Siamese Hierarchical Attention Networks for Extractive Summarization

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

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera sn, 46022, València, Spain
E-mail: {jogonba2, esegarra, fgarcia, esanchis, lhurtado}@dsic.upv.es

Abstract. In this paper, we present an extractive approach to document summarization based on Siamese Neural Networks. Specifically, we propose the use of Hierarchical Attention Networks to select the most relevant sentences of a text to make its summary. We train Siamese Neural Networks using document-summary pairs to determine whether the summary is appropriated for the document or not. By means of a sentence-level attention mechanism the most relevant sentences in the document can be identified. Hence, once the network is trained, it can be used to generate extractive summaries. The experimentation carried out using the CNN/DailyMail summarization corpus shows the adequacy of the proposal. In summary, we propose a novel end-to-end neural network to address extractive summarization as a binary classification problem which obtains promising results in-line with the state-of-the-art on the CNN/DailyMail corpus.

Keywords: Siamese Neural Networks, Hierarchical Attention Networks, Automatic Text Summarization

1. Introduction

Nowadays, automatic summarization is an important issue in the current world due to the great amount of information in different formats that is accessible. It is necessary to develop techniques that help us to tackle that huge amount of information. For this reason, there is an increasing interest in some areas of speech and text processing to develop techniques that allow the users to find, read, understand, or process the documents. In this context, automatic summarization can be an important aid because it provides a condensed version of documents that reduce the time to explore or analyze them.

Summarization techniques [29] [13] are usually classified as extractive, where some sentences (or other units) are selected from the documents, and abstractive, where the final summary is a sequence of generated sentences. Regarding the different approaches used to document summarization, some works are based on unsupervised learning techniques by considering statistical word features [2], topic modeling such as Latent Semantic Analysis [4], graph based approaches such as LexRank [5], among others [29] [13]. There are also systems based on supervised learning techniques such as Conditional Random Fields [27], Support Vector Machines [1] or Neural Networks [3] [20] [26] [22] [21].

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 [6] [12]. These systems have to tackle with specific problems derived from the errors generated by the speech recognition phase such as misrecognized words, or errors in punctuation marks.

Progress in summarization research has been influenced by the organization of evaluation conferences and the collection of corpora for training and test purposes. It can be highlighted the Document Understand-

*Corresponding author. E-mail: jogonba2@dsic.upv.es

ing Conferences (DUC)[1] which were integrated later in the Text Analysis Conference (TAC)[2]. These conferences were mainly oriented to evaluation tasks, therefore they provide corpora that were not large enough to be used in the estimation of some corpus-based models. This is the case of deep learning models, that are based on supervised learning techniques. Unfortunately, the construction of an appropriate corpus 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. An important resource for the corpus-based models is the recently created CNN/DailyMail summarization corpus, originally constructed by [7] for the passage-based question answering task, and adapted for the document summarization task [3] [21]. It consists of news stories from CNN and DailyMail and contains 312,085 document-summary pairs.

In the last few years, approaches based on Neural Networks have been applied to summarization, taking advantage of their powerful capabilities to learn extremely complex functions. The most widely used approaches are based on encoder-decoder architectures modeled by Recurrent Neural Networks that had provided good results in translation tasks. Generally, in these approaches, the encoder processes the source sequence as a list of continuous-space representations and the decoder generates the target sequence. Some of these approaches also incorporate attention mechanisms. In particular, Cheng and Lapata [3] proposed an attentional encoder-decoder approach for extractive single-document summarization and Nallapati, Zhai and Zhou [20] presented an extractive summarization approach based on sentence classification using Neural Networks. In both works, they applied their approaches to the CNN/DailyMail corpus since its large size makes it attractive for training deep Neural Networks.

In this work, we propose an extractive approach to text summarization which is based on Siamese Neural Networks with Hierarchical Attention mechanisms using distributed vector representation of words. Siamese Neural Networks are capable of learning from positive and negative samples. In our approach, we provide the network with positive and negative document-summary pairs; a positive pair is a document and its summary from the training set and a negative pair is a document and a summary of other different document randomly extracted from the training set. The Siamese Network is trained as a classifier to distinguish whether a summary is correct for a document or not. Furthermore, this model is enriched with an attention mechanism that provides the final score associated to each sentence of the input document. This way, given a document, the model assigns weights to the sentences, which allows us to establish a ranking, and to select the most salient sentences to build the summary. In summary, we propose a novel end-to-end neural network to address extractive summarization as a binary classification problem. In comparison to other deep learning approaches, our system requires a training time of only a few hours. On the other hand, some deep learning approaches require adapting the training corpus before training their models e.g. to convert the abstractive reference summaries to extractive labels as in [20]. This adaptation is not necessary in our approach that uses the training corpus in a straightforward way. We have performed some experiments on the CNN/DailyMail corpus that confirm the promising behaviour of our approach to the summarization problem.

