applied sciences

MDPI

Article

NASCA and NASES: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish

Vicent Ahuir *, Lluís-F. Hurtado, José Ángel González * and Encarna Segarra

VRAIN: Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Camí de Vera sn, 46022 València, Spain; lhurtado@dsic.upv.es (L.-F.H.); esegarra@dsic.upv.es (E.S.)
* Correspondence: viahes@eui.upv.es (V.A.); jogonba2@dsic.upv.es (J.Á.G.)

Abstract: Most of the models proposed in the literature for abstractive summarization are generally suitable for the English language but not for other languages. Multilingual models were introduced to address that language constraint, but despite their applicability being broader than that of the monolingual models, their performance is typically lower, especially for minority languages like Catalan. In this paper, we present a monolingual model for abstractive summarization of textual content in the Catalan language. The model is a Transformer encoder-decoder which is pretrained and fine-tuned specifically for the Catalan language using a corpus of newspaper articles. In the pretraining phase, we introduced several self-supervised tasks to specialize the model on the summarization task and to increase the abstractivity of the generated summaries. To study the performance of our proposal in languages with higher resources than Catalan, we replicate the model and the experimentation for the Spanish language. The usual evaluation metrics, not only the most used ROUGE measure but also other more semantic ones such as BertScore, do not allow to correctly evaluate the abstractivity of the generated summaries. In this work, we also present a new metric, called *content reordering*, to evaluate one of the most common characteristics of abstractive summaries, the rearrangement of the original content. We carried out an exhaustive experimentation to compare the performance of the monolingual models proposed in this work with two of the most widely used multilingual models in text summarization, mBART and mT5. The experimentation results support the quality of our monolingual models, especially considering that the multilingual models were pretrained with many more resources than those used in our models. Likewise, it is shown that the pretraining tasks helped to increase the degree of abstractivity of the generated summaries. To our knowledge, this is the first work that explores a monolingual approach for abstractive summarization both in Catalan and Spanish.

Keywords: abstractive summarization; monolingual models; multilingual models; transformer models; transfer learning

Citation: Ahuir, V.; Hurtado, L.-F.; González, J.Á.; Segarra, E. NASCA and NASES: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish. *Appl. Sci.* **2021**, *11*, 9872. https://doi.org/10.3390/app11219872

Academic Editors: Arturo Montejo-Ráez and Salud María Jiménez-Zafra

Received: 17 September 2021
Accepted: 20 October 2021
Published: 22 October 2021

1. Introduction

The purpose of the summarization process is to condense the most relevant information from a document or a set of documents into a small number of sentences. This process can be performed in an extractive or an abstractive way. While extractive summarization consists of identifying and copying those sentences in the original document that contain the most remarkable and useful information, abstractive summaries require abstractive actions that must be mastered. In this way, summaries are not mere clippings of the original documents; rather, abstractive summarizations are created by choosing the most important phrases of the documents and paraphrasing that content, creating a combination of some phrases, introducing new words, searching for synonyms, creating generalizations or specifications of some words or reordering content. All these actions must be done while preserving the linguistic cohesion and the coherence of the information [1–5].

Nowadays, Transformer-based language models excel in text generation, especially due to the transfer learning paradigm, by means of self-supervised pretraining on large text

corpora, and later fine-tuning on downstream tasks. The generation capabilities achieved by these models boosted the state of the art in automatic summarization. However, most of the models proposed in the literature, such as BART [6], PEGASUS [7], or T5 [8] are intended to the English language and are not directly applicable to other languages. Multilingual models such as mBART [9] or mT5 [10] were also studied in the literature to address that language constraint, but despite their applicability being broader than that of the monolingual models, their performance is typically lower, especially on languages that are underrepresented in the pretraining corpora, or differ so much in linguistic terms from the most represented languages [11–14]

For minority languages like Catalan, the data resources available are much lower than other languages like English, Chinese, or Spanish. Additionally, the multilingual models typically either do not include data of minority languages, or if they do, its proportion in the pretraining sets is much lower than those of the majority languages. In this work, we hypothesize that monolingual models are a better choice for those minority languages, such as the Catalan language, which are underrepresented in the pretraining datasets of the multilingual models, but for which reasonable amounts of data are available.

In this work, a BART-like summarization model for the Catalan language is pretrained from scratch, and then fine-tuned on the summarization task. During the pretraining step, we include several self-supervised tasks to enhance the of the degree of abstractivity of the generated summaries. Furthermore, to test our hypothesis about monolingual models, we compare the performance of our proposal against well-known pretrained multilingual models such as mBART and mT5. It is also interesting to study the performance of our proposal in languages with higher resources than Catalan. For this reason, we replicate the model and the experimentation for the Spanish language to extract conclusions about abstractivity and monolingual models in two different languages.

We performed experimentation on the Dataset for Automatic summarization of Catalan and Spanish newspaper Articles (DACSA) corpus [15] This corpus provides pairs of news article and its summary from different journals in the Catalan and the Spanish languages. The experimental results show that the monolingual models generalize better than the multilingual ones, obtaining a more stable summarization performance on the test partitions of the DACSA dataset. The provided experimentation also illustrate the improvements in abstractivity as a result of the addition of the pretraining tasks. We analyze the abstractivity of the models through the use of abstractivity indicators [2]. Following some of these indicators, which correspond to actions done by professional summary writers, we quantify the degree of abstractivity of the generated summaries as the summaries generated by the models. One of the common actions when a person writes an abstractive summary is to rearrange the information from the original document. To our knowledge, no metrics were proposed for this specific action. For this reason, in this work the *content reordering* metric, which aims to quantify the rearrangement degree of the information in the summary with respect to the document, is proposed.

The contributions of this work are the following:

- A monolingual abstractive text summarization model, News Abstract Summarization for Catalan (NASCA), is proposed. This model, based on the BART architecture [6], is pretrained with several self-supervised tasks to improve the abstractivity of the generated summaries. For fine-tuning the model, a corpus of online newspapers is used (DACSA).
- An evaluation of the performance of the model on the summarization task and an evaluation of the degree of abstractivity of its generated summaries are presented. We compare the results of each NAS model with the results obtained by the summarization models based on well-known multilingual language models (mBART [9] and mT5 [10]) fine-tuned for the summarization task for each language using the DACSA corpus.

- A text summarization model with the same pretraining process than NASCA is also trained and evaluated for Spanish, News Abstract Summarization for Spanish (NASES).
- The *content reordering* metric is proposed, which helps to quantify if the extractive content within the abstractive summary is written in a different order than in the document.

The monolingual models, NASCA (https://huggingface.co/ELiRF/NASCA, accessed on 19 October 2021) and NASES (https://huggingface.co/ELiRF/NASES, accessed on 19 October 2021), proposed in this work were publicly release through HuggingFace model hub [16].

