

Filtering Candidate Triples for Wikidata Enrichment through Joint Relation Extraction across Text and Tables

Mouna Kamel^{1,2,*}, Nathalie Aussenac-Gilles^{2,†} and Véronique Moriceau^{2,†}

¹Espace-Dev, Université de Perpignan, France

²Université de Toulouse, Toulouse INP, CNRS, IRIT, Toulouse, France

Abstract

Wikidata is a collaborative, structured, and multilingual knowledge base that plays a central role in the Semantic Web ecosystem. Its continuous enrichment relies on information collected from a variety of sources, including Wikipedia pages. While many relation extraction approaches have been proposed to support knowledge base construction, most of them focus on a single modality, such as textual content, tables, or infoboxes. In previous work, we introduced the ReTaT dataset to evaluate the extraction of cross-modal triples in which one argument is mentioned in a table and the other in the surrounding text. This paper investigates to what extent such text-table triples, when automatically extracted, can contribute to Wikidata enrichment. We propose a pipeline that aligns such triples with Wikidata entities and properties, then combines structural, semantic, and contextual constraints to classify them as eligible for automatic enrichment, requiring human validation, or rejected. An evaluation on the BizRel dataset shows that 46.5% of the triples can be automatically integrated, 39.5% require human validation, and 12.7% are rejected. These results demonstrate the value of combining Wikipedia text and tables for both automatic and human-in-the-loop Wikidata enrichment.

Keywords

Joint Relation Extraction, Text-table triple, Wikidata enrichment, LLM

1. Introduction

Wikidata is a collaborative, structured, and multilingual knowledge base that plays a central role in the Semantic Web ecosystem. It is continuously enriched with knowledge drawn from various sources, among which Wikipedia constitutes a particularly valuable resource. In Wikidata, information is represented as semantic triples, typically composed of two arguments connected by a predicate: (argument_1, predicate, argument_2). This representation has motivated a substantial body of research on the extraction of triples from Wikipedia pages to support the construction and enrichment of knowledge bases. However, most existing approaches remain limited to a single type of source within Wikipedia pages, focusing either on infoboxes, textual content, or tables, instead of jointly exploiting the complementary information conveyed across these modalities.

However, tables in a document rarely convey meaning in isolation; they are generally supported by surrounding text that offers necessary context. Conversely, the narrative text frequently introduces and expands upon information that is synthesized into tables. This strong semantic interplay suggests that considering tables and text jointly rather than separately could significantly improve the quality and completeness of relation extraction. This is illustrated by the example in Figure 1. In the paragraph, we can identify lexically expressed relations such as "Roman J. Israel, Esq. is a product of the Topic Studio company", "the film Roman J. Israel, Esq. was directed by Dan Gilroy", "Denzel Washington stars in Roman J. Israel, Esq.", "Roman J. Israel, Esq. was directed by Columbia Pictures", etc. The table, in turn, carries relationships expressed between the values in

Wikidata'26: Wikidata Workshop at ISWC 2026

*Corresponding author and main contributor.

†These authors contributed equally.

✉ mouna.kamel@univ-perp.fr (M. Kamel); aussenac@irit.fr (N. Aussenac-Gilles); veronique.moriceau@irit.fr (V. Moriceau)

🌐 https://hal.science/search/index/q/*/authIdHal_s/mouna-kamel (M. Kamel); <https://www.irit.fr/~Nathalie.Aussenac-Gilles/> (N. Aussenac-Gilles); <http://www.irit.fr/~Veronique.Moriceau/> (V. Moriceau)

🆔 0000-0002-8739-0747 (M. Kamel); 0000-0003-3653-3223 (N. Aussenac-Gilles); 0000-0001-9641-0714 (V. Moriceau)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

the **Title** and **Notes** columns, such as "The film *Roman J. Israel, Esq.* is distributed by Sony Pictures Releasing", "The film *Leave No Trace* is distributed by Bleecker Street", etc. While the paragraph mentions only one Topic Studio production, the table enriches this information by presenting, in chronological order, other productions by the studio (including the one cited in the paragraph), without the name of the company being explicitly mentioned in the table. Collaborative relationships between Topic Studio and Roadside Attractions or Neon can also be identified.

Only by jointly processing the two modalities text and table can one extract complete and semantically rich relations such as:

(Leave No Trace, product_or_service_of, Topic Studio),
 (The Oath, product_or_service_of, Topic Studio),
 (Topic Studio, collaboration, Roadside Attractions),
 (Topic Studio, collaboration, Neon), and so on.

In June 2017, it was announced [First Look Media](#) would be launching a film and television production and distribution company [Topic Studios](#).^[1] The company's first film was *[Roman J. Israel, Esq.](#)* directed by [Dan Gilroy](#), starring [Denzel Washington](#) and released in November 2017, by [Columbia Pictures](#).^[2]

Release Date	Title	Notes
November 17, 2017	<i>Roman J. Israel, Esq.</i>	distributed by Sony Pictures Releasing
June 29, 2018	<i>Leave No Trace</i>	distributed by Bleecker Street ^[6]
October 12, 2018	<i>The Oath</i>	co-distributed with Roadside Attractions ^[7]
June 7, 2019	<i>XY Chelsea</i>	distributed by Showtime ^[8]
August 2, 2019	<i>Luce</i>	co-distributed with Neon ^[9]
October 4, 2019	<i>Wrinkles the Clown</i>	distributed by Magnet Releasing
October 18, 2019	<i>The Laundromat</i>	distributed by Netflix
November 15, 2019	<i>The Report</i>	distributed by Amazon Studios ^[10]

Figure 1: Screenshot of a wikipedia page to illustrate relation expression across text and tables https://en.wikipedia.org/wiki/Topic_Studios. Key triples are only recoverable by combining textual and tabular information.

