Deep Learning based Text Summarization: Approaches, Databases and Evaluation Measures

Akanksha Joshi[1,2,3], E. Fidalgo[1,2], E. Alegre[1,2]
[1] Universidad de León, Spain
[2] Researcher at INCIBE (Spanish National Institute of Cybersecurity), León, Spain
[3]CDAC, Mumbai, India
{ajos,eduardo.fidalgo, enrique.alegre}@unileon.es

Abstract. The ever-growing volume of documents available online have compelled the need for smart text summarization systems and due to their state-of-the-art performances, deep learning approaches for text summarization has gained momentum recently. The objectives of this paper are threefold; first, we present a comprehensive review of the state-of-the-art approaches in automatic summarization research. Next, we provide an overview of the databases available for training and benchmarking the text summarization algorithms. Furthermore, we discuss the available evaluation measures for quality assessment of system generated summary. A performance comparison of the methods demonstrates that Convolutional Neural Network (CNN) based approach achieve better accuracy for extractive summarization whereas, for abstractive summarization, Deep Recurrent Generative Networks (DRGN) provides state-of-the-art results.

Keywords. Deep Learning, Text Summarization, Abstractive, Extractive, CNN

1. Introduction

There are billions of documents available online such as news, scientific articles, and blogs that contain longer pieces of text which can be summarized to provide short and meaningful information to the users. The text summary can be extractive or abstractive; in extractive the summary is generated by selecting salient sentences of the document and in abstractive, a summary is generated by paraphrasing the key concepts of the document. The main purpose of this survey is to provide a review of three major aspects of text summarization research including recent approaches, databases and evaluation measures. To the best of our knowledge, no such survey is available discussing all this at one place. The major contributions of this paper include:

- A survey of most influential work (based on yearly citations[1]) based on deep learning (DL) approaches for automatic text summarization.
- A review of well-known research databases for training and evaluation of summarization algorithms.
- An overview of available famous evaluation measures to benchmark the quality of system generated summary against the ground truth data.

[1] Obtained from Google Scholar

The rest of the paper is organized as follows: Section 2 describes summarization algorithms for text. In Section 3, the databases are discussed. Next, we explain the available evaluation measures in Section 4. Afterwards, we provide a comparative analysis of discussed methods in Section 5, followed by conclusion in Section 6.

2. Deep Learning based Automatic Text Summarization Approaches

In Table 1, we provide the summary of the automatic summarization algorithms.

Table 1: Summary of deep learning based approaches

Authors	DL	Summary of Approaches
Cheng et al. [10]	CNN, RNN	Authors proposed a hierarchical framework which makes use of CNN to generate sentence representation and RNN to represent the document.
Nallapati et al. [12]	RNN	The model consists of a two-layer bi-directional GRU-RNN sequence model for extractive summarization of documents.
Cao et al. [6]	CNN	The CNN is used to generate sentence embeddings and pooling applied with attention mechanism to combine sentence embeddings to form document embeddings in same latent space.
Cao et al. [8]	CNN	Authors developed a novel system called PriorSum that applied the enhanced Convolutional Neural Networks to capture the summary prior and concatenated them with document dependent features under a regression framework.
Yin and Pei [5]	CNN	The CNN based Language models are used for sentence representation and an optimization based on prestige and diversity is used to generate the summary.
Cao et al [7]	R2N2	Their approach is based on Recursive Neural Networks to rank sentences for multi document summarization.
Kageback [11]	RAE, FFNN	Authors exploited continuous vector representations for semantically aware sentence representations to generate an optimal summary.
Rush et al [3]	NNLM	They utilized Neural attention model with a contextual input encoder to generate abstractive summaries.
Chopra et al.[4]	RNN	Authors proposed a novel convolutional attention-based conditional recurrent neural network model for abstractive sentence summarization.
Nallpati et al. [1]	RNN	The encoder consists of bidirectional GRU-RNN and decoder consists of unidirectional GRU-RNN with an attention mechanism over source hidden states.
Ma et al.[9]	NPLM	Authors proposed a Neural Document Model based on Word and paragraph vectors.
Li at al. [13]	VAE, DRGN	To improve the performance and quality of abstractive summarization, the approach is based on the sequence to sequence oriented encoder-decoder framework equipped with a latent structure modeling component of Variational Auto-Encoders.
DL- Deep Learning, CNN- Convolutional Neural Network, RNN- Recurrent Neural Network, VAE-Variational Autoencoder, NNLM-Neural Network Language Model, DRGN-Deep Recurrent Generative Decoder, NPLM- Neural Probabilistic Language Model		

