Overview of the CLEF 2016 Cultural micro-blog Contextualization Workshop

Lorraine Goeuriot[1], Josiane Mothe[2], Philippe Mulhem[1], Fionn Murtagh[3], and Eric SanJuan[4]

[1] LIG, Université de Grenoble, France
[2] IRIT, UMR5505 CNRS, ESPE, Université de Toulouse, France
[3] University of Derby, UK, and Goldsmiths University of London, UK
[4] LIA, Université d'Avignon, France
josiane.mothe@irit.fr eric.sanjuan@univ-avignon.fr

Abstract. CLEF Cultural micro-blog Contextualization Workshop is aiming at providing the research community with data sets to gather, organize and deliver relevant social data related to events generating a large number of micro-blog posts and web documents. It is also devoted to discussing tasks to be run from this data set and that could serve applications.

1 Introduction

1.1 Context

Many statistical studies have shown the importance of social media; they seem to be now the main Internet activity for Americans, even when compared to email [5], and most of the social media. Chinese users spend an average of almost 90 minutes per day on social networks [6]. Social media is thus a key media for any company or organization, specifically in Business Intelligence related activities. Companies use social data to gather insights on customer satisfaction, but can also relate this data to key performance indicators [1], forecast product or services revenues [2] or measure and optimize their marketing. On the other hand, there are several levers that make social media popular in such ways. In the context of Twitter, Liu *et al.* mention content gratification ("content of the information carried through Twitter") and technology gratification ("easy to use") as the main gratifications influencing user intentions to continue to use Twitter; other gratifications being process ("searching for something or to pass time") and social ("interactivity with other parties through media") gratifications [3].

With regard to events such as festivals, social media is now widely used, and gathers various communities at cultural events: organizers, media, attendees, general public not attending the event. These communities are generally interested in different aspects of the generated information:

[5] http://www.socialmediatoday.com/content/17-statistics-show-social-media-future-customer-service, http://www.businessinsider.com/social-media-engagement-statistics-2013-12?IR=T

[6] http://www.setupablogtoday.com/chinese-social-media-statistics/

– the organizers: social media is a nice way to promote an event because it is community-driven. Social media is also useful during the event to get feedback from attendees and because it allows short and timely updates. After the event, data analytics on the discussion is also a useful feedback;
– the media: other media make use of the content put by organizers and attendees to report the event, as well as to inform the public;
– the public attending a festival: social media is a means to get information on the event, and communicate with other attendees on the event itself or related topics;
– the public not attending a festival: to get attendees and media feedback about the event using social media.

Social media is becoming a core component of communication for any event either professional or cultural.

Mining and organizing the information surrounding a cultural event can help broadening the perception and visualization of its social impact. In particular, micro-blogging is increasingly used in cultural events like festivals. For instance, more than 10 million twitts containing the keyword festival were sent and shared this summer 2015. On one hand this massive social activity can transform a local cultural event into an international event buzz. On the other hand, major festivals that do not follow the social mainstream could fail in attracting and renewing the public. Several national public scientific programs, such as "Tourism Australia's Social Media Program" or "The Travel Michigan Social Media Workshop Series", at the crossroads of computer science and humanities aim at studying this phenomenon, and its impact on the tourism industry as well as its impact on major national public institutions and society.

1.2 Aims of the Workshop

The aims of the Workshop are (1) to build a collection of twitts on targeted topics; we choose the case of festivals and collected millions of twitts. (2) to analyze the automatically built data set in order to extract the data set characteristics and to know better what is in the data collection. (3) to run the pre-defined tasks during several months and to define new tasks (during the Workshop day itself in September 2016).

In this paper, we present the corpus compiled for the CLEF Cultural micro-blog Contextualization and experimental tasks that make use of this corpus. This corpus has been built to study the social media sphere surrounding a cultural event, and contains micro-blog posts, a knowledge source, as well as all the web pages linked from the micro-blog posts.

More precisely, we first introduce use cases in Section 2, then we describe the data sets in Section 3. Section 4 gives some insights of the corpus while Section 5 depicts the experimented tasks which correspond to the pre-defined tasks. Section 6 concludes this paper.

2 Use Case Scenario

The goal of the CLEF Cultural micro-blog Lab is to develop processing methods and resources to mine the social media sphere surrounding cultural events such as festivals. twitts linked to an event makes a dense, rich but very noisy corpus: informal language, out of the language phrases and symbols, hashtags, hyperlinks... The information is also often imprecise, duplicate, or non-informative. The interest of mining such data is to extract relevant, and informative content, as well as to potentially discover new information.

