A Short Answer Grading System in Chinese by Support Vector Approach

Shih-Hung Wu, Wen-Feng Shih
Dept. of CSIE, Chaoyang University of Technology
168, Jifeng E.Rd. Wufeng District,
Taichung, 41349, Taiwan (R.O.C)
shwu@cyut.edu.tw, wu0fu491@gmail.com

Abstract

In this paper, we report a short answer grading system in Chinese. We build a system based on standard machine learning approaches and test it with translated corpus from two publicly available corpus in English. The experiment results show similar results on two different corpus as in English.

1 Introduction

To assess the learning outcomes of students with tests in various question types and grading methods, short answer question is one type of test that can test the level of students' understanding of specific concepts in a subject domain. Since grading short answer question requires natural language understanding, the test was manually graded by teachers.

Although technically similar to automatic essay grading, automatic short answer grading is not as mature as automatic essay grading. (Burrows et al., 2015) gives a survey on how the automatic short answer grading is dealt by various researchers. The traditional approach is string matching, which could be very efficient but not very effective.

Early work relied on regular expression patterns which were manually extracted from reference answers (Mitchell et al., 2002). The patterns included keywords in the reference answers. Patterns could also be learnt from the reference answers (Ramachandran et al., 2015). (Sultan et al., 2016) adopted the simpler notion of semantic alignment to avoid explicitly generating complicated patterns.

Semantic matching had also been proposed in early work (Leacock and Chodorow, 2003). This approach was also used by many researchers (Mohler et al., 2009; Mohler et al., 2011; Heilman and Madnani, 2013) in supervised learning machine learning. A large set of similarity measures is defined as features for a supervised learning model. Features range from word level n-gram overlap to deeper semantic similarity measures based on dictionary and distributional methods.

The short-text grading in SemEval Semantic Textual Similarity (STS) task (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015) drew the attention of many researchers and provided an evaluation platform. Since then, several systems have been proposed for short answer grading based on the semantic similarity with given reference answers (Mohler and Mihalcea, 2009; Mohler et al., 2011; Heilman and Madnani, 2013; Ramachandran et al., 2015). (Sultan et al., 2016) presented a simple short answer grading system for short answer in English. Given a question and its reference answers, a system measures the correctness of a student answer by calculating the similarity with the correct answers.

Comparing to the field in English, there are very little research projects on short answer grading in Chinese, and there is no publicly available corpus for short answering grading in Chinese.

In this paper we report how we build a system and how to test it with a translated corpus from two publicly available English corpus.

The system first extracts the text similarity features, and the features are used in a support vector model. In the first corpus, answers are graded from 0 to 5; we use support vector regression (SVR) model to learn the grading. In the second corpus, answers are graded as correct/incorrect; we use a support vector machine (SVM) classifier approach to deal with it. In the following sections, we will show the system architecture and experimental results.

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 125–129
Melbourne, Australia, July 19, 2018.

2 System Architecture

We adopt the previous works on the textual entailment (TE) as our prototype to tackle the short answer grading problem in Chinese. TE can be briefly defined as: "Given a pair of sentences (Student Answer, Reference answer), a program has to decide whether the information in Reference answer can be inferred by the Student answer". TE can be used in various applications, such as question answering system, information extraction, information retrieval, and machine translation. Once a system is able to decide whether T1 entails T2 or not, it can be regarded as an information filter to help users find useful information. Traditional approaches to TE are based on the semantic and syntactic similarities of the words in the sentences.

2.1 Support Vector Machines

Support vector machines (SVM) is a supervised machine learning classification algorithm, which can be used for classifying problem in n-dimension space. It is used widely in various natural language processing research projects and generally generates good results. Comparing to other classification algorithms, SVM algorithm usually has better result when the number of features is quite large and the data is sparse.稀疏

SVM uses $g(x) = w^T \emptyset(x) + b$ as the linear separation hyperplane, where w is the weight vector, b is the bias, $\emptyset(\cdot)$ is a set of high dimensional non-linear transformation function, where w and b is determined by training data that optimizes the following formulas:

$$\min \frac{1}{2} W^t W + C \sum_{i=1}^{N} \xi_i \qquad (1)$$
$$s.t. \begin{cases} y_i g(x_i) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1 \cdots N \end{cases}$$

where ξ_I is the slack variables, and C is the penalty coefficient for all the training samples (x_i, y_i).

