# Extraction of Implicit Quantity Relations for Arithmetic Word Problems in Chinese

Xinguo Yu, Pengpeng Jian
National Engineering Research Center for E-Learning
Central China Normal University, Wuhan, China
xgyu@mail.ccnu.edu.cn, jianpengpeng@mails.ccnu.edu.cn

Mingshu Wang, Shuang Wu
National Engineering Research Center for E-Learning
Central China Normal University, Wuhan, China
{mingshuwang, shuangwu}@mails.ccnu.edu.cn

***Abstract*—The extraction of implicit quantity relations is one of critical steps of building the machine solver for solving arithmetic word problems. Extracting implicit quantity relations is a challenge research problem as partial problem of natural language understanding, which has no mature solution method. To address this challenge, the existing work in the extraction of quantity relations mainly relies on creating problem solving framework and extracting system of linear equations. However, these existing methods can solve the relatively simple problems due to they have no methods to extract implicit quality relations. This paper proposes a novel method for extracting implicit quantity relations, which is achieved by the process of Chinese phrase parse, classification and instantiation method of required general implicit quantity relations with semantic models. A set of experiments were conducted to classify the arithmetic word problems, and an example was taken to illustrate the proposed method of identifying general implicit quantity relations. The experimental results demonstrate that the proposed method is effective in extracting implicit quantity relations.**

***Keyword-implicit quantity relation, classification, semantic models***

## I. INTRODUCTION

Extraction of implicit quantity relations is a challenge and interesting research problem because it is a critical step in understanding arithmetic word problems. Developing machine solver has been being active research topic since 60s of the last century and recently it has become a hot topic under the influence of fast advance of artificial intelligence (AI), multimedia, machine learning, and pattern recognition. Facing this challenge problem, this paper proposes a novel method for extracting implicit quantity relations and it takes arithmetic word problems in Chinese as example to present and to illustrate the proposed method. Based on this method a procedure is developed to produce the instantiated implicit relations. The proposed procedure consists of three main steps: Chinese phrase parse, SVM classification and instantiation method of required general implicit quantity relations with semantic models. Chinese phrase parse is the first step in parsing arithmetic word problems. A list of semantic models is prepared for different kinds of implicit quantity relations. In order to bridge the matching relationship between arithmetic word problems and semantic models, the problems classification is achieved by SVM. The feature input to SVM is a bag of words extracted from the given problem for recognizing its class. Once a semantic model matched with a portion of the text of the given problem, a corresponding equation will be added into the set of quality relations to represent the given problem.

Developing machine solver is an outstanding representative research problem in artificial intelligence and education information technology, which is an application -oriented problem. This problem relates with several AI research problems such as natural language understanding, theorem proving, machine inference and it has a good potential of application in building the intelligent tutoring system. The problem understanding provides core technology for intelligent tutoring system, which aims to automatically extract quality relations by combining semantic model and machine learning [1,2]. There are two approaches in problem understanding. The first approach aims to produce an equivalent expression for a given problem in term of finding its solution. And the target expression is required to be easily understood and operated by computer. The first achievement is to adopt a polynomial group to represent a plane geometry theorem. And the variable elimination method is used to judge whether the polynomial group is no contradiction, therefore to know whether the given theorem is correct or not [3]. Another group of works employ a set of primitive geometry relations to represent a plane geometry theorem and use point-eliminating method to infer the set of primitive geometry relations [4]. The given theorem is proved if the inference process builds an inference chain from the known relations to the target relation. MIT artificial intelligence laboratory introduced the technology of machine learning to extract system of linear equations from the text of problems [5].

