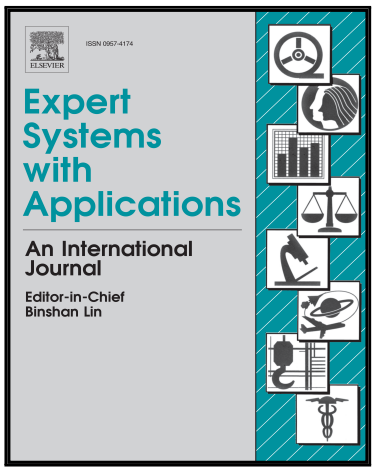

Opinion Summarization Methods: Comparing and Extending Extractive and Abstractive Approaches

Roque Enrique López Condori, Thiago Alexandre Salgueiro Pardo

PII: S0957-4174(17)30082-9
DOI: 10.1016/j.eswa.2017.02.006
Reference: ESWA 11113

To appear in: *Expert Systems With Applications*

Received date: 27 September 2016
Revised date: 20 December 2016
Accepted date: 2 February 2017

Please cite this article as: Roque Enrique López Condori, Thiago Alexandre Salgueiro Pardo, Opinion Summarization Methods: Comparing and Extending Extractive and Abstractive Approaches, *Expert Systems With Applications* (2017), doi: 10.1016/j.eswa.2017.02.006

Highlights

- Research about some aspect-based opinion summarization methods.
- A new content selection strategy to produce extractive summaries is proposed.
- A novel NLG template-based system to generate abstractive summaries is proposed.
- Extractive and abstractive opinion summarization methods are compared.

Opinion Summarization Methods: Comparing and Extending Extractive and Abstractive Approaches

Roque Enrique López Condori[a,*], Thiago Alexandre Salgueiro Pardo[a,*]

[a] *University of São Paulo, Institute of Mathematical and Computer Sciences, Interinstitutional Center for Computational Linguistics (NILC). Av. Trabalhador São-carlense, No. 400, São Carlos, São Paulo, Brazil*

Abstract

In the last years, the opinion summarization task has gained much importance because of the large amount of online information and the increasing interest in learning the user evaluation about products, services, companies, and people. Although there are many works in this area, there is room for improvement, as the results are far from ideal. In this paper, we present our investigations to generate extractive and abstractive summaries of opinions. We study some well-known methods in the area and compare them. Besides using these methods, we also develop new methods that consider the main advantages of the ones before. We evaluate them according to three traditional summarization evaluation measures: informativeness, linguistic quality, and utility of the summary. We show that we produce interesting results and that our methods outperform some methods from literature.

Keywords: Opinion Summarization, Aspect-Based Approach, Extractive and Abstractive Summarization

1. Introduction

"What other people think" has always been an important information for most of us during the decision-making process (Pang & Lee, 2008). Nowadays,

*Corresponding author

Email addresses: rlopez@icmc.usp.br (Roque Enrique López Condori), taspardo@icmc.usp.br (Thiago Alexandre Salgueiro Pardo)

before purchasing a product or service, it is very common to search, read and analyze online reviews that other people wrote about a product or service of interest.

The quantity of subjective texts, such as posts, tweets, and comments, among others, in which internet users express their opinions about products, people, events, businesses, etc., has been growing very quickly. For example, in a simple search in Google for "opinions about Samsung Galaxy S5", we found approximately 7,870,000 results. In the Buscap website (platform of product reviews), we found over 845 comments for the same product[1]. This amount is much smaller than the previous one, but is still too much to read. A person may hardly read one tenth of these comments.

In this scenario, to read and analyze that huge amount of reviews is a big problem, not only because of the quantity, but also because many of the reviews may be not relevant. The opinion summarization area aims to assist in this problem, extracting, synthesizing and showing the most relevant information. Although there are many works in this area, there is room for improvement, as the results are far from ideal.

More formally, opinion summarization is the task of automatically generating summaries for a set of opinions about a specific target (Conrad et al., 2009). One of the main approaches to generate opinion summaries is the aspect-based opinion summarization. This approach generates summaries of opinions for the main evaluated aspects of an entity. Entities may be any evaluated object, as organizations, services and products (e.g., a smartphone), and aspects are attributes or components of them (such as the battery or the screen of a smartphone).

In the last years, the interest in opinion summarization has been growing due to the importance of this task in the marketing community. With automatic opinion summaries, buyers have the best information to decide to carry out a particular purchase or not; it is possible to compare products and services

[1]Both searches performed on December 10, 2015

of interest; companies may improve their products and make better decisions; government agencies and entities may monitor and deal with crises and demands of the population; etc.

According to Mithun & Kosseim (2009), in comparison to the traditional summarization (summarization of objective documents, e.g., news), opinion summarization presents worse performance. This difference may be caused by many factors. Unlike traditional summarization, the sentiment (subjectivity) has a very important role in opinion summarization. Most of the reviews are written in informal language and sometimes have no relation with the main topics of the opinions. Besides informality, reviews usually do not follow standard language rules, introducing a lot of noisy in the process of summarization.

Most automatic methods in opinion summarization produce extractive summaries, which are created selecting the most representative text segments (usually sentences) from the original opinions. An opinion summary might also be abstractive, in which the content of the summary is rewritten using new text segments. There are few works that produce abstractive summaries, because they require some complex Natural Language Processing (NLP) tasks such as text generation or sentence fusion.

Extractive approaches are relatively easier to adapt to different domains, since they are limited to the extraction of sentences or expressions, however, these summaries may be significantly less coherent. On the other hand, abstractive approaches produce more sophisticated summaries, which usually contain material that improve the original content (Hahn & Mani, 2000). There are few studies comparing extractive and abstractive opinion summarization methods. According to Carenini et al. (2006), extractive methods generally perform better for traditional summarization, but these methods are not suitable for capturing subjective information because, in there pure form, they may not express the distribution of opinions. In Ganesan et al. (2010), it is also stated that, due to small variations of redundant opinions, extractive methods are often inadequate to summarize opinions. However, when the output format of extractive summaries is structured in aspects, summaries may provide the most important

information of every aspect in a specific and detailed way (Hu & Liu, 2004).

In this paper, we investigate some state of the art aspect-based opinion summarization methods, which produce both extractive and abstractive summaries, as well as structured and not structured summaries. Initially, four well-known methods from literature were analyzed and tested. During their analyses, we identified some gaps that affected their performance. Thus, in order to improve them, we proposed two more methods, which we refer by the names Opizer-E (an extractive method) and Opizer-A (an abstractive method).

In order to analyze the performance of the methods, quantitative and qualitative analysis were carried out according to three traditional evaluation measures in the area: informativeness, linguistic quality and utility of the summary. The results show that the methods produce interesting results and that our new methods outperform some other methods from literature.

In general, this paper makes the following contributions: i) we propose a new content selection strategy to produce extractive summaries of reviews, ii) we propose a novel Natural Language Generation (NLG) template-based system to generate abstractive summaries of opinions, and iii) we compare extractive and abstractive opinion summarization methods, as well as a traditional multi-document summarization method applied to opinions, using quantitative and qualitative measures, in order to assess the contribution of each one to the area and the best available approaches.

The remaining of this paper is organized as follows: in Section 2, we briefly introduce some basic concepts and some main related works; the proposed methods are described in Section 3; in Section 4, we describe the dataset and the evaluation measures used in this work; the results of the experiments are reported in Section 5; finally, in Section 6, we present our conclusions.

