Corpus Building and Evaluation of Aspect-based Opinion Summaries from Tweets in Spanish

Daniel Peñaloza, Rodrigo López, Juanjosé Tenorio, Héctor Gómez,
Arturo Oncevay-Marcos, Marco A. Sobrevilla Cabezudo
Research Group on Artificial Intelligence (IA-PUCP)
Departamento de Ingeniería, Pontificia Universidad Católica del Perú, Lima, Peru
{a20110645,a20112387,juanjose.tenorio,hector.gomez,arturo.oncevay,msobrevilla}@pucp.edu.pe

Abstract

This project involves the presentation and analysis of a corpus of Spanish extractive and abstractive summaries of opinions. The purpose of this work is to display a corpus of diverse summaries that could be used as a reference for academic research as we have not found one for the Spanish language as far as we know. We have analyzed the summaries based on the agreement between them as this shows how different they are written between each other and on aspect coverage and sentiment orientation as this proves the difference between the content that each summary tries to express. After the experimentation, we have found that even if each annotator uses a different expression to summarize a text, all of them contain similar messages. Furthermore, when writing, all of them prioritize on common aspects that are more representative of the corpus.

Keywords: aspect-based opinion summarization, corpora building, Spanish

1. Introduction

Automatic summarization is one of the most challenging problems in Natural Language Processing (NLP). The task is even more difficult if the summarized content is generated entirely by the program, and it does not present any extract of the original text. Theres are two kinds of summaries: abstractive and extractive (Labbé and Portet, 2012). In extractive summarization, we try to select the principal ideas or sentences and grouping them in the summary. In abstractive summarization, we try to write a text which contains all ideas.

The relevance of the summarization task increases if the context of the high variety and volume of information is considered. Social networks, such as micro-blogging sites, nowadays are offering the possibility to exploit any kind of publications, where user opinions are greatly valuable. Opinion mining (OM) is the NLP subfield advocated to this problem. The main goal is to extract a subjective value from the text, that might include a certain polarity degree, and it could be focused on a particular entity or aspect (Liu, 2012).

Therefore, it is likely to be a special difficulty in summarizing user opinions from social media, as there is not an specific order among all the opinion posts grouped by the same topic. Besides, each post may address a different aspect or entity related to the main topic, so there is a need to identify them first. After that, the subjectivity of each aspect must be measured, such as a positive or negative polarity (Wu et al., 2016). There is also a quantitative scope by considering the distribution of different positions regarding an opinion, as no one should be treated as more relevant than others (Liu, 2012).

Aspect-based opinion summarization is an attractive task in this context. However, the primary research is centralized in some specific languages or domains (mainly English), due to the lack of summaries corpus to evaluate the process in different ones. For that reason, this study proposes the development of a new opinion summaries corpus for aspect-based opinion mining in Spanish. The documents/posts are extracted from Twitter, and they are related to the political context in Spain. In addition, there is an evaluation performed over the new corpus. It is important to highlight this corpus will be available publicly for anyone to do further researches.

The paper is organized as follows: Section 2 describes the related work for building of opinion summaries corpora. Then, the corpus named STOMPOL is presented in Section 3. After that, the annotation process for extractive and abstractive summaries is detailed in Section 4, while the corpus analysis is reported and analyzed in Section 5. Finally, conclusions and future work are discussed.

2. Related Work

There are several works about Opinion Summarization for English. One problem is that every work used its own corpora to do their experiments.

The dataset proposed in (Hu and Liu, 2004) is the principal resource in aspect-based opinion summarization. Nevertheless, that corpus did not contain manual summaries, but aspects annotated and their associated sentiment.

Another works where opinion summaries were built are presented in (Tadano et al., 2010) (related to videogames), (Xu et al., 2011) (related to restaurants) and (Carenini et al., 2006) (related to digital camera and DVD player). They generated 25, 30 and 28 summaries respectively.

To do a quantitative analysis of the summaries, Tadano (Tadano et al., 2010) utilized ROUGE-1 to compare them as it indicates an n-gram recall between summaries. We have to highlight these works did not present any kind of qualitative analysis of opinion summaries based on aspects.

