

Chapter 4

Reconciliation-Based Data Enrichment



Flavio De Paoli, Matteo Palmonari, and Federico Belotti

Abstract Data enrichment is a crucial component of data preparation, aimed at improving the quality of input datasets by integrating external information from third-party sources. One approach to support this process is reconciliation-based data enrichment. In this approach, data providers organize their information around entities and expose reconciliation services for data consumers. Data consumers use these services to match tabular values in their data with entities in the source they want to access, enabling them to fetch additional information by querying the source with canonical identifiers. This approach reduces information asymmetry and facilitates access to external data, making enrichment more effective and user-friendly. The chapter introduces reconciliation-based enrichment, reviews current techniques for entity reconciliation in tabular data, and highlights the emerging role of large language models in supporting such tasks. Finally, it discusses the advantages of reconciliation-based enrichment with respect to sustainability principles, including reducing computational effort, limiting human intervention, and promoting resource efficiency in data preparation pipelines.

4.1 Introduction

Data collected from real-world sources is often raw, heterogeneous, and incomplete, and therefore cannot be directly exploited in AI applications or analytical workflows. Before meaningful analyses can be performed, data must undergo a preprocessing and enrichment phase aimed at improving quality, consistency, and usability. Data enrichment is the process by which existing datasets are augmented with additional

F. De Paoli (✉) · M. Palmonari · F. Belotti
University of Milan-Bicocca—DISCo, Milano, Italy
e-mail: flavio.depaoli@unimib.it

M. Palmonari
e-mail: matteo.palmonari@unimib.it

F. Belotti
e-mail: federico.belotti@unimib.it

information derived from external sources, thereby making them more complete, informative, and actionable.

A common approach to enrich a source dataset with content from one or more third-party data sources is to obtain a local copy of that data and apply data integration techniques [2]. An alternative approach, aligned with sustainability strategies, is based on *entity reconciliation* and fetches only the data relevant to the input table, possibly through an API, thus avoiding the assumption of full control over the third-party data sources.

A reconciliation-based process involves identifying the entities present in a dataset—such as people, locations, or organizations—and annotating them with unique identifiers drawn from reference vocabularies. Using vocabularies in this way provides a shared semantic framework, which facilitates the integration of datasets managed by different organizations and structured according to heterogeneous schemas. In other words, vocabularies act as standardized “glue” that aligns data elements across sources.

Vocabularies define terms, labels, and sometimes relationships among concepts, and are critical for standardizing language and reducing lexical ambiguity. When combined with ontologies, which provide a formal, structured representation of domain knowledge including classes, relationships, and logical constraints, ambiguities can be further reduced and dataset quality enhanced. The adoption of reference vocabularies and ontologies also significantly increases the reusability of data across domains and applications. Examples of widely used vocabularies include Schema.org, which provides terms for describing web resources, and FOAF (Friend Of A Friend), which models social connections and person-related information.

Of particular importance in modern data-driven applications are Knowledge Graphs (KGs), which represent entities as graph nodes and relationships among them as labeled edges. Entities in a KG are often defined according to an ontology, providing semantic structure that enables reasoning and inference. Prominent examples include Wikidata, DBpedia, and the Google Knowledge Graph, which have been widely employed for data integration, enrichment, and semantic search.

A critical aspect of data enrichment is entity disambiguation, which resolves ambiguities when the same name or label may refer to different real-world entities. This process typically involves matching entities in source datasets to a target KG to assign unique identifiers, ensuring that each entity can be clearly distinguished. Although automated algorithms can perform much of this reconciliation, in many cases human expertise is necessary to resolve ambiguous or complex situations. Once entities are reconciled, datasets can be integrated seamlessly. For example, a dataset containing weather observations may be linked to a KG of cities, allowing weather data to be enriched with additional geographic, demographic, or temporal information through identifier alignment.

While there exists a wide array of tools to support data preprocessing, annotation, and enrichment, these tasks remain challenging and resource-intensive, often requiring substantial expertise, time, and computational resources. Many existing tools are designed for domain experts, limiting accessibility for less experienced users. Nonetheless, such enrichment processes are essential for any high-quality data

analysis pipeline, as they provide the semantic foundation upon which AI models, analytics, and decision-making systems can operate effectively.

This chapter introduces reconciliation-based enrichment, reviews current techniques for entity reconciliation in tabular data, and highlights the emerging role of large language models in supporting such tasks. Finally, it discusses the advantages of reconciliation-based enrichment with respect to sustainability principles, including reducing computational effort, limiting human intervention, and promoting resource efficiency in data preparation pipelines.

Section 4.2 introduces the enrichment process and its main tasks—Modification, Reconciliation, and Extension—through an illustrative example. Section 4.3 presents the semantic annotation of tabular data, with particular focus on techniques addressing the key problem of entity reconciliation. Section 4.4 compares reconciliation techniques in terms of accuracy, execution time, memory usage, and sustainability. Finally, Sect. 4.5 concludes the chapter.

4.2 Enrichment Process

Data enrichment can be regarded as a specialized form of data integration aimed at augmenting a source dataset with additional information retrieved from external data sources. Unlike classical integration scenarios, enrichment is typically asymmetric: only the source dataset is under the control of the integrator, while external sources are autonomous, independently evolving, and only partially known. Moreover, data originating from different sources are often highly heterogeneous in terms of structure, representation, and semantics. This asymmetry significantly affects both the problem formulation and the set of applicable techniques.

As a consequence, classical data integration techniques—such as *Record Linkage* and schema-level integration approaches—are often insufficient or not directly applicable to enrichment tasks. Instead, data enrichment requires methods capable of coping with limited knowledge about external sources, semantic heterogeneity, and the absence of global schemas, motivating the adoption of semantic-based approaches.

To this end, semantic enrichment can be decomposed into two core tasks: *reconciliation* and *extension*. Reconciliation can be regarded as a form of entity linking applied to tabular data, consisting in mapping table mentions to entities in a target system of identifiers (e.g., geographical coordinates for locations) or a Knowledge Graph (e.g., Wikidata IDs for popular persons or organizations) so as to assign unique and persistent identifiers. The task is typically carried out through automatic algorithms; however, due to the inherent ambiguity and contextual dependence of textual mentions, the resulting mappings may be uncertain or ambiguous, thus requiring human intervention to validate, disambiguate, or refine the associations.

