Tweets analysis for event detection

Soumaya Cherichi[1], **Rim Faiz**[2]

1. Larodec, ISG Tunis, Université de Tunis
Bouchoucha, 2000 Le Bardo, Tunisie
soumayacherichi@gmail.com

2. Larodec, IHEC Carthage, Université de Carthage
Carthage Présidence, Tunisie
Rim.Faiz@ihec.rnu.tn

ABSTRACT. *Social media systems have been proven to be valuable platforms for information and communication, particularly during events; in case of natural disaster like earthquakes tsunami and states of nuclear emergencies in Japan in 2011. The behavior leads to an accumulation of an enormous amount of information. However, finding relevant posts can be a challenging task, since the relevance of a post is dependent both on its content, author and tweet's characteristics. Besides identifying tweets that describe a specific type of event is also challenging due to the high complexity and variety of event descriptions. These challenges present a big opportunity for Natural Language Processing (NLP) and Information Extraction (IE) technology to enable new large-scale data-analysis applications. Taking to account all the difficulties, this paper proposes a new metric to improve the results of the searches in microblogs. It combines content relevance, tweet relevance and author relevance, and develops a Natural Language Processing method for extracting temporal information of events from posts more specifically tweets. Our approach is based on a methodology of temporal markers classes and on a contextual exploration method. To evaluate our model, we built a knowledge management system. Actually, we used a collection of 10 thousand of tweets talking about the current events in 2014 and 2015.*

RÉSUMÉ. *Les médias sociaux ont été révélés être des plates-formes utiles pour les publications informations et la communication, en particulier lors des événements; par exemple en cas de catastrophe naturelle comme les tremblements de terre, tsunami et dans les pays d'urgence nucléaire comme au Japon en 2011. Ce comportement conduit à une accumulation d'une énorme quantité d'informations. Cependant, pour trouver les messages pertinents peut être une tâche difficile, puisque la pertinence d'un poste dépend à la fois de son contenu, de l'auteur et des caractéristiques du* tweet. *Outre l'identification des tweets qui décrivent un type spécifique d'événement est aussi difficile en raison de la grande complexité et la variété des descriptions d'événements. Ces défis présentent une grande opportunité pour le traitement automatique du langage naturel (NLP) et la technologie de l'extraction d'information (IE) pour construire à grande échelle de nouvelles applications d'analyse de données. En tenant compte de toutes ces difficultés, cet article propose une nouvelle mesure pour améliorer les résultats des recherches dans les microblogs. Il combine la pertinence du contenu, la pertinence de l'auteur et la pertinence du* tweet, *et développe une méthode de traitement automatique des langages naturels pour extraire des informations temporelle des événements dans les messages plus spécifiquement* tweets. *Notre approche est basée sur une* *méthodologie des classes de marqueurs temporels et sur une méthode d'exploration contextuelle. Pour évaluer notre modèle, nous avons construit un système de gestion des connaissances. En fait, nous avons utilisé une collection de 10 mille* tweets *qui parlent de l'actualité entre 2014 et 2015.*

KEYWORDS: *microblogs, relevant information, NLP, tweets search, information retrieval.*

MOTS-CLÉS : *microblogs, infromation pertinente, TALN, recherche de tweets, recherche d'informations.*

DOI:10.3166/ISI.21.1.61-80 © 2016 Lavoisier

1. Introduction

Recent years have revealed an important increase of interactive media, which gave birth to a huge volume of data in blogs and more precisely microblogs. These micro-blogs, like Twitter and Facebook, attract more and more users due to the easiness and the speed of information shared instantly. During the "Arab Spring Movement", Twitter was used as an information source to coordinate protests and to bring awareness to the atrocities (Huang, 2011). In recent world events, social media data has been shown to be effective in detecting earthquakes (Sakaki *et al.*, 2010; Mendoza *et al.*, 2010), rumors (Ou *et al.*, 2011), and identifying characteristics of information propagation (Xianghan *et al.,* 2015). This incites us to study the problem of event detection, which is an interesting and important task in such circumstances.

Much work has been dedicated to the extraction of information, essentially unstructured, on tweets, e.g. for named entities (Ritter *et al.*, 2011), detection of new topics (Petrovic *et al.*, 2010), events (Osborne *et al.*, 2014), automatic summarization (Xin Zhao *et al.,* 2011), the detection of emotions (Volkova *et al.*, 2013), user modeling (Li *et al.*, 2014)... Among the closest work of those proposed in this work, (Panem *et al.*, 2014) extracted a structure in the form of triplets and diagrams attribute values from the tweets, but using an analysis of dependencies and only for natural disaster events. Other work focuses instead on an open field and a large-scale treatment (Ruchi *et al.,* 2013), but often at the expense of structured approaches and global optimization. Bayesian models, however, also have been proposed for the extraction of events on Twitter, for example, in (Ritter *et al.*, 2012). These studies provide the basis for research in this work.

Despite the wealth of research on this issue, work from the literature focus on the textual content of the messages exchanged and neglect the social aspect. However, users often insert of non-text content in their messages. In particular, users have the ability to insert (*i.e.* mention) in their messages from other users pseudonyms. This practice called "mentioning" is found on most social media, notably Twitter, Facebook and Google+ that meet all three of the same syntax, namely "pseudonym". These statements are actually hotlinked intentionally to initiate the discussion with specific users or automatically when replying to a message, or even on Twitter, during a re-tweet. Considering this particular type of bond as dynamic as it is tied to a specific time period, *i.e.* the lifetime of the message and a specific subject, *i.e.* the one addressed by the message.

