Neural Extractive Document Summarization

Lilly Kumari
ECE Dept
University of Washington
lkumari@uw.edu

Abstract

In this work, we present two Recurrent Neural Network (RNN) based sequence models for extractive summarization of documents. The models differ in how they encode the sentence. The first one (RNN-RNN) uses a bidirectional RNN to learn sentence embeddings at word level while the second one (CNN-RNN) uses a two-layer Convolutional Neural Network (CNN) with a max-over-time pooling operation. Both the models use a bidirectional RNN to learn document embeddings by functioning at sentence level. The formulation of this problem as a sequence classification task gives the additional advantage of easy interpretations in terms of scoring predictions based on abstract features such as information content, salience and novelty. We show that the CNN-RNN model performs better than the RNN-RNN model for the given task.

1 Introduction

With the immense growth of data owing to the accessibility of information to almost everyone, there is a great need for automatic summarization techniques. The news data is mostly textual and one can do extractive or abstractive summarization. Extractive summarization methods generally comprise of two tasks: sentence scoring and sentence selection. The summary is created by selecting the salient sentences without any modification. While abstractive summarization techniques are more complex because of their need to generate natural language to rewrite the sentences.

Despite the popularity of abstractive summarization techniques, extractive summarization methods are still attractive because they are less expensive, less complex and produce semantically correct summaries.

While using traditional features like (word probability, TF*IDF weights, sentence position and sentence length) to score sentences is quite conventional, modeling the very concept of sentence scoring using a deep neural network could be complicated and that too, when there's no way to interpret what exactly is happening.

In this work, the whole process of summarization is formulated as a sequence classification task which assigns scores to each sentences based on abstract features such as information content, salience, novelty (or redundancy given the current summary), absolute and relative position importance with respect to the document. In order to do this classification, we learn the document level representation using a bidirectional RNN. This RNN takes in the sentence representations which are generated in two ways: 1) using a two-layer CNN network followed by max-over-time pooling operation 2) using a bidirectional RNN. The classification task makes the process very interpretable and one can visualize the scores respective to different abstract features that we defined earlier. Finally, we show that the CNN-RNN model does out-perform the RNN-RNN model (addressed as SummaRuNNer in) (Nallapati et al., 2017).

2 Related Work

Extractive document summarization is extremely popular because of its less complex nature. Traditional extractive summarization techniques focus on two sub-tasks: sentence scoring and sentence selection. Sentence scoring aims at assigning a weight to a sentence based on it's salience and information content. Then, sentence selection determines which sentences should be extracted based on their scores, which is generally done heuristically.

There is a vast amount of literature on defining handcrafted surface features which could be used even in unsupervised setting that doesn't need any data annotation. Such features include term frequency (Luhn, 1958), sentence length (Cao et al., 2015), TF*IDF weights (Erkan and Radev, 2004), sentence position (Ren et al., 2017). These features can be used alone or combined with weights satisfying an objective function.

Machine learning techniques have also been used extensively, For example, (Kupiec et al., 1995) used a Naive Bayes classifier to learn surface features combinations to score sentences. (Conroy and Oleary, 2001) used a Hidden Markov Model for single document document summarization. (Carbonell and Goldstein, 1998) proposed the Maximal Marginal Relevance (MMR) as a heuristic for selecting and scoring sentences. (Lin and Bilmes, 2011) treated the sentence selection as a task for finding the optimal subset of sentences in a document and they proposed to use submodular function maximization using greedy algorithm to find that optimal subset of sentences which represent the summary of the document.

Some research is also done in formulating the summarization as a sequence classification problem. For example, (Shen et al., 2007) used conditional random fields to classify sentences sequentially into two classes: whether to include in summary or not. This was done using handcrafted surface features.

Recently, deep neural networks have been used widely for extractive summarization. (Yin and Pei, 2015) applied Convolutional Neural Networks for the task of multi-document extractive summarization. They used CNN to project sentences to a common continuous vector space and then select sentences by minimizing the cost based on their prestige and diverseness. (Cheng and Lapata, 2016) used an attention based encoder-decoder framework for single-document extractive summarization. They also released their dataset where each document sentence was tagged with a binary label based on its match with the highlighted summary.

In this work, we also use a neural network based approach for single document extractive summarization. We formulate the task of extractive summarization as a sequence classification problem, the same way (Nallapati et al., 2017) did. Each sentence in the document is visited sequentially and a binary decision about whether to include it in the summary or not is made based on its score with respect to several abstract features such as information content, salience with respect to the document, redundancy with respect to the dynamic summary etc.

