Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Query-Oriented Multi-Document Summarization via Unsupervised Deep Learning

Yan Liu, Sheng-hua Zhong, Wenjie Li
Department of Computing, The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong, P.R. China
{csyliu, csshzhong, cswjli}@comp.polyu.edu.hk

Abstract

Extractive style query oriented multi document summariza tion generates the summary by extracting a proper set of sentences from multiple documents based on the pre given query. This paper proposes a novel multi document summa rization framework via deep learning model. This uniform framework consists of three parts: concepts extraction, summary generation, and reconstruction validation, which work together to achieve the largest coverage of the docu ments content. A new query oriented extraction technique is proposed to concentrate distributed information to hidden units layer by layer. Then, the whole deep architecture is fi ne tuned by minimizing the information loss of reconstruc tion validation. According to the concentrated information, dynamic programming is used to seek most informative set of sentences as the summary. Experiments on three bench mark datasets demonstrate the effectiveness of the proposed framework and algorithms.

Introduction

Automatically generating summaries from large text corpora has long been studied in both information retrieval and natural language processing, which could be dated back to the 1950s and 1960s (Luhn, 1958; Baxendale, 1958; Edmundson, 1969). Automatic text summarization is the process of creating shortened version of texts, has arisen to help users to catch the important information in the original text with bearable time costs (Khanpour, 2009). Early studies on text summarization aim at summarizing from pre-given documents without other requirements, which is usually referred 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, attracts more and more attention (Tang, et al., 2009). According to the size of the input to text summarization task, single-document and multi-document summarization tasks can be differentiated (Wan & Xiao, 2008) (Shen, et al., 2007). Taking into account the writing style of the output summary, text summarization techniques can be divided into extractive approaches and abstractive approaches (Wong, et al., 2008). Due to the limitation of current nature language generation techniques, extractive approaches, which select a number of indicative text fragments from the input documents to form a summary instead of rewriting an abstract (Chen, et al., 2008), are the mainstream in the area. In the paper, we follow the extractive style to develop techniques for query-oriented multi-document summarization.

Almost all extractive summarization methods face two key problems: the first problem is how to rank textual units, and the second one is how to select a subset of those ranked units (Jin, et al., 2010). The ranking problem requires systems model the relevance of a textual unit to a topic or a query. The selection problem requires systems improve diversity or remove redundancy so that more relevant information can be covered by the summary as its length is limited.

There have been a variety of studies to approach the ranking problem, including: surface feature based sentence ranking (Luhn, 1958; Radev, et al., 2004), graph-based sentence ranking (Wan & Xiao, 2009), (Wan, 2009), (Wei, et al., 2010), and supervised learning based sentence ranking (Cao, et al., 2007; Ouyang, et al., 2011). Even given a list of ranked sentences, it is not trivial to select a subset of sentences to form a good summary which includes diverse information within a length limit. Goldstein et al. (Goldstein, et al., 2000) presented one of the first global models through the use of the maximum marginal relevance (MMR) criteria, which scored 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 (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.

Despite more than fifteen years of extensive research, query-oriented multi-document summarization remains a well-known challenge in the field of nature language pro-

cessing because it is very difficult to bridge the gap between the semantic meanings of the documents and the basic textual units. So this paper intends to propose a novel framework by referencing the architecture of the human neocortex and the procedure of intelligent perception via deep learning. Different from shallow learning models such as support vector machine (SVM), deep learning, like deep belief network (DBN), models the learning task using deep architectures composed of multiple layers of parameterized nonlinear modules. To our knowledge, this is the first paper that utilizes deep learning in query-oriented multi-document summarization task.

In the following parts of this paper, we first discuss the motivation of utilizing deep learning to text summarization task. Then, a novel deep architecture with three parts of query-oriented concepts extraction, reconstruction validation for global adjustment, and summary generation via dynamic programming are introduced. In the experiment part, we demonstrate the performance of the proposed framework and the new algorithms on three benchmark datasets. The paper is closed with conclusion.

Deep Learning for Query-oriented Multi-documents Summarization

The rationale of utilizing deep learning in query-oriented multi-documents summarization is to provide human-like judgment by referencing human's neocortex and the procedure of intelligent perception. 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). All functional areas of neocortex can be roughly differentiated into six functionally distinct horizontal layers (Leuba & Kraftsik, 1994). When it is taken into consideration that many different neocortex areas, such as Broca's and Wernicke's areas, and other lexical-semantic processing areas, are involved in lexical-semantic processing, dozens of cortical layers are involved in generating even the simplest lexical-semantic processing. Therefore, deep learning model shows potentials to provide human-like judgment using a human-like system in tasks of nature language processing.