A preliminary study was conducted on paragraph–table pairs annotated at three levels: text only, table only, and jointly across both modalities [1]. It showed that 70.7% of the jointly identified triples were novel, as they emerged only from the combined interpretation of text and tables. This result suggests that multimodal integration can reveal relations missed by conventional unimodal approaches. Building on these findings, the present work investigates whether such triples can enrich Wikidata by providing novel, relevant, and reusable knowledge not yet explicitly represented in the knowledge base. Our work does not address triple extraction methods but rather Wikidata automatic enrichment with extracted triples. It contributes to the triple extraction task by evaluating the amount of added triples to Wikidata. The main contributions of this work can be summarized as follows:

1. Assessment of the Wikidata enrichment rate enabled by text-table relations. Our first contribution consists in evaluating the hypothesis that relations jointly expressed across text and tables can enrich Wikidata. Using manually annotated text–table triples, we measure their enrichment potential by estimating the proportions corresponding to knowledge already represented in Wikidata, missing but potentially reusable knowledge, and knowledge whose validity cannot be established automatically. We also assess whether this enrichment involves new entities,

properties, or triples.

2. Semi-automatic classification of automatically extracted text–table triples. Our second contribution consists in proposing a semi-automatic process for assessing the suitability of automatically extracted text–table triples for Wikidata enrichment. The process first aligns their subjects, predicates, and objects with Wikidata entities and properties, then applies structural and semantic constraints together with contextual heuristics. Based on these criteria, triples are classified as eligible for automatic integration, requiring human validation, or unsuitable for enrichment.

The remainder of this paper is organized as follows. Section 2 reviews the state of the art on relation extraction from textual and tabular content, as well as methods for aligning extracted relations with knowledge graphs. In Section 3, we assess the Wikidata enrichment potential of manually annotated text-table triples. Building on these findings, Section 4 presents the proposed method for automatically classifying triples as eligible for automatic enrichment, requiring human validation, or rejected. This method is evaluated in Section 5 on the BizRel dataset [1]. We then discuss the main findings, limitations, and perspectives of this work before concluding the paper in Section 6.

2. Related Work

Many previous studies have focused on extracting semantic relations from textual structures such as sentences [2], paragraphs [3], or tables [4, 5], to name just a few. Less numerous, a few approaches are capable of reasoning over both text and structured tabular data such as tables. TaBERT [6], an extension of the BERT model [7], is a language model designed to understand both natural language and tabular data. It was pre-trained on a corpus of 26 million tables and their textual contexts in English. TaBERT’s performance has been validated on question answering (QA) and SQL query generation tasks, after fine-tuning using manually annotated QA datasets. Zayats et al. [8] introduced a novel mechanism to enrich table representations based on the surrounding text, using an adaptation of Graph Neural Networks within a transformer framework. Their model was pre-trained on 1.6 million Wikipedia tables and a corpus of real-world natural language questions. Significant performance gains on QA tasks were observed. A closer step towards our setting is ReSel [9], which tackles N-ary relation extraction from scientific documents using both text and tables. The authors pose the N-ary relation extraction task as a question-answering problem: given a query structured as a partial N-ary tuple (e.g., <Task, Method, Dataset, Metric, Score>), a BERT-based model is trained to identify the missing final entity (Score) within the scientific document. Thus, it is indeed a relation extraction task, but only the tail entity (argument2) is retrieved, and the relation label is not explicitly identified. In 2025, the TRIPLET challenge¹ gathered original LLM-based contributions to extract relations expressed jointly on a table and the surrounding text.

In the scope of building or enriching a robust knowledge graph, extraction tools must ensure that extracted entities are meaningful, well-defined, and interoperable with already identified parts of the graph. This requires at first linking the triple head and tail to entities in the graph or to assign them a type using the classes of an ontology schema. Thanks to the quality of its design in compliance with a clear schema, Wikidata is one of most popular target knowledge graphs to evaluate relation extraction. For instance, ReBEL uses a seq2seq model to generate triples where entities and predicates are mapped to those of a knowledge graph, such as FreeBase or Wikidata [10]. The schema of Wikidata can be used as a guideline to either generate triples or to filter out generated triples. In [11], a transformer encoder-decoder model takes unstructured text as input and autoregressively generates triples thanks to a beam search constrained with the Wikidata schema. More recently, [12] proposed a task symmetric to relation extraction: a given Wikidata triple and a web page, they want to extract and validate the text entailing this triple in this page. Their approach uses an LLM with few-shot in-context learning and Chain-Of-Thought prompting.

