Geração de consultas SPARQL a partir de linguagem natural

Heber Gustavo Xavier de Castro
heber.castro@usp.br

Cléver Ricardo Guareis de Farias
farias@ffclrp.usp.br

Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP)
Universidade de São Paulo (USP)
Ribeirão Preto, SP, Brasil

Resumo

The Semantic Web aims to make web data understandable not only to humans but also to machines, enabling more efficient data integration, sharing, and reuse. Linked Open Data (LOD) initiatives have supported this vision by promoting the publication of semantically annotated and interconnected data. However, querying LOD repositories typically requires knowledge of SPARQL, a complex query language that limits access for non-expert users. Although several approaches have been proposed to automatically generate SPARQL queries from natural-language questions, most are designed for English and are tightly coupled to specific domains, which hinders reuse. This article presents a generic, domain-independent approach for generating SPARQL queries from questions written in Portuguese. The proposed method uses reference questions, parameterized query templates, and a synonym dictionary enriched by lexical resources and similarity metrics. The implementation is supported by the Natural2SPARQL tool, and the approach is validated through a case study in the financial domain using real data from the Brazilian stock exchange (B3). The results indicate that the method enables flexible and semantically accurate SPARQL query generation from natural-language input. Unlike learning-based approaches, our method avoids retraining and achieves up to 93.3% end-to-end success in controlled settings, demonstrating robustness and low adaptation cost.

1 Introdução

A evolução da Web para uma plataforma dinâmica gerou um crescimento exponencial de dados heterogêneos. Contudo, grande parte dessas informações permanece inacessível ao processamento computacional automatizado, devido à ausência de uma estrutura semântica explícita que permita sua interpretação por máquinas.

Com o objetivo de superar esse desafio, a *Web Semântica* foi proposta por Tim Berners-Lee (Berners-Lee et al., 2001) como uma extensão da web atual. Essa visão introduz um conjunto de tecnologias e padrões que viabilizam a representação, interconexão e consulta de dados de forma estruturada e semântica. Um dos pilares dessa iniciativa são os *Dados Abertos Ligados* (Linked Open Data — LOD), que consistem em conjuntos de dados interligados e semanticamente anotados, publicados na web com padrões abertos.

A exploração desses dados via SPARQL (Harris e Seaborne, 2013), embora poderosa, exige domínio de uma sintaxe complexa, restringindo o acesso a especialistas. Para superar essa barreira técnica, diversas abordagens propõem a geração automática de consultas SPARQL a partir de perguntas em linguagem natural, visando democratizar o acesso aos dados (Yahya et al., 2012; Paredes-Valverde et al., 2016; Dorabăţ e Posea, 2020; Chen et al., 2021; Miranda et al., 2024; Varagnolo et al., 2023).

Entretanto, a maioria dessas abordagens apresenta duas limitações recorrentes. A primeira refere-se à sua restrição ao idioma inglês, dificultando sua aplicação em línguas como o português. A segunda diz respeito à forte dependência de domínios específicos, o que inviabiliza sua reutilização em contextos distintos sem modificações substanciais.

Neste trabalho, propomos uma abordagem genérica e independente de domínio para a geração de consultas SPARQL a partir de consultas formuladas em linguagem natural em português. A independência de domínio refere-se à arquitetura e ao processo de transformação propostos, e não à validação empírica em múltiplos domínios. A proposta baseia-se em cinco etapas: i) elaboração de perguntas de referência; ii) definição de templates de consulta parametrizados; iii) enriquecimento da ontologia com sinônimos e termos relaciona-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 676–686

dos; iv) associação automática entre pergunta e template; e v) geração da consulta SPARQL final. A implementação da abordagem é realizada por meio da ferramenta *Natural2SPARQL*, que incorpora técnicas de Processamento de Linguagem Natural (PLN), como análise sintática, reconhecimento de entidades nomeadas e cálculo de similaridade semântica.

Para validar a proposta, conduzimos um estudo de caso no domínio financeiro, com foco em dados de negociação de ações na bolsa de valores brasileira (B3). O estudo envolveu o desenvolvimento de uma ontologia de domínio, a criação de um dicionário de sinônimos e de entidades nomeadas, com o apoio de WordNet (Miller, 1995) e ChatGPT-4 (OpenAI, 2025a), a geração de templates SPARQL e a construção de triplas RDF a partir de dados reais de negociação da B3.

Os resultados demonstram que a abordagem permite transformar perguntas complexas em consultas SPARQL executáveis de maneira flexível e com reduzido esforço de adaptação. A modularidade da arquitetura facilita sua reutilização em novos domínios, desde que acompanhada pela definição dos artefatos correspondentes.

Nesse contexto, abordagens baseadas exclusivamente em aprendizado de máquina, embora eficazes em cenários específicos, tendem a impor elevado custo de adaptação, dependência de retreinamento e limitada interpretabilidade, dificultando sua reutilização em novos domínios e idiomas. Assim, este trabalho parte da premissa de que uma abordagem baseada em engenharia de conhecimento reutilizável, fundamentada em perguntas de referência, templates parametrizados e recursos léxico-semânticos, pode oferecer uma alternativa de baixo custo, modular e robusta para a geração de consultas SPARQL em português por usuários não especialistas.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta os principais trabalhos relacionados; a Seção 3 descreve a abordagem de transformação proposta; a Seção 4 detalha os principais aspectos da ferramenta de suporte *Natural2SPARQL*; a Seção 5 apresenta o estudo de caso e os resultados da validação; e, finalmente, a Seção 7 apresenta as conclusões, limitações e sugestões para trabalhos futuros.

