KC99: A Prediction System for Chinese Textual Entailment Relation using Decision Tree

Tao-Hsing Chang
Dept. of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences
changth@kuas.edu.tw

Yao-Chi Hsu
Dept. of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences
1099108121@cc.kuas.edu.tw

Chung-Wei Chang
Dept. of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences
1099108140@cc.kuas.edu.tw

Yao-Chuan Hsu
Dept. of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences
1099108120@cc.kuas.edu.tw

Jen-I Chang
Dept. of Educational Psychology & Counseling, National Taiwan Normal University
jenichang111@ntnu.edu.tw

ABSTRACT
The aim of the current study is to propose a system, which can automatically deduce entailment relations of textual pairs. The system mainly uses seven features and a decision tree is utilized as a prediction model of the system and seven features of textual pairs are employed to be input of the prediction model. The experimental results for dataset Formal-run based on our proposed method are evaluated by NTCIR. In CT-BC task, Macro-F1 of the proposed method is 57.67%. In CT-MC task, Macro-F1 is 43.73%.

Keywords
Textual entailment, decision tree, features, RITE-2.

1. INTRODUCTION
Textual entailment is an important issue in the study of natural language understanding. Due to early and well-developed English language analysis tools like WordNet or grammar parser, many approaches for English textual entailment were proposed. In contrast, there are fewer tools for Chinese language analysis and the performance of these tools is not so good like that for English. Hence, some methods of textural entailment for Chinese have been proposed, but the performance of those methods is less than that of textual entailment for English. Apparently, Chinese textual entailment is still quite a difficult issue.

Entailment relations of a textual pair discussed in this study include four relations- forward, bidrection, contradiction or independence, which are defined by RITE 2 of NTCIR-10. An example of a textual pairs is as follows.

S1: 韭菜原產於中國，是常見的蔬菜之一。
(Chives, which originate in China, are one of the most commonly seen vegetables.)

S2：韭菜原產於中國。
(Chives originate in China.)

S3：韭菜原產於日本。
(Chives originate in Japan.)

S4 : 水菜原產於日本。
(Potherb mustard originates in Japan.)

S5: 水菜原產地為日本。
(Japan is the country of origin for Potherb mustard.)

As far as textual pair (S1,S2) is concerned, all information content in S2 can be inferred from S1, but the content 'one of the most commonly seen vegetables' in S1 cannot be inferred from S2, thereby classifying this textual pairs as forward relations. As for textual pair (S2,S3), there is a contradiction in the two sentences, so this textual pair is classified as contradiction relations. Regarding textual pair (S3,S4), though S4 resemble S3, their themes are different. Thus, that is classified as independence relations. Concerning textual pair (S4,S5), two sentences illustrate the same matter. Both textual pair (S4,S5) and textual pair (S5,S4) are forward relations. Therefore, both of the two textual pairs are classified as bidirection relations.

The aim of the current study is to propose a system, which can automatically deduce entailment relations of textual pairs. The system mainly uses seven features and a decision tree is utilized as a prediction model of the system and seven features of textual pairs are employed to be input of the prediction model. The rest of this paper is organized as follows. Section 2 discusses previous related studies and indicates relations between our proposed method and previous methods. Section 3 introduces features and prediction model developed by this study. Section 4 illustrates experimental result of adopting our proposed method in NTCIR Rite 2 data set. Finally, this paper displays discussions of our proposed method and experimental result as well as conclusion.

2. RELATED WORKS
Many research studies concerning English textual entailment inference have been proposed. Androutsopoulos and Malakasiotis [1] used surface string similarity to judge the textual entailment relations. The method can measure the similarity between two sentences more precisely if two sentences use the same words to address the same meaning; however, the method cannot deal with synonym pairs. Bos and Markert [2] proposed sifting rules of shallow semantic features to overcome this problem. [2] used WordNet as background knowledge to interpret if different words have the same or opposite meaning. For example, words 'murder' and 'slay' have the same meaning. 'Murderer' is derivative from 'murder' and antonym for the word 'murderer' is 'victim'. All

these words can be grouped together as related words of word 'murder'.

