FP7-ICT Strategic Targeted Research Project (STREP) TrendMiner (No. 287863)

Large-scale, Cross-lingual Trend Mining and Summarisation of Real-time Media Streams

D4.1.1 Multi-Lingual Summarisation of Stream Media Software - v1

Mark A. Greenwood (USFD), Kalina Bontcheva (USFD)

Abstract.
FP7-ICT Strategic Targeted Research Project (STREP) ICT-2011-287863 TrendMiner
Deliverable D1.1 (WP1.1)

In this deliverable we describe two baseline systems for summarizing multi-lingual stream media based upon term clouds and micropinions. These two approaches allow for different types of summarization dependeny upon the requirements from the use cases.

Keyword list: summarization, term clouds, micropinions

Project	TrendMiner No. 287863
Delivery Date	November 1, 2012
Contractual Date	October 31, 2012
Nature	Prototype
Reviewed By	Mauro Navarra (Eurokleis)
Web links	`http://demos.gate.ac.uk/trendminer/summarization/`
Dissemination	PU

TrendMiner Consortium

This document is part of the TrendMiner research project (No. 287863), partially funded by the FP7-ICT Programme.

DFKI GmbH
Language Technology Lab
Stuhlsatzenhausweg 3
D-66123 Saarbrcken
Germany
Contact person: Thierry Declerck
E-mail: declerck@dfki.de

University of Southampton
Southampton SO17 1BJ
UK
Contact person: Mahensan Niranjan
E-mail: mn@ecs.soton.ac.uk

Internet Memory Research
45 ter rue de la Rvolution
F-93100 Montreuil
France
Contact person: France Lafarges
E-mail: contact@internetmemory.org

Eurokleis S.R.L.
Via Giorgio Baglivi, 3
Roma RM
00161 Italy
Contact person: Francesco Bellini
E-mail: info@eurokleis.com

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
UK
Contact person: Kalina Bontcheva
E-mail: K.Bontcheva@dcs.shef.ac.uk

Ontotext AD
Polygraphia Office Center fl.4,
47A Tsarigradsko Shosse,
Sofia 1504, Bulgaria
Contact person: Atanas Kiryakov
E-mail: naso@sirma.bg

Sora Ogris and Hofinger GmbH
Linke Wienzeile 246
A-1150 Wien
Austria
Contact person: Christoph Hofinger
E-mail: ch@sora.at

Hardik Fintrade Pvt Ltd.
227, Shree Ram Cloth Market,
Opposite Manilal Mansion,
Revdi Bazar, Ahmedabad 380002
India
Contact person: Suresh Aswani
E-mail: m.aswani@hardikgroup.com

Changes

Version	Date	Author	Changes
1.0	15.10.2012	Mark A. Greenwood	initial draft
1.1	26.10.2012	Mark A. Greenwood	released for internal review
1.2	30.10.2012	Mark A. Greenwood	incorporated review comments from Mauro Navarra
1.3	30.10.2012	Mark A. Greenwood	changes to the underlying style file
1.4	1.11.2012	Mark A. Greenwood	finalised the cover sheet

Executive Summary

This deliverable describes the first prototypes for the multi-lingual summarization of stream media. We present two different approaches to summarization: term clouds and micropinions. These two approaches differ considerably in the type of summarizations they produce, which will allow us to provide differing styles of summarization to fulfil the use case requirements.

Contents

Chapter 1

Introduction

This deliverable introduces the first prototypes of two baseline approaches to the summarizarion of multi-lingual streaming media. The two GATE [CMB^{+}11] based approaches we have adopted for this deliverable are term clouds and micropinions and will be described in Chapters 2 and 3 respectively. Both approaches currently assume that a collection of media has been pre-selected via stream windowing or classification etc. and therefore focus purely on summarization. Chapter 4 details the software which accompanies this deliverable including information on usage and examples. Our choice of baseline systems was motivated by a thorough review of the current literature which is included with this deliverable as Appendix A.

It should be noted that as this is a prototype deliverable (i.e. focused on software delivery) only a brief overview of each system is provided. Full details are available by following the references given throughout this deliverable.

1.1 Relevance to TrendMiner

TrendMiner aims to work over large volumes of streaming media. Even after sub-set selection (windowing, sentiment classification etc.) the large volumes of data will prohibit users from manually examining every piece of relevant text. Summarization software will allow users to quickly get a sense for the content of a set of streamed media items.

1.1.1 Relevance to project objectives

The work reported in this deliverable provides the software processing required for the summarization of stream media.

1.1.2 Relation to other workpackages

Summarization of stream media is a requirement of the two use cases (WP6 and WP7) and the work reported in this deliverable will form the basis of that work.

Chapter 2

Summarization via Term Clouds

A term cloud is a form of weighted list, which allows text to be visualized by equating font size with importance[1]. In their simplest form term clouds can be generated from the raw tokens present in a document, however, it is usually more useful to perform some level of filtering to select appropriate terms from a corpus which can then be visualized. For example, it is common to form a term cloud from document tags rather than from the document content in order to ensure that only representative terms are included.

Our prototype term cloud summarization system uses TermRaider[2] [DMT$^+$12], which has been developed within the EU funded ARCOMEM[3] project, to extract and score relevant terms from a set of tweets. Terms are sorted alphabetically and scaled based upon their TF.IDF weighting [BS09].

