An automated system for grammatical analysis of Twitter messages. A learning task application

M. Oussalah*,**, B. Escallier, D. Daher

University of Birmingham, School of Electronics, Electrical and Computer Engineering, Edgbaston, B15 2TT Birmingham, UK

A B S T R A C T

This paper describes an educational study involving the use of Twitter as a way to enhance High School students' interaction while improving the linguistic quality of their messages. For this purpose, an interactive system has been developed for Twitter collection and analysis from grammatical perspective. The automated system involves a comprehensive data normalization phase, which allows us to identify any unknown token, and a grammatical analysis system. The latter makes use of a logical reasoning on bi-gram token representation as well as a simple rule-based reasoning in case of named-entity detection. The developed system allows the user to perform spatial, topic-based or identity-based search functionalities. Besides, the system generates interrupt to moderator (s) together with some statistical parameters related to user activity as soon as a linguistic inconsistency has been detected in order to take relevant course of actions. The automated system allows us to identify both the text normalization issues and the grammatical inconsistencies. The latter makes use of logical reasoning using bi-gram Wikipedia matching. A statistical analysis of tweet messages gathered from students that took part to this study has been carried out. Besides, the contribution of the peers to the improvement of the linguistic quality of users' messages has been quantified and investigated. The study demonstrates the interest of the participants to this new learning experience and evaluates the influence of the peers on their writing skills. Especially, the visibility and noticeability of Twitter messages to a large audience have been found to contribute widely to raise students' awareness about the linguistic quality of their messages. The study has also revealed the predominance of the slang language in their daily Twitter writings. Such abbreviations have shown to pose the greatest challenge for any automatic text analysis. Similarly named-entity identification and handling have also been shown to be very challenging, especially, given the nature of Twitter messages where capitalizing is often employed for emphasize as well.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The boom in social networking sites in the recent decade has provided unprecedented amount of information to various users, especially, teenagers and student community. This, in turn, has influenced and challenged the standards approaches to learning. Among these social network tools, one shall focus on Twitter [6,17], which, since its introduction in 2007 has gained more than half billion active users and more than two billions tweet searches made every day! Typically, each tweet is made of up to 140 characters, which can also be embedded with precise geo-location information to carry short conversations [25]. This microblogging service has been found particularly useful in

disaster monitoring, opinion influence as demonstrated in events like the 2009 US Airways Flight 1549 incident, the 2009 Iranian presidential election, Japan's earthquake/Tsunami, recent Arab uprising [27] and happiness quantification [30]. In all previous examples, the efficiency of Twitter as a way to convey real time information about the status of the underlying scenes, and, possibly, influence the public opinion, was clearly demonstrated. A key feature of Twitter is its open access, such that whenever a user enables unrestricted access, his/her tweets can appear on the public timeline, a running stream of tweets by registered users observable by anyone [18]. Especially, this offers the users the possibility to follow other (more popular) users, e.g., stars, corporates, which, in turn, enable the users to get the last update information regarding key events or news that are of interest to the user. Nevertheless, the tweet size restriction of 140 characters, which becomes even smaller when markup syntax and URLs were used together with high frequency of tweet messages sent by active users, some of which were only spam, renders the usage of correct English very difficult and very challenging. On the other hand,

---

* Corresponding author. Tel.:+44 121 4143128; fax: +44 121 414 4291.
** The author is on leave with Oulu University, Centre for Ubiquitous Computing, 90014 Oulu.

E-mail address: M.Oussalah@bham.ac.uk (M. Oussalah).