Besides the evidences from neuroscience, some theoretical analyses from machine learning also provide support for the argument that deep models are more compact and expressive than shallow models in representing learning functions, especially highly variable ones. Obviously, query-oriented multi-documents summarization is a highly intelligent task, even not easy for human beings. The mapping between the semantic meaning of multiple documents and the basic textual units is not straightforward. Fortunately, deep learning has two attractive characters. First, because of the nonlinear structure of multiple hidden layers, deep models can represent hard problem in more concise way, which is well adapted the essentials of summarization that includes information as much as it can with bearable length. Second, because of the pair-wise hidden layers reconstruction learning in most deep models, distributed information can be concentrated gradually layer by layer even if under unsupervised situation. This character will benefit the learning in large dataset, just like multi-document summarization.

Although deep learning has never be used in document summarization, many empirical validations have demonstrated that deep models have notable ability of multimedia data abstraction (Taylor, et al., 2010) (Liu, et al., 2009) in various tasks, such as image classification (Zhong, et al., 2011), image generation (Dahl, et al., 2010), and audio event classification (Ballan, et al. 2009). Hence, deep models are also promising to abstract text data effectively for query-oriented multi-documents summarization.

Deep Architecture

This paper intends to provide a uniform framework of generating text summary automatically from the original multiple documents according to the query. As mentioned, this is the first paper of utilizing deep learning in document summarization. To adapt the characters of this new application, a novel unsupervised deep learning model Query-oriented Deep Extraction (QODE) with the new deep architecture is shown in Figure 1.

The feature vector $\mathbf{f}^d = [f_1^d, f_2^d, \cdots, f_v^d, \cdots, f_V^d]$, the *tf* value of word in the vocabulary of $\mathbf{D}$ calculated in document $\mathbf{d}_m$, is input into deep architecture. V is the length of the vocabulary of $\mathbf{D}$. The output is a summary $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_t, \cdots, \mathbf{s}_T]$. For the hidden layer, Restricted Boltzmann Machines (RBMs) are used as building blocks (Smolensky, 1986). RBM is a two-layer recurrent neural network in which stochastic binary inputs and outputs are connected using symmetrically weighted connections. RBMs are utilized as the building blocks of deep models because the bottom-up connections can be used to infer the more compact high-level representations from low-level features and the top-down connections can be used to validate the effectiveness of the generated compact representations. The parameter space of the deep architecture is initialized randomly except for the input layer. The initial parameters of the first RBM are also determined by the query words.

Based on the new deep architecture, the deep learning procedure can be partitioned into three stages: concept extraction, reconstruction validation, and summary generation. In the concept extraction stage, three hidden layers H^1, H^2, and H^3 are used to abstract the documents using greedy layer-wise extraction algorithm. In our implementation, H^1 is used to filter out the words appearing accidental-

ly. Hidden layer H^2 is supposed to discover the key words; reconstruction validation part intends to reconstruct the data distribution by fine-tuning the whole deep architecture globally. Finally, the dynamic programming (DP) is utilized to maximize the importance of the summary with the length constraint. After these three stages, the final optimized summary S^* is generated. In the following session, we will discuss the detailed learning procedure of each part.

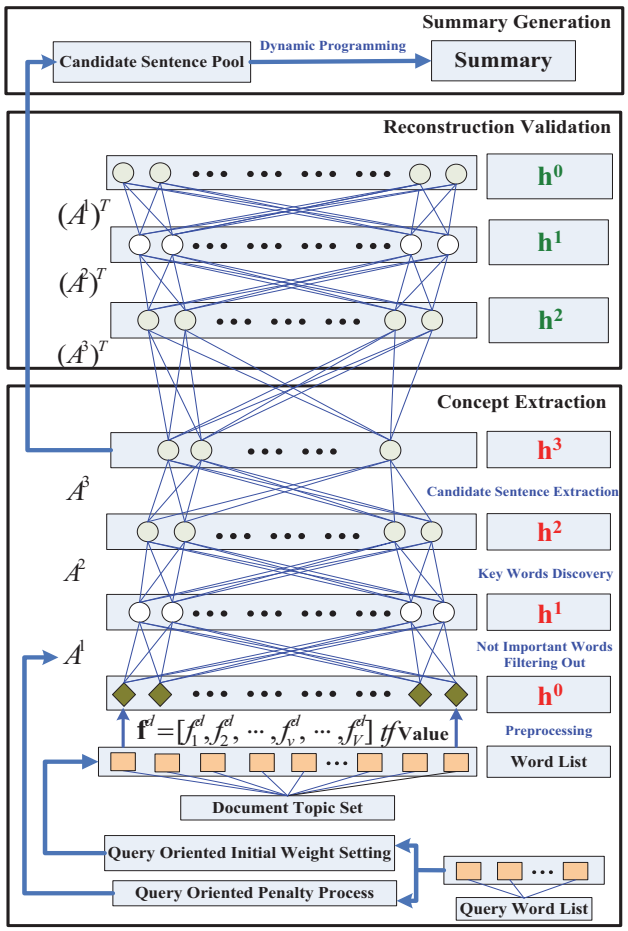


Figure 1: Deep architecture of QODE technique

Query-oriented Concept Extraction

First, we generate a vocabulary with length V based on the words appearing in the document topic set $\mathbf{D}$. The feature vectors $\mathbf{f}^D = [f_1^D, f_2^D, \cdots, f_v^D, \cdots, f_V^D]$ of the document set $\mathbf{D}$ and $\mathbf{f}^d = [f_1^d, f_2^d, \cdots, f_v^d, \cdots, f_V^d]$ of the single document $\mathbf{d}_m$ are calculated. Here, f_v^D is the *tf* value of v^{th} word in the vocabulary of $\mathbf{D}$ calculated in all documents. f_v^d is the *tf* value of v^{th} word in the vocabulary of $\mathbf{D}$ calculated in $\mathbf{d}_m$.

