IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-9 RITE

Min-Yuh Day [1, *], Re-Yuan Lee [1, 2], Cheng-Tai Liu [2], Chun Tu[1] , Chin-Sheng Tseng [1], Loong Tern Yap [1], Allen-Green C.L. Huang [1], Yu-Hsuan Chiu [1], Wei-Ze Hong [1]

[1] Department of Information Management, Tamkang University, New Taipei City, Taiwan
[2] Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
Phone: 886-2-26215656#2347

myday@mail.tku.edu.tw, {re.yuan.lee, tkuterry.liu, kevincncod2}@gmail.com, {496631473, 496636035}@s96.tku.edu.tw, {wivx.com, boloage, over1125}@gmail.com

ABSTRACT

In this paper, we describe the IMTKU (Information Management at TamKang University) textual entailment system for recognizing inference in text at NTCIR-9 RITE (Recognizing Inference in Text). We proposed a textual entailment system using a hybrid approach that integrate knowledge based and machine learning techniques for recognizing inference in text at NTCIR-9 RITE task. We submitted 3 official runs for both BC and MC subtask. In NTCIR-9 RITE task, IMTKU team achieved 0.522 in the CT-MC subtask and 0.556 in the CT-BC subtask.

Categories and Subject Descriptors

H.3.3 **[Information Search and Retrieval]**: Retrieval models and Search process

General Terms

Algorithms, Documentation, Experimentation

Keywords

IMTKU, Textual Entailment, Recognizing Textual Entailment in Chinese, Recognizing Inference in TExt (RITE), NTCIR, Hybrid Approach, Machine Learning, Knowledge-Based

1. INTRODUCTION

IMTKU participated in NTCIR-9 RITE Binary-class (BC) subtask and Multi-class (MC) subtask in Traditional Chinese (CT). We submitted 3 official runs for both BC and MC subtask. In addition, we also participate in RITE4QA subtask in both Traditional Chinese (CT) and Simplified Chinese (CS). We also submitted 3 official runs for RITE4QA subtask in both CT and CS language. In this paper, we described the algorithms, tools and resources used in IMTKU RITE system.

Recognizing Textual Entailment (RTE) is a PASCAL/TAC task of deciding given two text fragments, whether the meaning of one text is entailed (can be inferred) from another text which is mainly focused on English [4, 5] RITE (Recognizing Inference in TExt), however, is a generic benchmark task organized by NTCIR-9 that addresses major text understanding needs in various NLP/Information Access research areas which is mainly focused on Japanese and Chinese [8, 9].

RITE is a benchmark task for evaluating systems which automatically detect entailment, paraphrase, and contradiction in texts written in Japanese, Simplified Chinese, or Traditional Chinese. There are three task settings, namely, Binary-class (BC) subtask, Multi-class (MC) subtask, and RITE4QA subtask in RITE. In all subtasks, a system input is two texts and an output is one of two or five labels [8].

For instance, in the BC subtask, an input text pair would be something like the following [8]:

> t1: Yasunari Kawabata won the Nobel Prize in Literature for his novel "Snow Country"
>
> t2: Yasunari Kawabata is the writer of "Snow Country"

The system output for the BC subtask is YES for the above T1, T2 pair.

For the NTCIR-9 RITE binary classification (BC), given a text pair (t1, T2), a system judges if the hypothesis t2 can be inferred from t1 or not. It's an intrinsic evaluation task similar to PASCAL/TAC RTE 1-5 Main Task, but it's a new attempt that evaluates Asian languages. t1 is a sentence/paragraph-long text and t2 is a short sentence [8].

For the Multi-class Classification (MC) in NTCIR-9 RITE, given a text pair (t1, t2), a system detects entailment in more detail. The class would be yes (forward entailment, reverse entailment, paraphrase), no (contradiction, independence). It's also an intrinsic evaluation with more challenging setting than the BC subtask. The length of t1 and t2 is about the same [8].

