Model Construction for Question Retrieval in Community Question Answering Systems Based on HNC Sentence Categories (Postprint)

Authors: Wang Yu, Wang Fang

Date: 2019-05-10T00:00:00+00:00

Abstract

Community Q&A systems are inundated with substantial noise, posing challenges for users in information retrieval. Previous question retrieval models have predominantly focused on the lexical level. To address these issues, we construct a sentence-level question retrieval model. The novel model leverages sentence category knowledge from the Hierarchical Network of Concepts (HNC) theory to compute inter-question similarity from pragmatic, syntactic, and semantic perspectives. A question classification algorithm determines the categories of query and candidate questions to derive pragmatic similarity, while the structure of sentence category expressions and the composition of semantic chunks are employed to calculate syntactic and semantic similarities, respectively. Experimental results on real-world datasets demonstrate that the proposed HNC sentence category-based model enhances the accuracy of question retrieval outcomes.

Full Text

Preamble

Vol. 37 No. 6
Application Research of Computers
ChinaXiv Partner Journal

Construction of a Question Retrieval Model in Community Question Answering Systems Based on HNC Sentence-Category

WANG Yu, WANG Fang†
(Faculty of Management & Economics, Dalian University of Technology, Dalian, Liaoning 116024, China)

Abstract: Community question answering systems are plagued by substantial noise, creating difficulties for users seeking information. Previous question retrieval models have predominantly focused on the lexical level. To address these limitations, this paper constructs a sentence-level question retrieval model. The new model calculates similarity between questions from pragmatic, grammatical, and semantic perspectives based on sentence-category knowledge from Hierarchical Network of Concept (HNC) theory. Through a question classification algorithm, the model determines the question categories of both query and candidate questions to obtain pragmatic similarity. It then calculates grammatical and semantic similarity using the structure of sentence-category expressions and the composition of semantic chunks. Experiments on real datasets demonstrate that the new model based on HNC sentence-category improves the accuracy of question retrieval results.

Keywords: community question answering system; question retrieval; HNC theory; sentence category analysis; similarity calculation

0 Introduction

With the continuous development of the Internet, vast amounts of information have accumulated online. While ordinary users can obtain desired information through search engines, these engines return a series of relevant documents rather than direct answers to users' questions, requiring users to further screen the information. Additionally, users cannot describe their problems in natural language, and the sparse keyword sequences that fail to form complete semantics also lead to suboptimal retrieval performance from search engines.

Question answering systems provide a new information retrieval process from user questions to exact answers, simplifying the process of searching through uncertain documents and partially solving the problems of search engines. With the development of Web 2.0, ordinary users have gradually transformed from content recipients to content providers, resulting in abundant user-generated content (UGC) online. The main participants in question answering systems have also become community users, forming community question answering systems. Taking the Chinese Q&A community "Zhihu" as an example, since its launch, it has accumulated over 10 million questions and 34 million answers. Numerous questions have been resolved, becoming a valuable and knowledge-intensive online resource. How to find content similar or relevant to user questions from these extensive historical Q&A resources has become a primary research focus in question answering systems. Retrieving similar questions can reduce users' waiting time for answers and alleviate system redundancy caused by repeated similar content questions.

The retrieval objects in community question answering systems are questions and answers, which are shorter than typical documents and suffer from data sparsity issues. Moreover, the content exhibits casual natural language expression, with most words having polysemy and synonym phenomena, making strict term matching (such as vector space models) ineffective for identification.

To address these problems, Song et al. combined semantic and statistical information of questions to comprehensively calculate question similarity. Cai et al. utilized latent semantic information to solve the vocabulary gap problem in question retrieval, bridging the semantic gap through latent semantic information. Jeon et al. applied language models to community question answering system question retrieval, using a unigram model to model Q&A pairs for discovering similar questions. Xue et al. proposed a translation model-based language model on top of language models, effectively solving the word mismatch problem in retrieval. Reference [7] improved the retrieval performance of translation models by obtaining latent topic information of questions. However, the accuracy of translation models is susceptible to training corpora and text semantic expression, resulting in less-than-ideal retrieval performance. Xia et al. introduced the word knowledge base from Hierarchical Network of Concept (HNC) theory to modify translation probabilities, constructing a new question retrieval model and providing implementation algorithms for the model.