2. System Description

Our system addresses the extractive summarization task as a classification problem. Specifically, it learns whether a summary x' is correct for a document x or not. We consider that a summary x' is correct for a document x when they have similar semantics. In order to represent such semantics, we use Hierarchical Attention Networks (HAN) composed by Bidirectional Long Short Term Memory (BLSTM) [8] [25] networks.

This kind of network allows us to extract a vector representation of documents from the representations of their sentences. Moreover, the representation of each sentence is obtained from the representations of their words. As word representation we used Word2Vec word embeddings estimated from Google News [18] [17].

We use HAN to process both documents and summaries, where all the BLSTM share their weights, i.e. we use the same BLSTM to process documents and summaries, but the attention mechanisms have different weights. Finally, once the representations of documents and summaries are obtained, we follow an approach similar to [19] where the representation of doc-

[1] http://www-nlpir.nist.gov/projects/duc/
[2] http://www.nist.gov/tac/

uments and summaries with the difference between them are concatenated. We call this approach Siamese Hierarchical Neural Networks (SHA-NN) and its architecture scheme is shown in Figure 1.

2.1. Word Level

Let $\boldsymbol{x} = \{\boldsymbol{w_{11}}, ..., \boldsymbol{w_{1W}}, ..., \boldsymbol{w_{T1}}, ..., \boldsymbol{w_{TW}}\}$ and $\boldsymbol{x'} = \{\boldsymbol{v_{11}}, ..., \boldsymbol{v_{1V}}, ..., \boldsymbol{v_{Q1}}, ..., \boldsymbol{v_{QV}}\}$ be the input document and the input summary respectively, where $\boldsymbol{w_{ij}} \in \mathbb{R}^d$ is the d-dimensional embedding of the word j in the sentence i of the document $\boldsymbol{x}$, $\boldsymbol{v_{ij}} \in \mathbb{R}^d$ is the d-dimensional embedding of the word j in the sentence i of the summary $\boldsymbol{x'}$, W and V are the maximum number of words in the sentences of the document and the summary respectively, and T and Q are the maximum number of sentences in the document and the summary respectively.

From $\boldsymbol{x}$ and $\boldsymbol{x'}$ the networks compute the vector representation of the sentences in the same way for the document ($\boldsymbol{s_i} : 1 \leq i \leq T$) and the summary ($\boldsymbol{q_i} : 1 \leq i \leq Q$), as we show in Eqs. (1) and (2).

$$\boldsymbol{s_i} = \sum_{j=1}^{W} \alpha_{ij}[\overleftarrow{\boldsymbol{h_{ij}}}, \overrightarrow{\boldsymbol{h_{ij}}}] \tag{1}$$

$$\boldsymbol{q_i} = \sum_{j=1}^{V} \beta_{ij}[\overleftarrow{\boldsymbol{g_{ij}}}, \overrightarrow{\boldsymbol{g_{ij}}}] \tag{2}$$

where $\overleftarrow{\boldsymbol{h_{ij}}} \in \mathbb{R}^{d_1}$, $\overrightarrow{\boldsymbol{h_{ij}}} \in \mathbb{R}^{d_1}$, $\overleftarrow{\boldsymbol{g_{ij}}} \in \mathbb{R}^{d_1}$ and $\overrightarrow{\boldsymbol{g_{ij}}} \in \mathbb{R}^{d_1}$ are the outputs of the BLSTM at word level, both for document and summary respectively. And where $\boldsymbol{\alpha_i}$ and $\boldsymbol{\beta_i}$ are calculated as shown in Eqs. from (3) to (6).

