The CNN-Corpus in Spanish: a Large Corpus for Extractive Text Summarization in the Spanish Language

Rafael Dueire Lins
rdl.ufpe@gmail.com
UFPE/UFRPE
Recife, Brazil

Hilario Oliveira, Luciano Cabral,
Jamilson Batista, Bruno Tenorio,
Diego A. Salcedo
UFPE, Recife, Brazil

Rafael Ferreira, Rinaldo Lima
Gabriel de França Pereira e Silva
UFRPE, Recife, Brazil

Steven J. Simske
Colorado State University
Fort Collins, CO, USA

ABSTRACT

This paper details the development and features of the CNN-corpus in Spanish, possibly the largest test corpus for single document extractive text summarization in the Spanish language. Its current version encompasses 1,117 well-written texts in Spanish, each of them has an abstractive and an extractive summary. The development methodology adopted allows good-quality qualitative and quantitative assessments of summarization strategies for tools developed in the Spanish language.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Single-document Summarization, CNN Corpus, Spanish, Extractive Summarization, Multi-language Summarization

ACM Reference Format:
Rafael Dueire Lins, Hilario Oliveira, Luciano Cabral,, Jamilson Batista, Bruno Tenorio,, Diego A. Salcedo, Rafael Ferreira, Rinaldo Lima, Gabriel de França Pereira e Silva, and Steven J. Simske. 2019. The CNN-Corpus in Spanish: a Large Corpus for Extractive Text Summarization in the Spanish Language. In *ACM Symposium on Document Engineering 2019 (DocEng '19), September 23–26, 2019, Berlin, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3342558.3345423

1 INTRODUCTION

Automatic Text Summarization (ATS) is computer process of creating a shorter version (summary) from one or more documents, providing to the reader the most relevant information. The origin of this research area dates back to the works of Luhn in 1958 [5] and Edmundson in 1969 [2], but the fundamental challenges in the area remain unsolved. The Internet has raised interest in ATS, as readers cannot keep track of the large and quickly expanding volume of data in any area of knowledge.

The simplest form of ATS is *extractive* summarization, in which a number of sentences are selected and copied *verbatim* from the original text to form the summary. Such a summary does not necessarily offer coherent information, as the selected sentences may either not be self-contained or may have redundant or irrelevant information. Despite that, extractive summarization is the starting point for more sophisticated methods.

Evaluating the quality of a summary in objective ways is fundamental for increasing the validity of the results obtained in this area. The creation of a good-quality large test corpora is justified by the exhaustive manual effort required in the annotation task. Several people must read the entire content at least once and manually create a summary of reference. Subjectivity, the informational user needs, the wideness and richness of the vocabulary used, etc. are some of the complicating factors in such a task.

An important step ahead in meeting such an urgent demand for a good-quality test corpus for extractive summarization is the CNN-corpus in English [4], a team work that followed a rigorous development methodology, originally based on news articles from the CNN website[1]. The articles selected meet the high linguistic standards needed for the documents in a test corpus and have a good quality abstractive summary written by their original authors, the *highlights*. Such an abstractive summary served as the basis for generating an extractive summary of reference for each text, the *gold standard*, which allows the qualitative and quantitative assessment of automatic summarization techniques.

The absence of tools and test corpora for ATS in languages other than English is even more critical. In Spanish, reference [1] says that the only existing summarization test corpus available is the Gigaword corpus, which is not free and whose features such as size, development methodology, subject area, etc. were not found in the Internet. Apart from that, one can only find the Multilingual summary evaluation data from the Joint Research Centre (JRC) [6], which is a set of only twenty text documents in Spanish and six other languages (Arabic, Czech, English, French, German, and Russian). Thus, to the best of the knowledge of the authors of this paper, the corpus presented here is the largest test corpus for summarization for the Spanish language.

DocEng '19, September 23–26, 2019, Berlin, Germany

ACM ISBN 978-1-4503-6887-2/19/09.
https://doi.org/10.1145/3342558.3345423

[1] www.cnn.com

2 BUILDING THE CNN-CORPUS IN SPANISH

The development of CNN-corpus in Spanish has as the starting point the set of news articles harvested from the CNN Mexico website[2], such as the one shown in Figure 1. The CNN news articles in Spanish, similarly to the one in English, are high-quality, grammatically correct texts, report on subjects of general interest, and use standard vocabulary. Besides those fundamental features, there was one particularity of many of such texts that make them especially valuable for automatic text summarization: *Lo mas importante*, an abstractive summary written by the original authors of the texts, that corresponds to the *highlights* of the CNN texts in English. The data collected from the CNN website in Spanish were the original text, the story highlights and several other metadata such as the name of the authors, title, subject classification, date, and others. The development of the CNN-corpus in Spanish followed the same

Figure 1: News article collected from the CNN Mexico portal

methodology adopted in the development of the second version of the CNN-corpus in English [4]. The Web crawler gathered 7,466 articles from May 2014 to June 2014, applying the same filters of the English corpus to the Spanish articles, except for the case that articles with only one sentence in the story highlight were discarded. Similarly, news articles whose text were not "self contained" and depended on any sort of external link, graphic element, table, etc. for its understanding were also discarded.