The focus of question retrieval models is examining sentence similarity. Many previous question retrieval models concentrated on lexical matching. Ignoring the pragmatic characteristics of questions and simply treating questions as collections of word sequences may lead to retrieved questions whose corresponding answers are not what users actually want. Community question answering systems contain rich exploitable information such as user social attributes and question answers. Reference [12] utilized the questioner' s social network attributes and query question content to construct latent representations of questions, integrating textual information with user social network information to rank similarity between questions. Question categories describe question types from a general pragmatic perspective, where questions belonging to the same category are more likely to have the same answer than those in different categories. Reference [13] extracted features from lexical, grammatical, and semantic levels, then constructed a semantic kernel function based on question syntactic trees, using Support Vector Machines (SVM) for question classification to improve accuracy. Tian et al. argued that question similarity calculation should consider not only the similarity of questions themselves but also the similarity of their answers, and since question categories can constrain answers to some extent, they added question type similarity as a dimension in question similarity calculation, providing a new perspective.

This paper employs HNC theory' s sentence-category analysis method to obtain deeper grammatical and semantic information about questions. From a computer automation perspective, it uses hierarchical analysis to determine question types and constructs a question retrieval model from grammatical, semantic, and pragmatic aspects at the sentence level to improve retrieval accuracy.

1 HNC Sentence-Category Analysis

HNC (Hierarchical Network of Concept) theory is a natural language processing system proposed by researcher Huang Zengyang from the Institute of Acoustics, Chinese Academy of Sciences. It describes specific patterns of brain cognitive structures oriented toward entire natural languages, dividing cognitive structures into local and global associative contexts. The local associative context is a conceptual representation system at the lexical level, while the global associative context involves associations at the sentence and discourse level, mainly including semantic chunk and sentence-category theories. Semantic chunks describe a sentence from the deep linguistic level, solving problems where sentence semantics are difficult to define at the word or phrase level. Any natural language sentence can be represented by a formal sentence-category expression, which consists of semantic chunks—semantic chunks are functions of sentence-categories. Common HNC sentence-categories include action sentences, process sentences, and transfer sentences. Sentence-category expressions reveal the semantic chunk sequence of a sentence at the grammatical level, where sentences with high similarity often share the same semantic chunk sequence. Therefore, grammatical similarity between questions can be derived from the similarity of their sentence-category expressions. Semantic chunks can be words, word sequences, or even sentences, but all are composed of words. Thus, semantic chunk similarity can be represented by lexical semantic similarity. HNC theory also has a lexical conceptual semantic network at the lexical level, demonstrating excellent performance in expressing lexical semantic completeness. Semantic-level similarity between questions can be determined by lexical conceptual semantics.

HNC constructs a novel natural language processing system. The HNC system can complete comprehensive semantic analysis through sentence-category analysis, obtaining a series of sentence-category knowledge including primary and auxiliary semantic chunks. Such knowledge provides convenient conditions for computer-automated sentence-related calculations. Sentence-category expressions obtained from analysis represent sentence-level syntactic analysis. HNC clustering includes 57 basic sentence-categories and 3,192 mixed sentence-categories, completely describing almost all grammatical collocation phenomena in natural language sentences. The corresponding semantic chunks composing sentence-category expressions contain relevant semantic information. HNC names the next-level semantic units of sentences as semantic chunks, which can range from words to phrases or even sentences. Semantic chunks are divided into primary and auxiliary semantic chunks based on their function in sentences. Primary semantic chunks are named feature elements, actors, objects, and content, while auxiliary semantic chunks include seven types: means, tool, path, comparison, condition, cause, result, and purpose.

