Scientific Research Publishing

Journal of Computer and Communications, 2021, 9, 99-107
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Method of English Test Knowledge Graph Construction

Yuan Sun[1,2]*, Jiayi Tang[1,2], Zhen Zhu[1,2]

[1]School of Information Engineering, Minzu University of China, Beijing, China
[2]Minority Languages Branch, National Language Resource and Monitoring Research Center, Beijing, China
Email: *tracy.yuan.sun@gmail.com

How to cite this paper: Sun, Y., Tang, J.Y. and Zhu, Z. (2021) A Method of English Test Knowledge Graph Construction. *Journal of Computer and Communications*, **9**, 99-107.
https://doi.org/10.4236/jcc.2021.99007

Received: April 12, 2021
Accepted: September 27, 2021
Published: September 30, 2021

Abstract

English is one of the key subjects of basic education in many countries; more and more students tend to learn English online. This paper takes middle school English texts as the research object and proposes a method of English test knowledge graph construction. Through acquiring the data, preprocessing the corpus, and designing feature vectors, this paper realizes to extract the English knowledge points from the tests based on SVM model and construct an English test knowledge graph. It is important to the standardization, automation and systematization of online education learning.

Keywords

English Tests, Knowledge Graph, Knowledge Extraction

1. Introduction

As a new normal of education, online education is changing the traditional teaching mode. The number of online education users in China has reached 342 million in December 2020, what is more the number of mobile online education users has reached 341 million. As an important part of computer-assisted instruction, online exercise practice system is one of the important areas of modern educational technology research.

Knowledge Graph is a knowledge carrier to describe things and their relationships in the real world. The application of knowledge graph in education is a hot topic. Based on the knowledge of the course, the knowledge graph forms a complete knowledge network graph through the dependence and association characteristics of the knowledge. Focus on the mastery of knowledge in real time, we can recommend the exercise corresponding to weak knowledge.

The extraction of knowledge in questions is an important part of the implementation of the question selection system, which directly determines the degree of system automation. This paper applies natural language processing technology to middle school English questions to complete the extraction of knowledge in the questions. It is of great significance for the modernization and development of English education in junior high schools. To sum up, the main contributions of this paper are as follows.

1) After the corpus preprocessing is completed, this paper uses a supervised feature-based SVM classification method to extract English knowledge.

2) Constructed the knowledge graph of English tests. Visually display the knowledge graph of the constructed knowledge.

2. Related Works

So far in the field of education, there has been a lot of research on the test paper generation algorithm in the test management system at home and abroad. In composing test paper, some scholars use the method of fuzzy reasoning to design the level evaluation algorithm [1].

Some scholars define the learning situation of learners as a four-element model. This algorithm can obtain the relationship between knowledge according to the most recent learning information of learners, but it does not take the overall situation and historical data of students into comprehensive consideration to make title recommendation [2]. In addition, Li Haiyan *et al.* proposed the application of uncertain reasoning algorithm in student model by using concept map model [3]. Furthermore, Yang Qinglin took the task of automatic test paper generation as an objective optimization problem with restrictive conditions [4]. Using the genetic algorithm to generate test paper, which improves the coding method and fitness function [5].

Although the above methods can make students practice the wrong questions repeatedly, they don't fully consider the overall learning situation of students and pay attention to the knowledge behind the wrong questions.

In order to achieve information extraction on large data sets, many scholars have introduced various machine learning algorithm, the most widely used one is the support vector algorithm [6] [7]. As a branch of information extraction, in relation to extraction tasks, people usually use feature vector-based methods and introduce statistical methods [8]. Among the statistical methods based on feature vectors, the most classic are the maximum entropy model [9] and support vector machines [10]. At the same time, as the key link of these methods, feature vector construction, many methods have appeared.

Miller *et al.* adds multiple types of semantic information related to entity relationships, proposed a tree model of dependency syntax. Including lexical analysis results and syntactic analysis results and some language rules, these semantic relationship features provide a basis for relationship extraction [11]. In addition, Culotta and others constructed the kernel function based on the dependency tree, and used it as the input of the machine learning algorithm [12].