Once mentions have been reconciled, the extension task exploits the assigned identifiers to retrieve additional information from external data sources and incorporate it into the original table. This process is commonly based on Linked Data princi-

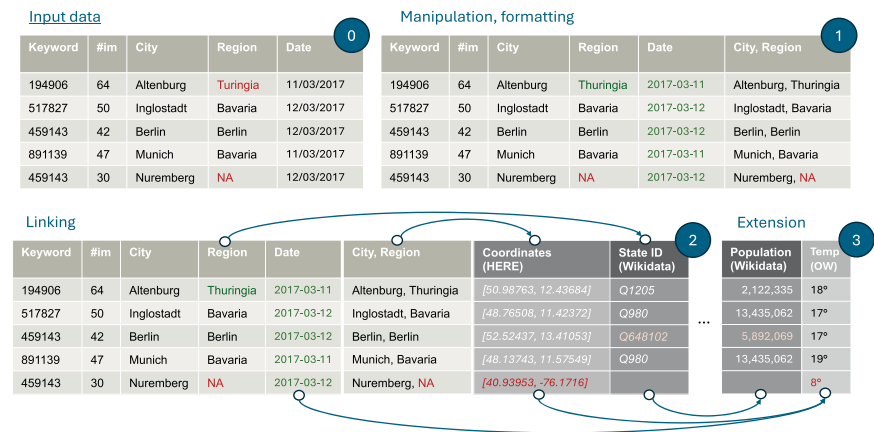


Fig. 4.1 Example of data preparation for optimizing digital marketing campaigns. Numbers identify data preparation steps. The background colors indicate data from different sources. Arrows trace how columns contribute to derived content. Text colors indicate data quality: red for missing or incorrect values; green for improved, reliable values

ples [5], allowing heterogeneous datasets to be combined through shared identifiers and semantic relationships.

To illustrate the process of enrichment, we consider the data preparation task in the digital marketing domain that has been discussed in [16]. With the objective of maximizing the performance of keyword-based marketing campaigns, we plan to study the correlation between impressions—the number of times the advertisement (AD) associated with the campaign has been displayed in search results—and weather conditions. The available data are the reports from the online advertising platforms, e.g., Google Ads, that include keyword identifiers, number of impressions, locations (city-region), and dates of the collection (Step 0 of Fig. 4.1). In this context, data preparation is a massive operation involving the management of 5 billion active keywords and the generation of more than 1M clicks per day in 234 countries and 15 languages.

As shown in the figure, accuracy and completeness issues may occur in the input data, such as misspellings, missing values, and so on. To prepare the data for analysis, assume that two data-enrichment tasks are required: (1) retrieving location information to normalize the data by population (state population in this case); and (2) retrieving weather information for the location on the given day (temperature in this example), as campaign optimization depends on weather conditions, which are known to influence user purchasing behavior [8].

Data on the population at the state level can be collected from open data sources such as Wikidata, while weather data can be obtained from weather services such

Output data

Keyword	#i	City	Region	Date	City, Region	Coordinates (HERE)	State ID (Wikidata)	Population (Wikidata)	Temp (OW)
194906	64	Altenburg	Thuringia	2017-03-11	Altenburg, Thuringia	[50.98763, 12.43684]	Q1205	2,122,335	18°
517827	50	Inglostadt	Bavaria	2017-03-12	Inglostadt, Bavaria	[48.76508, 11.42372]	Q980	13,435,062	17°
459143	42	Berlin	Berlin	2017-03-12	Berlin, Berlin	[52.52437, 13.41053]	Q64	3,782,202	17°
891139	47	Munich	Bavaria	2017-03-11	Munich, Bavaria	[48.13743, 11.57549]	Q980	13,435,062	19°
459143	30	Nuremberg	Bavaria	2017-03-12	Nuremberg, Bavaria	[49.45421, 11.07752]	Q980	13,435,062	12°

Fig. 4.2 The output table for the example of data preparation for digital marketing campaigns

as OpenWeather.¹ Population data may suffer from timeliness issues, whereas both temperature and population data can be affected by the presence of outliers.

Additional data issues related to weather services concern the use of specific formats—geocoordinates for locations and ISO 8601 for dates—which may differ from those adopted in the input data. Moreover, locality information can be incomplete or inaccurate, thus requiring disambiguation actions. For instance, geocoding services can be used to retrieve city coordinates (e.g., HERE²), while Wikidata reconciliation services such as OpenRefine³ can support the disambiguation of regions. These matching operations are inherently subject to uncertainty and potential errors.

As illustrated in Fig. 4.1, the data preparation process is structured as a pipeline composed of three main tasks—Manipulation and Formatting, Reconciliation, and Extension—followed by a validation and refinement phase aimed at detecting, assessing, and correcting anomalies in order to produce a high-quality enriched table.

In the remainder of this section, each phase is described in detail to clarify how the final output reported in Fig. 4.2 is systematically obtained through successive enrichment and refinement cycles.

4.2.1 Step 1: Manipulation and Formatting

The first step of the enrichment process focuses on the manipulation and formatting of the input data to improve data quality and ensure compatibility with subsequent enrichment operations. This step primarily addresses syntactic aspects of data preparation, targeting quality dimensions such as accuracy, consistency, and interoperability, which are prerequisites for effective semantic linking and data extension.

Manipulation activities include the correction of spelling errors in textual attributes, the normalization of character encoding, and the harmonization of naming conventions for locations. City and region names, in particular, may exhibit substantial lexical variability due to abbreviations, alternative spellings, or language-dependent forms. Standardizing these values mitigates ambiguity and increases the

¹ <https://openweathermap.org/>.
² <https://www.here.com/platform/geocoding>.
³ <https://wikidata.reconci.link/>.

robustness of downstream matching procedures. When location information is provided in separate columns (e.g., city and region), these attributes can be concatenated to form a composite location field, thereby facilitating subsequent disambiguation and geocoding.