2. Basic Concepts and Related Work

According to Kim et al. (2011), aspect-based or feature-based opinion summarization is made up of three phases: aspect identification, sentiment predic-

tion, and summary generation. The aspect identification phase aims to find the important topics present in opinions. The sentiment prediction determines the sentimental orientation (positive or negative) of the aspects found in the first phase. Finally, the summary generation is the phase responsible for determining which information will be included in the summary. This work is mainly focused in the third phase, i.e., investigating methods for summary generation.

In traditional multi-document summarization, most works have primarily focused on generating textual summaries. However, in aspect-based opinion summarization, many researchers have proposed several output formats for the automatic summaries. The main output formats of opinion summaries based on aspects may be classified as: structured in aspects, textual, and visual.

In structured summaries, the content of the summary is grouped into aspects and polarities (positive or negative). Thus, for some aspects, the most relevant opinions or sentences are shown according to some heuristic for selecting them. Additionally, these summaries show the number of positive/negative reviews or the number of stars (as some e-commerce sites do). Figure 1 shows a summary that is structured in aspects about opinions of a camera. As we may see, the summary is organized according to the picture aspect. For this aspect, sentences extracted from reviews are grouped in positive and negative sections.

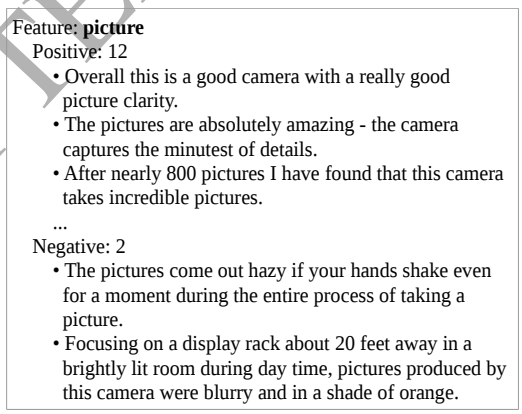

Figure 1: Opinion summary structured in aspects (reproduced from Hu & Liu (2004))

Visual summaries provide an overview of the users sentiment about a product or aspect according to their polarity (positive or negative). With these summaries, people may quickly understand what users like and dislike about a product, despite the large number of reviews. Figure 2 shows a visual summary of aspects of a digital camera using vertical bars. In the figure, each bar indicates the ratio of positive and negative reviews about the aspects (picture, battery, lens, etc.).

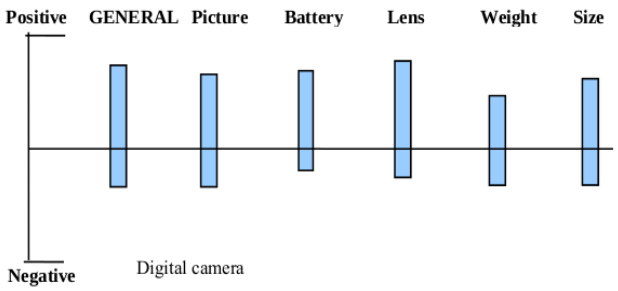

Figure 2: Visual opinion summary (reproduced from Liu (2012))

Textual summaries, in many cases, are generated listing some reviews or sentences. These reviews or sentences must be representative of the relevant aspects. In Figure 3, as examples, we show extractive and abstractive textual summaries, at the top and bottom of the image, respectively. Both summaries are about Canon G3 product. In the extractive summary, a simple list of sentences, separated by a period symbol, is shown. However, for the abstractive summary, the preferences of the users, expressed in the opinions, are analyzed and synthesized.

Most of the proposed works to summarize opinions follow extractive approaches. Hu & Liu (2004), considered the most classical method in the area, propose an architecture of summarization divided into three main phases: i) identification of aspects/features of products in opinions, ii) identification of positive and negative sentences for each aspect, and iii) summary generation using the previous information. For the last phase, the authors propose an extractive method that generates a structured summary (similar to the one

Extract. Bottom line, well made camera, easy to use, very flexible and powerful features to include the ability to use external flash and lens/filters choices. It has a beautiful design, lots of features, very easy to use, very configurable and customizable, and the battery duration is amazing! Great colors, pictures and white balance. The camera is a dream to operate in automode, but also gives tremendous flexibility in aperture priority, shutter priority, and manual modes. I'd highly recommend this camera for anyone who is looking for excellent quality pictures and a combination of ease of use and the flexibility to get advanced with many options to adjust if you like.

Abstract. Almost all users loved the Canon G3 possibly because some users thought the physical appearance was very good. Furthermore, several users found the manual features and the special features to be very good. Also, some users liked the convenience because some users thought the battery was excellent. Finally, some users found the editing/viewing interface to be good despite the fact that several customers really disliked the viewfinder. However, there were some negative evaluations. Some customers thought the lens was poor even though some customers found the optical zoom capability to be excellent. Most customers thought the quality of the images was very good.

Figure 3: Textual opinion summary (reproduced from Carenini et al. (2006))

in Figure 1), in which sentences are clustered according to their aspects and polarities. For each aspect (ranked according to its frequency), the quantity of positive and negative reviews with some randomly selected sentences are shown. Hu and Liu did not directly evaluate the quality of the summaries, but, instead of this, they evaluated the previous phases (aspect identification and sentiment prediction), stating that good results in these phases may improve the quality of summaries. The main contribution of this work is the proposed architecture for aspect-based opinion summarization. Additionally, it is the first study that proposes structured summaries of opinions.

Another extractive method is proposed by Tadano et al. (2010), which use the information of three features: rating of aspects (number of stars for the review), TF-IDF value, and the number of sentences with similar topics (through clustering of sentences using K-means algorithm (Steinhaus, 1956). Using these features, the authors proposed three methods of summarization: Meth1, which only considers the TF-IDF value of sentences; Meth2, which uses the TF-IDF value and the importance of clusters; and Meth3, which considers the same of Meth2 and the information of aspect rating. In all these methods, it is presented as a summary the sentence with the best score for each aspect. In the experiments, using ROUGE summary informativeness measure (Lin, 2004) (which is

introduced later in this paper), Meth3 achieved the best results, showing that the three features are important in the summarization of opinions. For ROUGE-1 and ROUGE-2, Meth3 achieved approximately 0.367 and 0.151, respectively.

There are few works that use abstractive approaches to generate summaries of opinions. Ganesan et al. (2010) propose a graph-based method that assumes no domain knowledge to produce abstractive summaries. The main idea of this method is to build a graph that represents the opinions to be summarized and, then, using some graph properties, to find appropriate paths that help in generating abstractive summaries. Basically, each (non-repeated) word in the opinions gives origin to a node in the graph, and weighted edges are established among the nodes of adjacent words. To compose the corresponding summary, the graph is traversed, looking for the most relevant paths that encode passages that, for instance, are more frequent and that present certain typical qualification word sequences (as the sequence "noun verb adjective"). For the experiments, the authors propose a measure of readability in which, given N sentences of an automatic summary and M sentences of a human summary (manually created), all these sentences are mixed and then other human judges are asked to choose N sentences from the summary that are less readable. If judges choose many sentences generated by the system, the readability is low. The readability tests showed that more than 60% of the sentences generated by the system are not different from sentences generated by humans. In relation to ROUGE-1 measure, the proposed method by Ganesan et al. achieved 0.327 in f-measure.

