Improving Opinion Summarization by Assessing Sentence Importance in On-line Reviews

Rafael T. Anchiêta, Rogério F. de Sousa, Raimundo S. Moura, Thiago A. S. Pardo

[1] Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
São Carlos/SP, Brasil

rta@usp.br, rfigsousa@ifpi.edu.br, rsm@ufpi.edu.br, taspardo@icmc.usp.br

Abstract. *This paper describes an approach for improving a state of the art opinion summarization method, incorporating the assessment of sentence importance in on-line reviews. We compare the enriched method to its original version and show that we significantly outperform it, producing more informative summaries.*

1. Introduction

According to [Conrad et al. 2009], opinion summarization is the task of automatically generating summaries for a set of opinions about a specific target. Such task is useful for several purposes. Imagine, for instance, a user that needs to decide which smartphone to buy. Summaries of hundreds of opinions about the several aspects of each device would be very helpful in this decision.

Extractive summarization methods are currently the most adopted ones. They create summaries by selecting and juxtaposing representative sentences from the source documents/opinions, using features as sentence position, word frequency, opinion size, and so on. These approaches usually produce reasonable summaries, but the results are still far from ideal (which would be summaries that humans produce). In this context, there is room for improvement.

In this paper, we enrich a state of the art extractive opinion summarization method, incorporating knowledge about the importance of the sentences in the on-line source opinions (in reviews) in order to better select the content to compose the summary. The summarization method is the one proposed by [Condori and Pardo 2017], which has already outperformed other well-known aspect-based methods in the area. The sentence importance assessment is carried out by the TOP(X) method [de Sousa et al. 2015]. We evaluate the enriched summarization method on part of a corpus related to electronic products, and measure summary informativeness using the traditional ROUGE measure [Lin 2004]. Our results show that our enriched summarization method significantly outperforms the original method.

The remaining of this paper is organized as follows. Section 2 briefly introduces some related work and the original summarization method of [Condori and Pardo 2017]. In Section 3, we introduce the TOP(X) method, which was used to enrich the summarization method. In Section 4, we present the corpus used in our evaluation. Section 5 reports the achieved results. Finally, some conclusions are presented in Section 6.

2. Related work on opinion summarization

Although the area of opinion summarization is relatively new, there are already many methods for performing the task. We briefly introduce here the most relevant related work.

[Beineke et al. 2003] was the first to tackle opinion summarization. The authors proposed an extractive method for selecting a single sentence that reflects the full opinion of its author. In order to select the most representative sentence, the authors used machine learning algorithms, using word frequency and sentence position as features.

[Hu and Liu 2004] proposed a summarization architecture organized in three steps: (i) identification of the evaluated aspects in the reviews, (ii) classification of each aspect as being positively or negatively evaluated in the review, and (iii) generation of the summary. The system receives the name of the product of interest as input and the web pages with opinions, producing an structured summary, which is a summary that shows relevant positive and negative sentences (indicating the total amount of sentences) for each relevant aspect of the product.

[Condori and Pardo 2017] developed and compared extractive and abstractive methods for opinion summarization. The extractive method, named Opizer-E, extracts a few sentences on the main aspects of the entity under evaluation. For this, the authors group similar sentences and rank them. They used the position of the sentence in the review and the proximity of the aspects to their qualifiers to rank the sentences. The proposed abstractive method uses templates to generate summaries, reusing text passages from the opinions. The authors used the OpiSums-PT corpus (in Brazilian Portuguese) to evaluate their methods, outperforming some previous approaches to the task.

3. The TOP(X) method

In order to improve sentence selection to compose opinion summaries, we used the TOP(X) method [de Sousa et al. 2015]. The TOP(X) method estimates the degree of importance of sentences in on-line reviews using a fuzzy inference system that has three input variables: author reputation, number of tuples <*aspect, qualifiers*>, and percentage of correctly spelled words. Based on these, sentence importance is given in a range of 0 to 10.

