Scalable Understanding of Multilingual MediA (SUMMA)

http://www.summa-project.eu

H2020 Research and Innovation Action
Number: 688139
D3.2 – Release of stream processing tools, version 1

Nature	Report	**Work Package**	WP3
Due Date	30/04/2017	**Submission Date**	04/05/2017
Main authors	Philip N. Garner		
Co-authors	Lexi Birch, Peter Bell & Andrei Popescu-Belis		
Reviewers	Steve Renals & Hervé Bourlard		
Keywords	Multilingual speech recognition, machine translation, metadata extraction, clustering and topic detection		
Version Control			
v0.1	**Status**	Draft	24-April-2017
v1.0	**Status**	Draft	28-April-2017
v1.1	**Status**	Final	02-May-2017

Contents

Abstract

WP3 is concerned with taking large volumes of multimodal media input, and providing streams of text (in both the source language and automatically translated to English), segmented into stories and clustered into topics. In this deliverable we report progress in WP3 in terms of the creation and release of tools to enable these component technologies, and to enable their integration with the SUMMA platform.

1 Introduction

SUMMA WP3 takes large volumes of multimodal media input, and provides streams of text (in both the source language and automatically translated to English), segmented into stories and clustered into topics. Key to the success of this workpackage is the ability to operate over large volumes of streaming data, using methods which are scalable and efficient.

The final output of this WP is the entry point to the media monitoring platform. It provides the essential processing necessary for further analysis by WP4 and WP5, including:

- Accurate, adaptive multilingual automatic speech recognition allowing spoken content to be made available as text to the higher semantic level processing in the project;
- Automatic machine translation of streaming transcription output;
- Streaming metadata extraction;
- Story clustering.

After careful analysis of all the system requirements, the SUMMA consortium agreed to follow a fully scalable and distributed data stream processing architecture based on Docker components (www.docker.com), as illustrated in Figure 1. Technologies are thus released from WP3 to the wider project in the form of Docker containers. Within WP3, however, lower level tools are shared to allow the individual partners to collaborate on creating the technologies.

This deliverable reports on the initial, consortium-internal software release, building on the core research and development on shallow stream processing carried out in WP3 (to be reported in D3.1 – Initial progress report on shallow stream processing). The main goal of the current report is to provide an overview of the SUMMA stream processing architecture, together with the delivery status of the components.

Note: This document is not intended as a full technical description of the technologies developed in WP3 – that information is contained in Deliverable D3.1 (M18). The current deliverable details those technologies that comprise the initial release of shallow stream processing tools. This deliverable is related to Deliverable D8.3, which lists components for IP purposes at a coarser granularity.