In fact event detection approaches designed for documents cannot be stricly applied to tweets due to their specific characteristics. Our work consists in suggesting a new metric, which allows studying the impact of each feature on impact on the quality of search results. We also intend to develop a Natural Language Processing method to extract temporal information from tweets.

We gathered the features on three groups: those related to content, those related to tweet and those related to the author. We used the coefficient of correlation with human judgment to define our score. For processing the content of tweets, we intend to use resources and linguistic methods. In our task of identifying event information from tweets, we are interested in identifying four classes of linguistic markers (keywords) namely temporal markers (calendar term, occurrence indicator, relative pronoun, transitive verb). Our experimental result uses a corpus of 10 thousand of subjective tweets, which are neither answers nor retweets.

The remainder of this paper is organized as follows. In Section 2, we give an overview of related works. In Section 3 we present our new metric measure. In Section 4, we present our approach of Event information extraction and discuss experiments and obtained results in Section 5. Finally, Section 6 concludes this paper and outlines future work.

2. Related works

A micro-blogging service is at the same time a communication means and a collaboration system that allows sharing and distributing text messages. In comparison with other social networks on the Web (for example Facebook, Myspace), the microblogs articles are particularly short and submitted in real time to report a recent event. At the time of this writing, several micro-blogging services exist. In this paper, we will focus on the micro-blogging service Twitter that is the most popular and widely used. Especially since certain features and functionalities characterize Twitter. The main one consists in social relationship that may follow. This directed association enables users to express their interest in other microbloggers' posts, called tweets, which doesn't exceed 140 characters. Moreover, Twitter is marked by the retweet feature, which gives users the ability to forward an interesting tweet to their followers.

Several works have focused on the analysis of data posted on microblogs, particularly in Twitter like (Ben Kraiem *et al.*, 2014). (Barbosa, Feng, 2010) and (Jiang *et al.*, 2011) propose approaches for sentiment classification regarding Twitter messages *i.e.* determine whether tweets express a positive, negative or neutral feeling. Positive and negative polarities correspond respectively to a favorable and unfavorable opinion as well. To solve this task the authors have used natural language processing and machine learning techniques. (Cha *et al*., 2010) proposes an approach to measure user u-influence in twitter.

Many studies have found that there is a high correlation between the information posted on the web and present results. (Doan *et al*., 2011) have used tweets to analyze awareness and anxiety levels of Tokyo inhabitants during the events of earthquakes tsunami and the sites of nuclear emergencies in Japan in 2011. (Lampos, Cristianini, 2010) have presented a method to measure the prevalence of H1N1 disease in the population of United Kingdom. They also sought in the tweets the symptoms related to the disease and obtained results, which were compared with real results from the Health Protection Agency. Besides (O'Connor *et al*., 2010) analyzed the tweets to predict public opinion and then compared the results with the surveys.

Twitter messages reflect useful event information for a variety of events of different types and scale. These event messages can provide a set of unique perspectives, regardless of the event type (Duan *et al*., 2010; Yardi, Boyd, 2010), reflecting the points of view of users who are interested or who participate in an event. In case of unexpected events such as Earthquakes, Twitter users sometimes spread news prior to the traditional news media (Kwak *et al*., 2010; Sakaki *et al.,* 2010). A for planned events (e.g., the 2010 Apple Developers conference), Twitter users often post messages in anticipation of the event, which can lead to early identification of interest in these events. Additionally, Twitter users often post information on local, community-specific events (e.g., a local choir concert), where traditional news coverage is low or non-existent.

Previous work on event extraction (Allan *et al*., 1998; Chambers, Jurafsky, 2011) and (Faiz, 2006) have focused largely on news articles, as historically this genre of text has been the main source of information on current events. In the meantime, social networking sites such as Facebook and Twitter have become an important complementary source of such information. While status messages contain a wealth of useful information, they are very disorganized fostering the need for automatic extraction, aggregation and categorization. Although there has been much interest in tracking trends or memes in social media (Lin *et al*., 2011; Leskovec *et al.,* 2009), little work has addressed the challenges arising from extracting structured representations of events from short or informal texts.

Several research efforts have focused on identifying events in social media in general and on Twitter in particular (Becker *et al.,* 2010; Metzler *et al*., 2012; Sankaranarayanan *et al*., 2009). Recent work on Twitter has started to process data as a stream, as it is produced, but has mainly focused on identifying events of a particular type, e.g., news events (Sankaranarayanan *et al*., 2009), earthquakes. Other works identify the first Twitter message associated to an event as soon at it happens (Petrovi´c *et al*., 2010).

In the context of event extraction from tweets, (Chakrabarti, Punera, 2011) have developed a framework that takes a keyword related to a particular event, returns a summary that responds to the request. The summary contains the time of the beginning that indicates when the event began to be discussed, a term that specifies how long the event was discussed, and a small number of posts during this time interval. In the same context, (O'Connor *et al.,* 2010) proposed a method to generate summaries from tweets (in real time) covering an event e.