According to [Jindal and Liu 2008, Xu 2013], author reputation is relevant to estimate validity and importance of reviews. The hypothesis is that people who regularly write messages have a better reputation than occasional authors. Thus, the method counts the number of reviews for each author in the corpus to find his/her reputation.

In reviews, it is usual to find the cited aspects near to their respective qualifications, for example, in "*the screen is very good*", where the aspect is "screen" and the qualification is "very good". In this context, the method extracts the tuple <*screen, very good*> by identifying the subject and the predicate in the sentence.

Some authors indicate that misspelled words become a problem when reviews are analyzed in sentiment analysis tasks [Tumitan and Becker 2013, Paltoglou and Giachanou 2014]. Thus, the more correct a review is, the more relevant it should be. To calculate the percentage of correct words, the method consults

Wiktionary[1] for the Portuguese language.

Having the above values, the TOP(X) method associates to each input variable three possible linguistic values: low, medium and high. For output value, four linguistic values were used: excellent, good, sufficient, and insufficient. These values were set in a discourse universe $U[0, 10]$. In order to map these input values to output values, a fuzzy rule base composed of a set of production fuzzy rules was used. The typical structure of a fuzzy rule is: **IF**$(x = a)$**AND**$(y = b)$**AND**$(z = c)$, **THEN**$(k = d)$, where x, y and z are the input variables and k is the output variable. Then, for instance, for the input values *low*, *low* and *low*, the output k would be *insufficient*.

The TOP(X) method was evaluated on a sentiment classification task. Using the method in order to select the best sentences in a corpus, the authors improved a lexicon-based classification in approximately 10% and 20% of f-measure to positive and negative sentiments, respectively.

4. The corpus

The OpiSums-PT corpus [Lopez et al. 2015] contains groups of reviews and their manually produced summaries for two domains: books and electronic products. The first domain is composed by reviews from the ReLi corpus [Freitas et al. 2013], consisting in a collection of opinions about 13 famous books. The second domain is composed by reviews of 4 electronic products collected from Buscapé[2] website. The sentences in the corpus were also manually annotated with their polarity and aspects.

In this paper, we use 4 groups of reviews of the electronic product domain. Each group, with 10 reviews, contains 5 extractive and 5 abstractive manually produced summaries. Each summary is composed by 100 words, approximately. We use only 4 groups because they were the only ones with the necessary information to TOP(X) method to work.

5. Experiments and results

We used the TOP(X) method to estimate the importance of the sentences of the OpiSums-PT corpus, i.e., for each sentence in the corpus the method assigns an importance value in a range of 0 to 10. Then, inside the summarization method of [Condori and Pardo 2017], such values are used to select which sentences to include for each aspect in the summary.

We generated summaries for the electronic products domain in four groups, regarding the products Galaxy SIII, Iphone 5, Samsung Smart TV, and LG Smart TV. Figure 1 shows an example of an automatically generated summary. It is possible to see two aspects (the entity itself - which is generally referred as an aspect in the area - and "price"), with positive and negative sentences for each one (accompanied by the total number of existent sentences for each case).

In order to evaluate the generated summaries, we used the traditional ROUGE measure [Lin 2004]. ROUGE automatically compares the n-grams in an automatic summary to the ones in one or more human summaries (the reference summaries), producing precision, recall, and f-measure results. It is considered an summary informativeness

[1] `http://pt.wiktionary.org`

[2] `http://www.buscape.com.br/`

measure and, as its authors have shown, it is as good as humans in ranking summaries. ROUGE measure is widely used in the evaluation of automatic summaries because it is reliable and quickly and easily applicable. In our evaluation, we used only the extractive reference summaries in the corpus (since our method produces extractive summaries).

