FP7-ICT Strategic Targeted Research Project PHEME (No. 611233)
Computing Veracity across Media, Languages, and Social Networks

D3.1 Cross-Media and Cross-Language Linking Algorithm

Piroska Lendvai (Universitaet des Saarlandes)

Thierry Declerck (Universitaet des Saarlandes)

Abstract

FP7-ICT Strategic Targeted Research Project PHEME (No. 611233)
Deliverable D3.1 (WP 3)

The context of creating a Cross-Media and Cross-Lingual (CM-CL) linking algorithm in the PHEME project is the need for a procedure that connects User-Generated Content (UGC) to topically relevant information in complementary media. For the purpose of the current deliverable, media that is complementary to the original media of UGC (i.e., a tweet) is defined as texts in news articles from the entire web. The goal of the CM algorithm is tweet-to-document linking: finding and linking a web document that contains information that overlaps with information in a tweet.
The CL feature of the implemented algorithm exposes that several datasets in the PHEME project are inherently multilingual; e.g. in the Journalism use case (WP8) the tweets collected contain tweets in predominantly English and German, but also French and Dutch. Another property of the data that accounts for cross-linguality is that the language of documents that are linked from tweets may be in a language that is not the referring tweet's language. Finally, cross-linguality can also occur when the algorithm is extended with a document-to-document linking procedure.
In this deliverable we describe the current implementation of the CM-CL algorithm.

Keyword list: cross-media linking, cross-lingual linking, user-generated content to authoritative content, authoritative content

Nature: Prototype
Dissemination: PU
Contractual date of delivery: 30.06.2015
Actual date of delivery: 13.07.2015
Reviewed By: Kalina Bontcheva (USFD), Laura Tolosi-Halacheva (ONTO)
Web links: http://www.dfki.de/lt/publication_show.php?id=8221

PHEME Consortium

This document is part of the PHEME research project (No. 611233), partially funded by the FP7-ICT Programme.

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP, UK
Tel: +44 114 222 1930
Fax: +44 114 222 1810
Contact person: Kalina Bontcheva
E-mail: K.Bontcheva@dcs.shef.ac.uk

MODUL University Vienna GMBH
Am Kahlenberg 1
1190 Wien
Austria
Contact person: Arno Scharl
E-mail: scharl@modul.ac.at

ATOS Spain SA
Calle de Albarracin 25
28037 Madrid
Spain
Contact person: Tomás Pariente Lobo
E-mail: tomas.parientelobo@atos.net

iHub Ltd.
NGONG, Road Bishop Magua Building
4th floor
00200 Nairobi
Kenya
Contact person: Rob Baker
E-mail: robbaker@ushahidi.com

The University of Warwick
Kirby Corner Road
University House
CV4 8UW Coventry
United Kingdom
Contact person: Rob Procter
E-mail: Rob.Procter@warwick.ac.uk

Universitaet des Saarlandes
Computational Linguistics Lab
Campus
D-66041 Saarbrücken
Germany
Contact person: Thierry Declerck
E-mail: declerck@dfki.de

Ontotext AD
Polygraphia Office Center fl.4,
47A Tsarigradsko Shosse,
Sofia 1504, Bulgaria
Contact person: Georgi Georgiev
E-mail: georgiev@ontotext.com

King's College London
Strand
WC2R 2LS London
United Kingdom
Contact person: Robert Stewart
E-mail: robert.stewart@kcl.ac.uk

SwissInfo.ch
Giacomettistrasse 3
3000 Bern
Switzerland
Contact person: Peter Schibli
E-mail: Peter.Schibli@swissinfo.ch

Executive Summary

The context of creating a CM-CL linking algorithm in the PHEME project is the need for a procedure that links User-Generated Content (UGC) to topically relevant information in *complementary media*. For the purpose of the current deliverable, media that is complementary to the original media of UGC (i.e., a tweet) is defined to be news article texts from the entire web. The CM-CL linking algorithm differs from general CM-Information Retrieval (CM-IR), where the goal is the ranked retrieval of both authoritative and user-generated content, based on some query. The CM-CL linking algorithm's direction always originates in the UGC, as it uses elements present in a tweet with minimal abstraction only, whereas retrieval is always directed at a web document text. The algorithm implements this directionality; it has two instantiations that are described in this deliverable. The first is based on URLs that are present in tweets, and that point to web documents. The second is based on hashtags that are present in tweets.

The goal of the CM algorithm is *tweet-to-document linking*: *finding and linking* a web document that contains information overlapping with information in the referring tweet. *Topic relevance* that a CM-CL algorithm in PHEME targets is defined such that *the tweet text and the linked text share at least one piece of information.* Topic relevance can later be redefined and directed towards the goals of PHEME's Work Package 4. The narrower goal in WP4 is to obtain corroborative information from external documents for enabling more informed judgement on tweet content reliability. Information gained from the resources linked by the CM-CL algorithm will be put to use in WP4 to corroborate information already present in a tweet.

The *cross-lingual component* of the implemented algorithm has several motivations. One source for cross-linguality is that several datasets in the PHEME project are inherently multilingual, e.g. in the Journalism use case (WP8) tweets are predominantly English (88%) and German (9%). Other languages like French, Spanish, Dutch, Italian, Russian, are less than 2%. Twitter does provide a language identifier via their API, so automatic language identification is not needed, but entity co-reference (see e.g. [Rao et al., 2010]) across languages still needs resolution. The algorithm therefore needs to target the linking of concepts that pertain to one and the same entity across languages, e.g. the *hashtags* '#Schweiz' (DE) and '#Suisse' (FR) that are both frequent tokens in the WP8 collected data. *Cross-lingual entity reference* does not have to be limited to hashtags, but the prototype CM-CL algorithm reported here yields better performance if it processes *tokens from the hashtag level* only. The reason is that hashtags typically denote the most important topics and entities in a tweet, the set of hashtags is user-coded and it comprises a relatively small (and grammatically more homogeneous) set of tokens.

Another cross-lingual property of our data is that the language of documents that are linked from tweets may be in a language that is not the tweet's language: a tweet in German can make a reference to an English article, for example because at the time of posting the tweet *authoritative content* was only available in English. The CM-CL algorithm makes use of the fact that URL references may inherently be cross-lingual in the data.

