CLEF 2017 Microblog Cultural Contextualization Lab Overview

Liana Ermakova[1], Lorraine Goeuriot[3], Josiane Mothe[2], Philippe Mulhem[3], Jian-Yun Nie[4], and Eric SanJuan[5]

[1] LISIS (UPEM, INRA, ESIEE, CNRS), Université de Lorraine, France
[2] IRIT, UMR5505 CNRS, ESPE, Université de Toulouse, France
[3] LIG, Université de Grenoble, France
[4] RALI, Université de Montréal, Québec, Canada
[5] LIA, Université d'Avignon, France
liana.ermakova@univ-lorraine.fr, josiane.mothe@irit.fr,
eric.sanjuan@univ-avignon.fr

Abstract. MC2 CLEF 2017 lab deals with how cultural context of a microblog affects its social impact at large. This involves microblog search, classification, filtering, language recognition, localization, entity extraction, linking open data, and summarization. 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 resulting mircroblog stream and related URLs is appropriate to experiment advanced social media search and mining methods. A collection of 70,000,000 microblogs over 18 months dealing with cultural events in all languages has been released to test multilingual content analysis and microblog search. For content analysis topics were in any language and results were expected in four languages: English, Spanish, French, and Portuguese. For microblog search topics were in four languages: Arabic, English, French and Spanish, and results were expected in any language.

1 Introduction: from Microblog to Cultural Contextualization

Microblog Contextualization was introduced as a Question Answering task of INEX 2011 [1]. The main idea was to help Twitter users to understand a tweet by providing some context associated to it. It has evolved in a Focus IR task over WikiPedia [2].

The CLEF 2016 Cultural Microblg Contextualization Workshop considered specific cultural twitter feeds [3]. 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 dealt with how cultural context of a microblog affects its social impact at large [4]. This involved microblog search, classification, filtering, language recognition, localization, entity extraction, linking open data, and summarization.

Regular Lab participants had access to the private massive multilingual microblog stream of *The Festival Galleries* project[6]. Festivals have a large presence on social media. The resulting microblog stream and related URLs were appropriate to experiment advanced social media search and mining methods.

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 scenari lead to three tasks lab participants could answer to:

- Content analysis
- Microblog search
- Timeline illustration

These tasks are detailed in Section 3 to 5. Section 2 depicts the data used in the various tasks and Section 6 describes the evaluation. Finally Section 7 draws some conclusions.

2 Data

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

It allows researchers in IR (Information Retrieval) and NLP (Natural Language Processing) to experiment a broad variety of multilingual microblog search techniques (Wikipedia entity search, automatic summarization, language identification, text localization, etc.).

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[7].

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 Potuguses for Wikipedia search.

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

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

2.1 Microblog Collection

The document collection is an updated extension of the microblog stream presented at the CLEF 2016 workshop [5](see also [6]).

It was provided to registered participants by ANR GAFES project[8]. 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" since June 2015. These microblogs were collected using private archive services based on streaming API. The average 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 (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 privacy agreement. The MC2 lab provides 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.

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 Wikipedia XML Corpus for Summary Generation

Wikipedia is under Creative Commons license, and its content can be used to contextualize tweets or to build complex queries referring to Wikipedia entities.

We have extracted an average of 10 million XML documents from Wikipedia 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 content of the main Wikipedia pages: title, abstract, section, and subsections as well as Wikipedia internal links. Other content such as images, footnotes and external links is stripped out in order to obtain a corpus easier to process using standard NLP tools.

By comparing contents over the years, it was possible to detect long term trends.

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

3 Content Analysis

The content analysis task has been inspired by [7–9].

Given a stream of microblogs, the task consists in:

- filtering microblogs dealing with festivals;
- language(s) identification;
- event localization;
- author categorization (official account, participant, follower or scam);
- Wikipedia entity recognition and translation in four target languages: English, Spanish, Portuguese, and French.
- automatic summarization of linked Wikipedia pages in the four target languages.

Each item has been evaluated independently, however, language identification could impact Wikipedia linking and the resulting summaries. The filtering and author categorization subtasks were inspired by the filtering and priority tasks at RepLab 2014[10].

Opinion mining was not initially considered, however two participants did apply binary opinion classifiers and detection of controversies on the provided corpus. It appears that like for the Reputation task in RepLab[10], microblog content needs to be contextualized and expanded to accurately identify opinions, especially for cultural events where nouns and adjectives considered as negative can reflect highly positive opinions for specific communities.

Language(s) identification is challenging over short contents that tend to mix several languages. Festival names over tweets often appear in English but the rest of the content can be in any other languages. Moreover festival attendees tend to add terms from various dialects to highlight the local context.

