applied sciences

MDPI

Article

Attentional Extractive Summarization †

José Ángel González ‡, Encarna Segarra ‡, Fernando García-Granada *,‡, Emilio Sanchis ‡ and Lluís-F. Hurtado ‡

Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, 46022 Valencia, Spain

* Correspondence: fgarcia@dsic.upv.es

† This paper is an extended version of our paper published in IberSpeech 2022.

‡ These authors contributed equally to this work.

Abstract: In this work, a general theoretical framework for extractive summarization is proposed—the Attentional Extractive Summarization framework. Although abstractive approaches are generally used in text summarization today, extractive methods can be especially suitable for some applications, and they can help with other tasks such as Text Classification, Question Answering, and Information Extraction. The proposed approach is based on the interpretation of the attention mechanisms of hierarchical neural networks, which compute document-level representations of documents and summaries from sentence-level representations, which, in turn, are computed from word-level representations. The models proposed under this framework are able to automatically learn relationships among document and summary sentences, without requiring Oracle systems to compute the reference labels for each sentence before the training phase. These relationships are obtained as a result of a binary classification process, the goal of which is to distinguish correct summaries for documents. Two different systems, formalized under the proposed framework, were evaluated on the CNN/DailyMail and the NewsRoom corpora, which are some of the reference corpora in the most relevant works on text summarization. The results obtained during the evaluation support the adequacy of our proposal and suggest that there is still room for the improvement of our attentional framework.

Keywords: siamese neural networks; hierarchical neural networks; attention mechanisms; extractive summarization

Citation: González, J.Á.; Segarra, E.; García-Granada, F.; Sanchis, E.; Hurtado, L.-F. Attentional Extractive Summarization. *Appl. Sci.* **2023**, *13*, 1458. https://doi.org/10.3390/app13031458

Academic Editors: Francesc Alías, José Luis Pérez Córdoba, Zoraida Callejas Carrión and António Joaquim da Silva Teixeira

Received: 22 December 2022
Revised: 10 January 2023
Accepted: 17 January 2023
Published: 22 January 2023

1. Introduction

In recent years, automatic text summarization has made strides mainly due to two factors: the success of Deep Learning models and the use of a large amount of information available on the web for building large corpora in order to train the Deep Learning models. The automatic text summarization problem has been addressed in the literature using abstractive, extractive, or mixed approaches. Extractive approaches compose summaries by selecting sentences or words directly from the documents, whereas abstractive approaches build the summaries by paraphrasing/rewriting the sentences of the documents. Furthermore, there are also mixed strategies that combine extractive and abstractive techniques, performed in a decoupled way or simultaneously during the training of the models. Recently, an important effort has been made to develop abstractive methods. However, extractive approaches are still important in summarization, since they maintain the coherence, the factuality, and do not hallucinate like abstractive approaches do. Additionally, selecting sentences is also important for other tasks such as Text Classification, Question Answering and Information Extraction.

Some successful approaches to extractive summarization are based on graph representations of the documents. This is the case with LexRank [1] and TextRank [2–5], among others. Other approaches are based on Neural Networks. Typically, these neural network-based approaches have been addressed as a sequential binary sentence classification problem [6–14]. However, the available corpora do not directly provide this kind

of labeling for training purposes since, in general, corpora only consist of (document, summary) pairs. In order to label the document sentences, prior to the training of the model, the most common strategy consists of using suboptimal extractive oracles [6–9]. Additionally, several unsupervised approaches for extractive summarization have been proposed by Joshi et al. [15], and Mohd et al. [16]. Recently, Reinforcement Learning strategies have been extensively applied [10–13] in order to dispense with the sentence labeling and optimizing directly the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric [17].

The research objective of this work is to present in detail the general theoretical framework for extractive summarization called *Attentional Extractive Summarization*. Our proposal dispenses with the sentence labeling, avoiding the large computational cost required to compute near-optimal solutions, and allowing us to address the summarization problem in a simpler way than Reinforcement Learning techniques. Specifically, our proposal is based on the interpretation of the attention mechanisms of neural models that are trained to distinguish correct summaries for documents. It should be noted that it is only required a binary signal in order to train the model instead of sentence labeling. This training allows our system to learn relationships among document and summary sentences, and to replace the binary sequential sentence classification with a binary classification among documents and summaries. After the training of the model, it is possible to select the most attended sentences by focusing on the document sentence attentions computed by the model.

The *Attentional Extractive Summarization* framework was proposed with the aim of generalizing our previous proposals in extractive summarization and boosting future works and improvements under this framework. Gonzalez et al. [18] proposed a Siamese neural model based on Hierarchical Attention Networks [19]. Later, in [20], an extension to other types of attention mechanisms, in particular, to multi-head self-attention mechanisms of the Transformer Encoders [21], was proposed by Gonzalez et al. In this work, these two systems are also instantiated under the proposed framework, replacing each theoretical component by concrete neural network-based approaches such as Hierarchical Attention Networks or Hierarchical Transformer Encoders to compute sentence and document representations, and attention mechanisms to compute sentence scores. Therefore, the proposed framework also allows the development of novel summarization systems, in addition to those presented in this work. For example, a system based on Hierarchical Convolutional Neural Networks, trained to distinguish correct summaries for documents and that relies on statistics such as the norm of the activations in order to compute sentence scores, would also fall under the umbrella of our framework. The performances of our systems were evaluated and studied on the CNN/DailyMail [22] and NewsRoom [23] corpora, comparing them with more recent systems based on diverse strategies (extractive and mixed summarization systems based on oracles or reinforcement learning). A preliminary version of this work has been presented in Gonzalez et al. [24].

In this paper, the theoretical framework is presented in detail: two systems ([18,20]) previously proposed by Gonzalez et al. were instantiated under the *Attentional Extractive Summarization* framework, an extensive evaluation of the systems on CNN/DailyMail and NewsRoom corpora was performed, showing that our systems are competitive with other extractive and mixed state-of-the-art systems. A detailed analysis of the results was performed, including the convergence of our models and the word-length distribution of system generated summaries and, finally, several examples are provided to illustrate the generated summaries and the attention weights used to score the sentences.

This paper is organized as follows. In Section 2, the state-of-the-art systems for extractive and mixed summarization, which were compared to our proposal for the CNN/DailyMail and NewsRoom corpora, are described. In Section 3, the main characteristics of our *Attentional Extractive Summarization* framework are introduced, to be formalized in Section 4. In Section 5, two systems that fall under the umbrella of our framework are defined. In Sections 6 and 7, the corpora and the hyper-parameters of the

systems are presented, respectively. In Section 8, our systems were compared to other state-of-the-art systems for the CNN/DailyMail and NewsRoom corpora. In Section 9, several analyses of the behavior of the proposed systems were performed. Finally, in Section 10, some conclusions and future works are presented.

2. Related Work

In this section, different approaches to extractive and mixed summarization are described, in particular those state-of-the-art systems used in the experimental comparison of this work. Recently, the construction of large corpora [22,23,25,26] has allowed the training of Deep Learning systems for the automatic summarization problem.

A very robust baseline used commonly in newspaper summarization is the Lead system. It is based on extracting the first k sentences of the documents to compose a summary. Although it seems naive, it is especially robust when it is applied to articles in newspapers, since in this domain, generally, the first sentences are dedicated to condensing the information of the whole document and they are used to attract the reader's attention.

Extractive approaches can be divided into two different categories: those which use an oracle algorithm to label the sentences of the documents before training the models, and those which directly optimize the ROUGE evaluation metric by means of Reinforcement Learning strategies.

Regarding the extractive systems based on oracles, the first approaches were proposed by Cheng et al. [6] and Nallapati et al. [7]. In [6], an encoder-decoder approach for extractive single-document summarization was proposed. In [7] (SummaRunner), Nallapati et al. presented two versions of Hierarchical Attention Networks to select sentences from the documents as a binary sequence classification problem. One of these versions is trained using the samples provided by the corpus without a previous sentence labeling. The other version requires a greedy algorithm as an oracle for labeling the corpus at sentence level, selecting as a reference summary the set of sentences from the document that maximize the similarity with respect to the reference summary. Recently, the great impact of the Transformer architecture [21] in Natural Language Processing tasks, and particularly in language modeling [27], has boosted the results in automatic summarization by fine-tuning powerful pre-trained language models. The most relevant example is the BertSumEXT system [8], which is based on the fine-tuning of pre-trained Bidirectional Encoder Representations from Transformers (BERT) models [27]. Liu et al. [8] also proposed abstractive and mixed strategies for generating summaries starting from the pre-trained BERT.

Reinforcement Learning strategies for automatic summarization have received great interest from the research community. Despite the first works on Reinforcement Learning being intended to perform abstractive summarization by Paulus et al. [28], recently these strategies have been widely used for extractive text summarization, directly optimizing the ROUGE evaluation measure [10–14]. Narayan et al. [10] argued about the application of cross-entropy with ground-truth sentence labels to optimize neural summarization models, and they proposed the application of the REINFORCE algorithm [29] for extractive summarization in order to train a hierarchical encoder-decoder. Zhang et al. [11] also discussed the suboptimal nature of the labels obtained using oracles. They presented a latent variable extractive model, which can also be viewed as a Reinforcement Learning approach, where the reward is defined as a weighted sum of two measures related to the precision and the recall. These measures were computed from the likelihood of a summary sentence and a document sentence, estimated using an attention-based sequence-to-sequence sentence compression model. This system can be trained in an extractive (Latent) or in a compressive way (Latent-Comp). A theoretically grounded method (BanditSum) was proposed by Dong et al. [12] to model the extractive summarization problem by means of a bandit formalism. They proposed a novel structure for computing the conditional probability of a subset of document sentences given the document, which avoids privileging early sentences over later ones. An approach based on Deep Q Learning (DQN) was proposed by Yaok et al. [13]. This approach is based on an iterative decision problem, where a sentence is selected at

each step. After each sentence selection, the state of the model is updated and the selected sentence is added to the summary state.

