2015 8th International Symposium on Computational Intelligence and Design

Automatic Document Summarization via Deep Neural Networks

Chengwei Yao
College of Computer Science and Technology
Zhejiang University
Hangzhou, China
yaochw@zju.edu.cn

Jianfen Shen
The Health Information Center of Zhejiang
Hangzhou, China
sjf_hz@126.com

Gencai Chen
College of Computer Science and Technology
Zhejiang University
Hangzhou, China
chengc@zju.edu.cn

Abstract—Automatic document summarization aim to extracting sentences which might cover the main content of a document or documents. To achieve this, many algorithms have been tried to rank the sentences by using task-specific features in a shallow architecture. The main challenge of these approaches is to keep balance between information coverage and redundancy because of absence of discovering the intrinsic semantic representation. Inspired by the recent successful achievement of Deep Learning, this paper proposes a new framework of document summarization via Deep Neural Networks (DNNs). Specifically, we feed the sentences as the input to the visible layer of DNNs. After pretraining layer by layer and fine-tuning, the lower dimensional semantic space can be revealed. Based on this space, we design sentences extraction algorithm to construct the summary. Experiments on the DUC2006 and DUC2007 dataset show that our framework works better than state-of-the-art methods.

Keywords-Document Summarization; Deep Learning; Deep Brief Network; Restricted Boltzmann Machine; Auto-encoder; ROUGE

I. INTRODUCTION

Automatic document summarization has been addressed by Natural Language Processing (NLP) and Text Mining community for more than half century. Although many frameworks and algorithms have achieved improvement in many benchmarks or task-specific applications, it is still the challenge job to summarize texts automatically in an intelligent way.

Many of the existing generic approaches use ranking models to extract sentences from a candidate set. To approach this, different kinds of task-specific features have been proposed, such as term-frequency, sentence position, sentence length, word pairs, lexical chains, rhetorical structure, proper names, etc. [15]. These kinds of explicit representations by the handcraft features make these methods quite difficult to leverage the good information coverage and less information redundancy.

In recent years, Document Summarization Based on Data Reconstruct (DSDR) [1] has been proposed. Instead of providing some handcraft features, DSDR aim to select sentences by minimizing reconstruction error through linear and nonlinear reconstruct of documents. Although this model works as end-to-end fashion and get state-to-the-art performance, it is still based on a shallow architecture, which limits its capacity to discover the intrinsic representation of semantic features.

Inspired by the successful achievement of Deep Learning, this paper proposed a new framework of document summarization via Deep Neural Networks (DNNs). More specifically, we build a multi-layer neural network with one visible layer and four hidden layers which contain 1000, 500, 500, 128 stochastic binary units respectively. Every two adjacent layers form a structure of Restricted Boltzmann Machine (RBM) [2,3]. We feed the sentences as input to the visible layer. After pretraining layer by layer and fine-tuning the weights [6,9], the top hidden layer with 128 units gets the sentence codes, which preserve intrinsic semantic features of sentences in a lower dimensional space. In this new semantic space, similar sentences will stay close, and we extract sentences which have high density of neighbors based on k-Nearest Neighbor Search. The qualitative and quantitative analysis in the experiments on the datasets of DUC2006 and DUC2007 shows that this document summarization framework works better than state-of-the-art methods.

II. RELATED WORK

A. Document Summarization

Automatic summarization from a single document or multiple documents has been studied from many perspectives, and by using various paradigms in different domains. Most of the works in the early years aim to design some task-specific features, and then apply ranking algorithms to select the sentences based on these features, such as Naïve-Bayes methods [16], Decision Tree [17], Hidden Markov Model(HMM) [18], log-linear models [19], Maximal Marginal Relevance (MMR) [20], Graph Spreading Activation [21], etc.. Although these models have made improvement, they suffer from the problem that top ranked sentences usually share much redundant information.

In recent years, there are some new approaches, which less depend on task-specific features [13,14]. Wan and

978-1-4673-9587-8/15 $31.00 © 2015 IEEE
DOI 10.1109/ISCID.2015.83

Yang [22] proposed a graph based model that can improve the document summarization by integrating topic modeling. DSDR [1,12] cast the summarization task to data reconstruction problem. There are also lexical chain based model [15], which discover the semantic relations of terms in the same semantic role by using WordNet. However, they are based on shallow architectures, which limit their capacity to discover the distributed representation of intrinsic features.

