FRASE: Frame-based Structured Representations for Generalizable SPARQL Query Generation

Papa Abdou Karim Karou Diallo, Amal Zouaq
LAMA-WeST, Polytechnique Montreal, Mila - Quebec AI Institute
{diallokarou28, amal.zouaq}@polymtl.ca

Abstract

Translating natural language questions into SPARQL queries enables Knowledge Base querying for factual and up-to-date responses. However, existing datasets for this task are predominantly template-based, leading models to learn superficial mappings between question and query templates rather than developing true generalization capabilities. As a result, models struggle when encountering naturally phrased, template-free questions. This paper introduces FRASE (FRAme-based Semantic Enhancement), a novel approach that leverages Frame Semantic Role Labeling (FSRL) to overcome this limitation. In addition, we present LCQ1-Frame, LCQ2-Frame, and QALD-10-Frame, a suite of new datasets derived from LC-QuAD 1.0, LC-QuAD 2.0, and QALD-10, where each question is enriched using FRASE through frame detection and the mapping of frame-elements to their corresponding arguments. We evaluate our approach for the Question-2-SPARQL task through extensive experiments using recent large language models (LLMs) under different fine-tuning configurations. Our results demonstrate that integrating frame-based structured representations consistently improves SPARQL generation performance, particularly in challenging generalization scenarios when test questions feature unseen templates (unknown template splits) and when they are all naturally phrased (reformulated questions). Our code and data are available at https://github.com/Lama-West/FRASE-LREC-2026.

Keywords: Question-2-SPARQL, LLMs, Generalization, Lexical variation, Structured representation

1. Introduction

Information democratization aims to ease access to the vast amount of factual information stored in Knowledge Bases (KBs) and relational databases. A crucial step towards this goal is the translation of natural language questions into structured queries such as SPARQL, SQL or S-expression (Diallo et al., 2024; Banerjee et al., 2022; Sharma et al., 2025). Recent advances in Large Language Models (LLMs), particularly decoder-only architectures, have significantly improved the semantic representation of natural language, achieving state-of-the-art (SOTA) results on a variety of natural language understanding tasks.

Nevertheless, despite these successes, LLMs remain sensitive to input phrasing and prompt formulation, resulting in limited robustness and generalization across diverse question formulations and unseen patterns (Leidinger et al., 2023; Zheng et al., 2023b). A key limitation of current Question-2-SPARQL systems lies in the datasets on which they are trained (Diallo et al., 2024; Reyd and Zouaq, 2023). Many benchmarks rely on rigid, template-based constructions, which lead models to learn surface-level mappings between input questions and query structures. This shortcut learning hinders their ability to generalize to naturally phrased, template-free questions, especially when the input deviates from patterns seen during training. This underscores the importance of adopting structured representations that unify diverse question formulations in the Question-2-SPARQL task, with frame-based representations offering a promising solution.

To address these aforementioned challenges, we introduce FRASE (FRAme-based Semantic Enhancement), a method that augments natural language questions with structured semantic information derived from Frame Semantics and use this enriched question representation for an improved Question-2-SPARQL parsing. These frames have proven valuable in enhancing semantic understanding in tasks such as machine reading comprehension (Guo et al., 2020; Flanigan et al., 2022; Bonn et al., 2024) and information extraction (Su et al., 2021; Li et al., 2024b; Chanin, 2023; Su et al., 2023).

To evaluate this architecture, we introduce LCQ1-Frame, LCQ2-Frame and QALD-10-Frame, a collection of new datasets derived from LC-QuAD 1.0 (Trivedi et al., 2017), LC-QuAD 2.0 (Dubey et al., 2019) and QALD-10 (Usbeck et al., 2024), in which each question is annotated with its corresponding frames and frame-elements and argument mapping using the FRASE pipeline. We conduct comprehensive experiments with several recent LLMs under different fine-tuning settings, assessing the impact of our semantic augmentation across various dataset splits, including out-of-distribution (OOD) settings. This allows us to address the following research questions:

RQ1: How does model performance vary when training and testing on template-based questions

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1303–1319
11–16 May 2026.

versus testing with naturally phrased questions or questions with unseen templates?

RQ2: To what extent does incorporating structured semantic representations based on frames improve performance across these different training and evaluation configurations?

RQ3: Can combining template-based and template-free questions during training improve generalization ?

Thus, our contributions are: (1) We propose FRASE, a new method for frame detection and arguments identification that does not rely on manually identified target spans, leveraging a retriever grounded in ontology relations and classes and FrameNet (Fillmore, 1976) frames. (2) We demonstrate that enriching questions with frame-based structured representations improves generalization in SPARQL query generation and make the representations of paraphrases closer in the embedding space. (3) We show that this improvement holds not only for unseen-template test sets but also for challenging, naturally phrased reformulations. (4) We enrich three datasets LC-QuAD 1.0, LC-QuAD 2.0, and QALD-10 with semantic frames using our approach, and make them publicly available as open resources for the research community.

2. Related Works

2.1. SPARQL query generation

SPARQL generation from question has been extensively studied using both Small and Large Language Models (SLMs and LLMs) (Diallo et al., 2024; Reyd and Zouaq, 2023; Sharma et al., 2025; Banerjee et al., 2022; Emonet et al., 2024; Zahera et al., 2024). These models are typically fine-tuned end-to-end, often with enhancements such as copy mechanisms (Banerjee et al., 2022; Diallo et al., 2024) or non-parametric memory modules (Sharma et al., 2025) to reduce URI-related errors.

While many systems explicitly generate SPARQL queries, others bypass query generation altogether by having LLMs directly produce answers grounded in the knowledge base (Shavarani and Sarkar, 2024; Alawwad et al., 2024; Muennighoff, 2022). Prompt engineering has also become a prominent strategy, using few-shot examples or explicit URI context to guide generation (Luo et al., 2023; Muennighoff, 2022; Diallo et al., 2024). Models such as Code Llama v2 (Roziere et al., 2023), Mistral 7B (Jiang et al., 2023), and Mistral 7B Instruct[1] have been widely adopted for such tasks.

Agent-based approaches such as SPINACH (Liu et al., 2024) and KB-Binder (Li et al., 2023) leverage LLMs as interactive agents, iteratively exploring the knowledge graph and refining candidate queries through tool use and binding mechanisms. In contrast, fine-tuning strategies exemplified by (Alekseev et al., 2025) rely on smaller LLMs trained with entity linking and predicate matching pipelines, aiming for robust handling of complex and temporal queries. Other works focus on in-context learning, with (D'Abramo et al., 2025) showing that pretrained LLMs can generate queries when provided with gold entities and relations without additional training. Finally, pretraining-oriented methods like TSET (Qi et al., 2024) emphasize structural correction during intermediate stages to enforce the syntactic and semantic validity of generated SPARQL. Together, these approaches highlight the field's dual objectives: increasing reliability in query generation while balancing scalability, supervision requirements, and reasoning flexibility.

Despite promising results, existing methods struggle with naturally phrased or paraphrased questions. For instance, LC-QuAD 2.0 studies show a marked performance drop on reformulated (template-free) queries, revealing poor generalization to unseen linguistic patterns (Diallo et al., 2024; Reyd and Zouaq, 2023).

2.2. Structured representation of text sequences

To structurally represent the semantics of text, several formalisms have been developed, including Abstract Meaning Representation (AMR) (Langkilde and Knight, 1998), Universal Conceptual Cognitive Annotation (UCCA) (Abend and Rappoport, 2013), Elementary Dependency Structures (EDS) (Oepen et al., 2002), Prague Tectogrammatical Graphs (PTG) (Zeman and Hajic, 2020) , Discourse Representation Structures (DRS) (Kamp and Reyle, 2019; Liu et al., 2018), and Universal Decompositional Semantics (UDS) (White et al., 2019). Each provides a unique perspective on meaning. For instance, AMR abstracts semantics into concept graphs while UCCA captures cognitively motivated event structures.