The method proposed by Gerani et al. (2014) is one of the first works that use discourse analysis to generate abstractive summaries. This method uses the Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) to structure each opinion. In this (discourse) structure, text segments are related by rhetorical/discourse relations in a tree format, and each branch of the tree is classified as more important (the nucleus of the relation) or less important (the satellite). Then, having the discourse structures, the method simplifies the structure of the opinions, keeping only the aspects of the segments and the corresponding

relations. The simplified structures of all the reviews are merged into a single graph, to which is applied a weighted version of the PageRank algorithm to select the most important subgraph, that is then mapped to natural language, which is done by filling some manually created templates with the main aspects and their polarities. In the experiments, Gerani et al. evaluated two extractive methods (MEAD-LexRank (Erkan & Radev, 2004) and MEAD* (Carenini et al., 2006)) and two abstractive methods (an adaptation of this method considering only the information of polarity, not considering RST information, and the complete method). Quantitative evaluation results, based on two pairwise preference user studies, showed substantial improvement of this method over the other ones. The biggest drawback of such method is the rigidity of the templates, because they do not permit to know exactly what is mentioned in the original opinions.

To address some of the limitations mentioned before, in this paper we propose novel extractive and abstractive methods that generate aspect-based summaries from multiple reviews. Our methods are introduced in what follows.

3. Opinion Summarization Methods

In this paper, we focus our research in the four methods from literature cited before (the ones of Hu and Liu, Tadano et al., Ganesan et al., and Gerani et al.) because of their diversity and the quality of their reported results. During their analyses, we identified some gaps that affected their performance. Among the main problems we found, we highlight the use of little information to select the representative sentences and the usage of pure static templates to produce the summaries. Specifically, many works that produced extractive summaries have not used different sources of information to select representative sentences, which may harm the quality and diversity of the summary. In the case of abstractive methods, many of them have generated summaries by selecting static templates, without dynamicity and without using explicit information provided by the reviews. These are important disadvantages because the summaries get a bit

artificial and users do not have access to what is explicitly said in the opinions. Thus, in order to improve them, we propose two more methods, which we refer by Opizer-E (an extractive method) and Opizer-A (an abstractive method).

In this work, we used a corpus that has aspects and their polarities manually annotated. Thus, we were exclusively focused on the summary generation phase, assuming as ready the phases of aspect identification and sentiment prediction, which are already extensively studied in the area.

3.1. The Extractive Method: Opizer-E

This method aims to generate summaries by extracting a small number of sentences about the main aspects, in such a way that aspect coverage and the distribution of polarity become preserved as much as possible. We considered these two criteria for the following reasons. Firstly, aspect coverage is an indicator of how many aspects of the source opinions are preserved in the generated summary. Thus, with this criterion, we intend to have diversity in the summary, focusing on the most important aspects and not only in one of them as other works sometimes do (Blair-Goldensohn et al. (2008), Xu et al. (2011)). Secondly, to communicate to the readers of the summary what is the sentiment in the opinions about the aspects is not simply a matter of classifying the summary as positive or negative. The readers want to know if all opinions that evaluate the aspects made it in a similar way or if there was variance. Thus, opinion summaries must preserve the polarity distribution as much as possible to reflect the overall sentiment about the aspects. For such purpose, we propose two phases: sentence clustering and sentence ranking. In general, differently of the extractive methods explained above and other approaches in the literature, Opizer-E combines two types of information (sentence position and qualifier proximity) that have not been used together in other works.

The output format of the summary used in this proposal is the one structured in aspects. The selection of this type of format was mainly due to two advantages that it offers. Firstly, a structured summary explicitly shows what is the general sentiment in the opinions, indicating the number of positive and

negative reviews or the amount of sentences with these sentiments. This is a good advantage for the summary readers, because they will know exactly how many reviews/sentences are positive or negative. The second benefit of this format is that it may facilitate the reading of the summary. As it is reported in the literature, extractive summaries often have the problem of lack of cohesion among sentences. Structured summaries reduce this problem by clustering the sentences by aspects and showing them as a list of items. The phases of Opizer-E are explained with more details in what follows.

3.1.1. Sentence Clustering

In this research, we worked with a collection of reviews that contain multiple aspects. In this context, some sentences might be similar to other ones. This phase aims to cluster review sentences that contain the same aspect and have the same polarity. More specifically, the idea is to cluster sentences at the level of aspect and polarity. To perform it, we used the information of the aspects and polarities manually annotated in the corpus, as we commented before.

For example, in the sentences (1) "the battery of my Iphone 5 lasts a short time", (2) "The iPhone 5 battery is light, I like this" and (3) "the duration of the iPhone 5 battery is short", the first and third sentences should be clustered in the same group because they mention the same aspect (battery) and have the same polarity (negative). On the other hand, the second sentence would be in another group, because, although it also cites the battery aspect[2], its polarity is positive.

After aspect-based sentence clustering, the occurrence frequency of the aspects is used to calculate their importance. In the resulting ranking, also used

[2]To the attentive reader, it is interesting to notice that, in reality, battery is not the true aspect. In fact, the aspects that the sentences cite are the "duration of the battery" and the "weight of the battery". In both cases, these true aspects are implicit and "battery" may be simply assumed as the main aspect. Some authors would go on and classify "duration" and "weight" as attributes (not aspects themselves) of the "battery" aspect. This is still an open question that is not fully solved in the area.

by Hu & Liu (2004) and Wang et al. (2013), the more often one aspect is, the more important it is.

3.1.2. Sentence Ranking

Extractive summaries based on aspects aim to list the most relevant sentences for each aspect. However, to establish a relevance criterion is not an easy task. Many studies suggest different criteria. In Beineke et al. (2003), the authors performed sentence ranking using their positions in the reviews. In Blair-Goldensohn et al. (2008), the sentences are ordered according to the polarity of sentences, that is, the authors suggest selecting the "most" positive and most negative sentences and using them in the final summary. Sharifi et al. (2010) uses the TF-IDF value for sentence ranking.

The purpose of this sentence ranking phase is to produce a rank of sentences for an aspect and show the best ones in the final summary according to the default size of the summary. For this, we conducted experiments with some of the previously mentioned criteria and with other two types of rankings proposed in this work. The first ranking is based on the sentence position in the review and the second one uses the proximity (distance) that exists between aspects and their qualifiers (textual segments with polarity) that evaluate them. The sentence position ranking is based on the idea that the first sentence is more important than the others in a review. This criterion has been used in several studies of news summarization (Seki (2002), Ouyang et al. (2010), Nóbrega et al. (2014)) with satisfactory results. Beineke et al. (2003) also used this criterion in opinion summarization, giving a big weight to the first sentence and other unique weight to the other sentences. Unlike Beineke et al. (2003), in this proposal, each sentence has a different weight depending on its position and the number of sentences in the review.

Therefore, following the above strategy, in Opizer-E, sentence position criterion is based on the idea that, when users give an opinion, they prefer to express their most important issue in the initial part of the text, which was empirically observed in our experiments with the corpus we work on. Experiments were

also carried out giving more importance to intermediate and final sentences, but the best results were obtained when the initial sentences had more weight. To calculate the weight of each sentence, we used Equation 1.

$$WPos(s_i) = \frac{len(r) - pos(s_i)}{len(r)} \quad (1)$$

where $WPos(s_i)$ is the weight of a sentence s_i and its value ranges between 0 and 1, with 1 being the biggest weight. $len(r)$ is the number of sentences in review r and $pos(s_i)$ indicates the position of sentence s_i in this review, with the first sentence having the value 0.