In this work, we develop two different deep neural network based framework to do this sequence classification: RNN-RNN model and CNN-RNN model. Each model consists of two sub-components: word level encoder and sentence level encoder. The two models differ with respect to the first subcomponent while the second subcomponent is same in both of them. RNN-RNN model uses a bidirectional-GRU (Gated Recurrent Unit) based RNN to learn sentence representations while the CNN-RNN model uses a two-layer CNN (Convolutional Neural Network) to learn sentence representations by deriving both local and global features via the convolution and max-pooling operation respectively.

We use the same corpus (Daily Mail news) used by (Cheng and Lapata, 2016) and (Nallapati et al., 2017) for all our experiments as the corpus is huge which makes it really lucrative for training deep neural networks. We evaluate our models automatically using ROUGE metrics (Lin, 2004) on the Daily Mail test set. Experimental results show that the CNN-RNN model performs slightly better than the RNN-RNN model.

3 Approach

In this work, we treat extractive summarization as a sequence classification problem where each sentence in the document is visited sequentially and a binary decision to include it in summary is made accounting for the previous decisions made for that document. This approach was formulated in (Nallapati et al., 2017) who proposed using a GRU based RNN-RNN model for this task. Here, we design a second model called CNN-RNN model where the sentence representation is learnt using a two-layer convolutional neural network. The model architectures are discussed in the next sec-

tion.

4 Models

We experiment with two different models to perform the sequence classification and assign binary scores to the document sentences. The first model is a GRU based RNN-RNN model and the second one is a CNN-RNN model. Each of these models have two subcomponents: Sentence Encoder which works at word level and Document Encoder which works at sentence level. For word embeddings, we used Google News 100-dimensional word2vec pre-trained embeddings (wor).

4.1 RNN-RNN Model

In the RNN-RNN model, we use a GRU (Chung et al., 2014) based recurrent neural network (RNN) in both the sentence encoder and the document encoder. Figure 1 presents the network architecture of the RNN-RNN model. GRU-RNN is a recurrent network with two gates, **u** called the update rule and **r** called the reset gate. The update gate helps the model to determine how much of the past information (from previous time steps) needs to be passed along to the future. This eliminates the risk of vanishing gradient problem because the model can now decide to copy all the information from the past. The reset gate decides how much of the past information to forget. The formal equations of what exactly happens inside a GRU are described below:

$$\mathbf{u}_j = \sigma\left(\mathbf{W}_{ux}\mathbf{x}_j + \mathbf{W}_{uh}\mathbf{h}_{j-1} + \mathbf{b}_u\right) \quad (1)$$

$$\mathbf{r}_j = \sigma\left(\mathbf{W}_{rx}\mathbf{x}_j + \mathbf{W}_{rh}\mathbf{h}_{j-1} + \mathbf{b}_r\right) \quad (2)$$

$$\mathbf{h}'_j = \tanh\left(\mathbf{W}_{hx}\mathbf{x}_j + \mathbf{W}_{hh}\left(\mathbf{r}_j \odot \mathbf{h}_{j-1}\right) + \mathbf{b}_h\right) \quad (3)$$

$$\mathbf{h}_j = \left(1 - \mathbf{u}_j\right) \odot \mathbf{h}'_j + \mathbf{u}_j \odot \mathbf{h}_{j-1} \quad (4)$$

where **W**'s and **b**'s are the weight and bias parameters of the GRU-RNN. $\boldsymbol{h}_j$ is the real value hidden-state representation/vector of the GRU at time step *j* and $\boldsymbol{x}_j$ is the input vector and $\odot$ represents the Hadamard product. We choose to use GRU instead of LSTM (Long short term memory networks) because of less tensor operations involved in the former case, so the process is quite speedier.

The sentence encoder working at the word level takes the word embeddings of the sentence words as inputs. It computes the hidden state representation at each word position sequentially based on current word embeddings and previous hidden state, first from start to end i.e. in forward manner. It then does the same operation running in a backward manner, from the last word to the first. That's why this pair of forward and backward GRU-RNN's are referred to as bidirectional RNN.

The hidden states of the sentence encoder are then max-pooled and the resultant forward and backward layer representations are concatenated and fed as input to the document encoder which is also a bidirectional GRU-RNN functioning at sentence level. The hidden states of this RNN encode the sentence representations of the document. The document embedding is then modeled as the non-linear transformation of average pooling of the concatenated (forward and backward layers) hidden-state representation of the sentence-level RNN.

$$\mathbf{d} = \tanh\left(W_d \frac{1}{N_d} \sum\nolimits_{j=1}^{N_d} \left[\mathbf{h}_j^f, \mathbf{h}_j^b\right] + \mathbf{b}\right) \quad (5)$$

where $\mathbf{h}_j^f$ and $\mathbf{h}_j^b$ are the hidden state representations of the j^{th} sentence of the forward and backward sentence level encoder RNN's respectively. N_d is the number of sentences in the document.

