An Attentional Extractive Summarization Framework
José Ángel González, Encarna Segarra, Fernando García-Granada, Emilio Sanchis, Lluís-F. Hurtado

VRAIN: Valencian Research Institute for Artificial Intelligence
Universitat Politècnica de València, Valencia, Spain
@dsic.upv.es
{jogonba2, esegarra, fgarcia, esanchis, lhurtado}

Abstract
Although currently, works on text summarization generally use abstractive approaches, extractive methods can be specially adequate for some applications, and they can help with other tasks such as Question Answering or Information Extraction. In this paper, we propose a general framework for extractive summarization, the Attentional Extractive Summarization framework. The proposed approach is based on the interpretation of the attention mechanisms of hierarchical neural networks, that 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 reference labels for each sentence before the training phase. We evaluate two different systems, formalized under the proposed framework, on the CNN/DailyMail and the NewsRoom corpora, which are some of the reference corpora in the most relevant works in text summarization. The results obtained during the evaluation support the adequacy of our proposal and they suggest that there is still room for the improvement of our attentional framework.

Index Terms: Siamese Neural Networks, Hierarchical Neural Networks, Attention Mechanisms, Extractive Summarization

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 phrasing/rewriting the sentences of the documents. Recently, an important effort has been done in developing abstractive methods. However, extractive approaches are still important in summarization, since they maintain the coherence, the factuality, and do not hallucinate such as abstractive approaches. Besides, selecting sentences is important also in other tasks such as question answering or information extraction.

Typically, supervised extractive summarization has been addressed as a sequential binary sentence classification problem [16], [12], [17], [23], [2], [22]. However, the available corpora do not provide directly this kind of labeling for training purposes, since in general, corpora only consist in (document, summary) pairs. In order to label the document sentences, previously to the training of the model, the most common strategy consists in using suboptimal extractive oracles [16], [12]. Also, several unsupervised approaches for extractive summarization have been proposed [9], [15]. The research objective of this work is to develop a framework for extractive summarization that we call Attentional Extractive Summarization. Our proposal dispenses with the sentence labeling, avoiding the large computational cost required to compute near-optimal solutions, and allowing 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 only a binary signal is required in order to train the model instead of a sentence labeling. This training allows our system to learn relationships among document and summary sentences, and to replace the binary sequential sentence classification by 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.

In this work, we instantiate two systems 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.

We evaluate and study the performance of these systems on the CNN/DailyMail [7] and NewsRoom [6] corpora, comparing them with other systems based on diverse strategies (extractive and mixed summarization systems based on oracles or reinforcement learning).

2. Attentional Extractive Summarization Framework
Approaches that do not rely on Reinforcement Learning strategies to optimize directly the ROUGE evaluation metric, are mainly based on the use of suboptimal oracle algorithms. Due to this, they require a binary sentence labeling in order to be trained. These approaches typically consist in using oracle systems to label the sentences by following some evaluation measures like ROUGE. In [17], 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 [17], the problem with 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, due to 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 [16], [14], [12].

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 [23]. 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 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 documents1. 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.

The proposed framework has some requirements. 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 use hierarchical models in order to model document-level representations from the sentence-level representations, which are built from the word-level representations. Second, a mechanism to distinguish correct summaries for documents, from the document-level representation, 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 is required to compute relationships among document and summary sentences. In our proposal, we focus 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.

A scheme of our framework can be seen in Fig. 1. Let $D = \{(X_k, X'_k)\}_{k=1}^M$ be a corpus of M documents (document, summary) pairs, where all documents and summaries are defined according to a vocabulary $V$, let $X_k = \{(x_{ij})_{ij}^W\}_{j=1}^T$ be a document composed by T sentences of W words, $X'_k \in V^{T \times W}$, let $X'_k = \{(x'_ij)_{ij}^Q\}_{j=1}^V$ be a summary composed by Q sentences of V words, $X'_k \in V^{Q \times V}$ and let $f : V^{T \times W} \times V^{Q \times V} \rightarrow \mathbb{R}^2$ be a model whose input is a (document, summary) pair and whose output is a probability distribution of the summary correctness over $C = \{0, 1\}$, 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 $X_k$, while for the $(X_k, X'_k)$ pair, the class computed will be 0, as $X'_k$ is the reference summary for another document from the corpus $D$, different from $X_k$. 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 $g : V^{T \times W} \rightarrow \mathbb{R}^d$ and in a summary encoder $g'_1 : V^{Q \times V} \rightarrow \mathbb{R}^{d'}$. 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.

