ECNU at MC2 2018 Task 2: Mining Opinion Argumentation

Jie Zhou, Qi Zhang, Qinmin Hu, and Liang He

Shanghai Key Laboratory of Multidimensional Information Processing
East China Normal University, 500 Dongchuan Road, Shanghai, 200241, China
{jzhou,qizhang}@ica.stc.sh.cn
{huqinmin}@gmail.cn
{lhe}@cs.ecnu.edu.cn

Abstract. This paper describes our participation in MC2 2018 task2: mining opinion argumentation. We build a tweet retrieval system, which is mainly composed by four parts: data preprocessing, retrieval, redundancy detection and reranking. Only the highly relevant and argumentive tweets are sent to the user based on the topics. In addition, three state-of-the-art information retrieval models as BB2 model, PL2 model and DFR model are utilized. The retrieval results are combined for final delivery.

Keywords: opinion argumentation, retrieval, argumentative ranking

1 Introduction

An argumentation is, broadly speaking, a claim supported by evidence [6]. In corpus-based text analysis, argumentation mining is a new problem 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, on-line debates, product reviews, newspaper articles and court cases, as well as in dialogical domains [8, 10, 6].

There are situations where the information we need to retrieve from a set of documents is expressed in the form of arguments. Recent advances in argumentation mining pave the way for a new type of ranking that addresses such situations and can positively reduce the set of documents one needs to access in order to obtain a satisfactory overview of a given topic. We build a proof-of-concept argumentative ranking prototype. We found that the results it provides significantly differ from and possibly improve those returned by an argumentation-agnostic search engine. Argumentative ranking does indeed provide results that are quite different from those that are obtained by a “traditional” search engine. In this task, relevant information is expressed in the form of arguments [6].

Success of such argumentation ranking will require interdisciplinary approaches based on the combination of different research issues. In fact, to better understand a short text and be able to detect the argumentative structures within a microblog, we could restore a “text contextualization” as a way to provide more information on the corresponding text [3]. Providing such information in order to detect argumentative tweets would highlight relevant ones. In other words, tweets expressed in the form of arguments. Thus, argumentation mining in this situation will tend to act in the same way of an Information Retrieval (IR) system where potential argumentative tweets had to come first. A similar approach that addresses such a purpose is presented in [2], where the output of the priority task will be a ranking of tweets according to their probability of being a potential threat to the reputation of some entity.

In this task, given a set of festivals name from most popular festivals on FlickR English and French language, participants have to search for the most argumentative tweets in a collection covering 18 months of news about festivals in different languages [4]. The identified tweets have to be a summary of ranked tweets according to their probability of being argumentative tweets. Such sets of tweets could be treated easier by priority, by a festival organiser. For each language (English and French), a monolingual scenario is expected : Given a festival name from a topic file, participants have to to search for the set of most argumentative tweets in the same query language within the microblog collection.

The reminder of the paper is organized as follows. Section 2 describes our approach. In Section 3, experimental results are presented. Finally, the paper is concluded in Section 4.

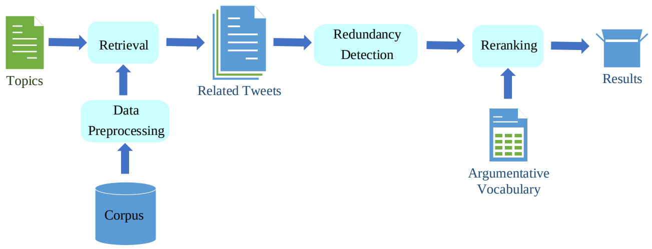

Fig. 1. The architecture of our system.

2 Our Approach

In this section, we demonstrate the architecture of our system, which is shown in Figure 1. It shows that our system mainly consists of four parts, namely data preprocessing, retrieval, redundancy detection and reranking. The details of each part are demonstrated in the following sections.

Data Preprocessing Before we start to run the system, we preprocess the dataset. We first solve the tweets as follow steps:

- Converting the letter in tweet to lowercase letters.
- Turning several spaces into one space.
- Replacing http:// or https:// in tweet with '<URL>'.
- Replacing @USERNAME in tweet with '<USERNAME>'.
- Replacing number in tweet with '<NUMBER>'.
- Replacing repeated character sequences of length 3 or greater with sequences of length 3.
- Removing punctuation in tweet

Then, we use NLTK for tokenization, stemming and splitting the sentences.

