applied sciences

MDPI

Article

Summarization of Spanish Talk Shows with Siamese Hierarchical Attention Networks

J.-A. González, L.-F. Hurtado, E. Segarra *, F. García-Granada and E. Sanchis

VRAIN: Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Camino de Vera s/n, 46022 València, Spain

* Correspondence: esegarra@dsic.upv.es

Received: 12 July 2019; Accepted: 9 September 2019; Published: 12 September 2019

Abstract: In this paper, we present an approach to Spanish talk shows summarization. Our approach is based on the use of Siamese Neural Networks on the transcription of the show audios. Specifically, we propose to use Hierarchical Attention Networks to select the most relevant sentences for each speaker about a given topic in the show, in order to summarize his opinion about the topic. We train these networks in a siamese way to determine whether a summary is appropriate or not. Previous evaluation of this approach on summarization task of English newspapers achieved performances similar to other state-of-the-art systems. In the absence of enough transcribed or recognized speech data to train our system for talk show summarization in Spanish, we acquire a large corpus of document-summary pairs from Spanish newspapers and we use it to train our system. We choose this newspapers domain due to its high similarity with the topics addressed in talk shows. A preliminary evaluation of our summarization system on Spanish TV programs shows the adequacy of the proposal.

Keywords: siamese hierarchical attention neural networks; extractive summarization; spanish talk shows summarization

1. Introduction

Nowadays, the development of automatic summarization systems is an important issue due to the great amount of information in different formats that is accessible in the web or in other repositories. Some summarization systems are based on unsupervised learning approaches by considering statistical features of words [1], topic modeling such as Latent Semantic Analysis [2], graph based approaches such as LexRank [3,4] (for a more exhaustive review see [5,6]). There are also systems based on supervised learning techniques such as Conditional Random Fields [7], Support Vector Machines [8] and Neural Networks [9–15].

Although most of the works are focused on the application of automatic summarization techniques to collections of purely textual documents, summarization systems are not limited to text input tasks. There are some other works that address the problem of adapting these techniques to audio recordings as input, typically broadcast news, lectures or meetings [16–18]. These systems have to tackle with specific problems derived from the errors generated by the speech recognition phase such as misrecognized words and errors in punctuation marks.

Just as the volume of textual documents available on the web has grown dramatically in recent years, the same is true in the case of TV programs collections. The television channels make available to the public the programs of their own production, generating with it large collections of videos. For the audience who could not follow the broadcast of the live programs, it is interesting the possibility of accessing them. Therefore, in addition to an adequate information retrieval system to perform the search in the programs collections, an automatic summarization system applied to these programs

will be helpful for searching information. This is especially interesting for talk shows, which consist of several speakers giving opinions on various topics introduced by the program's presenter.

The application of supervised methods to automatic summarization, as those based on Neural Networks, implies the availability of adequate corpora consisting of a set of document-summary pairs. The construction of large and high quality corpora for this purpose is not an easy task, because it is necessary a great human effort to generate thousands of manual summaries, or to design new approaches to obtain these summaries in a semiautomatic way. The first important resource for learning corpus-based summarization models is the CNN/DailyMail summarization corpus, originally constructed by Reference [19] for the task of passage-based question answering and adapted to the document summarization task [9,10]. It consists of news stories from CNN and DailyMail and contains 312,077 document-summary pairs.

It have been developed appropriate corpora for English, however this is not the same for other languages, such as Spanish. With the aim of building a corpus for Spanish, a strategy similar to the proposed in Reference [20] for the construction of the NEWSROOM corpus has been followed in this work. In Reference [20], they take advantage of the highlights or summaries, provided by authors or editors in the newsroom, in order to obtain the summaries. A crawling process on the newspaper websites extracts articles and summaries in a straightforward way. The NEWSROOM corpus contains news, sports, entertainment, financial and other kind of publications from several English newspapers.

In this work, we have built a corpus of Spanish newspapers—the ES-NEWS corpus. It consists of a set of 277,675 article-summary pairs extracted from 11 different Spanish newspapers. ES-NEWS corpus contains articles and summaries of news, sports, politics, culture and other topics. The use of this corpus in this work is two-fold—on the one hand, we evaluate our summarization system [15] with it in order to study the transferability of our system from English to Spanish and on the other hand, we use it to train the system that we apply to the summarization of Spanish talk shows.

Another contribution of this work is the study of the transferability of our summarization system to the domain of talk shows in Spanish. In this case, the documents are fragments of audio transcriptions of TV programs and the summaries consist of sentences written manually. Since we do not have a sufficiently large corpus of TV programs to train our summarization system for the talk shows, we use the ES-NEWS corpus. It should be noted that the characteristics of the corpus used for training are different from those of the TV programs. In the case of newspaper articles, there are some sentences that usually appear at the beginning of the article, that contain the main ideas or the relevant information underlying the article. However, the TV programs are conversations where the ideas and information are more scattered in the speaker turns. Even so, both of them expose content of similar topics.

