# DACSA: A large-scale Dataset for Automatic summarization of Catalan and Spanish newspaper Articles

**Encarna Segarra** **Vicent Ahuir**[§] **Lluís-F. Hurtado** **José Ángel González**
VRAIN: Valencian Research Institute for Artificial Intelligence
Universitat Politècnica de València, Spain
{esegarra, lhurtado, jogomba2}@dsic.upv.es
[§]viahes@eui.upv.es

## Abstract

The application of supervised methods to automatic summarization requires the availability of adequate corpora consisting of a set of document-summary pairs. As in most Natural Language Processing tasks, the great majority of available datasets for summarization are in English, making it difficult to develop automatic summarization models for other languages. Although Spanish is gradually forming part of some recent summarization corpora, it is not the same for minority languages such as Catalan. In this work, we describe the construction of a corpus of Catalan and Spanish newspapers, the Dataset for Automatic summarization of Catalan and Spanish newspaper Articles (DACSA) corpus. It is a high-quality large-scale corpus that can be used to train summarization models for Catalan and Spanish. We have carried out an analysis of the corpus, both in terms of the style of the summaries and the difficulty of the summarization task. In particular, we have used a set of well-known metrics in the summarization field in order to characterize the corpus. Additionally, we have evaluated the performance of some extractive and abstractive summarization systems on the DACSA corpus for benchmarking purposes.

## 1 Introduction

Automatic summarization is one of the central problems in Natural Language Processing (NLP). The development of automatic summarization systems is an important issue due to the great amount of information in different formats that is accessible on the web or in other repositories. It is necessary to develop techniques that help us to tackle that huge amount of information. For this reason, there is an increasing interest in the NLP community to develop techniques that allow the users to find, read, understand, or process the documents. In this context, automatic summarization can be an important aid because it provides a condensed version of documents that reduce the time to explore or analyze them.

Access to large-scale high-quality data is an essential prerequisite for making substantial progress in summarization. The application of supervised methods to automatic summarization, as those based on Neural Networks, requires the availability of adequate corpora consisting of document-summary pairs. The construction of large-scale and high-quality corpora for learning neural summarization models is not an easy task. It is necessary a great human effort to generate thousands of manual summaries, or to design new approaches to obtain these summaries in a semiautomatic way. The first important resource for learning corpus-based summarization models was the CNN/DailyMail summarization corpus (Hermann et al., 2015), originally constructed for the task of passage-based question answering and adapted to the document summarization task. It consists of news stories from CNN and DailyMail and contains 312,077 article-summary pairs. Afterwards, another English corpus was provided to the research community for summarization purposes, the NewsRoom corpus (Grusky et al., 2018). It consists of 1.3 million article-summary pairs that have been written by the authors and the editors of 38 different major news publications. The corpus was created through a web-scale crawling of over 100 million pages from a set of online publishers by gathering the news and using the summaries provided in the HTML metadata. The summaries contained in this corpus combine both extractive and abstractive strategies to describe the content of the articles. Also in 2018, the XSUM corpus (Narayan et al., 2018a) was presented, it is a large scale dataset obtained by harvesting online articles from the British Broadcasting Corporation (BBC) with one-sentence news summary.

As in most NLP tasks, the great majority of available datasets for summarization are in English. The



*Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5931 - 5943

lack of this kind of resources for other languages is an encumbrance to modeling that constraints the impact of language technologies on minority language communities. The creation of a large-scale Indonesian summarization dataset of 215,827 document-summary pairs, has just been published (Koto et al., 2020). Recently, some datasets that aim to fill the gap among English and other languages for the automatic summarization task have been proposed: MLSUM (Scialom et al., 2020), MassiveSumm (Varab and Schluter, 2021), and XL-Summ (Hasan et al., 2021). Although Spanish is the world's second-most spoken native language and is the official language in 21 countries, it has only recently been considered in general domain summarization datasets, as the aforementioned, and in specific domains as in (González et al., 2019). The situation is worse for Catalan, although it is not an endangered language, it is spoken by 10 million people in Spain and other three European countries, it is minority worldwide and is underrepresented or even not considered in summarization corpora.

In this work, we describe the construction of a corpus of Catalan and Spanish newspapers, the Dataset for Automatic summarization of Catalan and Spanish newspaper Articles (DACSA) corpus. With the aim of building a quality large-scale corpus that could be used to learn automatic summarization neural models for Catalan and Spanish, we used a strategy inspired by the construction of the NewsRoom corpus (Grusky et al., 2018). We conducted a crawling process on 30 different newspaper websites to extract articles and summaries in a straightforward way. The crawling included from Spanish mass media to regional newspapers. In order to obtain the summaries, we took advantage of the highlights and summaries, provided by authors or editors of the articles.

To ensure the quality of the DACSA corpus, we perform two subsequent filtering processes on the downloaded articles. The first filter was used to ensure, at least, a minimum length in both the article and the abstract. All the articles or summaries that were considered too short were discarded. Obviously, an article or summary too short implies discarding the article-summary pair. The second filter was used to ensure that the summaries were not almost verbatim copies of the first sentences of the articles. To do this, the article-summary pairs in which the overlapping between the summary and the article prefix of the summary length was high were also discarded. This way, we try to avoid a positional bias in the summaries by discarding those samples in which the summary is reduced to select the first sentences of the article.