Other sources of cross-linguality may arise from *document-to-document linking*. After tweet-to-document linking has taken place, a service can be invoked that returns a set of new documents that are similar to the originally linked document. Such an extension to the algorithm simulates a service that is going to be available in a later phase of the PHEME project. In the current deliverable, we exemplify the extension by calling an external web service, Event Registry[1], which performs large-scale, multilingual indexing and retrieval of news articles on the web.

Deliverable D3.1 "Cross-Media and Cross-Language Linking Algorithm" describes the current implementation of the CM-CL algorithm. Linking that is based on URLs and hashtags makes use of the wisdom of the crowd that is available in social media content in a semi-structured format. Our core assumption is that URL presence in tweets is a relevance signal analogous to landing page information in click data, utilizable in developing retrieval functions from observed user behaviour (see e.g. [Joachims, 2002]). The incorporation of such 'silver-standard' knowledge (cf. [Wissler et al, 2014]) is an extremely valuable asset in big data analytics.

We implemented a procedure in the domain of daily news that extracts and ranks key phrases via tweet-to-document linking, based on token similarity, Twitter metadata, and manually assigned event categories in tweets. For tweets that do not refer to any external documents, the algorithm links and evaluates how relevant document candidates are. This is achieved by story-based linking of documents to tweets, key phrase extraction from the tweets, and the assignment of phrase-document similarity weights for relevance ranking. This focus is motivated by the content discovery/recommendation scenario: a user who does not refer to external sources may be unaware of the cross-media context of their own content. Referring to external sources is a multi-purpose activity in social media practices, e.g. for content framing and verification, as well as content enrichment (i.e., guiding to extended information).

First, longest common subsequences (LCS) are identified between tweets and web documents referred to in tweets. Document-based LCS similarity metrics are applied to extract candidate key phrases, which get aggregated on the event level. The metrics are then computed for the same document base, but paired with tweets that did not link external references. The workflow yields complementary document rankings and key phrases from the two setups that collectively describe a shared event.

[1] http://eventregistry.org/

Contents

1 Introduction

The context of creating a CM-CL linking algorithm in the PHEME project is the need for a procedure that links User-Generated Content (UGC) to topically relevant information in *complementary media*. For the purpose of the current deliverable, media that is complementary to the original media of UGC (i.e., a tweet) is defined to be news article texts from the entire web. The CM-CL linking algorithm differs from general CM-Information Retrieval (CM-IR), where the goal is the ranked retrieval of both authoritative and user-generated content, based on some query. The CM-CL linking algorithm's direction always originates in the UGC, as it uses elements present in a tweet with minimal abstraction only, whereas retrieval is always directed at a web document text. The algorithm implements this directionality; it has two instantiations that are described in this deliverable. The first is based on URLs that are present in tweets, and that point to web documents. The second is based on hashtags that are present in tweets.

The goal of the CM algorithm is *tweet-to-document linking*: *finding and linking* a web document that contains information overlapping with information in the referring tweet. *Topic relevance* that targets a CM-CL algorithm in PHEME is defined such that *the tweet text and the linked text share at least one piece of information.* Topic relevance can later be redefined and directed towards the goals of PHEME's Work Package 4. The narrower goal in WP4 is to obtain corroborative information from external documents for enabling more informed judgement on tweet content reliability. Information gained from the resources linked by the CM-CL algorithm will be put to use in WP4 to corroborate information already present in a tweet.

The *cross-lingual component* of the implemented algorithm has several motivations. One source for cross-linguality is that several datasets in the PHEME project are inherently multilingual, e.g. in the Journalism use case (WP8) tweets are predominantly English (88%) and German (9%). Other languages like French, Spanish, Dutch, Italian, Russian, are less than 2%. Twitter does provide a language identifier via their API, so automatic language identification is not needed, but entity co-reference (see e.g. [Rao et al., 2010]) across languages still needs resolution. The algorithm therefore needs to target the linking of concepts that pertain to one and the same entity across languages, e.g. the *hashtags* '#Schweiz' (DE) and '#Suisse' (FR) that are both frequent tokens in the WP8 collected data. *Cross-lingual entity reference* does not have to be limited to hashtags, but the prototype CM-CL algorithm reported here yields better performance if it processes *tokens from the hashtag level* only. The reason is that hashtags typically denote the most important topics and entities in a tweet, the set of hashtags is user-coded and it comprises a relatively small (and grammatically more homogeneous) set of tokens.

Another cross-lingual property of our data is that the language of documents that are linked from tweets may be in a language that is not the tweet's language: a tweet in German can make a reference to an English article, for example because at the time of posting the tweet *authoritative content* was only available in English. The CM-CL algorithm makes use of the fact that URL references may inherently be cross-lingual in the data.

Other sources of cross-linguality may arise from *document-to-document linking*. After tweet-to-document linking has taken place, a service can be invoked that returns a set of new documents that are similar to the originally linked document. Such an extension to the algorithm simulates a service that is going to be available in a later phase of the PHEME project. In the current deliverable, we exemplify the extension by calling an external web service, Event Registry[2], which performs large-scale, multilingual indexing and retrieval of news articles on the web.

Deliverable D3.1 "Cross-Media and Cross-Language Linking Algorithm" describes the current implementation of the CM-CL algorithm. Linking that is based on URLs and hashtags makes use of the wisdom of the crowd that is available in social media content in a semi-structured format. Our core assumption is that URL presence in tweets is a relevance signal analogous to landing page information in click data, utilizable in developing retrieval functions from observed user behaviour (see e.g. [Joachims, 2002]). The incorporation of such 'silver-standard' knowledge (cf. [Wissler et al, 2014]) is an extremely valuable asset in big data analytics.

We implemented a procedure in the domain of daily news that extracts and ranks key phrases via tweet-to-document linking, based on token similarity, Twitter metadata, and manually assigned event categories in tweets. For tweets that do not refer to any external documents, the algorithm links and evaluates how relevant document candidates are. This is achieved by story-based linking of documents to tweets, key phrase extraction from the tweets, and the assignment of phrase-document similarity weights for relevance ranking. This focus is motivated by the content discovery and recommendation scenario: a user who does not refer to external sources may be unaware of the cross-media context of their own content. Referring to external sources is a multi-purpose activity in social media practices, e.g. for content framing and verification, as well as content enrichment (i.e., guiding to extended information).

