The SUMMA Platform Prototype

Renars Liepins[†]
renars.liepins@leta.lv

Ulrich Germann[♂]
ugermann@inf.ed.ac.uk

Guntis Barzdins[†] · **Alexandra Birch**[♂] · **Steve Renals**[♂] · **Susanne Weber**[♀]
Peggy van der Kreeft[♁] · **Hervé Bourlard**[♁] · **João Prieto**[♁] · **Ondřej Klejch**[♂]
Peter Bell[♂] · **Alexandros Lazaridis**[♁] · **Alfonso Mendes**[♁] · **Sebastian Riedel**[♀]
Mariana S. C. Almeida[♁] · **Pedro Balage**[♁] · **Shay Cohen**[♂] · **Tomasz Dwojak**[♂]
Phil Garner[♁] · **Andreas Giefer**[♁] · **Marcin Junczys-Dowmunt**[♂] · **Hina Imran**[♁]
David Nogueira[♁] · **Ahmed Ali**[♁] · **Sebastião Miranda**[♁] · **Andrei Popescu-Belis**[♁]
Lesly Miculicich Werlen[♁] · **Nikos Papasarantopoulos**[♂] · **Abiola Obamuyide**[‡]
Clive Jones[♀] · **Fahim Dalvi**[♁] · **Andreas Vlachos**[‡] · **Yang Wang**[♁] · **Sibo Tong**[♁]
Rico Sennrich[♂] · **Nikolaos Pappas**[♁] · **Shashi Narayan**[♂] · **Marco Damonte**[♂]
Nadir Durrani[♁] · **Sameer Khurana**[♁] · **Ahmed Abdelali**[♁] · **Hassan Sajjad**[♁]
Stephan Vogel[♁] · **David Sheppey**[♀] · **Chris Hernon**[♀] · **Jeff Mitchell**[♀]

[†]Latvian News Agency [♂]University of Edinburgh [♁]Deutsche Welle [♀]BBC
[♁]Idiap Research Institute [♁]Priberam Informatica S.A. [♀]University College London
[‡]University of Sheffield [♁]Qatar Computing Research Institute

Abstract

We present the first prototype of the *SUMMA* Platform: an integrated platform for multilingual media monitoring. The platform contains a rich suite of low-level and high-level natural language processing technologies: automatic speech recognition of broadcast media, machine translation, automated tagging and classification of named entities, semantic parsing to detect relationships between entities, and automatic construction / augmentation of factual knowledge bases. Implemented on the *Docker* platform, it can easily be deployed, customised, and scaled to large volumes of incoming media streams.

1 Introduction

SUMMA (Scalable Understanding of Multilingual Media)[1] is a three-year *Research and Innovation Action* (February 2016 through January 2019), supported by the European Union's *Horizon 2020* research and innovation programme. *SUMMA* is developing a highly scalable, integrated web-based platform to automatically monitor an arbitrarily large number of public broadcast and web-based news sources.

Two concrete use cases and an envisioned third use case drive the project.

1.1 Monitoring of External News Coverage

BBC Monitoring, a division of the *British Broadcasting Corporation* (BBC), monitors a broad variety of news sources from all over the world on behalf of the BBC and external customers. About 300[2] staff journalists and analysts track TV, radio, internet, and social media sources in order to detect trends and changing media behaviour, and to flag breaking news events. A single monitoring journalist typically monitors four TV channels and several online sources simultaneously. This is about the maximum that any person can cope with mentally and physically. Assuming 8-hour shifts, this limits the capacity of *BBC Monitoring* to monitoring about 400 TV channels at any given time on average. At the same time, *BBC Monitoring* has access to about 13,600 distinct sources,

[1] www.summa-project.eu

[2] To be reduced to 200 by the end of March, 2017.