$$\boldsymbol{\alpha_i} = \frac{e^{\boldsymbol{u_i}}}{\sum_{j=1}^{W} e^{u_{ij}}} \tag{3}$$

$$\boldsymbol{\beta_i} = \frac{e^{\boldsymbol{v_i}}}{\sum_{j=1}^{V} e^{v_{ij}}} \tag{4}$$

$$u_{ij} = tanh(W_u[\overleftarrow{\boldsymbol{h_{ij}}}, \overrightarrow{\boldsymbol{h_{ij}}}] + b_u) \tag{5}$$

$$v_{ij} = tanh(W_v[\overleftarrow{\boldsymbol{g_{ij}}}, \overrightarrow{\boldsymbol{g_{ij}}}] + b_v) \tag{6}$$

where $W_u \in \mathbb{R}^{2 \cdot d_1}$, $b_u \in \mathbb{R}$, $W_v \in \mathbb{R}^{2 \cdot d_1}$ and $b_v \in \mathbb{R}$ are the weights and the bias of the attention mechanism at word level both for document and summary respectively. $\alpha_{ij} \in \mathbb{R}$ is the relevance of the word j in the document sentence i, $\beta_{ij} \in \mathbb{R}$ is the relevance of the word j in the summary sentence i, $\boldsymbol{s_i} \in \mathbb{R}^{d_1}$ is the vector representation of the sentence i of the document, and $\boldsymbol{q_i} \in \mathbb{R}^{d_1}$ is the vector representation of the sentence i of the summary.

2.2. Sentence Level

From $\boldsymbol{s} = \{\boldsymbol{s_1}, \boldsymbol{s_2}, ..., \boldsymbol{s_T}\}$ and $\boldsymbol{q} = \{\boldsymbol{q_1}, \boldsymbol{q_2}, ..., \boldsymbol{q_Q}\}$, we compute the vector representations of the document and the summary, $\boldsymbol{r}$ and $\boldsymbol{p}$ respectively, as we show in Eq. (7) and Eq. (8).

$$\boldsymbol{r} = \sum_{i=1}^{T} \widehat{\alpha}_i[\overleftarrow{\boldsymbol{h_i}}, \overrightarrow{\boldsymbol{h_i}}] \tag{7}$$

$$\boldsymbol{p} = \sum_{i=1}^{Q} \widehat{\beta}_i[\overleftarrow{\boldsymbol{g_i}}, \overrightarrow{\boldsymbol{g_i}}] \tag{8}$$

where $\overleftarrow{\boldsymbol{h_i}} \in \mathbb{R}^{d_2}$, $\overrightarrow{\boldsymbol{h_i}} \in \mathbb{R}^{d_2}$, $\overleftarrow{\boldsymbol{g_i}} \in \mathbb{R}^{d_2}$ and $\overrightarrow{\boldsymbol{g_i}} \in \mathbb{R}^{d_2}$ are the outputs of the BLSTM at sentence level both for document and summary respectively. And where $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ are calculated as shown in Eqs. from (9) to (12).

$$\widehat{\boldsymbol{\alpha}} = \frac{e^{\widehat{\boldsymbol{u}}}}{\sum_{i=1}^{T} e^{\widehat{u}_i}} \tag{9}$$

$$\widehat{\boldsymbol{\beta}} = \frac{e^{\widehat{\boldsymbol{v}}}}{\sum_{i=1}^{Q} e^{\widehat{v}_i}} \tag{10}$$

$$\widehat{u}_i = tanh(W_{\widehat{u}}[\overleftarrow{\boldsymbol{h_i}}, \overrightarrow{\boldsymbol{h_i}}] + b_{\widehat{u}}) \tag{11}$$

$$\widehat{v}_i = tanh(W_{\widehat{v}}[\overleftarrow{\boldsymbol{g_i}}, \overrightarrow{\boldsymbol{g_i}}] + b_{\widehat{v}}) \tag{12}$$