Our work consists in examining the role and impact of social networks, in particular microblogs, on public opinion. We aim to analyze the behavior of users through the texts they post in order to extract the events that reflect the interests and opinions of a population. We introduce in this paper our approach for tweet search that integrates different criteria namely the social authority of micro-bloggers, the content relevance, the tweeting features as well as the hashtag's presence. Once we selected relevant tweets, we move to the step of identifying event information from these tweets. In the addition we want to identify four classes of linguistic markers (keywords) namely temporal markers (calendar term, occurrence indicator, relative pronoun, cause-consequence verb). This way our work can be seen different and unlike the work of (Doan *et al*., 2011) and (Lampos, Cristianini, 2010) which use only sets of keywords to detect events known in advance. In addition to the previous works we intend to detect events "not previously known" that can be stimulating for users at the same time.

3. Metric Measure of the impact of criteria to improve search results

We introduce a research model that combines tweets relevant content, the specificities of tweets and the authority of bloggers. This model considers the specificities of tweets and the authority of bloggers as important factors, which contribute to the relevance of the results. The search for tweets is a task of information retrieval whose goal is to select the relevant sections in response to a user's request. To present an accurate list of articles, our model combines a score of content's relevance, a score of author's authority and a score of tweets' specificities. The objective of this combination is to provide a list of tweets that cover the subject of the request and are posted by major bloggers.

3.1. Content relevance features

The criterion "Content" refers to the thematic relevance traditionally calculated by IR systems standards. The thematic relevance is generally measured by one of several IR models. One of the models reference Information Retrieval IR is the probabilistic model (Jones *et al*., 2000) with the weighting scheme BM25 as matching request document function. For this reason, we have adopted this model for the calculation of the thematic relevance. Of course, it is made possible to calculate using any other IR model. BM25 is a search function based "bag of words"; it allows us to organize all documents based on the occurrences of the query terms given in the documents (cf. Section 2).

We used four content relevance features:

a) Relevance(T,Q): we used OKAPI BM25; an algorithm developed by (Robertson, Jones, 1976) which measures the content relevance between the query Q and tweet T.

$$\text{Tf-Idf}(Q,ti) = \text{Tf}(Q,ti).\text{Idf}(Q,ti) = \text{Tf}(Q,ti)\ .\ \text{Log}(N/(\text{DfQ}+1)) \qquad (1)$$

Knowing that: w is a term in the query Q and Ti is the tweet i.

b) Popularity(Ti,Tj,Q): with i and j in n and i≠j: it used to calculate the popularity of a tweet from the corpus. It measures the similarity between the tweets in the context of the tweet's topic. We used cosine similarity, according to a study done by (Akermi, Faiz, 2012) cosine similarity is the most efficient similarity measure, in addition, it is not sensitive to the size of each tweet:

Popularity(Ti,Tj,Q)= Cosine (ti,tj,Q)=

$$\frac{\sum_{Q\in(ti\cap tj)} TfidfQ,ti * TfidfQ,tj}{\sqrt{\sum_{Q\in(ti)} (TfidfQ,ti)^2 * \sum_{Q\in(tj)} (TfidfQ,tj)^2}} \qquad (2)$$

Knowing that w is a term in the query Q, Ti is tweet i, Tj is tweet j, i and j n and i≠j.

c) Length of tweet (Lg(Ti,Q)): Length is measured by the number of characters that a tweet contains. Tweets more long, contains more information.

$$\text{Lg (ti)}= \frac{\text{Lg(ti)}-\text{MinLg(ti)}}{\text{MaxLg(ti)}} \qquad (3)$$

d) Out of Vocabulary (OOV(Ti)): This feature is used to roughly approximate the language quality of tweets. Words out of vocabulary in Twitter include spelling errors and named entities. This feature aims to measure the quality language of tweet. The smaller the number of out of vocabulary; the better the quality of tweet is.

3.2. Tweet relevance features

We note that the thematic relevance depends solely on the item and query. Each tweet has many technical features, and each feature form selection criteria that we have exploited.

a) Retweet (Ti,Q): is defined according to the number of times a tweet is retweeted. In a rational manner, the most retweeted tweets are most relevant. Retweets are forwarding of corresponding original tweets, sometimes with comments of retweeters. According to (Duan *et al.*, 2010), they are supposed to contain no more information than the original tweets.

$$\text{Retweet (ti)=}\ \frac{\text{Retweet (ti)}-\text{Minretweet}}{\text{Maxretweet}} \tag{4}$$

b) Reply(Ti): An @reply is any update posted by clicking the "Reply" button on a Tweet; it will always begin with @username. This feature aims to calculate the number of reply to a tweet. Ultimately tweets that receive the most response are the most relevant.

$$\text{Reply (ti)=}\frac{Reply\ (ti)-Minreply}{Maxreply} \tag{5}$$

c) Favor(Ti): this feature aims to calculate the number of times a tweet is classified as the favorite. According to (Sankaranarayanan *et al.*, 2009), if many followers consider a message as a favorite, it means that it is relevant.

$$\text{Favor (ti)=}\frac{Favor\ (ti)-Minfavor}{Maxfavor} \tag{6}$$

d) Hashtag Count(Ti):The # symbol, called a hashtag, is used to mark keywords or topics in a Tweet. Twitter users created it organically as a way to categorize messages. Generally, users do not use the same hashtag for a particular subject. That's why we initially performed the following tasks to normalize hashtags

– We standardized hashtags that relate to the same Topics.

– We noticed that hashtags that appear with the same hyperlinks are similar.

– The presence of two different hashtags with the same hypertext means that these hashtags are related to the same thing.

– We grouped hashtags that deal with the same topic to sum-up the reference using similar re-occurrent words.