the size restriction substantially increases the usage of abbreviations and slang words in tweet messages, some of which are ill-known by the community. This triggered an open debate whether Twitter can be of any benefit to learning community from linguistic perspective. Independent journalist Sirucek [33] claimed that Twitter is where "grammar comes to die". Indeed, many argued that the size restriction caused users to sacrifice normative grammar to communicate tweets as a sign of language degradation. In the same context, Borau et al. [8] found that the character limit and dictionary usage limited the use of communication strategies, thus, concluding Twitter does not help in building strategic competence. Typically, shortening message is achieved through intensive use of abbreviations, phonetic substitutions, deletion of selected words and characters, which obviously can seriously hamper the understandability of the content of the message. Studies such as [32] concluded that the high noise level in the content of Twitter messages renders the use of any standard natural language processing technique for content mining pretty challenging and almost impossible as they contain highly non-standard orthography that would make commonly employed automatic text processing tools nearly void. However such opinion is also counterbalanced by many other findings, which expressed strong positive influence of Twitter on learning task. Arguments for such trend are well founded and seem very intuitive. First, a general scepticism about new technology was often raised by some linguistic scholars who initially claimed, for instance, that telegraph, next phone messages would kill correct English, but none of these did happen [12]. Second, as a communication tool which can generate growth and value, traditional business writers have already accommodated the inherent restriction to send clear and well-understood messages to (potential) customers. Third, with the advances in e-learning technology and related academic programmes, the role of (online) recommended system is highly stressed [29]. Therefore interaction tools like Twitter would potentially be beneficial. Fourth, since the importance of peers-interaction in learning cannot be ignored and given the role of social network, including Twitter, in peers' formation and interaction, the impact of Twitter in students' academic performance is straightforward [11]. Indeed, peer networks have been demonstrated to promote prosocial behaviours, such as extracurricular participation and leadership [10]. Fifth, although it has been acknowledged that the research into the applicability of microblogging like Twitter in the language classroom is currently in its embryonic phase, its role in improving English for non-English speakers has been reported by many studies, see for instance [1] and references therein. Especially, it has been reported that given the short encoded messages, learners can follow daily conversations of native speakers as well as their peers, and engage in the underlying learning process. Antenos-Conforti [4] and Godwin-Jones [19] reported that social networking systems create new opportunities for teaching and learning given their ability to instantly engage the whole learning community. In this respect, one shall also mention the popular experience of Brazilian Red Balloon English School where pupils learned Grammar through identifying and correcting mistakes of their stars' (followers') tweet messages [9]. Another study performed by Grosseck and Holotescu [20] claimed that Twitter is the best place to practice a wide variety of expressions and fixed phrases. In the same spirit, Antenos-Conforti [4] argued that Twitter is effective because of two main reasons. First, a single tweet can trigger communication between socially connected users. Second, Twitter lowers affective filters so that any intentional thought can be transmitted. Ulrich et al. [36] analysed student interaction using a microblogging network designed for English language learning in a Chinese university. The authors found that students tended to interact with those of the same gender, to self-initiate replies to tweets, and to favor public communication. In his book on the use of Twitter for Japanese

learners, Homma [24] identified four reasons why Twitter is a good platform to study English. First, Twitters systems appears to be intuitive because of its simplicity and easy classification capability through hash-tags for instance, which provides a sound overall picture, as well as the possibility of initiating new communication based on previous state. Second, it increases users' motivation for participation to convey their daily desires and needs. Third, the size restriction makes it easy to follow up and replying accordingly as it takes only few seconds to do so. Fourth, it presents a visual medium analogy in the sense that after using Twitter for a while, people can always look back to what they have written, and see the extent of their improvements, if any.

Strictly speaking, the previous review indicates that the large extent of the studies involved in the use of Twitter for English studies have been primarily performed with non-native English speakers, although *bad* writing is found to be occurring with native English speakers as well. This partly motivates our current study to design and implement an educational case study involving Twitter with reasonably good English speakers. Besides, given the public nature of Twitter, our goal is to enlarge the domain of interaction of the learners beyond the school or classroom boundaries. For this purpose, an automated system has been designed to capture the geolocated tweets in the West Midland area, including the tweets submitted by the users not involved in this study, and automatically query the tweet messages for incorrect grammatical constructions and word inconsistencies. More specifically, in the light of our previous work [31], an open architecture allowing us to collect geolocated tweets together with all student participants is put forward. Next, standard natural language processing is extended through the use of a comprehensive normalization stage, while a bigram-representation model in conjunction with corpus matching were employed for grammatical analysis. Finally, the contribution of the peers' interaction in the learning process is quantified and discussed. Typically, the main research questions that we aim to investigate throughout this study are:

(i) Will students perceive Twitter as beneficial for practice, proceduralization and memorization of new grammatical constructions?
(ii) Does Twitter encourage new forms of linguistic constructions and language distortion?
(iii) How does interaction among learners take place?
(iv) How to design the software architecture to deal with grammatical analysis of the tweet messages?

Section 2 of this paper describes the experimental setup highlighting the participant population and the design experiment. Section 3 emphasizes the Twitter collection software. Next, basis of textual analysis system and text normalization are described in Section 4. Section 5 highlights the grammatical analysis system, while results in terms of data statistics, users' interaction and related discussions are reported in Section 6.

2. Experimental setup

2.1. Participants

The sample for this study consists of high school student population in West Midland area. We initially targeted schools with good achievement records in order to maximize our chance of recruiting good students. We meant by *good* students those who already master basic rules of English (normative) grammar and willing to positively collaborate in this study. The participants are selected on voluntary basis through a general questionnaire sent via Twitter, school internal notice board announcements, and link through Birmingham City Council. The students were asked to communicate the detail of their twitter accounts (usernames,