Recently, the interest in mixed strategies has increased. These approaches are typically based on first extracting a set of sentences and later adapting them to the reference summaries, e.g., Mendes et al. proposed compressing [30] and Chen et al. proposed paraphrasing [14]. In [30], they proposed a compressive approach that removes unnecessary words while keeping the summaries informative, concise, and grammatically correct. The model can be trained in an extractive way (ExConSumm-Ext) and in a compressive way (ExConSumm-Comp). In [14], they proposed a sentence-level policy gradient method to first select salient sentences and then to paraphrase them (Fast-RL). Other types of mixed strategies are those based on selecting or generating a new word at each step as in See et al. [31], or Ive et al. [32]. The most relevant example is [31], where an approach based on Pointer Networks and encoder-decoder models with attention mechanisms is proposed. Moreover, in order to address the word repetition problem, the authors enrich their system by using a coverage mechanism based on the encoder attentions of previous steps, for each decoder step (PointerGen+Cov). This system has been modified by the authors of [32], replacing Long Short Term Memories [33] with Transformers [21].

In this work, a theoretical framework for extractive summarization is proposed. It is based on the interpretability of the attention mechanisms proposed by Vaswani et al. [34] of Siamese hierarchical networks trained for distinguishing correct summaries for documents. Differently from the extractive approaches discussed before in this section, our approach is able to learn directly relationships among document and summary sentences, dispensing with extractive oracles and with the sequential sentence labeling. (A similar paradigm that addresses extractive summarization as a semantic matching problem has been explored recently in the literature [9]). Two systems that fall under the umbrella of the *Attentional Extractive Summarization* framework, were previously proposed by Gonzalez et al.: Siamese Hierarchical Attention Neural Networks [18] (SHA-NN) and Siamese Hierarchical Transformer Encoders [20] (SHTE). They are discussed in more detail in Section 5.

3. Attentional Extractive Summarization Framework

As pointed out before, approaches that do not rely on Reinforcement Learning strategies to directly optimize the ROUGE evaluation metric, are mainly based on the use of suboptimal oracle algorithms since they require a binary sentence labeling in order to be trained. These approaches typically consist of using oracle systems to label the sentences by following some evaluation measures such as ROUGE. In the paper of Narayan et al. [10], two types of oracles are distinguished: individual oracles, that label each sentence independently (e.g., semantic similarity above a threshold) and collective oracles that consider dependencies among sentences (e.g., greedy algorithms to search combinations of document sentences that maximize the ROUGE with respect to the reference summary). As stated in [10], the problem of the first type of oracles is that they often generate too many positive labels, causing the model to overfit the data. In the other case, the main problem is related to the underfitting, since the models trained with cross-entropy loss on collective labels will only maximize probabilities for the sentences in the selected sets. Collective oracles are common in the literature [7,8,30,35,36].

To require a sentence labeling for training the systems has several drawbacks. First, the labeling is suboptimal and it can fall in local optimum, leading the model to be trained with non-relevant sentences or missing relevant ones as it is shown in Zhang et al. [11]. Second, this problem becomes more complex for large corpora, where obtaining oracles can be computationally intensive if near-optimal solutions are preferred. Furthermore, the sequential classification, where each sentence is classified taking into account its dependencies with all the other sentences in the document, is a complex problem that can be simplified.

Our proposal allows the systems to learn by themselves relationships among the sentences of documents and reference summaries. These relationships are learned by

attention mechanisms, that are interpreted to extract the most relevant document sentences. In order to learn these relationships and to avoid training with a sequential classification problem on sequences of labeled sentences, we propose to address the summarization task as a binary classification problem where correct summaries are distinguished from incorrect summaries for documents. (We consider as incorrect summaries, for a given document, the reference summaries of other documents in the corpora.) This way, it is only required a binary signal in order to train the models, instead of sentence labeling. We call this proposal *Attentional Extractive Summarization* framework.

It is possible to identify the required mechanisms for designing systems based on the proposed framework. First, it is required to learn representations for documents and summaries that can be used to distinguish if a summary is correct for a given document. Regarding this point, we used hierarchical models in order to compute document-level representations from the sentence-level representations, which were built from the word-level representations. Second, a mechanism to distinguish correct summaries for documents, from the document-level representations, has to be designed. In our framework, this mechanism is based on Siamese networks, which use the document-level representations to address the summarization task as a binary classification problem, where a probability distribution of the summary correctness is computed. Finally, an interpretable mechanism to compute relationships among document and summary sentences is required. In our proposal, we focused on the attention mechanisms of the hierarchical models at document level in order to compute the relevance of the document sentences. In this way, it is possible to assign a score to each sentence (based on its relevance when distinguishing correct and incorrect summaries) and rank these scores to extract the k most relevant sentences.

4. Framework Definition

A scheme of our framework can be seen in Figure 1. Let $\mathcal{D} = \{(X_k, X'_k)\}_{k=1}^{M}$ be a corpus of M (document, summary) pairs, where all documents and summaries are defined according to a vocabulary $\mathcal{V}$, let $X_k = \{\{x_{ij}\}_{i=1}^{W}\}_{j=1}^{T}$ be a document composed by T sentences of W words, $X_k \in \mathcal{V}^{T\times W}$, let $X'_k = \{\{x'_{ij}\}_{i=1}^{V}\}_{j=1}^{Q}$ be a summary composed by Q sentences of V words, (Although documents and summaries of the dataset can have arbitrary lengths, the models based on neural networks digest fixed-length representations achieved by means of truncating or padding. So, we used a maximum number of sentences (T and Q) and a maximum number of words per sentence (W and V) to better reflect it. Regarding the notation, $\mathcal{V}^{A\times B}$ is intended to represent the set of A lists with B words each one, each word belonging to $\mathcal{V}$). $X'_k \in \mathcal{V}^{Q\times V}$ and let $f : \mathcal{V}^{T\times W} \times \mathcal{V}^{Q\times V} \to \mathbb{R}^2$ be a model whose input is a (document, summary) pair and whose output is a probability distribution of the summary correctness over $\mathbb{C} = \{0, 1\}$, where 0 is for incorrect summaries and 1 is for correct summaries.

The objective is that the model $f(.,.;\Theta)$ has to be able to determine if a (X, X') pair is correct or incorrect. This way, the class computed from the output of the model for the (X_k, X'_k) pair will be 1, as X'_k is the reference summary for the document X, while for the $(X_k, X'_{j\neq k})$ pair, the class computed will be 0, as $X'_{j\neq k}$ is the reference summary for another document from the corpus $\mathcal{D}$, different from X. In order to do that, the model must represent documents and summaries in a proper way to distinguish each case. Thus, $f(.,.;\Theta)$ relies on a document encoder $g : \mathcal{V}^{T\times W} \to \mathbb{R}^{d_g}$ and in a summary encoder $g' : \mathcal{V}^{Q\times V} \to \mathbb{R}^{d_{g'}}$. These encoders have to be able to model the hierarchical structure of documents and summaries, so that $g(.;\theta_1)$ and $g'(.;\theta_2)$ are decomposed in two different levels.

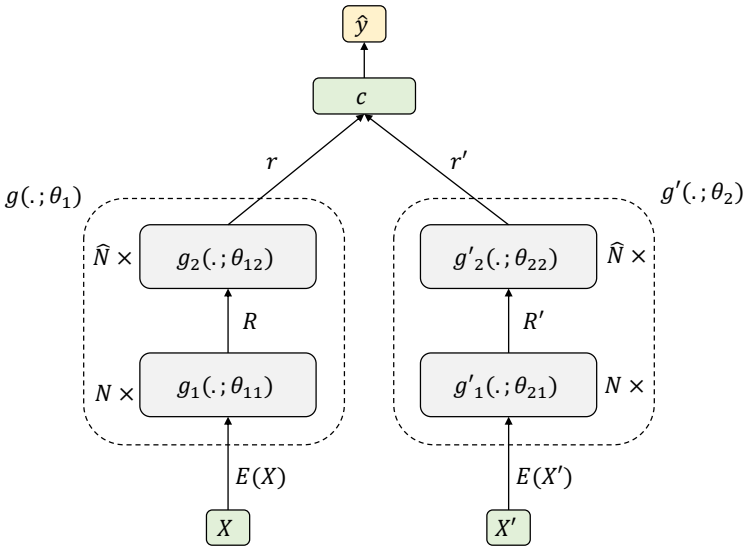

Figure 1. General scheme of the Attentional Extractive Summarization framework.

First, $g_1 : \mathcal{V}^{T\times W} \to \mathbb{R}^{T\times d_{g_1}}$ and $g'_1 : \mathcal{V}^{Q\times V} \to \mathbb{R}^{Q\times d_{g'_1}}$ that are applied independently on each sentence (of documents and summaries respectively) to obtain sentence-level representations from the word-level representations. The encoders can be composed of N hidden layers. In practice, the words are represented by means of a d_e-dimensional embedding model $\mathrm{E} : \mathcal{V} \to \mathbb{R}^{d_e}$, typically pretrained and applied to arbitrary-length (P) word sequences, i.e., $\mathrm{E} : \mathcal{V}^P \to \mathbb{R}^{P\times d_e}$. Therefore, $g_1 : \mathbb{R}^{T\times W\times d_e} \to \mathbb{R}^{T\times d_{g_1}}$ and $g'_1 : \mathbb{R}^{Q\times V\times d_e} \to \mathbb{R}^{Q\times d_{g'_1}}$. Second, in order to represent documents and summaries from the representation of their sentences, $g_2 : \mathbb{R}^{T\times d_{g_1}} \to \mathbb{R}^{d_g}$ and $g'_2 : \mathbb{R}^{T\times d_{g'_1}} \to \mathbb{R}^{d_{g'}}$ are defined. These encoders can have $\hat{N}$ hidden layers. Basically, the sentence encoders can be any function that digests a three-dimensional tensor of word embeddings representing the words inside the sentences of a text, and generates a vector representation for each sentence of the text. Similarly, the document encoders, can be any function that digests a matrix of sentence representations to generate a vector representation of the whole text.