In KBQA, AMR has been used as both an intermediate representation and a reasoning space, supporting semantic comparisons and query generation (Wang et al., 2023; Shi et al., 2024). Approaches rely on rule-based pipelines or seq2seq models (Regan et al., 2024; Bornea et al., 2021), but face challenges from parsing errors and mismatches with formal query languages (Jin et al., 2024). Recent works leverage AMR to improve robustness against hallucinations in LLM outputs (Regan et al., 2024) and extend its use to multi-hop QA (Wang et al., 2023), open-domain QA (Shi et al., 2024), data augmentation (Jin et al., 2024; Zhang

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

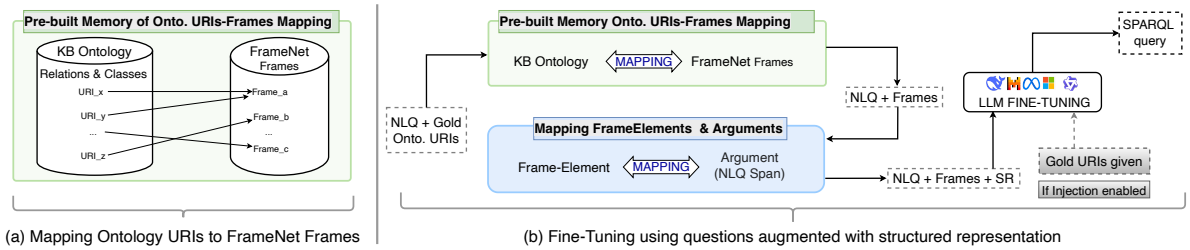


Figure 1: Overview of the proposed FRASE architecture. (a) Pre-construction of a retrieval-based memory that maps ontology (onto.) relations and classes to their corresponding FrameNet frames. (b) Fine-tuning stage, where natural language questions (NLQ) are augmented with structured representations (SR) derived from the pre-built memory and instantiated based on the questions' content. The gold ontology URIs in the beginning of Section (b) are used only to retrieve the frames and are then discarded to prevent data leakage. The gray connection indicates the optional URIs-Injection setting, in which gold URIs are explicitly provided (see Section 3.2.2).

et al., 2025), and various NLP tasks (Shivashankar et al., 2022). While AMR often shows neutral impact overall (Jin et al., 2024), (Zhang et al., 2025) report positive effects when augmenting structured representations with natural language descriptions.

Unlike previous SPARQL generation methods that rely on template learning or AMR-style intermediate representations, FRASE introduces a frame-semantic layer as a lightweight yet expressive bridge between natural language and knowledge base structures. It aligns KB relations and classes with FrameNet (Baker et al., 1998a) frames through a retriever and grounds each frame-element in its textual argument, eliminating the need for gold lexical units (An et al., 2023) or handcrafted alignment rules. This offers an effective solution to the lexical and phrasing variation issues that remain challenging for SOTA systems such as SPINACH(Liu et al., 2024) and TSET(Qi et al., 2024).

2.3. Our structured representation: frame-base semantic parsing

We adopt a frame-based structured representation as a middle ground between rigid code-like semantic formalisms and the more flexible representations aligned with the natural language data used to pre-train LLMs. Frames capture meaning through structured descriptions of prototypical situations, consisting of a frame-label, a natural language definition, core semantic roles (frame-elements), and the *lexical units* that evoke them.

Frame semantic parsing approaches fall into two main categories: generative sequence-to-sequence methods and representation learning techniques leveraging FrameNet's structured knowledge. Seq2Seq models treat frame parsing as a generation task that includes frame identification and argument extraction (Sutskever, 2014; Raffel et al., 2020; Kalyanpur et al., 2020; Chanin, 2023). Architectures such as T5 (Raffel et al., 2020), pre-trained on PropBank (Kingsbury and Palmer, 2002) and FrameNet, are often used with lexical units and data augmentation to improve robustness. Some models adopt multi-task learning with shared encoders and task-specific decoders (Kalyanpur et al., 2020). Representation learning methods, in contrast, focus on aligning sentence-level or span-level embeddings with candidate frames (Jiang and Riloff, 2021). These approaches often use Graph Neural Networks (Wu et al., 2020) or contrastive learning (Ju et al., 2024; An et al., 2023) to integrate semantic relations among frames, elements, and lexical triggers (Su et al., 2021; Zheng et al., 2022; Tamburini, 2022).

A major limitation of both paradigms is their reliance on an explicit target lexical unit to which the frame should be assigned—an annotation absent in datasets like LC-QuAD 2.0. Without such target, SOTA models for frame-based parsing become inapplicable, and structured semantic representations cannot be extracted or evaluated in downstream tasks such as SPARQL generation.

3. Methodology

3.1. Datasets

To ensure a fair comparison with previous studies, we evaluate our approach on three benchmark datasets: QALD-10, LC-QuAD 1.0, and LC-QuAD 2.0. Among these, QALD-10 and LC-QuAD 2.0 are built on the Wikidata knowledge base, while LC-QuAD 1.0 is based on DBpedia.

3.1.1. QALD-10 and LC-QuAD 1.0

QALD (Question Answering over Linked Data) datasets are part of the QALD challenge series, where questions are paired with SPARQL queries over linked data sources. In particular, QALD-10

(Usbeck et al., 2024), is built on the Wikidata knowledge base and addresses several shortcomings of earlier QALD benchmarks, such as low-quality translations for non-English questions and limited SPARQL query complexity. QALD-10 offers a more challenging and realistic testbed with semantically richer queries and diverse question formulations, making it one of the most comprehensive and practical datasets in the QALD series. We use the English version, which contains 412 training and 394 test examples. The dataset has approximately 58% of multi-hops questions (between 1 and 8 and 2.1 on average).

LC-QuAD 1.0 (Trivedi et al., 2017) (LCQ1 for short) includes 5,000 questions answerable using the DBpedia knowledge base, with human-readable entity URIs that simplify model processing. Each question has two versions—one auto-generated and one human-rewritten—and is linked to a query template via a template ID. Following prior works (Ding et al., 2019; Diallo et al., 2024; Banerjee et al., 2022), we use 4,000 questions for training, 500 for validation, and 500 for testing.

3.1.2. LC-QuAD 2.0

Abbreviated as LCQ2, this dataset (Dubey et al., 2019) is composed of a total of 30,225 questions paired with SPARQL queries. Each entry includes two semantically equivalent versions of the question: one generated from a predefined template and another reformulated manually to resemble more natural human phrasing. This dual-question format enables nuanced evaluation of model generalization to both synthetic and naturally expressed inputs. To further analyze the influence of question formulation on model performance, we exploit the dual-question structure of LCQ2 to construct three dataset variants: (1) Raw Questions: only the original (template-based) questions are retained. (2) Reformulated Questions: only the human-written, naturally phrased questions are used. (3) Combined Questions: both question versions are treated as distinct entries but associated to the same SPARQL query, effectively doubling the dataset size and increasing question variety.

3.2. FRASE main architecture

The motivation behind incorporating structured semantic representations via Frame Semantic Role Labeling (FSRL) lies in the observation that different surface formulations of a question share the same underlying meaning. Regardless of phrasing, such questions typically evoke the same core event or concept, along with a consistent set of participants and their roles. This intuition aligns closely with the theory of frame semantics, where each frame represents a conceptual structure that encapsulates an event or situation, and frame-elements denote the roles associated with its participants. Given a natural language question (NLQ), FRASE aims to extract its structured semantic representation by (1) identifying the frames it evokes and (2) mapping the associated frame-elements to their corresponding spans within NLQ.