A translation step from Spanish into English was introduced in the process as the tools used in the sentence similarity algorithm work only in English. Two automatic translation tools were tested: Google Translator and the Microsoft Translator API. The preliminary assessment performed showed that the performance of the latter was better than the former tool for the case of news articles. Thus, in the translation step, each text documents and story highlights in Spanish are translated into English using the Microsoft Translator API (http://www.microsoft.com/transla- tor/api/). Then, the preprocessing and classification steps are applied as in the process described in [4] for the building of the corpus in English, including the sentence similarity algorithm and the values adopted for the thresholds. There is a biunivocal correspondence between each sentence in Spanish and in English, and their indices are maintained unchanged. The translation process is used only as a way of finding which sentences in the text have the highest degree of similarity with the sentences in the highlights. The tool developed selected 1,700 documents in Spanish as "easier to match" the highlights onto the original text.

In the mapping and divergence resolution tasks, a team of four non-native Spanish annotators and one additional "referee" were used to map and solve the divergent cases using the same quality policies and values as in the process in English [4]. Two randomly chosen pairs of experts were used per week took 17 weeks to complete both tasks. At the end of the 18th week, there were 1,135 (66.8%) documents mapped without divergence and agreed on by two experts and 565 (33.2%) documents marked as error and discarded from the process. For those mapped documents, a manual inspection was executed which eliminated 18 documents. The resulting corpus has 1,117 Spanish documents with highlights, gold standards, and metadata. The statistics for the Spanish corpus are presented in Table 1. The 1,117 documents are not uniformly distributed in the 4 categories assined by CNN: sports (*Deportes*), entertaiment (*Entretenimiento*), Mexico (*national*), and world (*Mundo*). The category *Mundo* encompasses 60.0% of the articles while *Entretenimiento* has only 6.5%. The highlights of the CNN-corpus in Spanish has a total of 3,041 sentences with an average of 2.7 sentences per highlight.

The number of sentences per text ranges between 10 and 60 sentences with a mean of 17.3 and the mean number of words per sentence is 27.6. The mean compression rate between the original text and the story highlight is 15.6% . The mean number of sentences per gold-standard is 2.5 and the mean number of words per sentence in the gold-standard is 41.0. In Spanish, the sentences in the gold standard are approximately 2.5 times longer than the sentences in the highlights.

3 SOME DEVELOPMENT STATISTICS

This section evaluates some of the steps performed during the construction of the CNN-corpus in Spanish, allowing one to better understand the challenges faced on its development. Section 3.1 shows the experimental results of the assessment of the sentence similarity measure (SSM) proposed by Ferreira *et al.* (2014), used to select documents and point out the most similar sentences for each highlight.

3.1 Evaluating the SSM

The SSM algorithm [3] was applied to the texts translated from Spanish into English. As each sentence in the highlights ("*Lo mas importante*") was mapped onto one or more sentences of the document, and this process was reviewed by humans, it is possible to assess the accuracy of the similarity algorithm in estimating the most similar sentences correctly for each highlight. The experimental results are presented in Table 3. The first column indicates

[2]http://mexico.cnn.com

Table 1: Overview statistics of the Spanish corpus.

Categories	Articles	Avg.Sentence/ High. Summary	Avg.Words / Story High.	Avg.Sentences/ Text	Avg.Words/ Sentences	Avg.Sentences/ Gold. Summary	Avg.Words/ Gold. Sentence
Deportes	235	2.6	14.7	15.3	27.8	2.3	45.7
Entretenimiento	72	2.5	14.9	18.9	24.2	2.3	38.8
Nacional	140	2.8	14.9	17.0	32.1	2.6	46.1
Mundo	670	2.8	15.0	18.0	27.0	2.5	39.7
Total/Average	1,117	2.7	14.9	17.3	27.6	2.5	41.0

how many sentences from the original text were mapped onto a single sentence of the highlights. It is possible to note that the vast majority of sentences in the highlights (94.72%) were mapped onto one sentence of the original text. The maximum number of mapped sentences for a single sentence from the highlights was four (observed only six times). Accuracy measurements are also provided in Table 3. These are based on the agreement between the human annotators and the algorithm for spotting the key sentences in the summary. As more sentences are mapped onto a single sentence of the highlights, the accuracy of the similarity algorithm decreased. The results of this experiment showed that in 95.89% of the sentences in the highlights, i.e., 10,235 of the 10,674 highlights mapped, the human annotator needed only to read the five more similar sentences indicated by the similarity algorithm to find the set of most suitable phrases in the original text to be selected. These results demonstrate that the inclusion of the selection step impacted positively on the construction process of the CNN-corpus, reducing the manual effort spent for the creation of extractive summaries.

