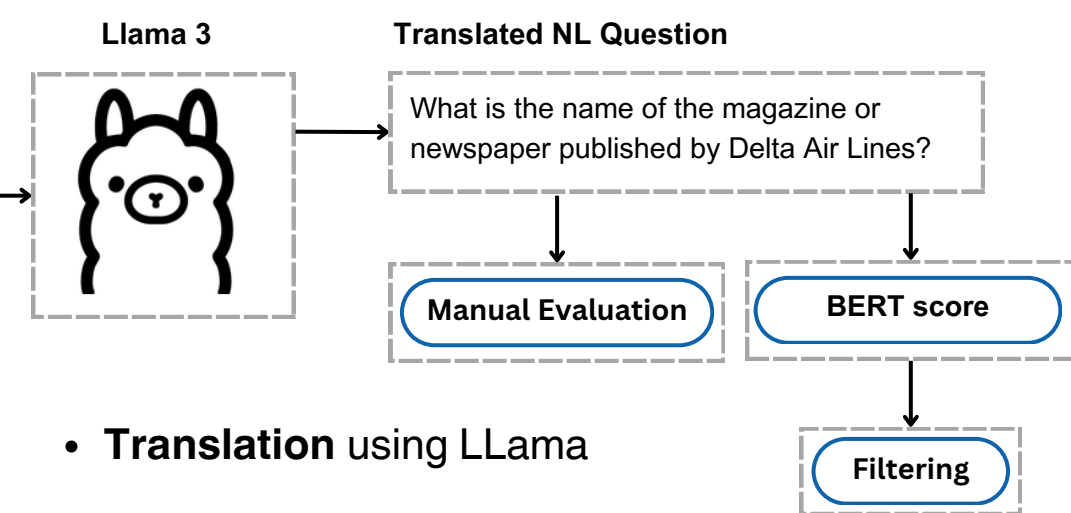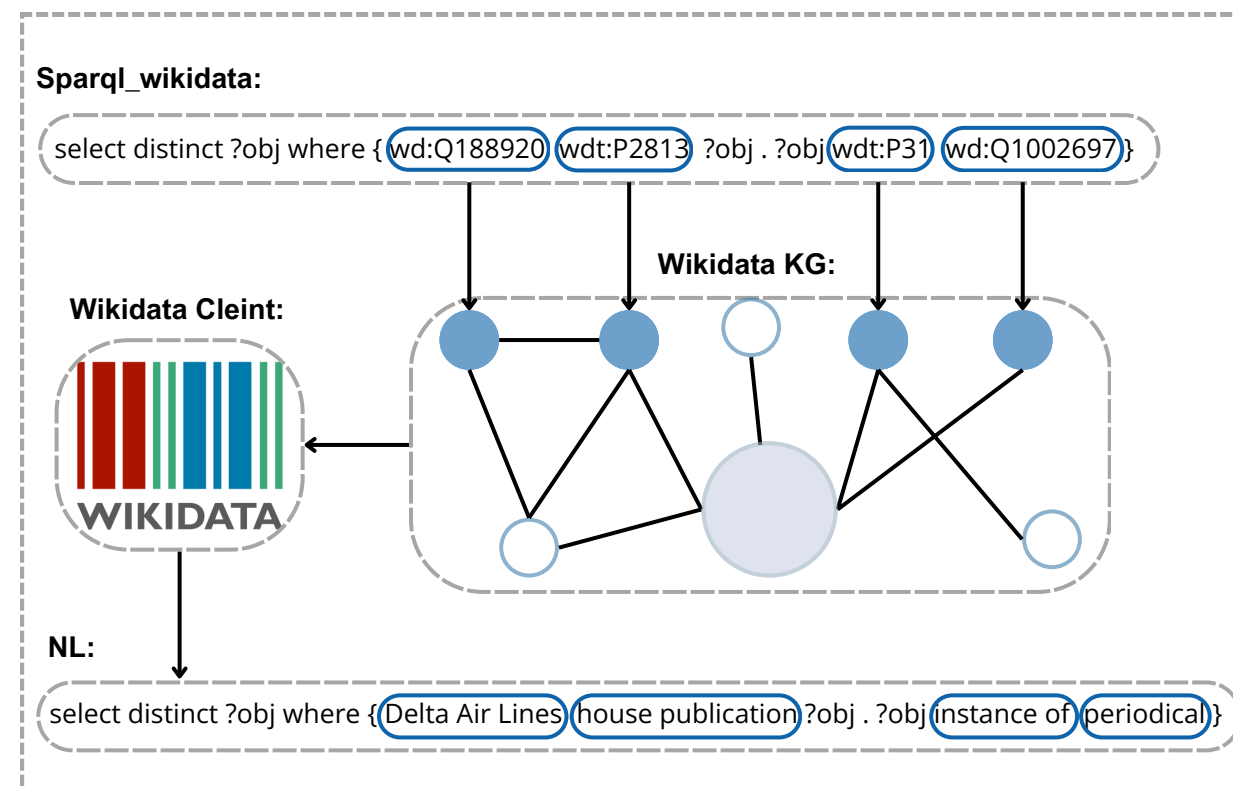# Structured Query Verbalization with Large Language Models