3.2.1. Stage 1: mapping ontology elements to FrameNet frames

In this stage, we construct a retrieval-based mapping between ontology URIs and their corresponding FrameNet frames. This process produces a structured memory that systematically associates ontology URIs with the appropriate frames across all datasets.

Frames and KB elements representation To enable effective semantic alignment between FrameNet frames and Knowledge Base (KB) relations/classes, we represent a frame in the most expressive method that combines three components: the frame-label, the definition (or description), and the list of its frame-elements. This representation captures both the semantic core of the frame and the structure of its participant roles, which improves retrieval performance. In contrast, representing KB relations/classes is more straightforward. For each relation/class URI in Wikidata, we concatenate the relation label (i.e., its name) with its textual description as provided by the KB. This representation succinctly captures the intended semantics of the relation/class and is well-suited for embedding-based similarity search.

Embedding and semantic retrieval For every question paired with a SPARQL query, we extract the corresponding ontology URIs (relations and classes). We assume that for each KB relation or class, there exists a semantically equivalent FrameNet frame expressing a similar conceptual meaning. To build the alignment memory between FrameNet frames and ontology elements, we begin by generating the embeddings for all frames using the BGE embedding model[2], which are then stored in a vector database. Then, for each URI of the ontology, we use its label and textual description from the KB as an input to the retriever which returns the top-1 most similar frame based on cosine similarity in the embedding space. This establishes an alignment between KB relations/classes and FrameNet frames. Leveraging this semantic alignment, we now have the frames evoked by each question, effectively bridging structured KB relations and linguistic semantic frames. These identified frames

[2]https://huggingface.co/BAAI

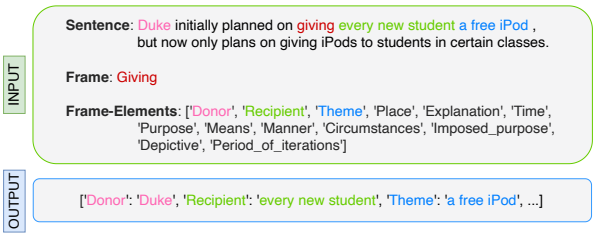


Figure 2: An example of FrameNet exemplars data used to fine-tune the Qwen model for argument identification.

form the foundation for the subsequent stages of structured semantic representation and argument identification.

3.2.2. Stage 2: Fine-Tuning using questions augmented with structured representation

Arguments identification After retrieving the most semantically aligned frame for each KB relation/class, we map frame-elements to text spans in the question as shown in the blue box in Section (b) of Figure 1. To this end, we fine-tuned Qwen2.5-7B (Yang et al., 2024), following (Devasier et al., 2025), on FrameNet exemplars ($3,353$ training and $1,247$ test sentences). The model learns to output structured JSON mappings between frame-elements and spans, enabling argument extraction as shown in Figure 2. The final augmented input[3] is presented in the format shown in Figure 3.

Frame-based dataset augmentation By applying FRASE to each dataset entry, we generate structured semantic representations and construct LCQ2-Frame, LCQ1-Frame, and QALD-10-Frame, which extend the original datasets with frame-based annotations. This enriched representation captures the overall meaning of a question through the identified frame and clarifies the semantic role of its entities or phrases by linking them to frame-elements. Although these extracted frames are not directly evaluated, we measure their impact through downstream SPARQL query generation performance.

SPARQL Query Generation All LLMs are fine-tuned using an INSTRUCTION-INPUT-OUTPUT format. In this setup, the instruction indicates to the model that it must generate a SPARQL query corresponding to the input question and provides the required context to successfully carry it out. The input consists of the NLQ, optionally augmented with its corresponding structured semantic representation

[3]The models we used have sufficient context length to process the entire input.

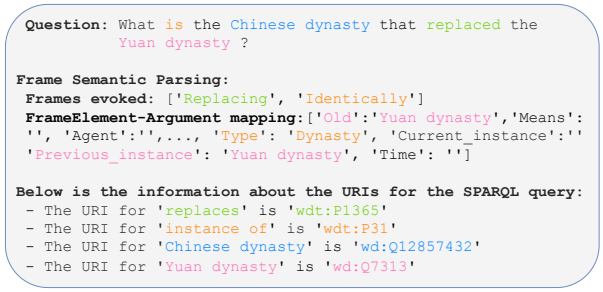


Figure 3: An example of input NLQ augmented with SR and Uris (the URIs-Injection is enabled here).

along with the URIs as shown in Section (b) of Figure 1. The output is a SPARQL query whose execution produces the answer to the question. For experiments involving the structured representation, this additional information is appended to the end of the question. This allows the model to condition its generation on both the NLQ and its enriched semantic context.

URIs-Injection mechanism As this work focuses on addressing lexical variation, our experiments also include scenarios in which the URIs required in the SPARQL queries are supplied via a URIs-Injection mechanism. This enables us to isolate issues arising from out-of-scope factors unrelated to lexical variation, such as the frequent presence of incorrect URIs (Diallo et al., 2024). This URIs-Injection mechanism used only during inference involves explicitly incorporating URIs into the input as supplementary context. Unlike the tagging methods in (Diallo et al., 2024; Banerjee et al., 2022; Reyd and Zouaq, 2023) that give the exact textual spans corresponding to each URI, we only provide the URIs' labels as shown in Figure 3. Moreover, a key advantage of this tagging method is its flexibility making it suitable for any type of question without relying on templates that enforce fixed positions of KB elements.

3.3. Models and evaluation metrics

For consistent results across different model architectures and sizes, we selected several recent and competitive LLMs: LLaMA 3.2 3B, Phi-4 14B, DeepSeek-R1 37B, and Mistral 7B GritLM that is a model with a unique training objective that combines cross-entropy loss with contrastive loss, aiming to optimize both text generation and embedding quality—making it potentially suitable for this test on generalization capability (Dubey et al., 2024; Abdin et al., 2024; Guo et al., 2025; Muennighoff et al., 2024).

For evaluation, we report BLEU scores by comparing the generated SPARQL queries with the gold-standard references. Additionally, we use

the F1 score as an execution-based metric computed by executing both the predicted and reference queries against the KB and comparing their returned answers.

4. Experiments and results

Since the Question-2-SPARQL datasets lack frame and frame-elements annotations, we evaluate the structured representation indirectly via its effect on SPARQL query generation performance.

4.1. SPARQL query generation - baselines using LCQ1 and LCQ2

Table 1 reports the performance of several recent LLMs fine-tuned on the Raw Questions subset of the LCQ1 and LCQ2 datasets. These constitute our baseline results. To evaluate generalization capabilities, we consider only one training configuration (template-based questions or Raw Questions) but for testing we consider two test scenarios: (1) the Raw Questions test set (left side of each dataset's bloc in Table 1), and (2) the Reformulated Questions test set, composed of human-written paraphrases (right side). Across models and question types, BLEU scores remain consistent, suggesting similar surface-level quality. However, the execution-based metric F1 has more variation. Notably, models exhibit a performance drop when evaluated on the reformulated test set, suggesting limited generalization to linguistic variations not seen during training. Somewhat unexpectedly, Mistral GritLM (Muennighoff et al., 2024) does not outperform the other models, despite its training objective. Among all models evaluated, Phi-4 and DeepSeek-R1 achieve the best overall performance, particularly in execution-based metrics. Consequently, we select them for subsequent experiments that analyze the impact of structured frame-based representations on generalization. Given the relatively close performance across models in this task, we hypothesize that trends observed using these two models are likely to extend to other models with comparable capacity.

4.2. Generalization experiments

4.2.1. LCQ2

Since LCQ2 is the most challenging dataset with two splits for assessing generalization, we evaluate model performance across the two levels of difficulty of the splits obtained following the protocol established by (Reyd and Zouaq, 2023):

Unknown Template Splits As described in Section 3.1.2, the Unknown Template Split (UTS) is constructed so that the test set contains only questions whose templates are not seen during training. In this experiment, both training and testing data are drawn from the UTS. For each, we consider two variants: (1) the original natural language questions alone, and (2) the same questions enriched with frame-based semantic annotations.