The 2016 CMC Workshop is centred on festival participants, and focusing on, but is not limited to, the following use cases:

- An insider participant gets a micro-blog post about the cultural event in which he or she is taking part but needs context to understand it (micro-blog posts often contain implicit information). He or she needs also this background information before clicking on the link if any because the network activity is low or to avoid leaving the micro-blog application. The contextualization systems to be experimented with in this lab have to provide a short highly informative summary extracted from the Wikipedia that explains the background of the micro-blog post text.
- A participant in a specific location wants to know what is going on in surrounding events relative to artists, music or shows that he or she would like to see. Starting from a list of bookmarks in the Wikipedia app, the participant seeks a short list of micro-blog posts summarizing the current trends about related cultural events. She/he is more interested in micro-blogs from insiders than outsiders

While our goal is to build data sets that will help research centred on the use cases above, we can foresee new research challenges that could be investigated with this data set: cultural events are often facing a big data challenge: direct stakeholders (organizers, artists, attendees), as well as indirect ones (media, public) can express themselves about the event, in different ways, media, and even languages. This data can be seen as a virtual sphere surrounding the event itself. Mining and organizing such data could bring very useful information on the events and their content. Besides the use cases given above, we believe such a corpus could lead to solve many other challenges in the domain.

For example, in the official web site for the Festival de Cannes 2015, in the part dedicated to the opening ceremony the May 13th, gives some excerpts of the speech of L. Wilson, but the tweet id=598999849091084288, sent the May 15th, is about a TV talkshow commenting the opening ceremony with the actor S. Baker. This tweet does not depicts the ceremony and should not be relevant to describe the opening ceremony. On the contrary, the tweet id=598636861280489473 lists some actress names (C. Deneuve, Noémie Lenoir, Natalie Portman) which are not in the official site but that give interesting information about the actual ceremony and is then relevant to the opening ceremony.

The biggest foreseen problem encountered in this scenario is the "mapping" between events and posts. It is not one-to-many, but many-to-many: one micro-

blog may be relevant to several events, and most of the time, a single event is mentioned in many posts. Moreover, one post may not be related to the events at all. Messages may be indirectly related to one event (a reply in a conversation for instance).

3 Datasets

3.1 micro-blog posts collection

We collect all public micro-blog posts from Twitter containing the keyword festival since June 2015 using a private archive service with Twitter agreement based on streaming API[7]. The average of unique micro-blog posts (i.e. without re-twitts) between June and September is $2,616,008$ per month.The total number of collected micro-blog posts after one year (from May 2015 to May 2016) is 50, 490, 815 (24, 684, 975 without re-posts).

These micro-blog posts are available online on a relational database with associated fields, among them 12 are listed in Table 1. The "Comments" row in Table 1 gives some figures about the existing corpus.

Table 1. Fields of the micro-blog posts collection.

Name	Description	Comments
text	text of the twitt	99% of the twitts contain a non empty text 66% contain an external compressed URL
from_user	author of twitt (string)	$62,105$ organizations among $11,928,952$ users.
id	unique id of micro-blog	total so far: $50,490,815$ posts.
iso_language_code	encoding of the twitt	the most frequent tags: en (57%), es (15%), fr (6%) and pt (5%).
source	interface used for posting the twitt	frequent tags: Twitter Web Client (16%) iPhone and Twitterfeed clients (11% each).
<geo_type, geo_coordinates_0, geo_coordinates_1>	geolocalization	triplet valued in 2.3% of the twitts.

Because of privacy issues, they cannot be publicly released but can be analyzed inside the organization that purchases these archives and among collaborators under privacy agreement. CLEF 2016 CMC Workshop will provide this opportunity to share this data among academic participants. These archives can be indexed, analyzed and general results acquired from them can be published without restriction. The Workshop will organize a scientific peer reviewed process among participants to discuss and to check the validity and reproducibility of results.

3.2 Linked web pages

66% of the collected micro-blog posts contain Twitter *t.co* compressed URLs. Sometimes these URLs refer to other online services like *adf.ly, cur.lv, dlvr.it*,

[7] `https://dev.twitter.com/streaming/public`

ow.ly, *thenews.uni.me* and *twrr.co.vu* that hide the real URL. We used the spider mode to get the real URL, this process can require several DNS requests. The number of unique uncompressed urls collected in one year is $11,580,788$ from $641,042$ distinct domains. Most frequent domains are: twitter.com (23%), www.facebook.com (5.7%), www.instagram.com (5.0%), www.youtube.com (4.5%), item.ticketcamp.net (1.1%) and g1.globo.com (1%)

3.3 Wikipedia Crawl

Unlike twitts and web pages, Wikipedia is under Creative Common license, and its contents can be used to contextualize twitts or to build complex queries referring to Wikipedia entities. Using the tools from INEX twitt conceptualization track[8] we extract from Wikipedia an average of 10 million XML documents per year since 2012 in the four main Twitter languages: English (en), Spanish (es), French (fr), and Portuguese (pt). These documents reproduce in an easy to use XML structure the contents of the main Wikipedia pages: title, abstract, section and subsections as well as Wikipedia internal links. Other contents as images, footnotes and external links are stripped out in order to obtain a corpus easy to process by standard NLP tools. By comparing contents over the years, it is possible to detect long term trends.

