Research on Classification of Kazakh Questions Integrate with Multi-feature Embedding

Gulizada HAISA
College of Information Science and Engineering, Xinjiang University; The Base of Kazakh and Kirghiz Language of National Language Resource Monitoring and Research Center on Minority Languages; Xinjiang Laboratory of Multi-language Information Technology ,Ürümqi, Xinjiang. P.R. China
E-mail: gulzada@stu.xju.edu.cn

Gulila ALTENBEK
(Corresponding Author)
College of Information Science and Engineering, Xinjiang University; The Base of Kazakh and Kirghiz Language of National Language Resource Monitoring and Research Center on Minority Languages; Xinjiang Laboratory of Multi-language Information Technology ,Ürümqi, Xinjiang. P.R. China
E-mail: gla@xju.edu.cn

Hayinaer AIERZHATI
School of Computer Science and Technology, University of Science and Technology of China, Hefei. P.R. China
E-mail: hynr@mail.ustc.edu.cn

Kaden kenzhekhan
Almaty University of Power Engineering and Telecommunications, Almaty. Kazakhstan
E-mail: K.kaden@aues.kz

***Abstract*—Kazakh is an agglutinative language, and this feature results in data sparseness to some extent. In addition, the Kazakh question sentences lack strict grammatical rules, and the word forms are changeable and irregular. Due to this, this article proposes a CNN+BiGRU question classification model and attention mechanism that integrate multi-features based on the Kazakh language characteristics. Kazakh words and language features are used as input for the neural network. The CNN network generates high-dimensional semantic features and transmits the output to the BiGRU network. The BiGRU layer models the context information, then the Attention layer concentrates on the input features, filters out the unnecessary information, and completes the classification with SoftMax. The research in this paper shows that our model effectively integrates language features, avoids data sparsity, improves the model's performance during training, and has higher performance on the classification of Kazakh questions.**

Keywords-Kazakh; question classification; multi-feature embeddings

I. INTRODUCTION

The question answering(QA) system is an important area of natural language processing, which can answer people's questions quickly and accurately[1]. The QA comprises three parts, namely question analysis, information retrieval, and answer acquisition[2]. Question classification, as the initial link between the question and answer system, directly impacts the subsequent acquisition of answers[3]. The question classification task can also determine the user's intentions and identify useful information in the question sentence, which is very useful for downstream tasks. For low-resource languages, question classification is still a challenging task.

Kazakh belongs to the low-resource languages. It is an agglutinative language with a large number of morphemes. When multiple affixes are added to a word, it will bring several morphology variations, causing data sparsity. Because of complex morphological structures and insufficient datasets, it is still in the preliminary research stage to classify Kazakh question sentences.

As a result of the peculiar characteristics of the Kazakh language, using the currently popular neural networks alone cannot provide comprehensive semantic information. To achieve this, the integration of linguistic features and attention mechanisms is proposed in the form of a hybrid neural network model. The contributions of this article are as follows:

(1) Due to the complexity of the Kazakh language and poor question standardization, this article combines the multi-features of the Kazakh language with current neural network technology. It fully considers the lexical and syntactic features, which effectively solve the problem of data sparseness.

(2) Using the hybrid network model proposed in this paper has a better performance on the question set in the Kazakh tourism field.

This paper is organized as follows: The following section discusses related work; Section three discusses Model architecture; Section four the experiment and analysis; Finally, there is a conclusion and future work section.

II. RELATED WORKS

Question classification is a classic natural language processing task. Researchers have investigated several approaches to this problem. Currently, traditional machine learning and deep learning methods are commonly used for question classification. In traditional machine learning techniques, SVM (Support Vector Machine) models use binary classification methods that combine many identical features to cope with multi-classification problems[4].

2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT) | 978-1-6654-3757-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/CECIT53797.2021.00169

978-1-6654-3757-8/21/$31.00 ©2021 IEEE
DOI 10.1109/CECIT53797.2021.00169

Metzler et al. [5] classified English questions using a radial basis kernel function and various feature fusion methods. Chen et al. [6] used the NBM (Naïve Bayesian Model) to study question classification. Jia et al. [7] used a weighted KNN (K-Nearest Neighbor) classifier to classify Chinese questions.

