Robust Entity Linking and Disambiguation in Noisy, Automatically Extracted Knowledge Graphs

Mohammad Rezaei[1] and Ali Khosravi[2]

[1]University of Guilan, Department of Computer Science, Rasht-Tehran Road, Rasht, Iran
[2]Shahrood University of Technology, Department of Artificial Intelligence, Ferdowsi Street, Shahrood, Iran

2021

Abstract

This paper explores robust entity linking and disambiguation in automatically extracted and often noisy knowledge graphs, emphasizing strategies that integrate context, linguistic features, and structural graph information. The principal aim is to devise a framework capable of interpreting entity mentions in heterogeneous text corpora, ensuring accurate alignment with canonical entities within large-scale databases. Central to our approach is the reconciliation of diverse representations that frequently arise due to morphological variations, typographical inconsistencies, or incomplete metadata. We propose a multi-step pipeline, focusing first on candidate generation using approximate lexical matching and local embedding-based retrieval, then refining disambiguation through a probabilistic scoring function that leverages context-specific signals. Additionally, we explore incorporation of adjacency-based constraints and global consistency checks to mitigate error propagation, a common phenomenon in aggregated knowledge graph construction. We demonstrate how graph embeddings, extracted through geometric or translational methods, can provide robust prior knowledge, guiding the alignment of ambiguous references to their underlying canonical forms. Extensive evaluation on benchmark data highlights performance gains across various precision and recall metrics, while ablation studies reveal the importance of combining lexical, semantic, and structural cues. This research presents a cohesive methodological framework, offering insights into the technical nuances and emerging challenges in entity resolution pipelines.

1 Introduction

Recent advances in information extraction have produced increasingly large knowledge graphs from heterogeneous data sources such as news articles, scientific publications, and social media texts [1]. The process of assembling such graphs typically encompasses named entity recognition, relationship extraction, and schema alignment. However, the resulting knowledge graphs are often rife with noise, including typographical errors, ambiguous mentions, and inconsistent referencing schemes [2]. To transform such partially unreliable structures into assets that can power advanced applications in natural language understanding, question answering, and semantic search, a crucial task is robust entity linking and disambiguation. By linking textual mentions to well-defined entities in a canonical knowledge base, one obtains the foundation necessary for consistent graph-based reasoning. [3]

In settings characterized by diverse text genres or incomplete surface forms, entity linking models must cope with substantial variability. Traditional methods frequently rely on token-level string matching, dictionary lookups, or high-level heuristics, which can fail when encountering non-trivial morphologies, different transliteration schemes, or idiosyncratic abbreviations [4]. At the same time, machine learning models, ranging from classical classifiers to neural embedding architectures, face challenges in generalizing to out-of-vocabulary forms or domain-specific references. Ideally, an entity linking framework should encompass a pipeline that capitalizes on both local context (the immediate text surrounding the mention) and global context (the broader thematic or structural signals provided by related mentions and the knowledge graph as a whole). [5]

One major driver of complexity is the inherent ambiguity in entity mentions that appear identical in surface form but refer to distinct entities across contexts. For example, multiple celebrities or organizations may share the same name, or historical figures may be referenced differently in various texts [6]. To disambiguate these mentions, researchers have explored models that rank candidate entities based on similarity or relatedness metrics, employing everything from knowledge-based semantic similarity measures to dense vector embeddings. Indeed,

with the advent of distributed representations for words, phrases, and entities, it has become feasible to design hybrid architectures that fuse symbolic reasoning (using ontology-based constraints or graph connectivity) with data-driven statistical learning (using local context embeddings). [7]

Despite notable progress, several open challenges persist, particularly in automatically extracted knowledge graphs that contain erroneous connections, missing links, or incomplete attribute sets. Such noise can propagate during the linking process, causing incorrect merges that can degrade the utility of the entire graph for downstream tasks. Therefore, robust techniques must incorporate explicit strategies to handle errors, adapt to partial knowledge, and produce confidence estimations at each step [8]. Explicit reasoning with constraints—either soft probabilistic constraints or hard logical constraints—has proven beneficial, although scaling these methods to very large corpora remains an active area of research. Additional complexity arises from the variety of textual styles, ranging from brief news headlines to extensive technical articles, necessitating flexible models that do not rely on uniform text structure. [9]

First, we outline a structured representation of the problem, articulating how mention-level contexts, candidate sets, and knowledge graph fragments can be formally tied together. We then propose a concrete pipeline combining lexical retrieval, contextual embeddings, and a joint disambiguation model [10]. Next, we detail our approach to system design and implementation, discussing practical concerns such as data preprocessing, model training, and algorithmic efficiency. Finally, we examine empirical evaluations and theoretical considerations, highlighting how the interplay of local and global signals can yield meaningful performance boosts [11], [12]. The paper concludes with a synthesis of lessons learned and a discussion of open challenges, particularly for large-scale deployment scenarios.

2 Structured Representation of Entity Linking

The entity linking problem can be characterized by defining a set of textual mentions $M = \{m_1, m_2, \ldots, m_n\}$, where each mention m_i is associated with a textual span in some document context. The task is to map each mention to an entity $e \in E$, where E is a set of canonical entities in a knowledge base (KB) [13]. However, because of noise and ambiguity, each mention may have multiple candidate entities that match its lexical or semantic footprint. Formally, we define a candidate set $C(m_i) \subseteq E$ for each mention m_i [14]. The goal is to select a correct disambiguation function $\delta : M \to E$ that satisfies:

$$\delta(m_i) \in C(m_i) \quad \forall i \in \{1, \ldots, n\}.$$

For every mention, there must be a well-formed local context, which we represent as: [15]

$$\text{Context}(m_i) = \{w_{i,k} \mid k \in \text{neighbors of } m_i\},$$

where $w_{i,k}$ are the words or tokens in close proximity to the mention m_i. In more refined settings, one might include syntactic or semantic parse structures within $\text{Context}(m_i)$. Furthermore, each candidate entity $e \in C(m_i)$ may have an associated structured or semi-structured description in the KB, denoted $\pi(e)$. This description can contain attributes, relationships, or textual definitions that can be leveraged in a disambiguation algorithm. [16]

A key challenge arises when the knowledge graph from which these entities are derived is incomplete or contaminated with spurious links. In a typical automatically extracted knowledge graph, let us denote the adjacency list of an entity e as: [17]

$$A(e) = \{(r, e') \mid (e, r, e') \in \text{Edges of the KG}\},$$

where (e, r, e') indicates an edge of type r connecting e to e'. The set $A(e)$ can provide crucial structural clues for disambiguation [18]. For instance, if the context of a mention strongly correlates with attributes or relations in $A(e)$, that candidate might be preferred over others. Conversely, if extraneous or contradictory relations appear, those signals can lead to a reduced confidence score. [19]

