CLEF MC2 2018 Lab Overview

Malek Hajjem[1,2], Jean Valére Cossu[3], Chiraz Latiri[2], and Eric SanJuan[1](✉)

[1] LIA, Avignon University, Avignon, France
{malek.hajjem,eric.sanjuan}@univ-avignon.fr
[2] LIPAH, Tunis Manar University, Tunis, Tunisia
chiraz.latiri@gnet.tn
[3] MyLI, My Local Influence, Marseille, France
jvcossu@gmail.com

Abstract. MC2 lab mainly focuses on developing processing methods and resources to mine the social media (SM) sphere surrounding cultural events such as festivals, music, books, movies and museums. Following previous editions (CMC 2016 and MC2 2017), the 2018 edition focused on argumentative mining and multilingual cross SM search. Public microblogs about cultural events like festivals are promotional announcements by organizers or artists, very few are personal and argumentative, the challenge is to find them before they eventually become viral. We report the main lessons learned from this 2018 CLEF task.

Keywords: Argumentation mining · Microblogs
Information Retrieval · Ranking

1 Introduction

Following previous editions, MC2 Lab 2018 was centered on multilingual culture mining and retrieval process over the large corpus of cultural microblogs [7] considered in the two previous editions [6,8]. Two main tasks were considered: cross language cultural microblog search and argumentation mining.

The initial challenge for 2018 was, given a short movie review on the French VodKaster[1] Social Media, find related microblogs in the MC2 corpus in four different target languages (French, English, Spanish and Portuguese). Indeed, browsing the VodKaster website, French readers get personal short comments about movies. Since similar posts can be found on twitter we decided to display to the reader a concise summary of microblogs related to the comment he/she is reading, considering bilingual and trilingual users that would read microblogs in other languages than French. In this user's context, personal and argumentative microblogs are expected to be more relevant than news or official announcements. Microblogs sharing similar arguments can be considered as highly relevant even though they are about different movies. From this initial task, came the idea of

[1] http://www.vodkaster.com/.

P. Bellot et al. (Eds.): CLEF 2018, LNCS 11018, pp. 302–308, 2018.
https://doi.org/10.1007/978-3-319-98932-7_27

a second one focusing on argument mining in a multilingual collection. It consisted in finding personal and argumentative microblogs in the corpus. Public posts about cultural events like festivals are mostly promotional announcements by organizers or artists. Personal argumentative microblogs about specific festivals provide real insights into public reception but both their variety and rarity make them difficult to seek. Therefore, argumentative mining captured most of participant efforts during this lab edition. The cold start scenario of finding them without any specific learning resource motivated the use of IR approaches based on language model or specialized linguistic resources.

The rest of this paper focus on this specific task. Related work is presented in Sect. 2. Section 3 is devoted to task thorough description an motivations. Data including a baseline run is fully described in Sect. 4. Result and participant approaches are reported in Sect. 5.

2 Related Work

Argumentation (or argument) mining is the automatic extraction of structured arguments from unstructured textual corpora [10]. This task represents a new problem in corpus-based text analysis that addresses the challenging task [13] of automatically identifying the justifications provided by opinion holders for their judgments. The initial research of argumentation mining has been proposed for legal documents, on-line debates, product reviews, political debates and newspaper articles, court cases, as well as in the dialogical domain [3,12,13].

As a result of the advent of social media platforms, argumentation mining for social media text and user generated content has been proposed [5,14]. The goal of argumentation mining with short and unstructured data is to improve our ability to process and infer meaning from social media text. In fact, this kind of data is characterized to be ambiguous by nature which makes it hard for a user to effectively understand what the opinion tweet is about. Generally, such tweets are indispensable to form a view about a new topic or make a decision based on users feedback. In such a case, expressed argument is all what we are looking for.

Regarding short texts, developed approaches for microblogs differ from techniques dedicated to other genres. These are usually longer, such as forums, product reviews, blogs and news. In fact high quality social media data sets annotated with argumentation structure are rare which affects the use of machine learning techniques. In this context we cite DART [4], a dataset to support the development of frameworks addressing the argument mining pipeline on Twitter.

This lack of resources and challenges to extract arguments from social media text could be explained by the fact that social media platforms such as comment boards on news portals, product review sites, or microblogs are less controlled communication environments where the communicative intention is not to engage in an argumentative discussion but rather to simply express an opinion on the subject matter [14]. To solve this issue, argumentation mining within social media text has to deal with several sets of features to capture the above mentioned characteristics for persuasive comment identification from user generated data. This was the case of [17] where authors propose and evaluate other features to rank comments for their persuasive scores, including textual information in the comments and social interaction related features.