4.2 CNN-RNN Model

This model uses a convolutional sentence encoder which works at word level. We designed a CNN based encoder for two reasons. Firstly, CNNs can be trained effectively and quickly without any long term dependencies in the model and secondly they are very successful at sentence level classification tasks such as sentiment classification (Kim, 2014). The CNN extracts local features via convolution and the max pooling layer generates global feature from the same. When multiple kernel widths are used, the local features can be captured very nicely in a n-gram fashion.

In the convolutional sentence encoder of the CNN-RNN model consisting of two *conv* blocks, we apply a temporal narrow convolution between $\mathbf{W} \in \mathbf{R}^{n \times n'}$ (where n' denotes the dimension of word embeddings and *n* is the number of words in the

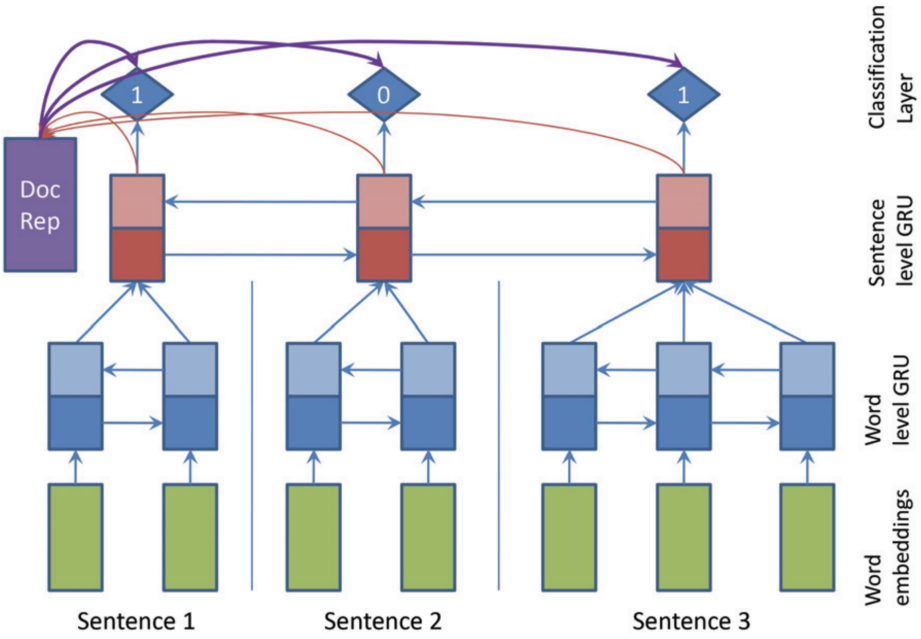

Figure 1: RNN-RNN Model

sentence) and a kernel $\mathbf{K} \in R^{c \times n'}$ of width c as follows:

$$\mathbf{f}_j^i = LeakyRelu\left(\mathbf{W}_{j:j+c-1} \odot \mathbf{K} + \mathbf{b}\right) \quad (6)$$

We then extract a single feature representing the sentence under kernel $\mathbf{K}$ using max-pooling over time as follows:

$$\mathbf{s}_{i,\mathbf{K}} = \max_j \mathbf{f}_j^i \quad (7)$$

Over all the feature maps, we perform this max-pooling over time operation which gives us a sentence representation (vector) under the Kernel $\mathbf{K}$ of width c. We use multiple kernels with different widths whose resultant representations are finally concatenated to get a final sentence representation.

For document encoder, we use a GRU based bidirectional Recurrent Neural Network (Chung et al., 2014) which runs at sentence level and takes the final sentence representations of the CNN-based sentence encoder as input. The functioning and modeling of document representation is same as described in the section 4.1 in equation 5. The network architecture is presented in Figure 2.

4.3 Classification Task

Once we get the sentence representations and the document embedding, we move on to the classification task. Each sentence in the document is visited sequentially during a second pass, where a logistic layer with sigmoid activation makes a binary decision about whether the sentence is a part of the summary or not. This logistic layer is broken into several components which represent separate abstract features as shown in Figure 3.

where y_j is the binary label indicating whether j^{th} sentence belongs to the summary or not. $\mathbf{h}_j$ is the sentence representation given by a non-linear transformation of the concatenated hidden states (forward and backward) representations of the sentence level RNN. $\mathbf{S}_j$ is the dynamic summary calculated till the j^{th} sentence which is simply a weighted summation of all sentence level hidden state embeddings visited till the j^{th} sentence. The weights here represent the probabilities of summary membership of respective sentences. Its formulated as follows:

$$\mathbf{s}_j = \sum_{i=1}^{j-1} \mathbf{h}_i P\left(y_i = 1 | \mathbf{h}_i, \mathbf{s}_i, \mathbf{d}\right) \quad (8)$$