In the qualifier proximity ranking, the main idea is to give more importance to the sentences that have aspects and the corresponding qualifiers (textual segments with polarity) closer. Thus, we intend to select sentences that express more specific and direct opinions for an aspect, which could allow to the users to quickly understand the main issues of an aspect in a concise summary, without many details. For this purpose, the distance between the aspects and their qualifiers was calculated.

In this work, qualifiers are textual segments or n-grams that express a sentiment about an aspect. The distance between an aspect and its qualifiers is calculated according to the positions that they have within the sentence. In this ranking, the smaller the distance is, the more important the sentence is. In the case where qualifiers are in another sentence, we consider the distance between the aspect and the beginning of the sentence where the qualifier is (in general, there were few cases as this in our experiments). In this ranking, we used Equation 2 to calculate the weight of a sentence for an aspect.

$$WPro(s_i) = 1 - \min_{\forall j \in a} \left(\frac{|pos(a) - pos(q_j)|}{len(s_i)} \right) \quad (2)$$

where $WPro(s_i)$ is the weight of a sentence s_i and its value ranges between 0 and 1, with 1 as the biggest weight. $len(s_i)$ is the size of the sentence s_i, $pos(a)$ is the position of the aspect a (that is being analyzed) and $pos(q_j)$ is the position of the qualifier q_j of the aspect a within the sentence.

As an example, consider the sentences (1) "the **camera** has a *great resolution* to capture images in movements" and (2) "the **camera** built into the top of the device with digital zoom *works well*". In the first sentence, the *great resolution* qualifier appears very close to the aspect/entity (the camera), expressing a specific and direct opinion about it. However, in the second sentence, the *works well* qualifier is far from the aspect. This sentence presents many unnecessary details for the aspect thay may be not essential to be included in the final summary. Thus, for the qualifier proximity ranking, the first sentence is more important than the second one.

Finally, to calculate the importance of the sentences for each aspect, we used Equation 3, which considers the rankings explained above. Experiments were also performed using these rankings separately, but the best results were achieved using both criteria in a weighted combination.

$$Imp(s_i) = \alpha \times WPos(s_i) + (1 - \alpha) \times WPro(s_i) \quad (3)$$

In Equation 3, $Imp(s_i)$ indicates the importance of a sentence s_i for a particular aspect. The α coefficient keeps the trade-off between the two rankings. The best value for α in our experiments was 0.65.

In Figure 4, we present a summary produced by Opizer-E. As it may be observed, it shows a positive and a negative sentence for each aspect. In the default configuration of Opizer-E, every aspect should have a positive and a negative sentence in the summary, in order to increase the coverage of aspects in the text. However, the number of sentences for each aspect may increase up to cover the desired number of words of the summary.

3.2. The Abstractive Method: Opizer-A

Due to the fact that abstractive summarization has been little explored because of its complexity, an extra motivation of this work was to go one step further on the state of the art for this type of summarization and to generate more natural summaries. In this paper, we studied some approaches to

```
Aspect: Samsung Smart TV
   Positive Sentences: 16
   - The best tv nowadays!
   Negative Sentences: 11
   - Its screen burned 1 month after the warranty.
Aspect: Price
   Positive Sentences: 0
   Negative Sentences: 2
   - What I did not like: Price too high
Aspect: Durability
   Positive Sentences: 0
   Negative Sentences: 2
   - Durability 0
Aspect: Camera
   Positive Sentences: 1
   - The camera with motion sensor works well and impresses
     who do not know it.
   Negative Sentences: 1
   - What I did not like: The camera  could have horizontal
     movement of vision and not only just vertical.
Aspect: Image Quality
   Positive Sentences: 1
   - Excellent TV with great image quality and features.
   Negative Sentences: 0
```

Figure 4: Extractive summary produced by Opizer-E

generate abstractive summaries, e.g., methods based on natural language generation (Radev & McKeown, 1998), sentence compression (Zajic et al., 2007), sentences fusion (Barzilay & McKeown, 2005) and methods based on templates (Jung & Jo, 2003). After this study, the latter approach was selected for the implementation of Opizer-A.

The choice of an abstractive system based on templates was motivated by the discussions in the work of Van Deemter et al. (2005) about the practicality and effectiveness of this approach. Moreover, according to Embar et al. (2013), the fact that templates are manually specified provides an advantage in scenarios that need specific information. Another important reason is that, recently, other works of opinion summarization, such as Gerani et al. (2014) and Fabbrizio et al. (2014), obtained satisfactory results using an approach based on templates.

In general, the main difference between Opizer-A and other abstractive methods is that Opizer-A, beyond using the polarity information of the most important aspects, also uses the information of the most representative qualifiers of

aspects. Opizer-A has two phases: clustering of textual segments and text generation based on templates.

3.2.1. Clustering of Textual Segments

In automatic text summarization, one of the initial tasks is to group the most similar information presented in the source texts, because it allows identifying the most mentioned topics in the texts and, consequently, the most representative information. This information usually are sentences or n-grams of the texts. Given that this approach should generate abstractive summaries, we opted to use n-grams or textual segments, instead of full sentences. Specifically, we used textual segments with explicit polarity about some aspects, because in the reviews they allow to know, more accurately, what users talk about. Therefore, this phase aims to cluster textual segments or qualifiers that are more similar and representative for each aspect.

As a first step in this phase, all aspects were scored to determine their order in the summary. For this purpose, we used the strategy proposed by Gerani et al. (2014), which uses RST (Rhetorical Structure Theory) (Mann & Thompson, 1988) relations[3] among aspects in the ranking. The main reason to use this strategy is that it allows determining the most relevant aspects in a group of opinions, following the users rhetoric in their reviews.

After that, textual segments were clustered according to the aspects that they contain and their respective polarity. In other words, to each aspect was associated the textual segments that mentioned it in two groups of polarity: positive and negative. Then, textual segments of each aspect and polarity were clustered according to their content, using K-means algorithm (Steinhaus, 1956), with a K value equal to 3 or 4. A small value was chosen because the idea is to guarantee the selection of just the main topics mentioned in the textual

[3]To identify the RST relations, we used the discourse parser proposed by Maziero et al. (2011), which has 62.5% of accuracy in the detection of RST relations and 81% in the identification of nuclearity (i.e., which segments are nuclei and satellites of the relations) in scientific texts.

segments. Moreover, we believe that there are not many topics for a particular aspect in the reviews. Finally, with these values for K, we achieved the best results in the experiments. In the clustering with K-means algorithm, stopwords were not considered.

Finally, for each aspect and polarity, the cluster that contained the highest number of elements (textual segments) was selected. With that, we intended to find the most repeated/redundant information and, therefore, the most important one in the group of texts. Given the chosen cluster, the textual segment with the highest normalized TF-IDF value (see Equation 4) was selected. This value was calculated using the partial TF-IDF values of the words that formed the textual segment, normalized by its size (number of words). The idea was to penalize long textual segments, because the summary may quickly reach the number of allowed words, reducing the possibility of putting other information in the text.

$$TSeg = \max_{\forall s \in a} \left(\frac{\sum_{i=1}^{n_s} TF\text{-}IDF_i}{n_s} \right) \quad (4)$$

In Equation 4, $TSeg$ is the selected textual segment to form the summary, n_s is the size of the textual segment s (in number of words) and $TF - IDF_i$ is the TF-IDF value of the word i of textual segment s in relation to the aspect a.