Therefore, the encoders $g(.;\theta_1)$ and $g'(.;\theta_2)$ are defined as a composition of two levels, $g = g_2(R;\theta_{12})$ and $g' = g'_2(R';\theta_{22})$, where $R = g_1(.;\theta_{11})$ and $R' = g'_1(.;\theta_{21})$. Because both documents and summaries come from the same domain, they could be represented in the same way through the use of the same set of parameters in both cases, i.e., $\theta_{11} = \theta_{21}$ and $\theta_{12} = \theta_{22}$, leading to Siamese architectures. Although this is possible, the θ parameters are not constrained to be always shared, so, for the sake of simplicity, we also refer to these architectures as Siamese networks. The parameters of the documents and summaries encoders are defined as $\theta_1 = [\theta_{11}, \theta_{12}]$ and $\theta_2 = [\theta_{21}, \theta_{22}]$.

As stated before, the document encoder $g(.;\theta_1)$ must be interpretable so that it must assign relevance scores both to words, in order to compute sentence representations, and to sentences, in order to compute document representations. Our approach consists in designing these encoders by means of attention mechanisms that assign scores to words and sentences. Then, document representations are computed as an average of their sentence representations, using the document level attention mechanism. At the same time, the sentence representations are computed as an average of their words, using the sentence level attention mechanism. The application of these mechanisms is diverse and they can be applied as auxiliary functions on top of the encoders [37,38] as in [18] or as main mechanisms to compute representations [21] as in [20].

Let $r = g(.;\theta_1)$ and $r' = g'(.;\theta_2)$ be the representations of document and summary respectively, the system must be able to determine if the summary is correct for the document, by using r and r'. In order to do this, a classifier $c(.,.;\theta_3)$ whose output is a probability distribution over $\mathbb{C}$, $c : \mathbb{R}^{d_g} \times \mathbb{R}^{d_{g'}} \to \mathbb{R}^2$, is applied. Therefore, the model $f(.,.;\Theta)$ can be

seen as a classifier $c(.,.;\theta_3)$ applied on top of the encoder outputs, both for document, r, and summary, r', i.e., $f(.,.;\Theta) = c(r, r';\theta_3)$. The parameters of the model are determined by the parameters of each subpart: encoders for documents and summaries and the classifier, $\Theta = [\theta_1, \theta_2, \theta_3]$.

The objective is that the model $f(.,.;\Theta)$ must be able to classify correctly the largest number of pairs, both the positives (extracted directly from the corpora) and the negatives (for a given document, reference summaries from all the other documents in the corpora, sampled by following a distribution p). Therefore, the objective is determined by the minimization of Equation (1).

$$\mathcal{L}(\Theta) = \sum_{k=1}^{|\mathcal{D}|} \mathrm{L}(f(X_k, X'_k;\Theta), y = 1) + \mathop{\mathbb{E}}_{p(X_{j\neq k}|X_k)} [\mathrm{L}(f(X_k, X'_j;\Theta), y = 0)], \tag{1}$$

where L is a loss function, and $\mathbb{E}_{p(X_{j\neq k}|X_k)}$ denotes expectation with respect to a Bernoulli distribution with parameter p.

It is interesting to highlight that, once the system is trained for minimizing the training objective, the encoders $g(.;\theta_1)$ and $g'(.;\theta_2)$ must compute proper representations of documents and summaries, respectively. In this way, the document representations, computed from their sentences by using the attention mechanism of $g_2(.;\theta_{12})$, are useful to distinguish correct and incorrect (document, summary) pairs. Moreover, this attention mechanism is able to assign a relevance score to each document sentence. Thus, it is possible to determine, focusing on the $g_2(.;\theta_{12})$ attentions, which document sentences have a greater impact on the document representation, being these sentences the most related with the reference summary.

Finally, it is also interesting to highlight that the attention mechanism of $g_1(.;\theta_{11})$ can be used to extract keywords from the documents, being the most attended words inside a sentence those mostly related with respect to the reference summary. We have not experimented in this work with these attentions, but it opens the door for future improvements by considering the words along with the sentences during the summarization process.

5. Proposed Systems

From the definition of the general framework, presented in the previous section, it is possible to design systems based on it for extractive summarization. To do this, it is necessary to define the encoders both for documents and summaries and both at sentence ($g_1(.;\theta_{11})$ and $g'_1(.;\theta_{21})$) and document level ($g_2(.;\theta_{12})$ and $g'_2(.;\theta_{22})$). Furthermore, it is also required to define a strategy for sentence scoring based on the attention mechanisms of document encoder $g_2(.;\theta_{12})$. In the following subsections, a formalization of two systems proposed inside the *Attentional Extractive Summarization* framework is defined [18,20].

5.1. Siamese Hierarchical Attention Networks

The Siamese Hierarchical Attention Neural Network (SHA-NN) [18] is an instance of the general attentional framework when the encoders are Hierarchical Attention Networks [19] based on Bidirectional Long Short Term Memory (BLSTM) [33,39] with attention mechanisms, i.e., $g_1(.;\theta_{11}) = \mathrm{BLSTM}_1(.;\theta_1)$, $g'_1(.;\theta_{21}) = \mathrm{BLSTM}_1(.;\theta_1)$, $g_2(.;\theta_{12}) = \mathrm{BLSTM}_2(.;\theta_2)$ and $g'_2(.;\theta_{22}) = \mathrm{BLSTM}_2(.;\theta_2)$. The BLSTM layers are shared for documents and summaries, both at sentence level (BLSTM_1 with dimensionality d_w) and at document level (BLSTM_2 with dimensionality d_s). However, the attention mechanisms for both branches of the Siamese model are not shared. Regarding classifier c, it is a feed-forward network. The architecture can be seen in Figure 2.

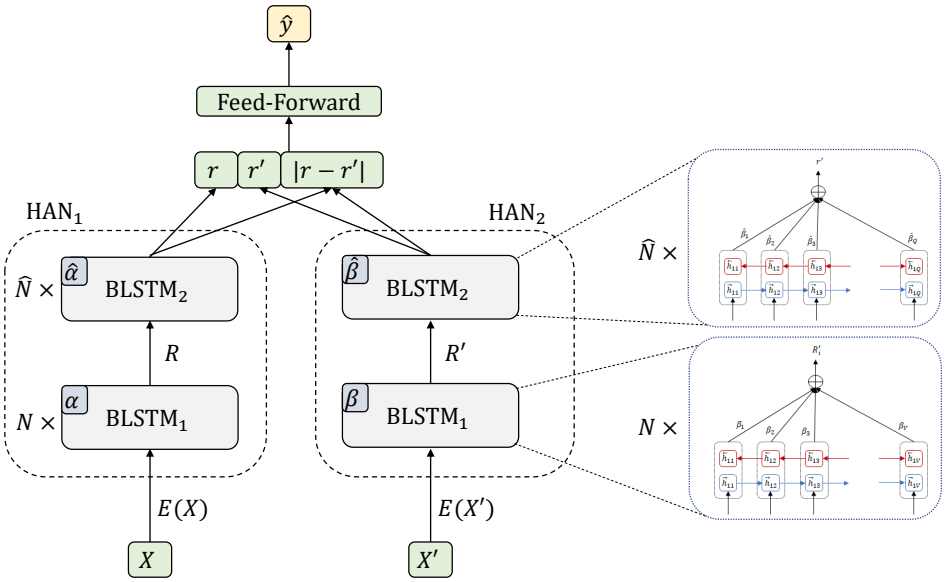

Figure 2. SHA-NN Architecture.

For this approach, $R \in \mathbb{R}^{T \times d_w}$ and $R' \in \mathbb{R}^{Q \times d_w}$ are computed, following Equations (2) and (4), as proposed in [37]. They are the output from the sentence level d_w-dimensional BLSTM$_1$ with attention, where each row i is computed as the average of the hidden vectors of the sentence i attended by $\alpha \in \mathbb{R}^{T \times W}$ (Equation (3)) and $\beta \in \mathbb{R}^{Q \times V}$ (Equation (5)) for document and summary respectively. This process is applied independently to each word embedding matrix that represents each sentence both for document and summary ($R_i : 1 \leq i \leq T$ and $R'_i : 1 \leq i \leq Q$). The following equations show a sentence encoder composed by $N = 1$ BLSTM network.

$$R_i = \sum_{j=1}^{W} \text{BLSTM}_1(\text{E}(X_i))_j \cdot \alpha_{ij} \tag{2}$$

$$\alpha_{ij} = \frac{e^{\tanh(W_u \text{BLSTM}_1(\text{E}(X_i))_j + b_u)}}{\sum_{k=1}^{W} e^{\tanh(W_u \text{BLSTM}_1(\text{E}(X_i))_k + b_u)}} \tag{3}$$

$$R'_i = \sum_{j=1}^{V} \text{BLSTM}_1(\text{E}(X'_i))_j \cdot \beta_{ij} \tag{4}$$

$$\beta_{ij} = \frac{e^{\tanh(W_v \text{BLSTM}_1(\text{E}(X'_i))_j + b_v)}}{\sum_{k=1}^{V} e^{\tanh(W_v \text{BLSTM}_1(\text{E}(X'_i))_k + b_v)}}, \tag{5}$$

where $W_u \in \mathbb{R}^{d_w}$, $W_v \in \mathbb{R}^{d_w}$, are the weights of the attention mechanism for document and summary at word level.

