Automatic Classification of Japanese Question Intention Based on Deep Learning

Lan Jin¹, Yancong Su², Jianxun Li²,*

¹Chinese Language and Culture College, Huaqiao University, No.8, Jiageng Road, Xiamen, 361021, Fujian, China
²School of Design Art, Xiamen University of Technology, No.600, Ligong Road, Xiamen, 361024, Fujian, China

* jianxun_li@xmut.edu.cn

Abstract. In order to improve the effect of deep learning, this paper puts forward the method of deep learning of Japanese. Before classification, we need to preprocess the word segmentation and delete the termination word in the question, and then use the feature vector to represent it. By learning to store the information in the network, we use heuristic rules to classify the question intention, extract the feature vectors representing different types of questions, and then make statistical analysis on the corpus of actual marked questions, establish a classification system, and bind features based on word packets, Realization intention classification. The experimental results show that the classification accuracy and influence parameters are relatively higher after deep learning, and the automatic classification method of Japanese question intention should be unique.

Keywords: Deep learning; Japanese; Question intention; Automatic classification.

1. Introduction

Language is a social phenomenon. It is one of the important means for people to exchange ideas, transmit information and communicate. The purpose of verbal communication is to understand other people's ideas and express their own ideas. Understanding other people's ideas is analytical, and expressing your own ideas is comprehensive. Through the automatic classification of Japanese question intention, we can understand the various pragmatic factors that affect the use of Japanese interrogative sentences and the internal factors that affect their pragmatic functions. It is helpful to reveal the nature of Japanese interrogative sentences, thus enlightening us to analyse Japanese interrogative sentences and their functions, and apply this new and systematic viewpoint to the study of other sentence patterns. In the process of automatic classification of Japanese question intention, question classification is the primary task. The process is described as: questions are expressed in the form of data features, and the output of classification is used as the classification mark. This paper mainly studies the deep learning classification method and its improvement, and applies it to the question classification of Chinese question answering system. Feature extraction and classifier design are the focus and difficulty of this paper. Understanding and mastering interrogative sentences in Japanese is conducive to improving the communicative competence. The study also provides some perspectives and methodological references for the pragmatic function analysis of other sentences.
2. Automatic classification of Japanese question intention

2.1 Research on the category characteristics of Japanese questions

Japanese interrogative sentences include interrogative sentences and interrogative sentences, which are inseparable. The question is the speaker's own question, and the question is to ask the other person's speech act. Questioning refers to that one party expresses doubts about the other party or raises questions to the other party, including self-questioning and self-answering[3]. The formal meanings of interrogative sentences are divided into the following categories: (1) the formal meanings of general interrogative sentences; (2) the formal meanings of general interrogative sentences. For example: surprise, sadness, regret, joy and so on. (3) Affirmation and irony. First of all, we should meet the following requirements: (1) there is a listener; (2) the speaker asks questions to the hearer. The audience can answer questions, and the speaker can get the information from the audience. In other words, only when questions and answers exist can they be regarded as interrogative sentences. Further, Japanese interrogative sentences can be divided into four forms: affirmative, judgmental, selective and explanatory. Like other grammatical problems, Japanese interrogative sentences have also attracted the attention of researchers. Japanese interrogative sentence is one of the modern expressions. Its function is mainly determined by the end of the sentence, interrogative words or interrogative intonation. It has three functions: expression, answer and non-expression[4]. Other functions are expressed in the form of interrogative sentences, such as declarative and praying functions. However, no matter what kind of interrogative sentence is, its questioning function is the most typical. In Japanese interrogative sentences, pragmatic factors mainly include contextual knowledge, encyclopaedia information, cognitive rules, communicative attitude, communicative intention, etc. they constitute grammatical categories, and the change of pragmatic factors will affect the choice of grammatical category structure. Questions, expectations, sincerity, politeness and other pragmatic factors also directly affect the use of Japanese interrogative sentences[5]. It is not only the realistic basis for the construction of grammatical category of Japanese interrogative sentences, but also an important reason for the structural changes of interrogative sentences. These pragmatic elements are analysed as follows:

| Japanese interrogative sentences expressing declarative functions | Japanese interrogative sentences expressing imperative function | Japanese interrogative sentences expressing exclamatory functions | Japanese interrogative sentences expressing greetings |
|---------------------------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------------|-----------------------------------------------------|
| Means to provide information                                  | A polite, polite request                                     | Express surprise                                            | Abstract, courtesy of social entertainment, no clear purpose, the answer is more general |
| Means explanation                                              | Express advice or advice                                     | Express joy                                                 | Communicate feelings and friendship                   |
| Express judgment or opinion (It means to state one's own judgment or opinion in a euphemistic way; to refute the other party's) | advice or advice without speaker involvement               | To express exclamation and admiration                        | There is no inclination to the information inquired  |
opinion or opinion tactfully, so as to seek
the change of opinion)

| The function of expressing negation and refutation | Signifying an order and the difference from indicating a request, suggestion, or advice | Express hope and expectation | Subjectivity, positive expression |
|---------------------------------------------------|-------------------------------------------------------------------------------------|-----------------------------|---------------------------------|
| It expresses the function of striving for understanding and calling for sympathy Speaking to oneself | Express urge | Express regret | Honesty, politeness |
| Expressing doubt | Indicates prohibition | Express worry, uneasiness, apprehension | Be dubious about the information inquired |
| To express a rhetorical question | Denounce, warn | To express ridicule or satire | The questioner has no doubt about the information inquired, or the degree of doubt is very low |
| | - | Express anger or dissatisfaction; express sadness | - |

Tab.1 Features of Japanese question categories

From the perspective of grammar, the interrogative function of Japanese interrogative sentences gradually weakens, and eventually transforms into other pragmatic functions. However, due to the change of grammar, the pragmatic functions of Japanese interrogative sentences have changed, making their syntactic features disappear gradually and finally degenerate into "no doubt" interrogative sentences, which is also the grammatical change of Japanese interrogative sentences. Language ability and grammar ability are always changing, so we call them "variables". The advantages and disadvantages of Japanese interrogative sentences are different. Pragmatic wisdom and grammatical wisdom can form their own "wisdom chain", and their functions and sentence patterns determine the relationship between them. Pragmatic variants are related to grammatical variants. Pragmatic variation comes from pragmatic variation, pragmatic variation is the first and grammatical variation is caused by the continuity of pragmatic factors in grammar. From the function of expressing questions to other functions, Japanese interrogative sentences have experienced a process from strong to weak, from weak to no doubt, from weak to weak, from question to question[6]. Therefore, to study the functional conversion principle of Japanese interrogative sentence, we should start from its change law, from strong to weak, and then to no doubt. Before classification, we need to pre-process the problem (segmentation and deletion of terminator), and then use the feature vector to represent. First, we should know what problems we have and determine their types according to the classification system adopted. At present, there is no unified standard for the classification system of questions[7]. According to the classification basis, most researchers divide the questions into three types: answer type, semantic type and mixed information type. The classification method based on answer type is widely used in question answering system. It has the advantages of simple structure, small classification granularity and wide coverage. Especially, hierarchical classification can provide higher
classification accuracy and more constraints. Its classification system is mainly English[8]. The following table is the definition of the Chinese question classification system, including 7 categories. Each category defines several sub categories according to the actual situation, with a total of 60 sub categories.

| Major categories   | Subclass                                                                 |
|--------------------|--------------------------------------------------------------------------|
| Human (hum)        | Specific person group organization character description character list character others |
| Location (LOC)     | Planet City continent State Province rivers lakes mountains ocean island location list address location other |
| Number (Num)       | Number quantity price percentage distance weight temperature age area frequency speed range ordinal number list number other |
| Time (TME)         | Year month day time range time list time others                          |
| Entity (obj)       | Animals, plants, food, color, currency, language, language, material, mechanical means of transportation, religious entertainment entity list entity other |
| Description (DES)  | Abbreviated meaning method cause definition description other             |
| Unknown            | unknown                                                                   |

Tab.2 Classification system of Japanese questions

Most of the research on problem classification draws on the idea of text classification. By analysing the information contained in the text, we can judge which kind of text belongs to. However, different from the text, interrogative sentences are usually short sentences with little lexical information and rich context[9]. In order to obtain better feature information, it is necessary to analyse the problem deeply, so as to improve the accuracy of classification.