According to the task description of NTCIR-9 RITE task [8], BC Subtask is defined as "Given a text pair *(t1, t2)* identify where *t1* entails (infers) a hypothesis *t2* or not", the expected system output label of RITE BC subtask is "{Y, N}". In addition, MC Subtask is defined as "A 5-way labeling subtask to detect (forward / reverse / bidirection) entailment or no entailment (contradiction / independence) in a text pair", the expected system output label of RITE MC subtask is "{F,R,B,C,I}", where F means "forward entailment (t1 entails t2 AND t2 does not entails t1)"; R means " reverse entailment (t2 entails t1 AND t1 does not entails t2)", B means "bidirectional entailment (t1 entails t2 AND t2 entails t1)"; C means "contradiction (t1 and t2 contradicts, or cannot be true at the same time)"; I means "independence (otherwise)". The evaluation of RITE system is the accuracy of labels predicted [8].

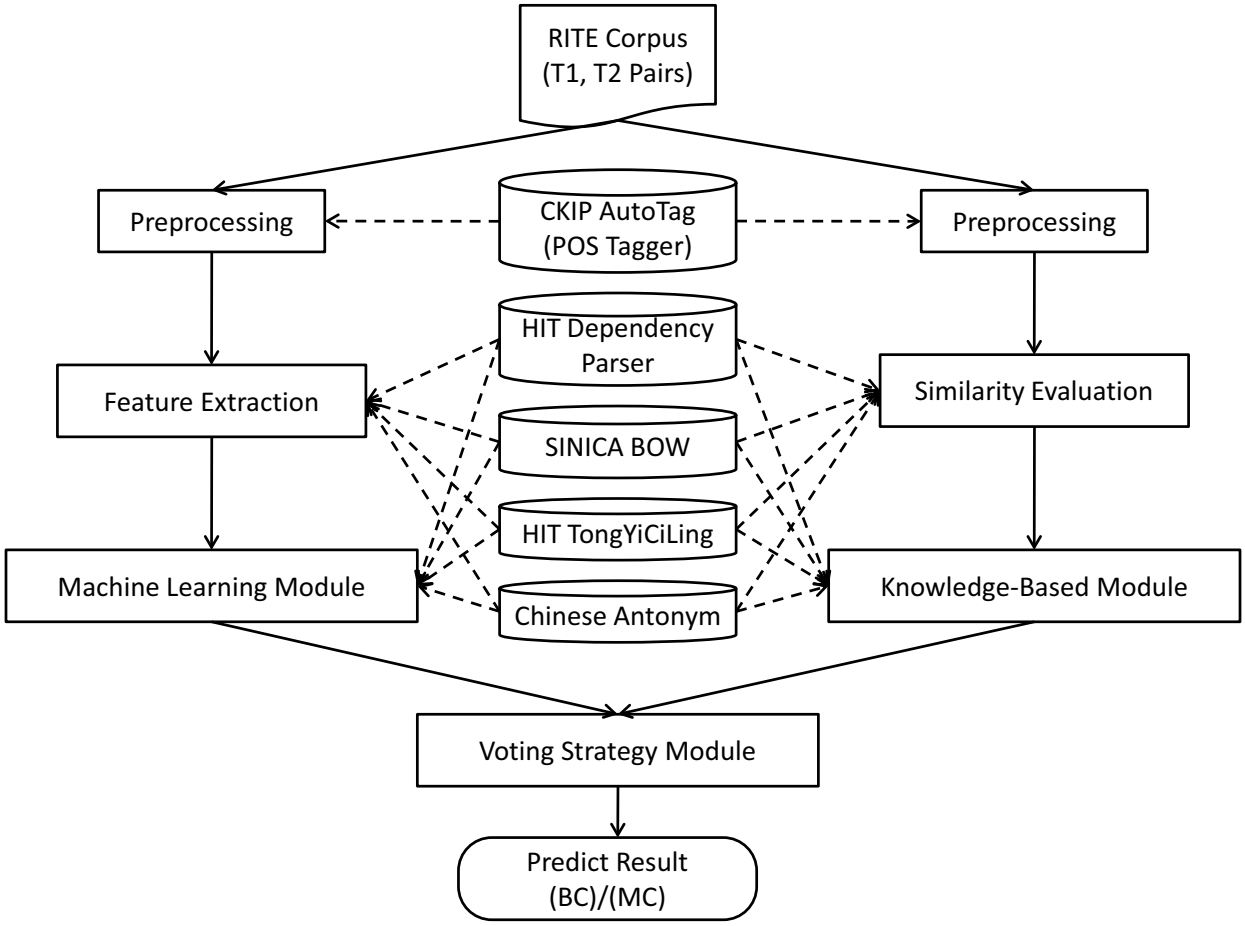

Figure 1. System Architecture of IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-9 RITE

2. SYSTEM ARCHITECTURE

Figure 1 shows the proposed system architecture of IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-9 RITE. We developed a textual entailment system using a hybrid approach that integrate knowledge based and machine learning techniques for recognizing inference in text at NTCIR-9 RITE task. There are three main modules in the proposed hybrid system architecture, namely, (1) machine learning module, (2) knowledge-based module, and (3) voting strategy module. We used LibSVM [1] for Machine Learning Module. We use syntactic and semantic resources for feature extraction in machine learning approach. The syntactic resources and tools we used are CKIP AutoTag (for Chinese POS tagging) [3] and HIT Dependency Parser [2]. We use Academia Sinica Bilingual Ontological WordNet (SINICA BOW) [6] and HIT TongYiCiLing (HIT TYCL) [2, 7] as major semantic resources. In addition, we also compiled a list of Chinese antonym (509 pairs) as additional semantic resources.

