A Qualitative Analysis of a Corpus of Opinion Summaries based on Aspects

Roque E. López[1], Lucas V. Avanço[1], Pedro P. B. Filho[1], Alessandro Y. Bokan[1], Paula C. F. Cardoso[1], Márcio S. Dias[1], Fernando A. A. Nóbrega[1], Marco A. S. Cabezudo[1], Jackson W. C. Souza[2], Andressa C. I. Zacarias[2], Eloize M. R. Seno[3], Ariani Di Felippo[2], Thiago A. S. Pardo[1]

Interinstitutional Center for Computational Linguistics (NILC)

Institute of Mathematical and Computer Sciences, University of São Paulo[1]
Av. Trabalhador São-Carlense, 400 - Centro, São Carlos, Brazil

Federal University of São Carlos[2]
Rodovia Washington Luís, Km 235, P.O.Box 676, São Carlos, Brazil

Federal Institute of São Paulo[3]
Rodovia Washington Luís, Km 235, AT-6, Room 119, São Carlos, Brazil

Abstract

Aspect-based opinion summarization is the task of automatically generating a summary for some aspects of a specific topic from a set of opinions. In most cases, to evaluate the quality of the automatic summaries, it is necessary to have a reference corpus of human summaries to analyze how similar they are. The scarcity of corpora in that task has been a limiting factor for many research works. In this paper, we introduce OpiSums-PT, a corpus of extractive and abstractive summaries of opinions written in Brazilian Portuguese. We use this corpus to analyze how similar human summaries are and how people take into account the issues of aspect coverage and sentiment orientation to generate manual summaries. The results of these analyses show that human summaries are diversified and people generate summaries only for some aspects, keeping the overall sentiment orientation with little variation.

1 Introduction

Opinion summarization, also known as sentiment summarization, is the task of automatically generating summaries for a set of opinions about a specific target (Conrad et al., 2009). According to Liu (2012), there are three main approaches to generate summaries of opinions: traditional summarization, contrastive view summarization and aspect-based summarization. Most of the works in opinion summarization follows the aspect-based approach, because it produces summaries with more information (Hu and Liu, 2004).

Aspect-based opinion summarization generates summaries of opinions for the main aspects of an object or entity. Objects could be products, services, organizations (e.g., a smartphone), and aspects are attributes or components of them (such as the battery or the screen for a smartphone). An automatic system of aspect-based opinion summarization receives as input a set of opinions about an object and produces a summary that expresses the sentiment for some relevant aspects.

Opinion summaries could be extractive or abstractive. Most automatic methods in opinion summarization produces extractive summaries, which are created selecting the most representative text segments (usually sentences) from the original opinions (Mani, 1999) (Radev et al., 2004). An opinion summary could also be abstractive, in which the content of the summary is rewritten using new text segments (Radev and McKeown, 1998) (Lin and Hovy, 2000). There are few works that produce abstractive summaries, because they require some complex Natural Language Processing tasks such as text generation or sentence fusion.

In both cases, to evaluate the performance of au-

Proceedings of LAW IX - The 9th Linguistic Annotation Workshop, pages 62–71,
Denver, Colorado, June 5, 2015.

tomatic methods, it is usually necessary to have a reference corpus of human summaries. With a corpus, automatic and human summaries could be compared to know how similar they are. Through that comparison, we could identify the errors of these automatic methods and, consequently, improve their performance. Moreover, a corpus of opinion summaries could be used in machine learning methods as training data to learn patterns for extracting important information from opinions.

Unfortunately, there are few available corpora for aspect-based opinion summarization (Ganesan et al., 2010) (Zhu et al., 2013) (Kim and Zhai, 2009), which difficults the progress of this task. Most of these corpora have focused on English. For Brazilian Portuguese language, to the best of our knowledge, there is no available corpus of opinion summaries.

In this paper, we present OpiSums-PT (**Opi**nion **Sum**marie**s** in **P**or**t**uguese), a corpus of opinion summaries based on aspects, written in Brazilian Portuguese. OpiSums-PT contains multiple human summaries, in which each summary comes from the analysis of 10 opinions. The building of this corpus was motivated by two main reasons: (i) to address the absence of a corpus of opinion summaries in Brazilian Portuguese and (ii) to evaluate how people generate summaries of opinions. Particularly, we analyze how similar human summaries are (for the same set of opinions) and how important the information of aspect coverage and sentiment orientation are.