5 Experiments

5.1 Dataset

We use the Daily Mail, a large news dataset released by (Hermann et al., 2015) which consists of $\approx$ 220k news articles. (Cheng and Lapata, 2016)

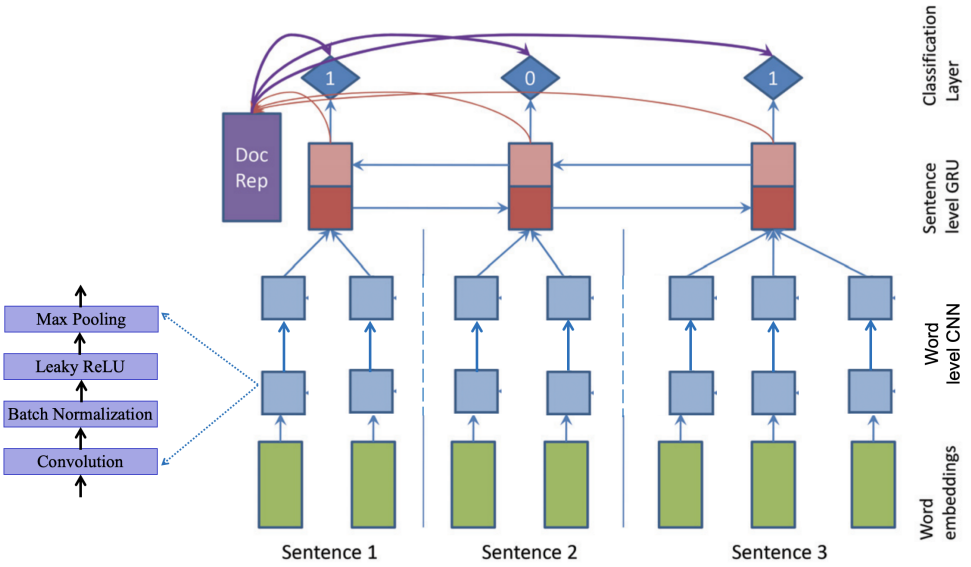

Figure 2: CNN-RNN Model

$$\underbrace{P\left(y_j = 1 \mid \mathbf{h}_j, \mathbf{s}_j, \mathbf{d}\right)}_{\text{Prob. of } j^{th} \text{ sentence to be a part of summary}} = \sigma\Big(\underbrace{W_c \mathbf{h}_j}_{\text{Information content of the sentence}} + \underbrace{\mathbf{h}_j^T W_s \mathbf{d}}_{\text{Salience of sentence wrt document}} - \underbrace{\mathbf{h}_j^T W_r \tanh\left(\mathbf{s_j}\right)}_{\text{Redundancy of sentence given current summary}} + \underbrace{W_{ap} \mathbf{p}_j^a + W_{rp} \mathbf{p}_j^r}_{\text{Absolute \& Relative position importance of sentence wrt document}} + b\Big)$$

Figure 3: Classification task for both the models

annotated this dataset with binary labels (0 or 1) for the task of extractive summarization. They designed a heuristics-based system which determined whether a docuement sentence matched a highlight and should be labeled with 1 (that is in the summary) or 0 (if not a part of summary). The heuristics account for the position of the sentence in the document, the unigram and bigram overlap between the document sentences and the highlights, the number of named entities which appear together in the document sentence and the highlights etc. We have used their dataset for our use-case as it matches with their task of scoring each sentence and assigning a binary label to it about being a part of summary or not.

Table 1 shows the data statistics of the Daily Mail dataset. Data preprocessing - the sentences were tokenized using *nltk* word tokenizer. All words were transformed into lower case, and Google News 100-dimensional word2vec pre-trained embeddings (wor) were used to encode the words for initialization.

5.2 Implementation Details

Data Preprocessing - we tokenized the sentences using the *nltk* word-tokenizer. We then transformed all the words into lower case. From the Google News 100-d word embeddings, we compiled a list of all the words which were mapped to an ID. The words in the Daily Mail refined corpus which were unseen in the word2vec words list were marked *unknown*. This way the vocabulary was limited to $\approx$ 150K words. The sentences over a length of 50 words were truncated and the maximum number of sentences in a document was limited to 100, in order to speed up the computation.

We implemented the neural networks in PyTorch (Paszke et al., 2017). The hidden state size for all the GRU based RNN was fixed at 200. For the CNN based word level encoder, we used three different kernel widths (3, 4, 5). We used a batch size of 128 for training. The learning rate was set at 0.001. For optimization, we used *adam* optimizer and in order to regularize the weights, we employed gradient clipping with a max norm of 1.0.