Event localization requires external resources. For large festivals WikiPedia often contains the information and it can be retrieved based on state-of-the-art QA approaches. However for small events it is necessary to query the public web or social networks. In this lab the 18 months feed of tweets about festivals allows to search for microblogs by festival organizers about venues.

The two subtasks Wikipedia Entity Recognition and Automatic Summarization refer to previous experiments around Tweet Contextualization[2]. Most efficient methods proceed in two steps: 1) retrieve most relevant Wikipedia pages, 2) propose a multidocument summary of them. WikiFying tweets is complex due to the lexical gap between tweets and Wikipedia pages. Extracting summaries looked easier by aggregating sentences from pages but ensuring and evaluating readability is an issue, specially on languages with less resources than English.

4 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 [11]. We considered the content of all the tweets dealing with festivals during the Arab Spring period and the task consisted in searching for traces of these festivals or artists in the lab corpus two years after. The use case was to follow up artists involved in the Arab spring festivals two or three years later. Indeed, most of the festivals before the Arab spring were relying on tourism and have been stopped after 2014, so they don't appear as festivals in the MC2 corpus of microblogs which spans from 2015 to 2016. However, most famous artists and film makers involved in Arabic festival have been invited in European and Canadian Festivals in 2015. The microblog search task in English and Arabic consisted in retrieving those indirectly related microblogs.

The difficulty of this task relies in generating a query based on a microblog content. Tweets are too short to apply standard name entity extraction algorithms but they are too long to be considered as queries for an IR system without a robust preprocessing that removes empty words and keeps only informative terms. Corpus and queries being encoded in utf8, systems can handle multiple languages, however language stop word lists and lemmatizers are specific to each language. For Arabic another difficulty appeared with dialects. Dealing with them require extra linguistic resources.

French queries were extracted from the VodKaster Micro Film Reviews [12]. VodKaster is a French social network about films. Users can post and share micro reviews in French about movies as they watch them. They can score films but also reviews written by others. We extracted as queries all micro reviews dealing with festivals during the period of the lab corpus (2015-2016). Most of them where posted from phones by festival attendees. Film micro reviews are easier to process than tweets because most of them contain well formed sentences. Searching for related tweets could be improved by considering the date of the micro review to identify the film festival. However microblogs about other festivals mentioning the same films or actors were also considered as relevant.

Spanish queries are a representative sample of sentences dealing with festivals from the Mexican newspaper *La jornada*[9]. We considered all the sentences from the newspaper mentioning a festival and extracted a random sample from this pool. These are well formed sentences easy to analyze but much harder to contextualize. Extracting queries about these sentences often requires to find the source article and neighboring sentences. The use case was that a reader highlights a sentence while reading the newspaper and sees related microblogs. Like for Arabic, it was also necessary to deal with the multiple variants of Spanish language.

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

This XML structure permits to work with complex queries like:

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

```
<!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

```
\# 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).

5 Timeline Illustration

The goal of this task was to retrieve all relevant tweets dedicated to each event of a festival, according to the program provided. We were really looking here at a kind of ”total recall” retrieval, based on initial artists’ names and 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 timeslot (date/time begin and end), and one venue location.

An excerpt from the topic list is:

```
<topic>
  <id>5</id>
  <title></title>
  <artist>Klangstof</artist>
  <festival>transmusicales</festival>
  <startdate>04/12/16-17:45</startdate>
  <enddate>04/12/16-18:30</enddate>
  <venue>UBU</venue>
</topic>
```

The id was an integer ranging from 1 to 664. We see from the excerpt above that, for a live music show without any specific title, the title field was empty. The artist name was a single artist, a list of artist names, an artistic company name or orchestra name, as they appear in the official programs of the festivals.

The festival labels were:

– *charrues* for Vielles Charrues 2015;
– *transmusicales* for Transmusicales 2015;
– *avignon* for Avignon 2016;
– *edinburgh* for Edinburgh 2016.
– For the date/time fields, the format is : *DD/MM/YY-HH:MM*.
– The *venue* is a string corresponding to the name of the location, given by the official programs.

If the start or end time is unknown, they are replaced with: *DD/MM/YY-xx:xx*. If the day is unknown, the date format is the following: *-HH:MM* (day is omitted).

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 precision at 1% and recall values at 5, 10, 25, 50 and 100 documents.

6 Results

Overall, 53 teams involving 72 individuals registered to the lab. Among them 12 teams from Brazil (1), France (4), Tunisia (2), Mexico (1), India (1), and Mongolia (1) submitted 42 valid runs. 11 teams submitted a working note to CLEF 2017.

