Lexical and Syntactic Feature Fusion for Question Type Classification

Kuei-Mei Lin *Jui-Feng Yeh*
Department of Computer Science and Information Engineering
National Chiayi University
Chiayi, Taiwan
s1120311@mail.ncyu.edu.tw ralph@mail.ncyu.edu.tw

Abstract

In question answering (QA) system, the question intent determines the direction of answer prediction. Accurately identifying question intent is crucial for answer prediction. The recent question intent identification method uses interrogative words as the main feature for reasoning. However, when explicit interrogative words are absent or when interrogative words are ambiguous, these approaches are lead to confusion and get wrong classification. To address this limitation, this paper proposes a question classification framework that integrates lexical and syntactic features. The multiple features are extracted and the proposed framework infers question intent by the features. Experimental results demonstrate the effectiveness of the proposed approach.

Keywords: Question Classification, Feature Fusion , Intent Recognition

1 Introduction

With the recent rise of Large Language Models (LLMs), the answer has not limit to find data from search engine. The answers can find by interactive engagement through question-answering systems. The main function of a QA system is to provide precise answers for users. Thus, question intent identification called as question classification is an important foundational upstream task. Identifying question classification is a critical process for predicting accurate answers. This process directly reduces the search space for possible answer spaces. Consequently, it significantly enhances the overall accuracy of the QA pipeline. Consequently, enhancing the precision of the question classification is a critical direction.

However, a critical limitation is in question classification. The question classification heavily relies on explicit interrogative words to predict the question types. This over-reliance on surface-level question words creates a severe path dependency. When explicit interrogative words are absent or contextually ambiguous, the approaches which heavily over-reliance interrogative words lead to semantic confusion and make misclassifications. For instance, in the question "What color is her skirt?". The surface wh-word is "What" but the real semantic focus on the combination of the POS-tagged structures and the word "color". If only using the interrogative words "What", the intent may loss and fail to predict the correct question classification.

To solve this limitation, we propose a question classification framework based on the fusion of lexical and syntactic features. We recognize that the question intent can not just decide by the interrogative tokens. Thus, the lexical features and syntactic features are extracted. The lexical features include character, word, word shape, and phrase features. The syntactic features include Part-of-Speech (POS) tags and interrogative tokens. These heterogeneous features are combined to a feature set and used a feature selection algorithm to filter the best feature set in our system. Finally, the linear SVM model is trained and infer to the correct classification.

2 Related Works

Question Answering (QA) is an important task in natural language processing. Question classification plays a crucial role to define the aspect of the precise answer. Shi et al proposed a question type guided model in visual question answering system. The question type guide model detect the numeric or temporal type and achieve better performance in the task. Lampert et al. and Moldovan et al. extracted the semantic features for question classification by machine learning methods. Mallikarjuna et al. proposed a data

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 279–283
November 6-7, 2026 ©2026 Association for Computational Linguistics

augmentation in question classification task. The deep learning-based model had a limitation on data sparse. Thus, the authors used back translation, paraphrasing and AEDA algorithm to add the number of labelled data. Mishra et al. proposed a model that integrated dual attention mechanisms and question types. The dual attention mechanism provided cross-modal representations of both visual and textual information. Liu et al. proposed combining question types with adaptive entropy minimization to improve the identification of fine-grained and unreliable samples. Li and Roth proposed a machine learning approach to question classification. The authors designed layered semantic hierarchy and classified questions into fine grained classes. Aburass et al. proposed an ensemble model for question classification. The model combined the advantage of Electra, GloVe, and LSTM models. Electra provided complex language understanding, GloVe offered word-level vector representations, and LSTM contributed long-term dependencies.

3 Methodology

Integrated lexical and syntactic features is adopted to identify the intent of questions. Interrogative words usually are used to identify the intent of question in the previous studies. However, the single feature cannot be applied to all type of the questions. Thus, we propose a feature fusion-based question classification and the framework is illustrated in Figure 1. The proposed method extracts the lexical and syntactic features to solve the limitation of single feature reasoning in question type classification. These features from the question sentence are defined as the foundation for semantic interpretation. Then, A Linear Support Vector Machine (Linear_SVM) model is used to infer the question type of the questions.