2. Related Work

Abstractive summarization works normally focused on the creation of models using approaches different to those used for extractive summarization [17–22]. Recently, abstractive summarizers became ubiquitous due to their powerful generation capabilities, achieved by using encoder-decoder architectures with Transformers [23] as backbone, and by pretraining them with self-supervised language modeling tasks on massive text corpora. This kind of models, especially PEGASUS [7], BART [6], T5 [8] and ProphetNet [24], fine-tuned for summarization tasks, are the state of the art in abstractive summarization benchmarks.

While all these models are nearly identical regarding their architecture, they mainly differ in the self-supervised tasks used in the pretraining stage. In some cases, such as BART, T5, and ProphetNet, these tasks aims the models to learn general aspects of the language, e.g., by masking tokens or reordering sentences. More specifically, BART is pretrained to reconstruct masked spans (text infilling) and to arrange sentences in the original order after being permuted (sentence permutation). Similarly, T5 is pretrained on encoder-decoder masked language modeling, in order to address universally all text-based language problems in a text-to-text format. Regarding ProphetNet, it is pretrained on future n-gram prediction to encourage the model to plan for future tokens instead of the next token, which prevents overfitting on strong local correlations. However, in other cases such as PEGASUS, the self-supervised tasks intentionally resemble the summarization task to encourage whole-document understanding and summary-like generation. In contrast to the previous models, PEGASUS is trained with Gap Sentences Generation (GSG), which consists of reconstructing the sentences that maximize the ROUGE with respect to the whole document. In this way, the authors of PEGASUS hypothesize that GSG is more suitable for abstractive summarization than other pretraining strategies, as it closely resembles the downstream task.

Other works are also based on strategies that involve pretraining to improve the abstractivity of the generated summaries. For instance, in [25], domain transfer and data synthesis techniques by using pretrained models are explored to improve the performance of abstractive summarization models in low-resource scenarios. Also, the authors of [26] propose to use pretrained language models to incorporate prior knowledge about language generation, which provides results comparable to state-of-the-art models in terms of ROUGE, while increasing the level of abstraction of the generated summaries, measured in terms of n-gram overlapping. Finally, in [27] a combination of several pretraining tasks is introduced to tailor the models to abstractive summarization, improving performance upon other Transformer-based models with significantly less pretraining data. Specifically, three tasks were proposed for pretraining: sentence reordering, next segment generation and masked document generation. While sentence reordering and masked document generation are identical to the text infilling and sentence permutation tasks used in BART, next segment generation aims to complete a document given a prefix of that document. Therefore, our work is similar to [27] in the sense that we combine the pretraining tasks of BART and PEGASUS to improve the abstractive skills of monolingual models trained for Catalan and Spanish.

All the models and proposals discussed in this section are intended for the English language, however, there are many other languages that deserve attention. Some efforts were done to consider other languages along with the English language by means of multilingual models such as mBART [9] or mT5 [10]. Although these efforts are very convenient and useful in many cases, the performance of the multilingual models is typically lower on languages that are underrepresented in the pretraining data or differ so much, in linguistic terms, from the most represented languages [13,14]. Learning monolingual models from scratch was extensively explored for language understanding by means of pretraining monolingual BERT models, with excellent results in many languages such as French [12,28], Dutch [29], or Spanish [11,30]. However, monolingual pretraining in languages other than English is still unexplored for language generation tasks such as abstractive summarization. To our knowledge, this is the first work that explores a monolingual approach for abstractive summarization both in Catalan and Spanish.

3. Newspapers Summarization Corpus

As stated above, the models proposed in this work are focused on the specific domain of newspaper articles. To train the models, the Dataset for Automatic summarization of Catalan and Spanish newspaper Articles (DACSA) [15] corpus was used. 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, and their distribution is shown in the Table 1:

Table 1. Statistics of Catalan set. Sources marked with * were not used for training the models.

			Article			Summary		
Source	**Docs**	**Tokens**	|V|	**Sents Per Doc**	**Words Per Sent**	|V|	**Sents Per Doc**	**Words Per Sent**
#1	238,233	114,500,016	614,146	17.68	27.19	115,954	1.14	20.16
#2	194,697	105,119,526	621,612	19.99	27.01	112,904	1.28	19.14
#3	137,447	63,683,416	485,286	14.99	30.92	91,975	1.05	22.65
#4	56,827	24,891,291	276,720	14.84	29.52	58,071	1.21	17.52
#5	44,381	26,977,332	277,225	18.04	33.69	55,216	1.15	23.86
#6	35,763	17,181,460	202,931	11.31	42.49	42,289	1.05	22.79
#7 *	7104	3,800,842	83,942	18.04	29.66	19,267	1.02	26.51
#8 *	5882	9,414,192	185,977	66.04	24.24	31,006	2.54	24.84
#9 *	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

Regarding the Spanish set, the corpus provides 2,120,649 sample pairs from 21 newspapers, distributed as it is detailed in the Table 2:

When the distributions of the samples on both subsets are analyzed, the amount of samples by source is far from being homogeneous. If these distributions preserve over the partitions (training, validation, and test set), the models will focus their learning on the newspapers that are predominant. To avoid this bias and achieve more general models, the test and validation sets were created in a way that ensured that all newspapers had roughly the same number of samples on those sets. To achieve this balance in the validation and test sets, the sources with less samples were discarded. In this way, it is guaranteed that all sources represent at least 5% of samples in each one of these two sets. The sources that were excluded are marked with an asterisk in the Tables 1 and 2.

The three sets for Catalan include 6 of the 9 newspapers, creating a training set that contains 636,596 samples and 35,376 samples for validation and test sets. In the case of Spanish, the three sets are composed of 13 of the 21 newspapers provided in the Spanish set of DACSA: the training set contains 1,802,919 samples, and the validation and test sets contain 104,052 samples each.

Table 2. Statistics of Spanish set. Sources marked with * were not used for training the models.