This feature aims to calculate the number of hashtags in tweet.

$$\text{Hashtagcount(Ti)=} \sum \text{of occurrences of hashtag} \tag{7}$$

e) Url count(Ti):Twitter allows users to include URL as a supplement in their tweets. This feature aims to estimates the number of times that the URL appears in the tweet corpus. According to (Cherichi, Faiz, 2014 and 2013a), tweets containing URLs are more informative.

$$\text{URLcount(Ti)=} \sum \text{of occurences of URL} \tag{8}$$

3.3. Author Relevance Features

Each blogger has specific characteristics such as number of follower and number of mention According to (Cherichi, Faiz, 2013b, 2013c); users who have more followers and have been mentioned in more tweets, listed in more lists and retweeted by more important users are thought to be more authoritative.

a) TweetCount(a):this feature represents the number of tweet posted by the author (a).

$$\text{Tweetcount(a)} = \sum \text{of tweets(a)} \tag{9}$$

b) Mention Count (author): A mention is any Twitter update that contains “@username” anywhere in the body of the Tweet; this means that @replies are also considered mentions. This feature aims to calculate the number of times an author (a) is mentioned.

$$\text{Mentioncount(a)} = \sum \text{of mention(a)} \tag{10}$$

c) Follower(a):this feature represents the number of followers to the author (a).

d) Following(a): this feature represents the number of subscriptions of the author (a) to other authors

e) Expertise(a): this feature was found by conducting a survey that asks people to rate the expertise of the blogger/author (a) from 0 to 10

f) RetweetRank(a): Retweet Rank looks up all recent retweets, number of followers, friends and lists of an author (a). It then compares these numbers with those of other bloggers’ and assigns a rank. Retweet Rank tracks both RTs posted using the Retweet button and other RTs (ReTweets) (e.g. RT@username). This feature is an indicator of how a blogger is influential on twitter

g) TwitterPageRank(a): this feature represents the rank of an author (a) of the total twitter users using PageRank Algorithm (Page, 1997) which works by counting the number and quality of links to a page to determine a rough estimate of how important the twitter’s blogger is.

h) Audience (a): is the size of the potential audience for a message. What is the maximum number of people who could have been exposed to a message?

$$\text{Audience (a)} = \sum_{f\ \in abonnés\ (a)} \frac{1+p*\ audience\ (f)}{\|Abonnements\ (f)\ \|} \tag{11}$$

3.4. Final score

After normalizing the feature scores, these three scores are combined linearly using the following formula:

$$\text{Score(Ti,Q)} = \text{Scorecontent(Ti,Q)} + \beta\ \text{ScoreTweet(Ti,Q)} + \gamma\ \text{ScoreAuthor(a,Q)} \tag{12}$$

With

– Scorecontent(Ti,Q) on [0, 1] because tweet content should deal with the topic of request. In fact, being able to measure the content relevance of a tweet is essential from a semantic perspective, since it enables distinguishing between noise and pertinent tweets: pertinent tweets must have a content score that goes beyond a threshold value which is the mean of the scores, otherwise it is considered non pertinent and can't be considered for the second filtering step. Once we selected the most relevant tweets according to the most reliable scorecontent, then we calculate the score of tweets according to scoretweet and scoreauthor mentioned above

– Scorecontent (Ti, Q) is the normalized score of the relevance of content;

– ScoreTweet (Ti,Q) is the normalized score of the specificity of the tweet Ti;

– ScoreAuthor (a, Q) is the normalized score of the importance of the author (a) corresponds to the blogger who published the tweet Ti;

– $\beta + \gamma = 1$;

We note that:

– Scorecontent(Ti,Q)=Relevance(T,Q) + Lg(Ti) + Popularity(Ti,Tj,Q) + Quality(Ti);

– ScoreTweet(Ti,Q)= Url count(Ti) + Hashtag Count(Ti) + Retweet(Ti) + Reply(Ti) + Favor(Ti);

– ScoreAuthor(a,Q)= TwitterPageRank(a) + Audience(a) + Tweet Count(a) + Mention Count(a) + Expertise(a) + RetweetRank(a) + Follower(a) + Following(a);

4. Event information extraction

To detect a target event from Twitter, we search from Twitter and find useful tweets. Our method of acquiring useful tweets for target event detection is portrayed in Figure 1.

Specifically, we automatically extract all information about events from tweets and specify more information about these events: associations, locations, temporal settings, etc. We propose an event extraction system which aims at automatic extracting of significant tweets bearing information with temporal knowledge from news articles as well as identifying the agent, the location, and the temporal setting of those events.

Our system (cf. Figure 1) is divided into five modules:

1) A lexical analysis module allowing the chunking of a tweet into words.

2) A morphological analysis module identifying words while triggering functions that deal with morphological inflexions and generate a morpho-syntactic code for each word.

3) A syntactic analysis module that re-establishes the order of the morpho-syntactic codes generated by the morphological analyzer with the aim of building some morpho-syntactic structures.

4) An extraction module, which allows us to pick out markers in order to identify distinctive sentences, which represent events.

5) A module for interpretation of the extracted tweets to identify “Who did what?”, “to whom?” and “where?”.

In the following sections, we briefly introduce the different phases of the event information extraction that are based typically on Natural Language Processing techniques.

