

TravelSum: A Spanish Summarization Application focused on the Tourism Sector

TravelSum: Aplicación de generación de resúmenes enfocado al sector del turismo

Alberto Esteban, Elena Lloret

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
{aesteban,eloret}@dlsi.ua.es

Abstract: This demo showcases a Web application that allows users to easily obtain a summary that is automatically generated taking into account the information provided by other users on the Internet. The application integrates several types of summaries, outlining the most relevant positive opinions, negative and both about restaurants and hotels. In addition, it provides multimodal information, such as graphics, maps or pictures. The results obtained from an on-line questionnaire conducted with real users reveals the potential and usefulness of such an application in the current society.

Keywords: Natural language processing, web application, text summarisation, multi-genre, abstractive summarisation, tourism

Resumen: Esta demostración presenta una aplicación web a través de la cual los usuarios pueden obtener un resumen generado automáticamente en español teniendo en cuenta la información proporcionada por otros usuarios en la Web. La aplicación integra varios tipos de resúmenes en los que se describen las opiniones positivas, negativas y neutrales sobre hoteles y restaurantes, junto con información multimodal, como gráficos, mapas o imágenes. Los resultados obtenidos a partir de un cuestionario realizado a usuarios reales revelan el potencial y la utilidad de tal aplicación en la sociedad actual.

Palabras clave: Procesamiento del lenguaje natural, aplicación web, generación de resúmenes, multigénero, resúmenes abstractivos, turismo

1 Introduction and Motivation

The Web is a valuable mechanism when users have to make a decision about the purchase of a product, the hiring of a service, the booking of a hotel, going to a restaurant, visiting a place, etc. It is very common for users to search and rely on others' opinions, resulting in the so-called Electronic Word of Mouth (Cheung and Thadani, 2012), which is gaining more and more importance, partly evidenced by the increasing number of review websites and their popularity. We can find either general review sites, e.g., TripAdvisor¹, or more specialized ones, such as Rotten

Tomatoes², Consumer Reports³ or Zomato⁴.

At the same time, information on the Web increases at an exponential rate since users act as digital content creators as well, thus being more and more difficult to read and process all this information in an efficient and effective manner. Taking this fact into consideration, what would a user prefer: to read 1,000 opinions about the product or service a user is interested in, or to have a tool that automatically processes all these opinions and provides a brief summary? In the former situation, a user would be only able to read a limited number of them, which may

¹<https://goo.gl/WBngac>

²<https://goo.gl/xrN8d>

³<https://goo.gl/MIUQF>

⁴<https://goo.gl/V23Vsa>

result in biased and not well-informed decisions (López-López and Parra, 2016). In the latter case, the system could be updated as long as new opinions are found regardless the information source, and the summary could be personalised with respect to the users' interests.

Given this context, this paper presents a Web application focused on the tourism sector. The proposed application works for Spanish and provides three types of abstractive summaries automatically generated from users' opinions about hotels and restaurants: i) a summary with the best-values aspects; ii) another with the worst-valued aspects; and iii) a final one with a combination of both to provide both the cons and pros of the hotel/restaurant. In addition, the resulting summaries are combined with supplementary multimodal information, such as maps, graphics, and pictures to provide users with extra information.

Although opinion summarization has been previously addressed in the literature, this has been focused only on one type of information source, providing only one type of summary (either positive, or negative) and mainly from an extractive point of view (Suzuki, 2012; Di Fabrizio, Stent, and Gaizauskas, 2014; Gerani et al., 2014; Ding and Jiang, 2015). To the best of our knowledge, our proposed Web application is the first one developed for Spanish that: i) integrates information from multiple and different sources; ii) generates abstractive summaries from different perspectives; and iii) provides a ready-to-use graphical interface that includes multimodal information.

This type of application can benefit several types of users. On the one hand, it can be used by users who want to easily and quickly summarize opinions from the Web about the tourism sector (i.e., hotels and restaurants), without having to read millions of them. On the other hand, the types of summaries created from different perspectives can be used by the companies (or managers) in charge of such hotels/restaurants for carrying out SWOT analysis, to identify strengths and weaknesses of their services, and be able to act accordingly.