B. Deep Learning

Theoretical results show that the shallow architectures are incapable of extracting certain types of complex structure from rich sensory input [8,10]. By contrast, study on visual cortex of human show that object recognition uses many layers of nonlinear processing and requires very little label input [30]. However, models with multilayer neural networks are very hard to train due to non-convex of their associated loss functions, until Hinton et al. [2] discover an effective way that learning deep neural network by greedy layer by layer strategy. This learning algorithm is based RBMs, in which gradient descent can approximately be estimated by Contrastive Divergence (CD) [3]. Based on that, Deep Belief Networks (DBNs) [2] and Auto-Encoder [31] have been proposed, and get surprising empirical results on many machine learning tasks.[9,11] It is the reason this paper follows this learning strategy to extract meaningful sentences from documents.

Hinton and Salakhutdinov also propose Deep Boltzmann Machine [4], which is fully undirected model, and let itself have ability to learn internal representations from lower and higher layers. Another very successful Deep model is Deep Convolutional Networks [7,23], which beat many shallow architecture models in the competitions on handwritten recognition and object recognition.

III. The Proposed Framework

This paper tries to discover the intrinsic features of sentences without using any kinds of sophisticated features. Instead, we just create word count vector for sentences as inputs to the DNN (Fig. 1). The framework of automatic document summarization via DNN is as below:

- After stemming and stop-word elimination, we decompose documents into individual sentences and create sentence vectors by word count. We feed the sentence vectors to the first layer of DNN, whose size is decided by the size of dictionary.
- Up the first layer, there are multi-hiddenlayers, which contains 1000, 500, 500, 128 units respectively. At pre-training stage, each two adjacent layers are connected by symmetric weights and build a RBM (Fig.1: solid boxes), which is undirected graphical model. Let W1, W2, W3, W4 be weight matrices between each two adjacent layers.
- After the weights of RBMs have been pre-trained by using Contrastive Divergence(CD), the network is unrolled to form a deep feed forward neural network (Fig.1: solid boxes and dashed boxes) to reconstruct sentence vectors, so that weights can be fine-tuned by back-propagation.
- After fine-tuning, we obtain the sentence codes at 128-units hidden layer by using these optimized weights. The sentence codes span a new lower dimensional space, where sentences with similar semantics will stay close. Under the assumptions that good candidates for summarization will have more close neighbors in this new space, we propose a simple algorithm to extract sentences based on k-Nearest Neighbor Search.

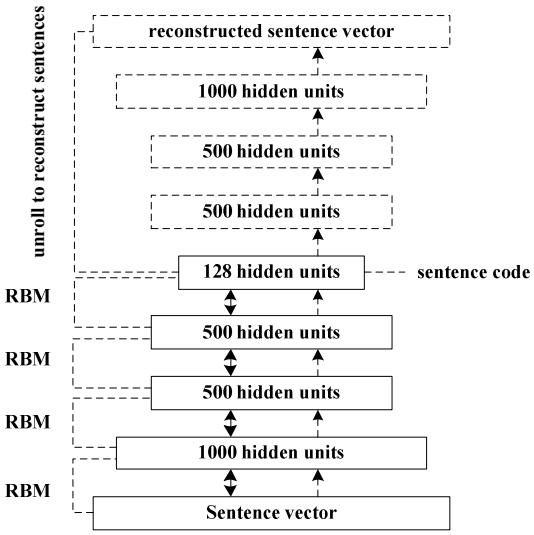

Figure 1. The framework of sentence encoding via DNN

A. Pretraining the Weights

In [5], Replicated Softmax Units (RSM) was proposed to model the input vectors of documents in DBN. RSM Units can handle the different length of documents, but its computation is intensive because it treats every word as the input to visible layer, and replicate D times by sharing the same weights, where D is the length of a document. This paper adopts a more simpler and straightforward way to build sentence vector. Let $v = [v_1, v_2, \dots, v_D] \in N^D$ be the sentence vector, where v_i is the count of words appeared in the sentence, D is the dictionary size, and let $h \in \{0,1\}^F$ be binary stochastic hidden features. $\{v, h\}$ form a RBM, which is Undirected Graphical Model, whose energy function is:

$$E(v,h) = -\sum_{i=1}^{D}\sum_{j=1}^{F} W_{ij}\, h_j v_i - \sum_{i=1}^{D} v_i b_i - \sum_{j=1}^{F} h_j a_j \,, \qquad (1)$$

