs from unstructured texts: Experiences in the e-commerce field for question answering., in: ASLD@ ISWC, 2020, pp. 56–71.

[2] Y. Cao, X. Wang, X. He, Z. Hu, T.-S. Chua, Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences, in: The world wide web conference, 2019, pp. 151–161.

[3] A. G. Regino, R. O. Caus, V. Hochgreb, J. C. d. Reis, Leveraging knowledge graphs for e-commerce product recommendations, SN Computer Science 4 (2023) 689. doi:10.1007/s42979-023-02149-6.

[4] T. T. Huynh, N. V. Do, T. N. Pham, N. T. Tran, A semantic document retrieval system with semantic search technique based on knowledge base and graph representation, in: New Trends in Intelligent Software Methodologies, Tools and Techniques, IOS Press, 2018, pp. 870–882.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[6] F.-L. Li, H. Chen, G. Xu, T. Qiu, F. Ji, J. Zhang, H. Chen, Alimekg: Domain knowledge graph construction and application in e-commerce, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 2581–2588.

[7] T. Xu, C. Guo, L. Du, J. Xu, P. Zhang, X. Feng, M. Li, A method for traditional chinese medicine knowledge graph dynamic construction, in: Proceedings of the 5th International Conference on Big Data Technologies, ICBDT '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 196–202. URL: https://doi.org/10.1145/3565291.3565323. doi:10.1145/3565291.3565323.

[8] J. Fei, W. Zeng, X. Zhao, X. Li, W. Xiao, Few-shot relational triple extraction with perspective transfer network, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 488–498. URL: https://doi.org/10.1145/3511808.3557323. doi:10.1145/3511808.3557323.

[9] H. Balabin, C. T. Hoyt, C. Birkenbihl, B. M. Gyori, J. Bachman, A. T. Kodamullil, P. G. Plöger, M. Hofmann-Apitius, D. Domingo-Fernández, Stonkgs: a sophisticated transformer trained on biomedical text and knowledge graphs, Bioinformatics 38 (2022) 1648–1656.

[10] D. Dessí, F. Osborne, D. R. Recupero, D. Buscaldi, E. Motta, Scicero: A deep learning and

nlp approach for generating scientific knowledge graphs in the computer science domain, Knowledge-Based Systems 258 (2022) 109945.

[11] A. Chessa, G. Fenu, E. Motta, F. Osborne, D. R. Recupero, A. Salatino, L. Secchi, Data-driven methodology for knowledge graph generation within the tourism domain, IEEE Access (2023).

[12] A. Tchechmedjiev, P. Fafalios, K. Boland, M. Gasquet, M. Zloch, B. Zapilko, S. Dietze, K. Todorov, Claimskg: A knowledge graph of fact-checked claims, in: The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18, Springer, 2019, pp. 309–324.

[13] N. Howard, E. Cambria, Intention awareness: improving upon situation awareness in human-centric environments, Human-centric Computing and Information Sciences 3 (2013) 1–17.

[14] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).