Then, $\mathbf{f}^d$ is input to the deep architecture as the visible layer H^0 to construct a RBM with hidden layer H^1. The energy of the state $(\mathbf{h}^0, \mathbf{h}^1)$ in the first RBM is:

$$E(\mathbf{h}^0, \mathbf{h}^1; \theta^1) = -((\mathbf{h}^0)^T A^1 \mathbf{h}^1 + (b^1)^T \mathbf{h}^0 + (c^1)^T \mathbf{h}^1) \quad (1)$$

where $\theta^1 = (A^1, b^1, c^1)$ are the model parameters between layer H^0 and layer H^1. A_{ij}^1 is the symmetric interaction term between visible unit i in H^0 and hidden unit j in H^1. b_i^1 is the i^{th} bias of layer H^0 and c_j^1 is the j^{th} bias of layer H^1. The first RBM has the following joint distribution:

$$P(\mathbf{h}^0, \mathbf{h}^1; \theta^1) = e^{-E(\mathbf{h}^0, \mathbf{h}^1; \theta^1)} / Z = (e^{-E(\mathbf{h}^0, \mathbf{h}^1; \theta^1)}) / (\sum_{\mathbf{h}^0} \sum_{\mathbf{h}^1} e^{-E(\mathbf{h}^0, \mathbf{h}^1; \theta^1)}) \quad (2)$$

where Z is the normalization constant. The log-likelihood probability of assigning to a visible vector to $\mathbf{h}^0$ in H^0 is:

$$\log P(\mathbf{h}^0) = \log \sum_{\mathbf{h}^1} e^{-E(\mathbf{h}^0, \mathbf{h}^1; \theta^1)} - \log \sum_{\mathbf{h}^0} \sum_{\mathbf{h}^1} e^{-E(\mathbf{h}^0, \mathbf{h}^1; \theta^1)} \quad (3)$$

Gibbs sampling from an RBM proceeds by sampling $\mathbf{h}^1$ given $\mathbf{h}^0$, then $\mathbf{h}^0$ given $\mathbf{h}^1$, etc. The conditional distributions over input state $\mathbf{h}^0$ in visible layer H^0 and hidden state $\mathbf{h}^1$ in hidden layer H^1 are given by Equation (4) and (5), where $\sigma(x) = 1/(1+\exp(-x))$.

$$p(\mathbf{h}^1 \mid \mathbf{h}^0) = \prod_j p(\mathrm{h}_j^1 \mid \mathbf{h}^0), \quad p(\mathrm{h}_j^1 = 1 \mid \mathbf{h}^0) = \sigma(\sum_j A_{ij} \mathrm{h}_i^0 + a_j) \quad (4)$$

$$p(\mathbf{h}^0 \mid \mathbf{h}^1) = \prod_i p(\mathrm{h}_i^0 \mid \mathbf{h}^1), \quad p(\mathrm{h}_i^0 = 1 \mid \mathbf{h}^1) = \sigma(\sum_j A_{ij} \mathrm{h}_j^1 + b_i) \quad (5)$$

Denote $\mathbf{h}^1(k)$ for the k-th $\mathbf{h}^1$ sample from the chain, starting at $k=0$ with $\mathbf{h}^1(0)$, which is the input observation for the RBM and $(\mathbf{h}^1(k), \mathbf{h}^0(k))$ for $k \longrightarrow \infty$ is a sample from the Markov chain. So we could calculate the derivative of Equation (3) with respect to the parameter $\theta^1 = (A^1, b^1, c^1)$ below:

$$\frac{\partial \log p(\mathbf{h}^1(0))}{\partial \theta^1} = -\sum_{\mathbf{h}^0(0)} p(\mathbf{h}^0(0) \mid \mathbf{h}^1(0)) \frac{\partial E(\mathbf{h}^1(0), \mathbf{h}^0(0))}{\partial \theta^1} + \sum_{\mathbf{h}^1(k)} \sum_{\mathbf{h}^0(k)} p(\mathbf{h}^1(k), \mathbf{h}^0(k)) \frac{\partial E(\mathbf{h}^1(k), \mathbf{h}^0(k))}{\partial \theta^1} \quad (6)$$