Template-free (Reformulated) questions We ran experiments with three question variants (Section 3.1.2): Raw, Reformulated, and Combined (treating raw and reformulated questions as separate training samples mapped to the same SPARQL query). We also assess performance when using the URIs-Injection mechanism described in Section 3.2.2 in comparison with setting where it's not used. For each dataset variant, we considered both standard and frame-augmented versions, resulting in a total of twelve training configurations and twenty-four test configurations for the two models considered (Phi-4 14B and DeepSeek-R1 37B). During inference, our evaluation is restricted to the Reformulated Questions subset of the test set. We examine both the original and frame-augmented variants to assess generalization, and we conduct experiments under two configurations: with and without the URIs-Injection mechanism. Table 2 reports the results of experiments on UTS while Tables 3 shows the results of the experiments using the reformulated questions.

4.2.2. LCQ1 and QALD-10

Following the same procedure as for LCQ2, we augmented the LCQ1 and QALD-10 datasets with frame-based semantic representations (Section 3.2) and conducted similar Question-2-SPARQL experiments. For LCQ1, we used reformulated questions for both training and testing; for QALD-10, we relied on the only available set of human-formulated questions. The results in Table 4, which also report SOTA baselines, confirm the benefits of incorporating structured frame information and highlight the importance of correct KB URIs. With the URIs-Injection mechanism—simulating ideal identifier retrieval—we achieved the best performance on both datasets.

4.3. Embedding quality via structured representations

We evaluate the effect of structured representations on embedding similarity between Raw and Reformulated Questions by computing cosine similarities over paraphrased pairs (Raw question and reformulated question) using Phi-4. Three pooling strategies were tested (Mean, EOS, Last Layer)

	Models	LCQ1				LCQ2			
		Raw Questions		Reformulated Questions		Raw Questions		Reformulated Questions	
		BLEU	F1	BLEU	F1	BLEU	F1	BLEU	F1
LLMs	Llama 3.2 3B	83	57	63	45	83	32	37	26
	Mistral 7B GritLM	88	61	66	49	85	25	43	24
	Phi-4 14B *	90	77	68	63	86	46	42	27
	DeepSeek-R1 37B	88	67	65	50	86	40	41	26

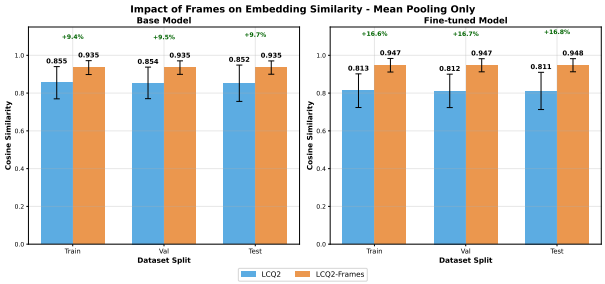
* Best overall result (baseline for comparison).

Table 1: Models performances with BLEU and F1 on **LCQ1** and **LCQ2** for two different types of questions (Template-based and template-free). The models are trained with template-based questions.

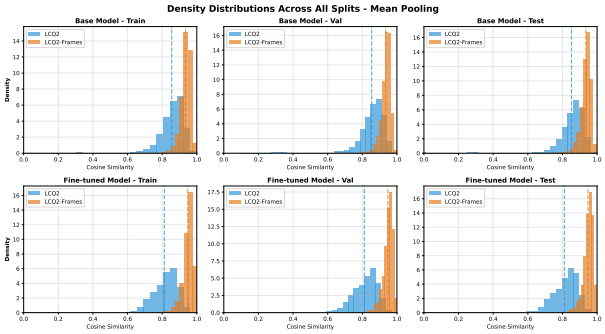
Model Used	Use of frame-based augmentation	Test on Unknown Template Split	
		BLEU (query-based)	**F1** (answer-based)
Phi-4 14B	✗	73	50
	✓	81	65
DeepSeek-R1 37B	✗	61	37
	✓	71	44

Table 2: Evaluation of Phi-4 14B and DeepSeek-R1 37B on raw questions from the Unknown Template Split of LCQ2, comparing models trained with and without frames.

and we report the best results obtained with mean pooling. Experiments were run on LCQ2 and LCQ2-Frames with both the pretrained base Phi-4 and the fine-tuned Phi-4 using Combined Questions. For the base model, we compare the embedding quality of the original questions with the embedding quality of questions augmented with frames. For the fine-tuned models, we evaluate two versions: one fine-tuned using only questions as input, and another fine-tuned using frame-augmented questions. We then compare their embedding quality in terms of paragraph detection performance using solely a raw question and its reformulated version (without adding frames). Results (Figure 4) shows that frames improve paraphrase detection by 9.5% on the base model and 16.6% after fine-tuning. Moreover, the fine-tuned model exhibits a clear rightward shift in similarity distributions, especially on LCQ2-Frames, reflecting stronger alignment of paraphrased questions, whereas the base model yields flatter distributions with peaks at lower similarity values.



(a) Effect of frames on cosine similarity between raw and natural questions across splits, using mean-pooled embeddings for base Phi-4 and its fine-tuned variant on LCQ2.



(b) Frames' impact on density distributions of cosine similarities across splits for base and fine-tuned Phi-4.

Figure 4: Frames effect on cosine similarity.

5. Discussion

Comparison with SOTA approaches We set new SOTA results on QALD-10 and LCQ1 with natural questions, showing the effectiveness of our approach in handling lexical variation. While our method surpasses prior systems on LCQ1 even without URIs-injection, this mechanism is crucial for exceeding SOTAs on QALD-10 (SPINACH and DFSL-MQ beam FS). However, these LLM-based methods have notable limitations. Most lack evaluation of generalization to natural questions, making our approach a strong alternative. Agent-based systems like SPINACH (Liu et al., 2024) suffer from high latency, exploration inefficiency, and poor scalability on large knowledge graphs (Li et al., 2023). In-context learning approaches (D'Abramo et al., 2025) cut training costs but their beam-search used to tackle the triple-flip problem (incorrectly swapping the subject and object in a SPARQL triple) reduces practicality in end-to-end KBQA. Pretraining-based frameworks such as TSET (Qi et al., 2024)

Model Used	Training Data	with Frames	Test on Reformulated Questions			
			BLEU (query-based)		**F1** (answer-based)	
			-	+ URIs	-	+ URIs
Phi-4 14B	Raw Questions	✗	42	44	27	34
		✓	54	70	30	49
	Reformulated Questions	✗	65	81	32	59
		✓	70	86	39	66
	Combined Questions	✗	67	77	40	53
		✓	**73**	**88**	**50**	**67**
DeepSeek-R1 37B	Raw Questions	✗	41	48	26	30
		✓	46	55	30	33
	Reformulated Questions	✗	62	53	26	31
		✓	64	73	35	48
	Combined Questions	✗	66	68	32	34
		✓	69	70	39	45

Table 3: Phi-4 and DeepSeek-R1 performance on reformulated questions of LCQ2's test set under different training configurations. Bold values indicate the best performance.