Training of the models was done on Azure VM NC6 (12 GB GPU memory size). The RNN-RNN model training took around four days for 5 epochs while the CNN-RNN model training took around

Daily Mail	Train	Val	Test
# (Documents)	193983	12147	10350
# (Summary / Doc)	1	1	1
Avg Doc Len (sentence)	26.12	25.23	25.55
Avg Summary Len (sentence)	3.81	4.27	4.03
Avg Doc Len (word)	750.98	729.34	739.34
Avg Summary Len (word)	55.4	60.96	57.77

Table 1: Data statistics of Daily Mail Dataset

three days for 5 epochs. While testing, instead of choosing a probability threshold for selecting sentences worthy of summary membership, we used word length based constraint. So, we selected sentences sorted by decreasing predicted probabilities (classification score which is mentioned in Figure 3) until we exceeded the length limit for the ROUGE evaluation metric.

5.3 Qualitative Results

Figure 4, 5, 6, 7 show the predicted scores against different abstract features and overall probability of summary membership for a sentence.

Figure 8, 9 show the predicted summary which is highlighed in the document against the gold summary in 1st column. The sentences highlighted in blue in the 2nd column are predicted to be a part of the final summary by the RNN-RNN model while the sentences highlighted in red in 3rd column are predicted to be a part of final summary by the CNN-RNN model.

In Figure 8, it can be seen that the 4th and 5th sentences from the beginning that are selected by the RNN-RNN model are redundant with respect to each other. While the CNN-RNN model does a better job at selecting the second sentence from the beginning which is a part of the gold summary/highlight. For another document in Figure 9, the CNN-RNN model again performs a better job by selecting more diverse and information rich sentences for summary membership.

5.4 Quantitative Results

We report the performance of both models with respect to ROUGE (Lin, 2004) Recall and F-score metrics at 75 words. The scores are reported in Table 2 from Rouge-1, Rouge-2 and Rouge-L which are computed using the matches of unigrams, bigrams and longest common sub-sequences respectively with respect to the ground truth summaries.

	RNN-RNN	CNN-RNN
Rouge-1 (R)	0.25840	**0.26192**
Rouge-1 (F)	0.25803	**0.26142**
Rouge-2 (R)	0.11316	**0.11701**
Rouge-2 (F)	0.11321	**0.11700**
Rouge-L (R)	0.13784	**0.14005**
Rouge-L (F)	0.16294	**0.16542**

Table 2: Performance of models on Daily Mail test set using limited length variants of Rouge at 75 bytes

It can be observed that the CNN-RNN model performs relatively better than the RNN-RNN model with respect to all the evaluation metrics.

6 Discussions

The CNN-RNN model is more efficient and fast to train because of lesser number of parameters. Since at a sentence level, there's not a great need to model long term dependencies within the sentence. So, one can simplify the process by using a convolutional neural network based word level encoder. The convolution operation captures the local features while max-over-time pooling operation helps in capturing the global representation. When this operation is carried out with different kernel widths, we can get multiple sentence representations for each kernel (with distributions distilled together just like a traditional n-gram language model) which could then be used together to represent the sentence.

It can be observed from Figure 5 and 7 that the CNN-RNN model assigns a higher salience (of a sentence with respect to the entire document representation) to the sentences which contain named entities.

Gold Summary: **London - born footballer Harry Kane scored just 80 seconds into his debut. The 21 - year - old scored a header in England's qualifier against Lithuania. Born in ridgeway , he played for ridgeway rovers and arsenal as a child. He was snapped up by Tottenham at the age of 11 and spent time on loan. Former coaches have described him as 'level headed' and 'humble'. He has been dating childhood sweetheart Katie Goodland for three years.**	**Content**	**Salience**	**Novelty**	**Probability**
Harry Kane lived every boy 's childhood dream tonight by scoring on debut for England with just his third touch of the ball.	0.56534576	0.95258963	0.5	**0.9706617**
He also wrote his name into the record books - it was the third fastest goal scored by an England debutant behind jack cock in 1919 and Bill Nicholson , who netted after just 19 seconds on his England debut against Portugal in 1951 to set the record.	0.57033545	0.93183666	0.34098586	0.9008612
He spent a year in the red half of north London before returning to his amateur childhood side after arsenal let him leave.	0.52774364	0.72542244	0.23136918	0.3957638
Born in Chingford , London , Harry Kane was playing for Ridgeway Rovers in east London - the same youth side David Beckham and Andros Townsend played for - when he was scouted by gunners spies and joined the club as an eight - year – old.	0.5741812	0.9552613	0.3360863	**0.9238422**
Last night Harry Kane wrote his name into the record books - his goal was the third fastest goal scored by an England debutant behind jack cock in 1919 and Bill Nicholson , who netted after just 19 seconds on his England debut against Portugal in 1951 to set the record.	0.5933621	0.9867106	0.17801444	**0.9451442**
His time is followed by Jack Cock (right) , who netted after 30 seconds in 1919.	0.50565374	0.36351898	0.05870257	0.03247365