Deep learning methods have proven to mine the depth information of texts. Many scholars currently use deep learning research methods due to neural networks ' solid nonlinear fitting capabilities. Commonly used models are network models such as FastText, CNN(Convolutional Neural Networks), LSTM (Long-Short Term Memory), RCNN(Region-based CNN), GRU(Gated Recurrent Unit), and BERT(Bidirectional Encoder Representation from Transformers). Xia et al. [8] designed a character-level LSTM network based on the attention mechanism to refine the classification of files. This structure can effectively capture local features and help to learn long-distance associations in the input sequence. Shi et al. [9] combined Bidirectional LSTM(BiLSTM) and CNN network models to extract question features for classification. Liu et al. [10] assigned the weight of the information obtained by Bidirectional GRU(BiGRU) to the attention mechanism in Chinese question classification and then completed the learning of local knowledge through CNN. Jin Ning et al. [11] put forward the BiGRU-MulCNN classification model for agricultural question classification tasks, using a bidirectional, recurrent unit neural network to obtain contextual feature information and parallel convolutional neural network for multi-granularity feature extraction. Mohammed et al. [12] used Bloom's classification as a basis to extract features and classify questions through TFPOS-IDF and word2vec.

There are very few researches related to Kazakh question sentences, but some researchers have researched for nearly ten years on Kazakh text classification tasks. Early studies used the traditional machine learning algorithm KNN [13], SVM [14], and other primitive algorithms to classify Kazakh text. Researchers have recently included fusion feature representation [15] and convolutional neural network [16] to type Kazakh short texts. In addition, as far as we know, this article is the first to study text classification in the field of Kazakh tourism, so this research has important research significance.

III. Model Architecture

This paper proposes a CNN+BiGRU model that integrates multi-linguistic features and attention mechanisms. The model comprises feature extraction, CNN, BiGRU, Attention, and Softmax layers. Figure1 shows the model structure diagram.

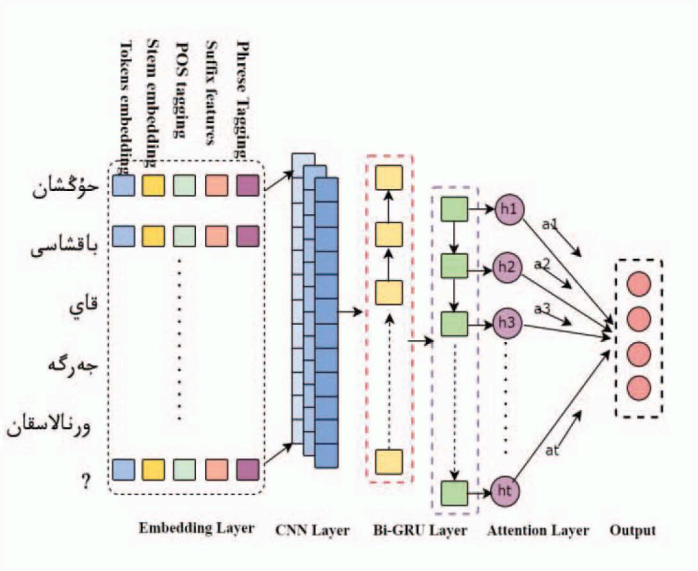

Figure 1. Model Architecture.

A. Feature Extraction Layer

Whether the extracted features are compelling directly affects the question classification model in model training. Here, according to the language characteristics of Kazakh language, we summarized a wealth of features that are conducive to question classification.

For example, the most common words (in the Kazakh tourist question) "قالاسندىعى" (on the city), "قالاسنان" (from the city), "قالاسى" (the city) and "قالاسىنىڭ" (in the city) are only "قالا" (city). The "سى" , "ىن" , "دا" , "عى" and "نىڭ" are all suffixes. Therefore, morphological analysis effectively solves the problem of data sparseness. In addition, the study of Kazakh morphological analysis [17]and stem extraction [18] has also done a lot of work and has been applied in practical research. The above research results of Kazakh morphology make it possible to apply morphological analysis to Kazakh question sentence classification. The features selected in this article are shown in TABLE Ⅰ.

