HAL open science

Research on the Key Techniques of Semantic Mining of Information Digest in the Field of Agricultural Major Crops Based on Deep Learning

Hao Gong, Yunpeng Cui, Ping Qian

To cite this version:

Hao Gong, Yunpeng Cui, Ping Qian. Research on the Key Techniques of Semantic Mining of Information Digest in the Field of Agricultural Major Crops Based on Deep Learning. 11th International Conference on Computer and Computing Technologies in Agriculture (CCTA), Aug 2017, Jilin, China. pp.496-507, 10.1007/978-3-030-06179-1_49 . hal-02111530

HAL Id: hal-02111530

https://inria.hal.science/hal-02111530v1

Submitted on 26 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Research on the Key Techniques of Semantic Mining of Information Digest in the Field of Agricultural Major Crops Based on Deep Learning

Hao G.J.M Gong[1], Yunpeng Cui[2(✉)], and Ping Qian[3]

[1, 2, 3] Agricultural Information Institute of CAAS, Beijing, China

[1] flygonghao@163.com

[2, 3] {cuiyunpeng,qianping}@caas.cn

Abstract. Nowadays application scopes of deep learning research in the machine learning subfield have been gradually expanded, mainly in the field of computer vision and natural language processing. However, in the latter NLP field, there is very little semantic excavation research on agricultural literature data. This paper bases on the attempting to combine relevant paradigms of semantic mining techniques and characteristics of agricultural digest data, for the service of providing new methods and technologies of information acquisition and analysis in the agricultural information domain. Data cleaning methods and data mining experiment are mainly based on deep learning algorithms, which are Seq2Seq and attention mechanism. Finally, through qualitative evaluation and quantitative evaluation of the experimental results, which based on the ROUGE evaluation index system, the experiment shows that the semantic mining model has reached the optimal level of model evaluation in the certain range.

Keywords: NLP; semantic mining; deep learning; ROUGE

1 Introduction

Now the network and other mediums produce more and more text contents, which include substantial agricultural science and technology information. Traditionally, agricultural researchers generally find a bunch of related agriculture literatures through web search engines, but the topics that they may contain may be more than just one, and the article may discuss the interactions and relationships between different themes. Those require researchers to have to browse the abstract of each article and even the entire contents, which stretches out searching time, makes the accuracy of access to information questionable and so on. Then, in the era of the gradual development of the natural language processing field, how to exactly grasp the gist of agricultural science and technology articles, accurately understand the themes of that, and better provide the knowledge service to the researchers become the research mainstream.

Although lack of rigorous theoretical basis, deep learning algorithms significantly reduce the threshold of application for the machine learning technology to bring the practice of engineering convenience. As for why it is popular at this time, there may be three basic reasons. First, the current development of the algorithm is more mature.

Second, different from the past, there are huge quantities of datasets right now. Third, computer hardware computing capacity is more powerful. Why not use Chinese abstract information for semantic mining is that the original automatic word segmentation and part of speech are in the initial stage of Chinese analysis and understanding process. It should provide initial information for the next step of syntactic and semantic analysis. But they need some sort of syntactic and semantic knowledge to be completed. In this way, automatic word segmentation needs to rely on certain results of syntactic and semantic analysis as a condition. The reason why this "circular argument" arises is that this research are using a mechanical process to simulate a human language process that is far from the real language of mankind and is too superficial for current level of research this research can make qualitative progress on this issue, so the automatic word segmentation and part of speech in Chinese automatically mark such a seemingly basic and simple problem, in a short time Chinese NLP researchers can not get to stand the test, and widely recognize results.

This paper will focus on the use of deep neural network techniques combined with the characteristics of information digest in the field of agricultural major crops to provide new methods and technologies of information acquisition and analysis in the agricultural information field.

2 The Relevant Theories of This Semantic Mining Research

2.1 Recursive Neural Network

Direct extraction of important sentences of the extraction method is relatively simple, such as PAM algorithm realizing(TextRank). While generating sentences (re-generate a new sentence) is more complex, but its effect is not satisfactory. In the traditional forward neural network, there exists basically no way to remember the state of a certain period of time. This is a problem when research needs to deal with timing activities, such as dealing with the state memory of sentence sequence in a segment, which is why researchers use recursive neural networks.

