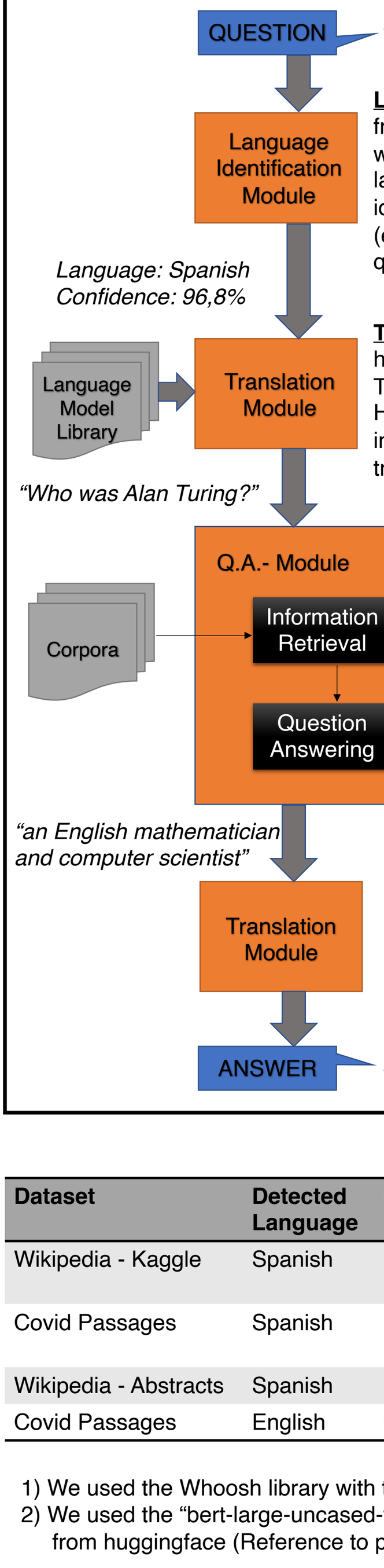
Multilingual Open Domain Question-Answering Pipeline based on different Corpora

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 or Covid corpora.

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.*

Q.A. Module: Q.A. model returns based on corpora and the regarding Information Retrieval an answer in English.

Information Retrieval: Extracts from the dataset the passage with the best fitting information overlap to the question. [1]
Question Answering: Extracts from the selected article the best start and end point of the answer using a BERT-model finetuned for Question-Answering. [2]

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

Corpora

Wikipedia Abstract Collection:

+ Covers large amount of topics
− Often very low quality content

Wikipedia abstracts provided directly by the Wikimedia foundation. Mostly not very well curated and in need of additional clean-up and pre-processing.

Kaggle simple - Wikipedia Abstracts:

+ Properly curated
− Comparably smaller and with random scope of topics

Dataset from Kaggle (V1) consisting of around 250k Wikipedia abstracts that were cleaned and prepared for NLP tasks.

COVID-19 Passages (CORD-19):

+ High information density on current topic
− Model not trained on technical language

Contains an early and filtered version of the COVID-19 Open Research Dataset (CORD-19), containing over 50000 scientific papers and excerpts.

Special Cases

English as Input Language:

Translation Module is ignored, Pipeline connects user question directly to Q.A. Module and gives output directly in English.

Input Language not recognized/ available:

If language detection module doesn't recognize a language, the recognition is below a threshold value or the used language is not available, the user is prompted to rephrase the question.

Examples

Dataset	Detected Language	Confidence Level	Question Input	Answer Output
Wikipedia - Kaggle	Spanish	97,2%	*¿Qué es Adobe Illustrator?*	*un programa de computadora para hacer diseño gráfico e ilustraciones*
Covid Passages	Spanish	99,3%	*¿Aumenta el tabaquismo el riesgo de contraer Corona?*	*podría aumentar la capacidad de las sarras - cov - 2 para entrar e infectar el cerebro*
Wikipedia - Abstracts	Spanish	99,8%	*¿Cómo se define una señal analógica?*	*amplitud, fase y frecuencia*
Covid Passages	English	99,9%	*How long do I have Corona symptoms?*	*long*

1) We used the Whoosh library with the Okapi BM25 algorithm (BM-> Best Matching)
2) 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