In addition to evaluate entailment relations via the similarity of vocabulary and semantic features, some methods [3-5] also utilized parsing tree to obtain syntactic structures as features for inferring textual entailment. Cabrio, Kouylekov and Magnini [3] proposed such methods as linear distance and tree edit distance to evaluate textual entailment relations based on tree distance [4-5]. [4] and [5] first employed parser to parse two sentences into syntactic parse trees. Through inserting, deleting and substituting several times, it adjusted two textual syntactic parse trees into the same graph. The number of occurrence of insertion, deletion and substitution are then utilized to compute the difference between the two trees. [3] used the difference to inferring the textural entailment relations.

Some research studies had proposed methods to infer Chinese textural entailment relations in competition for RITE of NTCIR-9 in 2011. Han and Ku [6] adopted shallow features, including sentence lengths, content of matched keywords, quantities of matched keywords and their parts of speech, to capture the difference between two sentences, were employed to judge textual entailment relations. Experimental results showed the approach performed well by only using shallow features to infer entailment relations. Huang et al. [7] proposed a method based on syntactic analysis. First, the method first used Stanford parser [8] to parse sentences into syntactic tree and to identify main verbs and nouns. Second, it then generalizes several syntactic features based on different types of main verbs and noun. Finally, it used these features to compute syntactic similarity between two sentences. Experimental result showed that the accuracy of methods in [7] was higher than that of methods which only used shallow features. Chang et al. [9] also indicated that using both shallow features and direct features enhanced the accuracy of prediction system for automatic essay scoring.

It is need to use a classification model to integrate different features mentioned above together and predict textual entailment relation [10]. Support vector machine (SVM) is the most commonly used model for classification. For instance, [2] used features, including word, parsing tree, sentiment polarity and the referred name entity, to transform a textual pair into a feature vector. Using SVM, the vector is classified into an entailment relation. Another commonly used model for classification is decision tree. Decision tree can be constructed by experts or such machine learning methods as ID3. ID3 can optimize the construction of decision tree via information gain.

Our proposed method also utilizes lexical features, semantic and syntactic features as well as decision tree as a prediction model to identify textual entailment relations.

3. Methodology

Seven features of textual pairs and decision trees are utilized in our proposed method to predict the entailment relations of textual pairs. Additionally, to compute features, textual pairs must undergo preprocessing stage. The details of preprocessing, the extraction of features and the use of decision tree illustrate as follows.

Preprocessing

Due to no delimiters separating words in Chinese sentences, it is necessary to undergo segmenting sentences into words. In addition, the use of part-of-speech of each word is also necessary for our proposed method. Many word segmentation and part-of-speech tagging systems have worked well. However, due to the need for analyzing features of textual pairs, WeCAn [11] is employed in this study to segment the sentences in textual pairs into words and tag the part-of-speech of these words.

In researches used Chinese textual pair, the unknown word is another problem especially needed to be dealt with. Because of no space between Chinese words, it is very difficult to identify all unknown words in text by Chinese word segmentation systems even those systems performed well. To reduce the probability of occurrence of the unknown word, proper nouns appear on Wiki are incorporated into dictionary in the proposed method. Moreover, SPLR method [12] is utilized to improve the ability to detect unknown words.

In addition, data format inconsistency often occurs in textual pairs. Such data as time and quantity in two sentences is the same but the representation for the data is not. There are three common situations for data format inconsistency. Firstly, it uses various ways to express the same data. For instance, half or 1/2 is used to represent as 0.5. The second is abbreviation. For example, year 2003 would be written as '03. The third is unit conversion. For instance, 1kg is written as 1000 grams. Although these problems occur frequently, these can be solved by only using finite rules to transform them into the same expression and unit. As a result, a module in our proposed system deals with data format inconsistency based on rule-based methods.

Feature 1: Word order exchange (WOE)

Some textual pairs which are contradiction relations use identical words in two sentences, but their semantic meanings are opposite due to different order of appearance of words. For instance, the following textual pair (S6,S7) is contradiction relation. Word "SARS virus" and "coronavirus" are used in both sentences, but it is different between the order of appearance of the words in the two sentences. Therefore, this study designs a feature called word order exchange (WOE). WOE of a textual pair is 1 if two sentences consist of the same words but have different word order; otherwise, that is -1. In some exception, two sentences have also different word order but the semantics of the sentences is the same. These exceptions usually possess conjunction. For example, in the textual pair (S8,S9), sentence S8 and S9 share the same semantic even through words "Ipsos" and "Associated Press" appear with different order in the sentences. This study designs a module to detect these exceptions. WOE of a textual pairs in exception is set to -1.