From R and R', $r \in \mathbb{R}^{d_s}$ and $r' \in \mathbb{R}^{d_s}$ can be obtained, following Equations (6) and (8), similarly to the sentence level but using BLSTM$_2$ and the attentions $\hat{\alpha} \in \mathbb{R}^{T}$ and $\hat{\beta} \in \mathbb{R}^{Q}$ for document and summary respectively. The following equations show a document encoder composed by the $\hat{N} = 1$ BLSTM network.

$$r = \sum_{j=1}^{T} \text{BLSTM}_2(R)_j \cdot \hat{\alpha}_j \tag{6}$$

$$\hat{\alpha}_j = \frac{e^{\tanh(W_{\hat{u}} \text{BLSTM}_2(R)_j + b_{\hat{u}})}}{\sum_{k=1}^{T} e^{\tanh(W_{\hat{u}} \text{BLSTM}_2(R)_k + b_{\hat{u}})}} \tag{7}$$

$$r' = \sum_{j=1}^{Q} \text{BLSTM}_2(R')_j \cdot \hat{\beta}_j \tag{8}$$

$$\hat{\beta}_j = \frac{e^{\tanh(W_{\hat{\partial}}\text{BLSTM}_2(R')_j + b_{\hat{\partial}})}}{\sum_{k=1}^{Q} e^{\tanh(W_{\hat{\partial}}\text{BLSTM}_2(R')_k + b_{\hat{\partial}})}}, \tag{9}$$

where $W_{\hat{u}} \in \mathbb{R}^{d_s}$, $W_{\hat{u}} \in \mathbb{R}^{d_s}$, are the weights of the attention mechanism for document and summary at document level.

These vector representations, r and r', capture bidirectional relationships among the sentence representations, which are obtained from the representations of their words. Then, they can be used to distinguish correct summaries for documents by forcing the attention mechanisms of the document branch to focus on the most relevant sentences. In order to do this, the vector representations of the document r, the summary r', and the difference between them $|r - r'|$ are concatenated and used as input to a feed-forward network with one softmax fully-connected layer, as defined in Equation (10), to compute a probability distribution over $\mathbb{C} = \{0, 1\}$.

$$\hat{y} = \text{softmax}(W_{\hat{y}}[r; r'; |r - r'|] + b_{\hat{y}}), \tag{10}$$

where $\hat{y}$ is the output of the classifier, $W_{\hat{y}} \in \mathbb{R}^{3d_s \times 2}$ is the weight matrix of the fully connected layer and $b_{\hat{y}} \in \mathbb{R}^2$ is the bias.

Once the network has been trained to distinguish correct summaries for documents, to carry out document summarization with SHA-NN, the attention mechanisms at document level can be directly used to rank sentences and then, to select the most relevant of them based on this rank. Specifically, for the summarization process, given a document X, a forward pass is performed on the document branch (left branch) of the Siamese network (HAN_1 in Figure 2) to obtain the attention score $\hat{\alpha}_j$ of each document sentence. From the ranking of the document sentences based on those scores, the top-k sentences with higher attention score are selected to build the summary.

5.2. Siamese Hierarchical Transformer Encoders

Siamese Hierarchical Transformer Encoders (SHTE) [20] is the instance of the general attentional framework when the encoders, both for sentence and document levels, are Transformer Encoders (TE) [21] shaped in a hierarchical way, i.e., $g_1(.;\theta_{11}) = \text{TE}_1(.;\theta_1)$, $g'_1(.;\theta_{21}) = \text{TE}_1(.;\theta_1)$, $g_2(.;\theta_{12}) = \text{TE}_2(.;\theta_2)$ y $g'_2(.;\theta_{22}) = \text{TE}_2(.;\theta_2)$. Additionally, in this case, all the weights are shared between the sentence and document levels of the two branches and classifier c is a feed-forward network. The scheme of this architecture can be seen in Figure 3.

The multi-head self-attention mechanism used in the Transformer Encoders is defined in Equations from (11) to (13).

$$\text{MultiHead}(A, B, C) = [\text{head}_1; ...; \text{head}_h]W^O \tag{11}$$

$$\text{head}_i = \text{Attention}(AW_i^Q, BW_i^K, CW_i^V) \tag{12}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\intercal}{\sqrt{d_k}}\right)V, \tag{13}$$

where A, B and C are the inputs of the multi-head attention, h is the number of attention heads, W_i^Q, W_i^K, W_i^V and W_i^O are the projection matrices for Query (Q), Key (K), Value (V) of the head i, and output (O) of the multi-head attention respectively. This mechanism is used both at sentence and document levels. Additionally, it is important to highlight that it does not consider the word order and, due to this fact, it is necessary to incorporate positional information into the system.

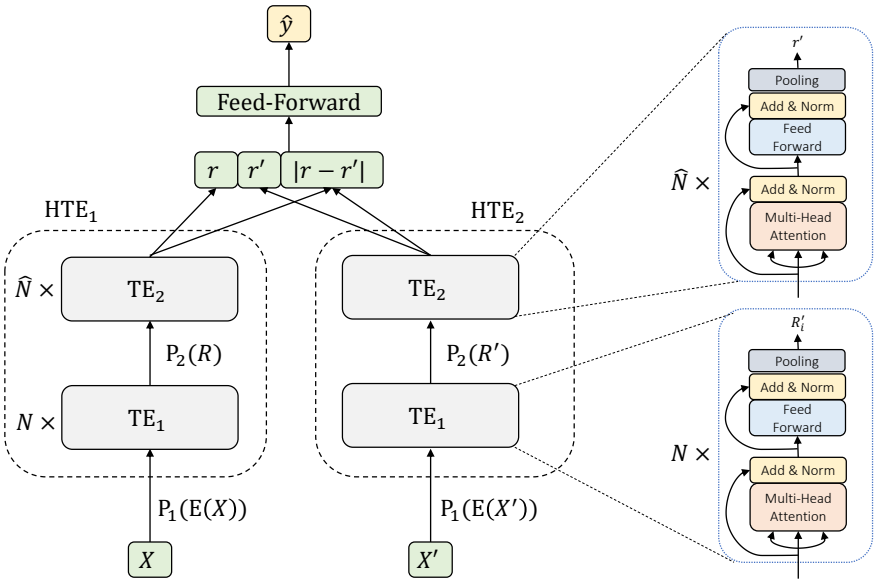

Figure 3. SHTE Architecture.

First, we define a function $\mathrm{P}_1 : \mathbb{R}^{d_e} \rightarrow \mathbb{R}^{d_e}$ that is applied independently to each word to identify its position in the input of the sentence encoder. Thus, from X and X', $R \in \mathbb{R}^{T \times d_w}$ for article and $R' \in \mathbb{R}^{Q \times d_w}$ for summary are computed by using Transformer Encoders as sentence encoders, following Equations (14) and (15).

$$R_i = \frac{1}{W} \sum_{j=1}^{W} \mathrm{TE}_1(\mathrm{P}_1(\mathrm{E}(X_i))) \tag{14}$$

$$R'_i = \frac{1}{V} \sum_{j=1}^{V} \mathrm{TE}_1(\mathrm{P}_1(\mathrm{E}(X'_i))), \tag{15}$$

where the $N = 1$ layered Transformer Encoder $\mathrm{TE}_1(.;\theta_1)$ is defined in Equation (16). Note that, if $N > 1$ Transformer Encoder layers are used, the output of TE_1 in layer i is used as input for TE_1 in layer $i + 1$.

$$\mathrm{TE}_1 = \mathrm{LayerNorm}(L + F) \tag{16}$$

$$F = \max(0, LW_1 + b_1)W_2 + b_2 \tag{17}$$

$$L = \mathrm{LayerNorm}(. + \mathrm{MultiHead}(.,.,.)), \tag{18}$$

where the weight matrices of the multi-head attention mechanism (Equations (11) and (12)) are defined for the sentence level as $W_i^Q \in \mathbb{R}^{d_e \times d_k}$, $W_i^K \in \mathbb{R}^{d_e \times d_k}$, $W_i^V \in \mathbb{R}^{d_e \times d_k}$ and $W_i^O \in \mathbb{R}^{(h \cdot d_k) \times d_w}$ and, additionally, are shared among the two branches; $W_1 \in \mathbb{R}^{d_w \times d_{fw}}$, $W_2 \in \mathbb{R}^{d_{fw} \times d_w}$, $b_1 \in \mathbb{R}^{d_{fw}}$ and $b_2 \in \mathbb{R}^{d_w}$ are the weights and the bias respectively of the position-wise feed-forward network; and LayerNorm refers to Layer Normalization [40]. This process is independently applied to each word embedding matrix that represents each sentence both for document and summary ($R_i : 1 \leq i \leq T$ and $R'_i : 1 \leq i \leq Q$).