First, longest common subsequences (LCS) are identified between tweets and web documents referred to in tweets. Document-based LCS similarity metrics are applied to extract candidate key phrases, which get aggregated on the event level. The metrics are then computed for the same document base, but paired with tweets that did not link external references. The workflow yields complementary document rankings and key phrases from the two setups that collectively describe a shared event.

2. Related work on cross-media linking

Some recent natural language processing studies present Cross-Media (CM) approaches with the purpose of aligning UGC and authoritative content. The goal of [Tanev et al, 2012] is to collect information about emergency situations from tweets that are complementary to mainstream media reports. The events that comprise the emergency situations are obtained from news releases. First, relevant keywords are determined from a centroid news article in a topically related article cluster; these are used in various query constructions to retrieve event-related tweets. The direction of linking proceeds from a centroid, authoritative article toward related tweets (UGC). In manual evaluation, 75% precision is reported on relatedness, while complementarity is judged based on several document-structure-level aspects, which provide

[2] http://eventregistry.org/

information about the location of the new information. The direction of the algorithm is motivated by the need to boost retrieval precision on established, which is *orthogonal to the mission of PHEME*, whose starting point are events that are discoverable from social media content, and might only later or not at all appear in mainstream news releases.

The [Tanev et al, 2012] algorithm is reused in [Balahur and Tanev, 2013], where the workflow is extended with further steps: based on a centroid article in an event cluster, related tweets are mined that contain URLs, using custom-threshold-based term-vector similarity. Then, relevance ranking takes place on these tweets, using platform-specific indicators (number of mentions, retweets, etc.). New, related articles on the web are retrieved based on the URLs of top-ranked tweets. Topical relevance in the last step, i.e. in tweet-to-document linking, is judged positive if the document „reports about the same news story or talks about facts, like effects, post developments and motivations, *directly related* to this news story". From our understanding, [Balahur and Tanev, 2013] do not report on the proportion of web articles found via the URL-based linking that were part of the query-originating news cluster. Such a metric would evaluate performance more transparently on *discovering additional information sources*, which is an important dimension of the PHEME CM-CL algorithm.

To improve retrieval in full-text search systems, query modelling within the language modelling framework has been investigated in the field of Information Retrieval. In these studies, documents were represented as generative probabilistic models; cf. Section 3 in [Meij et al., 2010]. As the basis for ranking, the difference can be computed between the language model of a document and that of a query, see e.g. [Lafferty and Zhai, 2003]. We see some parallel ideas between this framework and the CM-CL algorithm that scores the similarity between a tweet and a document, whereby the tweet is utilized as if it was a query. Its similarity score furthermore implicitly encodes important features of document content: e.g. term frequency, as well as valuable linguistic characteristics such as token proximity and syntax.

Very recently, creating systems for Semantic Textual Similarity judgements on Twitter data has been a Shared Task[3] in the Natural Language Processing community [Xu et al, 2015]. Given two sentences, the participating systems needed to determine a numerical score between 0 (no relation) and 1 (semantic equivalence) to indicate semantic similarity on the Twitter Paraphrase Corpus that was first presented in [Xu et al, 2014]. The sentences were linguistically pre-processed by tokenization, part-of-speech and named entity tagging. The system outputs are compared by Pearson correlation with human scores: the best systems reach above 0.80 Pearson correlation scores on well-formed texts. The organizers stress two main general findings. With respect to the technologies used, they note that "while the best performed systems are supervised, the best unsupervised system still outperforms some supervised systems and the state-of-the-art unsupervised baseline." With respect to the evaluation metrics used, an important outcome is that "the performance of the same system on the two tasks ("F1 vs. Pearson") are not necessarily related", which may indicate that statistical evaluations are easily biased by task setup and dataset design.

[3] http://alt.qcri.org/semeval2015/task1/

3. Algorithm structure

The CM-CL algorithm (henceforth: algorithm) follows the general retrieval scenario that is also present in the two studies treated above. In this section, we describe the main components of the algorithm.

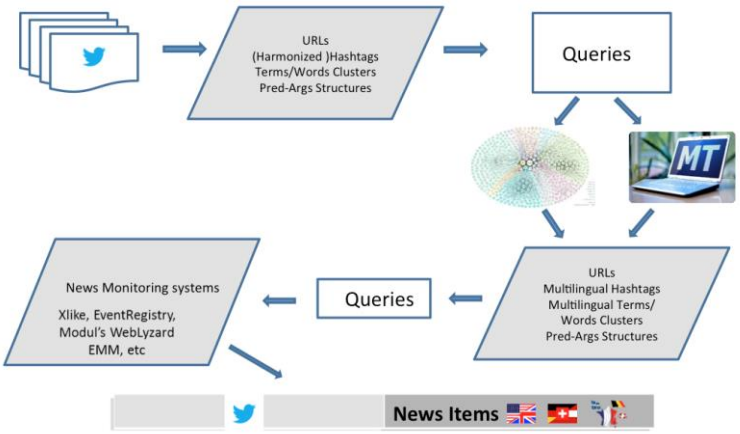

Figure 1. General workflow of the algorithm: extracted features from the UGC (i.e., tweets) are used as queries that can be first cross-lingually enriched by information retrieved from Linked Open Data (LOD), or directly by machine translation (MT) tools. The results are pairs of linked Twitter text and news items in multiple languages.

Event extraction

The datasets we use consist of so-called 'source tweets' relating to the broader events of (O) the Ottawa shooting[4] and (G) the Gurlitt art collection[5] that were annotated as rumours, in the same fashion as described in [Zubiaga et al., 2015]. The two sets are chosen out of the many available ones in Work Package 8 as they differ markedly in size, the amount of cross-linguality ("Gurlitt" predominantly in German and French), and the amount of labelled events. Figure 2 displays URL and hashtag statistics in (O) and (G).