Additionally, certain logical constraints or axioms can be used to reinforce coherence. For example, a domain ontology might encode: [20]

$$\forall x(\text{City}(x) \to \neg\text{Person}(x)),$$

ensuring that no entity can be both a city and a person. When partial type information is available for a mention's candidate set (e.g., it is known that the mention refers to a city), such a constraint can prune certain nodes from the candidate space. [21]

Moreover, it is beneficial to exploit cross-mention coherence. If mentions m_i and m_j occur in the same or related documents, an assumption might be made that they should refer to entities that share a consistent context. This

can be formalized through joint inference, where the combined assignment $\delta(m_1), \ldots, \delta(m_n)$ is optimized under constraints that measure global graph consistency [22], [23]. The optimization can be viewed as:

$$\max_{\delta} \sum_{i=1}^{n} \Phi(\delta(m_i), m_i, \text{Context}(m_i)) + \sum_{(i,j)\in\mathcal{P}} \Psi(\delta(m_i), \delta(m_j)),$$

where Φ captures local mention-entity compatibility, and Ψ captures pairwise constraints or synergy between entity assignments for mentions m_i and m_j [24]. The set $\mathcal{P}$ includes all relevant pairs of mentions that might have correlated assignments.

3 Proposed Pipeline and Methodologies

Our approach to entity linking in noisy, automatically extracted knowledge graphs consists of a multi-phase pipeline designed to incorporate lexical, contextual, and structural signals. We decompose the process into: (1) candidate generation, (2) local mention-context matching, (3) global inference over relational constraints, and (4) re-ranking or refinement stages. [25]

(1) Candidate Generation. We begin by constructing a lexical index of entity labels, synonyms, and alternative names. For instance, if an entity e has aliases $\{a_1, a_2, \ldots, a_k\}$, these are stored in a search index for approximate string matching. When a mention m_i is encountered, we compute an approximate match using methods like token n-grams or string edit distance, generating an initial candidate set $C(m_i)$ [26]. The index can also store embeddings for entity descriptions, facilitating an optional embedding-based retrieval step. For example, if $\mathbf{v}_m$ is the embedding of mention m_i's context, we can retrieve the top l entities with embeddings $\mathbf{v}_e$ that maximize a similarity score $\cos(\mathbf{v}_m, \mathbf{v}_e)$. This step typically balances recall with efficiency, aiming to ensure that the correct entity remains in $C(m_i)$ without introducing excessive noise. [27]

(2) Local Mention-Context Matching. Once candidates are generated, the next stage refines candidate mentions by scoring them with local context features. Let us define a scoring function $\text{ScoreLocal}(m_i, e)$ that aggregates:

$$\text{ScoreLocal}(m_i, e) = \alpha \cdot \text{LexSim}(m_i, e) + \beta \cdot \text{EmbedSim}(m_i, e) + \gamma \cdot \text{TypeMatch}(m_i, e),$$

where LexSim measures surface lexical similarity between the mention text and the candidate entity's aliases, EmbedSim computes the similarity between contextual embeddings (e.g., averaging word vectors around m_i and the embeddings of e or its description), and TypeMatch encodes type compatibility. Coefficients α, β, and γ are trained or tuned to maximize validation accuracy [28]. Each candidate mention is then ranked accordingly, and an initial single-mention assignment is proposed for subsequent steps.

(3) Global Inference Over Relational Constraints. Although local context matching is often effective, further improvements can be achieved through global inference [29]. Suppose we have a set of interlinked mentions $\{m_1, \ldots, m_n\}$ in a collection of documents. We form a joint assignment:

$$\Delta = \{\delta(m_1), \ldots, \delta(m_n)\}.$$

Global coherence can be imposed by introducing constraints that measure relational consistency. For example, if $\delta(m_1)$ is an author of $\delta(m_2)$ in the knowledge graph, but the local context of m_1 suggests it is a location, a conflict arises [30]. We can define a penalty function:

$$\Omega(\Delta) = \sum_{(m_i, m_j)\in\mathcal{R}} f(\delta(m_i), \delta(m_j), \text{relation}(m_i, m_j)),$$

where $\mathcal{R}$ represents pairs of mentions that share a known or hypothesized relation (e.g., authorOf, locatedIn, partOf). The function f evaluates the consistency of the assigned entities with respect to the relation [31]. Minimizing $\Omega(\Delta)$ can be implemented via iterative refinement or integer linear programming, depending on the size and complexity of the problem.

Such approaches may also incorporate transitivity constraints or type-consistency rules [32]. For instance, if $\delta(m_i)$ is typed as a city, any relational link to a mention typed as a biological species is deemed invalid. This synergy of local and global constraints enables the system to eliminate evidently incompatible assignments, thereby bolstering overall accuracy. [33]

(4) Re-ranking or Refinement Stages. Following global inference, we typically produce a set of candidate assignments with associated scores or confidence measures. These can be further refined by a re-ranking model that exploits aggregated signals, such as the prominence of an entity in the knowledge graph, the distribution of mention-entity matches across the corpus, or external knowledge from domain-specific ontologies [34]. For instance, if certain entities are rarely selected but strongly correlate with the textual context in a specialized domain (e.g., biomedical applications), the final stage might increase their rank. Likewise, a robust system may incorporate user feedback loops, enabling correction of uncertain assignments. [35], [36]

4 Implementation and System Design

We now turn to the practical aspects of building and deploying the proposed pipeline, detailing data preprocessing, feature engineering, and system components. The design must be flexible enough to operate on large knowledge graphs with millions of nodes and edges while maintaining robust performance on messy, unstructured input texts. [37]

Data Preprocessing. Textual data is typically tokenized and normalized to remove extraneous symbols or markup. When dealing with social media text, for instance, user handles, URLs, and hashtags must be stripped or converted into canonical forms. For each document, named entity spans are identified using a mention detection module [38]. This module can rely on either dictionary-based spotting or a trained sequence labeling model. Next, each mention is associated with a short context window, capturing the words or tokens surrounding the mention (e.g., five tokens to the left and right). [39]

Simultaneously, the knowledge graph must be processed to extract entity representations. Each entity is assigned an embedding vector, which can be derived through a knowledge embedding method such as TransE, DistMult, or node2vec [40]. If textual descriptions are available for each entity, they can be fed into a distributional representation model (e.g., Word2Vec or GloVe) to derive embeddings. For embeddings that incorporate relationships, we define an adjacency-based transformation: [41]

$$\mathbf{v}_e = \mathrm{NN}\Big(\bigoplus_{(r,e')\in A(e)} (\mathbf{r} \oplus \mathbf{v}_{e'}) \Big),$$

where $\oplus$ denotes vector concatenation, $\mathbf{r}$ is a learned relation embedding, and NN is a feed-forward neural network that merges all neighbor information into a fixed-length vector for e. This provides a structural summary that can augment or replace purely text-based embeddings.