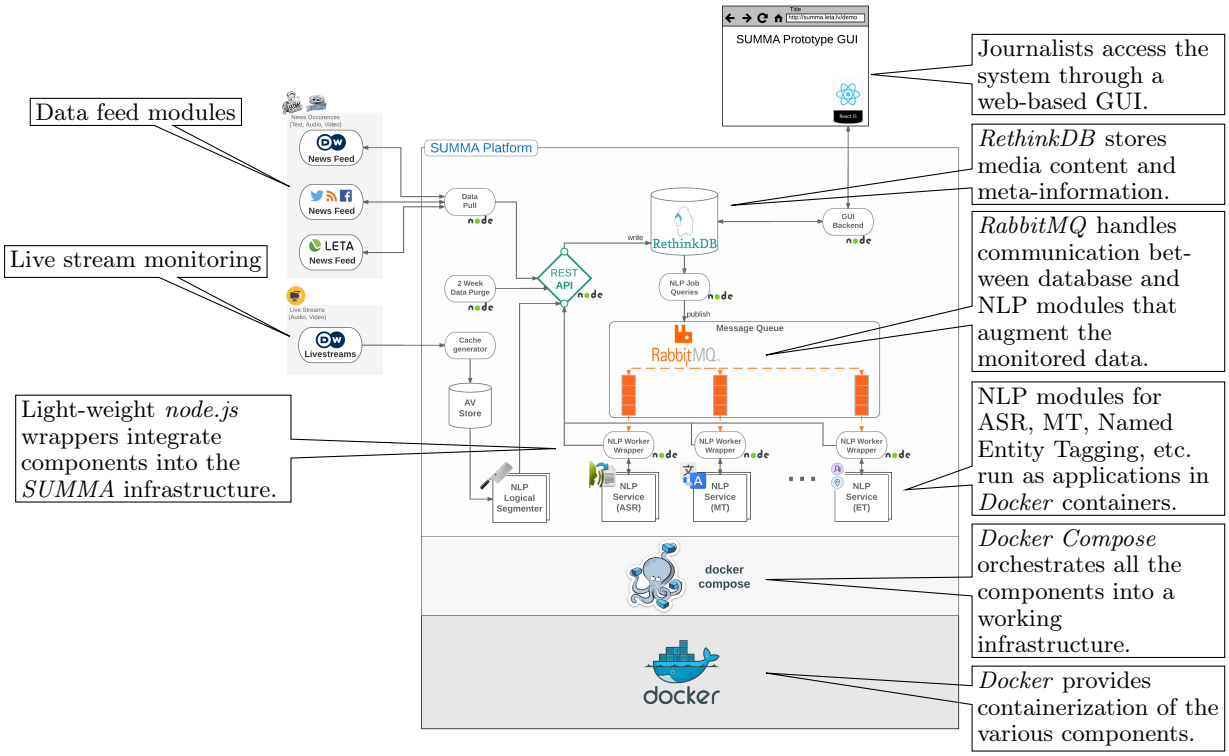

Figure 1: Architecture of the *SUMMA* Platform

including some 1,500 TV and 1,350 radio broadcasters. Automating the monitoring process not only allows the BBC to cover a broader spectrum of news sources, but also allows journalists to perform deeper analysis by enhancing their ability to search through broadcast media across languages in a way that other monitoring platforms do not support.

1.2 Monitoring Internal News Production

Deutsche Welle is Germany's international public service broadcaster. It provides international news and background information from a German perspective in 30 languages worldwide, 8 of which are used within *SUMMA*. News production within *Deutsche Welle* is organized by language and regional departments that operate and create content fairly independently. Interdepartmental collaboration and awareness, however, is important to ensure a broad, international perspective. Multilingual internal monitoring of world-wide news production (including underlying background research) helps to increase awareness of the work between the different language news rooms, decrease latency in reporting and reduce cost of news production within the service by allowing adaptation of existing news stories for particular target audiences rather than creating them from scratch.

1.3 Data Journalism

The third use case is data journalism. Measurable data is extracted from the content available in and produced by the *SUMMA* platform and graphics are created with such data. The data journalism dashboard will be able to provide, for instance, a graphical overview of trending topics over the past 24 hours or a heatmap of storylines. It can place geolocations of trending stories on a map. Customised dashboards can be used to follow particular storylines. For the internal monitoring use case, it will visualize statistics of content that was reused by other language departments.