Formatting operations focus on achieving syntactic interoperability with external services. Temporal attributes are converted into standardized representations, such as ISO 8601, to comply with the input requirements of weather data services. Similarly, spatial attributes may need to be reformatted to match the expected structure of geocoding services, for instance, by enforcing a consistent ordering of city and region names or by removing unnecessary textual elements. These transformations are essential to prevent query failures and misinterpretations caused by mismatches.

Although this step does not introduce new information, it plays a critical role in the enrichment pipeline. Issues left unresolved at this stage may propagate to later phases, negatively affecting the effectiveness of semantic linking and the reliability of the enriched data. For this reason, manipulation and formatting constitute a foundational preparation phase that directly impacts the quality and robustness of the linking and extension steps discussed in the following sections.

4.2.2 Step 2: Reconciliation

The second step of the enrichment process concerns the linking of values appearing in the input table to structured identifiers that enable integration with external data sources. Reconciliation establishes explicit connections between input values and entities, records, or identifiers provided by third-party systems, such as knowledge graphs, geocoding services, or domain-specific databases. As illustrated in Fig. 4.1, this step marks the transition from syntactic data preparation to semantic interpretation.

Reconciliation operations are typically performed using matching and linking techniques and are inherently subject to uncertainty. Ambiguity, lexical variability, and incompleteness in the input data may lead to incorrect or suboptimal associations. The downstream impact of such errors varies depending on the nature of the mismatch and on the role of the linked identifiers in subsequent enrichment steps.

For example, a geocoding service may mislink the city name “Nuremberg” to a location in the United States, resulting in coordinates that refer to a completely different geographical area and, consequently, to unrelated weather conditions. In contrast, Wikidata reconciliation may associate “Berlin” with the Berlin–Brandenburg Metropolitan Region⁴ rather than with the city entity.⁵ While the city could be considered a more precise match, the resulting discrepancy may have a limited effect on downstream analysis, as several attributes of the two entities are closely related.

⁴ <https://www.wikidata.org/wiki/Q648102>.

⁵ <https://www.wikidata.org/wiki/Q64>.

Different reconciliation strategies can be adopted depending on the characteristics of the input data and the target sources. *String-based approaches* rely on lexical similarity measures, such as exact matching, edit distance, or token-based similarity, and are generally efficient but sensitive to spelling variations and ambiguity. *Semantic approaches* exploit background knowledge, such as entity types, relationships, or contextual information available in knowledge graphs, to perform disambiguation beyond surface forms. *Hybrid approaches* combine lexical similarity with semantic constraints or contextual signals, aiming to balance efficiency and accuracy. These strategies differ in terms of robustness, computational cost, and error profiles, and are discussed in detail in the next section.

Given the inherent uncertainty of automated reconciliation, human-in-the-loop mechanisms play a crucial role in ensuring data quality. A first class of corrective techniques is based on rule-driven validation and correction. Domain rules, integrity constraints, and plausibility checks can be used not only to identify inconsistencies—such as links that violate geographic, temporal, or semantic constraints, or that produce implausible attribute values—but also to systematically correct them. In this way, human intervention can be captured in the form of explicit rules that prevent the recurrence of the same errors in future executions.

For example, in the case of the value “Berlin”, human feedback may establish a context-specific rule stating that, for the purposes of the considered analysis, the city entity should be preferred over the corresponding metropolitan region. Once defined, such a rule can be applied automatically to resolve similar ambiguities, thereby improving consistency and reducing the need for repeated manual intervention. Rules of this kind can be used either to guide automated re-linking or to constrain the set of candidate entities considered during the matching process.

Beyond rule-based validation, active learning techniques can be employed to guide human intervention more effectively. In this setting, the system selectively presents uncertain or ambiguous linking cases to human experts, whose feedback is then used to refine matching models, update similarity thresholds, or enrich training data. By focusing human effort on the most informative cases, active learning enables iterative improvement of linking quality while limiting manual effort. Over time, corrections provided by human operators can be exploited to improve automated linking strategies and reduce the frequency of similar errors in subsequent executions.

4.2.3 Step 3: Extension

The extension step enriches the input table by incorporating additional attributes retrieved from external data sources using the identifiers and normalized values produced in the previous steps. At this stage, new information is effectively introduced into the dataset, with the goal of supporting downstream analytical tasks that cannot be addressed using the original input alone.

Extension operations typically involve querying third-party services, knowledge graphs, or domain-specific datasets to retrieve attributes associated with the linked

identifiers. In the running example illustrated in Fig. 4.1, this includes the retrieval of population data at the state level and weather information for the specified location and date. More generally, extension can introduce a wide range of contextual, statistical, or descriptive attributes, depending on the analytical objectives and the available external sources.

Since extension relies entirely on the correctness of previously established links, errors introduced during the reconciliation step are propagated at this stage. Incorrect or imprecise identifiers may result in the retrieval of irrelevant, misleading, or inconsistent data, which can negatively affect downstream analysis. Therefore, extension represents a critical amplification point in the enrichment pipeline: small inaccuracies in reconciliation may systematically propagate through attribute retrieval and aggregation, producing structurally consistent but semantically incorrect data.

In addition to error propagation, extension introduces new sources of incompleteness and inconsistency. External data sources may differ in terms of temporal coverage, update frequency, granularity, and reliability, and may return missing, outdated, or partially conflicting values. These issues do not necessarily invalidate the enrichment process, but they require explicit treatment before the data can be reliably used for analysis.

For this reason, extension naturally leads to subsequent validation and refinement steps. Once additional attributes have been incorporated, the enriched dataset must be examined to identify missing values, outliers, and inconsistencies introduced or amplified during enrichment.

4.2.4 Step 4: Iterative Validation and Refinement

Following the extension step, the enriched dataset may exhibit missing values, inconsistencies, implausible values, or partially derived attributes that require further assessment before the data can be considered analytically reliable. Such issues may originate from incomplete external sources, temporal misalignment across datasets, service-level failures, ambiguity in entity linking, or error propagation from previous enrichment stages.