The idea of Contrastive Divergence (Hinton, 2002) algorithm use the difference between two Kullback-Leibler divergences is to take k small (typically $k=1$) to run the claim for only one step. When $k=1$, the derivative to the model parameter A^1 can be obtained by Equation (7),

$$\frac{\partial \log P(\mathbf{h}^1(0))}{\partial A^1} = <\mathbf{h}^1(0)^T \mathbf{h}^0(0)>_{data} - <\mathbf{h}^1(1)^T \mathbf{h}^0(1)>_{recon} \quad (7)$$

where $\langle \cdot \rangle_{data}$ denotes an expectation with respect to the data distribution and $\langle \cdot \rangle_{recon}$ denotes the "reconstruction" distribution of data after one step. This leads to a simple learning rule for performing stochastic steepest ascent in the log probability of the training data in Equation (8).

$$A^1 = \vartheta A^1 + \Delta A^1 = \vartheta A^1 + \varepsilon_A (<\mathbf{h}^1(0)^T \mathbf{h}^0(0)>_{data} - <\mathbf{h}^1(1)^T \mathbf{h}^0(1)>_{recon}) \quad (8)$$

Other parameters in θ^1 update function could be calculated in a similar manner, where ϑ is the momentum and $\varepsilon_A, \varepsilon_b, \varepsilon_c$ are the learning rate.

$$b^1 = \vartheta b^1 + \Delta b^1 = \vartheta b^1 + \varepsilon_b (\mathbf{h}^0(0) - \mathbf{h}^0(1)) \quad (9)$$

$$c^1 = \vartheta c^1 + \Delta c^1 = \vartheta c^1 + \varepsilon_c (\mathbf{h}^1(0) - \mathbf{h}^1(1)) \quad (10)$$

To integrate query information for document summarization, we have two different processes including: query oriented initial weight setting and query oriented penalty process. In classical deep network, the parameter matrix A^1

is initialized to small random values chosen from a zero-mean Gaussian with a standard deviation of about 0.01. Different from it, we strengthen the influence from query as Equation (11) after random initialization setting if the i^{th} node word v_i in H^0 belongs to the query.

$$A_{ij}^1 = \max(A^1) \quad \text{if } v_i \in \mathbf{q} \tag{11}$$

In the penalty process, the reconstruction error in query word is penalized more than others as below, where γ is the penalty factor.

$$\Delta A_{ij}^1 = \gamma \Delta A_{ij}^1 \quad \text{if } v_i \in \mathbf{q} \tag{12}$$

The above discussion is based on single document $\mathbf{d}_m$ for the first layer. Similar operations can be performed to the higher layer RBMs based on all documents in topic set $\mathbf{D}$.

After the concept extraction based on the deep architectures, the importance matrix AF is defined as Equation (13). The element AF_{in} of AF is the importance of i^{th} word in the vocabulary to the n^{th} node of hidden layer H^3, where K_3 is the number of unit in H^3, A^1, A^2, A^3 are the symmetric interaction term in layer pairs.

$$AF = [\underbrace{(\mathbf{f}^{D_1})^T, (\mathbf{f}^{D_2})^T, \cdots, (\mathbf{f}^{D_j})^T, \cdots, \mathbf{f}^{D_{K_3}})^T}_{K_3}](A^1 A^2 A^3) \tag{13}$$

In our implementation, hidden layer H^3 is assumed to extract the candidate sentences for the summary. Certainly, we could extract the candidate sentences of every node in H^3 only depend on how many unions of key words are in them according to AF_{in}. In our technique, after the reconstruction validation part globally adjust the deep network to find optimum parameters, the DP is utilized to maximize the query oriented importance of generated summary with the constraint of summary length.

Reconstruction Validation for Global Adjustment

In the first part, we use greedy layer-by-layer algorithm to learn a deep model for concept extraction. In this part, we use backpropagation through the whole deep model to fine-tune the parameters $\theta = [A, b, c]$ for optimal reconstruction.

The greedy layer-by-layer query-oriented concept extraction stage has performed a global search for a sensible and good region in the whole parameter space. Therefore, after the first part, we already construct a good data concept extraction model. Backpropagation is well known as a better local fine-tuning model than global search. So backpropagation is utilized to adjust the entire deep network to find good local optimum parameters $\theta^* = [A^*, b^*, c^*]$ which is used in summary generation via DP. And the learning algorithm in this stage is used to minimize the cross-entropy error $[-\sum_v f_v \log \hat{f}_v - \sum_v (1-f_v)\log(1-\hat{f}_v)]$, where f_v is the *tf* value of v^{th} word and $\hat{f}_v$ is the *tf* value of its reconstruction.

Summary Generation via Dynamic Programming

In this stage, DP is utilized to maximize the importance of the summary with the length constraint.

