International Journal of Database Theory and Application
Vol.9, No.5 (2016), pp.187-196
http://dx.doi.org/10.14257/ijdta.2016.9.5.19

Extracting Attributes of Named Entity from Unstructured Text with Deep Belief Network

Bei Zhong, Jin Liu, Yuanda Du, Yunlu Liaozheng and Jiachen Pu

College of Information Engineering, Shanghai Maritime University, 201306 Shanghai, China
{beizhong, jinliu, yddu, ylliaozheng, jcpu}@shmtu.edu.cn

Abstract

Entity attribute extraction is a challenging research topic with broad application prospects. Many researchers had proposed rule based or statistic based approaches to deal with the extraction task in a variety of application areas. Recently, deep learning had shown its capacity to model high-level abstractions in data by using multiple processing layers network with complex structures. However there has no research reported to conduct entity attribute extraction with deep learning method. In this paper, we propose a new approach to extract the entities' attributes from unstructured text corpus that was gathered from Web. The proposed method is an unsupervised machine learning method that extracts the entity attributes utilizing deep belief network (DBN). Experiment results show that, with our method, entity attributes can be extracted accurately and manual intervention can be reduced when compared with tradition methods.

Keywords*: information extraction, entity attribute extraction, DBN*

1. Introduction

The world has entered the era of big data. How to deal with massive text effectively and efficiently has become an urgent problem in front of us. Entity Attribute Extraction is an important technology in the field of natural language processing, and the information obtained can be not only provided to the users directly, but also used as the basis of building intelligent query and data mining. As an important aspect of information extraction, entity attribute can be used to define a new entity, conduct entity mining and other practical applications.

Natural language processing (NLP) technology first began in the meeting of MUC (message understanding conference) funded by the United States Department of Defense Advanced Research Projects Committee. The main purpose of the study on information extraction is to get the structured information from natural language text. With the development of technology, new NLP processing method emerges continuously. In order to utilize computer program to automatically define the characteristics of the numerous new concepts or things that is shown on the Web, and to explain the things through the features, entity attribute extraction which is to obtain the characteristic of an entity becomes a very important NLP technology. The task of entity attribute extraction is let computer fetch the attributes and their values by itself.

Entity refers to the independent existence of things. Each entity has its own characteristics, that is, different entities have their own specific attributes, and can be distinguished from other entities. The name of Entity often represents species, which have the same nature as other nouns. Generally, people specify a name for each entity, which is also known as Named Entity, NE. The entities with same category generally have similar attributes, but they are different in the value of the property. Different types of entities generally have different properties. Any concrete or abstract object can be called an entity. Different from other applications of information extraction, a user`s interest can

ISSN: 2005-4270 IJDTA

also be defined as an entity, such as people, institutions, products, *etc*. Moreover, things that appear in the corpus, can all be defined as an entity. Entities with different types have different attributes and information characteristics. While entities with same category generally have roughly the same attribute information structure, but the value of each attribute will be different. For example, there are general attributes of a people entity: full name, occupation, work units, mail, telephone, hobbies, *etc*.; the attributes of organization/unit entity: the name of institution/unit, address, department, responsible person, the nature of services, *etc*. And the typical attributes of a product entities: product name, manufacturer, product function, art, price, brand, characteristics, and so on [1].

In recent years, with the rapid development of search engine technology, searching has become more and more intelligent. The search engine has evolved from the "keywords search" to "SNS Search" and "Entity search".

Entity Search is more complicated than the keywords Search. Although the traditional keyword search has developed well, the results provided by the search engine can help users find the information, but in fact for the "Search Engine" system itself, it does not understand the meaning of the search. The primary focus on Entity search is not the "key words" but the object, such as people, institutions, organizations, *etc*. We hope a conversion from keywords to entity can help search engine understand and organize search results from a more subtle point of view. To a certain extent where can understand the meaning of query, and give their own answers, some of the more intelligent and personalized interaction is also depending on the entity which is the basis for upper layer applications.

Entity Search needs an entity-related information database. The information database not only includes massive entity information but also the relevant attributes which can accurately describe the entity. The construction of the Entity Database needs long-term accumulation and the relevant data mining technology.

This paper provides a novel entity attribute extraction method to conduct the entities and syntax analysis [2] , and the extraction of entity attribute with deep learning method. This method can be used in the task of entity attributes. The remaining of this paper is organized as follows. Section 2 presents the related work in the literature, Section 3 explains our proposed method and the last section presents the experiment results.

