CLEF 2017 MC2 Search and Timeline tasks Overview

, Lorraine Goeuriot[1], Philippe Mulhem[1], and Eric SanJuan[2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
[2] LIA, Université d'Avignon, France
lorraine.goeuriot@imag.fr, philippe Philipe.mulhem@imag.fr,
eric.sanjuan@univ-avignon.fr

Abstract. MC2 CLEF 2017 lab investigates the relationship between cultural microblogs and their social context. This involves microblog search, classification, filtering, language recognition, localization, entity extraction, linking open data, and summarization. The goal of the timeline illustration track is to study approaches that better retrieve microblogs issued during a cultural event, in order to get a glimpse of the attendees' perception. Regular Lab participants have access to the private massive multilingual microblog stream of *The Festival Galleries* project. Festivals have a large presence on social media. The topics were in four languages: Arabic, English, French and Spanish, and results were expected in any language.

1 Introduction

MC2 CLEF 2017 lab investigates the relationship between cultural microblogs and their social context. This involves microblog search, classification, filtering, language recognition, localization, entity extraction, linking open data, and summarization. The goal of the timeline illustration track is to study approaches that better retrieve microblogs issued during a cultural event in order to get a glimpse of the attendees' perception.

In 2016, the CLEF MC2 Workshop considered specific cultural twitter feeds [1]. In this context, restricted context, implicit localization and language identification appeared to be important issues. It also required identifying implicit timelines over long periods. The MC2 CLEF 2017 lab has been centered on Cultural Contextualization based on microblog feeds. It deals with how the cultural context of a microblog affects its social impact at large [2].

The overall usage scenario for the lab has been centered on festival attendees:

- an insider attendee who receives a microblog about the cultural event which he will participate in will need context to understand it (microblogs often contain implicit information);
- a participant in a specific location wants to know what is going on in surrounding events related to artists, music, or shows that he would like to see. Starting from a list of bookmarks in the Wikipedia app, the participant

will seek for a short list of microblogs summarizing the current trends about related cultural events. We hypothesize that she/he is more interested in microblogs from insiders than outsiders or officials.

These scenarios lead to three tasks lab participants could answer to: (1) Content analysis, (2) Microblog search, (3) Timeline illustration.

This paper describes the Timeline illustration task. The purpose of the task is to provide a glimpse of the atmosphere of a festival. To do so, participants are asked to retrieve for each events within a festival (concerts, plays, etc.) all the relevant tweets from the dataset.

Section 2 describes the datasets provided to participants. Participants' submissions are described in Section 3, and conclusions are given in Section 4.

2 Data

The lab gave registered participants access to a massive collection of microblogs and URLs related to cultural festivals around the world.

A personal login was required to acces the data. Once registered on CLEF each registered team can obtain up to 4 extra individual logins by writing to admin@talne.eu. This collection is still accessible on demand. Any usage requires to make a reference to the following paper: "L. Ermakova, L. Goeuriot, J. Mothe, P. Mulhem, J.-Y. Nie, and E. SanJuan, CLEF 2017 Microblog Cultural Contextualization Lab Overview, Proceedings of Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference of the CLEF Association, CLEF 2017, LNCS 10439, Dublin, Ireland, September 11-14, 2017". Updates will be frequently posted on the lab website[3].

An Indri index with a web interface is available to query the whole set of microblogs. Online Indri indexes are available in English, Spanish, French, and Portuguese for Wikipedia search.

2.1 Microblog Collection

The document collection is an updated extension of the microblog stream presented at the CLEF 2016 workshop [1](see also [3]).

It was provided to registered participants by ANR GAFES project[4]. It consists in a pool of more than 50M unique microblogs from different sources with their meta-information as well as ground truth for the evaluation.

The microblog collection contains a very large pool of public posts on Twitter using the keyword "festival" that have been collected since June 2015. These microblogs were collected using private archive services based on streaming API. The average number of unique microblog posts (i.e. without re-tweets) between June and September is 2,616,008 per month. The total number of collected microblog posts after one year (from May 2015 to May 2016) is 50,490,815

[3] `https://mc2.talne.eu/lab/`

[4] http://www.agence-nationale-recherche.fr/?Projet=ANR-14-CE24-0022

(24,684,975 without re-posts). These microblog posts are available online on a relational database with associated fields.