Once both filters were applied, we found that some newspaper sources had very few samples, less than 1000 in some cases. To balance the corpus partitions, we decided to remove the sources with few samples from the training, validation, and tests sets. Nevertheless, we joined together the samples from those sources to create a special test set, a test set with sources not present in the training process. Therefore, the corpus consists of four partitions per language: training, validation, and test sets along with an extra test set. Considering all the partitions, the DACSA corpus consists of a set of 725,184 article-summary pairs extracted from 9 different Catalan newspaper websites and 2,120,649 article-summary pairs extracted from 21 different Spanish newspaper websites. The DACSA corpus contains articles and summaries about politics, economics, sports, culture and other topics usually addressed in journalistic domains. To our knowledge, the DACSA corpus is the largest summarization dataset for both languages.

We have used four well known metrics in the summarization field in order to characterize the corpus. These metrics are: extractive fragment coverage and density (Grusky et al., 2018), abstractivity-p (Bommasani and Cardie, 2020), and novel n-grams (Kryściński et al., 2018). Additionally, for benchmarking purposes, we have evaluated the performance of 6 automatic summarization systems on the DACSA corpus. Concretely, we have used two unsupervised systems (lead-2 and textRank), an extractive summarization system, SHANN (González et al., 2019), two abstractive summarization systems, mBART (Liu et al., 2020) and mT5 (Xue et al., 2020), and one oracle to compute upper bounds of the performance in the DACSA corpus.

The DACSA corpus can be requested for research purposes at `https://xarrador.dsic.upv.es/resources/dacsa`.

## 2 Related Work

The automatic text summarization problem has been addressed in the literature using abstractive, extractive, or mixed approaches. On the one hand, extractive approaches compose summaries by selecting sentences or words directly from the documents (Cheng and Lapata, 2016; Nallapati et al.,

2017; Liu and Lapata, 2019; Narayan et al., 2018b; Zhang et al., 2018; Dong et al., 2018; Yao et al., 2018; Chen and Bansal, 2018). Most of these approaches address a sequential binary sentence classification problem in order to select the most salient sentences of the documents, following different criteria such as negative log likelihood on preselected sentences (Cheng and Lapata, 2016; Nallapati et al., 2017; Liu and Lapata, 2019) or ROUGE (Lin, 2004) rewards in reinforcement learning environments (Narayan et al., 2018b; Zhang et al., 2018; Dong et al., 2018; Yao et al., 2018). Other extractive architectures are based on siamese hierarchical attention networks built in terms of Long Short Term Memories and Transformer encoders (González et al., 2019, 2020). These models have been succesfully applied in summarization tasks of Spanish newspapers and talk shows (González et al., 2019). On the other hand, the abstractive approaches build the summaries by paraphrasing the sentences of the documents (See et al., 2017; Paulus et al., 2018; Ive et al., 2019). The vast majority of existing neural abstractive summarization models are based on encoder-decoder architectures (Sutskever et al., 2014). Finally, there are also mixed approaches that combine extractive and abstractive techniques, performed in a decoupled way or simultaneously inside the models (Mendes et al., 2019).

Due to the recent success of self-supervised learning, the focus of text summarization research has exhibited a gradual shift from extractive techniques to abstractive techniques (Lewis et al., 2020; Zhang et al., 2020a; Raffel et al., 2020). These kind of objectives allows to pretrain deep architectures (mainly Transformers) to learn vast amounts of general linguistic knowledge from large corpora, that can be transferred to downstream tasks by means of finetuning. The most successful model of this type is BERT (Devlin et al., 2019), that is pretrained with Masked Language Model and Next Sentence Prediction objectives on raw texts from English Wikipedia and BooksCorpus. Based on BERT, some architectural improvements have been proposed like RoBERTa (Liu et al., 2019) or ALBERT (Lan et al., 2020).

In some recent works, BERT and RoBERTa have been finetuned for extractive summarization (Liu and Lapata, 2019; Zhong et al., 2020), but, although it boosted the performance of the previous extractive approaches, the pretraining+finetuning philosophy has shown to be most effective for abstractive systems. Nowadays, the best performing abstractive models are BART (Lewis et al., 2020), T5 (Raffel et al., 2020) and PEGASUS (Zhang et al., 2020a), being all of them Transformers (Vaswani et al., 2017) pretrained self-supervisedly as denoising sequence to sequence autoencoders. Some multilingual variants of these models have been recently proposed, mBART (Liu et al., 2020) and mT5 (Xue et al., 2020). Both of them were pretrained following a multilingual denoising procedure on large-scale multilingual corpora. On the one hand, the mBART model was pretrained by using a corpus of 25 languages, extracted from the Common Crawl (Wenzek et al., 2020) (CC25). On the other hand, a multilingual variant of the Colossal Clean Crawled corpus (Raffel et al., 2020) was used to pretrain mT5.

Self-supervised pre-training requires obtaining large amounts of raw data in order to learn good initializations of deep models from denoising objectives. Also, the fine-tuning of these architectures in downstream tasks like text summarization implies the availability of adequate corpora consisting of document-summary pairs. As we mention above, the great majority of datasets for summarization are in English: CNN/DailyMail, NewsRoom, XSUM (Narayan et al., 2018a), and so forth. Although some multilingual datasets have been recently created, as MLSUM, MassiveSumm, and XL-Summ, they do not provide a large enough portion of Spanish data and only MassiveSumm provides a few samples for Catalan. It is in this context where we propose to build the DACSA corpus.

