CLEF MC2 Lab: Évaluation, Résultats, et Perspectives

Malek Hajjem[*,***] — **Jean Valère Cossu**[**] — **Chiraz Latiri**[***] — **Eric Sanjuan**[*]

[*] *LIA, Université d'Avignon et des Pays de Vaucluse*
339 chemin des Meinajaries, Agroparc BP 91228, 84911 Avignon cedex 9, France
{malek.hajjem, eric.sanjuan}@univ-avignon.fr
[**] *. My Local Influence, Aubagne, France*
jvcossu@gmail.com
[***] *. LIPAH, Tunis El Manar University, Tunisia*
chiraz.latiri@gnet.tn

RÉSUMÉ. *Le Lab MC2*[1] *2018 est une tâche de recherche d'information (RI) au sein de la campagne d'évaluation CLEF. L'objectif de cette tâche est de développer des méthodes et ressources textuelles pour l'analyse des médias sociaux autour d'évènements culturels. Cet atelier de recherche se focalise sur le corpus de microblogs "GAFES"*[2] *comme principale ressource. En exploitant ces données pendant trois ans, les organisateurs ont pu proposer au fil de temps des sous-tâches aux perspectives diverses. Les différents résultats prouvent que les meilleures performances sont parfois obtenues par l'application des techniques de TAL et d'autres fois réalisées à partir de méthodes plus basiques.*

Ce papier fournit un rapport de synthèse complet décrivant l'évolution de la tâche au fil de son exécution. Nous présentons dans cet article un rapport détaillé sur l'avancement du lab avec un focus particulier sur l'édition de 2018. Nous décrivons par la même occasion les approches proposées par les participants. Nous abordons en détails les mesures d'évaluations adoptées ainsi que les différentes sous-tâches proposées avec leurs objectifs. Nous donnons notamment un aperçu des ressources qui ont été construites, collectées et sont partagées en accès libre.

ABSTRACT. *MC2 CLEF Lab is a task in CLEF centered on mining the social media sphere surrounding cultural events. The objective of this task is to develop processing methods and resources to mine the social media (SM). The task focuses on the microblog collection of the*

1. `https://mc2.talne.eu/`
2. Corpus issu du projet ANR éponyme dit "Galerie des festivals (Gafes)" `https://anr-gafes.univ-avignon.fr/`

GAFES. Running for three years, the organizers have been able to propose over time many under tasks with various perspectives. The different results prove that the best systems are sometimes obtained by the application of NLP techniques and other times realized through traditional methods.

This paper provides a detailed overview of the lab. Unlike the annual overviews that focused on the results of participant's systems, we present in this article the report on the proposed approaches. We discuss in detail the evaluation measures and the different sub-tasks proposed and their goals. We give also an overview of the open access resources provided for registered participants.

MOTS-CLÉS : *Mesures d'évaluation, MC2 Lab, Recherche d'Information, Twitter*

KEYWORDS: *Evaluation measures, MC2 Lab, Information Retrieval, Twitter*

1. Introduction et motivations

MC2 Lab est une tâche de recherche d'information (RI) au sein de la campagne d'évaluation CLEF. L'objectif de cet atelier est de développer les méthodes et les ressources textuelles pour l'analyse des médias sociaux autour des évènements culturels. A travers cette fouille, les organisateurs et les participants visent à : (1) extraire des informations pertinentes[1], (2) produire un contenu informatif à partir d'un contenu plus ou moins bruité et (3) voire même à découvrir potentiellement de nouvelles informations. La tâche se focalise principalement sur le corpus multilingue des microblogs issu du projet "GAFES". Cet article décrit plus particulièrement la dernière édition de cette tâche : MC2 lab 2018. Durant cette édition deux tâches ont été proposées :

– La recherche multilingue d'information dans les microblogs culturels.

– La fouille d'argumentation dans les microblogs.

La contribution de ce papier consiste à fournir un rapport de synthèse complet sur la tâche tout au long de son exécution. La suite de l'article sera organisée comme suit. La section 2 expose le cadre général des contributions proposées. Dans la section 3, nous présentons une première tâche : la recherche d'information multilingue dans les microblogs ainsi que des travaux connexes. Dans la section 4, nous présentons la deuxième tâche : la fouille d'argumentation ainsi que son état de l'art. La section 5 décrit les ressources utilisées. La section 6 introduit les mesures d'évaluation adoptées et nous permet discuter des résultats obtenus par les participants avant de conclure cet article.

2. Cadre général des contributions

MC2 Lab 2018 prend la suite de l'atelier 2016 du CMC [2] ainsi que l'édition 2017[3]. Ce lab vise à fournir une variété des problématiques autour de l'analyse des données générées par les utilisateurs à travers les plateformes sociales dans un contexte multilingue.
Dans son édition de 2018 le MC2 Lab a une double intention :

– D'une part, il vise à fournir et partager des ressources textuelles de microblogs dans la perspective de fouille de ces données ;

– D'autre part, il cherche à instaurer un paradigme de méthode d'extraction de connaissances à partir de ces larges ressources textuelles et à travers les approches des participants.