After the optimum parameters are obtained in the reconstruction validation, we use them to calculate the importance matrix AF by Equation (13). Then we extract ten words with largest AF_{in} value in every n^{th} node of hidden layer H^3. The set of these unions words are denoted as $\mathbf{UN}$. The importance of every sentence In_t is calculated by Equation (14), where λ is the query word importance factor, μ_i is the word in sentence $\mathbf{s}_t$. And the importance of the generated summary could be denoted as $\text{In} = \sum_t \text{In}_t$.

$$\text{In}_t = \sum_i \omega_i, \begin{cases} \omega_i = \lambda & \text{if } (\mu_i \in \mathbf{UN}) \cap (\mu_i \in \mathbf{q}) \\ \omega_i = 1 & \text{if } \mu_i \in \mathbf{UN} \\ \omega_i = 0 & \text{others} \end{cases} \tag{14}$$

Taken the limited length of summary N_S into consideration, the summary length Le is defined as below, where l_t is the length of sentence $\mathbf{s}_t$.

$$\text{Le} = l_1 + ... + l_t + ... + l_T \le \text{N}_\text{S} \tag{15}$$

Based on the analysis above, we obtain the objective function aims to optimize with the constraint below. Because the task of Document Understanding Conference is to produce query-oriented multi-document summarization with allowance of 250 words, in our paper, N_S is equal to 250.

$$max \ \text{In} = \sum_t \text{In}_t, \ s.t. \ \text{Le} \le \text{N}_\text{S} \tag{16}$$

In context of mathematical optimization method, DP refers to simplifying a complicated problem by breaking it down into simpler sub-problems in a recursive manner. The optimization problem in (16) is classical knapsack problem which is often solved by DP. So we use DP to find the optimum solution.

The DP function is denoted in Equation (17). Here, $f_K(\lambda_K)$ is the maximum of the summary importance in stage K. K is the stage variable to describe the current sentence. The state variable λ_K is the remaining length before K starts. The decision variable u_K is the choice whether or not to put the current sentence $\mathbf{s}_t$ into the summary.

$$\begin{cases} f_K(\lambda_K) = \max\{u_K \text{In}_K + f_{K-1}(\lambda_{K-1})\} \\ \lambda_K = \lambda_{K+1} - u_{K+1} l_{K+1}, \quad K = t, 1 \le t \le T \\ \lambda_0 = 0, \lambda_T = 250, f_0(\lambda_0) = 0 \end{cases} \tag{17}$$

After solving the Equation (17) by positive sequence method, we obtain the optimized summary $S^* = \{\mathbf{s}_1^*, \mathbf{s}_2^*, ..., \mathbf{s}_t^*, ..., \mathbf{s}_T^*\}$, where $\mathbf{s}_t^*$ is the optimized sentence.

Empirical Validation

Evaluation Setup

In this section, we conduct several experiments for multi-document summarization task evaluation in the Document

Understanding Conference (DUC) on three open benchmark dataset DUC 2005, DUC 2006 and DUC 2007. There are altogether 50 topics, 50 topics, and 45 topics in DUC 2005, DUC 2006 and DUC 2007, respectively.

The task of DUC is to produce query-oriented multi-document summarization with generous allowance of 250 words. As a preprocessing step, the stop words in each sentence are removed and the remaining words are stemmed using the Porter's stemmer (Porter, 1980). In the evaluation step, the ROUGE (Lin, 2004) toolkit (i.e. ROUGEeval-1.5.5 in this study) is used for evaluation, which has been widely adopted by DUC tasks.

In performance comparison of three open datasets, we provide the results of the average recall scores of ROUGE-1, ROUGE-2 and ROUGE-SU4. We compare the performance of QODE with other representative multi-document summarization algorithms, including graph-based sentence ranking algorithms such as: Manifold-ranking model (Wan & Xiao, 2009), Multiple-modality model (Wan, 2009), and Document-sensitive model (Wei, et al., 2010); supervised learning based sentence ranking algorithms: SVM Classification (Vapnik, 1995), Ranking SVM (Jochims, et al., 2002) , Regression (Ouyang, et al., 2011); classical relevance and redundancy based selection algorithms: greedy search (Filatova & Hatzivassiloglou, 2004), maximum marginal relevance (MMR) (Goldstein, et al., 2000), integer linear program (ILP) (McDonald, 2007); and the NIST baseline system (Dang, 2005).

Performance Comparison

Firstly, we compare the performance of the proposed techniques with other representative ones on three standard datasets based on ROUGE scores. From the results of DUC 2005 shown in Table 1, it is obvious that our algorithm outperforms most of existing algorithms.

Table 1. Comparison to representative algorithms on the DUC 2005

System	ROUGE-1	ROUGE-2	ROUGE-SU4
QODE	0 3751	**0.0775***	**0.1341***
Manifold-ranking	**0.3839***	0.0737	0.1317
Multiple-modality	0 3718	0.0676	0.1293
Document-sensitive		0.0771	0.1337
SVM Classification	0 3663	0.0701	0.1243
Ranking SVM	0 3702	0.0711	0.1299
Regression	0 3770	0.0761	0.1329
Greedy search	0 3560	0.0610	
MMR	0 3701	0.0701	0.1289
ILP	0 3580	0.0610	
NIST Baseline		0.0403	0.0872

In proposed QODE, we integrate query information in concept extraction, layer-wise reconstruction, and summary generation. Table 2 shows the query oriented contribution analysis in three stages. Obviously, each step has its own contribution to the final summary generation.