Many datasets and benchmarks exist to evaluate the ability of relation extraction systems to identify Wikidata triples, such as the one defined for ReBEL [10]. [13] gathered a set of such datasets (available

¹<https://ecladatta.github.io/triplet2026/>

online² that they curated for the extraction of Wikidata entities and properties. Due to the difficulty of deciding whether a mention contains a relation or not, many benchmark datasets are of poor quality. To overcome this limitation, an alternative is to generate high quality synthetic data, like the SynthIE dataset proposed in [14]. The authors of SynthIE used LLMs for a symmetric task to relation extraction that LLMs perform very efficiently: the generation of sentences from Wikidata triples. In a similar way, the authors of [12] designed their own dataset to evaluate their LLM-based relation extraction approach. This dataset contains 80.5K triple-to-text alignments with diverse text covering 90K entities and 910 relations from Wikidata.

In the context of extracting relations expressed across text and tables, the organizers of the TRIPLET challenge associated with the ESWC2026 conference proposed the ReTaT benchmark [15]. ReTaT contains 3 domain-specific datasets (about Business, Telecommunications, and Female Celebrities). Each dataset consists of a collection of pairs (document, triple) where the document contains a table and surrounding text taken from Wikipedia pages of the 2024 Wikipedia dumps. The triple results from a manual annotation with domain-specific semantic relations between entities distributed across text and tables. For each corpus, the list of relation types is closed. A dedicated interface was used to support synchronized annotation and, when available, alignment with Wikidata QIDs and PIDs. The annotations were produced by knowledge-engineering experts, with an inter-annotator agreement of $\kappa = 0.70$. Further details on ReTaT are provided in [15].

3. Assessment of the Wikidata enrichment rate enabled by manually annotated text-table triples

To evaluate the potential of enriching Wikidata with text–table triples, we compiled several datasets of manually annotated triples, which we present here. Then we aligned these triples with Wikidata entities and properties. Finally we propose heuristics to filter out candidate triples to enrich Wikidata.

3.1. Datasets

For this study, we selected two of the three ReTaT subsets [15]: BizRel (business/economics) and TelCo (telecommunications) because of their contrasted and complementary characteristics. We did not consider the Female Celebrities corpus because annotations did not take into account qualifiers, whereas the historical nature of the content would require spatial and temporal dimensions to assert the validity of a triple.

First, the 35 target relations in TelCo already correspond to Wikidata properties, such as P2283, P2144, P880, and others. Therefore, the alignment of predicates with Wikidata properties is direct. In contrast, the 11 relations targeted in the BizRel dataset are derived from the CORE corpus [16], a high-quality resource specifically designed for the extraction of economic relations (i.e. *product of*, *service of*, *subsidiary of*, etc.). The alignment with Wikidata properties was then performed manually, highlighting that some of them have an exact Wikidata counterpart, such as *collaboration*, which corresponds to property P2652, whereas others have no exact and generic equivalent in Wikidata. This is the case for instance for *client_of*: several approximate Wikidata properties exist, such as P88 and P8956, but they are more specific or context-dependent and do not generally cover the exact meaning of the original relations defined in CORE. Next, BizRel focuses exclusively on entity–entity relations, whereas TelCo covers a broader range of Wikidata statements and is particularly rich in datatype properties, i.e., relations between an entity and a literal value. Finally, BizRel provides paragraph-level textual context in English, while TelCo relies on the full original Wikipedia page in English. Overall BizRel contains 255 text-table pairs, from which 2,774 triples were annotated, while TelCo contains 32 text-table pairs, from which 1,574 triples were annotated. Table 1 summarizes these characteristics.

²<https://github.com/Shimorina/relation-extraction-db-wikidata>

Table 1

Characteristics of the 2 corpora used for text–table triple extraction.

Dataset	BizRel	TelCo
Domain	Business	Telecommunication
Number of target relations	11	35
Number of text-table pairs	255	32
Number of annotated triples	2774	1574
Language	English	English
Types of relation	entity–entity	entity–entity & entity–literal
Set of target relations	closed set of business relations	closed set of Wikidata relations
Text size	paragraph	full-page

3.2. Quantitative analysis of Wikidata enrichment potential

We now evaluate the potential contribution of these annotated triples to Wikidata enrichment.

First of all, and for the remainder of this paper, we establish the following correspondence between triples extracted from each (text, table) document and Wikidata triples. In an extracted triple, the two elements connected by a relation are referred to as *Argument₁* and *Argument₂*, while the relation itself is called the predicate. In a Wikidata triple, *Argument₁* corresponds to the subject entity, the predicate corresponds to a Wikidata property, and *Argument₂* corresponds either to an object entity or, in the case of a datatype property, to a literal value. Thus, an extracted triple of the form (*Argument₁*, *predicate*, *Argument₂*) is represented in Wikidata as (*subject entity*, *property*, *object entity value*).

We identified three forms of potential Wikidata enrichment: the creation of new entities, the creation of new relations, and the addition of new triples. Table 2 describes the triple alignment patterns associated with each form: in each position, a QID or PID indicates an existing alignment, whereas an underscore (_) indicates that no corresponding Wikidata identifier was found. These missing elements are precisely those that indicate a potential opportunity for enrichment. In particular, new entities or a new predicate linking existing entities can ultimately enable Wikidata enrichment once the corresponding QIDs and PIDs have been created.