2.1.1 Simple Recursive Neural Network

A neural network takes a sequence $x_0,\ldots,x_N$ as input. Each element in the sequence is processed by the same node, and the node is called a state vector in a memory space and memorizes useful information. Then, during the output or all the time to end, this memory space is used to produce a solution to the problem. It can also be said that for each time step t, RNN has the state vector to be calculated:.

$$h_t = \sigma(U * x_t + W * h_{t-1}) \tag{1}$$

Here is a non-linear function, such as sigmoid function, tanh function or ReLU (modified linear unit) function, is initialized to a value, U and W are two weight vectors.

Then, an output can be calculated using the state vector at time t:

$$output_t = softmax(V * h_t) \tag{2}$$

V is the weight vector. Sometimes, only the final output will be used. Here, RNN normally takes the softmax normalization function. It is effective to use probability distributions between hierarchies. For example, if research has a set of terms, it will get the probability distribution of the group of words, and then choose one of them. More formally, if RNN has the K layer, each layer is represented by j, and then these

terms are converted into the following probability:

$$\mathrm{softmax}(z)_j = \frac{exp(z_j)}{\sum_{i=1}^{K} exp(z_i)} \tag{3}$$

Softmax function can be understood to be derived from the fact that if it is significantly larger than the other terms, then it is only one that needs to be considered (which has a value close to 1). It is one of the probability largest functions.

2.1.2 LSTM

LSTM (long and short memory network, Hochreiter et al., 1997) authors believe that the traditional RNN will be hidden as a model of the memory module, and other parts of the network has a direct connection, not only makes the model to expand the number of layers, resulting in the disappearance of the gradient problem, and make the effective historical information by a steady stream of new input data and can not save for a long time. LSTM redesigned the RNN memory module, consisting of a memory cell and a plurality of gate gates. LSTM uses the state of the memory cell to hold the history information. The use of input data for states updates and the operation of outputting state information are controlled by input gate and output gate respectively. When the input gate is closed, the history information is not disturbed by the new input data and is saved as is (constant error carrousel). Similarly, only the output gate is opened and the historical information in the memory cell is active.

2.1.3 Seq2seq (Sequence to Sequence)

Seq2seq is formally introduced in 2014 (Cho et al., 2014). The basic idea is to use the recursive unit to convert the input sequence into an output sequence. For example, this article is to enter a sequence of words or characters from the text of the article, which are converted into a summary of the paragraph (title sequence or small snippet of the summary). Seq2seq can be decomposed into two parts in general. The first part is called the encoding part, which translates the input into a separate vector (which can be understood as encoding the input), which is generated by using, for example, GRUs or LSTMs.

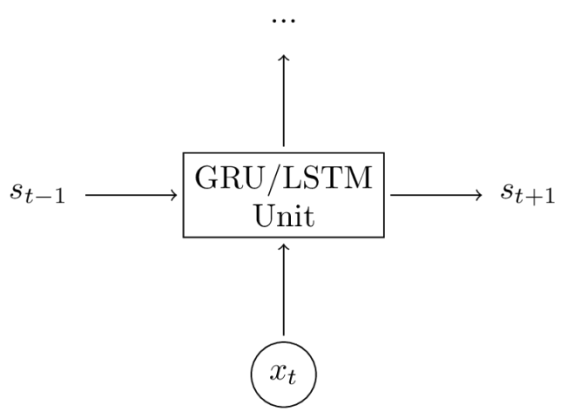

Fig. 1. Encoding input sequence x_t

The input is divided into several representations (in general terms), each representing an input as a gate. Then, once this research get the vector at the end of the encoding, this research assign it to the decoding section. The encoding part also has a recursive neural network, whose initial state is the final state of the code, and its input is the last generated output. Each recursive unit produces an output until the end of the sequence is generated. It should be noted that the coding part does not produce an output for each step. Instead, it enters a word at each step until the input sequence

is entered, and then the semantic correlation information betthis researchen the sequences is captured in its internal state memory. The final hidden layer's state vector is called a context vector or a "think" vector, which should be a semantic-related summary of the input sequence.