2 Trabalhos Relacionados

Diversas abordagens têm sido propostas para gerar consultas SPARQL via processamento de linguagem natural. Yahya et al. (Yahya et al., 2012) introduzem o framework DEANNA, que detecta frases na pergunta e as mapeia para itens semânticos. Utilizando análise de dependência, geram-se *Q-Units* (triplas potenciais). O sistema aplica uma desambiguação conjunta para resolver conflitos e agrupa os itens em triplas compatíveis com a base de conhecimento para a geração final do SPARQL.

Paredes-Valverde et al. (Paredes-Valverde et al., 2016) propõem uma extração baseada no processamento prévio da ontologia para criar um domínio léxico, permitindo buscas via descrições textuais. Diferente de outros métodos, as respostas são extraídas da ontologia e não de uma base RDF. A abordagem processa a pergunta para determinar relações semânticas e gera um modelo de pergunta para inferir o padrão SPARQL adequado.

Hamon et al. (Hamon et al., 2017) focam no domínio biomédico com um pipeline onde perguntas sofrem anotação linguística e semântica para extração de entidades. O processo segue com a abstração da pergunta para identificar elementos relevantes que, unidos, geram o padrão gráfico e a cadeia de consulta SPARQL final. Para auxiliar no mapeamento, a abordagem utiliza bancos de recursos semânticos externos específicos da área da saúde. Além disso, a lógica de construção permite a integração de predicados provenientes de diferentes conjuntos de dados por meio da propriedade *sameAs*, culminando na definição automática do tipo de cláusula de retorno, seja ela `SELECT` ou `ASK`.

Song et al. (Song et al., 2019) utilizam Grafos de Consulta Semântica (SQG). A abordagem inicia com a análise de dependência para construir o SQG, cujos vértices (entidades, variáveis) e arestas (relações) são associados à base de conhecimento via pontuação de similaridade. A consulta SPARQL é gerada percorrendo este grafo, onde cada aresta representa uma condição da consulta.

Dorabat e Posea (Dorabăţ e Posea, 2020) apresentam o framework onIQ, independente de ontologia. O método realiza análise sintática para identificar variáveis alvo e armazena triplas processadas em estruturas chamadas Metatriplas. Para evitar ambiguidades, verbos são mapeados para propriedades utilizando o WordNet (Miller, 1995), culminando na construção da consulta a partir das triplas filtradas.

Chen et al. (Chen et al., 2021) introduzem o QAWizard, uma abordagem que trata o mapeamento de linguagem natural para SPARQL como um problema de classificação multiclasse (Aly, 2005). O sistema opera em duas fases distintas: treinamento e consulta. Na fase de treinamento, utiliza-se um Modelo de Markov de Entropia Máxima (MEMM) (McCallum et al., 2000) para estimar a probabilidade de etiquetas semânticas, permitindo identificar tipos de entidades (como classes, propriedades ou variáveis) e seus respectivos papéis em triplas RDF (sujeito, predicado ou objeto). Durante a fase de consulta, o QAWizard realiza o pré-processamento linguístico e a anotação de termos em categorias, recorrendo a bases externas como DBpedia para a vinculação e desambiguação de entidades.

Smeros et al. (Smeros et al., 2025) introduziram o SPARQL-LLM, um sistema agnóstico a *triple-stores* projetado para gerar consultas em tempo real sobre grafos de conhecimento federados e descentralizados. O sistema utiliza metadados leves para fornecer ao modelo de linguagem uma compreensão precisa da estrutura do grafo, sem a necessidade de processar toda a base de dados. A abordagem incorpora um ciclo iterativo de validação e correção: quando uma falha na execução é detectada, as mensagens de erro do banco de dados são reinseridas no fluxo de geração, permitindo que o modelo refine a consulta em até três etapas de revisão.

No contexto da língua portuguesa, Miranda et al. (Miranda et al., 2024) propõem o sistema modular KGQAPT. A abordagem sequencial combina análise da pergunta, classificação do tipo de consulta e um mapeamento de frases que utiliza vinculação de entidades para relações. O trabalho utiliza um classificador *Random Forest* (Breiman, 2001) para identificar previamente a intenção da pergunta entre diferentes tipos de consultas SPARQL, como `SELECT` ou `ASK`. Além disso, a arquitetura emprega um modelo de *Tree-LSTM* (Tai et al., 2015) para ranquear as consultas candidatas, baseando-se na similaridade estrutural entre a árvore de dependência sintática da questão original e a representação lógica da consulta gerada.

Finalmente, Varagnolo et al. (Varagnolo et al., 2023) descreveram uma abordagem para a exploração de arquivos históricos baseada na ontologia CIDOC-CRM. O sistema utiliza análise sintática profunda para construir estruturas de representação de discurso (DRS), tratando a tradução para SPARQL como um problema de satisfação de restrições (CSP) resolvido por meio de otimização multi-objetivo. Essa modelagem matemática permite que o sistema identifique o caminho de navegação mais coerente dentro do grafo de conhecimento, garantindo alta fidelidade na tradução de sentenças.

Essas diferentes abordagens indicam que a geração de consultas SPARQL a partir de linguagem natural é predominantemente conduzida por soluções fortemente dependentes de domínio ou baseadas em aprendizado de máquina, as quais geralmente exigem re-treinamento, grandes volumes de dados anotados e apresentam limitada interpretabilidade. Adicionalmente, a maioria dessas abordagens é voltada ao idioma inglês, restringindo sua aplicabilidade ao português. Em contraste, a abordagem proposta neste trabalho adota uma estratégia baseada em engenharia de conhecimento reutilizável, combinando perguntas de referência, templates parametrizados e recursos léxico-semânticos, de forma independente de domínio e orientada ao português. Essa combinação permite reduzir o custo de adaptação a novos contextos, mantendo flexibilidade e robustez semântica, posicionando a proposta como uma alternativa complementar às soluções existentes.