Table 2

Cases of potential Wikidata enrichment

Case	Triple form	Type of enrichment
Case1: Missing subject or object entity or both	(_, PID, QID) or (QID, PID, _) or (_, PID, _)	new(s) entity(ies)
Case2: Missing value	(QID , PID, _)	new statement
Case3: Missing property	(QID, _, QID/value)	new relation
Case4: Complete triple not present in Wikidata	(QID, PID, QID/value)	new statement

We now evaluate each of these enrichment cases separately.

3.2.1. Wikidata enrichment with new entities

Candidate entities correspond to non-literal arguments that were not aligned with any Wikidata entity during the annotation process and therefore have no associated QID. Table 3 reports the number of entities that could serve as candidates for enriching Wikidata, in the two datasets.

As shown in Table 3, BizRel contains a much larger number of candidate entities than TelCo, with 810 candidates compared to 160. This difference is mainly due to the fact that TelCo is rich in triples involving datatype properties, whose objects are literal values rather than entities. The global enrichment rates, 35.57% for BizRel and 43.36% for TelCo, confirm the interest of the proposed approach for supporting Wikidata enrichment.

Table 3

Evaluation of entity-based enrichment for the BizRel and TelCo datasets.

Case	BizRel			TelCo		
	Number of entities	Number of unaligned entities	Enrichment rate	Number of entities	Number of unaligned entities	Enrichment rate
C1	2277	810	35.5%	369	160	43.3%

3.2.2. Wikidata enrichment with new properties

As said before, the BizRel corpus was annotated using a predefined set of relations derived from the CORE corpus [16]. Among these relations, three do not have a direct equivalent in Wikidata, namely *competitor of*, *client of*, and *acquired by*. These relations capture business interactions that are insufficiently represented in Wikidata and therefore offer a valuable opportunity to enrich its coverage of corporate ecosystems.

3.2.3. Wikidata enrichment from triples

Table 4 reports the triple-based enrichment results obtained for the BizRel and TelCo datasets. For triples involving a datatype property, we considered a triple to be complete as soon as the QID of the first argument and the PID of the property were successfully identified.

Table 4

Evaluation of triple-based enrichment for the BizRel and TelCo datasets.

Case	BizRel			TelCo		
	Number of triples	Number of complete triples not present in Wikidata	Enrichment rate	Number of triples	Number of complete triples not present in Wikidata	Enrichment rate
C2+C4	2774	1284	46.2%	1574	1237	78.5%
C1+C3		1490	53.7%		337	21.4%
Total		2754	99%		1561	99%

For the BizRel dataset, only 20 triples were already present in Wikidata, while 2754 were absent, including 1284 complete triples. Similarly, for the TelCo dataset, only 13 triples were already available in Wikidata, whereas 1561 were missing, including 1237 complete triples. Overall, approximately 99% of the manually annotated triples from both datasets are not represented in Wikidata. These findings demonstrate that relations extracted jointly from text and tables can reveal a substantial amount of previously unavailable knowledge and thus make a significant contribution to enrich Wikidata.

4. Filtering automatically extracted triples as candidates for Wikidata enrichment

The previous evaluation confirmed the relevance of our approach for enriching Wikidata with triples combining textual and tabular information. However, it relied on manually annotated data, where relation validity and Wikidata alignment were ensured by the provided QIDs and PIDs.

When triples are automatically extracted, the enrichment process involves a layer of uncertainty. Errors can therefore occur at different stages:

- during argument identification, entity mentions may be incorrectly detected or delimited;
- during predicate identification, the extracted relation may fall outside the targeted relation types;

- during alignment with Wikidata, the arguments and predicate may be incorrectly linked to Wikidata entities and properties.

The objective is therefore to construct two sets of triples, after discarding those identified as erroneous: one containing triples eligible for automatic Wikidata enrichment, and another containing triples that must be submitted to an expert for validation before enrichment. For this purpose, and taking these potential extraction errors into account, we propose a method for aligning extracted triples with Wikidata, complemented by filtering constraints and heuristics.

4.1. Method for Aligning Extracted Triples with Wikidata

Aligning triples with Wikidata consists of identifying the corresponding Wikidata elements for the components of each extracted triple. More precisely, given a triple of the form (*argument*₁, *predicate*, *argument*₂), the alignment process attempts to map the first argument to a Wikidata entity, the predicate to a Wikidata property, and the second argument to a Wikidata entity when the property range is not a literal datatype. We first describe predicate alignment, which is relatively direct, followed by the more complex process of argument alignment.

4.1.1. Predicate-Property alignment

The alignment of the extracted predicates with Wikidata properties is relatively straightforward, insofar as the predicates are targeted in advance. This reduces ambiguity, as each predicate is linked to a predefined relation type. We plan to deal with inverse properties in future work. Indeed, when an extracted triple does not match a Wikidata triple, and if the corresponding Wikidata property has an inverse property, then we should check the existence of a Wikidata triple linking the entities with the inverse relation.