For romance languages, we have found one work about corpora building. This work was proposed in (López et al., 2015). In this paper, authors built a corpora of abstractive and extractive opinion summaries for Brazilian Portuguese

and did a qualitative analysis by comparing the aspect coverage between summaries because his generated summaries covered many aspects of the same entity. Equally important, They made a sentiment orientation analysis as them must maintain the same sentiment expressed of the entity that the reference text had.

For Spanish, as far as we know, there is only one work about opinion summarization. In this work, authors built opinion summaries for Tourism sector (Esteban and Lloret, 2017) and then they implemented an abstractive opinion summarization method.

3. STOMPOL Corpus

The STOMPOL Corpus is one the corpora proposed since the TASS 2015 (Cámara et al., 2016). This corpus have been used in the task of polarity classification at aspect level. The STOMPOL corpus is a corpus composed by Spanish tweets related to a political aspects that appeared in the Spanish political campaign of regional and local elections that were held on 2015.

Each tweet presents one or more of the following aspects in its content: Economics, Health System, Education, Political party and other aspects. Also, each aspect is related to one or more of these sentiments: positive, negative and neutral (according the presented aspect).

In general, these corpus is composed by 784 tweets about 6 political parties. In Table 1 are presented the distribution of tweets by political party in STOMPOL corpus. We have to note that one tweet may contain comments about one or more political parties.

Political Party	Number of tweets
Ciudadanos	135
Izquierda Unida	118
Partido Popular	240
Partido Socialista Obrero Español	159
Podemos	114
Unión, Progreso y Democracia	103

Table 1: STOMPOL Corpus

4. Annotation Process

The main goal of this work was to create reference extractive and abstractive summaries in order to support future works about opinion summarization.

Due to the large number of tweets in STOMPOL corpus and the difficulty to read all of them to generate a summary, we chose to extract a few number of tweets for each political party. Table 2 shows the number of tweets considered for each one.

One point to highlight is the different number of tweets for each political party. It happened because we tried to select a sample which represented the real distribution of aspects in the STOMPOL corpus due to the imbalanced of corpus at aspect-level and, also, tried to cover all aspects.

In general, we created several reference summaries (abstractive and extractive) in order to reduce any possible bias and to understand the opinion summary generation in a better way. For each political party, we generated 3 extractive and 3 abstractive summaries, building 36 opinion

Political Party	Number of tweets
Ciudadanos	36
Izquierda Unida	26
Partido Popular	27
Partido Socialista Obrero Español	26
Podemos	30
Unión, Progreso y Democracia	33

Table 2: Sub corpus of STOMPOL Corpus

summaries in total. Although the number of summaries is small, it is enough for an analysis as seen on the related works. Table 3 shows the statistics of the corpus generated, in terms of number of summaries, sentences, tokens and its average.

	Abstractive	Extractive
Number of summaries	18	18
Number of sentences	77	130
Number of tokens	1864	1859
Average sentences by summary	4.27	7.22
Average tokens by summary	103.56	103.28

Table 3: Statistics of Corpus

Both kinds of summaries were handcrafted by 4 annotators. Each annotator had to generate one abstractive summary and one extractive summary for each political party. This task had a duration of 1 week and a half.

In order to improve to annotation process, we established some guidelines in relation to how to build the summary, the aspects coverage and the summary length.

In relation to how to build the summary, we suggested to the annotator to read all tweets and try to understand the principal ideas. Then, they had to build the summaries oriented to aspects.

In relation to aspects coverage, we provided a list of aspects (listed in Section 3) and the possible words related to them in order to facilitate the annotation process.

Finally, in relation to the summary length, we opted for generating summaries which contained between 90 and 110 words. We did not choose a compression rate because it is not used in the literature for this kind of task (López et al., 2015).

4.1. Extractive Summaries

In order to facilitate the building of extractive summaries, we had to do one previous step. This consisted in distinguish all tweets and the sentences in them. To do this, we assigned an identifier for each tweet of a political party. Then, we manually separated the sentences in the tweet and assigned an identifier to them too. We may see an example of extractive summary about the political party Podemos in Figure 1. In this Figure we may see the identifier <D5_S1>(first line), where D5 represents the tweet number 5 and S1 represents the sentence number 1 of that tweet. After this previous step, the annotators had to select some sentences which represented relevant sentimental information about each political party.