In a previous work [15], we proposed a supervised approach to text summarization which is based on Siamese Hierarchical Attention Neural Networks using distributed vector representation of words, the SHA-NN system. Siamese Neural Networks are capable of learning from positive and negative samples. During the training phase, we provide the network with positive and negative document-summary pairs; a positive pair is a document and its summary and a negative pair is a document and a summary of other different document randomly extracted from the training set. Furthermore, this model is enriched with an attention mechanism that provides the final score associated to each sentence of the input document, which allows us to establish a ranking and to select the most salient sentences to build the summary. We performed some experiments on the CNN/DailyMail corpus that confirmed the good behaviour of our approach to the summarization problem.

In this work, we address the application of the SHA-NN system to summarize TV programs, in particular, Spanish TV talk shows. First, in order to train our summarization system, the ES-NEWS corpus was built. Second, the SHA-NN system was evaluated on this text corpus. Third, a test corpus has been built with talk shows, the LN24-SUMM corpus. Finally, the summarization system trained with the ES-NEWS corpus has been applied to the LN24-SUMM test corpus. We present a preliminary

evaluation of our summarization system on the transcribed speech of the LN24-SUMM test corpus. Despite the different characteristics between the two corpora, the results of transferability between domains are promising.

2. Corpora

In this section we present the characteristics of the two corpora built for this summarization work, one of them for training purposes (ES-NEWS) and the other one for evaluation (LN24-SUMM).

2.1. ES-NEWS Corpus

We have built a corpus of written Spanish article-summary pairs—the ES-NEWS corpus. It consists of a set of 277,675 article-summary pairs extracted from 11 different Spanish online newspapers. ES-NEWS contains articles and summaries of news, sports, politics, culture and other subjects. We used this corpus to evaluate the transferability of the SHA-NN summarization system from English to Spanish and on the other hand, we used it to train the system that we apply to Spanish talk shows summarization.

The ES-NEWS corpus is composed by newspaper articles extracted from around 1 million URLs, which were collected during the last week of June 2018. To enforce the diversity of summarization styles, 11 websites of relevant newspapers of Spain have been used. These newspapers are: *Elconfidencial*, *EconomiaDigital*, *HuffingtonPost*, *ABC*, *FormulaTV*, *EldiarioCantabria*, *Publico*, *Vozpopuli*, *Rioja2*, *PeriodistaDigital* and *Eldiario*. It have been excluded newspapers that do not include highlights such as *ElMundo*, newspapers that only consider a single short highlight per article such as *ElPais* and newspapers whose crawled content are not articles but web content (advertising, keywords, etc.) such as *EuropaPress*.

Once all the URLs were crawled, following Reference [20], we have used the field *og:description* to extract the highlights that, concatenated, were considered as summaries. We made a preprocess in order to remove noise such as duplicated URLs, empty summaries and articles and non-journalistic articles. All the text was lowercased and tokenized by using the Spanish version of Stanford CoreNLP.

The corpus consists of 277,675 article-summary pairs. From it, training, development and test partitions have been defined, following similar proportions to CNN/Dailymail corpus [11] (90%, 5.5% and 4.5% respectively). Thus, resulting in a training set of 249,919 pairs, a development set of 15,266 pairs and a test set of 12,490 pairs. Some ES-NEWS corpus statistics are shown in Table 1.

Table 1. ES-NEWS corpus statistics.

Dataset Size	277,675 pairs
Mean Article Length	813.8 words
Mean Summary Length	46.0 words
Mean Word Overlapping	28.6 words
Mean Extractive Fragment Coverage	0.67
Mean Extractive Fragment Density	7.27
Mean Compression Ratio	20:1
Articles Vocabulary Size	961,485 words
Summaries Vocabulary Size	167,822 words
Overlapping Vocabulary	157,863 words

To make a comparison between the ES-NEWS and the NEWSROOM corpora, we have used the Extractive Fragment Coverage, Extractive Fragment Density and Compression Ratio measures. These metrics, proposed in Reference [20], aim to measure the overlapping between summaries and articles to analyze the diversity of summarization styles. The metrics are defined in Equations (1)–(3), where A is the sequence of words of the article, S is the sequence of words of the summary, $\mathcal{F}$ is the set of

common fragments, common sequences of words, in A and S computed using a greedy algorithm proposed in Reference [20] and $|\cdot|$ stands for the length, in terms of words, of the sequences.

$$\text{COVERAGE}(A, S) = \frac{1}{|S|} \sum_{f \in \mathcal{F}} |f| \tag{1}$$

$$\text{DENSITY}(A, S) = \frac{1}{|S|} \sum_{f \in \mathcal{F}} |f|^2 \tag{2}$$

$$\text{COMPRESSION}(A, S) = \frac{|A|}{|S|} \tag{3}$$

The Extractive Fragment Coverage is computed as the sum of the lengths of all the common fragments between the article and its summary divided by the size of the summary. Thus, the greater its value, the more common fragments or larger common fragments have been found between the article and its summary.