The results of these analyses indicate that agreement for human summaries, in terms of Kappa coefficient (Carletta, 1996) and ROUGE-1 measure (Lin, 2004), is low. The results also show that people generate summaries only for some aspects and they keep the overall sentiment orientation, with little variation, in the summaries.

The remaining of the paper is organized as follows: in Section 2, we introduce the main related works; in Section 3, we describe the resources used in this research; in Section 4, we explain how the corpus of summaries was created; the experiments and results of annotator agreement, aspect coverage and sentiment orientation are presented in Section 5; finally, in Section 6, we conclude this work.

2 Related Work

Many research works in aspect-based opinion summarization have created their own dataset crawling review websites or social networks. Of these resources, few could be considered as standard datasets. The dataset proposed in Hu and Liu (2004) is the most used resource in aspect-based opinion summarization. However, that corpus did not contain manual summaries, but aspects annotated and their associated sentiment. To evaluate automatic summaries in those works, the authors have used survey questions to select the best summaries.

In previous works in which opinion summaries were manually created, the annotation of the corpus has not been described in detail because it was not the main focus of these studies.

In Tadano et al. (2010), three participants annotated 25 reviews (approximately with 450 sentences) of opinions about a videogame. From the 25 reviews, 50 sentences were selected to the summary. In the experiments, ROUGE-1 measure between the annotator's summaries was 0.480, which shows that it is difficult to generate the same summary for opinions, even among humans.

Xu et al. (2011) crawled 32,007 reviews for three aspects (food, service and ambience) from 173 restaurants. From these reviews, 10 restaurants were chosen for evaluations and 7 restaurants to configure some parameters of the automatic method proposed by Xu et al. For each aspect of a restaurant, the authors created an extractive summary selecting several sentences with representative and diverse opinions. Each summary was composed by 100 words in average.

In Carenini et al. (2006), 28 annotators created abstractive summaries for a corpus of reviews about a digital camera and a DVD player. Each participant in the annotation received 20 reviews randomly selected from the corpus and generated a summary of 100 words. As instructions, the participants assumed that they worked for a manufacturer of products (either digital camera or DVD player). The purpose of these instructions was to motivate the user to look for the most important information worthy of summarization.

Ganesan et al. (2010) created a corpus of manual abstractive summaries using reviews of hotels,

cars and various electronic products. To collect the reviews, the authors used 51 "topic queries" (e.g., Ipod:sound and Toyota:comfort). Each "topic query" had 100 redundant sentences related to the query. Ganesan et al. used a crowdsourcing marketplace to get 5 human workers to create 5 different summaries for each "topic query". After the creation of the summaries, the authors reviewed each set of summaries and dropped summaries that had little or no correlation with the majority of them. Finally, each "topic query" had approximately 4 reference summaries.

Unlike these works, we performed a qualitative analysis of opinion summaries based on aspects. Besides that, we also compare extractive and abstractive summaries in terms of annotators agreement, aspect coverage and sentiment orientation. To the best of our knowledge, there are no similar works, most likely due to the difficulty of generating human-written summaries for opinions.

3 Corpora

To create the corpus of opinion summaries, we used reviews from two domains: books and electronic products. For the first one, we used the opinions of ReLi corpus (Freitas et al., 2013), a collection of opinions about 13 books. For the second domain, we collected reviews of 4 electronic products from Buscapé[1] website. The purpose of using these two domains is to have a corpus with different characteristics in the opinions. In the following sections, these two resources are explained in more detail.

3.1 Books

For book opinions, we used the ReLi corpus (Freitas et al., 2013). This corpus is composed of 1,600 reviews with 12,000 sentences about 13 books written by 7 famous authors of classical and contemporary literature. The opinions of ReLi were freely written by different users in specialized review websites.

The annotated opinions in ReLi are directly related to the books and their aspects (e.g., characters, chapters and story). Opinions about other books or movies of the books were not considered. In ReLi, reviews were annotated at the segment and sentence levels in three phases: (i) identification and annotation of the sentence polarity, (ii) identification of objects in sentences and (iii) identification of polarity in segments that contain sentiment. E.g., for the sentence "*The book is very interesting but its chapters are too long*", the polarity sentence is positive, the identified objects are *book* and *chapters*, and the polarities for the segments *very interesting* and *too long* are positive and negative, respectively.