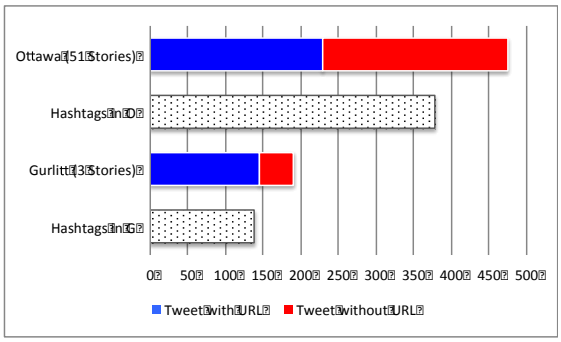

Figure 2: URL and hashtag statistics in the datasets "Ottawa" and "Gurlitt"

4 https://en.wikipedia.org/wiki/2014_shootings_at_Parliament_Hill,_Ottawa

5 https://en.wikipedia.org/wiki/2012_Munich_artworks_discovery

The two cross-media studies quoted in Section 2 delegate the first sentences of a centroid news article from a cluster of news articles to define the Event that they match to tweets. In PHEME, often a very typical tweet is promoted from a group of topically similar tweets to be the label for that group; the resulting tag can also be referred to as 'Story', 'category', 'annotation'; we reserve the usage of 'Event' to mark broader affairs, such as those that encompass all the Stories within a dataset. Stories are short, summarizing spans of texts (e.g. "Obama to speak with Harper today"). The PHEME datasets are manually categorized and labelled. This procedure will be automated later in the project.

Information retrieval

Our core assumption is that URL presence in tweets is a relevance signal analogous to landing page information in click data, utilizable in developing retrieval functions from observed user behaviour (see e.g. [Joachims, 2002]). This focus is motivated by the content discovery/recommendation scenario: a user who does not refer to external sources may be unaware of the cross-media context of their own content. Referring to external sources is a multi-purpose activity in social media practices, e.g. for content framing and verification, as well as content enrichment (i.e., guiding to extended information).

Similar to [Balahur and Tanev, 2013], the PHEME CM-CL algorithm performs tweet-to-document linking but its task is more difficult than merely retrieving topically related content. Topical relevance holds by definition in both instantiations of the algorithm (URL-based, hashtag-based): when a user posts a URL or designates a hashtag in their tweet, they would refer to (or mark-up) topically relevant content. Extreme cases such as trolling are part of Social Media data, but need not be addressed by our current implementation, as we are working on datasets that were filtered based on platform-specific metadata (e.g., a retweet threshold).

The CM-CL algorithm retrieves relevant content on the sub-event (Story) level, which is narrower than the topic level. Moreover, it needs to retrieve information that is shared with the information in the tweet. Topic relevance is defined such that the tweet text and the linked text share at least one piece of information. Our approach is to regard the content of the tweet as a free text query, and the externally linked page as the target document. The algorithm is powered by story-based linking of documents to tweets, key phrase extraction from the tweets, and the assignment of phrase-document similarity weights for relevance ranking. Performance is *evaluated qualitatively* in terms of (i) the nature and utility of key phrases extracted by the algorithm and (ii) the relevance of candidate web documents that the algorithm can supply for tweets that do not link to any external documents.

We explain two instantiations of the CM-CL algorithm: a URL-based and a hashtag-based one. The two instantiations are uniformly composed of the procedures of query construction, data retrieval, ranking, and the evaluation of linked content. Each instantiation focuses on user-provided meta-information, i.e. on URLs and on hashtags. URLs and hashtags are provided in a structured way by the Twitter API, which we make use of, but are also accurately locatable by regular expressions in the tweet body.

URL-based algorithm

In the pre-processing phase, all the tweets from the Twitter API are parsed for an 'expanded_url' item. Tweets that have no URL but are labelled with the same Story as the tweet with an URL are logged. In the content fetching phase, each URL is accessed using an HTTP network protocol API module, e.g. urllib2[6] for python. The fetched textual data can then be HTML-parsed with standard tools, e.g. the BeautifulSoup[7] library.

Not all URLs are possible to fetch technically, whereas it was also our intention to not to try to fetch every URL: if the number of tweets that refer to one and the same URL exceeds a threshold (currently set to 3), the URL is discarded. This filter is applied in order to separate dynamically grown live blog texts from static online news articles texts; the two genres serve different purposes which result in different surface text patterns, as well as topic patterns, in the data, and probably best be treated separately. Our manual observation is that there is relatively more textual and topical homogeneity in news articles, illustrated by an example from the (O) set:

Expanded url: http://cnn.it/ZGz1gu
Fetched headlines: "PM: Ottawa 'terrorist' killed soldier 'in cold blood' - CNN.com"
Tweets linking to this URL (In total: 2):
(1) 'Canadian media: Gunman shot soldier at war memorial. http://t.co/zNhxK6wBoy';
Annotated with *Story:* *A soldier has been shot at National War Memorial*
(2) 'Ottawa Police Service: There were "numerous gunmen" at the Canada War Memorial shooting. One person was shot. http://t.co/zNhxK6wBoy'
Annotated with *Story*: *There are multiple shooting suspects still at large*

As opposed to it, there is more textual and topical heterogeneity in a live reporting page, likewise from the (O) set:

Expanded url: http://bit.ly/ZNPRdO
Fetched headlines: Canada Shootings
Tweets linking to this URL (In total: 11):
(1) '@OttawaPolice: "Incidents occurred at National War Memorial, near the Rideau Centre and Parliament Hill." Live blog: http://t.co/q98AMohu7T'
Annotated with *Story:* *There was a shooting incident near/at the Rideau Centre*
(2) 'Witness tells CNN gunman shot one of two soldiers standing guard at war memorial in Ottawa. Live blog: http://t.co/q98AMohu7T'
Annotated with *Story:* *A soldier has been shot at National War Memorial*
(3) 'In response to Ottawa incidents, NORAD increased number of planes on higher alert status ready to respond if needed. http://t.co/q98AMohu7T'
Annotated with *Story:* *NORAD on high-alert posture*
(4) 'Senior U.S. official: Canadian government has informed U.S. that one shooter is dead in Ottawa. Live blog: http://t.co/q98AMohu7T'
Annotated with *Story:* *Suspected shooter has been killed/is dead*
(5) 'U.S. officials: Suspected shooter in Ottawa rampage identified as Canadian-born Michael Zehaf-Bibeau. Live blog: http://t.co/q98AMohu7T'
Annotated with *Story:* *The suspect's name is Michael Zehaf-Bibeau* (etc.)