In research applying HNC theory to sentence-category analysis, Chen Hong divided HNC sentence-category analysis into modules such as semantic chunk perception, sentence-category hypothesis, sentence-category verification, and semantic chunk composition analysis, implementing a complete HNC sentence-category analysis algorithm. Chi Zhejie calculated sentence similarity from two levels: primary/auxiliary semantic chunks and surface expression similarity, and deep semantic level, but did not consider the contribution of auxiliary semantic chunks to sentence similarity. Shi Yan calculated similarity by corresponding semantic chunks through sentence-category analysis and then averaging to compute final sentence similarity. Although this method utilized HNC theory, it overlooked various factors in semantic chunk composition and was therefore incomplete.

The ultimate goal of HNC sentence-category analysis is to obtain sentence-category expressions. However, the effectiveness of HNC sentence-category analysis depends on the HNC knowledge base, and the algorithm proposed in reference [18] has certain limitations in practical operation. Considering that sentence-categories are composed of semantic chunks, and semantic chunks serve as fundamental semantic units with pragmatic functions, we can derive sentence-category expressions for unknown sentences through similarity of part-of-speech tag sequences using known sentence-category expressions. For the sentence "The U.S. military airdropped relief supplies to victims of the Bosnian war," as a physical transfer sentence, it has the sentence-category expression: transfer initiator (U.S. military), transfer direction, transfer receiver (Bosnian victims), feature element (airdrop), and transfer content (relief supplies). The sentence also has a part-of-speech tag sequence: ns (U.S.), n (military), p (to), ns (Bosnia), n (war), n (victims), v (airdrop), n (relief supplies), with the correspondence shown in Table 1 , establishing the relationship between sentence-category expressions and sentence part-of-speech sequences.

By collecting sentences that have undergone HNC sentence-category analysis to obtain sentence-category expressions, then performing part-of-speech tagging on sentence word sequences, these sentences serve as prior knowledge. Sentence-category expressions for unanalyzed sentences can be obtained through similarity of sentence part-of-speech sequences.

The following sections will utilize HNC sentence-category analysis to obtain grammatical and semantic sentence information and the aforementioned pragmatic information represented by question types to comprehensively construct a question retrieval model.

Table 1 Correspondence between semantic block and part-of-speech sequence

2 Question Retrieval Model Based on HNC Sentence-Category Analysis

Based on the previous analysis of question categories and semantics, we construct a comprehensive question retrieval model from pragmatic, grammatical, and semantic levels. The effectiveness of a retrieval model must be verified through the accuracy of retrieval result ranking. In information retrieval tasks, metrics such as MRR, MAP, and AP@1 are typically used to demonstrate retrieval performance.

- **AP@1 (Average Precision):** The average percentage of relevant questions appearing in the first position of retrieval results for specific queries.
- **MAP (Mean Average Precision):** Represents the average precision of returned results. In our experiments, we calculate MAP for the top 10 results returned for each query, denoted as MAP@10.
- **MRR (Mean Reciprocal Rank):** While ensuring retrieval accuracy through MAP, we must also consider the ranking order of results. MRR incorporates the influence of ranking order into retrieval accuracy metrics.

Designing a ranking formula representing the retrieval model is essential for validating model effectiveness. This section designs a ranking formula for the retrieval model. The new question retrieval ranking mechanism is as follows:

$$Sim(Q,C) = \alpha \cdot Sim_{cat}(Q,C) + \beta \cdot Sim_{gse}(Q,C) + \gamma \cdot Sim_{sem}(Q,C)$$

where $Sim(Q,C)$ represents the overall similarity between query question Q and candidate question C; $Sim_{cat}(Q,C)$, $Sim_{gse}(Q,C)$, and $Sim_{sem}(Q,C)$ represent question type similarity, grammatical similarity, and semantic similarity, respectively; α, β, and γ are adjustment parameters representing the weights of the three similarity components in question retrieval. The algorithm flowchart is shown in Figure 1 [Figure 1: see original paper].