3 Abordagem de Transformação

Nossa abordagem para a transformação de consultas em linguagem natural em consultas SPARQL busca facilitar a interação entre usuários e bases de dados semânticas, permitindo que consultas complexas sejam expressas de forma intuitiva. A Figura 1 apresenta as diferentes etapas dessa abordagem.

A primeira etapa, chamada de *Elaboração de Perguntas de Referência*, produz um conjunto de perguntas que representam os tipos de intenção de consulta mais comuns dentro de um domínio de interesse.

A segunda etapa, denominada *Definição dos Templates de Consulta*, consiste em mapear categorias de perguntas em linguagem natural para estruturas correspondentes de consultas SPARQL. A partir do conjunto de *Perguntas de Referência*, é definido um template SPARQL parametrizado para cada tipo de pergunta. Esses templates incorporam a lógica de consulta ao grafo de conhecimento e contêm marcadores de posição que são preenchidos dinamicamente com as informações extraídas da *Pergunta do Usuário*.

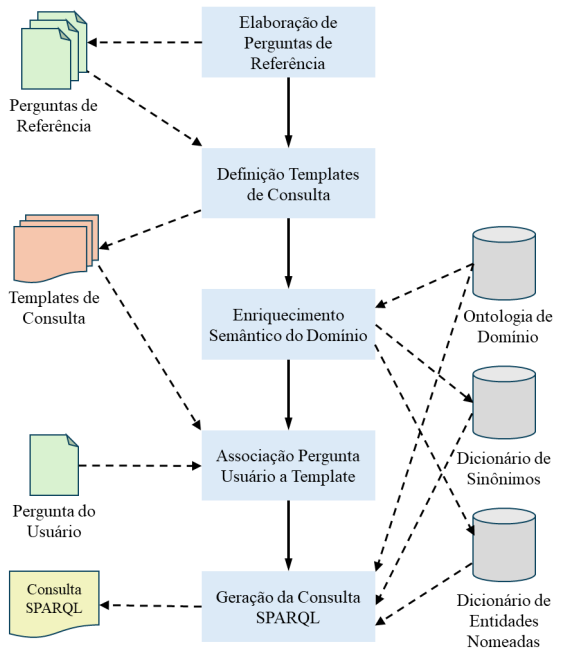

Figura 1: Abordagem de Transformação. Retângulos azuis representam atividades do processo de transformação. Linhas sólidas com ponta de seta conectando duas atividades indicam a ordem em que essas atividades são executadas. Retângulos verdes com o canto superior direito cortado representam perguntas. Cilindros cinza representam repositórios de conhecimento. Retângulos rosa com base curva representam templates de consulta. Por fim, um retângulo amarelo com base curva representam uma consulta SPARQL pronta para execução.

A terceira etapa, chamada de *Enriquecimento da Ontologia de Domínio*, busca produzir um *Dicionário de Sinônimos* e um *Dicionário de Entidades Nomeadas*. O *Dicionário de Sinônimos* pode ser criado associando-se cada termo definido na ontologia de domínio a um conjunto de sinônimos extraídos com base em informações providas por fontes externas de conhecimento, tais como WordNet (Miller, 1995) e o ChatGPT (OpenAI, 2025a). Já o *Dicionário de Entidades Nomeadas* pode ser criado com base em instâncias definidas na ontologia de domínio ou algum corpus específico contendo informações sobre o domínio de interesse.

Para cada sinônimo incluído no dicionário é atribuído um peso com base em dois critérios. O primeiro critério é a presença do sinônimo nas descrições das entidades e propriedades da ontologia de domínio (anotação *rdfs:comment*). Neste caso, o sinônimo recebe o peso máximo atribuído (1,0), pois é considerado um sinônimo mais próximo (semanticamente mais representativo) do conceito descrito na ontologia. O segundo critério baseia-se na aplicação de um algoritmo de similaridade semântica, que avalia o grau de proximidade entre a palavra principal e o sinônimo associado. Neste trabalho, propomos a utilização do algoritmo de Similaridade de Cosseno (Cosine Similarity) (Mikolov et al., 2013). Esse algoritmo transforma tanto a palavra principal quanto cada sinônimo em vetores numéricos de alta dimensão. Em seguida, calcula-se a similaridade entre esses vetores com base no cosseno do ângulo entre eles, resultando em um valor (ou score) entre 0 e 1, que representa o grau de semelhança semântica entre os termos. Esse valor é utilizado como o peso atribuído ao sinônimo: quanto mais próximo de 1, maior a similaridade conceitual em relação à palavra principal. Essa métrica foi adotada por oferecer baixo custo computacional, interpretabilidade e independência de modelos treinados, em alinhamento com o objetivo de uma abordagem leve, modular e de fácil adaptação.

A quarta etapa, denominada *Associação da Pergunta do Usuário ao Template*, tem como objetivo identificar o template de consulta mais adequado para responder à pergunta formulada pelo usuário. O processo se inicia com a comparação da pergunta do usuário com um conjunto de perguntas de referência associadas a cada template, utilizando o algoritmo de Similaridade de Cosseno. Com base nos scores obtidos, seleciona-se o template cuja pergunta de referência apresenta a maior similaridade com a entrada do usuário. Em seguida, a pergunta é submetida a uma análise sintática por meio de técnicas como *análise de dependência (dependency parsing)*, com o intuito de extrair suas estruturas gramaticais mais relevantes. Essa análise permite: 1) identificar os papéis sintáticos dos elementos da sentença, como sujeito, predicado e objeto, e 2) estabelecer as relações de dependência entre as palavras, viabilizando a construção de uma representação estruturada da pergunta.