3 Task

The proposed task is inspired from the field of focused retrieval. This later aims to provide users with direct access to relevant information in retrieved documents. For this task, a relevant information is expressed in the form of argument that supports or criticizes an event. So, we presume that the proposed method must perform:

1. a search process that focus on claims about a given topic out in a massive collection.
2. a ranking process that has a potential argumentative coming first.

Following such steps, a synthesis of many argument facets about a specific event is automatically constructed. Such an output could be treated more easily, on priority, by a festival organizer.

Argumentation mining is considered as an extension of the opinion mining issue from social network content. The main objective of this field is to automatically identify reason-conclusion structures that can lead to model social web user's positions about a service, product or event expressed through social media platforms. As explored in [10] most argumentation mining approaches have tackled the challenging task of extracting arguments based on machine learning methods. However, in case of argumentation mining from social media like Facebook and Twitter, the lack of labeled corpora with argumentation information and the informal nature of user-generated content make this task more complicated.

Argumentation mining in this task tend to act in the same way of an Information Retrieval (IR) system where potential argumentative microblogs had to come first. A similar approach that addresses such purpose was presented in RepLab task [2], where the output of the priority task will be a ranking of microblogs according to their probability of being a potential threat to the reputation of some entity.

Following the task proposition described above, the argumentation mining task of MC2 lab is then defined as **argumentation detection** combined with **priority ranking** of argumentative microblogs. The detection of argumentation content will depend on a search process that arranged microblogs based on the amount of claims about a given culture event or festival name.

The evidence related to such claims would be an invaluable information for festival organizers, journalists and communication departments. It would be useful even to normal festival spectator, since it would summarize all argumentation facets that one needs to access in order to obtain a satisfactory overview about a festival name.

Participants were welcome to present systems that attempt the whole task objective (argumentation detection + argumentation ranking). These two phases are explicitly considered in Argumentation mining task as following:

- Argumentation detection: Given a festival name as query (Topic), participants have to induce, from the microblog collection, the set of the most argumentative microblogs about this culture event.
- Argumentation ranking: Participants are asked to judge the relevance of each microblog of the set in term of argumentation.

4 Data

4.1 Corpus

The MC2 corpus is a microblog stream, covering 18 months from May 2015 to November 2016, about festivals in different languages [7]. This corpus was provided to registered participants by ANR GAFES project[2]. It consists of a pool with more than 50M unique microblogs from different sources with their meta-information.

4.2 Topics

Given a cultural query about festivals in English or French. The task proposes to search for the 100 most argumentative microblogs.

We chose to gather microblogs based on the most visible festival names on FlickR (the famous photos sharing site)[3] in order to avoid getting microblogs from official pages of festival organizers and getting a maximum of personal microblogs

Only the subset of festivals with at least 300 photos has been considered. The selection was done through a manual exploration on the microblog corpus to ensure providing queries with enough argumentation content for our target audience.

4.3 Baseline

The baseline approach consisted in using Indri language model to search for argumentative microblogs. For each festival, a query including lexical features expressing opinion and argumentation was defined following [1]. In argumentative microblogs, users usually use comparison language to compare and contrast ideas (*More, less*). Authors also tend to use pronouns like (*my, mine, myself,I*). Verbs like *believe, think, agree* and adverbs play an important role to identify argument components. They indicate the presence of a major claim and adverbs like *also,often or really* emphasize the importance of some premise [15]. Verbs like *should, could* are frequently used in argumentative context to express what users were expecting. In addition to this argumentative keywords list, we use a list expression opinion used in [9].

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

[3] https://www.flickr.com/.

5 Results

Argumentative mining received considerable interest with 31 registered participants, but only 5 teams submitted a total of 18 runs per language. Organizers baselines were added to this pool. The NDGC has been adopted as the main official measure, but precision at 100 could have been used since it provided the exact same rankings.

Two reference sets of argumentative structures represented as regular expresions have been assigned to each query (festival name). One has been exracted apriori from the manual interactive run provided as baseline. A second one has been extracted from participant runs. To avoid duplicated content, only microblog textual content has been considered. All meta-data like URLs, #hashtags and @replies were removed. Most argumentative phrases have been extracted from this material and been modeled as generic Regular Expressions. These steps were both applied to the English and French runs.

Table 1 describes average NDGC results for English queries. Results on French are similar but due to a smaller number of queries, differences are not statistically significant. All participant systems relied on an initial step of pretreatment to filter the original dataset by language and topic.