Figure 4: Classification output of RNN-RNN model on 1st chosen document, the sentences against probabilities highlighted in bold comprise the final summary

7 Conclusion

Conventional approaches to extractive single document summarization consist of two sub-tasks: sentence scoring followed by sentence selection. In this work, we treat extractive summarization as a sequence classification problem where we visit each sentence in a document sequentially and assign it scores on the basis of abstract features, the values of which determines its summary membership. The sentence selection is done based on these scores until word length constraint is satisfied. This approach makes the system very interpretable as one can visualize why a sentence is significant enough to be a part of the summary. We compare two different deep neural network based approaches and show that the CNN-RNN model slightly outperforms the RNN-RNN model.

8 Acknowledgement

We thank Prof. Yejin Choi and our TAs for their helpful comments and discussions.

References

https://code.google.com/archive/p/word2vec/.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *AAAI*, pages 2153–2159.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

John M Conroy and Dianne P Oleary. 2001. Text summarization via hidden markov models. In *SIGIR*, pages 406–407.

Gunes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. In *Journal of Artificial Intelligence Research*, pages 457–479.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems (NIPS)*.

Gold Summary: **London - born footballer Harry Kane scored just 80 seconds into his debut. The 21 - year - old scored a header in England's qualifier against Lithuania. Born in ridgeway , he played for ridgeway rovers and arsenal as a child. He was snapped up by Tottenham at the age of 11 and spent time on loan. Former coaches have described him as 'level headed' and 'humble'. He has been dating childhood sweetheart Katie Goodland for three years.**	**Content**	**Salience**	**Novelty**	**Probability**
Harry Kane lived every boy 's childhood dream tonight by scoring on debut for England with just his third touch of the ball.	0.94274765	0.7035911	0.5	**0.9400062**
He also wrote his name into the record books - it was the third fastest goal scored by an England debutant behind jack cock in 1919 and Bill Nicholson , who netted after just 19 seconds on his England debut against Portugal in 1951 to set the record.	0.976714	0.6921421	0.06830377	**0.9033845**
He spent a year in the red half of north London before returning to his amateur childhood side after arsenal let him leave.	0.9811002	0.30589187	0.01711672	0.482866
Born in Chingford , London , Harry Kane was playing for Ridgeway Rovers in east London - the same youth side David Beckham and Andros Townsend played for - when he was scouted by gunners spies and joined the club as an eight - year – old.	0.98096126	0.5182313	0.04245803	**0.8317192**
Last night Harry Kane wrote his name into the record books - his goal was the third fastest goal scored by an England debutant behind jack cock in 1919 and Bill Nicholson , who netted after just 19 seconds on his England debut against Portugal in 1951 to set the record.	0.992353	0.00042841	0.98364013	0.78702044
His time is followed by Jack Cock (right) , who netted after 30 seconds in 1919.	0.99056375	0.0000925	0.83754385	0.04996072

Figure 5: Classification output of CNN-RNN model on 1st chosen document, the sentences against probabilities highlighted in bold comprise the final summary

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *SIGIR*, pages 68–73.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *ACL*, pages 510–520.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. In *IBM Journal of research and development*, pages 159–165.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Pengjie Ren, Zhumin Chen, Zhaochun Ren, Furu Wei, Jun Ma, , and Maarten de Rijke. 2017. Leveraging contextual sentence relations for extractive summarization using a neural attention model. In *SIGIR*, pages 95–104.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In *IJCAI*, volume 7, pages 2862–2867.