The annotation of ReLi was conducted by linguists who attended a training process to be familiar with the task and instructions. According to Freitas et al. (2013), the agreement was calculate in a sample of 170 reviews and the obtained results were satisfactory. In the polarity identification of sentences, identification of objects and polarity identification in segments that contain sentiment, the agreement values were 98.3%, 72.6% and 99.8% in average, respectively.

For the annotation of our corpus, we randomly selected 10 reviews for each book of ReLi, taking as example other related works ((Carenini et al., 2006), (Tadano et al., 2010)) that have used a similar number of opinions as data source. In the selection of reviews, we determined that they contain at most 300 words. We used this filter because people prefer to read concise opinions instead of reviews with too many words. This criterion was also used in the selection of electronic product opinions.

3.2 Electronic Products

We collected opinions about electronic products from Buscapé, a website where users comment about different products (e.g. smartphones, clothes, videogames, etc.). These comments are written in a free format within a template with three sections: Pros, Cons, and Opinion.

To create the corpus of summaries, we collected a set of reviews about 4 electronic products: 2 smartphones (Samsung Galaxy S III and Iphone 5) and 2 televisions (LG Smart TV and Samsung Smart TV). For each product, we randomly selected 10 reviews.

This set of reviews was annotated by one person with strong knowledge in Sentiment Analysis. The annotation consisted in the identification of product aspects, e.g., battery and photo for smartphones, and sound and price for televisions. The identification of the polarity of segments that contain sentiment about the aspects was also annotated.

[1] http://www.buscape.com.br/

4 Corpus Annotation

According to Ulrich et al. (2008), abstractive summarization is the main goal of many research works, since it is what people naturally do, but extractive summarization has been more explored and effective since it is easier to compute. In this annotation, we generated both, extractive and abstractive summaries, to assistant different researches and to analyze how they are generated in opinions.

In OpiSums-PT, we created multiple reference summaries in order to reduce the overall subjectivity and any possible bias. For each book and electronic product, we generated 5 extractive and 5 abstractive summaries. In total, 170 summaries were manually created. Table 1 shows the content of OpiSums-PT in relation to the number of sentences, tokens, types and their average by summary.

Table 1: Content of OpiSums-PT

Features	**Extractive Summaries**	**Abstractive Summaries**
Summaries	85	85
Sentences	534	430
Tokens	8435	8611
Types	1702	1833
Average sentences by summary	6.3	5.1
Average tokens by summary	99.2	101.3
Average types by summary	71.1	72.4

This annotation was carried out by 14 participants with strong knowledge in Computational Linguistics and Natural Language Processing. Each participant created 12 summaries approximately during the annotation process. Each set of 5 summaries (extractive or abstractive) was generated by 5 different annotators.

To generate a summary, either extractive or abstractive, each annotator read 10 opinions about books or electronic products. This number of opinions was chosen because we believe that, when people look for opinions, they do not read large amounts of opinions, but a small sample of them.

The task of annotation was daily performed during 13 days, approximately. In the first meeting, the annotators received a training session together with the annotation manual document to be familiar with the task. In that document, we presented all instructions as well as the aspects identified in the opinions of ReLi and Buscapé. These aspects were taken from the annotation of these two data sources and were shown to the participants with the sole intention that annotators know them. Table 2 shows the objects and aspects presented to the participants in the annotation of OpiSums-PT.

Table 2: Objects and aspects identified in opinions

Objects	**Aspects**
Books	characters, story, chapters, dialogues, phrases, author's style, titles, images, vocabulary, text
Smartphones	battery, design, processor, screen, price, camera, weight, operating system, internet, photo, video, wi-fi, sound, size, headphones, speed, chip
TVs	design, price, camera, image quality, brightness, wi-fi, sound, durability, internet

In the other days of annotation, the annotators created summaries at home and sent them by email, as it was conducted in (Dias et al., 2014). Each day, an annotator generated only one summary (extractive or abstractive). We opted for this scheme in order to simplify the task for annotators and, consequently, to get good summaries.

