Expert Systems With Applications 68 (2017) 93–105

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Text summarization using unsupervised deep learning

Mahmood Yousefi-Azar*, Len Hamey

Department of Computing Faculty of Science and Engineering, Macquarie University, Sydney, NSW, Australia

ARTICLE INFO

Article history:
Received 19 May 2016
Revised 1 August 2016
Accepted 10 October 2016
Available online 12 October 2016

Keywords:
Deep Learning
Query-oriented Summarization
Extractive Summarization
Ensemble Noisy Auto-Encoder

ABSTRACT

We present methods of extractive query-oriented single-document summarization using a deep auto-encoder (AE) to compute a feature space from the term-frequency (*tf*) input. Our experiments explore both local and global vocabularies. We investigate the effect of adding small random noise to local *tf* as the input representation of AE, and propose an ensemble of such noisy AEs which we call the Ensemble Noisy Auto-Encoder (ENAE). ENAE is a stochastic version of an AE that adds noise to the input text and selects the top sentences from an ensemble of noisy runs. In each individual experiment of the ensemble, a different randomly generated noise is added to the input representation. This architecture changes the application of the AE from a deterministic feed-forward network to a stochastic runtime model. Experiments show that the AE using local vocabularies clearly provide a more discriminative feature space and improves the recall on average 11.2%. The ENAE can make further improvements, particularly in selecting informative sentences. To cover a wide range of topics and structures, we perform experiments on two different publicly available email corpora that are specifically designed for text summarization. We used ROUGE as a fully automatic metric in text summarization and we presented the average ROUGE-2 recall for all experiments.

1. Introduction

Text summarization is an automatic technique to generate a condensed version of the original documents. Manual summarization requires a considerable number of qualified unbiased experts, considerable time and budget and the application of the automatic techniques is inevitable with the increase of digital data available world-wide. The early technique to address with manual text summarization dates back as early as 1958 (Luhn, 1958) and the proposed techniques were reviewed extensively (Lloret & Palomar, 2012; Nenkova & McKeown, 2012).

Text summarization can be categorized into two distinct classes: abstractive and extractive. In the abstractive summarization, the summarizer has to re-generate either the extracted content or the text; however, in extractive category, the sentences have to be ranked based on the most salient information. In many research studies extractive summarization is equally known as sentence ranking (Edmundson, 1969; Mani & Maybury, 1999). In practice, specific text summarization algorithm is needed for different tasks. In particular, a summarization technique can be designed to work on a single document, or on a multi-document. Similarly, the purpose of summarization can be to produce a generic summary of the document, or to summarize the content that is most relevant to a user query. The focus of this paper is to propose an extractive query-oriented single-document summarization technique.

Deep learning showed strong promise in various areas, specifically in natural language processing (NLP) tasks (Collobert et al., 2011; Srivastava & Salakhutdinov, 2012). The pivot of our model is a deep auto-encoder (AE) (Hinton & Salakhutdinov, 2006a) as an unsupervised model. The AE learns the latent representations for both the query and the sentences in the document.

In this paper, the word “automatic” implies the feature learning process, that is, completely independent from human-crafted features. More clearly, the concept of deep learning (i.e. neural networks with more than one hidden layer) can be considered as a wide class of machine learning approaches and architectures in which the main characteristic is hierarchically using many layers of nonlinear information processing. The aim of the techniques is learning feature hierarchies with higher level features of the hierarchy extracted from lower level features (Bengio, 2009). In fact, with automatically learned features at multiple levels of abstraction a system may execute complex functions to directly transfer the input to the output.

The key factor of our model is the word representation. Typically, automatic text summarization systems use sparse input representations. Sparse representations can cause two problems for the model. First, not observing (enough) data in training process.

* Corresponding author.
E-mail addresses: mahmood.yousefiazar@hdr.mq.edu.au (M. Yousefi-Azar), len.hamey@mq.edu.au (L. Hamey).

http://dx.doi.org/10.1016/j.eswa.2016.10.017
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

This problem is intensified when the selected vocabulary consists of only a subpart of the total presented words in the training data. Second, too much zero in the input and output of AE.

To address these problems, we propose two techniques to reduce sparsity. First, we develop a local word representation in which each vocabulary is designed to construct the input representations of sentences in that document. Second, we add random noise to word representation vector, affecting both the inputs and outputs of the AE.

In our case, after ranking the sentences of a document based on the cosine similarity, they must be selected to generate the summary. Picking the top ranked sentences up is a straightforward selection strategy that is used for AE networks trained without added noise. However, we propose to use an ensemble approach that aggregates the rankings of different experiments, each the result of adding randomly generated noise. Each application uses different random noise added to the input word representation, producing a possibly different extractive summary. The final summary is obtained by selecting the sentences that occur most frequently in the individual summaries. This ensemble approach can make the summarization process robust against small differences between training methods and analogous with manual summarization where annotator(s) would produce different summaries for a document in each review of the document.