Wenpeng Yin and Yulong Pei. 2015. Optimizing sentence modeling and selection for document summarization. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Gold Summary: **Remains were found in the 3300 block of Bonnie hill drive at 4pm Tuesday. Police are trying to determine whether they are human.**	Content	Salience	Novelty	Probability
Police have established a crime scene in the Hollywood hills after a hiker discovered some skeletal remains in a wooded area on Tuesday.	0.51564693	0.6119396	0.5	**0.68630964**
The remains were found in the afternoon in the 3300 block of Bonnie hill drive.	0.52162933	0.7008587	0.45773074	0.67694616
A LAPD spokesperson told CBS Los Angeles that a call came in from the hiker around 4 p.m. scene : a hiker came across some skeletal remains in the 3300 block of Bonnie hill drive in the Hollywood hills around 4pm on Tuesday, police say police have not yet determined if the remains are definitely human.	0.57770115	0.9321809	0.4273235	**0.9188876**
They are searching the surrounding area for any clues.	0.4640495	0.14402336	0.39197418	0.06027255
The crime scene is close to the home of Andrew Getty, grandson of oil tycoon J Paul Getty , who was found dead inside his home today in apparently suspicious circumstances.	0.5430806	0.7598787	0.37755367	0.5795688
Investigation : police are now trying to verify whether the remains are human a hiker found skeletal remains this afternoon in the 3300 block of Bonnie hill drive in the Hollywood hills.	0.54729396	0.8477123	0.45758173	**0.7900325**

Figure 6: Classification output of RNN-RNN model on 2nd chosen document, the sentences against probabilities highlighted in bold comprise the final summary

Gold Summary: **Remains were found in the 3300 block of Bonnie hill drive at 4pm Tuesday. Police are trying to determine whether they are human.**	Content	Salience	Novelty	Probability
Police have established a crime scene in the Hollywood hills after a hiker discovered some skeletal remains in a wooded area on Tuesday.	0.9024977	0.18812832	0.5	**0.44430983**
The remains were found in the afternoon in the 3300 block of Bonnie hill drive.	0.9344763	0.25192922	0.2300925	**0.5597772**
A LAPD spokesperson told CBS Los Angeles that a call came in from the hiker around 4 p.m. scene : a hiker came across some skeletal remains in the 3300 block of Bonnie hill drive in the Hollywood hills around 4pm on Tuesday, police say police have not yet determined if the remains are definitely human.	0.9705156	0.3133538	0.3820622	**0.92638755**
They are searching the surrounding area for any clues.	0.94529176	0.03476408	0.01003024	0.01179064
The crime scene is close to the home of Andrew Getty, grandson of oil tycoon J Paul Getty , who was found dead inside his home today in apparently suspicious circumstances.	0.9696693	0.07102659	0.08936705	0.31846005
Investigation : police are now trying to verify whether the remains are human a hiker found skeletal remains this afternoon in the 3300 block of Bonnie hill drive in the Hollywood hills.	0.9765776	0.01294329	0.29380456	0.2812145

Figure 7: Classification output of CNN-RNN model on 2nd chosen document, the sentences against probabilities highlighted in bold comprise the final summary

Gold Summary: The show entitled The Seven Year Switch is an eight episode series in which couples live , eat , and even sleep with a new significant other. The show 's concept is based on the idea that after seven years of marriage , spouses often become restless and regret their decision to wed. After the switch is complete the couples will be able to decide whether or not they want to stay together or be with someone else.	**A new television show will attempt to rescue failed marriages by arranging a full - on two week spouse switch for couples , FYI announced yesterday**. The show entitled The Seven Year Switch is an eight episode series in which couples live , eat , and even sleep with a new significant other. The show is set to premiere this summer. **Deadline reports that the series is named after the 1955 Broadway show The Seven Year itch that was later developed into the film starring Marilyn Monroe and Tom Ewell. The Seven Year Itch : the series is named after the 1955 Broadway show the seven year itch that was later developed into the film starring Marilyn Monroe (left) and Tom Ewell (right) the show 's concept is based on the idea that after seven years of marriage , spouses often become restless and regret their decision to wed. The program chooses four couples who have reached the seven year mark and they guide them through the switch.** After the switch is complete the couples will be able to decide whether or not they want to stay together or be with someone else. Gena McCarthy , SVP Programming and Development at FYI , said the show is testing if separating people makes them miss one another even more than they imagined was possible. The show for FYI is produced by Kinetic Content who also produce the love - based show Married at First Sight. Married at First Sight pairs a couple based on their scientific match and they then live together for four weeks. At the end of the four weeks they decide if they want to stay together or divorce. Gena McCarthy, SVP Programming and Development at FYI , said the show is testing if separating people makes them miss one another even more than they imagined was possible.	**A new television show will attempt to rescue failed marriages by arranging a full - on two week spouse switch for couples , FYI announced yesterday. The show entitled The Seven Year Switch is an eight episode series in which couples live , eat , and even sleep with a new significant other.** The show is set to premiere this summer. Deadline reports that the series is named after the 1955 Broadway show The Seven Year itch that was later developed into the film starring Marilyn Monroe and Tom Ewell. **The Seven Year Itch : the series is named after the 1955 Broadway show the seven year itch that was later developed into the film starring Marilyn Monroe (left) and Tom Ewell (right) the show 's concept is based on the idea that after seven years of marriage , spouses often become restless and regret their decision to wed. The program chooses four couples who have reached the seven year mark and they guide them through the switch.** After the switch is complete the couples will be able to decide whether or not they want to stay together or be with someone else. Gena McCarthy , SVP Programming and Development at FYI , said the show is testing if separating people makes them miss one another even more than they imagined was possible. The show for FYI is produced by Kinetic Content who also produce the love - based show Married at First Sight. Married at First Sight pairs a couple based on their scientific match and they then live together for four weeks. At the end of the four weeks they decide if they want to stay together or divorce. Gena McCarthy, SVP Programming and Development at FYI , said the show is testing if separating people makes them miss one another even more than they imagined was possible.