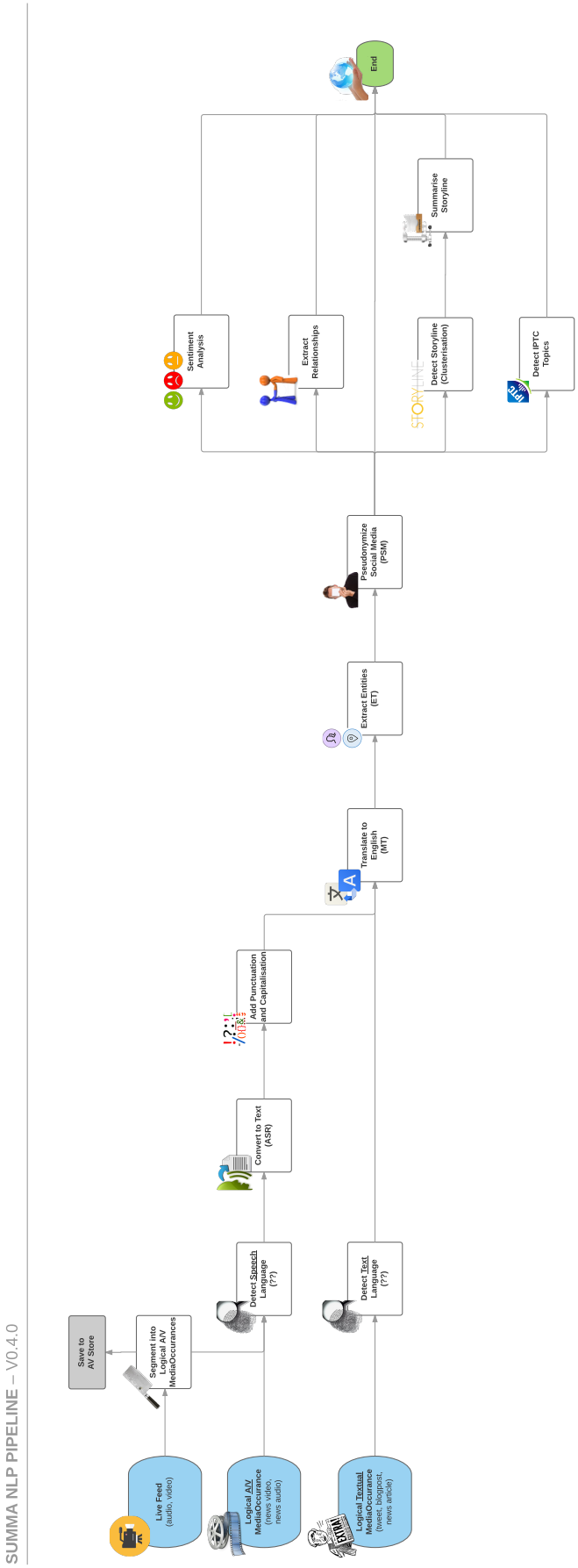

Figure 1: SUMMA data-flow as Docker components, indicating which technologies are expected from WP3. The diagram is taken from the SUMMA docker repository, where an electronic version can be found. Availability of the components are indicated in Section 4.2.

2 Software release by task

T3.1: Multilingual speech recognition

All automatic speech recognition (ASR) in SUMMA uses the Kaldi toolkit (Povey et al., 2011). The ASR Docker component used by the SUMMA platform is based on the Cloud ASR engine (Klejch et al., 2015) developed prior to the start of the project by a SUMMA researcher. Scripts to train models in each language are exchanged via a Git repository. The method for training models varies between different languages, according to type and quantity of language resources available. For example, for English, we have a hand-created pronunciation dictionary, whereas for Arabic, we train grapheme-based models because no dictionary is available. In English and German, we have large corpora of training data with time-aligned transcriptions available, whilst in other languages we must learn these alignments automatically as part of the model training process, or else use smaller, mismatched training corpora.

The full status of the models and training data is tracked on the SUMMA wiki page at https://www.wiki.ed.ac.uk/display/SUM/ASR, on a language-by-language basis. We break down the status as follows:

- **Initial:** A working system, but not necessarily expected to work well in SUMMA due to small or mismatched training data (for example, models trained on the GlobalPhone corpus of read speech)
- **Advanced:** A system that works well, and may also be expected to work well in SUMMA, for instance because it was trained on broadcast data.
- **Tuned:** A system that has been trained on or adapted to SUMMA-specific data.

We have currently released advanced components for English and Arabic, and initial components for German and Latvian.

T3.2: Machine translation:

SUMMA includes state-of-the-art machine translation capabilities. We base our models on our neural machine translation software which performed well in last year's shared task in the conference of machine translation (WMT). We came first in seven out of eight language pairs in which we competed.

Development work that has been done for the SUMMA project include making training and decoding fast by developing a C++ decoder called Marian/AMUNMT. This toolkit has the following features:

- Up to 15x faster translation than Nematus and similar toolkits on a single GPU
- Up to 2x faster training than toolkits based on Theano, Tensorflow, Torch on a single GPU
- Multi-GPU training and translation
- Batched translation on single and multiple GPUs
- Pure C++ implementation with minimal depedencies on external packages (cuda, boost)
- Optionally static compilation of binaries
- Permissive open source license (MIT)

Source languages that are currently supported in SUMMA are the following (all translation is into English): German, Arabic, Spanish and Russian.

T3.3: Metadata extraction

Punctuation prediction is a matter for research, but a path to docker integration has been identified. Two implementations are available, each via git repository.

