Expert Systems with Applications xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Query-oriented unsupervised multi-document summarization via deep learning

Sheng-hua Zhong [a,b], Yan Liu [b,*], Bin Li [c]

[a] College of Computer Science & Software Engineering, Shen Zhen University, Shen Zhen, Guang Dong 518060, China
[b] Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon 999077, Hong Kong, China
[c] Department of Linguistics and Translation, City University of Hong Kong, Kowloon 999077, Hong Kong, China

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Deep learning
Query-oriented summarization
Multi-document
Neocortex simulation

ABSTRACT

Capturing the compositional process from words to documents is a key challenge in natural language processing and information retrieval. Extractive style query-oriented multi-document summarization generates a summary by extracting a proper set of sentences from multiple documents based on pre-given query. This paper proposes a novel document summarization framework based on deep learning model, which has been shown outstanding extraction ability in many real-world applications. The framework consists of three parts: concepts extraction, summary generation, and reconstruction validation. A new query-oriented extraction technique is proposed to extract information distributed in multiple documents. Then, the whole deep architecture is fine-tuned by minimizing the information loss in reconstruction validation. According to the concepts extracted from deep architecture layer by layer, dynamic programming is used to seek most informative set of sentences for the summary. Experiment on three benchmark datasets (DUC 2005, 2006, and 2007) assess and confirm the effectiveness of the proposed framework and algorithms. Experiment results show that the proposed method outperforms state-of-the-art extractive summarization approaches. Moreover, we also provide the statistical analysis of query words based on Amazon's Mechanical Turk (MTurk) crowdsourcing platform. There exists underlying relationships from topic words to the content which can contribute to summarization task.

1. Introduction

Automatically generating summaries from large text corpora has long been attracting research attention from both information retrieval and natural language processing, the earlier studies of which could be dated back to the 1950s and 1960s (Baxendale, 1958; Edmundson, 1969; Luhn, 1958). Automatic generation of summaries creates shortened versions of texts to help users catch important information in the original text with bearable time costs (Khanpour, 2009). Currently, the creation of summaries is a task best handled by humans. However, with the explosion of textual data, especially in big data era, it is no longer financially possible, or feasible, to produce all types of summaries by hand. Earlier studies on text summarization aimed at summarizing from pre-given documents without requirements, which is usually referred to as generic summarization (Berger & Mittal, 2000). With the development of information retrieval, query-oriented summarization task, which requires summarizing from a set of document to answer a pre-given query, has started attracting more and more attention (Tang, Yao, & Chen, 2009). According to the size of the input, text summarization tasks can be grouped into single-document and multi-document summarization tasks (Shen, Sun, Li, Yang, & Chen, 2007). Based on the writing style of the output summary, text summarization techniques can be divided into extractive approaches and abstractive approaches (Song, Choi, Park, & Ding, 2011; Wong, Wu, & Li, 2008). Due to the limitation of current natural language generation techniques, extractive approaches are the mainstream in the field. An extractive approach selects a number of indicative text fragments from the input documents to form a summary instead of re-writing an abstract (Chen, Yang, Zha, Zhang, & Zhang, 2008) under a budget constraint. A budget constraint is natural in summarization task as the length of the summary is often restricted (Lin & Bilmes, 2010). In the paper, we adopt the extractive style to develop techniques for query-oriented multi-document summarization.

Almost all extractive summarization methods are faced with two key problems: how to rank textual units, and how to select a subset of those ranked units (Jin, Huang, & Zhu, 2010). The first

* Corresponding author.
E-mail addresses: csshzhong@szu.edu.cn (S.-h. Zhong), csyliu@comp.polyu.edu.hk (Y. Liu), binli2@cityu.edu.hk (B. Li).

http://dx.doi.org/10.1016/j.eswa.2015.05.034
0957-4174/

one on ranking requires systems to model the relevance of a textual unit to a topic or a query. The second one on selection requires systems to improve diversity or to remove redundancy so that more relevant information can be covered by the summary within a limited length.

Attempts to solutions of sentence ranking are varied. Some of solutions are based on surface features (Luhn, 1958; Radev, Jing, Stys, & Tam, 2004), some on graphs (Wan, 2009; Wan & Xiao, 2009; Wei, Li, Lu, & He, 2010), and some on supervised learning (Cao, Qin, Liu, Tsai, & Li, 2007; Ouyang, Li, Li, & Lu, 2011). After obtaining a list of ranked sentences, it is then important to select a subset of sentences to form a good summary that includes diverse information within a length limit. Goldstein, Mittal, Carbonell, and Kantrowitz (2000) were among the first to propose global models using the maximum marginal relevance (MMR) criteria. The models score sentences under consideration as a weighted combination of relevance plus redundancy with sentences already in the summary. Currently, greedy MMR style algorithms are the standard algorithms in document summarization. McDonald (2007) proposed to replace the greedy search of MMR with a globally optimal formulation, where the basic MMR framework can be expressed as a knapsack packing problem, and an integer linear program (ILP) solver can be used to maximize the resulting objective function.