The Extractive Fragment Density is a measure similar to the Extractive Fragment Coverage but using the square of common fragment lengths. Therefore, with the same number of common words, summaries with longer common fragments obtain higher values than those with more but shorter common fragments.

Figure 1 shows the density and coverage distributions along with the compression ratio for each newspaper in ES-NEWS corpus. Each box is a normalized bivariate density plot of Extractive Fragment Coverage (x-axis) and Extractive Fragment Density (y-axis) of a newspaper. Furthermore, the final distributions on full ES-NEWS is shown in the bottom-right box. As Figure 1 shows, the distribution of Extractive Fragment Density and Extractive Fragment Coverage of ES-NEWS is lead by *ElDiario* (coverage between 0.6 and 0.8 with a high density), due to that newspaper brings the largest number of articles to the corpus. Despite this, generally, the mean coverages and densities in the newspapers show that the introduction of new words in summaries and the use of long extractive fragments is moderate, although both are higher than in NEWSROOM corpus [20].

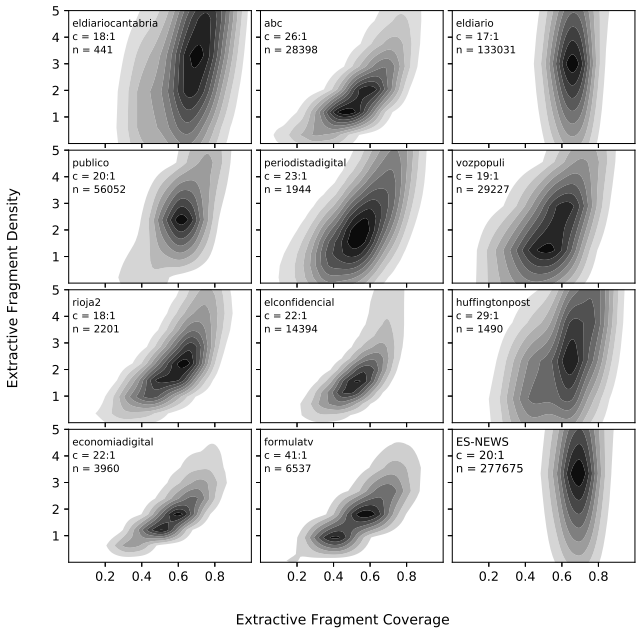

Figure 1. Extractive Fragment Density and Extractive Fragment Coverage distributions on ES-NEWS corpus, where c is the Mean Compression Ratio and n is the number of article-summary pairs.

2.2. LN24-SUMM Corpus

We have built a corpus for Spanish talk shows summarization—the LN24-SUMM corpus. It consists of a set of 30 document-summary pairs. Documents of this corpus are extracted from 5 talk shows of "La Noche en 24 horas", a program of the Spanish television (TVE). Documents have been obtained from the transcriptions of these TV programs, which have been manually segmented into pieces first from the Twitter hashtags appearing in the program videos, and second, from the interventions of the different speakers. This segmentation was made with the aim of summarize the opinion of a speaker about a given topic (hashtag). Four members of the research group generated the reference summaries. A common strategy based on paraphrasing the most representative sentences of each document was used. Consequently, the generated summaries, although they are abstractive, have very high Extractive Fragment Coverage and Density values, as it can be seen in Table 2. In this table some additional LN24-SUMM statistics are also shown.

It is interesting to see that the Mean Compression Ratio is lower in LN24-SUMM than in ES-NEWS corpus. Also, it can be seen that LN24-SUMM corpus has a very high mean Extractive Fragment Density and Coverage in comparison to ES-NEWS corpus. The differences could be because the newspaper summaries are made by many different journalists who are qualified in compressing information and in generating more diverse kind of summaries. In addition, extracting headlines from newspaper articles is simpler than from speaker interventions of talk shows. In the case of newspaper articles there are some sentences, that mainly appear at the beginning of the article, that contain the main ideas or information of the article. However, the talk shows are conversations where the relevant information is more scattered in the speaker turns and exhibits spontaneous speech phenomena, Therefore, they are more difficult to summarize. Additionally, the LN24-SUMM documents, that is, the transcribed and manually segmented talk shows, are very heterogeneous. Two examples to see the differences between both corpus are shown in Figure 2. In this Figure it is possible to see that the LN24-SUMM summaries are composed by more scattered sentences than the summaries of ES-NEWS.

Reference Summary (ES-NEWS): One of the best scientific minds in the world suffered from amyotrophic lateral sclerosis and lived 53 years longer than the doctors diagnosed him **(1/13)**. "Their courage and persistence with their brilliance and their humor inspired people all over the world," their children say in a statement **(4/13)**.