Another instruction in the annotation was related to the summary length. Both extractive and abstractive summaries should be composed by 100 words with a tolerance of ± 10 words, approximately. We choose the same number of words for these types of summaries to evaluate how they are generated under similar restrictions. A compression ratio in percentage (e.g., 25%) was not used because the vast majority of the works in aspect-based opinion summarization do not use this scheme (Carenini et al., 2006) (Ganesan et al., 2010) (Tadano et al., 2010).

4.1 Extractive Summaries

To create extractive summaries in our annotation, we asked the annotators to select the most important sentences from the original opinions. We did not establish a criterion to determine the importance of a sentence, it was a decision of each annotator. Likewise, we did not oblige to exclude sentences with dangling anaphora. We opted for this autonomy with the purpose that the creation of summaries to be as natural as possible. The number of aspects included in the final summary was chosen by each annotator.

The final summary was composed by complete

sentences. It was not allowed to rewrite the sentences of the original opinions. If a sentence presented misspellings and/or grammatical mistakes, they should not be corrected.

Each sentence of the source opinions had an identifier in the end part. This identifier allowed linking the summary sentence with the source opinion. Thus, for example, the identifier "<D20_S3>" indicates the third sentence of the opinion (document) 20. Figure 1 shows an example of an extractive summary (in bold, the identifiers of the sentences).

Um Smartphone quase Perfeito! **<D3_S1>**
O que gostei: Hoje é o melhor no mercado em relação ao seu processamento. **<D2_S3>**
A bateria dura bastante e os aplicativos ja instalados sao otimos. **<D7_S5>**
A camera é maravilhosa. **<D7_S4>**
O que não gostei: Ele esquenta um Pouco na parte de baixo mas não chega a incomodar, na cor branca ele parece ser muito frágil e o S Voice ainda não funciona em português. **<D3_S5>**
Esperava muito mais do Galaxy SIII pelo suspense que a Samsung promoveu. **<D2_S1>**
Depois dessa, quem tem coragem de investir em média R$ 1.700,00 no Galaxy SIII ou tentar a sorte com o Galaxy S4? **<D6_S9>**

[Translation]

A Smartphone almost perfect! **<D3_S1>**
What I liked: Today is the best on the market in relation to its processing. **<D2_S3>**
The battery lasts a lot and its installed applications are great. **<D7_S5>**
The camera is wonderful. **<D7_S4>**
What I did not like: It heats a little at the bottom but not enough to bother, in white color it seems very fragile and the S Voice does not work yet in Portuguese. **<D3_S5>**
I expected more of Galaxy SIII due to the suspense that Samsung promoted. **<D2_S1>**
After that, who has the courage to invest around R$ 1,700.00 in Galaxy SIII or try luck with the Galaxy S4? **<D6_S9>**

Figure 1: Example of Extractive Summary

As we can see in Figure 1, the extractive summary is composed by seven sentences from different opinions (D2, D3, D6 and D7). This happened frequently in our extractive summaries, indicating that relevant sentences for annotators were written by different web users. As consequence of this, the lack of cohesion between summary sentences was notorious.

4.2 Abstractive Summaries

To create abstractive summaries is more difficult than extractive summaries, since it implies generating new text. In our annotation, we asked the annotators to generate summaries as rewritten as possible in order to get more differentiated summaries in relation to the extractive summaries.

Abstractive summaries should indicate the actual scenario of source opinions (general predominant sentiment). Similar to the extractive summaries, the number of aspects to be included in abstractive summaries and the structure of the text were decisions of each annotator.

In Figure 2, we show an example of abstractive summary about Twilight book. In the first part of the text, the author's summary gives the overall sentiment for this book, and, then, describes the web user's sentiment for some book aspects. This structure was adopted by the majority of annotators.

A grande maioria dos leitores avaliaram negativamente o livro Crepúsculo, pois em geral, eles argumentaram que o livro tem um romance exagerado. Entre as principais desvantagens do livro, os leitores mencionaram que os personagens são superficiais, a escrita é péssima e a história é chata. Muitos dos usuários não conseguiram terminar de ler o livro e não recomendariam ele para outras pessoas. Por outro lado, outra pequena parte dos leitores acharam que o livro Crepúsculo é bom, pois consideraram que ele é intenso, romântico, cheio de mistérios e brilhante. Estes leitores afirmaram que, embora Crepúsculo seja um livro fictício, ele mostra a importância de um verdadeiro amor.