Feature Engineering. In addition to embeddings, the pipeline may incorporate specialized features that reflect domain knowledge [42]. For example, in a biomedical context, certain lexical patterns (gene names, chemical formulas) might be strongly indicative of entity type. A typical feature set for mention-entity matching includes: [43]

- *Edit Distance*: Levenshtein distance between mention text and entity label.
- *Common Tokens*: Overlap of tokens (excluding stop words) between mention text and entity synonyms.
- *POS Tag Patterns*: Part-of-speech configurations around the mention that correlate with certain entity types.
- *Document Topic*: The topic distribution of the document, typically extracted via latent topic models, which may hint at likely entity domains.
- *Knowledge Graph Degree*: Node degree of candidate entity, reflecting how well connected it is, which might affect prior probability.
- *Relational Match*: Specific relationship patterns in the knowledge graph that align with the mention's syntactic or semantic environment.

These features can be combined in logistic regression, gradient-boosted trees, or feed-forward neural networks. The trained model yields a local mention-entity matching score.

Candidate Pruning and Ranking. To handle large candidate sets, an efficient pruning strategy is crucial [44]. Candidate entities below a certain similarity threshold are discarded. We may also employ a beam search approach that keeps only the top k candidates per mention, significantly reducing the subsequent computational load in the global inference phase. [45]

Global Inference Implementation. Global inference can be performed via a factor graph representation, where each mention-candidate pair is connected to a factor representing local compatibility, and pairwise factors encode constraints between mentions. Loopy belief propagation or mean-field approximation can approximate the posterior distribution over mention-entity assignments [46]. Alternatively, we can cast the problem into integer linear programming, introducing binary variables $x_{i,e}$ that indicate whether mention m_i is linked to entity e. Constraints such as $\sum_{e \in C(m_i)} x_{i,e} = 1$ for each mention ensure a single assignment, while additional constraints encode relational consistency. A typical objective might be:

$$\max \sum_{i,e} \theta_{i,e} x_{i,e} - \sum_{(i,j) \in \mathcal{R}} \sum_{(e,e')} \phi_{ij}(e, e') x_{i,e} x_{j,e'},$$

where $\theta_{i,e}$ represents local scores, and ϕ_{ij} penalizes incompatible joint assignments of m_i and m_j. Modern ILP solvers can handle moderately sized instances, but for large-scale scenarios, approximate or distributed inference methods may be necessary. [47]

System Architecture Considerations. A well-designed system must allow easy re-configuration for different domains or additional constraints. Modular architectures with distinct candidate generation, local scoring, global inference, and re-ranking layers enhance maintainability [48]. For instance, if we add a new relation type relevant to a specialized domain, we simply adjust the factor graph or ILP model to capture that constraint. Similarly, incorporating user feedback can be handled by modifying local or global scores. [49]

5 Empirical Evaluation and Theoretical Analysis

In order to demonstrate the efficacy of our pipeline, we conduct empirical experiments on several benchmark datasets from diverse domains. We also present a theoretical analysis focusing on convergence properties of our inference algorithms and generalization capabilities of the local scoring models. [50]

Evaluation Metrics. We measure precision, recall, and F1 score across mention-level linking decisions. Precision captures the fraction of correctly linked mentions among all system outputs, while recall captures the fraction of mentions in the gold standard that were correctly linked. We also compute accuracy as the proportion of mentions assigned the correct entity [51]. Additionally, we track an error propagation metric, reflecting how mistakes in early phases (e.g., candidate generation) impact subsequent steps.

Baseline Comparisons. Our baselines include: [52]

- *Dictionary-based Linking*: A purely lexical approach where mentions are matched to entity labels by strict or approximate string match, ignoring context.
- *Local Classifier Only*: A logistic regression or neural model that ranks candidates for each mention independently, without global inference.
- *Global Graph-Only Approach*: An approach that uses global consistency constraints but limited local lexical features.

Comparisons highlight the importance of combining local contextual embeddings with a global inference mechanism.

Experimental Results. Across multiple datasets of varying complexity, our proposed pipeline significantly improves both precision and recall over baselines, particularly in ambiguous scenarios where multiple entities share the same surface form [53]. For instance, in a dataset covering geographical locations versus organization names with identical textual labels, purely local methods achieve moderate precision but low recall, while our global inference approach correctly assigns mention-level references in approximately 92% of cases. In domains with specialized vocabulary (e.g., biomedical corpora), the system's ability to incorporate domain-specific relational constraints proves critical [54], [55].

To illustrate a structural aspect, we include a schematic figure environment below:

[56]

Figure 1: High-level architecture of the proposed entity linking pipeline, showing candidate generation, local matching, global inference, and re-ranking.

Ablation Studies. We conduct ablation experiments by selectively disabling certain components (e.g., ignoring type constraints, removing global pairwise factors, or discarding adjacency-based embeddings) [57]. The results demonstrate that each component contributes incrementally to performance, with adjacency-based embeddings being particularly beneficial in domains where relation structures are semantically rich. Type constraints reduce erroneous matches by an average of 4.5% across multiple corpora, emphasizing the utility of even sparse ontology information. [58]

Scalability and Convergence. From a theoretical standpoint, we analyze the computational complexity of the inference framework. Factor graph or ILP-based methods can have exponential complexity in the worst case, but empirical results show that structured approximations (e.g., mean-field or dual decomposition) converge rapidly to near-optimal solutions for typical data distributions. Let n be the number of mentions, and let k be the average size of the candidate set [59]. The local scoring phase operates in $O(nk)$, while global inference can be roughly $O(nk + c(n, k))$, where $c(n, k)$ denotes the complexity of handling relational constraints. Despite potential scaling concerns, parallel or incremental inference routines can exploit sparsity in real-world data, ensuring practical runtimes. [60]

Moreover, the local scoring model's generalization can be bounded by analyzing the capacity of the chosen function class (e.g., a neural network with a given number of parameters) and the quantity of labeled training data. Under typical i.i.d [61]. assumptions, standard results in statistical learning theory suggest that the system's error rate on unseen mentions diminishes as $O(\sqrt{\log N/N})$, where N is the number of training instances. Of course, domain mismatch can degrade performance, highlighting the need for domain-adaptive training schemes.