1. Vis-à-vis d'une requête exprimée dans le contexte d'un festival. La notion de pertinence est liée à l'argumentativité des documents retournés, l'ensemble est détaillé par la suite.
2. Cultural Microblog Contextualization, un lab de contextualisation multilingue de microblogs concernant des évènnements culturels.
`https://mc2.talne.eu/lab/evenement-important`
3. `https://mc2.talne.eu/lab/tasks-2017/`

A partir de l'étude d'un état de l'art, les organisateurs proposent des problématiques innovantes pour lesquelles le besoin d'une solution reste insistant. Les participants à leur tour proposent des méthodes pour résoudre ces tâches données. A l'issue de cette campagne d'évaluation, on se retrouve avec différentes contributions qui présentent les techniques retenues et leurs stratégies.

Une fois qu'une série de contribution a été retournée, nous utilisons les mesures de performances adéquates pour évaluer les sorties de systèmes fournies. Ceci permet de dégager un classement des systèmes proposés par les participants. Cette année MC2 2018 considère deux principales tâches : **La recherche inter-langue des microblogs culturel (1) et la fouille d'argumentation dans un corpus à partir de Twitter (2)**.

3. Recherche d'information multilingue (RIM) dans les microblogs culturels

3.1. *RIM dans les campagnes d'évaluation*

Les plateformes sociales telles que Twitter permettent aux utilisateurs de partager et publier des contenus textuels avec un grand public. Face à cet énorme volume de données il est nécessaire de fournir aux utilisateurs de ces plateformes des outils d'assistance pour répondre à différents types de besoins, dont fouiller les opinions des autres utilisateurs sur un sujet donné, ou encore, obtenir plus d'information sur un évènement particulier (comme retrouver l'avis d'un expert etc.). C'est dans ce contexte que plusieurs campagnes d'évaluation ont souhaité répondre à cette problématique citons la plus connu d'entre elles : la campagne d'évaluation TREC avec la tâche Microblog Track (Soboroff *et al.*, 2012). On retrouve également la campagne INEX (Initiative for the Evaluation of XML Retrieval) qui orientait d'abord ses tâches de recherche vers des collections de documents structurés. En 2011, cette campagne a lancé pour la première fois une tâche de contextualisation de microblogs, tâche sociale proposant de fournir un résumé concis expliquant un peu plus un microblog donné (Bellot *et al.*, 2015). On cite dans ce contexte la campagne DEFT (Défi Fouille de Texte) : une campagne d'évaluation annuelle francophone qui propose des thématiques de recherche exploratoires axées sur la fouille de textes et dont les éditions les plus récentes se sont focalisées sur les microblogs à travers la classification de textes porteurs d'opinions (Paroubek *et al.*, 2018).

3.2. *Cas de CLEF : MC2 Lab*

Le défi initial de cette édition de MC2 consiste à fournir pour une courte critique de film en français, extraite du portail spécialisé Vodkaster [4], la liste des microblogs

4. `http://www.vodkaster.com/` Vodkaster est site Web qui permet aux internautes français d'écrire de courts commentaires personnels sur le cinéma, dites « micro critiques » (moins de 140 caractères). Ils peuvent noter les films mais aussi lire et commenter les critiques écrites par d'autres utilisateurs de la plateforme.

les plus pertinents par rapport à cette critique. Les microblogs retournés peuvent être en français, anglais, espagnol ou portugais (soit l'ensemble des langues supportés par le corpus GAFES). La majorité de ces critiques sont générées et consultées depuis des appareils mobiles des spectateurs et représentent un contenu personnel. Ces critiques étant brèves, il est donc difficile pour un lecteur d'en saisir tous les tenants et les aboutissants. A partir de ce contenu, nous avons extrait sous forme de requêtes tous les commentaires traitant de festivals sur la période 2015-2016 (la même période d'extraction de corpus GAFES).

Partant du principe qu'un contenu similaire peut être retrouvé sur Twitter, cette tâche propose de mettre à la disposition d'un utilisateur, qui souhaite découvrir ce contenu similaire, un résumé concis de ces messages (des microblogs). Un contenu similaire peut être, par exemple, des micro-blogs sur d'autres festivals mentionnant les mêmes réalisateurs ou acteurs. L'aspect multilingue est pris en considération étant donné que les messages les plus pertinents peuvent être en français ou dans d'autres langues. Cette tâche s'inscrit au cœur même de l'objectif du lab soit la mise en relation des ressources textuelles hétérogènes pour l'analyse d'un contenu social et purement personnel. Nous notons que cette tâche s'inspire de la tâche de contextualisation des microblogs culturels (Ermakova *et al.*, 2017).

La difficulté de cette tâche de recherche de microblogs consiste à réadapter la requête basée sur cette critique. En effet, ces « micro critiques » sont trop courtes pour appliquer des méthodes d'extraction des termes pertinents et utiles à un processus de RI classique et paradoxalement trop longues pour être considérées comme requêtes en tant que telles. Ainsi, le challenge réside dans la génération de la requête la plus appropriée aux systèmes de RI. La difficulté réside dans le choix de méthodes fiables de génération de requêtes ainsi que dans l'application des pré-traitements nécessaires afin de cibler le maximum de contenus similaires dans le corpus de microblogs.