This step consolidates value imputation, inference, augmentation, and validation activities within a unified iterative refinement process. The objective is to detect potential quality issues, apply corrective or enrichment operations, and reassess the dataset until a satisfactory level of consistency and completeness is achieved.

Value imputation techniques are adopted to handle missing or incomplete data. Basic strategies include default assignments, aggregation at higher granularity levels, or statistically derived central tendencies, whereas more advanced approaches may exploit correlations among attributes, temporal dependencies, or learned predictive models. Similarly, outlier detection mechanisms can be applied to identify anomalous or implausible values, which may subsequently be removed, corrected, or replaced through imputation. These operations are typically automated but may generate borderline cases where corrective actions are not unequivocally justified.

Human-in-the-loop intervention plays a central role in this iterative refinement phase. Human supervision may be triggered when confidence scores fall below predefined thresholds, when inferred or augmented values have a significant impact on downstream analytical results, or when domain expertise is necessary to resolve semantic ambiguities. The intervention may include identifying anomalies, validating candidate corrections, revising linking decisions, or re-executing portions of the enrichment pipeline under updated constraints. For example, reconciliation for “Nuremberg” should be performed again once the correct region is specified to address inaccuracies in coordinates and weather data. Importantly, validated corrections can be formalized into explicit rules or parameter adjustments, thereby progressively improving the automated components of the process.

Through successive validation cycles, inconsistencies are progressively reduced, confidence scores are recalibrated, and derived attributes are consolidated.

The outcome of this step is a semantically enriched and internally consistent table that integrates reconciled identifiers, extended attributes, inferred values, and validated augmentations. The final result of the iterative validation and refinement process is illustrated in Table 4.2, which represents the consolidated and quality-assessed output of the enrichment pipeline.

4.3 Semantic Annotation of Tables

Semantic technologies enable the identification of a shared space of identifiers across heterogeneous datasets, thereby supporting data linking, integration, and enrichment. When dealing with tabular data, semantics can be associated both with the content of table cells and with the table schema. According to the “Semantic Web Challenge on Tabular Data to Knowledge Graph Matching” (SemTab) [13], the former corresponds to Cell Entity Annotation, while the latter includes Cell Type Annotation and Cell Property Annotation.

Cell Entity Annotation (CEA) is a cell-level annotation task, also referred to as instance-level annotation. Its objective is to associate textual values occurring in table cells (mentions) with entities in a target Knowledge Graph (KG)—for instance, Wikipedia entities for real-world objects—or, more generally, with unique identifiers drawn from reference systems such as geographical coordinates for location entities. The task can be formulated as an entity linking problem, whose goal is to disambiguate mentions and map them to the corresponding entities in the target KG or to columns with classes defined in a KG ontology, thus capturing the semantic type of the values contained in each column. Cell Property Annotation (CPA) is another schema-level task that identifies semantic relations between pairs of columns by mapping them to properties defined in the KG ontology.

Schema-level annotations are typically inferred from the results of cell-level annotation, for instance, by analyzing the frequency and distribution of types and properties associated with the entities identified during reconciliation [7, 14]. Consequently, the quality of schema-level annotations—and, more broadly, of the overall

table enrichment process—critically depends on the accuracy of cell entity linking algorithms [6, 8, 9].

Entity linking, therefore, represents a central and critical step in the semantic annotation and enrichment process. Ambiguity, lexical variations, incomplete contextual information, and schema heterogeneity make reconciliation decisions inherently uncertain and error-prone, with potential cascading effects on downstream annotation and data extension tasks.

From an algorithmic perspective, existing entity linking approaches can be broadly categorized into three main families: heuristic-based, machine-learning-based, and language-model-based methods. These families differ in terms of decision mechanisms, supervision requirements, interpretability, and scalability characteristics. They are not mutually exclusive and do not represent orthogonal design choices; rather, they capture dominant modeling paradigms that are frequently combined with collective inference techniques, constraint-driven validation, and human-in-the-loop refinement strategies in practical systems.

4.3.1 Heuristic-Based Approaches to Reconciliation

Heuristic-based entity linking methods rely on manually designed rules and similarity functions to match mentions to candidate entities. Typical techniques include string similarity measures (e.g., edit distance, token overlap, phonetic similarity), normalization and canonicalization rules, and the exploitation of structured constraints such as type compatibility, geographic containment, or functional dependencies. These approaches are generally transparent, interpretable, and computationally efficient, which makes them particularly suitable for large-scale settings and interactive data preparation tools. However, their effectiveness is highly dependent on domain-specific tuning and they tend to degrade in the presence of noisy inputs or highly ambiguous mentions. Representative systems include SILK, which expresses linking rules using a declarative language combining similarity metrics and aggregation operators [18], as well as early reconciliation services integrated into tools such as OpenRefine.⁶ More often, they are used in combination with feature-based machine learning algorithms [14].

4.3.2 Machine-Learning-Based Approaches to Reconciliation

Machine-learning-based approaches formulate entity linking as a classification or ranking problem, in which candidate entities are scored using learned models. Features typically capture lexical similarity, contextual coherence, structural properties

⁶ <https://openrefine.org/>

of the knowledge graph, and, when available, collective signals across multiple mentions. Both supervised and semi-supervised techniques have been proposed, ranging from probabilistic models inspired by record linkage theory to more recent neural architectures based on distributed representations.

These approaches generally achieve higher accuracy than purely heuristic methods, particularly in ambiguous cases, but require labeled training data and introduce additional complexity in model training, deployment, and maintenance. A well-known example is DBpedia Spotlight, which combines statistical disambiguation with contextual evidence [15]. A more recent feature-based approach is Alligator [1], which uses a neural network operating on local feature vectors to transform engineered features into a normalized matching score. In practice, this step re-ranks a set of candidate entities by considering signals such as textual similarity and schema-level properties. Another approach explores deep learning architectures specifically designed for tabular data. One notable example is TURL [10], which adapts BERT [11] to explicitly model tabular structure and is subsequently fine-tuned to perform downstream tasks, including entity disambiguation.