3.2.2. Text Generation Based on Templates

The general idea of the second phase was to fill some templates with aspects and textual segments obtained in the previous phase according to their polarity. Templates are linguistic structures composed of slots that permit to collect information according to a specific target (Van Deemter et al., 2005).

We created templates for the main types of evaluations, such as when the majority of the users evaluate an aspect in the same way (positively or negatively) or when there are controversial opinions. To determine the structure and content of these templates, we made a study of some works that used a similar approach, e.g. Gerani et al. (2014) and Fabbrizio et al. (2014). In addition, we

analyzed the human summaries of our corpus in order to extract some common forms of writing.

In Table 1, the templates manually created that are used by Opizer-A are shown. As we may see, Table 1 is divided into five sections: first sentence, adjectives with polarity, expressions with verbs, connectives, and explainers. Templates of "first sentence" define the structure for the summary at the beginning, where symbols <> are the slots to be filled. Templates of the section "adjectives" and "expressions with verbs" allow expressing a sentiment according to the polarity of the aspects. "Connectives" are used to connect two sentences, taking into account their sentiments. Finally, "explainers" provide support to sentences through the qualifiers.

For instance, when many users evaluate an aspect in the same way, the template "Most of the opinions about <ASPECT> are" might be used. These templates should be filled with the relevant information obtained in the previous phase (clustering of textual segments) in order to create the sentences that compose the final summary.

The selection of templates of Table 1 is mainly based on the polarity of the aspects. For templates of the section "first sentence", "adjectives with polarity" and "expressions with verbs", we used a scale from -2 to 2 to determine the sentiment of aspects, similar to the work of Guzman & Maalej (2014) and Semantria tool[4], with the values -2, -1, 1, and 2 indicating very negative, negative, very positive, and positive sentiments, respectively.

The score for the sentiment of an aspect is determined by the frequency of their positive or negative qualifiers normalized by the total number of qualifiers. Thus, the scores in the interval [0, 0.25⟩ are interpreted as very negative, scores in the range [0.25, 0.40] are interpreted as negative, scores in the interval [0.60, 0.75] are interpreted as positive, and the scores in the range ⟨0.75, 1] are interpreted as very positive. When the score of polarity is in the range ⟨0.40, 0.60⟩, it is considered as a controversial, which indicates that there is a similar amount

[4]Available at `https://semantria.com`

First Sentence
Controversial: [‘The opinions of <ASPECT> shows controversial sentiment’, ‘The <ASPECT> has positive and negative opinions’...]
Polarity: [‘In general, the opinions about <ASPECT> are’, ‘Most of the opinions about <ASPECT> are’...]
Adjectives with Polarity
−2: [‘very negative’, ‘very unfavorable’...]
−1: [‘negative’, ‘unfavorable’...]
+1: [‘positive’, ‘favorable’...]
+2: [‘very positive’, ‘very favorable’...]
Expressions with Verbs
Controversial: [‘there are controversial opinions about it’, ‘are expressed controversial opinions about this feature’...]
−2: [‘users hate it’, ‘was rated as very bad’, ‘the opinions were very negative’...]
−1: [‘people dislike it’, ‘was rated as bad’, ‘the opinions were negative’...]
+1: [‘people like it’, ‘users consider it as satisfactory’...]
+2: [‘users loved it’, ‘was rated as excellent’, ‘the opinions were excellent’...]
Connectives
Agreement: [‘In the same way’, ‘Likewise’, ‘Similarly’...]
Non Agreement: [‘However’, ‘In contrast’, ‘Contrary’...]
Normal: [‘Furthermore’, ‘Moreover’, ‘Besides that’...]
General: [‘In relation to <ASPECT>’, ‘With respect to <ASPECT>’, ‘About <ASPECT>’...]
Explainers
[’because <QUALIFIER>’, ’forasmuch <QUALIFIER>’...]

Table 1: Templates used by Opizer-A

of positive and negative qualifiers about the same aspect (see Equation 5).

$$Sent_i = \begin{cases} -2 & score_i = [0, 0.25\rangle \\ -1 & score_i = [0.25, 0.40] \\ controversial & score_i = \langle 0.40, 0.60\rangle \\ +1 & score_i = [0.60, 0.75] \\ +2 & score_i = \langle 0.75, 1] \end{cases} \quad (5)$$

In templates of the "connectives" section, the selection is based on the agreement between the polarities of aspects present in continuous sentences. The selection of templates of "explainers" section is random. With the templates already defined, the next step consists in organizing the text in the summary. This is known as microplanning in Natural Language Generation area (Reiter, 1994).

In Opizer-A, the initial task in microplanning step structures the first sentence of the summary. This sentence is very important because it indicates which is the dominant sentiment for the main aspect. For this, the first template to be filled corresponds to the one in the "first sentence" section. Then, according to the polarity of the aspect, a template from "adjectives with polarity" section is chosen. With these two templates, the first sentence of the summary is ready.

From the second sentence on, the process is repeated for the other sentences. Thus, to create a new sentence in the summary, it is necessary to verify the polarity of the aspect in the previous sentence. Based on this, it is selected a template of the "connectives" section, and then a template of the "expressions with verbs" section. Finally, it is selected a template of the "explainers" section. This template is filled with the most representative textual segment (qualifier) identified in the previous phase.

In this last step, we identified a small set of rules to add some verbs in the initial part of the qualifiers, with the intention of improving the structure of the sentence. For example, given the template of Table 1 ("explainers" section)

"because <QUALIFIER>" and the qualifier "the best of today", it would be appropriate to add the verb "is" before the qualifier, so we would have "because [it] is the best of today". The defined rules basically analyze the morphosyntactic tag of the first word of the qualifier. For this, we used the tagger proposed by Fonseca & Rosa (2013)[5].

Figure 5 shows an abstractive summary generated by Opizer-A. In the first part of the text, the general sentiment for the main aspect (Samsung Smart TV) is mentioned, and then the users sentiment about other aspects of this product is described. In total, five aspects are described in the summary (Samsung Smart TV, price, design, camera, and image quality). However, this amount is configurable according to the desired number of words for the summary.

```
Most reviews of the Samsung Smart TV are favorable. On the
other hand, in relation to the price, users did not like it
because it is high. In contrast, regarding the design, it
was rated as excellent as it is sophistication and
modernity. Furthermore, with respect to the camera, they
expressed controversial opinions on this feature because it
works well and impresses who do not know it, but it could
have horizontal movement of vision and not only just
vertical. In addition, with respect to image quality, the
reviews were excellent because it is great.
```

Figure 5: Abstractive summary produced by Opizer-A

In general, to create this summary, Opizer-A ranked the aspects to determine their order in the text. In this case, the computed order was: Samsung Smart TV, price, design, camera, and image quality. For each aspect, its more representative textual segments were determined. Then, with this information, the templates were filled to create the summary. To create the first sentence, the template "Most reviews of the <ASPECT> are" from "first sentence" section (Polarity option) of Table 1 was selected, since the vast majority of the reviews were positives. This template was filled with the aspect "Samsung Smart TV",

[5]Tagger with 96.48% of accuracy (approximately) in news texts.

since it is the first in the ranking. To conclude the first sentence of the summary, a template from "adjectives with polarity" section was selected. As the polarity was positive, "favorable" was selected. For the second sentence, the template "On the other hand" from "connectives" section was selected because the polarity of the second aspect (price) is contrary to the first one (Samsung Smart TV). Then, templates of "connectives" and "expressions with verbs" sections were selected. Finally, the template "because <QUALIFIER>" of the "explainers" section was selected. This template was filled with the most representative textual segment (qualifier) identified for the aspect (price). In this case, the qualifier was "high". From the second sentence on, the process is repeated for the other sentences.