Another approach of problem understanding needs to extract problem solving information from the text of problems, and confirming a problem solving plan. In addition, more information should be extracted in the process of problem solving. This approach was firstly proposed by W.Kinsch, who proposed a framework of identifying the type of problems and extracting the knowledge frame mainly using for arithmetic word problems solving by machine [6]. Ma expanded W.Kinsch's knowledge frame, and achieved automatic solving for elementary school arithmetic word problems based on the cognitive model [7]. The group of Mohammad used the verb classification and the framework of problem solving process to solve arithmetic word problems [8]. Directly-stated method expresses a way of solving the given problems substituted for the system of high-order equations, which is formed by extracting quantity relations from math problems [9]. However, due to the difference in extract approach and framework of math problem solving, there exists two shortages in above-mentioned approach in the understanding of math problems. One is lacking of the function for the extraction of implicit quantity relations and the other is its limited ability of understanding arithmetic word problems is insufficient. So it is necessary to introduce more new methods for elevating the understanding ability of solving information.

This work in partially supported by National Key Technology Research and Development Program (NO.2014BAH22F01).

978-1-5090-6138-9/16 $31.00 © 2016 IEEE
DOI 10.1109/EITT.2016.55

The remainder of this paper is organized as follows. Section II describes the extraction of implicit quantity relations. Section III presents the experimental results. Finally, we conclude this paper in Section IV.

## II. METHOD AND PROCEDURE FOR EXTRACTING IMPLICIT QUANTITY RELATION

In this section, we present the model formulation of extracting implicit relations on elementary school arithmetic word problems. The general idea of the method is that our semantic model *T = (K, R, M, L)* contains four parameters. List *L* is the classification of arithmetic word problems. When having a parameter *L*, system will compare parameter *K* (keywords group) of semantic models with the part of speech in the problem. If one semantic model is matched, the quantity relation *R* will be used to construct equations. At last, utilizing map list *M* (a map list between *K* and *R*) to reconstruct the mapping relations between equations and entity variables. In the subsection *D*, an instantiation of the method is used to identify general implicit quantity relations.

### *A. Normalized Common Units*

The model formulation of extracting implicit quantity relations consists of Chinese phrase parse, SVM classification and instantiation method of required general implicit quantity relations. The instantiated method includes matching semantic models and constructing equations. The main modules of the proposed procedure for extracting implicit quantity relations are given in Figure 1. Before the subsection of Chinese phrase parse, normalized common units of arithmetic word problems is processed, which unified unit of length, unit of area, unit of volume, unit of capacity, unit of weight, unit of speed, unit of time and order of magnitude.

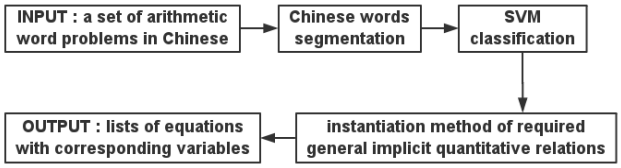


Figure 1. Main modules of the proposed procedure for extracting implicit quantity relations.

### *B. Chinese Phrase Parse*

Because of the meaning of Chinese phrase is unknown in arithmetic word problems for machine, the first procedure of data preprocessing is Chinese phrase parse. The structure of expressing a quality relation in arithmetic word problems consists of Chinese phrases (e.g., noun, verb, adjective, adverb, numeral and so on), so that the text of the given problem is required to be divided into meaningful phrases. A system of Chinese phrase parse is employed for Chinese text parse, which is called Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS), designed by The Chinese Academy of Science. In this step, given *S* as a set of arithmetic word problems. When the text *S* of a given arithmetic word problem is input into ICTCLAS, the output is the text *S′*. *S′* is the annotated form of *S,* which annotates the start and end character of phrases, part-of-speech of phrases.

| problem name | keywords | quantity |
|---|---|---|
| route | speed;distance;meet;depart;respective;route;kilometer;minute;time;away from;at the same time;round;two places;whole journey;hour;second | 58 |
| planting tree | plant;tree;quantifier for | 22 |
| circle | circle;diameter;radius | 33 |
| cuboid | cuboid;length;width;height | 30 |
| cube | cube;edge | 8 |
| triangle | triangle;side;bottom;height | 6 |
| trapezoid | trapezoid;upper base;bottom | 2 |
| parallelogram | parallelogram;side;bottom;height | 4 |
| rectangle | rectangle | 15 |
| square | square;side | 6 |
| percentage | %;fraction | 18 |
| average number | average | 39 |
| chicken&rabbit | leg | 5 |
| time | hour;minute;day;time | 19 |
| ratio | : | 18 |