At present, there are two kinds of mainstream classification organization structure. One is the one-to-k method. When K knowledge categories need to be divided, only K classifiers need to be constructed. However, the effect of this classifier is not good. When the individual classifier makes a wrong judgment, it will seriously affect the overall judgment effect. Another common classifier is a 1-to-1 discrimination method. Similarly, when K knowledge need to be divided, the number of classifications that need to be constructed can reached $K(K-1)/2$. When these judgments need to be compared, they need to be calculated as a whole. That is $K(K-1)/2$ judgment calculations. And then, by adding up the weights, the knowledge with the highest accumulated score is the final category. Although this method performs better than the former, the size of the classifier increases dramatically with the increase of the number of knowledge. In the knowledge extraction task, the applicability of this kind of method is insufficient.

3. Framework

The knowledge extraction framework for junior high school English text proposed in this article is shown in **Figure 1**.

4. Corpus Construction

There are two ways to generate corpus: English electronic documents provided by schools and free exercise resources on the Internet.

4.1. Corpus from English Exercise

The corpus is stored in MS Word format, including English exercise stems, options and analysis. This paper designs and implements a set of English text analysis tools. Extracting the question type, question stem, options, correct answer, and question analysis, formatted and stored in the local database (**Figure 2**).

Under the question number, the stem and option content can be extracted respectively, and the standard answer value can be extracted from the reference

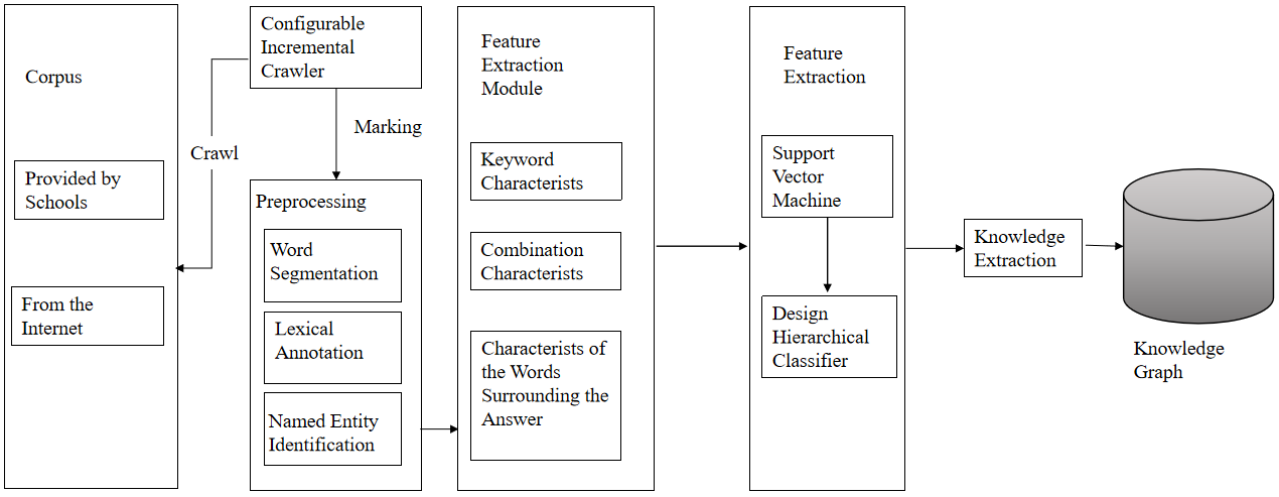

Figure 1. The framework for knowledge extraction.