4. Experiments

4.1. Dataset

For evaluating the methods, we used the OpiSums-PT corpus (López et al., 2015), composed of group of opinions and their extractive and abstractive summaries. The corpus is fully written in Brazilian Portuguese.

The corpus contains reviews of two domains: books and electronic products. For the first one, OpiSums-PT uses the reviews of ReLi corpus (Freitas et al., 2013), consisting in a collection of opinions about 13 books. For the second domain, OpiSums-PT uses reviews of 4 electronic products collected from a specialized website.

For each book and electronic product, OpiSums-PT has 5 extractive and 5 abstractive summaries, with each set of them being manually created by different human annotators. Each summary comes from the analysis of 10 reviews and is composed by 100 words, approximately. In total, OpiSums-PT has 170 human summaries (85 extractive and 85 abstractive summaries).

4.2. Evaluation Measures

Since the source opinions of OpiSums-PT already have manually identified aspects and polarities, the evaluation of the methods were focused exclusively

on the summary generation step, assuming as ready the steps of aspect identification and sentiment prediction. The methods were evaluated according to three traditional measures in the summarization area: informativeness, linguistic quality and utility.

4.2.1. Informativeness

Informativeness aims at assessing how much information from the source is preserved in the summary (Mani, 2001). To measure the informativeness of automatic summaries, one of the most popular measures in the summarization task is ROUGE measure (Lin, 2004), which was adopted in this work. ROUGE measure compares the amount of n-grams that an automatic summary and one or more human summaries, called reference summaries, have in common, producing precision, recall, and f-measure results.

ROUGE measure is widely used in the evaluation of automatic summaries, because it is quickly and easily applicable. However, the main disadvantage of this measure is that it only evaluates the n-gram matching, ignoring all aspects of the linguistic quality of summaries. To address these issues, there are other methods that may be used. We introduce the linguistic quality issues in what follows.

4.2.2. Linguistic Quality

The linguistic quality evaluation requires the participation of human assessors to judge the quality of a summary, considering some linguistic criteria. In DUC[6] and TAC[7] summarization conferences, summaries are evaluated considering five criteria: grammaticality, non-redundancy, referential clarity, focus, and structure and coherence (Dang, 2005).

This type of evaluation is manually performed by humans, who score the quality of a summary considering these criteria in a five-point scale: 1. Very Poor, 2. Poor, 3. Barely Acceptable, 4. Good, and 5. Very Good. In this work,

[6]Document Understanding Conference, at `http://duc.nist.gov/pubs.html`
[7]Text Analysis Conference, at `http://www.nist.gov/tac/`

26 people with strong knowledge in NLP participated in the evaluation of the linguistic quality of summaries. We calculated inter-annotator agreement using Kappa coefficient (Carletta, 1996). The Kappa values for grammaticality, non-redundancy, referential clarity, focus, and structure and coherence were 0.149, 0.088, 0.076, 0.100 and 0.127, respectively. It is possible to see that the values reflect low agreement among annotators. We believe it is likely due to the five-point scale, since annotators had with this scale some options to choose that are very similar (e.g. Good and Very Good). These values also indicate that manual evaluation of opinion summaries is a difficult task, even for annotators with strong knowledge in NLP.

Although this measure does not use reference summaries, it may benefit good automatic summaries that are quite different from the reference summaries. However, this measure does not assess how useful a summary may be. The following section explains the summary utility measure used in this work.

4.2.3. Utility

This criterion evaluates the usefulness of the opinion summaries for the readers. In particular, readers should evaluate how useful the summary is in a buying decision process. Since we used OpiSums-PT corpus, an item might be a book or an electronic product. For instance, to evaluate the usefulness of a summary about "Samsung Smart TV" product, the question asked was: "If you want to buy a Samsung Smart TV product, how do you rate the usefulness of the summary?".

To answer this question, we also used the scale described above, in which a score of "5" indicates that the summary is very useful (Very Good), "1" indicates that the summary is not useful (Very Poor), and scores "2" to "4" represent the gradation between the two extremes. The evaluation of the utility measure occurred together with the linguistic quality evaluation, also counting with the 26 evaluators, therefore. Similarly, for this measure the Kappa value was 0.133 which also indicate low agreement among annotators.

5. Results

For the experiments, in addition to Opizer-E and Opizer-A, four methods from literature (Hu and Liu, Tadano et al., Ganesan et al., and Gerani et al.) were adapted and evaluated with the same corpus in Portuguese. The summarizers proposed by Hu and Liu and Tadano et al. are completely independent of language, since they use the frequency and rating of aspects to generate extractive opinion summaries. Ganesan et al. and Gerani et al. are slightly language dependent, using the structure of graphs to rank aspects and to find relevant paths (word sequences). To adapt to Portuguese the method proposed by Ganesan et al., we modified some regular expressions (e.g., the sequence "adjective verb noun" for Portuguese) to find the most relevant word sequences in the graph. For Gerani et al., in order to create the graph, we used a RST parser for Portuguese (Maziero et al., 2011), and we translated the original templates used in that work.

Besides the opinion summarization methods, we also evaluated the results of a multi-document summarization, in order to check for the necessity of specialized methods for opinions. We tried the RSumm system (Ribaldo et al., 2012), a multi-document summarizer that is considered one of the best systems for the language of the corpus.

In what follows, the results are presented for each one of the metrics that we cited before.

5.1. Informativeness

Table 2 presents the average results for informativeness, with the best results in bold. To evaluate that, ROUGE measure was used, considering the human summaries of OpiSums-PT as reference summaries, producing recall (R), precision (P), and f-measure (F) measures. We show results for ROUGE-1, ROUGE-2 and ROUGE-L, as it is usually done in the area. ROUGE-1 stands for the unigram matching; ROUGE-2 for bigram matching; ROUGE-L for the longest common subsequence matching.

Automatic summaries produced by extractive methods were compared with human extractive summaries of OpiSums-PT, and likewise, automatic abstractive summaries were compared with human abstractive summaries of OpiSums-PT. We chose this strategy because the generation of these two types of summaries involves different processes.

Table 2: Results for Informativeness measure

	Methods	ROUGE-1			ROUGE-2			ROUGE-L		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Extr.	RSumm	0.306	0.291	0.296	0.165	0.152	0.157	0.282	0.267	0.273
	Hu and Liu	0.380	0.372	0.373	0.214	0.211	0.212	0.342	0.337	0.338
	Tadano et al.	0.386	0.382	0.382	0.251	0.249	0.249	0.372	0.368	0.368
	Opizer-E	**0.393**	**0.392**	**0.390**	**0.267**	**0.266**	**0.265**	**0.383**	**0.381**	**0.380**
Abstr.	Ganesan et al.	0.125	**0.291**	0.170	0.030	**0.070**	0.041	0.110	**0.257**	**0.149**
	Gerani et al.	0.138	0.199	0.158	0.029	0.040	0.032	0.103	0.153	0.119
	Opizer-A	**0.155**	0.280	**0.189**	**0.034**	0.062	**0.042**	**0.119**	0.225	0.148

For the extractive methods, the highest f-measure values were obtained by Opizer-E. We believe that both information (sentence position and qualifier proximity) used by Opizer-E helps to select the most representative sentences of the aspects considering their polarities. On the other hand, RSumm got the lowest performance, because it does not take into account the information about aspects or polarity. Using the Student's t-test with 95% of confidence (p-value ~ 0.00922), we found statistical difference in f-measure values between Opizer-E and RSumm. Among Opizer-E, Hu & Liu (2004) and Tadano et al. (2010), we did not find statistical difference. In the case of Tadano et al., we believe that using the number of stars assigned to the review to determine summary polarity may have slightly affected the system performance. Although the number of stars not always reflects what is written in the review, one may trust it, in general. Therefore, we believe that, if such evaluation discrepancy happened, it is probably insignificant.