TABLE 1: KEYWORDS OF IMPLICIT PROBLEMS TYPES

### *C. SVM Classification*

In this subsection, SVM classification presents the setting of keywords for implicit types of problems and the SVM algorithm which is used in the model.

(1) Keywords for implicit problem types

In this part, we summarized fifteen types of implicit quantity relations, and each type owned a set of keywords in Table 1. The keywords of each implication relation are obtained by multiple cross experiments. Besides, these keywords are used for SVM classification.

(2) SVM algorithm based on slack variable

In order to deal with situations of linear or nonlinear classification, tolerating noise and outliers, we introduce the slack variable for SVM classification algorithm. The purpose of algorithm is to get the hyper plane and the maximum margin in expression. By using Lagrange's transformation and Sequential Minimal Optimization, the target function is finally simplified as formula (1):

$$\max_{\alpha} \sum_{i=1}^{n} \alpha - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle L(x_i), L(x_j)\rangle \quad (1)$$
$$s.t.\ , 0 \le \alpha_i \le C, i = 1,2 \ldots, n, \sum_{i=1}^{n} \alpha_i y_i = 0$$

where $\alpha$ is the Lagrange multiplier, C is the weight number for slack variable and C is used to control the weight between finding hyper plane of the maximum margin and ensuring the minimum data point deviation in target function. With the classification of SVM, 283 arithmetic word problems were assigned into fifteen types of implicit quantity relations.

### *D. Instantiation Method of Required General Implicit Quantity Relations*

In this subsection, we present the solution to the extraction of implicit quantity relations by using semantic models. The implementation of extraction model is constructed by data mapping between classifications and semantic models. In order to denote each layer mapping relationship, a semantic model definition can be firstly formulated as *T = (K, R, M, L)*.

*K*: the keywords string of a semantic model.

*R*: the quantity relation that is needed to be extracted from the sentence in the arithmetic word problems. *R* is a math expression which is comprised of addition, subtraction, multiplication, division, and lower-case letters such as a, b, c ....., and they are formed in logical ways.

*M*: a list of matching variables between *K* and *R*, $M = \{\alpha_i, i = 1,2,3 \dots \dots n\}$ and *M* points out which entity information has to allocate a variable and participate in the process of constructing an equation.

*L*: the classification list of arithmetic word problems, and *L* belongs to a collection.

In order to understand the sentences of math word problems, semantic models should be constructed by present math word problems. The semantic model can be formulated as *T = (K, R, M, L)*, which includes model keywords group and corresponding quantity relations expression. When getting the keywords combination (*K*) from sub-sentence of math word problems, the extracted model will compare *K* with semantic models. Once there is a corresponding semantic model matches the *K*, a new map relation is constructed. The semantic model includes a series of implicit quantity relation models, which is identified by parameter *L* in the semantic model function. When a math word problem is classified by SVM, a classification list *L* will be obtained, and the extraction model will firstly compare *K* with semantic models according to the classification list *L*, and if there is no mapping result, the extraction model will compare with other semantic models without limited by parameter *L*. By using the same manner to deal with the whole sentence of a math word problem, some equations will be formed, and at last they form a system of equations. So far, a math word problem in Chinese is converted into a system of equations, and with the sentence of an arithmetic word problem can be converted into a system of equations, and each equation is maybe a high-order or a linear equation.