Because of privacy issues, they cannot be publicly released but can be analyzed inside the organization that purchased these archives and among collaborators under a privacy agreement. The MC2 lab provides this opportunity to share this data amongst academic participants. These archives can be indexed, analyzed and general results acquired from them can be published without restriction.

2.2 Linked Web Pages

66% of the collected microblog posts contain *Twittert.co* compressed URLs. Sometimes these URLs refer to other online services like *adf.ly, cur.lv, dlvr.it, ow.ly* that hide the real URL. We used the spider mode of the GNU *wget* tool to get the real URL, this process required multiple DNS requests.

The number of unique uncompressed URLs collected in one year is 11,580,788 from 641,042 distinct domains.

2.3 Topics for Microblog Search

Given a cultural query about festivals in Arabic, English, French, or Spanish, the task is to search for the 64 most relevant microblogs in a collection covering 18 months of news about festivals in all languages.

Queries have been extracted from resources suggested by participants.

Arabic and English queries were extracted from the Arab Spring Microblog corpus [4]. We considered the content of all the tweets dealing with festivals during the Arab Spring period. The task consisted in searching for traces of these festivals or artists in the lab corpus two years after this period. The usual case was to follow up artists involved in the Arab spring festivals two or three years later. There were 71 topics in arabic, 81 in English with an average of 10 tokens per topic and without URLs.

French queries were extracted from the Vodkaster Micro Film Reviews [5]. Vodkaster is a French social network about films. Users can post and share micro reviews in French about movies as they watch them. There were 233 topics in French with an average of 22 words per topic.

Spanish queries are a representative sample of sentences dealing with festivals from the Mexican newspaper *La jornada*[5]. We considered all the sentences from the newspaper mentioning a festival and extracted a random sample from this pool. These were well formed sentences that were easy to analyze but much harder to contextualize. There were 142 topics in Spanish with an average of 25 words per topic.

[5] `http://www.jornada.unam.mx`

2.4 Topics for Timeline Illustration

The goal of this task is to retrieve all relevant tweets dedicated to each event of a festival according to the program provided. In this case, we were looking at a kind of "total recall" retrieval based on the initial artists' names and the names, dates, and times of shows.

For this task, we focused on 4 festivals. Two French Music festivals, one French theater festival and one Great Britain theater festival:

- Vielles Charrues (2015),
- Transmusicales (2015),
- Avignon (2016),
- Edinburgh (2016).

Each topic was related to one cultural event. In our terminology, one event is one occurrence of a show (theater, music, ...). Several occurrences of the same show correspond then to several events (e.g. plays can be presented several times during theater festivals). More precisely, one topic is described by: one ID, one festival name, one title, one artist (or band) name, one time slot (date/time begin and end), and one venue location.

Participants were required to use the full dataset to conduct their experiments.

The runs were expected to respect the classical TREC top files format. Only the top 1000 results for each query run must be given. Each retrieved document is identified using its tweet ID. The evaluation is achieved on a subset of the full set of topics according to the richness of the results obtained. The official evaluation measures were interpolated at a precision of 1% and recall values at 5%, 10%, 25%, 50% and 100% .

3 Baselines

3.1 Microblog Search Task

A language model index powered by Indri and accessible through a web API has been provided. To deal with reposts, there was one document grouping all the users posts including his/her reposts. Each document has an XML structure (cf. Fig 1). Fig. 2 gives an example of such XML document.

This XML structure allows for one to work with complex queries like:

```
\# combine[m](
  Instagram.c es.l  \# 1(2016 05).d conduccin
  \# syn(pregoneros pregonero) \# syn(festivales festival))
```

This query will look for microblogs ([m]) posted from Instagram (.c) using Spanish locale (.l) in May 2016 (.d) dealing with pregonero(s) and festival(es).