3. EXPERIMENTAL RESULS AND ANALYSIS

We conduct several experiments using various datasets (sample data and develop data) to train and test models, as well as different combinations of features.

3.1 Official RITE1 Runs

In this section, we describe the algorithms and resources we used for generating the official runs. We also present the official results and discussions.

3.1.1 RITE-IMTKU-CT-BC Subtask

- **RITE-IMTKU-CT-BC-01**

Tools: CKIP AutoTag, HIT Dependency Parser, LibSVM

Resources: Bilingual Wordnet (SINICA BOW), HIT TongYiCiLing (HIT-TYCL)

Method: Hybrid approach (Integrated Knowledge Base and Machine Learning Approach) for NTCIR-9 RITE.

Feature Extraction from normalized t1 and t2.

Measure similarity match between t1 and t2.

Voting strategy from multiple prediction result.

- **RITE-IMTKU-CT-BC-02**

Tools: CKIP AutoTag, HIT Dependency Parser, LibSVM

Resources: Bilingual Wordnet (SINICA BOW), HIT TongYiCiLing (HIT-TYCL)

Method: Hybrid approach (Integrated Knowledge Base and Machine Learning Approach) for NTCIR-9 RITE.

Feature Extraction from normalized t1 and t2.

Measure similarity match between t1 and t2.

Multiple Features used (Word Based Similarity, Token Based Similarity, Lexical overlap, Text Pair Length, Token Length) in SVM.

- **RITE-IMTKU-CT-BC-03**

Tools: CKIP AutoTag, LibSVM

Resources: NONE

Method: Machine Learning Approach for NTCIR-9 RITE.

Feature Extraction from normalized t1 and t2.

4 Features used (Word Based Edit Distance, Token Based Edit Distance, Text Length, Identical POS Token Rate) in SVM.

3.1.2 RITE-IMTKU-CT-MC Subtask

- **RITE-IMTKU-CT-MC-01**

Tools: CKIP AutoTag, HIT Dependency Parser, LibSVM

Resources: Bilingual Wordnet (SINICA BOW), HIT TongYiCiLing (HIT-TYCL)

Method: Hybrid approach (Integrated Knowledge Base and Machine Learning Approach) for NTCIR-9 RITE.

Feature Extraction from normalized t1 and t2.