4 micro-blog Corpus Insights

An extended version of the analysis in this section is in [7]. Previous work on determining and analyzing narrative flows are available in [9], using Twitter data from a designed experiment; and in [8], using film script.

A perspective on an analysis carried out, is as follows. twitts between 11 May and 31 December 2015 were used.

The following festivals are selected: Cannes Film Festival (13–24 May 2015); Fèis Ìle, Islay (Scotland) Festival (23–31 May 2015); Berlin Film Festival (19–21 May 2015); CMA, Country Music Association (Nov. 2015); Yulin Dog (June 2015); and Avignon Theatre Festival (4–25 July 2015).

Figure 1 displays the planes of principal axes, i.e. factors, 1, 2 and of axes 3, 4. This is Correspondence Analysis providing analysis of semantics. We do see here how the principal factor plane is especially a contrasting engagement with the Cannes Film Festival for axis 1, and the Avignon Theatre Festival for axis 2. Meanwhile, both axes 3 and 4 can be said to be especially relevant for the Avignon Theatre Festival.

Such festivals are the central focus of interest in this micro-blog corpus. Dealing with the complete dataset would require:

1. semantic characteristics of words in the twitts (or abbreviations, named individuals like performers, political or other happenings, web addresses, language used, etc.).

[8] `http://tc.talne.eu`

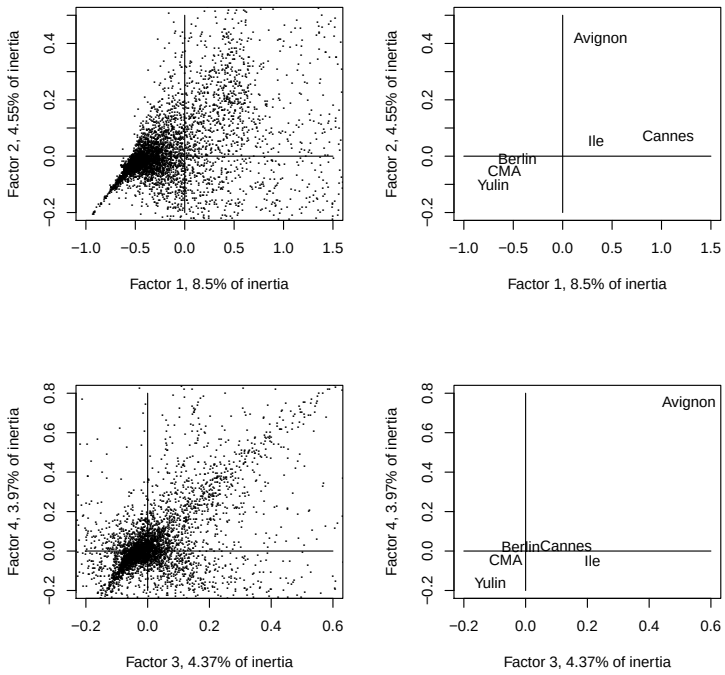

Fig. 1. The principal factor plane, in the top two panels, and the plane of factors 3,4 in the bottom two panels. The left panels display all words, with a dot at each word location. The right panels display the selected festivals.

2. pattern recognition in the data, and discovery of, and characterizing, trends.
3. predictive modeling and other approaches (e.g. quantitative measures of impact or performance).

Overall, we may have the foundations here for Bourdieu-based social research [10, 11].

5 Experimented tasks

Along with these three data sources (micro-blog posts, related Web and Wikipedia Encyclopedia), three types of search queries with related textual references will be provided to evaluate micro-blog systems:

– Contextualization based on Wikipedia where given a twitt as query the system has to provide a short summary extracted from the Wikipedia that provides all necessary background knowledge to fully understand the twitt.

– Summarization based on twitts where given a topic represented by a set of Wikipedia entities, extract a reduced number of twitts that summarizes main trends about that topic in festivals.
– Event link of a given festival program. Such information is useful for attendees of festivals, for people who are interested in knowing what happens in a festival, and for organizers to get feedback.