			Article			Summary		
Source	**Docs**	**Tokens**	|V|	**Sents Per Doc**	**Words Per Sent**	|V|	**Sents Per Doc**	**Words Per Sent**
#1	550,148	420,786,144	1,473,628	31.36	24.39	210,079	1.40	19.02
#2	342,045	174,411,220	907,312	16.66	30.61	148,271	1.06	22.34
#3	196,410	93,755,039	622,073	15.40	31.00	110,728	1.02	20.59
#4	168,065	105,628,806	659,054	23.35	26.92	112,908	1.09	22.30
#5	148,053	105,453,102	626,058	28.35	25.13	109,546	1.47	20.46
#6	116,561	93,956,373	524,177	26.16	30.81	169,025	1.27	43.20
#7	107,162	70,944,634	470,244	19.90	33.26	87,901	1.29	25.27
#8	99,098	65,352,628	495,495148148	25.03	26.35	81,654	1.25	18.38
#9	81,947	42,825,867	363,075	15.54	33.63	71,913	1.03	22.41
#10	74,024	57,782,514	470,826	30.28	25.78	81,793	1.31	20.23
#11 *	70,193	29,692,261	272,248	11.06	38.26	84,898	1.22	44.48
#12	57,235	28,198,002	294,175	16.06	30.68	58,580	1.21	19.49
#13	35,163	20,156,337	260,690	19.22	29.83	50,556	1.15	21.20
#14	35,112	28,408,974	309,194	30.48	26.55	78,751	1.18	28.35
#15 *	17,379	10,099,958	153,598	16.82	34.54	41,512	1.85	26.89
#16 *	16,965	13,791,564	166,446	28.26	28.77	29,955	1.07	25.18
#17 *	2450	4,545,924	135,761	74.97	24.75	23,588	3.16	26.72
#18 *	1374	641,752	39,094	17.08	27.34	12,365	1.98	29.43
#19 *	643	398,834	26,797	17.73	34.99	2495	1.04	16.02
#20 *	467	233,873	22,699	18.70	26.78	3857	1.22	24.23
#21 *	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

All the sources excluded were used as a separate test set. This partition allows to evaluate the generalization capabilities of the models. 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 3 and 4.

Table 3. Statistics of partitions for Catalan language.

				Article			Summary	
Partition	**Docs**	**Tokens**	|V|	**Sents Per Doc**	**Words Per Sent**	|V|	**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

Table 4. Statistics of partitions for Spanish language.

				Article			Summary	
Partition	**Docs**	**Tokens**	|V|	**Sents Per Doc**	**Words Per Sent**	|V|	**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

4. Summarization Models

In this work, a monolingual news summarization model is proposed: News Abstractive Summarization for Catalan (NASCA). It is a Transformer encoder-decoder model with the same architecture and hyper-parameters as BART [6]. Inspired by the work

of Zou et al. [27], we decided to combine several pretraining tasks to inject linguistic knowledge during the pretraining stage with the aim of increasing the abstractivity of the summaries generated by the model. Specifically, four tasks were combined: sentence permutation, text infilling [6], Gap Sentence Generation (GSG) [7], and Next Segment Generation (NSG) [27]. NASCA is pretrained simultaneously with the four tasks, which are randomly selected at each batch following a uniform distribution.

We hypothesize that the combination of these four pretraining tasks leads to improvements in the summarization task, especially concerning the abstractivity of the generated summaries. Firstly, with sentence permutation and text infilling, the model should acquire capabilities of content reordering and phrase replacements. Secondly, GSG should tailor the model to whole-document understanding, summary-like generation and paraphrasing. Finally, with NSG, the model could increase the cohesion of the whole summary, as the task consists of generating continuations of documents given a prefix.

NASCA was pretrained with the documents of the Catalan training set of the DACSA corpus (including some documents discarded in the corpora creation process [15]), the Catalan subset of the OSCAR corpus [31], and the dump from 20 April 2021 of the Catalan version of the Wikipedia. In total, 9.3 GB of raw text (2.5 millions of documents) were used to pretrain it.

Additionally, we replicated NASCA for the Spanish language. We refer to this model as News Abstractive Summarization for Spanish (NASES). NASES is identical to NASCA in terms of architecture and pretraining tasks, but they differ in the pretraining dataset. To pretrain NASES, we only used the Spanish documents of the DACSA corpus and the dump from 20 April 2021 of the Spanish version of the Wikipedia. We did not consider for NASES the Spanish subset of OSCAR corpus so as to not increase excessively the difference in the amount of data available for the Spanish model regarding the Catalan one. In total, 21 GB (8.5 million documents) were used to pretrain NASES. Note that even though we did not use the OSCAR corpus, the size of the pretraining dataset for Spanish is twice the size of the Catalan pretraining dataset.

In addition to the monolingual models, two multilingual models were used for the experimental comparison in the summarization task. We worked with two of the most widely used multilingual models in text summarization, mBART and mT5. Regarding the mBART model, we used the *mbart-large-cc25* version, released by Facebook and available online through HuggingFace (https://huggingface.co/facebook/mbart-large-cc25, accessed on 19 October 2021) [16]. For the mT5 model, we used the *mt5-base* version, published by Google, that is also available online (https://huggingface.co/google/mt5-base, accessed on 19 October 2021)).

All the monolingual and multilingual models were fine-tuned and evaluated for the summarization task using the DACSA corpus. The monolingual models proposed in this work were publicly released (https://huggingface.co/ELiRF/NASCA, accessed on 19 October 2021), (https://huggingface.co/ELiRF/NASES, accessed on 19 October 2021).

5. Metrics

To evaluate the performance of the summarization models we used the usual evaluation metrics, the most used ROUGE measure [32] which is based on n-grams, and a more semantic such as BertScore [33], which is based on contextual embeddings provided by a BERT language model. However, these metrics do not allow to correctly evaluate the abstractivity of the generated summaries.

Measuring the abstractivity of the summaries generated by the models is, except counting the introduced new words, not trivial. In some studies, abstractivity was measured as the absence of n-gram overlap [34,35], however, creating abstractive summaries is not just about solely of using different vocabulary [2]. In this work, we used a set of metrics as abstractivity indicators to asses the level of abstractivity. In particular, the following metrics were selected: *extractive fragment coverage* [34], *abstractivity*$_p$ [35], *novel 1-grams*, *novel 4-grams* [26]. Also in this work, we present a new metric, called *content reordering*,

to evaluate one of the most common characteristics of abstractive summaries, the rearrangement of the original content.

The *content reordering* metric was defined to quantify the percentage of reordering that the information in the summary suffered with respect to its original order in the document. This metric correlates positively with the abstractivity, and thus, by reordering the information, the summary increases its abstractivity.

The measure is based on the inversion concept. The inversion operation extracts all pairs of items that are out of order: $INV(\pi) = \{(a_i, a_j) | i < j \wedge a_i > a_j\}$, where π is a list of comparable elements [36]. For instance, with the list $[1, 5, 4, 2]$, the inverse operation results in $[(5, 4), (5, 2), (4, 2)]$.

Given a list of pairs (u, v), where u is the position of a maximum length segment in the original document, and v is the position in which such segment is placed in the summary, this list is sorted by u and the number of inversions that must be made to order the list of pairs by v is calculated. Thus, this allows us to quantify the disorder established in the list of the second component of the pairs when we take into account the order of the first component.