4. Fouille d'argumentation dans les microblogs

4.1. *Etat de l'art de la fouille d'argumentation*

La fouille d'argumentation est un nouveau domaine qui vise à définir les outils automatiques qui sont capables d'extraire à partir de textes en langue naturelle les justifications fournies par les détenteurs d'opinion pour argumenter leurs jugements.

L'étude des travaux d'état de l'art autour de fouille d'argumentation montre que plusieurs méthodes d'extraction d'argument ont été proposées jusqu'à présent dans les domaines tels que juridique, les débats en ligne, les commentaires de produits, les articles de journaux ou les articles scientifiques(Palau, 2011 ; Jiménez-Aleixandre et Erduran, 2007 ; Cabrio et Villata, 2012). Cependant, peu nombreux sont les travaux qui s'intéressent à la fouille d'opinion dans les textes courts et bruités. Toutefois, au fil du temps et avec le développement des plateformes sociales, l'extraction d'argumentation est considérée comme une extension du problème d'extraction d'opinions à partir du contenu textuel. L'objectif est d'identifier automatiquement les structures

justificatives qui peuvent exprimer la position des utilisateurs de réseaux sociaux sur un service, un produit ou un événement culturel.

Pour rendre les structures d'argumentation automatiquement détectables, dans le cas de Twitter, une fiabilité d'approche automatique robuste est requise. Dans ce contexte on cite (Ouertatani *et al.*, 2018) où les auteurs ont énuméré les caractéristiques d'une opinion argumentée selon les composantes d'arguments associés et par la suite ils ont mené des expériences en utilisant différents modèles de classification. Reste à noter que ces méthodes automatiques dépendent directement des ressources textuelles spécialement annotées pour cet objectif. Ces dernières doivent être créées de façon reproductible pour être fiables vis-à-vis d'une telle tâche. Cependant, l'ambiguïté des textes courts générés sur les médias sociaux, leurs styles d'écriture ainsi que leurs contenus hétérogènes entravent l'application de processus de détection automatique des argumentations à base de technique "d'Apprentissage Automatique". Nous expliquons par la suite comment la tâche de fouille d'argumentation à MC2 vise à contourner ces difficultés.

4.2. *CLEF MC2 Lab : RI au service de la fouille d'argumentation*

Le deuxième défi de cette édition de MC2 lab consiste à détecter *"Ce que pensent les spectateurs d'un festival donné"*. En effet, l'émergence des réseaux sociaux offre l'opportunité à chaque personne de donner son propre avis sur internet. Surveiller la réputation d'une manifestation culturelle pour analyser ce qui s'est bien déroulé de ce qui était moins bon représente une motivation intéressante de cette deuxième tâche. Cette analyse ne s'arrêtera pas à détecter les avis positifs ou négatifs de leurs spectateurs concernant un festival mais elle vise à aller plus loin pour comprendre "pourquoi" le spectateur a partagé une telle perception. Cette contribution s'inscrit dans ce qu'on appelle **la fouille d'argumentation**, une tâche qui vise à extraire automatiquement à partir de textes en langue naturelle les justifications fournies par les détenteurs d'opinions pour raisonner leur jugement.

Comme expliqué ci-dessus, l'intuition de cette tâche de fouille d'argumentation à MC2 est de contourner les difficultés concernant l'absence de ressources annotées en termes d'argumentation. MC2 2018 propose un autre moyen possible de détecter l'argumentation, à partir d'un corpus de microblogs génériques, à travers des approches basées sur l'extraction d'information. L'idée est d'adopter un processus de RI qui se focalise principalement sur les revendications des spectateurs à propos d'un festival donné à partir d'une collection massive de microblogs culturels. Cette approche invoque ce qu'on appelle le domaine de la recherche d'information **ciblée**, qui vise à fournir aux utilisateurs un accès direct aux informations pertinentes contenues dans les documents récupérés. Dans cette tâche, une information est considérée pertinente si elle est exprimée sous forme d'argument. Nous visons à l'issue de cette phase de pouvoir ordonner les microblogs argumentatifs en terme de pertinence d'argumentation exprimé à travers le message. Les microblogs les plus argumentatifs seront donc les mieux classés.

Pour ce faire, il est nécessaire de combiner des approches interdisciplinaires. Par exemple, pour mieux comprendre un texte court et pouvoir détecter sa structure argumentative, il est possible d'avoir recours à la « contextualisation de celui-ci » (Bellot *et al.*, 2016). Le but est de pouvoir fournir des informations supplémentaires au texte original afin d'aider à sa compréhension. Ces informations mettront en valeur les microblogs pertinents, dans notre cas les microblogs portant sur le même sujet et contenant des arguments. Ainsi, la fouille d'arguments dans cette situation aura tendance à agir de la même manière qu'un système de RI pour lequel des microblogs potentiellement argumentatifs doivent apparaître en premiers dans le classement de tous les documents retournés par un système de RI. Contrairement à un processus de classification binaire classique, qui se contente de détecter ce qui est argumentatif de ce qui ne l'est pas, cette tâche vise à ordonner en terme d'argumentation les microblogs pertinents concernant un festival donné.
En conclusion la tâche de fouille d'argumentation à partir de Twitter devrait agir en deux phases :

– Établir un processus de recherche qui se concentre sur les revendications à propos d'un festival donné ;

– Classer les résultats de recherche en se basant uniquement sur la pertinence en terme d'argumentation.