4.3.3 *Language-Model-Based Approaches to Reconciliation*

Recent approaches leverage large language models (LLMs) as a distinct paradigm for entity linking, exploiting their parametric knowledge and contextual reasoning capabilities. For instance, TableLlama [20] is based on the autoregressive LLaMA 2-7B model [17], which is fine-tuned using instruction tuning techniques [19] to perform table-related tasks. These models are typically trained and evaluated on benchmark datasets designed to support their general-purpose objectives, and they are reported to achieve state-of-the-art performance on the evaluation tasks.

In this setting, entity linking is often formulated as a prompt-based inference task, where mentions and their surrounding context are provided as input to the model, which then generates, ranks, or selects the most plausible entity identifiers. Such methods can be particularly effective in low-resource or zero-shot scenarios, as they are able to handle lexical variations and implicit contextual cues more robustly than traditional models.

At the same time, LLM-based approaches raise important concerns regarding reproducibility, explainability, controllability, and long-term stability, since their behavior may depend on prompt formulation, model versioning, and opaque internal representations. Moreover, scalability remains a key challenge when processing large tabular datasets, as commonly encountered in industrial data preparation pipelines. As a result, recent research increasingly explores hybrid solutions that combine LLM-based inference with symbolic constraints, candidate generation mechanisms, and validation steps.

Overall, the three families of approaches described above represent different trade-offs in terms of accuracy, interpretability, supervision requirements, and scalability.

For this reason, practical semantic annotation and data enrichment pipelines often adopt hybrid architectures that integrate heuristic filtering, learned ranking models, language-model-based components, and human-in-the-loop validation mechanisms.

4.4 Experimental Evaluation

In this section, we compare different families of methods for entity linking, focusing in particular on approaches based on machine learning algorithms and large language models. The objective of the evaluation is to assess the level of accuracy achievable by these approaches and to relate it to their operational sustainability in practical data enrichment workflows.⁷

To this end, we consider four tools that are representative of commonly adopted solutions in the literature: Alligator, TURL, TableLlama, and GPT-based approaches. Alligator represents feature-based machine-learning approaches relying on engineered features and ranking models; TURL exemplifies deep learning architectures specifically designed for tabular data; TableLlama represents recent approaches based on instruction-tuned large language models for table understanding; finally, GPT-based systems represent general-purpose language models used through prompt-based inference.

The first three tools are characterized by the availability of information about their training data and execution mechanisms. In contrast, models from the GPT family do not expose details about their internal architecture, training data, or optimization procedures. They therefore behave as black-box systems. At the same time, they represent some of the most powerful general-purpose large language models currently available, and can thus be reasonably expected to provide highly competitive performance.

By comparing these systems, our goal is not only to measure their performance in terms of entity linking accuracy but also to analyze the trade-offs they introduce with respect to computational requirements, reproducibility, and integration within scalable data enrichment pipelines.

4.4.1 Evaluation Datasets

Table 4.1 lists the general-purpose datasets used in the known pre-training phases of the considered approaches. For TableLlama, we use the model released by the authors on HuggingFace, while TURL and Alligator are trained on the same datasets used in their original papers.

For Alligator, the pre-training datasets are derived from the SemTab challenge and are primarily composed of tables synthetically generated by querying KGs. These tables can be large and closely resemble those used in business processes. However,

⁷ The content of this section is a revised version from [3].

Table 4.1 Lists of pre-training datasets

Approach	Dataset	Entities	Source	Entities domain
Alligator [1]	SemTab2021 - R2	47.4K	Graph queries	Cross
	SemTab2021 - R3	58.9K	Graph queries	
	SemTab2022 - R2 (2T)	994.9K	Graph queries/web tables	
	SemTab2020 - R4	667.2K	Graph queries	
	SemTab2019 - R3	390.4K	Graph queries	
	SemTab2019 - R1 (T2D)	8K	Graph queries/web tables	
	Total	2.16M		
TURL [10]	TURL WikiTable w/o WikiGS [12]	1.23M	Web Tables	Cross
TableLlama [20]	TableInstruct [20]	2.6M	Web Tables	Cross

Table 4.2 Statistics of the datasets used to fine-tune and evaluate models. † indicates datasets that have been sub-sampled so that they can be entirely processed by TableLlama within its 8192 context

Dataset	Split	Tables	Cols (min max $\bar{x}$)	Rows (min max $\bar{x}$)	Entities
SemTab2022 2T-red	Train†	91	1 8 4.86	5 369 98.61	14674
	Test†	26	1 8 4.65	7 264 74.58	4691
SemTab2022 HT-R1	Train	3691	2 5 2.56	4 8 5.68	26189
	Test	200	2 5 2.59	4 8 5.74	1406
SemTab2022 HT-R2	Train†	4344	2 5 2.56	4 8 5.57	20407
	Test	426	2 5 2.53	4 8 5.56	1829
SemTab2023 WikidataTables-R1	Train	9917	2 4 2.51	3 11 5.65	64542
	Test	500	2 4 2.46	3 11 6.95	4247
SemTab2021 BioDiv-red	Test†	11	1 26 17.45	26 100 58.90	1232
TURL-120k	Train	13061	1 43 5.43	1 624 12.98	120000
TURL-2k-red	Test	1295	1 14 1.03	6 257 32.95	1801

they generally lack the rich textual context typically associated with web tables, such as those found on Wikipedia. In contrast, generalistic LLM approaches have been trained primarily on web-based tables [4], which are relatively small, contain rich contextual information, and are crafted by humans for human use.

The term *pre-training* refers to the main training phase performed by the model authors, and we use it to distinguish this phase from the additional *fine-tuning* performed in our experiments. Thus, in these experimental settings, a model may first be pre-trained on one or more datasets, then further fine-tuned on different datasets, and finally evaluated on a third dataset. Table 4.2 reports the statistics of the datasets used for fine-tuning and testing the models.