The Idiap punctuation prediction is based on three "features" of the speech stream, being:

1. Pause detection, where length of pause is indicative of punctuation such as commas and sentence boundaries.
2. Language modelling, where punctuation is included in the language model such that punctuation marks appear at grammatically plausible places.
3. Ad-hoc rules, allowing language-specific peculiarities.

The punctuator was developed using (Swiss) French and German; it is currently available in English and German in SUMMA.

UEdin's punctuation system (Klejch et al., 2017) uses an approach based on neural machine translation (NMT), where the translation is from unpunctuated text to punctuated text. Acoustic information is incorporated by replacing the encoder part of a standard NMT system with a hierarchical encoder, which enables the frame-based acoustic features to be combined with word-based lexical features.

Text normalisation software has been made available by Idiap. This allows consistent removal of spurious characters, expansion of acronyms, conversion of numbers and the like. It has been used for normalisation of German.

Baseline speaker diarisation software is available from Idiap.

T3.4: Scalable news clustering and topic detection.

The initial baseline English clustering solution, inspired by Aggarwal and Yu (2006) was developed by LETA. It is based on TF-IDF feature extraction with single link hierarchical clustering and constant cut-off threshold. This clustering solution has been wrapped in a Docker container and used in the SUMMA Platform initial releases with promising performance results.

Following this initial baseline clustering release extensive research was conducted by UEDIN, Priberam and LETA to further improve the clustering results in multilingual setting. This research is described in the submissions Znotins et al. (2017) and Miranda et al. (2017), and resulted in a production-ready clustering system with the following specifications:

- Pure C++ production-ready module.
- SOTA online monolingual clustering results for English, Spanish and German.
- Online crosslingual clustering based on crosslingual embeddings and timestamps.
- Ready for all SUMMA languages, after training crosslingual embeddings for all languages.

This new monolingual and crosslingual clustering method has been wrapped in the Docker container recently, and will be integrated into the upcoming SUMMA Platform release.

IDIAP has developed a topic labelling solution based on hierarchical deep neural networks with attentions. The system was trained and tested using the Deutsche Welle topic labelled data. The system is wrapped in a Docker container and is integrated in the SUMMA Platform. This is a truly multilingual system, as it learns from a set of documents in several languages, with labels of various possible granularities in several languages. The system can thus label new documents in any language, with labels from all languages available in the training data. An important benefit

of sharing layers and/or attention models across languages is the capacity to transfer knowledge to low-resource languages and thus improve over monolingual models.

3 Release overview

3.1 Release mechanisms

Within WP3, the following release mechanisms are currently in use:

SUMMA wiki: Accessible to all partners; information only
https://www.wiki.ed.ac.uk/display/SUM/summa+Home

SUMMA GitHub: Private (project-only) GitHub repository containing software releases to be combined with the SUMMA platform

Idiap GitLab: Idiap's GitLab repository is closed source but some SUMMA partners have accounts enabling access.
https://gitlab.idiap.ch

Public websites: Some tools are available to the wider community.

Ad-hoc: Components are transferred between partners without them necessarily being available to the wider project.

The current release of WP3 components is within the project, for integration with the SUMMA platform demonstration system. By the end of the project we plan extensive open source releases, enabling an open source version of the SUMMA platform.

3.2 Release overview table

The table below lists the tools that have been made available to the other partners for creating of technologies. Where the technology is packaged as a Docker component, that is also indicated.

