Monitio – Large Scale MT for Multilingual Media Monitoring

Carlos Amaral
Priberam
Lisbon, Portugal
carlos@priberam.pt

Sebastião Miranda
Priberam
Lisbon, Portugal
ssm@priberam.pt

Abstract

Monitio is a real-time crosslingual global media monitoring platform which delivers actionable insights beyond human scale and capabilities. Our system continuously ingests a massive number of multilingual data sources that are automatically translated, filtered and categorized to generate intelligence reports specially geared towards media monitoring professionals' needs.

1 Origin

The starting point of Monitio was a multilingual media monitoring prototype developed between 2016–2019 in tight collaboration with the British Broadcast Corporation (BBC) and Deutsche Welle (DW).[2] Both broadcasters monitor a growing number of video streams in different languages, by assigning teams of human analysts that are grouped by languages. This approach is not scalable hence the need to reinvent media monitoring, tackling it globally and in a scalable fashion to break the current internationalization and scalability barriers.

The emergence of mature natural language processing (NLP) and artificial intelligence technologies gives European companies an opportunity to push Europe to the leadership of the media monitoring market, where multilinguality is a major issue and, simultaneously, a major opportunity.

By integrating machine translation (MT) in the ingestion and enrichment pipeline, Monitio enables the monitoring of sources in languages the human analyst is not fluent in, providing a truly global view of the events not culturally, geographically or politically biased for lack of access to a broader set of sources.

2 Challenges

A major challenge for Monitio is the huge volume of data ingested daily into the platform. One of Monitio's goals is to process 10 million new multilingual items per day. This has to be accomplished with a minimum delay to enable near real-time monitoring.

The automatic translation of this amount of documents is just one of the challenges for the platform's enrichment pipeline. All content entering the platform is subject to a series of NLP steps, namely its classification according to a standard topics taxonomy, the recognition of named entities like people, organizations, brands and places and linking them to external knowledge bases like Wikipedia or Wikidata, the production of a summary, and the clustering of all articles related to the same event in storylines.

Adding to the complexity, the large-scale dissemination of content on social media platforms, while ensuring a broad coverage of multiple connected viewpoints on the same subject of interest, stresses the problem of information verification which is a daunting task without the support of automation. Failing to address this problem leads to biased views on the subject and insufficient (or even wrong) insights.

[2] http://summa-project.eu

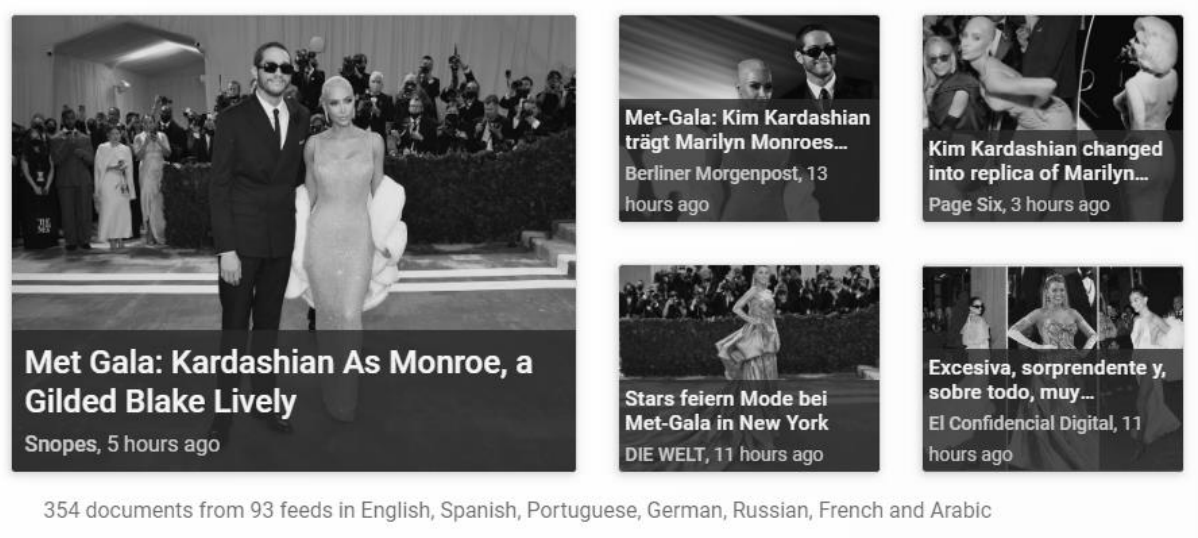

Figure 1: Multilingual Storyline

3 Solution

The scalability goals but also the cost implications led to the integration of free/open source MT models like EasyNMT from Hugging Face, deployed in a dedicated GPU infrastructure.

By employing MT in the beginning of the pipeline, Monitio enables indexing of the documents in all languages translated to a language users understand (e.g., English), thus allowing the users to search documents in other languages. On the other hand, Monitio does not employ MT before the NLP steps, which are executed in the source language of the documents to minimize error propagation.

To enable processing and organization of multilingual documents in different stories and topics, Monitio employs transfer learning through contextual multilingual DistilBERT sentence transformers[3] for crosslingual document clustering, topic detection and entity linking.

One NLP task which still relies heavily on language specific annotated corpora is named-entity recognition, which is one of the most difficult tasks to train generalized multilingual models.

4 Future

When a translation of better quality is needed, for instance, to be included in a report or to clear any doubt that may arise from the default MT, the user will be able to invoke third-party services on demand for a specific document.

Monitio will also integrate automatic speech recognition combined with MT to transcribe and translate video and audio content using wav2vec, an end-to-end deep learning model.

Another objective of the Monitio project is the creation of innovative tools for assisted fact checking. We are developing tools that help the users to verify a claim using the multilingual information available in the platform.

Acknowledgments

The European Union's Horizon 2020 FTI (Fast Track to Innovation) program is funding the productization and the implementation of the go-to-market strategy plan of Monitio under grant agreement No 965576, a project also named Monitio.[4]

References

Santos, João, Afonso Mendes and Sebastião Miranda, "Simplifying News Clustering Through Projection from a Shared Multilingual Space" in Proceedings of Text2Story (ECIR, 2022), pp. 15–24.

Ferreira, Pedro, Ruben Cardoso and Afonso Mendes, "Priberam Labs at the 3rd shared task on SlavNER", Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing (EACL 2021), pp. 86–92.

Miranda, Sebastião, David Nogueira, Afonso Mendes, Andreas Vlachos, Andrew Secker, Rebecca Garrett, Jeff Mitchel and Zita Marinho, "Automated Fact-Checking in the News Room", The Web Conference 2019, San Francisco, pp. 3579–358.

[3] https://huggingface.co/sentence-transformers

[4] https://monitio-project.eu/