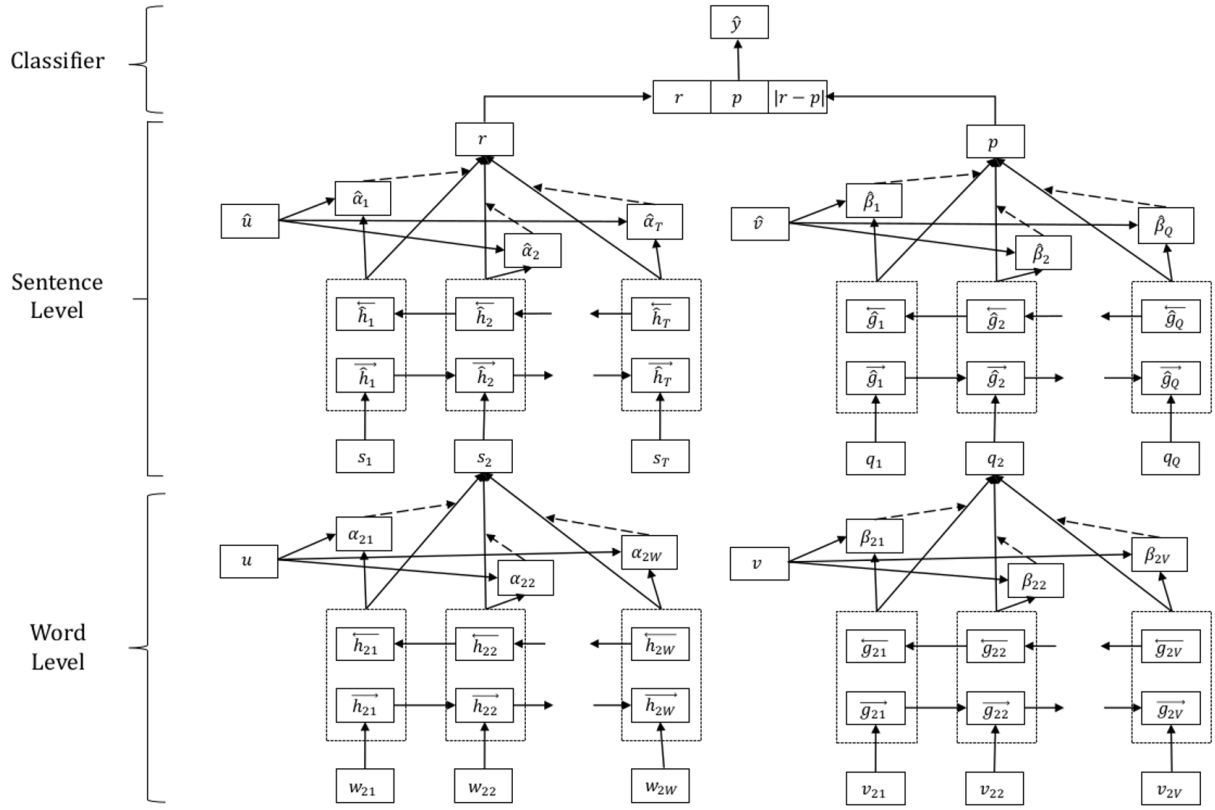

Fig. 1. The Siamese Hierarchical Neural Networks (SHA-NN) architecture scheme.

where $W_{\widehat{u}} \in \mathbb{R}^{2 \cdot d_2}$, $b_{\widehat{u}} \in \mathbb{R}$, $W_{\widehat{v}} \in \mathbb{R}^{2 \cdot d_2}$ and $b_{\widehat{v}} \in \mathbb{R}$ are the weights and the bias of the attention mechanism at sentence level both for document and summary. $\widehat{\alpha}_i \in \mathbb{R}$ is the relevance of the sentence i in the document, $\widehat{\beta}_i \in \mathbb{R}$ is the relevance of the sentence i in the summary, and $\boldsymbol{r} \in \mathbb{R}^{d_2}$ and $\boldsymbol{p} \in \mathbb{R}^{d_2}$ are the representations of document and summary respectively.

2.3. Classifier

The vector representations of the document $\boldsymbol{r}$, the summary $\boldsymbol{p}$, and the difference between them $|\boldsymbol{r} - \boldsymbol{p}|$ are concatenated to feed a fully-connected output layer with softmax activation function, as defined in Eq. (13).

$$\widehat{y} = softmax(W_{\widehat{y}}[\boldsymbol{p}, \boldsymbol{r}, |\boldsymbol{r} - \boldsymbol{p}|] + b_{\widehat{y}}) \tag{13}$$

In order to train the model, for each document we built a positive pair $(\boldsymbol{x_j}, \boldsymbol{x'_j})$, provided by the corpus, and a negative pair $(\boldsymbol{x_j}, \boldsymbol{x'_k}) : j \neq k$ where $\boldsymbol{x'_k}$ 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). Moreover, we considered that one train epoch was reached after processing 500 batches (32,000 samples in each epoch). We made this consideration with the aim of observing the model behaviour with finer granularity.

2.4. Summarization

In order to carry out document summarization with SHA-NN, once the network has been trained to distinguish correct summaries for documents, some of its estimated parameters could be used to select the document sentences that will compose the summary. That is, for the summarization process, given a document, a forward pass is performed to obtain the weight of each document sentence $\widehat{\alpha}_i$. From the ranking of the document sentences based on those weights the system considers the top sentences to build the summary.

Table 1
Average number of sentences and words (including words per sentence) in the training set.

