CLEF 2017 Microblog Cultural Contextualization Lab Overview

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

Avignon, Grenoble, Lorraine and Montréal Universities

https://mc2.talne.eu

September 11, 2017

Summary

1 Introduction

2 Data Insights

3 Queries

4 Results

5 Lab Sessions

Overview

Objective

Help Twitter users to understand a tweet by providing some context associated to it. The MC2 CLEF 2017 lab dealt with how cultural context of a microblog affects its social impact at large.

Tasks

- Content analysis
- Microblog search
- Timeline illustration

Data

Multilingual microblog stream of *The Festival Galleries* (ANR-14-CE24-0022).

Background: from Microblog to Cultural Contextualization

- Microblog Contextualization was introduced as a Question Answering task of INEX 2011. It has evolved in a Focus IR task over WikiPedia.
- CLEF 2016 **Cultural Microblog Contextualization** Workshop considered specific cultural twitter feeds.
- Restricted context implicit localization and language identification appeared to be important issues.
- The MC2 CLEF 2017 lab has been centered on Cultural Contextualization based on Microblog feeds.

Usage scenarios

Content analysis & Search motivation

An insider attendee who receives a microblog about the cultural event which he will participate in will need context to understand it since microblogs often contain implicit information.

Time Line Illustration

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.

18 months festival Microblog stream

Content

Public posts on Twitter using the keywords “festival” and some cities like Cannes and Edimbourgh between June 2015 and Novembre 2016 (more than 70 millions among which 1/2 are reposts).

URLs

66% of the collected microblog posts contain compressed URLs leading to 11,000,000 distinct uncompressed URLs from only 700,000 distinct domains.

Social Media User Oriented XML collection

Example

User Id: soulsurvivornl

Post id: 727389569688178688

- Atts: Lang:en, Client:Twitter for iPhone
- Date: 2016-05-03
- Text: @ndnl: Dit weekend begon het Soul Surivor Festival.

Post id: 727944506507669504

- Atts: Lang:en, Client:Facebook
- Date: 2016-05-04
- Text: Last van een festival-hangover?

Challenges

- Filtering microblogs really dealing with festivals;
- Language(s) identification;
- Event localization;
- Author categorization: official account, participant, follower or scam;
- Generating complex queries combining all features: users, clients, dates languages and synonyms.

Content Analysis

Wikifying MicroBlogs

Complex task due to the lexical gap between tweets and Wikipedia pages.

800 microblogs in all languages without URLs

Only text content was considered as query, no metadata.

Ressource

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).

MicroBlog Search

Arabic and English

1000 Microblogs posted within Arabic Spring (2010) about Festivals: the task consisted in following up festivals, movies and artists 5 years later.

French

French social network VodKaster about films. Users share micro reviews about movies as they watch them. Queries were all micro reviews dealing with some festival during the period 2015-2016.

Spanish

Sample of sentences dealing with festivals from the Mexican newspaper *La jornada*.

Time Line Illustration

Focus on 4 festivals

Two French Music festivals, one French theater festival and one Great Britain theater festival. Each topic was related to one cultural event (theater, music, ...).

- Vielles Charrues 2015;
- Transmusicales 2015;
- Avignon 2016;
- Edinburgh 2016.

12 active participants

Content analysis

6 teams on **language recognition** and **entity extraction** but only one on multilingual contextual summaries (LIA) and on localization task (Syllabs).

Microblog Search

5 teams. All did process the English set, three could process French queries, one Arabic queries and one Spanish queries.

Time Line Illustration

4 teams. Only one outperformed the BM25 baseline, some relevant microblogs didn't include the festival hashtag or were about videos posted by festivals later on after the event.

Most effective approaches

- 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.

Program Wednesday 13th September

Labs 4, 13:45-15:45, CMC, room 5039

Content analysis and Microblog Search: Detailed overview, 2 participant presentations and discussion towards *Cultural Image Queries over Social Media*.

Labs 5, 16:45-18:15, CMC, room 5039

Time Line Illustration: Detailed overview, evaluation material release and discussion towards *dealing with Language Dialects and Varieties in Mining and Search over Cultural Social Media posts*.