Neural Abstractive Text Summarization

Gaetano Rossiello

1st year PhD Student
Department of Computer Science
University of Bari "Aldo Moro"
gaetano.rossiello@uniba.it
Advisor: Giovanni Semeraro

Abstract. Abstractive text summarization is a complex task whose goal is to generate a concise version of a text without necessarily reusing the sentences from the original source, but still preserving the meaning and the key contents. We address this issue by modeling the problem as a sequence to sequence learning and exploiting Recurrent Neural Networks (RNNs). This work is a discussion about our ongoing research on abstractive text summarization, where our aim is to investigate methods to infuse prior knowledge into deep neural networks. We believe that these approaches can obtain better performance than the state-of-the-art models for generating well-formed and meaningful summaries.

Keywords: Natural Language Processing, Abstractive Text Summarization, Recurrent Neural Networks, Deep Learning

1 Introduction

Information overload is a problem in modern digital society caused by the explosion of the amount of information produced on both the World Wide Web and the enterprise environments. For textual information, this problem is even more significant due to the high cognitive load required for reading and understanding a text. Automatic text summarization tools are thus useful to quickly understand a large amount of information.

The goal of summarization is to produce a shorter version of a source text by preserving the meaning and the key contents of the original. This is a very complex problem since it requires to emulate the cognitive capacity of human beings to generate summaries. For this reason, text summarization poses open challenges in both natural language understanding and generation. Due to the difficulty of this task, research works in literature focused on the *extractive* aspect of summarization, where the generated summary is a selection of relevant sentences from the source text in a copy-paste fashion [16] [9]. Over the past years, few works have been proposed to solve the *abstractive* problem of summarization, which aims to produce from scratch a new cohesive text not necessarily present in the original source [17] [16].

Abstractive summarization requires deep understanding and reasoning over the text, determining the explicit or implicit meaning of each element, such as words, phrases, sentences and paragraphs, and making inferences about their properties [14] in order to generate new sentences which compose the summary.

Recently, riding the wave of prominent results of modern deep learning models in many natural language processing tasks [10], several groups have started to exploit deep neural networks for abstractive text summarization [15] [13]. These deep architectures share the idea of casting the summarization task as a neural machine translation problem [2], where the models, trained on a large amount of data, learn the alignments between the input text and the target summary through an attention encoder-decoder paradigm. In detail, in [15] the authors propose a feed-forward neural network based on neural language model [3] with an attention-based encoder, while the models proposed in [4] and [13] use the attention encoder into a sequence-to-sequence framework modeled by RNNs [18]. Once parametric models are trained, a decoder module greedily generates a summary, word by word, through a beam search algorithm.

The aim of these works based on neural networks is to provide a fully data-driven approach to solve the abstractive summarization task, where the models learn automatically the representation of relationships between the words in the input document and those in the output summary without using complex handcrafted linguistic features. Indeed, the experiments highlight significant improvements of these deep architectures compared to extractive and abstractive state-of-the-art methods evaluated on various datasets, including the gold-standard DUC-2004 [12] using several variants of ROUGE metric [11]. These results prove the effectiveness of the approaches based on modern deep learning architectures to solve the abstractive summarization task and this lays the foundation for a new promising area of research that we want to explore.

2 Motivation and Research Questions

The proposed neural attention-based models for abstractive summarization are still in an early stage, thus they show some limitations. Firstly, they require a large amount of training data in order to capture a good representation that properly maps good (soft) alignments between original text and the related summary. Moreover, since these deep models learn the linguistic regularities relying only on statistical co-occurrences of words over the training set, some grammar and semantic errors can occur in the generated summaries. Finally, these models work only at sentence level and are effective for sentence compression rather than document summarization, where both input text and target summary consist of several sentences.

In this work we argue about our ongoing research on abstractive text summarization. Taking up the idea of setting the summarization task as a sequence-to-sequence learning problem [18], we study approaches to infuse prior knowledge into a RNN in a unified manner in order to overtake the aforementioned limits.

In the first stage of our research we focus on methodologies to introduce linguistic features, such as part-of-speech and named entities tags, and relational semantic information coming from knowledge bases and thesaurus, such as

DBpedia and WordNet. We believe that informing the neural network about the specific syntactic and semantic role of each word (or concept) during the training phase may led several advantages as described below.

Introducing information about the syntactical role of each word, the neural network can tend to learn the right collocation of words by belonging to a certain part-of-speech class. Besides, for standard neural language models, the named entities are usually considered as rare words. This brings some shortcomings during new text generation, where the entities are treated as unknown words [6]. Thus, the linguistic features can improve the model avoiding grammar errors and producing well-formed summaries. Also, a jointly learning of word and knowledge embedding can induce the model to produce more meaningful summaries. Finally, the summarization task lacks of availability of data required to train the models, especially in specific domains. The introduction of prior knowledge can help to reduce the amount of data needed in the training phase.

Concretely, our research wants to answer the following questions:

RQ1 How to formalize the abstractive text summarization as a learning problem using deep neural networks?
RQ2 What is an effective way to introduce prior knowledge into deep neural models?
RQ3 How to combine the distributional and relational semantic in a unified model?
RQ4 Can the proposed models reduce the syntactic and semantic errors in the generated summaries?
RQ5 How to extend the models so that they can be applied to summarization of paragraphs, documents and multi-documents?
RQ6 How to evaluate the proposed models?

3 Methodology

At current state of our research, we are designing a novel approach to incorporate prior knowledge into deep neural networks. Our general deep architecture is inspired by [18], where we formalize the abstractive summarization as a sequence-to-sequence problem by adopting an encoder-decoder paradigm using RNNs. Figure 1 shows a graphical example.