Measure similarity match between t1 and t2

Multiple Features used (Word Based Similarity, Token Based Similarity, Lexical overlap, Text Pair Length, Token Length) in SVM.

- **RITE-IMTKU-CT-MC-02**

Tools: CKIP AutoTag, HIT Dependency Parser, LibSVM

Resources: Bilingual Wordnet (SINICA BOW), HIT TongYiCiLing (HIT-TYCL)

Method: Hybrid approach (Integrated Knowledge Base and Machine Learning Approach) for NTCIR-9 RITE.

Feature Extraction from normalized t1 and t2.

Measure similarity match between t1 and t2.

4 Features used (Word Based Similarity, Token Based Similarity, Lexical overlap, Text Pair Length) in SVM.

- **RITE-IMTKU-CT-MC-03**

Tools: CKIP AutoTag, LibSVM

Resources: NONE

Method: Machine Learning Approach for NTCIR-9 RITE.

Feature Extraction from normalized t1 and t2.

4 Features used (Word Based Edit Distance, Token Based Edit Distance, Text Length, Identical POS Token Rate) in SVM.

3.1.3 RITE1-IMTKU-CT-RITE4QA Subtask

- RITE1-IMTKU-CT-RITE4QA-01

Tools: CKIP AutoTag, HIT Dependency Parser, LibSVM

Resources: Bilingual Wordnet (SINICA BOW), HIT TongYiCiLing (HIT-TYCL)

Method: Hybrid approach (Integrated Knowledge Base and Machine Learning Approach) for NTCIR-9 RITE.

Feature Extraction from normalized t1 and t2.

Measure similarity match between t1 and t2

Multiple Features used (Word Based Siilarity, Token Based Similarity, Lexical overlap, Text Pair Length, Token Length) in SVM.

Machine Learning Training Dataset: Dev Dataset (421)

- RITE1-IMTKU-CT-RITE4QA-02

Tools: CKIP AutoTag, HIT Dependency Parser, LibSVM

Resources: Bilingual Wordnet (SINICA BOW), HIT TongYiCiLing (HIT-TYCL)

Method: Hybrid approach (Integrated Knowledge Base and Machine Learning Approach) for NTCIR-9 RITE.

Feature Extraction from normalized t1 and t2.

Measure similarity match between t1 and t2.

Multiple Features used (Word Based Siilarity, Token Based Similarity, Lexical overlap, Text Pair Length, Token Length) in SVM.

Machine Learning Training Dataset: Dev Dataset + Test Dataset (1321)

- RITE1-IMTKU-CT-RITE4QA-03

Tools: CKIP AutoTag, LibSVM

Resources: NONE

Method: Machine Learning Approach for NTCIR-9 RITE.

Feature Extraction from normalized t1 and t2.

13 Features used (Longest Common Substring, Word Based Edit Distance, T1 Token Based Edit Distance,T2 Token Based Edit Distance,T1 Noun Number-T2 Noun Number, T1 Verb Number-T2 Noun Number,T1 Text Length,T2 Text Length, Text Length Rate,T1 Text Length-T2 Text Length, Text Length Ratio, T1 Token Based Edit Distance-T2 Token Based Edit Distance, Identical POS Token Rate) in SVM.

3.1.4 RITE1-IMTKU-CS-RITE4QA Subtask

- RITE1-IMTKU-CS-RITE4QA-01

Tools: CKIP AutoTag, HIT Dependency Parser, LibSVM

Resources: Bilingual Wordnet (SINICA BOW), HIT TongYiCiLing (HIT-TYCL)

Method: Hybrid approach (Integrated Knowledge Base and Machine Learning Approach) for NTCIR-9 RITE.

Feature Extraction from normalized t1 and t2.

Measure similarity match between t1 and t2.

Multiple Features used (Word Based Siilarity, Token Based Similarity, Lexical overlap, Text Pair Length, Token Length) in SVM.