Name	Copyright	Licence	Docker
English ASR tools	UEdin	Apache 2.0	Y
Advanced	URL: http://www.mgb-challenge.org/recipe.html		
German ASR tools	Idiap	SUMMA only	Y
Initial	URL: https://gitlab.idiap.ch/speech/kaldi-bcn		
Arabic ASR tools	QCRI	SUMMA only	Y
Advanced	URL: https://github.com/qcri/ArabicASRChallenge2016		
Latvian ASR tools	LETA	SUMMA only	Y
Initial	URL: http://runa.korpuss.lv/lv_swbd_nnet2_ms_1e8.zip		
German-English MT	UEdin	MIT	Y
Advanced	URL: https://github.com/amunmt		
Arabic-English MT	QCRI	SUMMA only	Y
Advanced	URL: http://qatsdemo.cloudapp.net/qats/		
Spanish-English MT	UEdin	MIT	Y
Initial	URL: https://github.com/amunmt		
Russian-English MT	UEdin	MIT	Y
Initial	URL: https://github.com/amunmt		
Punctuation prediction	Idiap	SUMMA only	N
Advanced	URL: https://gitlab.idiap.ch/speech/punctuation		
Punctuation prediction	UEdin	Apache 2.0	Y
Advanced	URL: https://github.com/choko/acoustic_punctuation		
Text normalisation	Idiap	BSD 3-clause	N
Advanced	URL: https://github.com/idiap/asrt		
Speaker Diarisation	Idiap	GPL v.3	N
Advanced	URL: https://github.com/idiap/IBDiarization		
Clustering for English	LETA	MIT	Y
Advanced	URL: https://github.com/summa-leta/summa-platform		
Generic Clustering	Priberam	SUMMA only	Y
Advanced			
Multilingual Clustering	Idiap	SUMMA only	Y
Initial			

Based on these tools, the SUMMA demonstration platform now has operational pipelines for English, Arabic, and German.

4 Future plans

We envisage three classes of future actions:

1. **First priority: Moving components into the Docker architecture.** Some components are not yet available as Docker modules; this is indicated in the component table above. Release of a component as a Docker module is taken as that component being usable by the wider project, i.e., as a Docker module, or as part of another module (e.g., punctuation may be merged into speech recognition). We expect all components to be made available in Docker form by M18.
2. **Second priority: Development of more languages.** By M18 we plan to release the following baseline systems:
   - ASR: Persian (Farsi), Portuguese, Russian, Spanish, Ukranian
   - MT: Latvian, Persian (Farsi), Portuguese, Ukranian

   We also plan to improve the current baseline systems for German and Latvian ASR, and Spanish and Russian MT, which are currently baseline systems not adapted to the SUMMA domain. These systems will enable the SUMMA pipeline to operate in further languages by M18.
3. **Third priority: Development of new functionalities.** During 2017 we plan to release software tools for:
   - Text rewriting: multilingual tools to transform text produced by ASR or obtained from social media to a form better matched to the MT or other NLP components.
   - Topic segmentation: multilingual tools to segment real-time streams produced by the ASR system into semantically coherent segments.
4. **Fourth priority: Iterative improvement of all stream processing technology components.** Research and development advances in WP3 technologies will result in the release of updated software components.

References

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The Kaldi speech recognition toolkit. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 1–4, Hawaii, USA, December 2011.

Ondřej Klejch, Ondřej Plátek, Lukáš Žilka, and Filip Jurčíček. Cloudasr: Platform and service. In Pavel Král and Václav Matoušek, editors, *Text, Speech, and Dialogue: 18th International Conference, TSD 2015, Pilsen,Czech Republic, September 14-17, 2015, Proceedings*, pages 334–341. Springer International Publishing, 2015. ISBN 978-3-319-24033-6. doi: 10.1007/978-3-319-24033-6_38. URL http://dx.doi.org/10.1007/978-3-319-24033-6_38.

Ondřej Klejch, Peter Bell, and Steve Renals. Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

Charu C. Aggarwal and Philip S. Yu. A framework for clustering massive text and categorical data streams. In *Proceedings of the Sixth SIAM International Conference on Data Mining, April 20-22, 2006, Bethesda, MD, USA*, pages 479–483, 2006. doi: 10.1137/1.9781611972764.44. URL http://dx.doi.org/10.1137/1.9781611972764.44.

Arturs Znotins, Shay Cohen, Sebastiao Miranda, and Guntis Barzdins. A comparative study of online document clustering algorithms in a unified framework. In *ACL-2017 (submission)*, 2017. URL http://www.ltn.lv/~guntis/acl17news.pdf.

Sebastiao Miranda, Arturs Znotins, Shay Cohen, and Guntis Barzdins. Online monolingual and crosslingual clustering of news articles. In *EMNLP-2017 (submission)*, 2017. URL http://www.ltn.lv/~guntis/emnlp17news.pdf.

ENDPAGE

SUMMA

H2020-ICT-2015 688139

D3.2 Release of stream processing tools, version 1