The most used metrics in the literature to quantify the performance of the models in the summarization task are ROUGE (Lin, 2004) and BertScore (Zhang et al., 2020b). On the one hand, ROUGE measures the performance by counting exact matches. On the other hand, BertScore is a more semantic measure which is based on contextual embeddings provided by a BERT language model. These metrics are convenient to evaluate the performance, but they do not explicitly measure the abstractivity. Measuring the abstractivity of the summaries generated by the models is generally not trivial. In this work, we used a set of metrics as abstractivity indicators to asses the level of abstractivity: extractive fragment coverage and density (Grusky et al., 2018), abstractivity$_p$ (Bommasani and Cardie, 2020), and novel n-grams (Kryściński

et al., 2018). Additionally, we also used ROUGE and BertScore to compare the different summarization models.

## 3 Building the DACSA corpus

The DACSA corpus was collected using a distributed web crawler that captured over 6 million news articles, close to 2 million of articles published in Catalan, and more than 4 million written in Spanish. The articles were captured from 30 newspapers sources, 9 sources for Catalan and 21 sources for Spanish. The range of years of publication was between 2010 to 2020.

We divided the crawling process into two services. The first service was designed to retrieve the list of articles on the website of the newspapers source; we refer to this service as the *URLs extractor service*. The second one aims to extract the content (article content and summary) of the article; we refer to this service as the *content extractor service*. The whole crawler was developed with Python 3 and JavaScript (Node.js runtime) programming languages.

For the configurations (one per source) of the *content extractor service*, we used CSS selectors and the library *cheerio* (`https://cheerio.js.org/`). In order to capture the article and summary text, we designed the selectors that captured the visible information that a person would read, avoiding metadata. Using visual information instead of metadata is important because we detected that likely the metadata was automatically created by some naive process that could lose information, such as just extracting the first tokens of the article; meanwhile, the visual text is likely complete, readable and coherent.

We searched websites of electronic newspapers published in Spain, in Catalan or Spanish languages. To find the addresses of each article, we decided to use the list of news that electronic newspapers usually have on their website. The benefit of using the list of articles provided by these websites, contrary to the common crawling approach of following every link, was that we aimed the articles themselves, and there was no need to identify whether the web page is a news article or other kind of content. Thus, from the list of news in that newspapers source, we created two configurations, one for the *URLs extractor service* and another for the *content extractor service*.

We intended DACSA to be a large-scale, high-quality corpus for Catalan and Spanish. Thus, after the massive capture of samples, we defined two requirements that the articles and summaries must satisfy. We first established a threshold in the minimum number of words of the article and the summary, and second, a threshold in the maximum similarity between the summary and the first sentences of the article.

On the one hand, we discarded those samples with a short text in the article or the summary. Specifically, every sample inside the corpus contains at least 100 words in the article and 10 words in the summary. With this restriction, we ensure that the samples have enough content to generate a summary with a reasonable length.

On other hand, we rejected from the corpus those samples in which the summary is generated by simply extracting the first sentences of the article. Specifically, we restricted the overlapping between the summary and the starting sentences of the article by using a similarity metric based on the Levenshtein distance to quantify the degree of overlapping. The Equation (1) presents the definition of this metric.

$$f(A,S) = 1 - \frac{Levenshtein(A_{[1,|S|]}, S)}{|S|} \quad (1)$$

where $A$ is the sequence of words of the article text, $S$ is the sequence of words of the summary, $|S|$ is number of words of the summary, and $Levenshtein$ is the operation which returns the well-known Levenshtein distance between two texts. In this corpus, we established a maximum threshold of 0.9 of $f(A,S)$ between the article and the summary.

## 4 Dataset

After the above processes, the DACSA corpus was built. This corpus provides pairs of news article and its summary from different newspapers for both, the Catalan and the Spanish languages. Regarding the Catalan set, there are 725,184 sample pairs from 9 newspapers, regarding the Spanish set, the corpus provides 2,120,649 sample pairs from 21 newspapers.

The amount of samples by newspapers source is far from being homogeneous. If these distributions would be preserved over the partitions (training, validation, and test set), the models will focus their learning in the predominant newspapers. To avoid

| Source | #Docs | Tokens | Article | | | Summary | | |
|---|---|---|---|---|---|---|---|---|
| | | | Vocabulary size | Sents per doc | Words per sent | Vocabulary size | Sents per doc | Words per sent |
| Training | 636,596 | 316,817,625 | 1,206,292 | 17.39 | 28.62 | 206,616 | 1.17 | 20.36 |
| Validation | 35,376 | 17,831,029 | 258,999 | 16.17 | 31.17 | 51,940 | 1.15 | 20.93 |
| TESTI | 35,376 | 17,704,387 | 262,148 | 16.13 | 31.03 | 51,958 | 1.15 | 20.89 |
| TESTNI | 17,836 | 15,882,219 | 247,154 | 35.38 | 25.17 | 45,997 | 1.56 | 25.93 |
| Set | 725,184 | 368,235,260 | 1,326,343 | 17.71 | 28.67 | 223,978 | 1.17 | 20.59 |

Table 1: Statistics of the partitions for Catalan.

| Source | #Docs | Tokens | Article | | | Summary | | |
|---|---|---|---|---|---|---|---|---|
| | | | Vocabulary size | Sents per doc | Words per sent | Vocabulary size | Sents per doc | Words per sent |
| Training | 1,802,919 | 1,172,626,265 | 2,920,894 | 23.94 | 27.17 | 454,179 | 1.24 | 21.99 |
| Validation | 104,052 | 67,669,381 | 550,213 | 23.01 | 28.27 | 109,460 | 1.21 | 23.36 |
| TESTI | 104,052 | 67,363,994 | 550,910 | 22.93 | 28.23 | 109,706 | 1.21 | 23.34 |
| TESTNI | 109,626 | 59,603,306 | 447,679 | 16.25 | 33.46 | 116,201 | 1.35 | 36.84 |
| Set | 2,120,649 | 1,367,262,946 | 3,189,783 | 23.44 | 27.50 | 516,307 | 1.24 | 22.95 |

Table 2: Statistics of the partitions for Spanish.

this bias and achieve more general models, we propose that the test and validation sets be created in a way that all newspapers have roughly the same number of samples. To achieve this balance, we discarded some sources in order to guarantee that all sources represent at least 5% of samples in each one of these two sets. Additionally, we discarded those sources that have lower compression ratio than 10% in their summaries, since we considered these summaries too long compared to their corresponding articles.