Approaches	QALD-10 (Wikidata)		LCQ1 (DBpedia)	
	BLEU	F1	BLEU	F1
T5-Small + Tag within + Copy (Diallo et al., 2024)	–	–	88	73
TSET (Triplet Structure Enhanced T5) (Qi et al., 2024)	–	51	–	–
SPINACH (Liu et al., 2024)	–	69	–	–
DFSL-MQ beam FS (D'Abramo et al., 2025)	–	62	–	–
GPT-4o (Direct QA) (Alekseev et al., 2025)	–	37	–	–
Phi-4 14B (baseline)	32	47	87	74
Our Phi-4 14B + Frames	37	59	89	83
Our Phi-4 14B + URIs Injection	54	69	89	84
Our Phi-4 14B + Frames + URIs Injection	58	**76**	92	**89**
DeepSeek-R1 37B (baseline)	24	40	84	63
Our DeepSeek-R1 37B + Frames	25	47	86	68
Our DeepSeek-R1 37B + URIs Injection	42	61	89	73
Our DeepSeek-R1 37B + Frames + URIs Injection	44	67	90	75

Table 4: Performance comparison of our approach and SOTA methods on the natural questions in QALD-10 and LC-QuAD1.0. The reformulated questions are considered for both training and testing. Bold and underlined values indicate the best F1 and BLEU respectively.

improve structural validity but require carefully designed intermediate objectives and incur costly pretraining.

Impact of frame-based representations Across both Unknown Template Split (UTS) and Reformulated Questions settings, augmenting questions with frame-based structured representations consistently boosts execution performance and improves robustness to surface variation. In UTS, BLEU still declines but far less with frames, while execution metrics rise markedly, indicating better intent capture under unseen surface forms. In Reformulated Questions, the same trend holds across all dataset variants: models trained and evaluated with frame-enhanced inputs outperform those using only raw questions, underscoring the value of structured semantic context throughout the pipeline.

Impact of combination of different questions types The highest F1-score is achieved using the Combined Questions configuration together with the frame representation, as shown in Table 3. When URIs are injected, both reformulated and combined questions achieve comparable F1-scores. Although the two training setups—(1) using Combined Questions and (2) using Reformulated Questions—perform similarly when tested on Reformulated Questions, the Combined Questions configuration shows a more noticeable improvement when evaluated on Raw or Combined Questions. This setting reflects realistic usage scenarios, where some user queries follow predictable patterns while others are more complex and less structured. However, since the primary focus is on Reformulated Questions, and to maintain clarity, those additional results are omitted from the main manuscript.

Impact of structured representation on embedding quality Our results also demonstrate that the frame-based structured representation brings different paraphrases of the same question closer in the embedding space, thereby enhancing the model's ability to handle lexical variation.

6. Conclusion

In this paper, we introduced FRASE, a method that enriches natural language questions with structured semantic representations to improve LLM generalization in SPARQL query generation. By leveraging frame semantics as an intermediate layer, FRASE mitigates LLM brittleness to lexical and syntactic variation—two key challenges in Question-2-SPARQL task. Experiments across multiple recent LLMs show that FRASE consistently boosts performance, especially on unseen templates and naturally expressed questions. These results highlight the value of integrating structured semantics to enhance robustness and abstraction. Beyond question answering, FRASE points toward broader applications of semantic structuring in prompt engineering, which we plan to explore in future work.

7. Limitations

While our method outperforms baselines on Wikidata and DBpedia, it has limitations. The best F1-scores remain modest, warranting further analysis of cases poorly handled by frame-based representations. Performance also depends on the quality of textual descriptions for KB relations and classes, with URIs-Injection acting as an idealized retriever. When descriptions are sparse or inconsistent, frame detection may yield noisy alignments which is a limit of less curated KBs. Finally, our implementation relies solely on English FrameNet, limiting applicability to English. Extending to multilingual settings would require high-quality multilingual resources or robust cross-lingual mappings, which remain non-trivial.

Acknowledgments

This project was undertaken thanks to funding from IVADO[4] and the Canada First Research Excellence Fund. We also acknowledge the financial support of the NSERC Discovery Grant Program, which contributed in part to this research. The authors further gratefully recognize Compute Canada (Calcul Québec) for providing the computational resources that made this work possible.

8. Bibliographical References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report.

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (ucca). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238.

Hessa Abdulrahman Alawwad, Areej Alhothali, Usman Naseem, Ali Alkhathlan, and Amani Jamal. 2024. Enhancing textbook question answering task with large language models and retrieval augmented generation. *arXiv preprint arXiv:2402.05128*.

Artem Alekseev, Mikhail Chaichuk, Miron Butko, Alexander Panchenko, Elena Tutubalina, and Oleg Somov. 2025. The benefits of query-based kgqa systems for complex and temporal questions in llm era. In *International Conference on Applications of Natural Language to Information Systems*, pages 426–441. Springer.

Kaikai An, Ce Zheng, Bofei Gao, Haozhe Zhao, and Baobao Chang. 2023. Coarse-to-fine dual encoders are better frame identification learners. *arXiv preprint arXiv:2310.13316*.

Collin F Baker, Michael Ellsworth, and Katrin Erk. 2007. Semeval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998a. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998b. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Debayan Banerjee, Pranav Ajit Nair, Jivat Neet Kaur, Ricardo Usbeck, and Chris Biemann. 2022. Modern baselines for sparql semantic parsing. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2260–2265.

Nikita Baramiia, Alina Rogulina, Sergey Petrakov, Valerii Kornilov, and Anton Razzhigaev. 2022.

[4] https://ivado.ca/

Ranking approach to monolingual question answering over knowledge graphs. In *NLIWoD@ ESWC*, pages 32–37.

Julia Bonn, Jeffrey Flanigan, Jan Hajic, Ishan Jindal, Yunyao Li, and Nianwen Xue. 2024. Meaning representations for natural languages: Design, models and applications. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 13–18.

Mihaela Bornea, Ramon Fernandez Astudillo, Tahira Naseem, Nandana Mihindukulasooriya, Ibrahim Abdelaziz, Pavan Kapanipathi, Radu Florian, and Salim Roukos. 2021. Learning to transpile amr into sparql. *arXiv preprint arXiv:2112.07877*.

Manuel Borroto, Francesco Ricca, Bernardo Cuteri, and Vito Barbara. 2022. Sparql-qa enters the qald challenge. In *Proceedings of the 7th Natural Language Interfaces for the Web of Data (NLIWoD) co-located with the 19th European Semantic Web Conference, Hersonissos, Greece*, volume 3196, pages 25–31.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752.

Laura Caspari, Kanishka Ghosh Dastidar, Saber Zerhoudi, Jelena Mitrovic, and Michael Granitzer. 2024. Beyond benchmarks: Evaluating embedding model similarity for retrieval augmented generation systems. *arXiv preprint arXiv:2407.08275*.

David Chanin. 2023. Open-source frame semantic parsing. *arXiv preprint arXiv:2303.12788*.

K Clark. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Ganqu Cui, Jie Zhou, Cheng Yang, and Zhiyuan Liu. 2020. Adaptive graph encoder for attributed graph embedding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 976–985.

Dipanjan Das, Desai Chen, André FT Martins, Nathan Schneider, and Noah A Smith. 2014. Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A Smith. 2010. Probabilistic frame-semantic parsing. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 948–956.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devasier, Rishabh Mediratta, and Chengkai Li. 2025. Can llms extract frame-semantic arguments?

Papa Abdou Karim Karou Diallo, Samuel Reyd, and Amal Zouaq. 2024. A comprehensive evaluation of neural sparql query generation from natural language questions. *IEEE Access*.

Papa Abdou Karim Karou Diallo and Amal Zouaq. 2025. Enhancing frame detection with retrieval augmented generation. *arXiv preprint arXiv:2502.12210*.

Dennis Diefenbach, Kamal Singh, and Pierre Maret. 2017. Wdaqua-core0: A question answering component for the research community. In *Semantic Web Evaluation Challenge*, pages 84–89. Springer.

Jiwei Ding, Wei Hu, Qixin Xu, and Yuzhong Qu. 2019. Leveraging frequent query substructures to generate formal queries for complex question answering. *arXiv preprint arXiv:1908.11053*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II*, volume 11779 of *Lecture Notes in Computer Science*, pages 69–78. Springer.