If x represents the length of the input sequence, the encoding function is represented as f (representing a gate), then a hidden layer state can be calculated as follows:

$$h_i = f(h_{i-1}, x_i) \tag{4}$$

Here will h_0 be set to a fixed value (for example, 0 vector), and the decoding part will be the initial state. Then this research can compute the functions g and k:

$$\mathrm{p}(y_i | y_1, \ldots, y_{i-1}, x) = \mathrm{g}(y_{i-1}, s_i) \tag{5}$$

Here $s_i = k(s_{i-1}, y_{i-1})$, y represents the output sequence. Seq2seq model can adapt to different sequence structure, and the current more popular attention mechanism can also be applied to them. After adding the attention distribution mechanism, Decoder can generate the new Target Sequence to get the hidden information vector Hidden State of each character before the Encoder coding phase, so that the accuracy of generating the new sequence is improved.

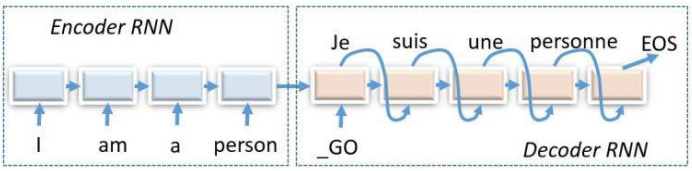

Fig. 2. An english to french translator with encoder and decoder Seq2Seq expansion, showing the flow of information (Source: https://esciencegroup.com/2016/03/04/fun-with recurrent-neural-nets-one-more-dive-into-cntk-and-tensorflow/)

2.2 Attention Mechanism

One problem that Seq2seq might exist is that it may not be able to effectively summarize the entire input sequence into a separate vector (this vector will be treated as a decoder input). In particular, the long term dependency may be more difficult to model, the solution is that refers to the "attention" mechanism. The basic idea is simple: to replace the sentence to build a vector, this research will keep the entire input sequence, and then in the recursive neural network every step to produce an output. What researchers are trying to do is to focus on the information that is useful to us in the input sequence at every time step, and this depends on the output of all the previous recursive neural networks.

"Attention" is a mechanism for machine translation proposed by Bahdanau for 2014. The "attention" mechanism does not directly use the input sequence, but directly uses the state vector, and then the state vector is combined with the weight produced by the previous step. This step, only useful information will be retained.

The state vector may be expressed exactly as (representing the length of the input sequence), which may be created from the input sequence using a bi-directional recurrent neural network such as the one mentioned above. Through these state vectors, the "attention" mechanism is used to construct the context vector at each time step of the output neural network. Enter a weight for the state vector and are:

$$c_i = \sum_{i=1}^{T_{in}} \alpha_{i,t} h_i \tag{6}$$

Then, the recursive neural network output layer uses this state vector. If it is recursive neural network state vector, but the output, this research can function f and h

said as follows:

$$p(y_i|y_1,\ldots,y_{i-1},x) = g(y_{i-1},s_i,c_i) \tag{6}$$

Where

$$s_i = f(s_{i-1},y_{i-1},c_i) \tag{7}$$

As mentioned above, the context vector is the weight of the input state vector that relies on all the forward output (by outputting the state vector). Thus, s (representing the "attention" vector) can be calculated by finding the probability of each state vector at each time step. Thanks to a function called a alignment model a, which is essentially a forward neural network:

$$e_{i,t} = a(s_{t-1},h_i) \tag{8}$$

Thanks to the softmax function, researchers can transfer $e_{i,t}$ to a probability αs,

