UNIVERSIDADE D
COIMBRA

Knowledge and Language 2025/2026 - Project Statement

Masters in Informatics Engineering
Natural Language Processing and Knowledge Representation

Membros:
Miguel Castela 2022212972
Miguel Martins 2022213951

October, 2025

1 Problem Statement and Project Goal

General LLMs can answer many questions but often lack domain or culturally specific Knowledge. The challenge we aim to address is building a system that can reliably and accurately answer complex, domain-specific questions about Portuguese cuisine and other recipes using structured data from the food.com dataset. The goal of the project is to develop a ChatBot that combines a structured Knowledge Graph (KG) of recipes with Natural Language Processing (NLP) techniques and Retrieval-Augmented Generation (RAG). The ChatBot will be able to interpret user queries in natural language, retrieve the most relevant recipes or recipe attributes, and provide accurate, context-aware responses.

2 Data, Tools and Proposed Approach

We will pre-process the dataset to include a bigger percentage of Portuguese recipes and keep the distribution of the others. The structured KG is built using the RDFS (Resource Description Framework Schema), an extension of RDF that provides a schema vocabulary for describing the relationships between resources. Each RDF triplet has the form:

- **subject**: a specific recipe or ingredient (e.g., "Bacalhau à Brás"),
- **predicate**: an attribute such as `hasIngredient` or `hasPrepTime`,
- **object**: the value of the attribute (e.g., "tomato", "15 minutes").

We will not use OWL (Web Ontology Language) as the dataset is already structured and does not require cardinality restrictions or complex class expressions. The system first uses a fine-tuned BERT model to identify the user's intent, such as find_recipe and retrieve_ingredients. Next, spaCy extracts relevant entities from the query, including ingredient, recipe_name and cooking_time, which are then linked to corresponding properties in the KG. The intent and linked entities are combined to generate a SPARQL query for precise retrieval of structured information from the graph. Each recipe has more textual attributes taken from the dataset, such as description, desciptive cooking steps and other tags (with different information like diet, food restrictions, etc.) added with a RAG step to give Ollama more context. This, as mentioned in the reviewed literature is achieved through an embeddings-based retrieval that identifies the top-k most relevant recipes with a similarity metric given the query and NLP retrieved context. Inspired by RecipeRAG, which employs only knowledge graph embeddings for retrieval, this work integrates a simpler, more queryable knowledge graph with a RAG component to retrieve richer external information, resulting in a more concise and flexible architecture. Finally, a React-based front-end presents the answers to the user in a chatbot interface.

3 checkpoint dates

Checkpoint 1 →18 November 2025, Checkpoint 2 →2 December 2025