Building Evaluation Datasets for Cultural Microblog Retrieval

Lorraine Goeuriot, Josiane Mothe, Philippe Mulhem, Eric SanJuan
LIG, IRIT, LIA
Univ. Grenoble Alpes, CNRS, Grenoble INP, France
Université de Toulouse, France
Université d'Avignon, France
{lorraine.goeuriot,philippe.mulhem}@univ-grenoble-alpes.fr
josiane.mothe@irit.fr, eric.sanjuan@univ-avignon.fr

Abstract

ECLEF Microblog Cultural Contextualization is an evaluation challenge aiming at providing the research community with datasets to gather, organize and deliver relevant social data related to events generating large number of microblogs and web documents. The evaluation challenges runs every year since 2016. We describe in this paper the resources built for the challenge, that can be used outside of the context of the challenge.

Keywords: microblog search, social network moining, cultural data, evaluation, informativeness

1. Introduction

Many statistic studies have shown the importance of social media; They seem to be now the main Internet activity for Americans, even when compared to email[1], and most of the social media. Chinese users spend an average of almost 90 minutes per day on social networks[2]. 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 forecast product or services revenues (Rui and Whinston, 2011) or measure and optimize their marketing. On the other hand, there are several levers that make social media such popular. 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 users'intention 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 (Liu et al., 2010).
With regards 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:

- 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 mean to get information on the event, and communicate with other attendees on the vent it-self 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, microblogs are increasingly used in cultural events like festivals. For instance, more than 10 million tweets containing the keyword *festival* were sent and shared over the summer 2015. On one hand this massive social activity can transform a local cultural event into an international buzz feed. 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 at the crossfield of computer science and humanities aim at studying this phenomena, and its impact on the tourism industry as well as its impact on major national public institutions and society.
We present in this paper the corpus compiled for the CLEF Cultural Microblog Contextualization challenge. This corpus has been built to study the social media sphere surrounding a cultural event, and contains microblogs, a knowledge source, as well as all the web pages linked from the microblogs. We introduce use cases in Section 3. and describe the datasets in Section 4.

2. State of the Art

There are two major trends when building data sets. In the one hand, reference collections are build in order to make it possible for researchers from all over the world to confront their methods and algorithms on common data. This view is the one the main conferences and programs for evaluation

[1] 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

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

follow. On the other hand, specific and owner-based collections are developed to answer or evaluate a very specific task or are based upon data with specific ownership.

In this paper, we do not intend to provide an exhaustive overview of the various datasets evaluation programs or researchers released. Rather, we focus on collections related either to microblog or events. Moreover, since usually collections are built in order to fit specific tasks, we focus on the ones related to IR tasks.

TREC (Text REtrieval Conference) is one of the major conference for IR evaluation. It runs every year since 1992 a set of tasks that aim at studying specific IR tasks.

TREC Contextual Suggestion track [3] is probably the collection most related to ours. This track aims at providing users with recommendations of "interesting places and activities based on the user's location, personal preferences, past history, and environmental factors such as weather and time" (Dean-Hall et al., 2015). In this collection, there is a set of attractions which consist in information on the attraction it-self and on its context (city, URL and title). This track started in 2012. The 2016 collection can be download at `http://145.100.59.205:8095/TREC2016_CS_Collection.zip`.

With regard to Twitter, while many papers refer to datasets that are built using the social network API, a few datasets have been developed and released for different tasks.

With regard to location extraction there are two public collections: the Ritter collection (Strauss et al., 2016) and the MSM2013 collection (Rowe et al., 2015), both of which are reference collections in the domain. The first collection was initially used by Ritter *et al.* (Strauss et al., 2016) while the second was the training dataset from Making Sense of Microposts 2013 (MSM2013). These two datasets are provided along with manual annotations on locations. Table 1 shows the number of tweets and their distribution (according to whether they mention a location or not) in both datasets. These data sets have been used for example in (Rowe et al., 2015; Ngoc and Mothe, 2018).

In TREC, the Microblog ran from 2011 until 2015 and targets an IR-oriented task. In 2011, this task addresses real-time adhoc search over 16M tweets (Soboroff et al., 2012). In 2015, a real time filtering task was introduced.

In NTCIR[4], the task is as follows: given a microblog post, retrieve or generate a coherent and useful response. Two Asian languages are considered: Chinese with post-comment pairs from Weibo, and Japanese on Twitter (Shang et al., 2016).

IN CLEF, the Tweet Contextualization task was first introduced as part of the INEX QA task (SanJuan et al., 2010) and became a full task the year after. The Tweet Contextualization task mainly focuses on helping a reader to understand a tweet by providing him a short summary of what the tweet is about. The organizers provided sets of tweets each year and the associated Wikipedia dump which was used for building the summaries. Bellot *et al.* provides an overview of the task and lessons learn (Bellot et al., 2016).