Por fim, a quinta e última etapa, denominada *Geração da Consulta SPARQL*, tem como objetivo materializar a consulta executável a partir dos artefatos produzidos nas etapas anteriores. Esse processo ocorre em duas fases distintas de substituição, garantindo uma separação clara entre a semântica da pergunta do usuário e a estrutura da ontologia de domínio.

Na primeira fase, os marcadores (*placeholders*) específicos associados ao template identificado na etapa anterior são dinamicamente substituídos por

valores concretos obtidos da pergunta do usuário, os quais foram também extraídos na etapa anterior. O resultado dessa etapa é uma consulta SPARQL intermediária, ainda genérica e independente dos detalhes específicos da ontologia. Na segunda fase, os valores concretos associados aos marcadores específicos são mapeados para os URIs ou literais específicos da ontologia e prefixos RDF são adicionados, resultando em consulta SPARQL pronta para ser executada. Este processo de mapeamento é realizado com o auxílio do Dicionário de Sinônimos e do Dicionário de Entidades Nomeadas produzidos na etapa de *Enriquecimento da Ontologia de Domínio*.

4 Ferramenta Natural2SPARQL

A ferramenta *Natural2SPARQL*[1] foi desenvolvida para prover suporte à transformação de uma pergunta em linguagem natural em uma consulta SPARQL, particularmente às duas últimas etapas da nossa abordagem de transformação. *Natural2SPARQL* foi concebida para ser executada em ambiente web seguindo uma arquitetura cliente-servidor clássica. Adicionalmente, prezou-se pela modularidade da solução, permitindo sua aplicação em diferentes domínios de conhecimento. A Figura 2 ilustra os principais elementos da arquitetura da ferramenta *Natural2SPARQL*.

Natural2SPARQL Web Interface representa o componente responsável pela interface gráfica com o usuário da ferramenta. Este componente basicamente permite que o usuário faça uma pergunta em linguagem natural que é então encaminhada para ser transformada. Após o processo de transformação, esse componente exibe o código SPARQL correspondente e disponibiliza a opção para executar a consulta internamente na ferramenta.

Natural Language Processing (NLP) Controller é o principal componente do back-end, sendo responsável por realizar a interação com *Natural2SPARQL Web Interface* e orquestrar internamente a transformação e execução de uma consulta SPARQL. Após receber uma consulta em linguagem natural enviada pelo componente *Natural2SPARQL Web Interface*, *NLP Controller* basicamente executa a etapa quatro da nossa abordagem de transformação.

NLP Controller inicialmente irá selecionar o template de consulta mais adequado para a geração da consulta SPARQL. Para isso, este componente compara a pergunta submetida com a pergunta de referência base de cada template de consulta disponível usando o algoritmo de *Similaridade de Cosseno*. *NLP Controller* irá então selecionar o template de consulta com o score mais alto. Na sequência, *NLP Controller* processa a pergunta do usuário de modo a extrair a estrutura gramatical da pergunta usando técnicas de etiquetagem gramatical (Cutting et al., 1992) e a análise das dependências sintáticas entre os elementos da sentença. Também realiza a extração de entidades presentes usando uma técnica de reconhecimento de entidades nomeadas (Mohit, 2014) e *Dicionário de Entidades Nomeadas*, próprio para o domínio de conhecimento.

Ao término deste processo, *NLP Controller* gera uma estrutura de dados JSON concisa, contendo o identificador do template mais apropriado e um mapeamento das entidades extraídas da pergunta. Estas informações são então encaminhadas para o componente *SPARQL Query Builder* para a execução da última etapa da abordagem de transformação. *SPARQL Query Builder* representa o componente responsável por materializar a consulta SPARQL em um formato executável. Esse processo baseia-se em um mecanismo de substituição em duas fases, projetado para oferecer máxima flexibilidade.

Na primeira fase, denominada *Substituição de Entidades*, *SPARQL Query Builder* seleciona o template base para a transformação e realiza a substituição inicial dos marcadores genéricos do template pelos valores extraídos por *NLP Controller*. O resultado é uma consulta ainda abstrata, usada como base para a etapa seguinte, chamada *Substituição de Marcadores RDF*. Nesta etapa, ocorre o mapeamento dos valores substituídos na primeira etapa para os termos RDF específicos definidos na ontologia. Este processo é realizado com o auxílio de um Dicionário de Sinônimos e mediante a consulta aos termos presentes na ontologia usando o componente *SPARQL Query*. Ao final deste processo, temos uma consulta SPARQL final, completa e sintaticamente correta, a qual é retornada para o usuário.

Finalmente, o componente *SPARQL Query* é responsável por acessar a ontologia base do processo de transformação de modo a auxiliar o componente *SPARQL Query Builder*, bem como acessar o conjunto de triplas RDF para permitir a execução da consulta SPARQL gerada caso o usuário assim o