Reference Summary (LN24-SUMM): More than 72 h have passed since the attack on Friday **(3/28)**. The city suffered a heinous attack on several fronts, leisure centers, in football and in a concert hall **(11/28)**. Terrorists have attacked our way of life, even children are accompanied by security forces **(16/28)**. Francois Hollande is going to meet in Paris with John Kerry **(18/28)**. The five terrorists have been identified, four were French and one would have a Syrian passport **(23/28)**.

Figure 2. Examples of summaries from ES-NEWS and LN24-SUMM corpora translated from Spanish. The position of the most related sentence in the document, for each sentence of the summary, is highlighted in bold at the end of the sentences.

Table 2. LN24-SUMM corpus statistics.

Dataset Size	30 pairs
Mean Article Length	921.7 words
Mean Summary Length	108.6 words
Mean Word Overlapping	64.8 words
Mean Extractive Fragment Coverage	0.90
Mean Extractive Fragment Density	19.31
Mean Compression Ratio	8:1
Articles Vocabulary Size	3969 words
Summaries Vocabulary Size	1103 words
Overlapping Vocabulary	1069 words

3. System Description

The SHA-NN system [15] is based on Hierarchical Attention Networks (HAN) [21] trained in a Siamese way, where its left branch extracts representations for whole documents and its right branch extracts representations for summaries. HAN allows us to extract a vector representation for documents and summaries from the representations of their sentences. Moreover, the representation of each sentence is obtained from representations of their words. These representations are trained to address a binary classification task that consists of determining if a summary S is correct for a document D. It acts as an intermediate task in order to extract the most relevant sentences of the document to make a summary. Figure 3 shows an outline of the system architecture.

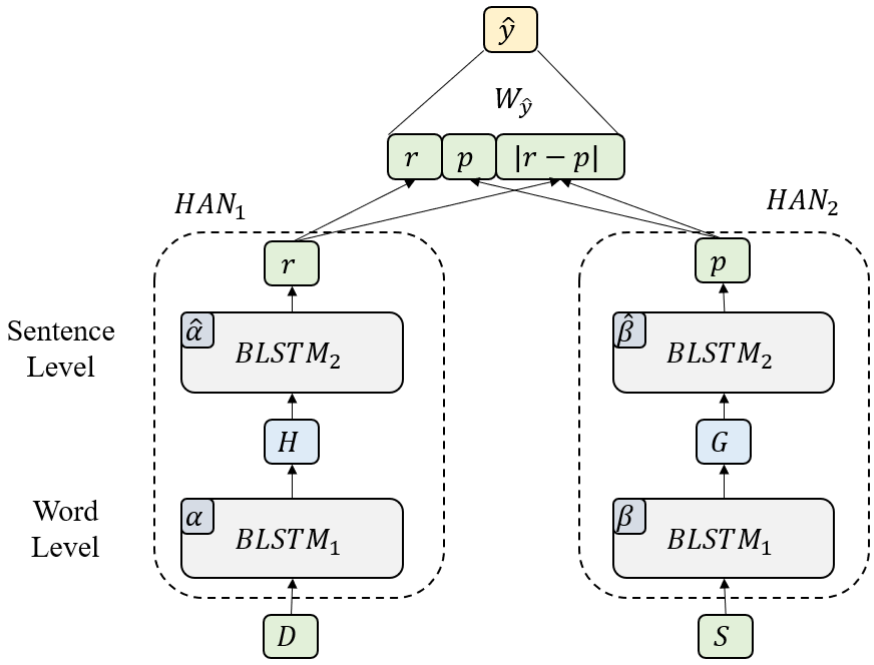

Figure 3. SHA-NN Architecture.

As input, the SHA-NN system uses d_e-dimensional skipgram word embeddings, trained from the training set of the ES-NEWS corpus, in order to represent both documents, $D \in \mathbb{R}^{T \times W \times d_e}$ and summaries, $S \in \mathbb{R}^{Q \times V \times d_e}$, where T and Q are the maximum number of sentences for document and summary and W and V are the maximum number of words per sentence for document and summary. These representations are used as input for the two Hierarchical Attention Networks (HAN_1 and HAN_2) whose BLSTM layers are shared between them, both at sentence level

($BLSTM_2$ with dimensionality $d_s = 512$) and at word level ($BLSTM_1$ with dimensionality $d_w = 512$). However, the attention mechanisms in both levels are not shared.