Algorithm of extracting quantity relations:

1 begin
2 define *T*: the library of semantic models
3 define *l* : the number of $T_j$
4 define *S*: a set of Chinese phrase parse
5 define $S_i$: sentence separated by punctuation
6 define *n* : the number of $S_i$
7 for(*i* in *n*)
8 for( *j* in *l*)
9 if sequence alignment $(S_i, S_i.K, S_i.L, T_j)$
10 else sequence alignment $(S_i, S_i.K, T_j)$
11 if $S_i.K$ is in library $T_j$
12 extract entity information from $S_i$ according to $T_jM$
13 allocate variables for entity information
14 replace corresponding elements by variables
15 construct an equation by $T_jR$
16 end if
17 end if
18 end for
19 end for
20 end

In step 9 above, the sequence alignment compares two sequences limited by keywords combination and classification list, and if condition is invalid, the sequence alignment of step 10 will be executed.

| An example of our method to extract quantity relations in arithmetic word problem in Chinese | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Arithmetic word problem | | 一块三角形的玻璃，它的底是1.25米，高是0.78米。这块玻璃的面积是多少平方米? | | | | | | | |
| Normalize common units | | None | | | | | | | |
| Chinese phrase parse | | 一/m 块/q 三角形/n 的/ude1 玻璃/n，/wd 它/rr 的/ude1 底/f 是/vshi1.25/m 米/q，/wd 高/a 是/vshi0.78/m 米/q。/wj 这/rzv 块/q 玻璃/n 的/ude1 面积/n 是/vshi 多少/ry 平方米/q？/ww | | | | | | | |
| SVM classification | | *L* = 6 | | | | | | | |
| Sentence separated by punctuation | | $S_1$ | | $S_2$ | | $S_3$ | | $S_4$ | |
| | | 一/m 块/q 三角形/n 的/ude1 玻璃/n，/wd | | 它 /rr 的 /ude1 底 /f 是/vshi1.25 /m 米/q，/wd | | 高/a 是/vshi0.78/m 米/q。/wj | | 这/rzv 块/q 玻璃/n 的/ude1 面积/n 是/vshi 多少/ry 平方米/q？/ww | |
| Semantic models | *T=(K, R, M , L)* | (mnn, c=a+b, m n n, 0) | | (f 是mq, a=b*c, f m q, 0) | | (a 是mq, a=b*c, a m q, 0) | | (面积ryq, c=a*b/2, ry q, 6) | |
| | *L* | 0 | | 0 | | 0 | | 6 | |
| | *K* | mnn | | f 是 | | a 是 mq | | 面积ryq | |
| | *R* | c=a+b | | a=b* | | a=b*c | | c=a*b/2 | |
| | *M* | m n n | | f m q | | a m q | | 面积 f a | |
| construct equations | equation | c=a+b => C=A+B | | a=b*c => C=1.25*D | | a=b*c => E=0.78*D | | c=f*a/2 => F=C*E/2 D=1 | |
| | map list | 三角形 | 玻璃 | 底 | 米 | 高 | | 多少 | |
| | | A | B | C | D | E | | F | |

TABLE 2: AN EXAMPLE OF OUR METHOD TO EXTRACT QUANTITY RELATIONS IN ARITHMETIC WORD PROBLEM IN CHINESE

### E. An Example

In this subsection, there is an example to illustrate the method of identified general implicit quantity relations (e.g., implicit or directly-state) in an arithmetic word problem in Table 2.

## III. EXPERIMENTS

In this section, we conduct experiments for SVM classification on elementary school arithmetic word problems and instantiating a method of identified general implicit quantity relations with an arithmetic word problem.

### A. Data Set

In SVM classification, the training data set were provided by Suzhou Education Publishing House, which contained different types of arithmetic word problems, and the arithmetic word problems were divided into 15 types. The targeting classification dataset were provided by Elementary school arithmetic application problem, People's Education Press, 2011(in Chinese), which totally contained 627 arithmetic word problems. The arithmetic word problem example of instantiating method was selected from the targeting data set.

