

CLEF MC2 2018 Lab: Technical Overview of Cross Language Microblog Search and Argumentative Mining

Jean Valre Cossu¹, Julio Gonzalo², Malek Hajjem^{3,4}, Olivier Hamon⁴, Chiraz Latiri⁴, and Eric SanJuan³

¹ My Local Influence, Aubagne, France

² UNED, Madrid, Spain

³ LIA, Avignon University, France

⁴ LIPAH, Tunis El Manar University, Tunisia

⁵ Syllabs, Paris, France

⁶ LIPAH, Tunis El Manar University, Tunisia

{malek.hajjem,eric.sanjuan}@univ-avignon.fr

Abstract. MC2 lab mainly focuses on developing processing methods and resources to mine the social media (SM) spheres surrounding cultural events such as festivals, music, books, movies and museums. Two main tasks and one pilot ran in 2018.

The first task was specific to movies. Topics were extracted from the French *VodKaster* website that allows readers to get personal short comments (microcritics) about movies. The challenge was to find related microblogs in four different languages in a large archive.

The second task, *argumentation mining*, aimed to automatically identify reason-conclusion structures that can lead to model social web users positions about a cultural event expressed via Twitter microblogs. The idea was to perform a search process on a massive microblog collection that focuses on claims about a given festival.

A pilot task was also launched on a new corpus, extending the 2017 language recognition task to handle also dialects.

Keywords: Argumentation Mining, Microblog Search, Cross Language Information Retrieval

1 Introduction

Following previous editions, MC2 Lab 2018 was centered on multilingual culture mining and retrieval processes over the large corpus of cultural microblogs[4] considered in the two previous editions[3, 5]. Two main tasks were considered: cross-language cultural microblog search (Task 1) and Argumentation Mining (Task 2).

Topics for Task 1 (microblog search) were extracted from the *VodKaster* website, that allows French readers to get personal short comments (microcritics) about movies. You can get similar and/or complementary opinions on Twitter;

however, they are less specific to movies and harder to find. The usual case is to display to the reader a concise summary of microblogs related to the microcritics he/she is reading, considering bilingual and trilingual users that would read microblogs in other languages than French. Summaries were exclusively made of extracts from microblog contents and should include authors' names if considered informative, and have to be readable. Codes like external URLs references to multimedia objects had to be removed as well. Summaries were intended to provide an idea of all relevant information included in the corpus, and diversity among top ranked microblogs was considered important.

Task 2 was about Argumentation Mining, a new problem in corpus-based text analysis that addresses the challenging task of automatically identifying the justifications provided by opinion holders for their judgment. Several approaches of argumentation mining have been proposed so far in areas such as legal documents, online debates, product reviews, newspaper articles and court cases, as well as in dialogical domains. With the popularization of social networks, argumentation mining is considered as an extension of the opinion mining issue from social network content. The aim is to automatically identify reason-conclusion structures that can lead to model social web users positions about a service or an event expressed through social network platforms like Twitter. Indeed, when we need to form an opinion on a new topic or make a decision, arguments is what we are looking for, rather than a mere aggregation of sentiment or stance. To make argumentation structures available, in the case of Twitter, robust automatic recognition is required. However, the ambiguity of natural language text produced in social media, the different writing styles, the lack of proper syntax, the large amount of implicit context and the heterogeneity of sources make argumentation mining on Twitter a very challenging problem.

Another possible way to identify the argumentation structures from a generic tweet corpus, is to use approaches based on information extraction. The idea is to perform a search process that focuses on claims about a given topic within a massive collection. This approach relates to the field of focused retrieval, that aims to provide users with direct access to relevant information in retrieved documents. In this task, relevant information is expressed in the form of arguments [7].

As in previous MC2 editions, registered participants were given access to the microblog collection[4] provided by ANR project GAFES⁷ with their meta-information and expanded URLs on a MySQL server. Due to legal terms, the access to this database is restricted to registered participants under a privacy agreement.

These two tasks are fully described in the remainder of the paper.

⁷ <http://anr-gafes.univ-avignon.fr/>

2 Task 1: Cross-Language cultural microblog search

Vodkaster⁸ is a French social network about movies where participants can share comments about movies under the form of microcritics no longer than a microblog. The main differences are the restricted cultural domains and the form. The objective of the task is for a given movie or microcritic language among French, English, Spanish, Portuguese and Arabic to provide a summary of the related microblogs.

Microblogs included in a summary should provide relevant information about at least one of the following aspects:

- The film mentioned in the microcritic includes a subject, genre, presence in festivals, reception, audience, critics or opinions, as well as actors and producers careers.
- Events such as festivals mentioned in the microcritics if any, including opinions and narratives.
- Comments and critics in Twitter similar to those in the microcritic if any. Extended summaries can include microblogs about closely related films and events.
- If promotional, automatic microblogs or retweets are not considered as relevant. However, retweets by movie aficionados or movie makers are considered relevant.