TABLE I. Examples of linguistic features of Kazakh words

Features	Description	Example
Token	Tokens	قالاسندىعى
Stem	Stem or root	قالا
POS	Part of speech tag	Noun
POS-1	The first suffix tag	S3
POS-1	The second suffix tag	C5
POS-1	The third suffix tag	LATT
Suffix-1	The first suffix	سى
Suffix-2	The second suffix	ىن
Suffix-3	The third suffix	دا/عى
Chunk	Phrase type	NP-I

The rich features discussed above serve as input vectors for the neural network in this article. The overall vector can be expressed as:

$$V = [V_{token}:V_{stem}:V_{POS}:V_{chunk}:V_{su}] \quad (1)$$

In above formula, V_{token} is the word vector, V_{stem} is the stem feature, V_{POS} is the POS tagging and V_{chuck} is the chunk vector.

B. CNN Layer

With CNNs (Convolutional Neural Networks), it is possible to retrieve local feature information, such as stems, suffixes, positional and tagging information, effectively acquiring word morphological information. As a rule of thumb, the average pooling or maximum pooling layer is usually added after the convolution layer to sample the data at this time. By doing so, the calculation efficiency may be improved, and the risk of overfitting may be reduced. It should be noted, however, that due to the low dimensionality of semantic features of the text data, the advantages of the pooling layer will be less pronounced. Therefore, in order to obtain sufficient semantic information, this paper adopts the CNN without the pooling layer. It also provides the local features of the convolution layer to the BiGRU layer for calculation.

C. BiGRU Layer

Many scholars have proved recurrent Neural Networks (RNN) suitable for text processing tasks, but they have problems such as disappearing gradients in the face of long text sequences. As a variant of RNN, GRU controls information by updating gate and resetting gate, effectively solving text's long sequence dependence. Compared with another type of RNN's variant network, LSTM, GRU has fewer parameters, so the training speed is fast, requiring fewer experimental samples. The calculation formula of GRU is as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (2)$$

$$\tilde{h}_t = \tanh(W_{\tilde{h}} \cdot [r_t \times x_t]) \quad (3)$$

$$r_t = \sigma(W_i \cdot [h_{t-1}, x_t]) \quad (4)$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \tilde{h}_t \quad (5)$$

In the above equations, x_t is the input information at the current moment, h_{t-1} is the state information at the previous moment, σ is the Sigmoid function, and $\tilde{h}_t$ is the candidate set W_z, $W_{\tilde{h}}$, and W_i as the parameters to be learned. "$\cdot$" means the matrix elements are multiplied, and h_t is the output of the GRU at time t. This paper uses BiGRU to model the contextual information of the output of the feature layer. The network captures the semantic features of the text from the front and back directions. The formula is shown in 6-8:

$$\overrightarrow{h_t} = \overrightarrow{GRU}(e_t) \quad t \in [1,T] \quad (6)$$

$$\overleftarrow{h_t} = \overleftarrow{GRU}(e_t) \quad t \in [1,T] \quad (7)$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad (8)$$

h_t represents the vectorized representation obtained by BiGRU.

D. Attention Layer

This paper uses the attention mechanism to utilize the feature matrix of BiGRU effectively. To make the output more accurate, the attention mechanism assigns different weights to the result of the BiGRU layer. The output value of the model at time i is determined by the context feature C_i of the current value and historical information. The formula is as follows:

$$C_i = \sum_{j=1}^{n} a_{ij} h_j \quad (9)$$

In the above formula: h_j represents the hidden layer state of the encoder at the j-th time; n represents the length of the input sentence; a_{ij} represents the attention distribution probability of the output at the j th time, which is then calculated using the Softmax method. The calculation formula is as follows:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{ik})} \quad (10)$$

$$e_{ik} = V_a^T \tanh(W_a S_{i-1} + U_a h_j) \quad (11)$$

In the above equations: e_{ij} represents the evaluation score of influence on i outputs at j moments; S_{i-1} represents the hidden state of the decoder at the previous moment; V_a, W_a and U_a are the weight matrices.

IV. Experiments

A. Dataset

We have used a Kazakh question classification dataset annotated by The Base of Kazakh and Kirghiz Language of National Language Resource Monitoring and Research Center on Minority Language. China. The data set has a total of 8690 question sentences, and specific examples of each category are illustrated in TABLE II.