where W_{ij} is a symmetric interaction weight between visible unit i and hidden unit j. F is the size of hidden layer, a_j , b_i is the bias of unit i and unit j,. The probability of visible units is:

$$P(v) = \frac{1}{Z}\sum_{h} exp\big(-E(v,h)\big), \qquad (2)$$

$$where: Z = \sum_{v}\sum_{h} exp\big(-E(v,h)\big).$$

Z is the partition function, which has exponential terms and is always intractable. The conditional probabilities of visible units and hidden units are given as below:

$$P(v_i|h) = \sigma\left(b_i + \sum_{j=1}^{F} h_j W_{ij}\right), \quad (3)$$

$$P(h_j|v) = \sigma\left(a_i + \sum_{i=1}^{D} v_i W_{ij}\right), \quad (4)$$

where $\sigma(x) = 1/(1 + exp(-x))$ is the sigmoid function, which let RBM be the nonlinear model.

Given a collection of N sentences: $\{v^{(n)}\}_{n=1}^{N}$, the derivative of the log-likelihood with respect to parameters W can be derived to the result as below[3]:

$$\frac{1}{N}\sum_{n-1}^{N}\frac{\partial log P(v^{(n)})}{\partial W_{ij}} = E_{P_{data}}[v_i h_j] - E_{P_{model}}[v_i h_j], \quad (5)$$

where $E_{P_{data}}[\cdot]$ denotes an expectation with respect to the data distribution, and $E_{P_{model}}[\cdot]$ denotes an expectation with respect to distribution defined by the model. Exact evaluate $E_{P_{model}}[\cdot]$ is intractable, because it takes time that is exponential to $\min\{D, F\}$. Therefore, here CD has been used to approximate to the gradient:

$$\Delta W_{ij} = \epsilon\left(E_{P_{data}}[v_i h_j] - E_{P_{recon}}[v_i h_j]\right), \quad (6)$$

where ϵ is the learning rate and P_{recon} is a distribution by running the Gibbs chain, which is initialized by the data, do the iteration T times(CD-T) between visible layer and hidden layer. When $T \to \infty$, the distribution will settle down to equilibrium and P_{recon} become a good approximation to the P_{model}. Many experiments in recent years show that CD-1 can also work quite well, and this paper choose CD-1 procedure to train the model, so that it can reduce learning time significantly.

A single hidden layer can not fully capture the features of visible data. So after one hidden layer has been trained, a new hidden layer is appended. Previous hidden layer now is treated as the visible layer, and a new RBM is built. This greedy layer by layer training is call pretraining, which can move weights to the proper region, where the features of inputs are mostly reserved.

B. Fine-tuning the Weights.

After pretraining, individual RBMs at each level are "unrolled" (Fig.1) to create a deep feed forward neural network, by which sentence vectors will be reconstructed. In this fine-tuning stage, the stochastic activity units in each hidden layer are replaced by deterministic, real-valued probabilities, so that we can back-propagate through the entire network to fine-tune the weights. For back-propagation, we use the cross-entropy error function as below:

$$C = -\sum_{n} v_{data}^{(n)} \, log \, v_{reconstruct}^{(n)}, \quad (7)$$

where v_{data} is input sentence vector, $v_{reconstruct}$ is the reconstructed sentence vector.

C. Details of the training.

Here, we use similar learning strategy of DBN. When pretraining, we divide the sentence vectors into mini-batches, and each mini-batch contains 100 sentences cases. It can speedup the training by upgrading the weights for each min-batch. And we choose the learning rate of 0.1, momentum of 0.9. We set 50 epochs for greedily pretraining. For fine-tuning, we make a larger mini-batches, whose size is 1000 cases. And we set 100 epochs for back-propagation when fine-tuning the weights.

D. Extract the sentences

After obtaining 128-dimensional sentence codes, the new semantic space has been revealed, where sentences with similar semantics will stay close together. To see what this space looks like, we visualize the sentence codes by t-SNE toolkit described in the next section.

