A Benchmark for the Detection of Metalinguistic Disagreements between LLMs and Knowledge Graphs: Problem Pitch

Bradley Allen and Paul Groth

ISWC Special Session on LLMs

2024-11-13

Problem statement

- Knowledge engineering is a process of formalization, i.e., translating knowledge expressed in natural language into a formal language
  - Formalization using a knowledge graph trades linguistic expressiveness for sound and tractable inference
  - Formalization using an LLM trades factuality for human-level linguistic competence

- Most work in GenAI/SemWeb harmonization focuses on how knowledge graphs can mitigate problems arising from lack of factuality in LLMs (q.v. Slot 2)

- There is little work studying how LLMs can mitigate problems arising from lack of expressiveness in knowledge graphs

Current focus: metalinguistic disagreement

Your task is to review a statement concerning the relation between two entities, the definition of the relation, and text that provides background information from which the statement was derived. You will determine if you disagree with the statement and, if so, whether your disagreement is factual or metalinguistic in nature.
...
Provide your answer as one of "AGREEMENT", "FACTUAL_DISAGREEMENT", or "METALINGUISTIC_DISAGREEMENT".

statement: "European Union / shares border with / Andorra"
definition: "shares border with": countries or administrative subdivisions, of equal level, that this item borders, either by land or water. A single common point is enough."
text: "Andorra ..., officially the Principality of Andorra (Catalan: Principat d'Andorra), also called the Principality of the Valleys of Andorra (Catalan: Principat de les Valls d'Andorra), is a sovereign landlocked microstate in Southwestern Europe, located in the eastern Pyrenees mountains and bordered by Spain and France. ... It is not a member of the European Union, but the euro is the de facto currency."

Zero-shot CoT prompt with example knowledge alignment from T-REx (Elsahar et al., 2018)
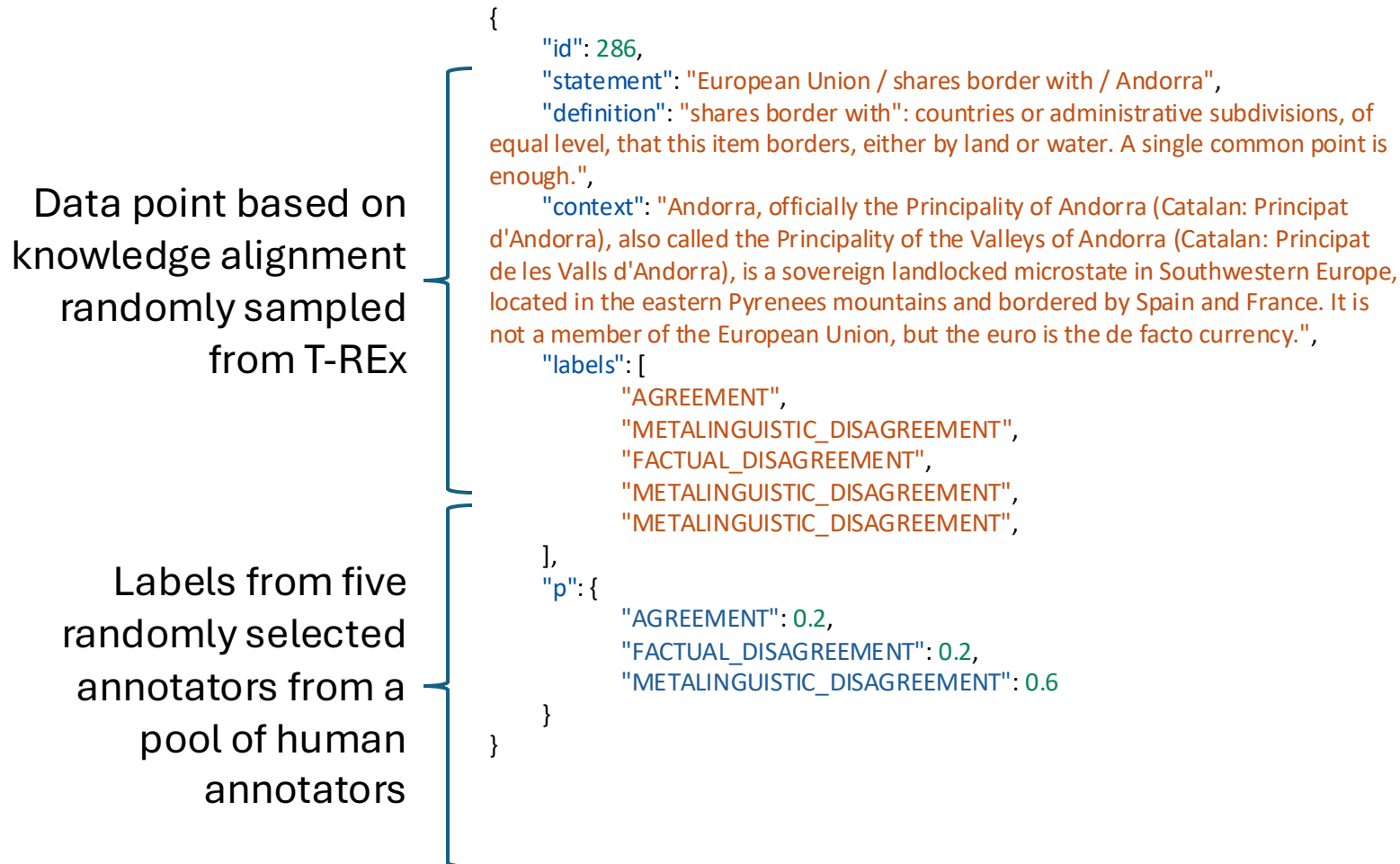
LLM
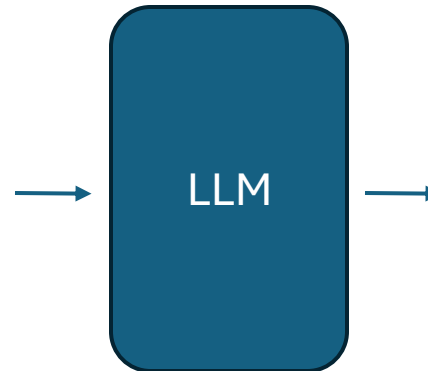
Evaluation:
- While Andorra does border EU member states (Spain and France)
- The EU itself is not a country or administrative subdivision of equal level to Andorra
- The EU is a supranational organization, while Andorra is a sovereign state
- These are not entities "of equal level" as required by the definition