Once terms have been extracted and scored the term cloud visualization is created by scaling each term based upon it's score using the following equation where f_i is the font size to use for term i with a score of t_i.

$$f_i = \frac{f_{max} \times (t_i - t_{min})}{t_{max} - t_{min}} \tag{2.1}$$

In the current implementation we convert f_i to an integer in the range 1 to 10 which are mapped to specific font sizes and colours for use in the term cloud display. The term cloud can be customized by controlling the number of terms to display as well as the base colour used (note that these customizations are currently missing from the simple web based demo described in Chapter 4, although they are available when viewing term clouds within the main GATE interface.).

An example term cloud is shown in Figure 2.1. This example was generated from a random sample of 450 tweets from the BBC, the Guardian, and CNN. As you can see

[1] For a comprehensive overview see `http://en.wikipedia.org/wiki/Tag_cloud`

[2] `https://gate.ac.uk/projects/arcomem/TermRaider.html`

[3] `http://www.arcomem.eu/`

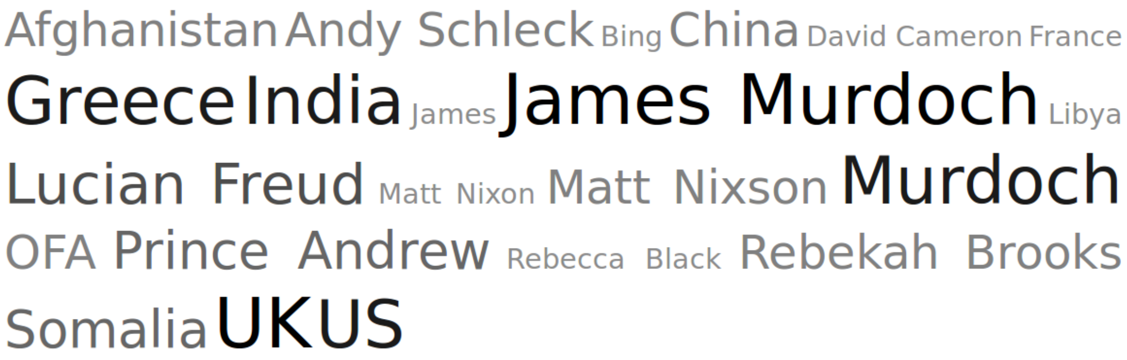

Figure 2.1: Example Term Cloud

there are a number of things that could be improved such as including normalising names (Nixon vs Nixson) and incorporating co-reference (Murdoch vs James Murdoch).

2.1 Multi-Lingual Support

In their simplest form, term clouds are essentially word based and as such require little (if any) adaptation for use across different languages. Clouds generated from more linguistically motivated terms (e.g. named entities) would require some level of adaptation for a given language. TermRaider has been developed with this in mind and already supports German. Support for other TrendMiner languages (Bulgarian, Italian and Hindi) will rest upon the ontology backed IE being developed in WP2 – term clouds can be generated from the extracted terms.

Chapter 3

Micropinion Based Summarization

While term clouds provide a quick visual way of summarising a text collection they can, in some situations, be misleading. For example, by weighting terms based upon frequency of occurrence, strongly held but infrequently expressed opinions can easily be overlooked. The second baseline system contained within this deliverable takes a very different approach to summarization and produces micropinions.

Micropinions are essentially concise phrases that represent the opinions expressed within a text collection. Micropionions are generally three or four words in length and are generated rather than extracted from text. For example, a collection of smartphone reviews may generate micropinions such as 'nice screen' or 'short battery life'.

Our current implementation follows the original approach [GZV12], and uses the Microsoft Web N-Gram service[1] to calculate readability, and point-wise mutual information (PMI) to determine representativeness. Current work is focused upon tuning the algorithm for use with tweets and to incorporate ontology backed IE into the phrase generation procedure.

3.1 Multi-Lingual Support

Whilst the underlying approach to micropinion generation is language independent it requires a large tri-gram language model in order to calculate readability. As noted above, the current prototype uses Microsoft's Web N-Gram corpus. While convenient and easy to use this resource was created using only English language documents (as was the equivalent Google derived corpus). Adaptation to support none English languages will therefore require us to locate an appropriate tri-gram corpora, although it is hopped that the incorporation of ontology backed IE into the algorithm will negate this issue somewhat (i.e. some multi-word entities will already be known to be readable). A possible alternative

[1] http://web-ngram.research.microsoft.com/info/

to a large tri-gram language model would be to utilise the text being summarised. Phrase readability could be estimated as the proportion of phrase bi-grams which also appear in the text being summarised. If successful this should also have the advantage of using readability to boost representativeness.

Chapter 4

Software Availability

Web based demos of both systems are available from
`http://demos.gate.ac.uk/trendminer/summarization/`

As previously mentioned, TermRaider is being developed within the ARCOMEM project and is not yet publicly available – an initial release is planned for sometime during November. This will include a public release of the term cloud generation software described within this deliverable.

Appendix A

State of the Art Review

The state-of-the-art review of stream media summarization is currently under review for publication and as such has currently been withheld from inclusion in this deliverable. Please contact the authors for access to a pre-print version.

Bibliography

[BS09] C. Buckley and G. Salton. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 2009.

[CMB+11] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M.A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. *Text Processing with GATE (Version 6)*. 2011.

[DMT+12] Stefan Dietze, Diana Maynard, Nina Tahmasebi, Yannis Stavrakas, Vassilis Plachouras, Elena Demidova, Jonathon Hare, David Dupplaw, Adam Funk, Wim Peters, and Patrick Siehndel. Extraction and Enrichment. Deliverable D3.2, ARCOMEM, 2012.

[GZV12] K. Ganesan, C. Zhai, and E. Viegas. Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st World Wide Web Conference*, pages 869–878, 2012.