netidee

STIPENDIEN

Elwin Huaman

Zwischenbericht | Call 19 | stip7335

Contents

1. Introduction

This report outlines the progress made of my research on building Knowledge Graphs (KGs) for Under-Resourced Languages (URLs). The aim of this work is to review existing literature on KGs, design and localize data collection instruments and analyze current practices on knowledge collection. This work sits at the intersection of natural language processing, semantic web technologies, and AI. In the following sections, we will discuss the three main writing-focused milestones: the introduction, fundamentals, and related work chapters.

2. Status

2.1 Milestone 1 - Introduction Chapter

The activities in this milestone were focused on drafting the thesis introduction chapter. This involved defining the scope of the research, refining the research questions, and providing a high level overview of the proposed thesis structure.

- The process of writing the introduction forced an early refinement of the project's scope. Initially, "under-resourced languages" was a broad term; it was necessary to define specific domains where under-resourced languages are applicable, e.g., from a linguistic, technological, social, perspectives to geopolitical aspects.
- A completed draft of the introduction chapter has been produced. It successfully established the research context, clearly states the main research question; *How can knowledge graphs be built to support lexicographical data of under-resourced languages?*.
- A minor problem was the initial over-scoping of languages to be studied, which was corrected through literature review and advisor feedback. I am focusing on Puno Quechua language as a use case.
- There were no major deviations from the initial plan.

2.2 Milestone 2 - Fundamentals Chapter

This milestone involved researching and writing the fundamentals chapter, which serves as the theoretical foundation for the thesis. The activities included a deep dive into three core areas: Knowledge Graphs's architecture, linguistic and technical challenges defining under-resourced languages, and the Responsible AI framework.

- A comprehensive draft of the fundamentals chapter is completed. The sections on knowledge graphs and under-resourced languages are finished. The responsible AI section is well-structured but requires further feedback and development with specific case studies relevant to the target language (i.e, Puno Quechua). The responsible AI section will be refined after the deployment of the Puno Quechua Knowledge Graph.

- A challenge has been the scarcity of literature that explicitly connects Responsible AI principles to knowledge graphs construction for under-resource language scenarios, requiring to produce a specific case study that can be discussed.
- There is a minor change in the timeline with respect to the responsible AI section, but it does not affect subsequents milestones, nor requires a change in the initial plan.

2.3 Milestone 3 - Related Work Chapter

This milestone involved a systematic literature review of three main areas: existing linguistic knowledge graphs (e.g., Wikidata, LLOD, WordNet), previous research efforts focused on under-resourced languages in NLP, and the use of knowledge graphs in digital lexicography.

- The literature review confirms a significant gap on knowledge graphs for under-resourced languages. While large multilingual knowledge graphs exist, their coverage of under-resourced languages is often shallow, incomplete, and English-centric, which introduces errors and cultural biases. This gap strongly validates the research problem I am tackling.
- A fundamental draft of the literature review has been accomplished. The literature review is an iterative process, so more literature might be included in the future as NLP technology, knowledge graph approaches, and AI applications advance.
- No major issues have been encountered at this stage, this work is proceeding as planned.

2.4 Milestone 4 - Knowledge Graph I: Creation and Hosting Chapter

This milestone is the core practical component of my research. This includes data acquisition, schema alignment, and the population and hosting of a knowledge graph.

- The schema alignment and knowledge graph creation has been started, and will continue. It is an iterative process and more knowledge can be ingested as they come. The advantage of Knowledge graphs is that even if you do not have a Graph database, it can still be maintained in structured format and files, e.g., JSON-LD format.
- Experiments are running into which Graph Database suits better to our purpose to Open Access and Data Sovereignty. A minor issue that do not affect the research objectives is the sustainability of those Graph Databases, they are expensive.
- This work is proceeding as planned.

3. Plan Update Summary

The initial plan was primarily composed of writing-oriented milestones. Given the solid theoretical foundation being established, I will continue with practical experimentation, knowledge graph deployment, and evaluation. This will involve setting up a development environment and running small-scale baseline experiments.