---

6 https://docs.python.org/2/howto/urllib2.html

7 http://www.crummy.com/software/BeautifulSoup/bs4/doc/

Live news blogs keep adding new text to the page as a report develops, while they are keeping the URL constant. As the focus of the blog's attention is shifting every time, the blog is updated with an unfolding sub-event, and tweet posts, still pointing to one and the same URL, would focus on different topics within the larger event. This introduces much more topical and textual heterogeneity in the tweets that all refer to a single page than in cases when static news articles are referenced from different tweet posts. Put differently, the linking mechanism in the Twitter platform is the same in both cases, but the underlying tweeter intent is different; in order to avoid tackling separate issues uniformly, we opted for applying the filter.

Similarity scoring

Similarity in the current implementation is based on the Longest Common Subsequence (LCS) metric. In our implementation, LCS is computed between an entire tweet and one sentence from a linked document. These two texts will be referred to as 'text pairs' or 'snippet pairs'. LCS is a language-independent, flexible-length, in-sequence n-gram matching method that we apply on the token level. LCS returns similarity based on the longest in-sequence common n-gram for each text pair, without the need for predefined n-gram length (cf. [Lin, 2004]) or full overlap of the shared string of tokens. LCS is used as an evaluative metric e.g. for scoring text pairs in automatic summarization and machine translation tasks.

Sentences in fetched documents were created using the NLTK tokenizer[8] and punctuation matching, while the LCS implementation was based on code from Wikibooks[9]. Snippet content was normalized: screen names and URLs that present no additional knowledge at this processing stage were removed to reduce sensitivity for string length in LCS. Spelling in texts was normalized by lowercasing and punctuation removal. We have experimented with stop word filtering as well but have not included it in the algorithm as tweets typically are created with carefully chosen content words to address the length limit.

Retrieval and content statistics

Table 1 displays retrieval statistics on documents linked from tweets. We still have to analyse in details the resulting differences for the two datasets.

Table 1: Retrieval and content statistics

	Ottawa	Gurlitt
#fetched documents	156	101
#unique top-level URL domains	58	60
%documents fetched per Story	5.2	33.6
%length fetched document body[10]	88.4	65.4

8 http://www.nltk.org/_modules/nltk/tokenize.html

9 https://en.wikibooks.org/wiki/Algorithm_Implementation/Strings/Longest_common_subsequence

10 Document length depends on the efficiency of HTML-parsing; the current implementation is suboptimal.

Similarity-based term extraction and retrieval

The algorithm extracts terms from each URL's content, based on similarity between the document fetched from this URL and
(Step 1) the tweet's content where this URL was referred from, as well as
(Step 2) other tweets' content, if they contained no URL but are labelled with the same Story.

In each of these two scenarios, similarity scores are computed for a large amount of cross-media textual pairs (i.e., a tweet and a document sentence). The mean document-based similarity score is used to rank the phrases extracted from tweets with respect to URLs (examples are supplied below).

In *Step 1*, tweet-document sentence LCS ratios are created for each document linked from a *URL-originating tweet (UOT)*. In effect, a weighted document frequency metric is computed that employs a per-tweet-metric. For each linked URL, it takes the LCS mean similarity score of the top-n most similar sentences between a document and its referring tweet. *n*=4 is used as a threshold, which is currently manually maximized by observing the output quality. The process produces a ranked list of n-best key phrases with respect to the document behind the URL. The larger the size of *n*, the more the score is smoothed over the entire document.
In *Step 2*, the same procedure is applied to these same document set, now paired with tweets that did *not* link external documents, but were labelled with the same Story as the tweet from which a candidate URL was referred from. A second list of ranked key phrases is extracted from these tweets -- with relation to each document and each Story. Creating the ranking for *non-URL tweets (NUT)* is computationally more expensive than for the UOT ranking (Step 1), because the amount of NUT tweets in a Story is always higher than the amount of tweets containing a URL that could be fetched. The value of the NUTs varies in the range of 1-29 in the Ottawa dataset, and between 0-41 in the Gurlitt dataset, while for UOTs it is almost always 1.

The example in Figure 3 illustrates that the length and amount of extracted phrases varies, and that the outputs from Step1 and Step2 are complementary.

Event: The FBI is assisting Canadians with the Ottawa shooting
Linked doc headlines: "Shots fired at Canadian War Memorial, Parliament; soldier killed, police scouring downtown Ottawa - The Washington Post"

Tweet with article URL: "FBI working with Canadian authorities to determine if Ottawa shooting was an act of terrorism http://t.co/oDy7eSbgFn"
phrase @1: 'fbi working with canadian authorities to determine if was an act of terrorism'
Extracted from: "The FBI was working with Canadian authorities Wednesday to determine if this was an act of terrorism" ; **Score**: 0.79
phrase @2: 'was an act of terrorism'
Extracted from: "that authorities said was an act of terrorism, with the Quebec police saying that it was a deliberate attack on a uniformed soldier"; **Score**: 0.31
[... phrase @3, phrase @4]
–> top-*4* LCS mean: 0.38

Tweet without article URL: "FBI assisting in the case of the Ottawa shooting, sources have confirmed to CTV News"
phrase @1: 'fbi assisting'
Extracted from: "the fbi is actively assisting canadian authorities" ; **Score**: 0.22
[... no more pairs]
–> top-*1* LCS mean: 0.22

=> **Top-ranked phrases for Event**:
'fbi working with canadian authorities to determine if was an act of terrorism', 'was an act of terrorism', 'shooting was', 'fbi assisting'

Figure 3: Similarity-based phrase extraction and ranking in (O)

We have also computed the Weighted Story Metric (WSM) for each dataset and both steps (see Table 2). WSM computes the mean LCS from all ranked document lists for all Stories. It is an implicitly weighted metric that accounts for Story frequency: the more tweets in a Story, the more that Story is represented in the WSM.