The following sections describe the similarity calculation process for each component.

2.1 Question Type Similarity Measurement

Traditional question answering systems primarily present factual questions, such as "Who was the first Chinese Nobel laureate?" or "What is a firewall?" However, community question answering systems feature multiple question forms from a pragmatic perspective. Reference [21] categorizes community question answering system questions into seven types: definition, fact, process, reason, opinion, yes/no, and description. Question types determine the answer focus of related questions. For instance, answers to process questions are completely different from those to definition questions. Since the purpose of question retrieval is to find similar questions in historical Q&A archives to obtain answers for newly submitted questions, pragmatic information represented by question types is crucial for question retrieval.

In Chinese text classification research, classification is generally performed by extracting classification features. Zhang et al. [22] built upon reference [23]' s proposal of applying keywords to document classification by adding keyword similarity calculation in short text classification, selecting K-Nearest Neighbors (KNN) and SVM as classifiers, which improved classification performance compared to traditional methods. Gao et al. [24] proposed integrating category clue words such as central words, subjects, interrogative words, and interrogative-related components on top of bag-of-words features, 挖掘 more useful classification information from syntactic or semantic perspectives. Yu et al. [25] proposed that interrogative words in questions are influential features, and constructing an interrogative word attention matrix can enhance the model' s focus on interrogative features to improve question classification accuracy. Reference [26] extracted question backbone words and interrogative words as classification features, then used a Bayesian classifier for question classification. This section builds upon reference [26]' s feature dimensions and considers the manual labeling attribute of community question answering systems, using question category labels as an additional feature for question classification, employing a Bayesian classifier.

First, we determine the question types of both the query question and candidate questions. The question classification method extracts classification features as described above and inputs them into a Bayesian classifier. When calculating question type similarity, we consider that although answers to different question types may have different foci, the background knowledge in answers may still be valuable to questioners. Therefore, question type similarity is determined by:

$$Sim_{cat}(Q,C) = \begin{cases} 1, & \text{if belonging to the same subcategory} \\ \alpha_1, & \text{if belonging to the same major category} \\ 0, & \text{if belonging to different categories} \end{cases}$$

where α_1 is the similarity coefficient for different question types, determined during parameter tuning.

The question category similarity calculation steps are as follows:

Algorithm 1: Question Category Similarity
Input: Questions Q and C
Output: Question category similarity $Sim_{cat}(Q,C)$

a) Extract the focus word sets of questions Q and C;

b) Compare the focus word sets with the word sets representing question categories through semantic similarity to determine the question categories of Q and C;

c) Calculate question category similarity using Equation (2).

2.2 Question Sentence-Category Expression Similarity Measurement

HNC sentence-category analysis yields the constituent elements at the sub-sentence level: semantic chunks. When examining similarity in sentence-category expressions composed of semantic chunks, we compare the semantic chunk compositions between the query question and candidate question, using quantitative methods to calculate sentence-category expression similarity. Semantic chunks are functions of sentence-categories. Primary semantic chunks that play a core role in sentences are called main semantic chunks, while explanatory semantic chunks are called auxiliary semantic chunks. Sentences with high similarity often share multiple identical semantic chunks. The sentence-category expression similarity calculation formula is defined as:

$$Sim_{gse}(Q,C) = \beta_1 \cdot Sim_{gsem}(Q,C) + \beta_2 \cdot Sim_{gsef}(Q,C)$$

where β_1 and β_2 are adjustment parameters representing the weights of main and auxiliary semantic chunk similarities in question expression similarity calculation; $Sim_{gsem}(Q,C)$ and $Sim_{gsef}(Q,C)$ represent main and auxiliary semantic chunk similarities, respectively. Drawing from reference [18]' s calculation method, similarity is determined according to the categories of main and auxiliary semantic chunks:

$$Sim_{gsem}(Q,C) = \begin{cases} 1, & \text{having identical sentence-category symbols} \\ \beta_{11}, & \text{belonging to generalized action or effect sentences} \\ 0, & \text{otherwise} \end{cases}$$

$$Sim_{gsef}(Q,C) = \begin{cases} \beta_{21}, & \text{if } FK_1 = FK_2 \text{ and } CFK_1 \neq CFK_2 \\ \beta_{22}, & \text{if } FK_1 \neq FK_2 \text{ and } CFK_1 \cap CFK_2 \neq \emptyset \\ 0, & \text{if } FK_1 \neq FK_2 \text{ and } CFK_1 \cap CFK_2 = \emptyset \end{cases}$$

where coefficients β_{11}, β_{21}, and β_{22} represent similarity degrees of main/auxiliary semantic chunks under different conditions, determined during parameter tuning; FK_i represents the auxiliary semantic chunk set of sentence i; CFK_i represents the content set of auxiliary semantic chunks of sentence i.

The question expression similarity calculation steps are as follows:

Algorithm 2: Question Sentence-Category Expression Similarity
Input: Questions Q and C
Output: Question sentence-category expression similarity $Sim_{gse}(Q,C)$

a) Perform HNC sentence-category analysis on Q and C to obtain corresponding sentence-category expression sets;

b) Calculate main and auxiliary semantic chunk similarities using Equations (4) and (5);

c) Substitute into Equation (3) to obtain the final sentence-category expression similarity.

2.3 Question Semantic Chunk Similarity Measurement

The calculation methods related to question sentence-category expression measurement in Section 2.2 represent surface-level similarity calculation. This section calculates sentence semantic-level similarity based on semantic chunks obtained from HNC sentence-category analysis. Semantic chunks can be words, phrases, or sentences, but all are word sequences. We can use the method introduced in reference [8] to calculate word sequence similarity. Sentence-category analysis yields two types of semantic chunks: main and auxiliary. Main semantic chunks play a dominant role in sentence semantics, while auxiliary semantic chunks represent explanatory components. General sentence similarity calculations ignore the contribution of auxiliary semantic chunks, which is reasonable for ordinary natural language sentences. However, in community question answering systems, due to limited question types, numerous questions share identical sentence-category expressions, making auxiliary semantic chunks crucial for measuring sentence similarity.

When calculating question semantic similarity, we correspond semantic chunks between questions. The calculation formula is:

$$Sim_{sem}(Q,C) = \gamma \cdot Sim_{semm}(Q,C) + (1-\gamma) \cdot Sim_{semf}(Q,C)$$

where $Sim_{semm}(Q,C)$ represents similarity between corresponding main semantic chunks; $Sim_{semf}(Q,C)$ represents similarity between corresponding auxiliary semantic chunks. Since main semantic chunks play the primary role in sentence semantics, coefficient γ should be set to a value greater than 0.5.

The question semantic chunk similarity calculation steps are as follows:

Algorithm 3: Question Semantic Chunk Similarity
Input: Questions Q and C
Output: Semantic chunk similarity $Sim_{sem}(Q,C)$

a) Correlate the main and auxiliary semantic chunks obtained from sentence-category analysis;

b) Use the lexical semantic similarity calculation method proposed in reference [8] to obtain similarities between main and auxiliary semantic chunks;

c) Substitute the semantic chunk similarities into Equation (6) to calculate the final similarity.

With all components of the question retrieval model's ranking formula now calculated, they are combined in Equation (1) to form the complete ranking formula. The following experiments validate the proposed algorithm' s effectiveness.

3 Experiments

To demonstrate the effectiveness of the question retrieval model proposed in Chapter 2, we selected three classic retrieval models for comparative analysis: vector space model-based, language model-based, and translation model-based retrieval models. The language model algorithm was implemented as described in reference [5], and the translation model algorithm as described in reference [6]. These three models were chosen because they are simple yet effective, and numerous previous studies have used them as foundations for designing retrieval models. The specific implementations from references [5] and [6] represent widely recognized high-performance implementations of these two retrieval models, and using them as reference models enhances the credibility of our experimental results.

