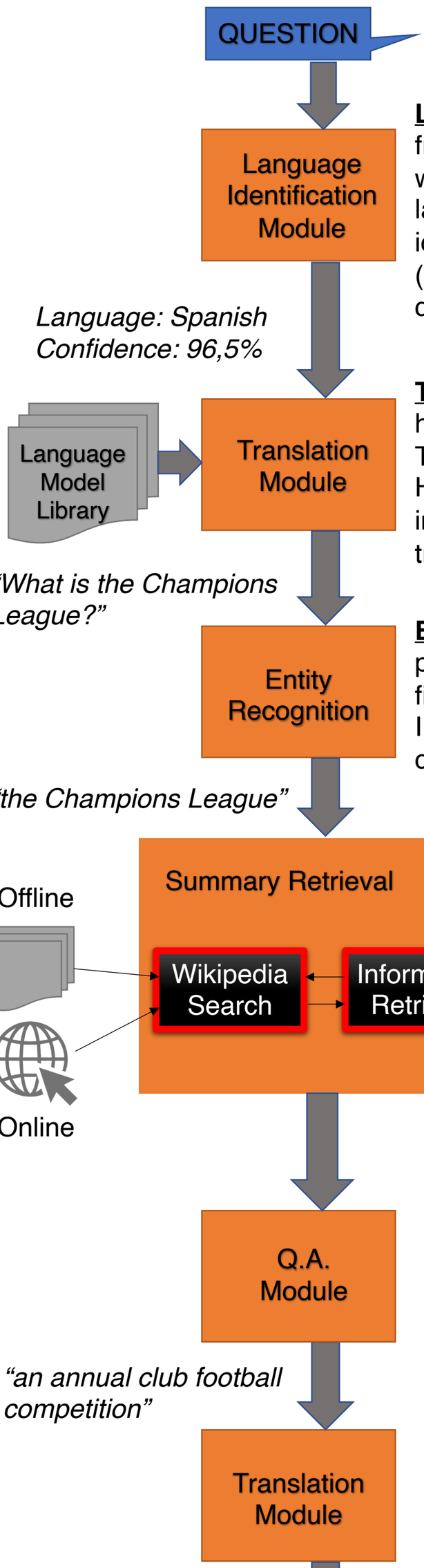
V2: Multilingual Open Domain Question-Answering Pipeline based on Wikipedia API

Nicolas Münster and Jan Jung
University of the Basque Country
nic.muenster@fau.de jan.jung@uni-freiburg.de

We create a pipeline including language recognition and a multilingual machine translation module in both directions to create a multilingual open domain Q.A. module, based on Wikipedia API.

Pipeline



Language Identification Module: We used FastText from Meta to recognize the language. The module works with 157 languages and returns the recognized language with the confidence level. When the identification score is below a defined threshold (e.g.70%), the user is prompted to rephrase the question.

Translation Module: Transformer model from the huggingface library, pretrained by the Language Technology Research Group at the University of Helsinki. Recognized language is taken from the included library and results in a bidirectional translator *src-eng* & *eng-src.*

Entity Recognition: Scan the question for possible entities. If one or more are found, the first one is used for the candidate generation and IR instead of the original question. Allowed candidates are either {PERSON, ORG, GPE} [1]

Summary Retrieval: receives summary from Wikipedia search and creates based in entity recognition the best article

Wikipedia Search: Returns a set of article titles (candidates) best fitting to input (recognized entity or question)

Information Retrieval: Find the best possible match between candidates and input. [2]

Question Answering: Extracts from the selected article the best start and end point of the answer using a BERT-model finetuned for Question-Answering. [3]

Translation Module: Module corresponding to the previously used Translation module, translation the answer back from English into the source language.

Wikipedia Search

Summary Retrieval:

Wikipedia API creates based on input a set of best summaries. The input is only the entity recognition and not the whole question, as filling words reduce quality of search results.

Online vs. Offline search:

+ Reduced memory consumption
+ Significantly faster
+ Better search results

− Needs internet connection

Remaining Problems

Small word count errors:

- Due to small word count, Language Identification module cannot recognize language with high confidence.
- Also IR module has problems:
  E.g. *Where is Spain -> Where is the love?*

→ *Partially mitigated through ER module*

- Question-Answering mismatch due to retrieval of non-related article (e.g. song title closely resembling the posed question)

Examples

Spanish (99,7%)

Q:	*¿Cuántas personas viven en Londres?*
A:	*9,002,488*

Basque (97,9%)

Q:	*Zein hizkuntza hitz egiten da Nigerian?*
A:	*ingelesa*

Spanish (99,9%)

Q:	*¿Cual es el sentido de la vida?*
A:	*Canciones de amor que parecen estar dirigidas tanto a una mujer como a una deidad*

English (99,7%)

Q:	*When was Germany reunited?*
A:	*3 October 1990*

1) We used the Entity Recognition from Spacy
2) We used the Whoosh library with the Okapi BM25 algorithm (BM-> Best Matching)
3) We used the "bert-large-uncased-whole-word-masking-finetuned-squad" model from huggingface (Reference to paper in GitHub-link)

Code to reproduce results and links to papers and data:
https://github.com/JanTJung/nlp2_project