Table 2: WSM scores for UOT and NUT expressed by for the data sets Ottawa and Gurlitt

	Weighted Story Metric (UOT)	Weighted Story Metric (NUT)
Ottawa	0.19	0.26
Gurlitt	0.29	0.50

It may be surprising to see that the similarity scores for the Gurlitt set are higher, as Gurlitt contains heavily cross-lingual data. One explanation can be that the Gurlitt set is skewed in terms of Story structure: only three Stories are labelled in this set, and one of them relates to the vast majority of the tweets. In addition, the size of the Gurlitt set is small, which can give rise to high standard deviation from the reported means.

Evaluation

Due to the task context, retrieval metrics such as precision and recall are not directly applicable to the output. Evaluation is made by comparison of the UOT and NUT ranked lists. Since the manually assigned Stories designate gold-standard labels with respect to tweet content similarity, UOT- and NUT-based rankings are directly comparable as if in A/B testing, where A (i.e., UOT ranking) is known to be a gold-standard reference ranking.

To create statistically solid evaluation scores, larger datasets need to be collected and processed by the algorithm, which is foreseen in a later phase of the PHEME project. Automatic evaluation will take place with standard IR evaluation metrics that are applicable to the CM-CL ranking. In particular, we will use the normalized discounted cumulative gain (NDCG) metric as it is "designed for situations of non-binary notions of relevance and is evaluated over some number *k* of top search results" (cf. Chapter 8, [Manning et al., 2008]). The newly gained structuring of the data will enable the analysis and learning of finer ranking patterns that have not been reported in previous cross-media content linking studies, or that have been unavailable for sub-event (i.e., Story) level similarity relations.

Below we illustrate the output of such an analysis on examples from both datasets. A pilot manual evaluation of the currently available datasets supplies an intriguing outcome: it is typically *not* the case that the top-ranked NUT tweet for a URL is an LCS-wise very similar tweet to the top-ranked UOT tweet.

Story: Shots fired on Parliament Hill

Headlines of fetched URL: *Shots fired at Canadian War Memorial, Parliament; soldier killed, police scouring downtown Ottawa - The Washington Post*
UOT @rank1:

0.33 ["Chaos broke out in Ottawa after a shooting at the war memorial and reports of gunfire in Parliament http://t.co/pVOh34l2ea"]
NUT highest rank with the same URL @rank3:
0.37 ["Witnesses say several dozen shots fired inside Parliament buildings after Canadian soldier shot at nearby War Memorial. #Ottawa #cdnpoli"]

Headlines of fetched URL: *Canadian Convert Suspected In Parliament Attack*
NUT @rank1:
0.45 ["More shots fired on Parliament Hill."]
UOT highest rank with the same URL @rank16:
0.13 ["Ottawa shootings reportedly at three locations - parliament, war memorial and shopping mall. http://t.co/fULkb6VpEv http://t.co/mL9Ivehy6Q"]

Story: The Leafs-Senators game in Ottawa has been postponed

Headlines of fetched URL: "*Leafs-Senators game postponed after shootings* - Sportsnet.ca"
UOT @rank1:
0.26 ["NHL postpones Wednesday's Leafs-Senators game due to tragedy in Ottawa http://t.co/Ohec0ceae7 http://t.co/sLeiCmoUN6"]
NUT @rank2:
0.42 ["NHL says date of rescheduled game TBD. NHL ``wishes to express its sympathy and prayers to all affected by the tragic events in Ottawa""]

Headlines of fetched URL: "*NHL postpones Leafs-Senators game after Ottawa shooting - NHL on CBC Sports - Hockey news, opinion, scores, stats, standings"*
NUT @rank1:
0.55 ["NHL says date of rescheduled game TBD. NHL ``wishes to express its sympathy and prayers to all affected by the tragic events in Ottawa""]
UOT @rank3:
0.16 ["NHL postpones tonight's Leafs-Senators game because of #OttawaShooting http://t.co/a6JxXm20nZ http://t.co/G80SbMBTlv"]

Results from the Gurlitt set are presented below.

Story: *The Bern Museum will accept the Gurlitt collection*

Headlines of fetched URL: "Bestätigt: Kunstmuseum Bern nimmt das Erbe des Kunstsammlers Cornelius Gurlitt an - KURIER.at"
UOT @rank1:
0.75 ["Bestätigt: Sammlung Gurlitt geht nach Bern http://t.co/FRCSHTU5hL"]
NUT @rank7:
0.82 ["RT @SWRinfo: Das Kunstmuseum Bern nimmt das Erbe des Kunstsammlers Cornelius #gurlitt an."]

Headlines of fetched URL: *"Die Entscheidung um Gurlitt-Erbe; Das Protokoll - News - Schweizer Radio und Fernsehen"*
NUT @rank1:
0.94 ["RT @SWRinfo: Das Kunstmuseum Bern nimmt das Erbe des Kunstsammlers Cornelius #gurlitt an."]
UOT @rank6:
0.58 ["Das @KunstmuseumBern nimmt das Erbe des Kunstsammlers #Gurlitt an. http://t.co/TFqkc1LcdV"]

Story: *Looted artworks will (initially) remain in Germany*

Headlines of fetched URL: "*Gurlitts Erbe: NS-Raubkunst bleibt in Deutschland*"
UOT @rank1:
0.76 ["Gurlitts #Erbe: NS-Raubkunst bleibt in #Deutschland http://t.co/nLM6tG9uHH"]
NUT @rank3:

0.11 ["Schäublin agreement in accepting Gurlitt collection: Objects with suspicion of being Nazi-looted art will initially remain in Germany."]