5. Présentation des ressources et du cadre d'évaluation du Lab MC2

5.1. *Le corpus de microblogs*

Le corpus de MC2 est un corpus des microblogs qui s'étend sur 18 mois entre mai 2015 et novembre 2016. Cette collection de microblogs comporte plus de 50 millions de microblogs avec leurs métadonnées. La collection de microblogs contient un large nombre de publications publiques sur Twitter contenant le terme "*festival*". Ces microblogs ont été collectés à l'aide de services d'archives privées basés sur des API en streaming. Le cadre de collecte s'inscrit dans le contexte d'un projet ANR dit "Galerie des festivals" (Gafes) qui est porté par deux laboratoires de l'Université d'Avignon (Centre Norbert Elias et Laboratoire Informatique d'Avignon). L'objectif global de ce projet est de former un observatoire des festivals, à partir du corpus de microblogs de festivals composé : des "Rencontres Trans Musicales de Rennes", du "Festival d'Avignon", du "Marché du film à Cannes", des "Vieilles Charrues" et du "festival Lumières". Du fait des conditions générales d'utilisations des services de Twitter, seuls les participants au lab MC2 auront la possibilité d'accéder à ce corpus donnant ainsi la possibilité d'indexer, fouiller et analyser la collection. Les résultats obtenus pourront être publiés sans restriction.

5.2. *Les requêtes*

5.2.1. *Tâche de recherche d'information multilingue*

Les requêtes pour la tâche 1 représentent une sélection de micro-critiques extraites de VodKaster en français mentionnant le terme "*festival*". Chaque requête contient :

– Un identifiant ;

– Un titre qui comprend le nom du film ;

– Le contenu textuel de la micro critique sur le film donné ;

– Une liste d'expressions dites "*nuggets*" définie manuellement à partir du contenu des micro-critiques[5].

5.2.2. *Tâche de fouille d'argumentation*

Partant d'une liste de noms de festival en anglais[6] et en français[7], la tâche de fouille d'argumentation propose de chercher les 100 messages Twitter les plus argumentatifs à propos de ces noms de festivals.
Le choix des noms de festival est basé sur le critère de visibilité[8]. Le but était, entre autre, de cibler un contenu de microblog purement personnel et d'éviter les microblogs publiés par les organisateurs de ces festivals. Seule la liste des festivals qui ont au moins 300 photos publiées ont été considérés.

6. Mesures d'évaluation et synthèse des résultats du Lab MC2

6.1. *Mesure d'informativité dans un cadre de recherche d'information dans les microblogs*

L'objectif de la tâche suit les traces de "contextualisation de texte court" introduite à INEX (Bellot *et al.*, 2015). A base d'un processus de RI, la finalité étant de générer un résumé permettant de remettre en contexte une critique qui, de par sa taille, ne couvre pas la totalité des éléments permettant à un lecteur de comprendre tout ou partie de son contenu. Ce résumé étant basé sur une ressource de données différente de celle des requêtes, est supposé fournir des informations complémentaires relatives au contenu de la critique et qui représente la continuité de cet événement culturel perçu par des spectateurs différents.

5. Cette liste contient les termes qui nous intéressent dans le contenu des critiques auxquelles on cherche les messages Twitter similaires.
6. `https://mc2.talne.eu/~t17malek/mc2_2018_t2/opinion/en/arg/data_t2_sample/English-topics.csv`
7. `https://mc2.talne.eu/~t17malek/mc2_2018_t2/opinion/en/arg/data_t2_sample/French_topics.csv`
8. Cette notion de festival populaire a été détecté à travers FlickR `https://www.flickr.com/photos/tags/flicker/` un site de partage de photos personnelles.

Notre contribution à travers cette tâche concerne plus particulièrement la mise en place d'un cadre d'évaluation de la RI des textes courts. Pour ce faire la tâche propose un protocole d'estimation de pertinence ainsi que des mesures d'évaluation adéquates à l'objectif final. L'étude de l'état de l'art concernant l'évaluation des résumes automatiques nous indique que la notion du contenu informatif de ce dernier est importante. En effet, les travaux de l'estimation de la qualité d'un résumé se base sur une comparaison avec un ou plusieurs résumés dits de "référence", ou directement avec le texte d'origine. Dans le cas d'utilisation de résumés de référence, ces derniers sont généralement construit manuellement. Ainsi, l'évaluation des résultats obtenus peut se faire en estimant le pourcentage d'information des résumés de référence présent dans le résumé construit automatiquement. Ceci peut se faire à base de n-grammes comme avec ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004). Citons également la métrique de pyramidale (Nenkova et Passonneau, 2004) dans laquelle, au lieu de comparer des distributions de n-grammes, on procède au préalable à l'identification manuelle de concepts clefs (nuggets) sur les résumés de référence, leur présence/absence affectera ensuite les résumés produits automatiquement.

