Similarity-Based Cross-Media Retrieval for Events

Piroska Lendvai and Thierry Declerck

Dept. of Computational Linguistics, Saarland University
Saarbrücken, Germany
piroska.r@gmail.com,declerck@dfki.de

Abstract. Our goal is to link social media content to contextually relevant information in complementary media in the domain of daily news. Web links from tweets with user-included URLs are transferred to URL-less tweets, using manually annotated events. The new cross-media ties establish authoritative feedback documents for unsupported social media content, and enable extracting an improved set of event-denoting terms based on longest common subsequences between tweets and documents.

Keywords: social media, information contextualization, similarity-based retrieval, cross-media feedback documents, term extraction

1 Introduction

We aim to create a cross-media (CM) linking algorithm in the *PHEME* project[1] to connect User-Generated Content (UGC) to topically relevant information in complementary media. Media that is complementary to UGC (in our pilot study, a tweet) is defined to be authoritative news releases on the web.

Recent natural language processing studies present some CM approaches with the purpose of aligning UGC and authoritative content. The goal of [5] is to collect information about emergency situations from tweets that are complementary to mainstream media reports. First, relevant keywords are determined from a centroid news article in a topically connected article cluster, and used in various query constructions to retrieve event-related tweets. The direction of linking is motivated by the need to boost retrieval precision on established events, which is orthogonal to the mission of the PHEME project – our targeted starting point is events that first emerge in social media and only later or not at all are covered in mainstream news releases. The algorithm of [5] is reused and extended in [2]: based on a centroid article in an event cluster, related tweets that contain URLs are mined, using custom-threshold-based term vector similarity. Then, relevance ranking takes place on these tweets, using platform-specific

Copyright © 2015 by the papers authors. Copying permitted only for private and academic purposes. In: R. Bergmann, S. Görg, G. Müller (Eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. Trier, Germany, 7.-9. October 2015, published at http://ceur-ws.org

[1] www.pheme.eu

indicators (number of mentions, retweets, etc). New, related articles on the web are retrieved based on the URLs of top-ranked tweets. [2] do not report on the proportion of web articles found that were already seen in the query-originating news cluster. Such information would evaluate the retrieval of complementary sources more transparently, and it forms an important part of our CM algorithm.

To implement CM linking for PHEME, our core assumption was that URL presence in tweets is a relevance feedback analogous to landing page information in click data, utilizable to develop retrieval functions from observed user behavior (see e.g. [3]). Referring to external sources is a multi-purpose activity in social media practices that may amalgamate among others intents of content framing (i.e., quoting authoritative sources) and content enrichment (i.e., guiding to extended information). Based on URLs that are present in tweets and point to web documents, we devised a method that transfers this explicit, user-included relevance signal to a collection of tweets that do not include explicit web links. The transfer is based on Events that have been manually annotated; each tweet is annotated with exactly one Event. Events are manually annotated situations or stories that describe smaller scale episodes than hashtag-denoted topics.

Our goal is to link URL-less tweets to a ranked list of web documents, where topic relevance is bootstrapped from event-based similarity between URL-including tweets and URL-less tweets, and ranking is based on aggregated n-gram similarity between tweet text and web document text. To this end, we extract and rank key phrases based on document–tweet similarity, and associate them with the Event the referring tweet is annotated with. As we focus on related content discovery and its use for rumour[2] verification purposes, our setup and results are more specific than the INEX tweet contextualization tasks (see e.g. [1]) to support a human reader.

2 Data and Algorithm

We worked with a dataset that consists of tweets relating to two broad events: (G) the Gurlitt art collection[3] and (O) the Ottawa shooting[4]. Tweets were pre-collected by filtering on event-related keywords (e.g. '*gurlitt*'), selecting events that meet the characteristics of a rumour. Each tweet was manually annotated for situations/stories (henceforth: Events[5]) that correspond to specific rumours, as described in [6]; for characteristics of the data see the top section of Table 1.

2.1 String similarity-based term extraction

For each URL-containing tweet within each Event, a tweet – document similarity calculation cycle is run. Similarity in the current implementation is based on

[2] defined in PHEME as *a circulating story of questionable veracity*

[3] https://de.wikipedia.org/wiki/Schwabinger_Kunstfund

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

[5] e.g. (G): `'The Bern Museum will accept the Gurlitt collection'`, `'Gurlitt was mentally unfit when he wrote his will'`; (O): `'There are snipers on the roof of the National Art Gallery'`, `'Shooter is still on the loose'`.