2.2 Classification of Japanese question intention

Deep learning is realized by using deep neural network. The neural network model is composed of a series of interconnected basic neurons, and the information is stored in the network by learning. Using heuristic rules to classify question intention is a rule-based classification method in Japanese questions[10]. The key words are understood as natural classification, so as to achieve the purpose of problem classification. At the same time, the problem intention classification based on rules also achieves good classification effect, but this classification method has its limitations. Japanese question questions can be expressed in many different ways. Through the analysis of this phenomenon, it is concluded that with the increase of the number of sentences, syntactic rules will increase sharply. Moreover, the ability to construct a large number of rules is not strong[11]. By analysing the corpus of a certain field, the rules constructed can only be used for data sets similar to the field, but not for other fields or other data sets. Therefore, it is difficult to construct a general rule framework to implement generalization. The main steps include the input of Japanese questions, data pre-processing, feature engineering, model training and question output, and understanding the classification items of Japanese questions. Based on this, the framework structure of Japanese question intention classification is optimized.
Before model training, data always needs to be pre-processed, which is a necessary step. Syntactic analysis is a hot topic in the field of grammar. The main component of syntactic structure is semantic dependence, that is, the binary relationship between pairs of phrases in a sentence. One of them is the core word and the other is the auxiliary word, which reflects the syntactic dependence between the core word and the auxiliary word. On the basis of analysing the results of dependency syntax, this paper proposes to classify sentences by subject, interrogative words and their adjuncts to improve the accuracy of classification. In order to classify the problem, we need to transform the feature into vector form and input it into the classifier[12]. At present, there are two kinds of problem classification methods: empirical rule method and statistical machine learning method. At present, machine learning method based on statistics is widely used because of its versatility, portability and scalability. Firstly, feature vectors representing different types of questions are extracted, and then the corpus of marked questions is statistically analysed to establish a classification system. Using this classifier, the problem can be classified and labelled. On this basis, a method based on feature vector extraction is proposed, which directly affects the classification accuracy[13]. According to the prior probability of the class and the distribution of the feature items, the conditional probability of the class is calculated to judge whether the text belongs to another category, and the maximum prior probability is taken as the text value of the class. In the aspect of problem classification, the improved deep learning model is usually used, and its mathematical form is as follows:

$$\arg \max_q P(q, Q_1, Q_2, \ldots, Q_n) = \arg \max_q \frac{P(q, Q_1, Q_2, \ldots, Q_n)}{P(Q_1, Q_2, \ldots, Q_n)}$$

(1)

Where q is a question type variable, Q1, Q2 QN is the participle variable of question. Because the denominator is not changed, it only involves the numerator. Therefore, the feature item after stopping using the word is deleted. According to the word pack model, questions can be simplified as follows:

$$\arg \max_q P(q, Q_1, Q_2, \ldots, Q_n) = \arg \max_q \left( P(q, Q_1) \times \cdots \times P(q, Q_n) \right)$$

(2)

Where (q, Qi) represents the number of times feature Qi appears in q class; N represents the number of problem types; counter (Qi) represents the total number of feature Qi in training set; M (M) represents the number of feature Qi in problem type M. In addition, the improved deep learning model uses TF-IDF weighted specific words to solve the problem of uneven distribution of corpus. In the process of feature extraction, the accuracy of feature extraction and part of speech tagging directly affect the accuracy of problem classification, so it is necessary to classify features. However, the current LTP platform is not ideal[14]. Many classification errors are caused by segmentation and part
of speech tagging errors. In the processing of Japanese interrogative sentences, we need to grasp the semantic information of the problem as a whole and analyse its syntax in order to improve the technology of word segmentation and part of speech tagging. In addition, different algorithms have different requirements for data and need different conversion. The performance of many machine learning algorithms is closely related to the amount of data. Normalization is a common normalization method, which adds all eigenvalues of each data to 1.