Ces modèles d'évaluation confirment le besoin d'une référence construite à partir des jugements de pertinence produits humainement. D'ailleurs selon la définition de la tâche, les évaluateurs sélectionnent les informations les plus pertinentes à partir des résultats de recherche restitués par les participants. Par conséquent, ces références ne peuvent pas être un passage particulier du corpus des microblogs, mais une sélection des expressions bien formées intéressant l'évaluateur. De ce fait, l'informativité ne peut pas être estimer à l'aide de mesures RI standard, car "la contextualisation" des critiques à partir de corpus des microblogs ne contient pas tous les passages pertinents, mais uniquement une sélection parmi tous ceux qui pourraient apporter une information complémentaire.

Pour construire la référence propre à cette tâche nous nous sommes basés sur les critères suivants :

– La référence doit contenir des aspects que les utilisateurs du réseau Vodkasters cherchent à retrouver sur les microblogs ;

– On trouve la trace du film mentionné dans le micro-critique dans l'un des festivals discutés dans le corpus de microblogs ou des opinions/avis à propos des carrières d'acteurs et de producteurs liés au film en question ;

– Les commentaires et les critiques dans les microblogs sont similaires à ce qui mentionné dans les micro-critiques, dans ce cas : les résumés résultants peuvent inclure des microblogs sur des films et des évènements étroitement liés ;

– Les microblogs ou les republications automatiques (non personnels) ne sont pas considérés comme pertinents. Cependant, les retweets des cinéphiles ou des cinéastes le sont ;

– La recherche des microblogs liés est également contrainte par la date de publication de la micro-critique.

Les résultats obtenus

Sept équipes se sont inscrites pour cette tâche, mais une seule équipe (qui consiste en une collaboration entre l'Institut Supérieur de Gestion, Université de Tunis, Tunisie, et l'Université Libre de Bruxelles) a réussi à soumettre 3 scénarios. Une référence (baseline) a été générée par les organisateurs sur la base de système Indri.
Dans un premier lieu, une référence multilingue de 2887 expressions textuelles uniques ayant suscité l'intérêt des utilisateurs de Vodkaster selon les responsables de la communauté. Tous les microblogs de cette référence contiennent des opinions personnelles sur des films ou des festivals connexes. Parmi eux, seuls 229 pourraient être liés aux requêtes proposées. Pour l'évaluation finale nous avons adaptés ces deux sources retenues :

– La large référence qui caractérise "l'intérêt" des utilisateurs de Vodkaster ;

– La référence la plus réduite qui caractérise la pertinence.

Nous utilisons la même méthodologie d'évaluation que celle utilisée lors de "INEX microblog Contextualisation" (Bellot *et al.*, 2016) pour comparer les soumissions des participants avec les références. Comme le témoigne le tableau 1, l'implémentation

Run	L'intérêt	La pertinence
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**

Tableau 1. *Évaluation des runs des participants à base de protocole d'évaluation d'INEX (mesure d'informativité skip-gramm).*

des trois scénarios produits par notre participant dépassent les performances du baseline produit par Indri. Les trois approches consistent en :

– **Un scénario Monolingue (FR-FR) :**

- Dans cette approche, le participant utilise l'extension de requête en français à base du modèle probabiliste BM25.

– **Deux Scénarios Multilingues (FR-EN)** : où le participant traduit les requêtes

- La première approche multilingue utilise une traduction à base de wordnet

- La deuxième approche multilingue utilise une trduction à base d'un dictionnaire bilingue

Les résultats obtenus montrent que le scénario monolingue est plus efficace. Ces résultats semblent logiques vu que les utilisateurs de Vodkaster sont majoritairement francophones. L'approche de traduction a donné à son tour des résultats intéressants en terme de pertinence, ce qui nous semble cohérent étant donné que la majorité de corpus de microblogs est en anglais.

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

Tableau 2. *Exemple des microblogs argumentatifs à propos du festival de Cannes.*

6.2. *Mesure de NDGC pour le classement des microblogs argumentatifs*

Le protocole d'évaluation de cette tâche sera imposé par la finalité de cette dernière. Nous rappelons que la contribution proposée suggère de contourner les approches automatiques pour détecter l'argumentation à partir d'un corpus des microblogs. En effet, la majorité de ces dernières exigent d'utiliser une ressource textuelle annotée en terme d'argumentation, de choisir un classifieur et des caractéristiques adéquates pour décrire un argument. L'extraction d'arguments dans notre cas reposera sur phase de RI qui a pour objectif de cibler les informations qui expriment une opinion argumentée (voir le tableau 2) dans un corpus de microblogs. Une large collection des microblogs non annotée est fourni aux participants. Nous n'imposons pas une façon de faire particulière pour l'identification des microblogs argumentés mais nous proposons de restituer pour un festival particulier la liste des microblogs argumentés ordonnée par pertinence argumentative. Nous notons qu'une telle tâche s'inspire d'une tâche présentée à RepLab (Amigó *et al.*, 2013) qui consiste à un classement des microblogs en fonction de leur probabilité d'avoir des propos impactant potentiellement la réputation d'une entité économique.