System outputs will be evaluated based on informativeness as in [4, 5]. Manual runs and Questionnaire data will be provided by the French ANR GAFES project[9].

6 Conclusion

We presented in this paper the Cultural micro-blog Contextualization (CMC) corpus, a temporal comprehensive representation of the virtual sphere surrounding cultural events. This corpus is composed of twitts, URLs linked to by these twitts, and of one knowledge source.

The built corpus has the big interest to provide a snapshot of: a) twitts, and, b) web pages pointed to by the twitts. From a scientific point of view, it will be possible to rerun experiments on the exact same sets of web documents, even years after the event took place. The topics covered by the corpus have several benefits:

– The amount of data in the corpus is manageable by academic research teams (around 50 millions of twitts and URLs, possibly split into smaller subsets depending on the task expected). This point is important as we expect numerous participants to experiment their ideas on the CMC corpus;
– Forcing the corpus perimeter to festival cultural events still covers a variety of festivals (cinema, music, theater, ...) that may have different features regarding their related social spheres;
– The cultural domain is usually well documented in resources like Wikipedia, so the CMC corpus will not suffer from the lack of knowledge that may be used during retrieval.

Without limiting the possible uses of this corpus, we foresee that the concurrent gathering of web pages and twitts may also pave the way to other studies inspired from [12], like co-evolutions of twitts and referred web pages over several occurrences of the same festival, or co-dynamics of topics in web pages and twitts.

We also presented some tasks associated with this data set. During the Workshop day at CLEF in September 2016, the collection will be discussed. We will discuss the quality of the data set based on analysis some participants have conducted, as well as the distribution of the corpus in accordance with the agreement we have with Twitter. During the Workshop day we will also discuss other possible tasks to be run over the data set.

[9] `http://www.agence-nationale-recherche.fr/?Project=ANR-14-CE24-0022`

References

1. Heijnen, J., de Reuver, M., Bouwman, H., Warnier, M., Horlings, H. : Social Media Data Relevant for Measuring Key Performance Indicators? A Content Analysis Approach. In Co-created Effective, Agile, and Trusted eServices, Lecture Notes in Business Information Processing, Vol. 155, Springer Berlin Heidelberg, 74–84, 2013.
2. Rui, H., Whinston, A. : Designing a Social-broadcasting-based Business Intelligence System, ACM Trans. Manage. Inf. Syst., ACM, New York, NY, USA, 2(4):1–19, 2011.
3. Liu, I., Cheung, C., Lee, M. : Understanding Twitter Usage: What Drive People Continue to twitt., PACIS, 92, 2010.
4. SanJuan, E., Bellot, P, Moriceau, V., Tannier, X., Overview of the INEX 2010 Question Answering Track (QA@INEX), in: S. Geva, J. Kamps, R. Schenkel, A. Trotman (Eds.), INEX, Vol. 6932 of Lecture Notes in Computer Science, Springer, 2010, pp. 269–281.
5. Bellot, P., Moriceau, V., Mothe, J., Tannier, X., SanJuan, E. : Mesures d'informativité et de lisibilité pour un cadre d'évaluation de la contextualisation de twitts in Document Numérique, vol.18(1), 2015, pp: 55–73.
6. Benz, D., Hotho, A., Jäschke, R, Krause, B., Mitzlaff, F., Schmitz, C., Stumme, G : The Social Bookmark and Publication Management System Bibsonomy. The VLBD Journal, Vol. 19, 849–875, 2010.
7. Murtagh, F. : Semantic mapping: Towards contextual and trend analysis of behaviours and practices, 2016. (Online proceedings.)
8. Murtagh, F., Ganz, A., McKie, S. : The structure of narrative: The case of film scripts, Pattern Recognition, 42, 302–312, 2009.
9. Murtagh, F., Pianosi, M., Bull, R. : Semantic mapping of discourse and activity, using Habermas's Theory of Communicative Action to analyze process, Quality and Quantity, 50(4), 1675–1694, 2016.
10. Le Roux, B., Rouanet, H. : Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis, Kluwer (Springer), Dordrecht, 2004.
11. Le Roux, B., Lebaron, F. : Idées-clefs de l'analyse géométriques des données. In F. Lebaron and B. Le Roux, editors, La Méthodologie de Pierre Bourdieu en Action: Espace Culturel, Espace Social et Analyse des Données, pp. 3–20, Dunod, Paris, 2015.
12. Leskovec, J., Backstrom, L., Kleinberg, J. : Meme-tracking and the dynamics of the news cycle. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09). ACM, New York, NY, USA, 497–506, 2009.