From R and R', $r \in \mathbb{R}^{d_s}$ and $r' \in \mathbb{R}^{d_s}$ can be obtained, following Equations (19) and (20), similarly to the sentence level but using TE_2 for document and summary respectively. Note that, due to Transformer Encoders are applied on top of the sentence representations, it is possible to include positional information also to take into account the position of the sentences both in documents and summaries. To do this, a function $\mathrm{P}_2 : \mathbb{R}^{d_w} \rightarrow \mathbb{R}^{d_w}$ is

defined, that is applied to each sentence independently to incorporate sentence positional information in the input of the encoder at document level.

$$r = \frac{1}{T}\sum_{j=1}^{T} \mathrm{TE}_2(\mathrm{P}_2(R)) \tag{19}$$

$$r' = \frac{1}{Q}\sum_{j=1}^{Q} \mathrm{TE}_2(\mathrm{P}_2(R')), \tag{20}$$

where $\mathrm{TE}_2(.;\theta_2)$, composed by $\hat{N} = 1$ layer is defined in the same way that TE_1, following Equation (21). If $\hat{N} > 1$, the output of TE_2 in layer i is used as input for the next layer $i+1$.

$$\mathrm{TE}_2 = \mathrm{LayerNorm}(\hat{L} + \hat{F}) \tag{21}$$

$$\hat{F} = \max(0, \hat{L}\hat{W}_1 + \hat{b}_1)\hat{W}_2 + \hat{b}_2 \tag{22}$$

$$\hat{L} = \mathrm{LayerNorm}(. + \mathrm{MultiHead}(.,.,.)), \tag{23}$$

where the weight matrices of the multi-head attention mechanism (Equations (11) and (12)) are defined for the document level as $W_i^Q \in \mathbb{R}^{d_w \times d_k}$, $W_i^K \in \mathbb{R}^{d_w \times d_k}$, $W_i^V \in \mathbb{R}^{d_w \times d_k}$ and $W_i^O \in \mathbb{R}^{hd_k \times d_s}$, and additionally, are shared among the two branches; $\hat{W}_1 \in \mathbb{R}^{d_s \times d_{fs}}$, $\hat{W}_2 \in \mathbb{R}^{d_{fs} \times d_s}$, $\hat{b}_1 \in \mathbb{R}^{d_{fs}}$ and $\hat{b}_2 \in \mathbb{R}^{d_s}$.

From the vectors r and r', the interaction between them is computed as their concatenation with their absolute difference. This interaction is used as input for a feed-forward network whose output is a probability distribution over $\mathbb{C} = \{0, 1\}$, as defined in Equation (10).

It is interesting to note the main difference of SHTE concerning SHA-NN. In SHA-NN, BLSTM are used to compute the representations, combined with attention mechanisms to average them. Due to the attention mechanism computes directly the impact of each sentence in the final representation, this score can be used directly to rank the sentences. However, in SHTE, the same attention mechanism computes both the representations and the relevance scores. Due to this fact, the relevance of each sentence is implicitly captured by the multi-head self-attention mechanism. This system considers that a document sentence is more relevant the more attended it is by all the sentences of the document. With the aim of building a ranking over the document sentences, we use the attention matrices of the last Transformer Encoder at document level, obtained after a forward pass on the left branch of the network from an input document, following Equations from (24) to (26).

$$G_i = \mathrm{softmax}\left(\frac{Q_i K_i^\intercal}{\sqrt{d_k}}\right) \tag{24}$$

$$H_{ij} = \frac{1}{h}\sum_{k=0}^{h} G_{kij} \tag{25}$$

$$\alpha_j = \frac{1}{T}\sum_{i=0}^{T} H_{ij}, \tag{26}$$

where $Q_i, K_i \in \mathbb{R}^{T \times d_k}$ are the Queries and Keys in head i, $G_i \in \mathbb{R}^{T \times T}$ is the attention matrix of head i, $H \in \mathbb{R}^{T \times T}$ is the averaged attention of all the heads, and $\alpha \in \mathbb{R}^T$ is the vector that contains the final score assigned to each sentence j.

The system is composed of h different attentions that explain different relationships among the sentences. As shown in Equation (25), we consider that all the relationships captured by the self-attention mechanism have the same relevance to obtain a score. For

this reason, the most attended sentences, on average among the different relationships (attentions), are considered as the most relevant, as it was stated in [20].

After computing the average attention of all the heads, H, the component H_{ij} represents the average attention that the model assigns to the sentence j when it is processing the sentence i. Then, it could be used to compute the relevance of a sentence j in the document based on the average attention that j receives of all the sentences of the document, following Equation (26). This process is used to compute the scores for all the sentences, and the scores are used to rank them for selecting the top-k most relevant document sentences in order to compose the summary.

6. Corpora

We carried out the experimentation by using two different corpora for newspaper summarization. On the one hand, the CNN/DailyMail (https://cs.nyu.edu/~kcho/DMQA/ (accessed on 16 January 2023)) corpus was used in this work. This corpus, which is a set of articles from the CNN and DailyMail news websites, was originally constructed for Question Answering [22] and was modified for abstractive and extractive summarization [6,41]. The CNN/DailyMail corpus was partitioned into 287,227 training (article, summary) pairs, 13,368 validation (article, summary) pairs and 11,490 test (article, summary) pairs. In order to compare our systems with most of the works on this corpus, we used the non-anonymized version. It should be noted that the ground truth summaries provided by this corpus are abstractive, and they were constructed by concatenation of the highlights associated with the documents.

On the other hand, the NewsRoom (https://lil.nlp.cornell.edu/newsroom/ (accessed on 16 January 2023)) corpus, proposed in [23] for the summarization task, was also used. It consists of 1.3 million articles and summaries 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. The NewsRoom corpus was partitioned into 995,041 training (article, summary) pairs, 108,837 validation (article, summary) pairs, and 108,862 test (article, summary) pairs.

Some characteristics of both corpora are presented in Table 1. It is important to note that the NewsRoom corpus is much bigger than the CNN/DailyMail corpus as stated before. Regarding the number of article sentences and words in all the sample sets, both corpora are very similar. However, reference summaries (Summ columns) are twice as long in CNN/DailyMail than in NewsRoom.

Table 1. Average number of sentences and words, including words per sentence, for both corpora.

		Sentences		**Words**		**Words/Sentence**	
Corpus	Set	Articles	Summ	Articles	Summ	Articles	Summ
CNN / DailyMail	Train	31.87	3.79	750.10	51.58	23.53	13.61
	Dev	26.77	4.11	737.06	57.57	27.53	14.00
	Test	27.11	3.88	745.59	54.65	27.51	14.07
NewsRoom	Train	29.91	1.40	773.57	30.37	25.86	24.65
	Dev	29.69	1.41	767.34	30.72	25.84	21.73
	Test	29.62	1.41	765.56	30.63	25.84	21.63

7. Experimental Setup

To carry out the experimentation, we maintained most of the hyper-parameters published both for SHA-NN [18] and SHTE [20] systems. All the experiments were performed in a single GPU NVIDIA GeForce RTX 2080.

On the one hand, for the SHA-NN system, we used pre-trained word embeddings, obtained by means of a $d_e = 300$-dimensional skip-gram architecture, trained from the

articles of the corpora. These embeddings were frozen during the training of the models. We used $N = 1$ sentence encoders and $\hat{N} = 1$ document encoders with $d_w = d_s = 512$. Adam [42] was used as update rule with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to optimize the cross-entropy. In order to train the model with both corpora, we used batches of 64 (article, summary) pairs (32 positive and 32 negative randomly sampled following a uniform distribution). To generate the summaries, the top-k most relevant sentences were selected by following directly the attention score of the document encoder.

On the other hand, for the SHTE system, we used randomly initialized word embeddings with $d_e = 128$ which were trained simultaneously with the model. Most of the hyper-parameters were also fixed, such as $N = 2$ word encoders and $\hat{N} = 2$ sentences encoders, $h = 6$ heads, $d_k = d_v = d_q = 64$, $d_w = d_s = d_{fw} = d_{fs} = d_e$, P_1 is the identity function (we do not add positional information to the words inside each sentence) and P_2 is the sine-cosine function defined in [21]. We only used positional information on the sentences due to the empirical results obtained in [20], where positional information in sentences seems to work better than positional information in words. Adam [42] was used as update rule with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to optimize the cross-entropy, and Noam [21] was used as learning rate schedule with $warmup\_steps$ = 4000. To train the model with CNN/DailyMail we used batches of 64 (article, summary) pairs (32 positive and 32 negative randomly sampled following an uniform distribution). For training with NewsRoom, we used batches of 128 (article, summary) pairs. In order to generate the summaries, the top-k most relevant sentences were selected by following the scoring mechanism presented in Section 5.2.

For both systems, we used early stopping with 20 epochs of patience during the training phase. For the summarization phase, both models extracted the $k = 3$ most relevant sentences for the CNN/DailyMail corpus and $k \in \{2, 3\}$ for the NewsRoom corpus.

8. Evaluation

In this section, we show and discuss the results obtained by the systems of the *Attentional Extractive Summarization* framework (SHA-NN and SHTE) on the CNN/DailyMail and NewsRoom corpora. We also performed comparisons with other extractive and mixed systems. (We considered as mixed systems those that combine extractive and abstractive strategies, either end-to-end or decoupled.)In order to reflect the categorization of the models, we show in the tables the category to which each model belongs. Specifically, these categories are five: *Heuristic* (simple rules to generate summaries), *Attentional* (models that fall under our framework for extractive summarization), *Oracle* (models that require a sentence labeling previously to the training phase), *Reinforcement* (models that use Reinforcement Learning to optimize ROUGE metrics during training) and *Text generation* (models that do not use oracles nor reinforcement learning, and are trained for text generation by maximizing the likelihood of each word in the reference summary, given the document and all the previous words in that reference summary). The evaluation of the systems' performance has been carried out by using three variants of the ROUGE measure [17]. Concretely, Rouge-N with unigrams and bigrams (R-1 and R-2) and Rouge-L (R-L) were used. It should be noted that the ROUGE measure is based on ngrams overlapping. Therefore it is adequate when reference and generated summaries are extractive; however, it is no longer as suitable when the reference or the generated summaries are abstractive.