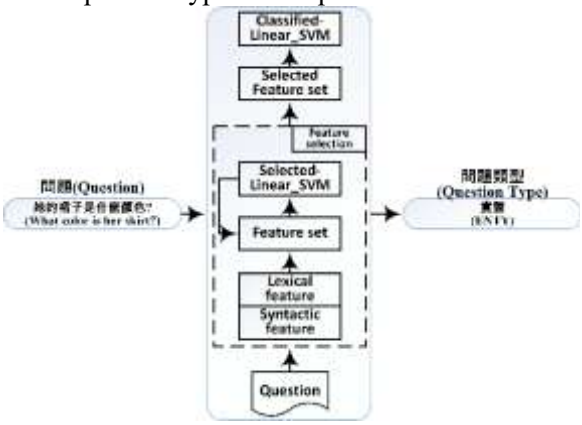

Figure 1: System framework

3.1 Feature extraction

To enrich the features and infer the question intent, lexical features and syntactic features are extracted. Lexical features are used to extract the fine-grain textual context and syntactic features are used to find the underlying grammatical structure of sentences.

Lexical features The character (Char), word, phrase, and word-shape features are the feature sets of lexical features. Character is the Chinese semantic smallest units. The individual characters provide foundational semantic and avoid the situation of Out-of-Vocabulary (OOV) problem. Thus, a question sentence is segmented to a sequence of single characters and be the character features. The words are a combination of characters and carry the more stabilized and concrete meanings. In word feature extraction, we use the CKIP word segmentation system to extract the words from the question sentences. For the character and word features, the BERT model is adopted to embed the segment result to an embedding. The formula (1)-(3) is defined the embedding process.

$$Q_c = [c_1 ... c_i], Q_w = [w_1 ... w_i] \quad (1)$$

$$V_c = BERT(c_i), i \in len(Q) \quad (2)$$

$$V_w = BERT(w_j), j \in len(Q_w) \quad (3)$$

V_c and V_w represent the embeddings of character and word features. The phrase features are constructed by the correlated words. The phrase features can capture the localized syntactic collocations and short-range contextual dependencies. In our proposed, the treebank and probabilistic context-free grammar (PCFG) are utilized to define the phrase features. The tree bank is used to label words with part-of-speech. The approach is defined by formula (4), and an example tree is provided in Figure. 2

$$p^{(1)} = f(W^{(B,C)} \begin{bmatrix} b \\ c \end{bmatrix})$$

$$p(P_1 \to B \quad C) = \frac{\exp(u_{BC}^{\mathrm{T}}([w_A; z]) + b_{BC})}{\sum_{B'C' \in M} \exp(u_{BC'}^{T}([w_A; z]) + b_{B'C'})} \quad (4)$$

$$s(p^{(1)}) = (v^{(B,C)})^T p^{(1)} + \log P(P_1 \to B \quad C)$$

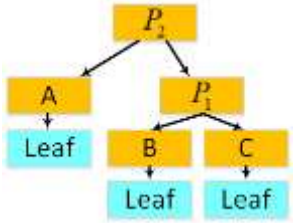

Figure 2: Parser tree diagram

The embedding of phrase features can not directly embed by the BERT model. Thus, the smooth inverse frequency (SIF) based algorithm is adopted. The SIF algorithm is based on the TF-IDF algorithm. The SIF algorithm performs a weighted average of character vectors based on their corpus frequency and subtracts the principal components of the text. The formula (5)-(6) are defined the phrase embedding by SIF algorithm.

$$v_{ph} = \frac{1}{|ph|}\sum_{c \in ph}\frac{a}{a + p(c)}v_c \quad (5)$$

$$v_{ph} = v_{ph} - uu^T v_{ph} \quad (6)$$

Where a is a free parameter typically set to 0.01. The frequency is defined as $p(c)$. v_{ph} represent the phrase feature embedding. Word-shape features is a surface properties. Word-shape features represent the normalization of question sequence. The features record uppercase and lowercase letters, numbers, and symbols in questions. Table 1 shows the example of lexical features.