Table 2. Query Oriented Contribution Analysis

Method			ROUGE-1	ROUGE-2	ROUGE-SU4
1	2	3			
√	√	√	0.3751	0.0775	0.1341
√	√		0.3731	0.0742	0.1315
	√	√	0.3734	0.0755	0.1329
√		√	0.3704	0.0740	0.1301

1. Query oriented initial weight setting, 2. Query oriented penalty process, 3. Summary importance maximization by DP

In Table 3 and 4, we also provide the performance comparison on DUC 2006 and DUC 2007. As an unsupervised learning algorithm, the performance of QODE is similar to the supervised learning based regression algorithm (Joachims, 2002). Therefore, we can still conclude that our system is able to achieve state-of-the-art performances giving the sufficient results listed above.

Table 3. Comparison to representative algorithms on the DUC 2006

System	ROUGE-1	ROUGE-2	ROUGE-SU4
QODE	0.4015	**0.0928***	0.1479
Manifold-ranking	**0.4101***	0.0886	0.1420
Multiple-modality	0.4031	0.0851	0.1400
Document-sensitive		0.0899	0.1427
SVM Classification		0.0834	0.1387
Ranking SVM		0.0890	0.1443
Regression		0.0926	**0.1485***
NIST Baseline		0.0491	0.0962

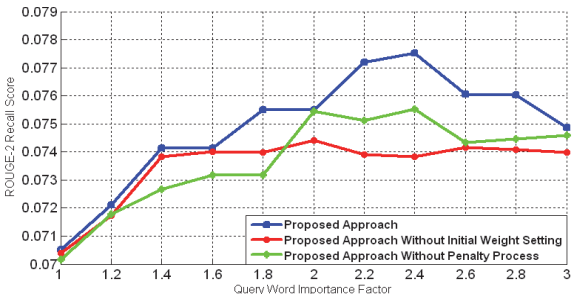
Table 4. Comparison to representative algorithms on the DUC 2007

System	ROUGE-1	ROUGE-2	ROUGE-SU4
QODE	0.4295	0.1163	**0.1685***
Manifold-ranking	0.4204	0.1030	0.1460
Multiple-modality		0.1123	0.1682
Document-sensitive	0.4211	0.1103	0.1628
SVM Classification		0.1075	0.1616
Ranking SVM	**0.4301***	**0.1175***	0.1682
NIST Baseline	0.3091	0.0599	0.1036

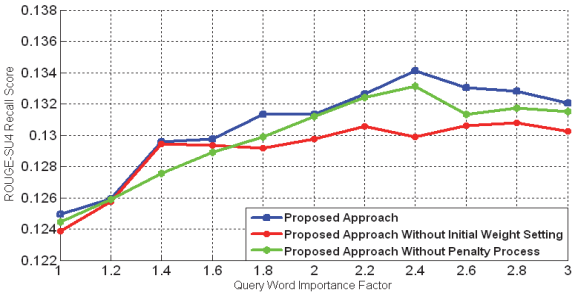
Parameter Tuning

There are numerical meta-parameters in the proposed techniques. For the parameters related with deep model, such as learning rate and the momentum, we simply follow the general setting (Hinton, 2010), although more careful choice may lead to better performance. The structure of deep learning model is another set of parameters, different with existing deep learning techniques that determine the structure such as the number of hidden layers based on intuition. We intend to provide more meaningful architecture by considering the characters of document summary task. In our implementation, the hidden layer H^1 is used to filter out the words appearing accidentally and 1000 hidden units are used in this paper. Hidden layer H^2 is supposed to discover the key words; therefore, the number of hidden units depends on the length of summary. In our experiment, the length is predetermined by the DUC tasks with allowance of 250 words, so we use 250 hidden units in H^2. Hidden layer H^3 is assumed to extract the candidate sentences for the summary. If the length of the summary is equal to 250 words, 10-hidden-unit is a reasonable setting of H^3.

For the parameter used in dynamic programming, we discuss the influence to ROUGE result on DUC 2005 from query word importance λ. Figure 2(a) shows the value of ROUGE-2 and Figure 2 (b) shows the value of ROUGE-SU4 when λ varies from 1 to 3. At most time, the proposed technique has the best performance. ROUGE-2 and ROUGE-SU4 peak together when λ is equal to 2.4. Similar to DUC 2005, the peak points of ROUGE-2 and ROUGE-SU4 curve can be obtained when λ is equal to 2.4 on DUC 2006 and DUC 2007.



(a) ROUGE 2 Recall performance vs. λ



(b) ROUGE SU4 Recall performance vs. λ

Figure 2: Performance comparison vs. λ.