[1]Disponível em: https://github.com/profhebercastro/NATURAL2SPARQL_V2

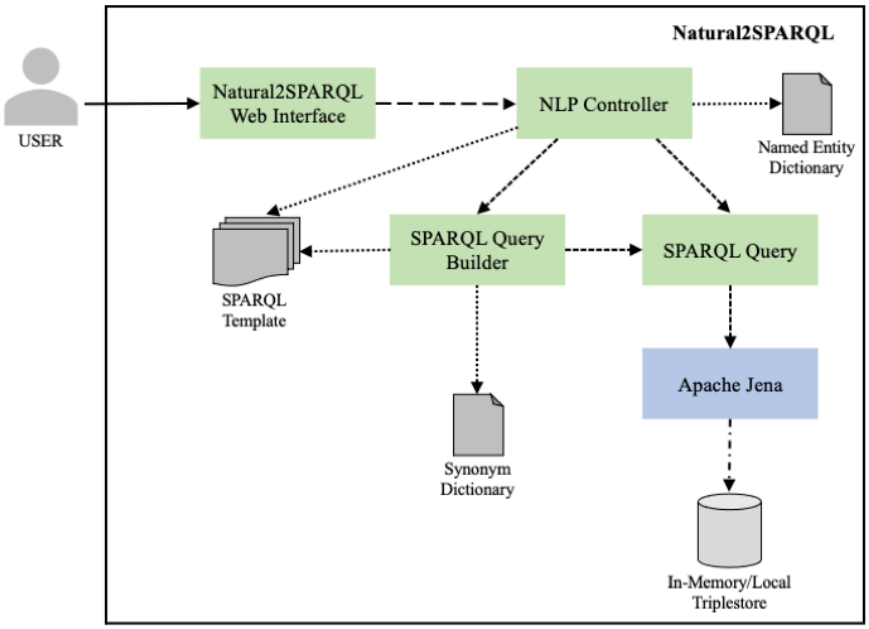

Figura 2: Arquitetura da ferramenta Natural2SPARQL. Retângulos verdes representam componentes desenvolvidos internamente, enquanto um retângulo azul representa um componente reutilizado. Elementos em cinza representam artefatos de conhecimento. Uma seta sólida indica a interação do usuário com a interface web. Uma seta com tracejado longo indica uma interação realizada via protocolo HTTP. Setas com tracejado curto indicam interações internas entre componentes da ferramenta. Setas pontilhadas indicam a interação de um componente com um artefato utilizado na construção da consulta. Por fim, setas tracejadas com ponto indicam a interação da ferramenta com conjuntos de triplas armazenados em memória ou em arquivos locais.

queira, concluindo assim o ciclo de transformação e execução da consulta. *SPARQL Query* utiliza a biblioteca *Apache Jena* para acessar a ontologia/triplas mantidas em memória.

5 Avaliação

Para validar a abordagem e a ferramenta *Natural2SPARQL*, foi conduzido um estudo de caso compreensivo em um cenário realista, seguido por uma avaliação quantitativa utilizando dois corpora de teste distintos. Esta seção detalha o estudo de caso que serviu de base para a avaliação, a metodologia utilizada para medir o sucesso da transformação e, por fim, apresenta e discute os resultados obtidos.

5.1 Estudo de Caso: O Domínio da B3

O domínio do mercado de ações brasileiro foi utilizado como ambiente de validação, com base em dados públicos da B3 (B3 S.A. – Brasil, Bolsa, Balcão, 2025). Esse domínio foi selecionado devido à sua complexidade, à riqueza terminológica e à sua relevância prática. O processo de engenharia de conhecimento para configurar a ferramenta nesse domínio envolveu a criação de uma ontologia, de uma base de conhecimento RDF e de um conjunto de templates de consulta.

Primeiramente, foram formuladas 18 *Perguntas de Referência* para delimitar o escopo informacional. Com base nelas, foi desenvolvida uma ontologia de domínio em OWL para modelar os conceitos centrais do mercado de ações. A ontologia resultante é composta por 101 classes, 13 propriedades de objeto e 18 propriedades de dados, cuja estrutura permite a formulação de consultas com diferentes níveis de complexidade.

Em seguida, dados reais de negociação da B3, referentes ao período de junho de 2025, foram processados e transformados em um conjunto de 74.112 triplas RDF no formato Turtle (TTL), que constitui a nossa base de conhecimento. Para o enriquecimento semântico, foram criados um Dicionário de Entidades (com empresas, tickers e índices) e um Dicionário de Sinônimos, que, com o apoio do WordNet e ChatGPT-4 (OpenAI, 2025a), resultou em 140 sinônimos mapeados para 30 termos-chave da ontologia.

Com base nas perguntas de referência, foram projetados 13 templates de consulta, classificados em três níveis de complexidade para cobrir difer-

entes intenções de consulta. Foram criados oito templates de *Simples* para recuperação direta de atributos; três templates *Intermediários* para consultas analíticas com filtros, agregações (‘SUM‘) e rankings (‘ORDER BY‘); e dois templates *Complexos* que utilizam subconsultas para responder a perguntas que exigem a correlação de múltiplas informações, como “Qual o volume da ação com a maior alta?”. O conjunto de artefatos formado pela ontologia, pela base de dados e pelos templates constituiu o ambiente empregado na avaliação da ferramenta.

A Figura 3 ilustra a estrutura gráfica de um template de busca simples criado para responder à pergunta “Qual o preço de fechamento da Petrobras em 10/06/2025?”. Nessa figura, *P* representa um *placeholder* para uma propriedade de dado definida entre um recurso e um literal. Esse mecanismo foi introduzido para permitir que um mesmo template seja utilizado na resposta a diferentes perguntas que compartilham a mesma estrutura. A Figura 4, por sua vez, apresenta a consulta gerada e executada pela ferramenta para responder a essa pergunta.

5.2 Metodologia de Avaliação

A avaliação rigorosa da transformação de perguntas em linguagem natural em consultas SPARQL, conforme sugerido na literatura (Singh et al., 2021), envolve um processo de ponta a ponta. Esse processo inclui a verificação da correta seleção do template, refletindo a intenção do usuário, a extração completa das entidades mencionadas, a validação sintática e semântica da consulta e, por fim, a obtenção do resultado esperado após sua execução. Uma pergunta é considerada bem-sucedida apenas quando todas essas etapas são realizadas corretamente. Neste trabalho, adotamos uma abordagem de avaliação consolidada que, embora mais direta, captura a essência desse processo. Consideramos como sucesso a transformação de uma pergunta em uma consulta SPARQL semanticamente equivalente à intenção original e executável sem erros de sintaxe ou de execução. Essa métrica pragmática funciona como um indicador robusto do sucesso das etapas anteriores, pois a geração de uma consulta correta depende diretamente da identificação adequada da intenção e das entidades presentes na pergunta.