	Sents	Words	Words/Sent
DailyMail Documents	27.0	773.1	28.6
DailyMail Summaries	3.9	56.1	14.5
CNN/DailyMail Documents	28.2	765.4	27.1
CNN/DailyMail Summaries	3.8	53.4	14.1

3. Corpus

For the experiments, the CNN/DailyMail[3] corpus, which is a combination of the CNN and the DailyMail corpora, was used. This corpus was originally constructed by [7] for question answering and modified by [3] and [21] for extractive and abstractive summarization respectively.

To make a fair comparison with other works, we used the anonymized version of DailyMail corpus, which consists of 196,961 training documents, 12,148 validation documents and 10,397 test documents. Additionally, we carried out other experiments with the combination of CNN/DailyMail corpus (anonymized version), which consists of 287,227 training documents, 13,368 validation documents and 11,490 test documents. In Table 1, several features of these two corpora are shown.

The ground truth summaries provided by this corpus are abstractive. They are built by the concatenation of all the highlights associated to the documents.

4. Related Work

Recently, due to the impact of neural networks in the Natural Language Processing community, a large number of approaches based on Deep Learning for Text Summarization have been proposed. Most of the proposed approaches address the summarization problem from an abstractive perspective and are based on encoder-decoder models with attention mechanisms [23] [28] [21] [15] [22]. The main problems of these systems are the generation of repeated words and the inability of producing words out of the training vocabulary (especially name entities). For this reason, more recent approaches propose coverage mechanisms and Pointer Networks to deal with these problems [26].

[3] https://cs.nyu.edu/~kcho/DMQA/

However, the extractive summarization has not been deeply explored and only a few works address it [20] [3]. ([26] can be seen as a hybrid between abstractive and extractive).

In this section, we discuss the works with which we have compared our system. To our knowledge, they are the only works that used the CNN/DailyMail corpus to perform summarization from an extractive perspective.

In [3], the authors propose an encoder-decoder model combined with attention mechanisms for extractive summarization. However, they report their results using a subset of 500 samples from DailyMail corpus. For this reason, we do not compare our system with them.

In [20], the authors present a Hierarchical Attention Network to choose sentences from the document as a sequence classification problem. They used two different summary sets to train their system. The SummaRunner-Abs system is trained using the summaries provided by the corpus. The SummaRunner-Ext system is trained from new summaries obtained by a greedy algorithm that transforms the CNN/DailyMail abstract summaries into extractive summaries, choosing a set of sentences from the document that maximize the similarity with respect to the abstractive summary.

SHA-NN and SummaRunner-Abs are similar systems since both are based on a sentence ranking without the need of converting the abstractive summaries in extractive summaries. However, the main difference between them is that they address different classification problems. In our case, the aim of the classification is to distinguish whether a summary is correct for a document or not, whereas SummaRunner-Abs system addresses a sequence classification problem in order to select sentences. Moreover, our criterion of sentence selection is based on a hierarchy of activations on the sentence level, instead of a classification for each sentence of the document. Furthermore, it is necessary to highlight that SHA-NN system is able to extract the most salient words of the documents due to the use of word and sentence level attention mechanisms.

Regarding [26], they proposed a hybrid approach based on Pointer Networks and encoder-decoder models with attention mechanisms. In the summarization process, the network can choose between generating a new word from the vocabulary or copying a word from the source document. Moreover, in order to address the word repetition problem, the authors enrich the system by using a coverage mechanism based on the attentions of previous timesteps, for each decoder timestep.

In addition to these systems, we compare the results of the SHA-NN system to another extractive approach which is not based on neural networks, TextRank [16]. Moreover, we provide results using two straightforward mechanisms: Lead-K, that extracts only the first K (tipically $K = 3$) sentences of the document, and Random-K, that randomly extracts K sentences from the document.

5. Experiments

We carried out two different experiments, one for each corpus. We evaluated the performance of our system in the experiments by using variants of the ROUGE measure [11]. Concretely, Rouge-N with unigrams and bigrams (Rouge-1 and Rouge-2) and Rouge-L were used. Although in the literature there are some proposals to evaluate automatic summarizations without using the gold standard [14] [24], in order to compare our system to other approaches in the same conditions, we evaluated it with ROUGE statistics using the gold standard provided by the DailyMail and CNN/DailyMail corpora.

In the experimentation we used $d_1 = d_2 = 512$, BatchNormalization [9] between each pair of layers, and Adam [10] as algorithm to minimize the cross entropy. From the ranking of the document sentences based on $\widehat{\alpha}_i$ the system considers the 3 top sentences (3 sentences with greater $\widehat{\alpha}_i$, sorted by their position in the document) to build the summary. We used zero padding to normalize the length of documents and sentences.