Due to limitations in HNC theoretical concept lexicon construction, our experimental dataset is in Chinese. First, we constructed a training corpus for the translation model to obtain prior knowledge. The training corpus construction process: randomly collect Q&A pairs from the Zhihu Q&A community, then identify similar questions among these pairs using Q&A system tagging features, and train the translation model using similar question pairs. For the test dataset, we used 1,140 manually labeled test questions, each with 20 candidate questions labeled as similar or dissimilar. We selected 20 test questions as a model parameter tuning set to determine the parameter settings in Chapter 2, as shown in Table 2 . The remaining test questions were used to validate all comparison models.

Table 2 Model parameter settings

Table 3 shows the comparison of models across evaluation metrics. "Improvement (IMPR)" indicates the improvement of each model over the vector space model for each metric.

Table 3 Comparison of models on MRR, MAP, and AP@1

The experimental results in Table 3 show that the simple vector space model lags behind other models across all retrieval metrics, demonstrating that term-matching models face data sparsity issues that prevent most similar questions from being recalled. Thus, bag-of-words models like the vector space model are less effective in UGC-rich text environments than in large-scale document resources.

Additionally, translation and language models show significant improvement across metrics because they have well-developed smoothing mechanisms, making their retrieval performance superior to ordinary vector space models. The translation model outperforms the language model because it can obtain background semantic knowledge from large-scale corpora that language models lack, and it replaces general text similarity calculation with translation probability.

Finally, the proposed HNC sentence-category analysis method demonstrates better improvement over previous methods in question retrieval effectiveness. The other three models simplify question similarity calculation in retrieval, making their results represent only semantic or lexical similarity rather than comprehensive similarity across semantics, grammar, and pragmatics, thus missing some similar questions. Experiments show that our syntax analysis-based retrieval model, combining multi-angle similarity from semantic, grammatical, and pragmatic perspectives, improves question retrieval effectiveness, validating our approach' s validity. During experimental verification, the HNC sentence-category analysis model involves more complex computation and multiple similarity dimensions than the previous three models, making it more robust, as errors in individual similarity dimensions have less impact on final accuracy.

4 Conclusion

Leveraging HNC natural language processing theory's sentence-category analysis method, this paper proposes a novel question retrieval model constructed from grammatical, semantic, and pragmatic perspectives based on sentence-category analysis results, improving question retrieval performance over previous models.

For a given query, we first determine the question categories of both query and candidate questions through a question classification algorithm to obtain question type similarity (pragmatic similarity). We then perform HNC sentence-category analysis on both questions and utilize the analysis results from grammatical and semantic perspectives to obtain final question similarity, which serves as the basis for ranking candidate questions. Experiments demonstrate that our proposed method significantly outperforms previous classic retrieval models. This paper considers question categories at the pragmatic level; user social attributes in community question answering systems are also important. Future work will consider incorporating user needs and interests into pragmatic similarity calculation.

References

[1] Yan Xia, Fan Shixi. On key techniques of massive question sentences retrieval based on community QA [J]. Computer Applications and Software, 2013, (7): 315-317.

[2] Su Xiaohu, Yang Sichun. Research on Chinese question answering system based on improved VSM [J]. Information Studies: Theory & Application, 2008, 31 (4): 624-627.

[3] Song W, Feng M, Gu N, et al. Question similarity calculation for FAQ answering [C]// Proc of International Conference on Semantics, Knowledge and Grid. 2007: 298-301.

[4] Cai L, Zhou G, Liu K, et al. Learning the latent topics for question retrieval in community QA [C]// Proc of International Joint Conference on Natural Language Processing. 2015.

[5] Jeon J, Croft W B, Lee J H. Finding similar questions in large question and answer archives [C]// Proc of ACM International Conference on Information and Knowledge Management. 2005: 84-90.