ERTIM Team found the highest number of argumentative microblogs using lexical data enrichment [16]. This resource associates a score to each lemma according to the affective. Besides these lexicon based measures, opinion was detected based on the proportion of adjectives among all part of speech tags. In addition to this opinion scoring process, ERTIM tackled the argumentation detection in the same way by scoring opinion tweets based on the number of conjunctions. Conjunctions are discourse connector commonly used to structure a text. This was a systematic approach applied to all microblogs in the corpus. Although they found a number of argumentative microblogs higher than other participants for almost all queries, there was no overlap with argumentative microblogs found in the baseline runs.

Teams relying on language model using queries mixing multiword terms with argumentative connectors found less argumentative microblogs but a larger overlap with the reference extracted from the baseline run.

Table 1. Best average NDGC scores for top participants (English)

Team	Organizer-Ref	Pooling-Ref
ERTIM	0.0092	**0.6011*****
ECNUica	0.03333	0.082
LIA-run2	**0.0609***	0.0632

6 Conclusion

Previous editions of the MC2 lab focused on contextualization [6] and timeline illustration [8,11] of cultural events over a 18 months period based on the ANR GaFes corpus [7]. In 2018 the main challenge has been to find authentic personal microblogs in this massive collection. This is required to portrait festival reputation among participants. Among them, public argumentative microblogs are the most important since they could have a direct impact on reputation. However, promotional microblogs by festival organizers tend to use similar syntax and form. The main finding of this year is that lexical filtering combined with part of speech analysis is the most efficient to detect these microblogs and rank them by priority. However, this extraction is not exhaustive. An interactive search using complex queries based on Indri language model[4] lead to discover undetected relevant personal argumentative microblogs.

References

1. Aker, A., et al.: What works and what does not: classifier and feature analysis for argument mining. In: Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017, pp. 91–96. Association for Computational Linguistics (2017)
2. Amigó, E., et al.: Overview of RepLab 2013: evaluating online reputation monitoring systems. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF 2013. LNCS, vol. 8138, pp. 333–352. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40802-1_31
3. Bal, B.K., Dizier, P.S.: Towards building annotated resources for analyzing opinions and argumentation in news editorials. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010). European Language Resources Association (ELRA), Valletta, Malta, May 2010
4. Bosc, T., Cabrio, E., Villata, S.: DART: a dataset of arguments and their relations on Twitter. In: European Language Resources Association (ELRA) (2016)
5. Dusmanu, M., Cabrio, E., Villata, S.: Argument mining on Twitter: arguments, facts and sources. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, pp. 2317–2322. Association for Computational Linguistics (2017)
6. Goeuriot, L., Mothe, J., Mulhem, P., Murtagh, F., SanJuan, E.: Overview of the CLEF 2016 cultural micro-blog contextualization workshop. In: Fuhr, N., et al. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 371–378. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_30
7. Goeuriot, L., Mothe, J., Mulhem, P., SanJuan, E.: Building evaluation datasets for cultural microblog retrieval. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018. European Language Resources Association (ELRA) (2018)
8. Goeuriot, L., Mulhem, P., SanJuan, E.: CLEF 2017 MC2 search and time line tasks overview. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, 11–14 September 2017, Dublin, Ireland (2017)

[4] http://www.cs.cmu.edu/lemur/3.1/IndriQueryLanguage.html.

9. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2004, pp. 168–177. ACM, New York (2004)
10. Lippi, M., Torroni, P.: Argumentation mining: state of the art and emerging trends. ACM Trans. Internet Technol. **16**(2), 10:1–10:25 (2016)
11. Mulhem, P., Goeuriot, L., Dogra, N., Ould Amer, N.: TimeLine illustration based on microblogs: when diversification meets metadata re-ranking. In: Jones, G.J.F., et al. (eds.) CLEF 2017. LNCS, vol. 10456, pp. 224–235. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_22
12. Newman, S.E., Marshall, C.C.: Pushing Toulmin too far: learning from an argument representation scheme (1992)
13. Palau, R.M., Moens, M.: Argumentation mining. Artif. Intell. Law **19**(1), 1–22 (2011)
14. Snajder, J.: Social media argumentation mining: the quest for deliberateness in raucousness. CoRR abs/1701.00168 (2017). http://arxiv.org/abs/1701.00168
15. Stab, C., Gurevych, I.: Identifying argumentative discourse structures in persuasive essays. In: EMNLP, pp. 46–56 (2014)
16. Warriner, A.B., Kuperman, V., Brysbaert, M.: Norms of valence, arousal, and dominance for 13,915 English lemmas. Behav. Res. Methods **45**(4), 1191–1207 (2013)
17. Wei, Z., Liu, Y., Li, Y.: Is this post persuasive? Ranking argumentative comments in online forum. In: ACL (2016)