From these inputs, $H \in \mathbb{R}^{T \times d_w}$ and $G \in \mathbb{R}^{T \times d_w}$ are computed, following Equations (4) and (6), as proposed in [22]. They are the output from the word 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 (5) and $\beta \in \mathbb{R}^{Q \times V}$ Equation (7) for document and summary respectively.

$$H_i = \sum_{j=0}^{W} BLSTM_1(D_{i1}, ..., D_{iW})_j \cdot \alpha_{ij} \tag{4}$$

$$\alpha_{ij} = \frac{e^{tanh(W_u BLSTM_1(D_{i1},...,D_{iW})_j + b_u)}}{\sum_{k=0}^{W} tanh(W_u BLSTM_1(D_{i1}, ..., D_{iW})_k + b_u)} \tag{5}$$

$$G_i = \sum_{j=0}^{V} BLSTM_1(S_{i1}, ..., S_{iV})_j \cdot \beta_{ij} \tag{6}$$

$$\beta_{ij} = \frac{e^{tanh(W_v BLSTM_1(S_{i1},...,S_{iV})_j + b_v)}}{\sum_{k=0}^{V} tanh(W_v BLSTM_1(S_{i1}, ..., S_{iV})_k + b_v)} \tag{7}$$

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. Once H and G are computed, $r \in \mathbb{R}^{d_s}$ and $p \in \mathbb{R}^{d_s}$ can be obtained, following Equations (8) and (10), similarly to the word 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.

$$r = \sum_{j=0}^{T} BLSTM_2(H_1, ..., H_T)_j \cdot \hat{\alpha}_j \tag{8}$$

$$\hat{\alpha}_j = \frac{e^{tanh(W_{\hat{u}} BLSTM_2(H_1,...,H_T)_j + b_{\hat{u}})}}{\sum_{k=0}^{T} tanh(W_{\hat{u}} BLSTM_2(H_1, ..., H_T)_k + b_{\hat{u}})} \tag{9}$$

$$p = \sum_{j=0}^{Q} BLSTM_2(G_1, ..., G_Q)_j \cdot \hat{\beta}_j \tag{10}$$

$$\hat{\beta}_j = \frac{e^{tanh(W_{\hat{v}} BLSTM_2(G_1,...,G_Q)_j + b_{\hat{v}})}}{\sum_{k=0}^{Q} tanh(W_{\hat{v}} BLSTM_2(G_1, ..., G_Q)_k + b_{\hat{v}})} \tag{11}$$

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 sentence level. These vector representations r and p, captures 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 which forces the attention mechanisms to focus on the most similar sentences of both. In order to do this, the vector representations of the document r, the summary p and the difference between them $|r - p|$ are concatenated to feed a fully-connected output layer with softmax activation function [23], as defined in Equation (12).

$$\hat{y} = softmax(W_{\hat{y}}[r; p; |r - p|] + b_{\hat{y}}) \tag{12}$$

In order to train the model, for each document we built positive pairs (D_j, S_j), provided by the ES-NEWS corpus and negative pairs $(D_j, S_k) : j \neq k$ where S_k is chosen randomly from the summaries of the remaining documents. For the positive pairs, the ground truth was $y_i = 1$ whereas for the negative pairs, the ground truth was $y_i = 0$. In this work, we used batches of 64 document-summary pairs (32 positive pairs and 32 negative pairs). The training objective consisted in minimizing the

cross-entropy between the prediction $\hat{y}$ and the class y. The model was trained for 20 epochs of 1500 batches on a GPU Titan X during 3 h.

To carry out document summarization with SHA-NN, once the network has been trained to distinguish correct summaries for documents, the attention mechanisms at sentence level can be used to rank sentences and then, to select the most relevant of them based on this rank. That is, for the summarization process, given a document D, a forward pass is performed on the left branch of the siamese network (HAN_1 in Figure 3) to obtain the attention score $\hat{\alpha}_j$ of each document sentence. From the ranking of the document sentences based on those scores, the $k = 3$ sentences with higher attention score are selected to build the summary.

4. Related Work

Many Neural Network based approaches for text summarization have been proposed in the last few years. Most of them are based on encoder-decoder architectures. Generally, in these approaches, the encoder reads the source sequence as a list of continuous-space representations from which the decoder generates the target sequence. Some of these approaches also incorporate attention mechanisms. In particular, Reference [9] proposed an attentional encoder-decoder approach for extractive single-document summarization and Reference [11] presented an extractive summarization approach based on sentence classification using Neural Networks and required a previous adaptation of the corpus based on ROUGE. In both works, they used the CNN/DailyMail corpus since its large size makes it attractive for training deep Neural Networks. Other Neural Network based works try to improve the behavior of the models incorporating more information. This is the case of Reference [24] where they jointly learn the attention mechanism, to obtain the score of the sentences and the selection mechanism to extract the most salient sentences. One recent trend consists in addressing the summarization problem as a sentence ranking task by considering Reinforcement Learning. This is typically done with the aim of optimizing discrete metrics that are not differentiable, such as the ROUGE evaluation metric [25,26].