5.3 Elaboração e Coleta dos Corpus de Teste

Para validar a robustez técnica e a aplicabilidade prática, criamos dois corpora de teste distintos:

Corpus de Teste Produzido por Agentes de IA. O primeiro conjunto, com 360 perguntas, foi formulado por seis agentes de IA generativa (ChatGPT-4 (OpenAI, 2025b), Claude 3 (Anthropic, 2025), Google Gemini (Google DeepMind, 2025), Perplexity AI (Perplexity AI, 2025), Microsoft Copilot (Microsoft, 2025) e Mistral Large (Mistral AI, 2025)). O objetivo foi avaliar como a ferramenta lida com diferentes formulações, sinônimos e estruturas sintáticas produzidas por cada IA, servindo como um proxy para a diversidade da linguagem humana em um ambiente controlado. Um prompt detalhado e estratificado foi utilizado para garantir que as perguntas cobrissem as 13 categorias de templates suportadas pelo sistema.

Corpus de Teste Produzido por Participantes Humanos. O segundo conjunto, com 356 perguntas, foi formulado por 30 participantes humanos voluntários através de um formulário online. O processo de coleta foi estruturado em etapas: consentimento, coleta do perfil do usuário (classificado como pouco/nenhum, intermediário ou avançado conhecimento do domínio), contextualização do domínio e, por fim, a produção de 10 a 20 perguntas por participante, alinhadas ao escopo do estudo de caso (dados de junho de 2025 da B3).

5.4 Análise dos Resultados

Análise do Corpus Produzido via Agentes de IA. A Tabela 1 detalha a quantidade e o percentual de perguntas transformadas corretamente e executadas com sucesso para cada modelo de linguagem produzido, bem como a contagem de erros. *Natural2SPARQL* alcançou uma taxa de sucesso média de **93,3%** no corpus gerado por IA. Este resultado demonstra a alta robustez técnica da arquitetura, que se mostrou capaz de lidar com uma grande variabilidade linguística para perguntas dentro do escopo. Os erros (6,7%) concentraram-se em dois cenários: 1) falha na extração da métrica de ranqueamento em consultas complexas com subconsultas, resultando em uma consulta inválida; e 2) seleção incorreta de template por ambiguidade semântica.

Análise do Corpus Produzido por Humanos. A Tabela 2 apresenta a análise de transformação do corpus produzido por humanos. Neste corpus, a taxa de transformação correta foi de **58,1%**, com uma taxa de execução bem-sucedida de **32,3%**. A discrepância entre os dois corpora é explicada pela natureza não controlada do corpus humano.

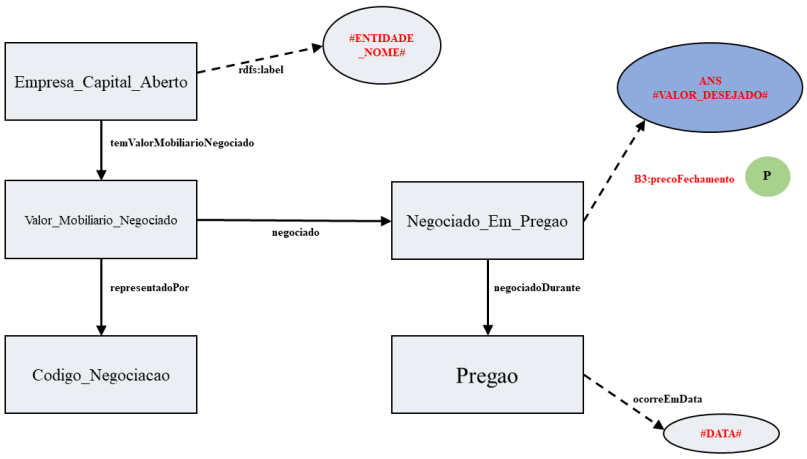

Figura 3: Estrutura gráfica de um template de busca simples. Retângulos representam recursos da ontologia, enquanto elipses representam valores de dados (literais). A elipse azul indica a variável correspondente à resposta esperada. Arestas sólidas indicam propriedades de objeto, enquanto arestas tracejadas indicam propriedades de dado.

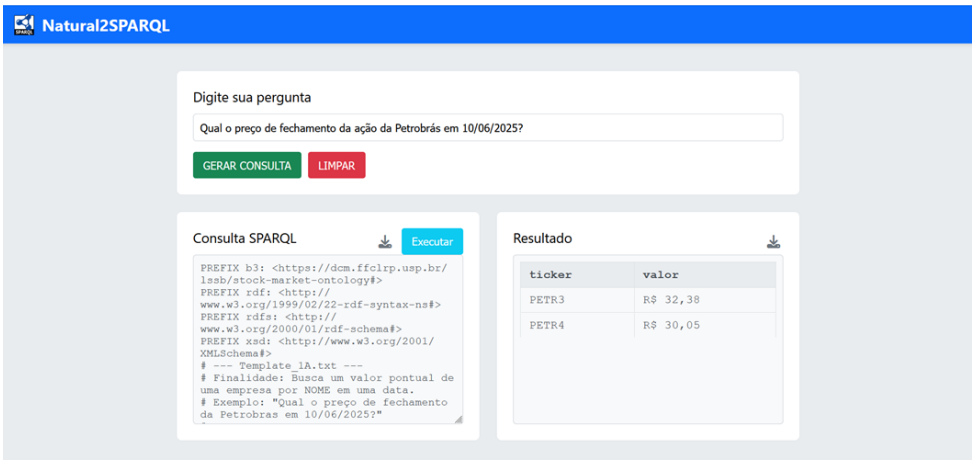

Figura 4: Consulta SPARQL gerada e executada pela ferramenta Natural2SPARQL.