To evaluate the models’ generalization capabilities, we consider the relationship between the data distribution of the datasets used for pre-training (see [3] for details)

Table 4.3 Test datasets categorisation in terms of In-Domain (IN), Out-Of-Domain (OOD), and Moderately-Out-Of-Domain (MOOD) w.r.t. the pre-trained datasets

Dataset	Source	Domain	Approaches		
			Alligator	TURL	TableLlama
SemTab2022 2T-red	Graph queries/web tables	Cross	IN	MOOD	MOOD
SemTab2022 HT-R1	Graph queries	Cross	IN	MOOD	MOOD
SemTab2022 HT-R2	Graph queries	Cross	IN	MOOD	MOOD
SemTab2023 WikidataTables-R1	Graph queries	Cross	IN	MOOD	MOOD
SemTab2021 BioDiv-red	Biodiversity Tables	Specific	OOD	OOD	OOD
TURL-120k	Web Tables	Cross	MOOD	IN	IN
TURL-2K-red	Web Tables	Cross	MOOD	IN	IN

and that of the datasets used for testing. When the pre-training and test datasets can be regarded as drawn from similar distributions (e.g., when both are constructed from graph queries over cross-domain data), we consider the evaluation to be *in-domain*. When the two distributions differ substantially (e.g., graph queries over cross-domain data versus biodiversity tables with domain-specific languages), we consider the evaluation to be *out-of-domain*. We further introduce a *moderately out-of-domain* setting for cases in which the two distributions are neither clearly similar nor clearly divergent, such as graph queries over cross-domain data versus web tables containing cross-domain data. Observe that GPT-4o models, as well as the backbone Llama model used in TableLlama, may have seen web tables in the test set during the language modeling pre-training phase (for TableLlama: before fine-tuning on tabular tasks). This potential data leakage cannot be controlled in our setting, as in prior experiments, and is orthogonal to the distribution-aware analysis. Table 4.3 reports the testing conditions for the different tools considered.

4.4.2 Evaluation Results

In this section, we evaluate and compare the considered approaches in terms of accuracy. In addition to predictive performance, we also analyse their computational efficiency along two dimensions: (i) execution time, measured as the average wall-clock time required to link a mention (excluding entity retrieval); and (ii) memory usage, measured as peak GPU memory for TURL and TableLlama, and as virtual memory size for Alligator.

All experiments were conducted on the following infrastructure. Testing and fine-tuning for TableLlama and TURL were performed on an Azure NC24ads_a100_v4

instance.⁸ Alligator was executed on a machine equipped with an Intel(R) Xeon(R) Gold 6238R CPU @ 2.20GHz with 64 cores and 96GB of RAM.

Table 4.4 reports the accuracy obtained by the evaluated algorithms on the considered datasets. Accuracy values are expressed as “ $Accuracy^{DataTypology}$ ”, where the superscript indicates the relationship between the evaluation dataset and the training data. In particular, ^{IN} denotes *in-domain* data, ^{MOOD} denotes *moderately out-of-domain* data, ^{OOD} denotes *out-of-domain* data, and ^{UNK} unknown data provenance. The superscript ^{MOOD → IN} indicates that the model has been fine-tuned in the MOOD setting, resulting in a transition of the evaluation condition from moderately out-of-domain to in-domain.

Overall, the results show that large language models achieve the highest accuracy across most datasets, with GPT-4o consistently ranking among the top-performing approaches. TableLlama also shows strong performance, confirming the effectiveness of LLM-based approaches specifically tuned for tabular tasks. In contrast, TURL performs well mainly when the evaluation data closely matches its training distribution, while the feature-based Alligator approach remains competitive in several in-domain settings. These results suggest that while LLM-based solutions currently provide the highest accuracy, traditional and feature-based approaches can still offer robust performance when the structure and domain of the data are well aligned with the training conditions.

To complement the accuracy analysis, we also compare the computational characteristics of the considered approaches. Figure 4.3 summarizes the average accuracy, execution time, and memory consumption across the evaluated datasets.

In this figure, the reported time corresponds to the elapsed time required to link a mention, excluding entity retrieval. The occupied memory corresponds to GPU memory for TURL and TableLlama, and to the “Virtual Memory Size” for Alligator. $\{Approach\}\text{-}FT$ denotes the respective approach fine-tuned in the MOOD setting.

Overall, the figure highlights a clear trade-off between predictive performance and computational cost. LLM-based approaches achieve the highest average accuracy but require significantly larger memory resources and longer execution times compared with the other methods. In contrast, Alligator exhibits very low computational requirements, running on CPU with limited memory usage, although with lower accuracy in several datasets. TURL represents an intermediate solution, showing moderate computational cost but also a stronger dependency on the similarity between training and evaluation data. These results suggest that the choice of approach should be guided not only by accuracy requirements but also by the available computational resources and the expected deployment scenario.

⁸ <https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/gpu-accelerated/nca100v4-series?tabs=sizebasic>.

Table 4.4 Accuracy of the evaluated algorithms. For each test dataset, the best result is highlighted in **bold**, while the second-best is *underlined*

Dataset	Alligator	TURL	TableLlama	GPT-4o	GPT-4o-mini	Alligator-FT	TURL-FT	TableLlama-FT
SemTab2022 2T-red	0.7156 ^{IN}	0.1343 ^{MOOD}	0.8243 ^{MOOD}	0.8783 UNK	0.7907 UNK	0.853 ^{IN}	0.3323 ^{MOOD} →IN	0.8400 ^{MOOD} →IN
SemTab2022 HT-R1	0.8890 ^{IN}	0.3997 ^{MOOD}	0.7873 ^{MOOD}	0.8542 UNK	0.8087 UNK	0.8165 ^{IN}	0.7454 ^{MOOD} →IN	0.8001 ^{MOOD} →IN
SemTab2022 HT-R2	0.8218 ^{IN}	0.2763 ^{MOOD}	0.6619 ^{MOOD}	0.7516 UNK	0.6565 UNK	0.7212 ^{IN}	0.6018 ^{MOOD} →IN	0.6778 ^{MOOD} →IN
SemTab2023 WikidataTables-R1	0.8248 ^{IN}	0.3391 ^{MOOD}	0.7426 ^{MOOD}	0.8034 UNK	0.7554 UNK	0.8086 ^{IN}	0.7061 ^{MOOD} →IN	0.7530 ^{MOOD} →IN
SemTab2021 BioDiv-red	0.5674 ^{OOD}	0.8109 ^{OOD}	0.9513 ^{OOD}	0.9748 UNK	0.9497 UNK	0.8547 ^{OOD}	0.6347 ^{OOD}	0.9610 ^{OOD}
TURL-2K-red	0.6017 ^{MOOD}	0.7118 ^{IN}	0.9051 ^{IN}	0.9095 UNK	0.8629 UNK	0.7456 ^{MOOD} →IN	0.5347 ^{IN}	0.9045 ^{IN}