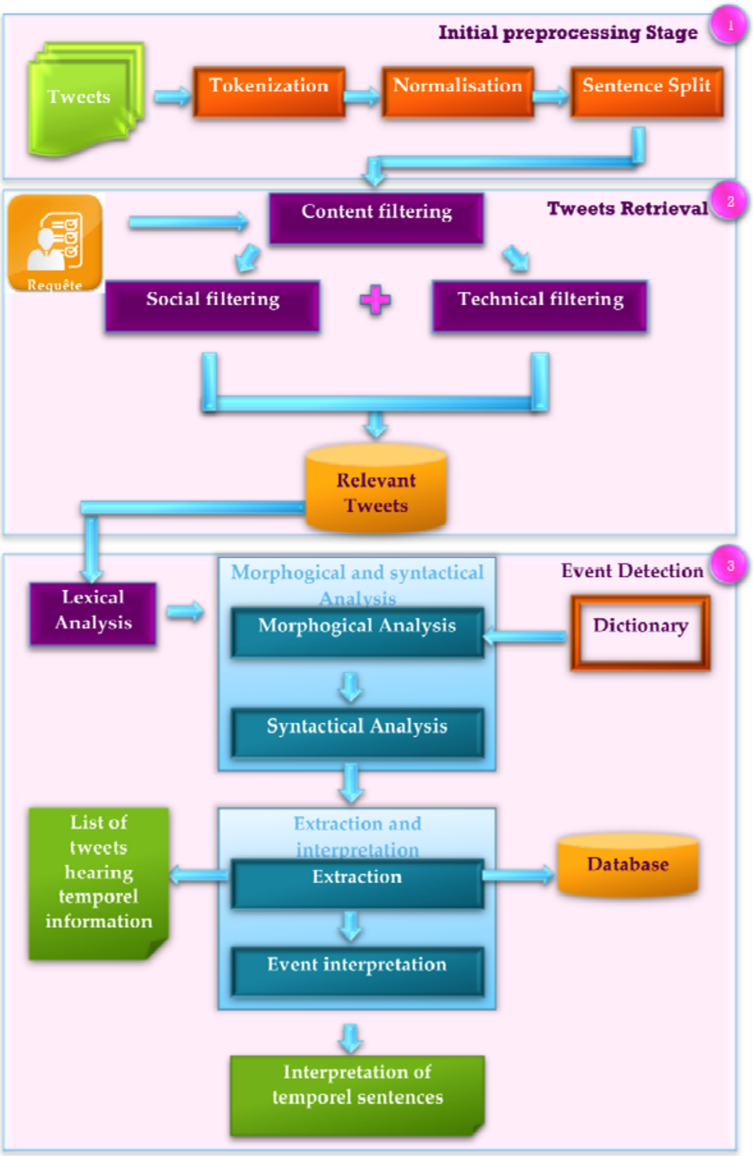

Figure 1. Architecture of our method

4.1. Lexical analysis

The lexical analysis module is an essential module for each text analysis whatever its type may be. Its fundamental task is the segmentation; our system splits the document into words by means of very precise definer detection. Example: “,” “?” “!”. We also have functions which deal with every case of ambiguity or nuance such as acronyms and abbreviations.

The role of the lexical marker is therefore to provide the “raw” basis to have access to the dictionary of the morphological analyzer and this will help the recognition of words.

4.2. Morphological analysis

The purpose of this analysis is to recognize the lexical unit provided by the lexical analyzer and to locate the linguistic data stored in the dictionary such as genre, syntactic category, etc.

This recognition requires, in the first place, the calculation of a possible valid inflexion starting from the raw basis which provides two variables: basis and inflexion respecting the following condition: **word = basis + inflexion**. The basis variable represents the key for dictionary research. This will be loaded in a chunking table to save time for research. Once we identify the basis, we will have information about syntactic category, root and inflectional model.

The inflexion variable, on the other hand, allows us to calculate other variables which are the following: verb tense, verb form, type and number for names and adjectives. The calculation will be possible if we refer to the inflectional model that has already been spotted thanks to the root variable.

When we finish the morphological analysis, we will have a syntactic category for each lexical unit such as genre, number, root and verb tense. To extract temporal information, we only need the root, the verb tense and the syntactic category, though we could extend the calculation process to other variables such as genre or number.

4.3. Syntactic analysis

The syntactic analysis of natural language cannot be directly achieved by means of the linguistic area. It needs regrouping rules. These rules imply that we first categorize the text form we intend to analyze; and this is the purpose of a morphological analysis.

On the other hand, it is worth mentioning that the syntactic analysis cannot be efficient unless it uses prediction in most of the cases. This can occur thanks to the morpho-syntactic structure, which is, sometimes, signaled by its form. Very often, these structures are spotted thanks to their precise morphological features. Their location is possible through the morphological analysis of the corresponding shapes (or forms).

The maximal analysis frame that we can logically adopt is the sentence. This seems to be logical since our task is to extract sentences. However, if we describe the case of using sentences, it is then easy to extend it to the case of using paragraphs.

Starting from the morpho-syntactic structures, which are the result of the morphological analysis, the syntactic analyzer will order and build tweets with structures that will allow us to apply the extraction process, it is what we are going to see in the next section.