$$\alpha_{i,t} = \frac{exp(e_{i,t})}{\sum_j exp(e_{j,t})} \tag{9}$$

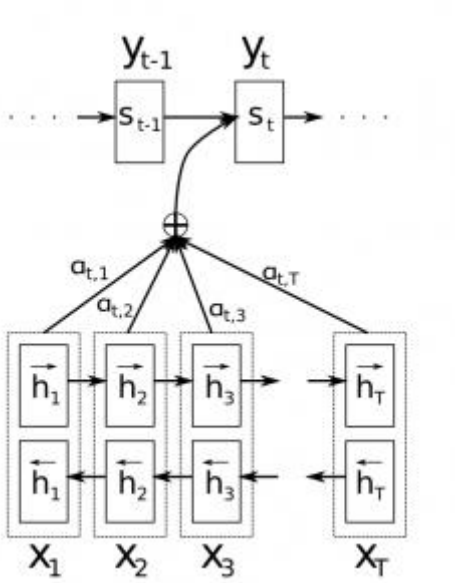

Fig. 3. Input sequence is tackled by bidirectional recurrent neural network, and 'attention' mechanism (Source: http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp/)

3 Rresearch Methodology of Semantic Mining on Agricultural Major Corps

NLP used the seq2seq and attention mechanism to solve the problem is in the field of sequence generating, which has been swept through a number of other methods.

The seq2seq model is generally structured as follows:

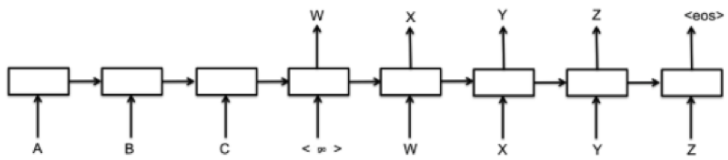

Fig. 4. Encoding-Decoding model

The encoder section encodes the input with a single layer or multiple layers of RNN/LSTM/GRU, and the decoder section is a language model used to generate the digest. This kind of generative problem can be summed up as a conditional probability problem P (word | context). Under the context condition, the probability of each word in the vocabulary is calculated, and the word with the highest probability is used word, and then generate all the words in the summary. The key here is how to express context, the biggest difference between each model is the context of the difference, where the context may either be the expression of the encoder, or be the expression of attention and encoder. The decoder part usually uses the beam search algorithm to do the build.

The research methodology in this paper is based on data-intensive machine learning techniques. Based on the deep learning theory mentioned in Section 2, the Seq2Seq and Attention mechanism is combined with specific methods to perform the agricultural science abstracts data processed in Section 4. Then use Objective function cross-entropy loss training and gradient descent algorithm to update parameters, adapt neural network parameters of the model to agricultural abstract corpus collection, and finally perform qualitative and quantitative model evaluation. The specific process is as follows:

1) Initially setting the model hyper parameter values, according to the Seq2Seq neural network model structure, completing the scripting, then defining the loss function, and selecting the optimization direction and optimization method;
2) Model initialization, model training and model parameter adjustment to achieve model benchmarks;
3) Model assessment, based on the qualitative assessment of the project carried out by the research team, and the next section will introduce the relevant assessment criteria for the implementation of qualitative assessment and ISI's ROUGE indicator system for quantitative assessment.

4 Model Experiment

4.1 The Corpus of Agricultural Science and Technology Information and Its Processing

4.1.1 The Dataset Composition

Based on the principle of research and implementation of cutting-edge technologies that have covered recent years, we selected English scientific and technological digest information on crop cultivation, molecular breeding, phenotypeomics, and new variety breeding harvested by the laboratory, which in the form of title, abstract, keywords, authors and institutions, and a total of 54,659 articles we captured. Where the article is originally marked with XML format, each article starts and ends with the <paper> ... </ paper> tag, and the document number begins with the <doc_id> </ doc_id> tag. The same article number is <paper_id> The journal number is <journal_id>, the title is <title>, the abstract is <abstract>, the keyword is <keyword>, the classification number is <classification>, the author is <authorlist> and so on.