One point to highlight in Figure 1 is that extractive summary is composed by seven sentences from different tweets (D5, D32, D8, D10, D17, D16 and D9). This indicates that

```
@lematelot49 @ahorapodemos llevo toda mi
vida acompañado de esa palabra "crisis" para que
me vengan con milongas la OCDE <D5_S1>
@Pablo_Iglesias_              @CorruptionTimes
@flameycitron y dicen que no hay dinero para
los enfermos de Hepatitis C <D32_S1>
@la_tuerka @Pablo_Iglesias_ dice conocer la
realidad social de los más necesitados pero calla
ante asesinato de miles de cristianos por ISIS
<D8_S1>
¿Perdiendo votos antes de empezar? <D10_S2>
@ahorapodemos pues aunque parezca increíble,
han vivido a costa de las minorías. <D17_S1>
@ahorapodemos @ierrejon que poca humanidad
y poca verguenza <D16_S1>
@andresysara10 en #España ???? , por desgracia
, para tener políticos honrados , debemos esperar
a q nazcan . <D9_S1>
```

Figure 1: Example of extractive summary

important sentences for annotators were written by different users. Finally, we may see the lack of coherence in this kind of summary.

4.2. Abstractive Summaries

In the case of abstractive summaries, the annotators had to understand the overall opinion and write the summary in their own words. The summary should contain the main aspects and to detail (if it is possible) the reasons which motivate each aspect-level sentiment.

We may see an example of abstractive summary about the political party Podemos in Figure 2.


```
En lo que se refiere a economía, se hace mención
a una crisis posiblemente asociada con el partido.
Mientras que lo que respecta a sanidad hay
pacientes que no pueden ser atendidos en los
hospitales. Además de esto, cuando se habla
sobre el partido la gente expresa abiertamente su
descontento con los políticos y con el partido en
sí, ya sea por sus posiciones u opiniones en
diversos temas, o también debido a que tan
involucrados están en temas de corrupción. Este
partido empieza a perder apoyo puesto que
ocultan información y su pasado es bastante
turbio.
```

Figure 2: Example of abstractive summary

5. Corpus Analysis

The analysis of the corpus consisted of three experiments. First, we checked the inter-annotator agreement between summaries to see how similar they are based on the words that each annotator used. Second, we checked the aspect coverage of each summary to see how related they are based on what elements are each summary talking about. Finally, we checked the sentiment orientation of each summary to see how comparable they are based on the feeling that each annotator wanted to express.

Summary	Abstractive	Extractive
Ciudadanos	0.310	0.692
Izquierda Unida	0.351	0.649
Partido Popular	0.282	0.568
Partido Socialista Obrero Español	0.311	0.609
Podemos	0.293	0.606
Unión Progreso y Democracia	0.270	0.587

Table 4: ROUGE metric of the summaries

Summary	No Agreement	Agreement	Total Agreement
Ciudadanos	0.818	0.000	0.090
Izquierda Unida	0.615	0.307	0.076
Partido Popular	0.789	0.105	0.105
Partido Socialista Obrero Español	0.705	0.235	0.058
Podemos	0.733	0.133	0.133
Union Progreso y Democracia	0.700	0.300	0.000

Table 5: Sentence Agreement on Extractive Summaries

The experiments were done to prove three affirmations. (i) There is not a perfect summary because each person has a different way of expressing themselves, (ii) there are aspects that are more relevant to the annotator when writing a summary and (iii) each annotator adds a sentiment to the summary they want to write.

5.1. Inter-Annotator Agreement

In order to obtain the inter-annotator agreement we used the ROUGE metric. The measure counts the number of overlapping words between the generated summary and a reference summary (Lin, 2004) For our experiment we used Rouge-1, this compares single words between summaries and is expressed as a recall. We used this metric to compare all the summaries of an specific entity with each other for both the extractive summaries and the abstractive summaries. The results of these experiments are in Table 4.

On every case, extractive summaries had better results than abstractive summaries. This is because on abstractive summaries, the annotators have the freedom to use any word they want and the metric compares that the words used between them are exactly the same.