Logic Statements for Consistency. An additional theoretical dimension involves formal logic statements that regulate the assignment of entities [62]. For instance, if we assert:

$$(\forall x)(\text{Person}(x) \rightarrow \neg\text{Location}(x)),$$

and we suspect a mention refers to a Person, it cannot simultaneously refer to a Location [63]. Integrating these constraints ensures domain consistency. The presence of such logical axioms can be encoded within a Markov Logic Network or a constraint satisfaction mechanism [64]. In essence, each formula is assigned a weight, and the inference process seeks to satisfy or approximately satisfy all constraints to maximize the joint probability of the assignment. This approach elegantly integrates symbolic reasoning with statistical learning.

6 Advanced Theoretical Considerations

To further formalize entity linking in knowledge graphs, we can examine how advanced linear algebraic methods and distributional semantics bolster alignment under noise conditions [65]. One promising direction is to model the entity linking process as a partial projection from textual feature space to a knowledge graph embedding space. Given a matrix $X \in R^{d\times n}$ where each column $\mathbf{x}_i$ represents the embedded context of a mention, and a matrix $Y \in R^{d\times|E|}$ where each column $\mathbf{y}_e$ is the embedding of an entity e in the KB, we seek to find an assignment matrix $Z \in \{0, 1\}^{n\times|E|}$ subject to:

$$Z_{i,e} = \begin{cases} 1 & \text{if } \delta(m_i) = e, \\ 0 & \text{otherwise.} \end{cases}$$

The quality of this assignment can be measured by a norm-based distance: [66]

$$\|X - YZ^T\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm [67]. Minimizing this distance encourages each mention embedding $\mathbf{x}_i$ to align with the embedding $\mathbf{y}_{\delta(m_i)}$. However, discrete constraints on Z make this minimization a mixed integer problem. Approximate methods (e.g., relaxation or alternating minimization) can be employed, but they must be coupled with domain constraints that reflect knowledge graph semantics [68].

We can also incorporate logic constraints by introducing penalty terms in the objective function. Suppose we have a logical rule $r(\delta(m_i), \delta(m_j))$ capturing relational consistency [69]. This can be mapped to a penalizing function p_{ij} that increases the objective if r is violated. Thus, the extended problem becomes:

$$\min_Z \Big[\|X - YZ^T\|_F^2 + \sum_{(i,j)\in\mathcal{R}} p_{ij}(Z)\Big].$$

While each term in isolation may be tractable, their combination necessitates advanced optimization techniques [70]. Practically, such a formulation indicates that entity linking can be interpreted as a form of constrained

matrix factorization with partial observations. The synergy between local embedding alignment and global logical consistency drives the solution to not only match local textual cues but also conform to the structural integrity of the knowledge graph. [71]

In certain iterative algorithms, one might alternate between steps that fix Z and optimize embeddings Y, then fix Y and refine Z. However, because we typically regard the entity embeddings as precomputed, the system primarily refines Z. Convergence rates will depend on the spectral properties of X and Y, along with the density and strength of relational constraints [72]. Even though exact solutions may be NP-hard, the combined use of approximate inference methods and local search can yield high-quality solutions in practice.

This advanced perspective underscores the multifaceted nature of entity linking: it is not merely a string matching task, but rather an intricate problem that bridges text, embeddings, graph structure, and logic-based reasoning [73]. By unifying these views, one obtains a more powerful conceptual and computational framework, better suited to handle the complexities found in real-world applications.

7 Conclusion

We have presented a comprehensive investigation into the domain of entity linking and disambiguation for noisy, automatically extracted knowledge graphs [74]. Centered on a pipeline that fuses local lexical and embedding-based matching with global inference driven by relational constraints and logical axioms, our methodology addresses the inherent challenges posed by ambiguous mentions, typographical errors, and incomplete graph structures. Through structured representations, we demonstrated how candidate sets, context windows, and graph adjacency can be systematically integrated to produce reliable alignment of textual mentions to canonical entities [75]. This integration not only strengthens the robustness of entity resolution processes but also ensures that information retrieval systems grounded in knowledge graphs remain semantically coherent and logically consistent. By leveraging a hybrid approach that combines statistical learning with symbolic reasoning, our framework mitigates the weaknesses of purely neural models that struggle with edge cases, rare entity occurrences, and insufficient training data. [76]

Empirical evaluations revealed that combining local and global signals yields significant performance gains, particularly in domains featuring high ambiguity or specialized terminologies. The synergy between context-aware embeddings and knowledge-driven constraints enables the system to resolve entity mentions with greater precision than methods that rely solely on surface form similarity or deep learning embeddings [77]. This improvement is particularly crucial in domains such as biomedical informatics, where ambiguous entity names often refer to vastly different concepts, and legal or financial texts, where fine-grained distinctions in terminology impact decision-making. The use of logical axioms further guarantees that the selected entity aligns with domain knowledge, thereby reducing the risk of erroneous linkages that might otherwise propagate through downstream applications [78], [79]. Furthermore, our analysis indicates that relational constraints play a crucial role in refining entity predictions by disallowing incorrect associations that contradict well-established ontological structures.

Theoretical considerations illustrated that scaling and convergence can be managed through approximate inference algorithms and factorization-driven perspectives, while logical constraints ensure robust semantic coherence. The computational complexity of incorporating logical axioms and global consistency checks is mitigated through efficient probabilistic inference techniques, such as variational approximations and sampling-based methods [80]. This balance between expressivity and efficiency allows the framework to scale to large datasets while maintaining high accuracy in entity linking tasks. The use of factorization-based approaches, such as tensor decomposition and knowledge graph embeddings, further enhances the tractability of the system, ensuring that inference remains computationally feasible even in large-scale applications [81]. Moreover, by structuring entity linking as a constraint satisfaction problem, we enable more interpretable decision-making processes, facilitating trust and adoption in critical real-world applications.

