Expert Systems with Applications 42 (2015) 7148–7156

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A novel concept-level approach for ultra-concise opinion summarization

Elena Lloret *, Ester Boldrini, Tatiana Vodolazova, Patricio Martínez-Barco, Rafael Muñoz, Manuel Palomar

Department of Software and Computing Systems, University of Alicante, E-03080 Alicante, Spain

ARTICLE INFO

Article history:
Available online 23 May 2015

Keywords:
Text summarization
Ultra-concise opinion summarization
Electronic Word of Mouth
Natural language generation

ABSTRACT

The Web 2.0 has resulted in a shift as to how users consume and interact with the information, and has introduced a wide range of new textual genres, such as reviews or microblogs, through which users communicate, exchange, and share opinions. The exploitation of all this user-generated content is of great value both for users and companies, in order to assist them in their decision-making processes. Given this context, the analysis and development of automatic methods that can help manage online information in a quicker manner are needed. Therefore, this article proposes and evaluates a novel concept-level approach for ultra-concise opinion abstractive summarization. Our approach is characterized by the integration of syntactic sentence simplification, sentence regeneration and internal concept representation into the summarization process, thus being able to generate abstractive summaries, which is one the most challenging issues for this task. In order to be able to analyze different settings for our approach, the use of the sentence regeneration module was made optional, leading to two different versions of the system (one with sentence regeneration and one without). For testing them, a corpus of 400 English texts, gathered from reviews and tweets belonging to two different domains, was used. Although both versions were shown to be reliable methods for generating this type of summaries, the results obtained indicate that the version without sentence regeneration yielded to better results, improving the results of a number of state-of-the-art systems by 9%, whereas the version with sentence regeneration proved to be more robust to noisy data.

1. Introduction and motivation

Each day more, within the framework of the current digital society of the Web 2.0, most of the information users produce can be found on-line and a great part of it is freely accessible. Thus we all have huge amounts of terabytes of user-generated content (e.g. in blogs, fora, review sites, social networks, etc.) that anybody can freely use (consciously or not) in our daily decision-making process. An example of this could be the typical situation when we are to buy a product; we first check the Web to find opinions, comments, information, on that item expressed by users that already have such product before deciding to acquire it or not. This phenomenon can be recognized as the modern (using data on the Web) version of the ancient Word of Mouth (WOM) originally developed in the streets many years ago when people relied on the opinions of others about a specific issue. However, the electronic Word of Mouth (eWOM) main difference with respect to the traditional one is that nowadays information is spread all over the world, in many languages, about a wide range of topics, services, places, etc., and there are vast amounts of information referring to the same topic. On the one hand, it is very useful to have all this information available; however, on the other hand, there is the risk in which important and key information may be lost or passed unnoticed, since it is impossible to manually manage all the information in an effective and efficient manner. In the current society there is a shortage of time for spending it and have an in-depth look at each and every single entry.

Given this context and these user needs, Natural Language Processing (NLP) provides the appropriate mechanisms that allow people to automatically deal with Web textual information, speeding up and simplifying the tasks of retrieving, extracting, classifying and summarizing, so that they can have just the necessary information users may need at a specific moment and in real time.

In this process and to obtain a high quality and reliable results, text summarization is a crucial process towards this goal; the objective of text summarization is to automatically produce a brief and concise fragment of text, containing only the essential and

* Corresponding author.
E-mail addresses: elloret@dlsi.ua.es (E. Lloret), eboldrini@dlsi.ua.es (E. Boldrini), naiadhe@gmail.com (T. Vodolazova), patricio@dlsi.ua.es (P. Martínez-Barco), rafael@dlsi.ua.es (R. Muñoz), mpalomar@dlsi.ua.es (M. Palomar).

http://dx.doi.org/10.1016/j.eswa.2015.05.026
0957-4174/

most relevant information about the sources that have been used (Spärck Jones, 1999). When it comes to text summarization, a wide range of summary types can be found (e.g., multi-document, indicative, informative, opinion-based, etc.). Thanks to this, this task becomes customizable to the users needs. In recent years, new summary types are emerging; one of this new types is ultra-concise opinion summarization, which poses greater challenges to the task, since it aims to just capture the main essence of an opinionated document with a few words. Taking into consideration the summary size, this task could be viewed as the task of headline extraction, but its main difference lies in the type of text where it is applied (Web 2.0 platforms, such as microblogs), thus having to address the particularities of the Web 2.0 textual genres, and meet the strict length restriction these services may have (e.g., 140 characters long).

The motivation of our research relies on the fact that at present microblogs are the trendy textual genre of the Web 2.0 that is increasing more in comparison with others and consequently, this communication channel is being widely used by different user types and considered as a point of reference for many users (Abdullah, Nishioka, Tanaka, & Murayama, 2014; Hennig-Thurau, Wiertz, & Feldhaus, 2014; Kim, Sung, & Kang, 2014). This implies that the generation of such brief summaries has the potential of reaching a huge number of users.