Additionally, we have also used the tags of the extractive summaries to obtain if the annotators are using the same sentences. This can be seen in Table 5. The results indicate that even when having the option to use the same words for the summary, the annotators do not do that except for some important sentences that are common between them.

5.2. Aspect Coverage

As we have seen on the results of the inter-Annotator Agreement, the annotators usually use different sentences in order to express themselves. For that reason, we have also reviewed the aspects of the entity that have been referenced by the annotators as they express more the ideas that the annotators wanted to include on their summaries.

For the extractive summaries we used the tags of the STOMPOL corpus that indicate which aspect has been referenced for a sentence. Furthermore, a sentence can reference multiple times the same aspect so we had only taken into consideration which one of the aspects are being referenced not the frequency of repetition.

For the abstractive summaries we did a manual review of each summary. Each annotator used their own words to

Summary	Extractive	Abstractive
Ciudadanos	**0.40**	**0.33**
Ciudadanos 1	0.40	0.40
Ciudadanos 2	0.40	0.40
Ciudadanos 3	0.40	0.20
Izquierda Unida	**0.93**	**0.46**
Izquierda Unida 1	1.00	0.60
Izquierda Unida 2	1.00	0.40
Izquierda Unida 3	0.80	0.40
Partido Popular	**0.66**	**0.60**
Partido Popular 1	0.60	0.60
Partido Popular 2	0.80	0.60
Partido Popular 3	0.60	0.60
Partido Socialista Obrero Español	**0.53**	**0.46**
Partido Socialista Obrero Español 1	0.40	0.40
Partido Socialista Obrero Español 2	0.60	0.40
Partido Socialista Obrero Español 3	0.60	0.60
Podemos	**0.66**	**0.60**
Podemos 1	0.60	0.60
Podemos 2	0.80	0.60
Podemos 3	0.60	0.60
Union Progreso y Democracia	**0.66**	**0.66**
Union Progreso y Democracia 1	0.80	0.80
Union Progreso y Democracia 2	0.80	0.60
Union Progreso y Democracia 3	0.40	0.60

Table 6: Aspect coverage of each summary

express themselves so in these case it is not as simple as looking for the mentions of the aspect. We had to read each sentence to find when the annotators used synonyms to talk about the aspects.

In Table 6, it may be seen the percentages of how much of every summary covers the aspects of a political party. For example, if a summary talks about the political party itself and the economy plan then it covers two aspects. As we are interested in five specific aspects then the summary will cover 40 percent. Here we may see that even if the average aspect coverage of an entity is low, it is because the annotators had decided to cover on a specific set of aspects. This is revealed by the aspect coverage of each annotator summary as usually they only differ by a small margin, meaning that there is an agreement on which aspects are more relevant to include on the summary.

On the other hand, we may see that usually extractive and abstractive summaries have similar results on the aspect coverage. This happens because the annotators try to express the same ideas for both summaries even if they can not use their own words. On the other hand, there was a case where the extractive summary covers more than the abstractive summary. This happens because when writing with his own words, the annotator has the possibility to focus on what is more relevant to him instead of using the sentence that seems better to him even if that sentence cover aspects than aren't important to the annotator.

5.3. Sentiment Orientation

Even though we have covered the elements that annotators where writing about, that is not complete enough as people express themselves with a sentiment in mind that is reflected on the way each annotator communicates and the words they use. Consequently, we had to analyze the sentiment orientation of each summary in order to get the com-

plete meaning of each summary. For these analysis we are considering that the sentences expressed on each summary are positive or negative as these are the most general sentiments that can be associated to a sentence.

Regarding the aspect coverage, we used the tags on the STOMPOL corpus that associates some words on a sentence to a polarity to analyze the extractive summaries. As we had only referenced if a sentence is positive or negative without taking into account the intensity of how it is expressed, we have not taken into consideration if a sentence has multiple occurrences of the same polarity on different words and only considered as it is positive or negative.

On the other hand, for the abstractive summaries we used Spanish Sentistrength (López et al., 2012) for the analysis. This is a dictionary of common words used on the Spanish language where each word is associated to a polarity.