This multi-faceted approach opens avenues for further research aimed at leveraging additional domain knowledge, adopting sophisticated optimization methods, and refining embedding strategies [82]. Future directions could explore the incorporation of probabilistic soft logic frameworks to enable more flexible reasoning over uncertain or incomplete data. Additionally, the adoption of graph neural networks (GNNs) could further enhance the propagation of relational information, allowing for richer contextual embeddings that better capture the dependencies between entities [83]. Moreover, the refinement of embedding strategies through contrastive learning and domain-adaptive fine-tuning holds promise for improving entity resolution in specialized fields with limited annotated data. Another compelling direction is the integration of cross-lingual entity linking approaches that leverage multilingual embeddings to bridge gaps between different languages, enhancing entity alignment across diverse corpora. [84], [85]

Ultimately, we anticipate that this fusion of lexical, embedding, and logical paradigms will continue to shape the evolution of entity linking systems in broad real-world settings. The increasing reliance on knowledge graphs across industries—from digital assistants and search engines to scientific knowledge management—demands methods that combine the statistical efficiency of machine learning with the interpretability and reliability of logical

reasoning [86]. By mitigating error propagation in automatically extracted knowledge graphs, our work paves the way for more accurate and semantically enriched applications in text analytics, information retrieval, and beyond. The intersection of deep learning and symbolic AI remains a promising research frontier, offering both practical advancements and theoretical insights into the nature of language understanding and structured knowledge representation. As entity linking continues to evolve, integrating domain-specific heuristics, probabilistic constraints, and richer contextual embeddings will be key to unlocking more precise and scalable knowledge-driven systems. [87]

References

[1] J. Wen, X. Meng, X. Hao, and J. Xu, "An efficient approach for continuous density queries," *Frontiers of Computer Science*, vol. 6, no. 5, pp. 581–595, Oct. 10, 2012. DOI: 10.1007/s11704-012-1120-4.

[2] J. Zhao, S. Sun, X. Liu, J. Sun, and A. Yang, "A novel biologically inspired visual saliency model," *Cognitive Computation*, vol. 6, no. 4, pp. 841–848, May 10, 2014. DOI: 10.1007/s12559-014-9266-z.

[3] H. Tan, X. Zhang, L. Lan, X. Huang, and Z. Luo, "Nonnegative constrained graph based canonical correlation analysis for multi-view feature learning," *Neural Processing Letters*, vol. 50, no. 2, pp. 1215–1240, Sep. 12, 2018. DOI: 10.1007/s11063-018-9904-7.

[4] S. Suh, S. Shin, J. Lee, C. K. Reddy, and J. Choo, "Localized user-driven topic discovery via boosted ensemble of nonnegative matrix factorization," *Knowledge and Information Systems*, vol. 56, no. 3, pp. 503–531, Jan. 8, 2018. DOI: 10.1007/s10115-017-1147-9.

[5] S. Zhang, J. Zhong, H. Yang, L. Zhantao, and G. Liu, "A study on pgep to evolve heuristic rules for fjssp considering the total cost of energy consumption and weighted tardiness," *Computational and Applied Mathematics*, vol. 38, no. 4, pp. 1–31, Oct. 10, 2019. DOI: 10.1007/s40314-019-0934-1.

[6] Y. Li, T. Yue, S. Ali, and L. Zhang, "Enabling automated requirements reuse and configuration," *Software & Systems Modeling*, vol. 18, no. 3, pp. 2177–2211, Nov. 29, 2017. DOI: 10.1007/s10270-017-0641-6.

[7] J. A. Fries, P. Varma, V. S. Chen, *et al.*, "Weakly supervised classification of aortic valve malformations using unlabeled cardiac mri sequences.," *Nature communications*, vol. 10, no. 1, pp. 3111–3111, Jul. 15, 2019. DOI: 10.1038/s41467-019-11012-3.

[8] J. Kim and A. Sim, "A new approach to multivariate network traffic analysis," *Journal of Computer Science and Technology*, vol. 34, no. 2, pp. 388–402, Mar. 22, 2019. DOI: 10.1007/s11390-019-1915-y.

[9] G. Mezzour, K. M. Carley, and L. R. Carley, "Remote assessment of countries' cyber weapon capabilities," *Social Network Analysis and Mining*, vol. 8, no. 1, pp. 1–15, Oct. 9, 2018. DOI: 10.1007/s13278-018-0539-5.

[10] C. Boucher, T. Gagie, A. Kuhnle, B. Langmead, G. Manzini, and T. Mun, "Prefix-free parsing for building big bwts," *Algorithms for molecular biology : AMB*, vol. 14, no. 1, pp. 13–13, May 24, 2019. DOI: 10.1186/s13015-019-0148-5.

[11] W. Liu and Z. Li, "An efficient parallel algorithm of n-hop neighborhoods on graphs in distributed environment," *Frontiers of Computer Science*, vol. 13, no. 6, pp. 1309–1325, Jul. 16, 2019. DOI: 10.1007/s11704-018-7167-0.

[12] Abhishek and V. Rajaraman, "A computer aided shorthand expander," *IETE Technical Review*, vol. 22, no. 4, pp. 267–272, 2005.

[13] Y. Lv, B. Cui, X. Chen, and J. Li, "Hat: An efficient buffer management method for flash-based hybrid storage systems," *Frontiers of Computer Science*, vol. 8, no. 3, pp. 440–455, Mar. 24, 2014. DOI: 10.1007/s11704-014-3364-7.

[14] T. Cai, M. Jiang, and G. Cao, "Multi-party quantum key agreement with five-qubit brown states," *Quantum Information Processing*, vol. 17, no. 5, pp. 103–, Mar. 19, 2018. DOI: 10.1007/s11128-018-1871-4.

[15] C. L. Goues, S. Forrest, and W. Weimer, "Current challenges in automatic software repair," *Software Quality Journal*, vol. 21, no. 3, pp. 421–443, Jun. 7, 2013. DOI: 10.1007/s11219-013-9208-0.

[16] M. L. Zeng and P. Mayr, "Knowledge organization systems (kos) in the semantic web: A multi-dimensional review," *International Journal on Digital Libraries*, vol. 20, no. 3, pp. 209–230, May 25, 2018. DOI: 10.1007/s00799-018-0241-2.

[17] Z. Xie, Z. Zeng, G. Zhou, and W. Wang, "Topic enhanced deep structured semantic models for knowledge base question answering," *Science China Information Sciences*, vol. 60, no. 11, pp. 110 103–, Sep. 21, 2017. DOI: 10.1007/s11432-017-9136-x.

[18] B. Dutta, A. Azhir, L.-H. Merino, *et al.*, "An interactive web-based application for comprehensive analysis of rnai-screen data.," *Nature communications*, vol. 7, no. 1, pp. 10 578–10 578, Feb. 23, 2016. DOI: 10.1038/ncomms10578.

[19] P. Sui and X. Yang, "A privacy-preserving compression storage method for large trajectory data in road network," *Journal of Grid Computing*, vol. 16, no. 2, pp. 229–245, Feb. 19, 2018. DOI: 10.1007/s10723-018-9435-5.

[20] R. Albertoni, M. D. Martino, P. Podestà, A. Abecker, R. Wössner, and K. Schnitter, "Lustre: A framework of linked environmental thesauri for metadata management," *Earth Science Informatics*, vol. 11, no. 4, pp. 525–544, Apr. 9, 2018. DOI: 10.1007/s12145-018-0344-8.