In Table 2, the results of our systems and other state-of-the-art systems for the CNN/DailyMail corpus are shown (ECS and PGen are the acronyms for ExConSumm and Pointer-Generator respectively). Our systems obtain similar results to those of Pointer-Gen+Cov [31], CopyCat [32], and SummaRunner [7]. The obtained results are worse in comparison, despite our systems sharing with it the same backbone architecture (Transformer Encoders). This is possibly due to BertSumEXT starts from a very powerful contextualized pre-trained language model [27]. Additionally, it is interesting to observe that the results obtained by our systems are better than those obtained by some Reinforcement Learning based systems such as DQN [13] and similar to Refresh [10]. Therefore, our extractive

summarization framework could be used as an alternative to Reinforcement Learning approaches and oracle-based systems.

Table 2. Results on CNN/DailyMail corpus for full-length Rouge. The suffix of Lead, SHA-NN, and SHTE models refers to the number of extracted sentences, *k*. Best results are in bold.

System	Strategy	Category	R-1	R-2	R-L
Lead-3	Ext	Heuristic	40.24	17.70	36.45
SHA-NN-3	Ext	Attentional	39.99	17.75	36.27
SHTE-3	Ext	Attentional	39.96	17.60	36.19
SummaRunner	Ext	Oracle	39.60	16.20	35.30
ECS-Ext	Ext	Oracle	41.70	18.60	37.80
BertSumEXT	Ext	Oracle	**43.25**	**20.24**	**39.63**
Refresh	Ext	Reinforcement	40.00	18.20	36.60
DQN	Ext	Reinforcement	39.40	16.10	35.60
Latent	Ext	Reinforcement	41.10	18.80	37.40
BanditSum	Ext	Reinforcement	41.50	18.70	37.60
ECS-Comp	Mix	Oracle	40.90	18.00	37.40
Latent-Comp	Mix	Reinforcement	36.70	15.40	34.30
PGen + Cov	Mix	Text generation	39.53	17.28	36.38
CopyCat	Mix	Text generation	39.15	17.60	36.17

Tables 3 and 4 show the results, in terms of ROUGE, on the NewsRoom corpus. Specifically, Table 3 shows the results on the full test set and Table 4 shows the results on each one of the three test subsets defined by Grusky et al. [23].

Each subset makes reference to the extractiveness degree of their summaries, measured in terms of the density metric proposed in [23]. There are 3 different subsets: NR-Ext (subset whose reference summaries have a high density of words that appear in the articles), NR-Mix (subset with a medium density), and NR-Abs (subset whose reference summaries have a low density and, then, it can be considered as abstractive).

Table 3. Results on the full test of NewsRoom. The suffix of Lead, SHA-NN, and SHTE models refers to the number of extracted sentences, *k*. Best results are in bold.

System	Strategy	Category	R-1	R-2	R-L
Lead-3	Ext	Heuristic	30.66	21.09	28.35
SHA-NN-3	Ext	Attentional	28.99	19.42	26.69
SHTE-3	Ext	Attentional	29.19	19.37	26.81
Lead-2	Ext	Heuristic	33.98	23.30	31.14
SHA-NN-2	Ext	Attentional	32.78	21.86	29.85
SHTE-2	Ext	Attentional	32.38	21.25	29.40
ECS-Ext	Ext	Oracle	**39.50**	**27.90**	**36.26**
PGen + Cov	Mix	Text generation	26.43	13.76	22.90
TLM	Mix	Text generation	33.30	20.06	29.26
FastRL	Mix	Reinforcement	21.93	9.37	19.61
ECS-Comp	Mix	Oracle	39.06	27.36	36.13

Table 4. Results on the three test subsets of NewsRoom (Extractive, Mixed, and Abstractive). The suffix of Lead, SHA-NN, and SHTE models refers to the number of extracted sentences, *k*. Best results are in bold.

	NR-Ext			**NR-Mix**			**NR-Abs**		
System	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Lead-3	51.98	47.85	51.20	25.62	13.00	22.30	14.57	2.62	11.73
SHA-NN-3	48.29	43.54	47.42	24.62	12.32	21.37	14.22	2.57	11.43
SHTE-3	48.62	43.35	47.65	24.76	12.20	21.43	14.33	2.53	11.51
Lead-2	57.87	53.03	56.83	28.60	14.33	24.46	15.68	2.77	12.35
SHA-NN-2	54.83	49.25	53.72	28.03	13.79	23.85	15.67	2.74	12.29
SHTE-2	53.97	47.87	52.59	27.78	13.41	23.56	15.57	2.67	12.22
ECS-Ext	**69.40**	**64.30**	**68.30**	**31.90**	**16.30**	26.90	17.20	3.10	13.60
PGen + Cov	39.10	28.00	36.20	25.50	11.00	21.10	14.70	2.30	11.40
TLM	53.30	44.20	50.10	28.10	12.10	23.00	**18.50**	**3.90**	**14.70**
ECS-Comp	68.40	62.90	67.30	31.70	16.10	**27.00**	17.10	3.10	14.10

It is possible to observe how extracting a number of sentences similar to the reference summary length (1.4 as shown in Table 1) improves notably the performance of the systems ($k = 2$ instead of $k = 3$). This behavior is observed especially in the NR-Ext and NR-Mix subsets, in comparison to the NR-Abs subset. This suggests that, when the reference summaries are extractive, in addition to determine the relevance of each sentence, it is also important to adjust correctly the length of the summaries. However, when the reference summaries are abstractive, the results by using $k = 2$ and $k = 3$ are very similar and clearly lower for all the systems. These bad results are due to the abstractiveness nature of this set of reference summaries, taking into account that the systems are extractive and mixed. Additionally, it is interesting to highlight that, although Lead is a robust baseline in the NR-Ext and NR-Mix subsets, it is not so good in the NR-Abs subset, where our systems obtain almost the same results.

In both cases, the results obtained by our systems are better than those obtained by widely used approaches such as Pointer-Gen+Cov [31] or by Reinforcement Learning systems such as FastRL [14,43]. Additionally, they obtain better results than TLM [44] in terms of ROUGE-2 and ROUGE-L on the full dataset, in spite of this system stands out in the abstractive subset NR-Abs. The only systems that consistently outperform the Lead systems are those based on ExConSumm [30] (both in the extractive and mixed variants), mainly due to they largely outperform the results on NR-Ext and NR-Mix subsets. Differently from our systems, these systems are able to generate variable-length summaries depending on the input text.

In Table 5, several details about the convergence of our systems are shown. Specifically, it shows the number of samples that each system has seen until convergence, the accuracy on the development set (for each sample in the development set, two samples are built, one positive and one negative randomly sampled), and the time until the convergence. It is possible to see how the SHTE model visited a large number of samples during the training until convergence, at the same time that obtains significantly worse results in terms of accuracy. However, the time required to train these models is significantly lower, requiring up to a four times shorter duration than SHA-NN for the NewsRoom corpus. Furthermore, as Tables 2 and 3 show, the results in terms of ROUGE on both corpora are very similar for both systems. Thus, SHTE constitutes an efficient alternative to SHA-NN since, with a lower training time, obtains very similar results in terms of ROUGE. In comparison to other systems such as BanditSum [12] (76 h in a single NVIDIA Geforce Titan Xp), DQN [13] (10 days on a single NVIDIA GeForce GTX 1080) or Refresh [10] (12 h "on a single GPU"), both systems require a significantly lower training time for the CNN/DailyMail corpus. Furthermore, they dispense with the computation of sentence oracles previously to the training step.

Table 5. Convergence statistics of our systems. Best results are in bold.

Corpora	System	Samples	Acc	Time (h)
CNN/DailyMail	SHA-NN	**2,624,000**	**99.62 ± 0.10**	3.51
	SHTE	4,160,000	91.92 ± 0.46	**2.38**
NewsRoom	SHA-NN	**5,088,000**	**96.16 ± 0.11**	6.45
	SHTE	5,760,000	90.61 ± 0.17	**1.65**

It is interesting to observe in Table 5 that, in spite of the significant differences in terms of accuracy during the evaluation with the development set, the results in terms of ROUGE in the evaluation of the test summaries are very similar. This clearly illustrates the mismatch discussed by Narayan et al. in [10], derived from the disconnection between the task definition and the training objective. This is the main drawback of the summarization systems based on optimizing the cross-entropy instead of the ROUGE measure. Due to this reason, it is interesting to search alternatives to Reinforcement Learning in order to optimize directly the evaluation measure.

Finally, despite of the systems based on the attentional framework obtained similar results to Lead in both corpora (mainly due to the bias to the first article sentences in almost all the extractive samples), those systems are capable of generalizing on unseen documents where the sentences are more scattered, as shown in Gonzalez et al. [45].

9. Analysis

Following the experimentation carried out by Mendes et al. in [30], we analyzed the lengths of the summaries generated by our proposals. Figures 4–6 show the word-length distributions of the summaries for Lead, SHA-NN, and SHTE systems (with $k \in \{2, 3\}$) applied on CNN/DailyMail corpus and NR-Ext subset of NewsRoom. We included also the word distribution of the human reference summaries for both corpora.