Algorithm 1: Extract sentences from the sentence semantic space

Input:

- The sentence codes set: $\boldsymbol{H} = \{\boldsymbol{h^{(1)}}, \boldsymbol{h^{(2)}}, \dots, \boldsymbol{h^{(N)}}\}$, where N is total number of sentences;
- The number of sentences to be extracted: m;

Output:

- The set of m sentences for summary: $\mathbf{S} = \{\mathbf{s^{(1)}}, \mathbf{s^{(2)}}, \dots, \mathbf{s^{(m)}}\} \subset \mathbf{H}$

1: for t=1 to m do
2: for each $\boldsymbol{h^{(i)}} \in \boldsymbol{H}$ do
3: $\{\boldsymbol{h^{(i_1)}}, \boldsymbol{h^{(i_2)}}, \dots, \boldsymbol{h^{(i_k)}}\} \leftarrow$ $\boldsymbol{k\ nearest\ neighbor\ search\ fuction}(\boldsymbol{h^{(i)}})$;
4: **compute**: $\mathbf{average\_dist}(\boldsymbol{h^{(i)}})$ between $\boldsymbol{h^{(i)}}$ $\boldsymbol{and}$ $\{\boldsymbol{h^{(i_1)}}, \boldsymbol{h^{(i_2)}}, \dots, \boldsymbol{h^{(i_k)}}\}$
5: end for
6: $\boldsymbol{s^{(t)}} = \mathbf{arg\ min}_{\boldsymbol{h^{(i)}}}\ \mathbf{average\_dist}(\boldsymbol{h^{(i)}})$;
7: record $\boldsymbol{k\ nearest\ neighbors}$ of $\boldsymbol{s^{(t)}}$: $\{\boldsymbol{s^{(t_1)}}, \boldsymbol{s^{(t_2)}}, \dots, \boldsymbol{s^{(t_k)}}\}$;
8: $\boldsymbol{S} \leftarrow \boldsymbol{S} \cup \{\boldsymbol{s^{(t)}}\}$;
9: $\boldsymbol{H} \leftarrow \boldsymbol{H} - \{\boldsymbol{s^{(t)}}\} \cup \{\boldsymbol{s^{(t_1)}}, \boldsymbol{s^{(t_2)}}, \dots, \boldsymbol{s^{(t_k)}}\}$;
10:end for

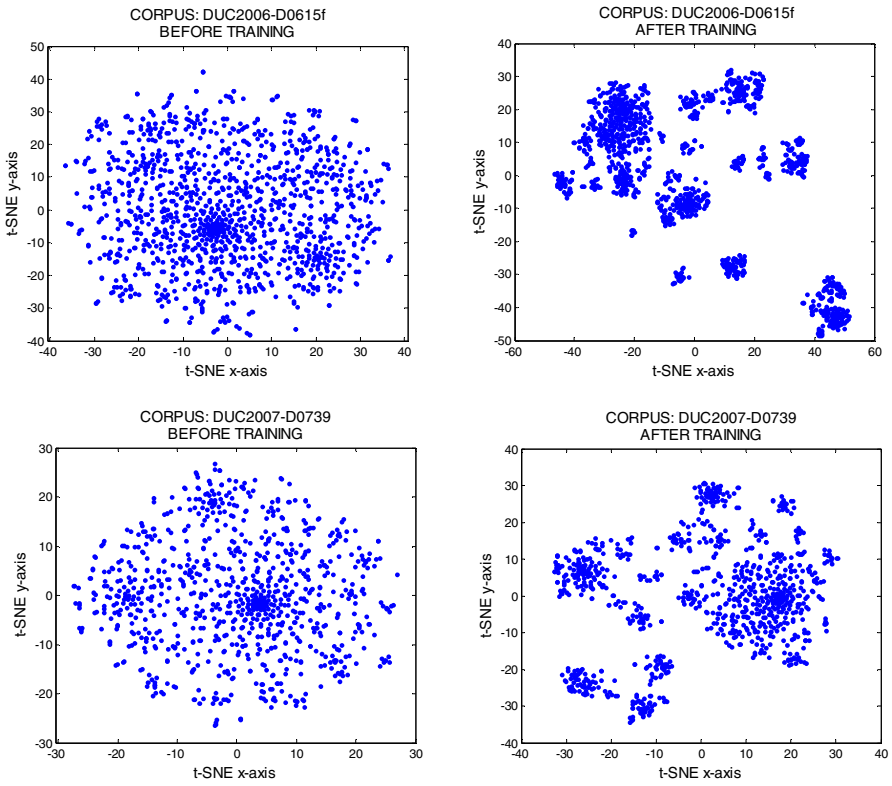

Figure 2. Sentences are visualized in 2D space by t-SNE toolkit, and each point is a sentence. Upper two figures: Visualization of DUC2006-D0615f corpus. Following two figures: Visualization of DUC2007-D0739 corpus. Left: Visualization of sentence vectors without training. Right: Visualization of sentence codes obtained at 128-unites hidden layer.