Figure 8: Summary results for a document against the ground truth highlight (1st column), 2nd column - RNN-RNN model, 3rd column - CNN-RNN model

Gold Summary: Mark Zuckerberg posted about new premises in Menlo Park , California. 430,000sq ft office will mainly be one enormous room for employees. Said that the cavernous open room will be perfect engineering space.	**Facebook has unveiled an enormous new office in Silicon Valley, which is large enough to hold a 9 - acre park on its roof , and will house 2,800 workers in a single room.** The new building in Menlo Park , California , measures 430,000 square feet and apparently has the ' largest open floor plan in the world. The social network 's newest location was revealed by founder Mark Zuckerberg , who posted an aerial view on the building on his public page. We wanted our space to create the same sense of community and connection among our teams that we try to enable with our services across the world. To do this , we designed the largest open floor plan in the world - a single room that fits thousands of people. **Lush : the office features a 9 - acre park on top of its 430,000 square feet of indoor space California skies : the office is not far from Facebook 's current Silicon Valley campus - to which it is connected by tunnel relaxing : the waterfront office was documented by a small army of instagram photographers ahead of opening day plans : this architect 's model of the office shows the single , long , open - plan room inside the office as well as the huge open space , which has yet to be pictured , there will be smaller spaces for meetings.** One Facebook employee posted an image of Mark Zuckerberg and Facebook Chief Technology Officer Mike Schroepfer meeting in one room , which had been filled with plastic balls - seemingly a first - day prank. According to local news outlet The Almanac , 2,800 engineers will eventually work in the main , open room. They have reportedly not all been hired yet. **Artsy : the warehouse has been crammed full of art pieces by bay area sculptors and painters , like the above trendy : the building will one day play host to 2,800 Facebook employees - but it is not yet at capacity construction on the new building , which is not far from Facebook 's existing campus , and will be connected via tunnel , was started in September 2014**. And despite the simple philosophy behind it , the new Facebook building still has plenty of capacity to impress. An army of instagram stars was unleashed on the building , and captured its quirky stairways , unusual paintjobs and sculpture pieces made by bay area artists .	**Facebook has unveiled an enormous new office in Silicon Valley, which is large enough to hold a 9 - acre park on its roof , and will house 2,800 workers in a single room. The new building in Menlo Park , California , measures 430,000 square feet and apparently has the ' largest open floor plan in the world.** The social network 's newest location was revealed by founder Mark Zuckerberg , who posted an aerial view on the building on his public page. We wanted our space to create the same sense of community and connection among our teams that we try to enable with our services across the world. To do this , we designed the largest open floor plan in the world - a single room that fits thousands of people. **Lush : the office features a 9 - acre park on top of its 430,000 square feet of indoor space California skies : the office is not far from Facebook 's current Silicon Valley campus - to which it is connected by tunnel relaxing : the waterfront office was documented by a small army of instagram photographers ahead of opening day plans : this architect 's model of the office shows the single , long , open - plan room inside the office as well as the huge open space , which has yet to be pictured , there will be smaller spaces for meetings.** One Facebook employee posted an image of Mark Zuckerberg and Facebook Chief Technology Officer Mike Schroepfer meeting in one room , which had been filled with plastic balls - seemingly a first - day prank. According to local news outlet The Almanac , 2,800 engineers will eventually work in the main , open room. They have reportedly not all been hired yet. Artsy : the warehouse has been crammed full of art pieces by bay area sculptors and painters , like the above trendy : the building will one day play host to 2,800 Facebook employees - but it is not yet at capacity construction on the new building , which is not far from Facebook 's existing campus , and will be connected via tunnel , was started in September 2014. And despite the simple philosophy behind it , the new Facebook building still has plenty of capacity to impress. An army of instagram stars was unleashed on the building , and captured its quirky stairways , unusual paintjobs and sculpture pieces made by bay area artists .

Figure 9: Summary results for a document against the ground truth highlight (1st column), 2nd column - RNN-RNN model, 3rd column - CNN-RNN model