Therefore, taking into account the importance of the eWOM and the potentiality of NLP technologies in the Web 2.0 scenario, this research work proposes the development and analysing of an innovative text summarization approach able to generate ultra-concise opinion abstractive summaries in the form of a tweet starting from different sources (reviews and microblogs). It is worth stressing that the main contribution of this article lies in the manner that this type of summaries is generated, since our approach performs abstractive summarization[1], which is one of the most challenging issues for current summarization approaches.

In this manner, our proposed approach goes beyond the mere identification and extraction of the most relevant sentence by including a sentence simplification and language regeneration stage that allow us to effectively condense the information. We thus identify the key parts of the sentences and then, generating the information, we are able to minimize the effects on its cohesion and coherence. In order to verify the effectiveness of the approach proposed, we tested it with two macro topics that are usually discussed considerably in the framework of the Web 2.0: technology and motor. After that we perform a qualitative evaluation to determine the appropriateness of our approach with respect to several quality criteria. The results show that the proposed simplification and regeneration stages are an added value to the current summarization strategies, having been proven to generate appropriate ultra-concise opinion summaries.

In this research article, our approach is presented as a single expert intelligent system, and evaluated intrinsically on its own, focusing on the quality of the opinion summaries that is able to generate. Undoubtedly, opinion summarization is one of the most valued and powerful NLP technologies. This type of technologies are acting as expert systems, given their power to advice or assist humans and/or other automatic processed when it comes to make any decision (Darling, 2014; Wang, Zhu, & Li, 2013), so more and more they are becoming essential in our current society (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014). Therefore, having shown the appropriateness of our method, this could be integrated and exploited as part of more complex applications, such as ERPs and other business intelligence process to help companies to deal with the eWOM, for instance, by disseminating ultra-concise messages through different channels to advertise and attract more clients or to make its potential customers aware of the high reputation of their products (Pai, Chu, Wang, & Chen, 2013). On the other hand, from a user point of view, this type of application could be very useful for providing the most outstanding feature for a product, service, etc. based on the on-line available information related to it.

As it can be seen, the technology proposed in this article, could be easily integrated in a real-life application that will allow users to save their time and effort since the system will do the job automatically analysing the texts selected and summarizing their content in a reliable way.

The remaining of this article is organized as follows. Section 2 covers the related work and puts our work in perspective. Section 3 explains our proposed approach for generating ultra-concise summaries, together with the tools and resources employed. Section 4 reports the corpus developed and the experimental framework. Section 5 contains the evaluation carried out and the results obtained providing a comparison with state-of-the-art approaches, as well as discussing the potentials and limitations of the approach. Finally, Section 6 concludes the article and outlines future work.

2. Related work

Text summarization was initiated by Luhn (1958) and Edmundson (1969) when analysed the first approaches to generate summaries automatically. Since then, most of the research in text summarization has focused on summarizing news documents, exploiting a wide range of techniques (Barzilay & McKeown, 2005; Huang, Wan, & Xiao, 2013; Kabadjov, Steinberger, & Steinberger, 2013; McKeown et al., 2002; Sarkar, 2012).

However, during the last decade, the expansion of the Web, and in particular, the birth of the Web 2.0 has raised the need for new types of summaries tightly related to the opinions users express about a specific topic, product, service, etc. In this context, new textual genres, such as blogs, reviews, or social networks have emerged, providing new forms of communication that highly differ from the traditional ones (i.e., news documents). Thus, the consequence is that there is the need to face new challenges due to different linguistic phenomena. For instance, the Web 2.0 is characterized by its informal nature (Mosquera & Moreda, 2012) that means new challenges for existing NLP tools and methodologies for language treatment.

Given the above context, there has been much interest in proposing summarization approaches for these new scenarios in recent years. Opinion or sentiment-based summarization are at a general level, a new type of summary that aim to provide the most relevant pros and cons of a specific product, service, etc. In Pang and Lee (2008), the importance of this type of summaries is acknowledged due to the vast amounts of opinions stated on the Internet. Having made an extensive analysis of the most relevant research on the field, we realised that almost nothing has been done on automatic summarization from different textual genres simultaneously and also with a content generation aspect. And here is the novelty of our research. We work with different textual genres (blogs and microblogs), retrieve data from these sources on a specific topic and we are able to generate a tweet that contains the most relevant content. Thus, the novelty is twofold: on the one hand we are able to treat texts from different sources, that present different linguistic challenges and on the other hand, we generate the tweet that is a summarization of the most relevant content.

[1] Abstractive summarization differs from extractive in that the summary does not limit to the selection and output of the most relevant sentences. In abstractive summarization a sentence transformation process is involved (e.g. sentence compression).