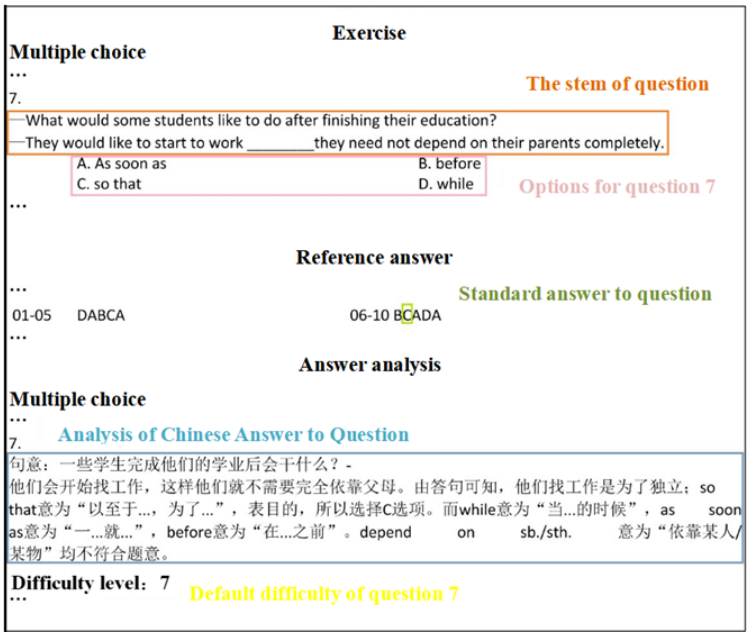

Figure 2. Schematic diagram of electronic document structured information analysis.

answer. Finally, obtain the analysis content and difficulty value from the answer analysis, which is saved in the database in structured form. This paper uses the open source POI toolkit to achieve extraction, writes a batch program to parse MSWord files, and uses regular expressions to parse out the key information in the text. The key matching rules are described in **Table 1**.

4.2. Data Extraction

In order to expand the scale of the knowledge graph and adaptively update it, this article first selects 7 websites with abundant free question corpus resources as the source of expanded knowledge graph data. Firstly, getting the text through the web crawler tool, analyze the HTML page of the topic in the website and extract the structured data of the junior high school English texts. Then save the configuration information in a file through json format. Using jsoup open source package to parse the HTML code on the page, different website configurations have different templates. Finally, the title and part of the title attribute data are extracted.

This article first uses Mongodb as the storage medium for the extraction results, and then associates them with the graph database after analyzing the knowledge information.

5. Preprocessing

The description of the structured information data is not completely enough to be directly used for topic recommendation of knowledge graph. Therefore, we need to further classify the topic attributes by knowledge. Before that, we need to preprocess the corpus. In this paper, we use word level processing tools to preprocess the topic entity object, such as word segmentation, part of speech

Table 1. Structured matching rules for text line questions.

Extract Content	Extraction Rule Description
Question Type	Use the list of question types to obtain by comparing strings.
Question	In two different categories, the question number is obtained by matching the question number, such as "[0-9]*.", and the content between the two question numbers is used as the whole question.
Stem	Get the content between the question number and the options, and remove the spaces.
Option	Get the content between option tags as an option and return it in the form of a list.
Analysis	Match the question number, and then parse the tag according to the answer to get the entire analysis content.
Question Difficulty	Get the value through the difficulty label behind "Analysis".

tagging, and named entity recognition and so on. The Stanford CorenNLP word segmentation tool is used for English texts, and the Hannlp tool is used for Chinese processing.

The first step of preprocessing is word segmentation. The corpus is manually checked to ensure the accuracy of the next round of work. Characters are the most basic unit of Chinese, and each character has an independent code. Single characters form words and words form sentences, with a period as the ending symbol. Unlike Chinese, the smallest unit of the English language is the word itself, and words are separated by spaces, so there is no need for word segmentation.

On the basis of word segmentation, it is also necessary to tag the words in the title. In this paper, the Chinese language processing uses the "ICTPOS3.0 Chinese Part-of-Speech Tag Set" tag set of the Institute of Computing Technology of the Chinese Academy of Sciences, and the "Penn Treebank" tagging system of the University of Pennsylvania [13] is used for English processing. Lexical annotation refers to the part-of-speech markers that mark each word, such as nouns, verbs, adjectives, etc. In order to ensure the accuracy of the corpus, manual proofreading is also carried out in this paper.

The recognition of named entities plays a very important role in the recognition of topic attributes. This article mainly focuses on the names of people, places, and organizations, recognize the named entities in the question stem and answer analysis respectively. Because the accuracy of entity relationship recognition is not high enough, this paper also uses manual proofreading.

