Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Analysing and evaluating the task of automatic tweet generation: Knowledge to business

Elena Lloret *, Manuel Palomar

University of Alicante, Department of Computing Systems, Apdo. de correo 99, E-03080 Alicante, Spain

ABSTRACT

In this paper a study concerning the evaluation and analysis of natural language tweets is presented. Based on our experience in text summarisation, we carry out a deep analysis on user's perception through the evaluation of tweets manual and automatically generated from news. Specifically, we consider two key issues of a tweet: its informativeness and its interestingness. Therefore, we analyse: (1) do users equally perceive manual and automatic tweets?; (2) what linguistic features a good tweet may have to be interesting, as well as informative? The main challenge of this proposal is the analysis of tweets to help companies in their positioning and reputation on the Web. Our results show that: (1) automatically informative and interesting natural language tweets can be generated as a result of summarisation approaches; and (2) we can characterise good and bad tweets based on specific linguistic features not present in other types of tweets.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction, context and motivation

In the current digital knowledge society, the overload of information has become a problem to companies, which cannot cope with all the available information. As a consequence, companies may not be exploiting the Web, and taking advantage of it accordingly, thus affecting key aspects, such as their visibility, reputation, marketing campaigns, customer's feedback, etc. With the birth of the Web 2.0, there has been a shift in the way the information is produced and consumed by users and companies. The Web 2.0 has established a wide range of on-line mechanisms and platforms through which companies can obtain direct feedback from users. These mechanisms (e.g., reviews, social networks) allow users to freely express their comments about companies and the products/services they offer, thus requiring the effective management of a large number of adapted contents, formats, and interaction patterns [1]. Companies have envisaged the great potentiality of the communication through the Web 2.0 and even there have been attempts to integrate these channels into ERP platforms [2]. Moreover, companies have created their own social network profiles, e.g., in Facebook or Twitter, in order to increase their visibility, and maximise their interaction with customers.

With more than 241 million active users per month,[1] 184 million of which uses Twitter through their mobile device, and more than 500 million tweets daily,[2] Twitter[3] has become an excellent social media for on-line real-time news attention.[4] The length restriction imposed on tweets (140 characters) force messages to be concise, though it is also possible to link out to external information to enrich the tweet. Moreover, hashtags (e.g., #UA_Universidad) allow to categorise information, to identify the trending topics, and more importantly to enable a rapid on-line information flow. According to [3] one of the key success factors of Twitter is that it is an appropriate channel to communicate in short messages and share information regardless of time and place. Moreover, Twitter has become a means of electronic Word of Mouth communication (eWOM) [4], where one of its main usages is information distribution [5], that is spread very quickly reaching a high number of users in real time.

Companies are concerned about what their customers think about them, and in this manner, it is really important for them,

* Corresponding author.
E-mail addresses: elloret@dlsi.ua.es (E. Lloret), mpalomar@dlsi.ua.es (M. Palomar).

[1] http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/#.U0zftVfjDYM [last access June 2015].
[2] http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html [last access June 2015].
[3] http://twitter.com/.
[4] http://pando.com/2014/02/06/facebook-vs-twitter-who-wins-the-battle-for-our-social-attention/ [last access June 2015].

what and how information is delivered on the Web, since this may have a direct influence on their popularity and branding, affecting their positioning and reputation, or attracting/discouraging new potential customers. In the context of Twitter, the information to be expressed should be to the point, very clear and concise. This will benefit the impact on their business strategy, and will improve the relationship with customers, thus being able to personalise the information, as well as to improve marketing campaigns.

The level of maturity reached by state-of-the-art Natural Language Processing (NLP) techniques can support companies in delivering, managing and analysing on-line textual information. Current NLP applications, such as information retrieval, sentiment analysis or text summarisation could help companies to monitor relevant information about them, classify it, and obtain the key ideas. Specifically, when it comes to information delivery, text summarisation techniques could be used for automatically generating candidate micro- or ultra-concise summaries in the form of a tweet [6,7]. This task would be similar to headline generation [8,9], but in the current context of the Web 2.0, and in particular applied to Twitter, thus obtaining natural language tweets.

In the process towards the automatic generation of natural language tweets, a crucial stage is to know how users perceive them, and whether there are any linguistic features leading to the best and worst generated tweets. This will allow companies to be aware of the suitable language that would help to catch users' attention without negatively affecting its informativeness. Furthermore, the analysis of both issues would benefit communication strategies for companies, who need to be strategic in designing and executing their tweets [10].