2. Related Work

Entity attribute extraction is a very important task in information extraction. While most research focus on attribute extraction of English text. Chinese entity attribute extraction is also an important task in information extraction. In our work, we focus on the entity attribute extraction from Chinese text.

2.1. Ways of Extract Attribution

Ding [3] pointed out the basic concept of attribute and the role of attribute, and the ways of extracting attribution: People attribute extraction, product attribute extraction, concept attribute extraction and enterprise attribute extraction.

2.1.1 People Attribute Extraction

Character attribute extraction is the extraction of the basic information of the characters, such as the birthday of the people, the place of birth and work, *etc*. Ye Zheng from Dalian University of Technology's extracted person's gender, position and other attributes information from the free text of the HowNet. He regarded the words describing the characters' attributes as entity, and treated Character attribute extraction, which can use to build character information database as the concrete application of the entity relation extraction [3]. Lu Wei, and some other scholars, who come from Wuhan

University, used the extraction of character information to construct the company's expert search system to facilitate the search for expert expertise. In addition, the extraction of character attributes can also be used for the development of character search engine, and also Meta Search Engine.

2.1.2 Product Attribute Extraction

With the development of electronic commerce, the Internet has accumulated hundreds of millions of product information. Product attribute extraction can not only increase the description of the goods, but also provide very valuable information for users and manufacturers such as the description of product, price and other information. Furthermore, product attribute extraction makes Suppliers to increase the retailer's merchandise database, and makes search engines to build a shopping search engine.

2.1.3 Concept Attribute Extraction

The concept attribute extraction includes the extraction of the popular concept and the academic concept. In the literature, Li Jing used domain properties in the concept of ontology to build domain ontology and made a research on the domain properties in the concept of domain ontology.

2.1.4 Enterprise Attribute Extraction

Attribute extraction from the enterprise level, can be used to extract the basic enterprise information. It constructs business directory database and timely tracks and extracts relevant information in the external environment of the enterprise, which meets the needs of corporate in term of public opinion monitoring. It is also an important part of enterprise management.

2.2. Methods of Extract Attribute

Attribute extraction methods can be divided into two categories: rule-based and based on statistics methods.

Rule-based approach is a method which contains pattern matching and focuses on areas of analysis and pattern matching. These methods often rely on specialized areas of background knowledge. When extracting character attributes, we need to match all possible patterns that describe the properties of the characters. Although it can also have a better extraction results, this model has many disadvantages: there are too many matching rules, not easy to sum up the integrity, and the large workload. When encountering new properties, the well-formed rules cannot handle them directly. Thus, there are some bottlenecks in the rule-based extraction: very complex rules, larger workload of the design process of rules and easily to make mistakes, difficult to cover all the linguistic phenomena, and requirements of domain experts' assistance to complete the work. Besides, the compatibility, flexibility and portability of the rule System is also poor. With this method it is difficult to cover with all language phenomenon, and experts in the field need to be involved. At the same time, it lacks compatibility, flexibility and portability.

Statistical methods are based on training corpus that is manually labeled to train a model. One of the most common methods is the semi-supervision, which requires a small amount of labeling as a seed, and then uses statistical knowledge to discover new candidate words. Typical methods of information extraction are conditional random method, hidden Markov model and others.

Most research in the literature is focused on extracting the English product attributes and character attributes. Santosh Raju and Katharina Probst proposed the method of extracting attributes and attribute-value pairs from the product description [4-5]. The method of product attributes extraction was unsupervised and semi-supervised. Rayid

Ghani and Wong T L proposed an unsupervised framework to extract attributes [6-7]. István N T [8] proposed a method that person attribute extraction from the textual parts of Web. Chun liang L [9] proposed an extraction method of attribute word in Chinese product reviews. Zhang Q [10] proposed a homepage character attribute extraction method that based on weak supervised learning for the characters of the characters. And Li Hong Liang [11] put forward an Encyclopedia character attribute extraction method based on the rules.

Sánchez D [12] proposed a new methodology for acquiring class attributes at an ontological level. In addition to object relations, one of the papers contributions is to address the discovery of data-properties and their associated data-types and value ranges. Results have been manually checked by means of an expert-based concept-per-concept evaluation for several well distinguished domains. And this method had been showing reliable results and a reasonable learning performance in the experiments.