3. Datasets for Text Summarization Research

Since deep learning systems require an enormous amount of data for training, many authors utilized Gigaword dataset [16] for abstractive as well as extractive summarization. Recently CNN and DailyMail news datasets [15] emerged as good source for training and testing purposes. DUC (Document Understanding Conference) datasets [14] such as DUC 2001, DUC 2002, DUC 2003, DUC 2004, DUC 2005, DUC 2006, and DUC 2007 are the benchmark datasets available in the literature for reporting accuracies on different summarization tasks.

4. Evaluation Measures

The evaluation approaches can be broadly divided into intrinsic and extrinsic [17] and various different types of measures under these categories are presented in Table 2 & 3.

Table 2: List of intrinsic evaluation measures

S.No	Intrinsic Evaluation	
1	Text Quality Measures	Grammaticality
2		Non-Redundancy
3		Reverential Clarity
4		Structure and Coherence
5	Content Evaluation	Precision, Recall and F-score
6		Relative utility
7		Cosine Similarity
8		Unit overlap
9		Longest common subsequence
10		n-gram matching (ROUGE)
11		Pyramids
12		LSA-based measure

Table 3: List of extrinsic evaluation measures

S.No	Extrinsic Evaluation
1	Document Categorization
2	Information retrieval
3	Question Answering

5. Performance Comparison of Text Summarization Algorithms

In this section, we conduct a performance analysis of deep learning based text summarization systems and report their ROUGE scores obtained from literature in Table 4. For the task of single document extractive text summarization, Cheng at al. [10] reports the best system on DUC 2002 single document summarization task. For DUC 2004 multi-document summarization, the language model based method of Yin and Pei [5] achieves the best result. For abstractive text summarization on DUC 2004 and Gigaword both, the Deep Recurrent Generative Networks proposed by Li et al.[13]

Table 4: Performance Comparison Deep Learning based Text Summarization Systems

Approaches	Single/ Multi/Sentence	Extractive/ Abstractive	Evaluation Dataset	Rouge-1	Rouge-2	Rouge-L	Rouge-SU4
Cheng et al. [10]	single	abstractive	CNN, DailyMail	42.20	17.30	-	34.8
		extractive	DUC2002	47.40	23.00	43.5	-
SummaRunner [12]		abstractive	CNN, DailyMail	42..20	17.3	34.4	-
		extractive	DUC2002	47.40	24.0	43.8	-
Kageback [11]			Opinosis	34.72	5.89	-	12.38
Nallpati et al.[1]		abstractive	DUC 2004	28.35	9.46	24.59	-
			CNN, DailyMail	32.49	11.84	29.47	-
DRGN[13]			DUC2004	31.79	10.75	27.48	-
			Gigaword	36.27	17.57	33.62	-
Attsum[6]	multi	extractive	DUC 2005	37.01	6.99	-	-
			DUC 2006	40.90	9.40	-	-
			DUC 2007	43.92	11.55	-	-
Cao et al[8]			DUC 2001	35.98	7.89	-	-
			DUC 2002	36.63	8.97	-	-
			DUC 2004	38.91	10.07	-	-
Yin and Pei [5]			DUC 2002	51.01	26.97	-	29.4
			DUC2004	40.90	10.70	-	14.9
Cao et al[7]			DUC2001	36.91	7.87	-	-
			DUC2002	37.96	8.88	-	-
			DUC2004	38.78	9.86	-	-
Ma et al.[9]			DUC2006	42.19	9.31	-	15.17
			DUC2007	43.42	10.50	-	16.24
Rush et al[3]	sentence	abstractive	DUC 2004	28.18	8.49	23.81	-
			Gigaword	31.00	12.65	28.34	-
Chopra et al.[4]			DUC2004	28.97	8.26	24.06	-
			Gigaword	33.78	15.97	31.15	-