The experiments consisted in two steps. First, the network is trained, and second, the trained network is used to process a new document by means of a forward pass, in order to obtain the weights $\widehat{\alpha}_i$ for each document sentence.

In the first step, we trained our system with the goal of distinguishing correct summaries for a given document, i.e. we used the SHA-NN system to solve the classification problem. In this step, we selected the best epoch for the model evaluated on the validation set. The cross entropy at each epoch for DailyMail and CNN/DailyMail corpora are shown in Figures 2 and 3.

Regarding to DailyMail corpus, the best model was obtained after 65 epochs, trained with 2,080,000 positive and negative pairs. This process took 5 hours on a single Nvidia Titan X GPU. With respect to CNN/DailyMail corpus, the best model was obtained after 70 epochs, trained with 2,240,000 positive and negative pairs. This process took 5.5 hours. Therefore, our model stands out in training speed compared to other systems such as [21] who took a few days to reach convergence on a single Tesla K-40 GPU.

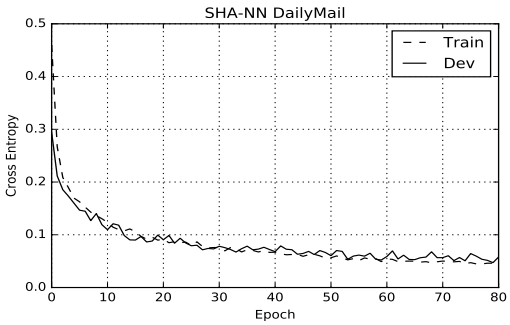

Fig. 2. Training and validation loss on DailyMail.

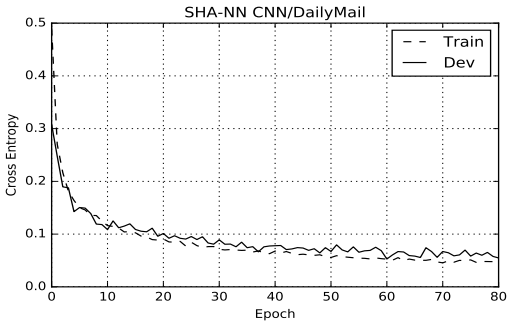

Fig. 3. Training and validation loss on CNN/DailyMail.

In the second step, once the network was trained, it was possible to use it for extractive summarization. With the aim of comparing our system to other systems, we carried out two experiments, both for DailyMail and CNN/DailyMail corpora.

In all the result tables, the results labeled with † were obtained in the experimentation we done in our laboratory, while the results labeled with ⋄ and ‡ are provided in [20] and [26] respectively. In [20], they use the anonymized and preprocessed version of the corpus, and in [26] they use the non-anonymized version of the corpus. It is important to mention that we have not used any kind of preprocess, for this reason, the experiments with Lead-3 do not get the same results compared to [20].

Tables 2 and 3 show the results of our system (SHA-NN) on the DailyMail corpus compared to other works, at 75 and 275 bytes for the evaluation with limited length Rouge recall.

Table 2

Results on DailyMail corpus with respect to the ground truth at 75 bytes (limited length Rouge recall).

	Rouge-1	Rouge-2	Rouge-L
Lead-3†	20.8	6.8	11.1
TextRank †	20.8	7.3	11.2
Random-3 †	12.5	2.2	6.5
SHA-NN †	**24.0**	**9.6**	**13.3**
Lead-3 ◇	21.9	7.2	11.6
SummaRunner-Abs ◇	23.8	9.6	13.3
SummaRunner-Ext ◇	**26.2**	**10.8**	**14.4**

Table 3

Results on DailyMail corpus with respect to the ground truth at 275 bytes (limited length Rouge recall).

	Rouge-1	Rouge-2	Rouge-L
Lead-3†	38.5	14.3	31.1
TextRank †	35.3	12.3	27.6
Random-3 †	27.0	6.6	21.4
SHA-NN †	**38.8**	**15.0**	**31.4**
Lead-3 ◇	40.5	14.9	32.6
SummaRunner-Abs ◇	40.4	15.5	32.0
SummaRunner-Ext ◇	**42.0**	**16.9**	**34.1**

Table 2 shows that our system provides results that are slightly worse than those provided by the best version of the SummaRunner system, however, it exceeds both the reference systems, the Lead-3 and the Random-3, as well as the results of the TextRank system. These are the results when the evaluation is made with respect to the ground truth at 75 bytes. When the evaluation is made with respect to the ground truth at 275 bytes, see Table 3, the results of our system are slightly worse: the results of TextRank and Random-3 systems are widely surpassed, however those of the Lead-3 system are only slightly improved.

