Ain Shams Engineering Journal (2014) **5**, 411–420

Ain Shams University

Ain Shams Engineering Journal

www.elsevier.com/locate/asej“
www.sciencedirect.com

ELECTRICAL ENGINEERING

Arabic summarization in Twitter social network

Nawal El-Fishawy [a], **Alaa Hamouda** [b], **Gamal M. Attiya** [a], **Mohammed Atef** [b,*]

[a] *Faculty of Electronic Engineering, Menoufia University, Menoufia , Egypt*
[b] *Faculty of Computer Engineering, Al-Azhar University, Cairo, Egypt*

Received 8 June 2013; revised 28 September 2013; accepted 22 November 2013
Available online 28 December 2013

KEYWORDS

Social networks;
Twitter;
Summarization;
Significance;
Similarity;
Feature selection

Abstract Twitter, an online micro blogs, enables its users to write and read text-based posts known as “tweets”. It became one of the most commonly used social networks. However, an important problem arises is that the returned tweets, when searching for a topic phrase, are only sorted by recency not relevancy. This makes the user to manually read through the tweets in order to understand what are primarily saying about the particular topic. Some strategies were developed for summarizing English micro blogs but Arabic micro blogs summarization is still an active research area. This paper presents a machine learning based solution for summarizing Arabic micro blogging posts and more specifically Egyptian dialect summarization. The goal is to produce short summary for Arabic tweets related to a specific topic in less time and effort. The proposed strategy is evaluated and the results are compared with that obtained by the well-known multi-document summarization algorithms including; SumBasic, TF-IDF, PageRank, MEAD, and human summaries.

1. Introduction

Twitter, the micro blogging site, has become a social phenomenon. It started in 2006 and became one of the most commonly used social networks. Twitter reached half billion user [1]. To help people who read Twitter posts or tweets, Twitter provides an interesting API that allows users to search for tweets that contain a topic phrase. A user can search for a topic phrase and retrieve a list of the most recent tweets that contain the topic phrase.

An important problem arises with Twitter is that the returned tweets are only sorted by recency, not relevancy. This behavior makes some difficulties in interpreting the retrieved results. Therefore, the user is forced to manually read through the returned tweets in order to understand what users are primarily saying about the particular topic. This process requires more effort and time from the Twitter users. To overcome this problem, tweets summarization should be performed automatically for the purpose of saving users time and effort. So, a summarization system is required to automate this process.

Several strategies were developed for automatic summarizing micro blogs. However, most of the proposed strategies are developed for summarizing English tweets [2,3]. But, no algorithms are developed for summarizing Arabic micro blogging posts and more specifically Egyptian dialect summarization,

* Corresponding author. Tel.: +20 0111 161 2226.
E-mail addresses: nelfishawy@hotmail.com (N. El-Fishawy), dr.alaa.hamouda@gmail.com (A. Hamouda), gamal.attiya@yahoo.com (G. M. Attiya), atef_plc_mox@yahoo.com (M. Atef).

Peer review under responsibility of Ain Shams University.

Production and hosting by Elsevier

2090-4479 © 2013 Production and hosting by Elsevier B.V. on behalf of Ain Shams University.
http://dx.doi.org/10.1016/j.asej.2013.11.002

although the Arabic language has become the sixth most widely used language on the Twitter social network.

This paper presents a machine learning based summarization system for summarizing Arabic posts in Twitter social network. The purpose is to produce short summary for Arabic tweets related to a specific topic in less time and effort. In the proposed strategy, the problem is formulated as a regression problem, not a binary classification based problem. That is, instead of classifying the tweets to be important and not important, each tweet is given a score that determines whether this tweet may candidate in the summary or not. This makes the system to generate the summary according to the user predefined compression rate. The proposed strategy is evaluated and the results are compared with that obtained by the well-known multi-document summarization algorithms including; SumBasic [4], TF-IDF [2], PageRank [5], MEAD [6], and human summaries.