2 System Architecture

Figure 1 shows an overview of the *SUMMA* Platform prototype. The Platform is implemented as an orchestra of independent components that run as individual containers on the *Docker* platform. This modular architecture gives the project partners a high level of independence in their development.

The system comprises the following individual processing components.

2.1 Data Feed Modules and Live Streams

These modules each monitor a specific news source for new content. Once new content is available, it is downloaded and fed into the database via a common REST API. Live streams are automatically

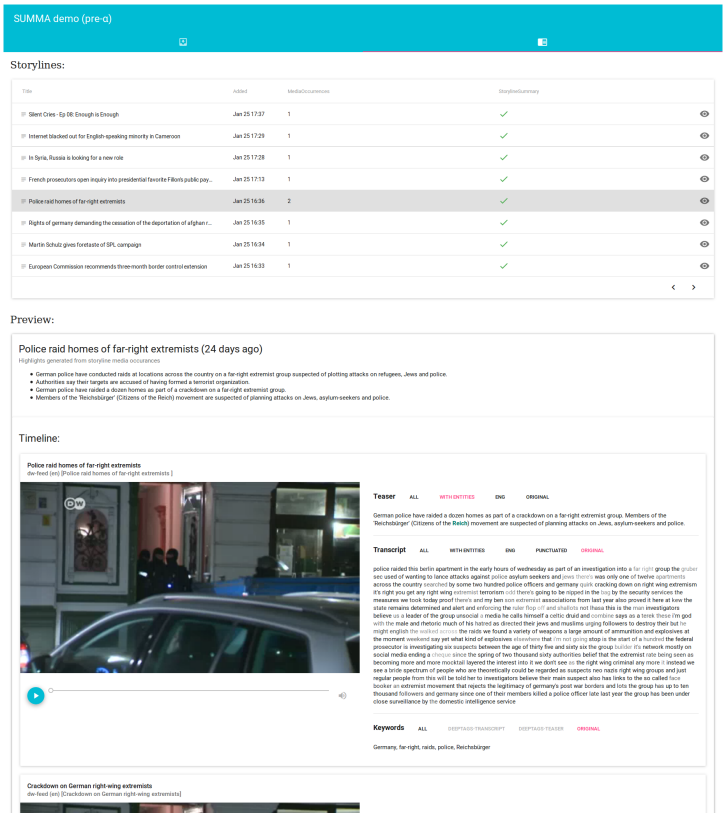

Figure 2: Web-based User Interface of the *SUMMA* Platform (Storyline View)

segmented into logical segments.

2.2 Database Back-end

Rethink-DB[3] serves as the database back-end. Once new content is added, *Rethink-DB* issues processing requests to the individual NLP processing modules via *RabbitMQ*.

2.3 Automatic Speech Recognition

Spoken language from audio and video streams is first processed by automatic speech recognition to turn it into text for further processing. Models are trained on speech from the broadcast domain using the *Kaldi* toolkit (Povey et al., 2011); speech recognition is performed using the *CloudASR* platform (Klejch et al., 2015).

2.4 Machine Translation

The lingua franca within *SUMMA* is English. Machine translation based on neural networks is used to translate content into English automatically. The back-end MT systems are trained with the *Nematus* Toolkit (Sennrich et al., 2017); translation is performed with *AmuNMT* (Junczys-Dowmunt et al., 2016).

2.5 Entity Tagging and Linking

Depending on the source language, Entity Tagging and Linking is performed either natively, or on the English translation. Entities are detected with *TurboEntityRecognizer*, a named entity recognizer within *TurboParser*[4] (Martins et al., 2009). Then, we link the detected mentions to the knowledge base with a system based on our submission to TAC-KBP 2016 (Paikens et al., 2016).