2.1 Use Case

The task's use case is to display a concise summary of microblogs to a (native French) reader that are related to the microcritics he/she is reading, considering bilingual and trilingual users that could read microblogs in other languages than French. Summaries are exclusively made of extracts from microblog contents and may include authors' names if this additional piece of information is considered as relevant and informative. Automatically produced summaries should be readable and coded items like external URLs and references to multimedia objects should be removed. Three different summary lengths in words are considered: 50, 150 and up to 250.

Summaries are intended to provide an idea of all relevant information included in the corpus. Diversity among top ranked microblogs is important. If the summary does not provide any microblog directly related to the topic, it is implicitly suggesting that there is no relevant information in the corpus.

2.2 Topics

Topics represent a selection from VodKaster microcritics in French mentioning the term festival. Each topic contains:

- A topic ID,

⁸ <http://www.vodkaster.com/>

- A title made of the movie name,
- A narrative showing a microcritic about the movie,
- A list of nuggets (i.e terms and expressions) manually extracted from microcritic.

To facilitate data exploration, an Indri index with a web interface has been provided to query the whole set of microblogs. Online Indri indexes are also available.

2.3 Results

Runs are evaluated according to their *informativeness* following INEX Tweets Contextualisation [2] guidelines. Seven teams registered for this task, but only one team (A collaboration between Chedi Bechikh Ali from the Institut Suprieur de Gestion, Universit de Tunis, Tunisia, and Hatem Haddad from the Universit Libre de Bruxelles) managed to submit 3 complete runs. A Baseline was generated based on Indri index. Both the baseline and the index were shared with participants.

A multilingual reference of 2887 unique textual contents that could be considered of interest by Vodkaster’s users according to community managers has been manually extracted from the corpus. All microblogs in this reference contain personal opinions about movies or related festivals. Among them, only 229 could be related to topics in the queries. We used a large textual reference characterizing interestingness and a reduced reference about relevant microblogs, and then applied INEX Tweet Contextualisation [2] methodology to compare participant runs with the provided baseline.

All three runs from the only participant outperformed the baseline. Three approaches were experimented. One (FR-FR) without translation, another with translation to English (FR-EN) and a third one using a French to English dictionary. In terms of interestingness, the monolingual approach (FR-FR) did better, which is coherent with the fact that the majority of Vodkaster users express themselves in French. However, the translation approach (FR-EN) outperformed all others on relevancy. This is again coherent with the fact that a majority of microblogs in the corpus are in English. Very specific relevant microblogs can be found but not in the query original language.

Table 2.3 shows interestingness and informativeness results for baseline and participant runs using the *context-eval.pl*⁹ program.

3 Task 2: Mining Opinion Argumentations

Topics for this task are a selection of festival names which are popular on Flickr in English (14) and French (4). Participants have to search for the most argumentative tweets in the same collection of microblogs used for Task 1. The identified

⁹ <http://tc.talne.eu>

Run	Interestingness	Relevance
Baseline	0.057	0.0062
Baseline	5.28	0.41
fr-en-dict	5.86	1.09
fr-fr	10.14	1.51
fr-en	6.89	2.02

Table 1. Evaluation of runs submitted by Chedi Bechikh Ali from the Institut Suprieur de Gestion, Universit de Tunis, Tunisia and Hatem Haddad from the Universit Libre de Bruxelles, based on INEX context-eval skip-gram informativeness measure.

microblogs must be ranked according to their probability of being argumentative. This use case was proposed to help festival organisers deal with online opinions about their festival finding out not only what people liked/disliked but, most importantly, why. For each language (English and French), a monolingual scenario is expected. Diversity in the rank is not required, because an argument that is frequently repeated is assumed to be of higher priority.

3.1 Evaluation

The official evaluation measure has been NDCG. This ranking measure gives a score for each retrieved microblog with a discount function over the rank.

These are examples of opinions about the "Cannes" festival name:

- I've seen some people saying they're boycotting Cannes *because of the high heels rule*. I'm not sure they'll notice.
- Not going to lie, one of my favorite things about the Cannes festival *is all of these handsome men in tuxedos*.
- Cannes is relevant because *movies get timed standing ovations*.

3.2 Baseline

To express argumentation, users tend to employ a specific list of argumentative keywords [1, 8, 6]. For example:

- *More, less*: to compare and contrast ideas
- Pronouns like *my, mine, myself, I* are used to make their statement sound more objective.
- Verbs like *believe, think, agree, should, could* play an important role to identify argument components and express what users were expecting.
- Adverbs like *also, often or really* emphasize the importance of some premise.

We also observed that some expressions (such as *because* \rightarrow *coz*) could be normalized to match a higher number of microblogs. These lexical features about opinion and argumentation were provided to participants.