The three sets for Catalan (training, validation and test set) are composed by 6 of the 9 newspapers, the training set contains 636,596 samples, and the validation and test sets have 35,376 samples each one. For Spanish, the three sets are composed by 13 of the 21 newspapers, the training set contains 1,802,919 samples, and the validation and test sets have 104,052 samples each one.

All the sources excluded were used as a separate test set. This partition allows evaluating the generalization capabilities of the summarization models against unseen newspaper sources. In this work, we refer to the test set with newspapers included in the training set as TESTI and to the test set that contains newspapers not included in the training set as TESTNI. The statistics of all the sets are shown in Tables 1 and 2.

In the Appendix A, Tables 7 and 8 show the distribution and the average lengths in terms of sentences and words of the articles and their summaries for Catalan and Spanish sets, detailed by the different newspaper sources.

## 5 Analysis of Dataset

In this section, an analysis of the level of abstractivity of the summaries of the corpus is done. First, the definition of the different measures used in this work is given, and second, we provide the application of these measures to the DACSA corpus.

### 5.1 Definition of Abstractivity Metrics

We used a set of metrics as abstractivity indicators to asses the level of abstractivity, they capture the degree of text overlapping between the summary and article. In particular, the following metrics have been selected: extractive fragment coverage and density, abstractivity$_p$, and novel n-grams.

**Extractive Fragment Coverage** (Grusky et al., 2018): the coverage measure quantifies the extent to which a summary is derivative of a text, that is, it measures the percentage of words in the summary that are part of an extractive fragment of the article.

**Extractive Fragment Density** (Grusky et al., 2018): contrary to the coverage, the density measure takes into account the length of the extractive fragments. A summary might contain many individual words from the article and therefore have a high coverage, however it might have a low density if the extractive fragments are short.

**Abstractivity$_p$** (Bommasani and Cardie, 2020): the *abstractivity$_p$* metric measures abstractivity as the absence of overlapping between the summary and the original text. Higher values indicate less overlapping and higher abstractivity. The $p$ parameter weights the length of each extractive fragment,

the higher value of $p$, the more the length of the extractive fragment is penalized.

**Novel n-grams**: (Kryściński et al., 2018) the *novel n-grams* metric quantifies the n-grams introduced in the summary that did not appear in the original text. The value of the metric is a percentage over the total of n-grams contained in the summary.

Additionally, we also used the **Compression Ratio**, that is, the ratio between the length of article and summary. Summarizing with higher compression is challenging as it requires capturing more precisely the critical aspects of the article text.

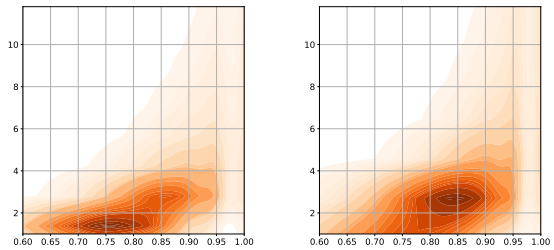
### 5.2 Dataset Abstractivity

This section presents the results of the abstractivity metrics described in Section 5.1 for the DACSA corpus. The results are shown separately for both languages; Table 3 shows the average values of the partitions for Catalan and Table 4 for Spanish. Tables 9 and 10 in the Appendix B also show these results for each newspaper source.

As Tables 3 and 4 show, the *training* and *validation* partitions have a similar type of summaries regarding their degree of abstractivity. The summaries in the test partitions, except the TESTNI set for Spanish, also show similar degree of abstractivity as the previous partitions.

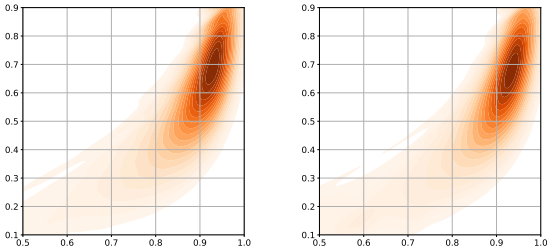
In order to better characterize the corpus, we also present in Figure 1 the distributions of the samples by combining the values of *extractive fragment coverage* and *extractive fragment density* of their summaries, and in Figure 2 the distribution of the samples by combining the values of *abstractivity$_p$ (p=2)* and *novel 2-grams*. These plots help to identify visually the degree of abstractivity of the summaries in the Catalan and Spanish sets. On the one hand, the metrics used in the first plots correlate negatively with the abstractitivy; thus, higher abstractivity is shown in the partition when the distribution is centered around the bottom left corner of the plot (where the values are lower on both metrics). On the other hand, the second plots correlate positively with the abstractivity; thus, the distributions are centered near the right top corner if the summaries are highly abstractive. Finally, we should point that due to the outliers, the distributions were hard to visualize. Hence, we exclude the 10% with the lowest values and the 10% with the highest values.