Partant de la proposition de tâche décrite ci-dessus, l'extraction d'argumentation sera une tâche définie comme étant une approche combinée entre la recherche d'information ciblée et le classement prioritaire des microblogs argumentatifs. Ceci explique notre choix concernant la mesure de classement NDGC. Une mesure qui donnera un score pour chaque microblog récupéré avec une fonction de réduction sur le rang. Vu que nous sommes principalement intéressés par les arguments les mieux classés, le NDGC, comme mesure de classement, répond forcément à nos attentes. On note que cette mesure a été aussi utilisée lors de la campagne d'évaluation TREC (Lin *et al.*, 2016). Un microblog est considéré très pertinent lorsque celui-ci est personnel et contient un argument qui fait directement référence au festival.

Enfin, la détection du contenu de l'argumentation dépendra du processus de classement des microblogs en fonction des appréciations relatives à un événement culturel ou à un nom de festival donné. Les arguments présentés par les spectateurs seraient une information précieuse pour les journalistes et les départements de la communication des organisateurs de festivals.

Génération du scénario de référence

Dans ce scénario nous visons à générer les microblogs argumentatifs à base d'un système RI. L'approche baseline consiste à utiliser Indri Index [9] afin de lancer une recherche des microblogs argumentatifs. Cette exécution basique repose sur une construction de requête simple à base des critères lexicaux qui expriment l'opinion et l'argumentation. En fait, nous supposons que pour exprimer des arguments, les utilisateurs ont tendance à employer une liste spécifique de ce qu'on peut appeler mots-clés argumentatifs. Les indicateurs lexicaux d'argumentation ont été inspirés par (Aker *et al.*, 2017). Nous notons aussi que l'exploration manuelle de corpus des microblogs nous a permis de cerner les critères lexicaux qu'utilisent les internautes pour argumenter leurs jugements. La liste non exhaustive de ces expressions qui expriment l'argumentation peut être :

– Les adverbes de comparaisons (*more, less*) ;

– Les prénoms personnels ou possessifs(*my, mine, myself, I*) : sont utilisés pour faire de leur déclaration, une déclaration plus objective.

– Des verbes comme (*believe, think, agree, should, could*) : jouent un rôle important dans l'identification de la structure d'un argument et expriment généralement les attentes d'un locuteur,

– Des adverbes comme (*also, often or really*) : soulignent l'importance de certains aspects.

Certaines autres expressions sont propres à la nature textuelle des microblogs. Dans ce contexte nous notons des expressions comme (*because* $->$ *coz)*) qui peuvent être normalisées pour une meilleure correspondance. La liste des critères argumentatifs sur la base du vocabulaire a été mise à la disposition des participants durant la campagne.

Les résultats et la méthodologie d'évaluation

Cette deuxième tâche de MC2 lab a suscité l'intérêt de plusieurs participants. En effet, nous avons reçu 31 demandes de participation. Cependant, seules 5 équipes ont finalement soumis leurs approches. Pour évaluer ces runs nous avons procédé à la construction des références. Deux types de références ont été proposées :

La construction de référence à base d'échantillonnage (pooling) s'inspire des méthodes d'évaluation adoptées par la campagne d'evaluation TREC (Sanderson, 2010). TREC est une campagne de référence en RI permettant d'évaluer la dimension thématique. La collection de jugement de pertinence (appelée aussi la vérité de terrain) associe à chaque *topic* (besoin d'information) l'ensemble des documents (dans notre cas les microblogs) pertinents. Étant donné que le corpus est trop volumineux pour être exhaustivement analysé dans le but d'identifier la pertinence de jugement, nous avons eu recourt à la technique du pooling. Ainsi, pour chaque requête, un pool des microblogs est constitué à partir des 100 premiers microblogs restitués par chacun des systèmes des participants. Par la suite, les doublons sont supprimés (une opération

9. https ://mc2.talne.eu/data/clef/api

L'expression régulière	L'argument ciblé	Type
.* c'est bien mais .*	clause qui exprime une contradiction	générique
.* super programmation	argument exact	specifique
(.*,){3,}	l'énumération d'un ensemble de justification	générique
.*delicious food .*	argument exact	spécifique

Tableau 3. *Échantillons d'expressions régulières en langue française et anglaise à faire correspondre aux "runs" des participants.*

d'union ensembliste est alors appliquée). L'hypothèse est que le nombre et la diversité des contributions au pool permettront de trouver un maximum de microblogs pertinents. Enfin, une vérification manuelle est effectuée pour examiner chaque microblog du pool afin d'identifier s'il répond ou non au besoin d'information spécifié dans la requête considérée (la structure argumentative dans notre cas). Le microblog est alors qualifié de pertinent ou de non pertinent.**Référence construite manuellement** : Cette collection représente le résultat de l'exploration manuelle du corpus, c'est-à-dire l'ensemble des microblogs argumentatifs qu'un annotateur humain avait annotés. Contrairement à l'étape précédente, aucune méthode automatique n'a été appliquée pour la construction de cette collection.