Let $\mathcal{F}(T, S)$[34] be the operation that returns the longest common extractive segments between a text T and its summary S, let $|S|$ be the number of words of the summary, and let $Reordered(T, S)$ be the operation that counts the number of extractive reordered segments; *content reordering* is defined as follows:

$$ContentReordering(T, S) = \begin{cases} \frac{\sum_{f \in \mathcal{F}(T,S)} |f|}{|S|} \cdot \frac{Reordered(T,S)}{|\mathcal{F}(T,S)|-1}, & |\mathcal{F}(T,S)|>1. \\ 0, & otherwise. \end{cases}$$

The output value range of the function is $[0, 1]$, where 1 is the highest degree of information rearrangement.

To illustrate this metric, we provide a full example with the following text (T):

> $_{1}$*Content reordering* *is a metric that* $_{7}$*quantifies* *how the extracted information from the original document is rearranged in the summary.* $_{21}$*Reorder the content* $_{24}$*is a common action* *used* $_{28}$*in abstractive summarization*.

and the following summary (S):

> $_{1}$*In abstractive summarization*, $_{4}$*reorder the content* $_{7}$*is a common action*, $_{11}$*content reordering* $_{13}$*quantifies* *it*.

The highlighted text are fragments in common between the original text and its summary. The subindex before the fragment indicates the starting position in words of the fragment. Thus, the list of the pairs (u, v) of the extractive fragments is the following one when it is ordered by u:

$$[(1, 11), (7, 13), (21, 4), (24, 7), (28, 1)]$$

The resulting list of the INV operation applied on the list made up with the second components of the pairs of the previous list is:

$$INV([11, 13, 4, 7, 1]) = [(\underline{11}, 4), (11, 7), (11, 1), (\underline{13}, 4), (13, 7), (13, 1), (\underline{4}, 1), (\underline{7}, 1)]$$

The $Reorder(T, S)$ operation is 4 since there are 4 extractive reordered segments. This value is computed as the unique values in the first components of the pairs in the previous list (11, 13, 4, 7). Additionally, the length (in words) of the summary is 14, there are 5

extractive fragments, and the sum of their length is 13. With all this information, the *content reordering* metric is calculated as follows:

$$ContentReordering(T,S) = \frac{13}{14} \cdot \frac{4}{5-1} = 0.93$$

With this result, we conclude that there is a certain degree of abstractivity in the summary introduced by a high degree of rearrangement of the information. This fact can be verified in the summary of the example. This abstractivity was introduced by the rearrangement of the extractive segments, and not due to the absence of text overlapping between the summary and the original text.

6. Results

In this section, we present the conducted experimentation with the summarization models. Firstly, we present the results of the performance obtained by the three models for Catalan in the summarization task: the NASCA model, the mBART model, and the mT5 model. Secondly, we show the results regarding the abstractivity of these models for Catalan. Additionally, we show the results for the three models for Spanish, the NASES model and the two multilingual ones. All the models were evaluated on the two test partitions, TESTI and TESTNI.

6.1. Summarization Performance of the Models for CATALAN

The performance of the models was evaluated using the ROUGE metrics [32] and BERTScore metric [33]. For each metric, we calculated the average F1 score and its 95% confidence interval by using bootstrapping. Results are shown in Table 5.

The average F1 scores are shown in a normal font size and their confidence intervals in a smaller font size, placed at the right-side of the score. The best average score for each metric within a test partition is remarked in bold style. The confidence intervals are shown in blue color if their range intersects with the confidence interval of the best score value of the metric within the same test partition; in other case, the confidence intervals are presented in black color.

Table 5. Average F1 scores and confidence intervals of models in summarization task in Catalan.

Partition	Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Ls	BERTScore
TESTI	NASCA	**28.84** (28.68, 29.01)	11.68 (11.51, 11.85)	22.78 (22.61, 22.94)	23.30 (23.13, 23.46)	71.85 (71.78, 71.92)
	mBART	28.59 (28.42, 28.77)	**11.89** (11.73, 12.06)	**23.00** (22.82, 23.16)	**23.39** (23.22, 23.56)	**72.03** (71.96, 72.10)
	mT5	27.01 (26.84, 27.18)	10.70 (10.54, 10.87)	21.81 (21.65, 21.97)	22.12 (21.98, 22.29)	71.55 (71.49, 71.61)
TESTNI	NASCA	**28.19** (27.97, 28.42)	11.20 (10.99, 11.43)	**21.45** (21.20, 21.65)	**22.44** (22.21, 22.67)	70.14 (70.05, 70.22)
	mBART	27.46 (27.24, 27.69)	11.04 (10.81, 11.29)	21.13 (20.93, 21.37)	22.01 (21.78, 22.24)	70.33 (70.25, 70.43)
	mT5	27.00 (26.77, 27.23)	**11.28** (11.04, 11.52)	21.27 (21.03, 21.51)	22.01 (21.78, 22.23)	**70.56** (70.47, 70.65)

The Table 5 shows, regarding the TESTI partition, that the NASCA model performs similarly compared to the multilingual mBART model. mBART presents significantly better BERTScore result than NASCA while there are overlappings in the confidence intervals in the ROUGE measures. The mT5 model has obtained a significant lower performance than the other two models, despite the fact that mT5 contains the Catalan language in its pretraining phase unlike the mBART model. We hypothesize that the pretraining dataset could influence the results. It could be that the data considered for Catalan to pretrain mT5 differs so much from our domain. Also, the proportion of languages similar to Catalan in the pretraining corpus could be related to this effect.

In the case of the TESTNI partition, there is a significant overall reduction of the performance in most of the metrics of the three models in comparison to the TESTI partition. Generally speaking, the NASCA model has significantly better performance in almost all ROUGE metrics compared to the multilingual models, although there is an

overlapping between the confidence interval of NASCA and that of mT5 in ROUGE-2. According to BERTScore, the mT5 model obtains significant differences in comparison to the scores of the NASCA and mBART models.

Taking into account the higher scores and the generalization capabilities, the results of the monolingual model are significant better than the multilingual ones. In one side, mBART has similar performance than NASCA model in the TESTI partition, however, the performance reduction in the second test partition indicates that the model generalizes worse than the other two models. On the other side, the mT5 model generalizes better than mBART, since the drop of the perfomance between the TESTI and the TESTNI is lower in mT5 than mBART, however, mT5 presents significantly lower performance than that of the NASCA model.

6.2. Abstractivity of the Summaries Generated by the Models for Catalan

To evaluate the abstractivity, 4 metrics were used: *extractive fragment coverage* [34] (henceforth, we refer to it simply as *coverage*), *abstractivity*$_p$ [35], *novel n-grams* [26]and *content reordering*. From now on, we refer those metrics as indicators, since each indicator complements, in some way, the other indicators to obtain a global perception of the level of abstractivity. The Table 6 shows the average scores and their confidence intervals. The scores are calculated by comparing the generated summaries against to their respective article text. The scores remarked in bold styles indicates the highest abstractivity. In this experimentation, the lowest value is emphasized in the *extractive fragment coverage* indicator since it correlates negatively with the abstractivity and the highest value is remarked in the remaining abstractivity indicators, since they correlate positively.

