PROPOSAL FOR PORQUE: A POLYLINGUAL HYBRID QUESTION ANSWERING SYSTEM

Victor Mireles[1], Artem Revenko[1], Nikit Srivastava[2], Daniel Vollmers[2], Anna Breit[1], Diego Moussallem[2]

1) SEMANTIC WEB COMPANY

2) UNIVERSITÄT PADERBORN

PORQUE

https://porque-project.eu/

Funded the Eureka Eurostars programme, Grant Number E114154

Motivational example: What types of human activities endanger birds?

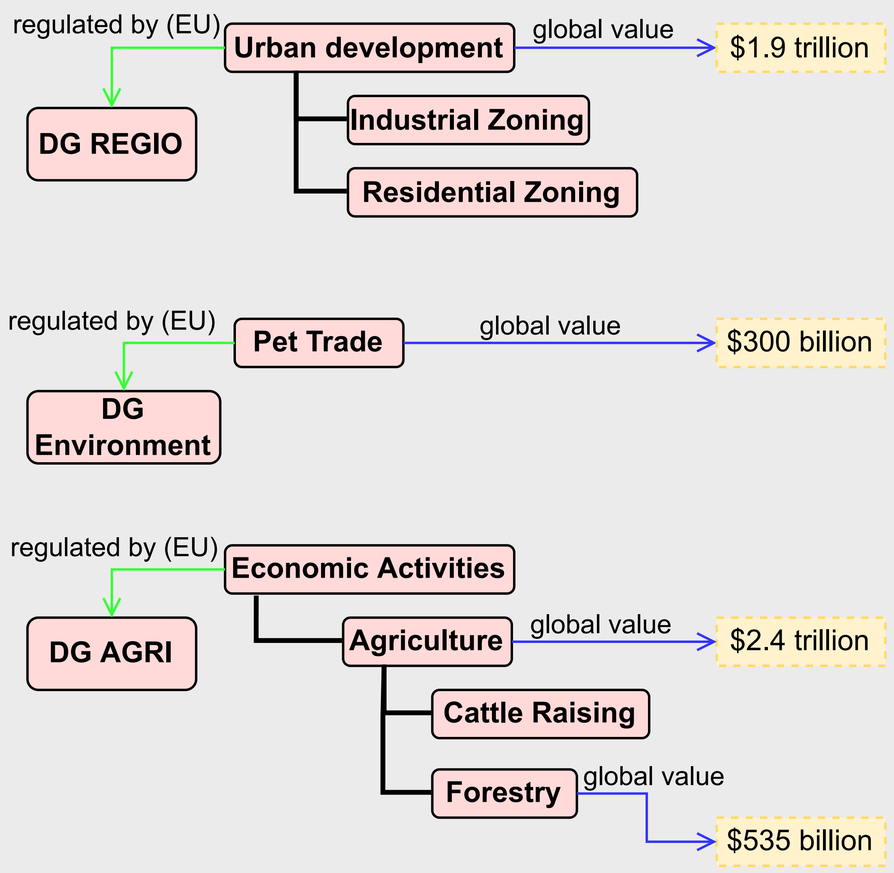
Corpus:

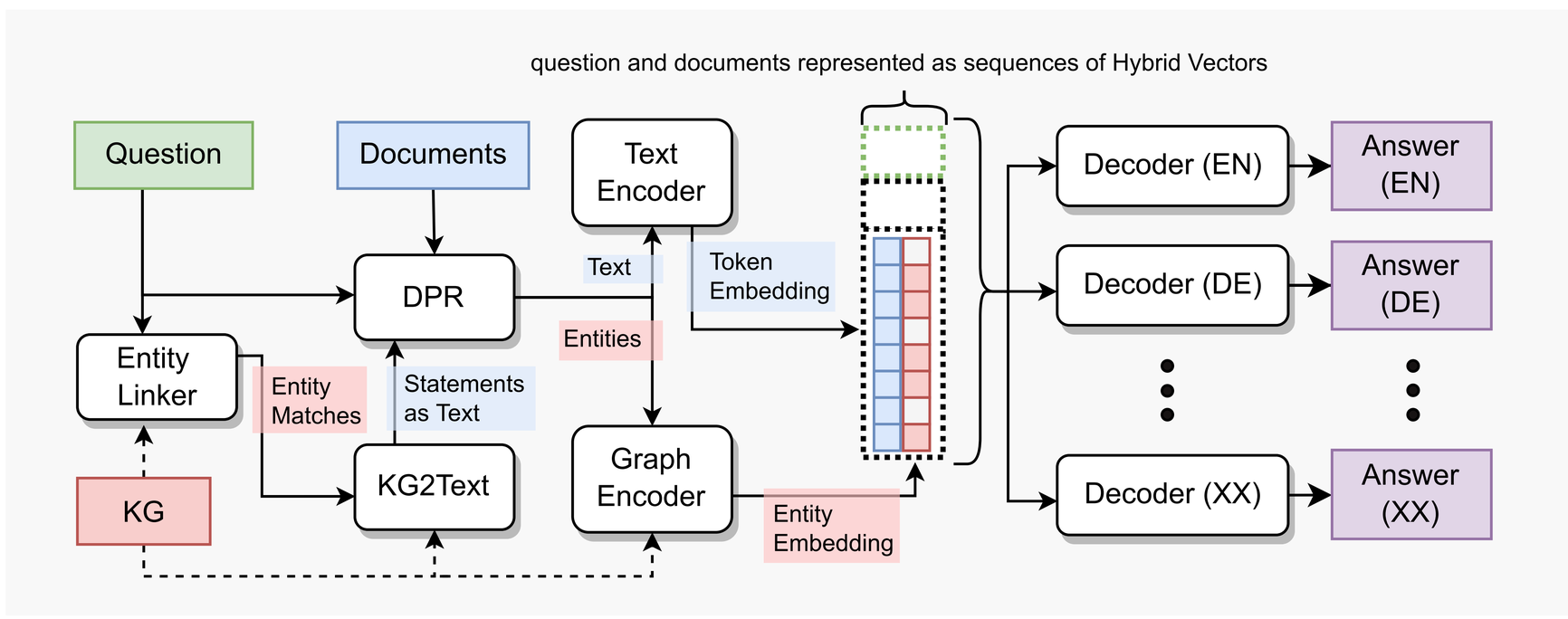
The Hispaniolan (Amazona ventralis), colloquially known as cuca, is a species of Amazon parrot in the family Psittacidae. **It is threatened in its home range by habitat loss and the capture of individuals for the pet trade.**

El hormiguerito del Paraná (Formicivora acutirostris) es una especie de ave paseriforme de la familia Thamnophilidae. **Mucho de su hábitat se encuentra en estado de degradación, y es amenzado por el uso del suelo para agricultura, industria y habitación**

The madanga (Anthus ruficollis) is a species of bird endemic to the Indonesian island Buru. Because the species are restricted to a single island and **its habitat is threatened by logging and land conversion for cattle farming**, it is listed as endangered by the IUCN.

Knowledge Graph:

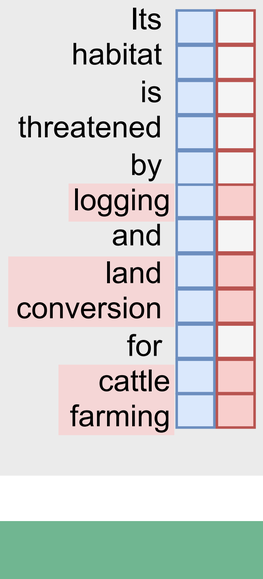




The Hybrid Vectors

In PORQUE, answers are generated based on the *hybrid vectors* that represent information present both in the documents and in the graph.
Each of these vectors corresponds to a token in a paragraph, and results from the combination (e.g., concatenation) of two components:

1. The contextualized, multilingual embedding of the token by the Text Encoder.
2. Either
   - the all-zeros vector in case the token is not part of any linked entity
   - the graph-embedding as provided by Graph Encoder in case it is the start of an entity mention



Other Approaches

- Late Fusion: Data-source specific ML systems are used to produce answers, which are then combined.
- Early fusion: Data sources is are integrated before the inference step.
  - **Text2KG**: Take all documents and convert it into triples. Then use methods for Question Answering over Linked Data to generate queries on the KG.
  - **KG2Text**: Generate natural language documents from triples. Then use methods for Question Answering over Documents to find spans containing the answer.

Our Proposal

- **Hybrid**: Information from both documents and knowledge graph is combined to generate the answer.
- **Polylingual**: The sources of information, the question, and the answer can all be in different languages.
- **Generative**: The answer is not explicitly contained in any of the sources of information, instead it is generated using a Language Mode.
- **Robust**: The approach also works with a single source of information.