There have been some attempts to address the problem of summarization of speech. Most of them are based on techniques successfully used for text summarization and they are directly applied to the output of the speech recognition process. Reference [16] presented an approach to this problem where salient sentences or segments, are extracted only using the textual information, by means of concatenation and reordering mecanisms the final summaries are generated. Experiments are performed on monologues such as lectures, presentations and news commentaries. In Reference [17], an extractive summarization approach also based on the textual representation of the audio is also presented. Salient sentences are extracted based on Language Model measures. Experiments are performed on the Mandarin broadcast news corpus MATBN [27] which was manually segmented and transcribed for evaluation purposes. It must be noted that this kind of corpus has a more regular structure than other speech programs as interviews or debates. Reference [18] also addresses the speech summarization problem but in this case, by using Convolutional Neural Networks for sentence selection. The system includes two convolutional networks, one of them working on the document and the other one working on a sentence of the same document. The system learns for each document sentence a score that represents its probability of belonging to the summary. They also use the MATBN corpus for evaluation purposes.

The SHA-NN system [15] is based on addressing a binary classification problem in order to select the most relevant sentences by means of the attention mechanisms. This system, differently from some Neural Networks based mentioned works, does not require the preparation of the corpus [11], being the system which learns the alignment between document and summary. Moreover, our system addresses the problem as a binary classification task in order to distinguish correct summaries for documents, instead of performing sentence classification to score the document sentences [9,11,18].

5. Experiments

We carried out two different experiments—first, we trained and evaluated the SHA-NN system with ES-NEWS corpus and second, we evaluated the trained system with the LN24-SUMM test corpus.

In order to evaluate our proposal, we performed a experimental comparison with 5 extractive unsupervised summarization systems. Concretely, they are Lead [11], LexRank [3], TextRank [4], Latent Semantic Analysis (LSA) [28] and SumBasic [29]. A short description of each system is shown below.

- *Lead* is a very popular and robust strategy to generate snippets and summaries of article newspapers that consists in extracting the first k sentences of the documents. This strategy is typically used as a baseline in the automatic summarization of newspaper articles, since in the writing style of this type of documents the most relevant information is usually condensed in the first paragraphs to capture the attention of the reader.
- *LexRank* is an unsupervised, graph-based summary generation system inspired by both PageRank and HITS. It is based on the idea that the relevance of a sentence depends on its similarity with the rest of the sentences in the text. The nodes of the graph are the document sentences and the edges measure the similarity between two sentences using a idf based cosine distance. Two sentences are connected if the cosine similarity between them is greater than a certain threshold. The summary is made with the most salient sentences. If a sentence is similar to many others, then it must be salient in the document.
- *TextRank*, like LexRank, is an unsupervised graph-based system inspired by PageRank. It uses a variation of PageRank to extract the most salient sentences of the document. Its most significant difference from LexRank is the way in which the weights of the edges are calculated. In this case, the edges measure the similarity between the different nodes based on the number of common words in the sentences.
- *LSA* is a method based on Singular Value Decomposition, where a word-sentence matrix is decomposed in three new matrices. One of these matrices represents the association of underlying topics to sentences. This matrix is used to select the more salient sentences.
- *SumBasic* exploits frequency related properties of the words to compose summaries, arguing that high frequency words in the documents are very likely to appear in the human generated summaries. It is a greedy search approximation where, first the probability distributions of the words are computed, second by using these probabilities a weight is assigned to each document sentence and later, the best scoring sentence is selected to fill the summary until the desired summary length has been reached.

In all the experimentation, we used the implementation of these systems provided by the Python *sumy* library (https://github.com/miso-belica/sumy) using the default configuration. All these systems extract 3 sentences in order to compose the summary.

The performance of the systems was evaluated by using variants of the ROUGE measure [30]. Concretely, Rouge-N with unigrams and bigrams (Rouge-1 and Rouge-2) and Rouge-L. Furthermore, the compression ratio of the generated summaries (Compression) was also analyzed. In order to compute the confidence intervals, we used the Bootstrap Confidence Intervals [31] approach. First, from the set of hypotheses provided by the system that we want to evaluate, we generated up to 1000 resamples by sampling with replacement from this original set of hypotheses. Each resample had the same size of the original set. Next, the value of the evaluation measure was calculated for each of the resamples. Finally, we computed the 95% confidence interval using the bootstrap distribution.

Table 3 shows the results of our system compared to other summarization systems using the test set of ES-NEWS corpus. All the results in this experimentation are statistically significant.

Table 3. Results on ES-NEWS corpus with respect to the ground truth (full length Rouge F_1).

