The Research on the Construction of Primary Mathematics Corpus Based on MATTER Cycle Method

Bo Song, Rui-Fu Wang, and Xiao-Mei Li

***Abstract*—Mathematics is an important subject which is the basis of all sciences, especially primary maths is the most basic and practical. It not only helps students to grow up in the future, but also cultivates students' psychological quality and mental flexibility. The study of mathematical language in primary schools can helps to improve the quality of teaching and deepens students' understanding and application of maths. In view of the computer lack of corpus support in primary mathematics problem solving, this paper analyzes the exam questions, review questions and the relevant knowledge, and constructs the primary mathematics corpus by using MATTER cycle method. This research not only provides support for machine solving, but also provides material for the related research of primary maths knowledge.**

***Index Terms*—Machine learning, mathematical language, corpus, MATTER cycle.**

I. Introduction

With the rapid rise of computer information technology and artificial intelligence, the rigid search function and question answering system have not met the needs of people's knowledge. People hope to communicate with computers through natural language. Since the 1980s, researchers for speech recognition have begun to organize a large number of spoken language data to create language models, using n-grams and hidden Markov models to identify words in the vocabulary based on the transliterated text. Siri in Iphone is an example. People can use ordinary natural language instead of clumsy keyboard input to ask questions, but it can only understand a subset of some key phrases and cannot fully understand natural language.

However, it is not feasible to input a large amount of data to a computer, and it is not feasible to learn to speak. It is necessary to prepare data that is easy for computer discovery mode and reasoning, and then achieve this by adding relevant metadata to the data set [1]. In order to make the algorithm learn more effectively, the annotation on the data must be accurate and related to the task to be executed. Thus, the language annotation is a key link in the development of intelligent human language technology [2]. This paper realizes the construction of primary school math corpus through MATTER development cycle, provides corpus support for machine solving technology, provides learners with a platform to consolidate knowledge points and improve learning efficiency, and provides materials for primary mathematics related knowledge researchers.

Manuscript received October 13, 2019; revised March 1, 2020. This work was supported in part by the U.S. Department of Commerce under Grant BS123456.

Bo Song and Rui-Fu Wang are with Software College of Shenyang Normal University, Shenyang, Liaoning, China (e-mail: songbo63@aliyun.com, 1642465067@qq.com).

Xiao-Mei Li is with Liaoning Basic Education Research and Training Center, Shenyang, Liaoning, China (e-mail: 1715500576@qq.com).

II. Technical Analysis

A. Corpus Research Status

The corpus is the basic resource for corpus linguistics research and the main resource for empirical language research methods. It is used in lexicography, language teaching, traditional language research, and statistical or case-based research in natural language processing. Since Francis developed the world's famous Brown corpus in 1960, the research on corpus has been gradually developed in China. The design and construction of Shanghai Jiaotong University Science and Technology English Corpus, University Learners Spoken English Corpus, Chinese Professional English Learners Corpus and so on [3]. The combination of corpus and computer technology has become an important means of modernization of language research. It is characterized by quantitative investigation + qualitative analysis and interpretation, which embodies the unity of rationalism and empiricism in corpus linguistics. It is the characteristic of contemporary linguistics research.

There are many types of corpora. According to its research purpose and purpose, the corpus is divided into heterogeneous, homogenous, systematic and special [4]. According to the language of the corpus, the corpus can also be divided into monolingual, bilingual and multilingual. According to the collection unit of corpus, the corpus can be divided into discourse, sentence, and phrase. Bilingual and multilingual corpora can be divided into parallel (aligned) corpus and comparative corpus according to the corpus organization form. The former corpus constitutes the translation relationship, which is mostly used in application fields such as machine translation and bilingual dictionary compilation. The latter will express the same content. Different language texts are collected together and used for language comparison studies. The application of the corpus method and the data provided will enable language research to be based on more reliable quantification, so that researchers can avoid subjective speculation on certain language phenomena to a certain extent, and make the research conclusion more objective and credible. In each parallel corpus, the instructional language has always been an essential component. For example, the National Modern Chinese Corpus and the Taiwan Academia Sinica corpus all contain teaching language, but the mathematical corpus

doi: 10.18178/ijiet.2020.10.4.1380

specially established for the study of mathematics teaching has not yet been seen [5]. Therefore, the purpose of this study is to build a primary school mathematical corpus, provide corpus support for machine solving techniques, and provide a platform for mathematics research and learning.