4.1.2. Argument-Entity alignment

This alignment process requires associating each extracted triple with its source Wikipedia page, which provides the contextual information needed to disambiguate the arguments and identify the corresponding Wikidata entities. Accordingly, the adopted method for aligning an argument of a triple with a Wikidata entity proceeds in three successive steps:

- The metadata of the Wikipedia page from which the triple containing the argument currently being processed was extracted are retrieved through the Wikipedia API³. When available, these metadata include the Wikidata identifier (QID) associated with the entity described by the page. When the argument matches the page title, it is linked to the corresponding Wikidata QID.
- Second, when the argument does not match the page title, the method examines the hyperlinks contained in the source Wikipedia page. If an argument matches the title of one of the linked Wikipedia pages, the Wikidata QID associated with that page is retrieved through the Wikipedia API, taking into account possible Wikipedia redirects.
- Finally, when the previous steps fail, the process queries the Wikidata API⁴ to retrieve candidate entities whose labels or aliases match the argument to be aligned. It then represents both the local Wikipedia context of the argument and the profile of each candidate entity, including its label, description, and aliases, as semantic embeddings. It selects the QID with the highest cosine similarity which is assigned to the argument, and returns no link when the score is too low or the result is too ambiguous.

³<https://en.wikipedia.org/w/api.php>

⁴<https://www.wikidata.org/w/api.php>

4.1.3. Evaluation of the alignment

Since predicate alignment is relatively direct without significant difficulty, we focus the evaluation exclusively on entity alignment on the BizRel and TelCo datasets. We compared the QIDs assigned to the extracted arguments with the reference one in the ReTaT benchmark, resulting from a manual annotation. Table 5 defines true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) based on whether the alignments proposed by the algorithm match the QIDs in the manual annotations. These categories are then used to compute the evaluation metrics.

Table 5
Definitions of TP, TN, FP, and FN for the alignment evaluation.

Category	Definition
TP	the reference value is a QID, and the method retrieves the same QID.
FP	the reference value is null, but the method proposes a QID.
FN	the reference value is a QID, but the method either returns null or retrieves an incorrect QID.
TN	the reference value is null, and the method does not propose any QID.

The evaluation results for the two datasets, BizRel and TelCo, are reported Table 6 using the standard performance metrics: precision, recall, accuracy, and F1-score.

Table 6
Performance of entity alignment in the BizRel and TelCo corpora

Dataset	Accuracy	Precision	Recall	F1-score
BizRel	75.71%	71.46%	81.25%	76.04%
TelCo	59.08%	46.38%	52.15%	49.10%

The results show a clear performance gap between the BizRel and TelCo datasets. For BizRel, the model achieves relatively strong performance, with an accuracy of 75.71% and an F1-score of 76.04%. The high recall indicates that the model is effective at identifying most positive instances, although the number of false positives remains significant, which slightly lowers precision. In contrast, the results for TelCo are much weaker, with an accuracy of 59.08% and an F1-score of 49.10%. These lower scores are likely due to the fact that the TelCo triples were extracted from highly technical Wikipedia pages, where entities are often domain-specific, abbreviated, or described in specialized terminology. This technical nature may make entity disambiguation more difficult and increase both false positives and false negatives.

4.2. Constraints and Heuristics

The objective is to divide the set of aligned candidate triples into three subsets:

- Rejected triples
- Triples eligible for automatic integration into Wikidata
- Triples requiring expert validation

To construct these subsets, we define the following four constraints and the heuristic described below.

Constraint C1 - Targeted relation. This constraint aims to determine if the PID associated to the predicate belongs to the predefined list of targeted relations/PID. Otherwise, the triple is discarded.

Constraint C2 – Cross-modal arguments. This constraint checks whether the two arguments of the triple originate from the two complementary sources: one from the table and the other from the surrounding text. Otherwise, the triple is discarded.

Constraint C3 — Novelty. The triple does not already exist in Wikidata. Otherwise, the triple is discarded.

Constraint C4 — Type compatibility. This constraint verifies whether each aligned argument, i.e. each argument associated with a Wikidata QID is compatible with the domain or range of the property PID, depending on whether it is the subject or the object of the triple. Otherwise, the triple is discarded.

Heuristic H1 — Intra-table consistency. This heuristic exploits the unique properties of text-table structures. It is based on the idea that triples sharing either the same predicate and the same first argument, or the same predicate and the same second argument, are likely to express a coherent semantic relationship within the table. The triples are therefore grouped according to these similarity criteria. If at least one triple in a group is found in Wikidata, this validation is propagated to all triples in the same group. As a result, triples belonging to this group and not already in Wikidata are considered eligible for automatic Wikidata enrichment.

Algorithm 1 specifies the order in which these constraints and heuristics are applied and, consequently, how sets of triples considered candidates for automatic or expert-validated Wikidata enrichment are being built. Altogether, these constraints and the heuristics eliminate erroneous triples among all automatically extracted ones.

5. Evaluation

We now evaluate our enrichment approach on a dataset composed of triples automatically extracted by a Large Language Model from the Wikipedia pages of the BizRel dataset.

5.1. LLM and dataset

Our evaluation is based on the complete set of triplets automatically extracted by a large language model from the Wikipedia pages that constitute the BizRel dataset. At this stage of our study, we have not evaluated the approach on the TelCo corpus because most of the extracted triples involve datatype properties, and the objects extracted by the LLM for such properties do not include qualifiers. For example, the LLM extracted the triple (“Bluetooth Low Energy”, “price (P2284)”, “US\$5.00”), in which “US\$5.00” is represented as a string, whereas the expected datatype for P2284 is Quantity. The price should instead be represented as a quantity (q) associated with a monetary unit, in this case the US dollar (Q4917). Consequently, constraint C4 would be violated for all triples of this type, even though addressing this issue merely requires additional preprocessing steps, which are part of our planned future work.