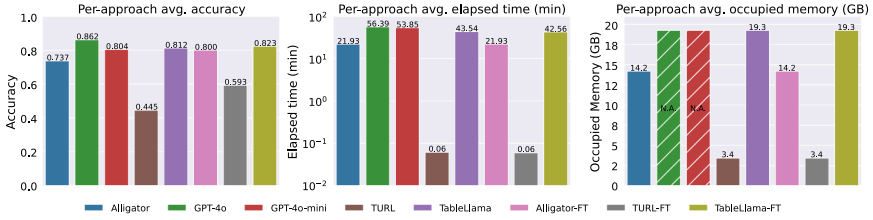


Fig. 4.3 Average accuracy, elapsed time (min), and memory usage (GB)

4.4.3 Sustainability

Beyond predictive performance, data enrichment approaches should also be evaluated in terms of their sustainability. In practice, the preparation and enrichment of tabular data may require substantial computational resources, human effort, and interaction with external services. These aspects become particularly relevant when enrichment pipelines operate at scale or must be repeatedly executed on evolving datasets.

Sustainability in data enrichment should account for the balance between improvements in data quality and the resources required to achieve them. Enrichment workflows typically involve two types of activities: *assessment*, aimed at identifying missing or erroneous values in the input data, and *improvement*, aimed at completing or correcting them.

Assessment costs quantify the effort needed to analyse the dataset and prepare enrichment actions, including understanding external sources and configuring matching processes. *Improvement costs* measure the resources required to execute the enrichment itself, such as performing matching and retrieving additional information. Since not all enrichment actions are fully effective, the portion of resources spent on unsuccessful operations or remaining errors is captured by *waste costs*.

The key characteristic of reconciliation-based enrichment pipelines—namely, linking table values to shared identifiers and using them to retrieve only the relevant information from external sources—reduces the need for extensive data discovery, schema understanding, and pairwise matching across heterogeneous datasets, which are typically the most resource-intensive phases of traditional data integration.

The comparison reported in Table 4.5 illustrates these trade-offs for the enrichment task described in Sect. 4.2. The table reports the application of the entity linking approaches discussed in Sect. 4.4, together with the specialized HERE API service, to link city names in the digital marketing dataset with geographic coordinates. For comparison, we also evaluated LLMs in a pairwise matching setting, reporting the corresponding results using the F1⁹ measure.

The evaluation follows the sustainability model introduced in [16], which assesses data preparation strategies by jointly considering their effectiveness and the resources required to achieve the improvement. Accordingly, the table reports the parameters

⁹ F1 score is the harmonic mean of precision and recall, providing a single measure of a method's accuracy in binary or multi-class classification tasks.

Table 4.5 Comparison of Methods, Assessment and Improvement Costs, and Waste. *Source* [16]

Method type	Matching method	LLM	Meas.	$p_{j,k}$	$c_{j,k}$ x1000	Ass. costs	Impr. costs	Total costs	Waste
Pairwise	GPT4o	✓	F1	0.8774	12.125	120	3891	4011	477
Pairwise	GPT4o-mini	✓	F1	0.8084	0.728	120	233	353	44
Reconc.	HERE	✗	Acc	0.9980	0.300	0	96	96	0
Reconc.	Alligator + Geo	✗	Acc	0.8000	0.266	0	85	85	17
Reconc.	TableLlama + Geo	✓	Acc	0.8230	1.367	0	438	438	77
Reconc.	GPT4o + Geo	✓	Acc	0.8620	5.200	0	1668	1668	230
Reconc.	GPT4o-mini + Geo	✓	Acc	0.8040	0.312	0	100	100	19

$p_{j,k}$ and $c_{j,k}$, together with the derived cost indicators. The parameter $p_{j,k}$ denotes the effectiveness of method k for quality dimension j , i.e., the proportion of erroneous or incomplete values that are successfully corrected or enriched by the method. The parameter $c_{j,k}$ represents the unit cost associated with performing the corresponding improvement action, capturing computational resources, external service usage, and human effort when applicable. The values of $c_{j,k}$ are based on assumed standard costs and should be interpreted for comparative purposes rather than as absolute measures.

As expected, the results indicate that reconciliation strategies differ significantly in their sustainability profiles. Lightweight methods and specialized APIs typically require minimal computational resources, offering a favourable trade-off between effectiveness and cost. In contrast, large language models achieve very high matching performance but impose substantially higher computational demands, which may limit their feasibility in large-scale or resource-constrained settings. Overall, reconciliation incurs lower costs than traditional integration methods such as pairwise matching.

These findings suggest that reconciliation-based enrichment represents a promising approach for sustainable data preparation pipelines. By leveraging shared identifiers and interoperable knowledge sources, these methods reduce redundant processing and enable more efficient reuse of existing information infrastructures.

4.5 Conclusions

Reconciliation-based data enrichment is a data preparation technique that does not require direct control or prior integration of the third-party datasets to be used. When such datasets expose access interfaces—such as APIs, SPARQL, or SQL endpoints—it is sufficient to reconcile the source data, i.e., link them to the same identifier system adopted by those datasets, in order to retrieve the information of interest.

The main limitation of reconciliation and linking techniques lies in the intrinsic ambiguity of the natural language expressions used in tabular data. For this reason, a variety of approaches—ranging from heuristic methods to machine learning and

language models—have been developed. As discussed in this chapter, these techniques achieve good levels of accuracy while exhibiting different cost–performance profiles.