Extract the original data set into the formatted "Title # Summary # Keyword / Keyword / Keyword ... Author / Author / Author ... # Organization / Organization ...", where the ellipsis represents there may be a number, and there will be no The item is marked as None; for example, an organization does not exist: "Title # Summary # Keywords / Keywords ... Author / Author ... # None". Which is the size of the abstract because of the different content of the length of the abstract, the title is also due to the lack of effective enough manpower and field experts (including agricultural rice field related to the technical vocabulary, to the corresponding human experts to mark the data is unlikely), So we chose to keep the data with a relatively objective summary of

the article and the title to generate the training corpus and discard the remaining ones so that the data set with the title plus the abstract can be generated, taking into account the same as in the future Look at the application, I used the abstract, title and keyword to generate the vocabulary Vocab.

The title of the article as the target, the first sentence of the abstract and the second sentence as the source, preprocessing, including: lower case, word, extract punctuation from the word, the end of the title and the end of the text will add a custom end tag <eos>, that no title or no content or title of the content less than 10 tokens or text over 50 tokens will be filtered out, the occurrence frequency sorted token, take the top 200,000 tokens as a dictionary, a low-frequency word symbol <unk> replaced.

4.1.2 Python's Natural Language Approach

Here the introduction of the relevant scientific data processing toolkit, needing to import the package numpy and collections, to use Stanford Natural Language Processing Toolkit to deal with the relevant statement information.

Because this research use the neural network model to carry out unregulated data processing, so here to simplify the relevant natural language processing methods, related to the pronunciation of the annotation, block, naming body recognition, relationship extraction and analysis of sentence structure and grammar and other omissions, only involved To the data segmentation and unrecognized vocabulary labeling and other simple natural language processing work, the remaining work to the model training and tuning skills. The data is placed in the path of the relevant file, and then converted using the encapsulated method, the data is abstracted as tensor tensor can recognize the tensor, for example, the data stored in the previous step is converted to the corresponding textsum model. Need to read the binary file format, for example, using the command ‘ `python textsum_data_convert.py --command text_to_binary --in_directories thesisdata / --out_files shuidao-train.bin, shuidao-validation.bin, shuidao-test.bin --split 0.8, 0.15,0.05` ’. Save the previous data set in proportion to the ratio of 8: 1.5: 0.5 to save the training binary file, verify the file and test set files. Here is the script to convert the binary file ANSI into an easy-to-read UTF-8 format, using the script data_convert_example.py conversion.

Among them, the model needs to read the sequence is marked as a sign of abstract, the value of the "=" behind the string, the output value is the value of the title tag, which contains the sequence <d> <p> <s> </ s> </ p> </ d> mark the article, paragraph, sentence. Where the labels are separated by tabs and abstract. Likewise, generating a validation set and a test set from the dataset. In this way, the data set can be used to complete the automatic generation of abstracts, title generation or article classification and so on.

4.2 Research progress of data mining semantic mining model

Research model code is built on the open source code of the Tensorflow project. This research changes session part of training code, part of the model in the `textsum/seq2seq_attention.py`, and make support for multi-core CPU by running `tf.ConfigProto`. In the model script running, Google's open source code is also vigorously pursuing its bazel run, this research makes one of the bazel running environments changed, to supported lab CPUs to run the thread unlocking protocol, so the terminal command changes from the original ‘ `$ bazel build -c opt --config = cuda textsum / ...` ’ to ‘ `$ bazel build -c opt --copt = -mavx --copt = -mavx2 --copt = -mfma -copt = -mfpmath = both --copt = -msse4.2 -copt = -march = native -config = opt textsum / ...` ’ to support more acceleration protocols AVX, SSE4.2 and so on. Finally, in order to show the data flow graph of our model in tensorboard dashboard, we added `sess.graph` to the original

`tf.train.FileWriter` builder, and of course researchers can use the `add_graph` method of `tf.train.FileWriter` builder to build the project data flow chart.

Here follows the model configuration：

Table 1. Model parameter configuration.