This project aims to provide a way of generating data for training LLMs for the specific use case of creating SPARQL queries from natural language.

Authors: Patrik Valach, Louisa Siebel

Advisors: M.Sc Tim Schwabe

SCAN ME

Approach

- Generating a **natural language question** from a **SPARQL query**

- Using the lc_Quad NLP-Query Database

- Map **ambiguous entities and properties** to their **natural language** form though the Wikidata client

Sparql_wikidata:

select distinct ?obj where { wd:Q188920 wdt:P2813 ?obj . ?obj wdt:P31 wd:Q1002697}

Wikidata Cleint:

WIKIDATA

Wikidata KG:

NL:

select distinct ?obj where { Delta Air Lines house publication ?obj . ?obj instance of periodical}

Llama 3

Translated NL Question:
What is the name of the magazine or newspaper published by Delta Air Lines?

Manual Evaluation   BERT score

Filtering

- **Translation** using LLama

- **Evaluation** of translation performance :
  - Manual
  - BERT score

- **Filtering** for **True Positives**

Results: BERT

- **BERT** has proven to be the most **reliable metric** in determining the **semantic equivalance**

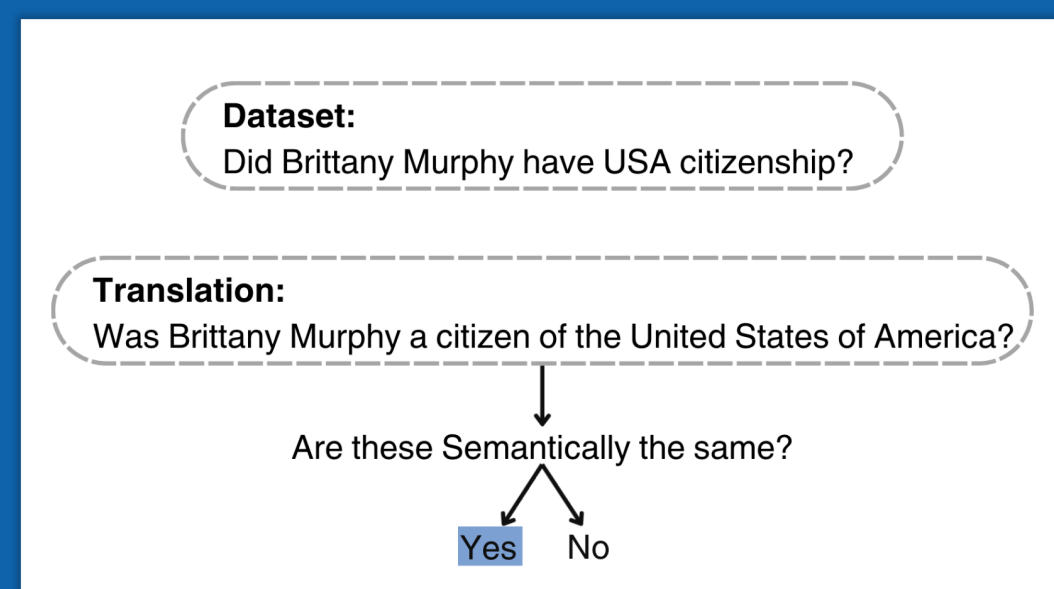Distribution of BERT Values
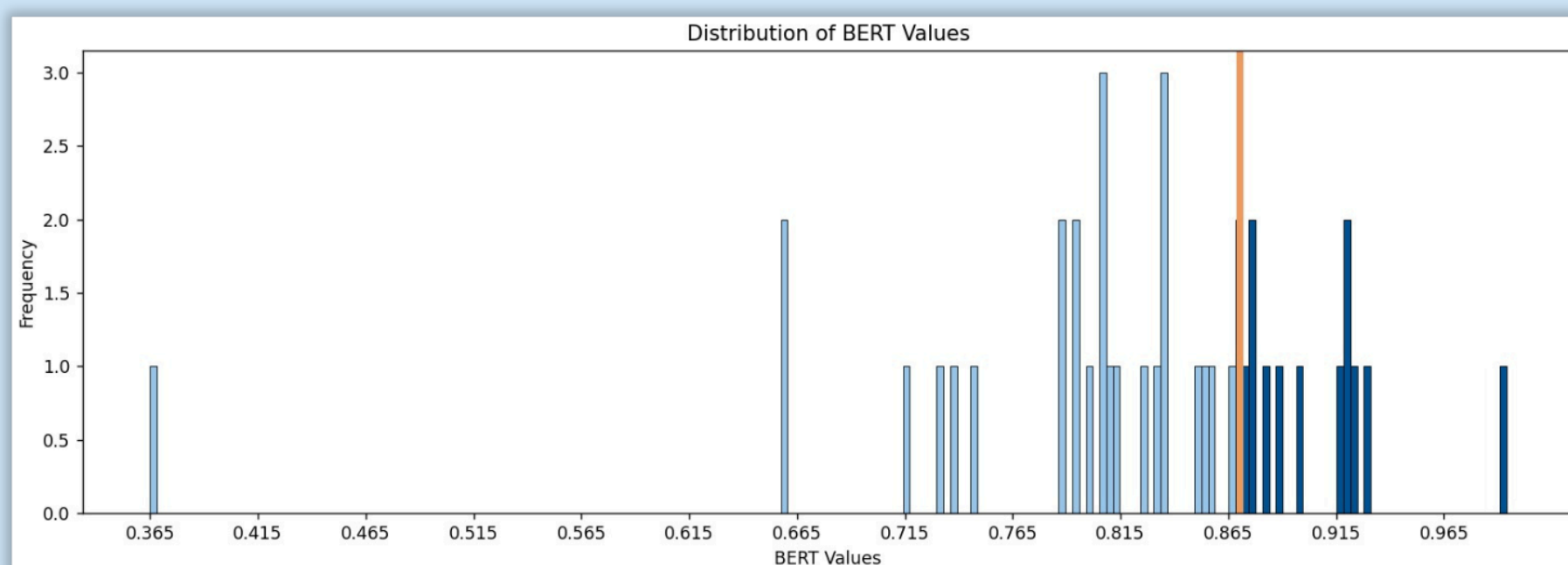
- **Minimum** value where we get only **True Positives** is **0,87**
 - When **filtering** with a **BERT** score **> 0.87** we obtain **only correct** translations