[Translation]

The vast majority of readers evaluated negatively Twilight book, because, in general, they argued that it has an exaggerated romance. Among the main disadvantages of this book, readers mentioned that characters are superficial, the writing is bad and the story is boring. Many users were not able to finish the reading of the book and they would not recommend it to other people. On the other hand, another small part of readers think that Twilight book is good, because they considered it intense, romantic, full of mysteries and amazing. These readers said that, although Twilight is a fictional book, it shows the importance of the true love.

Figure 2: Example of Abstractive Summary

In comparison with extractive summaries, these ones did not present the problem of lack of cohesion and show explicitly what was the predominant sentiment in the source opinions.

5 Experiments

After the annotation, we performed some experiments over OpiSums-PT. First, we calculated the annotators agreement to know how difficult this task is. Second, we analyzed the aspect coverage to estimate the proportion of aspects that is preserved in the summaries. Finally, the sentiment orientation in the summaries was computed to verify if it is proportional to the general sentiment in source opinions.

In this paper, we focused on these three issues. It is believed that (i) people generate not very similar opinion summaries, (ii) not all aspects are consid-

Table 3: Annotators agreement results

Books/ Electronic Products	Extractive Summary					Abstractive Summary
	Total Agreement	Majority Agreement	Minority Agreement	No Agreement	ROUGE-1	ROUGE-1
Capitães da Areia	0.000	0.267	0.200	0.533	0.405	0.218
Crepúsculo	0.000	0.286	0.357	0.357	0.414	0.239
Ensaio sobre a Cegueira	0.000	0.043	0.217	0.739	0.250	0.251
Fala sério. amiga!	0.077	0.154	0.154	0.615	0.606	0.299
Fala sério. amor!	0.118	0.118	0.294	0.471	0.600	0.287
Fala sério. mãe!	0.000	0.222	0.167	0.611	0.325	0.308
Fala sério. pai!	0.000	0.143	0.143	0.714	0.418	0.352
Fala sério. professor!	0.000	0.235	0.353	0.412	0.344	0.345
O Apanhador nos Campos de Centeio	0.000	0.091	0.409	0.500	0.360	0.253
O Outro lado da meia noite	0.000	0.136	0.182	0.682	0.392	0.232
O Reverso da Medalha	0.000	0.100	0.250	0.650	0.339	0.305
Se houver Amanhã	0.000	0.200	0.200	0.600	0.471	0.309
1984	0.000	0.263	0.316	0.421	0.366	0.238
Iphone 5	0.000	0.308	0.154	0.538	0.342	0.230
Samsung Galaxy S III	0.000	0.100	0.200	0.700	0.235	0.276
LG Smart TV	0.000	0.040	0.240	0.720	0.274	0.270
Samsung Smart TV	0.000	0.238	0.333	0.429	0.451	0.270
Average	**0.011**	**0.173**	**0.245**	**0.570**	**0.388**	**0.275**

ered in the final summary and (iii) humans consider the sentiment orientation to create an opinion summary. However, as far as we know, there are no previous works that proved these hypotheses. In this study, we explore these three hypotheses.

5.1 Inter-Annotator Agreement

We calculated the inter-annotator agreement for extractive and abstractive summaries. For both, we used the ROUGE score (Lin, 2004). For extractive summaries, Kappa coefficient (Carletta, 1996) was also calculated, as well as the percentage of common sentences in the summaries.

In extractive summaries, we calculated Kappa agreement for each book and electronic product, taking the sentences of source opinions and verifying which of them were included in the human summaries. In average, the Kappa value obtained in the experiments was 0.185. According to Liu and Liu (2008), the Kappa values reported for text and meeting summarization were 0.38 and 0.28 in average, respectively. Compared to these values, the Kappa agreement obtained by us in aspect-based opinion summarization is lower. This is likely due to the fact that in opinion summarization there are many different sentences that express the same meaning. Thus, different annotators could have chosen different sentences with similar content.

To compensate this problem of Kappa, we also used the ROUGE-N score. The ROUGE measure computes the n-gram overlap between summaries and, thus, could help to identify sentences that are similar in content. In our experiments, we used the ROUGE-1 score (unigram overlap).