Figure 1 shows that the Catalan set mainly contains summaries with short extractive fragments since the distribution centers in 75% of coverage and a density lower than 2. Also, we observe that the distribution tends to go up and right; thus, the samples of the set diversify to less abstractive summaries. In the case of Spanish, we observe that the extractive fragments are longer than in the first language due to the higher density, and also, the distribution centers in the 85% of coverage, which indicates that the summaries in the Spanish set reuse more words from the article than in the Catalan set. However, the distribution tends to go down and left, which indicates a big presence of abstractive summaries in this set.



(a) Catalan (b) Spanish

Figure 1: Distribution of the samples for the Catalan and Spanish sets. `x-axis`: Extractive Fragment Coverage, `y-axis`: Extractive Fragment Density.



(a) Catalan (b) Spanish

Figure 2: Distribution of the samples for the Catalan and Spanish sets. `x-axis`: Abstractivity$_p$ (p=2), `y-axis`: Novel 2-grams.

Figure 2 helps to show the diversity of the samples by combining abstractivity$_p$ ($p = 2$) and novel 2-grams, which brings us more information. Although in Figure 1 the distributions were different from language to language, in this figure, we observe that the two sets are similar regarding these two metrics; note that the darker zones follow the same pattern around the same range of values.

Based on Tables 3 and 4 and Figures 1 and 2, it can be concluded that the DACSA corpus provides samples that do not contain a predominance

| Source | Compression | Coverage | Density | Abstractivity$_p$ (p=2) | Novel 2-grams | Novel 3-grams | Novel 4-grams |
|---|---|---|---|---|---|---|---|
| Training | 23.12 | 80.87 | 3.52 | 84.13 | 55.55 | 73.08 | 81.26 |
| Validation | 22.85 | 81.16 | 3.96 | 82.50 | 54.02 | 70.99 | 79.02 |
| TESTI | 22.73 | 81.24 | 4.01 | 82.37 | 53.85 | 70.74 | 78.77 |
| TESTNI | 24.01 | 79.98 | 5.54 | 83.51 | 53.55 | 70.49 | 78.14 |
| Set | 23.11 | 80.09 | 3.62 | 83.95 | 55.35 | 72.80 | 80.96 |

Table 3: Average values of the metrics in the Catalan partitions.

| Source | Compression | Coverage | Density | Abstractivity$_p$ (p=2) | Novel 2-grams | Novel 3-grams | Novel 4-grams |
|---|---|---|---|---|---|---|---|
| Training | 27.73 | 82.84 | 5.64 | 80.92 | 51.33 | 68.74 | 76.57 |
| Validation | 26.32 | 83.02 | 5.53 | 80.07 | 49.60 | 66.23 | 73.94 |
| TESTI | 26.20 | 83.11 | 5.58 | 79.92 | 49.40 | 65.96 | 73.64 |
| TESTNI | 13.43 | 72.07 | 6.37 | 86.10 | 59.65 | 74.01 | 79.71 |
| Set | 26.85 | 82.31 | 5.67 | 81.10 | 51.58 | 68.76 | 76.46 |

Table 4: Average values of the metrics in the Spanish partitions.

of extractive summaries, and show great diversity regarding their degree of abstractivity.

## 6 Summarization models and performance results

We evaluate several summarization systems to understand the challenges posed by the DACSA dataset for summarization tasks. We consider both extractive and abstractive models, along with an extractive oracle to show an upper bound of the extractive performance in the corpora.

**Extractive systems**: Lead-k, TextRank (Mihalcea and Tarau, 2004) and SHANN (González et al., 2019) have been evaluated. Lead-k is a heuristic that extracts the first $k$ sentences of a text, being especially well suited to summarize newspaper articles. TextRank is a graph-based system inspired by PageRank, where nodes represent sentences, and edges measure similarities in terms of shared words. Finally, SHANN is a supervised system based on siamese hierarchical attentional networks. The document sentences are scored using sentence-level attentions and those with highest scores are extracted to build the summary. As the average number of sentences in the summaries of DACSA is near to two, we extracted two document sentences by using the extractive systems. We built the extractive systems upon code that is available on Github (Barrios et al., 2016), (González et al., 2019).

**Abstractive systems**: we considered two representative models with high performance on abstractive summarization, based on encoder-decoder architectures with Transformers as backbone: BART and T5. Due to there are neither BART nor T5 models pretrained from scratch for the Spanish and Catalan languages, we finetuned and evaluated their multilingual variants, mBART [1] and mT5[2]. It should be noted that, although both of them considered the Spanish language during pretraining, the Catalan language is not represented in the case of mBART, as this language is not contained in the CC25 dataset. We built the abstractive systems using the HugginFace toolkit (Wolf et al., 2020).

**Oracle**: we implemented an extractive oracle that aligns each summary sentence with the most similar document sentence using ROUGE. The aligned document sentences are concatenated to build the oracle summary.

In order to evaluate the models, we use ROUGE and BERTScore metrics. ROUGE-1, ROUGE-2 and ROUGE-L are reported to measure lexical overlapping, while BERTScore is used to measure semantic similarity.