Jacopo D'Abramo, Andrea Zugarini, and Paolo Torroni. 2025. Investigating large language models for text-to-sparql generation. In *Proceedings of*

the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing, pages 66–80.

Vincent Emonet, Jerven Bolleman, Severine Duvaud, Tarcisio Mendes de Farias, and Ana Claudia Sima. 2024. Llm-based sparql query generation from natural language over federated knowledge graphs. *arXiv preprint arXiv:2410.06062*.

Charles J Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Jeffrey Flanigan, Ishan Jindal, Yunyao Li, Tim O'Gorman, Martha Palmer, and Nianwen Xue. 2022. Meaning representations for natural languages: Design, models and applications. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 1–8.

Shuzheng Gao, Chaozheng Wang, Cuiyun Gao, Xiaoqian Jiao, Chun Yong Chong, Shan Gao, and Michael Lyu. 2025. The prompt alchemist: Automated llm-tailored prompt optimization for test case generation. *arXiv preprint arXiv:2501.01329*.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shaoru Guo, Yong Guan, Ru Li, Xiaoli Li, and Hongye Tan. 2020. Incorporating syntax and frame semantics in neural network for machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2635–2641.

Silvana Hartmann, Ilia Kuznetsov, M Teresa Martín-Valdivia, and Iryna Gurevych. 2017. Out-of-domain framenet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 471–482.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Tianyu Jiang and Ellen Riloff. 2021. Exploiting definitions for frame identification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2429–2434.

Zhijing Jin, Yuen Chen, Fernando Gonzalez, Jiarui Liu, Jiayi Zhang, Julian Michael, Bernhard Schölkopf, and Mona Diab. 2024. Analyzing the role of semantic representations in the era of large language models. *arXiv preprint arXiv:2405.01502*.

Wei Ju, Yifan Wang, Yifang Qin, Zhengyang Mao, Zhiping Xiao, Junyu Luo, Junwei Yang, Yiyang Gu, Dongjie Wang, Qingqing Long, et al. 2024. Towards graph contrastive learning: A survey and beyond. *arXiv preprint arXiv:2405.11868*.

Aditya Kalyanpur, Or Biran, Tom Breloff, Jennifer Chu-Carroll, Ariel Diertani, Owen Rambow, and Mark Sammons. 2020. Open-domain frame semantic parsing using transformers. *arXiv preprint arXiv:2010.10998*.

Hans Kamp and Uwe Reyle. 2019. 11. discourse representation theory. *Semantics–Theories*, page 321.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Paul Kingsbury and Martha Palmer. 2003. Propbank: the next level of treebank. In *Proceedings of Treebanks and lexical Theories*, volume 3. Citeseer.

Paul R Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *LREC*, pages 1989–1993.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *COLING 1998 Volume 1: The 17th*

International Conference on Computational Linguistics.

Jaebok Lee and Hyeonjeong Shin. 2024. Sparkle: Enhancing sparql generation with direct kg integration in decoding. *arXiv preprint arXiv:2407.01626*.

Young-Suk Lee, Ramon Fernandez Astudillo, Thanh Lam Hoang, Tahira Naseem, Radu Florian, and Salim Roukos. 2021. Maximum bayes smatch ensemble distillation for amr parsing. *arXiv preprint arXiv:2112.07790*.

Alina Leidinger, Robert Van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? *arXiv preprint arXiv:2311.01967*.

Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024a. Making text embedders few-shot learners.

Ru Li, Yunxiao Zhao, Zhiqiang Wang, Xuefeng Su, Shaoru Guo, Yong Guan, Xiaoqi Han, and Hongyan Zhao. 2024b. A comprehensive overview of cfn from a commonsense perspective. *Machine Intelligence Research*, 21(2):239–256.

Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. 2023. Few-shot in-context learning for knowledge base question answering. *arXiv preprint arXiv:2305.01750*.

Sam Lin, Wenyue Hua, Lingyao Li, Zhenting Wang, and Yongfeng Zhang. 2025. Ado: Automatic data optimization for inputs in llm prompts. *arXiv preprint arXiv:2502.11436*.

Jiangming Liu, Shay Cohen, and Maria Lapata. 2018. Discourse representation structure parsing. In *56th Annual Meeting of the Association for Computational Linguistics*, pages 429–439. Association for Computational Linguistics (ACL).

Shicheng Liu, Sina J Semnani, Harold Triedman, Jialiang Xu, Isaac Dan Zhao, and Monica S Lam. 2024. Spinach: Sparql-based information navigation for challenging real-world questions. *arXiv preprint arXiv:2407.11417*.

Jack Longwell, Mahdiyar Ali Akbar Alavi, Fattane Zarrinkalam, and Faezeh Ensan. 2024. Triple augmented generative language models for sparql query generation from natural language questions. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 269–273.

Haoran Luo, Zichen Tang, Shiyao Peng, Yikai Guo, Wentai Zhang, Chenghao Ma, Guanting Dong, Meina Song, Wei Lin, et al. 2023. Chatkbqa: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models. *arXiv preprint arXiv:2310.08975*.

Kangqi Luo, Fengli Lin, Xusheng Luo, and Kenny Zhu. 2018. Knowledge base question answering via encoding of complex query graphs. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2185–2194.

Seyed Mahed Mousavi, Simone Alghisi, and Giuseppe Riccardi. 2024. Is your llm outdated? benchmarking llms & alignment algorithms for time-sensitive knowledge. *arXiv preprint arXiv:2404.08700*.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Niklas Muennighoff, SU Hongjin, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. In *ICLR 2024 Workshop: How Far Are We From AGI*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark.

Stephan Oepen, Ezra Callahan, Dan Flickinger, Christoper D Manning, and Kristina Toutanova. 2002. Lingo redwoods. In *First Workshop on treebanks and linguistic theories*.

Aleksandr Perevalov, Dennis Diefenbach, Ricardo Usbeck, and Andreas Both. 2022a. Qald-9-plus: A multilingual dataset for question answering over dbpedia and wikidata translated by native speakers. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pages 229–234. IEEE.

Alexander Perevalov, Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Aina Hoffmann, Giuseppe Abrami, Chris Biemann, Andreas Both, Antje Fischer, Jan Hladky, Jan Kalo, Julian Krause, Christian Möller, Giulio Napolitano, Daniel Ruffinelli, Gëzim Sejdiu, Simon Walter, Nikita Zhiltsov, and Julius Zöllner. 2022b. Knowledge graph question answering leaderboard: A community resource to prevent a replication crisis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3004–3013, Marseille, France. European Language Resources Association.

Alexander Popov and Jennifer Sikos. 2019. Graph embeddings for frame identification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 939–948.

Jiexing Qi, Chang Su, Zhixin Guo, Lyuwen Wu, Zanwei Shen, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. 2024. Enhancing sparql query generation for knowledge base question answering systems by learning to correct triplets. *Applied Sciences*, 14(4):1521.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

MD Rahman, Feiroz Humayara, Syed Maudud E Rabbi, and Muhammad Mahbubur Rashid. 2024. Efficient medical image retrieval using densenet and faiss for birads classification. *arXiv preprint arXiv:2411.01473*.

Michael Regan, Shira Wein, George Baker, and Emilio Monti. 2024. Massive multilingual abstract meaning representation: A dataset and baselines for hallucination detection. *arXiv preprint arXiv:2405.19285*.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Samuel Reyd and Amal Zouaq. 2023. Assessing the generalization capabilities of neural machine translation models for sparql query generation. In *International Semantic Web Conference*, pages 484–501. Springer.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Cezar Sas, Meriem Beloucif, and Anders Søgaard. 2020. Wikibank: Using wikidata to improve multilingual frame-semantic parsing. In *The 12th Language Resources and Evaluation Conference*, pages 4183–4189.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, pages 593–607. Springer.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. *arXiv preprint arXiv:2109.08535*.

