SUMMA: Scalable Understanding of Multilingual Media

Ulrich Germann
University of Edinburgh
ugermann@ed.ac.uk

Alexandra Birch
University of Edinburgh
a.birch@ed.ac.uk

Peggy van der Kreeft
Deutsche Welle
peggy.van-der-kreeft@dw.com

Guntis Barzdins
University of Latvia
guntis.barzdins@lu.lv

Steve Renals
University of Edinburgh
s.renals@ec.ac.uk

ABSTRACT

SUMMA is a 3-year, multi-partner *Research and Innovation Action* (2/2016–1/2019), funded under the European Union's *Horizon 2020 Framework Programme*. The project's aim is to develop a highly scalable open-source software platform for monitoring live TV and radio broadcasts as well as internet-based content (text, video, and audio), and ingesting said content into a database of news reporting.

This extended abstract briefly describes the SUMMA Platform's functionality, architecture, and use cases.

ACM Reference Format:
Ulrich Germann, Alexandra Birch, Peggy van der Kreeft, Guntis Barzdins, and Steve Renals. 2018. SUMMA: Scalable Understanding of Multilingual Media. In *Proceedings of KDD Project Showcase Track (KDD '18).* ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

The open-source SUMMA Platform is a highly scalable, distributed architecture for monitoring a large number of media broadcasts in parallel, with a lag behind actual broadcast time of at most a few minutes. The Platform assembles state-of-the-art NLP technologies into a fully automated media ingestion pipeline that can record live broadcasts, detect and transcribe spoken content, translate from several languages (original text or transcribed speech) into English,[1] recognize Named Entities, detect topics, cluster and summarize documents across language barriers, and extract and store factual claims contained in these news items in a database.

Figure 1 shows the workflow. Incoming media streams are downloaded and/or recorded, depending on the source.[2] Audio is transcribed, non-English material is translated. The resulting text-based news items are then processed with downstream NLP modules: topic detection; named entity recognition and linking, and extraction of relations between named entities to build up a knowledge base of "facts" (i.e., factual claims made in news reporting). Similar documents are clustered into storylines, for which multi-document summaries are created via extractive summarisation.

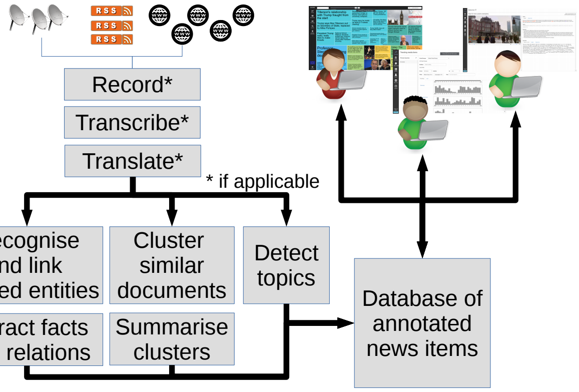
RSS
RSS
RSS
Record*
Transcribe*
Translate*
* if applicable
Recognise and link named entities
Cluster similar documents
Detect topics
Extract facts and relations
Summarise clusters
Database of annotated news items

Figure 1: The SUMMA Architecture

All information is stored in a central database that can be accessed by users via web-browser-based user interfaces and programmatic APIs.

2 IMPLEMENTATION

One explicit design goal in designing the Platform was to assemble state-of-the-art NLP technologies into an effective work flow while minimizing the integration effort required to do so. To this end, the Platform is designed as an assembly of microservices that each perform specific NLP tasks, such as speech recognition, machine translation, etc. Each service is encapsulated in a Docker[3] application container. This allows each service to specify its own software dependencies and eliminates conflicts between software dependencies for different services within the Platform. The overall Platform infrastructure is managed by Docker Compose (single-host deployment) or Docker Swarm (distributed deployment). A task scheduler orchestrates communication between the modules via a message queue.[4] The use of a message queue facilitates scaling of the platform: when all workers of the same type share the same message queues (one for open tasks, one for completed tasks), we can easily increase throughput by instantiating more task workers

[1] The choice of English as the lingua franca within the Platform is due to the working language of our use case partners; the highly modular design of the Platform allows the easy integration of custom translation engines, if required.

[2] Transient signals such as broadcasts received via Satellite are recorded; for sources with a persistent URL, only the URL is stored long-term, although all data may be stored locally temporarily during processing.