Retrieval With the daily tweet stream, we leverage the Terrier search engine [9] for indexing and retrieval. Three state-of-the-art information retrieval(IR) models, namely the BB2 model, the PL2 model and the DFR_BM25 model [1], are utilized for this task. Specifically, with the three IR models, we can obtain three scores for a tuple as (Topic, Tweet). Each IR model returns 3000 most related tweets.

By assuming that different retrieval models may compensate each other by combination, we do a linear combination of the scores to obtain better performance.

Redundancy Detection Since the pushed tweets are expected to cover a variety of arguments given by a user about a culture event, we delete identical tweets through the similarity between two tweets. Specifically, when a candidate tweets specific to a topic, we devise a redundancy detection strategy to determine whether it is redundant or not. To calculate the similarity score between two tweets, we first obtain the corresponding words set as $S(T_1)$ and $S(T_2)$. Then, the similarity score $Score(T_1, T_2)$ is formulated as:

$$Score(T_1, T_2) = \frac{|S(T_1) \cap S(T_2)|}{|S(T_1) \cup S(T_2)|} \tag{1}$$

where $S(T_1) \cap S(T_2)$ is the intersection of $S(T_1)$ and $S(T_2)$, $S(T_1) \cup S(T_2)$ represents the union of $S(T_1)$ and $S(T_2)$, $|\cdot|$ denotes the size of the set. If $Score(T_1, T_2)$ is large than the threshold Θ, we determine there are redundant.

Reranking We rerank the related tweets by considering whether the tweet contains the topic, the length of the tweets and the number of argumentative words in tweets. In order to obtain lexical feature, we download some English argumentative vocabularies (e.g. admirable,cool, admire, adorable adore, advantage and so on) and combine them together. For French, we translate the English vocabulary into French through Google translation API. Finally, we rerank the tweet T the for topic $Topic$ according to the following function:

$$f(T, Topic) = \xi + \alpha \cdot T_{length} + (1 - \alpha) \cdot N_{arg} \tag{2}$$

$$\xi = \begin{cases} 0, & \text{Topic is not in T} \\ 1, & \text{Topic is in T and is continuous} \\ \beta, & \text{Topic is in T and is not continuous} \end{cases} \tag{3}$$

where $\xi \in [0, 1]$ represents whether topic $Topic$ contained in tweet T and whether the topic is continuous in tweet, T_{length} is the length of the tweet T after normalizing, N_{arg} denotes the number of words in argumentative vocabulary after normalizing, $\alpha \in [0, 1]$ represents the weight between T_{length} and N_{arg}.

	NDCG-org+en	NDCG-pooling+en
final_run2_base_LIA_English	**0.06093**	0.046955
final_run1_LIA_English	0.06077	0.063217
en_1.run	0.00253	0.364993
en_2.run	0.00926	**0.601186**
en_3.run	0.00260	0.387928
English_run.run	-	0.053967
Our methods		
ans_0.6_en	0.03333	0.074550
ans_0.6_2_en	0.03333	0.074550
ans_0.4_en	0.02493	0.075260
ans_0.4_2_en	0.02440	0.075096
ans_0.2_en	0.01520	0.076618
ans_0.2_2_en	0.01520	0.076671
ans_0.6_3_en	0.01343	0.076590
ans_0.4_3_en	0.01196	0.079338
ans_0.0_en	0.01140	0.078299
ans_0.0_3_en	0.01057	0.092268
ans_0.0_2_en	0.01057	0.076736
ans_0.2_3_en	0.00977	0.082280
Baseline		
english_queries_red_m	0.00694	0.173046

Table 1. Performance of our submitted runs and the other published runs on English.

3 Experiments

3.1 Data

The complete stream of 70,000,000 microblogs is available. English and French are a respectively 12 and 4 festival name. They represent a set of some popular festivals on FlickR for which we have pictures. Topics were carefully selected by the organizer to ensure that selected topics have enough related argumentative tweets in our corpus. Such manual selection was conduct to to ensure a possible evaluation.