Table 2: Level of agreement (%) for the Spanish corpus.

Expert	Mapped Highlights	Avg. Annotator Agreement	Divergent Highlights	Avg. Judge Agreement
G	248	79.0	52	85.5
H	2,803	79.5	574	96.2
I	2,432	76.4	574	77.9
J	619	91.6	52	87.9

Table 3: Degree of easiness in the Spanish corpus.

j	Number of Mappings	Easiest j-Mappings	%	Easy j-Mappings	%
1	2,927	2,635	86.6	2,899	95.3
2	102	49	1.6	60	2.0
3	12	5	0.2	5	0.2

The experimental results for the Spanish corpus present the same behaviour as the English corpus [4], except there were no mapping for $j = 4$. It means that the translation of the text to English, in the translation step, did not introduce enough errors to affect the mapping process. In fact, the experiment shows that the multilingual process had practically the same results as the normal process. Therefore, the translation hypothesis, which states that the current translation technology is advanced enough so that it does not affect considerably the mapping process, is empirically validated. As consequence, the multilingual process can be applied to documents in any language that can be automatically translated into English. The impact of using only easy documents in the process is further detailed in respect to the time to map and to the quality of the final corpus.

3.2 Time Assessment

In Tables 4 and 5 one finds some account of the time taken for an expert to validate and fix the mapping of a sentence in story highlights onto the original text in the formation of the gold standard, using the platform developed. The overall and mapping times increased slightly with the number of sentences in the text. This may be so because there are few mappings that are not easy and the expert has to scan other sentences of the text to appropriately perform the task. The divergent time appears to be constant, however, which sounds reasonable because the annotators have to read only those sentences selected by the annotators to take a decision. The efficient and user-friendly interfaces developed in this project minimized the human effort and decreased the time elapsed in the tasks. The overall and mapping times oscillate heavily with the number of sentences in the text, not enabling the detection of a pattern. The mean divergent time was almost constant, as well.

3.3 Annotators' Agreement Level

The degree of agreement of the mappings between the experts may also be used as an indicator of the quality. If at least two annotators agreed on a mapping, then it was assumed that there is a higher probability that the summary was adequately mapped.

There were six experts involved in this task. The total number of mapped highlights is twice the number reported in the earlier Section, *Mapping Story Highlights onto the Original Text*, because each story highlight was mapped by two annotators according to the quality policy adopted. The overall annotator agreement rate is 89.2%, which means that 10.8% were discarded. The level of agreement between the six annotators is similar, ranging from 85.3% to 91.6%. After the judge decided on the divergent cases, the agreement level was also similar, ranging from 90.7% to 95.5%. For instance, the expert B annotated 3,772 story highlights, of which 89.9% of them were agreed by another expert (e.g. A, C, D, E or F) and 95.5% of them were agreed by another expert in the mapping task or by a judge in the divergent task. Specifically for the divergent cases, which correspond to 10.7% of the total number of mapped highlights (i.e. 10,674), the referees agreed with one of the annotators in 62.6%

Table 4: Spanish time assessment, in seconds, according to the number of sentences (m) in the text.

Number of Sentences (m)	Mapped Highlights	Overall Time (s)		Mapping Time (s)		Divergent Highlights	Divergent Time (s)	
		Average	Std. Dev.	Average	Std. Dev.		Average	Std. Dev.
[10..20]	2,068	44.2	47.8	20.5	22.7	361	4.8	7.8
(20..30]	746	47.6	48.9	21.2	21.7	192	4.6	8.1
(30..40]	196	47.1	46.7	20.7	19.6	56	4.9	7.2
(40..50]	27	67.9	46.8	28.8	21.4	9	4.4	7.4
(50..60]	8	26.8	18.4	11.1	8.6	2	3.0	0.0

Table 5: Spanish time assessment, in seconds, according to the degree of easiness.