S6: SARS 病毒屬於冠狀病毒
(SARS virus is a coronavirus.)

S7: 冠狀病毒屬於 SARS 病毒
(Coronavirus is a SARS virus.)

S8：美聯社和伊普索斯公司所進行的民調顯示，美國總統布希的施政滿意度已首次滑落到 39%
(Conducted by the Associated Press and Ipsos, the polls showed that the satisfaction of U.S. President George W. Bush's administration has fallen to 39% for the first time.)

S9：伊普索斯公司和美聯社所進行的民調顯示，美國總統布希的施政滿意度已首次滑落到 39%
(Conducted by Ipsos and the Associated Press, the polls showed that the satisfaction of U.S. President George W. Bush's administration has fallen to 39% for the first time.)

Feature 2: Consistency of the number of Nouns (CNN)
We observed that the same as the number of noun are used in two sentences; the higher probability a textual pair is bidirection or contradiction relation. This is owing to the fact that the number of noun is represented as the number of matters. The same matter may stand for that two sentences address the same event, or may lead to contradiction due to other negative words. For example, the following three sentences contain three nouns and the same matter is expressed in the sentences. Textual pair (S10,S11) is bidirection but textual pair (S11,S12) is contradiction.

S10: H5N1 型病毒株能透過禽類傳染給人體
(H5N1 virus strain can be transmitted to humans through poultry.)

S11: H5N1 型病毒株是藉由禽類傳染給人體
(H5N1 virus strain is transmitted to humans through poultry.)

S12: H5N1 型病毒株並非由禽類傳染給人體
(H5N1 virus strain is not transmitted to humans through poultry.)

As for the textual pair, this paper defines a feature as consistency of the number of nouns' (CNN). CNN of a textual pair is 1 if the number of noun within two sentences is the same; otherwise, that is -1.

Feature 3: Difference between rates of overlapping word (DRO)
The less the same words are used in two sentences within a textual pair, the higher probability a textual pair is independence relations. Thus, for a text pair (S_i, S_j), rate of word overlapping forward (RWF) and rate of word overlapping back (RWB) are defined as follows:

$$RWF = \frac{|W_i \cap W_j|}{|W_j|}, \qquad RWB = \frac{|W_i \cap W_j|}{|W_i|}$$

where W_k stands for a set of all words within sentence S_k, $|W_k|$ represents as the number of words in set W_k. If both RWF and RWB of a textual pair are lower, it means that too few information is shared by two sentences. Based on the observation, the textual pair should be independence because two sentences contain different contents. For instance, the following textual pair (S13,S14) is independence, where RWF is 0.05 and RWB is 0.16.

If the difference between RWF and RWB of a textual pair is higher, it represents that the sentences within the pair contain similar information but one includes a large amount of information and the other a less. Therefore, the textual pair may be forward. For example, a textual pair (S15,S16) is forward, where RWF is 0.30 and RWB is 0.75. As for other level of RWF and RWB, the textual pair may be independence because there are not sufficient evidences to identify the textual pair as one of the other three entailment relations.

S13: 馬來西亞原為日本電子業者眼中最佳的亞洲投資標，現被中國大陸取代
(As for Japanese electronics industry, Malaysia is the best investment opportunity in Asia, but is now replaced by China)

S14: 中國取代美國成為亞洲經濟核心
(China replaced the United States as the economic core of Asia.)

S15: 日本是投資馬來西亞的三大外商之一
(Japan is one of three major foreign companies that invests in Malaysia.)

S16: 日本有投資馬來西亞
(Japan invests in Malaysia.)

Based on observation mentioned above, feature 'difference between rates of overlapping word' (DRO) of a textual pair is defined as follows.

$$DRO = \begin{cases} 1, & if\ RWF \le TI\ and\ RWB \le TI \\ 0, & if\ RWB - RWF \ge TD\ and\ (RWF \ge TI\ or\ RWB \ge TI) \\ -1, & otherwise \end{cases}$$

where TI and TD are thresholds. According to the training data used in our experiment, TI and TD is 0.6 and 0.2, respectively.