Aditya Sharma, Luis Lara, Amal Zouaq, and Christopher J Pal. 2025. Reducing hallucinations in language model-based sparql query generation using post-generation memory retrieval. *arXiv preprint arXiv:2502.13369*.

Hassan S Shavarani and Anoop Sarkar. 2024. Entity retrieval for answering entity-centric questions. *arXiv preprint arXiv:2408.02795*.

Kaize Shi, Xueyao Sun, Qing Li, and Guandong Xu. 2024. Compressing long context for enhancing rag with amr-based concept distillation. *arXiv preprint arXiv:2405.03085*.

Kanchan Shivashankar, Khaoula Benmaarouf, and Nadine Steinmetz. 2022. From graph to graph: Amr to sparql. In *Proceedings of the 7th Natural Language Interfaces for the Web of Data (NLIWoD) co-located with the 19th European Semantic Web Conference (ESWC 2022), Hersonissos, Greece, 29th May*.

Yuqi Si and Kirk Roberts. 2018. A frame-based nlp system for cancer-related information extraction. In *AMIA annual symposium proceedings*, volume 2018, page 1524. American Medical Informatics Association.

Xuefeng Su, Ru Li, Xiaoli Li, Baobao Chang, Zhiwei Hu, Xiaoqi Han, and Zhichao Yan. 2023. A span-based target-aware relation model for frame-semantic parsing. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(3):1–24.

Xuefeng Su, Ru Li, Xiaoli Li, Jeff Z Pan, Hu Zhang, Qinghua Chai, and Xiaoqi Han. 2021. A knowledge-guided framework for frame identification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5230–5240.

I Sutskever. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *arXiv preprint arXiv:1706.09528*.

Fabio Tamburini. 2022. Combining electra and adaptive graph encoding for frame identification. In *Proceedings of the Thirteenth Language*

Resources and Evaluation Conference, pages 1671–1679.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023a. Towards benchmarking and improving the temporal reasoning capability of large language models. *arXiv preprint arXiv:2306.08952*.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023b. Towards robust temporal reasoning of large language models via a multi-hop qa dataset and pseudo-instruction tuning. *arXiv preprint arXiv:2311.09821*.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023c. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. In *International Semantic Web Conference*, pages 348–367. Springer.

Prashant Trivedi, Souradip Chakraborty, Avinash Reddy, Vaneet Aggarwal, Amrit Singh Bedi, and George K Atia. 2025. Align-pro: A principled approach to prompt optimization for llm alignment. *arXiv preprint arXiv:2501.03486*.

Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. 2017. Lc-quad: A corpus for complex question answering over knowledge graphs. In *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II*, volume 10588 of *Lecture Notes in Computer Science*, pages 210–218. Springer.

Ricardo Usbeck, Xi Yan, Aleksandr Perevalov, Longquan Jiang, Julius Schulz, Angelie Kraft, Cedric Möller, Junbo Huang, Jan Reineke, Axel-Cyrille Ngonga Ngomo, et al. 2024. Qald-10–the 10th challenge on question answering over linked data: Shifting from dbpedia to wikidata as a kg for kgqa. *Semantic Web*, 15(6):2193–2207.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Cunxiang Wang, Zhikun Xu, Qipeng Guo, Xiangkun Hu, Xuefeng Bai, Zheng Zhang, and Yue Zhang. 2023. Exploiting abstract meaning representation for open-domain question answering. *arXiv preprint arXiv:2305.17050*.

Wei-Hung Weng, Andrew Sellergen, Atilla P Kiraly, Alexander D'Amour, Jungyeon Park, Rory Pilgrim, Stephen Pfohl, Charles Lau, Vivek Natarajan, Shekoofeh Azizi, et al. 2024. An intentional approach to managing bias in general purpose embedding models. *The Lancet Digital Health*, 6(2):e126–e130.

Aaron Steven White, Elias Stengel-Eskin, Siddharth Vashishtha, Venkata Govindarajan, Dee Ann Reisinger, Tim Vieira, Keisuke Sakaguchi, Sheng Zhang, Francis Ferraro, Rachel Rudinger, et al. 2019. The universal decompositional semantics dataset and decomp toolkit. *arXiv preprint arXiv:1909.13851*.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Hamada M Zahera, Manzoor Ali, Mohamed Ahmed Sherif, Diego Moussallem, and A-C Ngonga Ngomo. 2024. Generating sparql from natural language using chain-of-thoughts prompting.

Daniel Zeman and Jan Hajic. 2020. Fgd at mrp 2020: Prague tectogrammatical graphs. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 33–39.

Jiahuan Zhang, Tianheng Wang, Hanqing Wu, Ziyi Huang, Yulong Wu, Dongbai Chen, Linfeng Song, Yue Zhang, Guozheng Rao, and Kaicheng Yu. 2025. Sr-llm: Rethinking the structured representation in large language model. *arXiv preprint arXiv:2502.14352*.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024a. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*.

Zihan Zhang, Meng Fang, and Ling Chen. 2024b. Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. *arXiv preprint arXiv:2402.16457*.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

Ce Zheng, Xudong Chen, Runxin Xu, and Baobao Chang. 2022. A double-graph based framework for frame semantic parsing. *arXiv preprint arXiv:2206.09158*.

Ce Zheng, Yiming Wang, and Baobao Chang. 2023a. Query your model with definitions in framenet: an effective method for frame semantic role labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14029–14037.

Hai-Tao Zheng, Zuotong Xie, Wenqiang Liu, Dongxiao Huang, Bei Wu, and Hong-Gee Kim. 2023b. Prompt learning with structured semantic knowledge makes pre-trained language models better. *Electronics*, 12(15):3281.

Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. 2021. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5822–5834.

A. Splits Statistics

Table 5 show the distribution of splits sizes for the different splits we have in LCQ2 and Figures 5a and 5b show the distribution of question lengths (in number of words) and the corresponding statistics (minimum, mean, maximum) for the training, validation, and test sets in the two LCQ2 splits: the **Original Split** and the **Unknown Template Split**. In the **Original Split**, the length distributions are relatively consistent across all subsets, with similar averages, indicating structural alignment between training and test data. In contrast, the **Unknown Template Split** reveals a clear discrepancy: test questions are, on average, significantly longer than those in the training and validation sets. This shift reflects the intended challenge of the split, where models must generalize to unseen question templates, which tend to be more complex and verbose. Such distributional differences likely contribute to the performance drop observed in this setting, particularly for surface-level generation metrics like BLEU.

Splits		Statistics				
		Total	**Train**	**Validation**	**Test**	**Unseen**
OS	Global templates	30	30	30	30	0
	Entries	30225	21761	2418	6046	6046
	Avg Query-Length	-	17	18	18	-
UTS	Global templates	30	24	6	6	6
	Entries	30225	24178	3023	3024	3024
	Avg Query-Length	-	16	16	36	-

Lengths are measured in term of number of words.
OS: Original Split - UTS: Unknown Template Split.

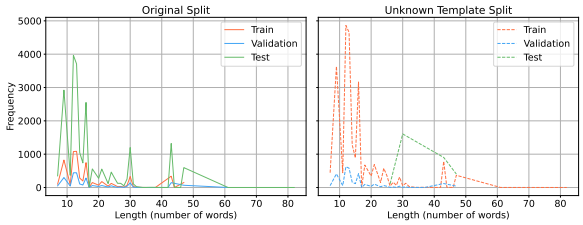
Table 5: LCQ2 statistics in terms of global templates and entries across data splits.

B. FRASE additional detail

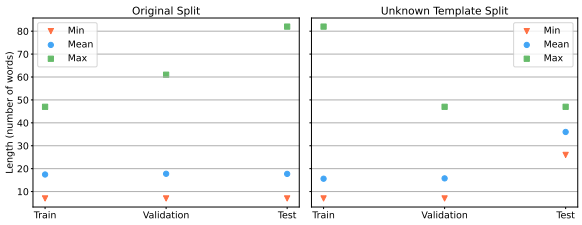
The Algorithm 1 shows the Stage 1 of FRASE that detects the frames evoked in any LCQ2 question as depicted in first part of Figure 6.