Table 4 shows the results on the CNN/DailyMail corpus in terms of Rouge F_1 using full-length summaries.

As Table 4 shows, only the best version of the SummaRunner system is able to slightly outperform the results of its Lead-3 reference system. Our results when the evaluation is made with respect to the ground truth (full length Rouge F_1) are worse than those provided by our Lead-3 reference system. That is true also for the best version of the Pointer Gen system.

When the length of the phrases of the summary considered for the calculation of the ROUGE, grows a deterioration of the results provided by all the compared systems is observed. We hypothesize that it can be due to the fact that it is the third of the sentences provided as the summary those that includes a greater number of errors with respect to the reference summary.

Table 4

Results on CNN/DailyMail corpus with respect to the ground truth (full length Rouge F_1).

	Rouge-1	Rouge-2	Rouge-L
Lead-3†	**37.3**	**15.1**	**34.0**
TextRank †	29.4	10.1	26.3
Random-3 †	26.7	7.3	23.9
SHA-NN †	35.4	14.7	33.2
Lead-3 ◇	39.2	15.7	**35.5**
SummaRunner-Abs ◇	37.5	14.5	33.4
SummaRunner-Ext ◇	**39.6**	**16.2**	35.3
Lead-3 ‡	**40.3**	**17.8**	**36.6**
Pointer Gen ‡	36.4	15.7	33.4
Pointer Gen + Cov ‡	39.5	17.3	36.4

Document: @entity0 want to sign @entity3 full - back @entity1 on a permanent deal . **the 25 - year - old signed for @entity3 from @entity6 for £ 2.1million but has been on loan with @entity0 this season . they have an option to make the deal permanent for £ 1.5million but @entity1 wants to see if there are other options before committing .** he has two years left on contract at @entity3 . **@entity0 defender @entity1 shields the ball from @entity15 forward @entity14 @entity0 , meanwhile , will not take up an option to sign @entity18 striker @entity17 on a permanent deal .** the @entity20 scored only one goal following his £ 8.5million move to @entity18 from @entity23 and has scored six times this season for @entity0 in 31 appearances . he will return to @entity26 at the end of the season . @entity1 (right) competes for a header with @entity29 forward @entity30 in december .

Ground Truth: the 25 - year - old signed for @entity3 from @entity6 for £ 2.1million . but the defender has been on loan with @entity0 this season . the club have an option to make the deal permanent for £ 1.5million . @entity0 will not take up an option to sign @entity18 striker @entity17 on a permanent deal .

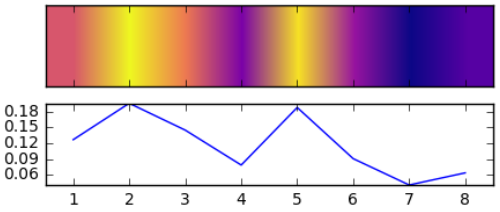

Fig. 4. Extractive summarization with a test sample of CNN/DailyMail corpus (DailyMail subset).

Document: **(@entity1) at least 54 people have died and 15 others are missing after a @entity4 fishing vessel sank off the @entity5 , according to @entity4 's state - run @entity7 news agency . more than 60 people were rescued thursday from the chilly waters in @entity4 's @entity10 . the @entity13 freezer trawler – a commercial fishing vessel – was carrying 132 people , the ministry said .** of the people on board , 78 were @entity4 . the 54 others were foreign nationals from @entity19 , @entity20 , @entity21 and @entity22 , according to the news agency , with the majority coming from @entity19. more than 20 fishing vessels are searching for the 15 people still thought to be missing , @entity7 said . the shipwreck was swift , with the trawler going down in the @entity30 within 15 minutes of getting into difficulties , the news agency reported . the most likely cause of the shipwreck was collision with an obstacle which damaged the hull , the official spokesman of @entity4 's @entity33 , @entity34 , is quoted as saying . the trawler is also thought to have keeled over as a result of hauling some 80 tons of fish on to the deck , the chairman of the emergencies commission in the @entity5 region , @entity43 , told @entity7 .

Ground Truth: fishing vessels are searching for 15 people still thought to be missing . there were 132 people on board the ship , 78 of them @entity4 , @entity7 news agency says . the rest were foreign nationals from @entity19 , @entity20 , @entity21 and @entity22 , it says .