Ces deux références de l'ensemble des structures argumentatives ont été par la suite représentées sous formes d'expressions régulières. Le but étant de pouvoir les assigner par la suite aux microblogs restitués par les participants pour découvrir le nombre des microblogs argumentatifs détectés. A partir de la première référence nous avons distingué 77 expressions régulières. La deuxième référence distingue 97 expressions régulières. Des prétraitements particuliers comme la suppression des doublons et des métadonnées (Hashtags et les URL) ont été appliqués. Seul le contenu textuel des microblogs a été considéré. On note que ces étapes d'évaluation ont été aussi bien appliquées pour les scénarios de recherche en anglais que pour ceux en français. Le tableau 3 expose des exemples d'expressions régulières en langue française utilisées.

Dans ce qui suit nous exposons les tableaux de l'ordonnancement des résultats obtenus en termes de NDGC. Le tableau 4 montre les 5 meilleurs résultats pour les requêtes en anglais, à base de la référence des organisateurs. Le tableau 5 montre les mêmes résultats en considérant cette fois la référence extraite par pooling des soumissions des participants. Les résultats pour les requêtes en français sont similaires, mais en raison du nombre réduit de requêtes (dans cette langue), les différences n'étaient pas statistiquement significatives.

Les soumissions des participants sont toutes basées sur une phase de prétraitements comprenant le filtrage de la collection des microblogs ainsi que l'identification de la langue. Nous pouvons grouper les soumissions des participants en deux ensembles :

– Les approches utilisant la même ressource de microblogs ;

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

Tableau 4. *Le classement de NDGC pour les cinq meilleures soumissions en français et en anglais (la référence de l'organisateur).* (*) *désigne significativité statistique avec* $p < 0.05$ *par rapport à la* 6^{ieme} *soumission.*

– Celles utilisatant d'autres ressources externes.

L'équipe ERTIM est celle qui a trouvé le plus grand nombre de microblogs argumentatifs par rapport à la référence construite par pooling. Cette équipe utilise l'enrichissement de données lexicales à base d'une ressource externe. Cette ressource associe un score à chaque lemme en fonction de sa nature affective (Tsytsarau et Palpanas, 2012). Outre ces mesures basées sur le lexique, l'opinion a été détectée sur la base des proportion d'adjectifs par rapport à toutes les balises d'étiquetage morphosyntaxique. En plus de ce processus de notation d'opinions, ERTIM a abordé la détection d'arguments de la même manière en notant des microblogs d'opinions à base du nombre de conjonctions. Bien que cette équipe soit parvenue à détecter la proportion des microblogs la plus élevée dans la référence construite par pooling, elle n'a pas réussi à avoir une performance similaire avec la référence manuelle.

Les équipes ayant utilisé des modèles de langue combinant reformulation de requêtes avec des connecteurs argumentatifs ont trouvé moins de nouveaux microblogs argumentatifs, mais un chevauchement plus important avec la référence extraite manuellement. D'ailleurs c'est le cas de l'équipe de LIA qui a trouvé la meilleure correspondance avec l'ensemble des microblogs argumentatifs construits manuellement. Cette équipe a utilisé des réseaux de neurones convolutifs. Étant donné que cette tâche ne donne pas accès à un corpus d'apprentissage, cette équipe a généré son propre corpus d'apprentissage. En ce qui concerne l'équipe ECNUica, ils ont expérimenté diverses stratégies de reclassement des microblogs restitués par leur système de RI. Enfin, l'équipe ISAMM a expérimenté une combinaison de techniques de RI, la représentation dans espace de thèmes (Topic model) et la fouille d'opinions.

7. Conclusion et perspectives du Lab MC2

Cette édition de MC2 a proposé deux tâches principales dont l'objectif était de révéler les limites des approches statistiques et linguistiques. Le premier challenge de MC2 2018 consistait à extraire les passages pertinents d'un corpus de microblog pour une requête délivrée à partir d'une collection de critiques de films issues de Vodkaster.

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

Tableau 5. *Le classement de NDGC pour les cinq meilleures soumissions (La référence pooling) (*) et (**) désigne la significativité statistique (avec respectivement* $p < 0.05$ *et* $p < 0.005$*) par rapport à la* 6^{ieme} *soumission.*

La contextualisation de ces critiques à partir de microblogs, consistait à produite un résumé concis des commentaires générés sur Twitter. Dans ce contenu nous attendions à avoir des microblogs argumentatifs. D'où l'idée de la deuxième tâche de fouille d'argumentation dans les microblogs. La détection d'argumentation dans le cadre proposé a tendance à cibler le contenu plus personnel plutôt que le contenu officiel publié par les structures d'organisation d'évènements culturels.