B. MATTER Annotation Cycle

Because the annotation process is not linear, multiple iterations can be required for defining the tasks, annotations, and evaluations, in order to achieve the best results for a particular goal [6].Most of today's annotation work is done by crowdsourcing, which divides the task into small tasks, and then sends it to many people. Each task has a small amount of compensation, which replaces a small part of the high salary labeler to mark the entire corpus. . The most famous Amazon Turkish robot uses this method. The researchers create HIT (Human Intelligence Tasks) and publish them on a bulletin board. Turkers can selectively accept tasks and get a small amount after completing the task [7].

Although the time-saving and low-cost approach of Amazon Turkey robots sounds ideal, the Human Intelligence (HIT) system is not absolutely perfect. First of all, it requires that each annotation task be split into "micro-tasks", but not all annotation tasks are suitable for splitting into micro-tasks, which will weaken the intuition of the labeling target to the labeling target, which may affect the annotation result [8]. Secondly, the data quality problem, it is difficult for researchers to ask the crowdsourced annotation personnel. Because a large part of Turker regards the human intelligence task as the main source of income, the quality of the collected data is uneven, and it is difficult to distinguish between good and bad.

Therefore, this paper uses the MATTER development cycle to realize the construction of the primary school math corpus. The specific steps of this process are: Model, Annotate, Train, Test, Evaluate, Revise. First, create models and specifications for specific language phenomena, and mark the knowledge points and review questions according to the specification. Then use the newly created annotation corpus for machine learning, evaluate the results and modify the models and algorithms. With the "modeling-labeling" and "training-testing" loops, once the model has been added and modified, the MATTER loop will be repeated from the beginning. Although this process is cumbersome and time consuming, it can greatly improve the performance of the algorithm and the accuracy of the data, providing a methodology for creating a gold standard corpus.

C. Development Environment

There are many annotation tools available on the market, such as GATE, Knowtator, MAE, MMAX2, etc. This article uses the MAE (Multipurpose Annotation Environment) to build a corpus. It provides a simple interface that requires very little configuration for easy operation and inspection by users. For the most part, we'll be using XML DTD (Document Type Definition) representations. XML is becoming the standard for representing annotation data, and DTDs are the simplest way to show an overview of the type of information that will be marked up in a document. By having a DTD, the XML in a file can be validated to ensure that the formatting is correct.

MAE's input form is similar to a DTD file. This type of file can specifically describe the name of the task. Elements and attributes. MAE can be used on Windows and UNIX systems. It can be run by Java program. It can be used to mark the file by first loading the definition task file (.dtd) and a file to be marked (.txt or .xml format).

This article uses the audit tool MAI (Multidocument Adjudication Interface) is an auxiliary program designed to audit the output of MAE. Its input form is the separated XML file of the MAE output and the DTD file used to generate the annotation. On Windows and Unix systems, it should be used with the most recent version of Java6 (it must have at least update 14 to run properly on Windows and Unix), though it can also be compiled under Java 5, so it can also be run on older Macs.

III. Construct Corpus

A. Corpus Selection

Since the implementation of the "Full-time Compulsory Education Mathematics Curriculum Standards (Experiment)" in 2001, various versions of primary school mathematics textbooks have emerged in an endless stream. In August 2008, Mr. Sun Xiaotian wrote a review of his work on the construction of mathematics textbooks for primary and secondary schools in China in recent years. The article pointed out: "To date, there have been 6 sets of primary schools, 9 sets of middle schools, and 6 sets of high school 6 sets of 21 sets of new mathematics textbooks compiled according to the requirements of the Standards." [9]. These textbooks belong to 13 publishers.