For each annotator, we computed ROUGE-1 scores using other annotators' summaries as references, and then we calculated the average between them. Table 3 shows the values of ROUGE-1 obtained for each book and electronic product in extractive and abstractive summaries. These results are better than Kappa results and may indicate that annotators choose different sentences that have similar content. The results for extractive summaries are better than abstractive summaries, because in abstracts annotators have independence to use different words, possibly synonyms and paraphrases.

For extractive summaries, we also computed the percentage of common sentences among the summaries created by annotators. In Table 3, we show the results. Total Agreement indicates the proportion of common sentences selected by five annotators; Majority Agreement, by four or three annotators; and Minority Agreement, by two annotators. No agreement indicates that annotators did not agree

in the selection of sentences.

On one hand, the results for these metrics indicate that annotators choose the same sentences in few cases. In average, only 1.1% (0.011) of sentences was selected by all annotators, and only 17.3% (0.173) of them by the majority of annotators. We believe that this is mainly due to the large number of sentences that annotators have to read to generate the summary (in average, 40 sentences). On the other hand, in many cases, annotators choose different sentences (see columns Minority and No Agreement), because, as it is reported in (Rath et al., 1961), in the summarization task, there is no single set of representative sentences chosen by humans. In addition, we believe that some especial linguistic characteristics of opinions, such as irony or usage of slangs, make this task more challenging.

In general, all results reported in Table 3 show that it is difficult to generate similar opinion summaries based on aspects (extractive or abstractive) even among humans. Although these results are low, they could be used as a topline performance to evaluate other automatic methods.

5.2 Aspect Coverage

An important issue in aspect-based opinion summarization is the aspect coverage. Aspect coverage is an indicator of how many aspects of the source opinions are preserved in the generated summary. Most research works have been focused on producing a summary for each aspect (Blair-Goldensohn et al., 2008) (Tadano et al., 2010) (Xu et al., 2011). However, if we want an overall summary, that approach could be not ideal.

In our work, we produced overall summaries based on aspects, i.e., a summary contains the most important aspects, according to the annotators, for a set of source opinions. In the experiments, to calculate the aspect coverage, we considered the objects or entities as aspects, similar to Gerani et al. (2014).

To estimate the aspect coverage for extractive summaries, we get the aspects annotated in the opinions of ReLi and Buscapé, and then it was verified how many of them are preserved in the summaries. In abstractive summaries, we used a semi-automatic search. We look for aspects using a list with their names. After that, we manually reviewed the summaries in order to add possible synonyms to the aspect list. For example, the word "*narrative*" was considered a synonym of the "*story*" aspect. Finally, we determined how many aspects were in the summaries. For each book and electronic product, we calculated the proportion of aspects preserved in the five summaries, and then we computed the average.

Table 4 shows the percentage of aspect coverage for extractive and abstractive summaries. As we can see, abstractive summaries have wider coverage than extractive summaries because annotators have less restriction to write an abstractive summary and, thus, they can include more aspects. On the other hand, in extractive summaries, annotators are limited to the content of the source opinion's sentences.

Table 4: Coverage of aspects in summaries

Books/ Electronic Products	**Extractive Summary**	**Abstractive Summary**
Capitães da Areia	0.450	0.700
Crepúsculo	0.467	0.567
Ensaio sobre a Cegueira	0.300	0.600
Fala sério, amiga!	1.000	1.000
Fala sério, amor!	0.550	0.550
Fala sério, mãe!	0.400	0.767
Fala sério, pai!	0.800	0.900
Fala sério, professor!	0.700	1.000
O Apanhador nos Campos de Centeio	0.550	0.800
O Outro lado da meia noite	0.800	0.760
O Reverso da Medalha	0.650	0.800
Se houver Amanhã	0.640	0.680
1984	0.600	0.760
Iphone 5	0.444	0.578
Samsung Galaxy S III	0.333	0.400
LG Smart TV	0.514	0.714
Samsung Smart TV	0.720	0.760
Average	**0.583**	**0.726**

There are few cases where all aspects are included in the summaries (books "Fala sério, amiga!" and "Fala sério, professor!"). In these cases, less than three aspects were presented in source opinions. By contrast, when the number of aspects in the source opinions was high, few of them were included in the summary (e.g., product Samsung Galaxy S III). It was most notorious in electronic products because they have more technical opinions that include many aspects.