- RITE1-IMTKU-CS-RITE4QA-02

Tools: CKIP AutoTag, LibSVM

Resources: NONE

Method: Machine Learning Approach for NTCIR-9 RITE.

Feature Extraction from normalized t1 and t2.

13 Features used (Longest Common Substring, Word Based Edit Distance, T1 Token Based Edit Distance,T2 Token Based Edit Distance,T1 Noun Number-T2 Noun Number, T1 Verb Number-T2 Noun Number,T1 Text Length,T2 Text Length, Text Length Rate,T1 Text Length-T2 Text Length, Text Length Ratio, T1 Token Based Edit Distance-T2 Token Based Edit Distance, Identical POS Token Rate) in SVM.

- RITE1-IMTKU-CS-RITE4QA-03

Tools: CKIP AutoTag, LibSVM

Resources: NONE

Method: Machine Learning Approach for NTCIR-9 RITE.

Feature Extraction from normalized t1 and t2.

9 Features used (Longest Common Substring, T1 Token Based Edit Distance,T2 Token Based Edit Distance,T1 Noun Number-T2 Noun Number, T1 Verb Number-T2 Noun Number,T1 Text Length-T2 Text Length, Text Length Ratio, T1 Token Based Edit Distance-T2 Token Based Edit Distance, Identical POS Token Rate) in SVM.

3.1.5 Summary of IMTKU Official Runs

We list the summary of IMTKU Official Runs for RITE CT BC, CT MC, CT RITE4QA, CS RITE4QA subtasks in Table 1, 2, 3, 4. Table 1 shows that the best performance of our submitted official runs for RITE CT BC Subtask is 0.556, which is "RITE1-IMTKU-CT-BC-02". Table 2 shows that the best performance of our submitted official runs for RITE CT BC Subtask is 0.552, which is "RITE1-IMTKU-CT-MC-01". Table 3 shows that the best performance of our submitted official runs for RITE CT RITE4QA Subtask is 0.4003, which is "RITE1-IMTKU-CT-RITE4QA-03". Table 4 shows that the best performance of our submitted official runs for RITE CS RITE4QA Subtask is 0.4716, which is "RITE1-IMTKU-CS-RITE4QA-01".

Table 1. Accuracy of IMTKU CT BC Subtask Official Runs

IMTKU BC Subtask Official Runs	Accuracy
RITE1-IMTKU-CT-BC-01	0.550
RITE1-IMTKU-CT-BC-02	**0.556**
RITE1-IMTKU-CT-BC-03	0.524

Table 2. Accuracy of IMTKU CT MC Subtask Official Runs

IMTKU MC Subtasks Official Runs	Accuracy
RITE1-IMTKU-CT-MC-01	**0.522**
RITE1-IMTKU-CT-MC-02	0.507
RITE1-IMTKU-CT-MC-03	0.268

Table 3. Accuracy of IMTKU CT RITE4QA Subtask Official Runs

IMTKU CT RITE4QA Subtask Official Runs	Accuracy
RITE1-IMTKU-CT-RITE4QA-01	0.3246
RITE1-IMTKU-CT-RITE4QA-02	0.3392
RITE1-IMTKU-CT-RITE4QA-03	**0.4003**

Table 4. Accuracy of IMTKU CS RITE4QA Subtask Official Runs

IMTKU CS RITE4QA Subtask Official Runs	Accuracy
RITE1-IMTKU-CS-RITE4QA-01	0.3319
RITE1-IMTKU-CS-RITE4QA-02	0.4090
RITE1-IMTKU-CS-RITE4QA-03	**0.4716**

The confusion matrices of RITE1 IMKTU CT BC subtask official runs are shown in Table 5, 6, 7. CT MC subtask official runs are shown in Table 8, 9, 10; CT RITE4QA subtask official runs are shown in Table 11, 12, 13. CS RITE4QA subtask official runs are shown in Table 14, 15, 16, respectively.