Table 6. Abstractivity indicators and confidence intervals for Catalan. Values are shown as percentages.

Partition	Model	Extractive Fragment Coverage	Content Reordering	Abstractivity$_p$ (p = 2)	Novel 1-Grams	Novel 4-Grams
TESTI	NASCA	**96.99** (96.94, 97.04)	46.17 (45.79, 46.55)	**47.19** (46.90, 47.46)	**03.21** (03.15, 03.26)	**28.65** (28.41, 28.92)
	mBART	97.73 (97.68, 97.77)	**47.85** (47.44, 48.23)	37.70 (37.42, 37.97)	02.40 (02.36, 02.45)	23.80 (23.55, 24.02)
	mT5	98.59 (98.55, 98.62)	41.25 (40.84, 41.67)	38.04 (37.78, 38.28)	01.51 (01.48, 01.55)	21.89 (21.71, 22.08)
TESTNI	NASCA	**96.66** (96.55, 96.77)	42.37 (41.84, 42.88)	**41.89** (41.44, 42.37)	**03.52** (03.40, 03.63)	**26.32** (25.91, 26.68)
	mBART	97.08 (96.99, 97.16)	**42.96** (42.40, 43.56)	36.98 (36.55, 37.41)	03.01 (02.92, 03.09)	24.32 (23.95, 24.70)
	mT5	98.31 (98.26, 98.36)	38.82 (38.24, 39.41)	39.18 (38.83, 39.54)	01.80 (01.74, 01.85)	23.20 (22.92, 23.48)

As it is shown in Table 6, all the models show a predominant extractivity behavior in the same way as the most abstractive models in the literature. All the scores of the abstractivity indicators denote low abstractivity. For instance, the *coverage* and *novel 1-grams* indicators show that the models reuses a lot of words from the original documents. Although all the models present high-extractivity in their generated summaries, there are significant differences among the models that can be analyzed.

Regarding the TESTI partition, the scores of most of the abstractivity indicators of the NASCA model reflect significantly better abstractivity than that of the multilingual models. Also, we can observe that the multilingual models have relatively similar scores in most of the indicators, although, the indicators of the mBART model show slightly more abstractivity than the mT5 model.

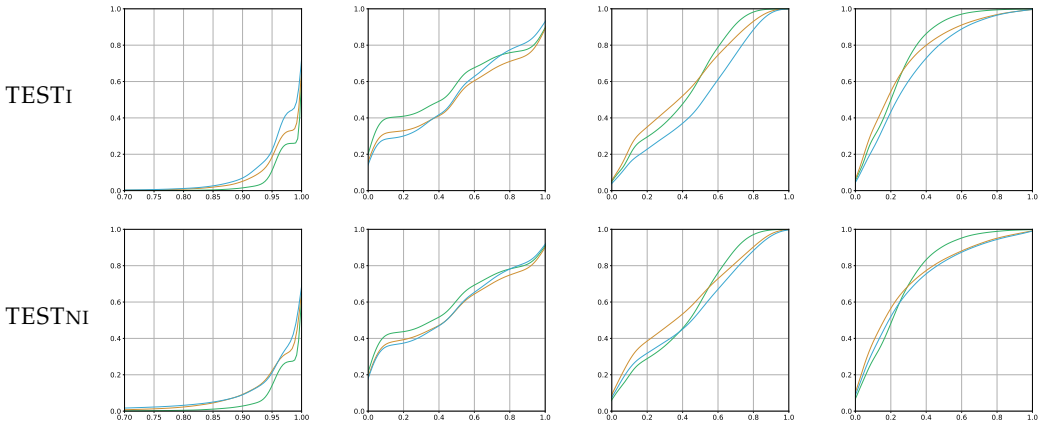
In the case of the TESTNI partition, the NASCA model indicators reflect better abstractivity than in the multilingual models. However, compared to the values in TESTI, NASCA reduced most of their abstractivity indicators scores except the *coverage* indicator, which is slightly better. In this partition, the differences in the values between the NASCA model and the multilingual models are lower than in the TESTI partition.

Overall, it is noticeable that the NASCA model reuses a lot of content from the original text. The model uses a lot of words from the original text which is reflected in the low value of the *novel 1-grams* indicator. However, despite the fact that the model reuses a lot of words, the extractive fragments tend to be shorter than in the multilingual models, since the *novel 4-grams* indicator shows a significantly higher value than in the multilingual models;

this fact is also exposed by the *abstractivity*$_p$ indicator, which presents a difference between the 5% and the 10% depending on the partition and the multilingual model. For all these observations in the indicators, we conclude that the NASCA model generates summaries with higher degree of abstractivity than the multilingual models.

With the aim of better analyzing the behavior of the models, we computed the cumulative distributions of the abstractivity indicators for each model and test partition. The results are presented in the Figure 1.

The plots show in the x-axis the indicator measured, and in the y-axis, the percentage of generated summaries that present less or equal score to the value in the x-axis. These plots are helpful to evaluate the abstractivity of the generated summaries by taking into account how they are distributed based on certain score. If a metric correlates negatively with the abstractivity, it is desired that the scores be lower; that is, the model accumulates the samples fast. In contrast, if the metric correlates positively, it is desired that the scores be higher. In this case, we say that the model accumulates the samples slowly.



(**a**) *Ext. Frag. Coverage* (**b**) *Content Reordering* (**c**) Abstractivity$_p$ ($p = 2$) (**d**) *Novel 4-grams*

Figure 1. Cumulative distribution of 4 abstractivity indicators for models NASCA, mBART, mT5 for Catalan.

In Figure 1, regarding the *coverage* indicator, which correlates negatively with abstractivity, we observe that the NASCA model stays always on top of the multilingual models, so this indicates that the samples are accumulated faster, which is a positive indication for the abstractivity. In the remaining indicators, which correlate positively with the abstractivity, the NASCA model tends to accumulate the samples slower than the multilingual models, which is also positive concerning the abstractivity, except the *content reordering* indicator. Regarding this indicator, although NASCA present a lower value than the mBART model in the Section 6.2, the NASCA model's distribution stays below the mBART until 40%, and later reaches and surpasses the multilingual models. This means that the *NASca* model, overall, introduces less *content reordering* on their summaries; however, the amount of summaries with rearrangement of the information is higher than in the ones generated by the multilingual models.