Results in Table 4 indicate that, for an overall aspect-based summary, humans consider only some aspects in the text. We did not find other works

Table 5: Sentiment orientation of summaries

Books/ Electronic Products	Actual Polarity		Extractive Summary		Abstractive Summary	
	Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Capitães da Areia	0.784	0.216	0.978	0.022	**0.370**	**0.630**
Crepúsculo	0.391	0.609	0.075	0.925	**0.510**	**0.490**
Ensaio sobre a Cegueira	0.812	0.188	0.880	0.120	**0.471**	**0.529**
Fala sério, amiga!	0.895	0.105	0.960	0.040	0.723	0.277
Fala sério, amor!	0.968	0.032	0.980	0.020	0.967	0.033
Fala sério, mãe!	0.510	0.490	0.680	0.320	0.569	0.431
Fala sério, pai!	0.842	0.158	0.877	0.123	0.950	0.050
Fala sério, professor!	0.621	0.379	0.791	0.209	0.686	0.314
O Apanhador nos Campos de Centeio	0.300	0.700	0.204	0.796	0.283	0.717
O Outro lado da meia noite	0.705	0.295	0.667	0.333	0.633	0.367
O Reverso da Medalha	0.667	0.333	0.521	0.479	0.558	0.442
Se houver Amanhã	0.867	0.133	0.952	0.048	0.716	0.284
1984	0.757	0.243	0.877	0.123	0.627	0.573
Iphone 5	0.975	0.025	0.971	0.029	0.810	0.190
Samsung Galaxy S III	0.584	0.416	**0.272**	**0.728**	**0.460**	**0.540**
LG Smart TV	0.622	0.378	0.674	0.326	0.753	0.247
Samsung Smart TV	0.556	0.444	0.502	0.498	0.536	0.464

to compare the results of aspect coverage, but we believe that our results show an approximation of how many aspects humans consider in a summary. Thus, automatic opinion summarization methods could use these results as indicator of how many aspects could be included in the summaries.

5.3 Sentiment Orientation

To communicate to summary's readers what is the sentiment in the opinions about the entity and its aspects is not simply a matter of classifying the summary as positive or negative. Summary's readers want to know if all opinions that evaluate the entity made it in a similar way or if they were varied. Thus, opinion summaries must preserve the polarity distribution as much as possible to reflect the overall sentiment about the entity and its aspects.

In our experiments, we evaluated how much humans (annotators) maintain the sentiment orientation in the manual summaries. To estimate the general sentiment presented in the source opinions, we extract the segments that contain sentiment with its polarities from the annotations of ReLi and Buscapé. We calculated the percentage of positive and negative segments. Table 5 shows the percentage of positive and negative sentiments presented in the source opinions (column "Actual Polarity") for each book and electronic product.

To calculate the sentiment in extractive summaries, we estimate the sentiment for positive and negative classes using the annotations of ReLi and Buscapé. For abstractive summaries, we calculated the sentiment with the automatic lexicon-based method proposed in Taboada et al. (2011) using the SentiLex lexicon (Silva et al., 2012), because, according to Balage Filho et al. (2013), it gets better results in comparison with other Brazilian Portuguese dictionaries.

Table 5 shows the results of the sentiment orientation for each book and electronic product. In general, annotators reflected the sentiment distribution of source opinions in the summaries. The proportions between positive and negative sentiments were not exactly the same, but were very similar. This shows that humans (annotators) take into account the sentiment to create the summary and consider both classes, positive and negative, according to how they appeared in the source opinions.

There are few cases where the sentiment orientation of summaries is opposite of the source opinions (marked in bold). This indicates that annotators focused only in one part of the source opinions ignoring the overall sentiment.

Extractive summaries got better correlations than abstractive summaries because the sentences of extractive summaries are the same of the source opinions and also because the sentiment in abstractive summaries was automatically calculated.

6 Conclusion

In this paper, we presented OpiSums-PT, a corpus of opinion summaries, extractive and abstractive, based on aspects written in Brazilian Portuguese. We also made a qualitative analysis about how people generate these types of summaries. As was previously showed, human summaries are diversified and people generate summaries only for some aspects keeping the overall sentiment orientation with little variation.

This work has been motivated, mainly, by the importance that a corpus has in this task and to assist future researches in the opinion summarization field.

The complete version of OpiSums-PT is available for download through the Sucinto project webpage[2] under a Creative Commons license.