Jia [13] proposed an approach to extract attribute and attribute value from Chinese wiki encyclopedia entry articles. Attribute values are viewed as named entities and class attributes are extracted based on frequent patterns mining and association analysis. The experiment results show that the method is feasible and effective.

Recently, deep learning is a frequently mentioned method. But most researchers are using it in the image processing. In the field of natural language processing, the application of deep learning just began recently. Collobert published a paper in 2011. This paper described the application of deep learning in the field of NLP. He proposed a model which had changed the traditional thinking of text categorization. Biggest advantage of the model was that no manual designed features in involved, and it just need the original text input vector to automatically extract features. Chen Yu applied deep learning to the Chinese entity relation extraction and Chinese named entity category discovery, and the proposed method achieved good results [14-15]. Relation extraction is a fundamental task in information extraction, which is to identify the semantic relationships between two entities in the text. The results at paper [14] show that DBN is a successful approach in the high-dimensional-feature-space information extraction. It outperforms state-of-the art learning models such as SVM and back-propagation network. In next section of this article, we present an entity extraction method based on deep belief network.

3. Entity Attribute Extraction Based on Deep Belief Network (EAEDB)

3.1. Deep Belief Network

Conditional random fields (CRFs) [16] are a classification model for tagging task. In this paper, it is also used to recognize named entities. CRFs are an undirected graphical model (also known as random field) which is used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes. Different from CRF, Deep Belief Networks (DBNs) are graphical models which learn to extract a deep hierarchical representation of the training data.

Deep Belief Network (DBN) [18] was proposed in 2006 by Geoffrey Hinton. By training the weights between neurons we can make the whole neural network generate the training data according to the maximum probability. In EAEDB, the structure of our DBN network is a kind of deep neural network that composed of several layers of Restricted Boltzmann Machines (RBM) and a layer of BP.

3.1.1. BP Network

Rumelhart, McClelland put forward the backward propagation of errors algorithm (error back pass learning algorithm) of BP network in 1985. As one of the neural network models, BP model, which is a kind of multilayer feed forward network trained by the error back propagation algorithm, has been used widely in not only academic but also

industrial applications. BP network can learn and store vast amount of mapping relation from input to output model without revealing the mathematical equations describing the mapping relationship in advance. Its learning rule is to use the fastest descent method, and constantly adjust the network weights and threshold by back propagation, so that the square-error of network is minimum. The topological structures of BP neural network model include input layer, hidden layer and output layer.

BP neural network is divided into two processes:

(1) Forward transfer sub-process.
(2) Reverse transfer of error signal sub-process.

Calculation Process:

(1) Calculating the input and output neurons in each layer.
(2) Using the desired output and the actual output to calculate partial derivative of error function for each input neurons.
(3) Using the connection weights between hidden layer and output layer, $\delta_o(k)$ and the output of hidden layer to calculate partial derivative $\delta_o(k)$ of error function for each hidden layer neurons.
(4) Using $\delta_o(k)$ of output layer neurons and output of hidden layer neurons to fix the connection weights $w_{oh}(k)$.

This "neuron" is a computational unit that takes as input x_1, x_2, x_3 (and a +1 intercept term). We will choose sigmoid function as the activation function.

Here are plots of the sigmoid and tanh functions:

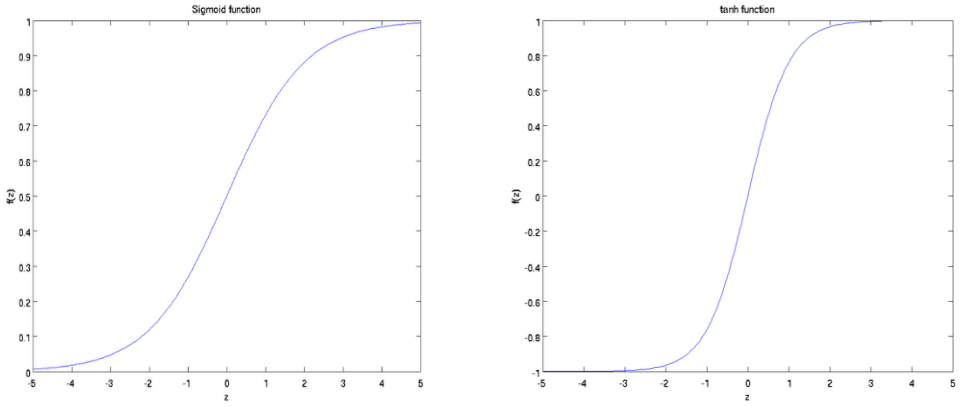

Figure 1. Activation Function

The tanh(z) function is a rescaled version of the sigmoid, and its output range is [−1,1]instead of [0-1].