[21] M. A. Z. Raja, A. Kiani, A. Shehzad, and A. Zameer, "Memetic computing through bio-inspired heuristics integration with sequential quadratic programming for nonlinear systems arising in different physical models," *SpringerPlus*, vol. 5, no. 1, pp. 2063–2063, Dec. 1, 2016. DOI: 10.1186/s40064-016-3750-8.

[22] J.-S. Kim, K.-Y. Whang, H.-Y. Kwon, and I.-Y. Song, "Paradise: Big data analytics using the dbms tightly integrated with the distributed file system," *World Wide Web*, vol. 19, no. 3, pp. 299–322, Dec. 11, 2014. DOI: 10.1007/s11280-014-0312-2.

[23] A. Basu *et al.*, "Iconic interfaces for assistive communication," in *Encyclopedia of Human Computer Interaction*, IGI Global, 2006, pp. 295–302.

[24] V. Kuleshov, J. Ding, C. Vo, *et al.*, "A machine-compiled database of genome-wide association studies.," *Nature communications*, vol. 10, no. 1, pp. 3341–3341, Jul. 26, 2019. DOI: 10.1038/s41467-019-11026-x.

[25] D. AL-Kafaf, D.-K. Kim, and L. Lu, "A three-phase decision making approach for self-adaptive systems using web services," *Complex Adaptive Systems Modeling*, vol. 6, no. 1, pp. 1–24, Oct. 23, 2018. DOI: 10.1186/s40294-018-0059-1.

[26] W. Chen, J. Zhou, J. Zhu, G. Wu, and J. Wei, "Semi-supervised learning based tag recommendation for docker repositories," *Journal of Computer Science and Technology*, vol. 34, no. 5, pp. 957–971, Sep. 6, 2019. DOI: 10.1007/s11390-019-1954-4.

[27] G. Wang and N. Shi, "Collaborative representation-based discriminant neighborhood projections for face recognition," *Neural Computing and Applications*, vol. 32, no. 10, pp. 5815–5832, Feb. 1, 2019. DOI: 10.1007/s00521-019-04055-6.

[28] R. Liu, T.-L. Kung, and K. K. Parhi, "Impulse noise correction in ofdm systems," *Journal of Signal Processing Systems*, vol. 74, no. 2, pp. 245–262, Jul. 7, 2013. DOI: 10.1007/s11265-013-0782-y.

[29] X. Meng, L. Cao, X. Zhang, and J. Shao, "Top-k coupled keyword recommendation for relational keyword queries," *Knowledge and Information Systems*, vol. 50, no. 3, pp. 883–916, Jun. 1, 2016. DOI: 10.1007/s10115-016-0959-3.

[30] P. Érdi, K. Makovi, Z. Somogyvári, *et al.*, "Prediction of emerging technologies based on analysis of the u.s. patent citation network," *Scientometrics*, vol. 95, no. 1, pp. 225–242, Jun. 21, 2012. DOI: 10.1007/s11192-012-0796-4.

[31] U. Hermjakob, Q. Li, D. Marcu, *et al.*, "Incident-driven machine translation and name tagging for low-resource languages," *Machine Translation*, vol. 32, no. 1, pp. 59–89, Oct. 23, 2017. DOI: 10.1007/s10590-017-9207-1.

[32] A. Hinze, D. Bainbridge, S. J. Cunningham, *et al.*, "Capisco: Low-cost concept-based access to digital libraries," *International Journal on Digital Libraries*, vol. 20, no. 4, pp. 307–334, Mar. 14, 2018. DOI: 10.1007/s00799-018-0232-3.

[33] Z.-T. Zhu, M.-H. Yu, and P. Riezebos, "A research framework of smart education," *Smart Learning Environments*, vol. 3, no. 1, pp. 1–17, Mar. 31, 2016. DOI: 10.1186/s40561-016-0026-2.

[34] E. N. Cinicioglu and P. P. Shenoy, "A new heuristic for learning bayesian networks from limited datasets: A real-time recommendation system application with rfid systems in grocery stores," *Annals of Operations Research*, vol. 244, no. 2, pp. 385–405, Jun. 21, 2012. DOI: 10.1007/s10479-012-1171-9.

[35] Y.-G. Song, H. Cui, and X. Feng, "Parallel incremental frequent itemset mining for large data," *Journal of Computer Science and Technology*, vol. 32, no. 2, pp. 368–385, Mar. 13, 2017. DOI: 10.1007/s11390-017-1726-y.

[36] A. Sharma, M. Witbrock, and K. Goolsbey, "Controlling search in very large commonsense knowledge bases: A machine learning approach," *arXiv preprint arXiv:1603.04402*, 2016.

[37] Y. Liu, J. Wang, L. Wei, *et al.*, "Droidleaks : A comprehensive database of resource leaks in android apps," *Empirical Software Engineering*, vol. 24, no. 6, pp. 3435–3483, May 16, 2019. DOI: 10.1007/s10664-019-09715-8.

[38] J. Zhao and X. Wang, "Vehicle-logo recognition based on modified hu invariant moments and svm," *Multimedia Tools and Applications*, vol. 78, no. 1, pp. 75–97, Oct. 16, 2017. DOI: 10.1007/s11042-017-5254-0.

[39] X. Yan, J. Zhang, Y. Xun, and X. Qin, "A parallel algorithm for mining constrained frequent patterns using mapreduce," *Soft Computing*, vol. 21, no. 9, pp. 2237–2249, Nov. 17, 2015. DOI: 10.1007/s00500-015-1930-z.

[40] Y. Zhang, P. Wang, T. Liu, *et al.*, "Active and intelligent control onto thermal behaviors of a motorized spindle unit," *The International Journal of Advanced Manufacturing Technology*, vol. 98, no. 9, pp. 3133–3146, Jul. 30, 2018. DOI: 10.1007/s00170-018-2425-8.

[41] B. Cowen, A. N. Saridena, and A. Choromanska, "Lsalsa: Accelerated source separation via learned sparse coding," *Machine Learning*, vol. 108, no. 8, pp. 1307–1327, May 21, 2019. DOI: 10.1007/s10994-019-05812-3.

[42] M. Zeng, B. Yao, Z.-J. Wang, *et al.*, "Catiri: An efficient method for content-and-text based image retrieval," *Journal of Computer Science and Technology*, vol. 34, no. 2, pp. 287–304, Mar. 22, 2019. DOI: 10.1007/s11390-019-1911-2.

[43] C. Wang, J. Liu, N. Desai, M. Danilevsky, and J. Han, "Constructing topical hierarchies in heterogeneous information networks," *Knowledge and Information Systems*, vol. 44, no. 3, pp. 529–558, Aug. 26, 2014. DOI: 10.1007/s10115-014-0777-4.