Figure 1. An example of an automatically generated opinion summary

Aspect: LG Smart TV
Positive sentences: 18
– What I liked: Image quality; 3D, Dual Player, Support for various video formats; Point-type remote control; Voice recognition; WiDI; Design.
Negative sentences: 13
– The quality drops a lot when the Dual Player function is used, however you can get fun.
Aspect: Price
Positive sentences: 1
– Excellent price and quality.
Negative sentences: 4
– Normal for its expensive price.

We compared our enriched method with the original one of [Condori and Pardo 2017] - the Opizer-E method, that is the state of the art in opinion summarization for the Portuguese language.

In Table 1, we show the achieved average results. We show results for comparisons of 1-grams (referenced by ROUGE-1), 2-grams (ROUGE-2) and the longest n-grams (ROUGE-L).

Table 1. Results of ROUGE measure

Methods	ROUGE-1			ROUGE-2			ROUGE-L		
	P	R	F	P	R	F	P	R	F
Opizer-E	0.395	0.415	0.403	0.187	0.199	0.192	0.367	0.386	0.376
Our method	**0.536**	**0.483**	**0.508**	**0.342**	**0.305**	**0.322**	**0.506**	**0.456**	**0.479**

One may see that our approach outperforms the Opizer-E method in electronic products domain for all ROUGE values. This shows that the TOP(X) method helped in the selection of more representative sentences, improving the informativeness of the summaries, which are now closer to the summaries generated by humans.

6. Conclusions and future work

In this paper, we have shown that incorporating sentence importance assessment in a state of the art opinion summarization method may improve its results, producing more informative summaries. Nonetheless, it is important to notice that our test corpus was very small. Future work includes testing the enriched method on bigger corpora and also for different domains (books, for instance). Future work also includes exploring more semantically-driven approaches to opinion summarization, for producing both extractive and abstractive summaries.

Acknowledgements

The authors are grateful to FAPESP and IFPI for supporting this work.

References

Beineke, P., Hastie, T., Manning, C., and Vaithyanathan, S. (2003). An exploration of sentiment summarization. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 12–15.

Condori, R. E. L. and Pardo, T. A. S. (2017). Opinion summarization methods: Comparing and extending extractive and abstractive approaches. *Expert Systems with Applications*, 78:124 – 134.

Conrad, J. G., Leidner, J. L., Schilder, F., and Kondadadi, R. (2009). Query-based opinion summarization for legal blog entries. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 167–176.

de Sousa, R. F., Rabêlo, R. A., and Moura, R. S. (2015). A fuzzy system-based approach to estimate the importance of online customer reviews. In *IEEE International Conference on Fuzzy Systems*, pages 1–8.

Freitas, C., Motta, E., Milidiú, R., and Cesar, J. (2013). Sparkle vampire lol! annotating opinions in a book review corpus. In *11th Corpus Linguistics Conference*, pages 128–146.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the International Conference on Web Search and Data Mining*, pages 219–230.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL workshop on Text Summarization Branches Out*, pages 74–81.

Lopez, R., Pardo, T., Avanço, L., Balage Filho, P. P., Bokan, A., Cardoso, P., Dias, M., Nóbrega, F., Cabezudo, M., Souza, J., Zacarias, A., Seno, E., and Di Felippo, A. (2015). A qualitative analysis of a corpus of opinion summaries based on aspects. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 62–71.

Paltoglou, G. and Giachanou, A. (2014). Opinion retrieval: Searching for opinions in social media. In Paltoglou, G., Loizides, F., and Hansen, P., editors, *Professional Search in the Modern World: COST Action IC1002 on Multilingual and Multifaceted Interactive Information Access*, pages 193–214.

Tumitan, D. and Becker, K. (2013). Tracking sentiment evolution on user-generated content: A case study on the brazilian political scene. In *Brazilian Symposium on Databases*, pages 1–6.

Xu, C. (2013). Detecting collusive spammers in online review communities. In *Proceedings of the 6th workshop on Ph. D. students in information and knowledge management*, pages 33–40.