Tables 5 and 6 show the performance results of the different models on the Catalan and Spanish DACSA TESTI and TESTNI sets in terms of ROUGE and BERTScore metrics. The oracle outperforms the other systems by a large margin. The worse results obtained by the oracle are in the Catalan TESTI, showing that this partition is the most abstractive test partition in the DACSA corpus. Generally, extractive systems are worse in

[1]HuggingFace finetuned mBART models:
ELiRF/mbart-large-cc25-dacsa-ca
ELiRF/mbart-large-cc25-dacsa-es

[2]HuggingFace finetuned mT5 models:
ELiRF/mt5-base-dacsa-ca
ELiRF/mt5-base-dacsa-es

| Partition | Model | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-Ls | BERTScore |
|---|---|---|---|---|---|---|
| TESTI | mBART | 28.59 | 11.89 | 23.00 | 23.39 | 72.03 |
| | mT5 | 27.01 | 10.70 | 21.81 | 22.12 | 71.55 |
| | SHANN | 23.56 | 9.64 | 17.31 | 19.16 | 68.85 |
| | TextRank | 16.54 | 5.62 | 11.98 | 15.33 | 58.90 |
| | Lead-2 | 23.41 | 9.33 | 17.28 | 19.04 | 68.96 |
| | Oracle | 41.68 | 25.53 | 36.29 | 36.64 | 75.87 |
| TESTNI | mBART | 27.46 | 11.04 | 21.13 | 22.01 | 70.33 |
| | mT5 | 27.00 | 11.28 | 21.27 | 22.01 | 70.56 |
| | SHANN | 30.40 | 9.64 | 17.31 | 19.16 | 69.72 |
| | TextRank | 17.16 | 5.83 | 12.27 | 15.93 | 60.36 |
| | Lead-2 | 31.44 | 15.74 | 23.63 | 26.32 | 70.30 |
| | Oracle | 47.16 | 29.44 | 40.23 | 41.82 | 75.86 |

Table 5: Average $F_1$ scores of the models in the summarization task in Catalan.

| Partition | Model | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-Ls | BERTScore |
|---|---|---|---|---|---|---|
| TESTI | mBART | 31.09 | 13.56 | 24.67 | 25.48 | 72.25 |
| | mT5 | 31.72 | 14.54 | 25.76 | 26.31 | 72.86 |
| | SHANN | 26.93 | 11.84 | 20.07 | 22.12 | 69.86 |
| | TextRank | 14.13 | 4.27 | 8.13 | 13.15 | 57.83 |
| | Lead-2 | 29.00 | 14.39 | 22.56 | 24.45 | 71.03 |
| | Oracle | 46.04 | 30.12 | 40.85 | 41.37 | 77.45 |
| TESTNI | mBART | 30.66 | 12.08 | 23.13 | 23.89 | 71.07 |
| | mT5 | 30.61 | 12.36 | 23.53 | 24.05 | 71.26 |
| | SHANN | 35.55 | 15.22 | 24.63 | 27.41 | 70.83 |
| | TextRank | 21.78 | 6.13 | 11.77 | 18.97 | 54.54 |
| | Lead-2 | 36.64 | 16.79 | 26.07 | 28.64 | 71.81 |
| | Oracle | 46.49 | 25.50 | 36.84 | 37.54 | 74.85 |

Table 6: Average $F_1$ scores of the models in the summarization task in Spanish

the TESTI than in the TESTNI, which suggests a higher extractivity in TESTNI than in TESTI. The high results of Lead-2, especially in the TESTNI sets, show that there is a positional bias in these sets.

## 7 Conclusions

Languages other than English have a lack of resources for learning models based on deep learning. This is true for endangered languages but it is also true even for those languages that have millions of speakers but are minority worldwide such as Catalan. In this work, we describe the construction of a corpus of Catalan and Spanish newspapers, the Dataset for Automatic summarization of Catalan and Spanish newspaper Articles (DACSA) corpus. We have included an analysis of the corpus using a set of well-known metrics in the summarization field in order to characterize the corpus. This characterization shows that DACSA provides samples that do not contain a predominance of extractive summaries, and show great diversity regarding their degree of abstractivity. We have also carried out an evaluation of the performance of some extractive and abstractive summarization systems on the DACSA corpus that could be used for benchmarking. To our knowledge, the DACSA corpus is the largest summarization dataset for Catalan and Spanish languages and is freely available for research purposes.

## Ethical considerations

The main objective of this work was to build a quality large-scale corpus that could be used to learn automatic summarization neural models for Catalan and Spanish. To achieve this objective, we selected a set of Spanish news sites, including from Spanish mass media to regional newspapers, and we collected as many data as possible from them. To increase the quality of the corpus, we filtered the article-summary pairs following basic statistics of the text. However, we did not apply any kind of content filtering. Therefore, our filtering could include biased content such as political tendency,

geographic imbalance or gender biases (Stanczak and Augenstein, 2021). A future direction towards improving the dataset quality would be to alleviate that biases, for example, by means of deduplicating content, augmenting artificially the samples to balance gender (Sun et al., 2019), politics, and geographic aspects, or either manually selecting an unbiased subset of the dataset.

The articles collected in the dataset are under Creative Common or private licenses. Nowadays, we are working on obtaining authorization for the distribution of all sources. Those newspaper sources under Creative Common license or the private ones with authorization are freely provided. DACSA can be requested at https://xarrador.dsic.upv.es/resources/dacsa.

## Acknowledgements

This work is part of the AMIC-PoC project (PDC2021-120846-C44), funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR". It is also partially supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València (PAID-11-21).

## References

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. *CoRR*, abs/1602.03606.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748, Brussels, Belgium. Association for Computational Linguistics.

José-Ángel González, Encarna Segarra, Fernando García-Granada, Emilio Sanchis, and Lluís-F Hurtado. 2019. Siamese hierarchical attention networks for extractive summarization. *Journal of Intelligent & Fuzzy Systems*, 36(5):4599–4607.

José Ángel González, Encarna Segarra, Fernando García-Granada, Emilio Sanchis, and Lluís-F. Hurtado. 2020. Extractive summarization using siamese hierarchical transformer encoders. *Journal of Intelligent & Fuzzy Systems*, 39:2409–2419. 2.