2.2 Support Vector Regression

Support Vector Regression (SVR) is using the SVM algorithm on regression problem. The goal of SVM is to find the separation hyperplane, and the goal of SVR is to find the regression hyperplane. For the given training set:

$$\{(x_1, z_1), \ldots, (x_1, z_1)\}$$

where $x_i \in R^n$ is a feature vector, and $z_i \in R^1$ is the target output. In order to find the hyperplane, two parameters C > 0, and $\varepsilon > 0$ must be given and the support vector regression can be defined:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i + \sum_{i=1}^{l} \xi_i^* \qquad (2)$$
$$\text{Subject to} \quad w^T \phi(x_i) + b - z_i \leq \epsilon + \xi_i ,$$
$$z_i - w^T \phi(x_i) - b \leq \epsilon + \xi_i^* ,$$
$$\xi_i, \xi_i^* \geq 0, i = 1, \ldots, l$$

In our experiment, we use a free SVM toolkit, LIBSVM, to train the SVR model.[1] (Chang and Lin, 2011)

2.3 Feature extraction

In this section, we briefly introduce the features used in SVM, which are the same as those used in previous work. Table 1 shows the ten features used in the experiments. The first three features are the numbers of common terms both in T1 and T2. The next three features are the BLEU scores. The rest four features are the numbers and differences of sentence length of T1 and T2.

No	Feature
1	unigram_recall
2	unigram_precision
3	unigram_F_measure
4	log_bleu_recall
5	log_bleu_precision
6	log_bleu_F_measure
7	difference in sentence length (character)
8	absolute difference in sentence length (character)
9	difference in sentence length (term)
10	absolute difference in sentence length (term)

Table 1: Features used in the system

3 Data Sets

3.1 Data Sets in English

SciEntBank:
This data set was used in SemEval-2013 and available via github[2]. The data set assigns one of five labels to a student response: correct, partially

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[2] https://github.com/leocomelli/score-freetext-answer/archive/master.zip

correct/incomplete, contradictory, irrelevant, and non-domain.

SciEntBank corpus in English contains 9,804 answers to 197 questions in 15 scientific domains. There is one reference answer for each question.

Data Structure Data Set:[3]

The data set is provided by (Mohler and Mihalcea, 2009), which is Data Structure questions and student responses graded by two judges. The data set assigns one of two labels to a student response: correct or incorrect. The questions are collected from ten assignments and two tests, and each one has a topic such as programming basics or sorting algorithms. A reference answer is also provided for each question. The interannotator agreement is 0.586 (Pearson's *r*) and .659 (RMSE on a 5-point scale). Average score of the two judges is used as the final gold score for each student answer.

3.2 Chinese Corpus Translation

Since there is no publicly available data set in Chinese, our experiments are conducted on the translated corpus. With the help of machine translation, we translate the two data set into Chinese and use them in our experiments. The sentences are then segmented into words by the Jieba[4] word segmentation toolkit. The quality of machine translation is not perfect, 12% of the sentences have to be corrected manually. The major error types are synonyms with improper usage in the context for both nouns and adjectives. There are also sentences with bad grammar.

4 Experiments

Since the SciEntBank data set has 5 way labelling, we use regression model to predict the scores of the student responses. And the Data Structure Data Set has 2 way labelling, we use the classification model to predict the scores of the student responses.

4.1 Metrics

For a regression result evaluation, we adopt the squared correlation coefficient and mean squared error. For a classification result evaluation, we adopt the accuracy.

Squared correlation coefficient, R^2

R^2 is the square of the Pearson correlation coefficient between the observed x and modeled (predicted) y data values of the score. Pearson's correlation coefficient is commonly represented by the letter *r*. So if we have one dataset $\{x_1,...,x_n\}$ containing n values and the prediction of the dataset $\{y_1,...,y_n\}$ containing *n* values, then that formula for *r* is:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

where *n* is the sample size, x_i is the sample indexed with *i*, y_i is the correspondent system prediction, and $\bar{x}$, $\bar{y}$ are the means of x_i, and y_i, respectively.