Headlines of fetched URL: "*Cornelius Gurlitt: Kunstmuseum Bern, das Erbe und Monika Grütters - SPIEGEL ONLINE"*
NUT @rank1:
0.13 ["Schäublin agreement in accepting Gurlitt collection: Objects with suspicion of being Nazi-looted art will initially remain in Germany."]
UOT @rank3:
0.38 ["Kunsterbe: Gurlitt-Sammlung geht in die Schweiz, Raubkunst bleibt in Deutschland... http://t.co/i8qqHu0LlS"]

Ranking of linked content

To sum up the results of the algorithm on retrieval and ranking of cross-media and cross-lingual content, we stress that our main focus is (i) the comparison the event-based key phrase lists and (ii) document rankings obtained for tweets with and without URL. Figure 4 shows that we obtain complementary phrases and document rankings from the two steps, which together refer to a shared Story. We conclude that the algorithm is able to (i) increase lexical variety by expanding the original set of index terms, and (ii) to link tweets with unsupported content to authoritative web documents.

Event: "Shooter is still on the loose"

Doc-Phrase rankings from Tweet with URL:
@1 Doc: "Sky News — Grabyo", LCS: 0.28
phrase: 'watch shots are fired inside ottawa s parliament building canadian police say shooting suspect is still at large '
@2 Doc: "Attack on Ottawa: PM Harper cites terrorist motive - The Globe and Mail", LCS: 0.25
phrase: 'soldier shot at war memorial'
@3 Doc: "Attack on Ottawa: PM Harper cites terrorist motive - The Globe and Mail", LCS: 0.20
phrase: 'soldier shot'

Doc-Phrase rankings from Tweet without URL:
@1 Doc: "Sky News — Grabyo", LCS: 0.30
phrase: 'police say suspect still at large'
@2 Doc: "Attack on Ottawa: PM Harper cites terrorist motive - The Globe and Mai", LCS: 0.29
phrases: 'canadian soldier at war memorial', 'parliament hill', 'parliament lockdown'
@3 Doc: "Canada's parliament attacked, soldier fatally shot nearby — Reuters", LCS: 0.25
phrase: 'shots fired parliament', 'at large'

Figure 4: Complementary document rankings and phrases from the two scenarios, referring to a shared Story in (O)

Outlook concerning the URL-based instantiation

Similarity in the current implementation is based on the Longest Common Subsequence (LCS) metric. The LCS similarity metric, which we used both for retrieval and ranking, works in a language-independent way. This requires more thorough pre-processing (e.g. stemming) for e.g. German, but works directly for English. On the other hand, one shortcoming of LCS is that it is ignorant about meaning, one of its major impacts being that the LCS similarity is not sensitive to modality whereas it is to lexical variation. LCS similarity may be high for text pairs in which one snippet has a negation marker in it, whereas it may be low in case the content in one snippet is paraphrased in the other snippet.

LCS implicitly encodes important features of document content: e.g. term frequency, as well as valuable linguistic characteristics such as token proximity and syntax. This can be advantageous when working with big data across languages and domains, as

foreseen in the PHEME project. The proposed LCS-based approach extracts key phrases, not words, which we plan to use to support semantically improved phrase auto-completion at query time. Other standard textual similarity metrics[11] are also suitable for the evaluation procedure; their technical integration in the algorithm will be investigated. We also plan to incorporate more string and frequency normalization, similarity weighting on the basis of document structure, as well as linked-open data-based entity and concept detection in the phrases.

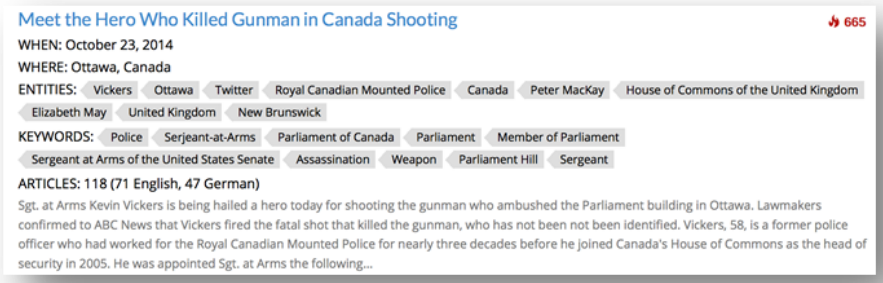

Figure 3: Example of a cluster of news articles fetched from Event Registry[12]. There are 118 articles available, and 47 are in German. The service also returns a list of detected entities and topic concepts, which can be used to further extend the cross-media and cross-lingual linking algorithm. The icon in the top right shows the amount of shares for this article on social media platforms[13].

Extension: Related news articles retrieval

Further statistics can be collected if a URL to a news article is submitted to a service that is able to return a set of similar news articles. Such an available service is Event Registry that allows programmatic querying via a python API[14]. The 'queryByUrl' method in the 'QueryArticle' class searches the Event Registry collection by an article URL: in case the article has been indexed in Event Registry, access to similar articles is possible based on various metrics. There are lots of potentials for the CM-CL algorithm in reusing the metadata -- exemplified by the screenshot illustration -- of a multilingual service like Event Registry, but these lie out of the scope of the current deliverable.

4. In progress

We are currently integrating the code written for the harmonization of hashtags (see the recent paper by (Declerck and Lendvai, 2015)), with the aim of proposing a topical clustering of hashtags that can then be sued for querying web documents, as this has been proposed for the URL-based approach described above. As hashtags are also used in other social media than Twitter, we expect also to establish a link across such social media. As the code for the harmonization of hashtags has been recently

11 e.g. https://code.google.com/p/dkpro-similarity-asl/, http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html#sklearn.feature_extraction.text.TfidfTransformer

12 http://eventregistry.org/

13 Counts are based on the http://www.sharedcount.com/ service.

14 https://github.com/gregorleban/event-registry-python

ported from Perl to Python, integration efforts with the code used for the algorithms described in this deliverable still have to be performed. We plan therefore an update of the code deliverable in Month 24.
We also plan to adapt the code used for accessing the Event Registry service to the services provided by the partner MODUL University, for which an API will be very soon be made available.
And finally, at the level of evaluation: we plan a new evaluation study to be applied to the finalized datasets of WP7 and WP8, which are due also at Month 18, the deadline for the actual version of D3.1, so that we didn't have the opportunity to evaluate our algorithms on the full range of the PHEME datasets yet.

5. Relevance to PHEME

The work described in this deliverable is relevant to the objectives of PHEME in various ways. Dealing with the topic of cross-media (cross-linking of UGC and other written document sources) is central is one wants to address the issues of *variety* of sources, which is one of the characteristics of Big Data.