Les résultats de ces deux tâches semblent être prometteurs. En effet, le protocole d'évaluation que nous avons choisi a réussi à montrer que les différents systèmes des participants ont pu découvrir des nouveaux arguments (référence pooling) que les organisateurs n'ont pas pu identifier (référence manuelle). Néanmoins ces mêmes résultats ont montré que les équipes qui ont réussi à décrypter ces nouveaux arguments, étaient moins à même de cibler les microblogs annotés comme argumentatifs par les organisateurs. Cette perception nous semble un peu vague mais nous les rapportons au fait que les soumissions se sont plus au moins focalisées sur le corpus des microblogs même et que d'autres ont eu recours à des ressources textuelles externes.

Pour conclure cette exploration argumentative telle que nous l'avons proposé se concentre sur les données textuelles et leurs structures. Nous visions par cette deuxième tâche à nous adapter le plus possible à la ressource textuelle. Le volet de production des méthodes complètement automatiques n'étant pas notre objectif ultime. D'ailleurs, les méthodes à base de RI ont donné une autre dimension à la problématique et peuvent être ré-implémenter pour toute sorte d'applications en cas d'absence de données annotées.

8. Bibliographie

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 », *Proceedings of the 4th Workshop on Argument Mining*, Association for Computational Linguistics, p. 91-96, 2017.

Amigó E., Carrillo de Albornoz J., 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* P. Forner, H. Müller, R. Paredes, P. Rosso, B. Stein (eds), *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 333-352, 2013.

Bellot P., Moriceau V., Mothe J., Juan E. S., Tannier X., « INEX Tweet Contextualization Task : Evaluation, Results and Lesson Learned », *Information Processing and Management*, vol. 52, n° 5, p. 801-819, September, 2016. Thanks to Elsevier editor. The definitive version is available at http ://www.sciencedirect.com The original PDF of the article can be found at Information Processing and Management (ISSN : 0306-4573) website website : http ://www.sciencedirect.com/science/article/pii/S0306457316300218.

Bellot P., Moriceau V., Mothe J., San Juan E., Tannier X., « Mesures d'informativité et de lisibilité pour un cadre d'évaluation de la contextualisation de tweets », *Document numérique, Evaluation en Recherche d'Information*, vol. 18, n° 1, p. 55-73, avril, 2015.

Cabrio E., Villata S., « Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions », *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Short Papers - Volume 2*, ACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 208-212, 2012.

Ermakova L., Goeuriot L., Mothe J., Mulhem P., Nie J.-Y., Sanjuan E., « CLEF 2017 Microblog Cultural Contextualization Lab Overview », *International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF 2017)*, vol. 10456, Springer, Dublin, IE, p. pp. 304-314, 2017. Thanks to Springer editor. This papers appears in Volume 10456 of Lecture Notes in Computer Science ISSN : 0302-9743 ISBN : 978-3-319-65812-4 The original PDF is available at :https ://link.springer.com/chapter/10.1007/978-3-319-65813-1_27.

Jiménez-Aleixandre M. P., Erduran S., *Argumentation in Science Education : An Overview*, Springer Netherlands, Dordrecht, p. 3-27, 2007.

Lin C.-Y., « ROUGE : A Package for Automatic Evaluation of summaries », *Proc. ACL workshop on Text Summarization Branches Out*, p. 10, 2004.

Lin J., Efron M., Wang Y., Sherman G., Voorhees E. M., « Overview of the TREC-2015 Microblog Track », 2016.

Nenkova A., Passonneau R., « Evaluating Content Selection in Summarization : The Pyramid Method », *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics : HLT-NAACL 2004*, 2004.

Ouertatani A., Gasmi G., Latiri C., « Détection d'opinion argumentée à partir de Twitter », *COnférence en Recherche d'Informations et Applications - CORIA 2018, 15th French Information Retrieval Conference, Rennes, France, May 16-18, 2018. Proceedings.*, 2018.

Palau R. M., Automatic Detection and Classification of Argumentation in a Legal Case (Automatische detectie en classificatie van de argumentatie in een juridische zaak), PhD thesis, Katholieke Universiteit Leuven, Belgium, 2011.

Paroubek P., Grouin C., Bellot P., Claveau V., Eshkol-Taravella I., Fraisse A., Jackiewicz A., Karoui J., Monceaux L., Juan-Manuel T.-M., « DEFT2018 : recherche d'information et analyse de sentiments dans des tweets concernant les transportsen Île de France. », *DEFT 2018 - 14ème atelier Défi Fouille de Texte*, vol. 2 of *Actes de la conférence Traitement Automatique des Langues, TALN 2018*, Rennes, France, p. 1-11, May, 2018.

Sanderson M., « Test Collection Based Evaluation of Information Retrieval Systems », *Foundations and Trends® in Information Retrieval*, vol. 4, n° 4, p. 247-375, 2010.

Soboroff I., Ounis I., Macdonald C., Lin J., « Overview of the trec 2012 microblog track », *In Proceedings of Text REtrieval Conference*, 2012.

Tsytsarau M., Palpanas T., « Survey on mining subjective data on the web », *Data Mining and Knowledge Discovery*, vol. 24, n° 3, p. 478-514, May, 2012.