Table 5. Confusion Matrix of RITE1-IMTKU-CT-BC-01 (Accuracy = 0.550)

	Y	N	
Y	**383**	338	721
N	67	**112**	179
	450	450	

Table 6. Confusion Matrix of RITE1-IMTKU-CT-BC-02 (Accuracy = 0.556)

	Y	N	
Y	**378**	328	706
N	72	**122**	194
	450	450	

Table 7. Confusion Matrix of RITE1-IMTKU-CT-BC-03 (Accuracy = 0.524)

	Y	N	
Y	**387**	365	752
N	63	**85**	148
	450	450	

Table 8. Confusion Matrix of RITE1-IMTKU-CT-MC-01 (Accuracy = 0.522)

	F	R	B	C	I	
F	**117**	4	17	32	43	213
R	1	**133**	17	23	23	197
B	34	23	**122**	85	33	297
C	16	1	19	**21**	4	61
I	12	19	5	19	**77**	132
	180	180	180	180	180	

Table 9. Confusion Matrix of RITE1-IMTKU-CT-MC-02 (Accuracy = 0.507)

	F	R	B	C	I	
F	**108**	4	14	20	34	180
R	1	**139**	16	29	38	223
B	33	28	**117**	88	37	303
C	19	1	21	**22**	1	64
I	19	8	12	21	**70**	130
	180	180	180	180	180	

Table 10. Confusion Matrix of RITE1-IMTKU-CT-MC-03 (Accuracy = 0.268)

	F	R	B	C	I	
F	**149**	156	107	100	143	655
R	16	**10**	8	7	13	54
B	3	2	**18**	11	6	40
C	4	0	39	**50**	4	97
I	8	12	8	12	**14**	54
	180	180	180	180	180	

Table 11. Confusion Matrix of RITE1-IMTKU-CT-RITE4QA-01 (Accuracy = 0.3246)

	Y	N	
Y	**116**	445	561
N	14	**107**	121
	130	552	

Table 12. Confusion Matrix of RITE1-IMTKU-CT-RITE4QA-02 (Accuracy = 0.3392)

	Y	N	
Y	**111**	430	541
N	19	**122**	141
	130	552	

Table 13. Confusion Matrix of RITE1-IMTKU-CT-RITE4QA-03 (Accuracy = 0.4003)

	Y	N	
Y	**107**	384	491
N	23	**168**	191
	130	552	

Table 14. Confusion Matrix of RITE1-IMTKU-CS-RITE4QA-01 (Accuracy = 0.3319)

	Y	N	
Y	**117**	441	558
N	13	**111**	124
	130	552	

Table 15. Confusion Matrix of RITE1-IMTKU-CS-RITE4QA-02 (Accuracy = 0.409)

	Y	N	
Y	**108**	379	487
N	22	**173**	195
	130	552	

Table 16. Confusion Matrix of RITE1-IMTKU-CS-RITE4QA-03 (Accuracy = 0.4716)

	Y	N	
Y	**82**	310	392
N	48	**242**	290
	130	552	

3.1.6 Discussions

It should be noted that we use MC dev dataset (dev_MC: 421 pairs) to convert BC dev dataset (dev_BC: 421 pairs), however, we regret that we made a mistake on the definition of BC subtask in our BC subtask formal run submission. We used to convert (F/R/B) as Y, and (C/I) as N in our formal BC training dataset. We found that we made a systematic mistake on transforming MC training data "RITE1_CT_dev_mc.txt" to BC training data "RITE1_CT_dev_bc.txt" after we received the MC/BC evaluation summary (evaluation-summary.txt) from organizers. That's why we had a relative good performance on MC subtask, however, we had a not good performance on BC subtask. The mistake we made is because of we used to use the definition use in MC subtask "A 5-way labeling subtask to detect (forward / reverse / bidirection) entailment or no entailment (contradiction / independence) in a text pair" (Incorrect conversion: "F/R/B"->Y, "C/I"->N).

We finally understand that the correct definition of BC subtask is "Given a text pair (t1,t2) identify whether t1 entails (infers) a hypothesis t2 or not." (Correct conversion: "F/B" -> Y, "R/C/I" ->N).