First, $g_1 : V^{T \times W} \rightarrow \mathbb{R}^{T \times d_1}$ and $g'_1 : V^{Q \times V} \rightarrow \mathbb{R}^{Q \times d'_1}$, 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_i-dimensional embedding model $E : V \rightarrow \mathbb{R}^{d_i}$, typically pretrained and applied to arbitrary-length (P) word sequences i.e. $E : V^P \rightarrow \mathbb{R}^{P \times d_i}$. Therefore, $g_1 : \mathbb{R}^{T \times P \times d_i} \rightarrow \mathbb{R}^{T \times d_1}$ and $g'_1 : \mathbb{R}^{Q \times P \times d_i} \rightarrow \mathbb{R}^{Q \times d'_1}$. Second, in order to represent documents and summaries from the representation of their sentences, $g_2 : \mathbb{R}^{T \times d_1} \rightarrow \mathbb{R}^d$ and $g'_2 : \mathbb{R}^{Q \times d'_1} \rightarrow \mathbb{R}^{d'}$ are defined. These encoders can have 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 $g_1(.; \theta_1)$ and $g'(.; \theta_2)$ are defined as a composition of two levels, $g = g_2(R; \theta_{22})$ and $g' = g'_2(R'; \theta_{22})$, where $R = g_1(.; \theta_{11})$ and $R' = g'_1(.; \theta_{21})$. Due to 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 $\theta$ 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 $\theta_2 = [\theta_{21}, \theta_{22}].$

![Figure 1: General scheme of the Attentional Extractive Summarization framework.](image)
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 [1][13] as in [3] or as main mechanisms to compute representations [20] as in [4].

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 \( \mathbb{C} \), \( c: \mathbb{R}^{d_w} \times \mathbb{R}^{d_s'} \rightarrow \mathbb{R}^2 \), 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 \( p \)). Then, the objective is determined by the minimization of Eq. 1.

\[
\mathcal{L}(\Theta) = \sum_{k=1}^{[P]} L(f(X_k, X_k'; \Theta), y = 1) + \mathbb{E}_{p(X_j|X_k)}[L(f(X_k, X_k'; \Theta), y = 0)]
\]  

(1)

where \( L \) is a loss function, and \( \mathbb{E}_{p(X_j|X_k)}[X_k] \) 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.

3. Proposed Systems

Based on this framework we have instantiated two systems.

3.1. Siamese Hierarchical Attention Networks

Siamese Hierarchical Attention Neural Networks (SHA-NN) [3] is the instance of the general attentional framework when the encoders are Hierarchical Attention Networks [21] based on Bidirectional Long Short Term Memory (BLSTM) [8][18] with attention mechanisms. The BLSTM layers are shared for documents and summaries, both at sentence level (BLSTM with dimensionality \( d_w \)) and at document level (BLSTM2 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.

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 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.

3.2. Siamese Hierarchical Transformer Encoders

Siamese Hierarchical Transformer Encoders (SHTE) [4] is the instance of the general attentional framework when the encoders, both for sentence and document levels, are Transformer Encoders (TE) [20] shaped in a hierarchical way. 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.

Figure 3: SHTE Architecture.

It should be noted that a function \( P_1 \) is applied to each word to identify its position in the input of the sentence encoder, and a function \( P_2 \) is applied to each sentence to incorporate sentence positional information in the input of the document encoder.
In the inference process, 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. 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.

4. Corpora

We carried out the experimentation by using two different corpora for newspaper summarization. On the one hand, the CNN/DailyMail\(^3\) corpus was used in this work. This corpus is a set of articles from the CNN and DailyMail news websites, that was partitioned into 287,227 training (article, summary) pairs, 13,368 validation (article, summary) pairs and 11,490 test (article, summary) pairs. On the other hand, the NewsRoom\(^4\) corpus 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 NewsRoom corpus was partitioned into 995,041 training (article, summary) pairs, 108,837 validation (article, summary) pairs, and 108,862 test (article, summary) pairs.

5. Evaluation

To carry out the experimentation we fixed the hyperparameters of the models as follows. On the one hand, for 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. 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) [3]. On the other hand, for 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 \( \bar{N} = 2 \) sentences encoders, \( h = 6 \) heads [4]. To train the model on CNN/DailyMail corpus we used batches of 64 (article, summary) pairs (32 positive and 32 negative randomly sampled following an uniform distribution). For training on NewsRoom corpus, we used batches of 128 (article, summary) pairs. \( P_1 \) was defined as the identity function (we do not add positional information to the words inside each sentence) and \( P_2 \) was defined as the sine-cosine function.