[3] www.docker.com

[4] www.rabbitmq.com

KDD '18, August 2018, London, UK

for the task in question, all of which will retrieve task assignments from and push completed tasks onto the same message queues.

3 USE CASES

Three use cases drive the project.

3.1 External Media Monitoring

In continuous operation since 1939, BBC Monitoring (BBCM) is a business unit within the BBC tasked with monitoring and digesting international news broadcasts and other media as an internal service to the BBC as well as a paid service to outside customers. Each of its currently 200 media monitoring journalists can monitor up to 4 live streams in parallel. However, with access to over 1,500 TV channels and ca. 1,350 radio stations, not to mention sources available on the internet, BBCM can currently only keep track of a small fraction of all the sources it has access to.

By alleviating BBCM's staff journalists from mundane monitoring tasks, the SUMMA platform will allow them to widen their monitoring coverage and focus on news interpretation and analysis rather than just keeping track of world-wide news coverage.

3.2 Internal Monitoring

Deutsche Welle (DW) is an international broadcaster covering news world-wide in 30 different languages. Regional news rooms produce and broadcast content independently. Monitoring DW's output with the SUMMA platform will enable DW as an organisation to better keep track of its own output and determine which stories have been covered where, and where there are gaps in the coverage.

3.3 Data Journalism

The SUMMA database will give journalists access to many thousands of news stories with additional metadata such as named entity tags provided by SUMMA's NLP processing modules, providing for large-scale analysis of the constantly evolving news landscape.

4 LANGUAGES COVERED

The SUMMA Platform currently can transcribe English, German, Arabic, Russian, Spanish, Latvian and offers translation from German, Arabic, Russian, Spanish, and Latvian into English. Coverage of Farsi, Portuguese, and Ukranian is work in progress. Due to licensing restrictions, unfortunately not all NLP models will be available under open-source licenses.

5 COMPONENT PERFORMANCE

Due to the large number of components within the system (languages covered times NLP processing capabilities), and the space limitations for extended abstract, we are not able to provide a performance evaluation of the individual components of the system here. Details about system performance as of late 2017 can be found in the project delivarables D3.1 [1] (speech recognition, machine translation, metadata extraction, clustering, and topic detection), D4.1 [3] (knowledge base construction), and D5.1 [2] (semantic parsing, summarization), which are available from the project's web site.[5] Further evaluations are planned for late 2018 and will be available as project deliverables from the project's web site in early 2019.

6 SCALABILITY

In a recent scalability test (Jan. 2018), we were able to ingest 400 TV channels in parallel on the Amazon AWS EC2 web cloud service infrastructure.

7 AVAILABILITY

The SUMMA Platform is currently scheduled to be released as open-source software by the end of August 2018 and will be available through the project's web site at http://www.summa-project.eu.

8 PROJECT CONSORTIUM

The project consortium comprises seven technical partners — the University of Edinburgh (co-ordinator; UK), the Latvian Information Agency (LETA; LV), University College London (UK), the University of Sheffield (UK), Idiap Research Institute (CH), Priberam Labs (PT), the Qatar Computing Research Institute (QA) —, and two use-case partners: BBC Monitoring (UK) and Deutsche Welle (DE).

ACKNOWLEDGEMENTS

This work was conducted within the scope of the Research and Innovation Action *SUMMA*, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688139.

REFERENCES

[1] Philip N. Garner, Alexandra Birch, Andrei Popescu-Belis, Peter Bell, Herve Bourlard, Steve Renals, Sebastião Miranda, and Ulrich Germann. 2017. *SUMMA Deliverable D3.1: Initial Progress Report on Shallow Stream Processing.* Technical Report. The SUMMA Consortium.

[2] Afonso Mendes, Pedro Balage, Mariana Almeida, Sebastião Miranda, Nikos Papasarantopoulos, Shashi Narayan, and Shay Cohen. 2017. *SUMMA Deliverable 5.1: Initial Progress Report on Natural Language Understanding.* Technical Report. The SUMMA Consortium.

[3] Abiola Obamuyide, Andreas Vlachos, Jeff Mitchell, David Nogueira, Sebastian Riedel, Filipe Aleixo, Samuel Broscheit, Andre Martins, Mariana Almeida, Sebastião Miranda, Afonso Mendes, and Andrei Popescu-Belis. 2017. *SUMMA Deliverable D4.1: Initial Progress Report on Automatic Knowledge Base Creation.* Technical Report. The SUMMA Consortium.

[5] www.summa-project.eu/deliverables