The evaluation highlights several trade-offs among current approaches to entity linking in tabular data. Extra-large API-based LLMs show strong generalization capabilities and consistently achieve the highest accuracy across datasets, albeit at substantially higher computational and economic costs. Medium-sized API models such as GPT-4o-mini provide an effective compromise, reaching performance levels close to larger models while maintaining a more favourable accuracy–cost balance. Non-LLM approaches remain relevant in specific scenarios: when the structure and domain of the data align with the assumptions of the method, they can offer a competitive trade-off between effectiveness and resource consumption.

Recent reasoning-oriented models suggest that higher accuracy may be achievable with fewer parameters; however, current implementations suffer from very large inference times, which may limit their applicability in practical pipelines. Finally, the experiments highlight the importance of task-specific training: matching the performance of a fine-tuned 7B model may require a general-purpose model of roughly 70B parameters when no task-specific training is available.

Acknowledgements This research was partially funded by the European Innovation Actions projects *enRichMyData* (HE 101070284) and *DataPACT* (HE 101189771).

References

1. Avogadro, R., Ciavotta, M., De Paoli, F., Palmonari, M., Roman, D.: Estimating link confidence for human-in-the-loop table annotation. In: 2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 142–149 (2023)
2. Batini, C., Scannapieco, M.: Data and Information Quality - Dimensions, Principles and Techniques. Data-Centric Systems and Applications. Springer, Cham (2016). <https://doi.org/10.1007/978-3-319-24106-7>. <https://doi.org/10.1007/978-3-319-24106-7>
3. Belotti, F., Cremaschi, M., Dadda, F., Avogadro, R., Palmonari, M.: How good are LLMs in disambiguating entities in tabular data? A comprehensive study. Data & Knowledge Engineering, p. 102596 (2026). <https://doi.org/10.1016/j.datak.2026.102596>. <https://www.sciencedirect.com/science/article/pii/S0169023X26000431>
4. Bhagavatula, C.S., Noraset, T., Downey, D.: Tabel: entity linking in web tables. In: International Semantic Web Conference, pp. 425–441. Springer (2015)
5. Bizer, C.: The emerging web of linked data. IEEE Intell. Syst. **24**(5), 87–92 (2009). <https://doi.org/10.1109/MIS.2009.102>
6. Ciavotta, M., Cutrona, V., De Paoli, F., Nikolov, N., Palmonari, M., Roman, D.: Supporting semantic data enrichment at scale. In: Technologies and Applications for Big Data Value, pp. 19–39. Springer (2022)
7. Cremaschi, M., Avogadro, R., Chierigato, D.: s-elBat: a semantic interpretation approach for messy tables. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab@ISWC), CEUR-WS. org, pp. 59–71. (2022)
8. Cutrona, V., De Paoli, F., Košmerlj, A., Nikolov, N., Palmonari, M., Perales, F., Roman, D.: Semantically-enabled optimization of digital marketing campaigns. In: Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I.F., Hogan, A., Song, J., Lefrançois, M., Gandon, F. (eds.)

- The Semantic Web - ISWC 2019—18th International Semantic Web Conference Proceedings, Part II, pp. 345–362. Springer, Cham (2019)
9. De Paoli, F., Ciavotta, M., Avogadro, R., Hristov, E., Borukova, M., Petrova-Antonova, D., Krasteva, I.: An interactive approach to semantic enrichment with geospatial data. *Data & Knowl. Eng.* **153**, 102341 (2024)
 10. Deng, X., Sun, H., Lees, A., Wu, Y., Yu, C.: Turl: table understanding through representation learning. *ACM SIGMOD Rec.* **51**(1), 33–40 (2022)
 11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (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)*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019)
 12. Efthymiou, V., Hassanzadeh, O., Rodriguez-Muro, M., Christophides, V.: Matching web tables with knowledge base entities: from entity lookups to entity embeddings. In: *The Semantic Web – ISWC 2017: 16th International Semantic Web Conference*. Vienna, Austria, 21–25 Oct 2017, *Proceedings, Part I*, pp. 260–277. Springer-Verlag, Berlin, Heidelberg (2017)
 13. Jiménez-Ruiz, E., Hassanzadeh, O., Efthymiou, V., Chen, J., Srinivas, K.: Semtab 2019: resources to benchmark tabular data to knowledge graph matching systems. In: *European Semantic Web Conference*, Springer (2020)
 14. Liu, J., Chabot, Y., Troncy, R., Huynh, V.P., Labbé, T., Monnin, P.: From tabular data to knowledge graphs: a survey of semantic table interpretation tasks and methods. *J. Web Semant.* **76**, 100761 (2023)
 15. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia spotlight: shedding light on the web of documents. In: *Proceedings of the 7th International Conference on Semantic Systems, I-Semantics '11*, p. 1–8. Association for Computing Machinery, New York, NY, USA (2011). <https://doi.org/10.1145/2063518.2063519>. <https://doi.org/10.1145/2063518.2063519>
 16. Pernici, B., Cappiello, C., Bono, C., Sancricca, C., Catarci, T., Angelini, M., Filosa, M., Palmonari, M., Paoli, F.D., Bergamaschi, S., Simonini, G., Mozzillo, A., Zecchini, L.: Sustainable quality in data preparation. *ACM J. Data Inf. Qual.* **17**(4:23), 1–33 (2025)
 17. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardaş, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023). <https://arxiv.org/abs/2307.09288>
 18. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and maintaining links on the web of data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) *The Semantic Web - ISWC 2009*, pp. 650–665. Springer, Berlin Heidelberg, Berlin, Heidelberg (2009)
 19. Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. In: *International Conference on Learning Representations*, (2022)
 20. Zhang, T., Yue, X., Li, Y., Sun, H.: TableLlama: towards open large generalist models for tables. In: Duh, K., Gómez-Adorno, H., Bethard, S. (eds.) *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL 2024, Mexico City, Mexico, June 16–21, 2024, pp. 6024–6044. Association for Computational Linguistics (2024)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