According to a survey conducted on the Internet on the use of mathematics textbooks in Liaoning, Shandong, Guangdong, and Henan provinces, most of the regions use the People's Education Publishing House and Beijing Normal University Press. Published primary school mathematics textbooks. This paper uses the mathematics textbook of Beijing Normal University to analyze the mathematics language, and downloads the mathematics papers of Xiaoshengchu in the provinces from 2015 to 2018 on the Internet. Take the year as the dividing line, put the type of survey questions into a Word documents for storage.

B. Model and Specification

In the MATTER development cycle, the first step is "phenomenon modeling" ("M" in the MATTER cycle), based on the collected data, to model the annotation task of the primary school mathematical corpus. First of all, how to classify a raw corpus into a detailed knowledge system [10], taking the four arithmetic operations as an example, according to the requirements of the pedagogical syllabus, the terms of the four compositing operations are mined out and organized to form the corresponding EXCEL document. Some are shown in Table I:

Then based on the above knowledge points, this article uses the XML Document Type Definition (DTD) to represent the model. This type of file can clearly describe the name,

elements and attributes of the task. The specific code is as follows:

<!ELEMENT operation (#PCDATA)>

<!ATTLIST operation label (add | sub | multi | divi | mix)>

TABLE I: MATHEMATICAL TERMS

Term	Meaning
Addition	Combine two numbers into one number
Subtraction	Know the sum of two numbers and one of the addends, and find the other addend
Multiplication	A simple operation to find the sum of several identical addends
Division	Knowing the product of two numbers and one of the products, the operation of finding another product
mixing	An equation contains any two or more of the four operations of addition, subtraction, multiplication, and division

C. Annotation and Review

After creating a corpus and model, the actual labeling process ("A" in the MATTER cycle) begins. The raw corpus is labeled according to the annotation model guide. This article uses the Multipurpose Annotatin Environment (MAE) to mark the primary school corpus. This manual annotation tool provides a simple interface that requires very little configuration, and its input form is similar to the DTD file mentioned above.

However, problems may occur when entering information during the annotation process. If the annotation is too tired or the concentration is not enough, the wrong label may be accidentally filled. Therefore, after the labeling is completed, the labeled data is calculated to have an IAA score [11]. If the scores are lower, the model is modified and then relabeled; if the score is ideal, the data can be reviewed to generate gold. Standard corpus, then use it to train and test machine learning algorithms. This phase is called the MAMA (Modeling - Annotation - Modeling - Annotation) loop, as shown in Fig. 1. The key part is to review the annotation results of the annotation personnel and use it to generate a gold standard corpus for machine learning.

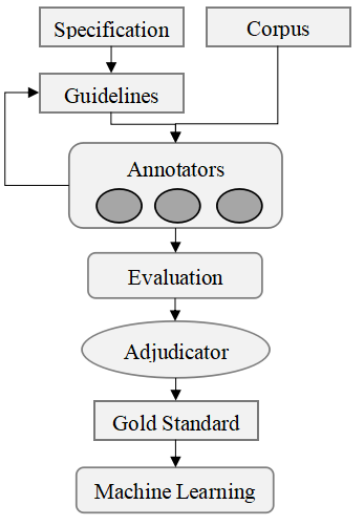

Fig. 1. Annotation cycle.

IV. ANNOTATION PROCESS

A. Machine Learning Algorithm

Creating an annotated corpus and applying it to a suitable machine learning algorithm is a difficult task. This paper uses a supervised learning algorithm to gradually improve the performance of the system through the annotated data, thus automatically classifying and annotate the text (ie MATTER). "T" in the cycle).