It can be seen that the word-length distributions of the summaries extracted by our proposals are almost identical to the distribution of the Lead system. This similarity can be observed also in other systems, based on Reinforcement Learning which dispenses of oracles, such as Latent [11] and Refresh [10], as shown in [30]. These results suggest that the extractive systems that do not use oracles are biased towards selecting the first sentences to a higher extent than oracle based systems. For both corpora, all the system distributions are shifted considerably to the right in comparison to the distribution of the human reference summaries. Thus, our systems seem not to be able to generate summaries in lower length ranges (12–50 for CNN/DailyMail, 5–25 for NR-Ext with $k = 2$, and 20–50 for NR-Ext with $k = 3$). This is mainly due to they are not able to build variable-length summaries and they are limited to select all the words of a fixed number of sentences without making word-level operations e.g., compression [30] or selection [31].

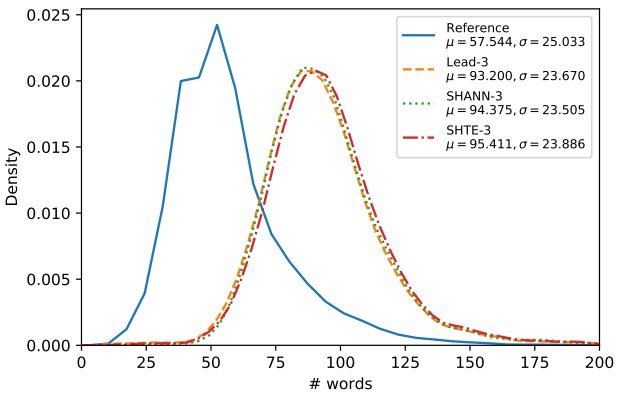

Figure 4. Word-length distribution of system generated summaries in comparison to human reference summaries for CNN/DailyMail.

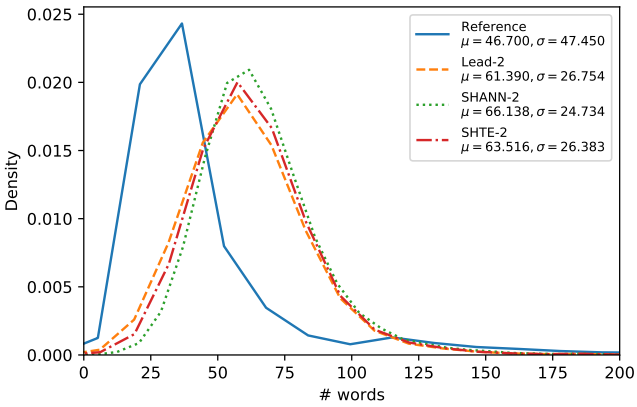

Figure 5. Word-length distribution of system generated summaries with $k = 2$ in comparison to human reference summaries for NR-Ext.

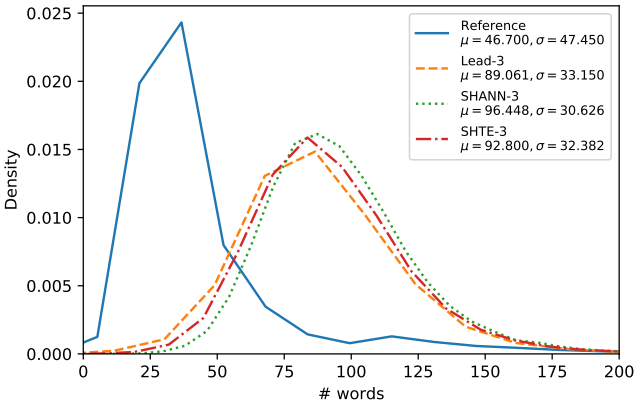

Figure 6. Word-length distribution of system generated summaries with $k = 3$ in comparison to human reference summaries for NR-Ext.

In Figures 7 and 8, we show two examples of summaries generated by the SHTE and SHA-NN systems both for NewsRoom and CNN/DailyMail respectively. In the NewsRoom example, the SHTE model generates a shorter summary than SHA-NN, extracting the first sentence of the document as a short and direct lead. Additionally, it is the only model that makes explicit the name of the analyst (*technical analyst andrew keene*, in the 5th sentence), like in the reference summary. Differently, SHA-NN prefers the sentence that contains the introduction to the analyst's statements (4th sentence), but does not contain its name. Only SHA-NN mentions the *copper* (*red metal*), the *mining* (*mining giant*), and the *end of the stock* (*rally could be over*). It should be noted that, in this example, SHA-NN prefers the longest sentences that appear before in the document (3rd and 4th), differently from SHTE (1st and 5th). For the CNN/DailyMail example, both systems extract the first article sentence. Along with it, SHTE extracts a sentence related to the reference summary and one irrelevant sentence. In the same way, this behavior is also observed in SHA-NN where the related sentence extracted is different from the one extracted by SHTE. If the 3rd sentence extracted by SHTE and the 2nd extracted by SHA-NN were selected, almost all the semantic of the reference summary would be covered. It is also interesting to note that the generated summaries are much longer than the reference summaries, due to our systems being restricted to selecting full article sentences; however, the reference summaries could be composed of simplified sentences.

• **Article**: forget gold and oil . copper prices is the real winner this year . the red metal is up more than 20 percent from its late january low—and that 's given one stock a big boost : freeport-mcmoran . the mining giant is up 40 percent in the same period , but one trader who relies heavily on the technicals and options market , is cautious on the stock , and he warned that the rally could be over . " i think we 're about to see some serious selling pressure in freeport , " said technical analyst andrew keene on cnbc 's " trading nation " on thursday . despite the recent strength , freeport shares are still lower on the year , down roughly 2 percent . what concerns keene most is that on a one-year chart , freeport-mcmoran shares have shown what 's called a " rounding top . " technicians often see this particular pattern as a reversal of a long-term trend . " to me , this looks like a weak stock that is about to roll over and i think it could head back to the low around $ 16.50 , " said keene , founder of keene on the market . keene also noted that fcx will run into resistance at its 150-day moving average and 200-day moving average .

• **Reference**: technical analyst andrew keene explains why the rally in one copper-mining stock is about to end .

• **SHTE**: forget gold and oil . " i think we 're about to see some serious selling pressure in freeport , " said technical analyst andrew keene on cnbc 's " trading nation " on thursday .

• **SHA-NN**: the red metal is up more than 20 percent from its late january low—and that 's given one stock a big boost : freeport-mcmoran . the mining giant is up 40 percent in the same period , but one trader who relies heavily on the technicals and options market , is cautious on the stock , and he warned that the rally could be over .

Figure 7. Summarization of a NewsRoom test sample.

• **Article**: allan donald has confirmed he is to step down as south africa bowling coach . the 48-year-old former test paceman has served his country as part of the coaching team since 2011 . he said : ' i have had some time to reflect after the world cup and have come to the conclusion that the time is right to move on . allan donald has confirmed he is to step down as south africa bowling coach after four years in the role . ' it was always a big dream of mine to work in the south african cricket environment after my playing days , and i was incredibly honoured to be given the opportunity .' the last four years have been the best of my life , and being involved with the proteas in the 2015 cricket world cup was a goal i had set my sights on . 'cricket south africa chief executive haroon lorgat added : ' allan was and always will be a stalwart in south africa , having served his country with distinction both on and off the field . ' he brought great knowledge and international experience to the proteas set-up , and we wish him well in his future endeavours .

• **Reference**: allan donald served as south africa bowling coach since 2011 . donald said ' it was always a big dream ' to work in south african cricket . chief executive haroon lorgat said donald will ' always be a stalwart ' .

• **SHTE**: allan donald has confirmed he is to step down as south africa bowling coach . he said : ' i have had some time to reflect after the world cup and have come to the conclusion that the time is right to move on . ' it was always a big dream of mine to work in the south african cricket environment after my playing days , and i was incredibly honoured to be given the opportunity .

• **SHA-NN**: allan donald has confirmed he is to step down as south africa bowling coach . the 48-year-old former test paceman has served his country as part of the coaching team since 2011 . allan donald has confirmed he is to step down as south africa bowling coach after four years in the role .

Figure 8. Summarization of a CNN/DailyMail test sample.

For the previous examples, in Figure 9 we show the attentions that each system assign to each sentence (the lighter the more relevant is a sentence). In this figure, the first column refers to the systems SHTE and SHA-NN when they are applied on the NewsRoom example, whereas the second column refers to their application on the CNN/DailyMail example. SHTE_H is the averaged matrix shown in Equation (25) for the SHTE system, SHTE_α are the relevance scores assigned to each sentence by the SHTE system following Equation (26), and SHA-NN_α are the relevance scores assigned to each sentence by the SHA-NN system.

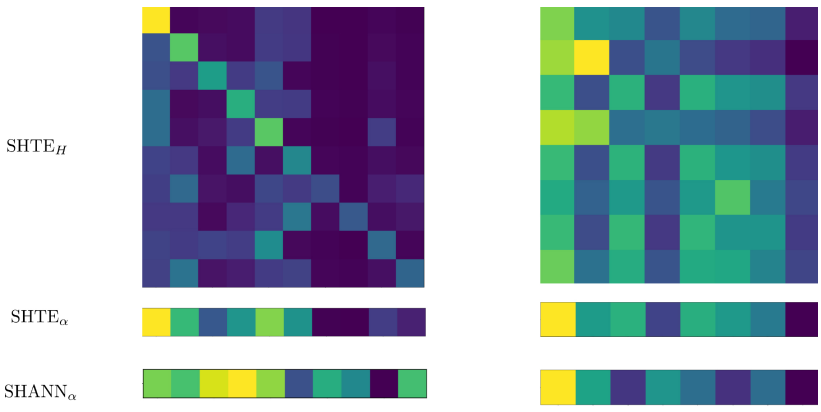

Figure 9. Attentions for the NewsRoom and CNN/DailyMail test examples for both SHTE and SHA-NN.