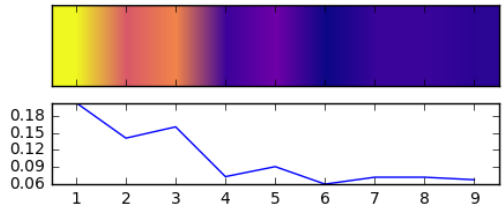

Fig. 5. Extractive summarization with a test sample of CNN/DailyMail corpus (CNN subset).

When evaluating the performance of the systems, it must be taken into account that we have compared extractive summaries against those of the reference that are abstractive. Moreover, we suspect that the abstractive summaries have been elaborated mainly using the first three sentences of the documents.

Figures 4 and 5 show two examples of summarization using the proposed SHA-NN system. We provide the *Document*, its *Ground Truth* summary, the three sentences extracted by our system (bold font), and the weights assigned to each sentence by our system. Figure 4 shows a good summarization example, where the sentences with highest weights contain the most information of the *Ground Truth*. Figure 5 shows another example, where the two sentences with the highest weights (sentences 1 and 3) are correct compared to the *Ground Truth*. However, the sentence with lowest weight (sentence 2) is not correct. Additionally, if our system considered more sentences for the summary, in this example the sentences 4 and 5, then the summary would contain almost all the information of the *Ground Truth*.

6. Conclusions

We have presented an approach based on Siamese Neural Networks for summarization tasks. It has been shown the adequacy of the proposed learning methodology to capture the relevance of words and sentences in order to extract the more salient sentences of documents. Our system also allows for the use of the training corpus in an easy way, taking advantage of positive and negative training samples. Experimental results confirm the promising behaviour of our proposal, they are in-line with the state-of-the-art. Additionally, our system requires a low training time.

As future works, we will study other kinds of deep learning architectures in order to process documents and summaries. In this work, we only experimented with extractive summarization at sentence level, however, as the model also assigns weights to words, we can take advantage of these information to enrich the sentence selection criterion. We will study the evolution of our proposal to tackle with abstractive summarization based on the weights obtained at word level. Furthermore, it could be interesting to explore the use of SHA-NN as sentence selector to feed abstractive models. Additionally, when we have adequate corpora in other languages we will also study the portability of our approach.

Acknowledgements

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 also financed by Universitat Politècnica de València under grant PAID-01-17.

References

[1] N. Begum, M. Fattah, and F. Ren. Automatic text summarization using support vector machine. 5:1987–1996, 07 2009.

[2] J. Carbonell and J. Goldstein. 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*, SIGIR '98, pages 335–336, New York, NY, USA, 1998. ACM.

[3] J. Cheng and M. Lapata. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[5] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479, Dec. 2004.

[6] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing*, 12(4):401–408, 2004.

[7] K. M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. *CoRR*, abs/1506.03340, 2015.

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

[9] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[11] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In S. S. Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[12] S. H. Liu, K. Y. Chen, B. Chen, H. M. Wang, H. C. Yen, and W. L. Hsu. Combining relevance language modeling and clarity measure for extractive speech summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(6):957–969, June 2015.

[13] E. Lloret and M. Palomar. Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37(1):1–41, 2012.

[14] A. Louis and A. Nenkova. Automatically assessing machine summary content without a gold standard. *Comput. Linguist.*, 39(2):267–300, June 2013.

[15] Y. Miao and P. Blunsom. Language as a latent variable: Discrete generative models for sentence compression. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 319–328. Association for Computational Linguistics, 2016.

[16] R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

[17] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc.

[19] S. Minaee and Z. Liu. Automatic question-answering using a deep similarity neural network. *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 923–927, 2017.

[20] R. Nallapati, F. Zhai, and B. Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3075–3081, 2017.

[21] R. Nallapati, B. Zhou, C. N. dos Santos, a. G. Çagl[Takase, and B. Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*, pages 280–290. ACL, 2016.

[22] R. Paulus, C. Xiong, and R. Socher. A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304, 2017.

[23] A. M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389. Association for Computational Linguistics, 2015.

[24] H. Saggion, J.-M. Torres-Moreno, I. d. Cunha, and E. SanJuan. Multilingual summarization evaluation without human models. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1059–1067, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[25] M. Schuster and K. Paliwal. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681, Nov. 1997.

[26] A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics, 2017.

[27] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen. Document summarization using conditional random fields. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2862–2867, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[28] S. Takase, J. Suzuki, N. Okazaki, T. Hirao, and M. Nagata. Neural headline generation on abstract meaning representation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1054–1059. Association for Computational Linguistics, 2016.

[29] G. Tur and R. De Mori. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.