6.1 Evaluation

For *Content Analysis*, q-rels based on pooling from participant submissions appeared to be unstable due to the variety of languages involved in this task and the variety of participants' approaches that could be efficient on different subsets of languages. Results were to be provided in four different languages; however, the submitted runs were extremely multilingual. Reaching stable q-rels by pooling would have required to stratify by language on input and output which leads to very sparse matrices of results. All results had to be extracted from the four WikiPedias in English, Spanish, French and Portuguese to have a common document ground base, but even the WikiPedia appeared to be highly redundant with multiple pages with similar content referring to the same cultural event from different perspectives.

By contrast, extensive textual references by organizers manually built on a reduced random subset of topics using the one powered by Indri provided to participants[10] and the aggregator DuckDuckGo[11] runs and on the four targeted different languages appeared to be more stable to rank runs based on token overlapping following the same methodology as in [2].

For Multilingual Microblog Search, we applied the same methodology based on textual references instead of document q-rels. Seven trilingual annotators fluently speaking 13 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 Corean, 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.

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

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

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

[10] http://tc.talne.eu

[11] https://ducduckgo.com

- 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 q-rels were expanded to any microblog containing the text of one previously assessed as relevant. this way, the q-rels were expanded to all reposts. Participant runs have then be ranked using treceval program provided by NIST TREC[12]. All measures have been provided since they lead to different rankings.

6.2 Participant Approaches

Among the 12 active participants, 6 teams participated to Content Analysis but only one (LIA) managed to produce multilingual contextual summaries in four languages on all queries (microblogs without urls mixing more than 30 languages) and only one managed to deal with the localization task (Syllabs). 5 teams participated to the multilingual microblog search but none managed to process the four sets of queries. All did process the English set, three could process French queries, one Arabic queries and one Spanish queries. Building realable multilingual stop word lists was a major issue and required linguistic expertise. 4 teams participated to 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 only based on festival names since some relevant microblogs didn't include the festival hashtag, neither on the dates since microblogs about videos posted by festivals later on after the event were considered as relevant.

The most effective approaches have been:

- Language Identification: Syllabs enterprise based on linguistic resources on Latin languages.
- Entity Extraction: FELTS system based on string matching over very large lexicons.
- MultiLingual Contextualization: LIA team based on automatic multidocument summarization using Deep Learning.
- MIcroblog Search: LIPAH based on LDA query reformulation for Language Model.
- Timeline Illustration: IITH using BM25 and DRF based on artist name, festival name, top hashtags of each event features.

7 Conclusion

Dealing with a massive multilingual multicultural corpus of microblogs reveals the limits of both statistical and linguistic approaches. Raw utf8 text needs to

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

be indexed without chunking. Synonyms and ambiguous terms over multiple languages have to be managed at query level. This requires positional index but 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 official tasks.

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 problematics 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 ISAMM team experimented opinion polarity detection in Twitter data combining sequence mining and topic modeling.
- 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. SanJuan, E., Moriceau, V., Tannier, X., Bellot, P., Mothe, J.: Overview of the INEX 2011 question answering track (qa@inex). In Geva, S., Kamps, J., Schenkel, R., eds.: Focused Retrieval of Content and Structure, 10th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2011, Saarbrücken, Germany, December 12-14, 2011, Revised Selected Papers. Volume 7424 of Lecture Notes in Computer Science., Springer (2011) 188–206
2. 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
3. Ermakova, L., Goeuriot, L., Mothe, J., Mulhem, P., Nie, J., SanJuan, E.: Cultural micro-blog Contextualization 2016 Workshop Overview: data and pilot tasks. In: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016. (2016) 1197–1200
4. 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
5. 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
6. 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)

7. Ngoc, H.T.B., Mothe, J.: Building a knowledge base using microblogs: the case of cultural microblog contextualization collection. In: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016. (2016) 1226–1237
8. Scohy, C., Chaham, Y.R., Déjean, S., Mothe, J.: Tweet data mining : the cultural microblog contextualization data set. In: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016. (2016) 1246–1259
9. Pontes, E.L., Torres-Moreno, J., Huet, S., Linhares, A.C.: Tweet contextualization using continuous space vectors: Automatic summarization of cultural documents. In: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016. (2016) 1238–1245
10. Amigó, E., Carrillo de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M., Spina, D.: Overview of replab 2014: Author profiling and reputation dimensions for online reputation management. In Kanoulas, E., Lupu, M., Clough, P.D., Sanderson, M., Hall, M.M., Hanbury, A., Toms, E.G., eds.: Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings. Volume 8685 of Lecture Notes in Computer Science., Springer (2014) 307–322
11. Features Extraction To Improve Comparable Tweet corpora Building, JADT (2016)
12. 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)