Features	Example
Char	{(她)(的)(裙)(子)(是)(什)(麼)(顏)(色)(?)}
	–
Word	{(她)(的)(裙子)(是)(什麼)(顏色)(?)}
	{(What)(color)(is)(her)(skirt)(?)}
Phrase	{(她的)(裙子)(是)(什麼顏色)(的)(?)}
	{(What color)(is)(her skirt)(?)}
Word-shape	–
	{(lowercase,4)(mix,1)(other,1)}

– means the features without using

Table 1: The Example of Lexical Features

Syntactic features The part of speech (POS) features and interrogative words (wh-word) features are the feature sets of syntactic features. An example of syntactic features is shown as Table 2. Syntactic features represent the structural framework and grammatical relations of the question sentence. POS tags define the different function of the words and provide linguistic constraints. CKIP parsers are utilized to extract the POS features. Interrogative word is one of the key point to detect question classification. Thus, interrogative word is adopted as wh-word features and provide intuitive type hints.

Features	Example
POS	{(她, PN)(的, DEG)(裙子, NN)(是, VC)(什麼, VV)(顏色, VV)(的, SP)(?, PU)}
	{(What, WDT)(color, NN)(is, VBZ)(her, PRP)(skirt, NN)(?, .)}
Wh-word	{(什麼)}
	{(What)}

Table 2: The Example of Syntactic Features

3.2 Feature Selection and Identification

Although we extract many features, using all of features may not be a best optimal approach. Too many features can add redundant noise and lead to overfitting. To reduce the situation of redundant noise, an iterative wrapper-based greedy forward feature selection algorithm is implemented. This feature selection algorithm lead to find the most discriminative feature combinations.

In this feature selection algorithm, the optimal feature set is initialized to an empty set ($\mathbf{F} = \varnothing$), and the candidate feature set is defined as $SF = \{SF_1, SF_2, \ldots, SF_i\}$. When it executing, the algorithm sequentially examines each candidate feature subset by linear SVM classifier. Linear SVM classifier is training on candidate feature subset and testing the model to obtain a localized classification accuracy. The localized classification accuracy is used to compare the global maximum accuracy A. If localized classification accuracy is greater than A, localized classification accuracy replace the global maximum accuracy (A) and the candidate feature subset (SF) replace the optimal feature set (F). The iteration continues until all features have been explored. The feature set in optimal feature set (F) is the final feature set. The final feature is fed into the linear SVM classifier to train and predict the question classification.

4 System Evaluations

4.1 Dataset and Metric

Dataset The Text REtrieval Conference (TREC) is used to evaluate the proposed model for question type classification. The Text Retrieval Conference (TREC) dataset contain 6 classes : abbreviation (ABBR), description (DESC), entity(ENTY), human(HUM), location(LOC) and numeric(NUM). The TREC dataset is partitioned into training (5,500 pairs), validation (550 pairs), and test sets (550 pairs).

Metric The accuracy is adopted as the standard evaluation in question classification task. In the TREC dataset, the question classification task is a six-class classification. Thus, the accuracy is defined as the formulation (7).

$$\text{Accuracy} = \frac{\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i)}{N} \times 100\% \quad (7)$$

N denotes the total number of question-answer pairs in the testing dataset. $y_i \in \{1,2,...,6\}$ represents

the ground truth from six question classes. $\hat{y}_i$ represents classes prediction.

4.2 Experimental Setup

In this study, the model is implemented using Python (versions 3.6.9 and 2.7.12) and PyTorch (version 1.6.0) on Ubuntu 18.04. during the pre-training phase. Hardware configurations and experimental environment adopted an Intel Core i7-6700 CPU and an NVIDIA GeForce RTX 2080Ti GPU with 11GB of GDDR6 memory. LinearSVM is implemented using L2 regularization, the regularization parameter (*C*) is set to 1.0, and the squared hinge loss function. Logistic regression is also implemented with L2 regularization, *C* is 1.0 and the L-BFGS optimizer. The kernel-based SVM employs the radial basis function kernel. NN is a 3-layer neural network consisting of one fully connected hidden layer with 100 units and a Softmax output layer was trained using the Adam optimizer with a learning rate of 0.001 for 100 epochs and a batch size of 16.

4.3 Results

We design the two kind of the experiments. First one is ablation study. In the ablation study experiment, the different combination feature set is evaluated and analyzed them. Another one is comparison with other classifiers. We use different kind of question type classifiers and compare them in our approach.

Ablation Study In this subsection, we compare the different feature sets with the extracted lexical and syntactic features. Table 3 illustrates the classification accuracy of the Speech Act Identification framework under five different feature configurations.