For each set of queries two sets of queries have been generated, one retrieved authors with all their posts, the other focused on the posts themselves. For

```
<!ELEMENT xml (f, m)+>
<!ELEMENT f ($\#$ user\_id)>
<!ELEMENT m (i, u, l, c d, t)>
<!ELEMENT i ($\#$ microblog\_id)>
<!ELEMENT u ($\#$ user)>
<!ELEMENT l ($\#$ ISO\_language\_code)>
<!ELEMENT c ($\#$ client>
<!ELEMENT d ($\#$ date)>
<!ELEMENT t ($\#$ PCDATA)>
```

Fig. 1. XML DTD for microblog search

```
<xml><f>20666489</f>
 <m><i>727389569688178688</i>
  <u>soulsurvivornl</u>
  <l>en</l>
  <c>Twitter for iPhone</c>
  <d>2016-05-03</d>
  <t>RT @ndnl: Dit weekend begon het Soul Surivor Festival.</t>
 </m>
 <m><i>727944506507669504</i>
  <u>soulsurvivornl</u>
  <l>en</l>
  <c>Facebook</c>
  <d>2016-05-04</d>
  <t>Last van een festival-hangover?</t>
 </m>
</xml>
```

Fig. 2. An example of document for microblog search

English, Spanish and French no preprocessing or stop word list was applied. This resulted in long queries that were long to process, especially in the case of Focus retrieval. For Arabic, a stop word list was applied which improved efficiency. Table 1 provides the statistics about authors and microblogs retrieved using this baseline index powered by Indri over plain bag of words queries with language model and default Dirichlet model. Average numbers per queries are indicated into parenthesis.

Arabic and English topics are tweets about festivals during the Arabic spring period, although they are comparable, English topics cover a larger number of reposts and a wider range of languages. Arabic tweets are also posted by a reduce number of authors. French topics led to an extraction of an even greater number of noisy reposts than English. This is due to the fact that the majority are micro critics about films and part of them refer to the Cannes Festival which generates a massive number of tweets in French. Finally, it is the Spanish corpus that is

Topics	Queries	Microblogs	Reposts	Authors	Languages
Arabic	71	5,685 (80)	2,463 (34)	2,472 (35)	17 (0.24)
English	81	14,133 (175)	5,402 (67)	3,566 (44)	32 (0.40)
French	75	38,680 (516)	10,114 (134)	3,184 (43)	16 (0.21)
Spanish	114	51,856 (455)	10,984 (9)	4,811 (42)	16 (0.14)

Table 1. Microblog search baseline statistics

composed of sentences from news about festivals that appears to be the most specific since it encompasses a reduced number of different languages.

3.2 Timeline illustration task

The timeline illustration task provided a baseline based on the Terrier system [6]. The microblogs indexed were filtered: the filtering is based on the tweets' timestamp (which corresponds to the dates of the festivals), and text matching patterns (location or festival name for instance). The subset obtained consists of 243,643 tweets. We chose to keep the entire text of the initial tweets: we removed the '@' and '#' characters, and used a classical stoplisting process and Porter stemmer.

The content-based retrieval uses the BM25 model with the default parameters (stoplist, Porter stemming, $b = 0.75$).

4 Participant approaches and evaluation

For Multilingual Microblog Search, we applied the same methodology based on textual references instead of document qrels. Seven trilingual annotators (whom all together were fluent in 13 different languages: Arabic, Hebrew, Euskadi, Catalan, Mandarin Chinese, English, French, German, Italian, Portuguese, Russian, Spanish and Turkish) produced an initial textual reference. This reference was extended to Korean, Japanese and Persian based on Google translate. However this automatic extension appeared to be noisy and had to be dropped out from the reference. Only results in one of the assessors language could then be evaluated. Final textual references to evaluate microblog search run informativity will be presented at CLEF 2017 CLEF lab sessions in Dublin. Informativity is defined and computed based on [7].

For *Timeline Illustration* it was anticipated that re-tweets would be excluded from the pools. But the fact that it was a recall-oriented task led participants to return all re-tweets. Excluding re-tweets would have disqualified recall oriented runs that missed one original tweet. Moreover, it emerged during the evaluation that re-tweets are often more interesting than original ones. Indeed, original tweets are often posted by festival organizers, meanwhile reposts by individuals are more informative about attendees' participation in the festival.