\begin{equation}
\|X\|_p = \arg \max_q P\left(q|Q_1, Q_2, \ldots, Q_n\right) \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}
\end{equation}

If the mean value of the feature is 0 and the variance is 1, then

\begin{equation}
\chi_{std}^{(i)} = \frac{\chi^{(i)} - \mu}{\sigma} \|X\|_p
\end{equation}

For each randomly selected nearest neighbor \(x_n\), a new sample is constructed:

\begin{equation}
x_{\text{new}} = \chi_{std}^{(i)} + \chi_{n} \ast \text{and}(0, 1) \ast (x_n - x)
\end{equation}

Two value cross entropy cost function and multi value cross entropy cost function are used as the model loss index, and the two value cross entropy is optimized by gradient descent method:

\begin{equation}
\text{Repeat } \theta_F = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j} = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_{\theta}(x^{(i)}) - y^{(i)})\right) \chi^{(i)} n
\end{equation}

In the problem classification, each independent feature does not play a very important role. It is necessary to combine multiple features to improve the accuracy of classification. There are mainly two feature combination methods: direct feature connection method and feature binding method based on word package. This method is simple and intuitive, adding one or more features directly after the word pack features, but it will increase the dimension of feature space, and some features themselves have no obvious effect[15]. The extra features will produce certain noise and reduce the classification accuracy, so which features should be selected according to the specific situation of the features. Feature binding based on word package means that part of speech, named entity, meaning and dependency are regarded as the attributes of vocabulary, and they are associated with lexical features to reflect the meaning of vocabulary more effectively. The classification of Japanese question intention based on the above method is relatively more detailed, which makes the extracted answers more accurate and improves the accuracy of classification.

3. Analysis of experimental results

On the basis of experiments, this paper studies the application of deep learning based automatic classification algorithm for Japanese problems and problem intention, and compares it with other models. The data set used in this experiment is from the log data collected by the third-generation robot DUWA of idepwise company. Japanese topic 68850, involving appointment, weather, music and other 13 aspects. Data pre-processing mainly includes missing value processing, wrong word correction, data classification and merging, hyphenation processing, de duplication and so on. Finally, 53117 samples were obtained, and the number of samples for each classification and the corresponding classification labels were obtained. The chatting class is divided into one class, and the
other 12 classes are divided into one class, and two deep learning models are established. On this basis, combined with two deep learning models, hierarchical classification technology is used to identify natural language problems. In the process of experiment, in order to verify the effectiveness of the method, two kinds of models and multi class models are randomly divided. 80% data is selected for training set and 20% for test set. This is a test on a company server. The hardware and software configuration details of the selected R & D are shown in the table.

| parameter                  | Dichotomous model | Multi classification model | Version information |
|----------------------------|-------------------|----------------------------|---------------------|
| input                      | Word vector       | Word vector                | 2017.3              |
| Word vector dimension      | 200               | 260                        | 1.4                 |
| Activation function        | RELU/SIGMOID      | RELU/SIGMOID               | Ubunte 16.04        |
| Optimization function      | adam              | adam                       | 128GB               |
| loss function              | Pycharm-professional | Tensorflow-gpu            | 4TB                 |
| Training rounds            | 40                | 60                         | TITAN               |
| Training rounds            | 20                | 20                         | X-Pascal            |
| Batch size                 |                   |                            |                     |

**Tab.3** Experimental parameter setting

Based on the above experimental environment and parameters, comparative detection was carried out and the test results were recorded, as shown in Fig.2.

**Fig.2** Comparison test results

Based on the above detection results, compared with the traditional methods, the classification accuracy and influence parameters of the proposed method are relatively higher in the practical application process. The automatic classification method of Japanese question intention should be unique, because the meaning of the question in this paper comes from the answer entity category, so a group of answers is obtained. After the case, we need to determine the unique answer category from
this set to improve the classification accuracy. It is proved that the automatic classification of Japanese question intention based on deep learning is better, which fully meets the research requirements.