The results presented in the Table 6 and the Figure 1 show enough evidences to conclude that the NASCA model presents better abstractivity than the rest of the trained models. Additionally, to verify if the improvement in the abstractivity indicators is due to the pretraining tasks, we pretrained a BART model specifically for Catalan using only the pretraining tasks proposed in the original work [6]. The results show that both models, NASCA and BART, have a similar performance in the summarization task, however, the NASCA model presents significant higher abstractivity indicators. For instance, in the *coverage* indicator of the TESTNI partition, the NASCA model scores 96.99 (96.94, 97.04) and BART 97.29 (97.24, 98.41). In the case of novel 4-grams, and also for TESTNI, the NASCA model scores 26.65 (25.91, 26.68) and BART 25.48 (25.12, 25.82).

An example of an article and the summaries generated by the three models is shown in Appendix A.

6.3. Summarization Performance and Abstractivity of the Summaries Generated by the Models for Spanish

It is also interesting to study the performance of our proposal in languages with higher resources than Catalan. For this reason, we replicated the model and the experimentation for the Spanish language. The summarization performance results and the results related to the abstractivity indicators are shown in Tables 7 and 8, respectively. In addition, the cumulative distributions of the abstractivity indicators are presented in Figure 2.

Table 7. Average F1 scores and confidence intervals of models in summarization task in Spanish.

Partition	Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Ls	BERTScore
TESTI	NASES	**33.24** (33.12, 33.38)	**15.79** (15.63, 15.93)	**26.76** (26.63, 26.89)	**27.56** (27.43, 27.69)	**73.11** (73.05, 73.16)
	mBART	31.09 (30.98, 31.20)	13.56 (13.44, 13.68)	24.67 (24.56, 24.78)	25.48 (25.37, 25.58)	72.25 (72.21, 72.30)
	mT5	31.72 (31.60, 31.85)	14.54 (14.39, 14.67)	25.76 (25.63, 25.89)	26.31 (26.18, 26.44)	72.86 (72.82, 72.91)
TESTNI	NASES	30.60 (30.52, 30.68)	10.75 (10.66, 10.83)	22.29 (22.21, 22.37)	23.06 (22.99, 23.15)	70.66 (70.62, 70.69)
	mBART	**30.66** (30.58, 30.74)	12.08 (11.98, 12.18)	23.13 (23.06, 23.22)	23.89 (23.81, 23.98)	71.07 (71.04, 71.10)
	mT5	30.61 (30.51, 30.70)	**12.36** (12.25, 12.47)	**23.53** (23.43, 23.62)	**24.05** (23.95, 24.14)	**71.26** (71.22, 71.30)

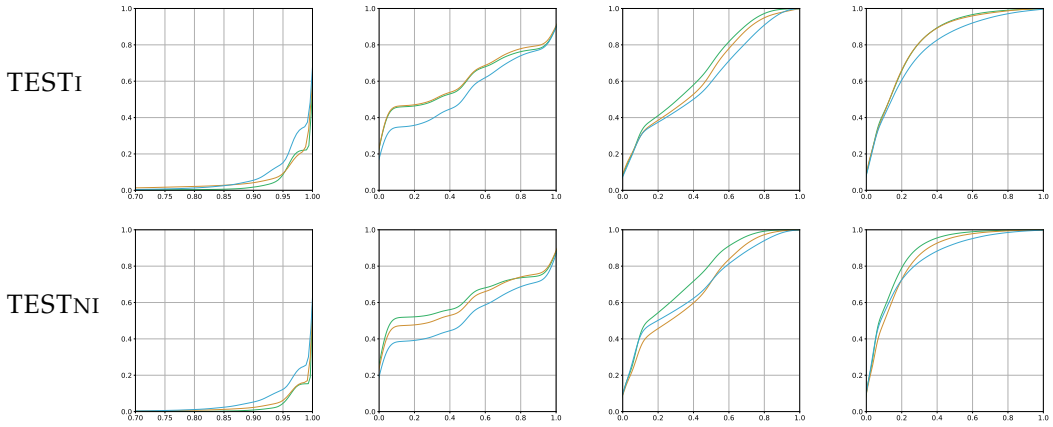
Table 7 shows that the NASES model presents the best performance of the three models in the TESTI partition. All the scores obtained by the NASES model are significantly better compared to those of the multilingual models. Specifically, the NASES model achieve, on average, 8.2% higher performance than mBART and 4.5% higher than mT5. Regarding the TESTNI partition, the NASES model reduces its performance in average, while mT5 achieves the best results in almost all the metrics.

The results show that the NASES excelled in the TESTI partition, which contains newspapers included in the training partition. However, NASES presents lower generalization capabilities than the multilingual models due to the noticeable performance reduction in the TESTNI partition, which contains newspapers not included in the training partition.

Table 8. Abstractivity indicators and confidence intervals for Spanish. Values are shown as percentages.

Partition	Model	Extractive Fragment Coverage	Content Reordering	Abstractivity$_p$ (p = 2)	Novel 1-Grams	Novel 4-Grams
TESTI	NASES	**97.65** (97.62, 97.68)	**45.27** (45.04, 45.50)	**38.15** (37.97, 38.31)	**02.55** (02.52, 02.58)	**21.17** (21.04, 21.31)
	mBART	98.14 (98.10, 98.18)	37.70 (37.45, 37.92)	35.17 (35.00, 35.32)	01.85 (01.81, 01.89)	17.58 (17.47, 17.70)
	mT5	98.74 (98.72, 98.76)	38.67 (38.42, 38.92)	32.41 (32.25, 32.58)	01.36 (01.34, 01.38)	17.39 (17.29, 17.49)
TESTNI	NASES	**98.16** (98.13, 98.19)	**46.58** (46.33, 46.82)	29.76 (29.60, 29.92)	**02.00** (01.97, 02.03)	**15.76** (15.65, 15.88)
	mBART	98.92 (98.90, 98.94)	39.38 (39.13, 39.61)	**30.48** (30.33, 30.64)	01.03 (01.01, 01.05)	14.68 (14.59, 14.78)
	mT5	99.24 (99.23, 99.26)	37.17 (36.91, 37.43)	24.19 (24.06, 24.32)	00.83 (00.81, 00.84)	12.08 (12.00, 12.16)

Regarding the abstractivity indicators on the TESTI partition, presented in Table 8, all the scores of the NASES model are significantly better than those of the multilingual models. In the TESTNI partition, the models present less abstractivity in comparison to the TESTI partition. Also in TESTNI, the NASES model shows significant differences compared to the multilingual models in all the indicators, excluding *abstractivity*$_p$ where mBART obtains better scores than NASES and the mT5 models. We also computed the cumulative distributions of the abstractivity indicators for each model and test partition. The results are presented in the Figure 2.



(**a**) *Ext. Frag. Coverage* (**b**) *Content Reordering* (**c**) Abstractivity$_p$ ($p = 2$) (**d**) *Novel 4-grams*

Figure 2. Cumulative distribution of 4 abstractivity indicators for models NASES, mBART, mT5 for Spanish.

