Cultural micro-blog Contextualization 2016 Workshop Overview: data and pilot tasks

Liana Ermakova, Lorraine Goeuriot, Josiane Mothe, Philippe Mulhem, Jian-Yun Nie, Eric Sanjuan

Cultural micro-blog Contextualization 2016 Workshop Overview: data and pilot tasks

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

[1] IRIT, UMR5505 CNRS, ESPE, Université de Toulouse, France
[2] LIG, Université de Grenoble, France
[3] RALI, Université de Montréal, Québec, Canada
[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

Festivals are gaining an increasing success and some of them, such as Cannes, Edinburgh or Avignon, produce a significant activity on the Web. This WorkShop proposes to develop access methods for the contents generated during and around the festivals to better understand the festival practices [1]. The underlying scientific problems concern both IR and humanities. This WorkShop focuses on two axes: the contextualization of the data collected on the Web and the search of content captured or produced by internet users [2].

For its first edition, it gives access for registered participants to a massive collection of microblogs and related urls[5][3].

In this overview we report the main tentative tasks that have been suggested to be discussed and experimented during the workshop. First §2 we introduce the contextualization task based on the Wikipedia. Then in §3 we discuss a possible microblog search task over a long time period. In §4 a timeline search task over several festivals is proposed.

2 Microblog Contextualization based on Wikipedia

This initial task aimed at generating a short summary providing the background information of a tweet to help a user to understand it following[4]. Given a microblog announcing some cultural event, participants have to provide a short

[5] All resources are available online:http://cmc.talne.eu

summary extracted from Wikipedia that provides -extensive -background about this event. The summary must contain some context information about the event in order to help answering questions of the form "what is this tweet about?" using a recent cleaned dump of Wikipedia. The context should take the form of a readable summary, not exceeding 500 words, composed of passages from the provided Wikipedia corpus.

Any open access resource could be used in addition to the data provided to participants subject to describing it and providing a valid URL.

2.1 Datasets

A restricted set of public micro-blogs in English were collected from a set of public on Twitter, all related to the keyword festival. The micro-blogs are in UTF8 csv format with various fields. In this task, the tweets do not contain URL. The other suggested tasks would use additional information.

Unlike tweets, Wikipedia is under Creative Common license, and its contents can be used to contextualize tweets or to build complex queries referring to Wikipedia entities. 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 such 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

2.2 Evaluation

Following [5] the summaries would be evaluated according to informativeness and readability.

Informativeness is the way they overlap with relevant passages (number of them, vocabulary and bi-grams included or missing). For each tweet, all passages from all participants will be merged and displayed to the assessor in alphabetical order. Therefore, each passages informativeness will be evaluated independently from others, even in the same summary. Assessors will only have to provide a binary judgment on whether the passage is worth appearing in a summary on the topic, or not.

Readability can only be accurately assessed by humans. A small panel of scholars in humanities will have to evaluate readability for a pool of summaries using on an online web interface. Each summary consists of a set of passages and for each passage, assessors will have to tick four kinds of check boxes:

- Syntax (S): tick the box if the passage contains a syntactic problem (bad segmentation for example),
- Anaphora (A): tick the box if the passage contains an unsolved anaphora,

– Redundancy (R): tick the box if the passage contains redundant information, i.e. information that has already been given in a previous passage,
– Trash (T): tick the box if the passage does not make any sense in its context (i.e. after reading the previous passages). These passages must then be considered as trashed, and the readability of following passages must be assessed as if these passages were not present.

3 Cultural MicroBlog Search based on WikiPedia entities

Given a cultural entity as a set of WikiPedia pages (typically a set of places to visit, artists to see on stage, festivals of interest etc.), the proposed task would be to provide a double extensive summary of relevant microblogs from insiders and outsiders. This task will involved two sub-tasks:
Task 2a: Retrieval of relevant microblogs for an entity (described by its wikipedia page)
Task 2b: Summarization of the most informative tweets (and comparison to manually built summaries)

3.1 Micro-blog collection

The document collection is provided to registered participants by ANR GAFES[6] project and consists in a pool of more than 50M unique micro-blogs from different sources with their meta-information as well as ground truth for the evaluation.

The micro-blog collection contains among other sources, all public posts on Twitter using the keyword festival since June 2015. These micro-blogs are collected using private archive service 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.

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 provided 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.

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*,

[6] `http://anr-gafes.univ-avignon.fr/demo.html`
[7] `https://dev.twitter.com/streaming/public`

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.

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

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%)

4 TimeLine illustration based on Microblogs

The goal of this task is to link the events of a given festival program to related microblog posts. Such information is useful for attendees of festivals, for people that are interested in knowing what happens in a festival, and for organizers to get feedback[6].

Microblog posts are provided with their timestamps, which are crucial as a basis for the requested linking. However, such timestamps must be use with care: they do not necessarily give accurate enough information (for instance in the case of parallel sessions), or might even generate perturbations (microblogs about one event may be posted before, during, or after the actual event).

Participants would be required to provide, for each event of the program, the 10 best tweets based on their relevance and diversity. In this task, diversity is a must because retrieving several times the same post is not beneficial in our case.

4.1 Data

Participants for this task would use a subset of the microblogs collection, matching the months the targeted festivals were organized at (July and December 2015).

In its tentative form, Festival programmes are provided in French: Two French music festivals have been selected: the festival des vieilles charrues and the transmusicales de Rennes. The timelines provided are selected subset of each festival program: the organizers selected a subset of the whole festival program (for each stage and time, list of artists playing).

The participants would be free to use any additional data to provide results: social (popularity,) or not (knowledge bases,); it should be described in the related paper and specified when submitting the runs.

4.2 Evaluation

The evaluation would be carried out on selected parts of the program chosen by the task organizers depending on the number of relevant tweets per event. The evaluation measures planned would be recall/precision based. Several types of runs will be proposed: time-only, content-only, time&content.

5 Conclusion

Cultural Microblog Contextualization CLEF 2016 WorkShop aims at developing processing methods for social media mining. Our focus is around festivals that are organized or that have a large presence on social media. Micro-blogs linked to an event make a dense, rich but very noisy corpus. Content is often imprecise, duplicate or non-informative.

We also envisage to provide an extra corpus of Images related to cultural festivals in the world. This access would allow researchers in IR and NLP to experiment a broad variety of multilingual microblog search techniques (WikiPedia entity search, and automatic summarization). Extensive textual references would be provided by scholars in humanities involved in the ANR GAFES project.

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. : INEX Tweet Contextualization task: Evaluation, results and lesson learned in Information Processing & Management, in press, 2016.
6. 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.