perform better than other approaches [1, 3, 4].

6. Conclusion

In this paper, we presented a review of deep learning based famous text summarization systems, benchmark databases for training and evaluation, and well-known measures for summary quality assessment. We also provided an analysis of accuracies of various text summarization methods and reported their ROUGE scores obtained from the literature. One of the major findings of this study is that people have exploited all major deep learning techniques for summarization task such as auto encoder, CNN, RNN. It is also observed that the CNN based approaches provide good accuracy for extractive summarization, and RNN techniques give better results for abstractive summarization. Another key observation is that mainly the news datasets are available and used in various researches, thus performance in other domains is still questionable.

References

[1] Nallapati, R.; Zhou, B.; dos Santos, C.; Gulcehre, C.; and Xiang, B. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. The SIGNLL Conference on Computational Natural Language Learning.

[2] Nallapati, R.; Zhou, B.; and Xiang, B. 2016. Sequence-to-sequence rnns for text summarization. International Conference on Learning Representations, Workshop track A.N. Author, Article title, Journal Title 66 (1993), 856–890.

[3] Rush, A. M.; Chopra, S.; and Weston, J. 2015. A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685.

[4] Chopra S, Auli M, Rush AM. 2016 Abstractive sentence summarization with attentive recurrent neural networks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, pp 93–98

[5] Yin, W., and Pei, Y. 2015. Optimizing sentence modeling and selection for document summarization. In Proceedings of the 24th International Conference on Artificial Intelligence, 1383–1389. AAAI Press.

[6] Cao, Z.; Li, W.; Li, S.; and Wei, F. 2016. Attsum: Joint learning of focusing and summarization with neural attention. arXiv preprint arXiv:1604.00125.

[7] Cao Z, Wei F, Dong L, Li S, Zhou M (2015) Ranking with recursive neural networks and its application to multi-document summarization. In: AAAI conference on artificial intelligence

[8] Cao Z, Wei F, Li S, Li W, Zhou M, Wang H (2015) Learning summary prior representation for extractive summarization. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers). Association for Computational Linguistics, Beijing, pp 829–833

[9] Ma, S., Deng, Z. H., & Yang, Y. (2016). An Unsupervised Multi-Document Summarization Framework Based on Neural Document Model. In COLING (pp. 1514-1523).

[10] Cheng, J., and Lapata, M. 2016. Neural summarization by extracting sentences and words. 54th Annual Meeting of the Association for Computational Linguistics.

[11] Kageback, M.; Mogren, O.; Tahmasebi, N.; and Dubhashi, D. 2014. Extractive summarization using continuous vector space models. 31–39

[12] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In Association for the Advancement of Artificial Intelligence

[13] Li, P., Lam, W., Bing, L., & Wang, Z. 2017. Deep Recurrent Generative Decoder for Abstractive Text Summarization. arXiv preprint arXiv:1708.00625.

[14] http://duc.nist.gov/data.html (Accessed on September 20, 2017)

[15] http://cs.nyu.edu/~kcho/DMQA/ (Accessed on September 20, 2017)

[16] https://catalog.ldc.upenn.edu/ldc2003t05 (Accessed on September 20, 2017)

[17] Spark-Jones, K. and J. R. Galliers (1995). Evaluating Natural Language Processing Systems: An Analysis and Review. Number 1083 in Lecture Notes in Artificial Intelligence. Springer.