	Rouge-1	Rouge-2	Rouge-L	Compression
SHA-NN	30.1 ± 0.19	14.6 ± 0.20	25.6 ± 0.19	6.5
Lead	**32.8** ± 0.21	**16.2** ± 0.24	**27.5** ± 0.21	8.7
LexRank	28.4 ± 0.18	12.4 ± 0.18	23.6 ± 0.16	6.3
TextRank	24.0 ± 0.18	10.7 ± 0.17	20.2 ± 0.16	4.7
LSA	27.1 ± 0.16	8.4 ± 0.17	21.7 ± 0.15	8.5
SumBasic	29.9 ± 0.19	10.1 ± 0.19	24.5 ± 0.18	10.7

The results of all the systems in Table 3 on the ES-NEWS corpus (in Spanish) considering Rouge-2 measure are in line with those on the English CNN/DailyMail corpus [15]; for instance, SHA-NN and Lead achieved 14.7 and 15.1 respectively on the CNN/DailyMail corpus. However, in terms of Rouge-1 and Rouge-L all the systems present a slight decrease of performances on the ES-NEWS corpus with respect to CNN/DailyMail corpus; for instance, SHA-NN and Lead achieved 35.4 and 37.3 respectively on the CNN/DailyMail corpus considering Rouge-1. This fact illustrates small differences between the two corpora that have affected the results of all the compared systems. Regarding SHA-NN system, the results show a good transferability between languages as we hypothesized.

Using the summarization system trained in the above experimentation, we evaluated it with the LN24-SUMM test corpus, which contains 30 document-summary pairs (a small test set compared to the training set). Table 4 shows the results of the SHA-NN system compared to other summarization systems.

Table 4. Results on LN24-SUMM corpus with respect to the ground truth (full length Rouge F_1).

	Rouge-1	Rouge-2	Rouge-L	Compression
SHA-NN	**46.0** ± 4.38	**29.0** ± 5.94	**42.2** ± 4.46	7.0
Lead	33.2 ± 4.86	17.4 ± 5.86	29.9 ± 5.17	13.4
LexRank	39.7 ± 3.64	20.6 ± 4.73	35.2 ± 4.03	5.9
TextRank	43.3 ± 5.76	27.0 ± 7.65	39.7 ± 6.10	3.9
LSA	36.1 ± 4.60	15.8 ± 5.86	31.2 ± 5.10	9.4
SumBasic	31.8 ± 5.21	14.4 ± 5.45	28.7 ± 4.94	24.7

This table shows that the results of our summarization system are better than those of the other systems at all levels of ROUGE, although it should be considered that the small size of the test set does not allow to obtain statistically significant results. It should be noted that when working with the ES-NEWS corpus, the Lead system, which consists of extracting the first 3 sentences of the article as a summary, outperforms the rest of the systems, including ours, as Table 3 shows. However, this system performed worse on LN24-SUMM corpus. This is due to the fact that in ES-NEWS corpus, unlike the LN24-SUMM corpus, the summary tends to be a very approximate version of the first sentences of the articles. Also, it is interesting that, although SHA-NN was trained under the bias to the first sentences (ES-NEWS), it is capable of generalizing when the relevant sentences are more scattered in the document (LN24-SUMM).

In relation to the transferability between domains, it is possible to see that all the results, in terms of ROUGE, obtained on the LN24-SUMM corpus are higher than those obtained on the ES-NEWS corpus. That is because the reference summaries of the LN24-SUMM corpus have a very high density (i.e., they are composed by long extractive fragments of the transcribed talk shows) and a very low compression ratio in comparison to the ES-NEWS corpus, as it can be seen in Tables 1 and 2.

Furthermore, it is interesting to see in Table 4 that in general, when the compression ratio of the generated summaries increases, the results in terms of ROUGE decrease. Our system provide the best trade-off Rouge/Compression among all systems. Moreover, although TextRank obtains the most similar results with respect to SHA-NN, it suffers from a very low compression ratio due to it tends to extract the longest sentences.

6. Conclusions

We have studied the transferability of the SHA-NN summarization system, which is based on Siamese Hierarchical Attention Neural Networks, between languages and between application domains. Regarding the languages, the results of our system on the ES-NEWS corpus, in Spanish, are in line with those on the CNN/DailyMail corpus, in English. Regarding the application domains, we trained our summarization system on the ES-NEWS corpus, a text corpus of newspaper articles and we applied it to the summarization of transcribed speech of talk shows. The experimental results confirm the good behaviour of our proposal. We presented experiments on transcribed speech and as future work we will address its application to recognized speech. We will also study the evolution of our proposal to tackle with abstractive summarization based on the weights provided by SHA-NN in order to reduce the impact of recognition errors in the generation of summaries.

Author Contributions: Conceptualization, L.-F.H. and E.S. (E. Segarra); Data curation, J.-A.G., L.-F.H., E.S. (E. Segarra) and F.G.-G.; Formal analysis, J.-A.G., L.-F.H., F.G.-G. and E.S. (E. Sanchis); Funding acquisition, L.-F.H.; Project administration, L.-F.H.; Investigation, J.-A.G., E.S. (E. Segarra), F.G.-G. and E.S. (E. Sanchis); Methodology, E.S. (E. Segarra) and F.G.-G.; Software, J.-A.G.; Validation, E.S. (E. Sanchis); Writing—original draft, J.-A.G., E.S. (E. Segarra), E.S. (E. Sanchis), F.G.-G.; Writing—review & editing, J.-A.G., L.-F.H., E.S. (E. Segarra), F.G.-G. and E.S. (E. Sanchis).