TABLE II. Classification System of Tourist Questions in Kazakh

Category	Example
Description	وكتابرده نارات جايلاۋىنىڭ اۋرايى قانداي؟
Human	باسباي شولاق ۇلىنىڭ قانداي مەكبەگى بار؟
Location	التاي قالاسىنا قانداي سايا‌حات وربندارى بار؟
Numeric	امايلۇڭ ويىن الاگىنىڭ بەلەت باعاسى قانشا؟
Time	تاگبالى ساحاراسى قاي جىلى سايا‌حات اشتى؟
Entity	قىز قۇار ويىنىڭ قاعيدالارى قالاي؟

The data is divided into three at a ratio of 8:1:1: training set, validation set, and test set. In TABLE Ⅲ, the specific distribution information is shown:

TABLE III. DATA SET DISTRIBUTION

Category	Train	Dev	Test
Description	1362	170	170
Human	333	42	42
Location	1511	189	189
Numeric	1391	174	174
Time	1128	141	141
Entity	1227	153	153

B. *Experimental Setup*

Word2vec[19], an open-source tool developed by Google, is used in this study for word vector training.To calculate word vectors, CBOW is used. The word feature window size is 8, and the minimum word frequency is 2. The detailed experimental parameters of the question classification model integrate multiple language features are shown in TABLE Ⅳ.

TABLE IV. EXPERIMENTAL PARAMETER SETTING

Parameter	The parameter value
Token size	128
Root size	64
Suffix size	64
POS tagging size	64
Phrase tagging size	64
CNN window size	3
Number of filters	50
BiGRU hidden size	256
Attention hidden size	100
Batch size	32
Epochs	50
Dropout	0.5
learning rate	0.001
Optimizer	Adam

C. *Parameter Setup*

Precision (P), recall (R), and F1-score (F1) are used as evaluation indices in this paper. Specific calculations are shown in equations (12), (13) and (14).

$$P = \frac{TP}{TP+FP} \times 100\% \tag{12}$$

$$R = \frac{TP}{TP+FN} \times 100\% \tag{13}$$

$$F1 = \frac{2*P*R}{P+R} \times 100\% \tag{14}$$

where the meanings of TP are true positives, TN is true negatives, FP is false-positives and FN is false negatives.

D. *Models for comparison*

To compare the performance of the proposed model in Kazakh tourism questions, this article examines several text classification models. The experimental results are provided in TABLE Ⅴ.

TABLE V. COMPARATIVE EXPERIMENTAL RESULTS OF DIFFERENT MODELS

Models	P(%)	R(%)	F1(%)
SVM	79.52	79.71	79.61
KNN	73.45	73.69	73.58
CNN	80.96	81.47	81.21
BiGRU	85.07	85.12	85.09
BiGRU+CNN	85.49	85.76	85.62
BiGRU+CNN+Att	85.86	86.19	86.02
BiGRU+CNN+Att+features	90.10	90.20	90.15

The results of the classical machine learning algorithms KNN and SVM in our Kazakh question classification experiment were significantly lower than those of the deep learning algorithm. The core idea behind this machine learning algorithm is to calculate the conditional probability by assuming that the words appearing in the question are independent of each other, so the training and testing time is shorter. Compared with the small sample data of KNN, SVM has better adaptability, and the indexes of P, R, and F1 are improved by 6.07, 6.02, and 6.04, respectively.

CNN and BiGRU, deep learning algorithms, have achieved better results than traditional machine learning algorithms KNN and SVM. The model based on deep learning networks strengthens the dependence on semantic features and context.

By combining BiGRU+CNNt, the Attention mechanism (BiGRU+CNN+Att) can be implemented, which weights the acquired features, filters useless data, and increases the P, R, and F1 indicators by 0.37, 0.43, and 0.40, respectively.

BiGRU+CNN+Att+features, a model that integrates multiple features proposed in this article, first uses word vector information with various features such as stems, affixes, parts of speech, and phrases to represent text input, to solve the sparsity of Kazakh tourism data effectively. This model performs better than BiGRU+CNN+Att, which uses only word vector input on P, R, and F1 indicators, respectively, by 4.24, 4.01, and 4.13.

E. *Ablation Experiment*

Further comparative experiments were designed to explore the influence of different features on the practical effect. Removing certain features and performing question classification are shown in the TABLE Ⅵ, where the complete hybrid model contains the multi-features proposed in this article (BiGRU+CNN+Att+ features).

TABLE VI. ABLATION STUDY ON DIFFERENT FEATURES

Different features	P(%)	R(%)	F1(%)
Complete model	90.10	90.20	90.15
W/O stems	87.07	87.61	87.34
W/O suffixes	87.09	87.02	87.05
W/O POS	88.56	88.87	88.71
W/O chunk	89.24	89.32	89.28

According to TABLE VI, when using the complete model, the model achieves the best results. However, if one of the linguistic features is removed, the model will gain less semantic information, and its performance will suffer. In particular, the embedding of suffixes and stem features has a more significant influence on the model. Thus, the morphological features have a more substantial impact on the Kazakh tourist question classification model.