Some problems in the primary school mathematics test questions will include the assessment of two types of knowledge points. Ordinary decision tree learning may lead to inaccurate annotation [12]. According to the model and annotation of the created primary school mathematical corpus, this paper chooses to have general Sexual and widely applied Bayesian learning algorithms for automatic annotation. Bayesian is a generative classifier that predicts classification by considering feature probability [13]. Correspondence between conditional probabilities according to Bayes' theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{1}$$

The formula is usually expressed in terms of nominal probabilities such as a priori, likelihood, and posterior, and is re-expressed as:

$$P(C \mid X_{F_1},\ldots,X_{F_N}) = \frac{P(X_{F_1},\ldots,X_{F_N} \mid C)P(C)}{P(X_{F_1},\ldots,X_{F_N})} \tag{2}$$

A common and reasonable step in training the classifier is to assume that the available evidence for the approximation procedure remains the same, so the evidence can be completely ignored and the non-standardized conditional probability can be used for the calculation. If we assume that the conditions between all the features $F1,\ldots,Fn$ are independent of each other, then a simpler formula can be obtained that can be used to calculate the probability estimate, namely:

$$P(C \mid X_{F_1},\ldots,X_{F_n}) \propto P(C)\prod_{i=1}^{n} P(X_{F_1} \mid C) \tag{3}$$

The re-presented learner approximation function can accept the actual calculation of the data set, and a strategy is needed to explain how to compare the hypothesis about class division under given data conditions [14]. Learning algorithms need to consider a number of candidate hypotheses, each of which involves dividing the data into a class from which to choose the one that is most likely, the Maximum A Posteriori (MAP) hypothesis [15]. Using the learner as the Bayesian classifier described above, the maximum posterior probability is returned based on the conditional independence assumption on the feature set:

$$Classify(f_1,\ldots,f_n) = \arg\max_{c \in C} P(C=c)\Pi P(X_{F_1} = f_1 \mid C=c) \tag{4}$$

Which can be regarded as a description, with the target C as a condition, how to generate a random instance of X.

B. Testing and Evaluation

After selecting the features used by the algorithm and algorithm, it is possible to actually test the algorithm with the gold standard corpus and evaluate the test results ("TE" in the MATTER cycle). The most common practice is to divide the corpus into two parts: the development corpus and the test corpus. The development corpus is further divided into two parts: the training set and the development-test set. The training set is used to train the algorithms used in the task, and the development-test set is used for error analysis [16]. After the algorithm training is finished, it can be run on the development-test set. If the algorithm does not correctly mark the corpus, adjust and retrain the algorithm and then test it again, and repeat the above process until a satisfactory result is obtained. After the training is completed, the algorithm will run on the reserved test corpus. These data have never been used in training and development and development tests, and the results of the algorithm on the new data can be obtained. This phase is called the TTER (training-evaluation) cycle, as shown in Fig. 2:

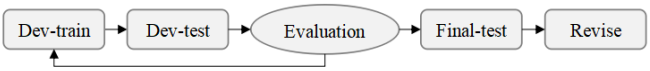

Fig. 2. Training-evaluation cycle.

C. Modification and Summary

In the training, testing and evaluation stages of machine learning, it may be necessary to modify the project all the time, but the adjustment of each step is only for the modification of the current step. Therefore, in this stage, this paper uses a step backward method to examine some “key” items that need to be modified in the project. This includes corpus modification, labeling model and label specification modification and algorithm implementation (ie "R" in the MATTER cycle).

Creating a gold standard corpus and training machine learning is a difficult task, and because many variables affect the outcome of the project, the representation and balance of the corpus is guaranteed at this stage. A corpus is a selective subset of a language, rather than all possible examples of a language, and the proportions of the different types of text it contains should be consistent with evidence-based and intuitive-based judgments.

V. Conclusion

Based on the MATTER annotation development cycle methodology, this paper constructs the primary school mathematical corpus, realizes the automatic annotation of the corpus, and provides the corpus support for the machine solving of the primary school mathematics problem, which helps the quality of primary school mathematics teaching and improves the students' ability. The understanding and application of mathematics, as well as materials for primary mathematics related knowledge researchers.

However, due to limited time and ability, there are still some problems in the corpus. The most prominent point is that compared with other existing corpora, the corpus is far from the level, and will continue to expand the source of corpus in the subsequent research process. The mathematics knowledge system of primary schools is constantly updated and updated. It is also necessary to update and improve existing algorithms to improve the accuracy of automatic labeling.