	Gurlitt	**Ottawa**
languages	DE, FR, EN	EN
events	3	51
tweets without URL	43	182
tweets with URL	147	341
unique URLs	143	187
fetchable web documents [by authoritative sources]	61 [61]	107 [107]
terms extracted from URLed tweets	110	169
terms extracted from URLless tweets	96	190
terms unseen in URLed tweets	83	143

Table 1. Characteristics of tweet data and of terms extracted from fetched web documents.

the Longest Common Subsequence (LCS) metric (cf. [4]). LCS is a language-independent, flexible-length skip-gram matching method that we apply on the token level for each tweet – document sentence pair[6]. No linguistic information is used, except for stopword filtering by the NLTK toolkit[7]. The process produces a ranked list of tweets based on LCS similarity with their linked document (which is in effect a user-coded feedback document) for all URL-providing tweets for a given Event, and outputs the longest common subsequence tokens between tweet and document body.

In the second pass, the cycle is applied to the same feedback web document set, now paired with tweets that did *not* link external documents but are hand-labeled with the same Events as the tweets from which web documents are referred from. This boosts the pool of linked authoritative[8] documents and tweets by 105% for *G* and 294% for *O*; extracted top-5 LCS phrases[9] grow qualitatively[10] by 75% for *G* and by 85% for *O*; cf. the bottom section of Table 1. An example output is provided below for the focus Event '`The Bern Museum will accept the Gurlitt collection`'.

Focus document's **headlines:** "Bestätigt: Kunstmuseum Bern nimmt das Erbe des Kunstsammlers Cornelius Gurlitt an - KURIER.at"
Top tweet with URL to focus document: *Bestätigt: Sammlung Gurlitt geht nach Bern http://t.co/FRCSHTU5hL*
LCS term of top URL-ed tweet and focus document: `'bestätigt sammlung gurlitt geht bern'`; Similarity **score:** 1.00
Top URL-less tweet labeled with focus Event: *RT @SWRinfo: Das Kunstmuseum Bern nimmt das Erbe des Kunstsammlers Cornelius #gurlitt an.*

[6] Casing is normalized, the retweet token, screen names and punctuation are removed
[7] nltk.org
[8] Based on a list of 25k authoritative news sources collected by PHEME.
[9] We keep the 5 most similar LCS phrases for each tweet–web document pair.
[10] I.e., in terms of obtaining new phrases that were unseen in the pool of URL-ed tweets–linked web documents.

LCS term for top URL-less tweet and focus document: `'kunstmuseum bern nimmt erbe kunstsammlers cornelius gurlitt'`; Similarity **score:** 0.79

3 Evaluation and Outlook

We presented a pilot study on transferring feedback document relevance for social media posts, based on manually annotated, fine-grained events. We used the LCS similarity metric to extract descriptive phrases for each Event; the obtained multi-word terms implicitly encode token proximity and word order, valuable for query- and document language modeling and indexing. LCS was also used to assign term-, respectively document weights to each Event, independent of a fixed document collection. Tweets with unsupported claims could be linked to authoritative web documents by utilizing hand-coded tweet–tweet similarity information; automatically obtaining this information is currently ongoing.

The findings suggest that LCS is advantageous when working with big data across languages and domains, as foreseen in the PHEME project. In future work we plan to compare LCS with other similarity metrics, as well as evaluate the obtained term, respectively document rankings in a retrieval scenario for information verification purposes. The major impact of Event-based bootstrapping of cross-media links is that we obtain a much larger set of cross-media context pairs, enabling the extraction of an improved list of event descriptors that can be put to use in fact checking and contextual document ranking, on which we plan to report in follow-up studies.

Acknowledgments We are grateful to two anonymous reviewers for their insightful comments. Work presented in this paper has been supported by the PHEME FP7 project (grant No. 611233).

References

1. Bellot, P., Moriceau, V., Mothe, J., Sanjuan, E., Tannier, X.: Overview of INEX tweet contextualization 2013 track. CLEF (2013)
2. Balahur, A., Tanev, C.: Detecting Event-Related Links and Sentiments from Social Media Texts. ACL Conference System Demonstrations (2013)
3. Joachims, T.: Optimizing search engines using clickthrough data. Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (2002)
4. Lin, Ch. Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out: Proceedings of the ACL-04 workshop. Vol. 8 (2004)
5. Tanev, H., Ehrmann, M., Piskorski, J., Zavarella V.: Enhancing Event Descriptions through Twitter Mining. In: Proceedings of ICWSM (2012)
6. Zubiaga, A., Liakata, M., Procter, R. N., Bontcheva, K., Tolmie, P.: Towards detecting rumours in social media. In: AAAI Workshop on AI for Cities (2015)