Therefore, the main objective of this paper is to conduct a deep study on user's perception of tweets through the analysis of approximately 1600 tweets generated either by humans (i.e., manually), or by seven current text summarisers (i.e., automatically). Our study will be focused on analysing two key issues of a tweet: its informativeness and interestingness, as well as determining the set of linguistic features that contribute to produce good and not so good tweets. In particular, the research questions to answer are: (1) do users equally perceive tweets that have been manually generated in comparison to the automatic ones?; and (2) what linguistic features should or should not a good tweet have in order to be interesting, as well as informative for the user? Both, the identification of interesting and useful contents from large text-streams is a crucial issue in social media [11], and they have been widely employed for evaluation purposes in the context of Twitter [12–14]. Whereas for the first question, we use descriptive statistics for analysing in detail the assessment provided by different users, in the second question, we will collect a sample of several types of tweets and analyse in-depth their linguistic features, and main differences.

Moreover, our research work will be carried out from a multilingual perspective (for English and Spanish) with the purpose of determining if the language and the manner in which the tweet was generated have any influence on the user's perceptions. This intermediate research is framed within the overall research of automatically extracting and generating natural language tweets from external news documents talking about a company, product, etc. in order to help companies improve their positioning and reputation on the Web.

The results obtained from this research show that: (1) state-of-the-art summarisers are capable of generating good natural language tweets, that are informative as well as interesting, and that could be an alternative to manual generated tweets; and (2) it is possible to distinguish and characterise good and bad tweets based on different linguistic features that are not present in other types of tweets.

The remainder of this paper is organised as follows. Section 2 presents a literature review of the relevant research work related to the topic of this article. Section 3 explains the research methodology and questions that we want to analyse within the scope of this paper. It also provides information about the initial dataset that is used for conducting all the analysis, together with the NLP and statistical tools employed. Sections 4 and 5 show and discusses the findings and results of the analysis with regard to each of the proposed research questions. Finally, Section 6 presents the conclusions and final considerations, as well as several suggestions for future investigation.

2. Related work

Recently, Twitter has become a valuable source of data for research in NLP. The vast amount of data that each day millions of users and companies exchange through this platform has made it possible the analysis and processing of this textual genre, thus becoming necessary to analyse and exploit suitable techniques to filter out/discard irrelevant information, as well as to design effective and appealing communicative streams.

Natural language generation and text summarisation can help to achieve such challenging goals. The current difficulty associated to building natural language generation systems [15] and our considerable experience in text summarisation for extracting key ideas [16–18] has led us to address this study from a summarisation perspective rather than from natural language generation, even though generating natural language and applying it to Social Media (e.g., Twitter) would be our ultimate long-term goal.

In the literature, text summarisation techniques have been employed in the context of Twitter mainly for summarising tweet streams related to the same topic or event. Some examples of this type of approaches can be found in [19–23]. Different techniques such as phrase reinforcing algorithm or TF-IDF are employed, among others. Of all these approaches, we would like to highlight on the one hand, the approach proposed in [20], since it includes the novelty of taking into account not only the tweets themselves, but also the information linked by such tweets, and a combination of both of them. Whereas in most of the research works, tweets which come from user-generated content, are treated as they are, here, the authors apply a normalisation process to transform them into standard English. This is an important stage, because traditional NLP tools may fail when no standard language is provided [24]. Concerning the summarisation stage, the authors employed a concept-based optimisation approach for selecting informative sentences while minimising the redundancy. In this approach, the relevant sentences were determined based on the maximum number of concepts covered. The results indicate that the combination of normalised tweets and Web content was the best performing approach, beating the results obtained in [21]. On the other hand, [23] proposes an interesting novel aspect for Twitter event summarisation, which takes into account subjective information for generating a summary from different perspectives users may have on the same event. The authors focused on sport events, so they considered the fans' viewpoints for their approach. Given a set of tweets related to a sports event, the first step was to extract the ones referring to the teams involved in the event, and classify and group them with respect to the team it was supporting. Then, a topic detection algorithm was employed for returning up to ten topics for each event being considered, and they were compared with the comments related to the same event but belonging to an external information source (e.g., BBC comments). To select the closest topic, the cosine similarity measure was employed. Finally, for the selected topic, a small set of representative tweets of each of the groups was extracted.

Despite the number of research works producing summaries from Twitter data, there are only a few aiming at producing the opposite: a tweet as a summary of a heap of information. This task