In Table 2, we may see that values of ROUGE-1, ROUGE-2 and ROUGE-L of abstractive methods were much lower than the results obtained by extractive

methods, because these abstractive approaches do not select complete sentences, but create a new text from the information of the original sentences.

For the abstractive methods, Gerani et al. (2014) had the lowest performance, possibly due to the rigidity of the templates used in this method. These templates have been manually defined and do not share many common words with source reviews, which is heavily penalized by ROUGE measure. Opizer-A got slightly better results than the other abstractive methods. We believe that it is because Opizer-A fills the templates with some aspect qualifiers present in the original sentences. Nonetheless, Ganesan et al. (2010) had the best performance for ROUGE-L. To generate good abstractive summaries, this method needs highly redundant source reviews, otherwise the generated summaries will be more extractive than abstractive. In OpiSums-PT, opinions source are not too redundant. Because of this, in most cases the summaries generated by Ganesan et al. were formed by complete sentences, increasing the probability of finding more common large n-grams between automatic and reference summaries, thereby allowing good ROUGE-L measure values. The t-test was also applied to the abstractive methods. On one hand, differences among these methods for ROUGE-1 and ROUGE-2 measures are statistically irrelevant. On the other hand, for ROUGE-L, there is a significant difference between Opizer-A and Gerani et al method.

Considering both approaches (extractive and abstractive), Opizer-E achieved the best results. In the case of Opizer-A, although it did not overcome in all measures the other abstractive methods, it achieved results very close to the best performance of the method proposed by Ganesan et al.

5.2. Linguistic Quality

For linguistic quality and utility measures, we only evaluated the best opinion summarization methods (according to the informativeness results) and RSumm summarizer, because these measures require human participation, being more expensive and time-consuming, therefore. Likewise, we did not evaluate the complete collection of books and products, but a sample of three items of

each kind. Additionally, we evaluated some abstractive human summaries of OpiSums-PT in order to know how far automatic summaries are in relation to the human summaries.

As we may see in Table 3, human summaries achieved the best average results, as expected.

Table 3: Results for Linguistic Quality measure

Methods	**Grammaticality**	**Non Redundancy**	**Referential Clarity**	**Focus**	**Structure and Coherence**
RSumm	3.500	**4.269**	3.917	3.212	3.237
Opizer-E	**3.763**	3.731	**4.045**	3.750	**3.423**
Opizer-A	3.141	3.936	3.981	**3.910**	3.115
Human	4.647	4.692	4.763	4.705	4.660

In relation to grammaticality, the method with the best results was Opizer-E and, on the other hand, Opizer-A had the worst results. One of the major errors reported in the evaluation of this method was the selection of verbs that connect the fixed parts with the filled parts of the templates (e.g., the identification of the appropriate verbs number to create a sentence).

In the non-redundancy criterion, RSumm was the best method. Regarding to Opizer-E and Opizer-A, the abstractive method generated less redundant summaries. In some cases, Opizer-A was penalized by using the same discourse markers (e.g., "however", "moreover", and "furthermore", among others) in the summary. Opizer-E produced some summaries with repeated sentences. These cases occurred when two aspects were present in the same sentence and that sentence was the unique sentence for these two aspects. Thus, that sentence was repeated in both aspects, affecting the performance of Opizer-E.

Regarding the referential clarity criterion, the methods performed more similarly (in relation to the other criteria). RSumm, Opizer-A and Opizer-E achieved 3.917, 3.981, and 4.045, respectively.

Opizer-A achieved the best results for the focus criterion. The main disadvantage of extractive methods (Opizer-E and RSumm) is that they select complete sentences and put them in the summary. These sentences may con-

tain several different topics, which may deviate the text from the main focus of the summary. On the other hand, in Opizer-A, only parts of sentences (n-grams) are selected, which are oriented to a main focus.

In structure and coherence criterion, Opizer-E achieved the best results. The fact of organizing the summary by aspects and by positive and negative sentences helped Opizer-E to generate better structured and coherent summaries.

5.3. Utility

Similar to linguistic quality measure, in utility measure we only evaluated some methods using the same sample. Table 4 shows the average results obtained in the experiments for summaries of electronic products and books. It is also shown in this table the general average of these two kinds of items.

As we may see in Table 4, the performance of RSumm was lower than Opizer-A and Opizer-E. This is possibly due to the fact that RSumm is not a method oriented to summarize opinions. Thus, RSumm might have selected sentences that had no relation with aspects or the sentiment of them. In addition, RSumm does not consider the "quantitative side" of a summary, which indicates the percentage or amount of positive or negative reviews about aspects or entities, which is very useful in the purchasing process.

Opizer-A had the best result (3.474) in summaries of products, but, in summaries of books, the performance was lower (2.667). This was mainly because reviews of books contain fewer aspects in relation to reviews of electronic products. With few aspects, the size of the summary generated by Opizer-A was small, since there is little information (aspects) to produce the text.

In general, Opizer-E had the best results. We believe that this happened because this method clustered the sentences by aspects and polarity and it also indicates the amount of positive and negative sentences, which is a very useful information in a decision-making process, because it allows knowing what is the general sentiment for aspects and entities.

For this measure, we got better results for electronic product summaries than for book summaries. As book opinions contain fewer aspects, there are fewer

available options to select the representative information. In contrast, electronic products have more technical opinions that include many aspects, which allow making better use of the provided information.

Table 4: Results for Utility measure

Methods	Products	Books	Average
RSumm	2.974	2.705	2.840
Opizer-E	3.244	**3.103**	**3.173**
Opizer-A	**3.474**	2.667	3.070
Human	4.615	4.462	4.538

6. Conclusions

In this paper, we presented our research about some aspect-based opinion summarization methods (extractive and abstractive approaches). We tested some well-known methods in the area and proposed two new methods: Opizer-E and Opizer-A, which use the main advantages of the previous methods in order to generate better summaries. The demonstration versions of all methods investigated in this paper are freely available for use[8].

In general, Opizer-E obtained the best results in the experiments. Although Opizer-A did not achieve the best performance in all the evaluations, the obtained results for this method were very close to the best performance. The results of these evaluations also show that, based on the three evaluation measures used in this work, traditional summarization approaches are not suitable for generating opinion summaries. Other results show that summaries produced by extractive opinion summarization methods have better informativeness than abstractive methods (see Table 2). However, in relation to the non-redundancy criterion (Table 3), the abstractive approaches obtained better performances than the extractive ones. In relation to the utility of the opinion summaries, extractive methods were better than the abstractive ones (Table 4).

[8] http://nilc.icmc.usp.br/semanticnlp/opizer/sumarizadores/

Future work includes investigating more advanced abstracting operations, as sentence fusion, as well as preprocessing techniques, as sentence normalization and simplification. We believe that such technologies may significantly improve the quality of abstractive summaries.