Function	Purpose	Setting
Mode	train、test、decode	train
Working_directory	Vocabulary files and training, verification and test file storage addresses	data/
Pretrained_model	--	--
Batch_size	Data bundle size for training	64
Bidirection Encoding layer	The number of bidirectional RNN layers in the encoding section	3
Artilcle length	The maximum acceptable article input sequence length	120
Summary length	Generated sequence maximum length	30
Word embedding size	Word vector dimension size	128
LSTM hiden units	Number of LSTM hidden units	192
Sample softmax	Final normalization function using units	4096
Vocabulary size	Vocabulary for generating, taking the word frequency top 200k	200k

4.3 Study on Experimental Results of Model

Model of the original training step is 10,000,000 steps, but if the amount of data is not large enough, many steps are likely to cause too much training. The sample volume of Google original data is about 10,000,000, which is marked Gigaword English news dataset (also can used DUC2004 dataset to test), in each server is 4 core K series graphics card and a total of 10 servers, running a week, in view of the inability to achieve its experimental environment, this research will apply the moderate quantity of standard experimental agricultural major crops digest corpus to the changed models, and the training steps temporarily transferred to the 130500 steps, the effect shown in figure.

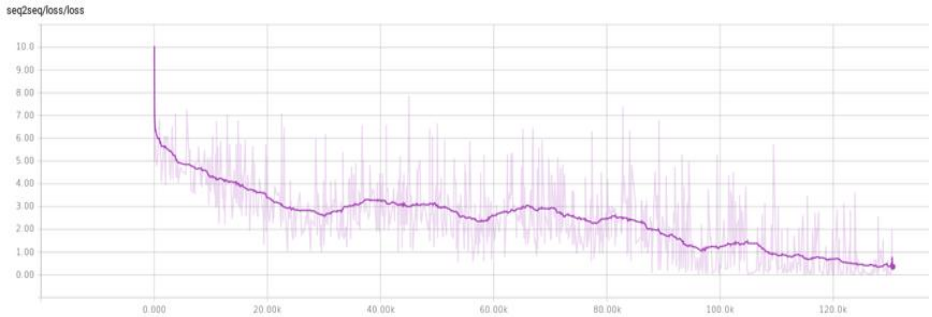

Fig. 5. Textsum model loss value trend on agricultural science and technology dataset. Smooth coefficient 0.88

This is agricultural corps digest data training renderings, after 9 days more than total recursive training about 130,500 steps, the loss value reached a standard level with the standard data set (below 1.0).

The following example illustrates the textum automatic digest model to generate the title sequence effect.

Article: To study whether the biomass of soil microorganisms in a boreal Pinus

sylvestris-Vaccinium vitis-idaea forest was limited by the availability of carbon or nitrogen, we applied sucrose from sugar cane, a C-4 plant, to the organic mor-layer of the C-3-C dominated soil. We can distinguish between microbial mineralization of the added sucrose and respiration of endogenous carbon (root and microbial) by using the C-4-sucrose as a tracer, exploiting the difference in natural abundance of C-13 between the added C-4-sucrose (delta(13)C -10.8parts per thousand) and the endogenous C-3-carbon (delta(13)C -26.6 parts per thousand). In addition to sucrose, NH4Cl (340 kg N ha(-1)) was added factorially to the mor-layer. We followed the microbial activity for nine days after the treatments, by in situ sampling of CO2 evolved from the soil and mass spectrometric analyses of delta(13)C in the CO2. We found that microbial biomass was limited by the availability of carbon, rather than nitrogen availability, since there was a 50% increase in soil respiration in situ between 1 h and 5 days after adding the sucrose. However, no further increase was observed unless nitrogen was also added. Analyses of the delta(13)C ratios of the evolved CO2 showed that increases in respiration observed between 1 h and 9 days after the additions could be accounted for by an increase in mineralization of the added C-4-C. (Data source: Ekblad et al., 2002)

Ref: Is growth of soil microorganisms in boreal forests limited by carbon or nitrogen availability? (Data source: Ekblad et al., 2002)

Decode: the effects of organic manure nutrient on the leaching by acid movement of soil.

This is the result of the abstractive automatic summary model decoder part training from the agricultural main corps corpus. The input sequence is article contents, the target sequence, title contents, is marked with Ref, and the output sequence is the decode flag.

Table 2. Textual automatic digest model ROUGE value contrast.