To extract the sentences properly, this paper assumes that a good candidate for summarization will have most density of neighbors nearby. That means this good candidate will cover more content of the document or the corpus. To confirm this assumption, we do the experiments on DUC2006_SCU and DUC2006_SCU (more details are in the next section). Under this assumption, we compute the average distance between each sentence and its k-nearest neighbors, and find the densest points to construct the summary. The distance metric we used here is 'cosine'.

Another problem is how to extract sentences with less information redundancy. In this new semantics space, it appears easy and straightforward to realize. After exacting a sentence, we remove this sentence and its k neighbors from the sentence set, and do the next extraction. The detail of the algorithm is shown in [Algorithm 1].

IV. EXPERIMENTS

It is difficult to evaluate the quality of automatic summarization, even evaluation by human will have very different result [28]. Therefore, to be comprehensive, this section presents our experiments to analyze the data and results from two perspectives: qualitative and quantitative. Before that, we describe the benchmark data set we used in the experiments.

A. Data Sets

In this experiment, we use the standard summarization benchmark data sets DUC 2006 and DUC 2007, which contain 50 and 45 corpuses respectively, with 25 news articles in each corpus. The sentences in each article have been separated by NIST[1]. More over, DUC also provide sub data set DUC2006_SCU and DUC2007_SCU, in which Summary Content Units (SCU) have been marked in the XML format documents by Pyramid evaluation activity [29]. This provides the useful resource with a measure of the degree to which some number of its individual sentences addresses the information.

B. Qualitative Analysis.

The framework proposed in this paper is unsupervised. To see what happens in the sentence space, we visualize the sentences of two corpuses randomly chosen from DUC2006 and DUC2007 respectively, which are 'DUC2006-D0615f' and 'DUC2007-D0739'. (Fig. 2) Each corpus contains 25 news articles with the same topic. That means the documents in the same corpus have quite same semantic features, but sentences in the documents are not.

The visualization toolkit we use here is t-Distributed Stochastic Neighbor Embedding (t-SNE) [24], which is widely used in machine learning area. It is particularly well suited for the visualization of high-dimensional datasets to 2 or 3 dimensions. We visualize sentence points before training and after training respectively, and as Fig. 2 shows, the DNN framework do well in clustering the sentences.

However, how can we be sure that the assumption we mention in the previous section works, which means in the new semantic space, the good candidate will or will not be located in the areas with highly density of sentence points? To confirm this assumption, we do experiment

[1] http://www.nist.gov/index.html

DUC2006_SCU and DUC2007_SCU, where sentences that contain the SCUs have been marked manually. We compare these marked sentences with the full set of sentences by computing the average distances between sentences and their k-nearest neighbors ($k \in [3,18]$) in the new semantic space. Fig. 3 shows, in the new semantic space, that the SCU marked sentences have quite shorter average distances to their k-nearest neighbors, which confirm our assumption.

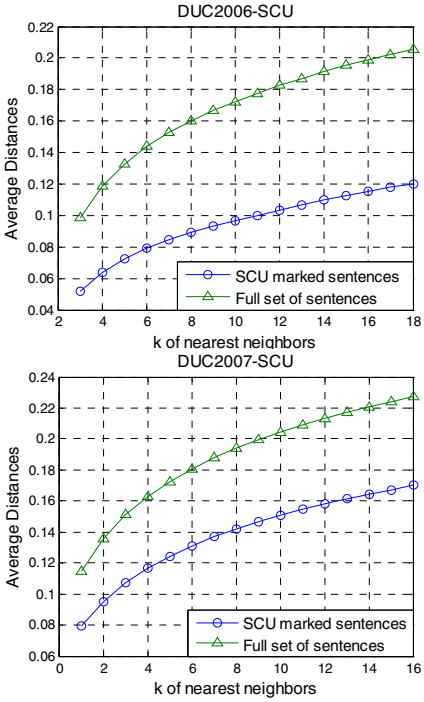

Figure 3: Average distances between sentences and their k-nearest neighbors in the new semantic space.