The LLM was selected based on a prior evaluation comparing **llama-4-maverick-17b-128e-instruct** and **deepseek-r1-distill-llama-70b** on the same corpora [1]. Both models were tested in zero-shot and few-shot settings, using three demonstrations in the latter case, and evaluated through exact matching against manually annotated triples. Since **deepseek-r1-distill-llama-70b** achieved the best F1-score on both datasets in the few-shot setting, this configuration was retained for extracting candidate triples. Table 7 summarizes the results.

5.2. Results and discussion

Our filtering method was applied to the 1274 candidate triples extracted by DeepSeek-R1 in the few-shot setting from the Wikipedia pages included in the BizRel dataset. The results obtained after applying the filtering algorithm are presented below:

T_{auto} : 593 triples eligible for automatic integration into Wikidata, representing 46.5% of the triples.

T_{expert} : 504 triples assigned to an expert validation, representing 39.5% of the triples.

T_{rejected} : 163 triples have been rejected, representing 12.7% of the triples.

Algorithm 1: Construction of candidate triples for Wikidata enrichment

Input : T , the set of automatically extracted triples; $\mathcal{R}$, the set of targeted relations;

Output: T_{expert} , the set of triples requiring expert validation; T_{auto} , the set of triples eligible for automatic Wikidata enrichment; T_{rejected} , the set of rejected triples;

Notation: For any triple t , $p(t)$ denotes its predicate, $\text{arg}_1(t)$ and $\text{arg}_2(t)$ its first and second arguments, respectively; $\text{SameTable}(\text{arg}_i(t), \text{arg}_i(t'))$ means that the two arguments originate from the same table.

$T_{\text{expert}} \leftarrow T$; $T_{\text{auto}} \leftarrow \emptyset$; $T_{\text{rejected}} \leftarrow \emptyset$;

foreach $t \in T$ **do**

“ // C1: Targeted relation constraint

if $p(t) \notin \mathcal{R}$ **then**

$T_{\text{expert}} \leftarrow T_{\text{expert}} \setminus \{t\}$;

$T_{\text{rejected}} \leftarrow T_{\text{rejected}} \cup \{t\}$;

continue;

// C2: Text-table complementarity constraint

if $\neg \left[(\text{arg}_1(t) \in \text{Text} \wedge \text{arg}_2(t) \in \text{Table}) \vee (\text{arg}_1(t) \in \text{Table} \wedge \text{arg}_2(t) \in \text{Text}) \right]$ **then**

$T_{\text{expert}} \leftarrow T_{\text{expert}} \setminus \{t\}$;

$T_{\text{rejected}} \leftarrow T_{\text{rejected}} \cup \{t\}$;

continue;

// H1: Grouping heuristic

$G \leftarrow \left\{ t' \in T_{\text{expert}} \mid \begin{array}{l} (\text{arg}_1(t) = \text{arg}_1(t') \wedge p(t) = p(t') \wedge \text{SameTable}(\text{arg}_2(t), \text{arg}_2(t'))) \\ \vee (\text{arg}_2(t) = \text{arg}_2(t') \wedge p(t) = p(t') \wedge \text{SameTable}(\text{arg}_1(t), \text{arg}_1(t'))) \end{array} \right\}$;

if $\exists t' \in G : \text{ExistsInWikidata}(t')$ **then**

$T_{\text{auto}} \leftarrow T_{\text{auto}} \cup G$;

$T_{\text{expert}} \leftarrow T_{\text{expert}} \setminus G$;

“

// C4: Type compatibility constraint

foreach $t \in \text{copy}(T_{\text{expert}})$ **do**

“ **if** $\text{Aligned}(\text{arg}_1(t)) \wedge \neg \text{Compatible}(\text{Type}(\text{arg}_1(t)), \text{Domain}(p(t)))$ **then**

$T_{\text{expert}} \leftarrow T_{\text{expert}} \setminus \{t\}$;

$T_{\text{rejected}} \leftarrow T_{\text{rejected}} \cup \{t\}$;

continue;

if $\text{Aligned}(\text{arg}_2(t)) \wedge \neg \text{Literal}(\text{arg}_2(t)) \wedge \neg \text{Compatible}(\text{Type}(\text{arg}_2(t)), \text{Range}(p(t)))$ **then**

$T_{\text{expert}} \leftarrow T_{\text{expert}} \setminus \{t\}$;

$T_{\text{rejected}} \leftarrow T_{\text{rejected}} \cup \{t\}$;

continue;

if $\text{Aligned}(\text{arg}_1(t)) \wedge \text{Compatible}(\text{Type}(\text{arg}_1(t)), \text{Domain}(p(t))) \wedge \left[\text{Literal}(\text{arg}_2(t)) \vee \right.$

$\left. \left(\text{Aligned}(\text{arg}_2(t)) \wedge \text{Compatible}(\text{Type}(\text{arg}_2(t)), \text{Range}(p(t))) \right) \right]$ **then**

$T_{\text{auto}} \leftarrow T_{\text{auto}} \cup \{t\}$;

$T_{\text{expert}} \leftarrow T_{\text{expert}} \setminus \{t\}$;

“

// Remove triples already present in Wikidata

foreach $t \in \text{copy}(T_{\text{expert}})$ **do**

if $\text{ExistsInWikidata}(t)$ **then**

$T_{\text{expert}} \leftarrow T_{\text{expert}} \setminus t$;

foreach $t \in \text{copy}(T_{\text{auto}})$ **do**

if $\text{ExistsInWikidata}(t)$ **then**

$T_{\text{auto}} \leftarrow T_{\text{auto}} \setminus t$;

return $T_{\text{expert}}, T_{\text{auto}}, T_{\text{rejected}}$;

Table 7

Results of Few-shot settings for DeepSeek model on the BizRel and TelCo datasets.