Pie chart: True Positives 30.0%, True Negatives 24.0%, False Negatives 46.0%
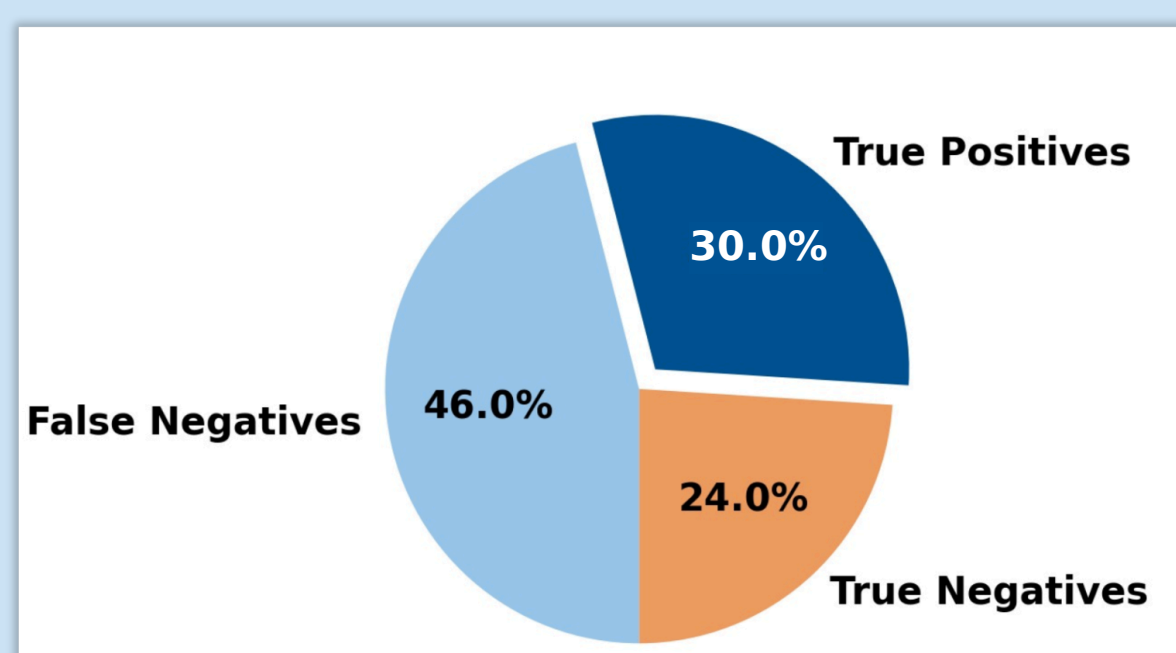
- Filter creates a set of **only correct** translations

- **Filtering result:** ≈1/3 of the original dataset size.

- Some **correct translations** are **sacrificed** for a truly **perfect result**

Results: Manual Inspection

Dataset:
Did Brittany Murphy have USA citizenship?

Translation:
Was Brittany Murphy a citizen of the United States of America?

Are these Semantically the same?

Yes   No

- **Result** from manual inspection is either **0 or 1**

- Helped **establish** a **metric** for finding the right **filter**

Pie chart: Different Semantic Meaning 24.0%, Same Semantic Meaning 76.0%
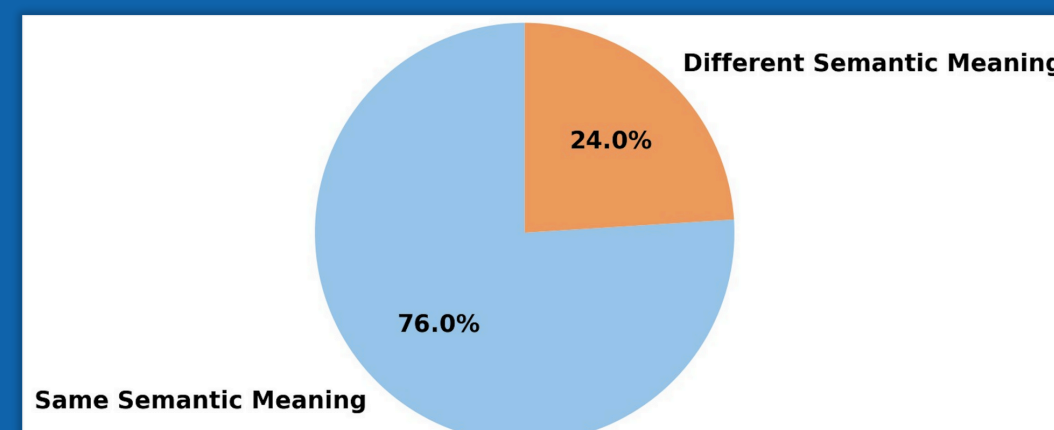
- Used to **evaluate** the models' **performance**
  - **76%** of translations are **correct**

- Used to **set** the **threshold** for BERT

Conclusion

- Translation task is very successful: **76%**,
- We have established **metric** by which we can filter out only **true positives** if we have the ground truth
- Altogether, the project successfully **fulfills** the **goal** and can be used to **create new dataset**

Future Work

- Develop a method to **automatically** specify if a translation is **correct** without relying on the **ground truth**
  - E.g. in a **production environment** where translations need to be **generated at scale**