Future work includes extending OpiSums-PT with other type of annotations, such as sentence alignment between summaries and identification of elementary discourse units.

Acknowledgments

Part of the results presented in this paper were obtained through research on a project titled "Semantic Processing of Texts in Brazilian Portuguese", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91. We would like to thank professor Lucia Rino and the other annotators for their valuable help in the building of the corpus.

References

Pedro Balage Filho, Thiago Pardo, and Sandra Aluísio. 2013. An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*, pages 215–219. Sociedade Brasileira de Computação.

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A Reis, and Jeff Reynar. 2008. Building a Sentiment Summarizer for Local Service Reviews. In *WWW Workshop on NLP in the Information Explosion Era*.

Giuseppe Carenini, Raymond Ng, and Adam Pauls. 2006. Multi-document Summarization of Evaluative Text. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 305–312.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.

Jack G. Conrad, Jochen L. Leidner, Frank Schilder, and Ravi Kondadadi. 2009. Query-based Opinion Summarization for Legal Blog Entries. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 167–176. ACM.

Márcio Dias, Alessandro Bokan, Carla Chuman, Cláudia Barros, Erick Maziero, Fernando Nobrega, Jackson Souza, Marco Sobrevilla, Marina Delege, Lucía Castro, Naira Silva, Paula Figueira, Pedro Balage, Roque López, Ariani Di Felippo, Maria das Graças Volpe, and Thiago Pardo. 2014. Enriquecendo o Córpus CSTNews - a Criação de Novos Sumários Multidocumento. In *Proceedings of the 1st Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish - ToRPorEsp*, pages 1–8.

Cláudia Freitas, Eduardo Motta, Ruy Milidiú, and Juliana Cesar. 2013. Sparkle Vampire LoL! Annotating Opinions in a Book Review Corpus. In *11th Corpus Linguistics Conference*.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 340–348. Association for Computational Linguistics.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita 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. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. ACM.

Hyun Duk Kim and ChengXiang Zhai. 2009. Generating Comparative Summaries of Contradictory Opinions in Text. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 385–394. ACM.

Chin-Yew Lin and Eduard Hovy. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. In *Proceedings of the 18th Conference on Computational Linguistics*, pages 495–501. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Looking for a Few Good Metrics: Automatic Summarization Evaluation-How

[2] http://www.icmc.usp.br/pessoas/taspardo/sucinto/

many Samples are Enough? In *Proceedings of the NTCIR Workshop*, pages 1–10.

Fei Liu and Yang Liu. 2008. What Are Meeting Summaries?: An Analysis of Human Extractive Summaries in Meeting Corpus. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 80–83. Association for Computational Linguistics.

Bing Liu. 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Inderjeet Mani. 1999. *Advances in Automatic Text Summarization*. MIT Press.

Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating Natural Language Summaries from Multiple On-line Sources. *Computational Linguistics*, 24(3):470–500.

Dragomir R. Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drábek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. 2004. MEAD - A Platform for Multidocument Multilingual Text Summarization. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*.

G.J. Rath, A. Resnick, and T.R. Savage. 1961. The Formation of Abstracts by the Selection of Sentences. *American Documentation*, 12(2):139–141.

Mário J. Silva, Paula Carvalho, and Luís Sarmento. 2012. Building a Sentiment Lexicon for Social Judgement Mining. In *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*, pages 218–228. Springer-Verlag.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307.

Ryosuke Tadano, Kazutaka Shimada, and Tsutomu Endo. 2010. Multi-aspects Review Summarization Based on Identification of Important Opinions and their Similarity. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 685–692.

Jan Ulrich, Gabriel Murray, and Giuseppe Carenini. 2008. A Publicly Available Annotated Corpus for Supervised Email Summarization. In *Proceedings of AAAI EMAIL Workshop*, pages 77–87.

Xueke Xu, Tao Meng, and Xueqi Cheng. 2011. Aspect-based Extractive Summarization of Online Reviews. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 968–975. ACM.

Linhong Zhu, Sheng Gao, Sinno Jialin Pan, Haizhou Li, Dingxiong Deng, and Cyrus Shahabi. 2013. Graph-Based Informative-Sentence Selection for Opinion Summarization. In *Advances in Social Networks Analysis and Mining (ASONAM)*, pages 408–412. IEEE.