A bias towards the first sentences can be seen. However, in spite of this bias, both systems are able to also assign high scores to late sentences of the documents. The matrix H of the SHTE system, in the NewsRoom example, is almost a lower triangular matrix, suggesting that the dependencies among the sentences are given only backward. This does not happen in the example of CNN/DailyMail where the attentions seem to compose patterns repeated at regular intervals within the same column.

10. Conclusions and Future Works

In this work, we presented a formalization of a general framework for extractive summarization that does not fall under the umbrella of the traditional extractive systems (based on suboptimal oracles or Reinforcement Learning to optimize the ROUGE). The main objective of this work is to favor the development of new models and techniques within our proposed framework. A future instantiation of this framework could be based on hierarchical BERT-like models, whose attentions could be interpreted to extract the most relevant sentences for the summary.

Under the proposed framework, the summarization systems are based on Siamese architectures to learn directly relationships among articles and summaries. Additionally, they are based on the interpretability of the attention mechanisms, to select the most relevant article sentences. For this reason, we referred to our extractive summarization framework as *Attentional Extractive Summarization*.

We have performed an extensive evaluation and several analyses of the systems in comparison to other Deep Learning extractive and mixed systems, both for the CNN/DailyMail and for the NewsRoom corpora. The obtained results are very promising and they suggest that there is still room for improvement in our attentional framework. This encourages us to continue with the research of this kind of systems.

As future work, several lines of research are open: the extraction of variable-length summaries, the use of the word attentions in order to perform post-process on the extracted sentences, and the inclusion of some abstractive mechanisms on top of the proposed extractive systems. Due to the similarity between the classification and summarization objectives, in the sense that they look for relevant segments of a text, it could be very interesting to study a strategy to approach a text classification system based on the output of a summarization system that provides the selected sentences.

Author Contributions: Conceptualization, L.-F.H. and E.S. (Encarna Segarra); Data curation, J.Á.G., L.-F.H., E.S. (Encarna Segarra) and F.G.-G.; Formal analysis, J.Á.G., L.-F.H., F.G.-G. and E.S. (Emilio Sanchis); Funding acquisition, L.-F.H.; Project administration, L.-F.H.; Investigation, J.Á.G., E.S. (Encarna Segarra), F.G.-G. and E.S. (Emilio Sanchis); Methodology, E.S. (Encarna Segarra) and F.G.-G.; Software, J.Á.G.; Validation, E.S. (Emilio Sanchis); Writing—original draft, J.Á.G., E.S. (Encarna Segarra), E.S. (Emilio Sanchis), F.G.-G.; Writing—review & editing, J.Á.G., L.-F.H., E.S. (Encarna Segarra), F.G.-G. and E.S. (Emilio Sanchis). All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR" under grants PDC2021-120846-C44 (AMIC-PoC-UPV) and PID2021-126061OB-C41 (BEWORD-UPV).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Erkan, G.; Radev, D.R. LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.* **2004**, *22*, 457–479. [CrossRef]
2. Mihalcea, R.; Tarau, P. TextRank: Bringing Order into Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 404–411.
3. Amancio, D.R.; Nunes, M.G.V.; Oliveira, O.N., Jr.; Costa, L.d.F. Extractive summarization using complex networks and syntactic dependency. *Phys. A Stat. Mech. Its Appl.* **2012**, *391*, 1855–1864. [CrossRef]
4. Ferreira, R.; Freitas, F.; de Souza Cabral, L.; Lins, R.D.; Lima, R.; França, G.; Simskez, S.J.; Favaro, L. A four dimension graph model for automatic text summarization. In Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) IEEE, Atlanta, GA, USA, 17–20 November 2013; Volume 1, pp. 389–396.
5. Tohalino, J.V.; Amancio, D.R. Extractive multi-document summarization using multilayer networks. *Phys. A Stat. Mech. Its Appl.* **2018**, *503*, 526–539. [CrossRef]
6. Cheng, J.; Lapata, M. Neural summarization by extracting sentences and words. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Association for Computational Linguistics: Berlin, Germany, 2016; Long Papers; Volume 1, pp. 484–494.
7. 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, AAAI'17, San Francisco, CA, USA, 4–9 February 2017; AAAI Press: Washington, DC, USA, 2017; pp. 3075–3081.
8. 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; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 3730–3740.
9. 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, Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6197–6208.
10. Narayan, S.; Cohen, S.B.; Lapata, M. 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, New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: New Orleans, LA, USA, 2018; Long Papers; Volume 1, pp. 1747–1759.
11. Zhang, X.; Lapata, M.; Wei, F.; Zhou, M. Neural latent extractive document summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 779–784.
12. Dong, Y.; Shen, Y.; Crawford, E.; van Hoof, H.; Cheung, J.C.K. BanditSum: Extractive summarization as a contextual bandit. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 3739–3748.
13. Yao, K.; Zhang, L.; Luo, T.; Wu, Y. Deep reinforcement learning for extractive document summarization. *Neurocomputing* **2018**, *284*, 52–62. [CrossRef]
14. Chen, Y.C.; Bansal, M. Fast abstractive summarization with reinforce-selected sentence rewriting. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Melbourne, Australia, 2018; Long Papers; Volume 1, pp. 675–686.
15. Joshi, A.; Fidalgo, E.; Alegre, E.; Fernández-Robles, L. Summcoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. *Expert Syst. Appl.* **2019**, *129*, 200–215. [CrossRef]
16. Mohd, M.; Jan R.; Shah, M. Text document summarization using word embedding. *Expert Syst. Appl.* **2020**, *143*, 112958. [CrossRef]
17. 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.
18. González, J.Á.; Segarra, E.; García-Granada, F.; Sanchis, E.; Hurtado, L.F. Siamese hierarchical attention networks for extractive summarization. *J. Intell. Fuzzy Syst.* **2019**, *36*, 4599–4607. [CrossRef]
19. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; Association for Computational Linguistics: San Diego, CA, USA, 2016; pp. 1480–1489.
20. González, J.Á.; Segarra, E.; García-Granada, F.; Sanchis, E.; Hurtado, L.F. Extractive summarization using Siamese hierarchical transformer encoders. *J. Intell. Fuzzy Syst.* **2020**, *39*, 2409–2419. [CrossRef]
21. 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, NIPS'17, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010.
22. Hermann, K.M.; Kočiský, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; Blunsom, P. Teaching machines to read and comprehend. In Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS'15, Cambridge, MA, USA, 7–12 December 2015; MIT Press: Cambridge, MA, USA, 2015; Volume 1, pp. 1693–1701.

23. 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, New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: New Orleans, LA, USA, 2018; Long Papers; Volume 1, pp. 708–719.
24. González, J.Á.; Segarra, E.; García-Granada, F.; Sanchis, E.; Hurtado, L.F. An Attentional Extractive Summarization Framework. In Proceedings of the IberSPEECH, Granada, Spain, 14–16 November 2022; pp. 106–110. [CrossRef]
25. Durrett, G.; Berg-Kirkpatrick, T.; Klein, D. Learning-based single-document summarization with compression and anaphoricity constraints. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Association for Computational Linguistics: Berlin, Germany, 2016; Long Papers; Volume 1, pp. 1998–2008.
26. Narayan, S.; Cohen, S.B.; Lapata, M. 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, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 1797–1807.
27. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; Long and Short Papers; Volume 1, pp. 4171–4186.
28. Paulus, R.; Xiong, C.; Socher, R. A deep reinforced model for abstractive summarization. In Proceedings of the 6th International Conference on Learning Representations ICLR, Vancouver, BC, Canada, 30 April–3 May 2018; OpenReview.net: Vancouver, BC, Canada, 2018; pp. 1–13.
29. Williams, R.J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **1992**, *8*, 229–256. [CrossRef]
30. Mendes, A.; Narayan, S.; Miranda, S.; Marinho, Z.; Martins, A.F.T.; Cohen, S.B. 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, 2–7 June 2019; Long and Short Papers; Volume 1, pp. 3955–3966.
31. 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; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; Long Papers; Volume 1, pp. 1073–1083.
32. Ive, J.; Madhyastha, P.; Specia, L. 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), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 3225–3234.
33. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
34. Wiegreffe, S.; Pinter, Y. Attention is not not explanation. 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; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 11–20.
35. Xiao, W.; Carenini, G. Extractive summarization of long documents by combining global and local context. 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; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 3009–3019.
36. Cao, Z.; Chen, C.; Li, W.; Li, S.; Wei, F.; Zhou, M. Tgsum: Build tweet guided multi-document summarization dataset. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 2906–2912.
37. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, 7–9 May 2015; pp. 1–15.
38. Luong, T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 1412–1421.
39. Schuster, M.; Paliwal, K. Bidirectional recurrent neural networks. *Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
40. Ba, L.J.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
41. Nallapati, R.; Zhou, B.; dos Santos, C.; Gulcehre, C.; 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; Association for Computational Linguistics: Berlin, Germany, 2016; pp. 280–290.
42. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015; pp. 1–15.
43. Keneshloo, Y.; Ramakrishnan, N.; Reddy, C.K. Deep transfer reinforcement learning for text summarization. In Proceedings of the 2019 SIAM International Conference on Data Mining (SDM), Calgary, AB, Canada, 2–4 May 2019; pp. 675–683.

44. Pilault, J.; Li, R.; Subramanian, S.; Pal, C. On extractive and abstractive neural document summarization with transformer language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 16–20 November 2020; pp. 9308–9319.
45. González, J.Á.; Hurtado, L.F.; Segarra, E.; García-Granada, F.; Sanchis, E. Summarization of spanish talk shows with Siamese hierarchical attention networks. *Appl. Sci.* **2019**, *9*, 3836. [CrossRef]