Dataset	Precision	Recall	F1-score
BizRel	0.71	0.31	0.43
TelCo	0.39	0.27	0.32

Among the triples classified into the automatic integration and expert validation sets, 14 were removed because they were already represented in Wikidata (1.3% of the triples). And among the rejected triples, 64 (39.3%) were discarded because their arguments did not exhibit the expected text-table complementarity. A further 60 triples (36.8%) were rejected because the type of the first argument was incompatible with the domain of the corresponding Wikidata property, while 39 triples (23.9%) were rejected because the type of the second argument was incompatible with the property range. Overall, these results show that more than 86% of the extracted triples were retained either for automatic integration or for expert validation.

Among the triples requiring expert validation, 388 contain at least one argument that could not be aligned with an existing Wikidata entity. Once the missing entities have been created in Wikidata, the corresponding triples can be resubmitted to the filtering algorithm. Their updated alignment status may allow some of them to satisfy the eligibility criteria for automatic integration, while others may still require expert validation or be rejected.

6. Conclusion and Perspectives

In this work, we first assessed the potential for enriching Wikidata using manually extracted reference triples whose relations are jointly expressed across textual and tabular content. The results showed that nearly 99% of these triples could contribute new knowledge to Wikidata. This particularly high proportion may also be explained by the fact that approximately 40% of the entities involved in these triples are not currently represented in Wikidata. These missing entities therefore constitute an additional source of enrichment in their own right.

In a second stage, we proposed a filtering process for identifying triples eligible for automatic Wikidata enrichment among those produced by an automatic relation extraction approach. This process first relies on an alignment algorithm that maps the components of the extracted triples ($argument_1$, $predicate$, $argument_2$) to the corresponding components of a Wikidata statement ($entity_1$, $property$, $entity_2/literal\ value$). The aligned triples are then filtered out using a set of constraints and heuristics designed to distinguish triples that can be automatically integrated from those requiring expert validation or those that should be rejected.

Applied to triples extracted in a few-shot setting by the **DeepSeek-R1-Distill-Llama-70B** model, the proposed process showed that 46.5% of the triples were eligible for automatic Wikidata enrichment, while 39.5% required expert validation. These findings highlight the substantial amount of knowledge conveyed jointly by textual and tabular content and demonstrate the importance of a controlled filtering process for enabling reliable and scalable Wikidata enrichment.

Several directions are planned for future work. Firstly, the algorithm evaluating Wikidata enrichment relies on various constraints and heuristics. Two of them, C2 and H1, are particular to the table-text setting. We plan to conduct new experiments to evaluate their effect on the overall results.

Our second main objective is to improve the ReTaT dataset by explicitly accounting, during manual annotation, for the qualifiers associated with datatype properties in the TelCo and Famous Women subsets. Accordingly, the LLM-based relation extraction process will need to be adapted so that it can extract not only datatype-property relations and their literal values, but also the qualifiers required to represent these statements accurately in Wikidata. These extensions will make it possible to evaluate Wikidata enrichment not only through object properties, as in the current study, but also through datatype properties. This evolution will consequently lead to the definition of new constraints and heuristics specifically designed to assess the validity, completeness, and integration eligibility of

datatype-property statements.

Furthermore, although the alignment process between extracted triples and Wikidata statements achieves an accuracy of 75.7% and an F1-score of 76% for triples involving object properties, future work will be dedicated to identify ways of improving its performance. In particular, further investigations could exploit additional contextual information from both textual and tabular content, enhance entity and property disambiguation, and make greater use of Wikidata type constraints and semantic similarities. Improving this alignment stage is essential, since the reliability of the subsequent filtering and enrichment decisions directly depends on the correctness of the mappings established between extracted arguments and predicates and their corresponding Wikidata entities and properties.

Acknowledgement

This work was carried out within the ECLADATTA project (2022-2027) funded by the French National Research Agency under grant ANR-22-CE23-0020. More information at <https://ecladatta.github.io/>.

Use of Generative AI

The authors used *Chat-GPT Plus (GPT5.6 Sol)* to assist with editing algorithm 1, and DeepL to support the translation into English. They reviewed and revised all generated text and take full responsibility for the content. Generative AI was not used to generate code or experimental results, fabricate citations, or produce figures.