Therefore, building a set of document qrels for time-line illustration was a two step process.

Firstly, tweet relevance on original tweets from baselines (each participant was asked to provide a baseline) was assessed on a 3-level scale:

- Not relevant: the tweet is not related to the topic,
- Partially relevant: the tweet is somehow related to the topic (e.g. the tweet is related to the artist, song, play but not to the event, or is related to a similar event with no possible way to check if they are the same).
- Relevant: the tweet is related to the event.

Secondly, the qrels were expanded to any microblog containing the text of a tweet previously assessed as relevant. In this manner, the qrels were expanded to all reposts. Participant runs were then ranked using the TRECEVAL program provided by NIST TREC[6]. All measures were provided since they lead to different rankings.

5 teams participated in the multilingual microblog search but none managed to process the four sets of queries. All teams were able to process the English set. Three of them manged to process French queries, one also processed Arabic queries and another, Spanish queries. Building reliable multilingual stop word lists was a major issue and required linguistic expertise. 4 teams participated in the timeline illustrations task but only one outperformed the BM25 baseline. The main issue was to identify microblogs related to one of the four festivals chosen by organizers. This selection couldn't be solely based on festival names since some relevant microblogs didn't include the festival hashtag. Nor could it be based on the dates of the festival since microblogs about videos posted by festivals later on after the event were considered as relevant.

The most effective approaches have been:

- MIcroblog Search: LIPAH based on LDA query reformulation for Language Model;
- Timeline Illustration: IITH using BM25 and DRF based on the artist's name, the festival name and the top hashtags of each of the events' features.

5 Conclusion

Dealing with a massive multilingual multicultural corpus of microblogs reveals the limits of both statistical and linguistic approaches. Raw utf8 text needs to be indexed without chunking. Synonyms and ambiguous terms over multiple languages have to be managed at query level. This requires positional indexes, however the usage of utf8 encoding makes them slow. It also requires linguistic resources for each language or for specific cultural events. Therefore, language and festival recognition appeared to be the key points of MC2 CLEF 2017's official tasks.

[6] http://trec.nist.gov/trec_eval/

The CLEF 2017 MC2 also expanded from a regular IR evaluation task to a task search. Almost all participants used the data and infrastructure to deal with topics that were beyond the initial scope of the lab. For example:

- the LSIS-EJCAM team used this data to analyze the role of social media in propagating controversies,
- the *My Local Influence* and U3ICM team experimented using sociological needs to characterize profiles and contents for Microblog search.

Researchers interested in using MC2 Lab data and infrastructure, but who didn't participate to the 2017 edition, can apply untill march 2019 to get access to the data and baseline system for their academic institution by contacting `eric.sanjuan@talne.eu`. Once the application accepted, they will get a personal private login to access lab resources for research purposes.

References

1. Goeuriot, L., Mothe, J., Mulhem, P., Murtagh, F., SanJuan, E.: Overview of the CLEF 2016 Cultural Micro-blog Contextualization Workshop. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings. (2016) 371–378
2. Murtagh, F.: 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. (2016) 1207–1225
3. Balog, K., Cappellato, L., Ferro, N., Macdonald, C., eds.: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016. Volume 1609 of CEUR Workshop Proceedings., CEUR-WS.org (2016)
4. Features Extraction To Improve Comparable Tweet corpora Building, JADT (2016)
5. Cossu, J.V., Gaillard, J., Juan-Manuel, T.M., El Bèze, M.: Contextualisation de messages courts :l'importance des métadonnées. In: EGC'2013 13e Conférence Francophone sur l'Extraction et la Gestion des connaissances, Toulouse, France (January 2013)
6. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: SIGIR'06 Workshop on Open Source Information Retrieval, (OSIR'06). (2006)
7. Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., Tannier, X.: INEX Tweet Contextualization task: Evaluation, results and lesson learned. Information Processing Management **52**(5) (2016) 801–819