3.1.2 RBM Network

RBM has only two layers of neurons, one is called visible layer which is composed of visible units and used for training data, the other is called hidden layer which is composed of hidden units and used as a feature detector. There is no neurons interconnection between the internal of visible layer and hidden layer but only symmetrical connecting line. The benefit is that the values of the hidden unit is unrelated in the case that the visible units' value has been given. The formula of given visible units value is as follows:

$$p(h|x) = \prod_{j=1}^{N} P(h_j|V) \quad (1)$$

In the same way, The formula of given hidden layer is:

$$p(v|h) = \prod_{I=1}^{M} P(V_I|V) \quad (2)$$

The h in the formula represents visible layer nodes while the v representing the hidden layer nodes.

Through the above formula, it is unnecessary to calculate only one value each time when calculate the value of each neuron, but compute paralleledly the whole layer neurons at the same time.

When getting the new data and clamp the data to visible layer, RBM will open or close the hidden units according to weights W. First, it will calculate activation value of each hidden layer unit (h=Wx). Then, using sigmoidal function to do standardization, and calculate the visible layer when the value of hidden layer has been given is the same.

$$P(h_j = 1) = \frac{1}{1+e^{-h_j}} \text{ (Open)} \quad (3)$$

$$P(h_j = 0) = 1 - P(h_j = 1) \text{(Close)} \quad (4)$$

$$f(x) = \frac{1}{1+e^{-x}} \quad (5)$$

3.2. Feature Extraction

The application of EAEDB is composed of following three steps:

1. Recognize named entities by CRFs, and then form the entity feature which is defined by the category of the recognized entity.
2. Analyze the text with semantic parser, get the Object Structure characteristics of the syntax tree.
3. Combine three features that have been discovered. Feed the integration of the feature set as parameters into DBN neural network, and then get the model of entity attributes extraction.

Table 1.1. Recognize Named Entities

Word	Tag
Confidence	B-NP
in	B-PP
The	B-NP
pound	I-NP
…	…

Feature vectors composition is as follows:

1. Etype: Entity category information feature.
2. Pos: Position features affects the relationship of words. In this paper, we concentrate on extracting nouns, verbs, quantifiers, prepositions and other features.
3. Tag: It is the result of the syntax parser.

Table 1.2. Characteristics of Parameters and Threshold Range

Feature	Threshold range
Etype	Names,Palce,Organization
Pos	Noun,Verb,Adjective,Numeral…
Tag	SBV(subject-verb),VOB(verb-object),HED(head),IOB(indirect-object),FOB(fronting-object),COO(coordinate) …

3.3. Process

Step 1: Since the corpus is obtained from the Internet which contains much noise text. At first, we must process the corpus to get pure unstructured text.

Step 2: Extract feature that is used to train the model. Before training DBN model, we must extract the feature vectors from the corpus. After the entity recognition, we can obtain a list of features, and then, combining the tag of syntactic tree that result parser syntactic as the feature vectors.

Step 3: After the previous two steps, we can obtain the feature vectors. Finally, the matrix of feature vectors is used to train the model of DBN.

DBN model training process is divided into two steps which is shown in Figure 3.1. Firstly, we must train each layer RBM network respectively, ensuring that the feature vectors are mapped to different feature space, and then the features are preserved. The final layer is set in a BP network, receiving the output of RBM as its input feature vector. In this model, each layer can only ensure that the weight is optimal to own RBM network layer, not for the entire DBN. Therefore, error message from back-propagation network will be propagated down to each top-level RBM. Fine-tuning the entire RBM network of training process, the result model can be seen as initialization parameters of BP network.

4. Experiment

We conducted experiment to test the effectiveness of EAEDB. The settings and the results are presented in this section.

4.1. Data Set

We gathered a large set of documents with a topic specific crawler [20] in shipping news website and travel website. Original corpus was the web page set that contains a lot of useless information such as ads and text only with html tags. After pre-processing the target text corpus, the size of the corpus was reduced to about 24MB.