2 TravelSum Web Application

Our proposed application is divided in two parts. The first part, the back-end, deals

with the retrieval, extraction and transformation of the information. The second part, the front-end, is related to the interface and service's options.

2.1 Back-end

The back-end was developed in Java and MySQL was used to store all the information. For the summarization process, we employed some natural language processing techniques and tools for retrieving and extracting information, as well as carrying out a linguistic analysis of the documents.

The whole process to create the summaries is depicted in Figure 1.

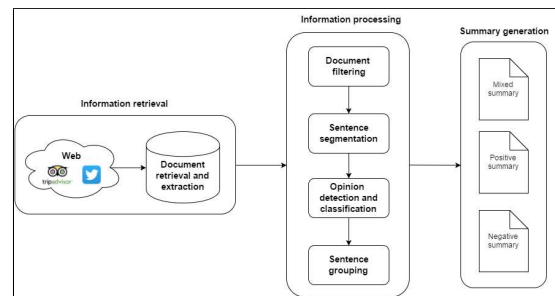


Figure 1: Overview of the summarization process

Next, each stage is briefly explained; more details can be found in (Esteban and Lloret, 2017).

- Document retrieval and extraction: The goal of this module is to retrieve all the necessary information for the creation of the summaries. On the retrieval phase, we extracted the information from: TripAdvisor, because is the world's largest travel site⁵ and Twitter, since it is one of the most popular social networks⁶. The retrieval was done developing specific crawlers for each of these sources.
- Document filtering: The goal of this stage is to discard information that is not in Spanish, or that does not give an opinion about the hotel/restaurant. This is very important in the case of information from Twitter, because the social network is more general and is not only focused on opinions. Different rules were developed for addressing this issue.
- Sentence segmentation: The aim of this stage is to split the reviews into sen-

⁵<https://goo.gl/CLFBj0>

⁶<https://goo.gl/aEbs0z>

tences. To do this process we used the Stanford’s software⁷.

- Opinion detection and classification: The goal of this stage is to classify the sentences with respect to their sentiment (neutral, positive and negative). For this, we relied on an existing tool⁸.
- Sentence grouping: The aim of this stage is to group similar sentences to avoid introducing redundancy in the summaries. To compute the similarity between sentences, the cosine metric⁹ was used.
- Summary generation: The aim of this module is to create the final summaries. We can divide this module in two tasks. In the first one, we rank the sentences of the groups in order to choose which sentences will be part of the summary. After a preliminary summary is created, we perform a post-processing task to improve the summaries coherence and readability. On the one hand, we employ some techniques and rules to change the summary to an impersonal style. On the other hand, we add some linking phrases in order to improve the cohesion of the summaries. Finally, we obtain three types of summaries, a mixed summary that shows good and bad aspects of the hotel/restaurant; a positive summary with the best aspects according to the customers, and a negative summary that highlights the worst aspects, i.e., the issues the customers did not like.

2.2 Front-end

The front-end is a Web application¹⁰ that allows any user to search a hotel/restaurant and get the summary generated from the reviews and tweets. An example of a summary together with the interface is shown in Figure 2. The technologies used for developing the application include HTML, CSS and JavaScript. Moreover, Bootstrap¹¹ was used to obtain a good design. The information about the hotels/restaurants was enriched with multimodal elements, such as graphics created with ChartJS¹², where we show

the ratings of some aspects extracted directly from TripAdvisor compared with the averaged aggregated score of the same aspects in such city establishments; a picture of the hotel/restaurant extracted from the Flickr service¹³; and a map with the location through Google Maps API¹⁴.

3 Evaluation and Results

To evaluate our application (some generated summaries, as well as the interface), we carried out a user evaluation by means of a questionnaire. We created a pool of 15 questions about different topics that were divided in 3 categories: i) the way the users searched touristic information on the Web; ii) the usefulness of the application together with its accessibility; and iii) their opinion about the generated summaries. 41 people answered the questionnaire, and we next discuss the results.

Concerning the first category, we asked the users the way they looked for information about hotels/restaurants, and we obtained that more than 95% used the Internet, using services as TripAdvisor, forums and specialized pages. Further on, we asked for the reliability of the information available on TripAdvisor or specialized pages. The result was that more than 97% thought that this information was very useful. In the case of Twitter, approximately 40% thought that the information in this social network was not useful.