V. Conclusion and Future Work

To tackle the question classification task on the morphologically complex language, this paper proposes a method of integrating morphological features with neural networks and effectively solving the problem of data sparsity. In addition, the Kazakh question classification model fused with multi-features can compensate for the inadequacy that cannot be fully obtained by merely using or splicing neural network models, thus effectively improving the Kazakh question classification process. The next step is to design a new neural network structure based on the existing one that still has some difficulty extracting features of complex morphological languages. Additionally, improving the accuracy of Kazakh morphological feature analysis would be helpful.

Acknowledgment

This work was supported by National Natural Science Foundation of P.R. China (NO： 62062062)， the National Key R&D Program of P.R. China (2020AAA0107902).

References

[1] Monisha S T A, Sarker S, Nahid M M H. Classification of bengali questions towards a factoid question answering system[C]//2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT). IEEE, 2019: 1-5.

[2] Yilmaz S, Toklu S. A deep learning analysis on question classification task using Word2vec representations[J]. Neural Computing and Applications, 2020: 1-20.

[3] Wang X, Wang H, Zhao G, et al. ALBERT over Match-LSTM Network for Intelligent Questions Classification in Chinese[J]. Agronomy, 2021, 11(8): 1530.

[4] Xuegong, Zhang. Introduction to statistical learning theory and support vector machines.Acta Automatica Sinica,2000.

[5] Metzler D, Croft W B. Analysis of statistical question classification for fact-based questions[J]. Information Retrieval, 2005, 8(3): 481-504.

[6] Yu Chen. An Algorithm of Question Classification in Question Answering[J]. Software engineering , 2015, 000(011):30-31.

[7] Keliang Jia,Xiaozhong Fan, Jinzhong Xu. Chinese Question Classification Based on KNN[J]. Microelectronics and Computer, 2008, 25(1): 156-158.

[8] Xia W, Zhu W, Liao B, et al. Novel architecture for long short-term memory used in question classification[J]. Neurocomputing, 2018, 299: 20-31.

[9] Mengfei Shi et al. "Community Q&A Question Classification Method Based on Bi-LSTM and CNN and Including Attention Mechanism." Computer System Applications 027.009(2018):157-162.

[10] Liu J, Yang Y, Lv S, et al. Attention-based BiGRU-CNN for Chinese question classification[J]. Journal of Ambient Intelligence and Humanized Computing, 2019: 1-12.

[11] Ning Jin, Chunjiang Zhao, Huarui Wu. Research on Classification Technology of Agricultural Questions and Answers Based on BiGRU-MulCNN[J]. Transactions of the Chinese Society of Agricultural Machinery, 2020, 051(005): 199-206

[12] Mohammed M, Omar N. Question classification based on Bloom’ s taxonomy cognitive domain using modified TF-IDF and word2vec[J]. PloS one, 2020, 15(3): e0230442.

[13] Hua Wang, Shouyong Wu. Kazakh text classification based on SVM[J]. Journal of Computer Applications, 2010, 30(06):1676-1678.

[14] GuLinazi, Tieli Sun, Yiliyaer. A Kazakh text classification method based on active learning support vector machine[J]. Journal of Intelligent Systems, 2011, 06(03):261-267.

[15] Yelibayeva G, Sharipbay A, Mukanova A, et al. Applied ontology for the automatic classification of simple sentences of the Kazakh language[C]//2020 5th International Conference on Computer Science and Engineering (UBMK). IEEE, 2020: 13-18.

[16] Parhat S, Ablimit M, Hamdulla A. A robust morpheme sequence and convolutional neural network-based Uyghur and Kazakh short text classification[J]. Information, 2019, 10(12): 387.

[17] Altenbek G, Abilhayer D, Niyazbek M, et al. A Study of Word Tagging Corpus for the Modern Kazakh Language[J]. Journal of Xinjiang University(Natural Science Edition), 2009.

[18] Altenbek G, Wang X. Kazakh segmentation system of inflectional affixes[C]//CIPS-SIGHAN Joint Conference on Chinese Language Processing. 2010.

[19] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.