4.4. Extraction and interpretation of event information

Based on the results of the research of (Faiz, 2006), we analyzed several tweets, we noticed that they may have one of the following forms:

Calendar term followed by an event. Example:

#**Tunisie**: et maintenant, le tour de l'**élection** #présidentielle; blog de @GeopolisFTV http://geopolis.francetvinfo.fr/tunisie-la-democratie-en-marche/2014/11/17/tunisie-apres-les-legislatives-la-presidentielle.html …

1) Preposition followed by a calendar term. Example :

Depuis le 25 mai, les Français de **Tunisie** sont représentés par 5 conseillers consulaires : leurs noms, leur mission. http://www.ambassadefrance-tn.org/Election-des-conseillers …

2) Event followed by a calendar term. Example :

Officiel : Le deuxième tour de l'**élection** présidentielle en #**Tunisie** aura lieu avant le 31 décembre 2014.
https://twitter.com/albawsalatn/status/481761209235279872

3) Subject followed by a relative pronoun, followed by a verb cause-consequence, followed by event. Example:

Jour historique en #**Tunisie** qui organise sa première **élection** présidentielle libre après 24 ans de benalisme

4) Subject followed by a verb cause-consequence, followed by event. Example :

L'Union européenne déploie 100 observateurs pour l'**élection** présidentielle en **Tunisie** http://fb.me/1J6uq02NQ

5) Subject followed by a verb cause-consequence, followed by event. Example :

En **Tunisie**, l'**élection** présidentielle s'achemine vers un second tour http://www.lemonde.fr/tunisie/article/2014/11/23/presidentielle-en-tunisie-vers-un-deuxieme-tour-entre-marzouki-et-essebsi_4528046_1466522.html … //

This representation has led us to draw the main **linguistic markers** and to sequence them according to their types.

1) The calendar term class

a) propo-num stands for preposition + number. Example: *Depui*s 2012

b) Cal-num stands for: calendar + number. Example: janvier 2010.

c) Prepo stands for preposition. Example: maintenant,

d) Num-cal-num stands for number + calendar + number Example: 17 janvier 2014.

2) The occurrence indicator class

a) Adj_occ stands for adjective + occurrence. Example: une autre fois, la dernière fois, la première fois

b) Adt_det_occ stands for tense adverb + determiner + occurrence. Example: encore une fois.

3) The relative pronoun class

a) Prr_aux_ppa: relative pronoun + auxiliary + past participle. Example: #Tunisie qui a organisé

b) Prr_aux_adv_ppa stands for relative pronoun + auxiliary + adverb + past participle. Example: qui a trop bu.

4) The cause-consequence verb

a) Verbconsq_subject : event + verb + event . Example: mini-tornade a provoqué des dégâts ;

b) Verbconsq_argument : subject + verb + event. Exemple: le Conseil de prévention et de lutte contre le dopage avait provoqué une petite crise avec l'Union cycliste.

5) As the temporal markers are independent from the language, our EXEV system can also be applied to English corpus and Arabic corpus. Examples of temporal markers:

6) maintenant (French), now (English), الآن (Arabic).

7) depuis 2012 (French), since 2012 (English), 2012 منذ عام (Arabic).

8) une autre fois (French), another time (English), مرة اخرى (Arabic).

9) avant (French), before (English), قبل (Arabic).

Table 1. List of cause-consequence verbs

Verbconsq_subject	Verbconsq_argument
avoir lieu	provoquer
se produire	organiser
provoquer	organiser
s'expliquer par	permettre
se traduire par	subir
affecter	déclencher
aboutir à	conduire à
précipiter	assister à
se passer	contribuer à
avoir pour origine	aboutir à
être entraîner	se traduire par
rendre à	donner lieu à
se donner	perpétrer
	inciter à

5. Experimental evaluation

We conducted a series of preliminary experiments on a collection of 10 thousand articles from Twitter, in order to evaluate the performance of our model.

We built a search engine that we have called "TWEETRIM", which allows to calculate all scores and display the most relevant tweets according to these score. It has as input a query composed of three keywords and as output a set of relevant tweets relative to the query.

To collect data, we implemented a Java program that used the Twitter4J library. This library provides access to data (tweets, user information...) Twitter via its programming interface, Twitter API. We mainly studied the content of tweets (their sizes, the most frequents words, words known by a French lexicon), the preoccupations users based on hashtags used, the behavior of users.

To perform queries and to collect the human judgment of relevance followed the following steps:

– we collected 1000 queries on recent actualities in Tunisia from users,

– then, we used the system that we have built which allows us to view the relevant 10 results according to the score of the content,

– and then, we asked 450 users to judge the 10 first results of each query.

We suppose that the content relevance already exists and we will improve our search result by varying our two other scores; ScoreTweet and ScoreAuthor. We calculate the correlation coefficient between our scores and the corpus, which allowed us to find our weighting coefficients β and γ.

5.1. Results

5.1.1. Estimation of weights

We make a comparison within the values of correlation coefficients and through the results, we observe that the best correlation coefficient between βScoreTweet+γScoreAuthor with human judgment score = 0,3842 when β = 0,8 and thus γ= 0,2.

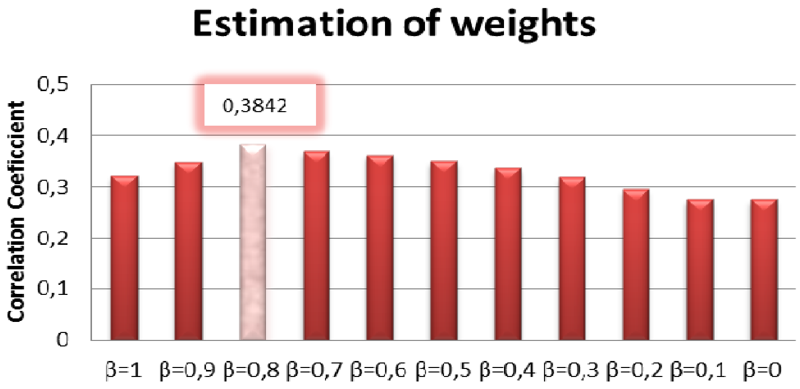

Figure 2. Estimation of weights

5.1.2. Evaluation of the system

We notice that the performance of the last 2 configurations are very close with a slight advantage for the combination “Tweet Features & Author Features” on the model based only on the specificities of the tweet and the importance of the author. We conclude that Author features have more impact on the search results than Tweet features.