At last, we used more than 10,000 Chinese sentences as training set and all of the text were processed using the method presented in Section3.3.

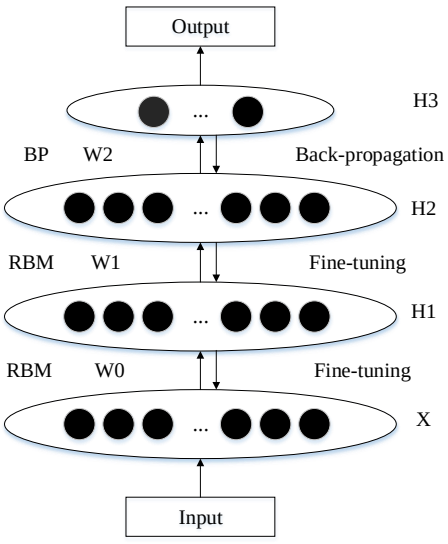

Figure 3.1. DBN

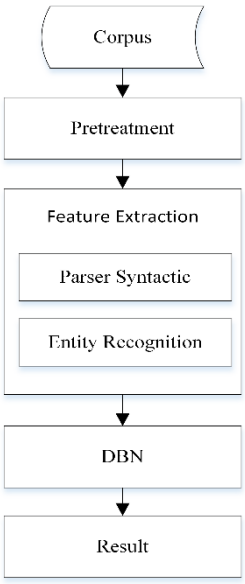

Figure 3.2. Flow Chart

4.2. Result and Analysis

With the different proportions, we randomly split the corpus into two sets. One is the development set and the other is the test set. We conducted experiments four times. Since the primary target of the experiments was to extract attributes, we analyzed the performance of results, and found that the method was feasible, though the preliminary results were not satisfactory. By comparing the results obtained from five experiments, it can be discovered that the feature dimension changes have great impact on the result, and more experiments would be needed to find what a suitable number is to get the best accuracy.

Table 4.1. Result of Entity Recognition

Word	Tag
中	B-LOC
国	I-LOC
首	N
都	N
是	N
北	B-LOC
京	B-LOC

Some of the tags that are used in the experiments are shown in Table 4.1, where the B-LOC represents the place names. We defined three types of entities that are place names, organization names, and personal names.

Entity cognition is expressed as classification tasks, hence metrics like Precision, Recall and F-Measure are used together for performance evaluation. They are defined as formula 6, formula 7 and formula 8 respectively.

$$\text{Precision } P = \frac{\text{Number of correctly extracted entity attributes}}{\text{Total number of extracted entity attributes}} \tag{6}$$

$$\text{Recall } R = \frac{\text{Number of correctly extracted entity attributes}}{\text{Actual number of extracted entity attributes}} \tag{7}$$

$$\text{F-Measure } F1 = \frac{2PR}{P+R} \tag{8}$$

Table 4.2. Result of Experiments

Category	Attributes of number
Port	8
Ship	10
Routes	3
View	5

Table 4.2 shows the experiments result that the entities attributes which are obtained from the testing corpus. The first column is category of attribute, and the second column is the number of attributes. Table 4.3 shows the precision, recall and F-measure results of the experiments. We conducted the experiments in four runs with different training and testing set ratio. The ratios are 9:1, 8:2, 7:3 and 6:4 respectively. And we can find with the 6:4 training and testing set ratio, the result's precision and F-measure are highest and there's no significant difference in terms of recall rate.

Table 4.3. Preliminary Results

Times	P/%	R/%	F/%
1	0.667	0.5	0.5714
2	0.600	0.5	0.5455
3	0.750	0.5	0.600
4	0.924	0.5	0.6538

5. Conclusions

In this paper, we presented a new unsupervised approach for attribute extraction from unstructured text. To start with, the famous CRFs model solve the Recognize named entities problem. And then, we extract attributes which can be discovered by using DBN model. In the future work, we will try this method with variable kinds of corpus. Also, we will add more text features and try the different ways to smooth the related RBM parameters to improve this method.

Acknowledgements

This work was supported by Shanghai Municipal Science &Technology Commission "Science and Technology Innovation Action Plan" project (14511107400), and by Shanghai Maritime University research fund project (No. 20130469), and by State Oceanic Administration China research fund project (201305026).