Feature 4: Difference between rates of overlapping POS (DOP)
The less the identical part-of-speech is used in two sentences within a textual pair, the higher probability a textual pair is independence relations. The phenomenon is similar to that for DRO. Referring to the definition of DRO, this paper defines a feature as 'Difference between rates of overlapping POS' (DOP) to represent the occurrence of the phenomenon. Given a textual pair (S_i,S_j), rate of part-of-speech overlapping forward (RPF) and ate of part-of-speech overlapping back (RPB) are as follows:

$$RPF = \frac{|P_i \cap P_j|}{|P_j|}, \qquad RPB = \frac{|P_i \cap P_j|}{|P_i|}$$

where P_k stands for a set of all words within sentence S_k, $|P_k|$ represents as the number of words in set P_k. If RPF of a textual pair is high enough and the difference between FPR and BPR of the pair is also higher, the textual pair may be forward. If not, the relation of the pair cannot be identified. Based on this observation, feature DOP is defined as follows.

$$DOP = \begin{cases} 1, & if\ RPF \ge TP\ \ and\ RPF - RPB \ge TK \\ -1, & otherwise \end{cases}$$

where TP and TK are thresholds. According to the training data used in our experiment, TP and TK is 0.7 and 0.2, respectively.

Feature 5: Occurrence of time (OOT)
In some textual pairs, time often appears in one sentence but does not appear in the other one. For example, in textual pair (S17,S18), S17 mentions a specific time '2001' but S18 does not mention any time. Hence, the entailment relation of the textual pair may be forward. This paper defines feature 'occurrence of time' (OOT) as follows. If the time occurs in both of sentences in a text pair, OOT of the pair is 1; otherwise, that is -1.

S17: 小泉純一郎 2001 年贏得自民黨總裁選戰
(Junichiro Koizumi won the LDP president election in 2001.)

S18: 小泉純一郎贏得自民黨總裁選戰
(Junichiro Koizumi won the LDP president election.)

Feature 6: Existence of negative word (ENW)
In some textual pairs, there is high similarity between two sentences but they express opposite meanings due to negative words. For instance, textual pair (S19,S20) is contradiction because of the word '未'(not). This paper defines feature 'existence of negative word' (ENW) as follows. If a textual pair possesses negative words, ENW of the pair is 1; otherwise, that is -1.

S19:小泉純一郎 2001 年贏得自民黨總裁選戰
(Junichiro Koizumi won the LDP president election in 2001.)

S20:小泉純一郎 2001 年未贏得自民黨總裁選戰
(Junichiro Koizumi did not win the LDP president election in 2001.)

Feature 7: Synonyms
Some textual pairs use most of the same words in two sentences, and only one word is different. If this word is synonyms, the textual pair may be bidirection relations. For example, in the following text pair (S21,S22), only word 'Holy See' and word 'Vatican' are different in these two sentences, but those words are the same meaning. Therefore, this paper defines feature 'synonyms' (SYN) as follows. Assume the most of the identical words are used in a textual pair and their word order is the same; but only one word is different. If these two different words are synonyms, SYN of the pair is 1; otherwise, that is -1.

To construct synonyms, each Chinese words in dictionary correspond to an English terms via Google translation. If two words correspond to the same English expression, two words are identified as synonyms each other.

S21: 若望保祿二世是教廷領導人
(John Paul II is the leader of the Holy See.)

S22: 若望保祿二世是梵諦岡領導人
(John Paul II is the leader of the Vatican.)

Among seven features, CNN and DRO are shallow features; WOE, ENW, OOT and SYN can be viewed as semantics feature; DOP can be treated as syntactical feature. Seven features of each textual pair in training data are utilized to construct a decision tree. For a testing textual pair, our proposed system will first compute the values of seven features of the pair and then employ the decision tree as well as the values to predict entailment relation of the testing textual pair.