[44] S. Khalid, M. U. Akram, T. Hassan, A. Jameel, and T. Khalil, "Automated segmentation and quantification of drusen in fundus and optical coherence tomography images for detection of armd.," *Journal of digital imaging*, vol. 31, no. 4, pp. 464–476, Dec. 4, 2017. DOI: 10.1007/s10278-017-0038-7.

[45] P. Maraver, R. Armañanzas, T. A. Gillette, and G. A. Ascoli, "Paperbot: Open-source web-based search and metadata organization of scientific literature," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–13, Jan. 24, 2019. DOI: 10.1186/s12859-019-2613-z.

[46] B. Bruening, "Word formation is syntactic: Adjectival passives in english," *Natural Language & Linguistic Theory*, vol. 32, no. 2, pp. 363–422, Feb. 8, 2014. DOI: 10.1007/s11049-014-9227-y.

[47] Z. Wang, S. Qiu, Q. Chen, *et al.*, "Anicode: Authoring coded artifacts for network-free personalized animations," *The Visual Computer*, vol. 35, no. 6, pp. 885–897, May 9, 2019. DOI: 10.1007/s00371-019-01681-y.

[48] J. F. Ferreira, C. Gherghina, G. He, S. Qin, and W.-N. Chin, "Automated verification of the freertos scheduler in hip/sleek," *International Journal on Software Tools for Technology Transfer*, vol. 16, no. 4, pp. 381–397, Mar. 18, 2014. DOI: 10.1007/s10009-014-0307-4.

[49] B. Bigio, D. Zelli, T. Lau, *et al.*, "Acnp 57th annual meeting: Poster session i," *Neuropsychopharmacology*, vol. 43, no. S1, pp. 77–227, Dec. 6, 2018. DOI: 10.1038/s41386-018-0266-7.

[50] A. Samareh and S. Huang, "Dl-chi: A dictionary learning-based contemporaneous health index for degenerative disease monitoring," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 1, pp. 17–, Mar. 5, 2018. DOI: 10.1186/s13634-018-0538-8.

[51] X. Pan and H.-B. Shen, "Rna-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach.," *BMC bioinformatics*, vol. 18, no. 1, pp. 136–136, Feb. 28, 2017. DOI: 10.1186/s12859-017-1561-8.

[52] H. Li, J. Cui, and J. Ma, "Social influence study in online networks: A three-level review," *Journal of Computer Science and Technology*, vol. 30, no. 1, pp. 184–199, Jan. 21, 2015. DOI: 10.1007/s11390-015-1512-7.

[53] X. Wu, G. Wang, X. Wang, and X. Yu, "Agency models based on different measures with comparison," *Soft Computing*, vol. 24, no. 9, pp. 6363–6373, May 15, 2019. DOI: 10.1007/s00500-019-04044-w.

[54] M. H. S. Segler, M. Preuss, and M. P. Waller, "Nature - planning chemical syntheses with deep neural networks and symbolic ai," *Nature*, vol. 555, no. 7698, pp. 604–610, Mar. 29, 2018. DOI: 10.1038/nature25978.

[55] A. Sharma, K. M. Goolsbey, and D. Schneider, "Disambiguation for semi-supervised extraction of complex relations in large commonsense knowledge bases," in *7th Annual Conference on Advances in Cognitive Systems*, 2019.

[56] K. M. Romagnoli, S. D. Nelson, L. E. Hines, P. E. Empey, R. D. Boyce, and H. Hochheiser, "Information needs for making clinical recommendations about potential drug-drug interactions: A synthesis of literature review and interviews.," *BMC medical informatics and decision making*, vol. 17, no. 1, pp. 21–21, Feb. 22, 2017. DOI: 10.1186/s12911-017-0419-3.

[57] Z. Rui, R.-h. Shi, Q. Jiaqi, and Z.-w. Peng, "An economic and feasible quantum sealed-bid auction protocol," *Quantum Information Processing*, vol. 17, no. 2, pp. 35–, Jan. 4, 2018. DOI: 10.1007/s11128-017-1805-6.

[58] S. Kumar, A. D. Vo, F. Qin, and H. Li, "Comparative assessment of methods for the fusion transcripts detection from rna-seq data.," *Scientific reports*, vol. 6, no. 1, pp. 21 597–21 597, Feb. 10, 2016. DOI: 10.1038/srep21597.

[59] L. Deng, Y. Jia, B. Zhou, J. Huang, and Y. Han, "User interest mining via tags and bidirectional interactions on sina weibo," *World Wide Web*, vol. 21, no. 2, pp. 515–536, Jun. 9, 2017. DOI: 10.1007/s11280-017-0469-6.

[60] L. Li, J. Ye, F. Deng, S. Xiong, and L. Zhong, "A comparison study of clustering algorithms for microblog posts," *Cluster Computing*, vol. 19, no. 3, pp. 1333–1345, Jul. 18, 2016. DOI: 10.1007/s10586-016-0589-2.

[61] Y. Tang, Z. Ren, W. Kong, and H. Jiang, "Compiler testing: A systematic literature analysis," *Frontiers of Computer Science*, vol. 14, no. 1, pp. 1–20, Mar. 7, 2019. DOI: 10.1007/s11704-019-8231-0.

[62] J. Hua, Y. Liu, Y. Shen, X. Tian, Y. Luo, and C. Jin, "A privacy-enhancing scheme against contextual knowledge-based attacks in location-based services," *Frontiers of Computer Science*, vol. 14, no. 3, pp. 143 605–, Dec. 19, 2019. DOI: 10.1007/s11704-019-8300-4.

[63] M. Olma, M. Karpathiotakis, I. Alagiannis, M. Athanassoulis, and A. Ailamaki, "Adaptive partitioning and indexing for in situ query processing," *The VLDB Journal*, vol. 29, no. 1, pp. 569–591, Nov. 15, 2019. DOI: 10.1007/s00778-019-00580-x.

[64] Y. Zhang, C. A. Phillips, G. L. Rogers, E. J. Baker, E. J. Chesler, and M. A. Langston, "On finding bicliques in bipartite graphs: A novel algorithm and its application to the integration of diverse biological data types," *BMC bioinformatics*, vol. 15, no. 1, pp. 110–110, Apr. 15, 2014. DOI: 10.1186/1471-2105-15-110.

[65] P. Peng, L. Zou, M. T. Özsu, L. Chen, and D. Zhao, "Processing sparql queries over distributed rdf graphs," *The VLDB Journal*, vol. 25, no. 2, pp. 243–268, Jan. 4, 2016. DOI: 10.1007/s00778-015-0415-0.