3. Use Case Scenario

The goal of the proposed corpus to help developing processing methods to mine the social media sphere surrounding cultural events such as festivals according to several points of views. Tweets linked to an event make a dense, rich but very noisy corpus. As described in (Gimpel et al., 2011), informal language, out of the language phrases and symbols, hashtags, hyperlinks, multi-words abbreviations, are all elements that lead to the fact that the information conveyed by a tweet is often imprecise. Additionally, many tweets are strict or near duplicates, leading to the fact that a special effort has to be put on their management during microblogs retrieval. Tweets also support interaction between users, leading to some of then to an interaction role without any topical content. The interest of mining such data is to extract relevant, and informative content, as well as to potentially discover new information.

The corpus provided is centered on festival participants, and therefore the use cases that we focus on are related to the tweets flow related to such cultural events. Tweets may focus on the whole festival, where others may concentrate on one specific show, or even a detail of such show. Typical use cases related to our corpus are:

1. A participant may get microblogs about the cultural event in which he is taking part, but one or several microblogs taken apart often contain implicit information: the ability to provide "contextual" information that is needed to understand the microblogs is then one interesting scenario. Such background information is important when the user is on the festival site and has a low bandwidth or because he does not want to switch between applications on his handphone. In this case, contextualization systems to be experimented have to provide with a short but highly informative summary extracted from Wikipedia that explains the background of one microblog text.

2. A participant in a specific location wants to know what is going on in surrounding events relatively to artists, music or shows that he would like to see. Starting from a list of bookmarks in the Wikipedia app, the participant seeks for a short list of microblogs summarizing the current trends about related cultural events. The idea of this scenario goes then from wikipedia to microblogs, and we are to what can be achieved by a personalized information retrieval system, in which the user profile is the user's Wikipedia app bookmark list. On important point though, is that we are more focusing on microblogs from insiders than outsiders, i.e., from real participants to the cultural event than from comments from people that are commenting the event from outside.

3. According to a given program of a festival (accessible through the official web site of the festival for instance), the organizers or any user may look for all the tweets related to the festival highlight. In this case,

[3] `https://sites.google.com/site/treccontext/trec-2016/trec-2016-contextual-suggestion-guidelines`

[4] NII Testbeds and Community for Information access Research

Table 1: Summary of the Ritter and MSM2013 datasets used to evaluate location extraction models on tweets.

	Ritter's dataset	MSM2013 dataset
# of tweets	2,394	2,815
# of tweets containing a location (TCL)	213 (8.8%)	496 (17.6%)
# of tweets without location (TNL)	2,181	2,319

the official program is used as a source for generating queries. Namely, the program is a list of triplets <title of highlight, date/time, location>, and each triplet, each triplet being a query. For the organizers of a festival, the interest of this scenario is to obtain an overview of the festival according to microblogs. For a attendee of a festival, the interest of this scenario is to get a recall about what he saw (for instance when a guest who joins a band on stage, we do not always know his name: this use case allows to gather such information *a posteriori*).

While our goal is to build datasets that will help research centered on the use cases above, we can foresee new research challenges that could be investigated with our dataset: 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 explore many other challenges in the domain, like the integration of time and localization in microblogs contextualization and retrieval.

4. Datasets

The dataset created for this evaluation lab contains several parts, described in the sections below.

4.1. Microblogs collection

We collected all public microblog from twitter containing the keyword festival from May 2015 to November 2016 using a private archive service with twitter agreement based on streaming API[5]. The average of unique microblogs (i.e. without re-tweets) between June and September is $2,616,008$ per month.

These microblogs are provided in UTF8 csv format with the 13 fields, among them 12 are listed in table 2. The "Comments" row in table 2 gives some figures about the existing corpus.

Because of privacy issues, they cannot be publicly released but can be analyzed inside the organization that purchase these archives and among collaborators under privacy agreement. CLEF Labs[6] 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.

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

[6] https://mc2.talne.eu

4.2. Linked web pages

66% of the collected microblogs contain Twitter *t.co* compressed urls. Sometimes these urls refer to other online services like *adf.ly, cur.lv, dlvr.it*, *ow.ly*, *thenews.uni.me* and *twrr.co.vu* that hide the real url.

4.3. Wikipedia Crawl

Unlike tweets and web pages, wikipedia is under Creative Common license, and its contents can be used to contextualize tweets or to build complex queries referring to wikipedia entities. Using the tools from INEX tweet conceptualization track[7] we extracted from wikipedia an average of 10 million XML documents per year since 2012 in the four main twitter languages: en, es, fr and 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 other 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.