Relevance to project objectives

Linking across media can help in detecting the type of "information" one source is spreading: factual statement, disinformation, or even disinformation? And this classification is at the core of the PHEME project.
Cross-linking supports the possibility to check incoming textual data against trusted sources.

Relation to other workpackages

As stated in the introduction of this deliverable, the work described here is a kind of generalisation of the work to be pursued in **WP4** "Detecting Rumours and Veracity". In D3.1 we describe how we can compare one source in UGC to other sources, and to establish if they are about the same topic. This is a preliminary step to the one that aims at establishing if contradictions between statements in different sources exist and in taking a decision on which sources to trust.
As for now D3.1 took as a basis for the application and the first evaluation of the algorithms data sets generated in **WP8** and this will be extended to the data sets generated for **WP7.** The use of WP8 datasets was easier in a first phase, since WP8 is also dealing with news media, and the focus of D3.1 was to link UGC with the content of news media in the Web. The next step will consist in applying and adapting the algorithm to the linking of UGC and patient records or scientific publications as those are building a core of the WP7 datasets.
Algorithms of D3.1 are and will increasingly be using the adapted multilingual pre-processing tools developed in the context of **WP2 (Task 2.3** Multilingual Pre-processing). Last but not least: the algorithm of task 3.1 will be included in **WP6** "Scalability, Integration, and Evaluation"

6. List of Abbreviations

API - Application Programming Interface
CM - Cross-Media
CL - Cross-Lingual
HTML - Hypertext Markup Language
HTTP - Hypertext Transfer Protocol
IR - Information Retrieval
LCS – Longest Common Subsequence
NDCG - Normalized Discounted Cumulative Gain
NLTK - Natural Language ToolKit
NUT - No-URL Tweet
PTM - Per-Tweet Metric
UOT - URL-Originating Tweet
URL - Uniform Resource Locator
WSM - Weighted Story Metric

7. Bibliography and references

Antenucci, D, Handy, G, Modi, A & Tinkerhess, M 2011, 'Classification of tweets via clustering of hashtags', EECS 545 Final Report.

Balahur, A & Tanev, H 2013, 'Detecting Event-Related Links and Sentiments from Social Media Texts', Proceedings of the 51st Annual
Meeting of the Association for Computational Linguistics (Conference System Demonstrations), Sofia, pp. 25-30.

Bansal, P, Bansal, R & Varma, V 2015, 'Towards Deep Semantic Analysis of Hashtags', Proceedings of the 37th European Conference on Information Retrieval *(ECIR 2015)*, Vienna, Austria.

Chin-Yew, C 2004, 'Rouge: A package for automatic evaluation of summaries', Proceedings of the ACL-04 workshop: Text summarization branches out. Barcelona, Spain.

Costa, C, Silva, C, Antunes, M & Ribeiro, B 2013, 'Defining Semantic Meta-Hashtags for Twitter Classification', Proceedings of the 11th International Conference on Adaptive and Natural Computing Algorithms, ICANNGA 2013, Lausanne, Switzerland.

Declerck, T & Lendvai, P 2015, 'Processing and Normalizing Hashtags', Proceedings of the Conference on Recent Advances in Natural Language Processing, Hissar, Bulgaria.

Gorrell, G, Petrak, J, & Bontcheva, K 2015, 'LOD-based Disambiguation of Named Entities in @tweets through Context #enrichment', Proceeding of the 12th European Semantic Web Conference (ESWC2015), Portoroz, Slovenia.

Joachims, T 2002, 'Optimizing search engines using clickthrough data'. Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD).

Kotsakos, D, Sakkos, P, Katakis, I & Gunopulos, D 2014, '#tag: Meme or Event?' In Proceedings of ASONAM 2014, Beijing, China.

Krokos, E & Samet, H 2014, 'A Look into Twitter Hashtag Discovery and Generation', Proceedings of the 7th ACM SIGSPATIAL Workshop on Location-Based Social Networks (LBSN'14), Dallas, Texas.

Lafferty, J & Zhai, C 2003, 'Probabilistic relevance models based on document and query generation', in *Language modeling for information retrieval*. Springer.

Laniado, D & Mika. P 2010, 'Making sense of Twitter', Proceedings of the 9th International Semantic Web Conference, Shanghai, China.

Llewellyn, C, Grover, C, Oberlander, J & Klein, E 2014, 'Re-using an Argument Corpus to Aid in the Curation of Social Media', Proceedings of the 9th Language Resources and Evaluation Conference, Reykjavik, Iceland

Manning, CD, Raghavan, P & Schütze, H 2008, *Introduction to Information Retrieval*, Cambridge University Press.

Meij, E, Trieschnigg, D, de Rijke, M & W. Kraaij 2010, 'Conceptual language models for domain-specific retrieval', *Information Processing and Management* 46, pp. 448–469, Elsevier.

Pöschko, J 2011, *Exploring Twitter Hashtags*. CoRR abs/1111.6553.

Rao, D, McNamee, P & Dredze, M 2010, 'Streaming Cross Document Entity Coreference Resolution', Proceedings of the Conference on Computational Linguistics (COLING)

Tanev, H, Ehrmann, M, Piskorski, J & Zavarella, V 2012, 'Enhancing Event Descriptions through Twitter Mining', Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM).

Wang, X, Wei, F, Liu, X, Zhou, M & Zhang, M 2011, 'Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach', Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1031-1040.

Wei, X, Callison-Burch, C & Dolan, B 2015, 'SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT).' Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval).

Weerkamp, W, Carter, S & Tsagkias, M 2011, 'How People use Twitter in Different Languages', Proceedings of Web Science, Koblenz Germany.

Wissler, L, Almashraee, M, Monett, D, & Paschke, A 2014 'The Gold Standard in Corpus Annotation', Proceedings of the IEEE Germany Student Conference, Passau, Germany.

Zubiaga, A, Liakata, M, Procter, R, Bontcheva, K & Tolmie, P 2015, 'Towards Detecting Rumours in Social Media', Proceedings of the AAAI Workshop on AI for Cities, Austin, Texas.