The plots presented in Figure 2 help us to reinforce the observations extracted from the numerical results showed in Table 8. The NASES model tends to accumulate slightly higher percentage of samples in the *coverage* indicator after the 90% of *coverage* is achieved. Regarding the remaining indicators, the accumulation tends to occur slower than in the other two models.

The abstractivity indicators analysis shows that the summaries generated by NASES have a significant higher abstractivity than those generated by the multilingual models, something that complements the observations made in the Sections 6.1 and 6.2 about the models for Catalan.

7. Conclusions

In this work, a monolingual model for abstractive summarization in Catalan, NASCA, was presented. The model was pretrained from scratch based on the BART architecture and using four self-supervised tasks with the aim of increasing the abstractivity of the generated summaries. The fine-tuning phase was carried out using the DACSA dataset, a corpus of articles obtained from online newspapers. The experimentation conducted supports the correctness of our proposal considering the three evaluated aspects: the performance of the model, the abstractivity of the generated summaries, and the generalization capabilities of the model.

Following the same architecture and the same training strategy, a model for abstractive summarization in Spanish, NASES, was also trained and evaluated, and it also provided very good results. To our knowledge, this is the first work that explores a monolingual approach for abstractive summarization both in Catalan and Spanish.

Additionally, in this work, we also proposed a new metric, *content reordering*, with the aim of helping to quantify the rearrangement of the original content within an abstractive summary. This characteristic is common in abstractive summaries, but it is not considered by the metrics in the literature.

Author Contributions: V.A. conceptualization, software, formal analysis, resources, data curation, writing—original draft preparation, and visualization. L.-F.H.: methodology, validation, formal analysis, resources, writing—review and editing, supervision, project administration, and funding acquisition. J.Á.G.: methodology, software, investigation, data curation, writing—original draft preparation, and visualization. E.S.: conceptualization, validation, investigation, writing—review and editing, supervision, project administration, and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Spanish Ministerio de Ciencia, Innovación y Universidades and FEDER founds under the project AMIC (TIN2017-85854-C4-2-R), and by the

Agencia Valenciana de la Innovació (AVI) of the Generalitat Valenciana under the GUAITA (INNVA1/2020/61) project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DACSA	Dataset for Automatic summarization of Catalan and Spanish newspaper Articles
GSG	Gap Sentences Generation
MDG	Masked Document Generation
NASCA	News Abstractive Summarization for Catalan
NASES	News Abstractive Summarization for Spanish
NSG	Next Segment Generation
SR	Sentence Reordering

Appendix A. Summarization Example

An example of an article, its reference summary, and the summaries generated by the three models are shown in Figure A1. It also shows the different metrics achieved by each summary. All the generated summaries are syntactically and semantically correct. Based on the low values of the ROUGE scores, we can affirm that all the generated summaries are very different from the reference one. Regarding the coverage indicator, although the three summaries are quite extractive, since they use several segments from the article, mT5 is by far the most extractive. Considering all the abstractive indicators, NASCA and mBART are better than mT5, and NASCA outperforms mBART especially in terms of novel n-grams and abstractivity$_p$.

Article: La clau va ser el ritme. El ritme amb què Marc Márquez va arrencar al Gran Premi de l'Argentina i amb què el va acabar. El pilot de Cervera, que sempre assegura que li agraden les curses en grup, va fer avançaments, va buscar els forats i va passar-s'ho bé dalt de la moto: a l'Argentina va decidir ser, per un dia, infidel al seu estil. Sabia que tenia ritme, ho havia demostrat durant totes les sessions d'entrenaments lliures i també als oficials (havia dominat cinc de les sis sessions), i a la cursa no va tenir rival. Va sortir, va posar el "mode creuer", com va dir, i va perdre de vista la resta de rivals. En una volta, un segon d'avantatge, i ja s'escapava de 12 segons dels perseguidors quan va decidir passar a controlar la cursa, sense prendre més riscos dels necessaris. "No és el meu estil, però després del que va passar l'any passat tenia ganes de fer una cursa així. Va passar el que va passar i volia demostrar el meu ritme", va assegurar després de baixar de la moto. Márquez va marcar la pole i la volta ràpida, i va ser líder des que es van apagar els semàfors fins al final. Va aconseguir el que es coneix com un Grand Chelem: el de Cervera, de fet, tan sols n'ha aconseguit cinc des que va debutar a MotoGP; tres a Austin (2014, 2016 i 2018), un a Jerez (2014) i el de diumenge a l'Argentina. "Pocs dies a l'any et trobes amb aquestes sensacions dalt de la moto. Calia aprofitar-ho, ha sigut perfecte", reconeixia. La manera més dolça de marcar el ritme. La victòria es va començar a coure molt abans de la sortida, al box, amb el seu equip, llegint els temps de les sessions d'entrenaments. "Els papers deien que era qui tenia més ritme. He intentat marcar les diferències en les set primeres voltes i, després, mantenir l'avantatge", explicava el català. Com si fos un rellotge, clavava volta a volta un 1:39. Al final, els 12 segons d'avantatge es van reduir a 9.816, que, si bé no és la distància més gran amb què Márquez ha guanyat una cursa (a Brno el 2017 va acabar primer amb 12.438 respecte a Pedrosa), sí que és la més gran que ha aconseguit el de Cervera en una cursa en sec: tant a Brno fa dos anys com a Sachsenring en fa tres, en què va acabar a 9.857 de Crutchlow, la pluja va marcar les curses. Lluny també queden els més de 37 segons d'avantatge amb què Dani Pedrosa va guanyar a Xest el 2012 sobre Nakasuga, també sota la pluja, després de la caiguda de Lorenzo. "Com que hem guanyat per deu segons, sembla que som en un altre món, però no, la distància és només de quatre punts respecte a Dovizioso", afegia Márquez. Just abans del podi es va veure segurament una de les imatges de l'any: Valentino Rossi, que va acabar segon, va encaixar la mà amb Márquez, un gest que no es veia des de feia un any, quan el de Cervera, precisament a Termas de Río Hondo, va tocar l'italià i el va fer caure, cosa que va comportar l'inici d'un terratrèmol. Diumenge, ja al podi, els dos campions van fer xocar les ampolles de xampany, però sense dirigir-se la paraula.

Reference: El triomf de Márquez a l'Argentina, el més ampli en sec del de Cervera a MotoGP.