4.4. Textual Assessments for Evaluation

Along with this three data sources (Microblogs, related Web and wikipedia Encyclopedia), two types of search queries with related textual references are provided to evaluate systems for Microblog:

- Contextualization based on Wikipedia where given a tweet as query the system has to provide a short summary extracted from the wikipedia that provides all necessary background knowledge to fully understand the tweet.
- Summarization based on tweets where given a topic represented by a set of wikipedia entities, extract a reduce number of tweets that summarizes main trends about that topic in festivals.

System outputs were evaluated based on informativeness as in(SanJuan et al., 2010). Manual runs and Questionnaire data were provided by the French ANR GAFES project.

5. Results obtained with the dataset - Challenge participation

Participation to these challenges have been reported in (Goeuriot et al., 2016) and (Ermakova et al., 2017). In our case, we shall focus on general results that show what is achievable over this data.

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

Table 2: Fields of the Microblogs collection.

Name	Description	Comments
text	text of the tweet	99% of the tweets contain a non empty text 66% contain an external compressed URL
id	unique id of tweet	total $80,641,580$ tweets.
from_user	author of tweet (string)	$16,128,316$ organizations among $3,577,724$ users.
iso_language_code	encoding of the tweet	133 tags: en (57%), es (15%), fr (6%) and pt (5%).
source	interface used for posting the tweet	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 tweets.
<created_at, time>	date/time of tweet	15.1% of the tweets are created on Sundays, and 13.3% on Thursdays.

The first outcome was that a microblog corpus over a such a long time period allows to represent tokens by temporal series: every token is represented by the vector of its occurrences in the microblogs grouped by week (78 weeks in total). This is like a temporal word embedding and allows to span along similar trends at concomitant periods which in the case of cultural events like festivals is essential. Following (Murtagh, 2016) presented at CLEF 2016 CMC workshop, an interface was set up in order to gain a better understanding of these correlations[8]. Figure 1 shows a hierarchical clustering among cities, music styles and other festival themes. It naturally clusters together Cannes, Hollywood, films, movies and film makers. However, it also reveals some proximity between Deezer, the main French music streaming service and free/trance music festivals. In a similar fashion, there appears to be a correlation between Apple and London due to the past Apple music festival in London (September 2015 and 2016[9]). Although it can also be noted that Spotify doesn't appear to be correlated to any specific festival event over this corpus.

The second main outcome was the use of WikiPedia as an exhaustive multilingual terminological resource over microblogs related to cultural events. Contextualizing microblogs appeared to be more effective than focus retrieval approaches to link microblogs with WikiPedia pages. That is, instead of considering the content as a query, WikiPedia text anchors were matched against it. Furthermore, by using systems like FELTS[10] based on state of the art Hash functions, it was even possible to upgrade this approach and apply it in real time on the stream of microblogs.

Another outcome was the difficulty to identify microblog languages in this corpus without using specialized lexical resources(Hamon et al., 2017). Not only does the term festival appears in microblogs in any known language, but microblogs can refer to several languages at a time, for example they can use a term that related languages (ex: Spanish covering 15% of the corpus and Catalan in only 0.21% of the microblogs) or language variants (ex: Mexican Spanish in 0.01% of the corpus). They can also mix two different languages like Parisian French (6%) and one Arabic dialect (0.09%).

On a further note, when looking at 16,769,807 collected unique urls in microblogs, it appears that less than 1/10 th refer to some accessible public content. Meanwhile, 2/3rds refer to content in private social networks and the rest refer to content that is hidden behind some pay wall. Therefore, it is not possible to automatically crawl the web sphere around these microblogs.

6. Conclusion

We presented in this paper the Cultural Microblog Challenge (MC2) corpus, a temporal comprehensive representation of the virtual sphere surrounding cultural events. This corpus is composed of tweets, web pages linked to by these tweets, and of one knowledge source.

The built corpus has the big interest to provide a snapshot of: a) tweets, and, b) web pages pointed to by the tweets, these pages being downloaded **as soon as the tweet is received**. 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. Reproducibility of results is then ensured, unlike with the Bibsonomy test collection (Benz et al., 2010), in which only URLs of web pages are provided, lowering the capacity to really compare systems, as web pages evolve in time. The topics covered by the corpus have several benefits:

- The amount of data in the corpus is manageable by academic research teams (around 20 millions of tweets, several millions of web documents, 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 MC2 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 MC2 corpus will not