Since the supervised SVM method is used as the knowledge classification algorithm, it is necessary to prepare a certain amount of topic data with classification labels as training corpus. According to the training corpus, part of the data with knowledge labels is directly used as the label of the training corpus, but a large number of corpora do not have knowledge labels, so manual intervention is needed to label the classification labels. Example 1 is preprocessing results.

In this paper, the preprocessing adopts Stanford's Core NLP program. Chi-

nese adopts the hannlp word-level processing toolkit, and the labeling system adopts the "ICTPOS3.0 Chinese Part-of-Speech Tag Set" labeling set, and the answer is analyzed and labeled as **Figure 3**. Separate the original word from the label by the "/" symbol.

6. Knowledge Extraction Based on the SVM Method

After the corpus preprocessing is completed, this paper needs to perform automatic topic knowledge extraction on a large corpus. In this paper, keyword features, combined features based on multi class word-level labeling and word labeling features around the correct option are respectively used in combination with the features of middle school English texts and knowledge.

For some linearly indivisible sample data, SVM transforms the original sample into a linearly divisible problem by mapping the original sample to a higher dimensional mapping space through the dimensional enhancement technique. At the same time, in order to avoid the "curse of dimensionality" caused by the rapid expansion of the computational scale due to dimensionality enhancement, the kernel function method is introduced to keep the computational process in the low-dimensional space. The extraction problem discussed in this paper is a multi-classification problem.

SVM was first used to solve the binary classification problem, but knowledge extraction is a complex multi classification problem, and the knowledge are divided into 10 categories according to the teaching requirements in this paper. In this paper, a knowledge hierarchical classifier structure with fast track is adopted.

This paper divides the whole hierarchical classifier into 4 layers. The first layer is mainly used to classify the major categories according to the word knowledge, phrase knowledge and sentence knowledge. Since the classification belongs to the triple classification problem, the 1-to-1 combination is used in a single classifier so as to ensure the accuracy of the classification. The maximum depth of the classifier is 4 layers, and the total number of classifiers using this combined structure is only 11. While the number of classifiers constructed using the 1-to-1 approach alone would be 45. While greatly reducing the number of classifiers, since each classifier has a different task, targeted feature vectors can be designed according to the characteristics of the specific pre-classified category. It can greatly reduce the difficulties caused by the need to select generic feature vectors.

[句意/n, ：/w, 一些/m, 学生/n, 完成/v, 他们/r, 的/uj, 学业/n, 后/f, 会/v, 干什么/v, ？/w, -/nx, 他们/r, 会/v, 开始/v, 找/v, 工作/vn, ，/w, 这样/r, 他们/r, 就/d, 不/d, 需要/v, 完全/ad, 依靠/v, 父母/n, 。/w, 由/p, 答/v, 句/q, 可知/v, ，/w, 他们/r, 找/v, 工作/vn, 是/v, 为了/p, 独立/v, ；/w, so/nx, /w, that/nx, 意/ng, 为/p, “/w, 以至于/c, .../w, ，/w, 为了/p, .../w, ”/w, ，/w, 表/n, 目的/n, ，/w, 所以/c, 选择/v, C/nx, 选项/n, 。/w, 而/c, while/nx, 意/ng, 为/p, “/w, 当/p, .../w, 的/uj, 时候/n, ”/w, ，/w, as/nx, /w, soon/nx, /w, as/nx, 意/ng, 为/p, “/w, 一./m, ../w, 就/d, .../w, ”/w, ，/w, before/nx, 意/ng, 为/p, “/w, 在/p, .../w, 之前/f, ”/w, 。/w, depend/nx, /w, on/nx, /w, sb/nx, .//w, sth/nx, . /w, 意/ng, 为/p, “/w, 依靠/v, 某人/r, //w, 某/r, 物/ng, ”/w, 均/d, 不/d, 符合/v, 题意/n, 。/w]

Figure 3. Preprocessing result of analysis.