References

[1] J. Y. Guo, Z. Li and Z. T. Yu, "Extraction and relation prediction of domain ontology concept instance, attribute and attribute value", Journal of Nanjing University, **(2012)**, pp. 383-389.
[2] L. Yuan, "Statistical syntactic parsing methods", Journal of Central South University, **(2014)**.
[3] D. J. Jun, Z. Y. Ning and H. B. Lin, "Survey on Attribute Extraction at Home and Abroad", Information Science", **(2011)**.
[4] S. Raju, P. Pingali and V. Varma, "An Unsupervised Approach to Product Attribute Extraction", Advances in Information Retrieval. Springer Berlin Heidelberg, **(2009)**, pp. 796-800.
[5] K. Probst, R. Ghani and M. Krema, "Semi-Supervised Learning Of Attribute-Value Pairs From Product Descriptions", In IJCAI-07, **(2007)**, pp. 2838-2843.
[6] R. Ghani, K. Probst, Y. Liu, M. Krema and A. Fano, "Text mining for product attribute extraction", SIGKDD Explorations, **(2006)**, pp. 41-48.
[7] T. L. Wong, W. Lam and T. S. Wong, "An Unsupervised Framework for Extracting and Normalizing Product Attributes from Multiple Web Sites", SIGIR '08 Proceedings of the 31st annual international ACM SIGIR conference on Research and develop, **(2008)**, pp. 35-42.
[8] N. T. István and F. R. Person, "Attribute Extraction from the Textual Parts of Web Pages", In Third Web People Search Evaluation Forum (WePS-3), CLEF 2010, **(2010)**.
[9] L. I. C. Liang, Y. H. Zhu and X. U. Y. Qiang, "Research of Attribute Word Extraction Method in Chinese Product Comment. Computer Engineering", **(2011)**, pp. 26-25.
[10] Q. Zhang, J. Xiong and X. Cheng, "Person Attributes Extraction Based on A Weakly Supervised Learning Method", Journal of Shanxi University, **(2015)**.
[11] L. H. Liang, Y. Yan, Y. H. Feng and J. Zhen, "Rules-Based Character Attributes Extraction from Baidu Encyclopedia", Journal of Integration Technology, **(2013)**.
[12] D. Sánchez, "A methodology to learn ontological attributes from the Web", Data & Knowledge Engineering, **(2010)**, pp. 573-597.
[13] Z. Jia, Y. Yang and H. E. Dake, "Attribute and Attribute Value Extracted from Chinese Online Encyclopedia", Acta Scientiarum Naturalium Universitatis Pekinensis, **(2014)**.
[14] C. Yu, Z. D. Quan and T. J. Zhao, "Chinese Relation Extraction Based on Deep Belief Nets", Journal of Software, **(2012)**, pp. 2572-2585.
[15] C. Yu, Z. Dequan and Z. Tiejun, "Study on Chinese Named Entity Categorization based on Deep Belief Nets. Intelligent Computer and Applications", **(2014)**, pp. 29-31.
[16] C. Sutton and A. Mccallum, "An Introduction to Conditional Random Fields for Relational Learning", In Introduction to Statistical Relational Learning, **(2006)**, pp. 93 - 127.
[17] J. Lafferty, A. McCallum and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", **(2001)**.
[18] http://deeplearning.net/tutorial/DBN.html#dbn
[19] P. Zhang, W. J. Li, F. R. Wei, Q. Lu and Y. X. Hou, "Exploiting the role of position feature in Chinese relation extraction", In: Proceeding of the 6th Int'l Conf. on Language Resources and Evaluation (LREC). Marrakech, **(2008)**, pp. 28-30.
[20] Y. U. Juan and Q. Liu, "Survey on topic-focused crawlers", Computer Engineering & Science, **(2015)**.

Authors

Bei Zhong, (1989-), male, born in Guangxi, master candidate, specialized in natural language processing research.

Jin Liu, (1975-), male, born in Sichuan, associate professor, specialized in web data mining, NLP and software engineering.

Yuanda Du, (1994-), male, born in Tianjin, bachelor candidate, specialized in natural language processing and machine learning research.

Yunlu Liaozheng, (1994-), male, born in Tianjin, bachelor candidate, specialized in natural language processing and machine learning research.

Jiachen Pu, (1994-), male, born in Tianjin, bachelor candidate, specialized in vision and language processing and machine learning research.