Decision: This is a **METALINGUISTIC_DISAGREEMENT** because:
- The disagreement centers on whether the EU and Andorra qualify as entities "of equal level" as specified in the definition
- It's not about the factual reality of physical borders, but about whether these specific entities fit the defined relationship type
- The issue is with the proper application of the term "shares border with" given its specific definition, not with the geographical facts

LLM response (Claude 3.5 Sonnet, accessed 2024-11-03)

Proposed benchmarking approach

Data point based on knowledge alignment randomly sampled from T-REx

```
{
    "id": 286,
    "statement": "European Union / shares border with / Andorra",
    "definition": "shares border with": countries or administrative subdivisions, of
equal level, that this item borders, either by land or water. A single common point is
enough.",
    "context": "Andorra, officially the Principality of Andorra (Catalan: Principat
d'Andorra), also called the Principality of the Valleys of Andorra (Catalan: Principat
de les Valls d'Andorra), is a sovereign landlocked microstate in Southwestern Europe,
located in the eastern Pyrenees mountains and bordered by Spain and France. It is
not a member of the European Union, but the euro is the de facto currency.",
    "labels": [
        "AGREEMENT",
        "METALINGUISTIC_DISAGREEMENT",
        "FACTUAL_DISAGREEMENT",
        "METALINGUISTIC_DISAGREEMENT",
        "METALINGUISTIC_DISAGREEMENT",
    ],
    "p": {

        "AGREEMENT": 0.2,
        "FACTUAL_DISAGREEMENT": 0.2,
        "METALINGUISTIC_DISAGREEMENT": 0.6
    }
}
```

Labels from five randomly selected annotators from a pool of human annotators

- Modeled after the SemEval-2024 Task 6 hallucination detection benchmark SHROOM (Mickus et al., 2024)

- Template-based verbalizations of Wikidata statement and Wikidata property definition

- Verbatim aligned Wikipedia content as context

- Human annotator label distribution as soft gold label

- Cohen's κ, Fliess' κ as metrics for LLM annotation evaluation

Research questions

- What problems arise from the lack of expressiveness in KGs?

- To what extent do these problems affect the trustworthiness of KGs as gold standards used in the evaluation of LLMs?

- Can we design benchmarks that measure LLMs' ability to detect such problems?