

The SUMMA Platform: Scalable Understanding of Multilingual Media

Ulrich Germann,[♦] Peggy van der Kreeft,[♣] Guntis Barzdins,^{♣♥} Alexandra Birch[♦]

[♦] University of Edinburgh; [♣] Deutsche Welle; [♣] LETA; [♥] University of Latvia

for the SUMMA Consortium*

corresponding authors: ugermann@ed.ac.uk, peggy.van-der-kreeft@dw.com

Abstract

We present the latest version of the SUMMA platform, an open-source software platform for monitoring and interpreting multi-lingual media, from written news published on the internet to live media broadcasts via satellite or internet streaming.

1 Introduction

The SUMMA platform is a highly scalable open-source infrastructure for monitoring and interpreting news streams in multiple languages,¹ and a variety of media formats, from written text published on the internet to live TV broadcasts via satellite.

Three use cases drive the project.

External Media Monitoring

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.

The SUMMA platform will allow BBCM’s staff journalists to widen their monitoring coverage and focus on news interpretation and analysis by alleviating them from mundane monitoring tasks.

Internal Monitoring

Deutsche Welle (DW) is an international broadcaster covering world-wide news 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.

Data Journalism

The SUMMA database will give journalists access to many thousands of news stories with additional

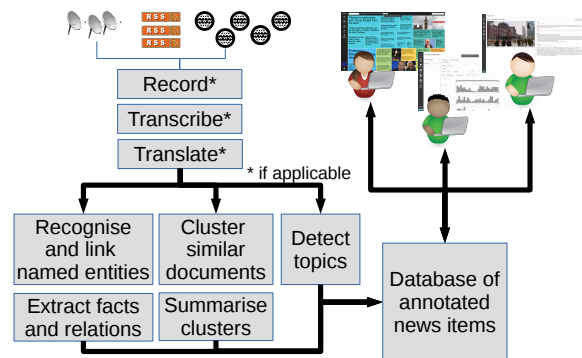


Figure 1: The SUMMA Platform Architecture


metadata such as named entity tags provided by SUMMA’s NLP processing modules, providing for large-scale analysis of the constantly evolving news landscape.

2 Architecture

The design of the SUMMA platform is shown in Fig. 1. Incoming media streams are downloaded and/or recorded, depending on the source. Audio is automatically transcribed, and non-English material is machine-translated into English. 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); and document clustering and multi-document cluster summarization.

News items, mentions of named entities, etc., are stored in a central database that can be accessed by users via web-browser-based user interfaces, or by programs via programmatic interfaces (APIs).

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.

* The SUMMA Consortium comprises the University of Edinburgh, LETA, Idiap Research Institute, Priberam, Qatar Computing Research Institute, University College London, and Sheffield University as research partners, and the BBC and Deutsche Welle as use case partners.

* © 2018 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

¹ Arabic, German, English, Farsi,^{*} Latvian,^{*} Portuguese,^{*} Russian, Spanish, Ukrainian^{*} (* planned for late 2018)