Degree of Easiness	Highlights	Overall Time (s)		Mapping Time (s)		Highlights	Divergent Time (s)	
		Average	Std. Dev.	Average	Std. Dev.		Average	Std. Dev.
(0.5..0.6]	593	52.0	54.7	23.7	25.4	193	4.9	8.5
(0.6..0.7]	1,212	47.8	50.3	21.7	23.1	257	4.8	8.1
(0.7..0.8]	774	43.4	45.0	20.0	21.2	123	4.3	5.8
(0.8..0.9]	367	34.0	33.8	15.6	16.0	31	5.1	7.8
(0.9..1.0]	99	35.1	34.1	16.0	15.9	16	5.8	8.0

of the cases. For the rest of the divergent highlights (37.4%), their documents were discarded from the final corpus. This means that the resulting corpus is formed by articles that are many levels easier than most articles in the original CNN news articles, besides being self-contained (there are no external web links) and text-only (there are no internal references to figures, tables, videos, etc.). For the Spanish corpus, the overall annotator agreement is 79.5%. Observe that the level of the annotator agreement between the four of them are disparate ranging from 76.4% to 91.6%. After the judge decided on the divergent cases, it was clear that the experts had different levels of proficiency of the Spanish language. For instance, the expert H had 79.5% of agreement with the other annotators, but after the divergent task, it had 96.2% of mapping validated in the final corpus, while expert I had 77.9%.

Specifically for the divergent cases, which correspond to 20.5% of the total number of mapped highlights (i.e. 3,041), the referees agreed with one of the annotators in 81.9% of the cases and, for the rest of the divergent highlights (18.1%), their documents were marked as error and discarded from the final corpus. Finally, the pairs G/J and H/I have the same values, because the pairs of experts were the same during the 17 weeks not varying between possible combinations due to availability issues.

4 CONCLUSIONS

This paper presents the CNN-corpus in Spanish, possibly the largest corpus for assessing algorithms for the extractive single-document automatic summarization in the Spanish language, the result of a large team effort

The starting point were 7,466 CNN news articles gathered from May 2014 to June 2014 at http://mexico.cnn.com, addressing subjects of general interest, following very high standards of vocabulary and grammar. A semi-automatic human supervised process was used to generate a *gold standard* extractive summary of each text. A user-friendly platform was developed with interfaces that minimized the human effort, decreased the time elapsed in the generation of the gold-standard, and lowered the chances of human errors. The rigorous development methodology used may be considered as an indicator of its quality.

Unfortunately, the news articles in Spanish from CNN websites no longer encompass their highlights, *"Lo mas importante"*, not making possible to use the methodology presented here to grow further the CNN-corpus in Spanish.

The CNN-corpus in Spanish, likewise the CNN-corpus in English, with all the original texts, their highlights, gold-standard summaries, and all XML-annotated versions is freely available for research purposes, under request to the authors.

5 ACKNOWLEDGMENTS

The authors are partly supported by a R&D project between HP Brazil R&D and UFPE originated from tax exemption (IPI - Law n^{o} 8.248, of 1991 and later updates). The authors are also grateful to PROPESQ-IFPE and CNPq by supporting this research.

REFERENCES

[1] Franck Dernoncourt, Mohammad Ghassemi, and Walter Chang. 2018. A Repository of Corpora for Summarization. In *Proceedings of the 11th Language Resources and Evaluation Conference*. European Language Resource Association, Miyazaki, Japan. https://www.aclweb.org/anthology/L18-1509

[2] H. P. Edmundson. 1969. New methods in automatic extracting. *Journal of the ACM (JACM)* 16, 2 (1969), 264–285.

[3] Rafael Ferreira, Luciano de Souza Cabral, Frederico Freitas, Rafael Dueire Lins, Gabriel de França Silva, Steven J. Simske, and Luciano Favaro. 2014. A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications* 41 (2014), 5780–5787.

[4] Rafael Dueire Lins, Hilario Oliveira, Luciano Cabral, Jamilson Batista, Bruno Tenorio, Rafael Ferreira, Rinaldo Lima, Gabriel de Franca Pereira e Silva, and Steven J. Simske. 2019. The CNN-Corpus: A Large Textual Corpus forSingle-Document Extractive Summarization. In *Proceedings of the 2019 ACM Symposium on Document Engineering (DocEng '19)*. ACM, New York, NY, USA, 201–210. https://doi.org/10.1145/1122445.1122456

[5] H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2, 2 (1958), 159–165.

[6] Steinberger J. Kabadjov M. Steinberger R. Turchi, M. 2010. Using parallel corpora for multilingual (multi-document) summarisation evaluation.. In *Multilingual and multimodal information access evaluation.*, Vol. LNCS 6360. Springer Verlag, 52–63.