At the same time, the fast-pass approach adopted in this paper can greatly increase the classification speed and improve the relevance of the classifier to a certain extent.

In order to train the knowledge prediction model and test the model effect, we manually label 3000 questions with knowledge labels, which are divided into ten knowledge categories. They are verb phrase (vp.), noun (n.), adjective (adj.), verb (v.), prepositional phrases (pp.), past simple passive (pasp.), present simple passive (prsp.), attributive clause (ac.), adverb (adv.), and preposition (prep.).

And 2000 of the questions are used as the training data and the remaining 1000 are used as the test data. In order to reduce the impact of errors on the prediction results, the radial basis function (RBF) is used as the kernel function method in this paper, and the experimental results are shown in **Figure 4**.

The P value represents the precision. The R value represents the recall. The F1 value represents the summed average of the correct and recalls rates. P, R and F1 are used to measure the performance of the classifier. The above experimental results show that SVM-based knowledge extraction of middle school English texts has a certain effect.

The Precision values of knowledge extraction in noun and in present simple passive are 76.09%, 74.68%. The highest value of Precision is 76.09% for nouns. The F1 values of knowledge extraction in present simple passive, in noun, and in adjective are 83.1%, 70.95%, 67.68%. The highest value of F1 is 83.1% for present simple passive. The average value of P, R, F1 is 65.54%, 65.11%, 65.07%.

It can be seen that noun extraction and passive voice extraction are more effective, which is considered to be due to the obvious features of this type of questions. The knowledge graph shows the relationship between the preposition knowledge graph and different texts, Different colors represent different texts.

7. Construction of English Test Knowledge Graph

This paper uses the topic knowledge extraction method proposed in this article to classify the topics into 10 main knowledge categories. Take preposition knowledge as an example. The topic exists in the form of a knowledge graph as shown in **Figure 5**.

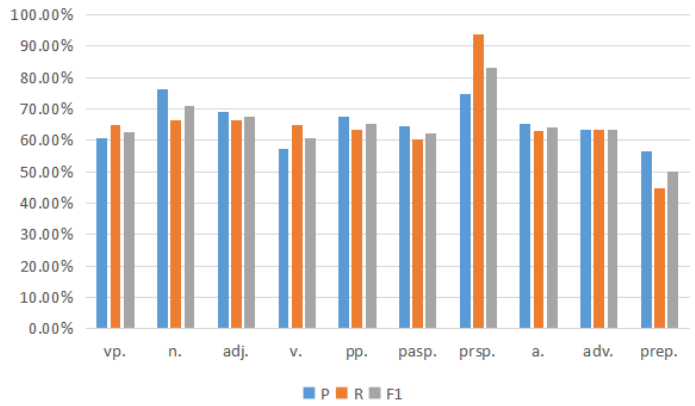

Figure 4. Result of the extraction.