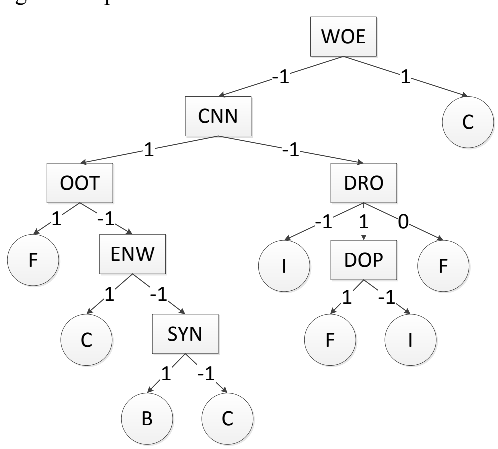

Fig. 1 Decision tree for textual entailment prediction

Decision tree has been widely used in various classification issues [13][14] and such optimization algorithms for decision tree as ID3 [15] were proposed. [2] employed decision tree to recognize textual entailment and achieved high accuracy scores on the RTE dataset. Owing to the fact that our system is only an experimental system, this study only uses simple decision tree to implement our methodology. To construct a decision tree, a set of training instances including several attributes and a target attribute are employed. The value of target attribute for a new unseen example can be predicted by the decision tree. The details of this concept can be referred to [13]. In this study, the instance, attribute and target attribute can be symbolized as a textual pair, seven features and entailment relation. Fig. 1 is the decision tree constructed by our proposed method.

4. Experimental Results

Experiments consist of two tasks of RITE-2 at the NTCIR-10 workshop: CT-BC (Chinese Traditional Binary Classification) and CT-MC (Chinese Traditional Multi-way Classification) [16]. Experimental material is composed of developed dataset and formal-run dataset. In CT-MC task, textual pairs should be classified into B (Bidirection), F (Forward), C (Contraction), I (Independent) relations. In CT-BC task, relations B and F are grouped to Y while relations C and I are integrated into N. Table 1 shows the detail of development and formal-run dataset.

Table 1. Statistics of the datasets [16]

Datasets	Relations				Total
	Y		N		
	B	F	C	I	
Development	82	184	74	81	421
Formal-run	151	328	114	288	881

There are two datasets provided by NTCIR-10. Dataset Development is used to train our proposed prediction model and evaluate the performance of features proposed by this paper. Dataset Formal-run is employed to test the performance of our proposed model. Table 2 is the prediction result for dataset Development based on our proposed method. Owing to the fact that categories B and F in CT-MC can be merged into category Y and categories C and I in CT-MC can be merged into category N in CT-BC, Table 2 thus only presents experimental results for CT-MC. The precision and recall of relation forward is higher than other relations and there is little difference in the recall and the precision of relations B and I; while it showed that the precision with higher relation C but lower recall. The overall precision is 61.20, and the recall is 49.43.

Table 2 Prediction results for dataset Development based on our proposed method

	Prediction				Precision	Recall
	B	C	F	I		
B	139	24	8	91	41.87	53.05
C	63	69	33	89	65.09	27.17
F	85	4	327	128	80.34	60.11
I	45	9	39	168	53.15	45.21
Overall	332	106	407	476	61.20	49.43

Table 3 Prediction results for dataset Formal-run based on our proposed method

Tasks / Indicator	CT-BC		CT-MC			
	Y	N	B	F	C	I
F1	66.42	48.93	45.48	63.61	16.67	49.24
Precision	60.45	57.58	42.94	57.00	15.87	66.08
Recall	73.70	42.54	48.34	71.95	17.54	39.24

The prediction results for dataset Formal-run based on our proposed method are evaluated by NTCIR. In CT-BC task, Macro-F1 of the proposed method is 57.67%. In CT-MC task, Macro-F1 is 43.73%. Table 3 is the evaluation of prediction performance for each type of textual entailment relation in dataset Formal-run based on our proposed method.

5. Discussion and further works

Table 3 showed a fairly good result when identifying textual pairs as relation F based on our proposed method, but very low performance when identifying the relation C. It is also the most important factor that influences the overall performance. Although the precision is pretty good in the training phase, the recall rate is quite low. By examining the result of training data against decision tree, several possible reasons are pointed out.

First, there are three nodes to distinguish F from decision tree. Moreover, the recall and precision of three features to distinguish F are higher. Therefore, the performance of identifying textual pairs as relation F is better. Second, although only one node for identifying textual pairs as relation B, it can be indeed found through four features. Thus, the performance of identifying textual pairs as relation B is acceptable.

Third, the results of identifying textual pairs as relation I seem pretty good. In the proposed decision tree, however, the textual pairs of which our system cannot confirm relations by the features are almost identified as I. It is very difficult to improve the performance of identifying textual pairs as relation I because identification to relation I is insensitive and insufficient. Finally, although there are also three nodes to distinguish C from decision tree, only few textual pairs which are classified into relation C can be distinguished by these nodes. Therefore, many textual pairs for relation C which cannot be classified are identified as relation I. It results in that prediction results are not excellent.