2.6 Topic Recognition and Labeling

This module labels incoming news documents and transcripts with a fine-grained set of topic labels. The labels are learned from a multilingual corpus of nearly 600k documents in 8 of the 9 *SUMMA* languages (all except Latvian), which were manually annotated by journalists at *Deutsche Welle*. The document model is a hierarchical attention network with attention at each level of the hierarchy, inspired by Yang et al. (2016), followed by a sigmoid classification layer.

2.7 Deep Semantic Tagging

The system also has a component that performs semantic parsing into Abstract Meaning Representations (Banarescu et al., 2013) with the aim to incorporate them into the storyline generation eventually. The parser was developed by Damonte et al. (2017).[5] It is an incremental left-to-right parser that builds an AMR graph structure using a neu-

[3] `www.rethinkdb.com`

[4] `https://github.com/andre-martins/TurboParser`

[5] Demo at `http://cohort.inf.ed.ac.uk/amreager.html`.

ral network controller. It also includes adaptations to German, Spanish, Italian and Chinese.

2.8 Knowledge Base Construction

This component provides a knowledge base of factual relations between entities, built with a model based on *Universal Schemas* (Riedel et al., 2013), a low-rank matrix factorization approach.The entity relations are extracted jointly across multiple languages, with entities pairs as rows and a set of structured relations and textual patterns as columns. The relations provide information about how various entities present in news documents are connected.

2.9 Storyline Construction and Summarization

Storylines are constructed via online clustering, i.e., by assigning storyline identifiers to incoming documents in a streaming fashion, following the work in Aggarwal and Yu (2006). The resulting storylines are subsequently summarized via an extractive system based on Almeida and Martins (2013).

3 User Interface

Figure 2 shows the current web-based *SUMMA* Platform user interface in the *storyline* view. A storyline is a collection of news items that concerning a particular "story" and how it develops over time. Details of the layout are explained in the figure annotations.

4 Future Work

The current version of the Platform is a prototype designed to demonstrate the orchestration and interaction of the individual processing components. The look and feel of the page may change significantly over the course of the project, in response to the needs and requirements and the feedback from the use case partners, the BBC and *Deutsche Welle*.

5 Availability

The public release of the *SUMMA* Platform as open source software is planned for April 2017.

6 Acknowledgments

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

Aggarwal, Charu C and Philip S Yu. 2006. "A framework for clustering massive text and categorical data streams." *SIAM Int'l. Conf. on Data Mining*, 479–483.

Almeida, Miguel B and Andre FT Martins. 2013. "Fast and robust compressive summarization with dual decomposition and multi-task learning." *ACL*, 196–206.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. "Abstract meaning representation for sembanking." *Linguistic Annotation Workshop*.

Damonte, Marco, Shay B. Cohen, and Giorgio Satta. 2017. "An incremental parser for abstract meaning representation." *EACL*.

Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. 2016. "Is neural machine translation ready for deployment? A case study on 30 translation directions." *CoRR*, abs/1610.01108.

Klejch, Ondřej, Ondřej Plátek, Lukáš Žilka, and Filip Jurčíček. 2015. "CloudASR: platform and service." *Int'l. Conf. on Text, Speech, and Dialogue*, 334–341.

Martins, André FT, Noah A Smith, and Eric P Xing. 2009. "Concise integer linear programming formulations for dependency parsing." *ACL*, 342–350.

Paikens, Peteris, Guntis Barzdins, Afonso Mendes, Daniel Ferreira, Samuel Broscheit, Mariana S. C. Almeida, Sebastião Miranda, David Nogueira, Pedro Balage, and André F. T. Martins. 2016. "SUMMA at TAC knowledge base population task 2016." *TAC*. Gaithersburg, Maryland, USA.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. "The Kaldi speech recognition toolkit." *ASRU*.

Riedel, Sebastian, Limin Yao, Benjamin M. Marlin, and Andrew McCallum. 2013. "Relation extraction with matrix factorization and universal schemas." *HLT-NAACL*.

Sennrich, Rico, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. "Nematus: a toolkit for neural machine translation." *EACL Demonstration Session*. Valencia, Spain.

Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. "Hierarchical attention networks for document classification." *NAACL*. San Diego, CA, USA.