[66] J. Hou, K.-l. Jia, and X.-j. Jiao, "Teaching evaluation on a webgis course based on dynamic self-adaptive teaching-learning-based optimization," *Journal of Central South University*, vol. 26, no. 3, pp. 640–653, Apr. 22, 2019. DOI: 10.1007/s11771-019-4035-5.

[67] C. Jin, Y. Kong, Q. Kang, W. Qian, and A. Zhou, "Benchmarking in-memory database," *Frontiers of Computer Science*, vol. 10, no. 6, pp. 1067–1081, May 6, 2016. DOI: 10.1007/s11704-016-5366-0.

[68] G. Peng, Y. Liu, J. Wang, and J. Gu, "Analysis of the prediction capability of web search data based on the he-tdc method – prediction of the volume of daily tourism visitors," *Journal of Systems Science and Systems Engineering*, vol. 26, no. 2, pp. 163–182, Jan. 24, 2017. DOI: 10.1007/s11518-016-5311-7.

[69] Y. Cheng, P. Ding, T. Wang, W. Lu, and X. Du, "Which category is better: Benchmarking relational and graph database management systems," *Data Science and Engineering*, vol. 4, no. 4, pp. 309–322, Nov. 11, 2019. DOI: 10.1007/s41019-019-00110-3.

[70] G. Li, P. Yang, Z. Liang, and S. Cui, "Intelligent design and group assembly of male and female dies for hole piercing of automotive stamping dies," *The International Journal of Advanced Manufacturing Technology*, vol. 103, no. 1, pp. 665–687, Mar. 25, 2019. DOI: 10.1007/s00170-019-03576-7.

[71] R. Shang, Y. Li, and L. Jiao, "Co-evolution-based immune clonal algorithm for clustering," *Soft Computing*, vol. 20, no. 4, pp. 1503–1519, Feb. 7, 2015. DOI: 10.1007/s00500-015-1602-z.

[72] Y. Lu, R. Xie, and S. Y. Liang, "Extraction of weak fault using combined dual-tree wavelet and improved mca for rolling bearings," *The International Journal of Advanced Manufacturing Technology*, vol. 104, no. 5, pp. 2389–2400, Jul. 5, 2019. DOI: 10.1007/s00170-019-04065-7.

[73] J. Corredor, M. Gaydos, and K. Squire, "Seeing change in time: Video games to teach about temporal change in scientific phenomena," *Journal of Science Education and Technology*, vol. 23, no. 3, pp. 324–343, Aug. 22, 2013. DOI: 10.1007/s10956-013-9466-4.

[74] M. Ertas, I. Yildirim, M. E. Kamasak, and A. Akan, "An iterative tomosynthesis reconstruction using total variation combined with non-local means filtering," *Biomedical engineering online*, vol. 13, no. 1, pp. 65–65, May 27, 2014. DOI: 10.1186/1475-925x-13-65.

[75] K. G. Talley and A. M. Ortiz, "Women's interest development and motivations to persist as college students in stem: A mixed methods analysis of views and voices from a hispanic-serving institution," *International Journal of STEM Education*, vol. 4, no. 1, pp. 1–24, Mar. 30, 2017. DOI: 10.1186/s40594-017-0059-2.

[76] Z. Chen, W. Zhang, B. Deng, H. Xie, and X. Gu, "Name-face association with web facial image supervision," *Multimedia Systems*, vol. 25, no. 1, pp. 1–20, Mar. 23, 2017. DOI: 10.1007/s00530-017-0544-y.

[77] Y. Jia, L. Bai, S. Liu, P. Wang, J. Guo, and Y. Xie, "Semantically-enhanced kernel canonical correlation analysis: A multi-label cross-modal retrieval," *Multimedia Tools and Applications*, vol. 78, no. 10, pp. 13 169–13 188, Feb. 27, 2018. DOI: 10.1007/s11042-018-5767-1.

[78] P. Peng, L. Zou, Z. Du, and D. Zhao, "Using partial evaluation in holistic subgraph search," *Frontiers of Computer Science*, vol. 12, no. 5, pp. 966–983, Sep. 18, 2018. DOI: 10.1007/s11704-016-5522-6.

[79] A. Sharma and K. M. Goolsbey, "Learning search policies in large commonsense knowledge bases by randomized exploration," 2018.

[80] C. C. Yan, H. Xie, B. Zhang, Y. Ma, Q. Dai, and Y. Liu, "Fast approximate matching of binary codes with distinctive bits," *Frontiers of Computer Science*, vol. 9, no. 5, pp. 741–750, Oct. 9, 2015. DOI: 10.1007/s11704-015-4192-0.

[81] H. Mei and L. Zhang, "Can big data bring a breakthrough for software automation," *Science China Information Sciences*, vol. 61, no. 5, pp. 056 101–, Apr. 9, 2018. DOI: 10.1007/s11432-017-9355-3.

[82] D. L. Strayer, J. M. Cooper, R. M. Goethe, M. M. McCarty, D. Getty, and F. Biondi, "Assessing the visual and cognitive demands of in-vehicle information systems," *Cognitive research: principles and implications*, vol. 4, no. 1, pp. 18–18, Jun. 21, 2019. DOI: 10.1186/s41235-019-0166-3.

[83] H. Xu, Y. Gu, J. Qi, J. He, and G. Yu, "Diversifying top-k routes with spatial constraints," *Journal of Computer Science and Technology*, vol. 34, no. 4, pp. 818–838, Jul. 19, 2019. DOI: 10.1007/s11390-019-1944-6.

[84] T. Liu, F. Wang, and G. Agrawal, "Stratified sampling for data mining on the deep web," *Frontiers of Computer Science*, vol. 6, no. 2, pp. 179–196, Mar. 31, 2012. DOI: 10.1007/s11704-012-2859-3.

[85] A. Abhishek and A. Basu, "A framework for disambiguation in ambiguous iconic environments," in *AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17*, Springer, 2005, pp. 1135–1140.

[86] S. D. Sen and J. A. Adams, "An influence diagram based multi-criteria decision making framework for multirobot coalition formation," *Autonomous Agents and Multi-Agent Systems*, vol. 29, no. 6, pp. 1061–1090, Oct. 15, 2014. DOI: 10.1007/s10458-014-9276-y.

[87] B. Chen, Y. Ding, and D. J. Wild, "Improving integrative searching of systems chemical biology data using semantic annotation," *Journal of cheminformatics*, vol. 4, no. 1, pp. 6–6, Mar. 8, 2012. DOI: 10.1186/1758-2946-4-6.