Tabela 1: Transformação do Corpus Agentes de IA.

Agente IA	**Perguntas Geradas**	**Transformadas Corretamente**	**Executadas Corretamente**	**Erros de Transformação**
ChatGPT-4	60 (100%)	60 (100%)	60 (100%)	0 (0%)
Claude 3	60 (100%)	57 (95%)	57 (95%)	3 (5%)
Google Gemini	60 (100%)	54 (90%)	54 (90%)	6 (10%)
Perplexity AI	60 (100%)	57 (95%)	57 (95%)	3 (5%)
Microsoft Copilot	60 (100%)	53 (88%)	53 (88%)	7 (12%)
Mistral Large	60 (100%)	55 (92%)	55 (92%)	5 (8%)
Total	**360 (100%)**	**336 (93,3%)**	**336 (93,3%)**	**24 (6,7%)**

A análise dos 149 erros de transformação revelou que a maioria não se deveu a falhas na lógica de tradução, mas sim a perguntas que estavam fora do escopo da base de conhecimento (e.g., conceitos não modelados como "dividendos" ou datas fora de junho de 2025).

Um resultado notável e contraintuitivo foi que os participantes *com pouco ou nenhum conheci-*

Tabela 2: Transformação do Corpus Humano, estratificada por nível de conhecimento auto-declarado.

Nível de Conhecimento	Perguntas Formuladas	Transformadas Corretamente	Executadas Corretamente
Pouco/Nenhum (26,1%)	93 (26,1%)	70 (75,3%)	48 (51,6%)
Intermediário (59,3%)	211 (59,3%)	111 (52,6%)	61 (28,9%)
Avançado (14,6%)	52 (14,6%)	26 (50,0%)	6 (11,5%)
Total	**356 (100%)**	**207 (58,1%)**	**115 (32,3%)**

mento do domínio tiveram a maior taxa de sucesso (75,3% de transformações corretas), enquanto participantes com *conhecimentos avançados* tiveram a menor (50,0%). Isso sugere que a clareza das instruções favoreceu a adesão dos iniciantes, enquanto a autoconfiança levou os mais experientes a extrapolarem o escopo definido.

A avaliação com os dois corpora complementares conferiu robustez técnica e relevância prática ao estudo. Os resultados do corpus de IA validaram o núcleo técnico da solução, demonstrando a eficácia da abordagem baseada em templates para traduzir intenções de consulta dentro de um escopo definido. Por outro lado, os resultados do corpus humano, embora com uma acurácia menor, reforçam a adequação da abordagem para o público-alvo (usuários leigos) e destacam que os principais gargalos para o desempenho de ponta a ponta são a cobertura ontológica e temporal da base de dados, e não falhas na lógica de tradução.

6 Discussão

Este trabalho apresentou o *Natural2SPARQL*, uma solução de engenharia de conhecimento para traduzir perguntas em português para consultas SPARQL. A avaliação demonstrou a viabilidade técnica da abordagem, alcançando uma acurácia de 93,3% ao lidar com a variabilidade linguística de múltiplos agentes de IA em um ambiente controlado. Mais importante, a análise do corpus humano revelou a adequação da ferramenta para o público-alvo, visto que usuários iniciantes obtiveram a maior taxa de sucesso, validando o objetivo de democratizar o acesso a dados semânticos. A principal contribuição reside em uma metodologia adaptável e de baixo custo, que desacopla a lógica de tradução do conhecimento de domínio. Ao contrário de abordagens que exigem o custoso retreinamento de modelos de ML para novos domínios, nossa solução se concentra na definição de artefatos reutilizáveis — templates e dicionários —, tornando-a um framework pragmático e acessível. A ferramenta, disponibilizada como open-source, destaca-se ainda pela sua flexibilidade para gerar consultas analíticas complexas, incluindo agregações, rankings e subconsultas, uma capacidade que vai além da simples recuperação de fatos. Apesar dos resultados promissores, o trabalho possui limitações inerentes à sua abordagem. A principal delas é a dependência de templates pré-definidos, o que restringe a capacidade do sistema a estruturas de perguntas previamente mapeadas e exige manutenção contínua para expandir sua cobertura. Adicionalmente, o mecanismo de seleção de templates, baseado puramente em similaridade de texto, mostrou-se insuficiente para resolver a ambiguidade semântica em perguntas complexas onde múltiplas intenções analíticas (e.g., ranqueamento vs. agregação) coexistem. Por fim, a validação, embora rigorosa, foi conduzida em um único domínio, o que aponta para a necessidade de futuros testes para aferir a generalização da abordagem em contextos com maior variabilidade semântica.

7 Conclusão

Este trabalho apresentou o *Natural2SPARQL*, demonstrando a viabilidade de uma abordagem baseada em engenharia de conhecimento para traduzir perguntas em português em consultas SPARQL. A avaliação indicou robustez técnica (93,3% de acurácia com agentes de IA) e adequação ao público-alvo, com melhor desempenho entre usuários iniciantes.

A principal contribuição deste trabalho é uma metodologia adaptável e de baixo custo para a geração de consultas SPARQL a partir de perguntas em linguagem natural, baseada em recursos reutilizáveis de engenharia de conhecimento. Implementada como software de código aberto, a ferramenta possibilita a geração flexível de consultas analíticas complexas, como agregações e rankings, ampliando o acesso a dados semânticos. Detalhes

adicionais sobre a abordagem de transformação, a ferramenta *Natural2SPARQL* e a avaliação realizada podem ser encontrados em (Castro, 2025).