Model	ROUGE-1	ROUGE-2
ABS (Rush et al., 2015)	30.88	12.22
ABS+	31.00	12.65
Char Level (Golub et al., 2016)	11.31	2.65
COMPRESS (Clarke et al., 2008)	19.63	5.13
TextRank (Mihalcea et al., 2014)	31.10	9.03
Textsum	36.99(up to **38.27**)	19.68(up to **20.58**)

Rouge (Recall-Oriented Understudy for Gisting Evaluation), the basic idea is to use the abstract generated by the model and the n-tuple contribution statistic of the reference abstract as the basis for judging.

It can be seen that the higher the ROUGE value, the better the effect of generating the abstract on behalf of the model, and conclusion from the table is that the textum model is superior to the current majority of text automatic summary models.

5 Related Research Progress

In academia, deep learning begins to receive attention from 2006, because in the tens of thousands of samples of medium-scale datasets, the deep learning learnt from the new sample at that time more than many popular algorithms and generalized better. Soon after, the deep learning in the industry has received more attention, because it provides a scalable way to train large data sets on the nonlinear model. LeCun, Bengio & Hinton, leaders in deep learning, are looking forward to the future development of deep learning in journal Nature (Lecun et al., 2015): first, although the recent unsupervised learning was restrained and supervised learned to snatch the limelight, but in the long run, unsupervised learning is still a more important issue, and Michael et al. also hold this idea (Michael et al., 2015). Second natural language processing will be the depth of learning in the future to achieve significant breakthroughs in the field, to better "understand" statements and text semantic system will appear. Third, combination of deep learning and symbolic artificial intelligence

will bring revolutionary changes in the field of artificial intelligence.

Direct extraction of important sentences of the extraction method is relatively simple, such as PAM algorithm realizing TextRank, while generating (re-generating a new sentence) is more complex, and the effect is not satisfactory. At present, the more popular Seq2Seq model, proposed by Sutskever et al., based on the structure of an Encoder-Decoder, the source sentence first is encoded into a vector of fixed dimension D, and then generate the target sentence one character at a time using the decoder part. After adding the 'Attention' mechanism, decoder can generate the new target Sequence to get the hidden information vector Hidden State of each character before the Encoder coding phase, so that the accuracy of generating the new sequence is improved. The increase in end-to-end training in neural networks has led to significant changes in many areas, including speech recognition, computer vision, and natural language processing. Recent work has shown that neural networks can do more than any sort of task, and they can be used to map complex structures to another complex structure. For example, a sequence is directly translated into another sequence, which has a direct application in natural language understanding (Luong et al., 2014).

6 Conclusion

The main advantage of this model is that when matches or transcends the current optimal benchmark, it requires little feature processing and specific domain-specific knowledge. This advantage, in my opinion, is to allow researchers to focus on tasks in certain unknown areas or in areas where it is too difficult to design artificial rules (e.g. ontology, etc.). This research will be devoted to the development of the automatic digest extraction system of semantic mining of agricultural science and technology literatures, and make a modest contribution to the development of semantic excavation of this field.

References

1. Hochreiter S, Schmidhuber J.: Long short-term memory. Neural computation. 9(8), 1735-1780 (1997)
2. Cho K, Merrienboer B Van, Gulcehre C, et al.: Learning Phrase Representations using RNN Encoder--Decoder for Statistical Machine Translation. empirical methods in natural language processing. 2, 1724-1734 (2014)
3. Rush A M, Chopra S, Weston J, et al.: Neural Attention Model for Abstractive Sentence Summarization. empirical methods in natural language processing. 2, 379-389 (2015)
4. Golub D, He X: Character-level question answering with attention. arXiv preprint arXiv:1604.00727 (2016)
5. Clarke J, Lapata M. Global inference for sentence compression: An integer linear programming approach. Journal of Artificial Intelligence Research. 31, 399-429 (2008)
6. Mihalcea R, Tarau P. TextRank: Bringing Order into Texts[J]. Emnlp. 6, 404-411 (2004)
7. Luong M T, Sutskever I, Le Q V, et al. Addressing the Rare Word Problem in Neural Machine Translation[J]. Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Veterinary Medicine. 27(2), 82-86 (2014)