[6] Xue X, Jeon J, Croft W B. Retrieval models for question and answer archives [C]// Proc of International ACM SIGIR Conference on Research and Development in Information Retrieval. 2008: 475-482.

[7] Zhang Weinan, Zhang Yu, Liu Ting. A topic inference based translation model for question retrieval in community-based question answering services [J]. Chinese Journal of Computers, 2015, 38 (2): 313-321.

[8] Xia Yuanyuan, Wang Yu. Construction of a query model of community question answering system based on HNC theory [J]. Computer Applications and Software, 2018, 35 (8): 98-101,169.

[9] Xiong Daping, Wang Jian, Lin Hongfei. An LDA-based approach to finding similar questions for community question answer [J]. Journal of Chinese Information Processing, 2012, 26 (5): 40-45.

[10] Chen Kang, Fan Xiaozhong, Liu Jie, et al. Calculation method of Chinese question semantic similarity based on question semantic representation [J]. Transactions of Beijing Institute of Technology, 27 (12): 1073-1076.

[11] Yang Haitian. The study on question retrieval technology in community question answer system [D]. Dalian: Dalian University of Technology, 2014.

[12] Chen Z, Zhang C, Zhao Z, et al. Question retrieval in community-based question answering via heterogeneous social influential network [J]. Neurocomputing, 2018, 285: 117-124.

[13] Jiang Longquan, Zhang Bo, Hu Zhipeng, et al. A semantic kernel function based question classification algorithm in question answering system [J]. Journal of Shanghai Normal University: Natural Sciences, 2018, 47(1): 53-56.

[14] Mishra A, Jain S K. A survey on question answering systems with classification [J]. Journal of King Saud University: Computer and Information Sciences, 2016, 28 (3): 345-361.

[15] Tian Weidong, Qiang Jipeng. Questions similarity computation based on question classification [J]. Application Research of Computers, 2014, 31 (4): 1090-1093.

[16] Huang Zengyang. HNC (hierarchical network of concept) theory [M]. Beijing: Tsinghua University Press, 1999.

[17] Huang Zengyang. HNC theory summary [J]. Journal of Chinese Information Processing, 1997, 11 (4): 11-20.

[18] Chen Hong. Natural language understanding–the study of sentence category analysis based on the theory of HNC [D]. Changchun: Changchun University of Science and Technology, 2004.

[19] Chi Zhejie. Research on key techniques of event text analysis using language concept knowledge [D]. Beijing: Institute of Acoustics, Chinese Academy of Sciences, 2017.

[20] Shi Yang. The research on chinese sentence similarity algorithm based on HNC [D]. Zhenjiang: Jiangsu University, 2009.

[21] Dong Caizheng, Liu Baisong. Community question answering-oriented Chinese question classification [J]. Journal of Computer Applications, 2016, 36 (4): 1060-1065.

[22] Zhang Zhihao, Guo Yi, Han Meiqi, et al. Research on short text classification based on keyword similarity [J/OL]. Application Research of Computers: 1-6 [2018-11-22]. https://doi.org/10.19734/j.issn.1001-3695.2018.04.0440.

[23] Onan A, Korukoğlu S, Bulut H. Ensemble of keyword extraction methods and classifiers in text classification [J]. Expert Systems with Applications, 2016, 57 (9): 232-247.

[24] Gao Chao, Yang Sichun, Wan Jiashan. Chinese question classification based on fusion category clue words [J]. Journal of Suzhou University of Science and Technology: Natural Science, 2018 35 (2): 73-78.

[25] Yu Bengong, Xu Qingtang, Zhang Peihang. Question classification based on MAC-LSTM [J/OL]. Application Research of Computers: 1-6 [2018-11-22]. https://doi.org/10.19734/j.issn.1001-3695.2018.05.0452.

[26] Wen Xu, Zhang Yu, Liu Ting, et al. Syntactic structure parsing based chinese question classification [J]. Journal of Chinese Information Processing, 2006, 20 (2): 33-39.