3.3 Results

Argumentative mining received considerable interest, with 31 registered participants. However, only 5 teams submitted a total of 18 runs per language. Organizer baselines were added to this pool as well. The NDGC has been adopted as the main official measure; however, precision at 100 gives the same rankings.

Two reference sets of argumentative structures were represented as regular expressions and have been assigned to each query (festival name). The first reference of 97 distinct regular expressions has been extracted a priori from the manual interactive run provided as baseline. The second one contains 77 expressions and has been extracted from participant runs. To avoid duplicated content, only microblog textual content has been considered. All meta-data such as URLs, #hashtags and @replies were removed.

These steps were both applied to the English and French runs. Table 3.3 provides examples of extracted regular expressions.

Regular Expression	Argument Matched	Type
.* c'est bien mais .*	contrasting clause	generic
.* super programmation	exact argument	specific
(.*,){3,}	enumeration of at least 3 elements	generic
.*delicious food .*	Exact argument	specific

Table 2. Sample of regular expressions used to match participants runs in French and English languages

Table 3.3 shows the five top runs based on NDGC results for English queries based on the organizers' reference. Table 3.3 shows the same for the reference extracted by pooling from participant submissions. Results over 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 the preliminary treatment to filter the original dataset by language and topic.

Participant runs can be grouped into two strategies: Runs based on the same lexical resource provided by organizers and runs which make use of external resources. ERTIM Team falls in the second category, and it is the group that found the highest number of argumentative microblogs using lexical data enrichment[9]. This resource associates a score to each lemma according to its affective nature.

Besides these lexicon-based measures, opinion was detected based on the proportion of adjectives with respect to 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 connectors 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 models using queries mixing multiword terms with argumentative connectors found less argumentative microblogs, but a larger overlap with the reference extracted from the baseline run. This was the case of the LIA Team, which found the best overlap with the reference of organizers by using a convolutional neural network. As no labeled data was provided, participants from this team constructed their own training dataset. Concerning ECNUica team, they experimented various re-ranking strategies. Finally the ISAMM team experimented with a combination of Information Retrieval, Topic Modeling and Opinion Mining techniques.

Run	Rank EN	Rank FR
LIA-run1	1 (*)	1
LIA-run2	2 (*)	2
ECNUica-0.6	3	6
ECNUica-0.6-2	4	8
ECNUica-0.4	5	3

Table 3. Average NDGC ranks for the five best runs for English on organizer’s reference. (*) denotes statistical significance with $p < 0.05$ with respect to the 6th run.

Run	Rank EN
Ertim-run2	1 (**)
Ertim-run3	2 (*)
Ertim-run1	3 (*)
ECNUica-0.0-3	4
Baseline	5

Table 4. Average NDGC ranks for the five best runs for English on pool (*) and (**) denote statistical significance (with $p < 0.05$ and $p < 0.005$, respectively) with respect to the 6th run.

4 Conclusion

The initial challenge for 2018 was, given a short movie review on the French VodKaster¹⁰ Social Media site, to find related microblogs in the MC2 corpus in four different target languages (French, English, Spanish and Portuguese). Browsing the VodKaster website, French readers got personal short comments about movies. Since similar posts can be found on Twitter, we decided to display

¹⁰ <http://www.vodkaster.com/>

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 scenario, 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. In addition, a second task was created focusing on argument mining in a multilingual collection. It consisted in finding personal and argumentative microblogs in the corpus. Public posts about cultural events such as festivals are frequently promotional announcements by organizers or artists. Personal argumentative microblogs about specific festivals, in contrast, provide real insights into public reception but both their variety and sparsity make them difficult to locate and aggregate. Argumentative mining attracted most of the participants' efforts in this edition of the MC2 CLEF Lab. The cold start scenario of finding them without any specific learning resources motivated the use of IR approaches based on language models or specialized linguistic resources.

References

1. Aker, A., Sliwa, A., Ma, Y., Lui, R., Borad, N., Ziyaei, S., Ghobadi, M.: 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. Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., Tannier, X.: INEX tweet contextualization task: Evaluation, results and lesson learned. *Inf. Process. Manage.* **52**(5), 801–819 (2016)
3. 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, Proceedings. Lecture Notes in Computer Science, vol. 9822, pp. 371–378. Springer (2016)
4. 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)
5. 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, Dublin, Ireland, September 11-14, 2017. (2017)
6. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 168–177. KDD '04, ACM, New York, NY, USA (2004)
7. Lippi, M., Torroni, P.: Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.* **16**(2), 10:1–10:25 (Mar 2016)
8. Stab, C., Gurevych, I.: Identifying argumentative discourse structures in persuasive essays. In: EMNLP. pp. 46–56 (2014)
9. Warriner, A.B., Kuperman, V., Brysbaert, M.: Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods* **45**(4), 1191–1207 (Dec 2013)