The reference model combines only the features linearly without weighting. This model gave us the correlation coefficient equal to 0,2459 and our model gave us the correlation coefficient of 0,3842. It can clearly be noticed that there is 56% improvement in the satisfaction of our human judgment.

The events information extraction were derived by running the system on tweets already selected through Twitterim from our dataset. The tweets covered different themes like weather reports, politics, statements of people and editorials.

On the whole, around 1000 tweets were used to ascertain that the extraction module (event information extraction) did work.

We have conducted experiments to verify the effectiveness of our proposed approach to event information extraction.

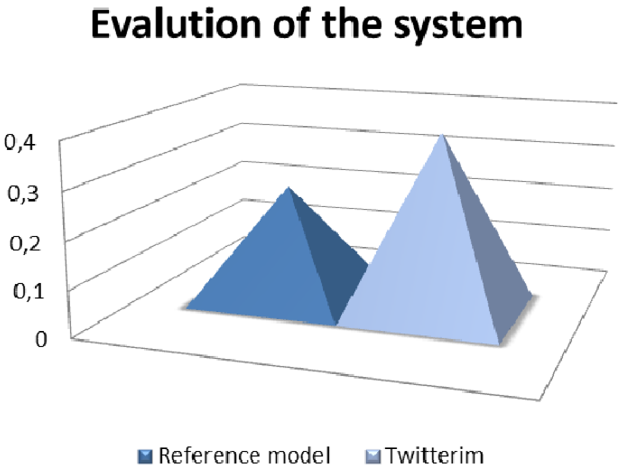

Figure 3. Comparing our model with reference model

Firstly, the extracted event sentences were evaluated by human experts and over 80% of them were deemed good event sentences.

Secondly, in order to measure the performance of the system, the results for the testing of the event extraction system were measured using standard information extraction units recall and precision where:

$$\text{Recall} = \frac{\text{No. of relevant event sentences indentified}}{\text{No. of event sentences sentences}} = 80\%$$

$$\text{Precision} = \frac{\text{No. of relevant event sentences identified}}{\text{No. of event sentences identified}} = 88.9\%$$

6. Conclusion

Research conducted under the auspices of knowledge management varies greatly in direction and scope. There are several approaches based on the features that have been proposed. Therefore the choice of characteristics is important to obtain a satisfactory result and close to the human judgment. We have proposed in this paper a new metric for Social Research on twitter. This has to integrate relevance of content, the specificities of tweets and the author’s importance in which we incorporate new features such as the audience. The primary experimental evaluation that we conducted on a collection of articles of Twitter shows the figures that we propose allow a better assessment of the bloggers' impact of and tweets' technical specificities.

We also presented several new techniques for identifying events and their related social media documents, by combining multiple context features of the document in a variety of disciplines.

Thanks to our morphological analyzer based on inflectional morphology, we were able to directly extract event type information as well as interpret the type of event itself (*i.e.,* future event or past event). We identified some classes of linguistic markers (keywords) namely temporal markers (calendar term, occurrence indicator, relative pronoun, and transitive verb) and number marker.

Looking ahead, we plan to conduct experiments under the Micro-blog using Text REtrieval Conference (TREC), evaluation framework, that will include a collection of many articles and queries on larger scale and whose relevance will be based on social judgments. We also intend to evaluate the influence of each feature independently. Besides, we plan to compare the performance of our model with other models for social information retrieval.

Our other future intentions will include investigation about whether the performance, in particular the recall, can be increased through extending our approach taking a second step which will allow the appearance of different events into one.

References

Akermi I., Faiz R. (2012). Hybrid method for computing word-pair similarity based on web content. *In Proceedings of the International Conference on Web Intelligence, Mining and Semantics, WIMS'12*, New York, NY, USA, ACM.

Barbosa L., Feng J. (2010). Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, p. 36-44. Association for Computational Linguistics.

Becker H., Naaman M. and Gravano L. (2010. Learning similarity metrics for event identification in social media. In *WSDM'10.*

Ben Kraiem M., Feki J., Khrouf K., Ravat F., Teste O. (2014). OLAP of the tweets: From modeling to exploitation*. IEEE International Conference on Research Challenges in Information Science (IEEE RCIS'14)*

Cha M., Haddadi H., Benevenuto Krishna F., Gummadi P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media 2010, ICWSM'10.*

Chakrabarti D., Punera K. (2011). Event Summarization using Tweets, *in ICWSM 2011*.

Chambers N., Jurafsky D. (2011). Template-based information extraction without the templates. *In Proceedings of ACL,* Portland, OR.

Cherichi S., Faiz R. (2014). Analyzing the behavior and text posted by users to extract knowledge. *In proceedings of the International Conference on Computational Collective Intelligence Technologies and Applications, ICCCI'14*, Seoul, Korea, ACM 2014 Lecture Notes in Artificial Intelligence of Springer-Verlag.