### B. Classification Result

In part C of section II, the detail implementation of SVM classification is shown with data set and the proposed method is feasible. Here, the result of selected 283 samples in arithmetic word problems is shown in Table 1. From Table 1, the first column is the 15 types of arithmetic word problems, and the quantity of each classification is shown in the last column. Table 2 shows an example of our method to extract quantity relations, the left column of which are the steps of the method in sequence, and the right column shows the each corresponding step of data transformation. At last, the map list between variables and entity information in Chinese is obtained and the equations of the arithmetic word problem example are constructed, in which implicit quantity relations are extracted in equations formation. So far SVM classifiers have not achieved a 100% of accuracy, although it makes correct classification for most of the data. SVM cannot identify the implicit for several cases in the process of identifying implicit quantity relations. The next subsection explains the reasons for classification errors and failure of extracting relations.

### C. Result Analysis

There are three reasons for the rest of classification errors in arithmetic word problems. Firstly, the keywords of each problems type have not yet attained the superior. Secondly, the training data chosen from Suzhou Education Publishing House are not perfect and all-around at present. Thirdly, some arithmetic word problems belong to multiple types, called cross-type problem. To address the reasons for the failure of extracting relations, one is lacking of enough semantic models in identifying implicit quantity relations, and the other is lacking uniqueness for every semantic model, which means one keywords group of semantic model maybe map two or more quantity relations. These reasons listed above lead to several errors in classification and extracting implicit quantity relations.

## IV. CONCLUSIONS AND FUTURE WORK

This paper has presented a novel method of extracting implicit quantity relations for the understanding of arithmetic word problems. The semantic model is used to extract the implicit quantity relations in arithmetic word problems. The method is achieved by the process of Chinese phrase parse, SVM classification and instantiation method of required general implicit quantity relations with semantic models. To evaluate the effectiveness of extracting relations, we conduct a set of experiments with SVM classification and taking an example to illustrate the method of identifying general implicit quantity relations. The experimental results show that the proposed method is effective for arithmetic word problems classification and implicit relations extraction.

In the future, we will study how to extend more effective semantic models of extracting implicit quantity relations and exploring the deep learning algorithms in the application of the implicit relationship extraction.

## Acknowledgment

National Key Technology Research and Development Program (NO.2014BAH22F01).

## References

[1] R. Battiti and M. Brunato, The Lion Way: Machine Learning plus Intelligent Optimization, version 2.0, 2015.

[2] P. Sermanet, K. Kavukcuoglu and Y. LeCun: EBLearn: Open-source energy-based learning in C++, IEEE Int'l Conference on Tools with Artificial Intelligence, 2009, pp. 693-697.

[3] Wenjun Wu, Elementary Geometry Decision Problem and Mechanized Demonstration [J]. Chinese Science (A), 1977, 6:507-516.

[4] J.Z. Zhang, S.C. Chou, X.S. Gao. Automated Production of Tradetional Proofs for Theorems in Euclidean Geometry. The Hilbert Intersection Point Theorems. Dept. of Computer Science, WSU, Tech Rep: TR-92-3, 1992.

[5] N.Kushman, Y. Artzi, L. ZettlemoyerandR. Barzilay. Learning to Automatically Solve Algebra Word Problems. Association for Computational Linguistics (ACL), 2014.

[6] W. Kinsch, J.G.Greeno. Understanding and Solving Word Arithmetic Problems, Psychology Review, 1985.

[7] Yuhui Ma, The Research for Automatic Solving for Arithmetic Application Problems Based on The Cognitive Model[J]. Central Compilation & Translation Press, 2012, 4.

[8] M. J. Hosseini, H. Hajishirzi,O. Etzioni, N. Kushman, Learning to Solve Arithmetic Word Problems with Verb Categorization. The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[9] X. Yu, Mingshui Wang. Solving Directly-stated Arithmetic Word Problems in Chinese, October 16-18, 2015, Wuhan, China, Int'l Conference of Educational Innovation through Technology.