Acknowledgments

To Samsung Eletrônica da Amazônia Ltda and FAPESP for supporting this work.

References

Barzilay, R., & McKeown, K. R. (2005). Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, *31*, pp. 297–328.

Beineke, P., Hastie, T., Manning, C., & Vaithyanathan, S. (2003). An Exploration of Sentiment Summarization. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications* (pp. 12–15). Stanford, US.

Blair-Goldensohn, S., Neylon, T., Hannan, K., Reis, G. A., Mcdonald, R., & Reynar, J. (2008). Building a Sentiment Summarizer for Local Service Reviews. In *Workshop NLP in the Information Explosion Era* (pp. 14–23).

Carenini, G., Ng, R., & Pauls, A. (2006). Multi-document Summarization of Evaluative Text. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 305–312).

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, *22*, 249–254.

Conrad, J. G., Leidner, J. L., Schilder, F., & Kondadadi, R. (2009). Query-based Opinion Summarization for Legal Blog Entries. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law* (pp. 167–176). ACM.

Dang, H. T. (2005). Overview of DUC 2005. In *Proceedings of the Document Understanding Conference*.

Embar, V. R., Deshpande, S. R., Vaishnavi, A. K., Jain, V., & Kallimani, J. S. (2013). sArAmsha - A Kannada Abstractive Summarizer. In *Advances in Computing, Communications and Informatics (ICACCI)* (pp. 540–544).

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479.

Fabbrizio, G. D., Stent, A., & Gaizauskas, R. (2014). A Hybrid Approach to Multi-document Summarization of Opinions in Reviews. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)* (pp. 54–63). Philadelphia, Pennsylvania, U.S.A.: Association for Computational Linguistics.

Fonseca, E. R., & Rosa, J. L. G. (2013). Mac-Morpho Revisited: Towards Robust Part-of-Speech Tagging. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology* (pp. 98–107).

Freitas, C., Motta, E., Milidi, R., & Cesar, J. (2013). Sparkle Vampire LoL! Annotating Opinions in a Book Review Corpus. In *11th Corpus Linguistics Conference* (pp. 128–146).

Ganesan, K., Zhai, C., & Han, J. (2010). Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)* (pp. 340–348). Beijing, China.

Gerani, S., Mehdad, Y., Carenini, G., Ng, R. T., & Nejat, B. (2014). Abstractive Summarization of Product Reviews Using Discourse Structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1602–1613). Association for Computational Linguistics.

Guzman, E., & Maalej, W. (2014). How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews. In *Requirements Engineering Conference (RE), 2014 IEEE 22nd International* (pp. 153–162).

Hahn, U., & Mani, I. (2000). The Challenges of Automatic Summarization. *Computer*, *33*, 29–36.

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '04 (pp. 168–177). New York, NY, USA: ACM.

Jung, J., & Jo, G. S. (2003). Template-Based E-mail Summarization for Wireless Devices. In A. Yazc, & C. ener (Eds.), *Computer and Information Sciences - ISCIS 2003* (pp. 99–106). Springer Berlin Heidelberg volume 2869 of *Lecture Notes in Computer Science*.

Kim, H. D., Ganesan, K., Sondhi, P., & Zhai, C. (2011). *Comprehensive Review Of Opinion Summarization*. Technical Report University of Illinois at Urbana-Champaign.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Digital Library of Engineering and Computer Science. Morgan & Claypool.

López, 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., & Di Felippo, A. (2015). A Qualitative Analysis of a Corpus of Opinion Summaries based on Aspects. In *Proceedings of The 9th Linguistic Annotation Workshop* (pp. 62–71). Denver, Colorado, USA: Association for Computational Linguistics.

Mani, I. (2001). Summarization Evaluation: An Overview. In *Proceedings of the Third Second Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, NTCIR-2*.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, *8*, 243–281.

Maziero, E., Pardo, T. A. S., da Cunha, I., Torres-Moreno, J.-M., & SanJuan, E. (2011). DiZer 2.0-An Adaptable On-line Discourse Parser. In *Proceedings of the III RST Meeting (8th Brazilian Symposium in Information and Human Language Technology)* (pp. 50–57).

Mithun, S., & Kosseim, L. (2009). Summarizing Blog Entries Versus News Texts. In *Proceedings of the Workshop on Events in Emerging Text Types* eETTs '09 (pp. 1–8). Stroudsburg, PA, USA: Association for Computational Linguistics.

Nóbrega, F. A. A., Agostini, V., Camargo, R. T., Felippo, A. D., & Pardo, T. A. S. (2014). Alignment-Based Sentence Position Policy in a News Corpus for Multi-document Summarization. In *Computational Processing of the Portuguese Language - 11th International Conference, PROPOR* (pp. 286–291).

Ouyang, Y., Li, W., Lu, Q., & Zhang, R. (2010). A Study on Position Information in Document Summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* COLING '10 (pp. 919–927). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. (pp. 1–135). Hanover, MA, USA: Now Publishers Inc. volume 2.

Radev, D. R., & McKeown, K. R. (1998). Generating Natural Language Summaries from Multiple On-line Sources. *Computational Linguistics*, *24*, 470–500.

Reiter, E. (1994). Has a Consensus NL Generation Architecture Appeared, and is It Psycholinguistically Plausible? In *Proceedings of the Seventh Interna-*

tional Workshop on Natural Language Generation INLG '94 (pp. 163–170). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ribaldo, R., Akabane, A. T., Rino, L. H. M., & Pardo, T. A. S. (2012). Graph-Based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information. In *Proceedings of the International Conference on Computational Processing of Portuguese* Lecture Notes in Computer Science (pp. 260–271). Springer.

Seki, Y. (2002). Sentence Extraction by TF/IDF and Position Weighting from Newspaper Articles. In *Proceedings of the 3rd NTCIR Workshop*.

Sharifi, B., Hutton, M.-A., & Kalita, J. K. (2010). Experiments in Microblog Summarization. In *SocialCom/PASSAT* (pp. 49–56). IEEE Computer Society.

Steinhaus, H. (1956). Sur la Division des Corps Matériels en Parties. In *Bulletin of the Polish Academy of Sciences* (pp. 801–804).

Tadano, R., Shimada, K., & Endo, T. (2010). Multi-aspects Review Summarization Based on Identification of Important Opinions and their Similarity. In *PACLIC* (pp. 685–692). Institute for Digital Enhancement of Cognitive Development, Waseda University.

Van Deemter, K., Krahmer, E., & Theune, M. (2005). Real Versus Template-Based Natural Language Generation: A False Opposition? (pp. 15–24). Cambridge, MA, USA: MIT Press volume 31.

Wang, D., Zhu, S., & Li, T. (2013). SumView: A Web-based Engine for Summarizing Product Reviews and Customer Opinions. (pp. 27–33). Tarrytown, NY, USA: Pergamon Press, Inc. volume 40.

Xu, X., Meng, T., & Cheng, X. (2011). Aspect-based Extractive Summarization of Online Reviews. In *Proceedings of the 2011 ACM Symposium on Applied Computing* (pp. 968–975). New York, NY, USA: ACM.

Zajic, D., Dorr, B. J., Lin, J., & Schwartz, R. (2007). Multi-candidate Reduction: Sentence Compression as a Tool for Document Summarization Tasks. (pp. 1549–1570). Tarrytown, NY, USA: Pergamon Press, Inc. volume 43.