C. Quantitative Analysis

1) Evaluation Metric: This paper adopts a widely used automatic documents summarization evaluation method: Recall-Oriented Understudy for Gisting Evaluation (ROUGE) toolkit[28], which measures quality of peer summary by counting overlapping units. The units have some kinds of metrics, such as n-gram, the word sequences and word pairs between the peer summary and reference. The standard ROUGE-N computation is as bellow:

$$ROUGE - N$$

$$= \frac{\sum_{S \in Ref} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in Ref} \sum_{gram_n \in S} Count(gram_n)} . \qquad (8)$$

In this experiment, we choose ROUGE-1,2,3 and GOUGE-L. ROUGE-L uses the longest common subsequence(LCS) metric.

2) Compared Methods: As this work is unsupervised, the quantitative evaluation is done by comparing this work with some traditional and state-of-the-art unsupervised summarization methods, which include:

- Random: selects sentences randomly.
- Lead [25]: orders the documents chronologically and takes the leading sentences one by one.
- LSA [26]: Latent Semantic Analysis, which apply SVD on terms by sentences matrix to select highest ranked sentences).
- ClusterHITS [27]: ranks the sentences with the authorities scores.
- DSDR-lin and DSDR-non [1]: treat summarization as linear and non-linear reconstruction, and minimize the reconstruction error.

3) Evaluation Results: ROUGE can generate three types of scores: recall, precision and F-measure. In this experiment, we use F-measure. Table 1 presents the experimental results, and shows that this paper get almost all the best results except DUC2007-ROUGE-1, which might be, in the sense, DNN is good at discover high level and implicit features, but not for the individual words co-occurrence.

TABLE I. AVERAGE F-MEASURE PERFORMANCE ON DUC2006

DUC2006				
Algorithm	Rouge-1	Rouge-2	Rouge-3	Rouge-L
Random	0.28496	0.04302	0.01101	0.25892
Lead	0.27411	0.04739	0.01179	0.23287
LSA	0.25804	0.03711	0.00912	0.23299
ClusterHITS	0.28802	0.05169	0.01298	0.25806
DSDR-lin	0.30801	0.05412	0.01286	0.26998
DSDR-non	0.32879	0.06012	0.01354	0.29679
This paper	***0.33201***	***0.06209***	***0.01656***	***0.31008***

TABLE II. AVERAGE F-MEASURE PERFORMANCE ON DUC2007

DUC2007				
Algorithm	Rouge-1	Rouge-2	Rouge-3	Rouge-L
Random	0.32012	0.05399	0.01289	0.29072
Lead	0.31438	0.06145	0.01709	0.26288
LSA	0.25947	0.03654	0.00912	0.22815
ClusterHITS	0.32901	0.06642	0.01938	0.29611
DSDR-lin	0.35662	0.07083	0.02009	0.31965
DSDR-non	***0.39128***	0.07376	0.01936	0.35199
This paper	0.37378	***0.07588***	***0.02653***	***0.36881***

V. CONCLUSION

Automatic document summarization and summary evaluation are highly challenging jobs despite of the long history of study. It may be because many approaches aim to handle the problem by ranking algorithm based on handcraft features in a shallow architecture. In recent years, Deep Learning has achieved encouraging improvement in many areas of automatic systems and Artificial Intelligence, which shows that multilayer unsupervised learning can discover the intrinsic and high lever distributed representation of the data. Inspired by this, this paper introduces the DNN to discover the sentence semantic space, and based on that, sentences with meaningful semantics can be extracted successfully and

information redundancy can be easily limited. Qualitative and quantitative analysis demonstrate that this deep framework can get the promising results.

ACKNOWLEDGEMENTS

This work was supported by Chinese National 863 Program of 'Demonstration of Digital Medical Service and Technology in Destined Region' (Grant No. 2012-AA02A614)

REFERENCES

[1] He Z.Y., Chen C., Bu J.J., Wang C., Zhang L.J., Cai D., He X. Document Summarization Based on Data Reconstruction, The International Conference of AAAI ,2012.

[2] Hinton G. E., Osindero S., and Y.. A fast learning algorithm for deep belief nets. Neural Computation, 18, pp 1527-1554. ,2006.

[3] Hinton G.E.. Training products of experts by minimizing contrastive divergence. Neural Computation, 14(8):1711-1800, ,2002.

[4] Salakhutdinov R. R. and Hinton G. E.. Deep Boltzmann machines. In Proceedings of the International Conference on Artificial Intelligence and Statistics, volume 12, 2009.