The specific contributions of the paper are as follow: We introduce an unsupervised approach for extractive summarization using AEs. Although AEs have previously been applied for summarization as a word filter (Zhong, Liu, Li, & Long, 2015), to the best of our knowledge we are the first to use the AE for summarization by sentence ranking. We will evaluate how AEs handle a sparse word representation such as *tf-idf* and a less sparse word representation based on a document-specific vocabulary, and also the impact of adding random noise to the local word representation vector. The addition of noise changes the AE from a deterministic feed-forward network to a stochastic model. To our best knowledge, this representation technique is not previously explored in the application of auto-encoders. We introduce the Ensemble Noisy Auto-Encoder (ENAE) in which the model is trained once and used multiple times on the same input, each time with different added noise. We show adding stochastic noise to the input and running an ensemble on the same input can make improvements. We expand the email summarization features beyond a set of features and develop to automatically extracted features of emails. Finally, our evaluation shows the proposed unsupervised model equalling performance of previously presented supervised models and exceeding comparable unsupervised techniques on BC3 dataset.

The next section describes recent related work. Section three is dedicated to our model in terms of topology and training algorithms, in particular, Restricted Boltzmann Machine (RBM) and AEs. The word representation is presented in section four. Section five presents sentence ranking measurement detail. The ENAE scheme is described in the six section. After providing experimental setup in the section seven, the results and discussion on the two different email datasets is presented in the section eight. The conclusion and future work is explained in the section nine. Bibliography is placed at the last section.

2. Related work

Machine learning techniques have widely used for text summarization (Conroy & O'leary, 2001; Corston-Oliver, Ringger, Gamon, & Campbell, 2004; Li, Zhou, Xue, Zha, & Yu, 2009; Ouyang, Li, Li, & Lu, 2011). They are trainable systems in which models learns how to generalize its parameters to extract salient points. Most of machine learning approaches in text summarization are inspired from information retrieval and adapted into the task such as Bayesian models, Hidden Markov Models (HMM), Support Vector Machines (SVM) and Support Vector Regression (SVR). In general, many supervised and unsupervised approaches have been proposed and they can be categorized into the following groups: Latent topic models as unsupervised techniques, classification and regression as the supervised techniques.

Kaikhah (2004) successfully introduced a shallow neural network for automatic text summarization. Also, Svore, Vanderwende, and Burges (2007) proposed a neural network-based system, called NetSum. This technique was inspired from Burges et al. (2005). Shardan and Kulkarni (2010) proposed a combination of the multilayer perceptron (MLP) with fuzzy logic and they reported that this combination improves the result. In addition to feed-forward neural networks, recurrent neural networks (RNNs) have been applied for text summarization (Prasad, Kulkarni, & Prasad, 2009). In spite of different topology, the approach was generally inspired from Kaikhah (2004).

Deep neural networks have been used for both abstractive summarization and extractive summarization. For abstractive summarization, IBM and Facebook companies developed successful models based on Recurrent Neural Network (RNN) and convolutional neural network (CNN) respectively (Nallapati, Zhou, Nogueira dos santos, Gulcehre, & Xiang, 2016; Rush, Chopra, & Weston, 2015).

For extractive summarization, Zhong et al. (2015) used a deep architecture that is similar to an AE. They used the learned representations for filtering out unimportant words of a document in the early layer and discovering key words in later layer. Their training method requires sample queries during the first stage. The concept space is to extract the candidate sentences for the summary. However, in our model both the term and concept space are developed based on a sentence of a document. We use the learned representations directly to represent the semantic similar of sentences and in the ranking function. In supervised models, Denil, Demiraj, and de Freitas (2014) proposed a model based on a CNN to extract candidate sentences to be included into the summary. A key contribution of the paper is that CNN is trained to classify sentiment labels that are either positive or negative (i.e. a binary label) and sentence extraction can affect the results of predicted a sentiment. Ha, Kang, Pyo, and Kim (2015) also used a CNN for summarizing Korean news articles and retrieving relevant images. Cao, Wei, Dong, Li, and Zhou (2015) used a Recursive Neural Network for text summarization using hand-crafted word features as inputs. With the capability of dealing with a variable-length input in a RNN, the proposed system formulates the sentence ranking task in a hierarchical regression fashion. Not using any hand-crafted word representations and labelling data are significant in our proposed model.

The proposed method of this paper in adding noise to the input can be analogous with the denoising auto-encoders (Vincent, Larochelle, Bengio, & Manzagol, 2008). To stop overfitting, denoising auto-encoders use noise in the input of AE. In a denoising AE, in order to prevent learning identity function, thereby capturing important information about the input in hidden layers, the input is corrupted and the network tries to undo the effect of this corruption. The intuition is that this rectification can occurs if the network can capture the dependences between the inputs. However, first, in most denoising AEs the input is corrupted using a random zero mask while in our model a small noise is added to the input. Second, in our model, the training algorithm adds very small noise to the input but the output is still the same as input. Third, the denoising AE adds noise to the inputs only during training, while we also add noise to input during test time.

Our earlier paper (Yousefi-Azar, Sirts, Aliod, & Hamey, 2015) presented preliminary experiments. In particular, the current paper presents results for the BC3 dataset compared with other related