Some researches can be developed and explored in the future. Firstly, although there are seven features are employed in the proposed method, many textual pairs is still not be classified by these features. It is necessary to explore more features, especially semantic and syntactical features. Second is to use more efficient synonyms extraction for feature SYN. As seen from experimental results, the performance of using feature synonyms is quite good, but the extraction used in our system is too rough so that many synonyms cannot be identified. Thirdly, the decision tree should be optimized by such algorithm as ID3. It can enhance the performance of decision tree.

6. Acknowledgements

This work is particularly supported by "Aim for the Top University Project" of the National Taiwan Normal University and the Ministry of Education, Taiwan, R.O.C.

7. REFERENCES

[1] Androutsopoulos, I. and Malakasiotis, P. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research.* 38, (May, 2010), 135-187.

[2] Bos, J. and Markert, K. 2005. Recognising textual entailment with logical inference. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistic*s (Oct. 2005), 628-635.

[3] Cabrio, E., Kouylekov, M., and Magnini, B. 2008. Combining specialized entailment engines for rte-4. In *Proceedings of TAC08, 4th PASCAL Challenges Workshop on Recognising Textual Entailment.* (2008).

[4] Kouylekov, M. and Magnini, B. 2005. Recognizing textual entailment with tree edit distance algorithms. In *Proceedings of the First Challenge Workshop Recognising Textual Entailment* (Apr. 2005), 17-20.

[5] Kouylekov, M. and Magnini, B. 2006. Tree edit distance for recognizing textual entailment: Estimating the cost of insertion. In *Proceedings of the PASCAL RTE-2 Challenge* (Apr. 2006), 68-73.

[6] Han, N. H. and Ku, L. W. 2011. The Yuntech system in NTCIR-9 RITE Task. In *Proceedings of the NTCIR-9 Workshop* (Dec. 2011), 345-348.

[7] Huang, H. H., Chang, K. C., Haver II, J. M., and Chen, H. H. NTU Textual Entailment System for NTCIR 9 RITE Task. In *Proceedings of the NTCIR-9 Workshop* (Dec. 2011), 349-352.

[8] Stanford Parser : A statistical parser. 2002. Available from : http://nlp.stanford.edu/software/lex-parser.shtml

[9] Chang, T. H., Lee, C. H., Tsai, P. Y. and Tam, H. P. 2009. Automated essay scoring using set of literary sememes. *Information: An International Interdisciplinary Journal*, 12, 2 (2011), 351-357.

[10] Akiba, Y., Taira, H., Fujita, S., KASAHARA, K., and Nagata, M. 2011. NTTCS textual entailment recognition system for the NTCIR-9 rite. In *Proceedings of the 9th NII Test Collection for Information Retrieval Workshop* (Dec. 2011), 330-334.

[11] Chang, T. H., Sung, Y. T., and Lee, Y. T. 2012. A Chinese word segmentation and POS tagging system for readability research. In *Proceedings of the 42nd Annual Meeting of the Society for Computers in Psychology*. (Minnesota, USA, 2012).

[12] Chang, T. H. and Lee, C. H. 2003. Automatic Chinese unknown word extraction using small-corpus-based method. *In Proceedings of International Conference on Natural Language Processing and Knowledge Engineering*. (Beijing, China, 2003), 459-464.

[13] Safavian, S.R. and Landgrebe, D. 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man and Cybernetics*. 21, 3, (May/Jun. 1991), 660-674.

[14] Chang, T. H., Lee, C. H., and Chang, Y. M. 2006. Enhancing automatic Chinese essay scoring system from figures-of-speech. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, (Wuhan, China, 2006), 28-34.

[15] Quinlan, J.R. 1986. Induction of Decision Tree. *Machine Learning*. 1, 1 (1986), 81-106.

[16] Watanabe, Y., Miyao, Y., Mizuno, J., Shibata, T., Kanayama, H., Lee, C. W., Lin, C. J., Shi, S., Mitamura, T., Kando, N., Shima, H., and Takeda, K. 2013. Overview of the Recognizing Inference in Text (RITE-2) at the NTCIR-10. In *Proceedings of the 10th NTCIR Conference*, (2013).