[8] https://mc2.talne.eu/shiny/gafes/ts_c/

[9] https://en.wikipedia.org/wiki/Apple_Music_Festival

[10] https://github.com/jourlin/FELTS

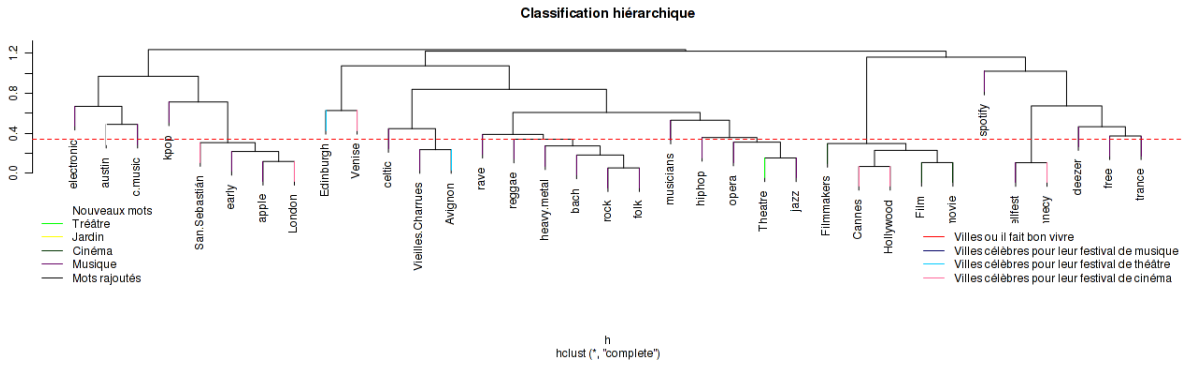

Figure 1: Correlations between temporal series associated with words

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 tweets may also pave the way to other studies, like co-evolutions of tweets and referred web pages over several occurrences of the same festival, or co-dynamics of topics in web pages and tweets.

7. Bibliographical References

Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., and Tannier, X. (2016). INEX tweet contextualization task: Evaluation, results and lesson learned. *Inf. Process. Manage.*, 52(5):801–819.

Benz, D., Hotho, A., Jäschke, R., Krause, B., Mitzlaff, F., Schmitz, C., and Stumme, G. (2010). The social bookmark and publication management system bibsonomy - A platform for evaluating and demonstrating web 2.0 research. *VLDB J.*, 19(6):849–875.

Dean-Hall, A., Clarke, C. L. A., Kamps, J., Kiseleva, J., and Voorhees, E. M. (2015). Overview of the TREC 2015 contextual suggestion track. In *TREC*, volume Special Publication 500-319. National Institute of Standards and Technology (NIST).

Ermakova, L., Goeuriot, L., Mothe, J., Mulhem, P., Nie, J., and SanJuan, E. (2017). CLEF 2017 microblog cultural contextualization lab overview. In *CLEF*, volume 10456 of *Lecture Notes in Computer Science*, pages 304–314. Springer.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *ACL (Short Papers)*, pages 42–47. The Association for Computer Linguistics.

Goeuriot, L., Mothe, J., Mulhem, P., Murtagh, F., and SanJuan, E. (2016). Overview of the CLEF 2016 cultural micro-blog contextualization workshop. In *CLEF*, volume 9822 of *Lecture Notes in Computer Science*, pages 371–378. Springer.

Hamon, O., Monnin, C., and de Loupy, C. (2017). Syllabs team at CLEF MC2 task 1: Content analysis. In *CLEF (Working Notes)*, volume 1866 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Liu, I. L. B., Cheung, C. M. K., and Lee, M. K. O. (2010). Understanding twitter usage: What drive people continue to tweet. In *PACIS*, page 92. AISeL.

Murtagh, F. (2016). Semantic mapping: Towards contextual and trend analysis of behaviours and practices. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 1207–1225.

Ngoc, H. T. B. and Mothe, J. (2018). Location extraction from tweets. *Inf. Process. Manage.*, 54(2):129–144.

Rowe, M., Stankovic, M., and Dadzie, A.-S. (2015). #microposts2015 - 5th workshop on 'making sense of microposts':big things come in small packages. In *WWW (Companion Volume)*, pages 1551–1552. ACM.

Rui, H. and Whinston, A. B. (2011). Designing a social-broadcasting-based business intelligence system. *ACM Trans. Management Inf. Syst.*, 2(4):22:1–22:19.

SanJuan, E., Bellot, P., Moriceau, V., and Tannier, X. (2010). Overview of the INEX 2010 question answering track (qa@inex). In *INEX*, volume 6932 of *Lecture Notes in Computer Science*, pages 269–281. Springer.

Shang, L., Sakai, T., Lu, Z., Li, H., Higashinaka, R., and Miyao, Y. (2016). Overview of the NTCIR-12 short text conversation task. In *NTCIR*. National Institute of Informatics (NII).

Soboroff, I., Ounis, I., Macdonald, C., and Lin, J. J. (2012). Overview of the TREC-2012 microblog track. In *TREC*, volume Special Publication 500-298. National Institute of Standards and Technology (NIST).

Strauss, B., Toma, B., Ritter, A., de Marneffe, M., and Xu, W. (2016). Results of the WNUT16 named entity recognition shared task. In *NUT@COLING*, pages 138–144. The COLING 2016 Organizing Committee.