Algorithm 1 Identify Frames for LCQ2 Questions

Require: LCQ2_question (natural language question), SPARQL_query (corresponding SPARQL query), VectorDatabase (stores frame vectors), KB (Knowledge Base with ontology and descriptions)
Ensure: EvokedFrames (set of frames evoked by the question)
Initialize an empty set `EvokedFrames`
Preprocess Frames:
for each frame in the set of available frames **do**
    Represent the frame using its name, description, and list of frame-elements
    Encode the frame representation into a vector using the embedding-model
    Store the frame vector in `VectorDatabase`
end for
Extract Relevant KB Elements:
Parse `SPARQL_query` to extract relevant KB element identifiers (URIs) corresponding to relation or class
Generate KB Element Representations:
for each URI in the extracted URIs **do**
    Fetch the label and textual description of the corresponding KB element from `KB`
    Encode the label and description using the embedding-model
end for
Align KB Elements with Frames:
for each vector representation of a KB element **do**
    Perform a similarity search in `VectorDatabase` to find the most similar frame vector(s)
    if a match is found (similarity score $\geq$ threshold) **then**
        Add the matched top-$k = 1$frame(s) to `EvokedFrames`
    end if
end for
return `EvokedFrames`



(a) Distribution of query length (number of words).



(b) AVG/Min/Max query length (number of words).

Figure 5: Statistics of query length in Original Split and Unknown Template Split.

C. LCQ2 Questions Annotation by FRASE

Figure 6 illustrates how our proposed FRASE pipeline semantically enriches a natural language question from LCQ2 using frame-based structured representations. In **Stage 1**, each relation URI in the associated SPARQL query is aligned with a corresponding FrameNet frame based on textual similarity. For instance, the relation 'wdt:P1365' ("replaces") is aligned with the **Replacing** frame, and "wdt:P31" ("instance of") is mapped to **Identicality**. In **Stage 2**, the system identifies the relevant **Frame Elements** and links them to corresponding spans in the question text. In this example, the element "Old" is mapped to "Yuan dynasty", and "Type" is inferred as "Dynasty". This structured representation captures the underlying semantic roles involved in the question and provides an interpretable abstraction that can be used to improve SPARQL generation and generalization.

D. Experimental details

We fine-tune all models using the QLoRA method, which combines 4-bit quantization with parameter-efficient fine-tuning. Specifically, we quantize the base model weights using NF4 quantization and bfloat16 computation via the `BitsAndBytesConfig`. We then apply a low-rank adaptation (LoRA) on key components of the transformer layers (e.g., `q_proj`, `k_proj`, `v_proj`, etc.) with rank 16, `lora_alpha` 16, and no dropout. The fine-tuning is performed using the `adamw_8bit` optimizer with a learning rate of 2×10^{-4}, and linear scheduling. We save and evaluate the model at the end of each epoch and report the best checkpoint based on validation loss. Further training parameters are detailed in Table 6.

Parameters	Values
Max Sequence Length	2048
Packing	False (for faster training)
Per Device Batch Size	8
Gradient Accumulation Steps	4
Warmup Steps	5
Number of Epochs	10 (adjustable)
Learning Rate	2e-4
Precision Mode	bfloat16
Quantization Type	4-bit (NF4)
LoRA Rank (r)	16
LoRA Alpha	16
LoRA Dropout	0
Target Modules	Attention + MLP Projections[1]
Optimizer	adamw_8bit
Weight Decay	0.01
Learning Rate Scheduler	Linear
Random Seed	1618
Evaluation Strategy	Epoch

[1] `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`

Table 6: Technical details of the fine-tuning

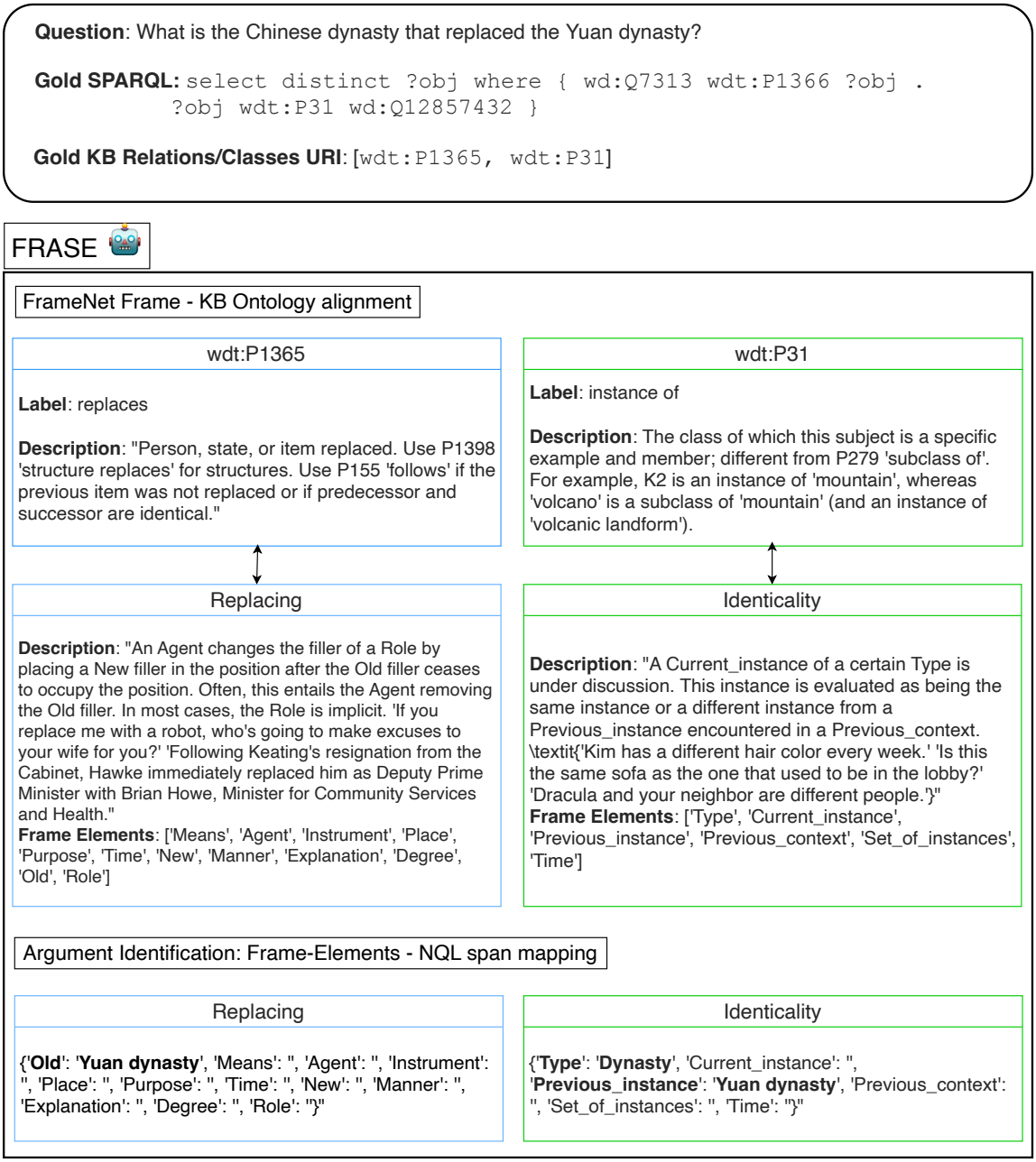


Figure 6: Illustration of the process of adding structured representations to a question: From KB–FrameNet alignment to argument identification.