Physical Information in Deep Network Analysis

Furthermore, we want to demonstrate the rationale of the proposed techniques, whether QODE really has advanced extraction ability. To demonstrate the extraction ability of proposed QODE, we analyze the information coverage in every layer using one document set D376e. In the document set D376e. There are 26 documents in this set and 9 human summaries are provided. For dataset D376e, the number of nodes in layer H^0 is equal to 2032. In our experiment, we set the number of hidden nodes in layer H^1 to 1000. So we keep 1000 words pushed out by H^1 with higher extraction weights and calculate how many of them appear in human's summary. We also calculate the percentage according to the filtering out 1032 words. From Table 5, obviously, deep networks intend to find the informative words.

Table 5. The statistical analysis of words in layer H^1

Words	Numbers	In Human Summary	Percentage
Filtering out words	1032	65	6.3%
Remaining words	1000	211	21.2%

In layer H^2, the number of hidden layer is reduced to 250. As previously, we calculate how many words pushed out by H^2 with higher extraction weights appear in human's summary in Table 6. The words of human summary coverage percentage is about 40%, which is nearly doubled to layer H^1. For the convenience of comparison, we randomly select 250 words from 2032 and calculate that how many of them appear in the human's summary. We repeat the experiments ten times and calculate the average percentage. Comparing these two results, the proposed techniques demonstrate the extraction ability again.

Table 6. The statistical analysis of words in layer H^2

Words	Number	In Human Summary	Percentage
Random words	250	34	13.6%
Key words	250	99	39.6%

There are ten hidden units in layer H^3 that corresponds to the ten sentences appearing in the 250-words summary. In Table 7, we list ten candidate sentences related with corresponding nodes. To compare with human summary, the ID numbers of human summary which has similar sentence are also listed. Therefore, inheriting the distinguished extraction ability from deep learning model, proposed QODE pushes out important concepts layer by layer effectively.

Table 7: Candidate sentence extracted in layer H^3

Sentence with Union of Key Words in Automatically Extracted Summary	Id of Human's Summary
An **international war** crimes **tribunal** covering the former **Yugoslavia** formally opens in The Hague **today** with a request for the extradiction from **Germany** of a Bosnian Serb **alleged** to have killed three Moslem **prisoners**.	A,B,C,D,E,G,H,I,J
The extradiction is important to the **tribunal** - the first **international war** crimes **court** since the Nuremberg trials after the second **world war** - because it has no power to try **suspects** in absentia.	B,C,D,E,G,H,I,J
World News in Brief: **Court rules** on **border**.	A,C,D,E,G,H,I,J
The International **Court** of Justice in The Hague ruled in Chad's favour in a 20-year **border** dispute with **Libya** which has **caused** two **wars**.	B,D,E,H,I,J
Maybe we'll go full circle; the World **Court** can condemn this action and then the **Soviets** can defy that body, just as the **United States** defied the **court's** condemnation of our **embargo** of Nicaragua.	C,D,G,I,J
Ever since the Reagan Administration walked out of the Hague to protest Nicaragua's **claim** of illegality in U.S. aid to the Contras, the **State** Department has opposed submitting to the **World Court** any **case** that involves the use of military force.	H,I,J
They refused to appear in the **World Court** 10 **years ago** when Washington sought the release of American hostages in Tehran.	H,I,J
A **year** after **Noriega's** capture, the **court** was still hearing arguments on whether Bush could be subpoenaed and the **World Court** was in preliminary hearings on Panama's complaint.	J
After six **months** of uproar, the U.S. district **court** judge in Miami **ordered** that the **case** proceed to trial.	Null
Mr Edwin Williamson, a **legal** adviser to the U.S. **State** Department who will address the court later in the proceedings, said **yesterday**, 'This (**court**) action in no way inhibits what the Security Council is doing. '	Null

Conclusion

This paper proposes a novel deep learning model for query-oriented multi-documents summarization. Inheriting the distinguished extraction ability from deep learning, the proposed framework pushes out important concepts layer by layer effectively. According to the empirical validation on three standard datasets, the results not only show the distinguishing extraction ability of QODE but also clearly demonstrate our intention of providing a human-like multi-document summarization for nature language processing.

Acknowledgments

This research was supported by HK PolyU 5183/11E.

References

Ballan, L.; Bazzica, A.; Bertini, M.; Bimbo, A. D. ; Serra, G. (2009). Deep Networks for Audio Event Classification in Soccer Videos. In Proc.of ICME.

Baxendale, P. B. (1958). Machine made index for technical literature an experiment. IBM Journal of Research Development.

Berger, A., and Mittal ,V. (2000). Query relevant summarization using FAQs. In Proc. of ACL.

Cao, Z., Qin. T., Liu, T., Tsai, M., Li, H., (2007). Learning to Rank: from Pairwise Approach to Listwise Approach. In Proc. of ICML.