The results of this experiment are presented on the Table 7. In particular, extractive summaries show disproportionate results compared to the actual polarity of the corpus. Conversely, abstractive summaries show better results. This happens because when doing extractive summaries, the annotator is limited by the sentences that they can use and gives more preference to covering the more relevant aspects instead of trying to express their opinion.

6. Conclusions and Future Work

In this paper, we presented a corpus of summaries of Spanish tweets related to political aspects. Additionally, we presented an analysis of this corpus to detail its contents. From the inter-annotator agreement we have demonstrated that the summaries are different from each other on the sentences they use to express their message. However, from the aspect coverage we can say that even though the summaries are different, they are about similar elements of the entity. Notably, the words used on the summary also change the message that tries to express as demonstrated on the sentiment orientation where the freedom of a annotator to use their own words give the abstractive summaries better results.

For future works, we will give more emphasis on the comparison between the summaries and the corpus as it can reveal more details on the distribution of the summaries as displayed on the sentiment orientation where the results of the abstractive summaries are more similar to the corpus than the extractive. Also, we might take into consideration the order in which the summaries are generated as writing first the extractive summaries could influence in the annotators to use the same words in the abstractive summaries.

Another essential point is to improve the tools used to measure the summaries. For example, we may find a sentiment dictionary that is more suited to the theme of the corpus that will obtain better results. In the same way, we can take into consideration the intensity of the polarity expressed on the sentences to get a more real grasp of what they are trying to express.

Finally, the corpus will be available in the following link: https://github.com/iapucp/spop-summ-lrec2018

Summary	Actual Polarity		Extractive		Abstractive	
	Positive	Negative	Positive	Negative	Positive	Negative
Ciudadanos	0.363	0.636	0.666	0.333	0.366	0.633
Izquierda Unida	0.347	0.652	0.300	0.700	0.253	0.746
Partido Popular	0.162	0.837	0.500	0.500	0.138	0.861
Partido Socialista Obrero Español	0.279	0.720	0.435	0.564	0.172	0.827
Podemos	0.207	0.792	0.500	0.500	0.228	0.771
Unión Progreso y Democracia	0.365	0.634	0.611	0.388	0.441	0.558

Table 7: Sentiment Orientation

7. Bibliographical References

Carenini, G., Ng, R., and Pauls, A. (2006). Multi-document summarization of evaluative text. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 305–312.

Cámara, E. M., García-Cumbreras, M. A., Villena-Román, J., and García-Morera, J. (2016). TASS 2015 - The evolution of the Spanish opinion mining systems. *Procesamiento del Lenguaje Natural*, 56:33–40.

Esteban, A. and Lloret, E. (2017). Travelsum: A spanish summarization application focused on the tourism sector. *Procesamiento del Lenguaje Natural*, 59:159–162.

Ganesan, K., Zhai, C., and 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*, pages 340–348. Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Labbé, C. and Portet, F. (2012). Towards an abstractive opinion summarisation of multiple reviews in the tourism domain. In *The First International Workshop on Sentiment Discovery from Affective Data (SDAD 2012)*, pages 87–94.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

López, R., Tejada, J., and Thelwall, M. (2012). Spanish sentistrength as a tool for opinion mining peruvian facebook and twitter. *Artificial Intelligence Driven Solutions to Business and Engineering Problems*, pages 82–85.

López, R. E., Avanço, L. V., Filho, P. P., Bokan, A. Y., Cardoso, M. S., Nóbrega, F. A., Cabezudo, M. A., Jackson, W., Souza, A. C., Seno, E. M., et al. (2015). A qualitative analysis of a corpus of opinion summaries based on aspects. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, pages 62–71. Association for Computational Linguistics.

Tadano, R., Shimada, K., and Endo, T. (2010). Multi-aspects review summarization based on identification of important opinions and their similarity. In *PACLIC*, pages 685–692.

Wu, H., Gu, Y., Sun, S., and Gu, X. (2016). Aspect-based opinion summarization with convolutional neural networks. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 3157–3163. IEEE.

Xu, X., Meng, T., and Cheng, X. (2011). Aspect-based extractive summarization of online reviews. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 968–975. ACM.