This paper presents a new method following the extractive style to summarize documents using deep techniques. Deep learning models the learning task using deep architectures composed of multiple layers of parameterized nonlinear modules. These models have been proved outstanding in feature extraction of visual data. To our knowledge, this is the first attempt that utilizes deep learning in query-oriented multi-document summarization task. Different from the existing methods, we neither directly rank the textual units based on the relevance to the topic or query, nor directly improve diversity or remove redundancy. The proposed deep learning algorithm is partitioned into three stages: concept extraction, reconstruction validation, and summary generation. In the concept extraction stage, hidden layers are used to abstract the documents layer by layer using greedy layer-wise extraction algorithm. The second stage of reconstruction validation intends to reconstruct the data distribution by fine-tuning the whole deep architecture globally. Finally, dynamic programming (DP) is utilized to maximize the importance of the summary with the length constraint. A novel framework with several new algorithms is proposed in the following part.

2. Related work on deep learning

Different from shallow learning models, deep learning is learning multiple levels of representation and abstraction so as to extract more senses out of data. Besides evidence from neuroscience, theoretical analyses from machine learning also confirmed that deep models are more compact and expressive than shallow models in representing most learning functions, especially highly variable ones. Many empirical validations are also reported to support that deep architectures are promising in solving hard learning problems (Larochelle, Erhan, Courville, Bergstra, & Bengio, 2007). Moreover, theoretical analysis shows that deep architectures are more efficient than shallow circuits such as a typical support vector machine (SVM), because the former can represent most common functions, especially highly-variable learning functions compactly and effectively.

However, it is difficult to learn the parameters of deep architectures with multiple hidden layers containing trainable weights at all levels. Back propagation, a well-known computationally efficient model for multilayer neural networks, also suffers from insufficient labeled data, high computational cost, and poor local optima when working under a deep model (Hinton, 2007). To reduce the difficulty of deep learning, Hinton and Salakhutdinov (2006) proposed deep belief network (DBN), i.e. a densely-connected, directed belief net with multiple hidden layers. DBN partitions the learning procedure to two stages: to abstract input information layer by layer and to fine-tune the whole deep network to the ultimate learning target (Hinton, Osindero, & Teh, 2006; Salakhutdinov & Hinton, 2007). The network pairs each feed-forward layer with a feed-back layer that attempts to reconstruct the input of the layer from the output. Such layer-wise generative models are implemented by a family of Restricted Boltzmann Machines (RBMs) (Smolensky, 1986). After a greedy unsupervised learning to each pair of layers, the lower-level features are progressively combined into more compact high-level representations. In the second stage, the whole deep network is refined using a contrastive version of the "wake-sleep" algorithm via a global gradient-based optimization strategy. Owing to this two-stage fast greedy learning, DBN exhibits notable performance in dimensionality reduction (Liu, Xu, Tsang, & Luo, 2009) and classification (Cao, Yu, Luo, & Huang, 2009) for different applications, such as image generation (Dahl, Ranzato, Mohamed, & Hinton, 2010), and audio event classification (Ballan, Bazzica, Bertini, Bimbo, & Serra, 2009).

The conference version of our work is the first attempt of deep learning methods for the query-oriented multi-document summarization task (Liu, Zhong, & Li, 2012). After we proposed deep learning models for document summarization task, more and more recent work focused on deep learning based methods. For example, Cao, Wei, Dong, Li, and Zhou (2015) introduced a ConvNet model to support introspection of the document structure. Their model is used to identify and extract task-specific salient sentences from documents. Denil, Demiraj, and Freditas (2014) developed a ranking framework upon Recursive Neural Networks to rank sentences for multi-document summarization. It formulates the sentence ranking task as a hierarchical regression process, which simultaneously measures the salience of a sentence and its constituents in the parsing tree.

3. Basic idea of proposed model

Humans do not have difficulty with summarizing documents based on given queries. Query-oriented multi-document summarization, however, has remained a well-known challenge in natural language processing in the past fifteen years of extensive research. In the evaluation of the summarization tasks in the Document Understanding Conference (DUC), the summaries created by human peers are much better than those extracted automatically. Motivated by this fact, we aimed at designing a proper deep architecture and corresponding unsupervised learning algorithms for query-oriented multi-document summarization. Latest research findings from neuroscience suggest that the deep learning model is consistent with the physical structure of human neocortex, evolution of intelligence, and propagation of information in the human neocortex. Thus, it has great potential to provide human-like judgment using a human-like system in tasks of natural language processing. A discussion of the deep learning model from three aspects is presented in the following sections.

(1) The deep architecture is identical to the multi-layer physical structure of the human cerebral cortex. The neocortex, which is associated with many cognitive abilities, has a complex multi-layer hierarchy (Lee & Mumford, 2003). The laminar structure and a multi-layer illustration of the neocortex are shown in Fig. 1. The neocortex can be roughly