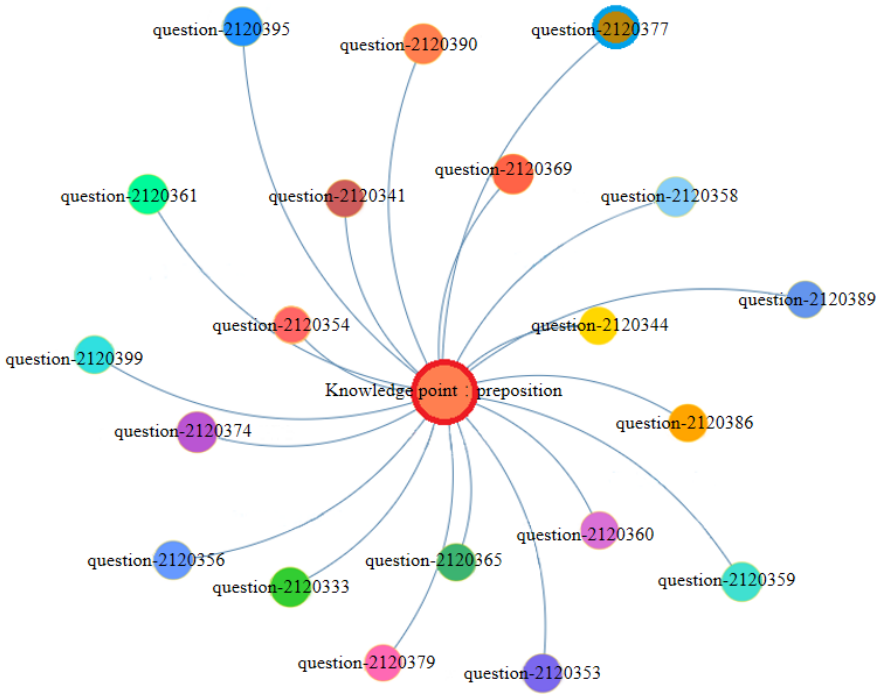

Figure 5. Preposition points knowledge graph.

8. Conclusion

In this paper, we take English texts in junior high school as the research object, introduce the method of collecting topic data, and adopt natural language processing techniques to pre-process the corpus. Then feature vectors are designed by combining the features of English texts and knowledge, and a hierarchical SVM knowledge classifier is designed to classify the knowledge so as to automatically obtain the knowledge information in the topics. The experiments show that SVM-based knowledge extraction of middle school English texts has certain effect, and it has a certain effect on the construction of English knowledge graphs.

Acknowledgements

This work is supported by National Nature Science Foundation (No. 61972436).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Ali, H.A., El-Dousky, A.I. and Ali, A.A.A. (2003) Exploitation of Intelligent Agent for Building Expandable and Flexible Computerized Tutoring System. *International Journal of Computers and Applications*, **25**, 119-129. https://doi.org/10.1080/1206212X.2003.11441693

[2] Ting, L. (2010) The Research about Intelligent Teaching System Based on Mul-

ti-agent Technology. *Computer Programming Skills & Maintenance*, **16**, 62.

[3] L, H.L. and F, H.X. (2004) Intelligent Network Teaching System Based on Multi-Agent. *Journal of University of South China* (*Science and Technology*), **18**, 71-74.

[4] Y, Q.L. (2020) Design and Implementation of Intelligent Paper Organization Examination System Based on Genetic Algorithm. Shandong Normal University.

[5] Qian, F. and Bo, S. (2004) Comparative Study of Knowledge Space Theory and Item Response Theory. *CET China Educational Technology*, No. 5, 75-76.

[6] Z, G.D., Jian, S., Jie, Z., *et al.* (2005) Exploring Various Knowledge in Relation Extraction. *Proceedings of the* 43*rd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 427-434. https://doi.org/10.3115/1219840.1219893

[7] Zelenko, D., Aone, C. and Richardella, A. (2003) Kernel Methods for Relation Extraction. *The Journal of Machine Learning Research*, **3**, 1083-1106.

[8] Ghamrawi, N. and McCallum, A. (2005) Collective Multi-Label Classification. *Proceedings of the* 14*th ACM International Conference on Information and Knowledge Management*, 195-200. https://doi.org/10.1145/1099554.1099591

[9] Banko, M., Cafarella, M.J., Soderland, S., *et al.* (2007) Open Information Extraction for the Web. *IJCAI*, **7**, 2670-2676.

[10] Zhao, S. and Grishman, R. (2005) Extracting Relations with Integrated Information Using Kernel Methods. *Proceedings of the* 43*rd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 419-426. https://doi.org/10.3115/1219840.1219892

[11] Miller, S., Fox, H., Ramshaw, L., *et al.* (2000) A Novel Use of Statistical Parsing to Extract Information from Text. *Proceedings of the* 1*st North American Chapter of the Association for Computational Linguistics conference*, Association for Computational Linguistics, 226-233.

[12] Culotta, A. and Sorensen, J. (2004) Dependency Tree Kernels for Relation Extraction. *Proceedings of the* 42*nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 423. https://doi.org/10.3115/1218955.1219009

[13] Marcus, M.P., Marcinkiewicz, M.A. and Santorini, B. (1993) Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, **19**, 313-330. https://doi.org/10.21236/ADA273556