Cherichi S., Faiz R. (2013a). New metric measure for the improvement of search results in microblogs. *Proc. of the International Conference on Web Intelligence, Mining and Semantics (WIMS'13),* New York, NY, USA, ACM.

Cherichi S., Faiz R. (2013b). Relevant information discovery in microblogs: new metric measure for the improvement of search results in microblogs. *Proc. of INSTICC International Conference on Knowledge Discovery and Information Retrieval (KDIR'13),* Vilamoura, Portugal, 19-22 September. ©SciTePress

Cherichi S., Faiz R. (2013c). Relevant information management in microblogs. In International Conference on Knowledge Management, Information and Knowledge *Systems (KMIKS 2013),* Hammamet, Tunisia, Avril.

Doan S., Vo B.K.H., Collier N. (2011). An analysis of Twitter messages in the 2011 Toho earthquake. Arxiv preprint arXiv:1109.1618,

Duan Y., Jiang L., Qin T., (2010). An empirical study on learning to rank of tweets. *COLING Proceedings of the 23rd International Conference on Computational Linguistics Proceedings of the Conference*, 23-27 August, Beijing, China, p. 295-303, Tsinghua University Press.

Faiz R. (2006). Identifying relevant sentences in news articles for event information extraction. *International Journal of Computer Processing of Oriental Languages (IJCPOL), World Scientific*, vol. 19, n° 1, p. 1-19.

Huang C. (2011). Facebook and Twitter key to Arab Spring uprisings: report. http://bit.ly/1bh6jV6. [Online; accessed 28-August-2013].

James A., Papka R., Lavrenko V. (1998). On-line new event detection and tracking. In *SIGIR.*

Jiang L., Yu M., Zhou M., Liu X., Zhao T. (2011). Target-dependent Twitter sentiment classification. *Proc. 49th ACL: HLT*, 1, p.151-160.

Lin J., Snow R., Morgan W. (2011). Smoothing techniques for adaptive online language models: Topic tracking in tweet streams. *In KDD.*

Li J., Ritter A., Hovy E. (2014). Weakly supervised user profile extraction from twitter. *In Proc. ACL.*

Jones S., Walker K., Robertson S. (2000). A probabilistic model of information retrieval: Development and comparative experiments. *Information Processing & Management*, vol. 36, n° 6, p. 779-808,

Jure L., Lars B., Kleinberg J. (2009). Meme-tracking and the dynamics of the news cycle. In *KDD.*

Kwak H., Lee C., Park H., Moon S. (2010). What is Twitter, a social network or a news media? In *WWW'10.*

Lampos V, Cristianini N. (2010). Tracking the u pandemic by monitoring the social web. *In Cognitive Information Processing (CIP), 2nd International Workshop on*, IEEE, p. 411-416.

Mendoza M., Poblete B., Castillo C. (2010). Twitter Under Crisis: Can we Trust What We RT? In Proceedings of the *First Workshop on Social Media Analytics*.

Metzler D., Cai C., Hovy E. (2012). Structured event retrieval over microblog archives. In *Proc. of HLT-NAACL*

Osborne M., Moran S., McCreadie R., der Von Lunen A., Sykora M., Cano E., Ireson N., Macdonald C., Ounis I., He Y., Jackson T., Ciravegna F., O'Brien A. (2014). Real-time detection, tracking, and monitoring of automatically discovered events in social media. *In Proc. ACL*.

O'Connor B., Balasubramanyan R., Routledge B.R., Smith N.A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the *International AAAI Conference on Weblogs and Social Media*, p. 122-129.

Page L. (1997). "PageRank: Bringing Order to the Web" at the Wayback Machine (archived May 6, 2002), Stanford Digital Library Project, talk. August 18, (archived 2002)

Panem S., Gupta M., Varma V. (2014). *Structured information extraction from natural disaster events on twitter.* August.

Parikh R., Karlapalem K. (2013). Et: Events from tweets. In Proc. companion WWW.

Petrovi´c S., Osborne M., Lavrenko V. (2010). Streaming first story detection with application to Twitter. In *NAACL'10.*

Qu Y., Zhang C.P., Zhang J. (2011). Microblogging after a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake. *In Proceedings of the ACM 2011 conference on Computer supported cooperative work*, p. 25-34.

Ritter A., S. Clark, Mausam, and O. Etzioni (2011). Named entity recognition in tweets: An experimental study. *In Proc. EMNLP.*

Ritter A., Mausam, O. Etzioni, S. Clark (2012). Open domain event extraction from twitter. *In Proc. KDD.*

Robertson S. E., Spärck J.K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science* vol. 27, n° 3, p. 129. doi:10.1002/asi.463 0270302

Sakaki T., Okazaki M., Matsuo Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. In *WWW'10.*

Sankaranarayanan J., Samet H., Teitler B.E., Lieberman M.D., Sperling J. (2009). Twitterstand: News in tweets. In *GIS'09.*

Volkova S., Wilson T., Yarowsky D. (2013). Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In *Proc. ACL* short paper.

Xin Zhao W., Jiang J., He J., Song Y., Achananuparp P., Lim E., Li X. (2011). Topical keyphrase extraction from twitter. In *Proc. ACL'11.*

Yard S., Boyd D. (2010). Tweeting from the town square: Measuring geographic local networks. *In ICWSM'10.*

Zheng X., Zeng Z., Chen Z., Yu Y., Rong C. (2015). *Detecting spammers on social networks Neurocomputing*, vol. 159, 2 July, p. 27-34