Table 1 and Table 2 show the results obtained by the systems on the CNN/DailyMail and the NewsRoom corpora respectively. We also performed comparisons with other extractive and mixed systems. In order to reflect the categorization of the models, we show in the tables the category to which each model belongs. 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). The evaluation of the systems’ performance has been done by using three variants of the ROUGE measure [11]. Concretely, Rouge-N with unigrams and bigrams (R-1 and R-2) and Rouge-L (R-L) were used.

Table 1: 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, \( k \).

<table>
<thead>
<tr>
<th>System</th>
<th>Strategy</th>
<th>Category</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lead-3</td>
<td>Ext</td>
<td>Heuristic</td>
<td>40.24</td>
<td>17.70</td>
<td>36.45</td>
</tr>
<tr>
<td>SHA-NN-3</td>
<td>Ext</td>
<td>Attentional</td>
<td>39.99</td>
<td>17.75</td>
<td>36.27</td>
</tr>
<tr>
<td>SHTE-3</td>
<td>Ext</td>
<td>Attentional</td>
<td>39.96</td>
<td>17.60</td>
<td>36.19</td>
</tr>
<tr>
<td>PGem+Cov</td>
<td>Mix</td>
<td>Text generation</td>
<td>39.53</td>
<td>17.28</td>
<td>36.38</td>
</tr>
<tr>
<td>FastRL</td>
<td>Mix</td>
<td>Reinforcement</td>
<td>38.23</td>
<td>16.31</td>
<td>34.66</td>
</tr>
<tr>
<td>ECS-Ext</td>
<td>Ext</td>
<td>Oracle</td>
<td>41.70</td>
<td>18.60</td>
<td>37.80</td>
</tr>
<tr>
<td>ECS-Comp</td>
<td>Mix</td>
<td>Oracle</td>
<td>40.90</td>
<td>18.00</td>
<td>37.40</td>
</tr>
</tbody>
</table>

Table 2: Results on the full test of NewsRoom. The suffix of Lead, SHA-NN, and SHTE models refers to the number of extracted sentences, \( k \).

<table>
<thead>
<tr>
<th>System</th>
<th>Strategy</th>
<th>Category</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lead-3</td>
<td>Ext</td>
<td>Heuristic</td>
<td>30.66</td>
<td>21.09</td>
<td>28.35</td>
</tr>
<tr>
<td>SHA-NN-3</td>
<td>Ext</td>
<td>Attentional</td>
<td>28.99</td>
<td>19.42</td>
<td>26.69</td>
</tr>
<tr>
<td>SHTE-3</td>
<td>Ext</td>
<td>Attentional</td>
<td>29.19</td>
<td>19.37</td>
<td>26.81</td>
</tr>
<tr>
<td>PGem+Cov</td>
<td>Mix</td>
<td>Text generation</td>
<td>26.43</td>
<td>13.76</td>
<td>22.90</td>
</tr>
<tr>
<td>ECS-Ext</td>
<td>Ext</td>
<td>Oracle</td>
<td>39.50</td>
<td>27.90</td>
<td>36.26</td>
</tr>
<tr>
<td>ECS-Comp</td>
<td>Mix</td>
<td>Oracle</td>
<td>39.06</td>
<td>27.36</td>
<td>36.13</td>
</tr>
</tbody>
</table>

In both corpora, the results obtained by our systems are better than those obtained by widely used approaches such as Pointer-Gen+Cov [19] or by Reinforcement Learning systems such as FastRL [10]. The only systems that consistently outperform Lead, SHANN, and SHTE are those based on ExConSumm [14] both in the extractive and mixed variants (ECS-Ext and ECS-Comp). Differently to our systems, ExConSumm models are able to generate variable-length summaries depending on the input text.

6. Conclusions

In this work, we presented a formalization of a framework for extractive summarization that does not fall under the umbrella of the traditional extractive systems. The main objective of this work is to favor the development of new models and techniques within our proposed framework. In particular, under the proposed framework, the summarization systems are based on siamese architectures to learn directly relationships among articles and summaries. Also, 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 evaluated our systems and compared them 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 the improvement of our attentional framework.

7. Acknowledgments

This work has been partially supported by the Spanish Ministerio de Ciencia e Innovación and FEDER funds under project BEWORD (PDM2021-1260610B-C41). It is also partially supported by the Vice-rectorado de Investigación de la Universitat Politècnica de Valencia (PAID-11-21)

---

3https://cs.nyu.edu/~kcho/DMQA/
4https://lil.nlp.cornell.edu/newsroom/
8. References