As the result of Table 3, only using word feature obtain the baseline accuracy of 89.75%. When combining the word-shape feature and word feature, the accuracy experiences a decrease to 89.51%. We suggests that word-shape attributes capture abstract orthographic regularities but using it may inject noise without linguistic constraints. Thus, the result become worse than only using word feature. However, the word-shape feature obtain the improvement when it integrates with other features. The POS features add the linguistic constraints by the syntactic structure. The accuracy improve to 90.08% with the inclusion of explicit interrogative word (wh-word) features. The improvement validates that integrating the interrogative word and syntactic structures can obtain the effective improvement. Finally, the integration of phrase features achieve the optimal performance. The experiments demonstrate that the multi-dimensional lexical and syntactic features are the positive effect on the classification of linear SVMs.

Comparison with Other Classifiers In this subsection, we compare the different classifier and evaluate them on the same question classification dataset – TREC dataset. The proposed linear SVM

Word	**Word-shape**	**POS**	**Wh-word**	**ph**	Accuracy
●					89.75
●	●				89.51
●	●	●			89.86
●	●	●	●		90.08
●	●	●	●	●	90.20

Table 3: Ablation study of different feature configurations

model achieves the highest classification accuracy. Specifically, linear SVM model is better than the kernel-based SVM (83.20%) and logistic regression (89.00%) by 7.00% and 1.20%, respectively. The extracted features from question sentences typically define high-dimensional and sparse representations. LinearSVM learns maximum-margin linear decision boundary in the original feature space, efficiently handles sparse inputs. Consequently, LinearSVM achieved superior classification performance compared with kernel-based SVM and logistic regression. We compare the deep learning-based architectures. The neural network and the complex hybrid model Electra+GloVe+LSTM. The neural network need the large-scale semantic contexts to fine-tune the model but the question sentences of dataset are short. The context of short question sentences provide the training limitation. The limitation of information has highly possibility to overfitting during the training process. Consequently, the deep learning model may not even match the performance of traditional machine learning approaches.

model	Accuracy
Kernel-based SVM	83.20
LR	89.00
Linear SVM	**90.20**
NN	75.82
Electra+GloVe+LSTM[54]	80.00

Table 4: Comparison with other models in the TREC dataset

5 Conclusion

To address the problem of relying on a single interrogative word in question classification task, we propose a question classification framework that integrates lexical and syntactic features. Lexical and syntactic features are extracted to enrich feature information. A feature selection algorithm is employed to avoid the influence of redundant feature noise and overfitting. Then, Linear SVM model is adopted to infer the correct question classification. The experiment demonstrates the effectiveness of our approach. The best accuracy achieve 90.20%. Furthermore, we observe that machine learning framework is better than deep learning model because of the limitation of contextual information of question sentences.

References

Aakansha Mishra, Ashish Anand, Prithwijit Guha. 2022. Dual attention and question categorization-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4(1):81-91. https://doi.org/10.1109/TAI.2022.3160418.

Andrew Lampert, Robert Dale, Cécile Paris. 2006. Classifying speech acts using verbal response modes. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 34-41.

Challa Mallikarjuna, Sangeetha Sivanesan. 2022. Question classification using limited labelled data. *Information Processing & Management*, 59(6):103094.https://doi.org/10.1016/j.ipm.2022.103094.

Cristian Moldovan, Vasile Rus, Arthur C. Graesser. 2011. Automated speech act classification for online chat. In *Proceedings of the Midwest Artificial Intelligence and Cognitive Science Conference*, 710 pages 23-29.

Jin Liu, Jiang Xie, Fengyu Zhou, and Shengfeng He. 2024. Question type-aware debiasing for test-time visual question answering model adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):10805-10816.

Shi Yang, Furlanello Tommaso, Zha Sheng, Anandkumar Animashree. 2018. Question type guided attention in visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 151-166. https://doi.org/10.1007/978-3-030-01225-0_10

Sondos Aburass, Osama Dorgham, Mohammad Ali Rumman. 2024. An ensemble approach to question classification: Integrating ELECTRA transformer, GloVe, and LSTM. *International Journal of Advanced Computer Science and Applications*, 15(1):112-121.

Xin Li, Dan Roth. 2002. Learning question classifiers. In *Proceedings of COLING 2002: The 19th International Conference on Computational Linguistics*, pages 1-7. https://doi.org/10.3115/1072228.1072378.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 104–120. Springer, Cham. https://doi.org/10.1007/978-3-030-58577-8_7.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 33:6616-6628.