Focusing our attention on the results of the questions related to the second category, we obtained that all the users (100%) considered useful to see a summary of the opinions from the travellers so that the best and the worst things could be highlighted, using the information available in TripAdvisor and Twitter. Another question evaluated the interface in terms of accessibility and usefulness. The interface obtained a great acceptance, around 8.64 out of 10.

In the third category of questions, we asked about the generated summaries. People evaluated the summaries in terms of coherence, utility and the presence of syntactical errors. The mark of the summaries was good; they obtained 7.45 out of 10.

Finally, it is important to note that there was one question related to the Turing test.

⁷<https://goo.gl/fUxNLJ>

⁸<https://goo.gl/SoQ4G0>

⁹<https://goo.gl/3rKzVT>

¹⁰<http://travelsum.gplsi.es/>

¹¹<https://goo.gl/03ur8>

¹²<https://goo.gl/zzG8m>

¹³<https://goo.gl/uDJQ1h>

¹⁴<https://goo.gl/5x4wT>

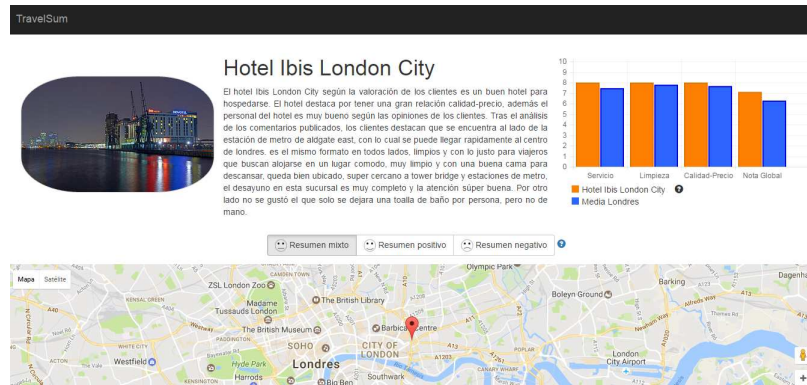


Figure 2: TravelSum Web application

We asked people how they thought summaries were created. The result was that 30% could not distinguish whether the summaries were produced by a computer or a person.

4 Conclusion and Future Work

We presented TravelSum, a Web application capable of producing three types of abstractive summaries about hotels and restaurants from users' opinions available on the Web.

The results of the evaluation showed that the summarization service is very useful. Although at the moment it only works for hotels or restaurants, it can be adapted to other domains or topics, such as products or shops.

As future work, we want to adapt the tool to work with other languages, e.g. English, as well as debugging possible grammatical errors on the summary process generation.

Acknowledgements

This research has been funded by the Valencian Government through the project PROMETEOII/2014/001, and by the Spanish Government through projects TIN2015-65100-R and TIN2015-65136-C2-2-R.

References

- Cheung, C. M. and D. R. Thadani. 2012. The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decision Support Systems*, 54(1):461 – 470.
- Di Fabbri, G., A. Stent, and R. Gaizauskas. 2014. A hybrid approach to multi-document summarization of opinions in reviews. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*,

pages 54–63, Philadelphia, Pennsylvania, U.S.A., June. Association for Computational Linguistics.

- Ding, Y. and J. Jiang. 2015. Towards opinion summarization from online forums. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 138–146, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, Bulgaria.

- Esteban, A. and E. Lloret. 2017. Propuesta y desarrollo de una aproximación de generación de resúmenes abstractivos multigénero. *Procesamiento del Lenguaje Natural*, 58:53–60.

- Gerani, S., Y. Mehdad, G. Carenini, R. T. Ng, and B. Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar, October. Association for Computational Linguistics.

- López-López, I. and J. F. Parra. 2016. Is a most helpful ewom review really helpful? the impact of conflicting aggregate valence and consumer's goals on product attitude. *Internet Research*, 26(4):827–844.

- Suzuki, Y. 2012. Classifying hotel reviews into criteria for review summarization. In *Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology*, pages 65–72, Mumbai, India, December. The COLING 2012 Organizing Committee.