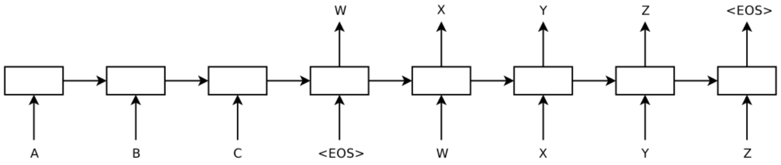

Fig. 1. An example of encoder-decoder paradigm for sequence to sequence learning [18].

These models learn soft alignments between source and target sentences from a training set composed by original text and target summary pairs. In detail, the encoder is a RNN that reads one token at time from the input source and returns a fixed-size vector representing the input text. The decoder is another RNN that generates words for the summary and it is conditioned by the vector representation returned by the first network. In order to find the best sequence of words that represent a summary, a beam search algorithm is commonly used. The RNNs of the encoder and decoder can be implemented with Elman RNN [7] or with a more sophisticated variants as Long-Short Term Memory (LSTM) [8] and Gated Recurrent Unit (GRU) [5] networks. In tasks involving language modeling, these variants have shown impressive performance and they solve the vanishing gradient problem that typically involves RNNs. Several tricks, such as to read the input sequence in reverse or bidirectional mode, are applied to the encoder to improve the performance. Moreover, some attention-based mechanisms [2] [15] [13] [4] are integrated into the encoder to help the network to remember certain aspects about the input. The good performance of the whole architecture often depends on how these attention-based components are modeled.

These models take in account only the distribution of the words in the training corpus. At first stage of our research, we want to incorporate lexical and syntactic features, such as part-of-speech and named entities tags, into RNNs. The core idea is to replace the softmax of each RNN layer with a log-linear model or a probabilistic graphical model, like factor graphs. This replacement does not arise any problem because the softmax function converts the output of the network into probability values, where the softmax can be seen as a special case of the extended version of RNN [6]. Thus, the use of probabilistic models allows to condition the probability value, given an extra feature vector that represents the lexical and syntactic information of each word. We believe that this approach can learn a better representation of the input vector during the training and it can help the decoder in the generation phase. In this way, the decoder can assign to the next word a probability value which is related to the specific lexical role of that word in the generated summary. This can allow the model to decrease the number of grammar errors in the summary, even using a smaller training set since the linguistic regularities are supported by the extra vector of syntactic features.

In the next step, we want to explore approaches that combine the distributional and relational semantic in a unified model. For this purpose, in the recent literature, several works [19] [20] [1] have been proposed principally to solve the automatic knowledge base construction task. Our idea is to adopt these methods into a sequence-to-sequence model to solve the abstractive text summarization problem. We believe that a jointly learning of the text and knowledge can improve the model by generating more abstractive and meaningful summaries.

Another promising direction that we want to investigate is the generation of abstractive summaries from documents or multiple documents using deep learning models. Broadly, the idea can involve a first extractive phase, where the relevant sentences are extracted from source text, followed by the abstractive phase, where a summary is generated only from these relevant sentences.

4 Evaluation Plans

We plan to evaluate our models on gold-standard datasets for the summarization task, such as DUC-2004 [12], Gigaword [15] and CNN/DailyMail [13] corpus, as well as on a local government dataset of documents made available by InnovaPuglia S.p.A. (consisting of projects and funding proposals) using several variants of ROUGE [11] metric.

ROUGE is a recall-based metric which assesses how many n-grams in generated summaries appear in the human reference summaries. This metric is designed to evaluate extractive methods rather than abstractive ones, thus the former would be advantaged. The evaluation in summarization is a complex problem and it is still an open challenge for three main reasons. First, given an input text, there are different summaries that preserve the original meaning. Furthermore, the words that compose the summary could not appear at all in the original source. Finally, ROUGE metric cannot measure the quality of grammar structure of the generated summary. To overcome these issues we plan an in-vivo experiment with a user study.

References

1. S. Ahn, H. Choi, T. Pärnamaa, and Y. Bengio. A Neural Knowledge Language Model. *CoRR*, abs/1608.00318, 2016.
2. D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.
3. Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, 2003.
4. S. Chopra, M. Auli, A. M. Rush, and S. Harvard. Abstractive sentence summarization with attentive recurrent neural networks. 2016.
5. J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.
6. M. Dymetman and C. Xiao. Log-linear rnns: Towards recurrent neural networks with flexible prior knowledge. *CoRR*, abs/1607.02467, 2016.
7. J. L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.
8. S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.
9. K. S. Jones. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481, 2007.
10. Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.
11. C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proc. of the ACL-04 Workshop*, page 10. Association for Computational Linguistics, 2004.
12. K. C. Litkowski. Summarization experiments in duc. In *Proc. of DUC 2004*, 2004.
13. R. Nallapati, B. Xiang, and B. Zhou. Sequence-to-sequence RNNs for text summarization. *CoRR*, abs/1602.06023, 2016.
14. P. Norvig. Inference in text understanding. In *AAAI*, pages 561–565, 1987.

15. A. M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. In *Proc. of EMNLP 2015, Lisbon, Portugal*, pages 379–389, 2015.
16. H. Saggion and T. Poibeau. Automatic text summarization: Past, present and future. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 3–21. Springer, 2013.
17. N. Salim. A review on abstractive summarization methods. *Journal of Theoretical and Applied Information Technology*, 59(1):64–72, 2014.
18. I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proc of NIPS*, pages 3104–3112, 2014.
19. P. Verga and A. McCallum. Row-less universal schema. *CoRR*, abs/1604.06361, 2016.
20. Z. Wang, J. Zhang, J. Feng, and Z. Chen. Knowledge graph and text jointly embedding. In *In Proc. of EMNLP*. ACL, 2014.