NASCA: El de Cervera va marcar la 'pole' a l'Argentina i va ser líder del Mundial en una volta.
(ROUGE-1: 5.97; ROUGE-2: 4.42; ROUGE-L: 4.72; BertScore: 67.08)
(Coverage: 85.00; Reordering: 85.00; Abstractivity$_p$: 87.75; Novel 1-grams: 15.79; Novel 4-grams: 94.12)

mBART: El de Cervera marca la 'pole' a l'Argentina i és líder des que es van apagar els semàfors.
(ROUGE-1: 6.28; ROUGE-2: 4.72; ROUGE-L: 5.97; BertScore: 69.17)
(Coverage: 85.00; Reordering: 85.00; Abstractivity$_p$: 79.75; Novel 1-grams: 15.00; Novel 4-grams: 70.59)

mT5: El pilot de Cervera, que sempre assegura que li agraden les curses en grup, va fer avançaments, va buscar els forats i va passar-se bé dalt de la moto.
(ROUGE-1: 9.58; ROUGE-2: 8.68; ROUGE-L: 9.27; BertScore: 72.96)
(Coverage: 96.97; Reordering: 48.48; Abstractivity$_p$: 35.54; Novel 1-grams: 3.70; Novel 4-grams: 13.33)

Figure A1. Text of the article, the reference summary, and the summaries generated by the models.

References

1. Rane, N.; Govilkar, S. Recent Trends in Deep Learning Based Abstractive Text Summarization. *Int. J. Recent Technol. Eng.* **2019**, *8*, 3108–3115. [CrossRef]
2. Jing, H. Using Hidden Markov Modeling to Decompose Human-Written Summaries. *Comput. Linguist.* **2002**, *28*, 527–543. [CrossRef]
3. Verma, P.; Pal, S.; Om, H. A Comparative Analysis on Hindi and English Extractive Text Summarization. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2019**, *18*, 1–39. [CrossRef]
4. Widyassari, A.P.; Rustad, S.; Shidik, G.F.; Noersasongko, E.; Syukur, A.; Affandy, A.; Setiadi, D.R.I.M. Review of automatic text summarization techniques & methods. *J. King Saud Univ. Comput. Inf. Sci.* **2020**. [CrossRef]
5. National Information Standards Organization. *Guidelines for Abstracts*; Standard, American National Standards Institute: Gaithersburg, MD, USA, 1997.
6. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. 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, Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7871–7880. [CrossRef]
7. Zhang, J.; Zhao, Y.; Saleh, M.; Liu, P. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, 13–18 July 2020; pp. 11328–11339.
8. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.

9. Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; Zettlemoyer, L. Multilingual Denoising Pre-training for Neural Machine Translation. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 726–742. [CrossRef]
10. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico City, Mexico, 6–11 June 2021; pp. 483–498. [CrossRef]
11. Cañete, J.; Chaperon, G.; Fuentes, R.; Ho, J.H.; Kang, H.; Pérez, J. Spanish Pre-Trained BERT Model and Evaluation Data. 2020. Available online: https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf (accessed on 19 October 2021).
12. Martin, L.; Muller, B.; Ortiz Suárez, P.J.; Dupont, Y.; Romary, L.; de la Clergerie, É.V.; Seddah, D.; Sagot, B. CamemBERT: A Tasty French Language Model. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020.
13. Virtanen, A.; Kanerva, J.; Ilo, R.; Luoma, J.; Luotolahti, J.; Salakoski, T.; Ginter, F.; Pyysalo, S. Multilingual is not enough: BERT for Finnish. *arXiv* **2019**, arXiv:1912.07076 [CrossRef]
14. Pires, T.; Schlinger, E.; Garrette, D. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 4996–5001. [CrossRef]
15. DACSA: A Dataset for Automatic summarization of Catalan and Spanish newspaper Articles. Unsubmitted.
16. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
17. Zhong, M.; Liu, P.; Chen, Y.; Wang, D.; Qiu, X.; Huang, X. Extractive Summarization as Text Matching. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 6–10 July 2020; pp. 6197–6208. [CrossRef]
18. Liu, Y.; Lapata, M. 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), Hong Kong, China, 3–7 November 2019; pp. 3730–3740. [CrossRef]
19. Nallapati, R.; Zhai, F.; Zhou, B. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; AAAI'17; p. 3075–3081.
20. Rush, A.M.; Chopra, S.; Weston, J. A Neural Attention Model for Abstractive Sentence Summarization. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 379–389. [CrossRef]
21. Nallapati, R.; Zhou, B.; dos Santos, C.; Gúllçehre, Ç.; Xiang, B. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, 11–12 August 2016; pp. 280–290. [CrossRef]
22. See, A.; Liu, P.J.; Manning, C.D. Get To The Point: Summarization with Pointer-Generator Networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1073–1083. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 20–22 June 2017; p. 6000–6010.
24. Qi, W.; Yan, Y.; Gong, Y.; Liu, D.; Duan, N.; Chen, J.; Zhang, R.; Zhou, M. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, Online, 16–20 November 2020; pp. 2401–2410.
25. Magooda, A.; Litman, D.J. Abstractive Summarization for Low Resource Data using Domain Transfer and Data Synthesis. In Proceedings of the The Thirty-Third International Flairs Conference, North Miami Beach, FL, USA , 17–20 May 2020.
26. Kryściński, W.; Paulus, R.; Xiong, C.; Socher, R. Improving Abstraction in Text Summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October 31–4 November 2018; pp. 1808–1817. [CrossRef]
27. Zou, Y.; Zhang, X.; Lu, W.; Wei, F.; Zhou, M. Pre-training for Abstractive Document Summarization by Reinstating Source Text. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 Novermber 2020; pp. 3646–3660. [CrossRef]
28. Le, H.; Vial, L.; Frej, J.; Segonne, V.; Coavoux, M.; Lecouteux, B.; Allauzen, A.; Crabbé, B.; Besacier, L.; Schwab, D. FlauBERT: Unsupervised Language Model Pre-training for French. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 2479–2490.
29. de Vries, W.; van Cranenburgh, A.; Bisazza, A.; Caselli, T.; van Noord, G.; Nissim, M. BERTje: A Dutch BERT Model. *arXiv* **2019**, arXiv:1912.09582.
30. Ángel González, J.; Hurtado, L.F.; Pla, F. TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter. *Neurocomputing* **2021**, *426*, 58–69. [CrossRef]
31. Ortiz Suárez, P.J.; Romary, L.; Sagot, B. A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 1703–1714.

32. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
33. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020.
34. Grusky, M.; Naaman, M.; Artzi, Y. 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), New Orleans, LA, USA, 1–6 June 2018; pp. 708–719. [CrossRef]
35. Bommasani, R.; Cardie, C. Intrinsic Evaluation of Summarization Datasets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 16–20 November 2020; pp. 8075–8096. [CrossRef]
36. Barth, W.; Mutzel, P.; Jünger, M. Simple and Efficient Bilayer Cross Counting. *J. Graph Algorithms Appl.* **2004**, *8*, 179–194. [CrossRef]