3.2 Evaluation

The official evaluation measures planned are: NDCG and Pyramid.

- **NDCG** This ranking measures will give a score for each retrieved tweet with a discount function over the rank. As we are mostly interested in top ranked arguments, this ranking measures meet our expectation. This measure was also used in TREC Microblog Track [5]. A tweet is :
  - Highly relevant when it is a personal tweet with an argument that directly referred to the festival noun (topic) and may contain more then one justification .
  - Relevant when it comportes at least two of graduation criteria cited above
  - Not relevant if no graduation criteria was found
  - Exemple of tweet gradution
- **Pyramid** [7] This evaluation protocol was chosen to evaluate how much the identified set of argumentative tweets about a festival name is diversified. In fact, participant results are expected to cover a variety of arguments given by a user about a culture event. Such an evaluation protocol will allow us to determine if the identified summary of ranked tweets expresses the same content in different words or involve different arguments about a given festival name.

	NDCG-org+fr	NDCG-pooling+fr
final_run2_base_LIA_French	2.885355	0.149578
final_run1_LIA_French	**2.893689**	0.067417
fr_1.run	2.597113	**2.057355**
fr_2.run	2.593689	1.394706
fr_3.run	2.593689	1.990625
French_run.run	2.592132	0.00000
Our methods		
ans_0.6_fr	2.602948	0.098308
ans_0.6_2_fr	2.602948	0.098031
ans_0.4_fr	2.605087	0.101283
ans_0.4_2_fr	2.605087	0.102899
ans_0.2_fr	2.601962	0.121148
ans_0.2_2_fr	2.601962	0.121148
ans_0.6_3_fr	2.602948	0.095363
ans_0.4_3_fr	2.605087	0.099080
ans_0.0_fr	2.600157	0.076990
ans_0.0_3_fr	2.600157	0.078816
ans_0.0_2_fr	2.600157	0.078750
ans_0.2_3_fr	2.601962	0.119515
Baseline		
French_queries_red_m	2.285177	0.048535

Table 2. Performance of our submitted runs and other published runs on French.

3.3 Experiment Results and Analysis

The experiment results are shown in Table 1 and Table 2. Our observation shows that the proposed model works better than baseline in most cases.

4 Conclusions

In this paper, we present our work in two scenarios of the MC2 2018 task2 mining opinion argumentation . We build a tweet retrieval system. It mainly performs four steps to determine whether to push a tweet or not. We apply three state-of-the-art IR models for search. Various retrieval results are combined for final delivery. Noting that the combination strategy does not work very well, we will extract more useful features and focus on the learning to rank approaches in the future.

References

1. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS) 20(4), 357–389 (2002)
2. Amigó, E., De Albornoz, J.C., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., De Rijke, M., Spina, D.: Overview of replab 2013: Evaluating online reputation monitoring systems. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 333–352. Springer (2013)
3. Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., Tannier, X.: Inex tweet contextualization task: Evaluation, results and lesson learned. Information Processing & Management 52(5), 801–819 (2016)
4. Latiri, M., Cossu, J.V., Latiri, C., SanJuan, E.: Clef 2018. International Conference of the Cross-Language Evaluation Forum for European Languages Proceedings LNCS (2018)
5. Lin, J., Efron, M., Wang, Y., Sherman, G., Voorhees, E.: Overview of the trec-2015 microblog track. Tech. rep. (2015)
6. Lippi, M., Sarti, P., Torroni, P.: Argumentative ranking
7. Nenkova, A., Passonneau, R., McKeown, K.: The pyramid method: Incorporating human content selection variation in summarization evaluation. ACM Transactions on Speech and Language Processing (TSLP) 4(2), 4 (2007)
8. Oraby, S., Reed, L., Compton, R., Riloff, E., Walker, M., Whittaker, S.: And that's a fact: Distinguishing factual and emotional argumentation in online dialogue. NAACL HLT 2015 p. 116 (2015)
9. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier information retrieval platform. In: European Conference on Information Retrieval. pp. 517–519. Springer (2005)
10. Schulz, C., Eger, S., Daxenberger, J., Kahse, T., Gurevych, I.: Multi-task learning for argumentation mining in low-resource settings (2018)