J.-A. González, L.-F. Hurtado, E. Segarra, F. García-Granada, and E. Sanchis. 2019. Summarization of spanish talk shows with siamese hierarchical attention networks. *Applied Sciences 2019*, 9(18).

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA. MIT Press.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Deep copycat networks for text-to-text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

*9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3225–3234, Hong Kong, China. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020. Liputan6: A large-scale Indonesian dataset for text summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 598–608, Suzhou, China. Association for Computational Linguistics.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Afonso Mendes, Shashi Narayan, Sebastião Miranda, Zita Marinho, André F. T. Martins, and Shay B. Cohen. 2019. Jointly extracting and compressing documents with summary state representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3955–3966.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3075–3081. AAAI Press.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *Proceedings of the 6th International Conference on Learning Representations ICLR*, pages 1–13, Vancouver, BC, Canada. OpenReview.net.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

*Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *CoRR*, abs/2112.14168.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, USA. Curran Associates Inc.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.

Kaichun Yao, Libo Zhang, Tiejian Luo, and Yanjun Wu. 2018. Deep reinforcement learning for extractive document summarization. *Neurocomputing*, 284:52–62.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 779–784, Brussels, Belgium. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

## A Statistics of DACSA

We show in Tables 7 and 8 a more detailed view of the statistics of the DACSA corpus, distinguishing among the sources from which it was built. The sources that were only considered in the TESTNI partitions are marked with an asterisk.

| | | | Article | | | Summary | | |
|---|---|---|---|---|---|---|---|---|
| Source | #Docs | Tokens | Vocabulary size | Sents per doc | Words per sent | Vocabulary size | Sents per doc | Words per sent |
| CA01 | 238,233 | 114,500,016 | 614,146 | 17.68 | 27.19 | 115,954 | 1.14 | 20.16 |
| CA02 | 194,697 | 105,119,526 | 621,612 | 19.99 | 27.01 | 112,904 | 1.28 | 19.14 |
| CA03 | 137,447 | 63,683,416 | 485,286 | 14.99 | 30.92 | 91,975 | 1.05 | 22.65 |
| CA04 | 56,827 | 24,891,291 | 276,720 | 14.84 | 29.52 | 58,071 | 1.21 | 17.52 |
| CA05 | 44,381 | 26,977,332 | 277,225 | 18.04 | 33.69 | 55,216 | 1.15 | 23.86 |
| CA06 | 35,763 | 17,181,460 | 202,931 | 11.31 | 42.49 | 42,289 | 1.05 | 22.79 |
| CA07* | 7104 | 3,800,842 | 83,942 | 18.04 | 29.66 | 19,267 | 1.02 | 26.51 |
| CA08* | 5882 | 9,414,192 | 185,977 | 66.04 | 24.24 | 31,006 | 2.54 | 24.84 |
| CA09* | 4850 | 2,667,185 | 102,024 | 23.61 | 23.29 | 19,584 | 1.16 | 28.05 |
| Set | 725,184 | 368,235,260 | 1,326,343 | 17.71 | 28.67 | 223,978 | 1.17 | 20.59 |

Table 7: Statistics by source in the Catalan set.

| | | | Article | | | Summary | | |
|---|---|---|---|---|---|---|---|---|
| Source | #Docs | Tokens | Vocabulary size | Sents per doc | Words per sent | Vocabulary size | Sents per doc | Words per sent |
| ES01 | 550,148 | 420,786,144 | 1,473,628 | 31.36 | 24.39 | 210,079 | 1.40 | 19.02 |
| ES02 | 342,045 | 174,411,220 | 907,312 | 16.66 | 30.61 | 148,271 | 1.06 | 22.34 |
| ES03 | 196,410 | 93,755,039 | 622,073 | 15.40 | 31.00 | 110,728 | 1.02 | 20.59 |
| ES04 | 168,065 | 105,628,806 | 659,054 | 23.35 | 26.92 | 112,908 | 1.09 | 22.30 |
| ES05 | 148,053 | 105,453,102 | 626,058 | 28.35 | 25.13 | 109,546 | 1.47 | 20.46 |
| ES06 | 116,561 | 93,956,373 | 524,177 | 26.16 | 30.81 | 169,025 | 1.27 | 43.20 |
| ES07 | 107,162 | 70,944,634 | 470,244 | 19.90 | 33.26 | 87,901 | 1.29 | 25.27 |
| ES08 | 99,098 | 65,352,628 | 495,148 | 25.03 | 26.35 | 81,654 | 1.25 | 18.38 |
| ES09 | 81,947 | 42,825,867 | 363,075 | 15.54 | 33.63 | 71,913 | 1.03 | 22.41 |
| ES10 | 74,024 | 57,782,514 | 470,826 | 30.28 | 25.78 | 81,793 | 1.31 | 20.23 |
| ES11* | 70,193 | 29,692,261 | 272,248 | 11.06 | 38.26 | 84,898 | 1.22 | 44.48 |
| ES12 | 57,235 | 28,198,002 | 294,175 | 16.06 | 30.68 | 58,580 | 1.21 | 19.49 |
| ES13 | 35,163 | 20,156,337 | 260,690 | 19.22 | 29.83 | 50,556 | 1.15 | 21.20 |
| ES14 | 35,112 | 28,408,974 | 309,194 | 30.48 | 26.55 | 78,751 | 1.18 | 28.35 |
| ES15* | 17,379 | 10,099,958 | 153,598 | 16.82 | 34.54 | 41,512 | 1.85 | 26.89 |
| ES16* | 16,965 | 13,791,564 | 166,446 | 28.26 | 28.77 | 29,955 | 1.07 | 25.18 |
| ES17* | 2450 | 4,545,924 | 135,761 | 74.97 | 24.75 | 23,588 | 3.16 | 26.72 |
| ES18* | 1374 | 641,752 | 39,094 | 17.08 | 27.34 | 12,365 | 1.98 | 29.43 |
| ES19* | 643 | 398,834 | 26,797 | 17.73 | 34.99 | 2495 | 1.04 | 16.02 |
| ES20* | 467 | 233,873 | 22,699 | 18.70 | 26.78 | 3857 | 1.22 | 24.23 |
| ES21* | 155 | 199,140 | 19,750 | 39.06 | 32.89 | 2098 | 1.91 | 21.79 |
| Set | 2,120,649 | 1,367,262,946 | 3,189,783 | 23.44 | 27.50 | 516,307 | 1.24 | 22.95 |