The rest of this paper is organized as follows. Section 2 presents an overview on related work. Section 3 states the summarization problem. The proposed system is described in Section 4 while the implementation of the proposed system is presented in Section 5. The evaluation and experimental results of the proposed system are discussed in Section 6. Finally, the conclusion is listed in Section 7.

2. Related work

Automatically summarizing micro blogging posts is a new area of research. A number of algorithms have been developed for English document summarization during recent years. MEAD [6] is a well-known flexible and extensible multi-document summarization system and was chosen to provide a comparison between the more structured document domains in which MEAD works fairly well. The platform implements multiple summarization algorithms such as position-based, centroid-based, largest common subsequence, and keywords. In [7], a LexRank algorithm is developed for computing the relative importance of sentences or other textual units in a document or a set of documents. It creates an adjacency matrix among the textual units and then computes the stationary distribution considering it to be a Markov chain. In [4], a SumBasic algorithm is proposed for document summarization. In this system, words that occur more frequently across documents have higher probability of being selected for human created multi-document summaries than words that occur less frequently. In [8], graph is applied for representing the structure of the text as well as the relationship between sentences of the document. Sentences in documents are presented as nodes. The edges between nodes illustrate connections between sentences. These connections are introduced by similarity relation between contents. The similarity between two sentences is calculated and each sentence is scored. All the scores for one sentence are combined to form a final score for each sentence. When the graph is processed, the sentences are categorized by their scores and sentences in higher orders are chosen for final summary. Other graph-based Phrase Reinforcement algorithm is developed in [3]. The algorithm first finds the most common phrase on one side of the search phrase, selects those posts that contain this phrase, and then finds posts with the most common phrase on the other side as well.

In [9], multi-document summarization system is developed for the Web context. The system is useful in combining information from multiple sources. Information may have to be extracted from many different articles and pieced together to form a comprehensive and coherent summary. One major difference between single document summarization and multi-document summarization is the potential redundancy that comes from using many source texts. The solution presented in [9] is based on clustering the important sentences picked out from the various source texts and using only a representative sentence from each cluster.

In [2], a hybrid TF-IDF algorithm developed. The idea of the algorithm is to assign each sentence within a document a weight that reflects the sentence's saliency within the document. The sentences are ordered by their weights from which the top sentences with the most weight are chosen as the summary. In order to avoid redundancy, the algorithm selects the next top tweet and checks it to make sure that it does not have a similarity above a given threshold with any of the other previously selected tweets because the top most weighted tweets may be very similar. Another method in [2] collects a set of Twitter posts, clusters the tweets into a number of clusters based on a similarity measure and then summarizes each cluster by picking the most weighted post as determined by TF-IDF algorithm.

Comments summarization over a collection of YouTube videos are studied in [10]. The system starts by clustering the comments and selecting the most representative comments of each cluster. Then, a precedence-based ranking framework is used for automatically selecting informative user-contributed comments.

Recently, some summarization systems are developed for Arabic text [11]. Rhetorical Structure Theory (RST) is one of the leading theories in computational linguistics [11]. It improved the ability of extracting the semantic behind the processed text. The applicability of RST to process and understand texts has been studied in Arabic language to extract the text structure, and then extract the semantic from this structure. In [12], an automatic extractive Arabic summarization system called *Ikhtasir* is developed. It integrates the advantages of an RST-based system with a scoring scheme which is a variant of the FarsiSumscoring formula. In [13], the summarizer, *Lakhas*, was developed using extracting techniques to produce ten words summaries of a new articles. *Lakhas* first summarizes the original Arabic document and then applies Machine Translation (MT), translating the summary into English. These systems support the single document summarization. A multi-document summarization system for Arabic comments was developed in [14].

3. Problem statement

Summarizing micro blogs can be viewed as an instance of the more general problem of automated text summarization, which is the problem of automatically generating a condensed version of the most important content from one or more documents.

The summarization problem can be simply described as follows: given a set of tweets that are related to a common search phrase (e.g., a topic), the problem is how to generate a