Root mean squared error (RMSE)

Features	R^2	RMSE
all features	0.083041	1.173427
only bleu	0.127850	1.102370

r = 0.357

Table 2: Performance on the Chinse version of the Mohler et al. (2011) dataset with in-domain training.

RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$$

同一个表复制一遍?

4.2 Results

Features	R^2	RMSE
all features	0.083041	1.173427
only bleu	0.127850	1.102370

Table 2 shows the regression results on the Chinese version of the Mohler et al. (2011) dataset. Where all features means the system uses all the features listed in Table 1, and only bleu means the system uses only the bleu features. The experiment result shows that more features can improve the performance.

不对啊，是下面行的指标好啊

Table 3 shows the classification result on the Chinese version of the SemEval-2013 dataset, where all features means the system uses all the features listed in Table 1, and only bleu means

Features	Accuracy(%)
all features	59.569
only bleu	59.568

Table 3: Performance on the Chinse version of the SemEval-2013 datasets.

[3] http://web.eecs.umich.edu/~mihalcea/downloads/ShortAnswerGrading_v1.0.tar.gz

[4] https://github.com/fxsjy/jieba

the system uses only the bleu features. In this experiment, the accuracy is almost the same. The result shows that more features do not improve the performance.

4.3 Discussions

Since the data sets are translated ones, it is not suitable to compare the results to the original ones. However, comparing to the result in English (Sultan et al., 2016), we find that the performance is similar.

5 Conclusion and Future Works

In this paper, we report a short answer grading system in Chinese based on a machine learning approach. We test it with translated corpus from two publicly available corpus in English. The experiment result shows that the results on the two different corpus is promising.

In the future, we will further develop the system with deep learning models. First at all, we will use distributed word embedding technique, such as word2vec, to improve the representation of the text. Then a recurrent neural network with long short term memory neuron is desired to replace the SVM model. Also curate corpus from native Chinese students is also important. Word segmentation is also important; instead of Jieba, we might use CKIP word segmentation service (Ma and Chen, 2003).

Most research projects require reference answers, and unsupervised automatic short answer grading is an interesting way to bypass the requirement (Adams et al., 2016)

6 Acknowledgment

This study is supported by the Ministry of Science under the grant numbers MOST106-2221-E-324-021-MY2.

References

Oliver Adams, Shourya Roy, Raghuram Krishnapuram. 2016. Distributed Vector Representations for Unsupervised Automatic Short Answer Grading, in Proceedings of The 3rd Workshop on Natural Language Processing Techniques for Educational Applications, Osaka, Japan.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A Pilot on Semantic Textual Similarity. In SemEval.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 Shared Task: Semantic Textual Similarity. In Second Joint Conference on Lexical and Computational Semantics (*SEM).

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre,Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In SemEval.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre,Weiwei Guo, I˜nigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In SemEval.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The Eras and Trends of Automatic Short Answer Grading. International Journal of Artificial Intelligence in Education 25: 60. https://doi.org/10.1007/s40593-014-0026-8

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27.

Michael Heilman and Nitin Madnani. 2013. ETS: Domain Adaptation and Stacking for Short Answer Scoring. In SemEval.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated Scoring of Short-Answer Questions. Computers and the Humanities, 37(04).

Ma, Wei-Yun and Keh-Jiann Chen, 2003, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff", Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, pp168-171.

Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge. 2002. Towards Robust Computerised Marking of Free-Text Responses. In Proceedings of the 6th International Computer Assisted Assessment (CAA) Conference.

Michael Mohler and Rada Mihalcea. 2009. Text-to-text Semantic Similarity for Automatic Short Answer Grading, in Proceedings of the European Chapter of the Association for Computational Linguistics (EACL 2009), Athens, Greece.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to Grade Short Answer Questions Using Semantic Similarity Measures and Dependency Graph Alignments. In ACL.

Lakshmi Ramachandran, Jian Cheng, and Peter Foltz. 2015. Identifying Patterns For Short An-

swer Scoring using Graph-based Lexico-Semantic Text Matching. In SemEval.

Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. 2016. Fast and Easy Short Answer Grading with High Accuracy. Proceedings of NAACL-HLT 2016, pages 1070–1075, San Diego, California, June 12-17, 2016.