Table 8: Statistics by source in the Spanish set.

## B Abstractivity in DACSA

We show in Tables 9 and 10 a fine-grained view of the abstractivity of the DACSA corpus, distinguishing among the sources from which it was built.

| Source | Compression | Coverage | Density | Abstractivity$_p$ (p=2) | Novel 2-grams | Novel 3-grams | Novel 4-grams |
|---|---|---|---|---|---|---|---|
| CA01 | 24.21 | 81.26 | 3.47 | 84.21 | 55.13 | 72.93 | 81.36 |
| CA02 | 23.62 | 80.71 | 3.28 | 85.43 | 56.90 | 74.99 | 82.98 |
| CA03 | 20.81 | 79.95 | 3.27 | 84.88 | 56.73 | 73.92 | 82.14 |
| CA04 | 21.50 | 79.54 | 3.27 | 83.51 | 57.03 | 73.92 | 81.85 |
| CA05 | 24.76 | 83.27 | 5.94 | 76.76 | 47.67 | 63.16 | 70.94 |
| CA06 | 21.88 | 82.45 | 4.53 | 80.48 | 50.73 | 67.09 | 75.12 |
| CA07* | 20.22 | 80.70 | 3.02 | 87.41 | 56.61 | 74.61 | 83.31 |
| CA08* | 31.01 | 72.49 | 2.04 | 95.75 | 65.60 | 85.19 | 92.28 |
| CA09* | 21.09 | 88.00 | 13.48 | 62.98 | 34.44 | 46.63 | 53.37 |
| Set | 23.11 | 80.09 | 3.62 | 83.95 | 55.35 | 72.80 | 80.96 |

Table 9: Average abstractivity metrics by source in the Catalan set.

| Source | Compression | Coverage | Density | Abstractivity$_p$ (p=2) | Novel 2-grams | Novel 3-grams | Novel 4-grams |
|---|---|---|---|---|---|---|---|
| ES01 | 35.07 | 83.64 | 7.26 | 81.25 | 52.32 | 71.16 | 79.22 |
| ES02 | 22.65 | 83.24 | 5.46 | 77.25 | 49.24 | 65.21 | 72.49 |
| ES03 | 23.89 | 81.52 | 3.60 | 82.53 | 54.06 | 71.48 | 79.90 |
| ES04 | 28.31 | 83.78 | 5.54 | 77.77 | 48.99 | 65.27 | 72.84 |
| ES05 | 25.88 | 79.10 | 3.55 | 86.94 | 57.40 | 75.30 | 82.86 |
| ES06 | 16.50 | 83.51 | 6.48 | 85.33 | 46.31 | 63.20 | 71.21 |
| ES07 | 22.55 | 85.31 | 6.53 | 79.31 | 44.69 | 61.50 | 69.70 |
| ES08 | 31.95 | 80.76 | 3.51 | 83.57 | 55.76 | 73.63 | 81.43 |
| ES09 | 24.04 | 80.37 | 3.07 | 85.79 | 56.72 | 74.92 | 83.32 |
| ES10 | 33.36 | 82.58 | 3.98 | 83.60 | 53.33 | 71.91 | 80.12 |
| ES11* | 8.50 | 63.03 | 1.65 | 96.53 | 73.02 | 88.20 | 93.65 |
| ES12 | 23.33 | 81.02 | 5.92 | 77.85 | 53.15 | 69.51 | 76.67 |
| ES13 | 26.35 | 85.67 | 7.90 | 67.78 | 42.31 | 55.97 | 62.51 |
| ES14 | 26.41 | 89.09 | 9.50 | 70.79 | 29.76 | 40.31 | 46.88 |
| ES15* | 11.94 | 94.27 | 24.19 | 51.47 | 20.16 | 27.35 | 30.80 |
| ES16* | 32.02 | 84.84 | 4.22 | 83.45 | 48.88 | 68.16 | 77.59 |
| ES17* | 28.10 | 68.50 | 11.03 | 86.13 | 61.74 | 76.20 | 80.81 |
| ES18* | 10.83 | 94.68 | 39.75 | 37.55 | 14.05 | 18.49 | 21.77 |
| ES19* | 38.80 | 76.20 | 5.07 | 68.91 | 53.72 | 64.12 | 67.99 |
| ES20* | 21.60 | 85.98 | 11.34 | 69.44 | 42.00 | 56.84 | 63.79 |
| ES21* | 39.51 | 78.64 | 4.10 | 90.11 | 56.33 | 73.82 | 81.75 |
| Set | 26.85 | 82.31 | 5.67 | 81.10 | 51.58 | 68.76 | 76.46 |

Table 10: Average abstractivity metrics by source in the Spanish set.