Clancey, W. J. (1979). Transfer of Rule Based Expertise through a Tutorial Dialogue. Ph.D. diss., Department of Computer Sci ence, Stanford University, Stanford, CA.

Cao, L.; Yu, J.; Luo, J. and Huang, T. S. (2009). Enhancing Se mantic and Geographic Annotation of Web Images via Logistic Canonical Correlation Regression. In Proc.of ACMMM.

Chen, E.K., Yang, X.K., Zha, H.Y., Zhang, R., and Zhang, W.J. (2008). Learning Object Classes from Image Thumbnails through Deep Neural Networks. In Proc. of ICASSP.

Dahl, G.; Ranzato, M.; Mohamed, A.; Hinton, G. E. (2010). Genearting more realistic images using gated MRF's. In Proc. of NIPS.

Dang, H.T. (2005). Overview of DUC 2005. In Proc. of DUC 2005.

Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM.

Filatova, E., and Hatzivassiloglou, V., (2004). A formal model for information selection in multisentence text extraction. In Proc. of COLING.

Porter, M.F. (1980). An algorithm for suffix stripping," Program.

Goldstein, J., Mittal, V., Carbonell, J., and Kantrowitz, M.. (2000). Multi Document Summarization by Sentence Extraction. In Proc. of the ANLP/NAACL Workshop on Automatic Summa rization.

Hinton, G.E. (2002). Training products of experts by minimizing contrastive divergence. Neural Computation.

Hinton, G. E.;Osindero, S.;The, Y. (2006). A Fast Learning Algo rithm for Deep Belief Nets. Neural Computation.

Hinton, G.E. (2010). A practical guide to training restricted Boltzmann machine. Tech. rep., 1 21, University of Toronto.

Joachims, T. (2002). Optimizing search engines using click through data. In Proc. of KDD.

Jin, F., Huang, M.L., and Zhu, X.Y. (2010). A Comparative Study on Ranking and Selection Strategies for Multi Document Summarization. In Proc. of COLING.

Lee, T., Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex", In JOSAA.

Leuba, G., Kraftsik, R. (1994). Changes in volume, surface esti mate, 3 dimensional shape and total number of neurons of the human primary visual cortex from midgestation until old age. In Inat. Embryol..

Larochelle, H.; Erhan, D.; Courville, A.; Bergstra, J.; and Bengio, Y. (2007). An Empirical Evaluation of Deep Architectures on Problems with Many Factors of Variation, In Proc.of ICML.

Lin, C.Y. (2004). Rouge: A package for automatic evaluation of summaries. In Proc. ACL.

Liu, Y.; Xu, D.; Tsang, I. W.; Luo, J. (2009). Using Large Scale Web Data to Facilitate Textual Query Based Retrieval of Con sumer Photos. In Proc.of ACMMM.

Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research Development.

Khanpour, H. (2009). Sentence Extraction for Summarization and Notetaking. PhD. Diss.. University of Malaya.

McDonald, R. (2007). A study of global inference algorithms in multi document summarization. Lecture Notes in Computer Sci ence.

Ouyang, Y. Li, W.J., Li, S.J., Lu, Q. (2011). Applying regression models to query focused multi document summarization. Infor mation Processing and Management.

Radev, D. R., Jing, H.Y., Stys, M., and Tam, D. (2004). Centroid based Summarization of Multiple Documents. Information Processing and Management.

Shen, D., Sun, J.T., Li, H., Yang, Q., and Chen, Z. (2007). Document Summarization using Conditional Random Fields. In Proc. of. IJCAI.

Smolensky, P. 1986. Information Processing in Dynamical Sys tems: Foundations of Harmony Theory. Parallel Distributed Pro cessing: Explorations in The Microstructure of Cognition.

Tang, J., Yao, L., and Chens, D. (2009). Multi topic based query oriented summarization. In Proc. of SLAM International Conference on Data Mining.

Taylor,G., Fergus, R., Cun, Y.L. and Bregler, C. 2010. Convolu tional learning of spatio temporal features. In Proc. of. ECCV.

Wan, X., Xiao, J.. (2008). Single document keyphrase extraction using neighborhood knowledge. In Proc. AAAI.

Wan, X. & Xiao, J.. (2009). Graph Based Multi Modality Learn ing for Topic Focused Multi Document Summarization. In Proc. of IJCAI.

Wan, X.. (2009). Topic Analysis for Topic Focused Multi Document Summarization. In Proc. of CIKM.

Vapnik, V. N. (1995). The nature of statistical learning theory. Springer.

Wei. F., W.J. Li, Q. Lu, and Y.X. He, (2010). A document sensitive graph model for multi document summarization. Knowledge and Information Systems.

Wong, K.F., Wu, M.J., and Li, W.J. (2008). Extractive Summarization Using Supervised and Semi supervised Learning. In Proc. of. ICCL.

Zhong, S.H., Liu, Y., Liu, Y.. 2011. Bilinear deep learning for image classification. In Prof. of ACMMM.