4. Conclusion
This paper proposes a question classification method based on category matching ranking. Question classification is an important module in question answering system. The main purpose of question classification is to determine the sentence pattern of user's question, which plays an indirect role in the subsequent answer extraction strategy. Therefore, improving the accuracy of question classification is also the key to improve the overall performance of the system. This paper summarizes the latest development of question classification in automatic question answering system from three aspects: question classification system, question feature extraction and classification model. Based on the analysis of the difficulties existing in question classification, the future development direction of question classification in language processing platform, question set and classification is proposed, which lays a foundation for further research on more effective question classification methods basics.

Acknowledgments
This research was partly supported by the Social Sciences Planning Projects of Fujian Province (No. FJ2016C163), the Xiamen Overseas Scholar Projects (No. XRS2016-314), and the High-level Talents Projects of Xiamen University of Technology (No.YKJ17011R).

References
[1] Tom Y, Devamanyu H and Soujanya P 2018 Recent trends in deep learning based natural language processing IEEE Computational Intelligence 13(3) pp 55-75
[2] Yoshida M, Yanuaryska R D and Shantiningsih R R 2019 Comparison of radiation risk perception and knowledge of radiation between Indonesian and Japanese dental students Journal of Environmental Radioactivity 204(8) pp 104-110
[3] Tang F, Mao B and Fadlullah Z M 2018 On removing routing protocol from future wireless networks: a real-time deep learning approach for intelligent traffic control IEEE Wireless Communications 99 pp 1-7
[4] Zhang Z, Xu S and Capinha C 2019 Using species distribution model to predict the impact of climate change on the potential distribution of Japanese whiting Sillago japonica Ecological Indicators 104(9) pp 333-340
[5] Chen F C and Jahanshahi R M R 2018 NB-CNN: deep learning-based crack detection using convolutional neural network and naïve bayes data fusion IEEE Transactions on Industrial Electronics 65(99) pp 4392-4400
[6] Yuan X, Xie L and Abouelenien M 2017 A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data Pattern Recognition 77(2) pp 160-172
[7] Vardan P, Yaniv R and Jeremias S 2018 Theoretical foundations of deep learning via sparse representations: a multilayer sparse model and its connection to convolutional neural networks IEEE Signal Processing 35(4) pp 72-89
[8] Kulin M, Kazaz T and Moerman I 2018 End-to-end learning from spectrum data: a deep learning approach for wireless signal identification in spectrum monitoring applications IEEE Access 6(4) pp 18484-18501
[9] Al-Antari M A, Al-Masni M A and Choi M T 2018 A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification Int. Journal of Medical Informatics 117(9) pp 44-54
[10] Sugawara S, Kushida and Iwagaki Y 2018 The 1975 type japanese diet improves lipid metabolic parameters in younger adults: a randomized controlled trial Journal of Oleo Science 67(5) pp 599-607
[11] Whang James 2018 Recoverability-driven coarticulation: acoustic evidence from Japanese high
vowel devoicing. *The Journal of the Acoustical Society of America* 143(2) pp 1159-1172

[12] Ono K, Kato E and Tsunemi K 2019 Does risk information change the acceptance of hydrogen refueling stations in the general Japanese population? *Int. Journal of Hydrogen Energy* 44(31) pp 16038-16047

[13] Vasconcellos L, Carvalho C T and Tavares R O 2018 Isolation, molecular and phenotypic characterization of Cronobacter spp. in ready-to-eat salads and foods from Japanese cuisine commercialized in Brazil *Food Research International* 107(5) p 353

[14] Pan J, Jia H and Shang M 2018 Physiochemical properties and tastes of gels from Japanese Spanish mackerel (Scomberomorus niphonius) urimi by different washing processes *Journal of Texture Studies* 49(6) pp 578-585

[15] Fuse S, Nirengi S and Amagasa S 2018 Brown adipose tissue density measured by near-infrared time-resolved spectroscopy in Japanese, across a wide age range *Journal of Biomedical Optics* 23(6) pp 1-9