[5] Salakhutdinov R. R. and Hinton G. E.. Replicated Softmax: an undirected topic model. In Advances in Neural Information Processing Systems 22, pages 1607-1614. ,2009.

[6] Salakhutdinov R. R. and Hinton G. E. Discovering Binary Codes for Documents by Learning Deep Generative Models. Topics in Cognitive Science, 2010.

[7] Krizhevsky A., Sutskever I., Hinton G. E. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems (NIPS'12), 2012.

[8] Bengio Y., Li Y., Alain G. and Vincent P., Generalized Denoising Auto-Encoders as Generative Models, in: Advances in Neural Information Processing Systems 26 (NIPS'13), 2013.

[9] [Hinton G. E., Learning multiple layers of representation. Trends in Cognitive Science, 11, 428–434, 2007.

[10] Hinton G. E., McClelland J. L., & Rumelhart, D. E.. Distributed representations. In Parallel distributed processing: explorations in the microstructure of cognition.Volume 1: Foundations (pp. 77–109). Cambridge, MA: MIT Press, 1986.

[11] Salakhutdinov R. R. & Hinton, G. E. Semantic hashing. In Proceedings of the SIGIR Workshop on Information Retrieval and Applications of Graphical Models, 2007.

[12] Cai, D., and He, X. Manifold adaptive experimental design for text categorization. IEEE Transactions on Knowledge and Data Engineering 24(4):707–719., 2012.

[13] Wan, X., and Yang, J. Multi-document summarization using cluster-based link analysis. In Proc. of the 31st ACM SIGIR, 299–306. ACM, 2008.

[14] Wang, D.; Li, T.; Zhu, S.; and Ding, C. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In Proc. of the 31st ACM SIGIR, 2008.

[15] D. Das and .F.T. Martins. A Survey on Automatic Text Summarization. Literature Survey for the Language and Statistics II course at Carnegie Mellon University, 2007

[16] J. Kupiec, J. Pedersen, and Chen, F.. A trainable document summarizer. In Proceedings SIGIR '95, pages 68-73, New York, NY, USA, 1995

[17] Lin C.Y. Training a selection function for extraction. In Proceedings of CIKM '99, pages 55-62, New York, NY, USA. 1999.

[18] J. M. Conroy and O'leary, D. P.. Text summarization via hidden markov models. In Proceedings of SIGIR '01, pages 406-407, New York, NY, USA. 2001.

[19] M. Osborne, Using maximum entropy for sentence extraction. In Proceedings of the ACL'02 Workshop on Automatic Summarization, pages 1-8, Morristown, NJ, USA. 2002.

[20] J. Carbonell and J. Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of SIGIR '98, pages 335-336, New York, NY, USA. 1998.

[21] I. Mani, and E. Bloedorn. Multi-document summarization by graph search and matching. In AAAI/IAAI, pages 622-628. 1997.

[22] Wan, X., and Yang, J.. CollabSum: exploiting multiple document clustering for collaborative single document summarizations. In Proc. of the 30th annual international ACM SIGIR, 150. ACM, 2007.

[23] Koray K., Pierre S., Y-Lan B., Karol G., Michaël M. and Yann L., Learning Convolutional Feature Hierachies for Visual Recognition, (NIPS), 2010.

[24] [24] L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

[25] Wasson, M.. Using leading text for news summaries: Evaluation results and implications for commercial summarization applications. In Proc. of the 17th international conference on Computational. linguistics-Volume 2, 1998.

[26] Gong, Y., and Liu, X.. Generic text summarization using relevance measure and latent semantic analysis. In Proc. of the 24th ACM SIGIR, 19–25. ACM, 2001.

[27] Wan, X., and Yang, J.. Multi-document summarization using cluster-based link analysis. In Proc. of the 31st ACM SIGIR, 299-306. ACM, 2008.

[28] Lin, C.Y., ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain, July 25 – 26. 2004.

[29] Copeck T., Inkpen D., Kazantseva A., Kennedy A., Kipp D., Szpakowicz S. Catch What You Can, Proceedings of the Document Understanding Conference(DUC), 2007.

[30] Lee TS, Mumford D, Romero R, Lamme V. The role of the primary visual cortex in higher level vision. Vision Res. 38:2429–54. 1998.

[31] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot and Yoshua Bengio, Contractive Auto-Encoders: Explicit invariance during feature extraction, in: Proceedings of theTwenty-eight International Conference on Machine Learning (ICML'11), 2011