Acknowledgment

The authors acknowledge — (1) The Basic Research Foundation of Universities in Liaoning (Grant: 2017L317); (2) The 17th Five-Year Plan of Education Science in Liaoning Province. (Grant: JG18DB451).

References

[1] Z. W. Feng, *Concise Course in Natural Language Processing*, Shanghai: Shanghai Foreign Language Education Press, 2012, pp. 240-284.
[2] H. Li, “Deep learning for natural language processing: advantages and challenges,” *National Science Review*, vol. 1, pp. 24-26, May 2018.
[3] C. D. Qi, “Study on mathematical definition based on corpus,” M.S. thesis, Xiamen University, 2009.
[4] J. Zheng, *Principles and Practice of NLP Chinese Natural Language Processing*, Beijing: Publishing House of Electronics Industry, 2017, pp. 311-321.
[5] Z. Z. Zheng, “Mathematics teaching material and corpus construction,” in *Proc. the Second National Educational Textbook Language Symposium, Concise Course in Natural Language Processing*, Shanghai: Shanghai Foreign Language Education Press, 2012, pp. 240-284.
[6] J. Pustejovsk and A. Stubbs, *Natural Language Annotation for Machine Learning*, O’Reilly Media, Inc, 2013, pp. 26-34.
[7] L. J. Zhou and T. Wang, “Amazon Turkey robot: A summary of research on crowdsourcing network platform for scientific research,” *Science and Technology Progress and Countermeasures*, vol. 8, pp. 156-160, January 2014.
[8] G. L. Zhao, Y. W. L. Hu, D. Zhu, S. Y. Zhao, T. Yang, and F. Chen, “The python-based general forum text extraction research,” *Computer Knowledge and Technology*, vol. 24, pp. 259-260, August 2018.
[9] X. T. Sun, “Review of the construction of mathematics teaching materials in primary and secondary schools in China in recent years,” *Journal of Mathematics Education*, vol. 4, pp. 6-10, August 2008.
[10] Z. Zheng, “Study on the construction of elementary mathematics probability and statistical corpus,” M.S. thesis, Huazhong Normal University, 2016.
[11] L. Tang and T. Y. He, “Design and implementation of python-based natural language data processing system,” *Electronic Technology and Software Engineering*, vol. 16, pp. 160-162, August 2018.
[12] S. Wu, “Web data mining and analysis based on python language,” *Computer Knowledge and Technology*, vol. 27, pp. 1-2, September 2018.
[13] S. Bird, E. Llein, and E. Loper, *Natural Language Processing with Python*, O’Reilly Media, Inc., 2012, pp. 263-271.
[14] D. S. Peng, “Several problems in natural language processing of mathematical academic documents,” M.S. thesis, Jilin University, 2018.
[15] D. M. Bikel, and I. Zitouni, *Multilingual Natural Processing Applications: From Theory to Practice*, Pearson Education, Inc., 2013, pp. 241-251.
[16] R. F. Liao and H. Liao, “An NLP-based robot query system,” *Computer Knowledge and Technology*, vol. 21, pp. 97-98, July 2018.

Bo Song obtained the master degree of technology of education from Fukuoka Education University in Japan in 1999. Since 2003, he has been a teacher in Software College of Shenyang Normal University in China and was a professor in 2011. His research interests include software engineering, NLP and deep-learning.

Rui-Fu Wang obtained the bachelor degree of software engineering from Shenyang Normal University in China in 2017. Since 2017, she is studying for a master degree at Shenyang Normal University in China. Her research interests include software engineering and NLP.

Xiao-Mei Li obtained the master degree of education from Shenyang Normal University in China in 2017. Since 2000, she has been a teacher in Liaoning Research and Training Centre for basic education in China and was a professor in 2011. She won the second-class and third-class of natural science achievements in Liaoning Province in China in 2014 and 2017 respectively. Her research interests mathematics education in primary schools and intelligent education information.