At the second phase of NTCIR-9 formal run, based on definition of RITE4QA substak, the correct label conversion from MC training data (RITE1_CT_dev_mc.txt) to BC training data (RITE1_CT_dev_bc.txt) was listed as follows:

F -> Y, B -> Y, R -> N, C -> N, I -> N

In order to test the consistence of difference dataset, we conduct experiments of cross validation focused on difference datasets. We used the gold standard dataset of CT BC subtask from organizers to train our machine learning model. Table 17 shows the experimental result of 10 fold cross validation (CV) of development and test datasets. The results show that the best performance of cross validation in BC subtask is 76.48%, which is development dataset with 421 pairs. However, we obtain 66.33% cross validation on the BC test dataset with 900 pairs by using the same features with same configuration in the machine learning model. In terms of consistence of dataset, we consider that the quality of development dataset is better than test dataset in BC subtask.

Table 17. Cross Validation of Development and Test datasets of CT BC Subtask

Datasets	10 Fold CV Accuracy
RITE1_CT_dev_bc_g.txt (gold standard) (BC Development Dataset: 421 pairs)	**76.48%**
RITE1_CT_test_bc_g.txt (BC Test Dataset: 900 pairs)	66.33%
RITE1_CT_dev_test_bc_g.txt (BC Dev+Test Dataset: 421+900 =1321 pairs)	67.67%

4. CONCLUSIONS

In this paper, we proposed a textual entailment system using a hybrid approach that integrate knowledge based and machine learning techniques for recognizing inference in text at NTCIR-9 RITE task. We submitted 3 official runs for both BC and MC subtask. In NTCIR-9 RITE task, IMTKU team achieved 0.522 in the CT-MC subtask and 0.556 in the CT-BC subtask.

The contributions of our study are as follows: (1) we proposed an RITE system by integrating knowledge-based and machine learning approach; (2) the machine learning approach used lexical and semantic features that measure the similarity of text pair to determine whether the text pair entails each other.

5. ACKNOWLEDGMENTS

This research was supported in part of TKU research grant. We would like to thank the support of IASL, IIS, Academia Sinica, Taiwan.

6. REFERENCES

[1] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology,* vol. 2, no. 3, 2011, pp. 27:1--27:27.

[2] W. Che, Z. Li, and T. Liu, “LTP: A Chinese Language Technology Platform,” in Proceedings of the Coling 2010: Demonstrations, Beijing, China., 2010, pp. 13-16.

[3] CKIP. "CKIP AutoTag," http://ckipsvr.iis.sinica.edu.tw/.

[4] I. Dagan, B. Dolan, B. Magnini, and D. Roth, “Recognizing textual entailment: Rational, evaluation and approaches (vol 15, pg 1, 2009),” *Natural Language Engineering,* vol. 16, 2010, pp. 105-+.

[5] I. Dagan, O. Glickman, and B. Magnini, “The PASCAL recognising textual entailment challenge,” *Machine Learning Challenges,* vol. 3944, 2006, pp. 177-190.

[6] C.-R. Huang, “Sinica BOW: integrating bilingual WordNet and SUMO ontology,” in International Conference on Natural Language Processing and Knowledge Engineering (NLPKE 2003), 2003, pp. 825-826.

[7] J.-J. Mei, Y.-M. Zhu, Y.-Q. Gao, and H.-X. Yin, *TongYiCi CiLin (Chinese Synonym Forest): Shanghai Press of Lexicon and Books*, 1983.

[8] H. Shima. "NTCIR9 RITE Main Page," http://artigas.lti.cs.cmu.edu/rite/Main_Page.

[9] H. Shima, H. Kanayama, C.-W. Lee, C.-J. Lin, T. Mitamura, Y. Miyao, S. Shi, and K. Takeda, “Overview of NTCIR-9 RITE: Recognizing Inference in TExt,” in Proceedings of NTCIR-8 Workshop Meeting, Tokyo, Japan, 2011.