Apesar dos resultados promissores, a abordagem apresenta limitações inerentes. A dependência de templates pré-definidos restringe o sistema a estruturas previamente mapeadas, exigindo manutenção para ampliar sua cobertura. Além disso, o mecanismo de seleção baseado em similaridade textual mostrou-se, em alguns casos, insuficiente para resolver ambiguidades semânticas quando coexistem múltiplas intenções analíticas (e.g., ranqueamento e agregação). Por fim, a validação em um único domínio reforça a necessidade de avaliações futuras para investigar a generalização em contextos mais diversos.

As limitações identificadas apontam oportunidades de aprofundamento desta pesquisa. Trabalhos futuros incluem o aprimoramento da interpretação das consultas por meio de mecanismos de desambiguação baseados em diálogo e de uma seleção de templates em dois estágios que integre análise semântica e sintática. Avalia-se também o uso de modelos de linguagem de grande porte para apoiar a geração de recursos de domínio, reduzindo o custo de engenharia. Adicionalmente, a extensão da arquitetura para suportar a composição dinâmica de consultas poderá permitir a combinação de múltiplos templates na resolução de perguntas que demandem múltiplos passos de raciocínio, ampliando a expressividade do sistema.

Referências

Mohamed Aly. 2005. Survey on multiclass classification methods. *Neural Networks*, 19(1):1–9.

Anthropic. 2025. Claude 3 model family. https://docs.anthropic.com/claude/docs/models-overview. Acesso em: 13 mar. 2026.

B3 S.A. – Brasil, Bolsa, Balcão. 2025. Portal da B3 - informações de mercado. https://www.b3.com.br/. Acesso em: 13 mar. 2026.

Tim Berners-Lee, James Hendler, e Ora Lassila. 2001. The semantic web. *Scientific American*, 284:34–43.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Heber Gustavo Xavier de Castro. 2025. Geração de consultas SPARQL a partir de linguagem natural. Tese de Mestrado, Universidade de São Paulo, Ribeirão Preto, SP.

Yi-Hui Chen, Eric Jui-Lin Lu, e Ting-An Ou. 2021. Intelligent sparql query generation for natural language processing systems. *IEEE Access*, 9:158638–158650.

Douglass Cutting, Julian Kupiec, Jan Pedersen, e Penelope Sibun. 1992. A practical part-of-speech tagger. Em *Proceedings of the Third Conference on Applied Natural Language Processing*, páginas 133–140.

Ionuţ Cristian Dorabăţ e Vlad Posea. 2020. onIQ: An ontology-independent natural language interface for building SPARQL queries. Em *Proceedings of the 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*, páginas 139–144. IEEE.

Google DeepMind. 2025. Gemini models. https://ai.google.dev/models/gemini. Acesso em: 13 mar. 2026.

Thierry Hamon, Natalia Grabar, e Fleur Mougin. 2017. Querying biomedical linked data with natural language questions. *Semantic Web*, 8(4):581–599.

Steve Harris e Andy Seaborne. 2013. SPARQL 1.1 query language. Recommendation, W3C. Disponível em: http://www.w3.org/TR/sparql11-query/.

Andrew McCallum, Dayne Freitag, e Fernando C. N. Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. Em *Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00)*, páginas 591–598.

Microsoft. 2025. Microsoft Copilot. https://learn.microsoft.com/copilot/. Acesso em: 13 mar. 2026.

Tomas Mikolov, Kai Chen, Greg Corrado, e Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Elbe Miranda, Aline Paes, e Daniel de Oliveira. 2024. SPARQL can also talk in Portuguese: answering natural language questions with knowledge graphs. Em *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*, páginas 56–66. Association for Computational Linguistics.

Mistral AI. 2025. Mistral Large. https://docs.mistral.ai/models/. Acesso em: 13 mar. 2026.

Behrang Mohit. 2014. Named entity recognition. Em *Natural language processing of semitic languages*, páginas 221–245. Springer.

OpenAI. 2025a. ChatGPT (versão baseada no modelo GPT-4). https://openai.com/chatgpt. Acesso em: 13 mar. 2026.

OpenAI. 2025b. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mario Andrés Paredes-Valverde, Rafael Valencia-García, Miguel Ángel Rodríguez-García, Ricardo Colomo-Palacios, e Giner Alor-Hernández. 2016. A semantic-based approach for querying linked data using natural language. *Journal of Information Science*, 42(6):851–862.

Perplexity AI. 2025. Perplexity AI. `https://www.perplexity.ai/about`. Acesso em: 13 mar. 2026.

Kuldeep Singh, Ankit Bhardwaj, Asif Ekbal, Pushpak Bhattacharyya, e Jens Lehmann. 2021. MTL-based question answering over knowledge bases. Em *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, páginas 2119–2123. ACM.

Panayiotis Smeros, Vincent Emonet, Ruijie Wang, Ana-Claudia Sima, e Tarcisio Mendes de Farias. 2025. SPARQL-LLM: Real-time SPARQL query generation from natural language questions. *arXiv preprint arXiv:2512.14277*.

Shengli Song, Wen Huang, e Yulong Sun. 2019. Semantic query graph based sparql generation from natural language questions. *Cluster Computing*, 22(1):847–858.

Kai Sheng Tai, Richard Socher, e Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. Em *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, páginas 1556–1566. Association for Computational Linguistics.

Davide Varagnolo, Dora Melo, e Irene Pimenta Rodrigues. 2023. An ontology-based question-answering, from natural language to SPARQL query. Em *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Ontology Development (IC3K 2023)*, páginas 174–181. SCITEPRESS.

Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, Maya Ramanath, Volker Tresp, e Gerhard Weikum. 2012. Natural language questions for the web of data. Em *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, páginas 379–390.