Funding: This work has been partially supported by the Spanish MINECO and FEDER founds under project AMIC (TIN2017-85854-C4-2-R). Work of José-Ángel González is financed by Universitat Politècnica de València under grant PAID-01-17.

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. Carbonell, J.; Goldstein, J. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24–28 August 1998; ACM: New York, NY, USA, 1998; pp. 335–336. [CrossRef]
2. Ozsoy, M.G.; Cicekli, I.; Alpaslan, F.N. Text Summarization of Turkish Texts Using Latent Semantic Analysis. In Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 23–27 August 2010; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010; pp. 869–876.
3. Erkan, G.; Radev, D.R. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. *J. Artif. Int. Res.* **2004**, *22*, 457–479. [CrossRef]
4. 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.
5. Tur, G.; De Mori, R. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
6. Lloret, E.; Palomar, M. Text summarisation in progress: A literature review. *Artif. Intell. Rev.* **2012**, *37*, 1–41. [CrossRef]
7. Shen, D.; Sun, J.T.; Li, H.; Yang, Q.; Chen, Z. Document Summarization Using Conditional Random Fields. In Proceedings of the 20th International Joint Conference on Artifical Intelligence, Hyderabad, India, 6–12 January 2007; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2007; pp. 2862–2867.
8. Begum, N.; Fattah, M.; Ren, F. Automatic text summarization using support vector machine. *IJICIC* **2009**. *5*, 1987–1996.
9. 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.
10. Nallapati, R.; Zhou, B.; dos Santos, C.N.; Çaglakase, G.; Xiang, B. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In Proceedings of the SIGNLL Conference on Computational Natural Language Learning (CoNLL), Berlin, Germany, 7–12 August 2016.

11. 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; pp. 3075–3081.
12. 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 (Volume 1: Long 316 Papers), Vancouver, CB, Canada, 30 July–4 August 2017; pp. 1073–1083. [CrossRef]
13. Paulus, R.; Xiong, C.; Socher, R. A Deep Reinforced Model for Abstractive Summarization. *CoRR* **2017**, arXiv: 1705.04304.
14. Narayan, S.; Cohen, S.B.; Lapata, M. Ranking Sentences for Extractive Summarization with Reinforcement Learning. In Proceedings of the NAACL-HLT. Association for Computational Linguistics, New Orleans, LA, USA, 1–6 June 2018; pp. 1747–1759. [CrossRef]
15. Ángel González, J.; 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]
16. Furui, S.; Kikuchi, T.; Shinnaka, Y.; Hori, C. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Trans. Speech Audio Process.* **2004**, *12*, 401–408. [CrossRef]
17. Liu, S.H.; Chen, K.Y.; Chen, B.; Wang, H.M.; Yen, H.C.; Hsu, W.L. Combining Relevance Language Modeling and Clarity Measure for Extractive Speech Summarization. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 957–969.
18. Tsai, C.I.; Hung, H.T.; Chen, K.Y.; Chen, B. Extractive speech summarization leveraging convolutional neural network techniques. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 13–16 December 2016; pp. 158–164.
19. 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—Volume 1, Montreal, QC, Canada, 7–12 December 2015; MIT Press: Cambridge, MA, USA, 2015; pp. 1693–1701.
20. 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; pp. 708–719.
21. 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; pp. 1480–1489. [CrossRef]
22. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* **2015**, arXiv: 1409.0473.
23. Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; Bordes, A. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 670–680. [CrossRef]
24. Zhou, Q.; Yang, N.; Wei, F.; Huang, S.; Zhou, M.; Zhao, T. Neural document summarization by jointly learning to score and select sentences. *arXiv* **2018**, arXiv:1807.02305.
25. Wu, Y.; Hu, B. Learning to extract coherent summary via deep reinforcement learning. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
26. Narayan, S.; Papasarantopoulos, N.; Cohen, S.B.; Lapata, M. Neural extractive summarization with side information. *arXiv* **2017**, arXiv:1704.04530.
27. Wang, H.M.; Chen, B.; Kuo, J.W.; Cheng, S.S. MATBN: A Mandarin Chinese broadcast news corpus. *Int. J. Comput. Linguist. Chin. Lang. Process.* **2005**, *10*, 219–236.
28. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
29. Nenkova, A.; Vanderwende, L. The impact of frequency on summarization. *Microsoft Res.* **2005**, *101*. CiteSeerX. Available online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.529.6099 (accessed on 11 September 2019).

30. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*; Marie-Francine Moens, S.S., Ed.; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
31. Moore, D.; McCabe, G.; Duckworth, W.; Sclove, S. *The Practice of Business Statistics Companion Chapter 18: Bootstrap Methods and Permutation Tests*; W. H. Freeman: New York, NY, USA, 2003.