CLEF 2017 Microblog Cultural Contextualization Content Analysis Task Overview

Liana Ermakova[1], Josiane Mothe[2], and Eric SanJuan[3]

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

Abstract. The MC2 CLEF 2017 Content Analysis task deals with classification, filtering, language recognition, localization, entity extraction, linking open data, and summarization. Festivals have a large presence on social media. The resulting microblog stream and related URLs are appropriate to experiment on advanced social media search and mining methods. For content analysis, topics were in any language and results were expected in four languages: English, Spanish, French, and Portuguese.

Keywords: Information retrieval, Tweet contextualization, Microblog analysis, CLEF evaluation forum

1 Introduction

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 (Information Retrieval) task over Wikipedia [2].

The CLEF 2016 Cultural Microblog Contextualization Workshop considered specific cultural Twitter feeds [3]. In this 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.

Given a stream of microblogs, the task consists in:

1. filtering microblogs dealing with festivals;
2. language identification;
3. event localization;
4. author categorization (official account, participant, follower or scam);

5. Wikipedia entity recognition and translation into four target languages: English, Spanish, Portuguese, and French;
6. automatic summarization of linked Wikipedia pages in the four target languages.

Each item is evaluated independently, however language identification could impact Wikipedia linking and the resulting summaries.

In this paper, Section 2 depicts the data used. Section 3 describes the baselines and state-of-the-art system. Section 4 describes participant approaches. Finally, Section 5 draws some conclusions.

2 Data

The MC2 Content Analysis 2017 task provides a set of 1,100 microblogs in 20 languages to be mapped into textual extracts from English, Spanish, French, and Portuguese Wikipedia.

2.1 Wikipedia XML Corpus

Wikipedia is under a 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). The corpus and tools to process them are available on the Tweet Contextualization website[4].

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 that is easier to process using standard NLP (Natural Language Processing) tools.

2.2 Queries

The query collection is a pool of 1,100 microblogs extracted from the microblog stream presented at the CLEF 2016 workshop [5](see also [6]). These microblogs have more than 80 characters, they do not contain URLs and are written in more than 20 different languages. The main languages are: en (60%), es (14%), fr (5%), pt (4%), it (2%). Other languages are: ja, de, nl, tr, id, ca, eu, zh, ru, sr, pl, ko, fi and ar.

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

3 Baselines and State-of-the-art System

For each Wikipedia we provided an XML retrieval system powered by Indri, a Perl API for the XML retrieval system using standard LWP (short for "Library for WWW in Perl"), the corpus in a single XML file (gzip compression), and the corpus split into 1,000 folders, one file per page (tgz archive). However these baselines did not provide text segmentation into sentences nor an automatic summarization tool. They only allowed to retrieve XML elements based on nested language models.

Based on these resources the available baselines are:

- filtering: based on the word "festival";
- language: based on Twitter local code;
- entity extraction: top ranked Wikipedia page titles based on a document language model;
- summarization: based on Wikipedia page abstracts.

A state-of-the-art contextualization system has also been used to generate a complete run available for active participants. This reference system is based on the Terrier platform[5]. Wikipedia pages in English, French, Spanish, and Portuguese were stemmed by the SnowBall stemmer. The pages retrieved by the InL2 model with Bo1 query expansion technique were interpreted as a baseline for the entity recognition subtask. Then, documents were parsed by Stanford CoreNLP in order to perform sentence chunking and lemmatization [6]. For the automatic summarization subtask we used the following baselines:

- the first passage from the top-scored Wikipedia page;
- the cosine similarity between a tweet and a candidate sentence;
- word2vec similarity between a tweet and a candidate sentence [7];
- the system based on local context analysis presented at CLEF-2015 [8].

4 Participants Approaches

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 [9].

4.1 Filtering and Opinion Mining

One participant (LIA-FR) scored all microblogs by proximity with a festival topic [10]. Opinion mining was not initially considered, however one participant (ISAMM-TN) did apply binary opinion classifiers [11]. It appeared that microblog interestingness about festivals assessed by organizers mostly relies on neutral microblogs because they are easier to understand without context.

[5] http://terrier.org/
[6] https://stanfordnlp.github.io/CoreNLP/

4.2 Language Identification

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

Using linguistic resources for main languages as Syllabs-FR did, allow to reach the best precision scores [12]. However, based on statistical approaches, the LIA-FR identified 121 errors in microblog local information among the 1,100 [13]. After evaluation, it appeared that 90 among the 121 were true errors: 30% about en, 20% about pt, 16% about es, 10% about id. The rest of true errors were about it, de, sh, fr, nl, ceb, ca, and sv.

4.3 Event Localization

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 (Question Answering) approaches. However for small events it is necessary to query the public web or social networks. The Syllabs-FR team managed to localize festivals in France using public information [12].

4.4 Entity Recognition and Automatic Summarization

The two subtasks: Wikipedia Entity Recognition and Automatic Summarization refer to previous experiments around Tweet Contextualization[2]. The most efficient methods proceed in two steps: 1) retrieve the 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, however ensuring and evaluating readability is an issue, especially with languages that have less resources than English.

The FELTS system managed to identify all Wikipedia page titles that explicitly appear in the 1,100 microblogs for the four target languages [13]. Multiword titles are often unambiguous. Among the 1,100 queries, 818 of them contained explicit references to unambiguous Wikipedia pages in English, 536 in Spanish, 485 in French, and 459 in Portuguese. By considering the Wikipedia abstract of these pages, it was then possible to directly extract high quality summaries contextualizing almost half of the topics in the four target languages. This approach has bean scaled to process microblog streams in real time.

5 Conclusion

Dealing with a massive multilingual multicultural corpus of microblogs reveals the limits of both statistical and linguistic approaches. 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 the overall MC2 CLEF 2017 lab official tasks.

Researchers interested in using MC2 Lab data and infrastructure, but who did not participate to the 2017 edition, can apply until March 2019 to get access to the data and baseline system for their academic institution by contacting `eric.sanjuan@talne.eu`. Once the application is accepted, they will get a personal private login to gain access to 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. Mikolov, T., Deoras, A., Povey, D., Burget, L., Černocký, J.: Strategies for training large scale neural network language models. 2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings (2011) 196–201
8. Ermakova, L.: A Method for Short Message Contextualization: Experiments at CLEF/INEX. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 6th International Conference of the CLEF Association, CLEF'15, Toulouse, France, September 8-11, 2015, Proceedings. Springer International Publishing, Cham (2015) 352–363
9. 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
10. Linhares Pontes, E., Huet, S., Torres-Moreno, J.M., Carneiro Linhares, A.: . (2017)
11. Ouertatani, A., Gasmi, G., Latiri, C.: Opinion polarity detection in Twitter data combining sequence mining and topic modeling. (2017)
12. Hamon, O., Monnin, C., de Loupy, C.: Syllabs Team at CLEF MC2 Task 1: Content Analysis. (2017)
13. Jourlin, P.: Entity Recognition and Language Identification with FELTS. (2017)