References

- [1] M. Ettaleb, V. Moriceau, M. Kamel, N. Aussenac-Gilles, The contribution of LLMs to relation extraction in the economic field, in: Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal) - Proceedings of the workshops, Association of Computational Linguistics, Abu Dhabi, United Arab Emirates, 2025, pp. 175–183. URL: <https://hal.science/hal-04940817>.
- [2] X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, R. Xu, A comprehensive survey on relation extraction: Recent advances and new frontiers, *ACM Comput. Surv.* 56 (2024). URL: <https://doi.org/10.1145/3674501>. doi:10.1145/3674501.
- [3] Y. Zheng, Y. Guo, Z. Luo, Z. Yu, K. Wang, H. Zhang, H. Zhao, A survey on document-level relation extraction: Methods and applications, in: Proceedings of the 3rd International Conference on Internet, Education and Information Technology (IEIT 2023), Atlantis Press, 2023, pp. 1061–1071. URL: https://doi.org/10.2991/978-94-6463-230-9_128. doi:10.2991/978-94-6463-230-9_128.
- [4] Z. Jiang, Y. Mao, P. He, G. Neubig, W. Chen, OmniTab: Pretraining with Natural and Synthetic Data for Few-shot Table-based Question Answering, in: M. Carpuat, M. de Marneffe, I. V. M. Ruíz (Eds.), Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL, Association for Computational Linguistics, 2022, pp. 932–942. URL: <https://doi.org/10.18653/v1/2022.naacl-main.68>.
- [5] J. Yang, A. Gupta, S. Upadhyay, L. He, R. Goel, S. Paul, Tableformer: Robust transformer modeling for table-text encoding, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), 60th Annual Meeting of the Association for Computational Linguistics ACL, Association for Computational Linguistics, 2022, pp. 528–537. URL: <https://doi.org/10.18653/v1/2022.acl-long.40>. doi:10.18653/v1/2022.acl-long.40.
- [6] P. Yin, G. Neubig, W. Yih, S. Riedel, TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), 58th Annual Meeting of the Association for Computational Linguistics ACL, Association for Computational Linguistics,

2020, pp. 8413–8426. URL: <https://doi.org/10.18653/v1/2020.acl-main.745>. doi:10.18653/v1/2020.ACL-MAIN.745.

- [7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423/>. doi:10.18653/v1/N19-1423.
- [8] V. Zayats, K. Toutanova, M. Ostendorf, Representations for question answering from documents with tables and text, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL, Association for Computational Linguistics, 2021, pp. 2895–2906. URL: <https://doi.org/10.18653/v1/2021.eacl-main.253>. doi:10.18653/v1/2021.EACL-MAIN.253.
- [9] Y. Zhuang, Y. Li, J. Zhang, Y. Yu, Y. Mou, X. Chen, L. Song, C. Zhang, Resel: N-ary relation extraction from scientific text and tables by learning to retrieve and select, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Conference on Empirical Methods in Natural Language Processing EMNLP, Association for Computational Linguistics, 2022, pp. 730–744. URL: <https://doi.org/10.18653/v1/2022.emnlp-main.46>. doi:10.18653/v1/2022.EMNLP-MAIN.46.
- [10] P.-L. Huguet Cabot, R. Navigli, REBEL: Relation extraction by end-to-end language generation, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2370–2381. URL: <https://aclanthology.org/2021.findings-emnlp.204/>. doi:10.18653/v1/2021.findings-emnlp.204.
- [11] M. Josifoski, N. D. Cao, M. Peyrard, F. Petroni, R. West, Genie: Generative information extraction, 2022. URL: <https://arxiv.org/abs/2112.08340>. arXiv:2112.08340.
- [12] N. Mihindukulasooriya, S. Hertling, Towards extracting triple-to-text alignments from wikidata references using llms, in: E. Curry, V. Presutti, J. McCrae, M. Alam, P. Colpaert, J. Xavier Parreira, D. Collarana, M. Sabou, A. Harth, P. Lisena (Eds.), The Semantic Web: ESWC 2025 Satellite Events, Springer Nature Switzerland, Cham, 2026, pp. 52–57.
- [13] A. Shimorina, J. Heinecke, F. Herledan, Knowledge extraction from texts based on Wikidata, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, Association for Computational Linguistics, Hybrid: Seattle, Washington + Online, 2022, pp. 297–304. URL: <https://aclanthology.org/2022.naacl-industry.33>.
- [14] M. Josifoski, M. Sakota, M. Peyrard, R. West, Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 1555–1574. URL: <https://aclanthology.org/2023.emnlp-main.96/>. doi:10.18653/v1/2023.emnlp-main.96.
- [15] M. Ettaleb, T. Ehrhart, N. Aussenac-Gilles, Y. Chabot, M. Kamel, V. Moriceau, R. Troncy, F. Wei, Retat: A unified benchmark for relation extraction across text and table, in: S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, A. Toral (Eds.), Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), Palma, Mallorca, Spain, 2026, pp. 1341–1351. doi:10.63317/4k8jqsbkjuir.
- [16] P. Borchert, J. De Weerd, K. Coussement, A. De Caigny, M.-F. Moens, CORE: A few-shot company relation classification dataset for robust domain adaptation., in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023, pp. 11792–11806. URL: <https://doi.org/10.18653/v1/2023.emnlp-main.722>. doi:10.18653/v1/2023.emnlp-main.722.

A. Online Resources

The sources (data) used in this paper are available at <https://www.codabench.org/competitions/12936/>.